面向GLUE任务的多模型对比与优化研究

摘要：

本文围绕GLUE基准中的RTE（文本蕴含识别）任务，系统对比了**BERT-base-uncased**、**CNN-LSTM+Transformer** 与 **RoBERTa-base** 三种模型在少样本场景下的性能与效率。

实验首先复现并微调BERT，在SST-2、STS-B、RTE三个数据集上分别获得93.12%、Pearson值为0.8403及58.12%的表现；针对RTE准确率过低的问题，通过增加训练轮数、调整学习率和序列长度等策略模型准确率仅提升至62.82%。

随后，构建融合**CNN**、**LSTM**与**Transformer**的轻量级模型，并使用**Spark**加速数据预处理，结果准确率仅约50%，且训练时间大幅增加，表明该结构不适用于小样本蕴含任务。

最后，引入预训练更充分的**RoBERTa-base**，并设计融合4种句子表示的增强分类头，在保持早停机制的前提下将RTE准确率提升至80.14%，相较BERT提升约27.6%，验证了预训练深度与任务特定结构对逻辑推理任务的关键作用。

关键词：**BERT-base-uncased**、**CNN-LSTM**、**Transformer**、**Spark**、**RoBERTa-base**

目录

1. 数据介绍

自然语言处理（NLP）主要自然语言理解（NLU）和自然语言生成（NLG）。为了让NLU任务发挥最大的作用，来自纽约大学、华盛顿大学等机构创建了一个多任务的自然语言理解基准和分析平台，也就是GLUE（General Language Understanding Evaluation）。

GLUE包含九项NLU任务，语言均为英语。GLUE九项任务涉及到自然语言推断、文本蕴含、情感分析、语义相似等多个任务。像BERT、XLNet、RoBERTa、ERINE、T5等知名模型都会在此基准上进行测试。GLUE的具体九项NLU任务如下，可分为三类，分别是单句任务，相似性和释义任务：

表1 GLUE任务表介绍

Corpus	Train	Test	Task	Metrics	Domain
单句任务					
CoLA	8.5k	1k	acceptability	Matthews corr	misc
SST-2	67k	1.8k	sentiment	acc	movie reviews
相似性任务					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
SST-B	7k	1.4k	Sentence similarity	Pearson/Spearman corr	misc
QQP	364k	391k	paraphrase	acc./F1	Social QA questions
释义任务					
MNLI	393k	20k	NLI	Matched acc./mismatched acc	misc
QNLI	105k	5.4k	QA/NLI	acc	Wikipedia
RTE	2.5k	3k	NLI	acc	News Wikipedia
WNLI	634	146	Coreference/NLI	acc	Fiction books

考虑到模型训练时间以及训练效率，因此在三项任务中分别选取一个数据量规模较小的数据集进行模型训练，分别是SST-2，SST-B，RTE。

SST-2是单句子分类任务，单句子分类任务，包含电影评论中的句子和它们情感的人类注释。这项任务是给定句子的情感，类别分为两类正面情感（positive，样本标签对应为1）和负面情感（negative，样本标签对应为0），并且只用句子级别的标签。也就是，本任务也是一个二分类任务，针对句子级别，分为正面和负面情感。

SST-B是相似性和释义任务，是从新闻标题、视频标题、图像标题以及自然语言推断数据中提取的句子对的集合，每对都是由人类注释的，其相似性评分为0-5。任务就是预测这些相似性得分，本质上是一个回归问题，但是依然可以用分类的方法，可以归类为句子对的文本五分类任务。

RTE是自然语言推断任务，它是将一系列的年度文本蕴含挑战赛的数据集进行整合合并而来的，包含RTE1，RTE2，RTE3，RTE5等，这些数据样本都从新闻和维基百科构建而来。将这些所有数据转换为二分类，对于三分类的数据，为了保持一致性，将中立和矛盾转换为不蕴含。

