情感分析 实验报告

致理书院 郭士尧 2022012406

代码链接： https://cloud.tsinghua.edu.cn/f/4f13dc3b2d454440bb79/

1 基本思路

本实验中要求我们基于给定的 word2vec 模型和分词后的标注数据集，完成文本情感（正/负）分析。

实现了三种模型：

- MLP：全连接神经网络
- CNN：卷积神经网络
- RNN-LSTM：基于长短时记忆网络的循环神经网络

2 实验结果

2.1 MLP

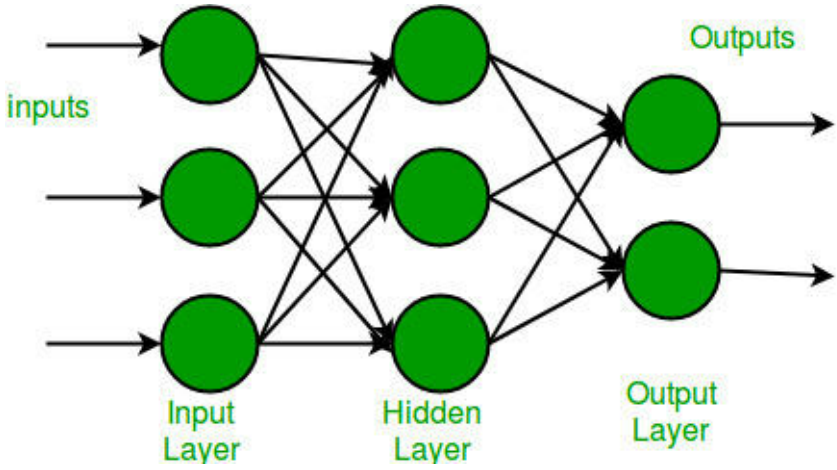
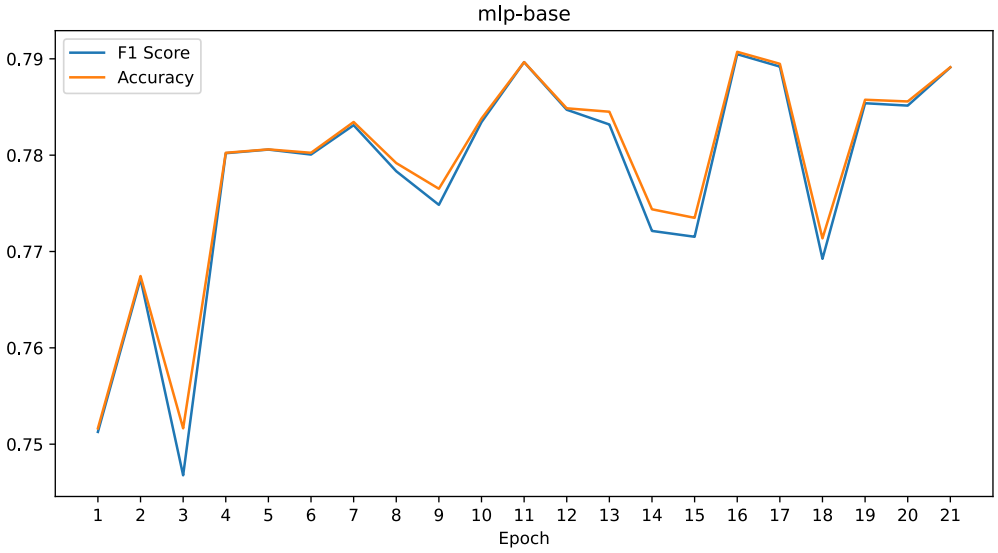


