计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(9):201−209 [doi: 10.15888/j.cnki.csa.008703]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于门控图注意力网络的归纳式文本分类①

王晨曦, 张莹祺

(华南师范大学 计算机学院, 广州 510631)
通信作者: 王晨曦, E-mail: chenxi.wang@m.scnu.edu.cn

摘　要: 为了有效地整合文本中的复杂特征和提取不同的上下文信息, 提出了基于门控图注意力网络的归纳式文本分类方法 (TextIGAT). 该方法首先为语料库中的每个文档进行单独构图, 并将其中所有的单词作为图中的节点, 以此保留完整的文本序列. 文本图中设计单向连接的文档节点, 使词节点能与全局信息交互, 并合并不同的上下文关系连接词节点, 从而在单个文本图中引入更多的文本信息. 然后, 方法基于图注意力网络 (GAT) 和门控循环单元 (GRU) 来更新词节点的表示, 并根据图中保留的文本序列应用双向门控循环单元 (Bi-GRU) 来增强节点的顺序表示. TextIGAT 能灵活地整合来自文本本身的信息, 因此能对包含新词和关系的文本进行归纳式学习. 在 4 个基准数据集 (MR、Ohsumed、R8、R52) 上的大量实验和详细分析表明了所提出的方法在文本分类任务上的有效性.

关键词: 文本分类; 图神经网络; 上下文信息; 归纳式学习; 自然语言处理

引用格式: 王晨曦,张莹祺.基于门控图注意力网络的归纳式文本分类.计算机系统应用,2022,31(9):201–209. http://www.c-s-a.org.cn/1003-3254/8703.html

Inductive Text Classification Based on Gated Graph Attention Network

WANG Chen-Xi, ZHANG Ying-Qi

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: To effectively integrate complex features in text and extract different contextual information, this study proposes an inductive text classification method based on a gated graph attention network (TextIGAT). This method constructs a graph structure for each document in the corpus and takes all the words as nodes in the graph to preserve the complete text sequence. One-way connected document-level nodes are designed in the text graph, so that word nodes can interact with global information, and different contextual connection word nodes are merged to introduce more text information in a single text graph. Then, the representations of word nodes are updated utilizing a graph attention network (GAT) and a gated recurrent unit (GRU), and the sequential representation of nodes is enhanced by a bi-directional gated recurrent unit (Bi-GRU) according to the text sequence retained in the graph. TextIGAT can flexibly integrate information from text, which thus allows inductive learning on text with new words and relations. Extensive experiments on four benchmark datasets (MR, Ohsumed, R8, and R52) and detailed analysis prove the effectiveness of our proposed method on text classification.

Key words: text classification; graph neural network (GNN); contextual information; inductive learning; natural language processing (NLP)

文本分类是自然语言处理 (natural language processing, NLP) 中的经典问题之一, 其目的是为文本单位, 例如句子、查询、段落或文档等分配标签[1]. 这项任务具有广泛的应用场景, 包括问题解答, 垃圾邮件

① 基金项目: 国家自然科学基金 (61772210)
收稿时间: 2021-12-16; 修改时间: 2022-01-29; 采用时间: 2022-02-15; csa 在线出版时间: 2022-06-17

检测, 情感分析, 新闻分类等. 其中的文本数据可以来源于网络数据, 电子邮件, 用户评论以及客户服务的问题和解答等不同场景. 近年来, 图神经网络 (graph neural network, GNN)[2–5] 的新型神经网络引起了广泛的关注, 并在文本分类中表现出卓越的性能[6–9]. 与卷积神经网络 (convolutional neural network, CNN)[10,11] 和循环神经网络 (recurrent neural network, RNN)[12,13] 等序列学习模型不同, GNN 可直接处理复杂的图结构化数据, 同时能优先考虑图的全局结构信息. 当基于 GNN 来进行文本学习时, 文本的结构特征和单词之间的长距离交互能有效地捕获, 以提高最终的文本分类性能.

基于 GNN 的文本分类包括两个阶段: 1) 根据文本信息构建图. 2) 建立对文本图的学习模型. 尽管现有的基于图的方法已经取得了令人满意的结果, 但存在两个局限性限制了分类性能的提升. 其一, 现有方法存在语序和歧义问题, 大多数模型中将文本中的词节点定义为一个集合 (set). 如在 TextING[6] 中, 将文本中多次出现的同一个单词视为唯一的词节点. 虽然这样能使图中的词节点紧密地连接, 但忽略了在文本中不同位置出现同一个词具有不同的语序和语义信息. 其二, 现有研究缺乏对文档中节点之间不同上下文信息进行有效交互和提取. 具体来说, 如在 TextGCN[9] 中构建了一个包含语料库中所有文档和词之间全局关系的单一图, 这样的方式忽略了文本中细粒度的词交互. 并且, 在整个语料库级别的大图中设计文档节点连接到所有相关词节点, 会使文本的结构信息在词节点进行消息传递 (message passing) 的过程中被文档节点模糊, 这也限制了模型对新文档进行归纳式学习 (inductive learning).

