华南理工大学
South China University of Technology

《统计学习与数据科学导论》课 程 论 文

(2023 -2024 学年第 2 学期)

论文题目：Exploring Advanced Sequence Classification Models for Sentiment Analysis using PyTorch: From Neural Bag of Words to Transformers

学生姓名： 潘玥

提交日期： 2024 年 8 月 28 日 学生签名：潘玥

学号	202130321189	座位编号	
学院	数学学院	专业班级	数学与应用数学
课程名称	统计学习与数据科学导论	任课教师	夏立

华南理工大學
South China University of Technology

教师评语：
本论文成绩评定： _____分

Exploring Advanced Sequence Classification Models for Sentiment Analysis using PyTorch: From Neural Bag of Words to Transformers

潘玥

摘要：随着社交媒体和在线评论的爆炸性增长，情感分析成为了自然语言处理领域的一个重要研究方向。本研究旨在通过深度学习技术，对电影评论进行情感倾向的自动检测。为此设计并实现了四种不同的深度学习模型：神经词袋（NBoW）模型、长短期记忆网络（LSTM）、卷积神经网络（CNN）和基于 BERT 的 Transformer 模型。每种模型都经过精心设计，以适应情感分析任务的需求。实验一采用了经典的神经词袋模型，该模型通过词嵌入和简单的池化操作捕捉文本特征，为后续实验提供了基线。实验二则转向了循环神经网络，特别是 LSTM 模型，它能够捕捉序列数据中的长期依赖关系，理论上更适合处理文本数据。实验三探讨了卷积神经网络在文本分类中的有效性。通过不同大小的卷积核提取局部特征，CNN 模型能够学习到文本的局部上下文信息。最后，实验四采用了最新的 Transformer 架构，利用预训练的 BERT 模型进行微调，以提高情感分类的准确性。所有模型均在 IMDb 电影评论数据集上进行了训练和测试。评估结果表明，基于 BERT 的 Transformer 模型在情感分析任务上表现最为出色，具有最高的准确率和最低的损失值。LSTM 模型和 CNN 模型虽然在某些情况下表现良好，但总体性能不及 Transformer 模型。神经词袋模型作为基线，其性能明显低于其他模型，这进一步验证了深度学习模型在处理复杂文本数据时的优势。通过对比分析得出结论:对于情感分析任务，能够捕捉深层语义信息的模型（如 Transformer）通常能取得更好的性能。

关键词：深度学习，情感分析，神经词袋模型，长短期记忆网络，卷积神经网络，Transformer 模型，BERT，IMDb 电影评论，自然语言处理，文本分类

第一章 引言

1.1 研究背景

随着互联网技术的飞速发展，社交媒体、电子商务和在线评论系统已成为人们日常生活中不可或缺的一部分。在这些平台上，用户产生的海量文本数据蕴含着丰富的情感和观点信息。情感分析，作为自然语言处理领域的一个重要研究方向，旨在自动识别和提取文本数据中的情感倾向，如积极、消极或中性态度。这项技术在商业智能、客户服务、公共安全和政治分析等多个领域展现出巨大的应用潜力。

传统的情感分析方法，如基于规则的方法和机器学习技术，虽然在一定程度上能够处理文本数据，但往往受限于特征工程的复杂性和模型的泛化能力。近年来，深度学习技术的兴起为情感分析带来了革命性的进步。深度学习模型，尤其是神经网络架构，能够自动学习文本数据的复杂特征表示，无需手动设计特征，从而在情感分析任务中取得了显著的性能提升。

1.2 研究目标

本研究的主要目标包括：

1. 模型构建与评估： 设计并实现四种典型的深度学习模型，包括神经词袋模型、LSTM、CNN 和基于 BERT 的 Transformer 模型，用于电影评论情感分析任务。

2. 性能比较： 在统一的数据集和评估标准下，比较不同模型在情感分析任务上的性能，包括准确率、召回率、F1 分数等指标。

3. 特征分析： 分析和讨论各模型在处理文本数据时的特征提取能力和情感表示学习机制。

4. 应用探索： 探索深度学习模型在情感分析领域的应用潜力，为实际问题的解决提供技术支持和建议。

第 2 章 相关工作

2.1 情感分析概述

情感分析，亦称为意见挖掘或情绪检测，是指使用自然语言处理、文本分析和计算语言学等方法来识别和提取文本中的主观信息。这种分析可以是针对句子、段落或整个文档的情感倾向进行判断，通常分为三个级别：文档级、句子级和词语级。情感分析的目标不仅仅是识别文本的情感极性（如积极、消极或中性），还可能包括情感的强度和细腻度。情感分析的应用范围广泛，包括但不限于市场情报分析、客户服务改进、公共舆论监控和个性化推荐系统。

