**对抗训练在N**LP**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。

1. **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：

那么扰动值和对抗样本定义为: 。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：

，且

* **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，一次是原始样本的，另一次是对抗样本的。而PGD则计算了次，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连续重复次训练，更新方式上和FGSM比较像，不过在计算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。的更新公式为：

1. **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](https://github.com/649453932/Chinese-Text-Classification-Pytorch)。

对抗训练的代码来源于github项目：[TextCNN-Adversarial-Training-in-NLP](https://github.com/shshlzh/TextCNN-Adversarial-Training-in-NLP.git)。

训练数据集来源于上述TextCNN作者从[THUCNews](http://thuctc.thunlp.org/)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 |  |
| PGD | 5分12秒 | 4 | 1分18秒 | 0.33 |  |
| Free | 3分51秒 | 3 | 1分57秒 | 0.39 |  |

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

1. **参考**
2. [Explaining and Harnessing Adversarial Examples](https://arxiv.org/abs/1412.6572)
3. [Adversarial Training Methods for Semi-Supervised Text Classification](https://arxiv.org/abs/1605.07725)
4. [Convolutional Neural Networks for Sentence Classification](https://arxiv.org/abs/1408.5882)
5. [Towards Deep Learning Models Resistant to Adversarial Attacks](https://arxiv.org/abs/1706.06083)
6. [Adversarial Training for Free!](https://arxiv.org/abs/1904.12843)