摘要

基于深度图神经网络的文本分类研究

随着迈入信息时代，在现实世界中，我们的周围充斥着大量文本。如何有效地处理这些文本并提取所需要的信息，这是一个有挑战性的问题。得益于文本挖掘领域的快速发展，文本分类吸引了来自学术界和工业界的广泛关注，并取得了十分显著的发展。

文本分类是自然语言处理一项十分经典和基础的任务，它的目的是为给定的文本赋予正确的标签，已经被广泛用于各种实际应用中，比如，问答系统、主题分类和论文发表地点预测等。与早期的基于规则的文本分类方法相比较，基于统计的文本分类方法有更好的准确率和稳定性。但是它们仍然严重依赖手工特征工程，这一个过程需要严谨的处理或者专业的领域知识，可能会消耗大量的时间并且费用高昂。同时，这些模型并没有充分利用大量的训练数据，因为相关的特征已经被提前定义。最近，深度学习的出现极大地改变了人工智能领域。这些深度学习方法能够自动地对复杂的特征进行建模并产生有语义和语境的文本表征，从而消除了繁琐复杂的手工特征设计过程，已经成为了包括文本分类在内的自然语言处理任务的主流范式。虽然之前基于序列的深度学习模型可以很好地捕捉局部连续词序列中的语义和句法信息，在文本分类任务中已经取得了令人印象深刻的进步，但是它们仍然有一些限制。首先，它们不能很好地捕获长距离的单词之间的交互，从而忽略了这些单词的全局共现信息。其次，它们忽略了文本中蕴含的语法或者句法结构，而这对正确理解文本是有帮助的。

最近，图神经网络逐渐变成一个研究热点，因为其在处理复杂结构数据和关系的强大表现能力。在文本分类任务中，一系列基于图神经网络的模型也取得了令人瞩目的表现。本文基于图神经网络首先研究了文本分类的一个具体应用场景，即论文发表场所预测。针对这个应用场景，之前提出的模型忽略了论文内部的结构信息，同时使用手工制作的特征来表示论文，而忽略了那些涉及

高级语义的特征。为了解决以上问题，本文提出为每个论文摘要构建语义图，并执行双重注意力消息传递神经网络以得到它们的辨别性论文摘要表征。在相关数据集上的大量实验结果表明，所提出的模型性能非常出色，显著优于现有的基线方法。

接下来，本文分析了之前提出的基于图神经网络的文本分类模型存在的缺陷。首先，这些模型仅仅考虑了文本内单词的一阶邻居，同时如果堆叠一定数量图神经网络层之后会出现过平滑问题。针对这些缺陷，本文提出一种深度图注意力扩散模型用于文本分类任务。具体来说，该模型首先使用注意力扩散技术来扩大文本中单词的感受野，这样做能够在每一层中捕获长距离的单词交互。此外，为了训练更深的神经网络以提取单词的隐藏语义，该模型将图神经网络的特征转换和特征传播过程解耦，这样也能缓解过平滑问题。在一系列基准数据集上的表现证明了本文提出的模型的优越性。

关键词：

图神经网络，文本分类，论文发表场所预测

Abstract

Deep Graph Neural Networks for Text Classification

As we move into the information age, we are surrounded by a large amount of texts in the real world. It is a challenging problem to process these texts effectively and extract the information we need. Benefit from the rapid development in text mining, text classification has attracted a lot of attention from both academia and industry and has made significant progress.

Text classification is a very classical and fundamental task in natural language processing, which aims to assign the correct label to a given text and has been widely used in various practical applications, such as question answering, topic classification, and paper publication venue prediction. Compared with earlier rule-based text classification methods, statistical-based text classification methods have better accuracy and stability. However, they still rely heavily on manual feature engineering, a process that requires rigorous processing or specialized domain knowledge and can be time-consuming and costly. At the same time, these models do not take full advantage of the extensive training data, as the relevant features are already defined in advance. Recently, the emergence of deep learning has dramatically changed the field of artificial intelligence. These deep learning methods can automatically model complex features and produce semantic and contextual representations of text, thus eliminating the tedious and complex manual feature design process. It has become the dominant paradigm for natural language processing tasks, including text classification. Although previous sequential-based deep learning models can capture semantic and syntactic information in local sequences of consecutive words well and have made impressive progress in text classification tasks, they still have some limitations. First, they do not capture long-distance word interactions well, thus ignoring the global co-occurrence information of these words. Second, they ignore the grammatical or syntactic structure within the text, which is helpful for correctly understanding the

text.

Recently, graph neural networks (GNNs) are becoming a research hotspot due to their powerful performance in handling complex structural data and relations. A series of GNN-based models have achieved impressive performance in text classification tasks. This work first investigates a specific application scenario of text classification, i.e., paper publication venue prediction. For this application scenario, the previously proposed models ignore the structural information inside the papers, using hand-crafted features to represent the papers while missing those features that involve high-level semantics. To address the above issues, this work proposes constructing semantic graphs for each paper abstract and executing a dual attention message passing neural network to obtain their discriminative paper abstract representations. Extensive experimental results on relevant datasets show that the performance of the proposed model is excellent, consistently outperforming existing baseline approaches.

Next, this work analyzes the drawbacks of previously proposed GNN-based text classification models, i.e., they only consider the one-hop neighbors of words within the text and suffer from oversmoothing problems if many GNN layers are stacked. To address these limitations, this work proposes a deep graph attention diffusion model for text classification tasks. Specifically, the model first uses attentional diffusion techniques to widen the receptive field of words in the text, which can capture long-range word interactions at each layer. In addition, to train a deeper network to extract the hidden semantics of words, the model decouples the feature transformation and feature propagation processes of GNNs, which can alleviate the oversmoothing problem. The performance on a series of benchmark datasets demonstrates the superiority of the model proposed in this work.

Keywords:

Graph Neural Network, Text classification, Paper-publication venue prediction

目录

第 1 章 绪论

1.1 研究背景和意义

1.1.1 文本分类的研究背景和意义

随着网络信息时代的来临，导致现实生活中数据量迅速增加。这些数据有各种表现形式，比如图像数据、语音数据、视频数据、文本数据等。其中，文本数据在互联网数据中占据着一定的比重，它有丰富的数据来源，具体形式包括聊天记录、商品评论、电子邮件、搜索片段等。这些文本数据有很高的利用价值，如果能针对特定场景有效地从中提取和分析有价值的信息，可以更好地满足人们的信息个性化需求。比如用户可能每天都会收到很多信息，其中垃圾信息可能占有一定的比例，这会降低用户的使用体验。运营商可以通过分析内容判断它是否为垃圾信息，帮助用户进行过滤，以满足用户需求。此外，电子商务平台可以通过商品评论挖掘用户的商品喜好，进行个性化的推荐；由于虚假信息在网络中的传播速度极快，为了降低它们的影响，电子社交平台需要准确辨别这些虚假信息，同时审核网络评论内容以帮助政府进行舆论监督。

文本分类是属于自然语言处理（natural language processing, NLP）领域的一种非常有效的信息处理技术[1]。它可以从丰富的文本内容中有效地提取有用的信息，并根据这些信息，将文本数据分类为不同类型，为这些文本指定预先定义的标签，从而满足个性化信息处理需求[2]。文本分类在实际中有广泛的应用场景，经典的场景包括信息检索[3]、新闻分类[4]、情感分类[5]、意图识别[6]、主题分类[7]等；最近提出的一些应用场景，如抽取式问答[8]和论文发表场所预测[9]，其本质上也属于文本分类的范畴。互联网时代创建文本数据的速度远远超过人工标注的速度，并且人工标注的结果很大程度上受到人的主观因素的影响，比如专业知识的限制。利用机器自动地从复杂文本中定位所需的信息并提取特征，实现分类的自动化，可以得到更可靠和无偏的结果。这意味着设计一个高效的

