Modeling and Simulation 建模与仿真, 2024, 13(2), 1322-1332
Published Online March 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.132124

Hans汉斯

基于自注意力机制改进GCNN模型的图书标签分类研究

张　健

上海理工大学管理学院，上海

收稿日期：2024年1月6日；录用日期：2024年3月13日；发布日期：2024年3月20日

摘　要

针对卷积神经网络聚焦于局部特征，不足以捕捉文本中长程依赖关系的问题，本文提出了一种基于CNN和自注意力机制改进的双通道图书标签分类模型(Gate Convolution Neural Network based on self-attention mechanism, GCNN-SAM)。该模型使用skip-gram将词嵌入成稠密低纬的向量，得到文本嵌入矩阵，分别输入到门卷积神经网络和自注意力机制，再经过逐点卷积，将两个通道中经过特征提取层得到的特征进行融合用于图书标签分类。在复旦大学中文文本分类数据集上进行对比实验，相较于SCNN、GCNN和其它改进的模型，测试集准确率达到96.21%，表明了GCNN-SAM模型在图书标签分类上具有优越性。同时，为验证GCNN-SAM模型的有效性，消融实验结果表明GCNN-SAM模型相较于CNN、GCNN和CNN-SAM在分类准确率上分别提升了5.9%、3.19%和3.66%。

关键词

图书标签分类，门卷积神经网络，自注意力机制，双通道

Research on Book Label Classification Based on Improved GCNN Model Based on Self-Attention Mechanism

Jian Zhang

Business School, University of Shanghai for Science and Technology, Shanghai

Received: Jan. 6th, 2024; accepted: Mar. 13th, 2024; published: Mar. 20th, 2024

Abstract

Aiming at the problem that convolutional neural networks focus on local features and are not enough to capture long-range dependencies in text, in order to improve this problem, this paper

文章引用: 张健. 基于自注意力机制改进 GCNN 模型的图书标签分类研究[J]. 建模与仿真, 2024, 13(2): 1322-1332.
DOI: 10.12677/mos.2024.132124

proposes an improved dual-channel book label classification model based on CNN and self-attention mechanism. The model uses skip-gram to embed words into dense low-latitude vectors to obtain a text embedding matrix, which is input into the gate convolutional neural network and self-attention mechanism respectively. After point-by-point convolution, the features obtained by the feature extraction layer in the two channels are fused for book label classification. A comparative experiment was conducted on the Fudan University Chinese text classification dataset. Compared with SCNN, GCNN and other improved models, the accuracy of the test set is 96.21%, which shows that the GCNN-SAM model has advantages in book label classification. At the same time, in order to verify the effectiveness of the GCNN-SAM model, ablation experiments were carried out. The results showed that the GCNN-SAM model improved the classification accuracy by 5.9%, 3.19% and 3.66% respectively compared with CNN, GCNN and CNN-SAM.

Keywords

Book Label Classification, Gate Convolutional Neural Network, Self-Attention Mechanism, Dual Channel

1. 引言

图书标签分类广泛应用于图书目录分类、推荐系统、信息检索和数字化等图书情报领域中[1]，其目的是将图书自动分配到一个或多个已定义好的类别中，可以帮助读者更容易地找到相关主题的书籍。图书标签是指添加到图书中的一组关键词或元数据，用于描述该图书的主题、内容、类型等丰富的语义信息[2]，然而由于图书标签的质量参差不齐，如标签语义信息与图书主题不一致，标签重复使用，同义、近义关系频现等[3]，当前大部分以《中国图书馆分类法》作为标签，主要是以基于特征工程和分类器的传统机器学习与基于自动编码器进行特征提取的深度学习文本分类两种方法[4]。

机器学习技术主要分为两步。第一步，需要手工对文本中的特征进行标注，对于图书标签分类，可以通过挖掘图书的元数据、目录、标签等信息，提取与图书内容和主题相关的特征。第二步，需要将提取的特征送入分类器，使用分类算法对图书进行分类。主流的分类算法有朴素贝叶斯分类器[5]、支持向量机[6]和多层感知机[7]，这些方法的优势在于计算量小，但往往都假设特征之间是独立的，无法利用特征之间的非线性关系将深层的重要语言特征纳入考量，例如句法歧义、句法多样性和主题适度等。由于CNN 在图像分类任务中表现良好，因此 Sergey 提出了使用一个简单而有效的 CNN 神经网络架构进行图书标签分类，这将有助于更好地组织和检索文本数据。

目前，深度学习模型已经成为解决图书标签分类问题的主流基础模型，许多学者将基于 CNN 改进的深度学习模型应用到图书标签分类中，不仅实现自动化特征提取，适应不同的输入尺寸和形状，还扩展到更大的数据集和更多的标签分类任务中。Wang 等人首先使用卷积神经网络(CNN)对文本使用手工特征提取器提取关键特征，使用循环神经网络(RNN)对提取的特征进行序列建模，最后引入注意力机制来加强模型对重要特征的关注度[8]。Zhao 等人提出在 TextCNN 的基础上增加一个层次化结构的自注意力卷积神经网络(HCNN-SAM)，将句子划分为单词和短语的序列，并使用 TextCNN 对每个序列进行特征提取，再通过层次化的自注意力机制以捕捉关键词和短语的文本特征[9]。相比于传统的基于序列模型的文本标签分类方法，Peng 等人提出基于门卷积神经网络(GCNN)使用文本数据的图来表示文本之间的关系，能

