idiap
RESEARCH INSTITUTE

PUBLICATIONS 2024

Scientific Report Annex

CONTENTS

AI FUNDAMENTALS

1. CONFERENCE PAPERS

OOD-Chameleon: Is Algorithm Selection for OOD Generalization Learnable?

Liangze Jiang and Damien Teney,

European Conference on Computer Vision (ECCV) Out Of Distribution Generalization in Computer Vision Workshop, 2024

Out-of-distribution (OOD) generalization is challenging because distribution shifts come in many forms. A multitude of learning algorithms exist and each can improve performance in specific OOD situations. We posit that much of the challenge of OOD generalization lies in choosing the right algorithm for the right dataset. However, such algorithm selection is often elusive under complex real-world shifts. In this work, we formalize the task of algorithm selection for OOD generalization and investigate whether it could be approached by learning. We propose a solution, dubbed OOD-Chameleon that treats the task as a supervised classification over candidate algorithms. We construct a dataset of datasets to learn from, which represents diverse types, magnitudes and combinations of shifts (covariate shift, label shift, spurious correlations). We train the model to predict the relative performance of algorithms given a dataset's characteristics. This enables a priori selection of the best learning strategy, i.e. without training various models as needed with traditional model selection. Our experiments show that the adaptive selection outperforms any individual algorithm and simple selection heuristics, on unseen datasets of controllable and realistic image data. Inspecting the model shows that it learns non-trivial data/algorithms interactions, and reveals the conditions for any one algorithm to surpass another. This opens new avenues for (1) enhancing OOD generalization with existing algorithms instead of designing new ones, and (2) gaining insights into the applicability of existing algorithms with respect to datasets' properties.

Are there identifiable structural parts in the sentence embedding whole?,

Vivi Nastase and Paola Merlo,

Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, 2024

Sentence embeddings from transformer models encode much linguistic information in a fixed-length vector. We investigate whether structural information – specifically, information about chunks and their structural and semantic properties – can be detected in these representations. We use a dataset consisting of sentences with known chunk structure, and two linguistic intelligence datasets, whose solution relies on detecting chunks and their grammatical number, and respectively, their semantic roles. Through an approach involving indirect supervision, and through analyses of the performance on the tasks and of the internal representations built during learning, we show that information about chunks and their properties can be obtained from sentence embeddings.

Formal Semantic Controls over Language Models,

Danilo Silva de Carvalho, Yingji Zhang, and André Freitas

Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), 2024

Text embeddings provide a concise representation of the semantics of sentences and larger spans of text, rather than individual words, capturing a wide range of linguistic features. They have found increasing application to a variety of NLP tasks, including machine translation and natural language inference. While most recent breakthroughs in task performance are being achieved by large scale distributional models, there is a growing disconnection between their knowledge representation and traditional semantics, which hinders efforts to capture such knowledge in human interpretable form or explain model inference behaviour. In this tutorial, we examine from basics to the cutting edge research on the analysis and control of text representations, aiming to shorten the gap between deep latent semantics and formal symbolics. This includes the considerations on knowledge formalisation, the linguistic information that can be extracted and measured from distributional models, and intervention techniques that enable explainable reasoning and controllable text generation, covering methods from pooling to LLM-based.

Empowering Cross-lingual Abilities of Instruction-tuned Large Language Models by Translation-following demonstrations,

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas,

62nd Annual Meeting of the Association for Computational Linguistics, 2024

The language ability of Large Language Models (LLMs) is often unbalanced towards English because of the imbalance in the distribution of the pre-training data. This disparity is demanded in further fine-tuning and affecting the cross-lingual abilities of LLMs. In this paper, we propose to empower Instructiontuned LLMs (It-LLMs) in languages other than English by building semantic alignment between them. Hence, we propose CrossAlpaca, an It-LLM with cross-lingual instruction-following and Translation-following demonstrations to improve semantic alignment between languages. We validate our approach on the multilingual Question Answering (QA) benchmarks XQUAD and MLQA and adapted versions of MMLU and BBH. Our models, tested over six different languages, outperform the It-LLMs tuned on monolingual data. The final results show that instruction tuning on non-English data is not enough and that semantic alignment can be further improved by Translation-following demonstrations.

Deep Clustering for Data Cleaning and Integration,

Hafiz Tayyab Rauf, Andre Freitas, and Norman W. Paton,

27th International Conference on Extending Database Technology, 2024

Deep Learning (DL) techniques now constitute the state-of-the-art for important problems in areas such as text and image processing, and there have been impactful results that deploy DL in several data management tasks. Deep Clustering (DC) has recently emerged as a sub-discipline of DL, in which data representations are learned in tandem with clustering, with a view to automatically identifying the features of the data that lead to improved clustering results. While DC has been used to good effect in several domains, particularly in image processing, the impact of DC on mainstream data management tasks remains unexplored. In this paper, we address this gap by investigating the impact of DC in data cleaning and integration tasks, specifically schema inference, entity resolution, and domain discovery, tasks that represent clustering from the perspective of tables, rows, and columns, respectively. In this setting, we compare and contrast several DC and non-DC clustering algorithms using standard benchmarks. The results show, among other things, that the most effective DC algorithms consistently outperform non-DC clustering algorithms for data integration tasks. However, we observed a significant correlation between the DC method and embedding approaches for rows, columns, and tables, highlighting that the suitable combination can enhance the efficiency of DC methods.

Learning Disentangled Semantic Spaces of Explanations via Invertible Neural Networks,

Yingji Zhang, Danilo S. Carvalho, Ian Pratt-Hartmann, and André Freitas,

62nd Annual Meeting of the Association for Computational Linguistics, 2024

Disentangled latent spaces usually have better semantic separability and geometrical properties, which leads to better interpretability and more controllable data generation. While this has been well investigated in Computer Vision, in tasks such as image disentanglement, in the NLP domain sentence disentanglement is still comparatively under-investigated. Most previous work have concentrated on disentangling task-specific generative factors, such as sentiment, within the context of style transfer. In this work, we focus on a more general form of sentence disentanglement, targeting the localised modification and control of more general sentence semantic features. To achieve this, we contribute to a novel notion of sentence semantic disentanglement and introduce a flow-based invertible neural network (INN) mechanism integrated with a transformer-based language Autoencoder (AE) in order to deliver latent spaces with better separability properties. Experimental results demonstrate that the model can conform the distributed latent space into a better semantically disentangled sentence space, leading to improved language interpretability and controlled generation when compared to the recent state-of-the-art language VAE models.

Graph Neural Flows for Unveiling Systemic Interactions Among Irregularly Sampled Time Series,

Giangiacomo Mercatali, Andre Freitas, and Jie Chen,

Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024

Interacting systems are prevalent in nature. It is challenging to accurately predict the dynamics of the system if its constituent components are analyzed independently. We develop a graph-based model that unveils the systemic interactions of time series observed at irregular time points, by using a directed acyclic graph to model the conditional dependencies (a form of causal notation) of the system components and learning this graph in tandem with a continuous-time model that parameterizes the solution curves of ordinary differential equations (ODEs). Our technique, a graph neural flow, leads to substantial enhancements over non-graph-based methods, as well as graph-based methods without the modeling of conditional dependencies. We validate our approach on several tasks, including time series classification and forecasting, to demonstrate its efficacy.

Diffusion Twigs with Loop Guidance for Conditional Graph Generation,

Giangiacomo Mercatali, Yogesh Verma, Andre Freitas, and Vikas Garg,

Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024

We introduce a novel score-based diffusion framework named Twigs that incorporates multiple co-evolving flows for enriching conditional generation tasks. Specifically, a central or trunk diffusion process is associated with a primary variable (e.g., graph structure), and additional offshoot or stem processes are dedicated to dependent variables (e.g., graph properties or labels). A new strategy, which we call loop guidance, effectively orchestrates the flow of information between the trunk and the stem processes during sampling. This approach allows us to uncover intricate interactions and dependencies, and unlock new generative capabilities. We provide extensive experiments to demonstrate strong performance gains of the proposed method over contemporary baselines in the context of conditional graph generation, underscoring the potential of Twigs in challenging generative tasks such as inverse molecular design and molecular optimization.

An LLM-based Knowledge Synthesis and Scientific Reasoning Framework for Biomedical Discovery,

Oskar Wysocki, Magdalena Wysocka, Danilo Carvalho, Alex Bogatu, Danilo Miranda, Maxime Delmas, Harriet Unsworth, and Andre Freitas,

62nd Annual Meeting of the Association for Computational Linguistics, 2024

We present BioLunar, developed using the Lunar framework, as a tool for supporting biological analyses, with a particular emphasis on molecular-level evidence enrichment for biomarker discovery in oncology. The platform integrates Large Language Models (LLMs) to facilitate complex scientific reasoning across distributed evidence spaces, enhancing the capability for harmonizing and reasoning over heterogeneous data sources. Demonstrating its utility in cancer research, BioLunar leverages modular design, reusable data access and data analysis components, and a low-code user interface, enabling researchers of all programming levels to construct LLM-enabled scientific workflows. By facilitating automatic scientific discovery and inference from heterogeneous evidence, BioLunar exemplifies the potential of the integration between LLMs, specialised databases and biomedical tools to support expert-level knowledge synthesis and discovery.

Consistent Autoformalization for Constructing Mathematical Libraries,

Lan Zhang, Xin Quan, and Andre Freitas,

Conference on Empirical Methods in Natural Language Processing, 2024

Autoformalization is the task of automatically translating mathematical content written in natural language to a formal language expression. The growing language interpretation capabilities of Large Language Models (LLMs), including in formal languages, are lowering the barriers for autoformalization. However, LLMs alone are not capable of consistently and reliably delivering autoformalization, in particular as the complexity and specialization of the target domain grows. As the field evolves into the direction of systematically applying autoformalization towards large mathematical libraries, the need to improve syntactic, terminological and semantic control increases. This paper proposes the coordinated use of three mechanisms, most-similar retrieval augmented generation (MS-RAG), denoising steps, and auto-correction with syntax error feedback (Auto-SEF) to improve autoformalization quality. The empirical analysis, across different models, demonstrates that these mechanisms can deliver autoformalizaton results which are syntactically, terminologically and semantically more consistent. These mechanisms can be applied across different LLMs and have shown to deliver improve results across different model types.

Reasoning with Natural Language Explanations,

Valentino Marco and Andre Freitas,

Conference on Empirical Methods in Natural Language Processing, 2024

Explanation constitutes an archetypal feature of human rationality, underpinning learning and generalisation, and representing one of the media supporting scientific discovery and communication. Due to the importance of explanations in human reasoning, an increasing amount of research in Natural Language Inference (NLI) has started reconsidering the role that explanations play in learning and inference, attempting to build explanation-based NLI models that can effectively encode and use natural language explanations on downstream tasks. Research in explanation-based NLI, however, presents specific challenges and opportunities, as explanatory reasoning reflects aspects of both material and formal inference, making it a particularly rich setting to model and deliver complex reasoning. In this tutorial, we provide a comprehensive introduction to the field of explanation-based NLI, grounding this discussion on the epistemological-linguistic foundations of explanations, systematically describing the main architectural trends and evaluation methodologies that can be used to build systems capable of explanatory reasoning.

Graph-Induced Syntactic-Semantic Spaces in Transformer-Based Variational AutoEncoders,

Yingji Zhang, Marco Valentino, Danilo S. Carvalho, Ian Pratt-Hartmann, and Andre Freitas,

Findings of the Association for Computational Linguistics: NAACL, 2024

The injection of syntactic information in Variational AutoEncoders (VAEs) has been shown to result in an overall improvement of performances and generalisation. An effective strategy to achieve such a goal is to separate the encoding of distributional semantic features and syntactic structures into heterogeneous latent spaces via multi-task learning or dual encoder architectures. However, existing works employing such techniques are limited to LSTM-based VAEs. In this paper, we investigate latent space separation methods for structural syntactic injection in Transformer-based VAE architectures (i.e., Optimus). Specifically, we explore how syntactic structures can be leveraged in the encoding stage through the integration of graph-based and sequential models, and how multiple, specialised latent representations can be injected into the decoder's attention mechanism via low-rank operators. Our empirical evaluation, carried out on natural language sentences and mathematical expressions, reveals that the proposed end-to-end VAE architecture can result in a better overall organisation of the latent space, alleviating the information loss occurring in standard VAE setups, resulting in enhanced performances on language modelling and downstream generation tasks.

Multi-Operational Mathematical Derivations in Latent Space,

Marco Valentino, Jordan Meadows, Lan Zhang, and Andre Freitas,

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024

This paper investigates the possibility of approximating multiple mathematical operations in latent space for expression derivation. To this end, we introduce different multi-operational representation paradigms, modelling mathematical operations as explicit geometric transformations. By leveraging a symbolic engine, we construct a large-scale dataset comprising 1.7M derivation steps stemming from 61K premises and 6 operators, analysing the properties of each paradigm when instantiated with state-of-the-art neural encoders. Specifically, we investigate how different encoding mechanisms can approximate expression manipulation in latent space, exploring the trade-off between learning different operators and specialising within single operations, as well as the ability to support multi-step derivations and out-of-distribution generalisation. Our empirical analysis reveals that the multi-operational paradigm is crucial for disentangling different operators, while discriminating the conclusions for a single operation is achievable in the original expression encoder. Moreover, we show that architectural choices can heavily affect the training dynamics, structural organisation, and generalisation of the latent space, resulting in significant variations across paradigms and classes of encoders.

Are there identifiable structural parts in the sentence embedding whole?,

Vivi Nastase and Paola Merlo,

7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, 2024

Sentence embeddings from transformer models encode much linguistic information in a fixed-length vector. We investigate whether structural information – specifically, information about chunks and their structural and semantic properties – can be detected in these representations. We use a dataset consisting of sentences with known chunk structure, and two linguistic intelligence datasets, whose solution relies on detecting chunks and their grammatical number, and respectively, their semantic roles. Through an approach involving indirect supervision, and through analyses of the performance on the tasks and of the internal representations built during learning, we show that information about chunks and their properties can be obtained from sentence embeddings.

HUMAN-AI TEAMING

1. JOURNAL PAPERS

A Probabilistic Approach to Multi-Modal Adaptive Virtual Fixtures,

Maximilian Mühlbauer, Thomas Hulin, Bernhard Weber, Sylvain Calinon, Freek Stulp, Alin Albu-Schäffer, and Joao Silverio,

IEEE Robotics and Automation Letters (RA-L), 2024

Virtual Fixtures (VFs) provide haptic feedback for teleoperation, typically requiring distinct input modalities for different phases of a task. This often results in vision- and position-based fixtures. Vision-based fixtures, particularly, re quire the handling of visual uncertainty, as well as target appearance/disappearance for increased flexibility. This creates the need for principled ways to add/remove fixtures, in addition to uncertainty-aware assistance regulation. Moreover, the arbitration of different modalities plays a crucial role in providing an optimal feedback to the user throughout the task. In this paper, we propose a Mixture of Experts (MoE) model that synthesizes visual serving fixtures, elegantly handling full pose detection uncertainties and teleoperation goals in a unified framework. An arbitration function combining multiple vision-based fixtures arises naturally from the MoE formulation, leveraging uncertainties to modulate fixture stiffness and thus the degree of assistance. The resulting visual serving fixtures are then fused with position-based fixtures using a Product of Experts (PoE) approach, achieving guidance throughout the complete workspace. Our results indicate that this approach not only permits human operators to accurately insert printed circuit boards (PCBs) but also offers added flexibility and retains the performance level of a baseline with carefully hand tuned VFs, without requiring the manual creation of VFs for individual connectors. An exemplary video showcasing our method is available at: https://youtu.be/6BDB3g0QyFg.

An Optimal Control Formulation of Tool Affordance Applied to Impact Tasks,

Boyang Ti, Yongsheng Gao, Jie Zhao, and Sylvain Calinon,

IEEE Transactions on Robotics (T-RO), 2024

Humans use tools to complete impact-aware tasks such as hammering a nail or playing tennis. The postures adopted to use these tools can significantly influence the performance of these tasks, where the force or velocity of the hand holding a tool plays a crucial role. The underlying motion planning challenge consists of grabbing the tool in preparation for the use of this tool with an optimal body posture. Directional manipulability de scribes the dexterity of force and velocity in a joint configuration along a specific direction. In order to take directional manipulability and tool affordances into account, we apply an optimal control method combining iterative linear quadratic regulator (iLQR) with the alternating direction method of multipliers (ADMM). Our approach considers the notion of tool affordances to solve motion planning problems, by introducing a cost based on directional velocity manipulability. The proposed approach is applied to impact tasks in simulation and on a real 7-axis robot, specifically in a nail-hammering task with the assistance of a pilot hole. Our comparison study demonstrates the importance of maximizing directional manipulability in impact-aware tasks.

gafro: Geometric Algebra for Robotics,

Tobias Löw, Philip Abbet, and Sylvain Calinon,

IEEE Robotics and Automation Magazine, 2024

Geometry is a fundamental part of robotics and there have been various frameworks of representation over the years. Recently, geometric algebra has gained attention for its property of unifying many of those previous ideas into one algebra. While there are already efficient open-source implementations of geometric algebra available, none of them is targeted at robotics applications. We want to address this shortcoming with our library gafro. This article presents an overview of the implementation details as well as a tutorial of gafro, an efficient C++ library targeting robotics applications using geometric algebra. The library focuses on using conformal geometric algebra. Hence, various geometric primitives are available for computation as well as rigid body transformations. The modeling of robotic systems is also an important aspect of the library. It implements various algorithms for calculating the kinematics and dynamics of such systems as well as objectives for optimization problems. The software stack is completed by Python bindings in pygafro and a ROS interface in gafro_ros.

Logic Learning from Demonstrations for Multi-step Manipulation Tasks in Dynamic Environments,

Yan Zhang, Teng Xue, Amirreza Razmjoo, and Sylvain Calinon,

IEEE Robotics and Automation Letters (RA-L), 2024

Learning from Demonstration (LfD) stands as an efficient framework for imparting human-like skills to robots. Nevertheless, designing an LfD framework capable of seamlessly imitating, generalizing, and reacting to disturbances for long horizon manipulation tasks in dynamic environments remains a challenge. To tackle this challenge, we present Logic-LfD, which combines Task and Motion Planning (TAMP) with an optimal control formulation of Dynamic Movement Primitives (DMP), allowing us to incorporate motion-level via-point specifications and to handle task-level variations or disturbances in dynamic environments. We conduct a comparative analysis of our proposed approach against several baselines, evaluating its generalization ability and reactivity across three long-horizon manipulation tasks. Our experiment demonstrates the fast generalization and reactivity of Logic-LfD for handling task-level variants and disturbances in long-horizon manipulation tasks.

Online Learning of Continuous Signed Distance Fields Using Piecewise Polynomials,

Ante Marić, Yiming Li and Sylvain Calinon,

IEEE Robotics and Automation Letters (RA-L), 2024

Reasoning about distance is indispensable for establishing or avoiding contact in manipulation tasks. To this end, we present an online approach for learning implicit representations of signed distance using piecewise polynomial basis functions. Starting from an arbitrary prior shape, our method incrementally constructs a continuous and smooth distance representation from incoming surface points, with analytical access to gradient information. The underlying model does not store training data for prediction, and its performance can be balanced through interpretable hyper parameters such as polynomial degree and number of segments. We assess the accuracy of the incrementally learned model on a set of household objects and compare it to neural network and Gaussian process counterparts. The utility of intermediate results and analytical gradients is further demonstrated in a physical experiment.

Online Multi-Contact Receding Horizon Planning via Value Function Approximation,

Jiayi Wang, Sanghyun Kim, Teguh Santoso Lembono, Wenqian Du, Jaehyun Shim, Saeid Samadi, Ke Wang, Vladimir Ivan, Sylvain Calinon, Sethu Vijayakumar, and Steve Tonneau,

IEEE Transactions on Robotics (T-RO), 2024

Planning multi-contact motions in a receding horizon fashion requires a value function to guide the planning with respect to the future, e.g., building momentum to traverse large obstacles. Traditionally, the value function is approximated by computing trajectories in a prediction horizon (never executed) that foresees the future beyond the execution horizon. However, given the non-convex dynamics of multi-contact motions, this approach is computationally expensive. To enable online Receding Horizon Planning (RHP) of multi-contact motions, we find efficient approximations of the value function. Specifically, we propose a trajectory-based and a learning-based approach. In the former, namely RHP with Multiple Levels of Model Fidelity, we approximate the value function by computing the prediction horizon with a convex relaxed model. In the latter, namely Locally-Guided RHP, we learn an oracle to predict local objectives for locomotion tasks, and we use these local objectives to construct local value functions for guiding a short-horizon RHP. We evaluate both approaches in simulation by planning centroidal trajectories of a humanoid robot walking on moderate slopes, and on large slopes where the robot cannot maintain static balance. Our results show that locally-guided RHP achieves the best computation efficiency (95%-98.6% cycles converge online). This computation advantage enables us to demonstrate online receding horizon planning of our real-world humanoid robot Talos walking in dynamic environments that change on-the-fly.

Tensor Train for Global Optimization Problems in Robotics,

Suhan Shetty, Teguh Lembono, Tobias Löw, and Sylvain Calinon,

International Journal of Robotics Research (IJRR), 2024

We propose an approach based on low-rank tensor approximation techniques to initialize the existing optimization solvers close to global optima. The approach uses only the definition of the cost function and does not need access to any database of good solutions. We first transform the cost function, which is a function of task parameters and optimization variables, into an unnormalized probability density function. Unlike existing approaches that set the task parameters as constant, we consider them as another set of random variables and approximate the joint probability distribution of the task parameters and the optimization variables using a surrogate probability model. For a given task, we then generate samples from the conditional distribution with respect to the given task parameter and use them as initialization for the optimization solver. As conditioning and sampling from an arbitrary density function are challenging, we use Tensor Train decomposition to obtain a surrogate probability model from which we can efficiently obtain the conditional model and the samples. The proposed method can produce multiple solutions coming from different modes (when they exist) for a given task. We first evaluate the approach by applying it to various challenging benchmark functions for numerical optimization that are difficult to solve using gradient-based optimization solvers with a naive initialization, showing that the proposed method can produce samples close to the global optima and coming from multiple modes. We then demonstrate the generality of the framework and its relevance to robotics by applying the proposed method to inverse kinematics and motion planning problems with a 7-DoF manipulator, as commonly encountered in robot manipulation.

Speech prosody enhances the neural processing of syntax,

Giulio Degano, Peter W. Donhauser, Laura Gwilliams, Paola Merlo, and Narly Golestani,

Communications Biology, 2024

Human language relies on the correct processing of syntactic information, as it is essential for successful communication between speakers. As an abstract level of language, syntax has often been studied separately from the physical form of the speech signal, thus often masking the interactions that can promote better syntactic processing in the human brain. However, behavioral and neural evidence from adults suggests the idea that prosody and syntax interact, and studies in infants support the notion that prosody assists language learning. Here we analyze a MEG dataset to investigate how acoustic cues, specifically prosody, interact with syntactic representations in the brains of native English speakers. More specifically, to examine whether prosody enhances the cortical encoding of syntactic representations, we decode syntactic phrase boundaries directly from brain activity, and evaluate possible modulations of this decoding by the prosodic boundaries. Our findings demonstrate that the presence of prosodic boundaries improves the neural representation of phrase boundaries, indicating the facilitative role of prosodic cues in processing abstract linguistic features. This work has implications for interactive models of how the brain processes different linguistic features. Future research is needed to establish the neural underpinnings of prosody-syntax interactions in languages with different typological characteristics.

Test-time adaptation for 6D pose tracking,

Long Tian, Changjae Oh, and Andrea Cavallaro,

Pattern Recognition, 2024

We propose a test-time adaptation for 6D object pose tracking that learns to adapt a pre-trained model to track the 6D pose of novel objects. We consider the problem of 6D object pose tracking as a 3D keypoint detection and matching task and present a model that extracts 3D keypoints. Given an RGB-D image and the mask of a target object for each frame, the proposed model consists of the self- and cross-attention modules to produce the features that aggregate the information within and across frames, respectively. By using the keypoints detected from the features for each frame, we estimate the pose changes between two frames, which enables 6D pose tracking when the 6D pose of a target object in the initial frame is given. Our model is first trained in a source domain, a category-level tracking dataset where the ground truth 6D pose of the object is available. To deploy this pre-trained model to track novel objects, we present a test-time adaptation strategy that trains the model to adapt to the target novel object by self-supervised learning. Given an RGB-D video sequence of the novel object, the proposed self-supervised losses encourage the model to estimate the 6D pose changes that can keep the photometric and geometric consistency of the object. We validate our method on the NOCS-REAL275 dataset and our collected dataset, and the results show the advantages of tracking novel objects. The collected dataset and visualisation of tracking results are available: https://qm-ipalab.github.io/TA-6DT/.

Tactile Ergodic Coverage on Curved Surfaces,

Cem Bilaloglu, Tobias Löw, and Sylvain Calinon,

IEEE Transactions on Robotics, 2024

In this article, we present a feedback control method for tactile coverage tasks, such as cleaning or surface inspection. These tasks are challenging to plan due to complex continuous physical interactions. In these tasks, the coverage target and progress can be easily measured using a camera and encoded in a point cloud. We propose an ergodic coverage method that operates directly on point clouds, guiding the robot to spend more time on regions requiring more coverage. For robot control and contact behavior, we use geometric algebra to formulate a task-space impedance controller that tracks a line while simultaneously exerting a desired force along that line. We evaluate the performance of our method in kinematic simulations and demonstrate its applicability in real-world experiments on kitchenware. Our source codes, experimental data, and videos are available as open access at https://sites.google.com/view/tactile-ergodic-control/.

Natural Language Understanding for Navigation of Service Robots in Low-Resource Domains and Languages: Scenarios in Spanish and Nahuatl

Amadeo Hernández, Rosa María Ortega-Mendoza, Esaú Villatoro-Tello, César Joel Camacho-Bello, and Obed Pérez-Cortés

Mathematics, 2024

Human–robot interaction is becoming increasingly common to perform useful tasks in everyday life. From the human–machine communication perspective, achieving effective interaction in natural language is one challenge. To address it, natural language processing strategies have recently been used, commonly following a supervised machine learning framework. In this context, most approaches rely on the use of linguistic resources (e.g., taggers or embeddings), including training corpora. Unfortunately, such resources are scarce for some languages in specific domains, increasing the complexity of solution approaches. Motivated by these challenges, this paper explores deep learning methods for understanding natural language commands emitted to service robots that guide their movements in low-resource scenarios, defined by the use of Spanish and Nahuatl languages, for which linguistic resources are scarcely unavailable for this specific task. Particularly, we applied natural language understanding (NLU) techniques using deep neural networks and transformers-based models. As part of the research methodology, we introduced a labeled dataset of movement commands in the mentioned languages. The results show that models based on transformers work well to recognize commands (intent classification task) and their parameters (e.g., quantities and movement units) in Spanish, achieving a performance of 98.70% (accuracy) and 96.96% (F1) for the intent classification and slot-filling tasks, respectively). In Nahuatl, the best performance obtained was 93.5% (accuracy) and 88.57% (F1) in these tasks, respectively. In general, this study shows that robot movements can be guided in natural language through machine learning models using neural models and cross-lingual transfer strategies, even in low-resource scenarios.