文本分类算法不仅仅是有用的，而且是绝对必要的，同时也能促进自然语言处理其他领域的发展。但是由于文本的非结构特性，从文本中提取有价值的信息具有一定的挑战性，对研究人员提出了很高的要求[10]。

1.1.2 论文发表场所预测的研究背景和意义

论文发表场所预测作为文本分类的一个重要应用场景，它的目的是预测一篇指定的论文最合适的发表场所[11]。这个任务对研究人员有很强的实际意义。首先，伴随着各国对科学技术发展的高度重视，学术界的科学活动与日俱增，产生的学术论文的数量呈指数型增长，论文发表场所迅速增加，因此选择一个合适的论文发表场所（期刊或者会议）发表工作有可能是非常耗时的。一个不合适的论文发表场所选择可能会导致研究者工作的延迟发表甚至有可能被直接拒绝[12]。因为一个存在的情况是论文和发表场所主题不匹配，而不是论文本身的质量问题，这就延迟了一个可能对特定领域的研究产生重大影响的工作的及时发表，阻碍了科学技术的发展。其次，一些研究者经常把自己局限于已知出版物的论文发表场所中，这就限制了其他合适场所的选择[13]。产生这种情况的一个可能原因是，即使有和研究工作更契合的发表场所，研究者们由于不能及时了解这些发表场所的具体信息，所以做出了次优的选择。第三，对于一个年轻研究者来说，在跨学科的时代背景下，选择适当的学术场所发表自己的工作是非常吃力的，因为研究工作可能同时适合多个科学领域[9]，论文发表场所预测的结果可以辅助研究者做出合理的决定。同时预测的结果对于期刊编辑也是有意义的，因为这些结果可以帮助期刊编辑确定该论文是否和某个期刊的主题匹配，这样就减少了编辑审核的工作量，加快了论文评审流程。最后，对于学术出版商来说，一个好的学术发表场所预测模型能够促进研究者在它们旗下的期刊上发表优秀的作品，从而形成一个良性循环，不断提升期刊的影响力。

1.2 国内外研究现状

1.2.1 文本分类研究现状

经典的基于机器学习的文本分类模型主要有三个步骤。首先是对文本进行预处理，将文本内容转化成计算机可以理解的形式。预处理操作是后面所有步骤的基础，因为文本分类的输入数据是原始的、非结构化的文本。与其他类型的数据如图像和时间序列不同，文本信息不具备内在的数值表示。比较流行的文本预处理操作有分词、去除停用词、词型还原等。这些操作执行完成之后，文本转化为分离的标记列表，其中的单词通过词汇表被映射成索引。第二个步骤是文本特征表示，也就是将文本映射至一个特征空间，常用的方法有 bag-of-words[14]（BoW）、N-gram[15]、TF-IDF[16]、word2vec[17]、GloVe[18]等。BoW 的核心思想是用和单词字典大小相等的向量表示每一个文本，其中向量中的第i个元素的值表示字典中第i个单词在该文本中的频率。这种方法由于将文本视作无序的词的集合进行简化表示，导致它没有考虑单词之间的语义关系和句子之间的上下文结构信息。N-gram[15]模型考虑了邻居单词的信息，并利用这些相邻词构建了一个字典。这个模型认为一个句子出现的概率可以用其中包含单词的联合概率来表示。为了方便，N-gram 模型引入了马尔可夫假设，也就是第n个单词仅仅依赖前面$n-1$个单词，而与其他单词没有关系。在设置了大小为 N 的滑动窗口之后，模型根据极大似然估计得到单词出现的条件概率。TF-IDF[16]同时利用单词频率和逆文档频率来表示文本，其中单词频率是一个文档中指定单词出现的次数，逆文档频率指的是包含该单词的文档数和语料库中所有文档数的对数比的倒数。在 TF-IDF 模型中，一个单词的重要性随着在一个文档中出现次数的增加而升高，随着在多个文档中出现次数的增加而降低。由于依赖于词汇表的规模，用 TF-IDF 方法得到的文本表示的维度通常会很多。为了缓解时间复杂度和模型消耗带来的问题，常见的做法包括限定单词的数量，或者通过主成分分析[19]（Principal Components Analysis，PCA）和线性判别分析[20]（Linear Discriminant Analysis，LDA）进行维度缩减。前面提及的文本表征方法侧重于

捕获单词的句法表示，而忽略了单词的语义。一个明显的例子是同义词，虽然它们在语义上是相似的，但是在之前的方法的特征空间下，这些单词的表征却是完全独立正交的，这显然是不合理的。word2vec[17]的出现则大大缓解了这一情况。word2vec 的核心思想是通过浅层神经网络利用单词局部上下文信息获得其良好的单词表征。与之前的模型得到的高维词向量不同，word2vec 得到的单词表征是低维连续的实值向量。它包含两个重要的模型变体，CBOW 和 skip-gram。CBOW 尝试通过周围单词来预测中心词，而 skip-gram 正好相反，它是根据中心词来预测周围的上下文单词。这两个模型变体真正的目标并不是学会如何完美的预测这些单词，而是得到这些单词有意义的嵌入。GloVe[18]模型同时考虑局部上下文信息和全局词共现统计信息，并尽可能地让训练得到的词向量包含更多的语法和句法信息。本质上，GloVe 是一个基于统计的模型，它通过构造语料库的词共现矩阵来学习词之间的语义相似性。word2vec 和 GloVe 这两种词嵌入模型和语言模型类似，但是它们更侧重于单个词的嵌入，并且都是假设一个词的含义可以从句子中周围的词中提取出来。上面提及的方法获得的文本特征表示全部都是固定的，也就是得到的词向量只能表达一个含义，而不能模拟一词多义现象，这是以上方法最大的缺陷。之后提出的基于上下文的文本表示方法则解决了上述这个问题。

第三个步骤是将得到的文本特征表示输入到分类器中做出预测。比较流行的分类器选择有支持向量机[21]（Support Vector Machine，SVM）、朴素贝叶斯[22]（Naïve Bayes, NB）、K 近邻[23]（K-Nearest Neighbors，KNN）、随机森林[24]（Random Forest，RF）。SVM[21]最初是用来处理模式识别的二分类任务，随后被引入到文本分类任务中作为文本表征的分类器。它尝试在一维特征空间中构造一个能最大化和不同类别的训练集数据距离的超平面，这样被证明能获得最大的泛化能力和最小的分类误差。如果需要进行非线性分类，SVM 则利用适当形式和参数的核函数将输入映射到高维空间，以更好地分离不同类别的训练数据。NB[22]基于这样一个假设，即各个特征是互相独立的，它们之间没有影响。这个前提让 NB 模型有效的同时也让其结构和计算变得十分简单，因此被广泛用在文本分类任务中。它实际上先在训练集中观察给定特征的类的先验概率，