够充分利用文本中的上下文信息和关系，同时避免了传统方法中基于 n-gram 的特征提取方式带来的稀疏性问题[10]。以上模型虽然在提升图书标签分类效果上表现良好，但大多数基于词向量进行训练，未能充分利用文本信息可能会出现过拟合或性能下降的问题。此外，它们在处理图书标签分类时只考虑了有限的上下文，因此，我们需要综合考虑文本信息和语境、背景等其他相关信息，并使用更加全面和深入的算法学习到高层次的语义特征，从而更好地理解文本的含义。

本文使用卷积神经网络(CNN)可以提取文本中的特征，并且使用多个不同大小的卷积核可以捕获不同长度的关键信息，从而更好地理解文本。门控卷积神经网络(Gate Convolution Neural Network, GCNN)相较于 TextCNN 和 HCNN，GCNN 可以捕捉文本中的更长程依赖关系。自注意力机制(Self-Attention Mechanism, SAM)相较于单头注意力机制和多头注意力机制，可以自由地对输入的不同部分进行加权，并进行多个头的计算和合并更加全面地提取特征信息，充分利用输入序列之间的相互关系，从而更加准确全面地进行预测。

因此，本文为了充分利用文本信息捕捉最值得关注的特征，进一步提高文本分类的效率，提出了一种基于 GCNN 的自注意力机制文本分类模型 GCNN-SAM (GCNN-Self-Attention Mechanism)。

本文的主要贡献总结如下：

1) 设计了双通道特征提取架构，对相同上下文的不同层次文本特征提取时，能够获取长距离依赖关系的重要文本表示。

2) 提出了利用逐点卷积来收集两个通道得到的重要特征，从而很好地整合了卷积操作分离的通道间信息。

3) 构建了基于自注意力机制的双通道文本分类模型，在 Fudan 数据集上进行实验，实验结果表明，该模型的整体性能达到 96.21%，超过所比对的基线模型。

2. 基于 GCNN-SAM 的文本分类模型

GCNN-SAM 模型首先由第一通道的 GCNN + PC 模型构成，提取上下文的层次特征，更容易捕获长距离依赖关系，提高模型非线性提取特征的能力[11]。第二通道由 SVM + PC 模型构成，提取上下文全局的文本信息，解决长距离依赖问题，并且可以平行的计算，大大提高了模型计算效率[12]。第一通道和第二通道提取到的重要特征拼接作为文本的特征表示，最后经过 softmax 函数输出文本类别的概率分布。GCNN-SAM 模型结构如图 1 所示：

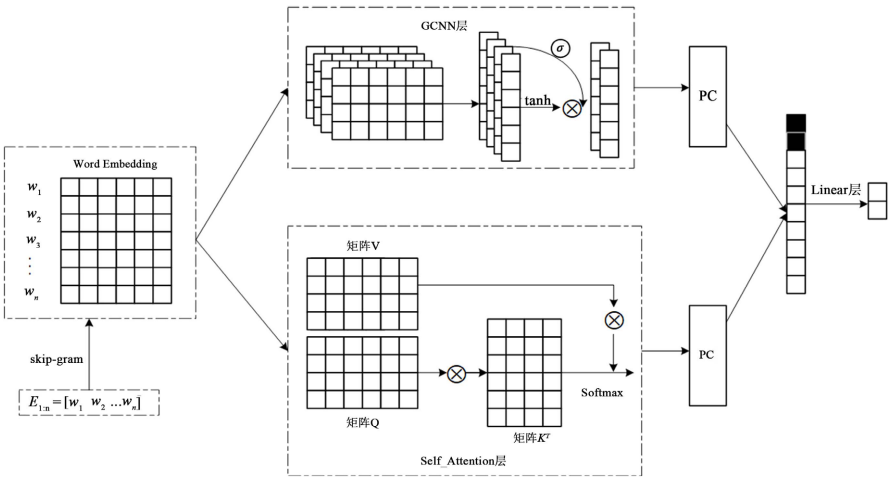

Figure 1. Improved GCNN model structure based on Self-Attention
图 1. 基于 Self-Attention 改进的 GCNN 模型结构

2.1. 词嵌入层

Mikolov 等在 2013 年同时提出了 CBOW 和 skip-gram 模型，这两个模型都是基于无监督的学习方式，是最常用的词嵌入技术之一，主要目的是将词用稠密低纬的向量表示，同时使得词具有了语义信息[13]。GCNN-SAM 模型使用 skip-gram 方法训练词向量。skip-gram 模型使用一段文本中的上下文词作为目标词，通过文本中上下文词来预测中间词。该模型没有隐藏层，由简单的神经网络构成。假设当前词为 $x_{(t)}$，则 skip-gram 的输出为 $x_{(t-2)}$、$x_{(t-1)}$、$x_{(t+1)}$、$x_{(t+2)}$，如图 2 所示。

