**对抗训练实验评估报告**

1. **实验任务**

理解论文ICLR2020《FAST IS BETTER THAN FREE: REVISITING ADVERSARIAL TRAINING》里的对抗训练方法PGD、“Free”、FGSM，并应用到TextCNN中，对比Baseline及引入这三种对抗训练方法后的性能，评价指标包括Precision、Recall、F1-score、Acc、Early stop时的迭代次数及用时。

1. **实验结果**

**表1 各对抗训练方法的实验结果**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模型 | Precision | Recall | F1 | Acc | Steps | Cost |
| Baseline | 90.84% | 90.82% | 90.81% | 90.82% | 5700 | 1min53s |
| +PGD | 91.69% | 91.62% | 91.62% | 91.62% | 10500 | 24min33s |
| +FGM | 91.37% | 91.33% | 91.34% | 91.33% | 8300 | 5min5s |
| +FGSM1 | 88.79% | 88.69% | 88.68% | 88.69% | 11000 | 6min58s |
| +FGSM2 | 82% | 79.77% | 80.01% | 79.77% | 11600 | 6min59s |
| +FreeAT1 | 82.45% | 81.64% | 81.63% | 81.64% | 2800 | 6min21s |
| +FreeAT2 | 44.02% | 23.06% | 14.94% | 23.06% | 1000 | 2min11s |

**表2 PGD在各类别上的实验效果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | Precision | Recall | F1 | 样例数 |
| finance | 93.06% | 88.50% | 90.72% | 1000 |
| realty | 93.33% | 93.80% | 93.57% | 1000 |
| stocks | 84.85% | 87.90% | 86.35% | 1000 |
| education | 95.70% | 95.70% | 95.70% | 1000 |
| science | 90.37% | 86.30% | 88.29% | 1000 |
| society | 89.52% | 92.30% | 90.89% | 1000 |
| politics | 90.46% | 90.10% | 90.28% | 1000 |
| sports | 95.51% | 95.70% | 95.60% | 1000 |
| game | 95.45% | 90.20% | 92.75% | 1000 |
| entertainment | 88.69% | 95.70% | 92.06% | 1000 |
|  | | | | |
| macro\_avg | 91.69% | 91.62% | 91.62% | 10000 |

其中FGSM\*的扰动策略是每个batch内先赋予初始扰动，梯度更新后再进行第二次扰动，不同的是FGSM1初始扰动仅对当前输入对应的Embedding扰动，FGSM2初始扰动是对全局Embedding进行扰动，第二次扰动后需要还原Embedding。FreeAT1和FreeAT2的不同之处在于对局部或全局的Embedding进行扰动。

1. **实验分析**

效果而言，PGD>FGM>TextCNN>FGSM\*>FreeAT\*，其中FGSM1（局部Embedding）>FGSM2（全局Embedding），FreeAT2（扰动作用全局Embedding，且不作还原）训练失效。PGD效果最好，比Baseline高0.6%。

性能而言，TextCNN>FGM> FGSM\*=FreeAT\*>PGD。

综合效果和性能而言，FGM是最好的策略，同时也说明引入对抗训练对模型效果的提升是有帮助的，但也取决于对抗策略的选择。另外本实验未对各参数展开更为详细的实验，各对抗训练策略展示出来的未必是最优水平。

1. **实验复现流程及代码**

见[github](https://github.com/Dsir123/Adversarial_Train)地址。