MEDIA RELEASE

Switzerland, July 27, 2023

TO BE RELEASED IMMEDIATELY

Towards energy-efficient artificial intelligence models

As artificial intelligence technologies reach everyday new performances, their energy cost is also increasing significantly. Idiap researchers are proposing a novel approach to address this challenge during a period of rising energy costs.

Artificial intelligence technologies are getting better and better at understanding the meaning of natural languages. This progress makes a crucial difference for real world applications using text analysis and speech recognition tools. Currently, the most cutting-edge technologies rely on models called transformers, which place significant demands on computing resources. As a result, the number of mathematical operations needed for analysing longer and longer text and speech signals becomes exceedingly burdensome even with more computing power. Acknowledging this flaw, Idiap's researchers came up with a strategy to reduce the computational and energy resources needed to run these technologies.

A new model

"When working with resource-demanding algorithms, researchers often have to artificially shorten their inputs to obtain results in a reasonable timeframe. Cutting computing costs is crucial for research and for its applications," Florian Mai, first author of the paper and research assistant in the Natural language understanding group, explains. To reduce these computing costs, the researchers decided to revisit a decades-old model called multi-layer perceptrons. A model that is usually considered unfit for processing language due to its inability to handle inputs of varying length. However, Idiap's researchers found that, by switching from a static to a dynamic processing model, language-related data can be processed effectively. They called this model HyperMixer.

More cost effective AI

In addition to their modelling improvement, the researchers were also able to demonstrate empirically that their model performs better or on par with traditional models. In comparison to transformers, HyperMixer achieves these results at substantially lower computing costs in terms of processing time, training data, and parameter tuning.

Beyond this scientific achievement, HyperMixer takes an important step in the direction of diminishing the environmental impact of AI technologies. In an era where energy costs are skyrocketing and resources are dwindling, it is crucial for research to contribute. "Idiap's motto AI for society must also be reflected in algorithms," James Henderson, head of the Natural language understanding group, concludes..

Scientific papers

- "HyperMixer: An MLP-based Low Cost Alternative to Transformers" Florian Mai, Arnaud Pannatier, Fabio Fehr, Haolin Chen, François Marelli, François Fleuret, James Henderson: https://publications.idiap.ch/publications/show/5073

- “HyperConformer: Multi-head HyperMixer for Efficient Speech Recognition” Florian Mai, Juan, Zuluaga-Gomez, Titouan Parcollet, Petr Motlicek:
https://publications.idiap.ch/publications/show/5054

- Code for both these models: https://github.com/idiap/hypermixing

Idiap Research Institute has been a world specialist in artificial intelligence for 30 years. Voice and visual recognition, human–computer interaction, robotics, and language analysis are just some of the Institute’s fields of expertise. Based in Martigny, Switzerland, Idiap is engaged in local, national, and international projects. The non-profit Idiap Foundation was created in 1991 by the city of Martigny, the State of Valais, the l’Ecole polytechnique fédérale de Lausanne (EPFL), the University of Geneva, and Swisscom.

Contacts

- Arnaud Pannatier, co-author of the paper, +41 77 439 30 16, arnaud.pannatier@idiap.ch
- Nicolas Filippov, head of communications, +41 79 139 92 65, nicolas.filippov@idiap.ch