JAIST Repository
https://dspace.jaist.ac.jp/

Title	Improving Robustness of Pre-trained Language Models by Counterfactual Explanations
Author(s)	LUU, Linh Hoai
Citation	
Issue Date	2023-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/18738
Rights	
Description	Supervisor:井之上 直也, 先端科学技術研究科, 修士(情報科学)

Japan Advanced Institute of Science and Technology

Improving Robustness of Pre-trained Language Models by Counterfactual Explanations

2110440 LUU, Linh Hoai

Recent natural language processing (NLP) techniques have achieved high performance on various NLP benchmark datasets, primarily due to the significant improvement of deep learning [2]. However, the research community has demonstrated that the NLP models are vulnerable to adversarial attacks [1] and that they are susceptible to adversarial examples and tend to make incorrect predictions.

One of the existing approaches for improving the robustness of a model is adversarial training: fine-tuning models on adversarially perturbed examples that are generated from training instances. However, these perturbed examples may not be optimal enough to fool target models because the perturbed examples were not guaranteed to be minimally edited from original instances and to change the target model's prediction.

Our hypothesis is adversarial training could make models more robust if the adversarially perturbed examples have such guarantees. In Explainable Artificial Intelligence, such perturbed examples are known as counterfactual explanations. In our work, we investigate the potential of counterfactual explanations for improving the robustness of NLP models.

Our contributions are summarized as follows:

- We introduce *counterfactual adversarial training*, a new approach to adversarial training–using counterfactual explanations to improve the robustness of NLP models.
- We show that the counterfactual adversarial training improves the robustness of the original model on Natural Language Inference (NLI) and Sentiment Analysis (SA), two representative NLP tasks, in both in-domain and out-of-domain settings.
- We provide an in-depth behavior analysis of counterfactual adversarial training.

References

[1] John X 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. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.*

[2] Marwan Omar, Soohyeon Choi, DaeHun Nyang, and David Mohaisen. 2022. Robust natural language processing: Recent advances, challenges, and future directions. *IEEE Access*.