# 实验3：使⽤卷积神经⽹络TextCNN进⾏文本识别报告

# 实验背景

实验使用THUCNews数据集进行文本分类任务。该数据来源于新浪新闻RSS订阅频道2005至2011年间的历史数据，经过筛选和过滤，共包含14个类别：财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐。实验使用了一个抽样后的子集，其中包含180,000条训练数据，10,000条验证数据和10,000条测试数据。

在文本分类任务中，卷积神经网络的核心思想是通过滑动窗口捕捉文本中的N-gram特征。这些N-gram特征对应于文本的局部特征，CNN通过自动组合和筛选这些特征，能够提取出更抽象的语义信息，从而实现文本分类的目标。

# 实验方案设计

实验使用PyTorch框架实现了一个TextCNN模型，并对THUCNews数据集进行训练和测试。

首先对数据集进行预处理，将文本转化为索引、构建词嵌入矩阵。在模型设计上，TextCNN由若干个卷积层和池化层组成，用于提取文本的局部特征。实验代码设计了三个不同的卷积核，分别为大小为2、3和4的卷积核，这样能够捕捉到不同粒度的N-gram特征。每个卷积核的输出通过最大池化层进行下采样，然后将所有池化层的输出拼接起来，最后通过一个全连接层实现分类。

实验过程中，使用了AdamW优化器和交叉熵损失函数进行模型训练。模型的超参数，如学习率、卷积核数量、隐藏层单元数、批次大小等，均根据实验要求进行合理选择。

实验的目标是通过多次迭代训练，最终在测试集上获得不低于75%的分类准确率。

# 实验过程

首先进行数据预处理，读取训练数据和测试数据，将文本转化为单词索引，并构建对应的词嵌入矩阵。此后，设计并实现了TextCNN模型，并在训练数据集上进行训练。训练过程包括多轮迭代，每一轮训练后，使用验证集进行模型性能评估，记录模型在验证集上的准确率。

在训练过程中，损失函数值逐渐降低，这表明模型正在逐步学习到有效的特征。与此同时，我们观察到验证集上的准确率逐渐提高，这意味着模型的泛化能力有所增强。每一轮训练后，如果验证集上的准确率超过先前记录的最佳成绩，就保存当前的模型参数为文件，到args预设的model文件夹中。

实验中，损失函数的下降过程和验证集性能的变化情况均表明模型在训练过程中稳定收敛。为了进一步验证模型的有效性，我们在训练结束后使用测试集对模型进行评估，并记录模型的最终准确率。

# 性能测试

通过多次迭代训练， TextCNN模型最终在测试集上实现了80.65% 的准确率，达到了实验要求的75%以上。实验输出结果见附录。

# 实验习题

1. 简述文本是如何转变为word\_id，再转变为word embedding的，并说明为什么要将文本从字符转化为word embedding（提示：built\_curpus函数实现了这一过程）

文本转化为word embedding的原因在于，字符级别的表示难以捕捉到上下文的语义信息，而词嵌入通过将单词映射到向量空间，能够更好地表示单词之间的语义关系，进而提高模型的分类效果。

1. 为什么本次实验要构建开发集dev？开发集有什么作用？

实验构建了开发集，其作用在于评估模型的泛化能力，防止模型过拟合。在每一轮训练后，模型会在开发集上进行评估，选择表现最优的模型参数用于最终测试。

1. 在Transformers等架构流行的今天，TextCNN在性能上的优势已经不显著，但是一些线上服务仍然使用TextCNN作为文本分类模型，请说明可能的理由

TextCNN模型结构简单，计算效率高，且对于中小型数据集，TextCNN能够在较短时间内实现较好的分类效果，这使得它在需要快速部署的应用场景中仍然具有优势。

# 附录：代码输出结果

[1/ 1000] Train\_Loss:1.7370

[1/ 2000] Train\_Loss:1.2561

[1/ 3000] Train\_Loss:1.0956

[1/ 4000] Train\_Loss:1.0269

[1/ 5000] Train\_Loss:0.9507

[1/ 5625] Train\_Loss:0.9180

dev acc:0.7071

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[2/ 1000] Train\_Loss:0.8667

[2/ 2000] Train\_Loss:0.8564

[2/ 3000] Train\_Loss:0.8323

[2/ 4000] Train\_Loss:0.8217

[2/ 5000] Train\_Loss:0.7942

[2/ 5625] Train\_Loss:0.7812

dev acc:0.7433

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[3/ 1000] Train\_Loss:0.7440

[3/ 2000] Train\_Loss:0.7428

[3/ 3000] Train\_Loss:0.7416

[3/ 4000] Train\_Loss:0.7323

[3/ 5000] Train\_Loss:0.7299

[3/ 5625] Train\_Loss:0.7135

dev acc:0.7599

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[4/ 1000] Train\_Loss:0.6772

