Robustness Evaluation on Deep Models: A Case Study of Adversarial Attacks on DeepMatcher

Daren Chao
University of Toronto
drchao@cs.toronto.edu

Bowen Zhang
University of Toronto
bwenzhang@cs.toronto.edu

Yibin Zhang
University of Toronto
ybzhang@cs.toronto.edu

Abstract

Although deep neural networks provide state-of-the-art results for most machine learning tasks, such as natural language processing, recent research has exposed that those are vulnerable to adversarial attacks. The existence of such adversarial examples implies the fragility of deep models. In this project, we will apply adversarial attacks on DeepMatcher, a deep learning (DL) model for entity resolution problem, to explore the methods of attacking a model based on natural language, instead of images, and evaluate its robustness based on the adversarial attack.

1 Introduction

Deep neural networks have become increasingly effective in many difficult artificial intelligence tasks. However, in the last few years, researchers have noticed that most of existing deep models are vulnerable to adversarial attacks. [4] firstly showed that it is possible to transform an image imperceptibly and thereby change how the image is classified. Therefore, a human-level accurate deep model might be fragile to adversarial attacks and incurs security issues as its actual behaviors might not follow our expectations.

In this project, we plan to find a proper method to generate adversarial attacks on a specific model, DeepMatcher [3], which proposed four DL solutions to tackle Entity Resolution (ER) problem. ER is a well-known research problem in the database community which aims at matching records that refer to the same real-world entity (e.g., to match products sold on multiple websites). DeepMatcher shows that it outperforms traditional solutions on the evaluation of precision, recall, and F-measure. Nevertheless, its robustness, i.e., the ability to defend against adversarial samples, is not reported yet.

The difficulty of this project is 1) generating text-based adversarial samples, as the input of DeepMatcher is two text-based tuples, and the output is a binary classification; 2) defining a new formulation of robustness; 3) proposing defense methods on text-based models.

2 Related works

While the research of adversarial attacks has a long history in machine learning, it has recently received a lot of attention in the deep learning community since [4] showed that adversarial attacks are powerful tools to investigate the vulnerabilities of a deep learning model. Studies on adversarial examples in the image domain have been well investigated, but in texts the research is not enough [5]. The ways to generate text-based adversarial attacks can be divided into four classes: char-level attacks, word-level attacks, sentence-level attacks, and multi-level attacks [2]. Instead of randomly perturbing text, we will use efficient hybrid multi-level methods to generate adversarial text, such as using the gradients of the model with respect to the input.

The text-based model cannot transplant most of the defense techniques from the image processing domain. Adversarial training interleaves training with the generation of adversarial examples , which

Preprint. Under review.

is the prevailing counter-measure to build a robust model. [6] proposed a general schema to block both word-level and character-level attacks on a text-based model.

In general, there are two different approaches one can take to evaluate the robustness of a neural network: prove a lower bound, or construct attacks that demonstrate an upper bound. However, [1] shows the need for better techniques to evaluate the robustness of neural networks.

3 Methodology

The ER is a text-based binary classification problem, with output *Match* or *NoMatch* and a pair of tuples (u, v) as input. In order to evaluate its robustness, we would like to perturb its input pairs and evaluate the accuracy of the perturbed examples. We pre-define a relative small *maximum perturbation* ϵ and assume the truth label does not change after the perturbation. Then, a test set is sufficient for our robustness evaluation.

There are different categories for text perturbations. Depends on the length of input tuple pairs, we mainly consider three of them: char-level, word-level, and hybrid perturbations. We apply an existing algorithm, the random flip, which is flipping each unit in the char or word level of the text with a probability, to generate text perturbations. The drawback of the random flip is very obvious, it is hard to find an adversarial example with minimal perturbation. In order to improve this, we would like to apply a gradient-based algorithm to generate perturbed examples. There are plenty of state-of-art gradient-based adversarial attack algorithms for attacking image classifiers, whose common idea is calculating the gradient of the loss function with respect to the input and do gradient ascent on input until the misclassification happens. Unfortunately, this idea may not be directly applicable to our project since our input space is non-differentiable discrete texts. As a result, one of our focuses will be developing a gradient-based algorithm for generating text perturbation effectively and efficiently.

In addition, we plan to boost DeepMatcher by adversarial training with the generated perturbation. We will analyze the changes from the original model to the boosted model and evaluate whether the boosted model is more robust.

4 Acknowledgement

We would like to thank the supervisor of the database group, Univ. of Toronto, Prof. Nick Koudas, for discussing the research direction[1]. Also, we would like to thank Yizheng Huang[2] for supporting us the background survey on this topic.

References

[1] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[2] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*, 2017.

[3] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra. Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 International Conference on Management of Data*, pages 19–34, 2018.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[5] W. Wang, L. Wang, B. Tang, R. Wang, and A. Ye. Towards a robust deep neural network in text domain a survey. *arXiv preprint arXiv:1902.07285*, 2019.

[6] Y. Zhou, J.-Y. Jiang, K.-W. Chang, and W. Wang. Learning to discriminate perturbations for blocking adversarial attacks in text classification. *arXiv preprint arXiv:1909.03084*, 2019.

[1]The authors are from the database group supervised by Prof. Nick Koudas. This project is related to their research interests but no previous work has been done prior to this proposal.

[2]Research staff of Nanyang Technological University, Singapore, yizheng.huang@ntu.edu.sg