图 1　MLP 模型结构图

MLP 模型相对简单，将所有词向量经过多层全连接层和激活函数进行处理，最后使用 Max Pooling 和一层分类头进行分类。

下面为不同参数规模下 MLP 模型的得分与 Epoch 的关系图：



最高 F1 Score 为 0.8105，最高准确率为 81.06%。

2.2 CNN

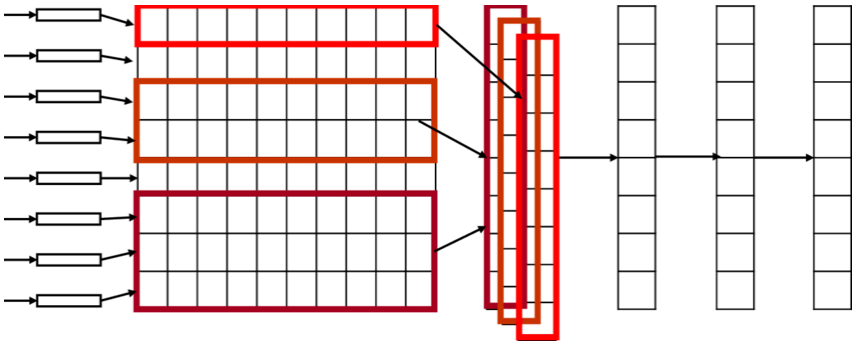
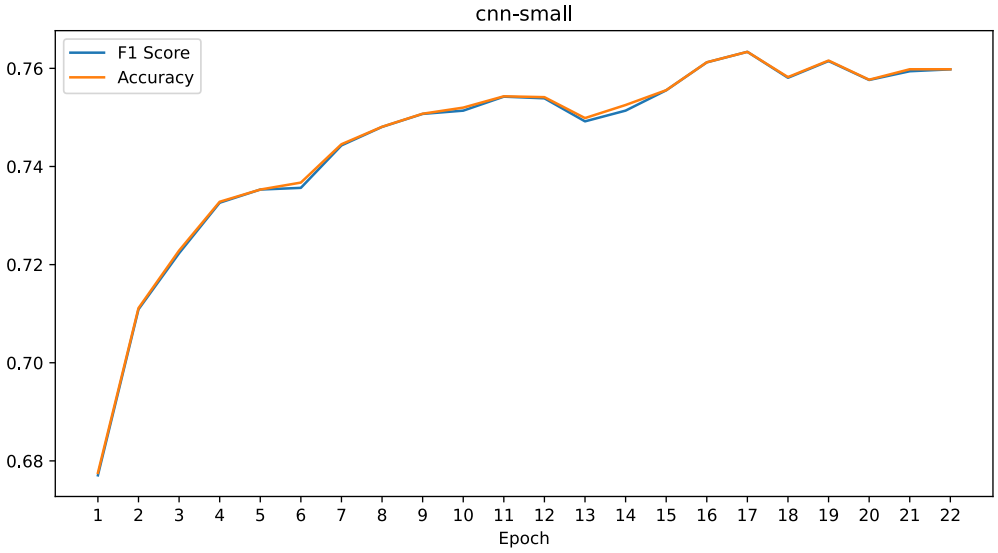
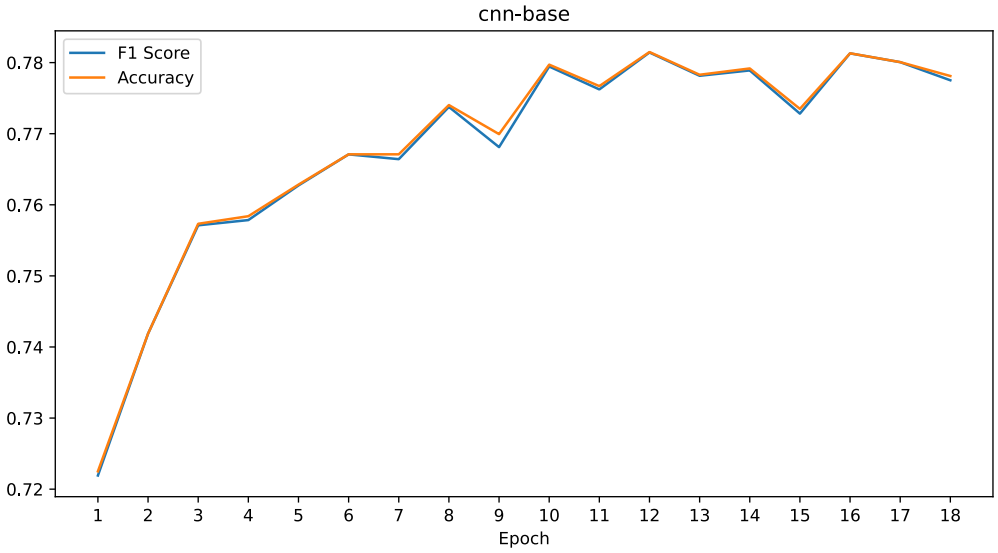


