计算机科学
COMPUTER SCIENCE

基于双向注意力机制和门控图卷积网络的文本分类方法

郑诚，梅亮，赵伊研，张苏航

引用本文

郑诚，梅亮，赵伊研，张苏航.基于双向注意力机制和门控图卷积网络的文本分类方法[J]．计算机科学，2023，50(1)：221-228.

ZHENG Cheng，MEI Liang，ZHAO Yiyan，ZHANG Suhang. Text Classification Method Based on Bidirectional Attention and Gated Graph Convolutional Networks [J]. Computer Science，2023，50(1)：221-228.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100095

基于双向注意力机制和门控图卷积网络的文本分类方法

郑 诚[1,2] 梅 亮[1,2] 赵伊研[1] 张苏航[1]
1 安徽大学计算机科学与技术学院 合肥 230601
2 计算智能与信号处理教育部重点实验室 合肥 230601

摘 要 现有基于图卷积网络的文本分类模型通常只是通过邻接矩阵简单地融合不同阶的邻域信息来更新节点表示,导致节点的词义信息表达不够充分。此外,基于常规注意力机制的模型只是对单词向量进行正向加权表示,忽略了产生消极作用的单词对最终分类的影响。为了解决上述问题,文中提出了一种基于双向注意力机制和门控图卷积网络的模型。该模型首先利用门控图卷积网络有选择地融合图中节点的多阶邻域信息,保留了之前阶的信息,以此丰富节点的特征表示;其次通过双向注意力机制学习不同单词对分类结果的影响,在给予对分类起积极作用的单词正向权重的同时,对产生消极作用的单词给予负向权重以削弱其在向量表示中的影响,从而提升模型对文档中不同性质节点的甄别能力;最后通过最大池化和平均池化融合单词的向量表示,得到文档表示用于最终分类。在4个基准数据集上进行了实验,结果表明,该方法明显优于基线模型。

关键词:文本分类;图卷积网络;注意力机制;文本表示;深度学习;自然语言处理
中图法分类号 TP391

Text Classification Method Based on Bidirectional Attention and Gated Graph Convolutional Networks

ZHENG Cheng[1,2],MEI Liang[1,2],ZHAO Yiyan[1] and ZHANG Suhang[1]
1 School of Computer Science and Technology,Anhui University,Hefei 230601,China
2 Key Laboratory of Intelligent Computing and Signal Processing,Ministry of Education,Hefei 230601,China

Abstract Existing text classification models based on graph convolutional networks usually simply fuse the neighborhood information of different orders through the adjacency matrix to update the representation of node in graph,resulting in insufficientrepresentation of the word sense information of the nodes. In addition,the model based on conventional attention mechanism only provides a positive weighted representation of the word embedding,ignoring the impact of words that produce negative effects on the final classification. To overcome the above problems,a model based on bidirectional attention mechanism and gated graph convolutional networks is proposed in the paper. Firstly,the model uses the gated graph convolutional networks to selectively fuse the multi-order neighborhood information of nodes in the graph,retaining the information of previous orders,to enrich the feature representation of nodes in graph. Secondly,the model learns the influence of different words on text classification results by the bidirectional attention mechanism,giving positive weights to words with positive effects on the classification and negative weights to words with negative effects to weaken their influence in the vector representation,to improve the model's ability to distinguish nodes with different properties in the document. Finally,the maximum pooling and average pooling are used to fuse the word representation in text to get the document representation for the final classification,where the average pooling can make each word play a role in generating a graph-level representation of the document and the maximum pooling can make the important words play a greater role in document embedding. Extensive experiments on four benchmark datasets show that the proposed model significantly outperforms the baseline model.

Keywords Text classification,Graph convolutional networks,Attention mechanism,Text representation,Deep learning,Natural language processing

到稿日期:2021-11-08 返修日期:2022-04-07
基金项目:安徽省重点研究与开发计划(202004d07020009)
This work was supported by the Key Research and Development Projects in Anhui Province(202004d07020009).
通信作者:郑诚(csahu@126.com)

1 引言

文本分类作为自然语言处理的基本任务和核心技术，得到了广泛的应用，包括垃圾邮件检测、情感分析、新闻分类等。传统基于机器学习的文本分类方法包括朴素贝叶斯[1]、支持向量机[2]和 k 近邻[3]，然而，这些方法需要手工提取特征，耗费大量的人力，效率低下。

近年来，随着深度学习的不断发展，卷积神经网络(Convolutional Neural Networks，CNN)和递归神经网络(Recurrent Neural Networks，RNN)在文本分类中得到了广泛应用。然而，这些模型大都集中在捕捉单词的局部信息和连续信息上，缺乏非连续和长距离词的交互信息。基于图卷积网络(Graphical Convolutional Networks，GCN)[4]的文本分类模型能够处理具有丰富结构关系的任务，在一定程度上解决了节点之间非连续的问题。但是，以往基于 GCN 的模型大多只是通过邻接矩阵简单地融合不同阶的邻域信息来更新节点表示，不能很好地生成单词表示。而基于注意力(Attention)[5]的模型通常采用 sigmoid 函数来生成一个介于 0 到 1 之间分布的正向注意力分数，不能对产生消极作用的单词给予负向权重，忽略了产生消极作用的单词对文档表示的影响。