然后通过贝叶斯定理计算其后验概率。同时，NB 模型对于缺失值也不敏感。但是，在实际中，模型所需要的独立性假设很难满足，可以会影响其表现。基于 KNN[23]的分类器同样十分简单快速，因为它没有任何参数，只需要在多次迭代中不断计算数据之间的距离。它的核心思想是为未标记的文本赋予最近的k个样本中包含实例最多的一个类别。但是，这种模型很大程度上受到所选择距离函数的影响。RF[24]是一种基于融合的方法，它使用随机抽取的特征子集，包含多个树分类器，其中的树共享一个数据分布。RF 的泛化误差取决于不同树之间的关系，并且随着森林中树木的增加而最终收敛。由于可以取得比较好的结果，RF 经常在实践中被使用。上面介绍的经典的机器学习文本分类模型从数据中进行学习。这些根据经验和知识从数据中预定义的特征对最终模型的表现至关重要。但是，特征工程是十分繁琐且艰巨的，且主观因素影响较大。并且由于特征是预定义的，因此不能泛化到新的任务中，也不能充分发挥大量训练数据带来的优势。如果受到计算资源和计算复杂度的限制，这些传统模型在小型文本数据集中表现可能会优于深度学习模型。

近年来，深度学习为包括自然语言处理在内的人工智能的各个领域带来了革命性的变化。由于深度学习不需要复杂的特征工程，且能够从大量数据中自动提取有意义的潜在信息和辨别性特征，因此获得了研究人员的广泛关注，并逐渐成为包括文本分类在内的自然语言处理各种任务的主流范式。研究人员已经提出了很多基于深度学习的模型用于处理文本分类任务，与之前传统的机器模型相比较，它们取得了令人满意的表现。最简单的深度学习架构是多层感知机（MultiLayer Perceptron，MLP）[25]，它通常包括输入层，带有激活函数的隐藏层和输出层。基于 MLP 的模型的输入通常用比如 BoW、TF-IDF 或者词嵌入等特征提取技术表示。卷积神经网络（Convolutional Neural Network，CNN）[26]和循环神经网络（Recurrent Neural Network，RNN）[27]作为深度学习的两种典型结构也被用来解决该任务。CNN 最开始应用在计算机视觉领域，它通过学习不同的卷积核来自动提取图像特征。由于 CNN 的平移不变、权重共享和局部连接等优秀性质，该网络架构受到了研究人员的欢迎。第一个基于 CNN 的文本分类模型叫做 DCNN[28]，它使用动态 k-max-pooling 技术来生成句子矩阵的特征图，

能够明确捕获单词和短语的长短期关系。接下来，一种更简单但是更加流行的TextCNN[29]模型被提出。TextCNN 首先输入预训练的词向量，然后经过一层带有不同大小卷积核的卷积层和池化层，最后输入全连接层进行分类操作。VDCNN[30]模型由于借鉴了残差网络的跳跃连接技术，使得它能够堆叠更多的卷积层而不出现性能退化。RNN 是用来建模序列结构数据的一个流行选择，这种架构能够提取句子结构信息，进而捕获上下文单词的潜在关系。TopicRNN[31]利用 RNN 构建单词的局部依赖关系，同时采用潜在主题模型捕获全局语义依赖。但是这种 RNN 架构训练过程不稳定，容易出现梯度消失和梯度爆炸的问题。因此，长短期记忆网络[32]（Long Short-term Memory，LSTM）被提出并成为一种流行的 RNN 架构。Tree-LSTM[33]将此架构应用在树结构上，它认为树能够更好地表示短语。MT-LSTM[34]利用多时间尺度的有价值信息对文档进行建模，并取得了优异的表现。受到人类只关注重要物体的启发，研究人员在深度学习模型中引入了注意力机制，同时根据得出的注意权重让模型具有一定的可解释性。在文本分类领域，注意力机制也被广泛应用。HAN[35]包含两个级别的注意力层，其中单词级别的注意力机制用于辨别句子中重要的单词，句子级别注意力机制用于辨别文档中重要的句子。通过这两种不同粒度的注意力机制，模型就能得到更加准确的文档表征用于分类。LSTMN[36]按照从左到右的顺序处理文本，并利用记忆模块和注意力模块进行推理。接下来，为了解决 RNN 只能顺序处理文本和 CNN 捕获单词之间关系能力欠佳的问题，研究人员提出了基于编码器-解码器架构的 Transformer[37]模型。为了让单词能够融合位置信息，在输入编码器之前，Transformer 在词嵌入的基础上融合了相对位置的嵌入，然后使用自注意力机制并行地计算句子中某个单词对其他单词的注意力分数，随后利用带有残差连接的前馈神经网络得到编码器的输出，并将其和$i-1$位置的解码器输出一起输入到解码器中。一系列基于 Transformer 架构的预训练模型取得了巨大的成功。这些模型都在大规模的语料中以自监督的方式进行训练来获得单词的上下文表征，然后在特定任务上进行微调，在很多下游任务中取得了令人惊讶的表现，其中最具代表性的是 GPT[38]和 BERT[39]模型。GPT 仅仅利用 Transformer 的解码器结构进行下一个单词预测，属于典型的自回归语言模型。在处理较长

文本序列或者执行语言生成任务时，GPT 能够发挥最大的优势。对 GPT 进行一定改进之后，也能将其用在包括文本分类的各种下游任务中。BERT 是一个利用双向 Transformer 架构获得文档的编码向量，并通过掩码语言模型（Masked Language Model，MLM）和下一句预测（Next Sentence Prediction，NSP）两个任务进行训练。前者是让模型根据上下文信息预测被随机遮蔽的句子中的单词，后者是判断给定的两个句子是否在文中是上下句关系。BERT 在包括文本分类的各种自然语言处理下游任务中都取得了令人瞩目的表现，说明模型能够融入大规模语料所蕴含的知识，同时证明了预训练-微调这种范式的有效性。

现实世界中，图是一种常见的数据结构，它能有效地刻画不同节点之间的结构关系。很多研究人员致力于将神经网络作用在图结构数据中，并提出了一种名为图神经网络（Graph Neural Network，GNN）的深度学习架构以自动地提取节点辨别性特征用于下游任务。虽然文本表现出一种序列性，但内部也包含图结构，比如定义了句子中单词之间语义和句法关系的解析树。如果能通过图神经网络将这些信息编码至设计的模型中，会有利于模型更好地捕获文本之间的各种丰富关系。具体到文本分类任务上，之前的 CNN 和 RNN 都无法捕获句子中远距离单词之间的交互和句法结构信息，而基于图神经网络的模型则可以很好地解决上面的问题。这些模型通常将文本构建成文本图或者语料图，然后将图神经网络作用在构建的图中得到单词或者文本的更新嵌入，随后将其输入到全连接层中输出标签概率，从而将文本分类问题转化为图中的节点分类问题。GraphCNN[40]在构建好的词图上进行卷积操作，这样既能通过卷积学习不同层次的语义，又能捕获长距离和非连续的语义信息。TextGCN[41]首先基于词共现和文档单词关系构建了一个包含单词和文本节点的异构图，其中存在单词-单词边和单词-文档边。然后它将图卷积网络（Graph Convolutional Network，GCN）作用在异构图中，并在文本标签的监督下共同学习单词和文本的表示。在文本分类的结果显示，TextGCN 的性能明显优于基于 CNN 和 RNN 的模型。但是 TextGCN 是为整个语料库构建图，这会占用很大的资源，同时以直推式的方式进行训练，这就说明该模型不能预测未见过的文本。接下来的一些工作要么通过降低模型复杂度要么通过改变模型训练策略来降低建模成本。SGC[42]模型就