为了解决上述限制, 本文提出了一种新的基于图的文本分类方法 TextIGAT, 以便准确地对文本自身特征进行提取整合, 来满足归纳式文本分类. 首先, 为了在文本图的构建中保留完整的文本序列, 将文档中所有重复出现的词保留为单独的词节点. 其次, 为每个输入的文本基于文中单词和上下文关系构建有向文本图, 此文本图通过词之间的共现和句法依赖关系构建, 且不计算任何边的权重来减少构图的操作. 然后, 模型使用自监督的图注意力网络 (self-supervised graph attention network, SuperGAT)[14] 和门控循环单元 (gated recurrent unit, GRU)[15] 作为基础, 对词节点与其邻接点进行消息传递, 并更新词节点的隐藏层状态. 模型中利用自注意力机制 (self-attention) 初始化整个文本的全局表示, 并设计单向连接的文档节点在更新过程中与局部词节点交互, 以此保留文本的结构信息并更好地融入全局信息. 由于本文提出的文本图保留了文本序列, 能通过 GRU 的双向模型 (Bi-GRU) 来增强词节点在整体文本中的顺序表示. 最后, 通过注意力机制对每个输入文本最终的词节点表示进行整合, 并训练和分类.

1 相关工作

1.1 基于深度学习的文本分类

现有基于深度学习的文本分类方法主要可以分为两类, 一类聚焦于对词嵌入的研究, 而另一类针对于深度神经网络模型的研究. 近些年来的研究表明, 深度学习在文本分类任务上的成功, 很大程度上取决于词嵌入的有效性[16]. 词嵌入是文本学习的数字表示模型, 它从大型未标记语料库中提取单词或短语的语义和句法信息, 并通过极大地降低向量的维数将词映射到实数的连续向量空间. 通过对词嵌入的研究, Joulin 等人[17] 提出一种简单有效的文本分类方法 fastText, 它将词或 n-gram 嵌入的平均值视为文档嵌入, 然后将文档嵌入使用线性分类器进行分类. Shen 等人[18] 使用池化 (pooling) 对词向量进行操作的模型 SWEM, 并在与深度神经网络的实验对比中取得更好的结果. 在对深度神经网络模型的研究中, CNN 和 RNN 这两种具有代表性的模型, 已经证明了在文本分类任务上的优越性. Kim[10] 提出了 TextCNN 模型来使用包含多个尺寸不一卷积核的卷积神经网络进行句子分类. Liu 等人[12] 基于长短期记忆网络 (long short-term memory, LSTM) 设计了 3 种不同的词表示共享机制, 并在文本分类中获得了较好的效果. Lai 等人[19] 提出了 CNN 和 RNN 的组合模型 TextRCNN, 结合两种网络的优势来对文本分类的效果进行提升. 为了进一步提高深度网络模型的表达能力, Yang 等人[20] 使用注意力机制 (attention mechanism) 为文本分类的模型的组成部分, 并取得了不错的分类结果. 这些深度学习模型在文本分类任务上实用且应用广泛, 但主要关注的是局部和顺序特征, 很少使用全局上下文信息, 也不能充分地捕捉词与词间的长期依赖关系.

1.2 基于图的文本分类

近年来, 由于在图结构数据的表示学习方面取得了巨大成功, 一些研究集中在基于图的方法来提高文本分类任务的性能. Rousseau 等人[21] 将文本分类视为

图分类问题, 通过对文本图进行子图挖掘来提取特征, 但通过子图挖掘的方式容易造成特征的损失. Peng 等人[22]先构建词的共现图, 再应用广度优先遍历来得到文本子图进行卷积操作, 以此达到分类的目的. 随着 GNN 的发展, 越来越多的基于 GNN 的模型被应用于文本分类中. Defferrard 等人[23] 提出了图卷积神经网络 (graph convolutional networks, GCN), 并且首先在文本分类任务中使用, 并优于传统的 CNN 模型. Yao 等人[6] 在整个语料库构建的单个大图上使用图卷积网络进行文本分类, 提出了 TextGCN 模型, 取得了出色的效果. 为了考虑更多的上下文特征, Liu 等人[7] 根据文本, 分别构造语义信息、句法信息和共现信息 3 种类型的文本图, 提出了 TensorGCN 模型. 其中, 节点信息可以在图内和图之间两种方式传递, 这样能充分挖掘文本的结构信息, 但需要占用较大的存储资源. Huang 等人[24] 为每个文本构建具有全局参数共享的图, 并提出 Text-level GNN 模型进行训练. 然而, 这些模型本质上是转导式学习 (transductive learning), 所以无法对含有新结构和新单词的文档进行分类. 在基于 GNN 的文本分类的归纳式学习中, Zhang 等人[9] 提出 TextING 用于归纳文本分类, 该模型为每个文档构建单独的图, 并使用门控图神经网络 (gated graph sequence neural networks, GGNN)[25] 来学习单词表示. 归纳式学习是基于图进行文本分析的趋势, Ding 等人[26] 使用超图 (hypergraph) 来基于注意力机制对每个文档进行建模, 并提出 HyperGAT 来支持文本超图上的表示学习. 与现有的方法不同, 本文通过增强词之间的上下文关系交互, 对基于 GNN 的方法缺少文本顺序特征的改进, 有效地整合文档中复杂的特征来提高最终文本分类的效果.