2. 模型介绍

2.1. BERT-base-uncased模型

本次实验首先选用**BERT-base-uncased模型**。BERT是一种预训练语言模型，在自然语言处理任务中展现了卓越的性能。BERT的设计特点使其在处理语义损失较大的翻译数据时，表现优于CNN、RNN等深度学习算法。其核心结构由多层Transformer Encoder组成，这使得它能够捕捉输入文本的上下文语义信息。BERT-base模型包含12层Transformer Encoder。每一层由多头自注意力机制（Multi-Head Self-Attention）和前馈神经网络（Feed-Forward Neural Network）组成，其具体结构如下图所示[3]。

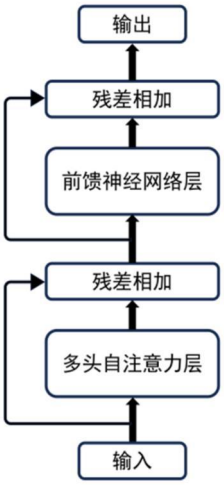

图1 BERT-base每层结构

本次实验在该模型的基础上进行超参数的微调以适应不同的数据集，其超参数如下：

表2 超参数设置

超参数	SST-2	RTE	RTE（改进）	STS-B
批次大小	32	32	8	48
学习率	2e-5	1e-5	2e-5	1e-5
训练轮数	2	5	15	3
最大序列长度	128	128	256	128
优化器	AdamW	AdamW	AdamW	AdamW
权重衰减	1e-4	1e-4	1e-4	1e-4
预热比例	10%	10%	10%	10%

训练完成后发现SST-2的准确率和SST-B的person值较好，而RTE的准确率较低，因此针对RTE进行单独的优化训练。对其参数进行调整，将原先的训练轮数从3轮增加至15轮，批大小从32减到8，增加参数更新次数，学习率改为2e-5，序列长度从128增长至256，以此捕捉更多的上下文信息，同时增强其耐心值为4，防止过拟合，使用更激进的梯度裁剪，最后得到的结果从0.5812提升至0.6282，提升并不大，以此考虑模型的替换。

2.2. CNN-LSTM+Transform模型

考虑到RTE任务数据量较少，因此尝试使用CNN-LSTM+Transform模型进行训练，寻找最佳模型，同时通过**Spark**对数据进行处理，增强模型性能。

Spark是

卷积神经网络（CNN）是一种深度学习模型，专用于处理网格状拓扑数据（如图像），能够自动提取空间层次特征，无需手动特征工程。其结构包括卷积层、池化层、全连接层和激活函数[1]。如下图，卷积层通过滤波器提取局部特征，池化层降低特征维度并增加不变性。

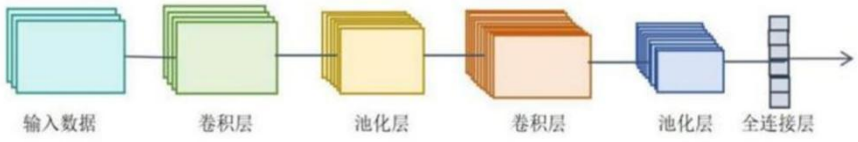

图 CNN神经网络结构示意

长短期记忆网络（LSTM）是循环神经网络（RNN）的变体，解决了传统RNN的梯度消失问题。LSTM通过遗忘门、输入门和输出门控制信息流动，捕捉长期依赖关系，适用于自然语言处理、语音识别和时间序列预测等领域。其核心公式包括细胞状态和门控机制的计算，能够有效管理时间序列数据中的信息传递[2]。

本模型同时在模型的中间层加入Transform模型，用于捕捉文本序列中的长距离依赖关系。通过自注意力机制并行计算序列中所有单词的关联，同时结合多头注意力捕捉多维度语义信息。这使得模型能更好地理解前提句和假设句之间的逻辑关系，最终提升文本蕴含识别（RTE）的准确率。

模型参数设置如下：

表3 参数设置

参数类别	取值	作用
数据处理	180	句子最大长度（控制输入序列规模）
数据处理	32	批次大小（影响训练速度和稳定性）
模型结构	300	词向量维度（与 GloVe 保持一致）
模型结构	384	LSTM 隐藏层维度（核心特征维度）
模型结构	4	Transformer 注意力头数