[4/ 2000] Train\_Loss:0.6800

[4/ 3000] Train\_Loss:0.6819

[4/ 4000] Train\_Loss:0.6848

[4/ 5000] Train\_Loss:0.6848

[4/ 5625] Train\_Loss:0.6767

dev acc:0.7660

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[5/ 1000] Train\_Loss:0.6308

[5/ 2000] Train\_Loss:0.6458

[5/ 3000] Train\_Loss:0.6533

[5/ 4000] Train\_Loss:0.6573

[5/ 5000] Train\_Loss:0.6388

[5/ 5625] Train\_Loss:0.6482

dev acc:0.7770

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[6/ 1000] Train\_Loss:0.5946

[6/ 2000] Train\_Loss:0.6208

[6/ 3000] Train\_Loss:0.6043

[6/ 4000] Train\_Loss:0.6219

[6/ 5000] Train\_Loss:0.6288

[6/ 5625] Train\_Loss:0.6404

dev acc:0.7762

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[7/ 1000] Train\_Loss:0.5758

[7/ 2000] Train\_Loss:0.5952

[7/ 3000] Train\_Loss:0.5899

[7/ 4000] Train\_Loss:0.6041

[7/ 5000] Train\_Loss:0.5978

[7/ 5625] Train\_Loss:0.6064

dev acc:0.7796

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[8/ 1000] Train\_Loss:0.5552

[8/ 2000] Train\_Loss:0.5820

[8/ 3000] Train\_Loss:0.5727

[8/ 4000] Train\_Loss:0.5817

[8/ 5000] Train\_Loss:0.5815

[8/ 5625] Train\_Loss:0.5810

dev acc:0.7858

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[9/ 1000] Train\_Loss:0.5508

[9/ 2000] Train\_Loss:0.5451

[9/ 3000] Train\_Loss:0.5667

[9/ 4000] Train\_Loss:0.5623

[9/ 5000] Train\_Loss:0.5701

[9/ 5625] Train\_Loss:0.5732

dev acc:0.7855

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[10/ 1000] Train\_Loss:0.5290

[10/ 2000] Train\_Loss:0.5375

[10/ 3000] Train\_Loss:0.5451

[10/ 4000] Train\_Loss:0.5530

[10/ 5000] Train\_Loss:0.5547

[10/ 5625] Train\_Loss:0.5725

dev acc:0.7884

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[11/ 1000] Train\_Loss:0.5227

[11/ 2000] Train\_Loss:0.5327

[11/ 3000] Train\_Loss:0.5329

[11/ 4000] Train\_Loss:0.5436

[11/ 5000] Train\_Loss:0.5390

[11/ 5625] Train\_Loss:0.5527

dev acc:0.7880

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[12/ 1000] Train\_Loss:0.5016

[12/ 2000] Train\_Loss:0.5246

[12/ 3000] Train\_Loss:0.5207

[12/ 4000] Train\_Loss:0.5382

[12/ 5000] Train\_Loss:0.5359

[12/ 5625] Train\_Loss:0.5372

dev acc:0.7919

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[13/ 1000] Train\_Loss:0.4878

[13/ 2000] Train\_Loss:0.4983

[13/ 3000] Train\_Loss:0.5207

[13/ 4000] Train\_Loss:0.5261

[13/ 5000] Train\_Loss:0.5411

[13/ 5625] Train\_Loss:0.5356

dev acc:0.7922

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[14/ 1000] Train\_Loss:0.4813

[14/ 2000] Train\_Loss:0.5062

[14/ 3000] Train\_Loss:0.5133

[14/ 4000] Train\_Loss:0.5099

[14/ 5000] Train\_Loss:0.5256

[14/ 5625] Train\_Loss:0.5212

dev acc:0.7924

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[15/ 1000] Train\_Loss:0.4813

[15/ 2000] Train\_Loss:0.4824

[15/ 3000] Train\_Loss:0.5107

[15/ 4000] Train\_Loss:0.5179

[15/ 5000] Train\_Loss:0.5155

[15/ 5625] Train\_Loss:0.5107

dev acc:0.7938

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[16/ 1000] Train\_Loss:0.4767

[16/ 2000] Train\_Loss:0.4848

[16/ 3000] Train\_Loss:0.4857

[16/ 4000] Train\_Loss:0.4976

[16/ 5000] Train\_Loss:0.5156

[16/ 5625] Train\_Loss:0.5210

dev acc:0.7910

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[17/ 1000] Train\_Loss:0.4668