2 文本图的构建

对于从语料库中给定的一个单词的文本$W = w_1, w_2, w_3, \cdots, w_n$, 其中每个单词的表示为$w_n$, n是文本长度. 为每个文本W构建一个有向文本图$\mathcal{G} = (\mathcal{V}, \mathcal{E})$, 其中节点$\mathcal{V}$表示文本中的单词, 而边$\mathcal{E}$表示词节点之间的关系. 文本图$\mathcal{G}$中将文本重复出现的单词被视为的不同节点, 所以保留了完整的文本序列和句法依赖关系. 为了增强文本中单词之间的交互, 本文提出的文本图$\mathcal{G}$通过共现 (co-occurrences) 和句法依赖 (syntactic dependency) 两种上下文关系, 以及全局文档节点构成.

首先, 通过固定大小的滑动窗口内出现单词来得到单词之间的共现关系, 其默认窗口大小为 3, 这种方式已广泛用于 GNN 的文本表示学习. 其次, 利用 Stanza 工具[27] 对文本中包含的句子和单词进行据句法分析, 以此来得到词节点间含有句法依赖关系的边. 其中, 边保留了提取句法关系的方向, 但删除了部分不重要和冗余的关系, 例如: 代表标点符号和单词之间关系的“punct”和代表助动词关系的“aux”. 在为每个文本构建有向图时, 部分词节点之间根据共现与句法依赖两种关系构建的边会出现重合. 然而, 在基于 GNN 的方法中每个图的节点只能通过唯一的边连接到其他节点, 所以本文提出的构图方法中将重合的边合并. 文本图中还针对每个输入文本建立文档节点, 并通过单一的边指向词节点. 最后, 每个词节点创建一条有向边来连接自身. 文本图构造如图 1 所示, 其中doc代表文档节点, w_n代表词节点, 图中只保留w_3和w_4窗口大小为 3 的共现关系与句法依赖关系合并, 便于阐明文本图的结构.

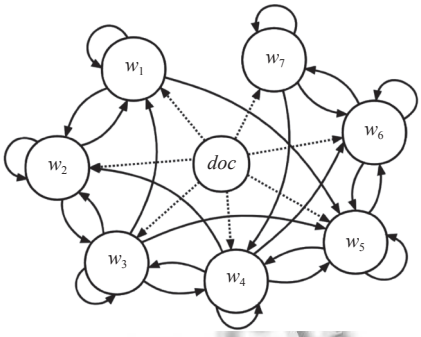

图 1 文本图的构建示例

本文提出的文本图$\mathcal{G}$保留了重复出现的词, 能对词节点的表示进行训练和更新来保留完整的序列和语义信息. 通过共现和句法依赖关系的合并, 使图中包含比单个关系文本图更多的边来连接词节点, 并增强了长依赖交互, 能够使用较浅的 GNN 学习到较好的文本表示, 从而避免过度参数化和因堆叠更多 GNN 层而导致过度平滑. 其次, 文档节点通过单向的边连接词节点可以避免在节点交互过程中, 全局信息对词结构信息的模糊. 此外, 连接节点自身的边能使词节点自身的特征参与传递. 文本图$\mathcal{G}$适用于对含有新结构和新单词的文档进行分类, 能够对文本进行归纳式学习.

3 TextIGAT 模型

TextIGAT 模型主要包括 3 个部分: 基于 GAT 的单词交互、文本序列特征交互, 以及文本表示学习, 整体架构如图 2 所示.

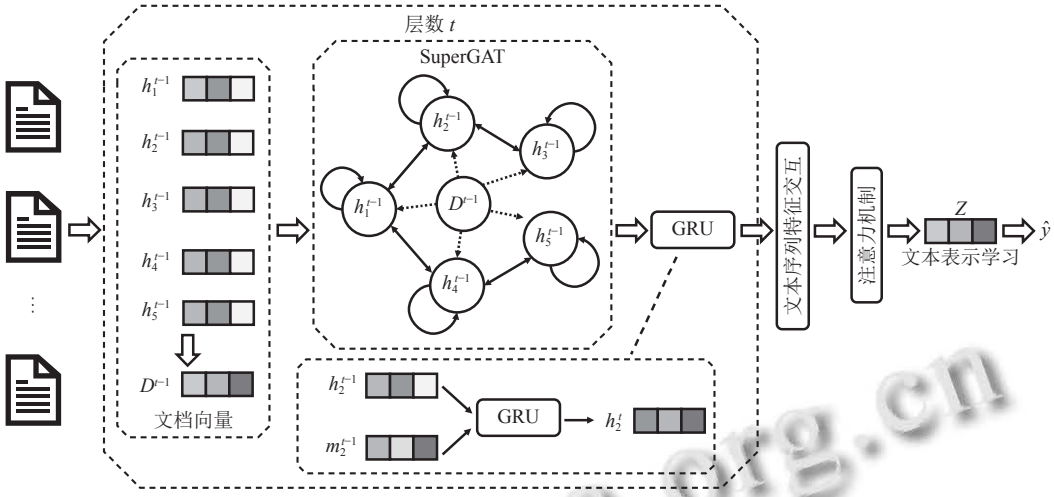

图 2 TextIGAT 模型示例图

3.1 基于 GAT 的单词交互

令输入文本的嵌入为$E=e_1,e_2,e_3,\cdots,e_n$,其中$e_n\in R^{d_w}$为词嵌入, d_w是词嵌入的维度. 当在第t层 GNN 网络时, 文本图中节点的隐藏状态表示为$H^t=h_1^t,h_2^t,\cdots,h_n^t$, 其中$h$是每个词节点的隐藏状态. 对于初始状态$H^0$将第$i$个节点的隐藏状态$h_i^0\in R^{d_w}$设置为它的嵌入$e_i$, 即$h_i^0=e_i$.