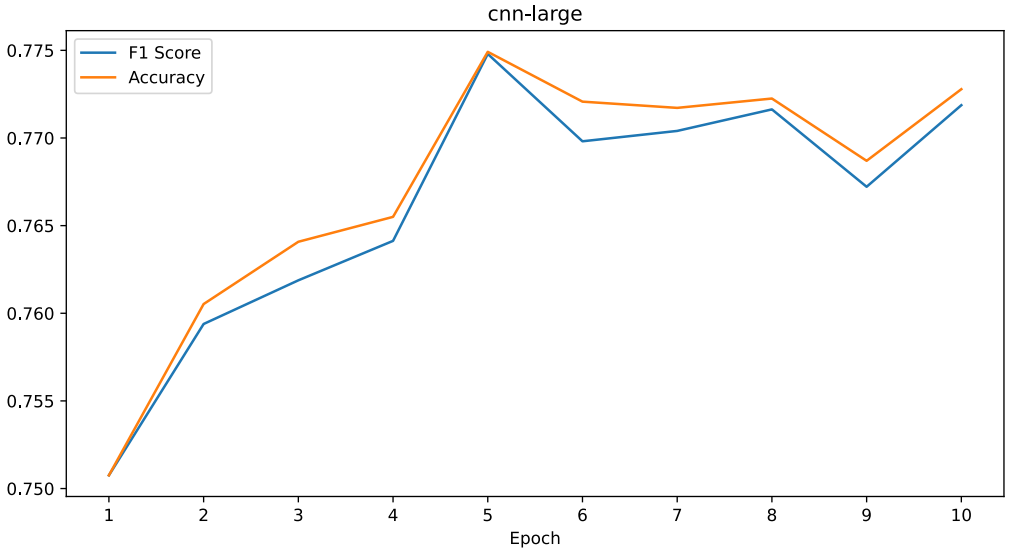
图 3　CNN 模型结构图

TextCNN 模型将句子视为二维矩阵，使用卷积核对矩阵进行卷积操作，本质是在提取相邻词语之间的局部特征。卷积操作后，使用 Max Pooling 层对特征进行降维处理，最后通过全连接层进行分类。

下面为不同参数规模下 CNN 模型的得分与 Epoch 的关系图：


cnn-small
F1 Score
Accuracy
0.76
0.74
0.72
0.70
0.68
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Epoch


cnn-base
F1 Score
Accuracy
0.78
0.77
0.76
0.75
0.74
0.73
0.72
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Epoch



