

Simple but Effective Techniques to Reduce Dataset Biases

Rabeeh Karimi^{1,2}, James Henderson¹

1. Idiap Research Institute
2. École Polytechnique Fédérale de Lausanne (EPFL)

November 13th, 2019

Overview

- 1 Introduction
- 2 Our Model
- 3 Experimental Results
- 4 Takeaways

Biases are a General Problem in NLP and Computer vision

How Much *Reading* Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks

Divyansh Kaushik

Zachary C. Lipton

Language Technology
Carnegie Mellon University
dkaushik

Making the V in VQA Matter:

Elevating the Role of Image Understanding in Visual Question Answering

Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

Devi Parikh³
Department of Computer Science
Georgia Institute of Technology
parikh@cs.gatech.edu

Jieyu Zhao[§]

Vicente Ordonez[§]

[§]University of California, Los Angeles

Annotation Artifacts in Natural Language Inference Data

Suchin Gururangan^{★◇} Swabha Swayamdipta^{★♥}

Omer Levy[★] Roy Schwartz^{★★} Samuel R. Bowman[†] Noah A. Smith[★]

Tackling the Story Ending Biases in The Story Cloze Test

Rishi Sharma¹, James F. Allen^{1,2}, Omid Bakhshandeh³, Nasrin Mostafazadeh^{4*}

1 University of Rochester, 2 Institute for Human and Machine Cognition, 3 Verneek.ai 4 Elemental Cognition

rishi.sharma@rochester.edu, nasrinm@cs.rochester.edu

Example: Biases in Visual Question Answering



Q: What color is the grass?

A: Green



Q: What color is the banana?

A: Yellow



Q: What color is the skye?

A: Blue

So what is the issue ... ?

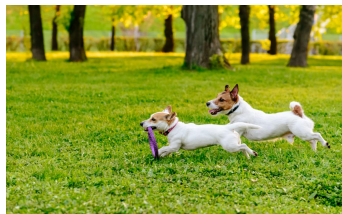
- A VQA system that fails to ground questions in image content would likely perform poorly in real-world settings



Q: What color is the banana?

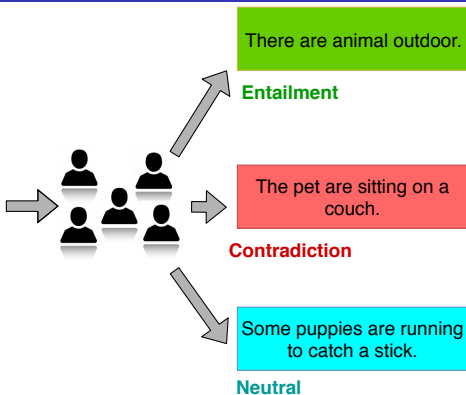
A: Yellow **X**

Example: Natural language Inference (NLI)



The dogs are running through the field.

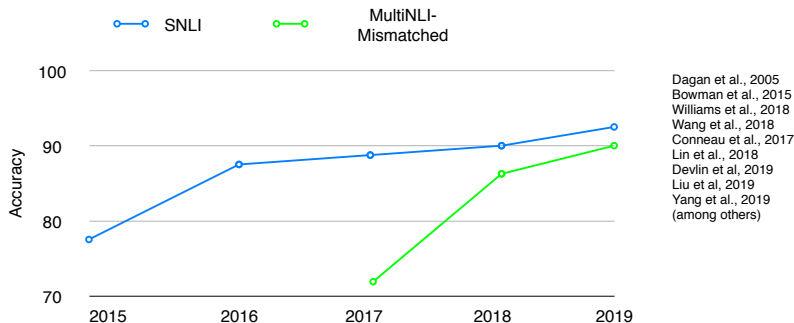
Premise



- SNLI (Bowman et. al, 2015) 570 K
- MultiNLI (Williams et. al., 2017) 433 K
- SNLI premises are Flickr captions.
- MultiNLI premises are collected from diverse genre.
- Hypotheses are crowdsource-generated.

Significant NLI Progress, almost human performance

- While NLI is a hard task, the community has made significant progress on large-scale NLI datasets.



Kicking out premises ...

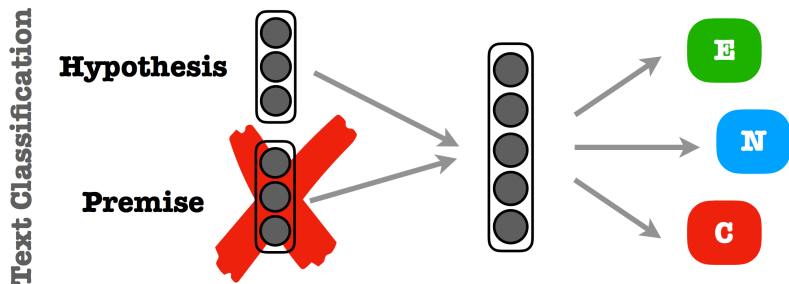


Figure: Figure from [GSL⁺18]

- Over 50% of NLI examples can be correctly classified without ever observing the premise!

