

Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics

Prajjwal Bhargava



UT Dallas

🏠 prajjwal1.github.io

🐦 [prajjwal_1](https://twitter.com/prajjwal_1)

Aleksandr Drozd



RIKEN CCS

🏠 blackbird.pw

🐦 [bkbrd](https://twitter.com/bkbrd)

Anna Rogers



U Copenhagen

🏠 annargrs.github.io

🐦 [annargrs](https://twitter.com/annargrs)

Workshop on Insights from Negative Results in NLP
EMNLP 2021

The Premise: Many popular NLP datasets contain spurious patterns, which get learned instead of the actual task.

[Gururangan et al., 2018, Belinkov et al., 2019, Rogers et al., 2020, Gardner et al., 2021].

The Hypothesis: A number of methods reported in the literature and suggested by us are expected to alleviate this problem. Do they really work?

Case Study

Natural Language Inference (NLI) - 3-class classification task: does the premise entails, contradicts, or is neutral with respect to the hypothesis?

The Dataset

MNLI [Williams et al., 2018] - one of the most popular resources for this task, but it has been shown to suffer from both annotation artifacts [Gururangan et al., 2018, Poliak et al., 2018] and annotator bias [Geva et al., 2019].

The Adversary

HANS [McCoy et al., 2019]:

- ▶ synthetic dataset targeting *lexical overlap*, *subsequence* and *constituent* heuristics.
- ▶ model trained on MNLI is likely to learn these heuristics and thus predict the "entailment" label for most HANS examples.
 - ▶ predict "The doctor was paid by the actor" entails "The doctor paid the actor", simply because these sentences contain the same words.

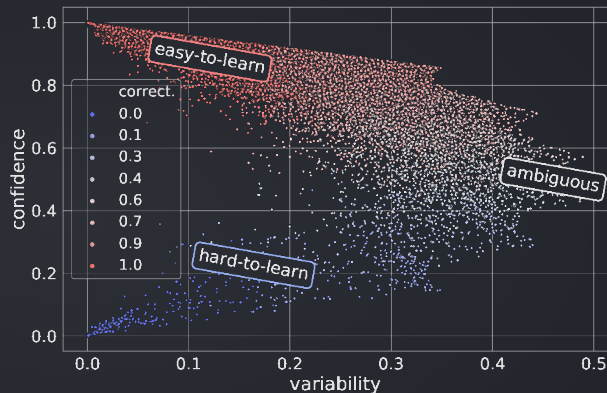


The Methods

- ▶ Sub-sampling based on dataset cartography
- ▶ Siamese Networks
- ▶ Adapters
- ▶ HEX debiasing
- ▶ Increasing model size

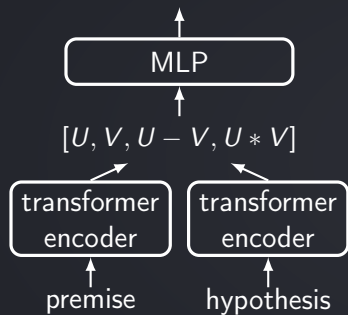
Dataset cartography [Swayamdipta et al., 2020]

characterize samples according to the model's confidence in the true class, and the variability of this confidence across epochs.



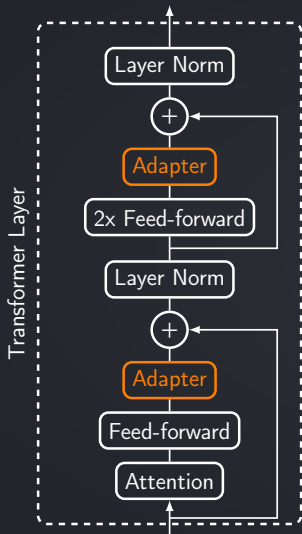
MNLI data map (with RoBERTa-large)

Siamese Networks



- ▶ learn representation for premise and hypothesis independently
- ▶ information bottleneck idea [Tishby et al., 2000, Alemi et al., 2016]

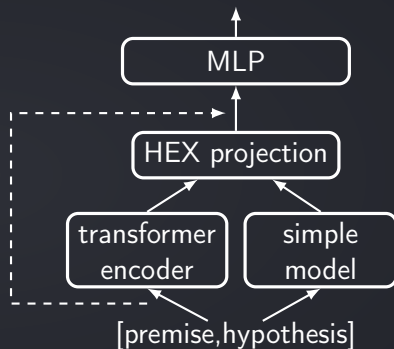
Adapters [Houlsby et al., 2019, Pfeiffer et al., 2020]:



- ▶ also tapping into the idea of bottlenecking.
- ▶ keep model weights, add trainable task-specific components in-between layers.