On the Nature of Explanation: An Epistemological-Linguistic Perspective for Explanation-Based Natural Language Inference,

Marco Valentino and Andre Freitas,

Philosophy & Technology, 2024

One of the fundamental research goals for explanation-based Natural Language Inference (NLI) is to build models that can reason in complex domains through the generation of natural language explanations. However, the methodologies to design and evaluate explanation-based inference models are still poorly informed by theoretical accounts on the nature of explanation. As an attempt to provide an epistemologically grounded characterisation for NLI, this paper focuses on the scientific domain, aiming to bridge the gap between theory and practice on the notion of a scientific explanation. Specifically, the paper combines a detailed survey of the modern accounts of scientific explanation in Philosophy of Science with a systematic analysis of corpora of natural language explanations, clarifying the nature and function of explanatory arguments from both a top-down (categorical) and a bottom-up (corpus-based) perspective. Through a mixture of quantitative and qualitative methodologies, the presented study allows deriving the following main conclusions: (1) Explanations cannot be entirely characterised in terms of inductive or deductive arguments as their main function is to perform unification; (2) An explanation typically cites causes and mechanisms that are responsible for the occurrence of the event to be explained; (3) While natural language explanations possess an intrinsic causal-mechanistic nature, they are not limited to causes and mechanisms, also accounting for pragmatic elements such as definitions, properties and taxonomic relations; (4) Patterns of unification naturally emerge in corpora of explanations even if not intentionally modelled; (5) Unification is realised through a process of abstraction, whose function is to provide the inference mechanism for subsuming the event to be explained under recurring patterns and high-level regularities. The paper contributes to addressing a fundamental gap in classical theoretical accounts on the nature of scientific explanations and their materialisation as linguistic artefacts. This characterisation can support a more principled design and evaluation of explanation-based AI systems which can better interpret, process, and generate natural language explanations.

Analysing the potential of open hotel review databases for IEQ assessment: A text mining approach,

Giulia Lamberti, Roberto Boghetti, Fabio Fantozzi, Francesco Leccese, and Giacomo Salvadori,

Building Research & Information, 2024

Indoor Environmental Quality (IEQ) significantly affects occupants' well-being and comfort. Assessing IEQ typically involves post-occupancy evaluation (POE), a method that can be time-consuming and particularly challenging in hotel settings, where guests may be disrupted by frequent requests for feedback. Hence, this paper investigates the capability of text mining to extract valuable information for IEQ assessment, such as identifying the main causes of IEQ dissatisfaction, detecting combined occurrences of IEQ aspects, and exploring the relationship between IEQ dissatisfaction and hotel attractiveness. To this aim, the study analysed 1494 five-star hotels in Europe, comprising 515,738 reviews. Among them, 13.1% contained references to keywords related to IEQ aspects. The major cause of dissatisfaction in hotels is acoustic (42.7% of the reviews), followed by thermal (35.7%), visual (11.1%) comfort, and IAQ (10.5%). Additionally, 9580 reviews demonstrated the co-occurrence of multiple IEQ aspects, highlighting the interplay between different aspects. Furthermore, the reviewer score, reflecting the hotel's attractiveness, showed an inverse relationship with the percentage of dissatisfied guests regarding IEQ, highlighting the impact of the indoor environment on the hotel rating. Overall, text mining is effective in supporting IEQ assessment and the study underscores the effect of addressing IEQ aspects on a facility's overall appeal.

A Minimum-Jerk Approach to Handle Singularities in Virtual Fixtures,

Giovanni Braglia , Sylvain Calinon, and Luigi Biagiotti

IEEE Robotics and Automation Letters, 2024

Implementing virtual fixtures in guiding tasks constrains the movement of the robot's end effector to specific curves within its workspace. However, incorporating guiding frameworks may encounter discontinuities when optimizing the reference target position to the nearest point relative to the current robot position. This article aims to give a geometric interpretation of such discontinuities, with specific reference to the commonly adopted Gauss-Newton algorithm. The effect of such discontinuities, defined as Euclidean Distance Singularities, is experimentally proved. We then propose a solution that is based on a linear quadratic tracking problem with minimum jerk command, then compare and validate the performances of the proposed framework in two different human-robot interaction scenarios.

Speech prosody enhances the neural processing of syntax,

Giulio Degano, Peter W. Donhauser, Laura Gwilliams, Paola Merlo, and Narly Golestani,

Communications Biology, 2024

Human language relies on the correct processing of syntactic information, as it is essential for successful communication between speakers. As an abstract level of language, syntax has often been studied separately from the physical form of the speech signal, thus often masking the interactions that can promote better syntactic processing in the human brain. However, behavioral and neural evidence from adults suggests the idea that prosody and syntax interact, and studies in infants support the notion that prosody assists language learning. Here we analyze a MEG dataset to investigate how acoustic cues, specifically prosody, interact with syntactic representations in the brains of native English speakers. More specifically, to examine whether prosody enhances the cortical encoding of syntactic representations, we decode syntactic phrase boundaries directly from brain activity, and evaluate possible modulations of this decoding by the prosodic boundaries. Our findings demonstrate that the presence of prosodic boundaries improves the neural representation of phrase boundaries, indicating the facilitative role of prosodic cues in processing abstract linguistic features. This work has implications for interactive models of how the brain processes different linguistic features. Future research is needed to establish the neural underpinnings of prosody-syntax interactions in languages with different typological characteristics.

Surprisal From Language Models Can Predict ERPs in Processing Predicate-Argument Structures Only if Enriched by an Agent Preference Principle,

Eva Huber, Sebastian Sauppe, Arrate Isasi-Isasmendi, Ina Bornkessel-Schlesewsky, Paola Merlo, and Balthasar Bickel,

Neurobiology of Language, 2024

Language models based on artificial neural networks increasingly capture key aspects of how humans process sentences. Most notably, model-based surprisals predict event-related potentials such as N400 amplitudes during parsing. Assuming that these models represent realistic estimates of human linguistic experience, their success in modeling language processing raises the possibility that the human processing system relies on no other principles than the general architecture of language models and on sufficient linguistic input. Here, we test this hypothesis on N400 effects observed during the processing of verb-final sentences in German, Basque, and Hindi. By stacking Bayesian generalised additive models, we show that, in each language, N400 amplitudes and topographies in the region of the verb are best predicted when model-based surprisals are complemented by an Agent Preference principle that transiently interprets initial role-ambiguous noun phrases as agents, leading to reanalysis when this interpretation fails. Our findings demonstrate the need for this principle independently of usage frequencies and structural differences between languages. The principle has an unequal force, however. Compared to surprisal, its effect is weakest in German, stronger in Hindi, and still stronger in Basque. This gradient is correlated with the extent to which grammars allow unmarked NPs to be patients, a structural feature that boosts reanalysis effects. We conclude that language models gain more neurobiological plausibility by incorporating an Agent Preference. Conversely, theories of human processing profit from incorporating surprisal estimates in addition to principles like the Agent Preference, which arguably have distinct evolutionary roots.

Beyond bigrams: call sequencing in the common marmoset (Callithrix jacchus) vocal system,

Alexandra B. Bosshard, Judith M. Burkart, Paola Merlo, Chundra Cathcart, Simon W. Townsend, and Balthasar Bickel,

Royal Society Open Science, 2024

Over the last two decades, an emerging body of research has demonstrated that non-human animals exhibit the ability to combine context-specific calls into larger sequences. These structures have frequently been compared with language's syntax, whereby linguistic units are combined to form larger structures, and leveraged to argue that syntax might not be unique to language. Currently, however, the overwhelming majority of examples of call combinations are limited to simple sequences comprising just two calls which differ dramatically from the open-ended hierarchical structuring of the syntax found in language. We revisit this issue by taking a whole-repertoire approach to investigate combinatoriality in common marmosets (Callithrix jacchus). We use Markov chain models to quantify the vocal sequences produced by marmosets providing evidence for structures beyond the bigram, including three-call and even combinations of up to eight or nine calls. Our analyses of these longer vocal sequences are suggestive of potential further internal organization, including some amount of recombination, nestedness and non-adjacent dependencies. We argue that data-driven, whole-repertoire analyses are fundamental to uncovering the combinatorial complexity of non-human animals and will further facilitate meaningful comparisons with language's combinatoriality.

2. Conference Papers

A Differentiable Integer Linear Programming Solver for Explanation-Based Natural Language Inference,

Mokanarangan Thayaparan, Marco Valentino, and Andre Freitas,

Joint International Conference on Computational Linguistics, Language Resources and Evaluation, 2024

Integer Linear Programming (ILP) has been proposed as a formalism for encoding precise structural and semantic constraints for Natural Language Inference (NLI). However, traditional ILP frameworks are non-differentiable, posing critical challenges for the integration of continuous language representations based on deep learning. In this paper, we introduce a novel approach, named Diff-Comb Explainer, a neuro-symbolic architecture for explanation-based NLI based on Differentiable BlackBox Combinatorial Solvers (DBCS). Differently from existing neuro-symbolic solvers, Diff-Comb Explainer does not necessitate a continuous relaxation of the semantic constraints, enabling a direct, more precise, and efficient incorporation of neural representations into the ILP formulation. Our experiments demonstrate that Diff-Comb Explainer achieves superior performance when compared to conventional ILP solvers, neuro-symbolic black-box solvers, and Transformer-based encoders. Moreover, a deeper analysis reveals that Diff-Comb Explainer can significantly improve the precision, consistency, and faithfulness of the constructed explanations, opening new opportunities for research on neuro-symbolic architectures for explainable and transparent NLI in complex domains.

A System for Human-Robot Teaming through End-User Programming and Shared Autonomy,

Michael Hagenow, Emmanuel Senft, Robert Radwin, Michael Gleicher, Michael Zinn, and Bilge Mutlu,

ACM/IEEE International Conference on Human-Robot Interaction, 2024

Many industrial tasks—such as sanding, installing fasteners, and wire harnessing—are difficult to automate due to task complexity and variability. We instead investigate deploying robots in an assistive role for these tasks, where the robot assumes the physical task burden and the skilled worker provides both the high-level task planning and low-level feedback necessary to effectively complete the task. In this article, we describe the development of a system for flexible human-robot teaming that combines state-of-the-art methods in end-user programming and shared autonomy and its implementation in sanding applications. We demonstrate the use of the system in two types of sanding tasks, situated in aircraft manufacturing, that highlight two potential workflows within the human-robot teaming setup. We conclude by discussing challenges and opportunities in human-robot teaming identified during the development, application, and demonstration of our system.

Sharingan: A Transformer Architecture for Multi-Person Gaze Following,

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez,

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

Gaze is a powerful form of non-verbal communication that humans develop from an early age. As such, modeling this behavior is an important task that can benefit a broad set of application domains ranging from robotics to sociology. In particular, the gaze following task in computer vision is defined as the prediction of the 2D pixel coordinates where a person in the image is looking. Previous attempts in this area have primarily centered on CNN-based architectures, but they have been constrained by the need to process one person at a time, which proves to be highly inefficient. In this paper, we introduce a novel and effective multi-person transformer-based architecture for gaze prediction. While there exist prior works using transformers for multi-person gaze prediction [38, 39], they use a fixed set of learnable embeddings to decode both the person and its gaze target, which requires a matching step afterward to link the predictions with the annotations. Thus, it is difficult to quantitatively evaluate these methods reliably with the available benchmarks, or integrate them into a larger human behavior understanding system. Instead, we are the first to propose a multi-person transformer-based architecture that maintains the original task formulation and ensures control over the people fed as input. Our main contribution lies in encoding the person-specific information into a single controlled token to be processed alongside image tokens and using its output for prediction based on a novel multiscale decoding mechanism. Our new architecture achieves state-of-the-art results on the GazeFollow, VideoAttentionTarget, and ChildPlay datasets and outperforms comparable multi-person architectures with a notable margin. Our code, checkpoints, and data extractions will be made publicly available soon.

A Unified Model for Gaze Following and Social Gaze Prediction,

Anshul Gupta, Samy Tafasca, Naravich Chutisilp, and Jean-Marc Odobez,

18th IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2024

Human gaze plays a crucial role in communication and social interaction. Many recent studies have focused on predicting the 2D pixel location of a person's gaze target in an image. However, this approach has limitations when it comes to studying gaze for downstream applications that require analysis of higher-level social gaze behaviors. Previous works have post-processed the predicted 2D gaze target for social gaze prediction, however, we show that this approach is insufficient. Our proposed method jointly predicts the gaze target and social gaze behavior, explicitly incorporating people interaction for state of the art results on three social gaze tasks- looking at heads, mutual gaze and shared attention. Additionally, we introduce evaluation protocols for these tasks, presenting a promising avenue for future research in gaze behavior analysis.

Configuration Space Distance Fields for Manipulation Planning,

Yiming Li, Xuemin Chi, Amirreza Razmjoo, and Sylvain Calinon,

Robotics: Science and Systems (RSS), 2024

The signed distance field (SDF) is a popular implicit shape representation in robotics, providing geometric information about objects and obstacles in a form that can easily be combined with control, optimization and learning techniques. Most often, SDFs are used to represent distances in task space, which corresponds to the familiar notion of distances that we perceive in our 3D world. However, SDFs can mathematically be used in other spaces, including robot configuration spaces. For a robot manipulator, this configuration space typically corresponds to the joint angles for each articulation of the robot. While it is customary in robot planning to express which portions of the configuration space are free from collision with obstacles, it is less common to think of this information as a distance field in the configuration space. In this paper, we demonstrate the potential of considering SDFs in the robot configuration space for optimization, which we call the configuration space distance field (or CDF for short). Similarly to the use of SDF in task space, CDF provides an efficient joint angle distance query and direct access to the derivatives (joint angle velocity). Most approaches split the overall computation with one part in task space followed by one part in configuration space (evaluating distances in task space and then computing actions with inverse kinematics). Instead, CDF allows the implicit structure to be leveraged by control, optimization, and learning problems in a unified manner. In particular, we propose an efficient algorithm to compute and fuse CDFs that can be generalized to arbitrary scenes. A corresponding neural CDF representation using multilayer perceptrons (MLPs) is also presented to obtain a compact and continuous representation while improving computation efficiency. We demonstrate the effectiveness of CDF with planar obstacle avoidance examples and with a 7-axis Franka robot in inverse kinematics and manipulation planning kinematics and manipulation planning tasks. Project page: https://sites.google.com/view/cdfmp/home.

Contextual Biasing Methods For Improving Rare Word Detection In Automatic Speech Recognition,

Mrinmoy Bhattacharjee, Nigmatulina Iuliia, Amrutha Prasad, Pradeep Rangappa, Srikanth Madikeri, Petr Motlicek, Hartmut Helmke, and Matthias Kleinert,

49th IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2024

In specialized domains like Air Traffic Control (ATC), a notable challenge in porting a deployed Automatic Speech Recognition (ASR) system from one airport to another is the alteration in the set of crucial words that must be accurately detected in the new environment. Typically, such words have limited occurrences in training data, making it impractical to retrain the ASR system. This paper explores innovative word-boosting techniques to improve the detection rate of such rare words in the ASR hypotheses for the ATC domain. Two acoustic models are investigated: a hybrid CNN-TDNNF model trained from scratch and a pre-trained wav2vec2-based XLSR model fine-tuned on a common ATC dataset. The word boosting is done in three ways. First, an out-of-vocabulary word addition method is explored. Second, G-boosting is explored, which amends the language model before building the decoding graph. Third, the boosting is performed on the fly during decoding using lattice re-scoring. The results indicate that the G-boosting method performs best and provides an approximately 30-43% relative improvement in recall of the boosted words. Moreover, a relative improvement of up to 48% is obtained upon combining G-boosting and lattice-rescoring.

Cross-transfer Knowledge between Speech and Text Encoders to Evaluate Customer Satisfaction,

Luis Felipe Parra-Gallego, Tilak Purohit, Bogdan Vlasenko, Juan Rafael Orozco-Arroyave, and Mathew Magimai.-Doss,

Annual Conference of the International Speech Communication Association (Interspeech), 2024

Customer Satisfaction (CS) in call centers influences customer loyalty and the company's reputation. Traditionally, CS evaluations were conducted manually or with classical machine learning algorithms; however, advancements in deep learning have led to automated systems that evaluate CS using speech and text analyses. Previous studies have shown the text approach to be more accurate but relies on an external ASR for transcription. This study introduces a cross-transfer knowledge technique, distilling knowledge from the BERT model into speech encoders like Wav2Vec2, WavLM, and Whisper. By enriching these encoders with BERT's linguistic information, we improve speech analysis performance and eliminate the need for an ASR. In evaluations on a dataset of customer opinions, our methods achieve over 92% accuracy in identifying CS categories, providing a faster and cost-effective solution compared to traditional text approaches.

D-LGP: Dynamic Logic-Geometric Program for Reactive Task and Motion Planning,

Teng Xue, Razmjoo Amirreza, and Sylvain Calinon,

IEEE International Conference on Robotics and Automation (ICRA), 2024

Many real-world sequential manipulation tasks involve a combination of discrete symbolic search and continuous motion planning, collectively known as combined task and motion planning (TAMP). However, prevailing methods often struggle with the computational burden and intricate combinatorial challenges, limiting their applications for online replanning in the real world. To address this, we propose Dynamic Logic-Geometric Program (D-LGP), a novel approach integrating Dynamic Tree Search and global optimization for efficient hybrid planning. Through empirical evaluation on three benchmarks, we demonstrate the efficacy of our approach, showcasing superior performance in comparison to state-of-the art techniques. We validate our approach through simulation and demonstrate its reactive capability to cope with online uncertainty and external disturbances in the real world. Project webpage: https://sites.google.com/view/dyn-lgp

Enhancing Ethical Explanations of Large Language Models through Iterative Symbolic Refinement,

Xin Quan, Marco Valentino, Louise A Dennis, and Andre Freitas,

18th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2024

An increasing amount of research in Natural Language Inference (NLI) focuses on the application and evaluation of Large Language Models (LLMs) and their reasoning capabilities. Despite their success, however, LLMs are still prone to factual errors and inconsistencies in their explanations, offering limited control and interpretability for inference in complex domains. In this paper, we focus on ethical NLI, investigating how hybrid neuro-symbolic techniques can enhance the logical validity and alignment of ethical explanations produced by LLMs. Specifically, we present an abductive-deductive framework named Logic-Explainer, which integrates LLMs with an external backward-chaining solver to refine step-wise natural language explanations and jointly verify their correctness, reduce incompleteness and minimise redundancy. An extensive empirical analysis demonstrates that Logic-Explainer can improve explanations generated via in-context learning methods and Chain-of-Thought (CoT) on challenging ethical NLI tasks, while, at the same time, producing formal proofs describing and supporting models' reasoning. As ethical NLI requires commonsense reasoning to identify underlying moral violations, our results suggest the effectiveness of neuro-symbolic methods for multi-step NLI more broadly, opening new opportunities to enhance the logical consistency, reliability, and alignment of LLMs.

Estimating the Causal Effects of Natural Logic Features in Transformer-Based NLI Models,

Julia Rozanova, Marco Valentino, and Andre Freitas,

Joint International Conference on Computational Linguistics, Language Resources and Evaluation, 2024

Rigorous evaluation of the causal effects of semantic features on language model predictions can be hard to achieve for natural language reasoning problems. However, this is such a desirable form of analysis from both an interpretability and model evaluation perspective, that it is valuable to investigate specific patterns of reasoning with enough structure and regularity to identify and quantify systematic reasoning failures in widely-used models. In this vein, we pick a portion of the NLI task for which an explicit causal diagram can be systematically constructed: the case where across two sentences (the premise and hypothesis), two related words/terms occur in a shared context. In this work, we apply causal effect estimation strategies to measure the effect of context interventions (whose effect on the entailment label is mediated by the semantic monotonicity characteristic) and interventions on the inserted word-pair (whose effect on the entailment label is mediated by the relation between these words). Extending related work on causal analysis of NLP models in different settings, we perform an extensive interventional study on the NLI task to investigate robustness to irrelevant changes and sensitivity to impactful changes of Transformers. The results strongly bolster the fact that similar benchmark accuracy scores may be observed for models that exhibit very different behavior. Moreover, our methodology reinforces previously suspected biases from a causal perspective, including biases in favor of upward-monotone contexts and ignoring the effects of negation markers.

Extending the Cooperative Dual-Task Space in Conformal Geometric Algebra,

Tobias Löw and Sylvain Calinon,

IEEE International Conference on Robotics and Automation (ICRA), 2024

In this work, we are presenting an extension of the cooperative dual-task space (CDTS) in conformal geometric algebra. The CDTS was first defined using dual quaternion algebra and is a well-established framework for the simplified definition of tasks using two manipulators. By integrating conformal geometric algebra, we aim to further enhance the geometric expressiveness and thus simplify the modeling of various tasks. We show this formulation by first presenting the CDTS and then its extension that is based around a cooperative point pair. This extension keeps all the benefits of the original formulation that is based on dual quaternions, but adds more tools for geometric modeling of the dual-arm tasks. We also present how this CGA CDTS can be seamlessly integrated with an optimal control framework in geometric algebra that was derived in previous work. In the experiments, we demonstrate how to model different objectives and constraints using the CGA-CDTS. Using a setup of two Franka Emika robots we then show the effectiveness of our approach using model predictive control in real world experiments.

Fine-tuning Self-Supervised Models For Language Identification Using Orthonormal Constraint,

Amrutha Prasad, Andrés Carofilis, Geoffroy Vanderreydt, Driss Khalil, Srikanth Madikeri, Petr Motlicek, and Christof Schüpbach,

49th IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2024

Self-supervised models trained with high linguistic diversity, such as the XLS-R model, can be effectively fine-tuned for the language recognition task. Typically, a back-end classifier followed by statistics pooling layer are added during training. Commonly used back-end classifiers require a large number of parameters to be trained, which is not ideal in limited data conditions. In this work, we explore smaller parameter back-ends using factorized Time Delay Neural Network (TDNN-F). The TDNN-F architecture is also integrated into Emphasized Channel Attention, Propagation and Aggregation-TDNN (ECAPA-TDNN) models, termed ECAPA-TDNN-F, reducing the number of parameters by 30 to 50% absolute, with competitive accuracies and no change in minimum cost. The results show that the ECAPA-TDNN-F can be extended to tasks where ECAPA-TDNN is suitable. We also test the effectiveness of a linear classifier and a variant, the Orthonormal linear classifier, previously used in x-vector type systems. The models are trained with NIST LRE17 data and evaluated on NIST LRE17, LRE22 and the ATCO2 LID datasets. Both linear classifiers outperform conventional back-ends with improvements in accuracy between 0.9% and 9.1%.

Generalized Policy Iteration using Tensor Approximation for Hybrid Control,

Suhan Shetty, Teng Xue, and Sylvain Calinon,

International Conference on Learning Representations (ICLR), 2024

Optimal Control of dynamic systems involving hybrid actions is a challenging task in robotics. To address this, we present a novel algorithm called Generalized Policy Iteration using Tensor Train (TTPI) that belongs to the class of Approximate Dynamic Programming (ADP). We use a low-rank tensor approximation technique called Tensor Train (TT) to approximate the state-value and advantage function which enables us to efficiently handle hybrid action space. We demonstrate the superiority of our approach over previous baselines for some benchmark problems with hybrid action spaces. Additionally, the robustness and generalization of the policy for hybrid systems are showcased through a real-world robotics experiment involving a non-prehensile manipulation task.

Improving Semantic Control in Discrete Latent Spaces with Transformer Quantized Variational Autoencoders,

Yingji Zhang, Danilo Carvalho, Marco Valentino, Ian Pratt-Hartmann, and Andre Freitas,

18th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2024

Achieving precise semantic control over the latent spaces of Variational AutoEncoders (VAEs) holds significant value for downstream tasks in NLP as the underlying generative mechanisms could be better localised, explained and improved upon. Recent research, however, has struggled to achieve consistent results, primarily due to the inevitable loss of semantic information in the variational bottleneck and limited control over the decoding mechanism. To overcome these challenges, we investigate discrete latent spaces in Vector Quantized Variational AutoEncoders (VQVAEs) to improve semantic control and generation in Transformer-based VAEs. In particular, We propose T5VQVAE, a novel model that leverages the controllability of VQVAEs to guide the self-attention mechanism in T5 at the token-level, exploiting its full generalization capabilities. Experimental results indicate that T5VQVAE outperforms existing state-of-the-art VAE models, including Optimus, in terms of controllability and preservation of semantic information across different tasks such as auto-encoding of sentences and mathematical expressions, text transfer, and inference. Moreover, T5VQVAE exhibits improved inference capabilities, suggesting potential applications for downstream natural language and symbolic reasoning tasks.

Logic-Skill Programming: An Optimization-based Approach to Sequential Skill Planning,

Teng Xue, Amirreza Razmjoo, Suhan Shetty, and Sylvain Calinon,

Robotics: Science and Systems (RSS), 2024

Recent advances in robot skill learning have un locked the potential to construct task-agnostic skill libraries, facilitating the seamless sequencing of multiple simple manipulation primitives (aka. skills) to tackle significantly more complex tasks. Nevertheless, determining the optimal sequence for independently learned skills remains an open problem, particularly when the objective is given solely in terms of the final geometric configuration rather than a symbolic goal. To address this challenge, we propose Logic-Skill Programming (LSP), an optimization based approach that sequences independently learned skills to solve long-horizon tasks. We formulate a first-order extension of a mathematical program to optimize the overall cumulative reward of all skills within a plan, abstracted by the sum of value functions. To solve such programs, we leverage the use of tensor train factorization to construct the value function space, and rely on alternations between symbolic search and skill value optimization to find the appropriate skill skeleton and optimal subgoal sequence. Experimental results indicate that the obtained value functions provide a superior approximation of cumulative rewards compared to state-of-the-art reinforcement learning methods. Furthermore, we validate LSP in three manipulation domains, encompassing both prehensile and non-prehensile primitives. The results demonstrate its capability to identify the optimal solution over the full logic and geometric path. The real-robot experiments showcase the effectiveness of our approach to cope with contact uncertainty and external disturbances in the real world.