针对上述问题，本文提出了一种基于门控图卷积网络和双向注意力机制的文本分类方法。该方法首先将每个文档构建为一个文档图，使模型可以实现归纳学习。其次，对图卷积网络进行改进，使用门控机制有选择地融合图中节点的多阶邻域信息，利用之前阶的信息和更新后的信息进行迭代来达到丰富节点特征的效果，从而更好地生成单词的隐层表示。然后使用双向注意力对传统的注意力机制进行了一定的改进，不同于传统的注意力机制，它使用 tanh 作为激活函数，在赋予对分类起积极作用的单词正向权重的同时，对产生消极作用的单词给予负向权重，以削弱其在文档表示中的影响，从而使模型能够区分对最终分类起着不同作用的单词。再将赋予权重的特征与原始特征相加，使模型对产生积极影响的单词进行特征增强，对产生消极影响的单词特征进行特征削弱，从而得到差异化更大的特征表示，帮助模型更好地分类。最后，利用最大池化和平均池化将单词表示融合为文档表示，用于最终的分类。总的来说，本文的主要贡献如下：

(1)利用门控机制对图卷积网络进行改进，提出了门控图卷积网络。该网络有选择地融合图中节点的多阶邻域信息，保留了上一层的邻域信息，更好地生成了单词节点表示。

(2)对注意力进行了一定的改进，使模型在给予对分类起积极作用的单词正向权重的同时，对产生消极作用的单词给予负向权重，以削弱其在文档表示中的影响，从而提升模型对文档中不同性质节点的甄别能力。

(3)在 4 个文本基准数据集上进行了大量实验，结果表明，本文模型明显优于基线模型，验证了本文模型的有效性。

2 相关工作

现有的采用深度学习进行文本分类的方法已经取得显著的进展。Kim 等[6]提出了 TextCNN 模型，用于提取文档中局部和位置不变的特征。Zhang 等[7]利用 CNN 提取字符级的特征表示，取得了不错的效果。Graves 等[8]使用 Bi-LSTM，利用门控机制来捕获文档中长距离的语义信息。Chen 等[9]利用优化的多通道 CNN 提取局部特征，弥补 BiGRU 忽视局部特征的不足。Liu 等[10]利用 Multi-timescale LSTM 来捕捉不同时间尺度的上下文信息，以对长文档进行建模。由于 CNN 和 RNN 优先考虑局部信息和顺序信息，这些模型虽然能够很好地捕获局部单词序列中的语义和句法信息，但会忽略非连续和长距离词的交互信息。因此，注意力机制和图卷积网络被广泛应用于文本分类领域，以解决这些问题，并取得了一定的成果。

注意力机制最初是在计算机视觉领域[11]提出的，而在自然语言处理领域，注意力机制最先被使用在基于解码器[12]的机器翻译任务中，随后扩展到其他任务中。Zhou 等[13]将 Attention 与 Bi-LSTM 相结合，用于关系分类任务。Yang 等[14]提出了分层注意力模型，使用单词和句子两个层面的注意力结构，使得模型可以给予词语和句子不同的关注度。Ding 等[15]利用注意力计算知识图谱每个概念的权重值，减小无关噪声概念对短文本分类的影响。然而，这些基于注意力机制的模型，通常只是对节点向量进行简单的正向加权来得到文档表示，不能很好地削弱产生负向影响的单词对文档表示的影响。

图卷积网络最近受到了越来越多的关注，Peng 等[16]提出了一种基于图的 CNN 模型，首次将文档转换为图的形式，并将其作为图卷积网络的输入。Yao 等[17]提出了 TextGCN，该模型首先将文档和单词作为图中的节点，接着使用滑动窗口生成单词之间边的关系，构建了一个基于语料库级别的大型异构图，最后使用图卷积网络对文档节点进行分类。Huang 等[18]则为每个文档单独构建一个文档图，单词之间边的权重是随机初始化以及全局共享的，并在训练中不断得到更新；对图中的每个单词节点使用消息传播机制(Message Passing Mechanism，MPM)[19]，先聚合其邻居节点的信息，然后更新节点表示。Yuan 等[20]利用异构图和主题模型对短文本进行分类。基于 GCN 的文本分类模型虽然能够在一定程度上解决文档中非连续和长距离词的交互信息问题，但是这些模型通常只是通过邻接矩阵简单地融合不同阶的邻域信息来更新节点表示，没有充分利用节点的多阶邻域信息，不能很好地更新单词节点表示。

3 基于双向注意力和门控图卷积网络的模型

本文模型由 4 个关键部分组成，分别为图构建层、门控图卷积网络层、双向注意力池化层和分类层。基于双向注意力和门控图卷积网络的模型的整体框架如图 1 所示。本节先给出模型的整体算法，然后对这 4 个部分进行了详细的介绍。

本文模型的具体描述如算法 1 所示。

算法 1　基于双向注意力和门控图卷积网络的模型

输入：文档 Text=($word_1$，$word_2$，…，$word_l$)