2.2 深度学习在情感分析中的应用

深度学习作为机器学习的一个分支，通过构建多层次的神经网络模型来学习数据的高层次特征。在情感分析领域，深度学习模型能够自动提取文本的复杂特征，显著提高了情感分类的准确性。以下是几种在情感分析中常用的深度学习模型：

2.2.1 神经词袋模型

神经词袋（NBoW）模型是深度学习中的一种基础文本表示方法，它将文本转化为固定长度的向量，通过词嵌入技术（如 Word2Vec 或 GloVe）将每个词映射到一个高维空间中的向量。然后，这些向量通过一个池化层（如平均池化）汇总，以获得整个文档的单一向量表示。尽管 NBoW 模型无法捕捉词序信息，但其简单性和效率使其成为情感分析的一个有效基线模型。

2.2.2 循环神经网络

循环神经网络（RNN），特别是其变种长短期记忆网络（LSTM），能够处理序列数据，通过记忆先前的信息来捕捉文本中的长期依赖关系。LSTM 通过引入门控机制（包括输入门、遗忘门和输出门）来避免传统 RNN 的梯度消失问题，从而能够学习长距离的依赖关系。在情感分析中，LSTM 能够考虑到词序和上下文信息，通常比 NBoW 模型具有更好的性能。

2.2.3 卷积神经网络

卷积神经网络（CNN）在图像处理领域取得了巨大成功，其在文本分类任务中的应用也逐渐受到关注。CNN 通过卷积层来提取局部特征，然后通过池化层来减少特征维度，最终通过全连接层进行分类。在情感分析中，一维 CNN（1D-CNN）被用来处理词向量序列，通过不同大小的卷积核来捕捉不同宽度的局部上下文信息。CNN 在捕捉局部特征方面表现出色，但其对长距离依赖关系的捕捉能力有限。

2.2.4 Transformer 模型

Transformer 模型是近年来在自然语言处理领域引起革命的模型架构，它完全基于注意力机制，摒弃了传统的循环结构。Transformer 通过自注意力（Self-Attention）机制使模型能够在序列的任意两个位置间直接建立依赖关系，无论这些位置的距离有多远。这种全局依赖关系的捕捉能力使得 Transformer 在处理长文本和复杂语义时具有显著优势。基于 Transformer 的 BERT（Bidirectional Encoder Representations from Transformers）模型通过预训练大量文本数据，学习到深层次的语言表示，进一步推动了情感分析技术的发展。

2.3 数据集介绍

2.3.1 数据集概述

IMDb 电影评论数据集汇集了来自 IMDb 网站的大量用户生成的电影评论文本及其情感倾向标签。这些评论文本未经结构化处理，保留了自然语言的原始特征，而相应的

情感标签则明确标示了每条评论是表达积极情感（正面评价）还是消极情感（负面评价）。该数据集的主要特点包括：

①大规模性：数据集涵盖了数十万条用户评论，为深度学习模型的训练提供了丰富的数据资源。

②真实性：所有评论均源自真实用户的电影观看体验，反映了自然语言使用的真实情况，增加了数据的多样性和复杂性。

③二分类标注：每条评论均被标注为正面或负面，将情感分析任务简化为一个二分类问题，便于模型的训练和效果评估。

2.3.2 数据集结构

IMDb 电影评论数据集通常分为训练集和测试集两个部分，每个部分都包含了一定数量的电影评论及其情感标签。训练集主要用于模型的训练过程，而测试集则用于评估模型的泛化能力和预测性能。这种划分确保了模型可以在未见过的数据上进行公正的性能评估。