[17/ 2000] Train\_Loss:0.4781

[17/ 3000] Train\_Loss:0.5048

[17/ 4000] Train\_Loss:0.4974

[17/ 5000] Train\_Loss:0.5026

[17/ 5625] Train\_Loss:0.4976

dev acc:0.7923

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[18/ 1000] Train\_Loss:0.4631

[18/ 2000] Train\_Loss:0.4774

[18/ 3000] Train\_Loss:0.4888

[18/ 4000] Train\_Loss:0.4908

[18/ 5000] Train\_Loss:0.5062

[18/ 5625] Train\_Loss:0.4943

dev acc:0.7943

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[19/ 1000] Train\_Loss:0.4595

[19/ 2000] Train\_Loss:0.4633

[19/ 3000] Train\_Loss:0.4832

[19/ 4000] Train\_Loss:0.4866

[19/ 5000] Train\_Loss:0.5035

[19/ 5625] Train\_Loss:0.4971

dev acc:0.7930

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[20/ 1000] Train\_Loss:0.4543

[20/ 2000] Train\_Loss:0.4637

[20/ 3000] Train\_Loss:0.4858

[20/ 4000] Train\_Loss:0.4953

[20/ 5000] Train\_Loss:0.4764

[20/ 5625] Train\_Loss:0.4985

dev acc:0.7946

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[21/ 1000] Train\_Loss:0.4507

[21/ 2000] Train\_Loss:0.4560

[21/ 3000] Train\_Loss:0.4811

[21/ 4000] Train\_Loss:0.4865

[21/ 5000] Train\_Loss:0.4813

[21/ 5625] Train\_Loss:0.4895

dev acc:0.7916

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[22/ 1000] Train\_Loss:0.4529

[22/ 2000] Train\_Loss:0.4587

[22/ 3000] Train\_Loss:0.4694

[22/ 4000] Train\_Loss:0.4786

[22/ 5000] Train\_Loss:0.4823

[22/ 5625] Train\_Loss:0.4867

dev acc:0.7911

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[23/ 1000] Train\_Loss:0.4370

[23/ 2000] Train\_Loss:0.4638

[23/ 3000] Train\_Loss:0.4638

[23/ 4000] Train\_Loss:0.4751

[23/ 5000] Train\_Loss:0.4797

[23/ 5625] Train\_Loss:0.4951

dev acc:0.7968

已保存最佳模型

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[24/ 1000] Train\_Loss:0.4399

[24/ 2000] Train\_Loss:0.4438

[24/ 3000] Train\_Loss:0.4664

[24/ 4000] Train\_Loss:0.4774

[24/ 5000] Train\_Loss:0.4831

[24/ 5625] Train\_Loss:0.4832

dev acc:0.7920

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[25/ 1000] Train\_Loss:0.4255

[25/ 2000] Train\_Loss:0.4520

[25/ 3000] Train\_Loss:0.4675

[25/ 4000] Train\_Loss:0.4702

[25/ 5000] Train\_Loss:0.4750

[25/ 5625] Train\_Loss:0.4859

dev acc:0.7967

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[26/ 1000] Train\_Loss:0.4375

[26/ 2000] Train\_Loss:0.4419

[26/ 3000] Train\_Loss:0.4598

[26/ 4000] Train\_Loss:0.4772

[26/ 5000] Train\_Loss:0.4700

[26/ 5625] Train\_Loss:0.4759

dev acc:0.7948

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[27/ 1000] Train\_Loss:0.4292

[27/ 2000] Train\_Loss:0.4419

[27/ 3000] Train\_Loss:0.4636

[27/ 4000] Train\_Loss:0.4672

[27/ 5000] Train\_Loss:0.4698

[27/ 5625] Train\_Loss:0.4738

dev acc:0.7962

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[28/ 1000] Train\_Loss:0.4204

[28/ 2000] Train\_Loss:0.4480

[28/ 3000] Train\_Loss:0.4608

[28/ 4000] Train\_Loss:0.4651

[28/ 5000] Train\_Loss:0.4685

[28/ 5625] Train\_Loss:0.4768

dev acc:0.7951

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[29/ 1000] Train\_Loss:0.4241

[29/ 2000] Train\_Loss:0.4394

[29/ 3000] Train\_Loss:0.4574

[29/ 4000] Train\_Loss:0.4727

[29/ 5000] Train\_Loss:0.4648

[29/ 5625] Train\_Loss:0.4659

dev acc:0.7931

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[30/ 1000] Train\_Loss:0.4313

[30/ 2000] Train\_Loss:0.4390

[30/ 3000] Train\_Loss:0.4542

[30/ 4000] Train\_Loss:0.4522

[30/ 5000] Train\_Loss:0.4708

[30/ 5625] Train\_Loss:0.4718

dev acc:0.7927

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

运行时间：24.8046 min

test dataset accuracy:0.8065