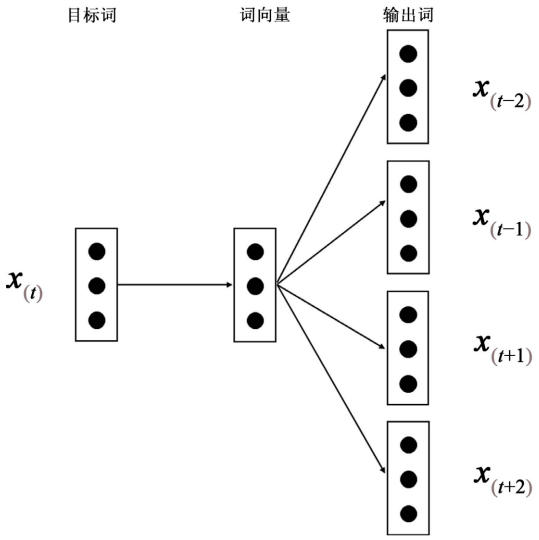

Figure 2. Skip-gram model structure
图 2. Skip-gram 模型结构

2.2. 特征提取层

特征提取层由 GCNN + PC 和 SAM + PC 两部分构成。

2.2.1. GCNN + PC

GCNN (Graph Convolutional Neural Network)是一种用于图像、视频、文本等数据的深度学习模型。在文本分类任务中，GCNN 主要是用来进行文本表示的。本文的 GCNN 由卷积层，门控层，池化层，全连接层组成，卷积层用于提取文本局部的特征，由于它的权值共享，因此可以降低学习的复杂度，门控层控制网络中信息的流动，同时可以提高该层非线性提取特征的能力，池化层用的是 1-max 池化，提取最重要的一个特征，降低了特征的维度，在一定程度上缓解了模型过拟合，全连接层用来连接所有的特征，对特征进一步提取，GCNN 模型如图 3 所示。

文本经过 skip-gram 嵌入后的特征矩阵表示为：$E_{1:n}=\left[x_1,x_2,\cdots,x_n\right]^{\mathrm{T}}$，$x_i\in R^m$，$E_{1:n}\in R^{n\times m}$。假设一个卷积核 $W\in R^{h\times m}$，该卷积核可以应用在窗口大小为 h 个词上，那么通过卷积核产生一个特征可以表示为：

$$s_i=f\left(W\times E_{i:i+h-1}+b\right),\ s_i\in R \tag{1}$$

f 是非线性的激活函数，b 是偏置项，$b\in R$，$E_{i:i+h-1}$ 为输入的词向量矩阵第 i 行到第 $i+h-1$ 行。卷积核通过在词嵌入矩阵维度方向上滑动来提取特征，词嵌入矩阵上窗口为 $\left\{E_{1:h},E_{2:h+1},\cdots,E_{n-h+1:n}\right\}$，卷积后得到的特征：

$$S=\left[s_1,s_2,s_3,\cdots,s_{n-h+1}\right],\ S\in R^{(n-h+1)\times 1} \tag{2}$$