Multi-Relational Hyperbolic Word Embeddings from Natural Language Definitions,

Marco Valentino, Danilo Carvalho, and Andre Freitas,

18th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2024

Natural language definitions possess a recursive, self-explanatory semantic structure that can support representation learning methods able to preserve explicit conceptual relations and constraints in the latent space. This paper presents a multi-relational model that explicitly leverages such a structure to derive word embeddings from definitions. By automatically extracting the relations linking defined and defining terms from dictionaries, we demonstrate how the problem of learning word embeddings can be formalised via a translational framework in Hyperbolic space and used as a proxy to capture the global semantic structure of definitions. An extensive empirical analysis demonstrates that the framework can help imposing the desired structural constraints while preserving the semantic mapping required for controllable and interpretable traversal. Moreover, the experiments reveal the superiority of the Hyperbolic word embeddings over the Euclidean counterparts and demonstrate that the multi-relational approach can obtain competitive results when compared to state-of-the-art neural models, with the advantage of being intrinsically more efficient and interpretable.

Multitask Speech Recognition and Speaker Change Detection for Unknown Number of Speakers,

Shashi Kumar, Srikanth Madikeri, Nigmatulina Iuliia, Esaú Villatoro-Tello, Petr Motlicek, Karthik Pandia D S, S. Pavankumar Dubagunta, and Aravind Ganapathiraju,

49th IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2024

Traditionally, automatic speech recognition (ASR) and speaker change detection (SCD) systems have been independently trained to generate comprehensive transcripts accompanied by speaker turns. Recently, joint training of ASR and SCD systems, by inserting speaker turn tokens in the ASR training text, has been shown to be successful. In this work, we present a multitask alternative to the joint training approach. Results obtained on the mix-headset audios of AMI corpus show that the proposed multitask training yields an absolute improvement of 1.8% in coverage and purity based F1 score on SCD task without ASR degradation. We also examine the trade-offs between the ASR and SCD performance when trained using multitask criteria. Additionally, we validate the speaker change information in the embedding spaces obtained after different transformer layers of a self-supervised pre-trained model, such as XLSR-53, by integrating an SCD classifier at the output of specific transformer layers. Results reveal that the use of different embedding spaces from XLSR-53 model for multitask ASR and SCD is advantageous.

Normalizing Flows for Speaker and Language Recognition Backend,

Aleix Espuña, Amrutha Prasad, Petr Motlicek, Srikanth Madikeri, and Christof Schüpbach,

Odyssey 2024: The Speaker and Language Recognition Workshop, 2024

In this paper, we address the Gaussian distribution assumption made in PLDA, a popular back-end classifier used in Speaker and Language recognition tasks. We study normalizing flows, which allow using non-linear transformations and still obtain a model that can explicitly represent a probability density. The model makes no assumption about the distribution of the observations. This alleviates the need for length normalization, a well known data preprocessing step used to boost PLDA performance. We demonstrate the effectiveness of this flow model on NIST SRE16, LRE17 and LRE22 datasets. We observe that when applying length normalization, both the flow model and PLDA achieve similar EERs for SRE16 (11.5% vs 11.8%). However, when length normalization is not applied, the flow shows more robustness and offers better EERs (13.1% vs 17.1%). For LRE17 and LRE22, the best classification accuracies (84.2%, 75.5%) are obtained by the flow model without any need for length normalization.

On the Utility of Speech and Audio Foundation Models for Marmoset Call Analysis,

Eklavya Sarkar, and Mathew Magimai-Doss,

4th International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots, 2024

Marmoset monkeys encode vital information in their calls and serve as a surrogate model for neuro-biologists to understand the evolutionary origins of human vocal communication. Traditionally analyzed with signal processing-based features, recent approaches have utilized self-supervised models pre-trained on human speech for feature extraction, capitalizing on their ability to learn a signal's intrinsic structure independently of its acoustic domain. However, the utility of such foundation models remains unclear for marmoset call analysis in terms of multi-class classification, bandwidth, and pre-training domain. This study assesses feature representations derived from speech and general audio domains, across pre-training bandwidths of 4, 8, and 16 kHz for marmoset call-type and caller classification tasks. Results show that models with higher bandwidth improve performance, and pre-training on speech or general audio yields comparable results, improving over a spectral baseline.

Probability-Aware Word-Confusion-Network-to-Text Alignment Approach for Intent Classification,

Esaú Villatoro-Tello, Srikanth Madikeri, Bidisha Sharma, Driss Khalil, Shashi Kumar, Iuliia Nigmatulina, Petr Motlicek, and Aravind Ganapathiraju,

49th IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2024

Spoken Language Understanding (SLU) technologies have seen a big improvement due to the effective pre-training of speech representations. A common requirement of industry-based solutions is the portability to deploy SLU models in voice-assistant devices. Thus, distilling knowledge from large text-based language models has become an attractive solution for achieving good performance and guaranteeing portability. In this paper, we introduce a novel architecture that uses a cross-modal attention mechanism to extract bin-level contextual embedding from a word-confusion network (WNC) encoding such that these can be directly compared and aligned with traditional text-based contextual embedding. This alignment is achieved using a recently proposed tokenwise contrastive loss function. We validated our architecture's effectiveness by fine-tuning our WCN-based pre-trained model to perform intent classification on the SLURP dataset. Obtained accuracy (81%), depicts a 9.4% relative improvement compared to a recent and equivalent E2E method.

Representing Robot Geometry as Distance Fields: Applications to Whole-body Manipulation,

Yiming Li, Yan Zhang, Amirreza Razmjoo, and Sylvain Calinon,

IEEE International Conference on Robotics and Automation (ICRA), 2024

In this work, we propose a novel approach to represent robot geometry as distance fields (RDF) that extends the principle of signed distance fields (SDFs) to articulated kinematic chains. Our method employs a combination of Bernstein polynomials to encode the signed distance for each robot link with high accuracy and efficiency while ensuring the mathematical continuity and differentiability of SDFs. We further leverage the kinematics chain of the robot to produce the SDF representation in joint space, allowing robust distance queries in arbitrary joint configurations. The proposed RDF representation is differentiable and smooth in both task and joint spaces, enabling its direct integration to optimization problems. Additionally, the 0-level set of the robot corresponds to the robot surface, which can be seamlessly integrated into whole-body manipulation tasks. We conduct various experiments in both simulations and with 7-axis Franka Emika robots, comparing against baseline methods, and demonstrating its effectiveness in collision avoidance and whole-body manipulation tasks. Project page: https://sites.google.com/view/lrdf/home

σ-GPTs: A New Approach to Autoregressive Models,

Arnaud Pannatier, Evann Courdier, and Francois Fleuret,

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2024

Autoregressive models, such as the GPT family, use a fixed order, usually left-to-right, to generate sequences. However, this is not a necessity. In this paper, we challenge this assumption and show that by simply adding a positional encoding for the output, this order can be modulated on-the-fly per-sample which offers key advantageous properties. It allows for the sampling of and conditioning on arbitrary subsets of tokens, and it also allows sampling in one shot multiple tokens dynamically according to a rejection strategy, leading to a sub-linear number of model evaluations. We evaluate our method across various domains, including language modeling, path-solving, and aircraft vertical rate prediction, decreasing the number of steps required for generation by an order of magnitude.

Speech and Language Recognition with Low-rank Adaptation of Pretrained Models,

Amrutha Prasad, Srikanth Madikeri, Driss Khalil, Petr Motlicek, and Christof Schüpbach,

Annual Conference of the International Speech Communication Association (Interspeech), 2024

Fine-tuning large pre-trained models demands considerable computational resources, posing practical constraints. Majority of the total number of parameters in these models are used by fully connected layers. In this work, we consider applying a semi-orthogonal constraint, followed by full fine-tuning to the fully connected layers reduces model parameters significantly without sacrificing efficacy in downstream tasks. Specifically, we consider wav2vec2.0 XLS-R and Whisper models for Automatic Speech Recognition and Language Recognition. Our results show that we can reduce the model size by approximately 24%during both training and inference time with 0.7% absolute drop in performance for XLS-R and no drop in performance for Whisper for ASR. In combination with performance-efficient training with low-rank adapters, the resource requirements for training can be further reduced by up to 90%.

Towards interfacing large language models with ASR systems using confidence measures and prompting,

Maryam Naderi, Enno Hermann, Alexandre Nanchen, Sevada Hovsepyan, and Mathew Magimai-Doss,

Annual Conference of the International Speech Communication Association (Interspeech), 2024

As large language models (LLMs) grow in parameter size and capabilities, such as interaction through prompting, they open up new ways of interfacing with automatic speech recognition (ASR) systems beyond rescoring n-best lists. This work investigates post-hoc correction of ASR transcripts with LLMs. To avoid introducing errors into likely accurate transcripts, we pro pose a range of confidence-based filtering methods. Our results indicate that this can improve the performance of less competitive ASR systems.

Towards Robo-Coach: Robot Interactive Stiffness/Position Adaptation for Human Strength and Conditioning Training,

Chenzui Li, Xi Wu, Tao Teng, Sylvain Calinon, and Fei Chen,

IEEE International Conference on Robotics and Automation (ICRA), 2024

Traditional strength and conditioning training relies on the utilization of free weights, such as weighted implements, to elicit external stimuli. However, this approach poses a significant challenge when attempting to modify or adjust the loads within a single training set. This paper introduces an innovative method for achieving adjustable loads during resistance training by leveraging physical Human-Robot Interaction (pHRI). The primary objective is to regulate targeted muscle activation through the use of Robo-Coach (robotic coach system). We first utilize a Task-Parameterized Gaussian Mixture Model (TP-GMM) to learn the motion of coach demonstration, which can be generalized for the trainees. The 3D path extracted from the generated trajectory is then projected onto a 2D plane with respect to the direction of the load. Furthermore, we propose a hybrid stiffness/position generator for online task execution. This generator determines the desired positions in the 2D plane according to the contact point displacements in the stimuli direction and, simultaneously, sets the desired stiffness based on the muscle activation feedback. Finally, the Robo-Coach is implemented with a variable impedance controller to achieve load-adjustable resistance training with the trainee. The biceps curl exercises were conducted and the results showed favorable performance, indicating the effectiveness of this approach.

Understanding the effects of language-specific class imbalance in multilingual fine-tuning,

Vincent Jung and Lonneke van der Plas,

Findings of the European chapter of Association for Computational Linguistics, 2024

We study the effect of one type of imbalance often present in real-life multilingual classification datasets: an uneven distribution of labels across languages. We show evidence that fine tuning a transformer-based Large Language Model (LLM) on a dataset with this imbalance leads to worse performance, a more pronounced separation of languages in the latent space, and the promotion of uninformative features. We modify the traditional class weighing approach to imbalance by calculating class weights separately for each language and show that this helps mitigate those detrimental effects. These results create awareness of the negative effects of language-specific class imbalance in multilingual fine-tuning and the way in which the model learns to rely on the separation of languages to perform the task.

Comparing Data-Driven And Handcrafted Features For Dimensional Emotion Recognition,

Bogdan Vlasenko, Sargam Vyas, and Mathew Magimai-Doss,

49th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024

Speech Emotion Recognition (SER) has garnered significant attention over the past two decades. In the early stages of SER technology, 'brute force'-based techniques led to a significant expansion in knowledge-based acoustic feature representation (FR) for modeling sparse emotional data. However, as deep learning techniques have become more powerful, their direct application has been limited by the scarcity of well-annotated emotional data. As a result, pre-trained neural embedding on large speech corpora have gained popularity for SER tasks. These embedding leverage existing transfer learning methods suitable for general-purpose self-supervised learning (SSL) representations. Recent studies on downstream SSL techniques for dimensional SER have shown promising results. In this research, we aim to evaluate the emotion-discriminative characteristics of neural embedding in general cases (out-of-domain) and when fine-tuned for SER (in-domain). Given that most SSL techniques are pre-trained primarily on English speech, we plan to use speech emotion corpora in both language-matched and mismatched conditions. We will assess the discriminative characteristics of both handcrafted and standalone neural embedding as FRs.

CCDb-HG: Novel Annotations and Gaze-Aware Representations for Head Gesture Recognition,

Pierre Vuillecard, Arya Farkhondeh, Michael Villamizar, and Jean-Marc Odobez,

18th IEEE Int. Conference on Automatic Face and Gesture Recognition (FG), Istanbul, 2024

Despite remarkable progress in various human behavior perception tasks, head gesture recognition (HGR) has received limited attention in terms of datasets, benchmarks, and methods. In this work, we aim to address this gap and make two main contributions. First, we densely annotated the existing large-scale conversational dataset CCDb with diverse head gesture categories. This results in the CCDb-HG dataset, which can serve as a comprehensive benchmark for HGR research. Secondly, while previous gesture recognition methods have largely relied on head pose or facial landmarks as input, we propose to explore in addition the use of gaze to resolve ambiguous cases. This follows from the fact that head dynamics in interactions is driven by two main functions: communication (i.e. head gestures) and attention (i.e. gazing at other people or objects of interest). In fact, the head dynamics associated with attention activities can be confused for communication gestures, even though the gaze patterns are quite different in the two cases. In addition, we study several geometric and temporal data augmentation techniques to improve the generalization across novel viewpoints, as well as different model architectures to establish baseline performance on CCDb-HG. Our findings provide insights into various aspects of HGR and motivate further research in this field. To facilitate reproducibility, we will release the CCDb-HG annotations, code, and HGR models.

Exploring the Zero-Shot Capabilities of Vision-Language Models for Improving Gaze Following,

Anshul Gupta, Pierre Vuillecard, Arya Farkhondeh, and Jean-Marc Odobez,

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Workshop on Gaze Estimation and Prediction in the Wild, 2024

Contextual cues related to a person's pose and interactions with objects and other people in the scene can provide valuable information for gaze following. While existing methods have focused on dedicated cue extraction methods, in this work we investigate the zero-shot capabilities of Vision-Language Models (VLMs) for extracting a wide array of contextual cues to improve gaze following performance. We first evaluate various VLMs, prompting strategies, and in-context learning (ICL) techniques for zero-shot cue recognition performance. We then use these insights to extract contextual cues for gaze following, and investigate their impact when incorporated into a state of the art model for the task. Our analysis indicates that BLIP-2 is the overall top performing VLM and that ICL can improve performance. We also observe that VLMs are sensitive to the choice of the text prompt although ensembling over multiple text prompts can provide more robust performance. Additionally, we discover that using the entire image along with an ellipse drawn around the target person is the most effective strategy for visual prompting. For gaze following, incorporating the extracted cues results in better generalization performance, especially when considering a larger set of cues, highlighting the potential of this approach.

Neural Redshift: Random Networks are not Random Functions,

Damien Teney, Armand Nicolicioiu, Valentin Hartmann, and Ehsan Abbasnejad,

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024

Our understanding of the generalization capabilities of neural networks (NNs) is still incomplete. Prevailing explanations are based on implicit biases of gradient descent (GD) but they cannot account for the capabilities of models from gradient-free methods nor the simplicity bias recently observed in untrained networks. This paper seeks other sources of generalization in NNs. To understand the inductive biases provided by architectures independently from GD, we examine untrained, random-weight networks. Even simple MLPs show strong inductive biases: uniform sampling in weight space yields a very biased distribution of functions in terms of complexity. But unlike common wisdom, NNs do not have an inherent "simplicity bias". This property depends on components such as ReLUs, residual connections, and layer normalizations. Alternative architectures can be built with a bias for any level of complexity. Transformers also inherit all these properties from their building blocks. We provide a fresh explanation for the success of deep learning independent from gradient-based training. It points at promising avenues for controlling the solutions implemented by trained models.

Bi-directional training for composed image retrieval via text prompt learning,

Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould,

Winter Conference on Applications of Computer Vision (WACV), 2024

Composed image retrieval searches for a target image based on a multi-modal user query comprised of a reference image and modification text describing the desired changes. Existing approaches to solving this challenging task learn a mapping from the (reference image, modification text)-pair to an image embedding that is then matched against a large image corpus. One area that has not yet been explored is the reverse direction, which asks the question, what reference image when modified as described by the text would produce the given target image? In this work we propose a bi-directional training scheme that leverages such reversed queries and can be applied to existing composed image retrieval architectures with minimum changes, which improves the performance of the model. To encode the bi-directional query we prepend a learnable token to the modification text that designates the direction of the query and then fine-tune the parameters of the text-embedding module. We make no other changes to the network architecture. Experiments on two standard datasets show that our novel approach achieves improved performance over a baseline BLIP-based model that itself already achieves competitive performance.

Tracking linguistic information in transformer-based sentence embeddings through targeted sparsification,

Vivi Nastase and Paola Merlo,

9th Workshop on Representation Learning for NLP (RepL4NLP)), Association for Computational Linguistics, 2024

Analyses of transformer-based models have shown that they encode a variety of linguistic information from their textual input. While these analyses have shed a light on the relation between linguistic information on one side, and internal architecture and parameters on the other, a question remains unanswered: how is this linguistic information reflected in sentence embeddings? Using datasets consisting of sentences with known structure, we test to what degree information about chunks (in particular noun, verb or prepositional phrases), such as grammatical number, or semantic role, can be localized in sentence embeddings. Our results show that such information is not distributed over the entire sentence embedding, but rather it is encoded in specific regions. Understanding how the information from an input text is compressed into sentence embeddings helps understand current transformer models and help build future explainable neural models.

Selective mixup helps with distribution shifts, but not (only) because of mixup,

Damien Teney, Jindong Wang, and Ehsan Abbasnejad,

International Conference on Machine Learning (ICML), 2024

Mixup is a highly successful technique to improve generalization of neural networks by augmenting the training data with combinations of random pairs. Selective mixup is a family of methods that apply mixup to specific pairs, e.g. only combining examples across classes or domains. These methods have claimed remarkable improvements on benchmarks with distribution shifts, but their mechanisms and limitations remain poorly understood. We examine an overlooked aspect of selective mixup that explains its success in a completely new light. We find that the non-random selection of pairs affects the training distribution and improve generalization by means completely unrelated to the mixing. For example, in binary classification, mixup across classes implicitly resamples the data for a uniform class distribution - a classical solution to label shift. We show empirically that this implicit resampling explains much of the improvements in prior work. Theoretically, these results rely on a regression toward the mean, an accidental property that we identify in several datasets. We have found a new equivalence between two successful methods: selective mixup and resampling. We identify limits of the former, confirm the effectiveness of the latter, and find better combinations of their respective benefits.

Explaining models relating objects and privacy,

Alessio Xompero, Myriam Bontonou, Jean-Michel Arbona, Emmanouil Benetos, and Andrea Cavallaro,

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024

Accurately predicting whether an image is private before sharing it online is difficult due to the vast variety of content and the subjective nature of privacy itself. In this paper, we evaluate privacy models that use objects extracted from an image to determine why the image is predicted as private. To explain the decision of these models, we use feature-attribution to identify and quantify which objects (and which of their features) are more relevant to privacy classification with respect to a reference input (i.e., no objects localised in an image) predicted as public. We show that the presence of the person category and its cardinality is the main factor for the privacy decision. Therefore, these models mostly fail to identify private images depicting documents with sensitive data, vehicle ownership, and internet activity, or public images with people (e.g., an outdoor concert or people walking in a public space next to a famous landmark). As baselines for future benchmarks, we also devise two strategies that are based on the person presence and cardinality and achieve comparable classification performance of the privacy models.

Sparse multi-view hand-object reconstruction for unseen environments,

Yik Lung Pang, Changjae Oh, and Andrea Cavallaro,

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024

Recent works in hand-object reconstruction mainly focus on the single-view and dense multi-view settings. On the one hand single-view methods can leverage learned shape priors to generalise to unseen objects but are prone to inaccuracies due to occlusions. On the other hand dense multi-view methods are very accurate but cannot easily adapt to unseen objects without further data collection. In contrast sparse multi-view methods can take advantage of the additional views to tackle occlusion while keeping the computational cost low compared to dense multi-view methods. In this paper we consider the problem of hand-object reconstruction with unseen objects in the sparse multi-view setting. Given multiple RGB images of the hand and object captured at the same time our model SVHO combines the predictions from each view into a unified reconstruction without optimisation across views. We train our model on a synthetic hand-object dataset and evaluate directly on a real world recorded hand-object dataset with unseen objects. We show that while reconstruction of unseen hands and objects from RGB is challenging additional views can help improve the reconstruction quality.

Open-Vocabulary Object 6D Pose Estimation,

Jaime Corsetti, Davide Boscaini, Changjae Oh, Andrea Cavallaro, and Fabio Poiesi,

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024

We introduce the new setting of open-vocabulary object 6D pose estimation, in which a textual prompt is used to specify the object of interest. In contrast to existing approaches, in our setting (i) the object of interest is specified solely through the textual prompt, (ii) no object model (e.g., CAD or video sequence) is required at inference, and (iii) the object is imaged from two RGBD viewpoints of different scenes. To operate in this setting, we introduce a novel approach that leverages a Vision-Language Model to segment the object of interest from the scenes and to estimate its relative 6D pose. The key of our approach is a carefully devised strategy to fuse object-level information provided by the prompt with local image features, resulting in a feature space that can generalize to novel concepts. We validate our approach on a new benchmark based on two popular datasets, REAL275 and Toyota-Light, which collectively encompass 34 object instances appearing in four thousand image pairs. The results demonstrate that our approach outperforms both a well-established hand-crafted method and a recent deep learning-based baseline in estimating the relative 6D pose of objects in different scenes. Code and dataset are available at https://jcorsetti.github.io/oryon.

Image-guided topic modeling for interpretable privacy classification,

Alina Elena Baia and Andrea Cavallaro,

European Conference on Computer Vision (ECCV) Workshops, 2024

Predicting and explaining the private information contained in an image in human-understandable terms is a complex and contextual task. This task is challenging even for large language models. To facilitate the understanding of privacy decisions, we propose to predict image privacy based on a set of natural language content descriptors. These content descriptors are associated with privacy scores that reflect how people perceive image content. We generate descriptors with our novel Image-guided Topic Modeling (ITM) approach. ITM leverages, via multimodality alignment, both vision information and image textual descriptions from a vision language model. We use the ITM-generated descriptors to learn a privacy predictor, Priv×ITM, whose decisions are interpretable by design. Our Priv×ITM, classifier outperforms the reference interpretable method by 5 percentage points in accuracy and performs comparably to the current non-interpretable state-of-the-art model. https://jcorsetti.github.io/oryon.

Hardware-effective Approaches for Skill Extraction in Job Offers and Resumes,
Laura Vásquez-Rodríguez, Bertrand Audrin, Samuel Michel, Samuele Galli, Julneth Rogenhofer, Jacopo Negro Cusa, and Lonneke van der Plas,
18th ACM Conference on Recommender Systems, 4th Workshop on Recommender Systems for Human Resources 2024

Recent work on the automatic extraction of skills has mainly focused on job offers and not resumes while using state-of-the-art resource-intensive methods and considerable amounts of annotated data. However, in real-life industrial contexts, the computational resources and the annotated data available can be limited, especially for resumes. In this paper, we present our experiments that use hardware-effective methods and circumvent the need for large amounts of annotated data. We experiment with various methods that vary in hardware requirements and complexity. We evaluate these systems both on public and commercial data, using gold-standard for evaluation. We find that standalone rule-based and semantic model performance on the skill extraction task is limited and variable between job offers and resumes. However, neural models can perform competitively and be more stable, even when using small datasets, with an improvement of ~30%. We present our experiments using minimal hardware, mostly CPU-based with less than 8 GB of RAM for rule-based and semantic methods and using GPUs for neural models with a maximum memory usage for both CPU and GPU of 24 GB, with less than 25 minutes of training time.

Fast Streaming Transducer ASR Prototyping via Knowledge Distillation with Whisper,
Iuliia Thorbecke,Juan Zuluaga-Gomez,Esaú Villatoro-Tello, Shashi Kumar, Pradeep Rangappa, Sergio Burdisso, Petr Motlicek, Pandia Karthik, and Aravind Ganapathiraju,
Conference on Empirical Methods in Natural Language Processing, 2024

The training of automatic speech recognition (ASR) with little to no supervised data remains an open question. In this work, we demonstrate that streaming Transformer-Transducer (TT) models can be trained from scratch in consumer and accessible GPUs in their entirety with pseudo-labeled (PL) speech from foundational speech models (FSM). This allows training a robust ASR model just in one stage and does not require large data and computational budget compared to the two-step scenario with pre-training and fine-tuning. We perform a comprehensive ablation on different aspects of PL-based streaming TT models such as the impact of (1) shallow fusion of n-gram LMs, (2) contextual biasing with named entities, (3) chunk-wise decoding for low-latency streaming applications, and (4) TT overall performance as the function of the FSM size. Our results demonstrate that TT can be trained from scratch without supervised data, even with very noisy PLs. We validate the proposed framework on 6 languages from CommonVoice and propose multiple heuristics to filter out hallucinated PLs.

TokenVerse: Towards Unifying Speech and NLP Tasks via Transducer-based ASR,
Kumar, Shashi, Srikanth Madikeri, Juan Zuluaga-Gomez, Iuliia Thorbecke, Esaú Villatoro-Tello, Sergio Burdisso, Petr Motlicek, Pandia Karthik, and Aravind Ganapathiraju,
Conference on Empirical Methods in Natural Language Processing, 2024

In traditional conversational intelligence from speech, a cascaded pipeline is used, involving tasks such as voice activity detection, diarization, transcription, and subsequent processing with different NLP models for tasks like semantic endpointing and named entity recognition (NER). Our paper introduces TokenVerse, a single Transducer-based model designed to handle multiple tasks. This is achieved by integrating task-specific tokens into the reference text during ASR model training, streamlining the inference and eliminating the need for separate NLP models. In addition to ASR, we conduct experiments on 3 different tasks: speaker change detection, endpointing, and NER. Our experiments on a public and a private dataset show that the proposed method improves ASR by up to 7.7% in relative WER while outperforming the cascaded pipeline approach in individual task performance. Our code is publicly available: https://github.com/idiap/tokenverse-unifying-speech-nlp

Dialog2Flow: Pre-training Soft-Contrastive Action-Driven Sentence Embeddings for Automatic Dialog Flow Extraction,

Sergio Burdisso, Srikanth Madikeri, and Petr Motlicek,

Conference on Empirical Methods in Natural Language Processing, 2024

Efficiently deriving structured workflows from unannotated dialogs remains an underexplored and formidable challenge in computational linguistics. Automating this process could significantly accelerate the manual design of workflows in new domains and enable the grounding of large language models in domain-specific flowcharts, enhancing transparency and controllability. In this paper, we introduce Dialog2Flow (D2F) embeddings, which differ from conventional sentence embeddings by mapping utterances to a latent space where they are grouped according to their communicative and informative functions (i.e., the actions they represent). D2F allows for modeling dialogs as continuous trajectories in a latent space with distinct action-related regions. By clustering D2F embeddings, the latent space is quantized, and dialogs can be converted into sequences of region/action IDs, facilitating the extraction of the underlying workflow. To pre-train D2F, we build a comprehensive dataset by unifying twenty task-oriented dialog datasets with normalized per-turn action annotations. We also introduce a novel soft contrastive loss that leverages the semantic information of these actions to guide the representation learning process, showing superior performance compared to standard supervised contrastive loss. Evaluation against various sentence embeddings, including dialog-specific ones, demonstrates that D2F yields superior qualitative and quantitative results across diverse domains.