模型结构	0.5	正则化比例（防止过拟合）
训练配置	80	最大训练轮次（控制训练时长）
训练配置	0.0001	初始学习率（影响收敛速度）

但其模型最终得到的效果并不好，因此考虑进行预训练模型的更换。

2.3. RoBERTa-base模型

鉴于RTE任务涉及复杂的语义推理和逻辑关系判断，我们选择了预训练更充分、性能通常更强的**RoBERTa-base模型**，并特别设计了融合多层面信息的增强分类头，以期提升模型对文本蕴含关系的捕捉能力。RoBERTa是BERT的改进版，通过更充分的训练和更大的批次大小，在各种NLP任务上表现优异。同时在其基础上添加了一个增强的分类头，同时使用4种不同表示（Pooler、CLS、平均池化、最后token），通过不同的表示来捕捉到不同层面的信息，融合后能够提供更全面的语义理解。与标准分类头对比如下：

表4 分类头对比

特征	标准分类头	本模型分类头	优势
输入	CLS token（768维）	4种特征融合（3072维）	更全面的信息
结构	单层线性	三层神经网络	更强的表示能力
激活函数	无/ReLU	GELU	更优的非线性特征
正则化	少量Dropout	双重Dropout	更强的抗过拟合
归一化	无	LayerNorm	稳定训练过程
参数初始化	默认	Xavier正态分布	更稳定的收敛

其模型关键的超参数列表如下：

表5 模型数值

超参数	数值	说明
批次大小	8	训练时每个批次的样本数，小批量防止小数据集过拟合
学习率	主干网络：1e-5 偏置/LN参数：1e-6	分层学习率设置，保护预训练知识
训练轮数	10	最大训练轮数10，早停机制实际提前终

		止训练
最大序列长度	192	输入序列的最大token长度，平衡上下文信息与计算效率
优化器	AdamW	带权重衰减的Adam优化器
权重衰减	0.01	L2正则化强度，应用于非偏置参数
预热比例	10%	训练开始阶段学习率线性增加的比例

通过以上参数的设置，RTE数据集的准确率从0.6282提升到0.8014，性能提升了27.57%。

3. 实验结果分析

3.1. BERT-base-uncased模型

BERT-base-uncased模型对于三种不同任务数据集训练结果如下：

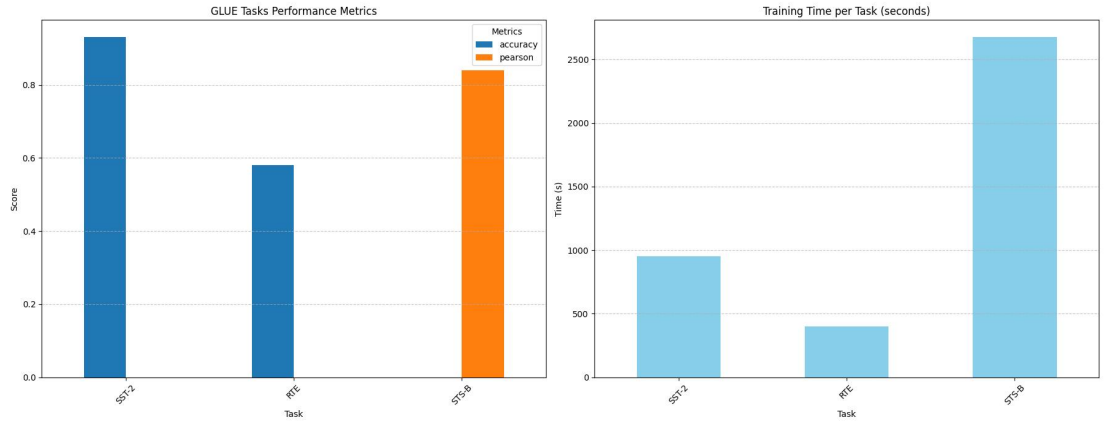

图2 模型训练结果

其中左侧为模型对于各个数据集精准率与person值的比较，由于无法登录到GLUE官网，因此本次实验的准确率其具体数值是通过dev.tsv（开发集）计算得到的。从各个数据集的**准确率与person值**中可以看出：