是通过去除连续层之间的非线性变化和折叠权重矩阵为一个线性变化来降低模型复杂度，并且在文本分类实验中证明了其和 TextGCN 相当的性能。TextING[43]模型在固定大小的滑动窗口上利用单词之间的共现为每个文本构建单独的图，然后执行门控图神经网络学习单词嵌入，并通过图池化操作得到文档的表示进行最后的分类。该模型是以一种归纳式的方式进行训练，并且不需要固定图结构，因此可以预测未见过的文本。

1.2.2 论文发表场所预测研究现状

一种典型的方法是将论文发表场所预测问题转化为分类问题。虽然考虑论文的全文更有利于正确预测它和哪些期刊或者会议更加匹配，但是论文的长度会对计算资源提出更高的要求。因此之前的模型采取的做法是提取被看作是一篇论文简述的摘要，重点是从中提取有用的特征信息以提高预测性能。这些模型收集一系列已经发表的论文，并把它们对应的出版物视作标签，通过使用有监督的机器学习模型构建预测器。目前存在的论文发表场所预测模型主要可以分为基于机器学习的方法和基于深度学习的方法。在基于机器学习的一系列方法中，代表性模型包括 VRS[44]、TMVR[9]、PRS[45]。VRS 提出同时考虑不同期刊的写作风格和主题，并区分不同类型的相邻论文的贡献以预测给定论文的发表场所。TMVR 提出了一种基于论文摘要的主题匹配模型，模型遵循的基本假设是每个发表场所都与潜在主题相关联，然后通过论文与发表场所之间的主题匹配进行预测。PRS 采用了传统的机器学习方法卡方统计提取论文特征，然后用词频向量表示论文，但是这样做忽略了单词之间的相关性和高层语义信息。

最近，基于深度学习的一系列模型被设计用来自动地从论文中学习单词表示以解决这个任务，并取得了好的表现。Pubmender[46]利用 CNN 学习论文摘要嵌入，能够有效地捕获连续单词的局部序列所包含的语义信息。CNAVER[47]提供了一个融合框架，采用基于排名的论文-论文网络模型和期刊-期刊网络模型的融合。同时，它还解决了由数据的稀疏性、稳定性和多样性带来的一些问题。HASVRec[48]提出通过使用双向 LSTM 和分层注意力网络来学习摘要嵌入，在预

测发表场所方面取得了令人鼓舞的表现。但是，这些模型都忽略了论文摘要的结构信息，如何有效地融入这些信息以增强模型预测性能，仍然值得探索。

1.3 本文完成的工作

本文主要研究了两个问题，首先是针对文本分类的一个具体应用场景，即论文发表场所预测，设计了一个基于图神经网络的模型提升预测性能。其次是针对文本分类这个经典的研究问题，本文首先总结了之前模型的缺陷，然后设计了一个基于深度图注意力扩散网络的模型，在多个基准数据集上取得了明显地提升。

（1）论文发表场所预测

本文设计了一个名为 VPALG 的模型。该模型首先将每个序列化的摘要表示为图以捕获单词之间的拓扑结构，进而从摘要中覆盖高级语义信息。然后，它通过执行消息传递神经网络从形成的摘要图中学习摘要表征。具体来说，它通过消息传递过程对单词的节点嵌入过程进行编码，并通过图读出过程将这些节点嵌入聚合成摘要图嵌入。为了同时捕获单词之间的多样性，模型在消息传递和图读出过程中采用了双重注意力机制。最后，模型可以从学习到的摘要嵌入及其相应的发表场所监督下进行训练。为了进一步提升模型，自训练方法被采用以获得更多高置信度的伪标签。提出的 VPALG 模型在三个收集的数据集上的表现显著高于之前的针对这个问题提出的基线模型。

（2）文本分类

本文设计了一个名为 DADGNN 的模型，用于解决之前基于图神经网络的文本分类模型的缺陷。具体来说，该模型首先使用注意力扩散技术来扩大文本中每个单词的感受野，从而可以在每一层中捕获长距离单词的交互。此外，为了提取单词的隐藏语义信息，该模型解耦了图神经网络的特征传播和特征转换过程以训练更深层的网络。最后，通过注意力机制计算每个节点的权重以获得精确的文本表示用于分类。在一系列基准数据集上的实验结果证明了提出模型的优越性。

1.4 论文组织安排

本文的组织结构具体如下所示：

第 1 章，首先介绍文本分类的研究背景和意义，然后介绍文本分类的一个重要应用场景论文发表场所预测的研究背景和意义。接下来分别回顾了文本分类和论文发表预测的研究现状，并指出之前的模型在这两种研究问题存在的缺陷和本文的研究内容。

第 2 章，本章主要介绍所研究的理论基础，主要包括图神经网络的介绍和典型的深度图神经网络架构。

第 3 章，本章详细介绍针对第一个研究问题所提出的基于图神经网络的论文发表场所预测模型。模型的主要组成部分，即摘要图构建，基于注意力机制的消息传递，基于注意力机制的图读出和模型预测层

第 4 章，本章介绍了论文发表场所预测模型进行的实验，主要包括实验数据集、基线模型、评估指标和实验设置。同时，本章还分析了实验结果、消融实验和参数敏感性。

第 5 章，本章详细介绍针对第二个研究问题所提出的基于图注意力扩散网络的文本分类模型。模型的主要组成部分，即文本图构建，特征转换层，特征传播层，图级别表征和图谱分析。

第 6 章，本章介绍了基于图注意力扩散网络的文本分类模型的实验，主要包括实验数据集、实验设置和基线模型。同时，本章还进行了实验结果分析，消融实验验证，模型参数使用，模型消耗内存，同时进行了可视化实验分析。

第 7 章，本章进行了工作总结，并提出后续研究计划。

第 2 章 相关理论介绍

2.1 图神经网络符号表示

图作为一种复杂的数据结构，由节点和边构成，可以十分有效地表示现实中的许多复杂系统，如社交网络、知识图谱和金融网络等[49]。由于深度学习技术的兴起，许多研究者探索如何使用这项技术自动提取图中的节点特征以更好地作用于下游任务。由于和图像、文本这种规则的网格结构相比较，图中不同的节点可能有不同数量的邻居节点，同时这些节点之间存在着依赖关系，它们不再服从独立同分布，因此直接应用以前开发的卷积神经网络是不可取的。图神经网络就是在这种情况下出现的，它已经发展成为一个普遍的和强大的计算框架，用于处理不规则的图结构数据。

为了更好地描述图神经网络，下面介绍经常使用的符号。一个图可以被表示为$G=(V,E)$，其中$V=\{v_1,\cdots,v_n\}$和$E=\{e_1,\cdots,e_m\}$分别表示图中的节点集合和边集合。每条边$e_i\in E$可以被表示为$e_i=(v_j,v_k)$，其中$v_j,v_k\in V$。如果图G为无向图，则它的邻接矩阵是对称的，用A进行表示，当且仅当v_i和v_j之间有边时$\mathrm{A}_{ij}=1$或者$\mathrm{A}_{ji}=1$；如果图G为带权图，则边上的值为相应的权重；如果图G为有向图，一般来说它的邻接矩阵是不对称的，也就是$\mathrm{A}_{ij}\neq\mathrm{A}_{ji}$。此外经常用$N(v_i)=\{v_j\,|\,(v_i,v_j)\in E\}$表示节点$v_i$的邻居节点集合，用$\mathrm{D}=\mathrm{diag}(\mathrm{D}_{11},\cdots,\mathrm{D}_{nn})$表示图$G$的度矩阵，且满足$\mathrm{D}_{ii}=\sum_j^n\mathrm{A}_{ij}$，其中$\mathrm{diag}=(\cdot)$代表只有矩阵的主对角线上有值，其他位置全为 0。如果图中的节点有属性值X，其中$\mathrm{X}\in\mathbb{R}^{n\times d}$是原始节点特征矩阵，$\mathrm{x}_i\in\mathbb{R}^d$是节点$i$的原始特征向量。

