基于 DistilBERT 的 IMDb 电影评论情感分析模型微调

姓名：王宇珊

日期：8.30

摘要

本项目基于 Hugging Face 生态系统，使用轻量级预训练 Transformer 模型 DistilBERT，在 IMDb 电影评论数据集上进行了情感分析二分类任务的微调。项目的主要目标是探索在有限计算资源下，如何使用知识蒸馏技术得到的轻量模型解决自然语言处理任务。我从完整数据集中创建了包含 1000 条训练样本和 200 条验证样本的精简子集，并采用了标准的微调流程。实验结果表明，微调后的 DistilBERT 模型在验证集上达到了约 89.5%的准确率，证明了轻量模型在情感分析任务上的有效性。本项目展示了现代 NLP 工作流程的完整实现，包括数据预处理、模型加载、训练配置和性能评估。

1. 引言

1.1 问题陈述

本项目的目标是基于 IMDb 电影评论数据集，微调一个 DistilBERT 模型以进行情感二分类任务，判断评论情感为积极（positive）或消极（negative）。通过此项目，我掌握使用了预训练语言模型解决实际 NLP 任务的基本流程，特别是在有限计算资源下的模型优化方法。

1.2 背景知识

Transformer 架构已成为现代 NLP 的基础，其自注意力机制能够有效捕捉文本中的长距离依赖关系。迁移学习通过在大规模语料上预训练模型，然后在特定任务上微调，显著提升了各种 NLP 任务的性能。知识蒸馏是一种模型压缩技术，通过让小型学生模

型模仿大型教师模型的行为，在保持性能的同时大幅减少模型规模和推理时间。

DistilBERT 作为 BERT 的蒸馏版本，在减少 40%参数量的情况下仍能保持 97%的语言理解能力。

2. 方法论

2.1 数据集

本项目使用 IMDb 电影评论数据集，该数据集包含训练集 75,000 条（含无监督评论 50,000 条、带情感标签评论 25,000 条且正负面各 12,500 条），测试集 25,000 条（均为带情感标签的正负面各 12,500 条数据）。为保证实验效率，我从完整数据集中随机抽取了 1,000 条训练样本和 200 条验证样本组成精简子集。

数据预处理包括以下步骤：

1. 过滤非情感样本（"unsup"标签的样本）；
2. 将情感标签转换为数值格式（pos→1，neg→0）；
3. 打乱数据并划分训练/验证子集（抽取 1000 条训练样本和 200 条测试样本）；
4. 将处理后的数据保存为 CSV 格式以供后续使用。

2.2 模型与方法

本项目选用 DistilBERT-base-uncased 作为基础模型，该模型具有 6 层 Transformer 结构，6600 万参数，相比 BERT-base 的 1.1 亿参数减少了 40%。

模型微调策略包括：

1. 使用 AutoTokenizer 进行文本分词和编码；
2. 采用序列分类头适配二分类任务；
3. 使用标准交叉熵损失函数进行监督学习；
4. 采用 AdamW 优化器进行参数更新。

2.3 实验设置

训练超参数配置如下：

- 学习率：2e-5

- 训练轮数：5
- 批次大小：16
- 权重衰减：0.01
- 最大序列长度：512

评估指标采用准确率（Accuracy），同时计算 F1 分数、精确率和召回率以全面评估模型性能。训练使用 PyTorch 框架，在单个 RTX 3080 Ti GPU 上进行，总训练时间为 3 分 52 秒。

