对抗训练在 NLP 中的应用实验报告

1. 背景

GAN 之父 Ian Goodfellow 在 15 年的 ICLR[1]第一次提出了对抗训练这个概念，简而言之，就是在原始输入样本上加一个扰动，得到对抗样本后，用其进行训练，提升模型的训练效果。为将其迁移到 NLP 任务中，Goodfellow 在 17 年的 ICLR[2]中提出了可以在连续的 embedding 上做扰动。本报告在文本分类模型 TextCNN[3]的基础上实现了 FGSM[1]、PGD[4]和 Free[5]这几种对抗训练的方法，并比较和分析实验结果。

2. 常用的对抗训练的方法

- **FGSM (Fast Gradient Sign Method)**[1]

FGSM 是 Goodfellow 提出的对抗训练时方法，假设当前输入的梯度为：

$$g = \Delta_x L(x, y; \theta)$$

那么扰动值和对抗样本定义为: $\delta = \varepsilon * sign(g)$，$\tilde{x} = x + \delta$。可以理解为将输入样本向着梯度的方向增加，这样得到的对抗样本就能造成损失的增加，从而促进模型更进一步的学习。

- **PGD (Projected Gradient Descent)**[4]

PGD 可以看作是对于 FGSM 或者 FGM 的近一步改进，FGSM 直接通过ε参数只经过了一步算出了扰动值，这样得到的扰动可能不是最优的。PGD 进行了改进，多迭代几次，慢慢找到最优的扰动值，具体的迭代公式：

$$\delta_{t+1} = \alpha * \frac{g_t}{||g_t||_2}$$，α为迭代的步长，且$||\delta_t||_2 \le \varepsilon$

- **Free (Free Adversarial Training)**[5]

从 FGSM 到 PGD，主要是优化对抗扰动的计算，虽然取得了更好的效果，但计算量也一步步增加。对于每个样本，FGSM 或 FGM 都是两次前后向的计算，一次是原始样本 x 的，另一次是对抗样本 $x + \delta$的。而 PGD 则计算了 $K + 1$次，消耗了更多的计算资源。因此 Free 在 PGD 的基础上进行了训练速度的优化。

Free 的思想是在对每个样本 x 连续重复 M 次训练，更新方式上和 FGSM 比较像，不过在计算δ时时复用了上一步的梯度，又和 PGD 一样，整体训练的 epoch 相当于乘以了 M。δ的更新公式为：

$$\delta_{t+1} = \delta_t + \varepsilon * sign(g)$$

3. 对抗训练实验和效果分析

- **实验设置**

TextCNN 的代码来源于 github 项目 Chinese-Text-Classification-Pytorch。

对抗训练的代码来源于 github 项目：TextCNN-Adversarial-Training-in-NLP。

训练数据集来源于上述 TextCNN 作者从 THUCNews 中抽取了 20 万条新闻标题，一共 10 个类别，每类 2 万条，文本长度在 20 到 30 之间

- **机器配置**

GPU: 16G V100,

其他: 16 核 CPU，128G 内存

- **实验结果**

指标数据:

方法	acc	micro-precison	micro-recall	micro-f1
Baseline	89.18%	0.8924	0.8918	0.8919
FGSM	90.87%	0.9089	0.9087	0.9086
PGD	89.81%	0.8989	0.8981	0.8982
Free	88.07%	0.8817	0.8807	0.8808

性能数据:

方法	训练时间	stop epoch	每个 epoch 时间	Test loss	参数配置
Baseline	1 分 23 秒	3	27 秒	0.34	
FGSM	5 分 11 秒	6	51 秒	0.3	$\varepsilon = 0.1$
PGD	5 分 12 秒	4	1 分 18 秒	0.33	$\varepsilon = 0.1, K = 3, \alpha = 0.1$
Free	3 分 51 秒	3	1 分 57 秒	0.39	$\varepsilon = 0.1, M = 3$

注：(1)各方法的参数配置见 models，(2)上述详细的实验指标见 log

- **数据分析**

✓ 从实验指标看，FGSM 方法的指标是最好的，有一个可能的解释是 TextCNN 模型结构太简单了，太复杂的方法反而不会带来提升

✓ PGD 方法因为训练速度比较慢，而且可以调的参数比较多，因此没有尝试太多组参数，多尝试几组应该还会有收益

✓ Free 方法由于每个样本需要连续的更新 M 次，所以整体的 epoch 是最多的，但是效果却是最差的。除了上述第一条原因外，Free 也有自己的缺点，Free 每次的扰动都是根据前一次样本的梯度计算出来的，对于当前样本不一定是最优的

- **后续展望**

✓ 这几种对抗训练的方法还不少参数可以调，后续时间充分可以进一步调整，上述结果已经初步证明了对抗训练在 NLP 中的效果；模型本身也有一些参数需要配合对抗训练去调整的，比如学习率、dropout 等

✓ 后续可以把 TextCNN 换成 Bert 等复杂的模型，增大模型的复杂度，让模型有更多的空间可以学习

✓ 以上几种方法有各自的缺点，后续可以尝试其他类 PGD 的改进方法，比如 FreeLB、YOPO、SMART 等方法

4. 参考

[1] Explaining and Harnessing Adversarial Examples

[2] Adversarial Training Methods for Semi-Supervised Text Classification

[3] Convolutional Neural Networks for Sentence Classification

[4] Towards Deep Learning Models Resistant to Adversarial Attacks

[5] Adversarial Training for Free!