**对抗训练评估报告**

# 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

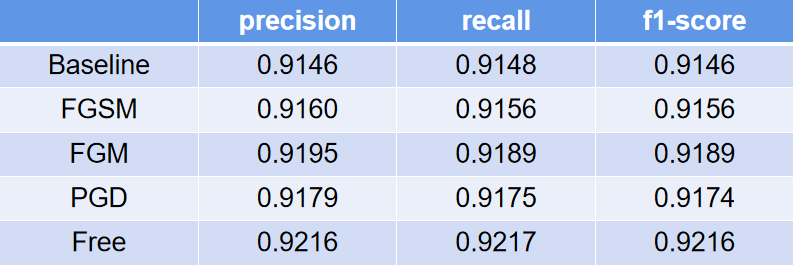
对抗训练的主要思想是提升模型对于对抗样本的错误率。因此对于对抗样本，需要朝着梯度增大的方向进行训练。下面是几种不同的对抗训练方式。

上述的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。下面是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。

# Expiremental Settings

从[THUCNews](https://link.zhihu.com/?target=http://thuctc.thunlp.org/" \t "https://zhuanlan.zhihu.com/p/_blank)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](https://link.zhihu.com/?target=https://github.com/Embedding/Chinese-Word-Vectors" \t "https://zhuanlan.zhihu.com/p/_blank)。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。

# Results & Analysis



Figue1综合对比图

由上图可以看出，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。

**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]。

# 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).