**对抗训练在TextCNN中的应用**

作者：王志伟，Github地址： https://github.com/neorecrown/TextCNN\_with\_Adversarial\_Training

**一．任务描述**

阅读ICLR2020 的一篇论文，Fast is better than free: Revisiting adversarial training，参考其源码，将图像领域的对抗训练技术应用到简单 TextCNN中。实现PGD、“Free”、FGSM 三种对抗训练方法。提供它们与baseline（原始的 TextCNN)的性能比较， 性能评价指标为precision, recall 以及f1 得分。

**二．复现流程**

1. 复现 TextCNN：下载源码，复现　Baseline　训练流程，得到　Baseline　结果
2. 对抗训练实现：根据论文中提供的对抗训练算法表（如下表），编写 adversarial\_strategy.py，其中编写方式借鉴知乎专栏文章[https://zhuanlan.zhihu.com/p/91269728]，以脚本方式编写
3. 对抗训练: 根据对应策略改写 TextCNN 源码中的 train\_eval().
4. 定义训练ｆｌｏｗ：整理 main.py，定义网络关键参数，定义flow．
5. 输出结果：分别在命令行内输入可得Baseline，PGD，Free，FGSM (random initial)的结果

```

# 训练并测试：

# TextCNN Baseline

python main.py --adversarial\_model Baseline

# TextRNN with PGD

python main.py --adversarial\_model PGD --pgd\_steps 7 #默认为7

# TextRNN with Free

python main.py --adversarial\_model Free --minibatch\_replays 10 #默认为10

# TextRCNN with FGSM

python main.py --adversarial\_model FGSM #alpha设为2

```

1. 分析结果

**三：任务输出**

**Table1:** precision, recall and f1-score of basline, PGD, Free and FGSM are shown below, respectively.





**Table1:** Comparison of the precision, recall and f1-score obtained from the basline, PGD, Free and FGSM models, respectively.

**四．任务总结**

在 TextCNN 下引入对抗训练，PGD 算法准确度最高； FGSM 虽然准确度稍差，但未下降太多，且相比 PGD 训练花费时间更快。