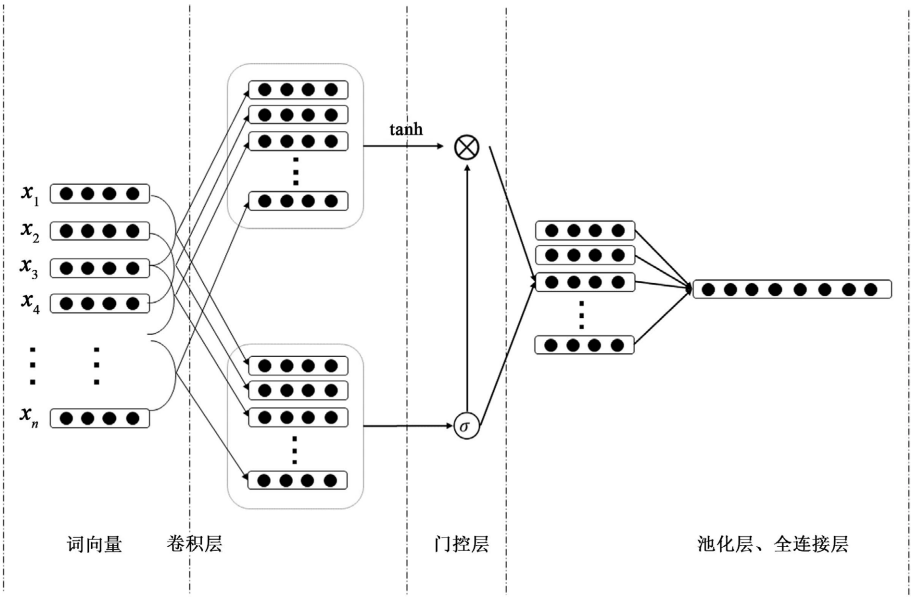

Figure 3. GCNN model structure
图 3. GCNN 模型结构

本文使用了大小相等的两个卷积核 $W \in R^{h\times m}$，词向量矩阵 $E_{1:n}$ 经过这两个卷积核后可以得到两个特征映射图 S_1, S_2，其中 $S_1 = \left[a_1, a_2, a_3, \cdots, a_{n-h+1}\right]$，$S_2 = \left[b_1, b_2, b_3, \cdots, b_{n-h+1}\right]$，然后把卷积后的结果做一个门控计算：

$$V = \tanh\left(S_1\right) \times \sigma\left(S_2\right),\ V \in R^{(n-h+1)\times 1} \tag{3}$$

其中 σ 为 sigmoid 激活函数，将(3)式得到的结果再经过 1-max 池化得到：

$$v = \max\left\{V\right\},\ v \in R \tag{4}$$

PC 收集来自 GCNN 提取到的局部特征，卷积核大小为 1，在每个词语上进行卷积操作，若输入序列长为 n，PC 可以通过如下公式定义：

$$PC\left(n\right) = \sigma\left(n * W^1 + b^1\right) \tag{5}$$

其中*代表卷积操作，W^1 是将要学习的参数矩阵，b^1 为卷积核的偏置项，σ 为 relu 激活函数，若经过 GCNN 后词向量维度为 d_{hid}，那么 $W^1 \in R^{d_{hid}\times d_{hid}}$，$b^1 \in h^{d_{hid}}$，经过 PC 后得到的特征表示为：

$$h^g = \left[h_1^g, h_2^g, \cdots, h_n^g\right] \tag{6}$$

2.2.2. SAM + PC

自注意力机制是在 Encoder-Decoder 框架下，让模型能够根据当前位置的输入关注到输入序列中与之相关的信息，从而提高模型的表现能力。编码器(encoder)将一个序列 $\left(x_1, x_2, x_3, \cdots, x_n\right)$ 映射到另外一个等长的序列 $\left(y_1, y_2, y_3, \cdots, y_n\right)$，其中 $x_i \in R^m$，$y_i \in R^m$。注意力机制的计算分为三个阶段：

阶段一，对于 Encoder 的每个输入 $E_{1:n} \in R^{n\times m}$，随机初始化三个服从均匀分布的矩阵 X，Y，Z 其中 $X \in R^{m\times m}$，$Y \in R^{m\times m}$，$Z \in R^{m\times m}$，通过对输入进行三个矩阵乘法分别计算 Q, K, V：

$$Q = E_{1:n}X \tag{7}$$

$$K = E_{1:n}Y \tag{8}$$

$$V = E_{1:n} Z \tag{9}$$

其中 $Q, K, V \in R^{n \times m}$，将 K 转置为 $K^T \in R^{m \times n}$。

阶段二，对于 Encoder 中的每个位置 $E_{1:n}$，将 Q 与 K^T 进行相乘得到相似度得分。为了防止结果过大，因此除以 $\sqrt{m}$ 得到放缩后的 S，为了让注意力机制能够关注到与当前位置相关的信息，我们需要对这些相似度得分进行归一化。通常使用 softmax 函数进行归一化，使其值映射在(0, 1)区间上，其中，Z_i 表示第 i 个值被分为的权重值。

阶段三，将得到的注意力权重与对应位置的矩阵 V 进行加权求和，得到 Encoder 对应位置的输出：

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{m}}\right)V \tag{10}$$

其中 $Attention(Q, K, V) \in R^{n \times m}$，词嵌入矩阵经过 Self-Attention 后形状没有改变。在 Decoder 中，我们可以采用类似的方法，使用自注意力机制关注到 Encoder 中的不同位置，从而得到与当前 Decoder 位置相关的 Encoder 输出信息，self-attention 模型结构如图 4 所示。

