

Does the model really learn semantic information?

Weizhe Tang

The University of Texas at Austin

wt4992@utexas.edu

Abstract

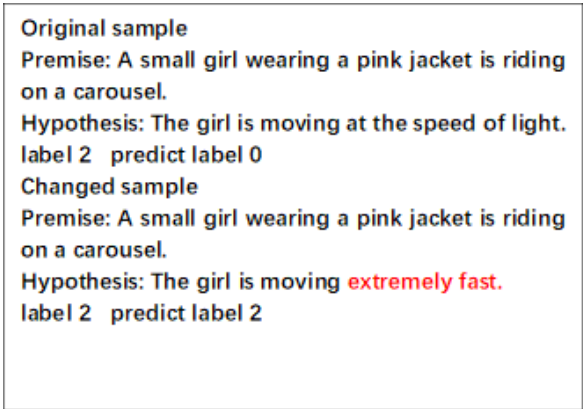
NLI is now a very popular research field, but a large number of researchers have raised questions about whether NLI has learned semantic information in the true sense.

We have conducted research on SNLI and ANLI, and found that the model cannot learn the semantic information of natural language processing well. And based on Textattack technology for data amplification and construction of new loss. The above two methods will increase the baseline ACC from 88.97% to 90.02% on the SNLI dataset, and from 37.9% to 42.9% on the ANLI dataset.

1 Introduction

NLI (Natural Language Inference) is a hot research field in recent years, and people have done a lot of research in this field, such as Semantics-aware BERT (SemBERT), which is capable of explicitly absorbing contextual semantics over a BERT backbone. (Zhang et al., 2020), multi-Task Learning (MTL) networks with transformers (Piliault et al., 2020). However, some scholars have proposed whether the NLI model can really learn natural language instead of statistics on the characteristics of the dataset. (Poliak et al., 2018)

We implemented the Electra-small (Clark et al., 2019) discriminator for finetuning as our baseline. and select the SNLI (Bowman et al., 2015) dataset as our dataset. In the process of analyzing the output results of the model, we found that the model will perform very poorly when the hypothetical sentence contains common sense, but when we use common sense to change the description with simpler sentences, the model can succeed identify it (Figure1).



Original sample
Premise: A small girl wearing a pink jacket is riding on a carousel.
Hypothesis: The girl is moving at the speed of light.
label 2 predict label 0

Changed sample
Premise: A small girl wearing a pink jacket is riding on a carousel.
Hypothesis: The girl is moving extremely fast.
label 2 predict label 2

Figure 1: An example that the original sentence contains some common sense information, which can be interpreted as simple facts. After explaining and transforming part of the sentence that contain common sense, the model can inference well

In order to better evaluate whether the model can solve the above situation, we manually constructed a subset of dataset, in which the sample is mainly constructed by sample from SNLI that contain common sense. It is used to evaluate whether the model can learn common sense information.

Due to the small number of such samples, we also introduced ANLI dataset, collected via an iterative, adversarial human-and-model-in-the-loop procedure (Nie et al., 2020).

After analyzing the inference results of the baseline model and the training set, we found that the samples containing common sense are less in the training set, accounting for about 2% of the original dataset (the sampling statistics are not complete statistics), so we use Textattack to train the samples in the collection have been data augmented (Morris et al., 2020).

On the other hand, by analyzing the results of the confusion matrix of the baseline, I found that the model performs poorly on the neutral label and is evenly distributed on both sides of entailment

and contradiction. However, the error samples of entailment are mostly distributed in neutral and less contradiction. Therefore, I think that these three labels have a certain linear strength relationship. In order to make the model perform better, I use hinge loss instead of the original cross entropy loss.

Our electra small based model achieved a score of 90.02% on SNLI, which is higher than (Kim et al., 2019), and my parameters are 3.13 times less than theirs. At the same time, this is also higher than 88.97% of our baseline in this article.

At the same time, in the ANLI test set, the baseline achieved a score of 28%, and our model achieved a score of 33%.

2 Adversarial NLI

2.1 Dataset

A growing body of evidence shows that most of models does not learning meaning in flexible and robust way, but learn to exploit spurious statistical patterns in datasets (Mccoy et al., 2019; Geva et al., 2019; Glockner et al., 2018; Tsuchiya, 2018; Poliak et al., 2018; Gururangan et al., 2018).

Therefore, for the NLI task, a difficult dataset constructed by human annotators is very important. We use ANLI as our test dataset. In order to test the true performance of the model, the model will not be trained on the ANLI training set, but will only be tested on the ANLI test set.

At the same time, we found that the baseline model does not perform well when it comes to common sense issues. We extracted some common sense issues from SNLI and added them to the ANLI dataset to jointly generate a new dataset.

Premise: Fido Dido is a cartoon character created by Joanna Ferrone and Sue Rose. Rose first developed the character in 1985, on a napkin in a restaurant. They later stenciled Fido on T-shirts with the credo: "Fido is for Fido, Fido is against no one". These T-shirts became very popular in New York.
Hypothesis: Fido Dido was first developed in the 20th century.
label 0 predict label 2

Figure 2: One of the ANLI test set, through an iterative, adversarial human-and-model-in-the-loop solution for NLU dataset collection, the model will significantly misjudge such a sample.