3.1.1 文档节点表示

模型中使用自注意力机制 (self-attention)[28] 对输入文档计算全局特征来初始化文档节点的向量, 使得局部词节点能与全局文本表示进行交互. 自注意力机制只需要通过对序列中的词向量进行计算, 因此不需要词节点之间关系的参与. 文档全局表示使用缩放的点积注意力机制 (scaled dot-product attention)[29] 进行计算, 这是点积注意力机制变体的一种, 注意力权重基于以下等式计算:

$$A=Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V$$

其中, 键 K (key) 和值 V (value) 的查询 Q (query) 是相同的向量并且等于H^t, 分母$\sqrt{d}$是控制注意力分数比例的比例因子. 然后对A进行的平均池化 (avg pooling) 得到文档全局表示$D\in R^{d_w}$, 其思想是从加权词节点中保留整体的特征来表示全局文本信息.

3.1.2 聚合

通过 GAT 的消息传递, 图中的每个节点都可以根据其相连接的节点中最具代表性的特征聚合成自身的节点特征. 为了在文本图上自适应的通过节点之间的重要程度进行消息传递, 本文提出的模型中使用 SuperGAT[14] 来聚合来自邻居节点和文档节点的信息. SuperGAT 是将 GAT 中注意力的计算方式与点积注意力机制相结合的图神经网络, 能够更有效为重要的相邻节点分配更大的权重. 令初始状态表示为h_i^0, 节点i在t层聚合来自邻接点特征的表示m_i^t可以如下获得:

$$e_{i,j}^t=a^{\mathrm{T}}\left([W_\ell h_i^{t-1}\|W_\ell h_j^{t-1}]\right)\cdot\sigma\left((h_i^{t-1})^{\mathrm{T}}h_j^{t-1}\right)$$

$$\alpha_{i,j}^t=\frac{\exp\left(LeakyReLU(e_{i,j}^t)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(LeakyReLU\left(e_{i,k}^t\right)\right)}$$

$$m_i^t=\sigma\left(\frac{1}{R}\sum_{r=1}^{R}\sum_{j\in\mathcal{N}_i}\alpha_{i,j}^t W_r h_j^{t-1}\right)$$

其中, a^{T}是注意力机制中的单层前馈神经网络, W_ℓ和W_r是模型中可训练的参数, σ为非线性函数 $Sigmoid$, $LeakyReLU$ 为激活函数, $\mathcal{N}_i$代表节点i在图中的邻节点, R为多头注意力 (multi-head attention) 机制中头 (head) 的数量, 默认大小为 3. 通过将 SuperGAT 应用于建立的文本图上, 单词节点可以灵活地聚合来自邻居节点和文档节点中更多重要特征.

3.1.3 更新

在聚合词节点信息之后, 模型基于 GGNN[25] 的思想来更新文本图中词节点的隐藏状态. GGNN 是采用循环神经网络思想来处理图结构数据和图嵌入的 GNN, 其通过门控机制有效地决定如何更新节点的隐藏状态. 本模块使用 GRU 的基础循环结构作为更新节

点状态的主要组件, 词节点更新公式如下:

$$z_i^t = \sigma\left(W_z b_i^t + U_z h_i^{t-1} + b_z\right)$$

$$r_i^t = \sigma\left(W_r b_i^t + U_r h_i^{t-1} + b_r\right)$$

$$\widetilde{h}_i^t = \tanh\left(W_h b_i^t + U_h(r_i^t \odot h_i^{t-1}) + b_h\right)$$

$$h_i^t = \widetilde{h}_i^t \odot z_i^t + h_i^{t-1} \odot \left(1 - z_i^t\right)$$

其中, W_x, U_x, b_x ($x \in \{z,r,h\}$)是可训练的权重和偏差值, tanh 表示双曲正切激活函数. z_i^t和r_i^t作为更新门和重置门分别控制有助于当前节点嵌入的信息. 由于第一层图神经网络能对一阶邻居进行操作, 通过堆叠 GNN 网络层数 T 后能获得每个节点的最终隐藏状态$H^T = h_1^T$, $h_2^T, h_3^T, \cdots, h_n^T$, n是节点数量.