PC 收集 SAM 模型提取到的来自文本的全局信息，卷积核大小为 1。PC 可通过公式(5)计算，经过 PC 后得到的特征表示为：

$$h^m = \left[h_1^m, h_2^m, \cdots, h_n^m\right] \tag{11}$$

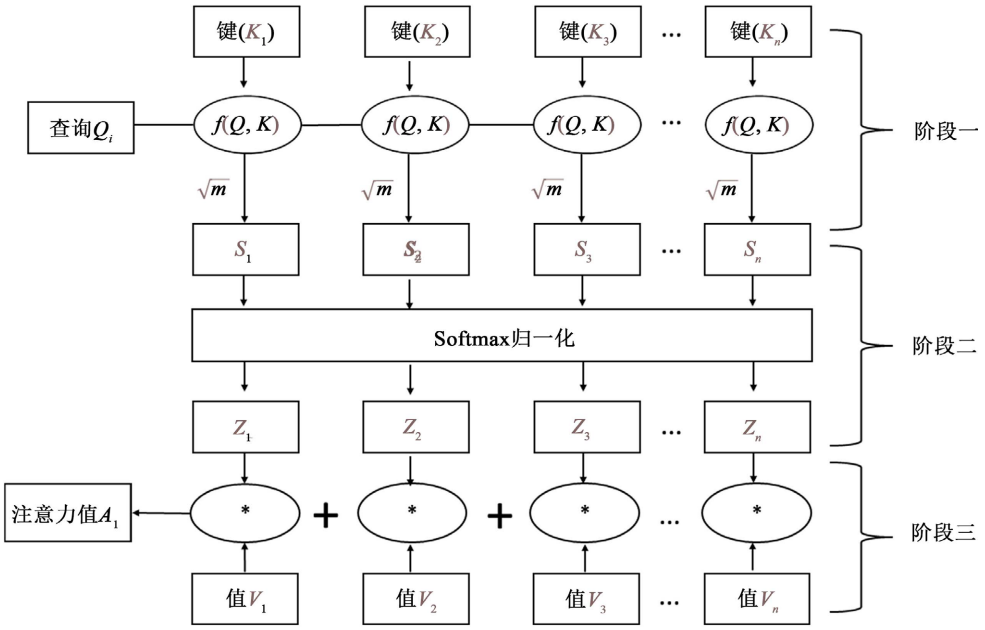

Figure 4. Self-Attention structure
图 4. Self-Attention 结构

2.3. 交叉熵损失函数

在分类任务中，交叉熵损失函数(Cross-entropy loss, CEL)是一种常用的损失函数，它能够有效地惩罚模型对错误类别的高概率预测，并促使模型学习产生正确类别的高概率预测[14]。其计算公式为：

$$L_{CEL} = -\sum_{i=1}^{N}\left(y_i \log\left(\hat{y}_i\right)\right) \tag{12}$$

其中，N 为样本数，y_i 是第 i 个样本的标签，$\hat{y}_i$ 是模型输出的第 i 个样本的预测概率。在多分类问题中，$\hat{y}_i$ 通常是 softmax 函数的输出。交叉熵损失函数越小，表示模型的预测结果与真实标签越接近。

2.4. 输出层

文本分类的输出层采用全连接结构及 Softmax 分类器进行分类，计算出单个文本属于图书类型的概率矩阵，计算公式如下：

$$O = soft\max\left(W_0 P + b\right) \tag{13}$$

其中，P 为全连接层的输入，O 为模型的输出结果，即概率矩阵，W_0 为权值矩阵，b 为偏置向量。

3. 实验

3.1. 实验数据集

在复旦大学中文文本分类数据集上进行实验，该数据集包含 20 个不同主题的新闻文本，包括体育、计算机、财经、时政、教育等领域，每个主题下有约 5000 篇文本，总共有 100,000 篇文本。每篇文本都被打上了一个预先定义好的类别标签。

随机地选取了其中 12 个类别循序打乱进行实验，12 个类别分别为体育，财经，法律，农业，教育，计算机，航天，时政，历史，地理，文学，艺术。然后根据需要的分级数量，并依据学科分类体系标准[15]，将 Fudan 语料库的标签进行重新划分，将原本的 12 个类别划分为 5 个标签，它们分别是自然科学，工程技术，医药卫生，人文社科和环境科学。

为了减少过拟合，最大化利用数据，数据集随机按 8:2 的比例被分为训练集和测试集，其中，训练集有 6019 篇文本，测试集有 1505 篇文本，为了方便地将数据批量化处理，从而加速训练过程，将数据变为固定长度 600，长了截断，短了补 0，将标签转换为 one-hot 编码表示。

将数据集进行了预处理，通过结巴分词将文本分词，去标点符号，去低频词，使用 skip-gram 模型来对单词进行嵌入，得到低纬稠密的单词向量。对于新词而言，结巴分词采用了隐马尔可夫模型(Hidden Markov Model, HMM)，通过对语料的大规模训练，得到模型的发射概率、起始概率和转移概率，进而通过维特比算法得到概率最大的隐藏序列，即 BEMS 标注序列，使用 BEMS 标注序列可以对语句进行分词并识别出其中的新词[16]。

3.2. 模型训练

GCNN 通道中的 CNN 选择三种滤波器，滤波器窗口大小分别为 2，3，4，采用了 2 个通道分别提取不同层次的特征，为了防止模型过拟合，在卷积或者经过多头注意力后加上 Dropout 函数，SAM 模型输出维度为 300，使用 Adam 优化器加快模型收敛速度，自动调整学习率。激活函数使用了 ReLU 函数，并在训练中采用了交叉熵损失函数。

当模型网络层数参数量较多时，容易造成过拟合，同时为了加快收敛速度，因此本实验探究了部分超参数对模型准确率的影响，结果如图 5 所示，固定其他超参数，其中图 5(a)和图 5(b)分别测试在 Fudan 数据集上不同学习率和 Dropout 值的准确率变化趋势。