输出：文档表示h_g

1. 以标准方式对文本进行预处理，包括标记化和停用词去除
2. 基于滑动窗口将文本构建为图，得到特征矩阵 $\mathbf{X}$ 和邻接矩阵$\widetilde{\mathbf{A}}$
3. for nodeiin graph：
4. 　V_i←Glove embedding/ * 将单词表示为向量形式 * /
5. 　for layer t in {2，3，…，T}：
6. 　$\widetilde{\mathbf{h}}^t$←tanh($\widetilde{\mathbf{A}}\mathbf{h}^{t-1}\mathbf{W}_b$)/ * 将节点的一阶邻域信息传递到自身节点 * /
7. 　$\boldsymbol{\alpha}$← $\sigma(\mathbf{W}_c\mathbf{h}^{t-1})$/ * 得到具有门控功能的选择矩阵 $\boldsymbol{\alpha}$，控制节点邻域信息的聚合 * /
8. 　$\mathbf{h}^t$←$\mathbf{h}^{t-1}\odot\boldsymbol{\alpha}+\widetilde{\mathbf{h}}^t\odot(1-\boldsymbol{\alpha})$/ * 有选择地融合不同阶邻域信息 * /
9. 　$\mathbf{h}^{t'}$←tanh($\mathbf{W}_n\mathbf{h}^t+\mathbf{b}_s$)/ * 获取深层节点表示 * /
10. 　score←tanh($\mathbf{W}_s\mathbf{h}^t+\mathbf{b}_s$)/ * 获取双向注意力分数 * /
11. 　$\bar{\mathbf{h}}$←score$\odot\mathbf{h}^{t'}$/ * 将节点表示赋予权重 * /
12. 　$\mathbf{h}_n$←$\bar{\mathbf{h}}+\mathbf{h}^{t'}$/ * 与原始特征相加，来对节点特征进行增强 * /
13. $\mathbf{h}_g$←MaxP($\mathbf{h}_1\cdots\mathbf{h}_n$)+MeanP($\mathbf{h}_1\cdots\mathbf{h}_n$)/ * 利用最大池化与平均池化获取文档表示，用于分类 * /

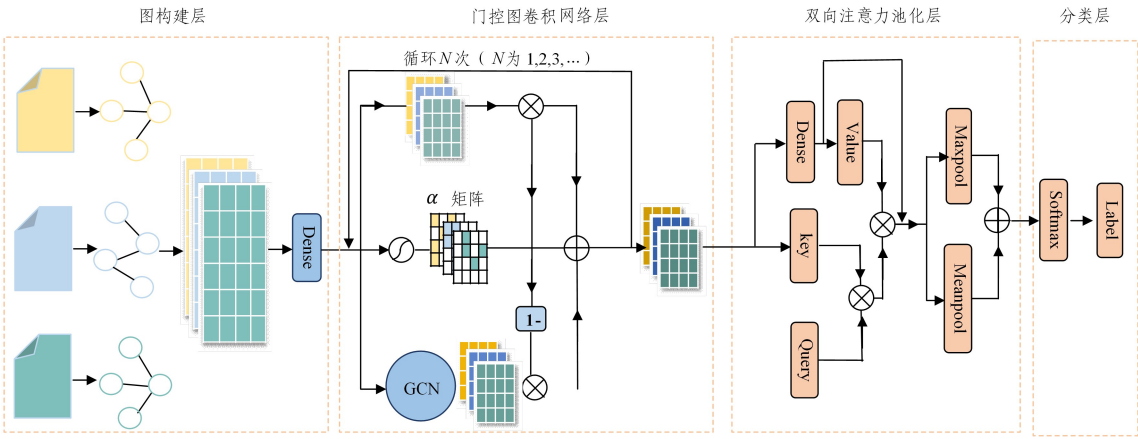

图 1　基于双向注意力和门控图卷积网络模型的整体结构

Fig. 1　Overall architecture of model based on bidirectional attention and gated graph convolutional networks

3.1　图构建层

为了将文档转换为图，本文首先将文档中的单词映射为图中的节点，接着使用滑动窗口得到单词之间边的关系，具体过程如图 2 所示，图中滑动窗口大小为 5，蓝色节点为中心节点。

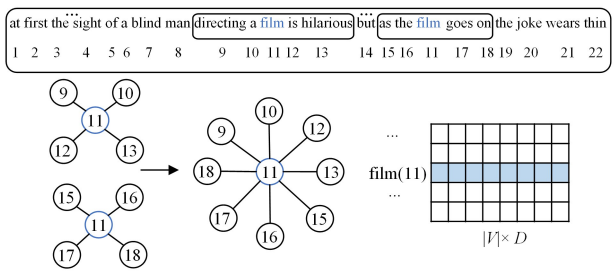

图 2　文档图（电子版为彩图）

Fig. 2　Text graph

滑动窗口内最中间的节点为中心节点（如图 2 中的“film”所示）。如果某个节点与中心节点共同出现在滑动窗口内，则在它们之间建立一条边（如“directing”与“film”之间存在一条边）。特别地，若一个单词在文档前后多次出现，则在图中其表示为同一个中心节点，并将邻居节点进行合并，这样可以使距离相对很远但与同一个单词共现的单词通过中心节点连接在一起。如图 2 中的“film”节点，它与“directing”“a”“is”“hilarious”这 4 个节点在图中第一个滑动窗口内共现，则“film”分别与“directing”“a”“is”“hilarious”存在一条边，而“film”又与“as”“the”“goes”“on”在另一个滑动窗口内共现，则“film”又会与“as”“the”“goes”“on”存在一条边。由于图 2 中节点是唯一存在的，通过“film”节点，上述的 8 个节点分别成为了各自的 2 阶邻居，即使其中某些节点之间的相对距离较远。

图 2 中单词节点使用 Glove[21] 进行初始化，将单词节点表示为固定维度的向量 $\boldsymbol{w}_i\in R^D$，D 为向量维度，则文档可表示为 $S=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_l)$。其中，$l$ 为文档中单词的个数。