3.2 文本序列特征交互

该模块旨在提取整个文档的词序信息, 进一步地增强词节点的表达能力. 循环神经网络可以根据其结构特征有效地提取文本序列特征, 因此使用双向门控循环神经网络 (Bi-GRU) 能够增强词节点的序列表示. Bi-GRU 模型由两个 GRU 组件组成, 可以同时处理来自前向和后向的输入. 在构建文本图的过程中, 由于将每个文档中重复出现的词作为不同的词节点, 而保留了完整的文本序列. 因此, 最终经过 GNN 的隐藏状态 $H^T = h_1^T, h_2^T, h_3^T, \cdots, h_n^T$可以看作是 Bi-GRU 的起始状态, Bi-GRU 的隐藏状态定义如下:

$$\overrightarrow{s}_i = GRU\left(\overrightarrow{s}_{i-1}, h_i^T\right)$$

$$\overleftarrow{s}_i = GRU\left(\overleftarrow{s}_{i-1}, h_i^T\right)$$

其中, $\overrightarrow{s}_i$和$\overleftarrow{s}_i$分别是第i个词节点在 Bi-GRU 中的前向和后向隐藏状态向量. 然后, 通过计算$\overrightarrow{s}_i$、$\overleftarrow{s}_i$和h_i^T的算术平均值, 来获得每个词节点的最终表示$S_i \in R^{d_w}$.

3.3 文本表示学习

模型最后通过所有的词节点的最终表示S_i计算文本表示 z 来对输入的文本进行分类. 基于注意力机制[30]使用以下公式计算每个词节点的注意力权重:

$$C_i = Softmax\left(w_\alpha^{\mathrm{T}} f(W_\alpha S_i) + b_\alpha\right)$$

其中, W_α表示权重矩阵, w_α是权重向量, b_α是偏差值, f是非线性激活函数, 如双曲正切变换. $Softmax$用于对词节点的注意力权重进行归一化. 注意力权重C_i通过计算词节点向量的加权和, 以此产生一个集成的文本向量表示 z:

$$z = \sum_{i=1}^{n} C_i S_i$$

最后, 通过将文本表示 z 输入到$Softmax$层来预测标签. 通过以下交叉熵函数来最小化损失:

$$\hat{y} = Softmax\left(W_y z + b_y\right)$$

$$\mathcal{L} = -\sum_k y_k \log\left(\hat{y}_k\right)$$

其中, W_y和b_y分别是权重和偏差值, $\hat{y}$是预测的标签分数, $\hat{y}_k$表示文本的第 k 个正确标签 (ground truth). 因此 TextIGAT 方法可以通过最小化所有标记文档的损失函数$\mathcal{L}$进行学习.

4 实验

4.1 数据集

为了实验结果对比的一致性, 本文对 TextING 模型中的 4 个广泛使用的英文文本数据集 (MR、Ohsumed、R8 和 R52) 进行实验.

MR[31]: 用于二元情感分类的电影评论数据集, 其中包含 5 331 条正面评论和 5 331 条负面评论, 每条评论只包含一个句子. 实验中训练集和测试集的划分来自于 Tang 等人[32].

Ohsumed: 临床数据集 MEDLINE 的子集, 涵盖了 5 年 (1987–1991 年) 的 270 种医学期刊的所有参考文献. 数据集中的每个医学摘要都有来自 23 个心血管疾病类别的一个或多个相关类别, Ohsumed 仅使用了属于一类的 7 400 份文件.

R8 和 R52: Reuters 21578 数据集的子集. R8 有 8 个类别, 可以拆分为 5 485 个训练和 2 189 个测试文档. R52 有 52 个类别, 可以拆分为 6 532 个训练文档和 2 568 个测试文档. 两个数据集中的每个文档只与一个主题相关联.

表 1 中展示了评估数据集的统计数据及其补充信息. 通过在文献 [10] 中的文本清理和标记为来预处理所有数据集, 并删除了 NLTK 中定义的停用词, 以及 Ohsumed、R8 和 R52 出现少于 5 次的低频词. 因为 MR 属于短文本, 因此不对其中的单词进行删减.

4.2 对比试验

实验中, 对比的基线模型可以分为 3 种类型:

(1) 基于序列的模型: 方法从局部连续词序列中捕获文本特征, 包括: CNN-non-static[10] 和 Bi-LSTM[33] 两

种基于 CNN 和 RNN 的代表性模型.

(2) 基于词嵌入的模型: 方法基于预先训练的词嵌入对文档进行分类, 包括: fastText[17] 和 SWEM[18] 两种对词嵌入使用平均池化或最大池化 (max pooling) 提取特征.

(3) 基于图的模型: 方法基于图形结构来进行单词之间的交互, 包括: Graph-CNN[22] 对文本子图进行卷积运算; TextGCN[6] 利用 GCN 来学习整个语料库级别图的单词和文档嵌入; TensorGCN[7] 在词张量图上采用图内和图之间传播; Text-level GNN[24] 构建具有全局参数共享的文档图进行文本表示学习; HyperGAT[26] 为每个文档构建超图并使用双重注意机制进行归纳分类; TextING[9] 是基于图的归纳式文本分类的 SOTA 模型.

表 1 评估数据集的统计表

Dataset	MR	Ohsumed	R8	R52
Docs	10 662	7 400	7 674	9 100
Train	7 108	3 357	5 485	6 532
Test	3 554	4 043	2 189	2 568
Words	18 764	14 157	7 688	8 892
Classes	2	23	8	52
Avg Len	20.39	135.82	65.72	69.82

4.3 实验设置

出于公平比较, 所有实验在 Intel Xeon E5-2 680 v4 CPU 和 RTX 3 090 GPU 上运行. 数据集选用文献 [9] 中提供的训练集和测试集, 随机选择训练集的 10% 作为验证集, 并根据验证集的性能调整超参数的最佳值. 本文基于 PyTorch 框架实现 TextIGAT 模型, 默认堆叠两层 GNN. 模型中采用 Adam 优化器[34] 对参数进行更新, Ohsumed 数据集的学习率为 0.001, 其他数据集的学习率为 0.000 3. 为防止模型训练中过拟合, Dropout 在每个模块中设置为 0.5, 并应用于词嵌入的初始状态. 文本图中随机删除节点间连接的边, 概率设置为 0.3, 以获得最佳性能. R8 数据集的 L2 正则化参数 (weight decay) 设置为 5E–5, 其他数据集为 5E–6.