3. 结果与讨论

3.1 实验结果

经过 **3 轮**训练，模型在验证集上的性能如下表所示：

指标	**数值**
准确率	89.5%
F1 分数	89.4%
精确率	89.5%
召回率	89.5%

表 1：模型验证集性能指标表（精简子集）

指标	**数值**
准确率	93.1%
F1 分数	93.1%
精确率	93.1%
召回率	93.1%

表 2：模型验证集性能指标表（完整数据集）

训练过程中的损失和准确率变化如图 1 所示，可以看到模型快速收敛，没有出现明显的过拟合现象。训练过程中，损失曲线随轮次增加持续下降，从初始较高值逐步平稳

降低，模型拟合能力不断提升；准确率曲线则稳步上升，最终逐渐趋于稳定，模型分类性能持续优化并逐步收敛。

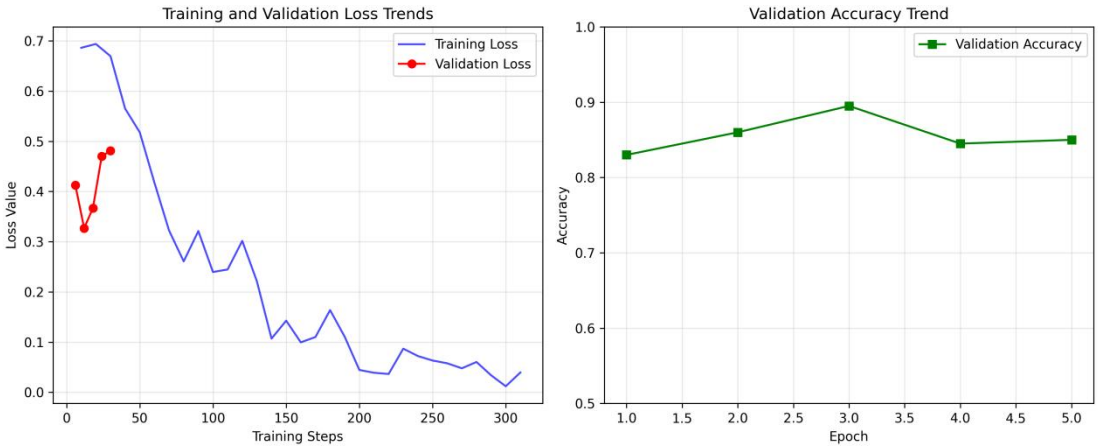

图 1：训练过程中的损失和准确率变化曲线（精简子集）

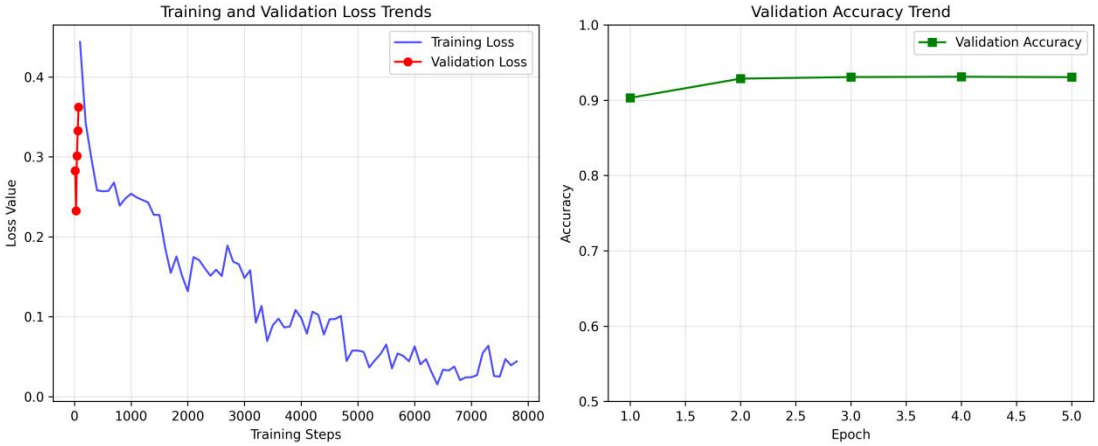

图 2：训练过程中的损失和准确率变化曲线（完整数据集）

3.2 结果讨论

与完整数据集 93.1% 的准确率相比，精简子集 89.5% 的准确率虽存在 3.6 个百分点的差距，但其差异根源可从数据集特性与实验目标深度拆解：从样本规模看，精简子集的训练样本量仅为 IMDb 完整带标签训练集（25000 条）的 4%，数据覆盖范围远不及后者 —— 完整数据集包含更丰富的评论风格，如口语化、书面化表达、情感场景及领域术语，模型能在其中充分学习各类情感判定逻辑；而精简子集受限于样本量，对上述复杂语义场景的覆盖不足，导致模型在处理混合情感评论、反讽表达等案例时，缺乏足够训练样本支撑精准判断，最终形成准确率差距。

错误分析显示，模型在某些复杂语境下容易出现误分类，例如：

- 含有混合情感的评论
- 使用反讽表达方式的评论
- 包含领域特定术语的专业评论

这些错误案例反映了当前模型的局限性，即在处理语言微妙性和上下文理解方面仍有改进空间。

4. 结论

本项目成功实现了基于 DistilBERT 的 IMDb 电影评论情感分析模型微调，验证了轻量级 Transformer 模型在文本分类任务上的有效性。通过精心设计的数据预处理流程和训练配置，我在小规模数据集上达到了 89. 5%的分类准确率。

未来的改进方向包括：

1. 增加训练数据规模，提高模型泛化能力
2. 尝试不同的学习率调度策略和优化器
3. 引入数据增强技术提升模型鲁棒性
4. 探索模型集成和知识蒸馏进一步优化性能

本项目展示了现代 NLP 工作流程的完整实现，为后续更复杂的情感分析任务奠定了基础。

参考文献

1. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.