DiffuCOMET: Contextual Commonsense Knowledge Diffusion,

Silin Gao, Mete Ismayilzada, Mengjie Zhao, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut,

62nd Annual Meeting of the Association for Computational Linguistics, 2024

Inferring contextually-relevant and diverse commonsense to understand narratives remains challenging for knowledge models. In this work, we develop a series of knowledge models, DiffuCOMET, that leverage diffusion to learn to reconstruct the implicit semantic connections between narrative contexts and relevant commonsense knowledge. Across multiple diffusion steps, our method progressively refines a representation of commonsense facts that is anchored to a narrative, producing contextually-relevant and diverse commonsense inferences for an input context. To evaluate DiffuCOMET, we introduce new metrics for commonsense inference that more closely measure knowledge diversity and contextual relevance. Our results on two different benchmarks, ComFact and WebNLG+, show that knowledge generated by DiffuCOMET achieves a better trade-off between commonsense diversity, contextual relevance and alignment to known gold references, compared to baseline knowledge models.

Nonparametric Variational Regularisation of Pretrained Transformers,

Fabio Fehr and James Henderson,

First conference on Language Modelling, 2024

Retrained transformers have demonstrated impressive abilities, but tend not to generalise well out-of-domain and are very expensive to fine-tune on new domain data. Nonparametric Variational Information Bottleneck (NVIB) has been proposed as a regulariser for training cross-attention in transformers, potentially addressing this domain overfitting problem. We extend the NVIB framework to replace all types of attention functions in transformers. We show that existing pretrained transformers can be reinterpreted as nonparametric variational models using an empirical prior distribution and identity initialisation with controllable hyperparameters. We then show that changing the initialisation introduces a novel, information-theoretic post-training regularisation in the attention mechanism, which improves out-of-domain generalisation on NLP tasks without any additional training. This success supports the hypothesis that the way pretrained transformer embeddings represent information is accurately characterised by nonparametric variational Bayesian models.

GADePo: Graph-Assisted Declarative Pooling Transformers for Document-Level Relation Extraction,

Andrei Catalin Coman, Christos Theodoropoulos, Marie-Francine Moens and James Henderson,

3rd Workshop on Knowledge Augmented Methods for NLP, 2024

Document-level relation extraction typically relies on text-based encoders and hand-coded pooling heuristics to aggregate information learned by the encoder. In this paper, we leverage the intrinsic graph processing capabilities of the Transformer model and propose replacing hand-coded pooling methods with new tokens in the input, which are designed to aggregate information via explicit graph relations in the computation of attention weights. We introduce a joint text-graph Transformer model and a graph-assisted declarative pooling (GADePo) specification of the input, which provides explicit and high-level instructions for information aggregation. GADePo allows the pooling process to be guided by domain-specific knowledge or desired outcomes but still learned by the Transformer, leading to more flexible and customizable pooling strategies. We evaluate our method across diverse datasets and models and show that our approach yields promising results that are consistently better than those achieved by the hand-coded pooling functions.

A Human Perspective to AI-based Candidate Screening,

Laura Vásquez-Rodríguez, Bertrand Audrin, Samuel Michel, Samuele Galli, Julneth Rogenhofer, Jacopo Negro Cusa and Lonneke van der Plas,

8th Hawaii International Conference on System Sciences (HICSS), 20244

Skill extraction is at the core of algorithmic hiring. It is based on identifying terms commonly found in both targets (i.e., resumes and job offers), aiming at identifying a “match” or correspondence between both. This paper focuses on skill extraction from resumes, as opposed to job offers, and considers this task both from the Human Resource Management (HRM) and AI points of view. We discuss challenges identified by both fields and explain how collaboration is instrumental for a successful digital transformation of HRM. We argue that annotation efforts are an ideal example of where collaboration between both fields is needed and present an annotation effort on 46 resumes with 41 trained annotators, resulting in a total of 116 annotations. We analyze the skills extracted by multiple different systems and compare those to the skills selected by the annotators, and find that the skills extracted differ a lot in terms of length and semantic content. The skills extracted with conversational Large Language Models (LLMs) tend to be very long and detailed, other systems are very concise, whereas humans are in the middle. In terms of semantic similarity, conversational LLMs are closer to human outputs than other systems. Our analysis proposes a different perspective to understand the well-studied, but still unsolved skill extraction task. Finally, we provide recommendations for the skill extraction task that aligns with both HR and computational perspectives.

Robust Manipulation Primitive Learning via Domain Contraction

Teng Xue, Amirreza Razmjoo Fard, Suhan Shetty and Sylvain Calinon,

Conference on Robot Learning, 2024

Contact-rich manipulation plays an important role in human daily ac tivities, but uncertain parameters pose significant challenges for robots to achieve comparable performance through planning and control. To address this issue, do main adaptation and domain randomization have been proposed for robust pol icy learning. However, they either lose the generalization ability across diverse instances or perform conservatively due to neglecting instance-specific informa tion. In this paper, we propose a bi-level approach to learn robust manipulation primitives, including parameter-augmented policy learning using multiple mod els, and parameter-conditioned policy retrieval through domain contraction. This approach unifies domain randomization and domain adaptation, providing optimal behaviors while keeping generalization ability. We validate the proposed method on three contact-rich manipulation primitives: hitting, pushing, and reorientation. The experimental results showcase the superior performance of our approach in generating robust policies for instances with diverse physical parameters.

Sparse Optical Sampling in the Close Proximity of a Robotic Arm,

Martin Laurenzis, Ante Marić, Emmanuel Bacher, Mateusz Pietrzak, Stéphane Schertzer, Francesco Grella, and Sylvain Calino,

Springer Proceedings in Advanced Robotics, 2024

Close collaboration between humans and robots needs a sensing infrastructure to monitor the robot environment and secure human-robot interaction. In this context, we investigate sparse optical range sampling using a distributed network of robot mounted Time-of-Flight (ToF) sensors. We present an evaluation of sensor candidates, provide experimental characterization of an early prototype and show strategies for environment modeling and object reconstruction.

TESS: Text-to-text selfconditioned simplex diffusion,

Rabeeh Karimi Mahabadi, Hamish Ivison, Jaesung Tae, James Henderson, Iz Beltagy, Matthew E. Peters and Arman Cohan,

18th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2024

Diffusion models have emerged as a powerful paradigm for generation, obtaining strong performance in various continuous domains. However, applying continuous diffusion models to natural language remains challenging due to its discrete nature and the need for a large number of diffusion steps to generate text, making diffusion-based generation expensive. In this work, we propose Text-to-text Self-conditioned Simplex Diffusion (TESS), a text diffusion model that is fully non-autoregressive, employs a new form of self-conditioning, and applies the diffusion process on the logit simplex space rather than the learned embedding space. Through extensive experiments on natural language understanding and generation tasks including summarization, text simplification, paraphrase generation, and question generation, we demonstrate that TESS outperforms state-of-the-art non-autoregressive models, requires fewer diffusion steps with minimal drop in performance, and is competitive with pretrained autoregressive sequence-to-sequence models. We publicly release our codebase at this https URL.

What Does it Take to Generalize SER Model Across Datasets? A Comprehensive Benchmark,

Adham Ibrahim, Shady Shehata, Ajinkya Kulkarni, Mukhtar Mohamed and Muhammad Abdul-Mageed,

Conference of the International Speech Communication Association (Interspeech), 2024

Speech emotion recognition (SER) is essential for enhanc ing human-computer interaction in speech-based applications. Despite improvements in specific emotional datasets, there is still a research gap in SER's capability to generalize across real world situations. In this paper, we investigate approaches to generalize the SER system across different emotion datasets. In particular, incorporate 11 emotional speech datasets and illus trate a comprehensive benchmark on the SER task. We also ad dress the challenge of imbalanced data distribution using over sampling methods when combining SER datasets for training. Furthermore, we explore various evaluation protocols for adept ness in the generalization of SER. Building on this, we explore the potential of Whisper for SER, emphasizing the importance of thorough evaluation. Our approach is designed to advance SER technology by integrating speaker-independent methods.

Inference to the Best Explanation in Large Language Models,

Dhairya Dalal, Marco Valentino, Andre Freitas and Paul Buitelaar,

62nd Annual Meeting of the Association for Computational Linguistics, 2024

While Large Language Models (LLMs) have found success in real-world applications, their underlying explanatory process is still poorly understood. This paper proposes IBE-Eval, a framework inspired by philosophical accounts on Inference to the Best Explanation (IBE) to advance the interpretation and evaluation of LLMs' explanations. IBE-Eval estimates the plausibility of natural language explanations through a combination of explicit logical and linguistic features including: consistency, parsimony, coherence, and uncertainty. Extensive experiments are conducted on Causal Question Answering (CQA), where IBE-Eval is tasked to select the most plausible causal explanation amongst competing ones generated by LLMs (i.e., GPT 3.5 and Llama 2). The experiments reveal that IBE-Eval can successfully identify the best explanation with up to 77% accuracy (≈ 27% above random), improving upon a GPT 3.5-as-a-Judge baseline (≈+17%) while being intrinsically more efficient and interpretable. Additional analyses suggest that, despite model-specific variances, LLM-generated explanations tend to conform to IBE criteria and that IBE-Eval is significantly correlated with human judgment, opening up opportunities for future development of automated explanation verification tools.

Verification and Refinement of Natural Language Explanations through LLM-Symbolic Theorem Proving,

Xin Quan, Marco Valentino, Louise A. Dennis and Andre Freitas,

Conference on Empirical Methods in Natural Language Processing, 2024

Natural language explanations represent a proxy for evaluating explanation-based and multi-step Natural Language Inference (NLI) models. However, assessing the validity of explanations for NLI is challenging as it typically involves the crowd-sourcing of apposite datasets, a process that is time-consuming and prone to logical errors. To address existing limitations, this paper investigates the verification and refinement of natural language explanations through the integration of Large Language Models (LLMs) and Theorem Provers (TPs). Specifically, we present a neuro-symbolic framework, named Explanation-Refiner, that integrates TPs with LLMs to generate and formalise explanatory sentences and suggest potential inference strategies for NLI. In turn, the TP is employed to provide formal guarantees on the logical validity of the explanations and to generate feedback for subsequent improvements. We demonstrate how Explanation-Refiner can be jointly used to evaluate explanatory reasoning, autoformalisation, and error correction mechanisms of state-of-the-art LLMs as well as to automatically enhance the quality of explanations of variable complexity in different domains.

Self-Refine Instruction-Tuning for Aligning Reasoning in Language Models,

Leonardo Ranaldi and Andre Freitas,

Conference on Empirical Methods in Natural Language Processing, 2024

The alignment of reasoning abilities between smaller and larger Language Models are largely conducted via supervised fine-tuning using demonstrations generated from robust Large Language Models (LLMs). Although these approaches deliver more performant models, they do not show sufficiently strong generaliza tion as the training only relies on the provided demonstrations. In this paper, we propose a self-refine Instruction-tuning method that allows for Smaller Language Models to self-improve their reasoning abilities. Our approach is based on a two-stage process, where reasoning abilities are f irst transferred between LLMs and Small Lan guage Models (SLMs) via Instruction-tuning onsynthetic demonstrations provided by LLMs, and then the instructed models self-improve through preference optimization strategies. In particular, the second phase operates refine ment heuristics based on Direct Preference Op timization, where the SLMs are prompted to deliver a series of reasoning paths by automat ically sampling the generated responses and providing rewards using ground truths from the LLMs. Results obtained on commonsense and math reasoning tasks show that this ap proach consistently outperforms Instruction tuning in both in-domain and out-domain scenarios, aligning the reasoning abilities of smaller and larger language models.

Aligning Large and Small Language Models via Chain-of-Thought Reasoning,

Leonardo Ranaldi and Andre Freitas,

18th Conference of the European Chapter of the Association for Computational Linguistics, 2024

Chain-of-Thought (CoT) prompting empowersthe reasoning abilities of Large Language Models (LLMs), eliciting them to solve complexreasoning tasks in a step-wise manner. However, these capabilities appear only in models with billions of parameters, which represent an entry barrier for many users who are constrained to operate on a smaller model scale, i.e., Small Language Models (SLMs). Although many companies are releasing LLMs of the same family with fewer parameters, these models tend not to preserve all the reasoning capabilities of the original models, including CoT reasoning.In this paper, we propose a method for aligning and transferring reasoning abilities between larger to smaller Language Models. By using an Instruction-tuning-CoT method, that is, an Instruction-tuning designed around CoT-Demonstrations, we enable the SLMs to generate multi-step controlled reasoned answers when they are elicited with the CoT mechanism. Hence, we instruct a smaller Language Model using outputs generated by more robust models belonging to the same family or not, evaluating the impact across different types of models. Results obtained on question-answering and mathematical reasoning benchmarks show that LMs instructed via the Instruction-tuning CoT method produced by LLMs outperform baselines within both in-domain and out-domain scenarios.

Learning Goal-oriented Bimanual Dough Rolling Using Dynamic Heterogeneous Graph Based on Human Demonstration,

Junjia Liu, Chenzui Li, Shixiong Wang, Zhipeng Dong, Sylvain Calinon, Miao Li and Fei Chen,

IEEE International Conference on Robotics and Biomimetics (ROBIO), 2024

Soft object manipulation poses significant challenges for robots, requiring effective techniques for state representation and manipulation policy learning. State representation involves capturing the dynamic changes in the environment, while manipulation policy learning focuses on establishing the relationship between robot actions and state transformations to achieve specific goals. To address these challenges, this research paper introduces a novel approach: a dynamic heterogeneous graph-based model for learning goal-oriented soft object manipulation policies. The proposed model utilizes graphs as a unified representation for both states and policy learning. By leveraging the dynamic graph, we can extract crucial information regarding object dynamics and manipulation policies. Furthermore, the model facilitates the integration of demonstrations, enabling guided policy learning. To evaluate the efficacy of our approach, we designed a dough rolling task and conducted experiments using both a differentiable simulator and a real-world humanoid robot. Additionally, several ablation studies were performed to analyze the effect of our method, demonstrating its superiority in achieving human-like behavior.

Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement,

Vivi Nastase, Chunyang Jiang, Giuseppe Samo and Paola Merlo,

Tenth Italian Conference on Computational Linguistics, 2024

In this paper, our goal is to investigate to what degree multilingual pretrained language models capture cross-linguistically valid abstract linguistic representations. We take the approach of developing curated synthetic data on a large scale, with specific properties, and using them to study sentence representations built using pretrained language models. We use a new multiple-choice task and datasets, Blackbird Language Matrices (BLMs), to focus on a specific grammatical structural phenomenon -- subject-verb agreement across a variety of sentence structures -- in several languages. Finding a solution to this task requires a system detecting complex linguistic patterns and paradigms in text representations. Using a two-level architecture that solves the problem in two steps -- detect syntactic objects and their properties in individual sentences, and find patterns across an input sequence of sentences -- we show that despite having been trained on multilingual texts in a consistent manner, multilingual pretrained language models have language-specific differences, and syntactic structure is not shared, even across closely related languages.

Exploring Italian sentence embeddings properties through multi-tasking,

Vivi Nastase, Giuseppe Samo, Chunyang Jiang and Paola Merlo,

Tenth Italian Conference on Computational Linguistics, 2024

We investigate to what degree existing LLMs encode abstract linguistic information in Italian in a multi-task setting. We exploit curated synthetic data on a large scale -- several Blackbird Language Matrices (BLMs) problems in Italian -- and use them to study how sentence representations built using pre-trained language models encode specific syntactic and semantic information. We use a two-level architecture to model separately a compression of the sentence embeddings into a representation that contains relevant information for a task, and a BLM task. We then investigate whether we can obtain compressed sentence representations that encode syntactic and semantic information relevant to several BLM tasks. While we expected that the sentence structure -- in terms of sequence of phrases/chunks -- and chunk properties could be shared across tasks, performance and error analysis show that the clues for the different tasks are encoded in different manners in the sentence embeddings, suggesting that abstract linguistic notions such as constituents or thematic roles does not seem to be present in the pretrained sentence embeddings.

BLM-It - Blackbird Language Matrices for Italian: A CALAMITA Challenge,

Chunyang Jiang, Giuseppe Samo, Vivi Nastase and Paola Merlo,

Tenth Italian Conference on Computational Linguistics, 2024

In this challenge, we propose Blackbird Language Matrices (BLMs), linguistic puzzles to learn language-related problems and investigate deeper formal and semantic properties of language, through a process of paradigm understanding. A BLM matrix consists of a context set and an answer set. The context is a sequence of sentences that encode implicitly an underlying generative linguistic rule. The contrastive multiple-choice answer set includes negative examples produced following corrupted generating rules. We propose three subtasks —agreement concord (Agr), causative (Caus) and object-drop (Od) alternation detection— each in two variants of increasing lexical complexity. The datasets comprise a few prompts for few-shot learning and a large test set.

Annotator-centric Active Learning for Subjective NLP Tasks,

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah and Enrico Liscio,

Conference on Empirical Methods in Natural Language Processing, 2024

Active Learning (AL) addresses the high costs of collecting human annotations by strategically annotating the most informative samples. However, for subjective NLP tasks, incorporating a wide range of perspectives in the annotation process is crucial to capture the variability in human judgments. We introduce Annotator-Centric Active Learning (ACAL), which incorporates an annotator selection strategy following data sampling. Our objective is two-fold: 1) to efficiently approximate the full diversity of human judgments, and 2) to assess model performance using annotator-centric metrics, which value minority and majority perspectives equally. We experiment with multiple annotator selection strategies across seven subjective NLP tasks, employing both traditional and novel, human-centered evaluation metrics. Our findings indicate that ACAL improves data efficiency and excels in annotator-centric performance evaluations. However, its success depends on the availability of a sufficiently large and diverse pool of annotators to sample from.

3. PhD Theses

Advancing Self-Supervised Deep Learning for 3D Scene Understanding,

Mohammad Mahdi Johari,

Ecole Polytechnique Fédérale de Lausanne, 2024

Recent advancements in deep learning have revolutionized 3D computer vision, enabling the extraction of intricate 3D information from 2D images and video sequences. This thesis explores the application of deep learning in three crucial challenges of 3D computer vision: Depth Estimation, Novel View Synthesis, and Simultaneous Localization and Mapping (SLAM). In the first part of the study, a self-supervised deep-learning method for depth estimation using a structured-light camera is proposed. Our method utilizes optical flow for improved edge preservation and reduced over-smoothing. In addition, we propose fusing depth maps from multiple video frames to enhance overall accuracy, particularly in occluded areas. Further, we demonstrate that these fused depth maps can be used for self-supervision to further improve the performance of a single-frame depth estimation network. Our models outperform state-of-the-art methods on both synthetic and real datasets. In the second part of the study, a generalizable photorealistic novel view synthesis method based on neural radiance fields (NeRF) is introduced. Our approach employs a geometry reasoner and a renderer to generate high-quality images from novel viewpoints. The geometry reasoner constructs cascaded cost volumes for each nearby source view, while the renderer utilizes a Transformer-based attention mechanism to integrate information from these cost volumes and render detailed images using volume rendering techniques. This architecture enables sophisticated occlusion reasoning and allows our method to render competitive results with per-scene optimized neural rendering methods while significantly reducing computational costs. Our experiments demonstrate superiority over state-of-the-art generalizable neural rendering models on various synthetic and real datasets. In the last part of the study, an efficient implicit neural representation method for dense visual SLAM is presented. The method reconstructs the scene representation while simultaneously estimating the camera position in a sequential manner from RGB-D frames with unknown poses. We incorporate recent advances in NeRF into the SLAM system, achieving both high accuracy and efficiency. The scene representation consists of multi-scale axis-aligned perpendicular feature planes and shallow decoders that decode the interpolated features into Truncated Signed Distance Field (TSDF) and RGB values. Extensive experiments on standard datasets demonstrate that our method outperforms state-of-the-art dense visual SLAM methods by more than 50% in 3D reconstruction and camera localization while running up to 10 times faster and eliminating the need for pre-training.

Safe Deep Neural Networks,

Kyle Matoba,

Ecole Polytechnique Fédérale de Lausanne, 2024

The capabilities of deep learning systems have advanced much faster than our ability to understand them. Whilst the gains from deep neural networks (DNNs) are significant, they are accompanied by a growing risk and gravity of a bad outcome. This is troubling because DNNs can perform well on a task most of the time, but can sometimes exhibit non-intuitive and nonsensical behavior for reasons that are not well understood. I begin this thesis arguing that closer alignment between human intuition and the operation of DNNs is massively beneficial. Next, I identify a class of DNNs that are particularly tractable and which play an important role in science and technology. Then I posit three dimensions on which alignment can be achieved–(1) philosophy: thought exercises to understand the fundamental considerations,(2) pedagogy: to help fallible humans interact effectively with neural networks, and (3) practice: methods to impose desired properties upon neural network, without degrading their performance. Then I present my work along these lines. Chapter 2 analyzes philosophically the issues of using penalty terms in criterion functions to avoid (negative) side effects via a three-way decomposition into the choice of (1) baseline, (2) deviation measure, and (3) scale of the penalty. Chapter 3 attempts to understand which inputs a DNN maps to an output class. I present two approaches to this problem, which can help users recognize unsafe behavior, even if they cannot formulate safety beforehand. Chapter 4 examines whether max pooling can be written as the composition of ReLU activations in order to investigate an open conjecture that max pooling is essentially redundant. These studies advance our pedagogical grasp of DNN modelling. Finally, Chapter 5 engages with practice by presenting a method for making DNNs more linear, and thereby more human-compatible.

Low-Resource Speech Recognition and Understanding for Challenging Applications,

Juan Pablo Zuluaga,

Ecole Polytechnique Fédérale de Lausanne, 2024

Automatic speech recognition (ASR) and spoken language understanding (SLU) is the core component of current voice-powered AI assistants such as Siri and Alexa. It involves speech transcription with ASR and its comprehension with natural language understanding (NLU) systems. Traditionally, SLU runs on a cascaded setting, where an in-domain ASR system automatically generates the transcripts with valuable semantic information, e.g., named entities and intents. These components have been generally based on statistical approaches with hand-crafted features. However, current trends have shifted towards large-scale end-to-end (E2E) deep neural networks (DNN), which have shown superior performance on a wide range of SLU tasks. For example, ASR has seen a rapid transition from traditional hybrid-based modeling to encoder-decoder and Transducer-based modeling. Even though there is an undeniable improvement in performance, other challenges have come into play, such as the urgency and need of large-scale supervised datasets; the need of additional modalities, such as contextual knowledge; massive GPU clusters for training large models; or high-performance and robust large models for complex applications. All of this leads to major challenges. This thesis explores solutions to these challenges that arise from complex settings. Specifically, we propose approaches: (1) to overcome the data scarcity on hybrid-based and E2E ASR models, i.e., low-resource applications; (2) for integration of contextual knowledge at decoding and training time, which leads to improved model quality; (3) to fast develop streaming ASR models from scratch for challenging domains without supervised data; (4) to reduce the computational budget required at training and inference time by proposing efficient alternatives w.r.t the state-of-the-art E2E architectures. Similarly, we explore solutions on the SLU domain, including analysis on the optimal representations to perform cascaded SLU, and other SLU tasks aside from intent and slot filing that can be performed in an E2E fashion. Finally, this thesis closes by covering STAC-ST and TokenVerse, two novel architectures that can handle ASR and SLU tasks seamlessly in a single model via special tokens.

A Stochastic Approach to Contact-rich Manipulation,

Julius Jankowski,

Ecole Polytechnique Fédérale de Lausanne, 2024

For robots to operate in unstructured environments, they are required to interact with objects through contact. Those contacts may be used to push objects to the side, deform objects, or manipulate objects in-hand. This thesis addresses the problem of controlling robots to exploit contacts to manipulate objects. Being able to anticipate the outcome of such physical interactions is essential for robots to gain true autonomy. However, contact interactions are particularly challenging to reason over in model- based control approaches due to the discontinuous nature of contacts. Moreover, interacting with objects the robot has not interacted with before will naturally lead to uncertainty in the prediction of contact dynamics. For instance, the robot can not anticipate the mass distribution of an object before making contact, which requires the robot to reason over possible outcomes before touching the object in a potentially unfavorable or unsafe way. Throughout this thesis, we formulate the problem of contact-rich manipulation with a robot manipulator as a model-predictive control problem. We explore stochas- tic optimization to plan for robot control trajectories in realtime. We show that the stochasticity in the optimization process enables the algorithm to explore the space of contacts without relying on local gradients or discretization of the contact space. We furthermore study how uncertainties in the physical properties of the object propagate through the contact dynamics and how the robot can actively reduce such uncertainty by exploiting favorable contact modes and sequences. We integrate the above contribu- tions into a planning and control framework for robots to manipulate objects through contacts in realtime. The framework is evaluated in a series of robot experiments, demonstrating robots autonomously performing dynamic hand-overs, push objects to a moving target, play air hockey, and manipulate objects robustly using two arms in the presence of uncertainty in the dynamics of the object.

Discovering Meaningful Units from Text Sequences,

Melika Behjati,

Ecole Polytechnique Fédérale de Lausanne, 2024

In recent years, the field of Natural Language Processing has seen significant revolution by the introduction of Transformers, a stack of multiple layers of attention and non-linearity, capable of performing almost any task and the backbone for large foundation models. In contrast to traditional NLP pipelines, this architecture is able to learn the features required to perform a specific task without any assumptions about the structure of language. The only remaining hard-coded aspect is the way the text input is fed to these models. This preprocessing step, known as the tokenization step, divides the input into chunks which could be as fine-grained as bytes or as coarse-grained as words. The most popular approach is to use subwords, such as Byte Pair Encodings or word pieces which lie between characters and words. However, it has been shown that hard-coding the input representations, has its own drawbacks. In particular, it would lead to sub-optimal performance in downstream tasks. In addition, to perform different tasks we need different levels of representations. In this thesis, we define and address the novel task of inducing units from text in an unsupervised manner. This work is a step towards completely end-to-end models which can decide which level of representation is the most suitable for them to perform a specific task. Our contributions are two-fold: First, we design models which are able to induce units without supervision at different levels. And second, since the task is novel, we need novel evaluations to show its effectiveness. Therefore, for every model we develop, we design and/or gather the set of tasks which evaluate and interpret the performance of our models. In the first chapter, we design a model to induce morpheme-like units from a sequence of characters. We adapt a method from object discovery in vision, called Slot Attention for our purpose. We propose to evaluate this model by introducing bi-directional probing evaluation. In the second chapter, we design a model which induces word-like units from a sequence of characters by integrating non-parametric variational information bottleneck in the last layers of a transformer encoder. In the next chapter, we move to the multi-modal domain and starting from subwords, we design a model which induces phrases from image captions by aligning them to the objects in the image. Lastly, we explore a task-driven approach towards inducing entities.

