Attention Intervention for Syntactic Knowledge Transfer

Cheongwoong Kang
KAIST, South Korea
cw.kang@kaist.ac.kr

Abstract

Pre-trained language models have led to recent success in various NLP tasks. Although there have been many efforts to analyze what linguistic properties are learned inside, how they are used for prediction is still unclear. In this work, we propose a way to analyze whether attention patterns encode syntactic knowledge or not. Specifically, we intervene attention weights to correct wrong predictions. We test the proposed method with the BERT base model on natural language inference tasks. The experimental results show that our method substantially improves the performance on a challenge set without additional training. This implies that attention weights play an important role in carrying syntactic knowledge.

1 Introduction

Pre-trained language models have shown outstanding performance in various NLP tasks (Devlin et al., 2019; Radford et al., 2018). BERT is one of the most popular pre-trained language models that leverages Transformer architecture (Vaswani et al., 2017) to learn text representation. Despite the success of BERT-based models (Lan et al., 2019; Liu et al., 2019), the exact mechanisms have not been well studied.

There has been ongoing debate in the field of artificial intelligence and machine learning about the role of attention mechanisms in neural networks and how they can be used to explain the behavior of these models (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019). One view is that attention mechanisms, which allow a model to focus on specific parts of its input while processing it, can provide a form of "explanation" for the model's predictions. According to this view, the attention weights learned by the model can be used to identify which parts of the input were most important in making a particular prediction, and this information can be used to understand and interpret the model's behavior. However, there are also counterarguments to this view. Some researchers have pointed out that attention weights may not always correspond to the true factors that influenced the model's prediction, and that other approaches, such as post-hoc explanations based on input perturbations or counterfactual examples, may be more effective at explaining model behavior.

In this work, we suggest a method for determining whether attention patterns in a machine learning model represent syntactic knowledge. To do this, we modify attention weights to see if wrong predictions can be corrected. We focus on syntactic heuristics in HANS dataset (McCoy et al., 2019).

2 Background

2.1 Natural Language Inference

Natural Language Inference (NLI) is a task that involves determining the relationship between a premise and a hypothesis. The premise is a statement that provides some context or background information, and the hypothesis is a statement that is either entailed by, contradicts, or is neutral with respect to the premise.

For example, given the premise "A cat is running on a crosswalk," the following hypotheses would have the following relationships to the premise:

- "A cat is crossing the street." (Entailment)
- "A cat is flying." (Contradiction)
- "A dog is running" (Neutral)

2.2 Heuristic Analysis for NLI Systems

HANS (Heuristic Analysis for NLI Systems) is a dataset proposed to diagnose syntactic heuristics in NLI. It is considered a challenge set since BERT fails to solve most of problems in HANS. To verify whether attention intervention can correct wrong predictions, we use HANS for experiments.

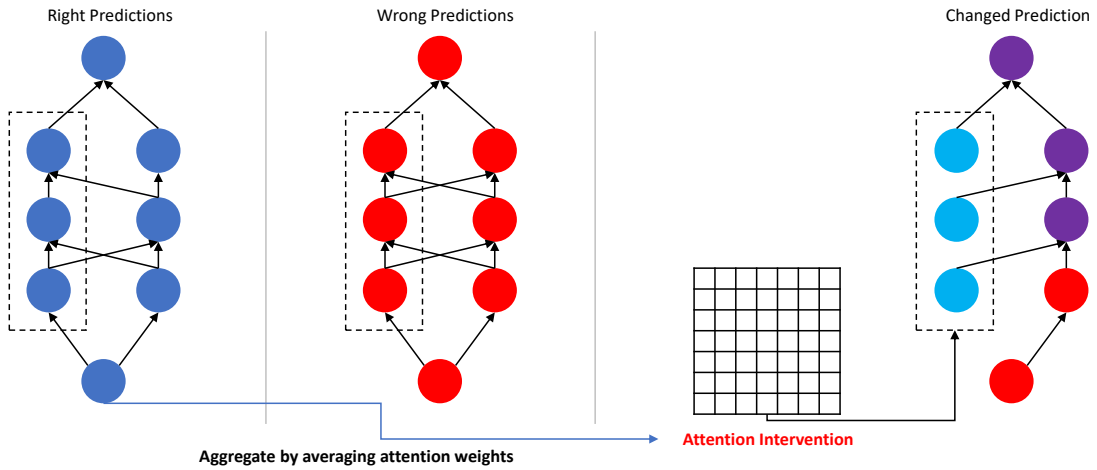

Figure 1: The overall procedure of attention intervention.

3 Methodology

3.1 Attention Patterns Collection

First, we categorize samples into two groups based on model predictions: (1) right, (2) wrong. Here, both "contradiction" and "neutral" are considered as "non-entailment" since the HANS dataset only contains two labels "entailment" and "non-entailment". Then, we collect attention weights in the "right" group. Since attention patterns may vary depending on input sequence lengths, we collect them separately for each template (there are 68 templates in total). Finally, we aggregate them by averaging, as shown in the left part of Figure 1.

3.2 Attention Intervention

We exploit the collected attention patterns with attention intervention. As shown in the right part of Figure 1, the attention weights are replaced with the collected attention patterns. We hypothesize that intervening attention weights may change the prediction towards a desired direction.

4 Experiments

4.1 Setting

We study a BERT base uncased model fine-tuned to MultiNLI (Wang et al., 2019), one of the most popular NLI datasets. We test the model on the validation set of HANS. We apply attention intervention on both "right" and "wrong" groups to analyze the effects on both groups.

	Entailment	Non-entailment
Baseline	0.99	0.14
+ Ours	0.96	0.43

Table 1: The accuracy before and after applying our proposed method (attention intervention).

4.2 Results

Table 1 shows the per-class accuracy before and after applying attention intervention. Here, "Baseline" denotes the BERT base uncased model and "Ours" denotes attention intervention. The experimental results show that attention intervention substantially improves accuracy, implying that syntactic knowledge indeed flows through attention weights.

5 Future Work

There are several possible future research directions. An example is filtering out samples on which the model makes ambiguous predictions when collecting attention patterns. Another direction is to analyze what makes the intervention successful or not. Another one is extending the proposed approach to analyze other types of knowledge, such as semantic knowledge.

6 Conclusion

In this work, we show that attention intervention is effective to change predictions to a desired direction.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20.