Biases in NLI - Patterns in the hypothesis



A group of female athletes are gathered together and excited.

Purpose clauses



Neutral

They are gathered together because they are working together.



Some men and boys are playing frisbee in a grassy area.

Generalization



Entailment

People play frisbee outdoors.



A man with a black cap is looking at the street.

Negation



Contradiction

Nobody wears a cap.

Can we avoid biases?

- This is hard to avoid biases during the creation of datasets.
- Constructing new datasets, specially in large-scale is costly and still could results in other artifacts.
- This is important to develop techniques which to prevent models from using known biases to be able to leverage existing datasets
- Goal: train robust model to improve their generalization performance on evaluation phase, where typical biases observed in the training data do not exist.

Overview of Our Model

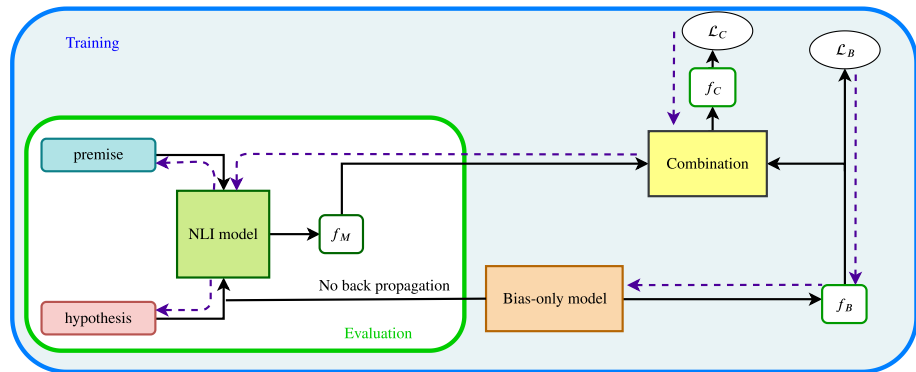
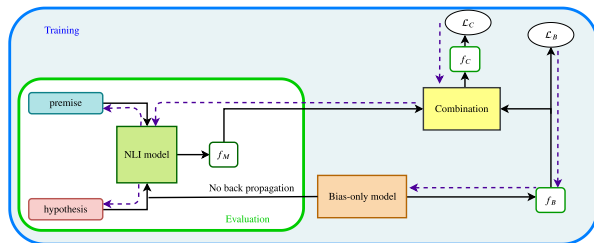


Figure: An illustration of our debiasing strategies on NLI. Solid arrows show the flow of input information, and dotted arrows show the back-propagation flow of error. Blue highlighted modules are removed after training. At test time, only the predictions of the base model f_M are used.

Steps to make the models robust to biases ...



- Identify the biases
- Train the bias-only branch f_B .
- Compute the combination of the two models f_C
 - Motivate the base model to learn different strategies than the ones used by the bias-only branch f_B .
- Remove the bias-only classifier and use the predictions of the base model.

Step 1: Bias-only Model

- Fortunately often times, we know what are the domain-specific biases
- Train the bias-only model using only biased features

Hypothesis

A woman is **not** taking money for any of her sticks.
A boy with **no** shirt on throws rocks.
A man is **asleep** and dreaming while sitting on a bench.
A **naked** man is posing on a ski board with snow in the background.

Labels ?

f_B

Contradiction

Step 2: Training a Robust Model

- Classical learning strategy:

$$\mathcal{L}(\theta_M) = -\frac{1}{N} \sum_{i=1}^N a_i \log(\text{softmax}(f_M(x_i))), \quad (1)$$

- Down-weighting the impact of the biased examples so that the model focuses on learning hard examples.
- Avoid major gradient updates from trivial predictions.
- Ensemble techniques:
 - Method 1: Product of experts [Hin02]
 - Method 2: RUBI [CDBy⁺19]
- Weight the loss of the base model depending on the accuracy of the bias-only model
 - Method 3: Debiased Focal Loss

Method 1: Product of Experts

- Combine multiple probabilistic models of the same data by multiplying the probabilities together and then renormalizing.
- Combine the bias-only and base model predictions:

$$f_C(x_i, x_i^b) = f_B(x_i^b) \odot f_M(x_i), \quad (2)$$

x_i^b is the biased features, and x_i is the whole input.

- Update the model parameters based on the cross-entropy loss of the combined classifier.