Robot Learning using Tensor Networks,

Suhan Shett,

Ecole Polytechnique Fédérale de Lausanne, 2024

In various robotics applications, the selection of function approximation methods greatly influences the feasibility and computational efficiency of algorithms. Tensor Networks (TNs), also referred to as tensor decomposition techniques, present a versatile approach for approximating functions involving continuous variables, discrete variables, or combinations of these variable types. Apart from their approximation capabilities, TNs offer efficient methods for conducting algebraic operations, calculus, probability modeling, and optimization, which are particularly essential in robotics applications. This thesis highlights the importance of a specific TN known as Tensor Train (TT) for function approximation in robotics by addressing a diverse range of previously challenging problems. Initially, utilizing TT, the thesis enhances the scalability and deployability of an ergodic exploration algorithm commonly employed in robotic exploration. Subsequently, the thesis introduces a novel numerical optimization algorithm named Tensor Train for Global Optimization (TTGO) to determine the optima of functions represented in TT format. Given that numerous robotics tasks are framed as numerical optimization problems, TTGO provides efficient solutions to several optimization-based problems in robotics, including inverse kinematics with obstacles, motion planning, and policy learning, as demonstrated in the thesis. In summary, this thesis underscores the promising potential of TNs as valuable tools in the field of robotics.

SUSTAINABLE AND RESILIENT SOCIETIES

1. JOURNAL PAPERS

EdgeFace : Efficient Face Recognition Model for Edge Devices,

Anjith George, Christophe Ecabert, Hatef Otroshi Shahreza, Ketan Kotwal and Sébastien Marcel,

IEEE Transactions on Biometrics, Behavior, and Identity Science, 2024

In this paper, we present EdgeFace- a lightweight and efficient face recognition network inspired by the hybrid architecture of EdgeNeXt. By effectively combining the strengths of both CNN and Transformer models, and a low rank linear layer, EdgeFace achieves excellent face recognition performance optimized for edge devices. The proposed EdgeFace network not only maintains low computational costs and compact storage, but also achieves high face recognition accuracy, making it suitable for deployment on edge devices. The proposed EdgeFace model achieved the top ranking among models with fewer than 2M parameters in the IJCB 2023 Efficient Face Recognition Competition. Extensive experiments on challenging benchmark face datasets demonstrate the effectiveness and efficiency of EdgeFace in comparison to state-of-the-art lightweight models and deep face recognition models. Our EdgeFace model with 1.77M parameters achieves state of the art results on LFW (99.73%), IJB-B (92.67%), and IJB-C (94.85%), outperforming other efficient models with larger computational complexities. The code to replicate the experiments will be made available publicly.

FRCSyn-onGoing: Benchmarking and Comprehensive Evaluation of Real and Synthetic Data to Improve Face Recognition Systems,

Alexander Unnervik, Anjith George, Hatef Otroshi Shahreza, Parsa Rahimi and Christophe Ecabert,

Information Fusion, 107:102322, 2024

This article presents FRCSyn-onGoing, an ongoing challenge for face recognition where researchers can easily benchmark their systems against the state of the art in an open common platform using large-scale public databases and standard experimental protocols. FRCSyn-onGoing is based on the Face Recognition Challenge in the Era of Synthetic Data (FRCSyn) organized at WACV 2024. This is the first face recognition international challenge aiming to explore the use of real and synthetic data independently, and also their fusion, in order to address existing limitations in the technology. Specifically, FRCSyn-onGoing targets concerns related to data privacy issues, demographic biases, generalization to unseen scenarios, and performance limitations in challenging scenarios, including significant age disparities between enrollment and testing, pose variations, and occlusions. To enhance face recognition performance, FRCSyn-onGoing strongly advocates for information fusion at various levels, starting from the input data, where a mix of real and synthetic domains is proposed for specific tasks of the challenge. Additionally, participating teams are allowed to fuse diverse networks within their proposed systems to improve the performance. In this article, we provide a comprehensive evaluation of the face recognition systems and results achieved so far in FRCSyn-onGoing. The results obtained in FRCSynonGoing, together with the proposed public ongoing benchmark, contribute significantly to the application of synthetic data to improve face recognition technology.

From Modalities to Styles: Rethinking the Domain Gap in Heterogeneous Face Recognition,

Anjith George and Sébastien Marcel,

IEEE Transactions on Biometrics, Behavior, and Identity Science, 2024

Heterogeneous Face Recognition (HFR) focuses on matching faces from different domains, for instance, thermal to visible images, making Face Recognition (FR) systems more versatile for challenging scenarios. However, the domain gap between these domains and the limited large-scale datasets in the target HFR modalities make it challenging to develop robust HFR models from scratch. In our work, we view different modalities as distinct styles and propose a method to modulate feature maps of the target modality to address the domain gap. We present a new Conditional Adaptive Instance Modulation (CAIM) module that seamlessly fits into existing FR networks, turning them into HFR-ready systems. The CAIM block modulates intermediate feature maps, efficiently adapting to the style of the source modality and bridging the domain gap. Our method enables end-to-end training using a small set of paired samples. We extensively evaluate the proposed approach on various challenging HFR benchmarks, showing that it outperforms state-of-the-art methods. The source code and protocols for reproducing the findings will be made publicly available.

Group Membership Verification via Nonlinear Sparsifying Transform Learning,

Behrooz Razeghi, Marzieh Gheisari, Amir Atashin, Dimche Kostadinov, Sébastien Marcel, Deniz Gunduz and Slava Voloshynovskiy,

IEEE Access, 2024

In today's digitally interconnected landscape, confirming the genuine associations between entities—whether they are items, devices, or individuals—and specific groups is critical. This paper introduces a new group membership verification method while ensuring minimal information loss, coupled with privacy-preservation and discrimination priors. Instead of verifying based on a similarity score in the original data space, we use a min-max functional measure in a transformed space. This method comprises two stages: (i) generating candidate nonlinear transform representations, and (ii) evaluating the min-max measure over these representations for both group assignment and transform selection. We simultaneously determine group membership and pick the appropriate representation from the candidate set based on the evaluation score. To solve within this framework, we employ an iterative alternating algorithm that both learns the parameters of candidate transforms and assigns group membership. Our method's efficacy is assessed on public datasets across various verification and identification scenarios and further tested on real-world image databases, CFP and LFW.

Template Inversion Attack Using Synthetic Face Images Against Real Face Recognition Systems,

Hatef Otroshi Shahreza and Sébastien Marcel,

IEEE Transactions on Biometrics, Behavior, and Identity Science, 2024

In this paper, we use synthetic data and propose a new method for template inversion attacks against face recognition systems. We use synthetic data to train a face reconstruction model to generate high-resolution (i.e., 1024×1024) face images from facial templates. To this end, we use a face generator network to generate synthetic face images and extract their facial templates using the face recognition model as our training set. Then, we use the synthesized dataset to learn a mapping from facial templates to the intermediate latent space of the same face generator network. We propose our method for both whitebox and blackbox TI attacks. Our experiments show that the trained model with synthetic data can be used to reconstruct face images from templates extracted from real face images. In our experiments, we compare our method with previous methods in the literature in attacks against different state-of-the-art face recognition models on four different face datasets, including the MOBIO, LFW, AgeDB, and IJB-C datasets, demonstrating the effectiveness of our proposed method on real face recognition datasets. Experimental results show our method outperforms previous methods on high-resolution 2D face reconstruction from facial templates and achieve competitive results with SOTA face reconstruction methods. Furthermore, we conduct practical presentation attacks using the generated face images in digital replay attacks against real face recognition systems, showing the vulnerability of face recognition systems to presentation attacks based on our TI attack (with synthetic train data) on real face datasets.

Vulnerability of State-of-the-Art Face Recognition Models to Template Inversion Attack,

Hatef Otroshi Shahreza, Vedrana Krivokuca and Sébastien Marcel,

IEEE Transactions on Information Forensics and Security, 2024

Face recognition systems use the templates (extracted from users' face images) stored in the system's database for recognition. In a template inversion attack, the adversary gains access to the stored templates and tries to enter the system using images reconstructed from those templates. In this paper, we propose a framework to evaluate the vulnerability of face recognition systems to template inversion attacks. We build our framework upon a real-world scenario and measure the vulnerability of the system in terms of the adversary's success attack rate in entering the system using the reconstructed face images. We propose a face reconstruction network based on a new block called "enhanced deconvolution using cascaded convolution and skip connections" (shortly, DSCasConv), and train it with a multi-term loss function. We use our framework to evaluate the vulnerability of state-of-the-art face recognition models, with different network structures and loss functions (in total 31 models), on the MOBIO, LFW, and AgeDB face datasets. Our experiments show that the reconstructed face images can be used to enter the system, which threatens the system's security. Additionally, the reconstructed face images may reveal important information about each user's identity, such as race, gender, and age, and hence jeopardize the users' privacy.

Why daylight should be a priority for urban planning,

Carlo Volf, Bruno Bueno, Peter Edwards, Richard Hobday, Stephan Mäder, Barbara S. Matusiak, Katharina Wulff, Werner Osterhaus, Gabriele Manoli, Christina Della Giustina, Jasmin Joshi, Jerome H. Kämpf, Kevin Vega, and Christoph Kueffer,

Journal of Urban Management, 2024

Daylight is essential for ecosystems and for the physical and mental well-being of people. In densely populated cities, only a small proportion of total daylight is available to support urban greenery and most people have little daily exposure to natural daylight. Despite this, many cities have followed a strategy of densification as a way of preventing urban sprawl and reducing energy consumption. In this article, we review the biological importance of daylight and show that urban densification leads to a reduction in the daylight available for both people and nature. We conclude that daylight in cities should be treated as a limiting resource that needs to be planned and managed carefully, much like water or energy. We suggest elements for a policy framework aimed at optimizing urban daylight, including how to determine daylight needs, how to determine the maximum viable urban density, and policy options for built and unbuilt areas.

PRIMIS: Privacy-Preserving Medical Image Sharing via Deep Sparsifying Transform Learning with Obfuscation,

Isaac Shiri, Behrooz Razeghi, Sohrab Ferdowsi, Yazdan Salimi, Deniz Gunduz, Douglas Teodoro, Slava Voloshynovskiy and Habib Zaidi,

Journal of Biomedical Informatics, 2024

The primary objective of our study is to address the challenge of confidentially sharing medical images across different centers. This is often a critical necessity in both clinical and research environments, yet restrictions typically exist due to privacy concerns. Our aim is to design a privacy-preserving data-sharing mechanism that allows medical images to be stored as encoded and obfuscated representations in the public domain without revealing any useful or recoverable content from the images. In tandem, we aim to provide authorized users with compact private keys that could be used to reconstruct the corresponding images. Method: Our approach involves utilizing a neural auto-encoder. The convolutional filter outputs are passed through sparsifying transformations to produce multiple compact codes. Each code is responsible for re constructing different attributes of the image. The key privacy-preserving element in this process is obfuscation through the use of specific pseudo-random noise. When applied to the codes, it becomes computationally infeasible for an attacker to guess the correct representation for all the codes, thereby preserving the privacy of the images. Results: The proposed framework was implemented and evaluated using chest X-ray images for different medical image analysis tasks, including classification, segmentation, and texture analysis. Additionally, we thoroughly assessed the robustness of our method against various attacks using both supervised and unsupervised algorithms. Conclusion: This study provides a novel, optimized, and privacy-assured data-sharing mechanism for medical images, enabling multi-party sharing in a secure manner. While we have demonstrated its effectiveness with chest X-ray images, the mechanism can be utilized in other medical images modalities as well.

Inference from Real-World Sparse Measurements,

Arnaud Pannatier, Kyle Matoba and Francois Fleuret,

Transactions on Machine Learning Research, 2024

Real-world problems often involve complex and unstructured sets of measurements, which occur when sensors are sparsely placed in either space or time. Being able to model this irregular spatiotemporal data and extract meaningful forecasts is crucial. Deep learning architectures capable of processing sets of measurements with positions varying from set to set, and extracting readouts anywhere are methodologically difficult. Current state-of-the-art models are graph neural networks and require domain-specific knowledge for proper setup. We propose an attention-based model focused on robustness and practical applicability, with two key design contributions. First, we adopt a ViT-like transformer that takes both context points and read-out positions as inputs, eliminating the need for an encoder-decoder structure. Second, we use a unified method for encoding both context and read-out positions. This approach is intentionally straightforward and integrates well with other systems. Compared to existing approaches, our model is simpler, requires less specialized knowledge, and does not suffer from a problematic bottleneck effect, all of which contribute to superior performance. We conduct in-depth ablation studies that characterize this problematic bottleneck in the latent representations of alternative models that inhibit information utilization and impede training efficiency. We also perform experiments across various problem domains, including high-altitude wind nowcasting, two-day weather forecasting, fluid dynamics, and heat diffusion. Our attention-based model consistently outperforms state-of-the-art models in handling irregularly sampled data. Notably, our model reduces the root mean square error (RMSE) for wind nowcasting from 9.24 to 7.98 and for heat diffusion tasks from 0.126 to 0.084.

Verification of an open-source Python library for the simulation of district heating networks with complex topologies,

Roberto Boghetti and Jérôme H. Kämpf,

Energy, Volume 290, 2024

District Heating Networks are seen as a popular solution for the decarbonation of space and domestic hot water heating by the use of renewables or waste heat. Their design and operation can be enhanced both environmentally and economically through the use of dedicated simulation tools.
This work presents a new multicomponent Python-based software library including an efficient steady-state model for the hydraulic simulation and a Lagrange-based dynamic model for the thermal part. The model is verified on real data from a meshed DHN with a granularity of 15 min. For the considered period of 6 days, the model estimated the return temperatures of two heat plants with a mean error of 0.08 K and 0.43 K and standard deviations of 0.56 K and 0.71 K respectively. On the hydraulic part, after applying a fixed correction factor, the error on the pressure differences was estimated in 3 peripheral locations of the network. In the worst case, the relative error distribution had a mean of 0.38% and a standard deviation of 8.74%. Finally, the implemented hydraulic solver achieved a speed improvement over the standard loop method of 17.5%.

Verification of an open-source Python library for the simulation of district heating networks with complex topologies,

Roberto Boghetti and Jérôme H. Kämpf,

Energy, Volume 290, 2024

District Heating Networks are seen as a popular solution for the decarbonation of space and domestic hot water heating by the use of renewables or waste heat. Their design and operation can be enhanced both environmentally and economically through the use of dedicated simulation tools.
This work presents a new multicomponent Python-based software library including an efficient steady-state model for the hydraulic simulation and a Lagrange-based dynamic model for the thermal part. The model is verified on real data from a meshed DHN with a granularity of 15 min. For the considered period of 6 days, the model estimated the return temperatures of two heat plants with a mean error of 0.08 K and 0.43 K and standard deviations of 0.56 K and 0.71 K respectively. On the hydraulic part, after applying a fixed correction factor, the error on the pressure differences was estimated in 3 peripheral locations of the network. In the worst case, the relative error distribution had a mean of 0.38% and a standard deviation of 8.74%. Finally, the implemented hydraulic solver achieved a speed improvement over the standard loop method of 17.5%.

ATCO2 corpus: A Large-Scale Dataset for Research on Automatic Speech Recognition and Natural Language Understanding of Air Traffic Control Communications,

Juan Zuluaga-Gomez, Karel Veselý, Igor Szöke, Alexander Blatt, Petr Motlicek, Martin Kocour, Mickael Rigault, Khalid Choukri, Amrutha Prasad, Seyyed Saeed Sarfjoo, Iuliia Nigmatulina, Claudia Cevenini, Pavel Kolčárek, Allan Tart, Jan Černocký, and Dietrich Klakow,

Journal of Data-centric Machine Learning Research, 2024

Personal assistants, automatic speech recognizers and dialogue understanding systems are becoming more critical in our interconnected digital world. A clear example is air traffic control (ATC) communications. ATC aims at guiding aircraft and controlling the airspace in a safe and optimal manner. These voice-based dialogues are carried between an air traffic controller (ATCO) and pilots via very-high frequency radio channels. In order to incorporate these novel technologies into ATC, large-scale annotated datasets are required to develop the data-driven AI systems. Two examples are automatic speech recognition (ASR) and natural language understanding (NLU). However, ATC is considered a low-resource domain. In this paper, we introduce the ATCO2 corpus, a dataset that aims at fostering research on the challenging ATC field, which has lagged behind due to lack of annotated data. In addition, we also open-source a GitHub repository that contains data preparation and training scripts useful to replicate our baselines related to ASR and NLU. The ATCO2 corpus covers 1) audio and radar data collection and pre-processing, 2) pseudo-transcriptions of speech audio, and 3) extraction of ATC-related named entities. The ATCO2 corpus is split into three subsets: (i) ATCO2-test-set corpus contains 4 hours of ATC speech with manual transcripts and a subset with gold transcriptions for named-entity recognition (callsign, command, value) and speaker role detection. (ii) The ATCO2-test-set-1h corpus is a one-hour open-sourced subset from the 4h test set.\footnote{Free to download, available at: https://www.atco2.org/data. (iii) The ATCO2-PL-set corpus consists of 5'281 hours of pseudo-transcribed ATC speech enriched with contextual information (list of relevant n-gram sequences per utterance), speaker turn information, signal-to-noise ratio estimate and English language detection score per sample. The whole ATCO2 corpus is publicly distributed through ELDA catalog (https://catalog.elra.info/en-us/repository/browse/ELRA-S0484/). We expect the corpus will foster research on robust ASR and NLU not only in the field of ATC communications but also in the general research community.

Heterogeneous Face Recognition with Prepended Domain Transformers,

Anjith George and Sebastien Marcel,

Springer Nature, 2024

Face recognition (FR) has become a very popular method for biomet ric authentication thanks to its non-contact nature and high accuracy. State-of-the-art face recognition systems are obtaining human parity even in unconstrained scenarios, thanks to the deep neural network architectures and the large datasets available for training them. How ever, there are several other types of face imaging modalities such as infrared, thermal, depth and so on which can boost the performance of FR systems even further. The main challenge in using these new modalities is the lack of availability of large-scale labeled datasets to train FR models. Heterogeneous face recognition provides a solution to this issue by leveraging the large sets of training data available for visible spectrum data. Heterogeneous Face Recognition (HFR) involves matching facial images from different domains, such as thermal to visi ble images (VIS), sketches to visible images, and near-infrared to visible images. This process is especially beneficial for aligning visible spec trum images with those from other modalities. However, HFR poses significant challenges due to the domain gap between the source and target images, compounded by the lack of large-scale, paired heteroge neous face image datasets for training HFR models. In this chapter, we introduce a lightweight and effective method for cross-modality face image matching. The core idea in our approach is to integrate a neu ral network component, known as the prepended domain transformer (PDT), at the beginning of an existing face recognition (FR) model to bridge the domain gap. By retraining this prepended module with a small number of paired samples in a contrastive learning framework, we achieved state-of-the-art results in various HFR benchmarks. The PDT blocks are versatile and can be retrained for different source-target com binations using our framework. This approach is compatible with any pre-trained FR model due to its architecture-agnostic nature. Addition ally, its modular design allows for training with a limited set of paired samples, making it easier for integration and deployment. The source code and protocols for reproducing the results are publicly available.

Knowledge Distillation for Face Recognition using Synthetic Data with Dynamic Latent Sampling,

Hatef Otroshi, Anjith George and Sébastien Marcel,

IEEE Access, 2024

State-of-the-art face recognition models are computationally expensive for mobile applications. Training lightweight face recognition models also requires large identity-labeled datasets, raising privacy and ethical concerns. Generating synthetic datasets for training is also challenging, and there is a significant gap in performance between models trained on real and synthetic face datasets. We propose a new framework (called SynthDistill) to train lightweight face recognition models by distilling the knowledge from a pretrained teacher model using synthetic data. We generate synthetic face images without identity labels, mitigating the problems in the intra-class variation generation of synthetic datasets, and dynamically sample from the intermediate latent space of a face generator network to generate new variations of the challenging images while further exploring new face images. The results on different benchmarking real face recognition datasets demonstrate the superiority of SynthDistill compared to training on previous synthetic datasets, achieving a verification accuracy of 99.52% on the LFW dataset with a lightweight network. The results also show that SynthDistill significantly narrows the gap between real and synthetic data training. The source code of our experiments is publicly available to facilitate the reproducibility of our work.

2. Conference Papers

A Novel and Responsible Dataset for Face Presentation Attack Detection on Mobile Devices,

Nathan Ramoly, Alain Komaty, Vedrana Krivokuca, Lara Younes, Ahmad-Montaser Awal and Sébastien Marcel,

IEEE International Joint Conference on Biometrics, 2024

Presentation Attack Detection (PAD) is essential for ensuring the security of face recognition (FR) systems, particularly in the context of mobile authentication in various sectors, such as online banking and government services. However, current PAD methods are often sensitive to the data domain, partly due to the limitations of training PAD datasets. In this paper, we introduce the SOTERIA dataset, which provides captures of bona-fide and diverse Presentation Attacks (PAs) recorded using smart phones. The dataset was collected responsibly from 70 consenting individuals, as opposed to web scraping. It includes face videos, motion data, and depth information (when available) as well as a novel projector-based replay attack. To demonstrate the utility of the SOTERIA dataset, we evaluate the vulnerability of a SOTA FR model (IRes Net100) to the PAs in the dataset. We also analyze the PAD capabilities of a SOTA PAD model (DeepPixBis) through cross-dataset experiments as well as on real attacks observed in an industrial application. Our findings show the effectiveness and versatility of the SOTERIA dataset in advancing PAD research, in particular toward generalization.

Assessing the Reliability of Biometric Authentication on Virtual Reality Devices,

Ketan Kotwal, Gokhan Ozbulak and Sébastien Marcel,

IEEE International Joint Conference on Biometrics, 2024

Recent developments in Virtual Reality (VR) headsets have unlocked a plethora of innovative use-cases, many of which were previously unimaginable. However, as these use-cases, such as personalized immersive experiences, necessitate user authentication, ensuring robustness and resistance to spoofing attacks becomes imperative. The absence of appropriate dataset has constrained our understanding and assessment of VR devices' susceptibility to presentation attacks. To address this research gap, we introduce VRBiom: a new periocular video dataset acquired from a VR headset (Meta Quest Pro), comprising 900 genuine and 1104 presentation attack videos, each spanning 10 seconds. The bona-fide videos consist of variations in terms of gaze and glasses; while the attacks are constructed with 6 different types of instruments. Additionally, we evaluate the performance of two prominent CNN architectures trained using various configurations for detecting presentation attacks in the newly created VRBiom dataset. Our benchmarking on VRBiom reveals the presence of spoofing threats in VR headsets. While baseline models exhibit considerable efficacy in attack detection, substantial scope exists for improvement in detecting attacks on periocular videos. Our dataset will be a useful resource for researchers aiming to enhance the security and reliability of VR-based authentication systems.

Breaking Template Protection: Reconstruction of Face Images from Protected Facial Templates,

Hatef Otroshi Shahreza and Sébastien Marcel,

18th International Conference on Automatic Face and Gesture Recognition (FG), 2024

Face recognition systems tend toward ubiquity and are commonly utilized for security purposes. These systems operate based on facial representations, called templates, extracted by a deep neural network from each face image. However, it has been shown that face recognition templates can be inverted to reconstruct underlying face images, posing new security and privacy threats to face recognition systems. To mitigate such attacks against face recognition systems, several biometric template protection schemes have been proposed in the literature. The ISO/IEC 24745 standard requires each biometric template protection scheme to fulfill several requirements, among which non-invertibility is of the utmost importance. Therefore, each of the proposed template protection schemes in the literature used an ad-hoc approach to investigate the invertibility of the protected templates. In this paper, we consider a scenario where an adversary gains knowledge of a template protection scheme as well as its secrets, and tries to reconstruct a face image using a leaked protected template. We consider different template protection schemes, including Bio Hashing, MLP-Hashing, and Homomorphic Encryption (HE), and reconstruct face images from protected templates. We also use different state-of-the-art face recognition models in both whitebox and blackbox scenarios. To our knowledge, this is the first work on learning-based reconstruction of face images from protected facial templates.

ChatGPT and biometrics: an assessment of face recognition, gender detection, and age estimation capabilities,

Ahmad Hassanpour, Yasamin Kowsari, Hatef Otroshi Shahreza, Bian Yang and Sébastien Marcel,

IEEE International Conference on Image Processing (ICIP), 2024

This paper explores the application of large language models (LLMs), like ChatGPT, for biometric tasks. We specifically examine the capabilities of ChatGPT in performing biometric-related tasks, with an emphasis on face recognition, gender detection, and age estimation. Since biometrics are considered as sensitive information, ChatGPT avoids answering direct prompts, and thus we crafted a prompting strategy to bypass its safeguard and evaluate the capabilities for biometrics tasks. Our study reveals that ChatGPT recognizes facial identities and differentiates between two facial images with considerable accuracy. Additionally, experimental results demonstrate remarkable performance in gender detection and reasonable accuracy for the age estimation tasks. Our findings shed light on the promising potentials in the application of LLMs and foundation models for biometrics.

Deep Variational Privacy Funnel: General Modeling with Applications in Face Recognition,

Behrooz Razeghi, Parsa Rahimi and Sébastien Marcel,

49th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024

In this study, we harness the information-theoretic Privacy Funnel (PF) model to develop a method for privacy-preserving representation learning using an end-to-end training framework. We rigorously address the trade-off between obfuscation and utility. Both are quantified through the logarithmic loss, a measure also recognized as self-information loss. This exploration deepens the interplay between information-theoretic privacy and representation learning, offering substantive insights into data protection mechanisms for both discriminative and generative models. Importantly, we apply our model to state-of-the-art face recognition systems. The model demonstrates adaptability across diverse inputs, from raw facial images to both derived or refined embeddings, and is competent in tasks such as classification, reconstruction, and generation.

Latent Enhancing AutoEncoder for Occluded Image Classification,

Ketan Kotwal, Tanay Deshmukh and Preeti Gopal,

IEEE International Conference on Image Processing (ICIP), 2024

Large occlusions result in a significant decline in image classification accuracy. During inference, diverse types of unseen occlusions introduce out-of-distribution data to the classification model, leading to accuracy dropping as low as 50%. As occlusions encompass spatially connected regions, conventional methods involving feature reconstruction are inadequate for enhancing classification performance. We introduce LEARN: Latent Enhancing feAture Reconstruction Network– An auto encoder based network that can be incorporated into the classification model before its classifier head without modifying the weights of classification model. In addition to reconstruction and classification losses, training of LEARN effectively combines intra- and inter-class losses calculated over its latent space—which lead to improvement in recovering latent space of occluded data, while preserving its class-specific discriminative information. On the OccludedPASCAL3D+ dataset, the proposed LEARN outperforms standard classification models (VGG16 and ResNet-50) by a large margin and up to 2% over state-of-the-art methods. In cross-dataset testing, our method improves the average classification accuracy by more than 5% over the state-of-the-art methods. In every experiment, our model consistently maintains excellent accuracy on in-distribution data.

