Supporting information: Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training

Joe Stacey
University College London
ucakjd0@ucl.ac.uk

Pasquale Minervini
University College London
p.minervini@ucl.ac.uk

Haim Dubossarsky
University of Cambridge
hd423@cam.ac.uk

Sebastian Riedel
University College London
s.riedel@cs.ucl.ac.uk

Tim Rocktäschel
University College London
t.rocktaschel@cs.ucl.ac.uk

1 Model Run Time

A single experiment implementing the adversarial training for 2,048 dimensions with 20 adversarial classifiers typically lasted for 5 days on a single GPU. This time includes the initial adversarial training, in addition to freezing the representations and trying to learn the bias from the resulting model sentence representations.

Due to the large number of experiments performed for this paper (repeated trials for statistical testing, implementing adversarial training for each combination of dimensions and the number of adversaries, and finally testing these de-biased models on different datasets), over 10,000 GPU hours were required in total to complete the experimentation.

2 Hyper-Parameter Ranges

Hyper-parameters for the adversarial training were tested across λ values of 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99, 0.999. After demonstrating that the adversarial training improves performance across any λ hyper-parameter, a λ value of 0.5 was used throughout the experimentation (the highest performing λ value tested in terms of reducing the bias which does not result in a drop in model performance).

When testing the performance of the de-biased models on the 12 different NLI datasets, as per Belinkov et al. (2019), hyper-parameters were selected based on the model accuracy for each datasets dev set. This involved a grid-search over the following α and β parameters described by (Belinkov et al., 2019): 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5 and 5.0.

3 Datasets

The following datasets were used within the model experimentation: ADD-ONE-RTE (Pavlick and Callison-Burch, 2016), GLUE (Wang et al., 2018), JOCI (Zhang et al., 2017), MNLI (Williams et al., 2018), MPE (Lai et al., 2017), SCITAIL (Khot et al., 2018), SICK (Marelli et al., 2014), SNLI-hard (Gururangan et al., 2018), and three datasets recast by White et al. (2017): DPR (Rahman and Ng, 2012), FN+ (Pavlick et al., 2015) and SPR (Reisinger et al., 2015).

The test and dev splits of each dataset are provided in the experiment code (please see 'Experiment code' below). The size of the train, dev and test sets for each dataset are also provided below:

- **SNLI** has 549,360 examples in the training set, 9,842 in the dev set and 9,824 in the test set.
- **SNLI-hard** has 3,261 examples in this test set.
- **ADD-ONE-RTE** has 4,481 examples in the training set, 510 in the dev set and 387 in the test set.
- **HANS** has 15,000 examples in the training set, dev set and in the test set.
- **JOCI** has 26,492 examples in the training set, 3,311 in the dev set and 3,311 in the test set.
- **MPE** has 32,000 examples in the training set, 4,000 in the dev set and 4,000 in the test set.
- **SICK** has 4,439 examples in the training set, 495 in the dev set and 4,906 in the test set.
- **SCITAIL** has 23,596 examples in the training set, 1,304 in the dev set and 2,126 in the test set.
- **DPR** has 2,080 examples in the training set, 486 in the dev set and 1,095 in the test set.
- **FN+** has 124,011 examples in the training set, 15,914 in the dev set and 14,679 in the test set.

- **SPR** has 123,0855 examples in the training set, 15,296 in the dev set and 15,456 in the test set.
- **MultiNLI matched** has 392,702 examples in the training set, 9,823 in the dev set and 9,815 in the test set.
- **GLUE** has 8,551 examples in the training set, 1,043 in the dev set and 1,063 in the test set.

4 Experiment code

The source code has been provided with details of how to repeat our experiments. This source code includes a link to download the datasets used in our experiments and the GloVe word embeddings that we use.

References

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander M. Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *ACL (1)*, pages 877–891. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT (2)*, pages 107–112. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5189–5197. AAAI Press.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 100–109. Asian Federation of Natural Language Processing.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223. European Language Resources Association (ELRA).

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. 2015. Framenet+: Fast paraphrastic tripling of framenet. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 408–413. The Association for Computer Linguistics.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 777–789. ACL.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Trans. Assoc. Comput. Linguistics*, 3:475–488.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 353–355. Association for Computational Linguistics.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005. Asian Federation of Natural Language Processing.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, pages 1112–1122. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal commonsense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.