最高 F1 Score 为 0.7814，最高准确率为 78.15%。

2.3 RNN-LSTM

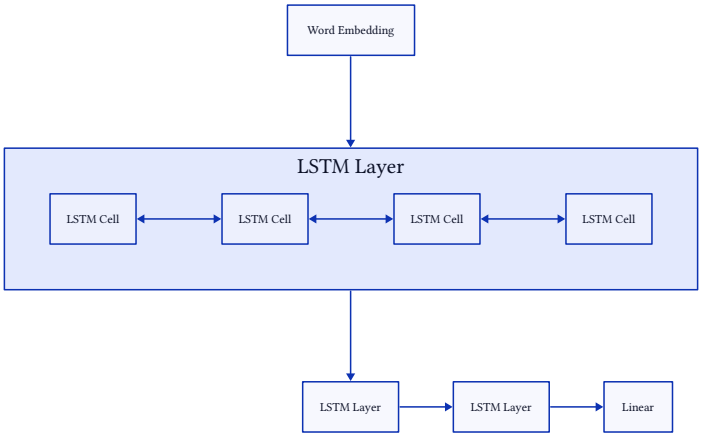
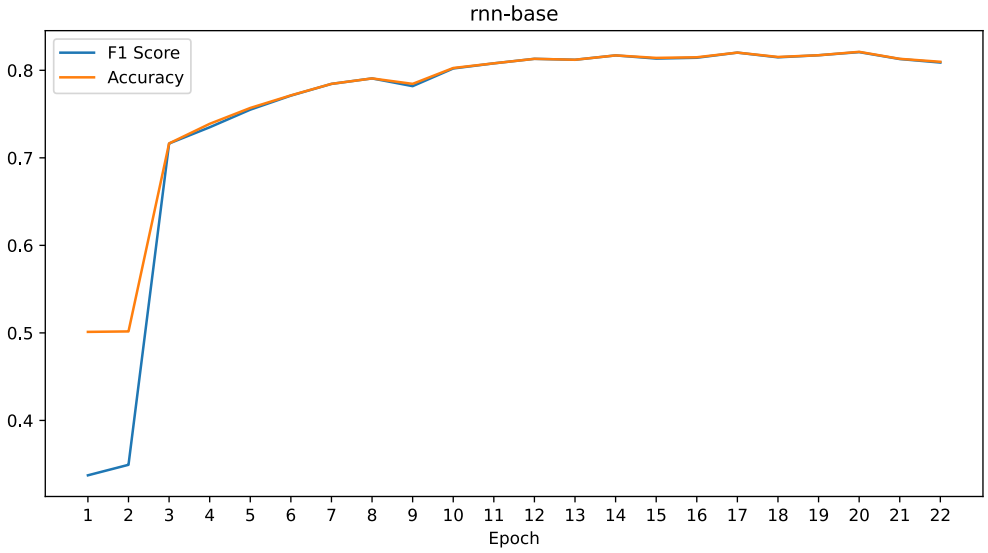
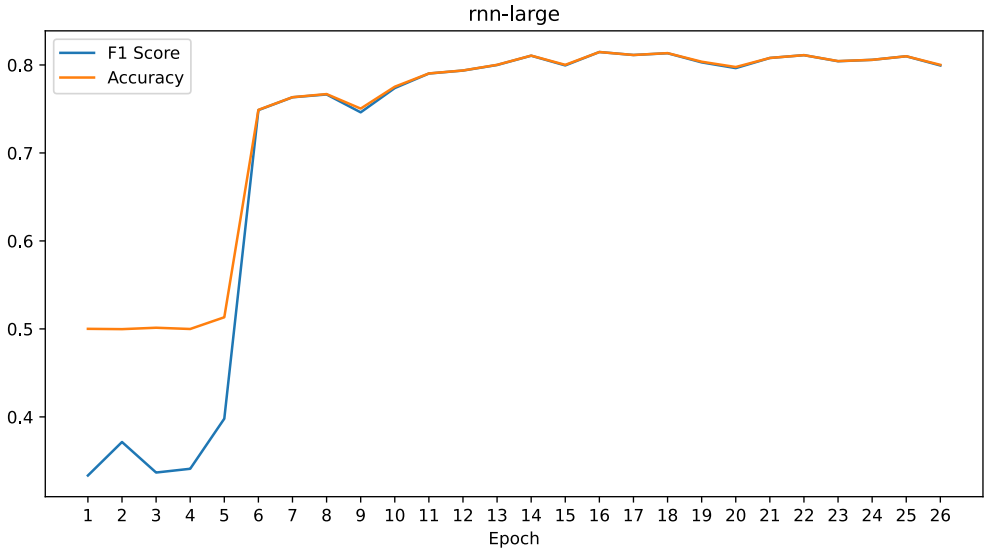