实验使用 300 维度的 GloVe 词向量[35] 来初始化词嵌入. 针对归纳式文本分类任务, 初始的词嵌入在模型训练期间不进行更新. 不包含在语料库中的词 (out-of-vocabulary, OOV) 在预处理中会被替换为 UNK, 并从均匀分布 [−0.01, 0.01] 中随机采样初始化词嵌入. 基线模型使用其原始论文和复现中的默认参数设置.

4.4 实验结果

表 2 展示了本文方法和其他基线方法的文本分类的准确度结果, 其中 TextIGAT 在 4 个评估数据集上均达到了较好的分类效果, 展示了其出色文本分类能力.

对比表 2 中的结果, 基于图的模型通常优于 CNN、LSTM 和 fastText 等传统神经网络模型. 这是由于图形结构的特性造成的, 图形结构有利于文本处理能使词节点能够通过不同的搭配来学习更准确的表示. 这一观察表明文本分类性能可以通过捕获长距离词节点交互提高. 其中, 基于 GCN 的方法本质上是转导式学习, 这会降低测试具有新词和关系的文档的性能. TextING 和 Text-level GNN 在基于图的归纳式文本分类取得了不错的结果, 但由于欠缺对词关系的进一步考虑, 性能受到了限制. HyperGAT 模型使用超图构建的文本图能够减少计算消耗, 但也忽视了整体文本中的结构特征. TextIGAT 模型中文本图含有的边比其他基线方法要多, 能够减少图的直径并增强了节点之间的消息传递.

表 2 评估数据集上各种模型的准确率 (%)

模型	MR	Ohsumed	R8	R52
CNN-non-static	77.75	58.44	95.71	90.54
Bi-LSTM	77.68	49.27	96.31	90.54
fastText	75.14	57.70	96.13	92.81
SWEM	76.65	63.12	95.32	92.94
Graph-CNN	77.22	63.86	96.99	92.75
TextGCN	76.74	68.36	97.07	93.56
TensorGCN	77.91	70.11	98.04	95.05
Text-level GNN	75.47	69.40	97.89	94.60
HyperGAT	78.32	69.90	97.97	94.98
TextING	79.82	70.42	98.04	95.48
TextIGAT	**80.56**	**71.31**	**98.13**	**96.04**

TextIGAT 模型通过结合循环结构和 GNN, 显示出比大多数基于图的基线方法更强大的分类能力. 上下文序列特征在短文本或情感分类中起着关键作用, 例如在 MR 电影评论数据集中. 现有方法通常采用词之间的共现来引入顺序信息, 但这样的方式在整个文本图中表达词序的能力有限. 采用循环结构来对 GNN 的最终隐藏状态进行转换, 可以有效地保持文本整体的顺序特征. 通过实验表明, 本文提出的 TextIGAT 模型在对比其他基于图的基线模型取得了不错的提升.

4.5 消融实验

实验通过消融实验对 TextIGAT 模型各模块进行分析, 结果如表 3 中所示. 其中, w/o global, w/o self-loop, w/o co-occurrence 和 w/o dependency 是文本图的变体, 分别指文本图中删除连接全局文档节点、自循环、共现信息和句法依赖的边. w/o sequential 是模型中去除文本顺序特征的交互. 从实验结果可见, 删除文本图中任何关系所连接的边都会导致准确率下降. 其中, w/o

global 和 TextIGAT 之间的性能差距显示了将文档节点单向与词节点更新相结合的重要性. 此外, 不同关系的边在不同的数据集中有不一样的作用. 如在不使用共现的 MR 数据集和不使用句法依赖的 Ohsumed 数据集的准确度结果较差. 这表明共现在 MR 数据集中起着重要作用, 而句法依赖在 Ohsumed 数据集中是必不可少的. 而 TextIGAT 模型通过共现和句法依赖关系的合并, 能让两种关系相互补充, 在不同的数据集中发挥出更好的分类效果. 其次, 通过将 w/o sequential 与原始模型的结果进行比较, 表明与序列交互的结合可以增强 TextIGAT 模型表达能力, 尤其是对于短文本的 MR 数据集. 因此, 所有模块对于本文所提出的 TextIGAT 模型都是必需的.

