对抗训练测试报告

1. 实验简介

本项目使用 FGSM、PGD、FreeLB 对情绪分类数据集进行对抗性训练的 tensorflow2 实现，baseline 模型为 TextCNN。

2. 实验环境

Python3
Tensorflow==2.6.2

3. 实验数据

Github 上下载的情感分类数据集，共 45339 条样本，共 8 个类别（样本类别分布差异较大），数据按 7:3 划分训练集和测试集。

4. 实验参数

```
maxlen = 100 # 最大句长（词数）
max_features = 10000 # 词表维度
test_size=0.3 # 测试集比例
embedding_dims = 128 # 词向量维度
batch_size = 64 # batch size
epochs = 20 # 最大训练 epoch
```

5. 实验结果

整体实验结果如下：

	Precision	Recall	F1
Baseline	0.61	0.66	0.62
FGSM	0.64	0.69	0.63
PGD	0.56	0.65	0.57
FreeLB	0.61	0.65	0.62

5.1. Baseline

训练时间：105 s
实验结果：

	Precision	**Recall**	**F1**	**Support**
anger	0.36	0.20	0.26	547
disgust	0.30	0.16	0.21	919
fear	0.60	0.03	0.07	86
happiness	0.39	0.30	0.34	833
like	0.45	0.28	0.35	1232
none	0.73	0.89	0.80	8982

sadness	0.35	0.21	0.26	774
surprise	0.25	0.04	0.07	226
avg / total	0.61	0.66	0.62	13599

5.2. FGSM

训练时间：463 s
实验结果：

	Precision	**Recall**	**F1**	**Support**
anger	0.51	0.15	0.24	547
disgust	0.35	0.14	0.20	919
fear	0.57	0.05	0.09	86
happiness	0.50	0.26	0.34	833
like	0.48	0.28	0.36	1232
none	0.72	0.93	0.82	8982
sadness	0.50	0.19	0.28	774
surprise	0.50	0.04	0.07	226
avg / total	0.64	0.69	0.63	13599

5.3. PGD

训练时间：841 s
实验结果：

	Precision	**Recall**	**F1**	**Support**
anger	0.33	0.05	0.09	547
disgust	0.28	0.09	0.14	919
fear	0.00	0.00	0.00	86
happiness	0.38	0.14	0.20	833
like	0.33	0.11	0.17	1232
none	0.69	0.94	0.79	8982
sadness	0.34	0.07	0.11	774
surprise	0.12	0.01	0.02	226
avg / total	0.56	0.65	0.57	13599

5.4. FreeLB

训练时间：262 s
实验结果：

	Precision	Recall	F1	Support
anger	0.36	0.21	0.27	547
disgust	0.30	0.15	0.20	919
fear	0.38	0.07	0.12	86
happiness	0.44	0.25	0.32	833
like	0.40	0.29	0.34	1232
none	0.74	0.87	0.80	8982
sadness	0.31	0.26	0.28	774
surprise	0.11	0.04	0.06	226
avg / total	0.61	0.65	0.62	13599

备注：因时间原因，本实验未做复杂调参，初步根据 3 种对抗算法原理实现对抗训练技术。