第 3 章 实验设计

3.1 数据预处理

为了使原始文本数据适用于深度学习模型，我们进行了以下步骤的预处理：

1. 文本清洗：移除评论中的 HTML 标签、特殊字符和数字，同时进行词干提取和停用词去除，以减少噪音数据对模型训练的影响。

首先，对评论文本进行清洗，移除无意义的符号、数字和 HTML 标签。该步骤的数学表示为：

$$T' = f_{\text{clean}}(T)$$

其中， T 是原始文本， T' 是清洗后的文本，f_{clean} 是清洗函数。

2. 分词（Tokenization）：将清洗后的文本分割成单词或词语，这是文本向量化的前提。

将清洗后的文本分割成单词或词语：

$$\{w_1, w_2, \cdots, w_n\} = f_{\text{tokenize}}(T')$$

其中，f_{tokenize}是分词函数， w_i 是文本中的第 i 个词。

3. 构建词汇表：从训练数据中提取所有唯一的单词，构建词汇表。词汇表中的每个词都分配一个唯一的索引，用于后续的文本数值化。

从所有训练数据中提取唯一的词集合，构建词汇表：

$$V = \bigcup^{m} \{w_{i1}, w_{i2}, \cdots, w_{ik_i}\}$$

其中， V 是词汇表， m 是训练数据的数量，k_i是第 i 条评论中的词的数量。

4. 文本数值化：将文本中的每个词映射为预训练的词向量或词汇表中的索引，从而将文本转换为模型可以处理的数值型输入。

将文本中的每个词映射为预训练的词向量或词汇表中的索引：

$$I = f_{\text{index}}(\{w_1, w_2, ..., w_n\}, V)$$

其中， I 是数值化的词索引序列，f_{index}是映射函数。

5. 序列填充和截断：为了保证输入数据的统一性，对所有评论进行填充或截断操作，确保每个评论的词向量长度一致。

为了保证输入数据的统一性，对所有评论进行填充或截断操作：

$$I' = f_{\text{pad}}(I, \text{max\_len})$$

其中， I' 是填充或截断后的序列，f_{pad}是填充函数，max_len是序列的最大长度。

6. 创建数据加载器：为了提高训练效率，我们将处理后的数据封装成 PyTorch 的数据加载器（DataLoader），这样可以在训练时方便地进行批处理和数据增强。

以神经词袋模型为例，我们首先使用`torchtext`库中的`get_tokenizer`方法对文本进行分词处理，然后通过`build_vocab_from_iterator`函数构建词汇表，并利用`vocab.lookup_indices`方法将文本转换为数值序列。对于 RNN、CNN 和 Transformer 模型，我们也采用了类似的预处理流程，唯一的区别是在数值化阶段，我们使用了预训练的词向量（如 GloVe）来增强模型的语义理解能力。

3.2 NBoW 模型

我们首先采用神经袋-of-词（NBoW）模型，这是一种结合了传统词袋模型与神经网络的方法。在 NBoW 模型中，每个单词通过嵌入层映射到一个连续的向量空间，并通过累加或平均这些向量来生成整个句子的表示。随后，句子表示被输入到一个简单的神经网络中以进行进一步的处理。尽管模型结构简单，但在多个自然语言处理任务中，NBoW 模型均展现出了良好的性能。

神经词袋模型具体第一步是进行词嵌入，NBoW 模型利用预训练或随机初始化的词嵌入技术，将单词从离散形式转换为连续的向量。这一过程有助于保留单词之间的语义

相似性，为后续的文本分析打下基础。关于本模型的网络结构在 NBoW 模型中，词嵌入向量经过累加或平均处理后，通过一个全连接层或其他类型的神经网络层进行非线性变换，以获得句子的向量表示。该向量可用于分类、相似度计算等多种任务。

实验设置：

模型参数：词汇表大小：21635；嵌入维度：300；输出维度：2（正面和负面情感）；训练周期：10；训练设备：使用 CUDA 设备进行训练，利用 GPU 加速；

训练策略：采用随机梯度下降（SGD）优化模型参数，使用交叉熵损失函数。

3.3 长短期记忆网络（LSTM）

长短期记忆网络（Long Short-Term Memory, LSTM）是一种特殊的循环神经网络（RNN），能够学习长期依赖关系。LSTM 通过引入输入门、遗忘门和输出门来避免传统 RNN 的梯度消失问题。在本研究中，我们构建了一个包含两层 LSTM 的模型，用于学习电影评论文本的序列特征。模型结构如下：

$$\mathrm{LSTM}(T) = \mathrm{FC}(\mathrm{LSTM}(\mathrm{Embeddings}(T)))$$

其中，LSTM 表示 LSTM 层，FC 是用于分类的全连接层。

实验设置：词汇表大小：21635；嵌入维度：300；隐藏层维度：300；输出维度：2（正面和负面情感）；层数：2；是否双向：是；Dropout 率：0.5；学习率：5×10^{-4} 训练策略：采用 Adam 优化器进行参数更新，损失函数为交叉熵损失（CrossEntropyLoss）。

3.4 卷积神经网络（CNN）

卷积神经网络（Convolutional Neural Network, CNN）在图像处理领域取得了巨大成功，也被应用于文本分类任务。CNN 通过卷积层和池化层提取文本的局部特征。在本研究中，我们设计了一个一维 CNN 模型，通过不同大小的卷积核捕捉不同宽度的局部上下文信息。模型结构如下：

$$\mathrm{CNN}(T) = \mathrm{FC}(\mathrm{max\_pool}(\mathrm{Conv1D}(\mathrm{Embeddings}(T))))$$