2.2 Performance

We first do not use the training set in the ANLI dataset for training, but train on the SNLI dataset and test on the ANLI test set. The results are shown in Table 1

In the SNLI test set, the Baseline model performed better, but in the ANLI test set, the performance of the Baseline model was severely degraded. We believe that he has not learned natural language information in a real sense, it can also be proved by observing the results of model inference (Figure 2).

On the other hand, we also studied the confusion matrix of model reasoning, as shown in Table 2,3. In SNLI datasets the error samples of entailment are mostly distributed in neutral and less contradiction the opposite is true in ANLI dataset. This shows that the specially designed dataset for confrontation has successfully deceived the model.

	Baseline
SNLI	88.97%
ANLI	28.9%

Table 1: Average ACC of Baseline training on SNLI

	Entailment	Neutral	Contradiction
Entailment	3021	235	73
Neutral	244	2707	284
Contradiction	75	213	2990

Table 2: confusion matrix of SNLI

	Entailment	Neutral	Contradiction
Entailment	86	83	165
Neutral	75	90	168
Contradiction	128	92	113

Table 3: confusion matrix of ANLI

3 Approach

3.1 Data augmentation

We use the wordnet technology (Figure3) in Textattack to augment the premise part of the dataset, and there are a large number of data augmentation methods to choose. For example, charswap, embedding (augments text by replacing words with neighbors in the counter-fitted embedding space, with a constraint to ensure their cosine similarity is at least 0.8). The reason for not choosing the embedding method is that even it ensures words

	Baseline	Electra-small + data augmentation	Electra-small + hinge loss	Electra-small +hinge loss + data augmentation	Electra-base + hinge loss + data augmentation
SNLI	88.97%	89.82%	89.94%	90.02%	90.76%
ANLI without train on ANLI	28.93%	-	-	34.77%	35.22%
ANLI	37.99%	-	-	42.39%	44.73%

Table 4: Experiment result

cosine similarity is at least 0.8. There is no guarantee that the semantics will not change greatly, which is particularly important in semantic inference. Charsawp is a kind of interference for the NLI task, even if such data appears in our dataset (Figure4).

Sample
Premise: Two men are engaging in some sort of combat sport.
Changed sample
Premise: Two men are engaging in some sort of combat play.

Figure 3: Sample after augmentation

Sample1
Premise: A woman huge a fluffy white dog.
Hypothesis: The dog is embraced by the woman
label 0
Sample2
Premise: A woman huge a fluffy white dog.
Hypothesis: The dog is embraced by the woman
label 1

Figure 4: Two example in SNLI dataset

3.2 Hinge loss

According to the observation results in Table 2, we can find that most of the error samples are distributed in the neutral category, so we hope to use hinge loss to better achieve the classification of these three situations.

Higeloss was first used in svm. In the current task, we expect that the three categories can be more classified and better. Therefore, in this task, hinge loss is a better choice than cross-entropy.

3.3 Bigger model

In addition to the above methods, we also tried to use a larger model to obtain better performance, we finally tried electra-base, and achieved good results. The ablation experiments show that the above methods have improved the model to varying degrees.

4 Experiments

Through the results of the SNLI dataset, we can see that different methods have different effects on the effect. Finally, using the Electra-base, hinge loss and data augmentation methods, our model got a score of 90.76%.

In the second line, that is not to retrain on the ANLI dataset, only the results of the test on the ANLI test set can be seen. Data expansion and the use of hinge loss can effectively improve the robustness of the model, and can better help the model learn semantic information instead of statistical features.

Finally, we also tried to experiment with ANLI as a new dataset. The model got a correct rate of 44.73%, far exceeding the 37.99% of the baseline.

5 Conclusion

In general, we have conducted certain research on the NLI task and confirmed that it is difficult for the model to learn semantic information on the SNLI dataset, but to learn statistical features. When we are migrating the dataset, the sharp drop in the accuracy rate can explain this problem.

For this we use wordnet, a data augmentation method, which makes the model more robust. At the same time, we also introduced hinge loss in this task, which can really help the model have better performance. Finally, we also tried to use the above methods on a larger model (Electra-base), and finally proved that these two methods are not

only effective on small models, but also effective on larger models.

However, these methods cannot perfectly solve the NLI problem. How to better enable the model to learn the real semantic information, rather than based on the statistical characteristics of the dataset, will be a significant research direction. The recent multi-task learning, I think it is an interesting and effective way.

Acknowledgments

We'd like to thank Dr. Durrett and the TAs for running such a great and helpful NLP class!

References

- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.
- Mor Geva, Yoav Goldberg, and Jonathan Berant. 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.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.
- Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6586–6593.
- R. T. McCoy, E. Pavlick, and T. Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.
- John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Jonathan Pilault, Christopher Pal, et al. 2020. Conditionally adaptive multi-task learning: Improving transfer learning in nlp using fewer parameters & less data. In *International Conference on Learning Representations*.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *NAACL HLT 2018*, page 180.
- Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.