Project LIT - Outline

Erik Imgrund, Niklas Loeser, Andre Trump

February 17, 2023

1 Task Description

What is the aim - in view of the specific phenomenon under consideration? Identify linguistic biases of current models for Natural Language Inference (NLI). Improve Language Models (LMs) by removing biases from the training process.

Hypotheses

- Current language models are biased for NLI in the zero-shot and fine-tuned settings.
- Removing biases from the dataset improves results in a less biased model.
- A less biased model results in worse accuracy.

2 Method

How will we approach this aim? First as a baseline we will test the bias and accuracy of a model in a zero-shot and fine-tuned manner. The bias is

tested by calculating plausibility and faithfulness (Attanasio et al. 2022) for the explanations provided as well as manual inspection of handpicked examples. By analyzing the baseline models and in particular their explanations we aim to identify linguistic biases in their predictions. Using those identified linguistic biases we want to remove heavily biased examples for the training dataset and train a model on the de-biased data. The resulting model is compared to the baseline models.

Which methods will we apply? A pretrained BERT model (Devlin et al. 2018) is used for the zero-shot tasks as well as basis for the fine-tuning tasks. Integrated Gradients (Sundararajan, Taly, and Yan 2017), LIME (Ribeiro, Singh, and Guestrin 2016) and Partition SHAP Values (Lundberg and Lee 2017) are used as explanation methods. Two approaches to removing the biased examples are tested (Clark, Yatskar, and Zettlemoyer 2019): The biased data can be reweighed based on how biased it is such that biased examples influence the fine-tuning less or more depending on how biased the example is. A different approach is based on ensembling the fine-tuning model during training with a biased model so that the model does not need to learn those.

How to probe or fine-tune for your task? For the zero-shot task two methods are tried. The next-sentence-prediction head of the pretrained BERT model is used or discourse relation markers between the premise and hypothesis are predicted using a pretrained DisSent model (Nie, Bennett, and Goodman 2017). Each discourse marker is assigned to either entailment, neutral or contradiction and the most probable discourse marker is used to assign a prediction of the model.

For fine-tuning the premise and hypothesis are both fed into a LM separated by a marker token and a text classification head is trained based on the embedding obtained from the model to predict the class of the combined text. That is either entailment, neutral or contradiction.

3 Models and Data Sets

Select suitable data and resources To test the bias of the model we plan to use e-SNLI (Camburu et al. 2018) for large-scale calculation of plausibility and faithfulness of the explanations the models provide. For testing the accuracy and manually analyzing the bias of the models we plan to use SICK (Marelli et al. 2014), as it is less biased than SNLI (Bowman et al. 2015) and MultiNLI (Williams, Nangia, and Bowman 2018). To fine-tune the model we use MultiNLI, as it is less biased than SNLI but is sufficiently large.

Additional resources to improve learning Additionally, the training set of SICK can be used to introduce data with less bias. Additional tests could be introduced to test specific biases identified by manual analysis.

4 Experiments

To evaluate the performance of the model on natural language inference we use the F1-Score and Matthews correlation coefficient (Matthews 1975). We evaluate the faithfulness and plausibility with the same metrics as described in Attanasio et al. 2022.

The models used for probing are Pretrained Language Models (PLMs) based on DisSent, which also provide the pretrained basis for the fine-tuned models.

References

Attanasio, Giuseppe et al. (2022). "ferret: a Framework for Benchmarking Explainers on Transformers". In: *arXiv preprint arXiv:2208.01575*.

Bowman, Samuel R et al. (2015). "A large annotated corpus for learning natural language inference". In: *arXiv preprint arXiv:1508.05326*.

Camburu, Oana-Maria et al. (2018). "e-snli: Natural language inference with natural language explanations". In: *Advances in Neural Information Processing Systems* 31.

Clark, Christopher, Mark Yatskar, and Luke Zettlemoyer (2019). "Don't take the easy way out: Ensemble based methods for avoiding known dataset biases". In: *arXiv preprint arXiv:1909.03683*.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Marelli, Marco et al. (2014). "A SICK cure for the evaluation of compositional distributional semantic models." In: *Lrec*. Reykjavik, pp. 216–223.

Matthews, Brian W (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2, pp. 442–451.

Nie, Allen, Erin D Bennett, and Noah D Goodman (2017). "Dissent: Sentence representation learning from explicit discourse relations". In: *arXiv preprint arXiv:1710.04334*.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR, pp. 3319–3328.

Williams, Adina, Nikita Nangia, and Samuel R Bowman (2018). "The multi-genre nli corpus". In.