- Apply a sigmoid function to the bias-only model's predictions to obtain a mask containing an importance weight between 0 and 1 for each possible label.

$$f_C(x_i, x_i^b) = f_M(x_i) \odot \sigma(f_B(x_i^b)), \quad (3)$$

Method 2: RUBI [CDBy⁺19]

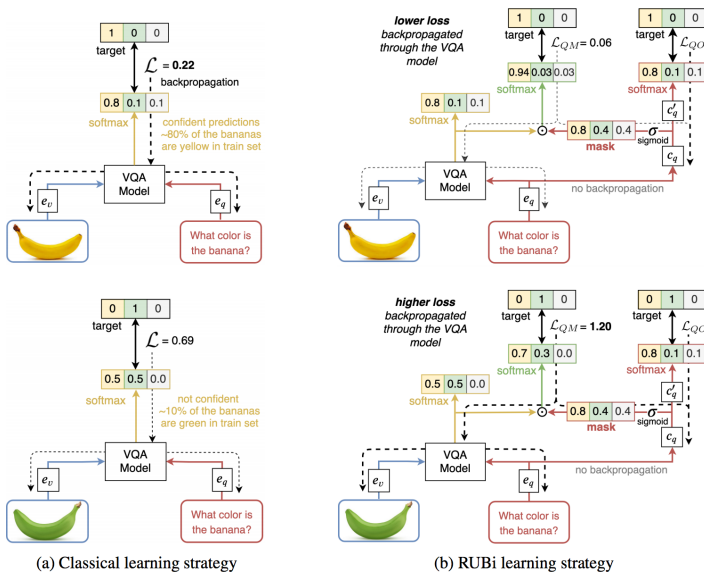


Figure: Detailed illustration of the RUBI impact on the learning [CDBy⁺19].

Debiased Focal Loss

- Explicitly modulating the loss depending on the accuracy of the bias-only model:

$$\mathcal{L}_C(\theta_M; \theta_B) = -\frac{1}{N} \sum_{i=1}^N a_i (1 - f_B(x_i^b))^{\gamma} \log(f_M(x_i)), \quad (4)$$

observations

- When the example is unbiased, and bias-only branch does not do well, $f_B(x_i^b)$ is small, and the loss remains unaffected.
- As the sample is more biased and $f_B(x_i^b)$ is closer to 1, the loss for the most biased examples is down-weighted.

Evaluation of Generalization Performance

- We train our models on two large-scale NLI datasets, namely SNLI and MNLI, and FEVER dataset.
- Evaluate performance on the challenging unbiased datasets.

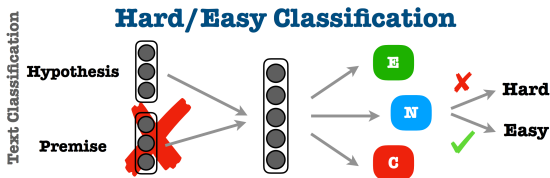


Figure: Figure from [GSL⁺18]

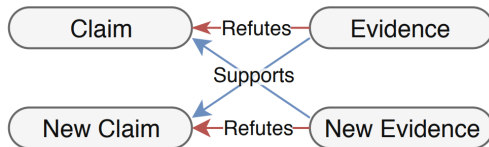


Figure: Figure from [SJSJSY⁺19]

Experimental Results - Fact Verification

- Obtaining 9.76 points gain on FEVER symmetric test set, improving the results of prior work by 4.65 points.

Table: Results on FEVER development (Dev) set and FEVER symmetric test set.

Debiasing method	Dev	Symmetric test set
None	85.99	56.49
RUBI	86.23	57.60
Debiased Focal Loss	83.07	64.02
Product of experts	86.46	66.25
[SJSJSY ⁺ 19]	84.6	61.6

Experimental Results - MNLI

Table: Results on MNLI matched (MNLI) and mismatched (MNLI-M) sets.

Debiasing Method	MNLI		MNLI-M	
	Test	Hard	Test	Hard
None	84.11	75.88	83.51	75.75
Product of experts	84.11	76.81	83.47	76.83

Table: Results on MNLI matched and HANS datasets

Debiasing Method	MNLI	HANS	Constituent	Lexical	Subsequence
None	83.99	61.10	61.11	68.97	53.21
RUBI	83.93	60.35	56.51	71.09	53.44
Debiased Focal Loss	84.33	64.99	62.42	74.45	58.11
Product of experts	84.04	66.55	64.29	77.61	57.75

Experimental Results - SNLI

- Gain of 4.78 points on SNLI hard set.

Table: Results on SNLI and SNLI hard sets.





Debiasing method	BERT		InferSent	
	Test	Hard	Test	Hard
None	90.53	80.53	84.24	68.91
RUBI	90.69	80.62	83.93	69.64
Debiased Focal Loss	89.57	83.01	73.54	73.05
Product of experts	90.11	82.15	80.35	73.69
AdvCls belinkov2019adversarial	-	-	83.56	66.27
AdvDat belinkov2019adversarial	-	-	78.30	55.60

Takeaways

- High performance of neural models could be due to leveraging superficial cues in the data.
- This is hard to avoid biases during creation of datasets.
- We need to develop methods robust to existing biases
- Let bias-only model capture the biases and we adjust cross-entropy loss to focus learning on the hard examples.
- Substantial improvement in the model robustness and better generalization performance.

Thank you. Any questions?

References I

-  Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh, *Rubi: Reducing unimodal biases in visual question answering*, Advances in neural information processing systems, 2019.
-  Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith, *Annotation artifacts in natural language inference data*, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, 2018.
-  Geoffrey E Hinton, *Training products of experts by minimizing contrastive divergence*, Neural computation (2002).
-  Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay, *Towards debiasing fact verification models*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019.