Evaluation Report on NLP Adversarial Training

Zhang De

Github: https://github.com/TMIFI/Adversarial-Chinese-Text-Classification-Pytorch

Abstract

Adversarial training is initially proposed by computer vision community to boost the model robustness against intentional attack(twisted) on the input data as well as to improve the model generalization ability. According to previous CV research, such training trick could indeed improve the model robustness meanwhile unfortunately decrease the ability to correctly classify the non-adversarial data. Recently, NLP community has discovered that if the attack is placed on continuous embeddings to disregard the actual mapping between words, the adversarial training could be a handy tool to improve the model generalization ability. This experiment is conducted to implement different adversarial training methods with TextCNN model on text classification task. Three adversarial training methods (FGSM, PGD, and FREE) are implemented. The rest of this report will discuss the performance evaluation of each method.

Evaluation

There are three adversarial training methods implemented in this report, FGSM, PGD, and FREE. FGSM is proposed and later optimized to achieve comparable predicting power with much shorted training time than PGD and FREE. PGD is regarded as the strongest one among three competitors, but the training time complexity is the highest. FREE is designed to boost the training time of standard PGD, but the complexity is still not comparable with FGSM. The baseline method is training TextCNN without any adversarial training, and TextCNN model is the only used classifier here.

The evaluation metrics using here are precision, recall, and f1. Since, this text classification is processed on a multi-class dataset. Three scores mentioned above are calculated using macro average over all the classes in this report.

Chart (Table 1) below illustrates the performance of each method in details. We could observe that the result is consistent with the paper (fast is better than free: revisiting adversarial training). The PGD has the best model performance on all three metrics due to strong regularization. The FGSM and FREE unsurprisingly surpass the baseline. From time complexity perspective, we could observe that FGSM has least time consumption among three methods. PGD takes the most time, and FREE alleviates time consumption from PGD.

In conclusion, the result shows that adversarial training could boost the model generalization ability on unseen datasets in NLP fields.

Methods	Precision	Recall	F1	Time
TextCNN (Baseline)	0.9064	0.9061	0.9060	Nan
TextCNN + FGSM	0.9085	0.9078	0.9079	1:10:42
TextCNN + PGD	0.9170	0.9169	0.9168	3:24:00
TextCNN + FREE	0.9094	0.9086	0.9088	2:45:00

Table 1: Performance evaluation