从图 5(a)可以看出学习率较大时模型在 Fudan 数据集上的影响较大，不易收敛，当学习率为 0.1 的时候，GCNN-SAM 模型在测试集上的准确率最低，模型性能最差，学习率为 0.0001 时，准确率最高，此时模型泛化性能最好，因此本实验使用 Adam 优化器对应的学习率选择 0.0001。从图 5(b)可以看出 Dropout 值低于 0.6 时在测试集上的准确率变化不大，大于 0.6 时准确率发生了较大的变化，Dropout 值为 0.5 时在测试集上的准确率最高，因此本文 Droupout 值选择 0.5。

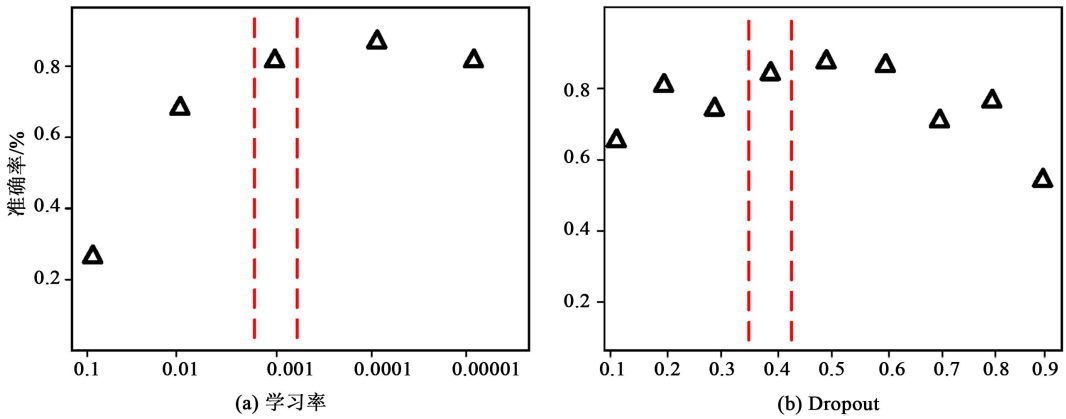

Figure 5. The influence of learning rate and Dropout value on the experimental results
图 5. 学习率和 Dropout 值对实验结果的影响

3.3. 模型评估指标

本文的评估指标采用大多数文本分类任务常用的三个指标，分别是查准率(Accuracy, Acc)，召回率(Recall, Rec)和 F_1 值(F_1-score)。通过这三个指标可以综合评估 GCNN-SAM 分类模型的性能，各个指标计算公式如式子(14)~(16)所示。

$$Acc = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{14}$$

$$\mathrm{Re}c = \frac{TP_i}{TP_i + FN_i} \tag{15}$$

$$F_1 = \frac{2 * Acc_i * \mathrm{Re}c_i}{Acc_i + \mathrm{Re}c_i} \tag{16}$$

对于图书标签类别 i，TP_i (True Positive)为真正例，指模型正确地预测出正例的样本数，FP_i (False Positive)为假正例，指模型错误地将负例预测为正例的样本数，TN_i (True Negative)为真负例，指模型正确地预测出负例的样本数，FN_i (False Negative)为假负例，指模型错误地将正例预测为负例的样本数。

3.4. 对比模型与结果

1) GCNN [17] + PC：GCNN 是在 CNN 的基础上加入一种简化的门控机制，提高了 CNN 非线性提取特征的能力，每两个 CNN 做一个门控机制来对特征进行提取，然后通过 PC 和 softmax 输出文本类别概率。

2) TextCNN [9]：使用三通道的卷积层，卷积核窗口大小分别为 2，3，4，词嵌入后的文本矩阵输入到 TextCNN，然后用 1-max 池化得到最重要的一个特征，再将特征进行融合，最后通过全局最大池化，全连接层，softmax 函数进行分类。

3) CNN-RNN [8]：首先将词嵌入得到词嵌入矩阵，然后分别用 CNN 和 RNN 提取局部特征和全局特征。

4) SAM + PC：将词嵌入矩阵输入到多头注意力模型中，通过全连接层，PC，softmax 函数得到文本的概率分布。

5) CNN：单通道的 CNN，卷积核窗口大小设置为 3，词嵌入后的文本矩阵输入到 CNN，提取短语级

别的特征，然后通过全连接层输出文本的类别。

6) SCNN：该模型有两个通道，将词嵌入后的向量分别通过 CNN 通道和 SAM 通道进行特征提取，将提取后的特征最后经过输出层进行输出。在模型训练过程中，使用交叉熵损失函数和 Adam 优化算法进行参数优化[18]。

Table 1. Comparison of experimental results of different neural network models
表 1. 不同神经网络模型的实验结果对比