在构建完文档图之后，为了得到节点的隐层表示并降低特征表示的维度，对于图中的每个单词节点，我们通过一个全连接神经网络和非线性激活函数，来对原始节点做一个非线性变换，如式(1)所示：

$$h^1=\tanh(\boldsymbol{W}_aS+b_a)\tag{1}$$

其中，$\boldsymbol{W}_a$ 为权重矩阵，b_a 为偏置项，tanh 为激活函数。

3.2　门控图卷积网络层

3.2.1　图卷积网络

GCN 是一种卷积神经网络，能够直接对图进行操作，它利用卷积运算使得节点的信息在图中进行传播，通过不断地聚合邻居节点信息，使得中心节点得到更新。

本文将图表示为 $G=(V,E)$，其中 $V(|V|=n)$ 是单词的集合，n 为图中节点的个数，E 是边的集合，并且规定每个节点都与其自身节点相连，即对于任意的节点 v，$(v,v)\in E$。令 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$，$\boldsymbol{X}\in R^{n\times k}$ 是包含 n 个节点的矩阵，其中 k 是特征向量的维度，$\boldsymbol{x}_n\in R^k$ 是第 n 个单词的特征向量。$\boldsymbol{A}\in R^{|V|\times|V|}$ 为 G 的邻接矩阵，$\boldsymbol{D}$ 为 G 的度矩阵，其中 $\boldsymbol{D}_{ii}=\sum_j\boldsymbol{A}_{ij}$。

由于自环的存在，$\boldsymbol{A}$ 的对角线元素都为 1。对于一层的 GCN，经过一层 GCN 之后得到新的节点表示为：

$$\boldsymbol{Z}=\boldsymbol{\sigma}(\widetilde{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}) \tag{2}$$

其中，$\widetilde{\boldsymbol{A}}=\widetilde{\boldsymbol{D}}^{-1/2}\boldsymbol{A}\widetilde{\boldsymbol{D}}^{-1/2}$ 为对称归一化的拉普拉斯矩阵，$\boldsymbol{W}$ 为权重矩阵，σ 为激活函数。

3.2.2 门控图卷积网络

现有基于图卷积网络的文本分类模型只是通过使用邻接矩阵简单地融合不同阶的邻域信息来更新节点表示，导致节点的词义信息表达不够充分。为了更好地融合节点的多阶邻域信息，生成了词义信息更加丰富的单词表示。受门控机制的启发，本文对 GCN 做出了一些改进，提出了有选择地融合邻域信息的门控图卷积网络。

如式(3)所示，我们首先将输入特征 h^{t-1} 进行一个线性变换，接着使用邻接矩阵 $\widetilde{\boldsymbol{A}}$ 将节点的一阶邻域信息传递到自身节点，通过结合周围节点和自身节点的信息来更新当前节点表示，再通过 tanh 激活函数进行非线性变换，得到 $\widetilde{\boldsymbol{h}}^{t}$。通过上述操作，图中的节点可以利用周围节点信息来更新自身节点信息。

$$\widetilde{\boldsymbol{h}}^{t}=\tanh(\widetilde{\boldsymbol{A}}\boldsymbol{h}^{t-1}\boldsymbol{W}_{b}) \tag{3}$$

其中，$\boldsymbol{W}_{b}$ 为权重矩阵，tanh 为激活函数。

受到 LSTM 门控机制的启发，我们将输入特征 h^{t-1} 通过投影矩阵投影到信息选择空间，并通过 sigmoid 函数将值压缩到 0～1 之间，得到具有门控功能的选择矩阵 $\boldsymbol{\alpha}$，如式(4)所示：

$$\boldsymbol{\alpha}=\sigma(\boldsymbol{W}_{c}\boldsymbol{h}^{t-1}) \tag{4}$$

其中，$\boldsymbol{W}_{c}$ 为投影矩阵，σ 为 sigmoid 函数。

如图 3 所示，$\boldsymbol{\alpha}$ 矩阵中的值为 $\boldsymbol{\alpha}_{i,j}$，取值介于 0 到 1 之间，表示第 i 个节点的第 j 个维度的选择系数。通过使用 $\boldsymbol{\alpha}_{i,j}$，可以更好地控制每一阶的邻域信息，让不同阶邻域信息进行有选择的融合。

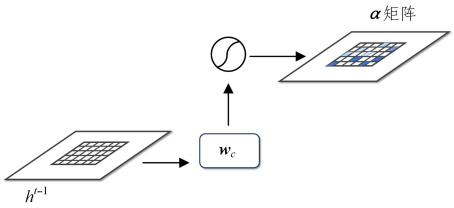

图 3 选择矩阵 $\boldsymbol{\alpha}$

Fig. 3 Select matrix $\boldsymbol{\alpha}$

如图 4 所示，我们使用选择矩阵 $\boldsymbol{\alpha}$，将它看作一个更新门，用于控制信息的传输，将之前阶的信息和当前阶的信息进行有选择的融合。当 α 值越接近 1，则代表上一阶的邻域信息保存得越多，越靠近 0，则表示融合了越多的当前阶信息。通过使用 $\boldsymbol{\alpha}$ 控制 $\boldsymbol{h}^{t-1}$，$1-\boldsymbol{\alpha}$ 控制 $\widetilde{\boldsymbol{h}}^{t}$，我们得到了更新后的节点表示 h^{t}，如式(5)、式(6)所示：

$$Gate(\widetilde{\boldsymbol{h}}^{t},\boldsymbol{h}^{t-1})=\boldsymbol{h}^{t-1}\odot\boldsymbol{\alpha}+\widetilde{\boldsymbol{h}}^{t}\odot(1-\boldsymbol{\alpha}) \tag{5}$$

$$\boldsymbol{h}^{t}=Gate(\widetilde{\boldsymbol{h}}^{t},\boldsymbol{h}^{t-1}) \tag{6}$$

其中，$\odot$ 为 Hadamard 乘积。

通过堆叠多层的门控图卷积网络，模型可以有选择地融合不同阶的邻域信息，使用不同阶邻域信息中的重要部分来更新中心节点的信息。而不同阶的邻域信息相当于文档中不同范围大小的局部消息，通过有选择地融合这些邻域信息，模型能够更好地生成中心节点的单词表示。由于这种融合不是简单的相加，而是利用具有门控机制的 $\boldsymbol{\alpha}$ 矩阵进行有选择的融合，因此可以保留不同阶邻域信息中重要的部分，去除噪声信息，从而达到丰富单词词义表示的效果。