SST-2（情感分类）的准确率达到0.9312，在glue分类任务中表现优异，说明BERT-base-uncased模型对电影评论情感（积极/消极）的判别能力强，契合SST-2数据集（短文本、情感倾向明确）的特点。

RTE（文本蕴含识别）的准确率只有0.5812，其准确率较低，接近于随机猜

测，说明该模型对文本当中蕴含的语义逻辑关系的捕捉能力有限，导致准确率较低。因此针对该训练集需要重新进行训练以提高准确率。

STS-B（语义文本相似度）的person值为0.8403，在回归任务当中表现良好，说明模型能够有效捕捉文本的语义相似程度，但因语义相似度的细微差异难精准量化导致其与理论最优值仍有一定的差距。

从各个数据集的**训练时间**对比来看：

SST-B的训练时间最久，虽然其训练集数据量并不是最多的，但其需学习文本之间的连续相似度，其数据分布更细粒度，模型需要更多迭代拟合，因此耗时最长。

SST-2是二分类任务，但样本集数量较多，模型平衡了任务复杂度和数据量，使得耗时中等。

RTE样本量最少，且任务目标相对简单，模型快速收敛，耗时最短，同时由于其样本数量少导致其准确率下降。

针对RTE进行**参数改进**，其模型训练图如下：

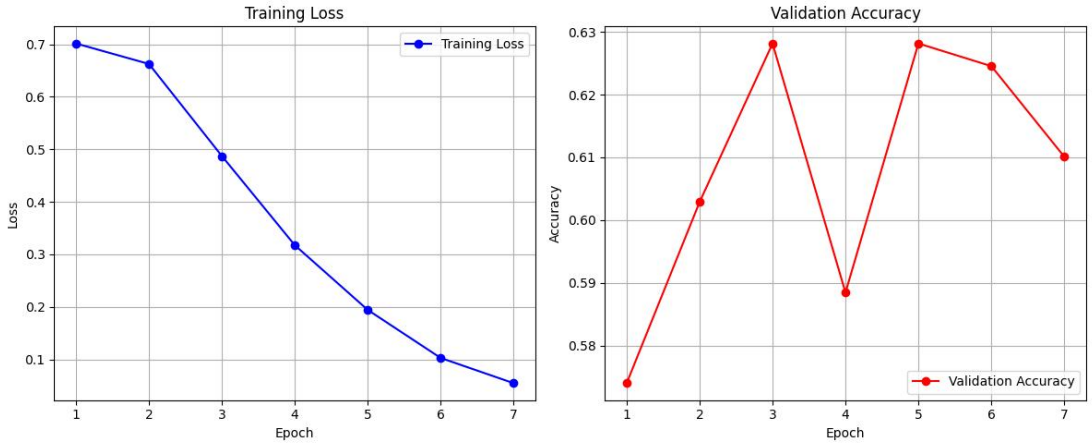

图3 RTE参数改进训练图

从图上可以看出，随着训练轮数的增加损失率得到下降，而模型在第3轮训练时得出最好的训练准确率，第3轮训练后连续4轮其准确率没有得到提升，因此模型进行了早停机制。同时经过参数改进过后其准确率提升到0.628，相较于原先有一定的提升，但是其依旧不可行，因此考虑其他方法提升RTE任务准确率。