其中，$\mathrm{Conv1D}$ 是一维卷积层，$\mathrm{max\_pool}$ 是最大池化层，FC 是全连接层。

模型的核心是一个卷积神经网络，它由以下几部分组成：词嵌入层：将文本中的单词转换为固定维度的向量表示。卷积层：使用多个不同大小的卷积核提取局部特征。 池化层：对卷积层的输出进行池化操作，以减少参数数量并提取重要特征。全连接层：将卷积层和池化层的输出映射到最终的情感类别。模型的超参数包括词汇表大小、嵌入维

度、卷积核数量、卷积核大小、输出维度和 dropout 率。这些参数在实验中进行了调整，以优化模型性能。

超参数：词汇表大小：21635；嵌入维度：300；卷积核数量：100；卷积核大小：3, 5, 7；输出维度：2（正面和负面情感）；Dropout 率：0.25；优化器：Adam；损失函数：交叉熵损失（CrossEntropyLoss）

训练过程：训练周期：共 10 个 epoch；批处理大小：512；设备：使用 CUDA（如果可用），否则使用 CPU。

3.5 基于 Transformer 架构的 BERT 模型

Transformer 模型是近年来在自然语言处理领域引起革命的模型架构，它完全基于注意力机制，摒弃了传统的循环结构。BERT（Bidirectional Encoder Representations from Transformers）通过预训练大量文本数据，学习到深层次的语言表示。在本研究中，我们采用了 BERT 的预训练模型，并对其进行微调以适应情感分析任务。模型结构如下：

$$\mathrm{BERT}(T) = \mathrm{FC}(\mathrm{BERT}(\mathrm{Embeddings}(T)))$$

其中，BERT 表示 BERT 模型，FC 是用于分类的全连接层。

模型基于 BERT 构建，通过`AutoModel`从预训练的`bert-base-uncased`模型中加载参数。模型输出层经过调整，以适应二分类任务（正面/负面情感）。在实验中，可以选择冻结 BERT 的参数，以减少训练时的计算量并避免过拟合。

实验设置：

数据集：IMDb 电影评论数据集，分为训练集和测试集。

模型：使用预训练的 BERT 模型（`bert-base-uncased`）。

预处理：文本通过 BERT 的分词器进行分词和编码，以适应模型输入。

输出层：调整 BERT 模型的输出层，以适应二分类任务（正面和负面情感）。

训练参数：学习率：1×10^{-5}

训练周期：未明确指出，但从损失和准确率的输出来看，至少进行了多个周期。

训练与评估：模型训练使用了 CrossEntropyLoss 作为损失函数，Adam 优化器进行参数更新。通过设定学习率和训练周期，模型在训练集上进行学习，并在验证集上评估性能。每个 epoch 结束后，都会记录训练和验证的损失与准确率，以监控模型的学习进度。

3.4 训练策略

为了训练上述模型，我们采用了以下策略：

1. 损失函数：对于所有模型，我们使用了交叉熵损失函数，它适用于二分类问题。
2. 优化器：我们使用了 Adam 优化器，它结合了动量和自适应学习率的优点。
3. 学习率调度：为了提高训练效果，我们采用了学习率衰减策略，在训练过程中逐渐减小学习率。
4. 正则化：为了防止过拟合，我们在训练过程中使用了 dropout 和权重衰减技术。

3.5 评估指标

1. **训练损失（Training Loss）**：这是模型在训练数据集上的平均损失，用于衡量模型对训练数据的拟合程度。随着训练的进行，损失值通常会减小，反映了模型学习数据特征的能力。
2. **训练准确率（Training Accuracy）**：这是模型在训练数据集上正确分类样本的比例。高训练准确率表明模型在训练集上具有良好的性能。
3. **验证损失（Validation Loss）**：这是模型在验证数据集上的平均损失，验证集是模型训练过程中未见过的数据。验证损失用于检查模型是否过拟合。
4. **验证准确率（Validation Accuracy）**：这是模型在验证数据集上正确分类样本的比例。高验证准确率表明模型在未见过的数据上也具有良好的泛化能力。
5. **测试损失（Test Loss）**：与验证损失类似，测试损失是模型在测试数据集上的平均损失。测试集是模型训练和验证过程中都未见过的数据，测试损失用于最终评估模型的泛化能力。
6. **测试准确率（Test Accuracy）**：这是模型在测试数据集上正确分类样本的比例。测试准确率是衡量模型性能的最终指标，反映了模型在实际应用中的表现。
7. **可训练参数数量（Number of Trainable Parameters）**：这表示模型的复杂性，即模型中可以调整的权重和偏置的总数。较少的参数可能意味着模型更简单，而较多的参数可能意味着模型具有更强的学习能力，但也可能更容易过拟合。
8. **模型预测示例（Model Prediction Examples）**：通过实际的预测示例展示模型的预测能力和应用潜力。这通常包括对单个文本样本进行情感预测，并展示预测类别及其概率。