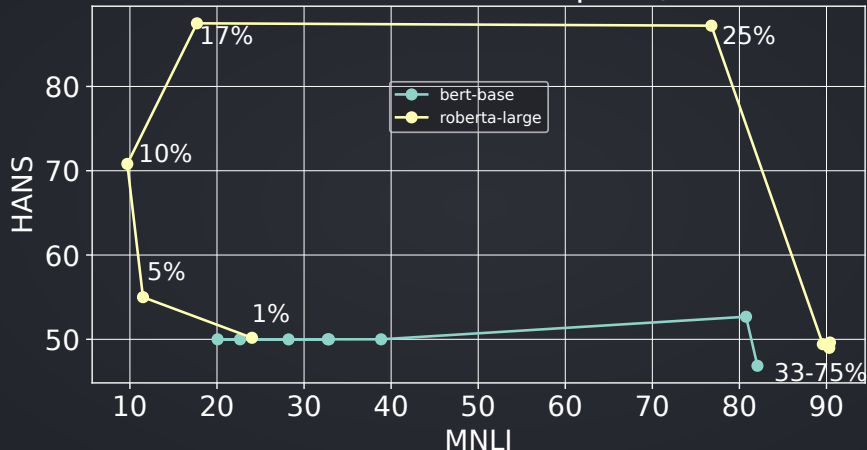
HEX debiasing [Wang et al., 2019]

- ▶ Model of less capacity is expected to rely more on superficial features
- ▶ Juxtapose representations learned by "small" and "big" models.
- ▶ some success reported in NLP [Zhou and Bansal, 2020]



Results: Cartography

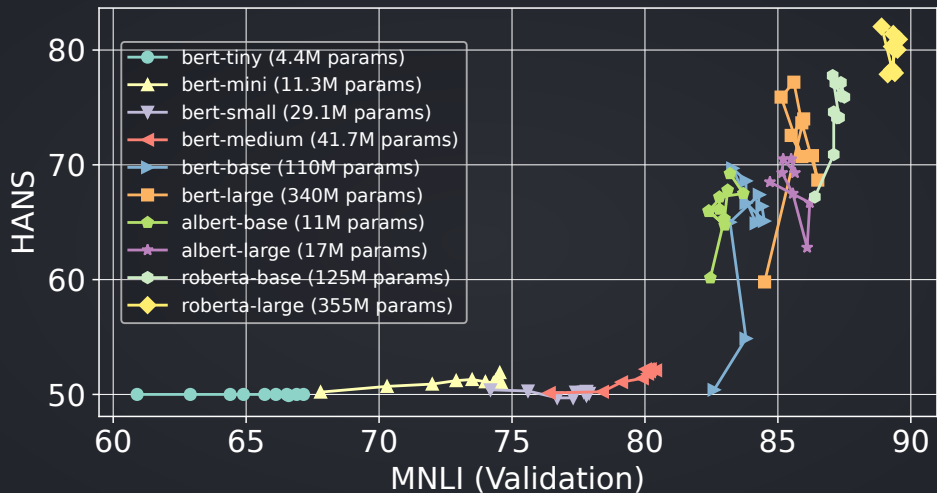
Models trained on hard MNLI samples (Lexical Overlap)



Results: Architectures

Architecture	Encoder	HF trainer			Custom Trainer		
		MNLI/std	HANS/std	runs	MNLI/std	HANS/std	runs
Siamese networks / frozen encoder	BERT-base	51.43	50.74	1	57.2 / 0.2	51.3 / 0.1	3
	BERT-large	51.72	51.12	1	61.4 / 0.1	51.6 / 0.1	5
Siamese networks / trainable encoder	BERT-base	58.9	52.79	1	76.5 / 0.03	51.3 / 0.03	3
	BERT-large	59.9	51.21	1	78.7	52.5	1
Adapter networks	BERT-base	82.6	50.97	1			
	BERT-large	84.75	57.17	1			
	RoBERTa-base	86.33	57.21	1			
	RoBERTa-large	90.4	75.93	1			
HEX debiasing	BERT-base	56.25	50.58	1			