3.2. CNN-LSTM+Transform模型

模型训练结果如下：

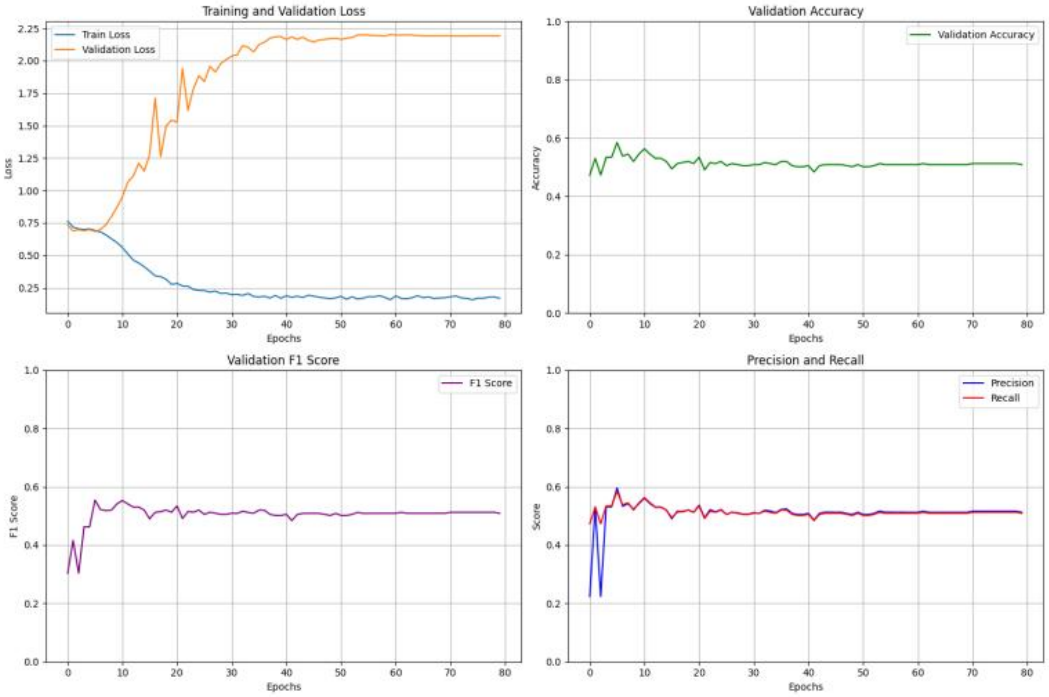

图4 模型训练图

从上图中可以看出随着训练轮次增加，训练损失逐渐下降并趋于稳定。这表明模型在训练数据上的拟合效果越来越好。同时验证损失在训练初期快速上升，然后趋于平稳。这可能意味着模型在训练数据上的表现很好，但在验证数据上的泛化能力有所下降。

而其验证准确率、F1 分数、精确率和召回率都在训练初期快速提升，然后趋于平稳，这表明模型可能已经达到了性能瓶颈，其准确率仅在0.5左右徘徊，因此该模型训练效果不好，同时其训练总时长达到1483.15s，相较原先的训练时间远远超过，但并没有实现时间换取准确率的功能，因此认为该模型不适合于RTE任务，应优先考虑更换预训练模型。