Entity Matching Across Small Networks Using Node Attributes,

Zahra Ahmadi, Zijian Zhang, Hoang H. Nguyen, Sergio Burdisso, Srikanth Madikeri, Petr Motlicek, Erinc Dikici, Gerhard Backfried, Marek Kovac and Daniel Kudenko,

27th European Conference on Artificial Intelligence (ECAI), Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS), 2024

Entity matching, also known as user identity linkage, is a critical task in data integration. While established techniques primarily focus on large-scale networks, there are several applications where small networks pose challenges due to limited training data and sparsity. This study addresses entity matching in the field of criminology, where small networks are common and the number of known matching nodes is restricted. To support this research, we exploit a multimodal dataset, collected as part of a security-related project, consisting of an intercepted telephone calls network (i.e., ROXSD data) and a network of social forum interactions (i.e., ROXHOOD data) collected in a simulated environment, although following real investigation scenario. To improve accuracy and efficiency, we propose a novel approach for entity matching across these two small networks using node attributes. Existing techniques often merely focus on topology consistency between two networks and overlook valuable information, such as network node attributes, making them vulnerable to structural changes. Inspired by the remarkable success of deep learning, we present UGC-DeepLink, an end-to-end semi-supervised learning framework that leverages user-generated content. UGC-DeepLink encodes network nodes into vector representations, capturing both local and global network structures to align anchor nodes using deep neural networks. A dual learning paradigm and the policy gradient method transfer knowledge and update the linkage. Additionally, node attributes, such as call contents and forum exchanged texts, enhance the ranking of matching nodes. Experimental results on ROXSD and ROXHOOD demonstrate that UGC-DeepLink surpasses baselines and state-of-the-art methods in terms of identity-match ranking.

Face Recognition Using Lensless Camera,

Hatef Otroshi Shahreza, Alexandre Veuthey and Sébastien Marcel,

49th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024

Coded aperture imaging is an emerging technique allowing thin form factor cameras that can be cheaply constructed. Many applications benefit from using such lensless cameras, such as face recognition. We propose a method for face recognition using coded aperture images that does not require retraining any component of the face recognition pipeline, but instead applies post-processing to the images with deep learning refinement so that they are compatible with existing face recognition for RGB images. We generate training data with a simulation process, based on the convolutional model of a lensless camera, and train a neural network to reconstruct face images. We train our network with a multi-term loss function to refine identity information in the reconstructed face image. We provide extensive experiments on different face recognition datasets, including LFW, CA-LFW, CP-LFW, AgeDB, FERET, and FRGC, showing the effectiveness and generalization of our proposed method. Our source code will be made available publicly to facilitate the reproducibility of our work.

Face Reconstruction from Partially Leaked Facial Embeddings,

Hatef Otroshi Shahreza and Sébastien Marcel,

49th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024

Face recognition systems are widely used in different applications. In such systems, some features (called templates) are extracted from each face image and stored in the system's database. In this paper, we propose an attack against face recognition systems where the adversary gains access to a portion of facial templates and aims to reconstruct the underlying face image. To this end, we train a face reconstruction network to invert partially leaked templates. In our experiments, we evaluate the vulnerability of state-of-the-art face recognition systems on different datasets, including MOBIO, LFW, and AgeDB. Our experiments demonstrate the vulnerability of face recognition systems to template inversion based on a portion of leaked templates. For example, with only 20% of facial templates, our experiments show that an adversary can achieve a success attack rate of 87% on a system based on ArcFace on the LFWdataset configured at the false match rate of 0.1%. To our knowledge, this paper is the first work on the inversion of partially leaked facial templates, and paves the way for future studies of attacks against face recognition systems based on partially leaked templates.

Heterogeneous Face Recognition Using Domain Invariant Units,

Anjith George and Sébastien Marcel,

49th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024

Heterogeneous Face Recognition (HFR) aims to expand the applicability of Face Recognition (FR) systems to challenging scenarios, enabling the matching of face images across different domains, such as matching thermal images to visible spectra. However, the development of HFR systems is challenging because of the significant domain gap between modalities and the lack of availability of large-scale paired multi-channel data. In this work, we leverage a pre-trained face recognition model as a teacher network to learn domain invariant network layers called Domain-Invariant Units (DIU) to reduce the domain gap. The proposed DIU can be trained effectively even with a limited amount of paired training data, in a contrastive distillation framework. This proposed approach has the potential to enhance pre-trained models, making them more adaptable to a wider range of variations in data. We extensively evaluate our approach on multiple challenging benchmarks, demonstrating superior performance compared to state-of-the-art methods.

Mapping the Media Landscape: Predicting Factual Reporting and Political Bias Through Web Interactions,

Dairazalia Sanchez-Cortes, Sergio Burdisso, Esaú Villatoro-Tello and Petr Motlicek,

14th International Conference of the CLEF Association, 2024

Bias assessment of news sources is paramount for professionals, organizations, and researchers who rely on truthful evidence for information gathering and reporting. While certain bias indicators are discernible from content analysis, descriptors like political bias and fake news pose greater challenges. In this paper, we propose an extension to a recently presented news media reliability estimation method that focuses on modeling outlets and their longitudinal web interactions. Concretely, we assess the classification performance of four reinforcement learning strategies on a large news media hyperlink graph. Our experiments, targeting two challenging bias descriptors, factual reporting and political bias, showed a significant performance improvement at the source media level. Additionally, we validate our methods on the CLEF 2023 CheckThat! Lab challenge, outperforming the reported results in both, F1-score and the official MAE metric. Furthermore, we contribute by releasing the largest annotated dataset of news source media, categorized with factual reporting and political bias labels. Our findings suggest that profiling news media sources based on their hyperlink interactions over time is feasible, offering a bird's-eye view of evolving media landscapes.

Mitigating Demographic Bias in Face Recognition via Regularized Score Calibration,

Ketan Kotwal and Sébastien Marcel,

IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, 2024

Demographic bias in deep learning-based face recognition systems has led to serious concerns. Several existing works attempt to mitigate bias by incorporating demographic-specific processing during inference, which requires knowledge or learning of demographic attribute with an additional cost. We propose to regularize training of the face recognition CNN, for demographic fairness, by imposing constraints on the distributions of matching scores. Our regularization term enforces the score distributions from different demographic groups to respect a predefined probability distribution, as well as it penalizes misalignment of distributions across demographic groups. The proposed method improves fairness of face recognition models without compromising the recognition accuracy, and does not require extra resources during inference. Our experiments indicate that in a cross-dataset testing, the regularized CNN can reduce the variation in accuracies (i.e., more fairness) of different demographic groups up to 25% while slightly improving recognition accuracy over baselines.

Modality Agnostic Heterogeneous Face Recognition with Switch Style Modulators,

Anjith George and Sébastien Marcel,

IEEE International Joint Conference on Biometrics, 2024

Heterogeneous Face Recognition (HFR) systems aim to enhance the capability of face recognition in challenging cross-modal authentication scenarios. However, the significant domain gap between the source and target modalities poses a considerable challenge for cross-domain matching. Existing literature primarily focuses on developing HFR approaches for specific pairs of face modalities, necessitating the explicit training of models for each source-target combination. In this work, we introduce a novel framework designed to train a modality-agnostic HFR method capable of handling multiple modalities during inference, all without explicit knowledge of the target modality labels. We achieve this by implementing a computationally efficient automatic routing mechanism called Switch Style Modulation Blocks (SSMB) that trains various domain expert modulators which transform the feature maps adaptively reducing the domain gap. Our proposed SSMB can be trained end-to end and seamlessly integrated into pre-trained face recognition models, transforming them into modality-agnostic HFR models. We have performed extensive evaluations on HFR benchmark datasets to demonstrate its effectiveness. The source code and protocols will be made publicly available.

ProGAP: Progressive Graph Neural Networks with Differential Privacy Guarantees,

Sina Sajadmanesh and Daniel Gatica-Perez,

17th ACM International Conference on Web Search and Data Mining, 2024

Graph Neural Networks (GNNs) have become a popular tool for learning on graphs, but their widespread use raises privacy concerns as graph data can contain personal or sensitive information. Differentially private GNN models have been recently proposed to preserve privacy while still allowing for effective learning over graph-structured datasets. However, achieving an ideal balance between accuracy and privacy in GNNs remains challenging due to the intrinsic structural connectivity of graphs. In this paper, we propose a new differentially private GNN called ProGAP that uses a progressive training scheme to improve such accuracy-privacy trade-offs. Combined with the aggregation perturbation technique to ensure differential privacy, ProGAP splits a GNN into a sequence of overlapping submodels that are trained progressively, expanding from the first submodel to the complete model. Specifically, each submodel is trained over the privately aggregated node embed dings learned and cached by the previous submodels, leading to an increased expressive power compared to previous approaches while limiting the incurred privacy costs. We formally prove that ProGAP ensures edge-level and node-level privacy guarantees for both training and inference stages, and evaluate its performance on benchmark graph datasets. Experimental results demonstrate that ProGAP can achieve up to 5-10% higher accuracy than existing state-of-the-art differentially private GNNs.

Reliability Estimation of News Media Sources: Birds of a Feather Flock Together,

Sergio Burdisso, Dairazalia Sanchez-Cortes, Esaú Villatoro-Tello and Petr Motlicek,

Conference of the North American Chapter of the Association for Computational Linguistics, 2024

Evaluating the reliability of news sources is a routine task for journalists and organizations committed to acquiring and disseminating accurate information. Recent research has shown that predicting sources' reliability represents an important first-prior step in addressing additional challenges such as fake news detection and fact-checking. In this paper, we introduce a novel approach for source reliability estimation that leverages reinforcement learning strategies for estimating the reliability degree of news sources. Contrary to previous research, our proposed approach models the problem as the estimation of a reliability degree, and not a reliability label, based on how all the news media sources interact with each other on the Web. We validated the effectiveness of our method on a news media reliability dataset that is an order of magnitude larger than comparable existing datasets. Results show that the estimated reliability degrees strongly correlates with journalists-provided scores (Spearman=0.80) and can effectively predict reliability labels (macro-avg. F1 score=81.05). We release our implementation and dataset, aiming to provide a valuable resource for the NLP community working on information verification.

ROXSD: The ROXANNE Multimodal and Simulated Dataset for Advancing Criminal Investigations,

Petr Motlicek, Erinç Dikici, Srikanth Madikeri, Pradeep Rangappa, Gerhard Backfried, Johan A. Rohdin, Petr Schwarz, Marek Kovac, Kvetoslav Maly, Dominik Bobos, Dietrich Klakow, Eleni Konstantina Sergidou et al.

Odyssey 2024: The Speaker and Language Recognition Workshop, 2024

The ROXANNE project, conducted under the European Union's Horizon 2020 Programme, aimed to revolutionize criminal investigations by integrating speech, language, and video technologies with criminal network analysis. Despite the success in technology development, the project faced evaluation challenges due to the scarcity and legal restrictions surrounding real-world criminal activity datasets. In response, we intro duce ROXSD, a simulated dataset of communication in organized crime. ROXSD is a set of wiretapped conversations (collected through communication service providers) between drug dealing suspects, following a realistic screenplay (incl. realis tic conditions and constraints of a real investigation) prepared by Law Enforcement Agencies (LEAs). With a focus on multimodality and multilinguality, the dataset comprises 20 hours of telephone and video conversations involving 104 speakers, and is further aligned with ground-truth annotations for each modality involved, enabling precise evaluation and development of technologies. In addition, the multimodal data are enhanced with metadata and prior knowledge (e.g., suspects' biometric profiles) which is typically available as a result of lawfully intercepted communication. This paper introduces ROXSD as a pivotal resource for advancing technology in criminal research (specifically in domain of speech, text and network analysis). ROXSD not only facilitates in the domain of technology development and evaluation but also showcases the potential of simulated datasets in advancing the field of organized crime analytics, emphasizing the importance of such datasets in the absence of comprehensive real-world alternatives.

SDFR: Synthetic Data for Face Recognition Competition,

Hatef Otroshi Shahreza, Christophe Ecabert, Anjith George, Alexander Unnervik and Sébastien Marcel,

IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), 2024

Large-scale face recognition datasets are collected by crawling the Internet and without individuals' consent, raising legal, ethical, and privacy concerns. With the recent advances in generative models, recently several works proposed generating synthetic face recognition datasets to mitigate concerns in web-crawled face recognition datasets. This paper presents the summary of the Synthetic Data for Face Recognition (SDFR) Competition held in conjunction with the 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024) and established to investigate the use of synthetic data for training face recognition models. The SDFR competition was split into two tasks, allowing participants to train face recognition systems using new synthetic datasets and/or existing ones. In the first task, the face recognition back bone was fixed and the dataset size was limited, while the second task provided almost complete freedom on the model backbone, the dataset, and the training pipeline. The submitted models were trained on existing and also new synthetic datasets and used clever methods to improve training with synthetic data. The submissions were evaluated and ranked on a diverse set of seven benchmarking datasets. The paper gives an overview of the submitted face recognition models and reports achieved performance compared to baseline models trained on real and synthetic datasets. Furthermore, the evaluation of submissions is extended to bias assessment across different demography groups. Lastly, an outlook on the current state of the research in training face recognition models using synthetic data is presented, and existing problems as well as potential future directions are also discussed.

Vascular Biometrics Experiments on Candy -- A New Contactless Finger-Vein Dataset,

Sushil Bhattacharjee, David Geissbuhler, Guillaume Clivaz, Ketan Kotwal and Sébastien Marcel,

International Conference on Pattern Recognition (ICPR), 2024

The topic of finger-vein (FV) biometrics is an active and growing topic of research. Most FV systems available today rely on contact sensors that capture vein patterns of a single finger at a time. We have recently completed a project aimed at designing a contactless vein sensing platform, named sweet. In this paper we present a new FV dataset collected using sweet. The dataset includes multiple FV samples from 120 subjects, and 280 presentation attack instruments (PAI), captured in a contactless manner. Further, we present baseline FV authentication (FVA) results achieved for proposed dataset. The sweet platform is equipped to capture a sequence of images suitable for photometric-stereo (PS) reconstruction of 3D surfaces. We present a FV presentation attack detection (PAD) method based on PS reconstruction, and the corresponding baseline FV PAD results on the proposed dataset.

Vulnerability of Face Age Verification to Replay Attacks,

Pavel Korshunov, Anjith George, Gokhan Ozbulak and Sébastien Marcel,

49th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024

Presentation attacks on biometric systems have long created significant security risks. The increase in the adoption of age verification systems, which ensure that only age-appropriate content is consumed online, raises the question of vulnerability of such systems to replay presentation attacks. In this paper, we analyze the vulnerability of face age verification to simple replay attacks and assess whether presentation attack detection (PAD) systems created for biometrics can be effective at detecting similar attacks on age verification. We used three types of attacks captured with iPhone 12, Galaxy S9, and Huawei Mate 30 phones from iPad Pro, which replayed the images from a commonly used UTKFace dataset of faces with true age labels. We evaluated four state of the art face age verification algorithms, including simple classification, distribution-based, regression via classification, and adaptive distribution approaches. We show that these algorithms are vulnerable to the attacks, since the accuracy of age verification on replayed images is only a couple of percentage points different compared to when the original images are used, which means an age verification system cannot distinguish attacks from bona fide images. Using two state of the art presentation attack detection systems, DeepPixBiS and CDCN, trained to detect similar attacks on biometrics, we demonstrate that they struggle to detect both: the types of attacks that are possible in age verification scenario and the type of bona fide images that are commonly used. These results highlight the need for the development of age verification specific attack detection systems for age verification to become practical.

GLoFool: global enhancements and local perturbations to craft adversarial images,

Mirko Agarla and Andrea Cavallaro,

18th European Conference on Computer Vision (ECCV) Workshops, 2024

Adversarial examples crafted in black-box scenarios are affected by unrealistic colors or spatial artifacts. To prevent these short-comings, we propose a novel strategy that generates adversarial images with low detectability and high transferability. The proposed black-box strategy, GLoFool, introduces global and local perturbations iteratively. First, a combination of image enhancement filters is applied globally to the clean image. Then, local color perturbations are generated on segmented image regions. These local perturbations are dynamically increased for each region over the iterations by sampling new colors on an expanding disc around the initial global enhancement. We propose a version of the method optimized for quality, GLoFool-Q, and one for transferability, GLoFool-T. Compared to state-of-the-art attacks that perturb colors, GLoFool-Q generates adversarial images with better color fidelity and perceptual quality. GLoFool-T outperforms all the black-box methods in terms of success rate and robustness, with a performance comparable to the best white-box methods.

Synthetic to Authentic: Transferring Realism to 3D Face Renderings for Boosting Face Recognition,

Parsa Rahimi, Behrooz Razeghi, and Sébastien Marcel,

18th European Conference on Computer Vision (ECCV) Workshops, 2024

In this paper, we investigate the potential of image-to-image translation (I2I) techniques for transferring realism to 3D-rendered facial images in the context of Face Recognition (FR) systems. The primary motivation for using 3D-rendered facial images lies in their ability to circumvent the challenges associated with collecting large real face datasets for training FR systems. These images are generated entirely by 3D rendering engines, facilitating the generation of synthetic identities. However, it has been observed that FR systems trained on such synthetic datasets underperform when compared to those trained on real datasets, on various FR benchmarks. In this work, we demonstrate that by transferring the realism to 3D-rendered images (i.e., making the 3D-rendered images look more real), we can boost the performance of FR systems trained on these more photorealistic images. This improvement is evident when these systems are evaluated against FR benchmarks like IJB-C, LFW which utilize real-world data by 2% to %5, thereby paving new pathways for employing synthetic data in real-world applications. The project page is available at: https://idiap.ch/paper/syn2auth.

Detecting Criminal Networks via Non-Content Communication Data Analysis Techniques from the TRACY Project,

Pradeep Rangappa, Amanda Muscat, Alejandra Sanchez Lara, Petr Motlicek, Michaela Antonopoulou, Ioannis Fourfouris, Antonios Skarlatos, Nikos Avgerinos, Manolis Tsangaris, and Kasia Kostka,

15th EAI International Conference on Digital Forensics & Cyber Crime, 2024

This paper explores the critical role of non-content data (NCD), provided by electronic communications service providers in aid ing criminal investigations. As highlighted by the Law Enforcement Agen cies (LEAs) and the European Commission, NCD plays a fundamental role in identifying suspects and discerning behavioral patterns. Despite its significance, LEAs encounter various challenges in effectively ana lyzing the extensive volume of NCD. To address this issue, this paper presents the importance of (although simulated but realistic) data collec tion, the technologies that can be built and the methods for detecting the suspect within the framework of the TRACY project. These techniques aim to enhance capabilities of LEAs by processing large-scale NCD and aligning it with existing evidence. By prioritizing the tracing of suspects movements and integrating data from diverse NCD sources, TRACY's initial approach on synthetic data promises to significantly advance the identification of offenders involved in serious and organized crime.

Second Edition FRCSyn Challenge at CVPR 2024: Face Recognition Challenge in the Era of Synthetic Data,

Hatef Otroshi Shahreza, Anjith George, Alexander Unnervik, Parsa Rahimi, and Sébastien Marcel,

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024

Synthetic data is gaining increasing relevance for training machine learning models. This is mainly motivated due to several factors such as the lack of real data and intra-class variability time and errors produced in manual labeling and in some cases privacy concerns among others. This paper presents an overview of the 2nd edition of the Face Recognition Challenge in the Era of Synthetic Data (FRCSyn) organized at CVPR 2024. FRCSyn aims to investigate the use of synthetic data in face recognition to address current technological limitations including data privacy concerns demographic biases generalization to novel scenarios and performance constraints in challenging situations such as aging pose variations and occlusions. Unlike the 1st edition in which synthetic data from DCFace and GANDiffFace methods was only allowed to train face recognition systems in this 2nd edition we propose new sub-tasks that allow participants to explore novel face generative methods. The outcomes of the 2nd FRCSyn Challenge along with the proposed experimental protocol and benchmarking contribute significantly to the application of synthetic data to face recognition.

FRCSyn Challenge at WACV 2024: Face Recognition Challenge in the Era of Synthetic Data,

Alexander Unnervik, Anjith George, Christophe Ecabert, Parsa Rahimi, Hatef Otroshi Shahreza, and Sébastien Marcel,

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, 2024

Synthetic data is gaining increasing relevance for training machine learning models. This is mainly motivated due to several factors such as the lack of real data and intra-class variability time and errors produced in manual labeling and in some cases privacy concerns among others. This paper presents an overview of the 2nd edition of the Face Recognition Challenge in the Era of Synthetic Data (FRCSyn) organized at CVPR 2024. FRCSyn aims to investigate the use of synthetic data in face recognition to address current technological limitations including data privacy concerns demographic biases generalization to novel scenarios and performance constraints in challenging situations such as aging pose variations and occlusions. Unlike the 1st edition in which synthetic data from DCFace and GANDiffFace methods was only allowed to train face recognition systems in this 2nd edition we propose new sub-tasks that allow participants to explore novel face generative methods. The outcomes of the 2nd FRCSyn Challenge along with the proposed experimental protocol and benchmarking contribute significantly to the application of synthetic data to face recognition.

Model Pairing Using Embedding Translation for Backdoor Attack Detection on Open-Set Classification Tasks,

Alexander Unnervik, Hatef Otroshi Shahreza, Anjith George and Sébastien Marcel,

Neural Information Processing Systems (NeurIPS), Safe Generative AI Workshop, 2024

Backdoor attacks allow an attacker to embed a specific vulnerability in a machine learning algorithm, activated when an attacker-chosen pattern is presented, causing a specific misprediction. The need to identify backdoors in biometric scenarios has led us to propose a novel technique with different trade-offs. In this paper we propose to use model pairs on open-set classification tasks for detecting backdoors. Using a simple linear operation to project embeddings from a probe model's embedding space to a reference model's embedding space, we can compare both embeddings and compute a similarity score. We show that this score, can be an indicator for the presence of a backdoor despite models being of different architectures, having been trained independently and on different datasets. This technique allows for the detection of backdoors on models designed for open-set classification tasks, which is little studied in the literature. Additionally, we show that backdoors can be detected even when both models are backdoored. The source code is made available for reproducibility purposes.

Mirror-based Full-View Finger Vein Authentication with Illumination Adaptation,

Junduan Huang, Zifeng Li, Sushil Bhattacharjee, Wenxiong Kang and Sébastien Marcel,

IEEE Transactions on Circuits and Systems for Video Technology, 2024

Full-view finger vein (FV) biometrics systems capture multiple FV images of the presented finger ensuring that the entire surface of the finger is covered. Existing full-view FV systems suffer from three common problems: large device size, high cost for multi-camera system, and sub-optimal illumination in the recorded FV images. To address the problem of device size, we propose a novel Mirror-based Full-view FV (MFFV) capture device. The MFFV device has a compact size by using mirror-reflection approach. We reduce the cost of the device by using low-cost components, in particular, consumer-grade cameras. To address the problems of lower-quality images captured by such cameras and obtain optimally illuminated FV images, we propose a two-step approach. The first step is a Multi-illumination Intensities FV (MIFV) capture strategy, which capture the FV image set with varying illumination intensities. In the second step, a FV illumination adaptation (FVIA) algorithm is proposed to select the optimally illuminated FV image from the MIFV image set. Using the proposed MFFV device, we collect a comprehensive dataset, namely MFFV dataset, along with reproducible baseline FV authentication results for both single-view and full-view FV. Our experimental results demonstrate that the MIFV capture strategy as well as the FVIA algorithm can effectively improve the authentication performance, and that the full-view FV authentication is significantly superior than the single-view FV authentication. The source-code and dataset for reproducing our experimental results are publicly available (https://github.com/SCUT-BIP-Lab/MFFV).

Unveiling Synthetic Faces: How Synthetic Datasets Can Expose Real Identities,

Hatef Otroshi Shahreza and Sébastien Marcel,

Neural Information Processing Systems (NeurIPS), Workshop on New Frontiers in Adversarial Machine Learning, 2024

Synthetic data generation is gaining increasing popularity in different computer vision applications. Existing state-of-the-art face recognition models are trained using large-scale face datasets, which are crawled from the Internet and raise privacy and ethical concerns. To address such concerns, several works have proposed generating synthetic face datasets to train face recognition models. However, these methods depend on generative models, which are trained on real face images. In this work, we design a simple yet effective membership inference attack to systematically study if any of the existing synthetic face recognition datasets leak any information from the real data used to train the generator model. We provide an extensive study on 6 state-of-the-art synthetic face recognition datasets, and show that in all these synthetic datasets, several samples from the original real dataset are leaked. To our knowledge, this paper is the first work which shows the leakage from training data of generator models into the generated synthetic face recognition datasets. Our study demonstrates privacy pitfalls in synthetic face recognition datasets and paves the way for future studies on generating responsible synthetic face datasets. Project page: https://www.idiap.ch/paper/unveiling_synthetic_faces/.

Temporal fine-tuning for early risk detection,

Horacio Thompson, Esaú Villatoro-Tello, Manuel Montes-y-Gómez, and Marcelo Errecalde

Memorias de las JAIIO, 2024

Early Risk Detection (ERD) on the Web aims to identify promptly users facing social and health issues. Users are analyzed post by-post, and it is necessary to guarantee correct and quick answers, which is particularly challenging in critical scenarios. ERD involves opti mizing classiőcation precision and minimizing detection delay. Standard classiőcation metrics may not suffice, resorting to speciőc metrics such as ERDEθ that explicitly consider precision and delay. The current re search focuses on applying a multi-objective approach, prioritizing clas siőcation performance and establishing a separate criterion for decision time. In this work, we propose a completely different strategy, temporal őne-tuning, which allows tuning transformer-based models by explicitly incorporating time within the learning process. Our method allows us to analyze complete user post histories, tune models considering differ ent contexts, and evaluate training performance using temporal metrics. We evaluated our proposal in the depression and eating disorders tasks for the Spanish language, achieving competitive results compared to the best models of MentalRiskES 2023. We found that temporal őne-tuning optimized decisions considering context and time progress. In this way, by properly taking advantage of the power of transformers, it is possible to address ERD by combining precision and speed as a single objective.