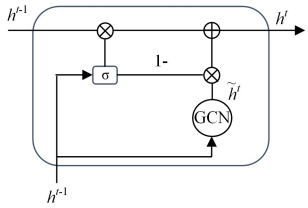

图 4 门控图卷积单元

Fig. 4 Gated graph convolutional unit

3.3 双向注意力池化层

对文档进行分类时，人们往往在某些单词上停留很久的时间，这些单词通常是对分类结果影响较大的单词，注意力机制则可以看作是在模拟人类的这一系列过程。由于句子中每个单词对生成文档表示的贡献程度是不一样的，在进行编码时，对分类越重要的单词模型应赋予越大的权重。因此，通常采用注意力机制给句子中的每个单词计算一个注意力得分，然后通过单词及其得分形成句子的向量表示。通过上述操作，模型可以只对文档中少部分的单词给予关注，而不是针对整篇文档，从而有效地过滤文档中的无效信息。通过对文档中单词给予不同的权重，来体现每个单词对文档分类的影响程度。

然而，传统的 Attention 在获取不同部分单词的权重时，使用 sigmoid 函数来计算注意力分数，得到的是介于 0 到 1 之间的注意力权重分布。它只考虑到每个单词对文档分类单向的积极影响，却忽略了单词可能会对分类产生反向的消极影响。如在电影评论中，“it is weird, wonderful, and not necessarily for kids”为正面评价，但是“not”对该文档分类结果产生了消极影响。

为了得到差异更大的特征表示，本文采用 tanh 函数来计算注意力分数。如式(7)所示，我们对经过多层门控图卷积网络更新之后的单词表示 h^{t} 做一个线性变换，得到语义信息更加丰富的单词表示 $h^{t'}$。式(8)中，使用 tanh 替换 sigmoid 函数作为激活函数得到单词的注意力分数，其值介于 -1 到 1 之间，即可获得双向的注意力权重。通过式(9)形成带权重的单词表示 $\bar{h}$：

$$\boldsymbol{h}^{t'}=\tanh(\boldsymbol{W}_{n}\boldsymbol{h}^{t}+\boldsymbol{b}_{n}) \tag{7}$$

$$score=\tanh(\boldsymbol{W}_{s}\boldsymbol{h}^{t}+\boldsymbol{b}_{s}) \tag{8}$$

$$\bar{\boldsymbol{h}}=score\odot\boldsymbol{h}^{t'} \tag{9}$$

其中，$\boldsymbol{W}_{n}$ 为权重矩阵，W_{s} 作为查询项(Query)，$\boldsymbol{h}^{t}$ 作为键值

(Key)，⊙代表 Hadamard 乘积。

$$\boldsymbol{h}_n=\bar{\boldsymbol{h}}+\boldsymbol{h}^{t'} \tag{10}$$

通过注意力分数的正负，模型可以更好地甄别出对生成文档表示产生不同影响的单词。如式(10)所示，将$h^{t'}$与$\bar{h}$相加。该操作对对文档分类有积极影响的单词进行了特征增强，对分类产生消极影响的单词进行了特征削弱，得到特征表示差异更大的h_n，从而有利于提高模型的分类效果。

另外，我们使用平均池化和最大池化融合单词表示，得到文档表示h_g。平均池化可以使每个单词在生成文档表示时都发挥作用，而最大池化可以使重要的单词发挥更大的作用。通过结合平均池化和最大池化，模型能够更好地生成文档表示，最终的预测基于文档表示。池化操作的定义如式(11)所示：

$$\boldsymbol{h}_g=MaxP(\boldsymbol{h}_1\cdots\boldsymbol{h}_n)+MeanP(\boldsymbol{h}_1\cdots\boldsymbol{h}_n) \tag{11}$$

其中，$MaxP$ 表示最大池化，$MeanP$ 表示平均池化。

3.4　分类层

最后，如式(12)所示，使用文档表示h_g来预测文档的标签。

$$\hat{y}=\mathrm{softmax}(\boldsymbol{W}_g\boldsymbol{h}_g+\boldsymbol{b}_g) \tag{12}$$

其中，$\boldsymbol{W}_g$是将文档表示映射到输出空间的权重矩阵，b_g为偏置项，将得到的结果传递到 softmax 进行归一化。通过加入 L2 正则化项来防止模型的过拟合。使用交叉熵(cross-entropy loss)作为损失函数，如式(13)所示：

$$L=-\sum_i \cdot\, y_i\log(\hat{y}_i)+\lambda\parallel\Theta\parallel_2 \tag{13}$$

其中，y_i为真实标签，Θ表示模型参数，λ 表示正则化因子。

4　实验

4.1　数据集

在 4 个广泛使用的基准数据集 Ohsumed，MR，R8，R52 上进行了实验，数据集的介绍如表 1 所列。

(1)Ohsumed：来自 MEDLINE 数据库，一个由医学图书馆维护的重要医学文献书目数据库。数据集中的每个文档都与 23 个疾病类别中的一个或多个相关。由于本文侧重于单标签文本分类，因此只留下属于一个类别的 7400 个数据。其中，训练集有 3357 个数据，测试集有 4043 个数据。

(2)MR：用于二元情感分类的电影评论[22]数据集，其中每个评论只包含一句话。数据集有 7108 条正面评价和 3554 条负面评价。本文使用 Tang 等[23]处理的训练集和测试集。