图 7　LSTM 模型结构图

LSTM 模型分为多层，每一层包含多个双向连接的 LSTM 单元。词向量经过 LSTM 层后，输出的隐藏状态被送入一个全连接层进行分类。LSTM 模型能够有效地捕捉文本中的长距离依赖关系，并且对上下文信息的建模也更为有效。

下面为不同参数规模下 RNN 模型的得分与 Epoch 的关系图：





最高 F1 Score 为 0.8208，最高准确率为 82.11%。

3 结果分析

和 MLP 对比，CNN 的效果相对较差，一方面可能是因为超参数选择不当，另一方面可能是任务本身相对简单，对局部信息依赖并不多，重要的是一些负面词汇的出现与否（如“坏”、“不好”、“烂”等），这种简单的特征使用 MLP 就有较好的结果。

可以看到，RNN-LSTM 的模型表现要优于其余两个模型，其原因在于 RNN-LSTM 模型通过引入记忆单元，能够有效地捕捉文本中的长距离依赖关系，同时对上下文信息的建模也更为有效，可以更好地理解文本的整体语义信息；而 CNN 模型则主要依赖于局部特征的提取，MLP 模型则是表达力整体较弱。

4 问题思考

4.1 实验训练什么时候停止是最合适的？简要陈述你的实现方式，并试分析固定迭代次数与通过验证集调整等方法的优缺点

在本实验中，我实现了基于 Loss 变化的早停策略。具体而言，当训练过程中发现在验证集上连续多次 Loss 大于最低 Loss $+\Delta$ 时，就停止训练。

固定迭代次数：

- 优点：简单易实现，训练时间可控。
- 缺点：可能会导致过拟合。

基于验证集调整：

- 优点：可以根据模型在验证集上的表现来动态调整训练过程，避免过拟合。
- 缺点：需要额外的验证集，增加了计算开销；同时不一定能取得最好的结果。

4.2 实验参数的初始化是怎么做的？不同的方法适合哪些地方？（现有的初始化方法为零均值初始化，高斯分布初始化，正交初始化等）

- 零均值初始化 & 高斯分布初始化：将参数初始化为均值为 0 的均匀分布（或高斯分布），相对于直接使用 0 初始化，能够避免神经元之间的对称性问题（引入随机性），适用于大部分场景。
- 正交初始化：将参数初始化为正交矩阵，能够保持输入和输出的方差不变，可以避免梯度消失或爆炸的问题，适用于深度网络。
- Xavier 初始化：根据输入和输出的神经元数量来初始化参数，能够保持每层的方差一致，适用于 Sigmoid 和 Tanh 激活函数。
- Kaiming 初始化：基于 Xavier 初始化，适用于 ReLU 激活函数，能够保持每层的方差一致。

4.3 过拟合是深度学习常见的问题，有什么方法可以防止训练过程陷入过拟合

一方面是使用前面提到的早停策略，另一方面是使用 Dropout 技术，通过在训练过程中随机丢弃一部分神经元来模拟同时训练多个模型，从而提高模型的泛化能力，防止过拟合。

4.4 试分析 CNN，RNN，全连接神经网络（MLP）三者的优缺点

- MLP
  - 优点：简单直接，容易实现
  - 缺点：模型表达能力有限
- CNN
  - 优点：能够提取局部特征，适合处理图像和文本数据
  - 缺点：对长距离依赖关系建模能力不足
- RNN
  - 优点：能够处理序列数据，适合处理长距离依赖关系
  - 缺点：训练时间较长，容易出现梯度消失或爆炸的问题