最近流行的图神经网络模型可以分为两类，分别是谱域图神经网络和空域图神经网络。谱域图神经网络从图信号处理的角度定义图卷积核，其中的图卷积操作可以被理解为平滑图信号，即从原始信号中去除噪声。空域图神经网络

则从信息传递的角度在空域直接定义图卷积操作，因该操作的灵活性、泛化性和高效率，得到研究者的青睐，发展十分迅速。下面分别详细介绍这两种图神经网络及代表性模型。

2.2 谱域图神经网络

谱域图神经网络模型假定图是无向的，其中谱图卷积操作是通过图傅里叶信号变换（Graph Fourier Transform）定义的。而图傅里叶变换依赖于图拉普拉斯矩阵的特征分解。通常图拉普拉斯矩阵被定义为L=D-A，它被对称归一化之后可以被表示为$\mathrm{L}=\mathrm{D}^{-1/2}\mathrm{L}\mathrm{D}^{-1/2}=\mathrm{I}_n-\mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2}$，其中$\mathrm{I}_n\in\mathbb{R}^{n\times n}$是单位矩阵。由于对称归一化的图拉普拉斯矩阵L是半正定的，因此它可以被特征分解为$\mathrm{L}=\mathrm{U}\Lambda\mathrm{U}^{\top}$，其中$\Lambda=\mathrm{diag}(\lambda_0,\lambda_1,\cdots,\lambda_n)\in\mathbb{R}^{n\times n}$是由特征值组成的对角矩阵，并且这些特征值满足$0=\lambda_0\le\lambda_1\le\cdots\le\lambda_{n-1}=\lambda_{\max}$。$\mathrm{U}=\{\mathrm{u}_0,\mathrm{u}_1,\cdots,\mathrm{u}_{\mathrm{n}\text{-}1}\}\in\mathbb{R}^{n\times n}$则是特征值对应的特征向量组成的矩阵。对称归一化的图拉普拉斯矩阵的特征向量可以组成一个正交空间，因为$\mathrm{U}\mathrm{U}^{\top}=\mathrm{I}_n$，因此这些特征向量可以被视为图傅里叶变换中的基。作用在图信号x上的图傅里叶变换可以被表示为$f(\mathrm{x})=\mathrm{U}^{\top}\mathrm{x}$，逆图傅里叶变换则被表示为$f^{-1}(\mathrm{x})=\mathrm{U}\mathrm{x}$，其中x是根据图傅里叶变换得到的信号。图傅里叶变换将输入图的信号映射到由对称归一化的图拉普拉斯矩阵的特征向量为基构成的正交空间上。因此，信号 x和滤波器g在图G上的图卷积可以被定义为：

$$g*_G\mathrm{x}=\mathrm{U}((\mathrm{U}^{\top}g)\odot(\mathrm{U}^{\top}\mathrm{x}))=\mathrm{U}g_\theta\mathrm{U}^{\top}\mathrm{x} \quad \text{.....(2.1)}$$

其中$g_\theta=\mathrm{diag}(\mathrm{U}^{\top}g)$表示由参数$\theta$组成的对角化的图滤波器。各种谱域图神经网络模型设计的最大区别在于参数θ选择的不同。

Spectral CNN[50]令$g_\theta=\Theta_{k,i,j}$为一组学习参数，并且具有类似卷积神经网络的k个通道的卷积核，图卷积操作具体可以表示为：

$$\mathrm{X}_{:,j}^{(k)}=\sigma(\sum_{i=1}^{d^{(k-1)}}\mathrm{U}\Theta_{i,j}^{(k)}\mathrm{U}^{\top}\mathrm{X}_{:,i}^{(k-1)})\ \ j=1,2,\cdots,d^{(k)} \quad \text{.....(2.2)}$$

其中 $\mathrm{X}^{(k-1)}\in\mathbb{R}^{n\times d^{(k-1)}}$ 是第 $k-1$ 层的输入图信号，并且 $\mathrm{X}^{(0)}=\mathrm{X}$。$d^{(k-1)}$ 和 $d^{(k)}$ 分别是输入和输出的特征维度。$\Theta_{i,j}^{(k)}$ 是一个由可训练参数填充的对角矩阵的图卷积核。由于 Spectral CNN 需要对图拉普拉斯矩阵进行特征分解，因此它的时间复杂度为 $O(n^3)$，这也就意味着，模型不能作用于规模特别大的图上。此外，该模型所学习的图滤波器是针对特定领域的，因此不能直接泛化到具有其它图结构的领域中。

为了解决上面的限制，ChebNet[51]提出利用切比雪夫多项式（Chebyshev polynomials）来近似代替图滤波器 g_θ。图滤波器和切比雪夫多项式的表达式如下所示：

$$g_\theta=\sum_{k=0}^{K}\theta_k T_k(\tilde{\Lambda}) \quad \text{.....(2.3)}$$

$$T_0(x)=1 \quad T_1(x)=x \quad T_i(x)=2xT_{i-1}(x)-T_{i-2}(x) \quad \text{.....(2.4)}$$

其中 $\tilde{\Lambda}=\frac{2}{\lambda_{\max}}\Lambda-\mathrm{I}_n$。采取该操作的原因是 $\Lambda\in[0,\lambda_{\max}]$，如果直接参与网络训练的话，有可能因为梯度问题造成训练不稳定，但是经过变换之后 $\hat{\Lambda}\in[-1,1]$。

因此，使用上述定义的图滤波器 g_θ 作用在图信号 x 的图卷积表达形式为：

$$g*_G\,\mathrm{x}=\mathrm{U}(\sum_{i=0}^{K}\theta_i T_i(\hat{\Lambda}))\mathrm{U}^{\top}\mathrm{x}=\sum_{i=0}^{K}\theta_i T_i(\mathrm{L})\mathrm{x} \quad \text{.....(2.5)}$$

其中 $T_i(\mathrm{L})=\mathrm{U}T_i(\hat{\Lambda})\mathrm{U}^{\top},\mathrm{L}=2\mathrm{L}/\lambda_{\max}-\mathrm{I}_n$。

在公式 2.3~2.5 的表达式中，可以得出图滤波器具有 K 阶局部空间性，这也就意味着它可以利用目标节点 K 阶以内的邻居节点，能够明确地利用图的结构信息。和 Spectral CNN 相比较，ChebNet 将卷积核中可训练的参数降低为 $K+1$，远远少于节点数目。同时，ChebNet 不需要进行图拉普拉斯矩阵的特征分解，因此降低了模型的时间复杂度。