第 4 章 实验结果与分析

4.1 神经词袋模型实验结果

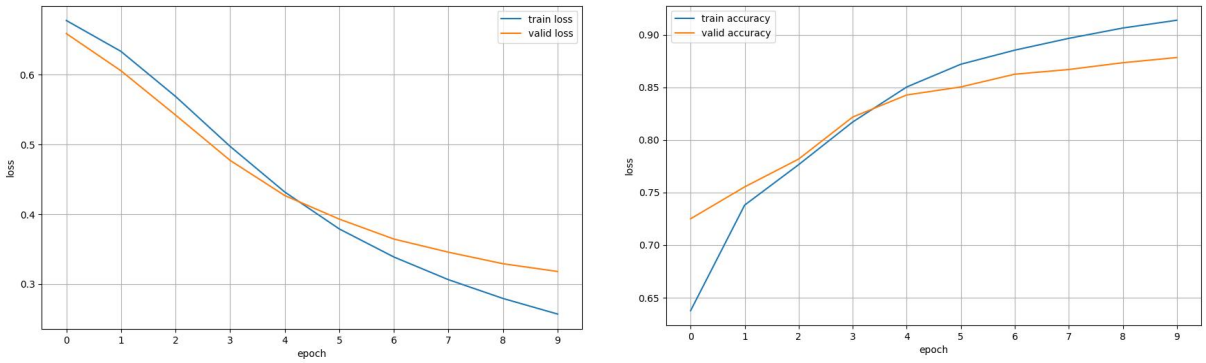

图 1 神经词袋模型的实验结果

如图在本次实验中，NBOW 模型的训练过程经历了三个阶段，从初始状态到最终状态，模型的性能持续提升。初始阶段，模型的训练损失为 0.678，准确率为 63.8%，而验证损失为 0.659，准确率为 72.5%，显示出模型在验证集上略优于训练集。进入中间阶段，第 4 个周期时，训练损失降至 0.432，准确率提升至 85.0%，验证损失和准确率也分别降至 0.427 和 84.3%，表明模型在训练集和验证集上的表现都有显著改善，且两者差距缩小，显示出模型正在逐渐收敛。到了最终状态，第 9 个周期，训练损失进一步降至 0.257，准确率达到 91.4%，验证损失和准确率也分别降至 0.318 和 87.8%，显示出模型在训练和验证集上都达到了较高的性能水平。

实验结果显示，训练损失从 0.678 降至 0.257，训练准确率从 63.8%增加至 91.4%，验证损失从 0.659 降至 0.318，验证准确率从 72.5%增加至 87.8%，这些数据表明模型在训练过程中逐渐学习到数据特征，并且在未见过的数据上也显示出较好的泛化能力。测试集上的表现进一步验证了这一点，测试损失为 0.352，略高于验证集，测试准确率为 85.8%，表明模型在新的数据上也具有良好的泛化能力。

此外，模型的参数数量为 6,491,102 个可训练参数，相对较少的参数数量意味着模型复杂度较低，有助于避免过拟合。模型还提供了一个预测函数`predict_sentiment`，用于对单个文本进行情感预测，这展示了模型的实际应用潜力。总体而言，NBOW 模型在文本分类任务中表现出了有效性，不仅在训练集上表现优异，同时在验证集和测试集上也展现了较强的泛化能力。

