工程实践与科技创新 IV-J 课程作业中期报告

张晨阳　肖真然

目录

1 课题简介

本团队探讨的是关于简单大语言模型在情感分类课题上的能力，经过微调训练，在多个现行的文本情感分类任务数据集上取得了一定的效果。在中期报告中，本团队将说明数据集选取与制作、模型训练与测试的具体过程，并展示在初步微调后的 GPT2 模型上的实际效果。

2 数据集选择

2.1 任务类型

在文本情感分析这一大课题中，文本情感分类是一项最为基础的工作。一般地，它接受一段文本的输入，并最终返回一个形如 joy、sadness 的情感类别描述词。在一些具体的任务上，数据集往往对文本情感划分为若干个大类别，并使用数字序号与之建立映射作为数据集 label。

在本次作业中，模型采用的是 117M 参数量版本的生成式模型 GPT-2。这一模型版本相对较老，参数量有限，不能期望模型能够做到分析极细粒度的丰富语义，故文本情感分析中情感主体提取、针对具体属性的细粒度情感分析等任务很难获得很好的效果。故至少在第一部分中，本项目仅对文本总体情感分类这一任务进行微调训练。

尽管 GPT-2 模型本身是一个生成模型，但通过提示词工程等一系列方法，可以让模型拥有执行分类任务的能力，从而完成本项目期望的文本情感分类任务。

2.2 数据集选择

文本情感分类在 NLP（自然语言处理）中是一个被广为考察的任务课题，现行有许多开源的相关数据集可供使用。在广泛地检索了各种相关数据集后，本项目基于作业任务分出以下几类比较有代表性的数据集类型。

倾向分析 这一类数据集是占比最多的，数据的文本来源主要是各类网站的评价文本，如商品评价、影片评价等等。这类数据的 labels 通常为二类或者三类，分别是正面情感（Positive）、负面情感（Negative），以及可能的中性情感/陈述。这种分类方法在市场调研等实际应用中会有很高的效益，但过于简单的分类并不适合本作业中的生成模型，且评判的力度有限。在本作业中只作为可供使用的文本集合。

细粒度情感分类 这一类数据相对较少，数据的文本来源主要是各类相对简单的日常表达文本，加上部分评论类文本，质量整体相对参差。这类数据集的 labels 通常为多类，比较常见的有 6 类、13 类等等，一般对应着 joy、sadness、love 等大类情感描述词。这

类数据集将作为本次作业的主力军，经过有效筛选，从 Kaggle 上选取出三类数据集作为作业使用：

1. `Emotions in text_refresh` 6 类情感分类数据共 21459 条
2. `Emotion Detection from Text_refresh` 13 类情感分类数据共 40000 条
3. `emotion analysis based on text_refresh` 13 类情感分类数据共 839555 条，其中 neutral 类占 80.3%，其余数据共 165017 条

超细粒度情感分类 这边特别提到一个数据集，Google 发布的大规模数据集 GoEmotions。这一数据集是上述细粒度情感分类的特例，文本来源于 Reddit 评论，而 labels 来到了 28 类，远超其他数据集。尽管数据集本身能够导向高精度地情感辨析分类，但处于 gpt-2 适应度、与其他数据集的兼容等多方面的问题，仅作为可供使用的文本集合，以及情感划分的参考性指标。

3 模型微调

3.1 实验环境

基础设施 Linux Ubuntu, CUDA used 4090

深度学习框架 Pytorch

预训练模型 `GPT2LMHeadModel.from_pretrained("gpt2")`

3.2 数据预处理

为了使得 gpt-2 生成模型能够适应分类问题的需求，本项目使用的数据集需要经过提示词的预处理加工。主要分为以下两个方面：适应分类问答的提示词包装工程，以及多个数据集在评价标准的统合工程。

提示词：一问一答 分类问题可以被显式地表示为一套选项固定的选择题。对于生成模型而言，为了方便后续测试的可扩展性，本项目暂时并不通过截取特征后加装分类网络结构，而是简单的通过提示词文本设计 question+answer 的文本数据组合。由此，训练所用的样本格式如下：

```
1 question = "Please read the given text, analyze its sentiment, and select one
    of the following emotion options: empty, sadness, excitement, neutral, fear
    , surprise, love, amusement, annoyance, joy, boredom, relief, anger. \nThe
    text is: "
2 answer = "\nThe answer is:"
```

```

full_text = question + texts[i] + answer + ' ' + emotions[i]
```

可以根据经验显见，这类提示词在 Chat GPT 等大模型上会有很有效的表现，由此推理在 gpt2 上应该有相对应地弱化表现。具体的进一步比较关系会在后续部分完成。

数据集统合 由于采用相同格式的输入文本，对于数据集需要进行统一的格式化，一方面是原数据读取格式读取的统合，另一方面是原标签标记描述词的统合。

经过比较数据集发现，本次采用的三个数据集均采用 csv 格式，通过简单地选列与列名统一，可使其结构达到一致。另一方面，两个 13 类数据集采用完全一致的标签构成，且完全包含了 6 类数据集的标签，仅需要对同义词的不同表达——例如 happiness-joy——进行统一，这里以谷歌的 go_emotion 数据集标签为标准，统一了三个数据集的标签。

具体而言，6 类标签分别是 anger、fear、joy、love、sadness、surprise，13 类标签则是在 6 类的基础上增加了 amusement、annoyance、boredom、empty、excitement、neutral、relief。这一点在上一节的 question 文本中已经有所体现。

此外，由于数据集本身规模较大且 13 类的模型存在部分类别的显著不平衡，在使用中本项目采取了截取小部分数据集（总数据条数约为 10k 级）的方式训练模型。

3.3 模型训练

针对以上数据集，本项目设计了训练与测试代码，训练使用了 Transformers 提供的 Trainer 类来简单实现训练过程，在后续项目中，不排除使用更加基础性的工作来支持更加复杂的深度学习设计方法。

采用部分参数如下

```
training_args = TrainingArguments(
    output_dir='./model/gpt2_trained',
    num_train_epochs=3,
    per_device_train_batch_size=1,
    save_steps=1000,
    learning_rate=1e-4,
    overwrite_output_dir=True,
)
```

完整代码见附件。

3.4 结果展示

在这一节将展示微调训练前后模型处理情感分类问题上性能的差异。作为初步比较，先使用手动判断 accuracy 作为唯一指标，在后续涉及更加精度的模型优化时，可能引入更加复杂的评判机制。

Dataset	Size (pre label)	Test	Accuracy
EIT	800	EIT	0.905
EIT	800	EDFT	0.149
EDFT	800	EIT	0.26
EDFT	800	EDFT	0.269
EDFT	800	EABOT	0.115
EABOT	800	EIT	0.068
EABOT	800	EDFT	0.213
EABOT	800	EABOT	0.914
ALL	800*3	EIT	0.895
ALL	800*3	EDFT	0.154
ALL	800*3	EABOT	0.069

从测试结果来看，gpt-2 能在 6 分类场景下表现出良好的性能，但在 13 分类的场景下能力显著下降。有关这部分的具体分析将在后续展开。

4 总结

在第一部分的课程作业中，本小组以简单大语言模型在情感分类课题上的能力为研究课题，以 117M 轻量级 gpt-2 模型为 baseline 开展了数据集与模型微调工作，最终在三个情感分类数据集的组合集上取得了一定的成果。