因为L的最大特征值小于等于 2，GCN[52]令$\lambda_{\max} \approx 2$和$K = 1$，通过一阶近似 ChebNet 进一步降低模型复杂度，则公式 2.5 可以被简化为：

$$g *_G \mathrm{x} = \theta_0 T_0(\mathrm{L})\mathrm{x} + \theta_1 T_1(\mathrm{L})\mathrm{x} = (\theta_0 + \theta_1 \mathrm{L})\mathrm{x} \quad (2.6)$$

因为$\lambda_{\max} \approx 2$，则$\mathrm{L} = 2\mathrm{L}/\lambda_{\max} - \mathrm{I}_n = \mathrm{L} - \mathrm{I}_n$，则公式 2.6 进一步可以表示为：

$$g *_G \mathrm{x} = (\theta_0 - \theta_1(\mathrm{I}_n - \mathrm{L}))\mathrm{x} = (\theta_0 - \theta_1(\mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2}))\mathrm{x} \quad (2.7)$$

为了进一步减少训练的参数，GCN 令$\theta = \theta_0 = -\theta_1$，则公式 2.7 被表示为：

$$g *_G \mathrm{x} = \theta(\mathrm{I}_n + \mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2})\mathrm{x} \quad (2.8)$$

由于$\mathrm{I}_n + \mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2} \in [1,2]$，如果在深度学习模型中使用该卷积算子的话，有可能会出现数值不稳定和梯度方面的问题。为了解决这个问题，GCN 引入了重归一化的技巧，即用$\mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2}$替代$\mathrm{I}_n + \mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2}$，其中$\mathrm{D} = \mathrm{D} + \mathrm{I}_n$和$\mathrm{A} = \mathrm{A} + \mathrm{I}_n$，则公式 2.8 可以被表示为$g *_G \mathrm{x} = \theta\mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2}\mathrm{x}$。如果进一步引入卷积神经网络多卷积核的思想，则 GCN 的最终公式定义如下：

$$\mathrm{H} = \mathrm{X} *_G g_\Theta = \sigma(\mathrm{A}\mathrm{X}\Theta) \quad (2.9)$$

其中$\mathrm{A} = \mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2}$是对称归一化邻接矩阵，X 是输入的图信号，Θ为可训练的权重矩阵。

GCN 通过一阶 ChebNet 近似，大大降低了模型可学习的参数量，如果不考虑多卷积核通道，则仅仅需要一个可训练的参数。虽然 GCN 只关注于节点的一阶邻居节点，但是通过堆叠多层图卷积，仍然可以扩大卷积核的感受野，提升模型的表现力。

2.3 空域图神经网络

基于空域的图神经网络模型根据图中节点之间的空间关系定义卷积操作，可类比于卷积神经网络对于图像的处理。它的核心思想是迭代地聚合邻居节点的信息以更新中心目标节点的嵌入。接下来介绍经典的空域图神经网络模型。

MPNN[53]定义了基于空域的图神经网络模型的统一框架，该框架分为两个阶段，即消息传递阶段和图读出阶段。其中消息传递阶段包含两个重要的函数，消息聚合函数和节点更新函数，具体过程可以表示为：

$$\begin{aligned} m_v^{(k+1)} &= \sum_{u \in N(v)} M_k(\mathrm{x}_v^{(k)}, \mathrm{x}_u^{(k)}, e_{uv}) \\ \mathrm{x}_v^{(k+1)} &= U_k(\mathrm{x}_v^{(k)}, m_v^{(k+1)}) \end{aligned} \quad \text{.....(2.10)}$$

其中e_{uv}表示边上的特征，其中边的两端分别为节点u和节点v。

节点的原始特征经过k次迭代更新之后能够得到节点的隐藏特征，这项特征被输入到下一神经网络层中执行节点级别的下游任务。当要执行图级别的任务时，需要使用图读出函数将更新之后的节点嵌入以生成整个图的输出向量，具体过程可以表示为：

$$u = R(\{\mathrm{x}_v^{(k)} : v \in V\}) \quad \text{.....(2.11)}$$

其中u代表得到的图的输出向量。需要注意的时，图读出函数R应该具有置换不变性。

如果指定M_k, U_k和R的函数的具体表达形式之后，就可以转换为很多具体的空域图神经网络模型。

GraphSAGE[54] 是一个通用的归纳式学习框架，能够利用节点特征信息，为以前从未见过的节点有效地生成节点嵌入向量。此外，图中节点的邻居数量可能变化很大，尤其是当图的规模特别大时，这种情况尤其明显。因此 GraphSAGE 提出不使用目标节点全部的邻居，而是为每个节点采样固定的邻居参与节点更新。具体来说，GraphSAGE 由消息聚合函数和节点更新函数组成，用公式表示为：

$$\begin{aligned} m_{N(v)}^{(k)} &= \mathrm{AGGREGATE}^{(k)}(\mathrm{x}_u^{(k-1)} : u \in N(v)) \\ \mathrm{x}_v^{(k)} &= \mathrm{UPDATE}^{(k)}(\mathrm{x}_v^{(k-1)}, m_{N(v)}^{(k)}) \end{aligned} \quad \text{.....(2.12)}$$

其中$N(v)$表示从节点v的全局邻居节点中均匀采样固定数量的邻居节点集合。$m_{N(v)}^{(k)}$是来自第k层的目标节点的邻居节点聚合的消息。

GAT[55]认为前面提出的图神经网络模型为中心节点的邻域提供相同的权重，

这样就无法体现出不同邻居节点的重要性，因此它根据节点的相对重要程度，采用注意力机制为不同的邻居节点赋予不同的权重。GAT 的图卷积层定义如下：

$$h_v^{(k)} = \sigma(\sum_{u \in N(v)} \alpha_{uv}^{(k)} \mathrm{W}^{(k)} h_u^{(k-1)}) \quad \text{.....(2.13)}$$

其中σ代表非线性激活函数，$h_v^{(0)} = \mathrm{x}_v$，$\alpha_{uv}^{(k)}$是节点u和节点v之间的注意力权重，计算公式如下：

$$\alpha_{ij}^{(k)} = \mathrm{softmax}(\rho(\mathrm{a}^{\top}[\mathrm{W}^{(k)} h_i^{(k)} \| \mathrm{W}^{(k)} h_j^{(k)}])) \quad \text{.....(2.14)}$$

其中$a^{\top}$和$\mathrm{W}^{(k)}$都是可训练的参数，ρ代表 LeakyReLU 激活函数。Softmax 操作能让得到的注意力值位于（0,1）。此外，GAT 还引入多头注意力机制来提升模型的表现能力以得到有用的节点的特征，具体执行过程通过扩展公式 2.13 表示如下：

$$h_v^{(k)} = \sigma(\frac{1}{Q}\sum_{q=1}^{Q}\sum_{u \in N(v)} \alpha_{uv}^{(k)} \mathrm{W}^{(k)} h_u^{(k-1)}) \quad \text{.....(2.15)}$$

其中Q表示注意力头的数量。通过对多头注意力机制得到的节点嵌入进行平均操作可以得到节点最终的嵌入向量。

2.4 本章小结

本章主要回顾了所研究问题的相关理论。为了更清晰地表述之后涉及的理论模型，首先介绍了图神经网络的常用符号表示；接下来详细介绍了谱域图神经网络的发展历程及代表性模型；最后介绍了最近受到特别多关注的空域图神经网络模型的发展和典型模型。

