对抗训练评估报告

1. Introduction

一般情况下，通过训练得到的网络可能是脆弱的，缺乏正确应对对抗例子的能力。对抗性训练是训练模型正确地分类未修改示例和对抗性示例的过程。该训练方式不仅提高了对对抗实例的鲁棒性，而且提高了对原始实例的泛化性能。

2 From CV to NLP

对抗训练的主要思想是提升模型对于对抗样本的错误率。因此对于对抗样本，需要朝着梯度增大的方向进行训练。对抗训练方式最开始都是基于 CV 领域的，对输入样本进行扰动从而得到对抗样本。和 NLP 不同的是，CV 输入是图像，可以视为普通的连续实数向量。在加上扰动 Δr 后仍然是有意义的图像。而 NLP 的输入是 one-hot 向量，因此无法将 CV 的对抗训练方法直接用于 NLP。为此，[1]中指出，可以对 embedding 层进行扰动，扰动得到的向量可以作为合理的对抗样本。关于几种不同的对抗训练方式，我们可以将其合理地应用到 NLP 上。

3.Expiremental Settings

从 THUCNews 中抽取了 20 万条新闻标题，文本长度在 20 到 30 之间。一共 10 个类别，每类 2 万条。以字为单位输入模型，使用了预训练词向量：搜狗新闻 Word+Character 300d。类别：财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐。Baseline 模型为 TextCNN。对抗训练方式包括 FGSM[2],FGM[1], PGD[3]和 Free[4]。使用 precision, recall 以及 f1-score 对训练结果进行评估。对于用于鲁棒性评估的测试集，参考了[6]中的方法。对测试集进行两次反向翻译，并随机替换或删除某些字符，得到受干扰后的测试集。

4. Results & Analysis

由 Figure1 可以看出，几种对抗训练的方式给模型带来了不同程度的提升，其中 Free 的提升最为明显，而 FGSM 的提升最小。

FGSM　　FGSM 对 baseline 的提升是几个对抗训练中最小的一个。表明简单地使用符号函数来增大梯度的方式对模型的影响是较小的。

FGM　　和 FGSM 的区别在于，FGM 相当于给扰动做了缩放，将扰动控制在一定范围内。从而参数的更新也在一定范围之内，更有可能达到某个局部最优。这样的做法从实验结果上看是比较有效的。

	precision	recall	f1-score
Baseline	0.9146	0.9148	0.9146
FGSM	0.9160	0.9156	0.9156
FGM	0.9195	0.9189	0.9189
PGD	0.9179	0.9175	0.9174
Free	0.9216	0.9217	0.9216

Figure1 综合性能对比

PGD PGD 在单步更新的基础上引入了多步。这样的扰动相比 FGSM 更为强烈。从实验结果来看，有理由相信 PGD 相比 FGSM 具有更好的对抗示例。

Free Free 实际上就是多步的 FGSM。类似 PGD，引入更好对抗示例的同时，简化了计算步骤。在实验过程中发现对于 Free 和 PGD 方式，需要设置合理的步数，步长和学习率，使得模型在多步计算中能够有效地逼近最优解。

可行性分析 从结果来看，在 embedding 层上添加扰动的方式的确为模型带来了更多可供学习的例子。通过对对抗示例和原始样本的学习，模型可以学习得更好，更具鲁棒性。

限制性分析 通过对比实验和分析，发现保留 dropout 层时对抗训练的效果不好。分析发现，每次计算扰动值时会用到 embedding 层全体神经元的信息，而 dropout 在每一次 forward 的时候会随机丢弃掉部分神经元。所以 dropout 和对抗训练无法混用[5]。

	precision	recall	f1-score
Baseline (without dropout)	0.8246	0.8159	0.8125
Baseline (with dropout)	0.8155	0.7987	0.8011
FGSM	0.8085	0.7921	0.7939
FGM	0.8248	0.8098	0.8125
PGD	0.8212	0.8038	0.8064
Free	0.8280	0.8176	0.8193

Figure2 鲁棒性对比

鲁棒性分析 从 Figure2 可以看出，与不带 dropout 的 Baseline 相比，FGM,PGD 以及 Free 都更加鲁棒。而有意思的是，与带 dropout 的 Baseline 相比，就只有 Free 一种对抗训练方式更加鲁棒。

从原理分析，因为 dropout 程序导致两个神经元不一定每次都在一个 dropout 网络中出现。这样权值的更新不再依赖于有固定关系的隐含节点的共同作用，阻止了某些特征仅仅在其它特定特征下才有效果的情况 。迫使网络去学习更加鲁棒的特征。虽然其并非是一种对抗训练方式，但也可以有效提升模型鲁棒性。

Github 项目地址：
https://github.com/Usaodon/Adversarial-training-on-Chinese-text-classificaion/tree/main

Reference

[1]Miyato, Takeru, Andrew M. Dai, and Ian Goodfellow. "Adversarial training methods for semi-supervised text classification." *arXiv preprint arXiv:1605.07725* (2016).
[2]Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
[3]Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).
[4]Shafahi, Ali, et al. "Adversarial training for free!." *Advances in Neural Information Processing Systems* 32 (2019).
[5]Zhu, Chen, et al. "Freelb: Enhanced adversarial training for language understanding." (2019).
[6]Lee, Junghoon, Jounghee Kim, and Pilsung Kang. "Back-Translated Task Adaptive Pretraining: Improving Accuracy and Robustness on Text Classification." *arXiv preprint arXiv:2107.10474* (2021).