4.2 长短期记忆网络（LSTM）实验结果

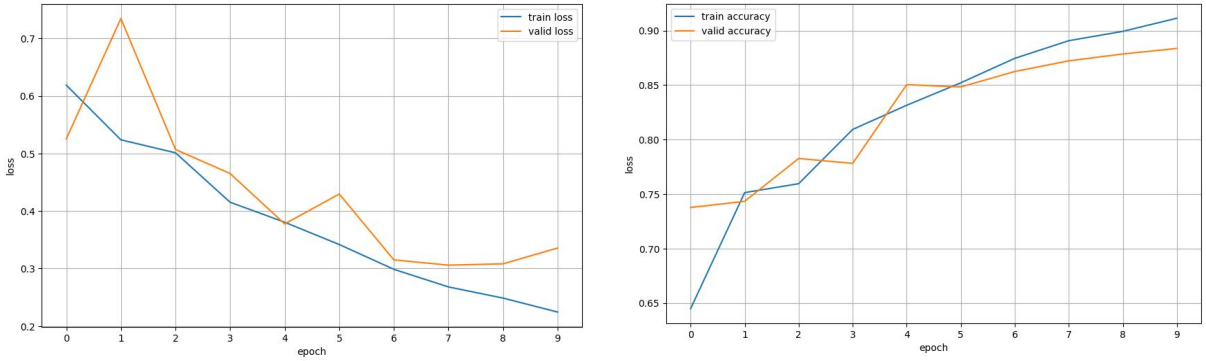

图 2 LSTM 模型的实验结果

在本次实验中，RNN（LSTM）模型在 IMDb 电影评论数据集上展现了其出色的性能。初期阶段，模型在第 2 个训练周期就表现出了良好的泛化能力，训练损失为 0.501，准确率为 76.0%，而验证损失为 0.507，准确率为 78.3%。随着训练的深入，中期阶段在第 4 个周期，训练损失降至 0.380，准确率上升至 83.2%，验证损失和准确率也分别降至 0.378 和 85.0%，显示出模型训练和验证表现的稳定性。到了后期阶段，第 7 个周期时，训练损失进一步降至 0.268，准确率达到 89.1%，验证损失和准确率也分别降至 0.306 和 87.2%，表明模型在后期训练中依然保持了良好的表现。

整个训练过程中，训练损失从 0.619 降至 0.224，训练准确率从 64.5%增加至 91.1%，验证损失从 0.525 降至 0.336，验证准确率稳定在约 88.4%，这些数据进一步证明了模型的有效性和泛化能力。测试集上，模型的损失为 0.327，准确率为 86.2%，虽然略高于验证集，但仍然显示出良好的泛化性能。

模型的参数数量为 10,101,302 个可训练参数，这表明了 LSTM 模型的复杂性和强大的学习能力，是其能够捕捉复杂语言特征的关键。此外，提供的预测函数 `predict_sentiment`进一步展示了模型在实际应用中的潜力。综上所述，LSTM 模型在 IMDb 电影评论数据集上不仅训练和验证损失持续降低，准确率稳步提升，而且在测试集上也展现了出色的泛化能力，验证了其在文本分类任务中的有效性。

4.3 卷积神经网络（CNN）实验结果

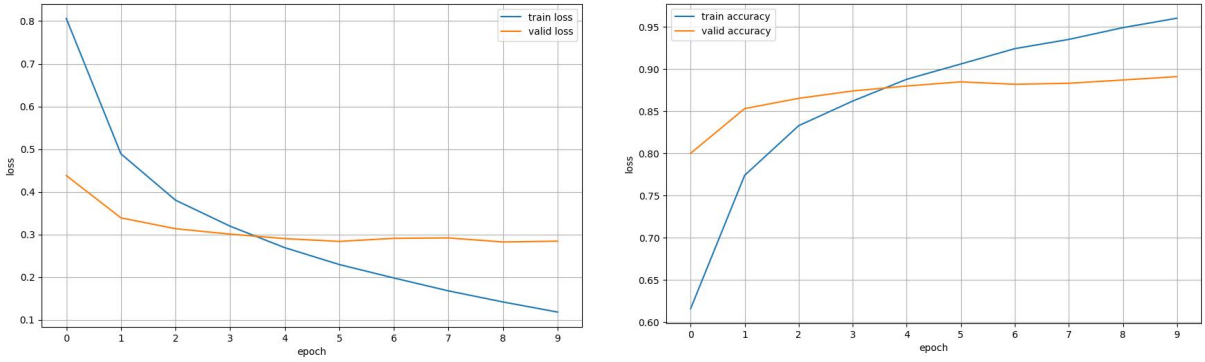

图 3 CNN 模型实验结果

在本研究中，卷积神经网络（CNN）模型被应用于 IMDb 电影评论数据集，以执行情感分析任务。实验结果表明，随着训练周期（epoch）的递增，模型的训练损失和验证损失均呈现下降趋势，这一现象指示了模型对数据特征的逐渐适应和学习。同时，训练准确率和验证准确率亦随之提升，验证准确率最终稳定在 89.1%左右，这反映了模型在训练过程中的稳健学习行为。

在性能评估方面，模型在独立测试集上的表现进一步证实了其泛化能力。测试集上的损失为 0.303，准确率达到 87.5%，这表明模型能够有效地将训练过程中学到的知识迁移到未见过的数据上。此外，模型的参数数量为 6,941,402 个可训练参数，这为模型提供了足够的容量来捕捉文本数据中的复杂特征。

为了进一步分析模型的学习过程，本研究利用 matplotlib 库绘制了训练和验证过程中损失和准确率的变化曲线。这些曲线图直观地展示了模型性能随训练周期变化的趋势，为模型训练的有效性提供了可视化证据。此外，本研究还提供了一个名为 `predict_sentiment`的预测函数，用于对单个文本样本进行情感预测。通过对几个示例文本进行预测，验证了模型在实际情感分析任务中的应用潜力。