第 3 章 基于图神经网络的论文发表场所预测模型

大多数现有的基于深度学习的论文发表场所预测模型使用卷积神经网络或者循环神经网络有效地捕获局部连续文本信息，因为这两种网络架构都侧重于局部性和序列性。虽然在这个任务上取得了好的表现，但现有工作的实用性在现实场景中可能会受到显著影响，主要由于以下几个因素。首先，论文摘要是高度结构化的，这有助于正确理解论文的内容[56]。然而，这种结构信息却不能被目前使用的深度学习模型捕捉。如果用图建模论文的话，由于模型对结构的学习能力更强，能够很好地解决上述问题。其次，这些论文与高级语义和精确的逻辑相关联，与网页与新闻等一般的文档有显著不同，而之前的模型使用 BoW 输入，并没有使用包含语义信息的词嵌入技术，例如 GloVe 和 BERT。已经有文献证实，好的词嵌入对表示文本是必不可少的[57]。第三，现有的方法经常使用这样一个假设，即一篇论文只与发表场所有关。然而，在一个指定场所发表的论文并不意味着另一个发表场所不合适。因此，用单一标签进行训练是不切合实际的。但是，人工标记不仅会耗费大量时间，而且会受到人为偏见的影响。基于上述讨论，本文提出了一种名为 VPALG 的模型，目标是为论文摘要提取更具辨别性的特征来推断发表场所，从而获得更好的预测性能。

3.1 模型框架

总的来说，模型的基本思想是将预测任务转化为分类任务，并且可以从在线发表的论文中学习。具体来说，本文首先从总共 l 个出版场所收集 n 篇发表的论文摘要。让 $\Omega=\{x_i,y_i\}_i^n$ 代表论文摘要的集合，其中 x_i 表示第 i 篇摘要的内容，$y_i\in\{0,1\}^l$ 表示其对应的发表场所。值得注意的是，每个 y 可以被当作 x 的不完全标签向量，可以通过自训练的方式更新。模型把 Ω 当作训练数据集。本文实际上要做的是从 Ω 中学习模型 f，可以预测论文摘要的潜在出版场所。受先前消息传递神经网络的启发[58]，本文提出 VPALG 有效地从 Ω 学习分类模型。更

准确地说，它首先将每个摘要表示为单词图。然后，模型通过基于注意力机制的消息传递过程对词的节点嵌入进行编码，并通过基于注意力机制的图读出过程将这些词的节点嵌入聚合到摘要的图嵌入中。最终目标是最小化摘要表示的预测结果与真实标签或伪标签之间的损失。具体来说，模型的目标函数可以表述如公式 3.1 所示：

$$\mathcal{L}(\Pi,\Phi,\Theta)=\frac{1}{n}\sum_{i}^{n}\ell(f_{\Pi}(f_{\Phi}(f_{\Theta}(G_i))),y_i) \quad \text{.....(3.1)}$$

其中 G_i 代表第 i 个摘要构成的语义图；$f_{\Pi}, f_{\Phi}, f_{\Theta}$ 分别代表由 Π、Φ 和 Θ 参数化的预测模型、基于注意力机制的图读出和基于注意力机制的消息传递。ℓ 是损失函数。为了更加清晰，提出的模型的总体框架如图 3.1 所示，最重要的符号展示在表 3.1 中。接下来，本文将从摘要图构建，基于注意力机制的消息传递和基于注意力机制的图读出详细介绍所提出的 VPALG 模型。

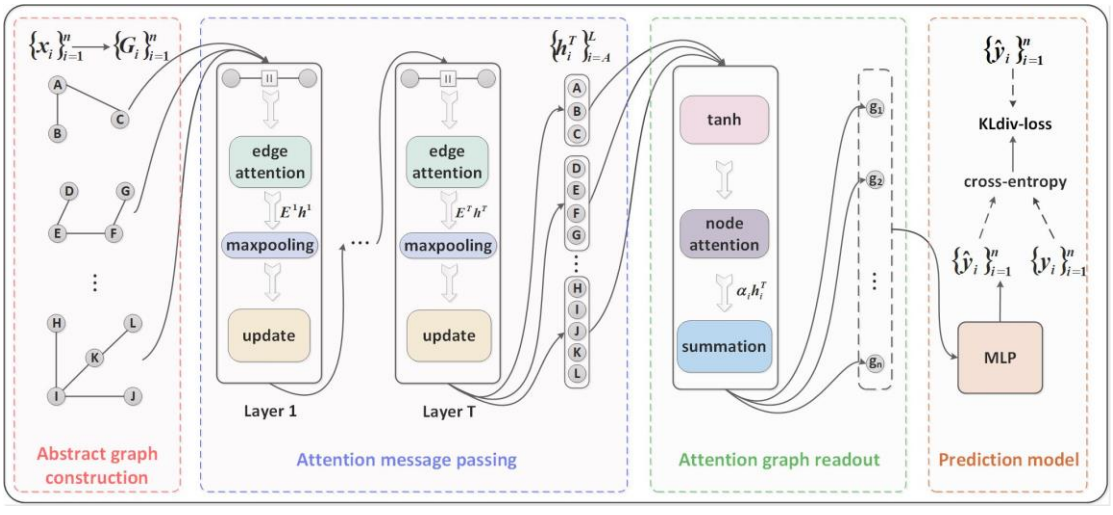

图 3.1 VPALG 的整体框架。

表 3.1 重要符号说明

符号	说明
n	训练集中论文摘要的数目
l	论文出版场所的数目
d	节点和图嵌入的维度
x	论文摘要的内容
$G=(\mathcal{V},\mathbf{E},\mathbf{H})$	论文摘要图
$\mathbf{y}\in\{0,1\}^l$	标签向量

表 3.1（续表）重要符号说明

符号	说明
$\Pi=\{\mathbf{W}_p\}$	预测模型的参数
$\Phi=\{\mathbf{W}_t,\mathbf{w}_e\}$	基于注意力机制的图池化过程的参数
T	基于注意力机制的消息传递的层数
$\Theta=\{(\mathbf{W}_a^t,\mathbf{W}_u^t,\gamma^t)\}_{t=1}^T$	基于注意力机制的消息传递的参数

3.2 摘要图构建

与基于 BoW 输入的模型不同，VPALG 将每一个摘要表示成一个单词语义图，以进一步捕获单词网络的拓扑结构。在摘要图中，模型使用大小为 S 的滑动窗口构建单词-单词边。形式上，对于每个摘要 x，对应的摘要图被表示为 $G=(\mathcal{V},\mathbf{E},\mathbf{H})$。首先，$\mathcal{V}$ 表示出现在 x_i 中的单词的节点集。其次，$\mathbf{E}\in\mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ 包含所有节点对之间的边权重，其中如果 x 中的第 i 个词和第 j 个词是邻居，则 $\mathbf{E}_{ij}=1$，否则为 0。最后，$\mathbf{H}=[\mathbf{h}_1,\cdots,\mathbf{h}_{|\mathcal{V}|}]^\top\in\mathbb{R}^{|\mathcal{V}|\times d}$ 包含由预训练词嵌入方法（例如 GloVe 和 BERT）得到的原始节点嵌入。值得注意的是，模型利用三种方式来构建单词节点之间的边，如图 3.2 所示：(a) 如果两个单词是邻居，则它们之间存在双向边；(b) 如果两个单词是邻居，则它们之间存在反向的单向边；(c) 如果两个单词是邻居，则它们之间存在前向的单向边。此外，这三种创建边的方式都包含所有单词的自循环边。

