**对抗训练在自然语言处理实验分析报告**

**1、目的：**

基于已经提出的对抗训练方法，包括，FGSM、FGM、PGD、Free，使用TextCNN网络架构，测试在文本分类上的效果和性能。测试的方式包括正常情况下和样本受到attack两种情况，attack的方式为PGD模式产生对抗样本作为测试样本。

**2、对抗训练简单介绍：**

对抗训练的核心是生成对抗样本。对抗样本是“迷惑”模型的样本，让模型给出错误的结果，也就是使得模型的loss增加。理论上，对于输入，会在给定范围内添加一个增量，在这个范围内使得模型loss增加最大的增量作为我们的对抗样本。和梯度下降法相反，对抗增量可以使用loss对于输入参数的梯度上升法。主要有下面的几种方式：

**FGSM**：最先被提出来的对抗训练方式，计算loss对输入参数梯度的符号值，输入加上一个缩放的符号值就可以产生对抗增量。

**FGM**：和FGSM相似，只是不是用梯度的符号值，使用的是梯度的norm后的值，这样最大程度保留了梯度的原始方向信息。

**PGD**：是上面两种方法的增量生成的迭代版本，会在指定迭代范围内，使用上一步的梯度生成当前迭代步的对抗增量，再计算模型的梯度，不断累加增量得到一个对抗样本总的对抗增量。使用这个总的对抗增量来训练模型参数。

**Free**：和PGD比较类似，不过在迭代增量过程中，会加入模型参数的训练，这样可以加速模型训练过程，提高训练效率。

**3、测试集实验结果**

**实验对象：**中文文本的分类，十个类别：财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐。训练数据集18万，验证数据集1万，测试集1万。测试集上模型结果如下：

**表1正常没有受到attack的结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **accuracy** | **precision** | **recall** | **f1** |
| **base** | 90.85% | 90.86% | 90.85% | 90.85% |
| **fgsm** | 91.27% | 91.29% | 91.27% | 91.27% |
| **fgm** | 91.45% | 91.45% | 91.45% | 91.44% |
| **pgd** | 91.29% | 91.29% | 91.29% | 91.28% |
| **free** | 90.54% | 90.57% | 90.54% | 90.54% |

**表2受到PGD累加attack的结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **accuracy** | **precision** | **recall** | **f1** |
| **base** | 51.19% | 50.11% | 51.19% | 47.53% |
| **fgsm** | 84.21% | 85.74% | 84.21% | 84.34% |
| **fgm** | 73.32% | 77.18% | 73.32% | 71.51% |
| **pgd** | 50.63% | 50.16% | 50.63% | 46.68% |
| **free** | 86.29% | 86.68% | 86.29% | 86.36% |

**效果分析：**

1. **正常测试情况下，效果差别不大，fgsm、fgm、pgd对比base有所提升。**
2. **受到attack情况下，fgsm、fgm、free都明显好于base, pgd可能受到超参数影响，需要进一步实验分析。**