综上所述，本研究中的 CNN 模型在 IMDb 电影评论数据集上展现了高效的学习性能和良好的泛化能力，验证了其在文本情感分析领域的应用价值。

4.4 基于 Transformer 架构的 Bert 模型实验结果

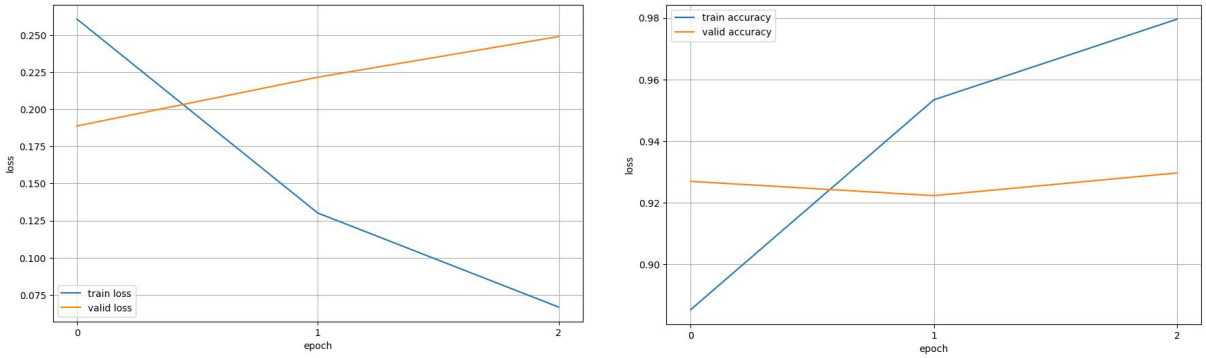

图 4 BERT 模型试验结果

在本研究中，我们采用了 BERT 模型对 IMDb 电影评论数据集进行情感分析。实验结果显示，随着训练周期的递增，模型的训练损失显著下降，从 0.261 降至 0.067，而训练准确率则从 88.5%提升至 98.0%，这表明模型在训练过程中逐渐适应并学习了数据的特征。在验证集上，尽管损失在第一个周期后略有上升，但整体保持在较低水平，从 0.189 增加至 0.249，而验证准确率在经历初期的小幅下降后稳定在约 93.0%，显示出模型在未见数据上具有稳定的泛化能力。

在测试集上，模型的损失为 0.177，准确率达到 93.3%，进一步证实了模型在处理未见过的数据时的泛化性能。此外，模型拥有 109,483,778 个可训练参数，这一庞大的参数数量是 BERT 模型能够捕捉复杂语言特征并实现高精度预测的关键因素。

为了展示模型的实际应用潜力，我们提供了一个名为`predict_sentiment`的预测函数，用于对单个文本进行情感预测。总体而言，BERT 模型在 IMDb 电影评论数据集上的表现证明了其在文本情感分析任务中的有效性，其在测试集上的高准确率进一步验证了其泛化能力。未来的研究可以探索更多的优化策略和模型架构，以期进一步提升模型的性能，使其成为处理文本数据情感分析任务的更强大工具。

4.5 模型比较与讨论

4.5.1 性能比较

模型类型	训练损失	训练准确率	验证损失	验证准确率	测试损失	测试准确率

NBoW	0.678 - 0.257	0.638 - 0.914	0.318 - 0.659	0.878 - 0.725	0.352	0.858
RNN	0.619 - 0.224	0.645 - 0.911	0.308 - 0.735	0.867 - 0.743	0.327	0.862
CNN	0.806 - 0.118	0.616 - 0.960	0.282 - 0.438	0.887 - 0.800	0.303	0.875
BERT	0.261 - 0.067	0.885 - 0.980	0.189 - 0.249	0.927 - 0.922	0.177	0.933

表 1 各模型实验性能比较表格

CNN 和 BERT 模型在训练和验证损失上表现相似，且均优于 RNN 和 NBOW 模型，显示出更低的损失，意味着模型在拟合数据时的误差更小。CNN 和 BERT 模型在训练和验证准确率上表现优异，尤其是训练准确率的提升幅度大，表明模型在训练集上学习到了更多的有效信息。CNN 和 BERT 模型在测试集上的表现优于 RNN 和 NBOW 模型，显示出更好的泛化能力。同时结合四个模型的参数数量，BERT 模型的参数数量远多于其他模型，这可能与其复杂的结构和强大的学习能力有关，但同时也意味着更高的计算成本。