(3)R8 和 R52：是路透社 21578 数据集的两个子集。R8 有 8 个类别，分为 5485 个训练数据和 2189 个测试数据，R52 数据集总共有 52 个类别，分为 6532 个训练数据和 2568 个测试数据，并且每个数据只与一个类别相关。

表 1　数据集介绍

Table 1　Datasets overview

Dataset	Doc	Train	Test	Class
Ohsumed	7400	3357	4043	23
MR	10662	7108	3554	2
R8	7674	5485	2189	8
R52	9100	6532	2568	52

表 2　实验结果比较(平均值±标准差)

Table 2　Test accuracy of various models on datasets(mean± standard deviation)

(单位：%)

Model	Ohsumed	MR	R8	R52
CNN-non-static	58.44±1.06	77.75±0.72	95.71±0.52	87.59±0.48
Bi-LSTM	49.27±1.07	77.68±0.86	96.31±0.33	90.54±0.91
PTE	53.58±0.29	70.23±0.36	96.69±0.13	90.71±0.14
fastText	57.70±0.49	75.14±0.20	96.13±0.21	92.81±0.09
SWEM	63.12±0.55	76.65±0.63	95.32±0.26	92.94±0.24
TextGCN	68.36±0.56	76.74±0.20	97.07±0.10	93.56±0.18
TextLevelGNN	69.40±0.60	—	97.80±0.20	94.60±0.30
DHTG	68.80±0.33	77.21±0.11	97.33±0.06	93.93±0.10
S^2GC	68.50±0.10	76.70±0.10	97.40±0.10	94.50±0.20
T-VGAE	70.02±0.14	78.05±0.11	97.66±0.09	95.00±0.12
Our Model	**71.12±0.38**	**79.65±0.31**	**97.84±0.12**	**95.42±0.14**

注：加粗数据表示最优值，'—'表示实验结果数据缺失

4.2　对比实验

为了验证本文模型的有效性，挑选了当前表现较好的一些模型进行对比实验。对比模型的介绍如下。

(1)CNN-non-static：卷积神经网络。使用预训练词向量的 CNN 模型，利用卷积操作和最大池化获得文本表示。

(2)Bi-LSTM：双向 LSTM。将预训练好的词向量输入到双向 LSTM，将模型最后一个隐藏状态作为文本的表示。

(3)PTE：预测性文本嵌入。它首先将单词、文档和标签作为节点构成了一个大型异构网络，在这个网络中学习单词表示，然后对单词表示进行平均池化得到文档表示，最后送入分类器进行分类。

(4)fastText[24]：一种简单高效的文本分类方法。它将单词向量的平均值作为文档向量，然后将文档向量送入线性分类器进行分类。

(5)SWEM[25]：简单词嵌入模型。该模型采用简单的词向量池化策略。

(6)TextGCN：一种基于图的文本分类模型。该模型使用单词和文档作为节点，为整个语料库构建一个大图，然后使用 GCN 对文档节点进行分类。

(7)TextLevelGNN：一种基于图的文本分类模型。该模型使用单词作为节点，建立一个基于文档的图，图中边的权重随机初始化，在训练中得到更新。

(8)DHTG[26]：将概率深度主题模型集成到图的构建中，并提出了一种全新的可训练层次主题图(HTG)模型，包括单词级、层次主题级和文档级节点。

(9)S^2GC[27]：利用改进的马尔可夫扩散核导出了简单谱图卷积，捕获了每个节点的全局和局部上下文信息。

(10)T-VGAE[28]：提出了一种归纳式主题变分图自动编码器模型，该模型继承了主题模型的可解释性和图变分编码器高效的信息传播机制，从而捕获文档和单词之间隐藏的语义信息。

4.3　实验环境介绍

操作系统为 Linux，内存为 64 GB，CPU 为 AMD EPYC 7302，显卡为 24 GB 的 NVIDIA GeForce RTX 3090。使用 Pytorch 框架实现本文模型。

4.4　实验设置

对于所有的数据集，本文将训练集按照 9∶1 的比例随机

划分为两个部分，分别作为训练集和验证集。接着使用 Adam[29] 训练器，并将 Adam 的学习率设置为 0.001，将 dropout 率设置为 0.5。由于 Yao 等对所有使用预训练的模型都采用 Glove 进行初始化，并给出了相应的实验结果，为了方便进行比较，本文也使用 Glove 词向量进行初始化，维度为 300 维。对于 Glove 中没有的词，使用随机初始化。

4.5 实验及结果分析

在 4 个基准数据集上进行了大量实验，实验结果如表 2 所列，结果为 10 次运行结果的平均值。实验结果表明，本文模型明显优于其他基线模型（一些基线模型的结果见文献[17]）。本文模型表现较好的原因主要有以下 3 点：1）通过堆叠多层门控图卷积网络，能够捕获高阶邻域信息，将不同阶邻域信息进行有选择的融合，生成词义信息更加丰富的单词表示；2）使用双向注意力机制对产生消极作用的单词给予负向权重，削弱其在文档表示中的影响，以更好地生成用于分类的文档表示；3）本文构建的图属于文档级别，占用的时间与显存资源都更少。

4.6 GPU 显存占用与模型运行时间分析

表 3 列出了在不同数据集上本文模型与其他模型在 GPU 显存与时间方面的开销（同一环境配置下的显存与时间消耗）。从表 3 中的数据可知，在显存占用和运行时间上，本文模型具有明显的优势，分析具体原因如下：TextGCN 需要使用训练文档和测试文档构建一个基于语料库级别的图，因此不可避免地建立了大量的边，消耗了一定的显存与时间；TextLevelGNN 随机初始化单词之间边的权重，该权重属于模型参数，需要在训练中不断迭代更新，因此需要消耗一部分时间与显存；本文构建的图则属于文档级别，边的权重在文档构建为图时就已确定，因此不会消耗过多的时间与显存。