表 3 TextIGAT 模型在消融实验上的准确率 (%)

数据集	MR	Ohsumed	R8	R52
w/o global	80.45	71.22	98.09	95.68
w/o self-loop	80.60	70.97	98.10	95.93
w/o co-occurrence	80.18	71.19	98.04	95.73
w/o dependency	**80.71**	71.08	97.97	95.89
w/o sequential	79.86	71.06	97.92	95.55
TextIGAT	80.56	**71.31**	**98.13**	**96.04**

4.6 参数分析

图 3 中记录不同数量的图层数的方法的模型性能. 在 MR 和 Ohsumed 数据集上, 当层数为两个或更多时, 模型达到了最佳的文本分类性能. 由于在文本图中引入了更多的边, 单词节点可以接收传递更多的信息, 且更准确地学习单词表示. 同时, 为了评估标记训练数据大小的影响, 实验比较了几个具有不同训练数据比例的基线模型. 图 4 记录在 MR 和 Ohsumed 数据集上使用不同训练集占比的测试准确率. 随着标记训练数据百分比的增加, 所有进行实验的模型都提高其分类性能. 而 TextIGAT 模型可以在标记文档有限的情况下显著优于其他基线, 证明了其在文本分类中的有效性.

4.7 个案分析

图 5 中展示了在 MR 数据集中正面评论和负面评论两个类别的注意力机制可视化效果. 其中, 突出显示的单词与注意力权重成正比, 并且在文本图中与之相连的词节点较为紧密, 它们与类别标签呈正相关. 通过本个案分析结果表示, TextIGAT 模型可以在注意力机制和图注意力模型的结合下, 准确地提取关键的文本信息来得到富有表现力的文本表示, 以此最终达到提升文本分类的目的.

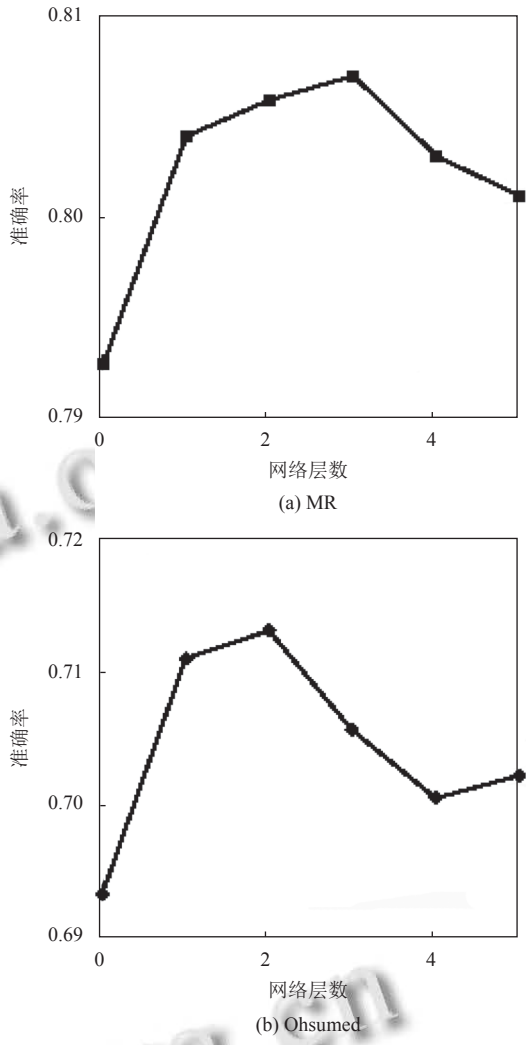

图 3 GNN 的层数对准确率的影响

5 结语

本文提出了一种基于门控图注意力网络的归纳式文本分类方法, 以有效地整合文本中的复杂特征和提取不同的上下文关系来提升分类效果. 该方法为语料库中每个输入文本构建图, 图中根据不同的上下文关系连接词节点, 增强了在单个的文本图中节点交互的距离和信息. 其次, 通过单向的文档节点引入全局文本信息的交互, 灵活避免对文本结构信息的模糊. 此外, 模型中对于 GRU 结构的利用, 提升了词节点的词顺序表示以及信息的更新. TextIGAT 模型专注于文本本身的特征, 能够对含有新单词和新结构的文本进行归纳式文本分类. 本文通过进行广泛的实验并与基线模型的对比, 实验结果证明了本文提出的 TextIGAT 模型有效提升了基于图神经网络的文本分类效果.