模型	查准率	召回率	F_1 值
GCNN + PC	92.27	91.50	91.69
TextCNN	92.73	92.65	92.69
SAM + PC	89.82	89.79	89.98
CNN	90.68	90.85	90.59
CNN-RNN	89.96	92.03	91.50
S_CNN	90.86	92.26	90.51
SCNN	91.77	92.48	91.75
GCNN-SAM	**96.21**	**96.21**	**96.23**

7) S_CNN：单通道的 CNN，首先将词向量输入到 SAM 提取句子的内部特征，再通过 CNN 提取局部特征，最后经过全连接层对文本进行分类。与传统的卷积神经网络不同，S_CNN 还使用了一种新的池化方法，称为 k-max pooling，可以选择输出前 k 个最大值[19]。

GCNN-SAM 模型和以上七个基准模型在 Fudan 数据集上的实验结果如表 1 所示。从表 1 可以看出本文提出的 GCNN-SAM 模型在文本分类中较其他七种模型有较好的分类效果，在 Fudan 数据集上，测试集准确率为 96.21%，比其他七种模型的准确率都高，表明了本文所提出的 GCNN-SAM 模型的优越性。

在 Fudan 数据集上双通道的 SCNN 比单通道的 S_CNN 准确率要高一点，准确率高出了 0.91%，因为多通道可以提取更加丰富的文本特征，帮助提高文本分类的效果；除了本文所提出的 GCNN-SAM 模型，TextCNN 模型比其他模型的准确率都高，是由于 TextCNN 用了三种窗口大小卷积核，提取到不同大小的 n-gram 局部特征，增强了 CNN 对局部特征的提取能力，然后通过 1-max 最大池化提取出最重要的特征，将不同粒度的特征融合在一起作为文本的表示。对比 GCNN + PC 模型和 CNN 模型，可以看出 GCNN + PC 模型比 CNN 模型的准确率高，这是因为 GCNN 模型在 CNN 模型的基础上加入了门控机制，提高了 CNN 模型非线性提取特征的能力并且 PC 收集了 GCNN 所得到的特征。对 S_CNN 模型和 CNN-RNN 模型分类结果分析，发现 S_CNN 模型的准确率比 CNN-RNN 模型的准确率高，这是因为 RNN 存在长距离依赖的问题，当文本过长时不能捕捉到有效的文本信息，并且可能发生梯度消失和梯度爆炸，模型参数无法正常更新，而对于自注意力，它可以关注到对文本分类有较大影响的部分，并且解决了 RNN 所遇到的问题，可以有效的处理长文本信息，验证了其在文本分类中的有效性。对比 SCNN 模型，SAM + PC 和 CNN 模型，可以发现 SCNN 模型的性能优于 SAM 模型和 CNN 模型，因为 SCNN 模型既利用 SAM 提取到句子级别的特征，又利用 CNN 提取到短语级别的局部特征，然后将这两种特征进行融合，有利于模型提取到更加丰富的特征，弥补了单个模型的不足，充分发挥各个模型的优点。通过观查 CNN 模型和 SAM 模型，可以发现 CNN 模型的准确率比 SAM + PC 高，这是由于 Fudan 数据集都是较短的文本，因此 CNN 可以利用较多数量的卷积核来提取文本中更多的局部特征。

总体上，GCNN-SAM 模型在准确率，召回率，精确率，F_1 值上较其他模型都有较大的提升，本文所提出的 GCNN-SAM 模型有效利用了 SAM，GCNN，PC 分类模型，充分发挥了各个模型的优势，进一步提升了模型的整体性能。

3.5. 消融实验与结果

为验证 GCNN-SAM 模型的有效性，进行消融实验。将 GCNN-SAM 模型分解，设置 CNN、GCNN 和 CNN-SAM，实验结果如表 2 所示。

Table 2. Fudan data set ablation experimental results
表 2. Fudan 数据集消融实验结果

模型	查准率	召回率	F_1 值
CNN	90.31	90.31	90.37
GCNN	93.02	93.02	93.02
CNN-SAM	92.55	92.55	92.58
GCNN-SAM	**96.21**	**96.21**	**96.23**

从表 2 可以看出 GCNN 与 CNN-SAM 分类效果都要明显优于 CNN，这是由于 GCNN 能够有效地利用文本中的语义信息和结构信息，还可以捕捉节点之间的依赖关系，从而提高了分类的准确性。而 CNN-SAM 相较于 CNN 多引入了自注意力机制，模型在卷积层中进行了子采样，使得它能够学习到更加抽象和高层次的特征，从而提高了模型的性能。GCNN 与 GCNN-SAM 的分类效果接近，尽管 GCNN-SAM 在 GCNN 的基础上引入了注意力机制，用于加强模型对于重要特征的关注，但其效果并不总是明显的，尤其是在任务较简单的情况下，注意力机制会引入额外的噪声反倒干扰模型的学习。因此，它们在文本分类任务中的表现相似并不奇怪。GCNN-SAM 的分类效果最佳，GCNN-SAM 相较于传统的 CNN 和 GCNN 具有更好的特征提取能力，忽略无关的特征，并且采用了双向卷积，使得模型能够考虑到整个文本的上下文信息，更加全面地学习到文本的语义信息，能够更好地拟合训练数据，进而提高了分类的准确性。