表 3 GPU 显存消耗与运行时间

Table 3 Consumption of GPU memory and run time

Dataset	Model	Memory/MB	Time/s
Ohsumed	TextLevelGNN	9728	496
	TextGCN	7192	368
	OurModel	2236	245
R8	TextLevelGNN	8456	582
	TextGCN	4638	158
	OurModel	1822	124
R52	TextLevelGNN	9171	630
	TextGCN	5185	190
	OurModel	1864	159

4.7 参数分析

图 5 给出了不同数据集上滑动窗口大小对测试集准确率的影响，可以看到，随着滑动窗口的不断增大，测试集准确率先是有一定的提高，当滑动窗口超过一定大小时，准确率就会慢慢下降。这表明，如果窗口太小，模型不能充分地捕获周围节点的邻域信息，而窗口太大则会在关系不紧密的节点之间添加边，使得更新节点表示时融入了噪声信息，从而影响分类效果。

图 6 给出了在 R52 数据集上门控图卷积网络层数对分类结果的影响，可以看出，随着层数的增加，分类准确率也有一定程度的增加，但是随着层数继续增加，结果有一定程度的下降，然后趋于平缓。通过分析，其原因可能是，随着层数的增加，模型在保留一部分原始单词信息的同时也捕获了较远距离单词的信息，通过将两者信息有选择地融合，生成了词义信息更加丰富的单词表示。随着网络层数进一步加深，各节点表示会趋于相同，但由于门控机制的存在，模型保留了一部分原始单词信息，使得模型在一定程度上缓解了图中节点过平滑的问题，因此准确率会有所下降，然后趋于平缓。

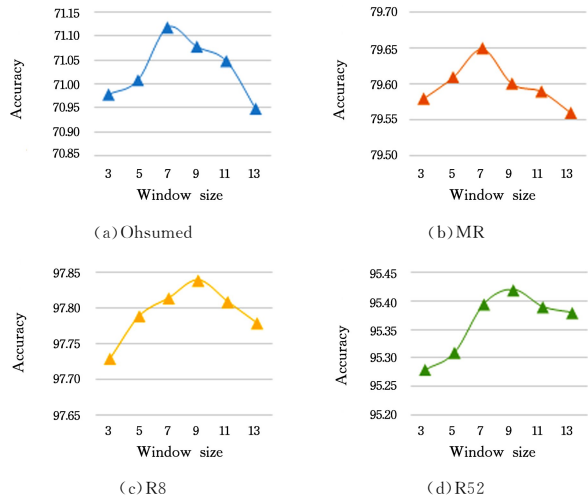

图 5 滑动窗口大小对实验结果的影响

Fig. 5 Influence of size of sliding window on experimental results

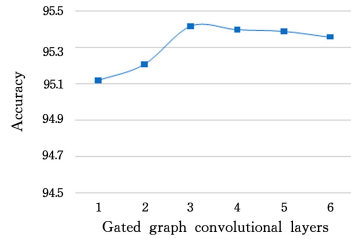

图 6 门控图卷积网络层数对实验结果的影响

Fig. 6 Influence of different layers of gated graph convolutional networks on experimental results

4.8 消融实验

为了进一步分析本文模型，本文进行了消融实验。表 4 列出了实验结果，其中 w/o $\boldsymbol{\alpha}$ 表示不使用选择矩阵 $\boldsymbol{\alpha}$，w/o Bi-Att 表示不使用双向注意力，而是使用普通注意力机制，Original 表示原始模型。从表 4 中的结果可以看出，替换部分模块的模型相比原来的模型在实验结果上都有着不同程度的下降，说明各个模块之间是相辅相成的，模型的每一个部分对生成最终的表示都起着非常重要的作用，也进一步验证了模型各个部分的合理性。

表 4 消融实验结果

Table 4 Results of ablation studies

Model	Ohsumed	MR	R8	R52
w/o $\boldsymbol{\alpha}$	70.31	79.26	97.24	95.03
w/o Bi-Att	70.82	79.09	97.13	95.19
Original	**71.12**	**79.65**	**97.84**	**95.42**

注：加粗数据表示最优值

为了验证门控图卷积网络融合多阶邻域信息的有效性，本文将图卷积网络生成的信息直接送入下一层，而不使用选择矩阵 $\boldsymbol{\alpha}$ 有选择地融合之前阶的信息和新生成的信息，其他

部分保持不变。从表 4 中可以看出，不使用选择矩阵的模型在所有数据集上的表现都最差。其主要原因是，虽然图中不同阶的邻域信息都对最终分类有用，但作用大小不同，若只是通过邻接矩阵进行简单的融合，则不能很好地捕获节点的邻域信息，应使用具有门控机制的选择矩阵 $\boldsymbol{\alpha}$ 进行有选择的融合。

另外，为了验证双向注意力机制对最终分类效果的影响，我们将注意力池化层中的双向注意力替换为普通的注意力，其他部分保持不变。从表 4 中可以看出，与使用双向注意力机制的模型相比，使用普通注意力机制的模型的表现更差。这表明，双向注意力能够识别出文本中对分类有消极影响的单词，从而削弱其产生的影响，得到差异更大的特征表示，帮助模型进行更好的分类。

4.9　双向注意力可视化分析

为了进一步说明双向注意力的作用，本文选取了 MR 数据集中的正面电影评论，对其注意力权重进行了可视化。如图 7 所示，其中第一列为普通注意力权重分布，第二列为双向注意力分布，黄色为正向权重，蓝色为负向权重。可以看到，使用普通注意力时，模型不能很好地识别出“not”这个对分类产生负向影响的单词，而使用双向注意力机制则会给“not”一个绝对值较大的负向权重，从而可以在一定程度上削弱其影响。