Results: Model Size




Results: Model Size


Architecture	Encoder	HF trainer			Custom Trainer		
		MNLI/std	HANS/std	runs	MNLI/std	HANS/std	runs
Vanilla finetuning: increased model size	BERT-tiny (4.4M)	64.48/0.24	50/0	3	67.4 / 0.2	50 / 0.02	5
	BERT-mini (11.3M)	72.3/0.29	50.97/0.04	3	76.3 / 1	52.3 / 0.3	10
	BERT-small (29.1M)	76.48/0.12	50.39/0.14	3	78.4 / 0.5	51.1 / 0.3	5
	BERT-medium (41.7M)	79.64/0.14	51.02/0.26	3	80 / 0.3	52 / 0.4	5
	BERT-base (110M)	83.74/0.04	53.98/0.78	3	84 / 0.2	69 / 4	16
	BERT-large (340M)	85.9/0.02	72.04/1.97	3	86.5 / 0.1	77.8 / 2.4	3
	RoBERTa-base (125M)	87.46/0.1	73.11/1.13	3	87.5 / 0.3	77.7 / 1.7	10
	RoBERTa-large (355M)	90.3/0.07	79.95/0.56	3	90 / 0.4	82.05 / 1	3
	ALBERT-base-v2 (11M)	83.06/0.13	66.6/0.78	3	84.2 / 0.6	69.2 / 2.2	4
	ALBERT-large-v2 (17M)	85.08/0.3	70.64/2.91	3	85.5 / 0.9	70.5 / 1.6	4


Few more observations

- ▶ Two independent implementations to address "negative results" concerns: the trend persists but difference in accuracy is noticeable.
- ▶ While MNLI validation reaches close to SOTA accuracy in one epoch, HANs benefits from training for up to 10 epochs.

Thank you for your attention!

Paper:  <https://arxiv.org/abs/2110.01518>

Code:  <https://github.com/vecto-ai/langmo>

Code:  https://github.com/prajjwal1/generalize_lm_nli



Bibliography I



Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2016).
Deep Variational Information Bottleneck.



Belinkov, Y., Poliak, A., Shieber, S., Van Durme, B., and Rush, A. (2019).
Don't take the premise for granted: Mitigating artifacts in natural language inference.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.



Gardner, M., Merrill, W., Dodge, J., Peters, M. E., Ross, A., Singh, S., and Smith, N. (2021).
Competency Problems: On Finding and Removing Artifacts in Language Data.
arXiv:2104.08646 [cs].



Geva, M., Goldberg, Y., and Berant, J. (2019).
Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.



Gururangan, S., Swamydipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018).
Annotation artifacts in natural language inference data.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.



Houlsby, N., Giurui, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019).
Parameter-efficient transfer learning for NLP.
In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799, Long Beach, California, USA. PMLR.

Bibliography II



McCoy, T., Pavlick, E., and Linzen, T. (2019).

Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.



Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., and Gurevych, I. (2020).

Adapterfusion: Non-destructive task composition for transfer learning.



Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018).

Hypothesis only baselines in natural language inference.

In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.



Rogers, A., Kovaleva, O., Downey, M., and Rumshisky, A. (2020).

Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8722–8731.



Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. (2020).

Dataset cartography: Mapping and diagnosing datasets with training dynamics.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.



Tishby, N., Pereira, F. C., and Bialek, W. (2000).

The information bottleneck method.

arXiv:physics/0004057.

Bibliography III



Wang, H., He, Z., Lipton, Z. L., and Xing, E. P. (2019).
Learning robust representations by projecting superficial statistics out.
In *International Conference on Learning Representations*.



Williams, A., Nangia, N., and Bowman, S. (2018).
A broad-coverage challenge corpus for sentence understanding through inference.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.



Zhou, X. and Bansal, M. (2020).
Towards robustifying NLI models against lexical dataset biases.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771, Online. Association for Computational Linguistics.