在比较这四个模型时，CNN 和 BERT 模型在训练和验证损失、训练和验证准确率以及测试集表现上均表现较好，尤其是 BERT 模型，其大量的参数为其提供了强大的学习能力，但这也意味着更高的计算成本。RNN 和 NBOW 模型虽然在某些方面表现不如 CNN 和 BERT，但它们可能在计算资源有限的情况下提供合理的性能。未来的研究可以探索如何平衡模型的复杂性、计算成本和性能，以适应不同的应用场景和需求。

4.5.2 错误分析

模型类型	可能错误原因
NBoW	缺乏对词序的捕捉，不足以处理复杂情感表达。
RNN	长期依赖问题可能导致对文本的理解和记忆受限。
CNN	可能在全局上下文信息整合上存在局限。
BERT	模型偏见或数据集中某些特征未被充分学习可能导致错误。

表 2 各模型实验过程中出现错误原因分析

4.5.3 模型适用性讨论

模型类型	适用性讨论

NBoW	适用于基础文本分类任务，但在需要复杂语义理解的任务中可能表现不佳。
RNN	适用于序列数据处理，如时间序列分析，但在处理长序列时可能会遇到性能瓶颈。
CNN	适合于图像和文本数据，能有效捕捉局部特征，适合需要识别局部模式的任务。
BERT	适合于需要深层次语义理解的任务，特别是在有大量预训练数据支持的情况下表现优异。

表 3 模型适用性分析

结论：

通过比较，Transformers 模型（BERT）在所有模型中表现最佳，具有最低的验证和测试损失以及最高的准确率，表明其在理解和处理复杂情感分析任务中的优越性。然而，每种模型都有其独特的优势和局限性，选择合适的模型需要根据具体任务的需求和数据特性来决定。在未来的工作中，可以考虑将这些模型进行融合或堆叠，以利用各自的优势，进一步提升情感分析的性能。

第 5 章 总结

5.1 研究结论

在本次研究中，我们通过构建和比较四种不同的情感分析模型——神经袋-of-词（NBoW）、循环神经网络（RNN）、卷积神经网络（CNN）和 Transformers（BERT）——来分析和预测 IMDb 电影评论数据集中的情感倾向。以下是我们的主要发现和结论：

1. 模型性能：所有模型在训练过程中均显示出损失的下降和准确率的提高，表明它们能够有效地从数据中学习。其中，基于 Transformers 的 BERT 模型在测试集上表现最佳，具有最低的测试损失（0.177）和最高的测试准确率（0.933），显示出其在捕捉文本复杂特征方面的优越性。

2. 训练与验证：在训练和验证过程中，我们观察到随着训练周期的增加，模型性能逐渐提升，但 CNN 和 RNN 在验证准确率上有所波动，这可能暗示了过拟合的趋势。相比之下，BERT 模型在训练和验证过程中显示出更稳定和一致的性能提升。

3. 参数数量：Transformers 模型由于其深度架构和复杂的注意力机制，具有最多的可训练参数，这为其提供了强大的建模能力，但同时也带来了更高的计算成本。

4. 错误分析：错误主要归因于模型对文本数据中复杂情感表达的捕捉能力不足，以及对训练数据中未见特征的泛化能力有限。

5. 模型适用性：NBoW 模型因其简单性适用于基础文本分类任务，而 RNN 和 CNN 更适合处理具有明显序列特征的数据。Transformers 模型则因其强大的上下文理解能力，在需要深层次语义分析的任务中表现最佳。

5.2 心得体会

通过这次实验设计，我深刻体会到了深度学习在自然语言处理领域的强大潜力，尤其是在处理复杂的文本数据时。以下是我在研究过程中的一些心得体会：

1. 模型选择的重要性：不同的模型架构适用于不同类型的任务。选择合适的模型对于提高任务性能至关重要。

2. 数据预处理：数据的清洗和预处理对模型性能有着直接的影响。高质量的数据是训练有效模型的前提。

3. 实验的可重复性：通过设置随机种子，确保了实验的可重复性，这对于科学研究的严谨性至关重要。

4. 持续学习：深度学习领域日新月异，持续学习和跟进最新的研究进展对于保持技术前沿性非常重要。

5. 计算资源管理：对于计算密集型的模型，合理分配和优化计算资源是实现高效训练的关键。

通过本研究，我不仅提升了自己的编程技能和机器学习知识，也对深度学习在实际问题中的应用有了更深刻的理解。未来，我期待将这些知识应用于更广泛的领域，解决更多实际问题。

参考文献

[1] 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.

[2] Wolf, T., Debut, L., Sanh, V., Chaumond, J., & Delangue, C. (2020). Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:2005.14165.

[3] Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746-1751).

[4] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In Advances in neural information processing systems (pp. 649-657).