图 7　注意力可视化（电子版为彩图）

Fig. 7　Attention visualization

结束语　本文提出了一种基于双向注意力机制和门控图卷积网络的文本分类模型。该模型首先将文档构建为图，其次使用门控图卷积网络对多阶邻域信息进行有选择的融合，然后使用双向注意力机制为不同性质的单词分配不同的权重，以获得差异更大的特征表示用于最终的分类。在 4 个基准数据集上的大量实验表明，本文方法明显优于基线模型，但由于本文构建的图还未充分利用文本中单词的统计信息，未来的工作会考虑引入外部知识，如语料库的一些内在统计特征等，来提高模型的性能。

参 考 文 献

[1] WANG Q, GARRITY G M, TIEDJE J M, et al. Naive Bayesian-Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy[J]. Applied and Environ-mental Microbiology, 2007, 73(16): 5261-5267.

[2] FORMAN G. BNS Feature Scaling: An Improved Representation over TF-IDF for SVM Text Classification[C]// Proceedings of the 17th ACM Conference on Informationand Knowledge Management. New York: ACM, 2008: 263-270.

[3] TAN S. An Effective Refinement Strategy for KNN Text Classifier[J]. Expert Systems with Applications, 2006, 30(2): 290-298.

[4] KIPF T N, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[C]// International Conference on Learning Representation. On Line: ICLR, 2017: 101-112.

[5] LUONG M T, PHAM H, MANNING C D. Effective Approaches to Attention-Based Neural Machine Translation[J]. Proceeding of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, 28(2): 1412-1421.

[6] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Empirical Method in Natural Language Processing. Stroudsburg: ACL, 2014: 1746-1751.

[7] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networksfor text classification[C]// Conference and Workshop on Neural Information Processing Systems. Montreal: NIPS, 2015: 649-657.

[8] GRAVES A, JAITLY N, MOHAMED A. Hybrid Speech Recognition with Deep Bi-Directional LSTM[C]// 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. New York: IEEE, 2013: 273-278.

[9] CHEN K J, LIU H. Chinese Text Classification Method Based on Improved BiGRU-CNN[J]. Computer Engineering, 2022, 48(5): 59-66, 73.

[10] LIU P, QIU X, CHEN X, et al. Multi-Timescale Long Short-Term Memory Neural Network for Modelling Sentences and Documents[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 2326-2335.

[11] MNIH V, HEES N, GRAVS A. Recurrent Models of Visual Attention[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2204-2212.

[12] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014, 15(3): 152-161.

[13] ZHOU P, SHI W, TIAN J, et al. Attention-Based Bidirectional Long Short-term Memory Networks for Relation Classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016: 207-212.

[14] YANG Z, YANG D, DYER C, et al. Hierarchical Attention Networks for Document Classification[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics. Washington: NAACL, 2016: 1480-1489.

[15] DING C H, XIA H B, LIU Y. Short Text Classification Model Combining Knowledge Graph and Attention Mechanism[J]. Computer Engineering, 2021, 47(1): 94-100.

[16] PENG H, LI J, HE Y, et al. Large-Scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN[C]// Proceedings of the 2018 World Wide Web Conference. New York: ACM, 2018: 1063-1072.

[17] YAO L, MAO C, LUO Y. Graph Convolutional Network for Text Classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2019: 7370-7377.

[18] HUANG L, MA D, LI S, et al. Text Level Graph Neural Net-

work for Text Classification [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2019: 2216-2225.

[19] GILMER J, SCHOENHOLZS, RILEY P F, et al. Neural Message Passing for Quantum Chemistry[C]// International Conference on Machine Learning. New York: ACM, 2017: 1263-1272.

[20] YUAN Z Y, GAO S, CAO J, et al. Method for Few-Shot Short Text Classification Based on Heterogeneous Graph Convolutional Network[J]. Computer Engineering, 2021, 47(12): 87-94.

[21] PENNINGTON J, SOCHER R, MANNIG C D. Glove: Global Vector for Word Representation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1532-1543.

[22] CER D, YANG Y, KONG S Y, et al. Universal Sentence Encoder[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg: ACL, 2018: 1422-1433.

[23] TANG J, QU M, MEI Q. Pte: Predictive Text Embedding throughLarge-Scale Heterogene-ous Text Network[C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015: 1165-1174.

[24] JOULIN A, GRAVE E, BOJANOWKI P, et al. Bag of Tricks for Efficient Text Classification [C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 427-431.

[25] SHEN D, WANG G, WANG W, et al. Baseline Needs More Love: On Simple Word-Embedding based Models and Associated Pooling Mechanisms[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 440-450.

[26] WANG Z, WANG C, ZHANG H, et al. Learning Dynamic Hierarchical Topic Graph with Graph Convolutional Network for Document Classification[C]// International Conference on Artificial Intelligence and Statistics. BOSTON: JMLR, 2020: 3959-3969.

[27] ZHU H, KONIUSZ P. Simple Spectral Graph Convolution[C]// International Conference on Learning Representation. On Line: ICLR, 2021: 151-163.

[28] XIE Q, HUANG J, DU P, et al. Inductive Topic Variational Graph Auto-Encoder for Text Classification[C]// Proceeding of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. Washington: NAACL, 2021: 4218-4227.

[29] KINGMA D, BA J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014, 8(2): 4123-4131.

ZHENG Cheng, born in 1964, Ph.D, associate professor. His main research interests include data mining and text analysis, natural language processing.

（责任编辑：喻藜）