3.3. RoBERTa-base模型

针对RTE准确率较低的情况，重新选择模型进行参数调整训练，得到结果如下：

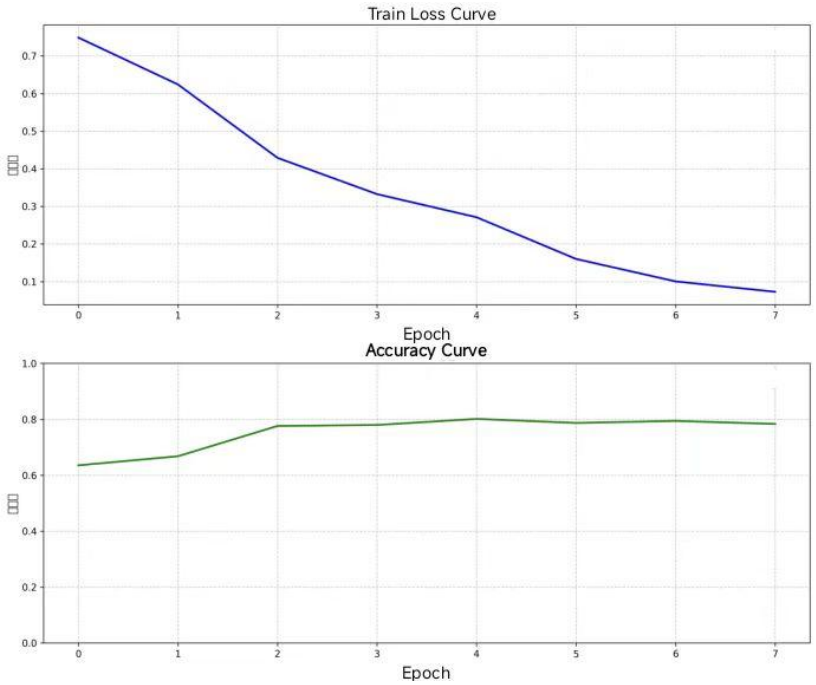

图5 模型训练图

从上图中看出随着训练轮次（Epoch）的增加，损失值是逐渐下降的。这说明模型在不断地学习，对训练数据的拟合程度在提高。而准确率是逐渐上升的，这和损失曲线的变化是对应的，损失降低，准确率提高，符合我们对模型训练的预期。

表6 性能对比

性能对比	BERT-base-uncased模型	RoBERTa-base模型
准确率	0.5812	0.8014
训练时间	399	883

从上表中可以看出，更换模型后RTE的准确率得到了显著的提升，从原先的58.12%提升到了80.14%，将其从随机猜测提升到具有实际应用价值的水平，但是与此同时训练时间得到增加，从原先的399秒增加到883秒，增加121%，其原因可能是由于RoBERTa在训练时动态生成掩码模式，增强泛化能力，使用了更多的数据以及更长的训练步数，进行更细粒度的分词处理。

4. 实验总结

本实验以GLUE-RTE任务为核心，先在BERT-base-uncased上进行SST-2、ST

S-B、RTE三个数据集的训练，得出SST-2、STS-B表现良好，而后针对RTE训练集准确率低进行改进，随后在CNN-LSTM+Transformer与RoBERTa-base两种模型进行预训练，在三个模型间展开系统对比与优化。综合结论如下：

1、预训练深度决定上限：在数据量受限的RTE任务上，RoBERTa凭借更大规模预训练与动态掩码策略显著优于BERT及轻量级组合模型。

2、轻量级模型并非万能：CNN-LSTM+Transformer在图像、时序等领域表现优异，但在需要深层语义推理的小样本文本任务上，训练成本与收益不成正比。

3、微调策略需任务定制：对BERT的简单超参数调整对RTE提升有限；而RoBERTa通过分层学习率、多重Dropout与3072维增强分类头，有效抑制过拟合并提升鲁棒性。

4、训练-性能权衡：RoBERTa训练时间较BERT增加121%，但将准确率从随机水平提升至可实用区间，证明在准确率优先场景下额外开销是可接受的。

5、后续方向：可继续探索Prompt-Tuning、数据增强及多任务联合训练，在保持高效率的同时进一步挖掘小样本蕴含任务的潜力。

5.参考文献

[1]陈金红,崔东文.基于深度学习神经网络超参数优化的入库径流预测方法研究——以云南省暮底河水库为例[J].三峡大学学报(自然科学版),2023,45(04):25-32.DOI:10.13393/j.cnki.issn.1672-948x.2023.04.005.

[2]李析男,朱飞燕.基于CNN-LSTM的贵州省水资源需水预测与趋势分析[J/OL].人民珠江,1-14[2025-07-27].http://kns-cnki-net.ez.zust.edu.cn/kcms/detail/44.1037.tv.20250723.1331.002.html.

[3]曲春来,高敏洁,李一飞,等.基于BERT模型的多语种谣言识别研究[J/OL].计算机工程与应用,1-16[2025-07-28].http://kns-cnki-net.ez.zust.edu.cn/kcms/detail/11.2127.tp.20250725.0944.002.html.

[4]李佳宁,褚丽莉,于清波.基于SSWE与RoBERTa的文本情感分析[J].长江信息通信,2025,38(05):47-51.DOI:10.20153/j.issn.2096-9759.2025.05.013.