HyperFace: Generating Synthetic Face Recognition Datasets by Exploring Face Embedding Hypersphere,

Hatef Otroshi-Shahreza, Sébastien Marcel,

Neural Information Processing Systems (NeurIPS), Safe Generative AI Workshop, 2024

Face recognition datasets are often collected by crawling Internet and without individuals' consents, raising ethical and privacy concerns. Generating synthetic datasets for training face recognition models has emerged as a promising alternative. However, the generation of synthetic datasets remains challenging as it entails adequate inter-class and intra-class variations. While advances in generative models have made it easier to increase intra-class variations in face datasets (such as pose, illumination, etc.), generating sufficient inter-class variation is still a difficult task. In this paper, we formulate the dataset generation as a packing problem on the embedding space (represented on a hypersphere) of a face recognition model and propose a new synthetic dataset generation approach, called HyperFace. We formalize our packing problem as an optimization problem and solve it with a gradient descent-based approach. Then, we use a conditional face generator model to synthesize face images from the optimized embeddings. We use our generated datasets to train face recognition models and evaluate the trained models on several benchmarking real datasets. Our experimental results show that models trained with HyperFace achieve state-of-the-art performance in training face recognition using synthetic datasets.

Exploring generalization to unseen audio data for spoofing: insights from SSL models,

Atharva Kulkarni, Hoan My Tran, Ajinkya Kulkarni, Sandipana Dowerah, Damien Lolive, and Mathew Magimai Doss,

Annual Conference of the International Speech Communication Association (Interspeech), 2024

Deep learning-based speech synthesis has significantly im proved realistic audio deepfakes. Despite advanced techniques such as self-supervised learning (SSL) and datasets, current state-of-the-art (SOTA) detection systems fail in out-of-domain scenarios due to the inability to generalize. This work explores the generalization problem through comprehensive experimen tation on cross-data evaluation. We observed how training data impacts model generalization, revealing that even SOTA sys tems struggle with consistent performance across different eval uation settings. This indicates a lack of extensive generalization abilities, especially in SSL approaches. To address this prob lem, we propose a multi-stage training framework alongside an ensemble of different systems to enhance the robustness and reliable detection in known and unknown out-of-domain sce narios. Experimental evaluation underscores the importance of an ensemble approach to mitigate the limitations in individual systems.

Unveiling Biases while Embracing Sustainability: Assessing the Dual Challenges of Automatic Speech Recognition Systems,

Ajinkya Kulkarni, Atharva Kulkarni, Miguel Couceiro and Isabel Trancoso,

Annual Conference of the International Speech Communication Association (Interspeech), 2024

In this paper, we present a bias and sustainability fo cused investigation of Automatic Speech Recognition (ASR) systems, namely Whisper and Massively Multilingual Speech (MMS), which have achieved state-of-the-art (SOTA) perfor mances. Despite their improved performance in controlled set tings, there remains a critical gap in understanding their effi cacy and equity in real-world scenarios. We analyze ASR biases w.r.t. gender, accent, and age group, as well as their effect on downstream tasks. In addition, we examine the environmental impact of ASR systems, scrutinizing the use of large acoustic models on carbon emission and energy consumption. We also provide insights into our empirical analyses, offering a valuable contribution to the claims surrounding bias and sustainability in ASRsystems.

Parametric point spread function estimation for thermal imaging systems using easy-to-manufacture random pattern targets,

Florian Piras, Edouard De Moura Presa, Peter Wellig, and Michael Liebling,

Target and Background Signatures X: Traditional Methods and Artificial Intelligence, 2024

Thermal and visible cameras can be characterized by their Point Spread Function (PSF), which captures the aberrations induced by the image formation process, which includes effects due to diffraction or motion. Various techniques for estimating the PSF based on a simple image of a target object that consists of a random pattern were shown to be effective. Here, we describe a computational pipeline for estimating parametric Gaussian PSFs characterized by their width, height, and orientation, based on binary random pattern targets that are suitable for thermal imaging and easy to manufacture. Specifically, we consider the influence of deviating from a strict random pattern so the targets can be manufactured with common cutting or 3D printing devices. We evaluate the estimation accuracy based on simulated patterns with varying dot, pitch, and target sizes for different values of the point spread function parameters. Finally, we show experimental examples of acquired on manufactured devices. Our results indicate that the proposed random pattern targets offer a simple and affordable approach to estimating local PSFs.

ChildPlay-Hand: A Dataset of Hand Manipulations in the Wild,

Arya Farkhondeh, Sami Tafasca, and Jean-Marc Odobez

European Conference on Computer Vision Workshop (ECCVW): Observing and Understanding Hands in Action, 2024

Hand-Object Interaction (HOI) is gaining significant atten tion, particularly with the creation of numerous egocentric datasets driven by AR/VR applications. However, third-person view HOI has received less attention, especially in terms of datasets. Most third-person view datasets are curated for action recognition tasks and feature pre-segmented clips of high-level daily activities, leaving a gap for in-the-wild datasets. To address this gap, we propose ChildPlay-Hand, a novel dataset that includes person and object bounding boxes, as well as manipulation ac tions. ChildPlay-Hand is unique in: (1) providing per-hand annotations; (2) featuring videos in uncontrolled settings with natural interactions, involving both adults and children; (3) including gaze labels from the ChildPlay-Gaze dataset for joint modeling of manipulations and gaze. The manipulation actions cover the main stages of an HOI cycle, such as grasping, holding or operating, and different types of releasing. To illus trate the interest of the dataset, we study two tasks: object in hand detec tion (OiH), i.e. if a person has an object in their hand, and manipulation stages (ManiS), which is more fine-grained and targets the main stages of manipulation. We benchmark various spatio-temporal and segmentation networks, exploring body vs. hand-region information and comparing pose and RGB modalities. Our findings suggest that ChildPlay-Hand is a challenging new benchmark for modeling HOI in the wild.

3. PHD THESES

Performing And Detecting Backdoor Attacks on Face Recognition Algorithms,

Alexander Unnervik,

Ecole Polytechnique Fédérale de Lausanne, 2024

The field of biometrics, and especially face recognition, has seen a wide-spread adoption the last few years, from access control on personal devices such as phones and laptops, to automated border controls such as in airports. The stakes are increasingly higher for these applications and thus the risks of succumbing to attacks are rising. More sophisticated algorithms typically require more data samples and larger models, leading to the need for more compute and expertise. These add up to making deep learning algorithms more a service provided by third parties, meaning more control and oversight of these algorithms are relinquished. When so much depends on these models working right, with nefarious actors gaining so much from them being circumvented, how does one then verify their integrity? This is the conundrum of integrity which is at the heart of the work presented here. One way by which face recognition algorithms (or more generally speaking, deep learning algorithms) fail, is by being vulnerable to backdoor attacks (BA): a type of attack involving a modification of the training set or the network weights to control the output behavior when exposed to specific samples. The detection of these backdoored networks (which we refer to as backdoor attack detection (BAD) is a challenging task, which is still an active field of research, particularly so when considering the constraints within which the literature considers the challenge (e.g. little to no consideration of open-set classification algorithms). In this thesis, we demonstrate that BAs can be performed on large face recognition algorithms and further the state of the art in BAD by providing with the following contributions: first, we study the vulnerability of face recognition algorithms to backdoor attacks and identify backdoor attack success with respect to the choice of identities and other variables. Second, we propose a first method by which backdoor attacks can be detected by studying weights distribution of clean models and comparing an unknown model to such distributions. This method is based on the principle of anomaly detection. Third, we propose a method for safely deploying models to make use of their clean behavior and detecting the activation of backdoors with a technique we call model pairing.

On the Information in Deep Biometric Templates: from Vulnerability of Unprotected Templates to Leakage in Protected Templates,

Hatef Otroshi Shahreza,

Ecole Polytechnique Fédérale de Lausanne, 2024

Biometric recognition systems tend toward ubiquity and are widely being used in different applications for authentication purposes. Compared to conventional authentication tools, such as PIN or password, which are always in danger of being forgotten or stolen, biometric authentication offers excellent convenience for the user. In contrast, in addition to security threats, biometric systems are also in danger of privacy issues. This is because biometric data include privacy-sensitive information of enrolled subjects, which causes privacy concerns in the application of biometric systems. Generally, in biometric systems, some features (also known as templates) are often extracted from biometric data, and are stored in the database of the system. Then, during recognition, similar templates are extracted and compared to the ones stored in the database. In this thesis, we focus on templates stored in biometric systems and investigate the vulnerability of systems to different attacks based on templates stored in the database of a biometric system. In the first part of the thesis, we consider the face recognition system as one of the popular biometric systems and show that if an adversary gains access to the database of a face recognition system, they may be able to reconstruct face images of underlying leaked facial templates. The reconstructed face images not only reveal privacy-sensitive information but also can be used to impersonate the systems that the user is enrolled in. We evaluate the adversary's successful attack rate in entering the system based on an injection attack by bypassing the camera. In addition, we consider the real-world scenario where the adversary may perform a practical presentation attack to impersonate and evaluate the attack rate. In the second part of the thesis, we propose new methods to protect biometric templates. We present MLP-Hash, a new cancelable biometric scheme that works based on random multi-layer perceptrons (MLP). We also discuss a hybrid template protection mechanism that leverages cancelable biometric and homomorphic encryption. Using cancelable biometric and homomorphic encryption not only boosts for a higher security of the protected templates but also reduces the required computation compared to applying homomorphic encryption only. The proposed template protection schemes can be used in systems of different biometric modalities (face, voice, finger vein, etc.). Finally, we present a new method to protect and enhance vascular biometric recognition methods using BioHashing and an auto-encoder network. The last part of this thesis is focused on the evaluation of template protection schemes. We first benchmark different template protection schemes based on the ISO/IEC 24745 standard requirements. We discuss the metrics to evaluate the leakage of information in the protected biometric templates. In particular, we investigate the invertibility of protected biometric templates and also propose a new measure to evaluate the linkability of protected templates. The proposed linkability metric is based on maximal leakage, which is a well-studied measure in information-theoretic literature. We show that the resulting linkability measure has a number of important theoretical properties and an operational interpretation in terms of statistical hypothesis testing. We further explore the application of our proposed method for the case that the adversary gains access to multiple protected templates.

AI FOR LIFE

1. JOURNAL PAPERS

Absolute retinal blood flow in healthy eyes and in eyes with retinal vein occlusion,
Thibaud Mautuit, Pierre Cunnac, Frédéric Truffer, André Anjos, Rebecca Dufrane, Gilbert Maître, Martial Geiser and Christophe Chiquet,
Microvascular Research, 152, 2024

To measure non-invasively retinal venous blood flow (RBF) in healthy subjects and patients with retinal venous occlusion (RVO). Methods: The prototype named AO-LDV (Adaptive Optics Laser Doppler Velocimeter), which combines a new absolute laser Doppler velocimeter with an adaptive optics fundus camera (rtx1, Imagine Eyes®, Orsay, France), was studied for the measurement of absolute RBF as a function of retinal vessel diameters and simultaneous measurement of red blood cell velocity. RBF was measured in healthy subjects (n = 15) and patients with retinal venous occlusion (RVO, n = 6). We also evaluated two softwares for the measurement of retinal vessel diameters: software 1 (automatic vessel detection, profile analysis) and software 2 (based on the use of deep neural networks for semantic segmentation of vessels, using a M2u-Net architecture). Results: Software 2 provided a higher rate of automatic retinal vessel measurement (99.5 % of 12,320 AO images) than software 1 (64.9 %) and wider measurements (75.5 ± 15.7 μm vs 70.9 ± 19.8 μm, p smaller than 0.001). For healthy subjects (n = 15), all the retinal veins in one eye were measured to obtain the total RBF. In healthy subjects, the total RBF was 37.8 ± 6.8 μl/min. There was a significant linear correlation between retinal vessel diameter and maximal velocity (slope = 0.1016; p smaller than 0.001; r2 = 0.8597) and a significant power curve correlation between retinal vessel diameter and blood flow (3.63 × 10−5 × D2.54; p smaller than 0.001; r2 = 0.7287). No significant relationship was found between total RBF and systolic and diastolic blood pressure, ocular perfusion pressure, heart rate, or hematocrit. For RVO patients (n = 6), a significant decrease in RBF was noted in occluded veins (3.51 ± 2.25 μl/min) compared with the contralateral healthy eye (11.07 ± 4.53 μl/min). For occluded vessels, the slope between diameter and velocity was 0.0195 (p smaller than 0.001; r2 = 0.6068) and the relation between diameter and flow was Q = 9.91 × 10−6 × D2.41 (p smaller than 0.01; r2 = 0.2526). Conclusion: This AO-LDV prototype offers new opportunity to study RBF in humans and to evaluate treatment in retinal vein diseases.

Developing 3D-Printed Wrist Splints for Distal Radius and Scaphoid Fractures,
Bernadette Tobler-Ammann, Frederic Schuind, Loïc Voillat, Théophile Gentilhomme, Esther Vögelin, Noé Murith and Bernard Masserey,
Journal of Wrist Surgery, 2024

The purpose of this study is to optimize conservative treatment of distal radius and scaphoid fracture, in terms of comfort, fracture stabilization, and prevention of cast complications. Advances in additive manufacturing have allowed the development of patient-specific anatomical braces (PSABs) which have the potential to fulfill this purpose. Our specific aims were to develop a model of PSAB, adapted to fracture care, to evaluate if this brace would be well tolerated by healthy volunteers and to determine its mechanical properties as compared with conventional methods of wrist immobilization. Several three-dimensional-printed splint prototypes were designed by mechanical engineers based on surgeons' and hand therapists' clinical expertise. These experimental braces underwent testing in a preclinical study involving 10 healthy volunteers, assessing comfort, satisfaction, and activities. The final prototype was mechanically compared with a conventional cast and a prefabricated splint, testing different closing systems. A mathematical algorithm was created to automatically adapt the final PSAB model to the patient's anatomy. The final prototype achieved an overall satisfaction score of 79%, weighing less than 90 g, made from polyamide, and fixed using hook and loop straps. The PSAB stiffness varied between 0.64 and 0.99 Nm/degree, surpassing the performance of both conventional plaster casts and prefabricated splints. The final wrist PSAB model, adapted for fracture treatment, is lightweight, comfortable, and provides anatomical contention. It is currently being tested for the treatment of stable distal radius and scaphoid fractures in comparison to conventional plaster cast.

Automated Sign Language Vocabulary Assessment: Comparing Human and Machine Ratings and Studying Learner Perceptions,

Franz Holzknecht, Sandrine Tornay, Alessia Battisti, Aaron Olaf Batty, Katja Tissi, Tobias Haug and Sarah Ebling,

Language Assessment Quarterly, 2024

Although automated spoken language assessment is rapidly growing, such systems have not been widely developed for signed languages. This study provides validity evidence for an automated web application that was developed to assess and give feedback on handshape and hand movement of L2 learners' Swiss German Sign Language signs. The study shows good machine-internal and human-machine agreement through many-facet Rasch analysis. Learner perceptions examined through questionnaire responses indicate that the automated system occasionally generated ratings which impacted the quality of feedback at the level of individual signs for individual learners. Implications are discussed from a learning-oriented assessment perspective.

Large Language Models, scientific knowledge and factuality: A framework to streamline human expert evaluation,

Magdalena Wysocka, Oskar Wysocki , Maxime Delmas, Vincent Mutel and André Freitas,

Journal of Biomedical Informatics, 2024

The paper introduces a framework for the evaluation of the encoding of factual scientific knowledge, designed to streamline the manual evaluation process typically conducted by domain experts. Inferring over and extracting information from Large Language Models (LLMs) trained on a large corpus of scientific literature can potentially define a step change in biomedical discovery, reducing the barriers for accessing and integrating existing medical evidence. This work explores the potential of LLMs for dialoguing with biomedical background knowledge, using the context of antibiotic discovery. Methods: The framework involves three evaluation steps, each assessing different aspects sequentially: fluency, prompt alignment, semantic coherence, factual knowledge, and specificity of the generated responses. By splitting these tasks between non-experts and experts, the framework reduces the effort required from the latter. The work provides a systematic assessment on the ability of eleven state-of-the-art LLMs, including ChatGPT, GPT-4 and Llama 2, in two prompting-based tasks: chemical compound definition generation and chemical compound–fungus relation determination. Results: Although recent models have improved in fluency, factual accuracy is still low and models are biased towards over-represented entities. The ability of LLMs to serve as biomedical knowledge bases is questioned, and the need for additional systematic evaluation frameworks is highlighted. Conclusion: While LLMs are currently not fit for purpose to be used as biomedical factual knowledge bases in a zero-shot setting, there is a promising emerging property in the direction of factuality as the models become domain specialised, scale up in size and level of human feedback.

Genome scale metabolic network modelling for metabolic profile predictions,

Juliette Cooke, Maxime Delmas, Cecilia Wieder, Pablo Rodríguez Mier, Clément Frainay, Florence Vinson, Timothy Ebbels, Nathalie Poupin, and Fabien Jourdan,

PLOS Computational Biology, 2024

Metabolic profiling (metabolomics) aims at measuring small molecules (metabolites) in complex samples like blood or urine for human health studies. While biomarker-based assessment often relies on a single molecule, metabolic profiling combines several metabolites to create a more complex and more specific fingerprint of the disease. However, in contrast to genomics, there is no unique metabolomics setup able to measure the entire metabolome. This challenge leads to tedious and resource consuming preliminary studies to be able to design the right metabolomics experiment. In that context, computer assisted metabolic profiling can be of strong added value to design metabolomics studies more quickly and efficiently. We propose a constraint-based modelling approach which predicts in silico profiles of metabolites that are more likely to be differentially abundant under a given metabolic perturbation (e.g. due to a genetic disease), using flux simulation. In genome-scale metabolic networks, the fluxes of exchange reactions, also known as the flow of metabolites through their external transport reactions, can be simulated and compared between control and disease conditions in order to calculate changes in metabolite import and export. These import/export flux differences would be expected to induce changes in circulating biofluid levels of those metabolites, which can then be interpreted as potential biomarkers or metabolites of interest. In this study, we present SAMBA (SAMpling Biomarker Analysis), an approach which simulates fluxes in exchange reactions following a metabolic perturbation using random sampling, compares the simulated flux distributions between the baseline and modulated conditions, and ranks predicted differentially exchanged metabolites as potential biomarkers for the perturbation. We show that there is a good fit between simulated metabolic exchange profiles and experimental differential metabolites detected in plasma, such as patient data from the disease database OMIM, and metabolic trait-SNP associations found in mGWAS studies. These biomarker recommendations can provide insight into the underlying mechanism or metabolic pathway perturbation lying behind observed metabolite differential abundances, and suggest new metabolites as potential avenues for further experimental analyses.

Relation Extraction in Underexplored Biomedical Domains: A Diversity-optimized Sampling and Synthetic Data Generation Approach,

Maxime Delmas, Magdalena Wysocka and Andre Freitas,

Computational Linguistics, 2024

The sparsity of labeled data is an obstacle to the development of Relation Extraction (RE) models and the completion of databases in various biomedical areas. While being of high interest in drug-discovery, the literature on natural products, reporting the identification of potential bioactive compounds from organisms, is a concrete example of such an overlooked topic. To mark the start of this new task, we created the first curated evaluation dataset and extracted literature items from the LOTUS database to build training sets. To this end, we developed a new sampler, inspired by diversity metrics in ecology, named Greedy Maximum Entropy sampler (https://github.com/idiap/gme-sampler). The strategic optimization of both balance and diversity of the selected items in the evaluation set is important given the resource-intensive nature of manual curation. After quantifying the noise in the training set, in the form of discrepancies between the text of input abstracts and the expected output labels, we explored different strategies accordingly. Framing the task as an end-to-end Relation Extraction, we evaluated the performance of standard fine-tuning (BioGPT, GPT-2, and Seq2rel) and few-shot learning with open Large Language Models (LLMs) (LLaMA 7B-65B). In addition to their evaluation in few-shot settings, we explore the potential of open LLMs as synthetic data generators and propose a new workflow for this purpose. All evaluated models exhibited substantial improvements when fine-tuned on synthetic abstracts rather than the original noisy data. We provide our best performing (F1-score = 59.0) BioGPT-Large model for end-to-end RE of natural products relationships along with all the training and evaluation datasets. See more details at https://github.com/idiap/abroad-re.

2. Conference Papers

Content-Based Objective Evaluation Of Artificially Generated Sign Language Videos,

Neha Tarigopula, Preyas Garg, Skanda Muralidhar, Sandrine Tornay, Dinesh Babu Jayagopi and Mathew Magimai-Doss,

49th IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2024

Sign language is vital for communication within the deaf and hard-of-hearing community. Avatar-based methods and deep learning techniques like Generative Adversarial Networks have shown promise in generating sign language video content. One of the challenges in sign language generation is the evaluation of the generated video content. One possible solution is to subjectively evaluate using human raters. This is time-consuming and costly. The other possible solution is objective evaluation. In the literature, video quality metrics such as PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) and skeleton-based measures such as MSE have been proposed. A limitation of these approaches is that they do not provide information about the generated video content. In this paper, we propose a novel phonology-based approach that evaluates the generated video along different channels, namely, hand movement and handshape, which convey the linguistic information in sign language. More precisely, in this approach an objective score is obtained by extracting sequences of hand movement sub-units and handshape sub-units class conditional probabilities (posterior features) from the source and generated videos and comparing them using dynamic time warping. Our experimental studies demonstrate that the proposed objective scoring method yields a better correlation to subjective human ratings than PSNR, SSIM, and MSE-based metrics.

DAIC-WOZ: On the Validity of Using the Therapist's prompts in Automatic Depression Detection from Clinical Interviews,

Sergio Burdisso, Ernesto A. Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, A. Pastor López-Monroy and Petr Motlicek,

6th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, 2024

Automatic depression detection from conversational data has gained significant interest in recent years. The DAIC-WOZ dataset, interviews conducted by a human-controlled virtual agent, has been widely used for this task. Recent studies have reported enhanced performance when incorporating interviewer's prompts into the model. In this work, we hypothesize that this improvement might be mainly due to a bias present in these prompts, rather than the proposed architectures and methods. Through ablation experiments and qualitative analysis, we discover that models using interviewer's prompts learn to focus on a specific region of the interviews, where questions about past experiences with mental health issues are asked, and use them as discriminative shortcuts to detect depressed participants. In contrast, models using participant responses gather evidence from across the entire interview. Finally, to highlight the magnitude of this bias, we achieve a 0.90 F1 score by intentionally exploiting it, the highest result reported to date on this dataset using only textual information. Our findings underline the need for caution when incorporating interviewers' prompts into models, as they may inadvertently learn to exploit targeted prompts, rather than learning to characterize the language and behavior that are genuinely indicative of the patient's mental health condition.

Predicting Heart Activity from Speech using Data-driven and Knowledge-based features,

Gasser Elbanna, Zohreh Mostaani and Mathew Magimai-Doss,

Annual Conference of the International Speech Communication Association (Interspeech), 2024

Accurately predicting heart activity and other biological signals is crucial for diagnosis and monitoring. Given that speech is an outcome of multiple physiological systems, a significant body of work studied the acoustic correlates of heart activity. Recently, self-supervised models have excelled in speech-related tasks compared to traditional acoustic methods. However, the robustness of data-driven representations in predicting heart activity remained unexplored. In this study, we demonstrate that self-supervised speech models outperform acoustic features in predicting heart activity parameters. We also emphasize the impact of individual variability on model generalizability. These findings underscore the value of data driven representations in such tasks and the need for more speech-based physiological data to mitigate speaker-related challenges.

Syllable Level Features For Parkinson's Disease Detection From Speech,

Sevada Hovsepyan and Mathew Magimai-Doss,

49th IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2024

Early detection of Parkinson's disease (PD), one of the most common neurodegenerative diseases, is crucial for successful treatment and symptom management. In this study, we propose a novel approach inspired by neurocomputational models of speech perception, for PD detection from speech samples. Our project emphasizes the importance of acoustic/linguistic markers to extract features at the syllable level, in contrast to conventional methods that extract features at the frame or state level. Through the use of syllable-level features (SLF), we successfully identify PD in recorded speech samples. Remarkably, the results not only match but potentially exceed the effectiveness of traditional feature sets used for this purpose. We hope that the proposed approach will provide a new basis for integrating linguistic insights into the identification of speech-related diseases.

Refining Tuberculosis Detection in CXR Imaging: Addressing Bias in Deep Neural Networks via Interpretability,

Özgür Güler, Manuel Günther, and André Anjos,

12th European Workshop on Visual Information Processing, 2024

Automatic classification of active tuberculosis from chest X-ray images has the potential to save lives, especially in low- and mid-income countries where skilled human experts can be scarce. Given the lack of available labeled data to train such systems and the unbalanced nature of publicly available datasets, we argue that the reliability of deep learning models is limited, even if they can be shown to obtain perfect classification accuracy on the test data. One way of evaluating the reliability of such systems is to ensure that models use the same regions of input images for predictions as medical experts would. In this paper, we show that pre-training a deep neural network on a large-scale proxy task, as well as using mixed objective optimization network (MOON), a technique to balance different classes during pre-training and fine-tuning, can improve the alignment of decision foundations between models and experts, as compared to a model directly trained on the target dataset. At the same time, these approaches keep perfect classification accuracy according to the area under the receiver operating characteristic curve (AUROC) on the test set, and improve generalization on an independent, unseen dataset. For the purpose of reproducibility, our source code is made available online.

Comparing Stability and Discriminatory Power of Hand-crafted Versus Deep Radiomics: A 3D-Printed Anthropomorphic Phantom Study,

Oscar Jimenez-del-Toro, Christoph Aberle, Roger Schaer, Michael Bach, Kyriakos Flouris, Ender Konukoglu, Bram Stieltjes, Markus M. Obmann, André Anjos, Henning Müller, and Adrien Depeursinge

12th European Workshop on Visual Information Processing, 2024

Radiomics have the ability to comprehensively quantify human tissue characteristics in medical imaging studies. However, standard radiomic features are highly unstable due to their sensitivity to scanner and reconstruction settings. We present an evaluation framework for the extraction of 3D deep radiomics features using a pre-trained neural network on real computed tomography (CT) scans for tissue characterization. We compare both the stability and discriminative power of the proposed 3D deep learning radiomic features versus standard hand-crafted radiomic features using 8 image acquisition protocols with a 3D-printed anthropomorphic phantom containing 4 classes of liver lesions and normal tissue. Even when the deep learning model was trained on an external dataset and for a different tissue characterization task, the resulting generic deep radiomics are at least twice more stable on 8 CT parameter variations than any category of hand-crafted features. Moreover, the 3D deep radiomics were also discriminative for the tissue characterization between 4 classes of liver tissue and lesions, with an average discriminative power of 93.5%.

Suppressing noise disparity in training data for automatic pathological speech detection,

Mahdi Amiri and Ina Kodrasi,
International Workshop on Acoustic Signal Enhancement, 2024

Although automatic pathological speech detection approaches show promising results when clean recordings are available, they are vulnerable to additive noise. Recently it has been shown that databases commonly used to develop and evaluate such approaches are noisy, with the noise characteristics between healthy and pathological recordings being different. Consequently, automatic approaches trained on these databases often learn to discriminate noise rather than speech pathology. This paper introduces a method to mitigate this noise disparity in training data. Using noise estimates from recordings from one group of speakers to augment recordings from the other group, the noise characteristics become consistent across all recordings. Experimental results demonstrate the efficacy of this approach in mitigating noise disparity in training data, thereby enabling automatic pathological speech detection to focus on pathology-discriminant cues rather than noise-discriminant ones.