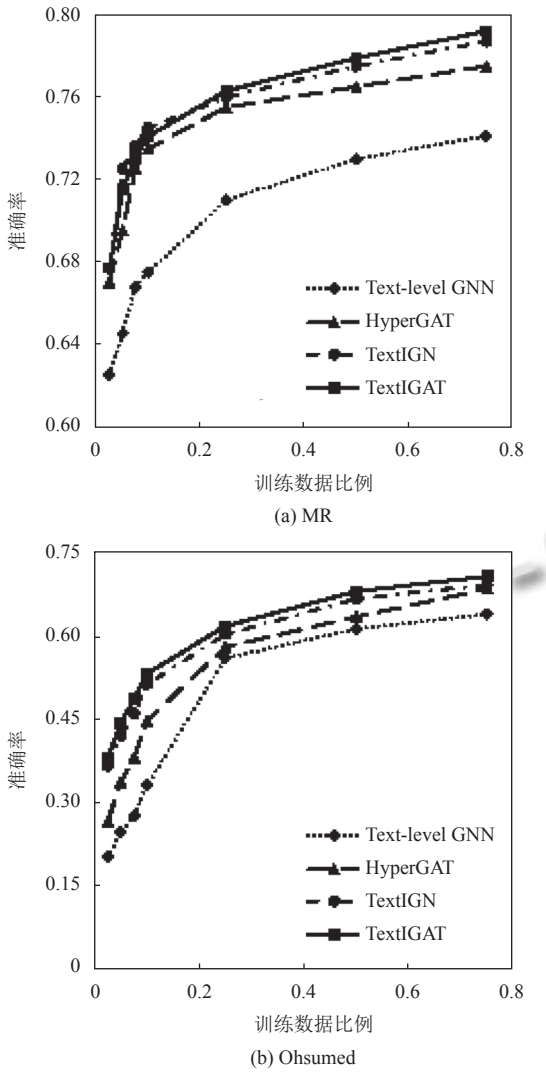

图 4 训练数据比例对准确率的影响

a) Positive reviews
the art direction and costumes are gorgeous and finely detailed , and kurys' direction is clever and insightful .

b) Negative reviews
theological matters aside , the movie is so clumsily sentimental and ineptly directed it may leave you speaking in tongues .

图 5 MR 数据集的注意力机制可视化

参考文献

1 Minaee S, Kalchbrenner N, Cambria E, *et al*. Deep learning-based text classification: A comprehensive review. ACM Computing Surveys, 2022, 54(3): 62.

2 Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains. Proceedings of 2005 IEEE International Joint Conference on Neural Networks. Montreal: IEEE, 2005. 729–734.

3 Scarselli F, Gori M, Tsoi AC, *et al*. The graph neural network model. IEEE Transactions on Neural Networks, 2009, 20(1): 61–80. [doi: 10.1109/TNN.2008.2005605]

4 Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR, 2017. 1–14.

5 Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR, 2018. 1–12.

6 Yao L, Mao CS, Luo Y. Graph convolutional networks for text classification. Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu: AAAI Press, 2019. 7370–7377.

7 Liu X, You XX, Zhang X, *et al*. Tensor graph convolutional networks for text classification. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020. 8409–8416.

8 Li W, Li SH, Ma SM, *et al*. Recursive graphical neural networks for text classification. arXiv: 1909.08166, 2019.

9 Zhang YF, Yu XL, Cui ZY, *et al*. Every document owns its structure: Inductive text classification via graph neural networks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL, 2020. 334–339.

10 Kim Y. Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: ACL, 2014. 1746–1751.

11 Zhang X, Zhao JB, LeCun Y. Character-level convolutional networks for text classification. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: NIPS, 2015. 649–657.

12 Liu PF, Qiu XP, Huang XJ. Recurrent neural network for text classification with multi-task learning. Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York: IJCAI Press, 2016. 2873–2879.

13 Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing: ACL, 2015. 1556–1566.

14 Kim D, Oh A. How to find your friendly neighborhood: Graph attention design with self-supervision. Proceedings of the 9th International Conference on Learning

Representations. Vienna: ICLR, 2021. 1–25.
15 Cho K, Van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014. 1724–1734.
16 Wang GY, Li CY, Wang WL, *et al*. Joint embedding of words and labels for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne: ACL, 2018. 2321–2331.
17 Joulin A, Grave E, Bojanowski P, *et al*. Bag of tricks for efficient text classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Valencia: EACL, 2017. 427–431.
18 Shen DH, Wang GY, Wang WL, *et al*. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne: ACL, 2018. 440–450.
19 Lai SW, Xu LH, Liu K, *et al*. Recurrent convolutional neural networks for text classification. Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI). Austin: AAAI Press, 2015. 2267–2273.
20 Yang ZC, Yang DY, Dyer C, *et al*. Hierarchical attention networks for document classification. Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: NAACL, 2016. 1480–1489.
21 Rousseau F, Kiagias E, Vazirgiannis M. Text categorization as a graph classification problem. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing: ACL, 2015. 1702–1712.
22 Peng H, Li JX, He Y, *et al*. Large-scale hierarchical text classification with recursively regularized deep graph-CNN. Proceedings of the 2018 World Wide Web Conference. Lyon: ACM, 2018. 1063–1072.
23 Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 3844–3852
24 Huang LZ, Ma DH, Li SJ, *et al*. Text level graph neural network for text classification. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 3442–3448.
25 Li YJ, Tarlow D, Brockschmidt M, *et al*. Gated graph sequence neural networks. Proceedings of the 4th International Conference on Learning Representations (ICLR). arXiv: 1511.05493, 2015.
26 Ding KZ, Wang JL, Li JD, *et al*. Be more with less: Hypergraph attention networks for inductive text classification. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: ACL, 2020. 4927–4936.
27 Qi P, Zhang YH, Zhang YH, *et al*. Stanza: A Python natural language processing toolkit for many human languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Online: ACL, 2020. 101–108.
28 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of Neural Information Processing Systems. Long Beach: NIPS, 2017. 5998–6008.
29 Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 1412–1421.
30 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473, 2014.
31 Pang B, Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor: ACL, 2005. 115–124.
32 Tang J, Qu M, Mei QZ. PTE: Predictive text embedding through large-scale heterogeneous text networks. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney: ACM, 2015. 1165–1174.
33 Huang ZH, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv: 1508.01991, 2015.
34 Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv: 1412.6980, 2014.
35 Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: ACL, 2014. 1532–1543.

(校对责编: 孙君艳)