4. 结论与展望

针对大数据时代下，图书自动标签分类在图书馆情报领域的重要性和必要性，本文设计了一种基于 CNN 和 PC 的自注意力机制文本分类模型 GCNN-SAM，首先通过词嵌入技术将词语嵌入成低纬稠密的向量，使得词具有了语义信息，其次利用 GCNN 通道提取 n-gram 粒度的局部特征，再利用 SAM 通道提取上下文全局的特征，使用两个通道可以提取出更加丰富的特征和上下文信息，接着在卷积层中对特征图进行 PC 逐点卷积操作，从而增强模型的非线性能力，最后通过将这三个部分结合起来，GCNN-SAM 模型有效地提取了文本重要特征，并将其用于文本标签分类任务中，达到较高的分类精度。

基金项目

本工作得到国家自然基金青年基金资助项目(71901144)、中国青少年研究会研究课题资助项目(2023B18)、尚理晨曦社科专项项目(22SLCX-ZD-005)的资助。

参考文献

[1] 李蕾, 王冕, 章成志. 区分标签类型的社会化标签质量测评研究[J]. 图书情报工作, 2013, 57(23): 11.

[2] 潘薇, 喻浩. 文献信息知识组织与内容揭示方法探究[J]. 图书馆研究, 2019, 39(3): 46-48.

[3] Cui, J.W, and Li, C.J. (2017) Identifying Semantic Relations of Clusters Based on Linked Data. *Data Analysis and Knowledge Discovery*, **1**, 57-66.

[4] Zeng, Z.Y., Wang, H., Zhang, Z.B., *et al*. (2021) A Study on Construction and Application of Text Classification Model Integrated with Associated Information from GCN. *Data Analysis and Knowledge Discovery*, **3**, 12-20.

[5] 朱靖波, 王会珍, 张希娟. 面向文本分类的混淆类判别技术[J]. 软件学报, 2008, 19(3): 630-639.

[6] 何嘉欣, 张涛, 陈旭岚, 等. 基于数据挖掘的 P2P 网贷个人信用评价模型研究[J]. 建模与仿真, 2021, 10(4): 991-1002.

[7] Yang, L., Hang, X.S. and Wang, J.Y. (2022) Identifying Subtypes of Clinical Trial Diseases with BERT-TextCNN. *Data Analysis and Knowledge Discovery*, **6**, 69-81.

[8] Wang, S., Liu, C., Wu, J. and Yao, X. (2019) Multi-Label Text Classification with Deep Learning. *IEEE Access*, **7**, 174906-174917.

[9] Gao, S., Ramanathan, A. and Tourassi, G. (2018) Hierarchical Convolutional Attention Networks for Text Classification. In: *Proceedings of the Third Workshop on Representation Learning for NLP*, Association for Computational Linguistics, Melbourne, 11-23. https://doi.org/10.18653/v1/W18-3002

[10] Ma, Y.F., Peng, Y.X. and Cambria, E. (2018) GCN-Based Joint Learning for Community Sentiment Analysis. *Proceedings of the* 27*th International Conference on Computational Linguistics*: *Technical Papers*, New Mexico, 20-26 August 2018, 1288-1297.

[11] Alshubaily, I. (2021) TextCNN with Attention for Text Classifi-Cation. ArXiv: 2108.01921.

[12] Chen, W. and Fan, J.S. (2022) MechPerformer: A General Deep Learning Model for History-Dependent Response Prediction in Structural Engineering. *Journal of Building Structures*, **43**, 209-219.

[13] Mikolov, T., Sutskever, I., Chen, K., *et al*. (2013) Distributed Representations of Words and Phrases and Their Compositionality. arXiv: 1310.4546.

[14] 李琳, 段围, 周栋,袁景凌. 基于深度语义匹配的法律条文推荐方法[J]. 软件学报, 2022, 33(7): 2618-2632.

[15] 杨灿, 董海龙. 基于国家标准学科分类的统计学科体系研究[J]. 统计研究, 2010, 27(1): 50-57.

[16] Cen, Y.H., Han, Z. and Ji, P.P. (2008) Chinese Term Recognition Based on Hidden Markov Model. 2008 *IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, Wuhan, 19-20 December 2008, 219-224.

[17] Yan, C.M. and Wang, C. (2021) Development and Application of Convolutional Neural Network Model. *Journal of Frontiers of Computer Science & Technology*, **15**, 27.

[18] Zagoruyko, S. and Komodakis, N. (2015) Learning to Compare Image Patches via Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 4353-4361. https://doi.org/10.1109/CVPR.2015.7299064

[19] Hai, H., Guo, Z.L., Cai, T.T. and He, Z.C. (2022) A Text Classification Method Based on a Convolutional and Bidirectional Long Short-Term Memory Model. *Connection Science*, **34**, 2108-2124. https://doi.org/10.1080/09540091.2022.2098926