Adversarial Robustness analysis in automatic pathological speech detection approaches,

Mahdi Amiri and Ina Kodrasi,
Annual Conference of the International Speech Communication Association (Interspeech), 2024

Automatic pathological speech detection relies on deep learning (DL), showing promising performance for various pathologies. Despite the critical importance of robustness in healthcare applications like pathological speech detection, the sensitivity of DL-based pathological speech detection approaches to adversarial attacks remains unexplored. This paper explores the impact of acoustically imperceptible adversarial perturbations on DL-based pathological speech detection. Imperceptibility of perturbations, generated using the projected gradient descent algorithm, is evaluated using speech enhancement metrics. Results reveal a high vulnerability of DL-based pathological speech detection to adversarial perturbations, with adversarial training ineffective in enhancing robustness. Analysis of the perturbations provide insights into the speech components that the approaches attend to. These findings highlight the need for research in robust pathological speech detection.

Test-time adaptation for automatic pathological speech detection in noisy environments,

Mahdi Amiri and Ina Kodrasi,
European Signal Processing Conference, 2024

Deep learning-based pathological speech detection approaches are gaining popularity as a diagnostic tool to support time-consuming and subjective clinical assessments. While these approaches perform well in controlled environments with clean recordings, their performance significantly degrades in realistic scenarios with background noise. In this paper, we propose a test time adaptation framework to increase the robustness of such approaches to background noise during inference. To this end, we use a voice activity detector to extract noise-only segments from the test signal. These segments are used to augment a portion of the training/validation data, which is then exploited to fine-tune the classification models. Extensive experimental results demonstrate the effectiveness of the proposed framework in increasing robustness to noise for state-of-the-art automatic pathological speech detection approaches.

Impact of speech mode in automatic pathological speech detection,

Shakeel Sheikh and Ina Kodrasi,
European Signal Processing Conference, 2024

Automatic pathological speech detection approaches yield promising results in identifying various pathologies. These approaches are typically designed and evaluated for phonetically controlled speech scenarios, where speakers are prompted to articulate identical phonetic content. While gathering controlled speech recordings can be laborious, spontaneous speech can be conveniently acquired as potential patients navigate their daily routines. Further, spontaneous speech can be valuable in detecting subtle and abstract cues of pathological speech. Nonetheless, the efficacy of automatic pathological speech detection for spontaneous speech remains unexplored. This paper analyzes the influence of speech mode on pathological speech detection approaches, examining two distinct categories of approaches, i.e., classical machine learning and deep learning. Results indicate that classical approaches may struggle to capture pathology discriminant cues in spontaneous speech. In contrast, deep learning approaches demonstrate superior performance, managing to extract additional cues that were previously inaccessible in nonspontaneous speech.

Investigating Semantic Segmentation Models to Assist Visually Impaired People,

Michael Villamizar, Olivier Canévet and Jean-Marc Odobez,

18th European Conference on Computer Vision Workshop (ECCVW), 2024

This paper addresses the semantic segmentation task with the purpose of allowing visually impaired people to comprehend their en vironments. To this end, we study and leverage convolutional networks trained on public automotive datasets and a new egocentric infrared dataset collected in urban areas. Domain adaptation, efficiency and seg mentation accuracy are the focus of our study.

GS: A Novel Framework for Multi-Person Temporal Gaze Following and Social Gaze Prediction,

Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard and Jean-Marc Odobez,

Neural Information Processing System (NeurIPS), 2024

Gaze following and social gaze prediction are fundamental tasks providing insights into human communication behaviors, intent, and social interactions. Most previous approaches addressed these tasks separately, either by designing highly specialized social gaze models that do not generalize to other social gaze tasks or by considering social gaze inference as an ad-hoc post-processing of the gaze following task. Furthermore, the vast majority of gaze following approaches have proposed models that can handle only one person at a time and are static, therefore failing to take advantage of social interactions and temporal dynamics. In this paper, we address these limitations and introduce a novel framework to jointly predict the gaze target and social gaze label for all people in the scene. It comprises (i) a temporal, transformer-based architecture that, in addition to frame tokens, handles person- specific tokens capturing the gaze information related to each individual; (ii) a new dataset, VSGaze, built from multiple gaze following and social gaze datasets by extending and validating head detections and tracks, and unifying annotation types. We demonstrate that our model can address and benefit from training on all tasks jointly, achieving state-of-the-art results for multi-person gaze following and social gaze prediction. Our annotations and code will be made publicly available.

Toward Semantic Gaze Target Detection,

Samy Tafasca, Anshul Gupta, Victor Bros and Jean-Marc Odobez,

Neural Information Processing System (NeurIPS), 2024

From the onset of infanthood, humans naturally develop the ability to closely observe and interpret the visual gaze of others. This skill, known as gaze following, holds significance in developmental theory as it enables us to grasp another person’s mental state, emotions, intentions, and more [6]. In computer vision, gaze following is defined as the prediction of the pixel coordinates where a person in the image is focusing their attention. Existing methods in this research area have predominantly centered on pinpointing the gaze target by predicting a gaze heatmap or gaze point. However, a notable drawback of this approach is its limited practical value in gaze applications, as mere localization may not fully capture our primary interest — understanding the underlying semantics, such as the nature of the gaze target, rather than just its 2D pixel location. To address this gap, we extend the gaze following task, and introduce a novel architecture that simultaneously predicts the localization and semantic label of the gaze target. We devise a pseudo-annotation pipeline for the GazeFollow dataset, propose a new benchmark, develop an experimental protocol and design a suitable baseline for comparison. Our method sets a new state-of-the-art on the main GazeFollow benchmark for localization and achieves competitive results in the recognition task on both datasets compared to the baseline, with 40% fewer parameters.

Weakly-supervised Autism Severity Assessment in Long Videos,

Abid Ali, Mahmoud Ali, Jean-Marc Odobez, Camilla Barbini, Séverine Dubuisson, Francois Bremond, and Susanne Thümmler,

International Conference on Content-based Multimedia Indexing, 2024

Autism Spectrum Disorder (ASD) is a diverse collection of neurobiological conditions marked by challenges in social communication and reciprocal interactions, as well as repetitive and stereotypical behaviors. Atypical behavior patterns in a long, untrimmed video can serve as biomarkers for children with ASD. In this paper, we propose a video-based weakly-supervised method that takes spatio-temporal features of long videos to learn typical and atypical behaviors for autism detection. On top of that, we propose a shallow TCN-MLP network, which is designed to further categorize the severity score. We evaluate our method on actual evaluation videos of children with autism collected and annotated (for severity score) by clinical professionals. Experimental results demonstrate the effectiveness of behavioral biomarkers that could help clinicians in autism spectrum analysis.

SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials,

Mael Jullien, Marco Valentino and André Freitas,

18th International Workshop on Semantic Evaluation (SemEval), 2024

Large Language Models (LLMs) are at the forefront of NLP achievements but fall short in dealing with shortcut learning, factual inconsistency, and vulnerability to adversarial inputs.These shortcomings are especially critical in medical contexts, where they can misrepresent actual model capabilities. Addressing this, we present SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for ClinicalTrials. Our contributions include the refined NLI4CT-P dataset (i.e., Natural Language Inference for Clinical Trials - Perturbed), designed to challenge LLMs with interventional and causal reasoning tasks, along with a comprehensive evaluation of methods and results for participant submissions. A total of 106 participants registered for the task contributing to over 1200 individual submissions and 25 system overview papers. This initiative aims to advance the robustness and applicability of NLI models in healthcare, ensuring safer and more dependable AI assistance in clinical decision-making. We anticipate that the dataset, models, and outcomes of this task can support future research in the field of biomedical NLI. The dataset, competition leaderboard, and website are publicly available.

An LLM-based Knowledge Synthesis and Scientific Reasoning Framework for Biomedical Discovery,

Oskar Wysocki, Magdalena Wysocka, Danilo Carvalho, Alex Bogatu, Danilo Miranda, Maxime Delmas, Harriet Unsworth, and Andre Freitas,

62nd Annual Meeting of the Association for Computational Linguistics, 2024

We present BioLunar, developed using the Lunar framework, as a tool for supporting biological analyses, with a particular emphasis on molecular-level evidence enrichment for biomarker discovery in oncology. The platform integrates Large Language Models (LLMs) to facilitate complex scientific reasoning across distributed evidence spaces, enhancing the capability for harmonizing and reasoning over heterogeneous data sources. Demonstrating its utility in cancer research, BioLunar leverages modular design, reusable data access and data analysis components, and a low-code user interface, enabling researchers of all programming levels to construct LLM-enabled scientific workflows. By facilitating automatic scientific discovery and inference from heterogeneous evidence, BioLunar exemplifies the potential of the integration between LLMs, specialised databases and biomedical tools to support expert-level knowledge synthesis and discovery.

OptoMechanical Modulation Tomography for Ungated Compressive Cardiac Light Sheet Microscopy,

François Marelli and Michael Liebling

IEEE International Symposium on Biomedical Imaging (ISBI), 2024

OptoMechanical Modulation Tomography (OMMT) is a compressed sensing optical microscopy method where measurements are obtained by scanning a light sheet through a sample while modulating its intensity over the course of the camera integration time. Because mechanical scanning is not instantaneous, this method was so far considered unsuitable for imaging dynamic samples. Yet living samples would particularly benefit from the method’s reduced light exposure. In this paper we extend OMMT to allow imaging of objects that have a periodic motion, such as the heart in transparent larvae. We derived a method in which measurements are obtained by integrating the space-phase domain along patterned paths. We implemented the reconstruction with an iterative solver, and demonstrated the feasibility of the method based on simulated data of a beating heart. We observed that compression factors up to 8 lead to reliable reconstruction, and that the method is robust to uncertain acquisition start phases. Our results confirm that OMMT can be extended to imaging dynamic samples opening up the possibility to apply this method in experimental settings where low light exposure is desirable.

Parametric point spread function estimation for thermal imaging systems using easy-to-manufacture random pattern targets,

Florian Pirasa, Edouard De Moura Presab, Peter Welligb, and Michael Lieblinga,

SPIE Defense + Commercial Sensing, 2024

Thermal and visible cameras can be characterized by their Point Spread Function (PSF), which captures the aberrations induced by the image formation process, which includes effects due to diffraction or motion. Various techniques for estimating the PSF based on a simple image of a target object that consists of a random pattern were shown to be effective. Here, we describe a computational pipeline for estimating parametric Gaussian PSFs characterized by their width, height, and orientation, based on binary random pattern targets that are suitable for thermal imaging and easy to manufacture. Specifically, we consider the influence of deviating from a strict random pattern so the targets can be manufactured with common cutting or 3D printing devices. We evaluate the estimation accuracy based on simulated patterns with varying dot, pitch, and target sizes for different values of the point spread function parameters. Finally, we show experimental examples of acquired on manufactured devices. Our results indicate that the proposed random pattern targets offer a simple and affordable approach to estimating local PSFs.

AI FOR EVERYONE

1. JOURNAL PAPERS

Generative AI Literacy: Twelve Defining Competencies,
Ravinithesh Annapureddy, Alessandro Fornaroli and Daniel Gatica-Perez,
ACM Digital Government: Research and Practice (DGOV), 2024

This paper introduces a competency-based model for generative artificial intelligence (AI) literacy covering essential skills and knowledge areas necessary to interact with generative AI. The competencies range from foundational AI literacy to prompt engineering and programming skills, including ethical and legal considerations. These twelve competencies offer a framework for individuals, policymakers, government officials, and educators looking to navigate and take advantage of the potential of generative AI responsibly. Embedding these competencies into educational programs and professional training initiatives can equip individuals to become responsible and informed users and creators of generative AI. The competencies follow a logical progression and serve as a roadmap for individuals seeking to get familiar with generative AI and for researchers and policymakers to develop assessments, educational programs, guidelines, and regulations.

M3BAT: Unsupervised Domain Adaptation for Multimodal Mobile Sensing with Multi-Branch Adversarial Training,
Lakmal Buddika Meegahapola, Hamza Hassoune and Daniel Gatica-Perez,
PACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies (IMWUT), 8(2):46, 2024

Over the years, multimodal mobile sensing has been used extensively for inferences regarding health and well-being, behavior, and context. However, a significant challenge hindering the widespread deployment of such models in real-world scenarios is the issue of distribution shift. This is the phenomenon where the distribution of data in the training set differs from the distribution of data in the real world---the deployment environment. While extensively explored in computer vision and natural language processing, and while prior research in mobile sensing briefly addresses this concern, current work primarily focuses on models dealing with a single modality of data, such as audio or accelerometer readings, and consequently, there is little research on unsupervised domain adaptation when dealing with multimodal sensor data. To address this gap, we did extensive experiments with domain adversarial neural networks (DANN) showing that they can effectively handle distribution shifts in multimodal sensor data. Moreover, we proposed a novel improvement over DANN, called M3BAT, unsupervised domain adaptation for multimodal mobile sensing with multi-branch adversarial training, to account for the multimodality of sensor data during domain adaptation with multiple branches. Through extensive experiments conducted on two multimodal mobile sensing datasets, three inference tasks, and 14 source-target domain pairs, including both regression and classification, we demonstrate that our approach performs effectively on unseen domains. Compared to directly deploying a model trained in the source domain to the target domain, the model shows performance increases up to 12% AUC (area under the receiver operating characteristics curves) on classification tasks, and up to 0.13 MAE (mean absolute error) on regression tasks.

2. Conference Papers

Score Normalization for Demographic Fairness in Face Recognition,
Yu Linghu, Tiago de Freitas Pereira, Christophe Ecabert, Sébastien Marcel and Manuel Günther,
IEEE International Joint Conference on Biometrics (IJCB), 2024

Fair biometric algorithms have similar verification performance across different demographic groups given a single decision threshold. Unfortunately, for state-of-the-art face recognition networks, score distributions differ between demographics. Contrary to work that tries to align those distributions by extra training or fine-tuning, we solely focus on score post-processing methods. As proved, well-known sample-centered score normalization techniques, Z-norm and T-norm, do not improve fairness for high-security operating points. Thus, we extend the standard Z/T-norm to integrate demographic information in normalization. Additionally, we investigate several possibilities to incorporate cohort similarities for both genuine and impostor pairs per demographic to improve fairness across different operating points. We run experiments on two datasets with different demographics (gender and ethnicity) and show that our techniques generally improve the overall fairness of five state-of-the-art pre-trained face recognition networks, without downgrading verification performance. We also indicate that an equal contribution of False Match Rate (FMR) and False Non-Match Rate (FNMR) in fairness evaluation is required for the highest gains. Code and protocols are available.

Demographic Fairness Transformer for Bias Mitigation in Face Recognition,
Ketan Kotwal and Sébastien Marcel,
IEEE International Joint Conference on Biometrics (IJCB), 2024

Demographic bias in deep learning-based face recognition systems has led to serious concerns. Often, the biased nature of models is attributed to severely imbalanced datasets used for training. However, several studies have shown that biased models can emerge even when trained on balanced data due to factors in the data acquisition process. Considering the impact of input data on demographic bias, we propose an image to image transformer for demographic fairness (DeFT). This transformer can be applied before the pretrained recognition CNN to selectively enhance the image representation with the goal of reducing the bias through overall recognition pipeline. The multi-head encoders of DeFT provide multiple transformation paths to the input which are then combined based on its demographic information implicitly inferred through soft-attention mechanism applied to intermittent layers of DeFT. We compute probabilistic weights for demographic information, as opposed to conventional hard labels, simplifying the learning process and enhancing the robustness of the DeFT. Our experiments demonstrate that in a cross-dataset testing (pretrained as well as locally trained models), integrating the DeFT leads to fairer models, reducing the variation in accuracies while often slightly improving average recognition accuracy over baselines.

Human Interest or Conflict? Leveraging LLMs for Automated Framing Analysis in TV Shows,
David Alonso del Barrio, Max Tiel and Daniel Gatica-Perez,
ACM International Conference on Interactive Media Experiences (IMX), 2024

In the current media landscape, understanding the framing of information is crucial for critical consumption and informed decision making. Framing analysis is a valuable tool for identifying the underlying perspectives used to present information, and has been applied to a variety of media formats, including television programs. However, manual analysis of framing can be time-consuming and labor-intensive. This is where large language models (LLMs) can play a key role. In this paper, we propose a novel approach to use prompt-engineering to identify the framing of spoken content in television programs. Our findings indicate that prompt-engineering LLMs can be used as a support tool to identify frames, with agreement rates between human and machine reaching up to 43%. As LLMs are still under development, we believe that our approach has the potential to be refined and further improved. The potential of this technology for interactive media applications is vast, including the development of support tools for journalists, educational resources for students of journalism learning about framing and related concepts, and interactive media experiences for audiences.

Learning About Social Context from Smartphone Data: Generalization Across Countries and Daily Life Moments,

Aurel Ruben Mader, Lakmal Buddika Meegahapola and Daniel Gatica-Perez,

ACM Conference on Human Factors in Computing Systems (CHI), 2024

Understanding how social situations unfold in people's daily lives is relevant to designing mobile systems that can support users in their personal goals, well-being, and activities. As an alternative to questionnaires, some studies have used passively collected smart phone sensor data to infer social context (i.e., being alone or not) with machine learning models. However, the few existing studies have focused on specific daily life occasions and limited geographic cohorts in one or two countries. This limits the understanding of how inference models work in terms of generalization to everyday life occasions and multiple countries. In this paper, we used a novel, large-scale, and multimodal smartphone sensing dataset with over 216K self-reports collected from 581 young adults in five countries (Mongolia, Italy, Denmark, UK, Paraguay), first to understand whether social context inference is feasible with sensor data, and then, to know how behavioral and country-level diversity affects inferences. We found that several sensors are informative of social context, that partially personalized multi-country models (trained and tested with data from all countries) and country-specific models (trained and tested within countries) can achieve similar performance above 90% AUC, and that models do not generalize well to unseen countries regardless of geographic proximity. These findings confirm the importance of the diversity of mobile data, to better understand social context inference models in different countries.

Towards Wine Tasting Activity Recognition for a Digital Sommelier,

Mario O. Parra, Jesus Favela, Luis A. Castro, and Daniel Gatica-Perez,

ACM ICMI Workshop on Exploring Innovative Technology for Commensality and Human-Food Interaction, 2024

In this study, we evaluated the feasibility of using zero-shot classification models for activity recognition in a Digital Sommelier. Our experiment involved preprocessing video data by extracting frames and categorizing user activities related to a wine-tasting scenario. Image classification models demonstrated high accuracy in distinguishing between "engaged" and "disengaged" states. However, video classification models presented a lower performance in classifying user activities such as "observing wine," "smelling wine," and "sipping wine" due to the interdependent nature of the activities. Despite these challenges, our findings highlight the potential of zero-shot classification models in enhancing virtual assistants' ability to recognize and respond to user activities.

Using Backbone Foundation Model for Evaluation Fairness Without Demographic Data in Chest Radiography,

Dilermando Queiroz Neto, André Anjos and Lilian Berton,

27th International Conference on Medical Image Computing and Computer Assisted Intervention, 2024

Ensuring consistent performance across diverse populations and incorporating fairness into machine learning models are crucial for advancing medical image diagnostics and promoting equitable healthcare. However, many databases do not provide protected attributes or contain unbalanced representations of demographic groups, complicating the evaluation of model performance across different demographics and the application of bias mitigation techniques that rely on these attributes. This study aims to investigate the effectiveness of using the backbone of Foundation Models as an embedding extractor for creating groups that represent protected attributes, such as gender and age. We propose utilizing these groups in different stages of bias mitigation, including pre-processing, in-processing, and evaluation. Using databases in and out-of-distribution scenarios, it is possible to identify that the method can create groups that represent gender in both databases and reduce in 4.44% the difference between the gender attribute in-distribution and 6.16% in out-of-distribution. However, the model lacks robustness in handling age attributes, underscoring the need for more fundamentally fair and robust Foundation models. These findings suggest a role in promoting fairness assessment in scenarios where we lack knowledge of attributes, contributing to the development of more equitable medical diagnostics.

Does Data-Efficient Generalization Exacerbate Bias in Foundation Models?,

Dilermando Queiroz Neto, André Anjos, and Lilian Berton,

18th European Conference on Computer Vision (ECCV), 2024

Foundation models have emerged as robust models with label efficiency in diverse domains. In medical imaging, these models contribute to the advancement of medical diagnoses due to the difficulty in obtaining labeled data. However, it is unclear whether using a large amount of unlabeled data, biased by the presence of sensitive attributes during pre-training, influences the fairness of the model. This research examines the bias in the Foundation model (RetFound) when it is applied to fine-tune the Brazilian Multilabel Ophthalmological Dataset (BRSET), which has a different population than the pre-training dataset. The model evaluation, in comparison with supervised learning, shows that the Foundation Model has the potential to reduce the gap between the maximum AUC and minimum AUC evaluations across gender and age groups. However, in a data-efficient generalization, the model increases the bias when the data amount decreases. These findings suggest that when deploying a Foundation Model in real-life scenarios with limited data, the possibility of fairness issues should be considered.

CulturePark: Boosting Cross-cultural Understanding in Large Language Models,

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie and Jindong Wang,

Neural Information Processing Systems (NeurIPS), 2024

Cultural bias is pervasive in many large language models (LLMs), largely due to the deficiency of data representative of different cultures. Typically, cultural datasets and benchmarks are constructed either by extracting subsets of existing datasets or by aggregating from platforms such as Wikipedia and social media. However, these approaches are highly dependent on real-world data and human annotations, making them costly and difficult to scale. Inspired by cognitive theories on social communication, this paper introduces CulturePark, an LLM-powered multi-agent communication framework for cultural data collection. CulturePark simulates cross-cultural human communication with LLM-based agents playing roles in different cultures. It generates high-quality cross-cultural dialogues encapsulating human beliefs, norms, and customs. Using CulturePark, we generated 41,000 cultural samples to fine-tune eight culture-specific LLMs. We evaluated these models across three downstream tasks: content moderation, cultural alignment, and cultural education. Results show that for content moderation, our GPT-3.5-based models either match or outperform GPT-4 on datasets. Regarding cultural alignment, our models surpass GPT-4 on Hofstede's VSM 13 framework. Furthermore, for cultural education of human participants, our models demonstrate superior outcomes in both learning efficacy and user experience compared to GPT-4. CulturePark proves an important step in addressing cultural bias and advancing the democratization of AI, highlighting the critical role of culturally inclusive data in model training. Code is released at this https URL.

3. PhD Thesis

Generalization and Personalization of Machine Learning for Multimodal Mobile Sensing in Everyday Life,

Lakmal Buddika Meegahapola,
Ecole Polytechnique Fédérale de Lausanne, 2024

A range of behavioral and contextual factors, including eating and drinking behavior, mood, social context, and other daily activities, can significantly impact an individual's quality of life and overall well- being. Therefore, inferring everyday life aspects with the use of smartphone and wearable sensors, also broadly known as mobile sensing, is gaining traction across both clinical and non-clinical populations due to the widespread use of smartphones around the world. Such inferences are of use in mobile health apps, mobile food diaries, and generic mobile apps. However, despite the long-standing promise in the domain, realizing the full potential of models, in the wild, is still far from reality due to two primary deployment challenges: the generalization and personalization of models. In addition, there are understudied domains, such as eating and drinking behavior modeling with multimodal mobile sensing and machine learning. Hence, this thesis delves into the realm of multimodal mobile sensing with an eye for the generalization and personalization of models, exploring a range of novel inferences at the intersection of eating and drinking behavior, mood, daily activities, and context. This thesis offers an extensive exploration of novel inferences and deployment challenges in multimodal mobile sensing. First, the thesis explores eating and drinking behavior and its interplay with mood, social context, and daily activities, viewed through the lens of both model personalization and generalization. Additionally, the thesis delves into the challenge of cross-country generalization for mobile sensing-based models and presents a novel deep learning architecture for unsupervised domain adaptation, yielding enhanced performance in unfamiliar domains. As a result, this thesis contributes both empirically and methodologically to the fields of ubiquitous and mobile computing and digital health.

Biologically Inspired Spiking Neural Networks for Speech Recognition,

Alexandre Bittar,
Ecole Polytechnique Fédérale de Lausanne, 2024

Biological neural networks, driving cognitive processes in the human brain, have long been a source of inspiration for computational models. Drawing from the physiology of neural dynamics, spiking neural networks stand out as prominent candidates for replicating and understanding the brain's functionality through efficient information processing. In this thesis, we investigate spiking neural networks during sequential processing by leveraging deep learning frameworks to train and evaluate them on speech recognition tasks. Focusing on the acoustic model, our approach captures the temporal patterns and phonetic features inherent to speech signals, providing insights into speech processing throughout the human auditory pathway. A first part is dedicated to conventional artificial neural networks and their utilisation of recurrence to address context dependencies and develop a form of working memory. Building upon a recent probabilistic derivation of recurrent neural networks, our research extends the approach and yields novel interpretable deep learning modules. While the main contributions remain theoretical, the resulting lightweight Bayesian recurrent units are shown to improve speech recognition performance compared to standard recurrent neural networks. In a second part, we shift to the main focus of spiking neural networks that encode and transmit information via sparse and binary spike sequences. Using the surrogate gradient method, we formulate physiologically inspired architectures as a special case of recurrent neural networks. This enables us to bootstrap a study of spiking neural networks from existing deep learning frameworks. Here we also explore the role of recurrence and memory in the form of different feedback mechanisms including layer-wise recurrent connections and unit-wise spike frequency adaptation in the neuron model. While the main aim is to develop the understanding of physiological processes, our results on speech recognition tasks also contribute to the field of energy-efficient neuromorphic technology. Lastly, an analysis of our trained spiking architectures reveals the replication of key features observed in biological networks, offering a novel and scalable approach for their study. In particular, we explore the phenomenon of neural oscillations, characteristic of cognitive processes in the brain. Our analysis confirms the presence of cross-frequency couplings in the trained networks during speech processing, notably between theta and gamma frequency bands. This synchronisation of the spiking activity is shown to arise naturally, simply through gradient descent training, and is enhanced by the incorporation of recurrent mechanisms. In summary, this thesis contributes to the field of neural network research by offering insights into the concepts of recurrence and spiking dynamics during sequential processing tasks, particularly in the context of speech recognition. Through a combination of theoretical analysis and practical experimentation, we develop novel methods, deep learning modules, and physiologically inspired architectures that advance our understanding of neural computation and its applications. By replicating key features observed in biological networks, our research contributes to future developments in neuromorphic computing and cognitive science.

Idiap Research Institute
Rue Marconi 19
1920 Martigny
Switzerland