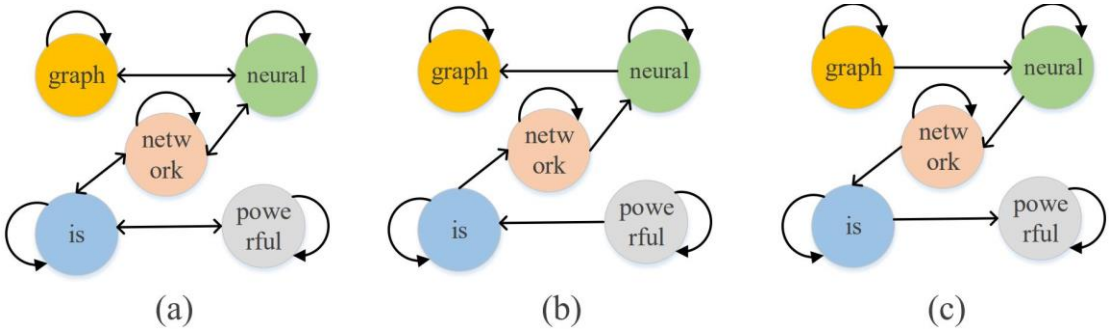

图 3.2 三种构建摘要图的示意图。采用的示例句子为，“graph neural network is powerful”：（a）双向图；（b）单向图（后向）；（c）单向图（前向）。

3.3 基于注意力机制的消息传递层

为了从每个摘要图中编码节点嵌入，模型执行了注意力消息传递过程。它的基本思想是通过从图中的相邻节点传播的消息来更新目标节点嵌入。具体来说，模型在总共T层上传播消息以及更新节点嵌入，这个过程可以表示为$f_{\Theta}(\mathbf{E},\mathbf{H})$，$\Theta=\{(\mathbf{W}_a^t,\mathbf{W}_u^t,\gamma^t)\}_{t=1}^{T}$。邻接矩阵和特征矩阵开始的值分别为$\mathbf{E}^0=\mathbf{E}$，$\mathbf{H}^0=\mathbf{H}$。在每一层$t$，本文首先使用加权最大池化操作计算聚合嵌入$\mathbf{P}^t=[\mathbf{p}_1^t,\cdots,\mathbf{p}_{|\mathcal{V}|}^t]^{\top}$，如公式 3.2 所示：

$$\mathbf{p}_i^t=\mathbf{Maxpooling}(\mathbf{E}_{i1}^t\mathbf{h}_1^{t-1},\cdots,\mathbf{E}_{i|\mathcal{V}|}^t\mathbf{h}_{|\mathcal{V}|}^{t-1}),\quad i=1,\cdots,|\mathcal{V}| \tag{3.2}$$

其中h^{t-1}表示第$t-1$层对应的节点嵌入，$\mathbf{E}^t=[\mathbf{E}_{ij}^t]_{|\mathcal{V}|\times|\mathcal{V}|}$包含着第$t$层的边上的权重，也就是边注意力值。边注意力是在第$t-1$层的节点嵌入计算的，它通过执行串联嵌入的线性变换然后应用 softmax 函数进行归一化来实现，具体过程如下所示：

$$\mathbf{E}_{ij}^t=\begin{cases}\mathbf{W}_a^t(\mathbf{h}_i^{t-1}\,\|\,\mathbf{h}_j^{t-1}), & \text{if } \mathbf{E}_{ij}^0=1\\ 0, & \text{otherwise}\end{cases}\quad i,j=1,\cdots,|\mathcal{V}| \tag{3.3}$$

$$\mathbf{E}_{ij}^t=\frac{\exp(\mathbf{E}_{ij}^t)}{\sum_{j=1}^{|\mathcal{V}|}\exp(\mathbf{E}_{ij}^t)},\quad i=1,\cdots,|\mathcal{V}| \tag{3.4}$$

其中“||”表示串联操作，$\mathbf{W}_a^t$表示第t层的线性注意变换参数。

在获得邻居消息聚合嵌入后，模型接着利用移动平均公式更新节点嵌入，然后应用 ReLU 激活函数进行非线性更新，具体过程如公式 3.5 和 3.6 所示：

$$\mathbf{h}_i^t=\gamma_i^t\mathbf{W}_u^t\mathbf{h}_i^t+(1-\gamma_i^t)\mathbf{W}_u^t\mathbf{p}_i^t \tag{3.5}$$

$$\mathbf{h}_i^t=\mathrm{ReLU}\left(\mathbf{h}_i^t\right),\quad i=1,\cdots,|\mathcal{V}| \tag{3.6}$$

其中操作$\mathrm{ReLU}(x)=\max(0,x)$；$\mathbf{W}_u^t$表示在第$t$层的线性更新转换权重；$\gamma_i^t$是对应$h_t^i$的自适应调节向量。在这里，$\gamma_i^t$也会在模型训练过程中进行更新。图注意

力网络[55]也采用了类似的注意力机制，但它只是迭代地向前传播聚集的邻域注意力信息。这也就是说，随着层数的增加，节点的原始信息会逐渐消失。在公式 3.5 中，模型引入了调节参数 γ_i^t，所以节点可以在传播过程中保持局部性，并且仍然可以获得有关邻居节点的更新信息。同时，在公式 3.5 中它还能够捕获使用移动平均更好的组合配置[59]，这在图表示学习中有强大的表现力。此外，由于每个节点及其连接节点之间的注意力分数是在局部网络结构中计算的，因此模型明确地包含了图的拓扑信息。

3.4 基于注意力机制的图读出层

给定节点嵌入 $\mathbf{H}^T=[\mathbf{h}_1^T,\cdots,\mathbf{h}_{|\mathcal{V}|}^T]^\top$，模型通过执行基于注意力机制的图读出过程获得摘要图表示。为了捕获出现在摘要中的单词的权重，受到之前工作的启示[60]，模型利用了一种节点级别注意机制。因此，模型可以将最终的图嵌入计算为节点嵌入的加权平均值。具体来说，注意力图读出过程可以被正式表达为 $f_{\mathbf{\Phi}}(\mathbf{H}^T)$，$\mathbf{\Phi}=\{\mathbf{W}_t,\mathbf{w}_e\}$。首先，注意力向量 $\boldsymbol{\alpha}\in\mathbb{R}^{|\mathcal{V}|}$ 通过以下公式计算：

$$\mathbf{V}_i=\tanh(\mathbf{W}_t\mathbf{h}_i^T) \quad (3.7)$$

$$\alpha_i=\frac{\exp(\mathbf{V}_i\cdot\mathbf{w}_e)}{\sum_{j=1}^{|\mathcal{V}|}\exp(\mathbf{V}_j\cdot\mathbf{w}_e)},\quad i=1,\cdots,|\mathcal{V}| \quad (3.8)$$

其中 $\mathbf{W}_t$ 是节点嵌入向量的注意力转换参数，$\mathbf{w}_e$ 是线性转换向量，$\mathbf{V}_i$ 是 $\mathbf{V}$ 的第 i 行。

其次，最终图嵌入 g 可以被计算为节点嵌入的加权平均值，计算过程如公式 3.9 所示：

$$\mathbf{g}=\sum_{i=1}^{|\mathcal{V}|}\alpha_i\mathbf{h}_i^T \quad (3.9)$$

3.5 模型预测层

本文将摘要的图嵌入视为新的表示，给出了一种新形式的训练数据集

$\hat{\mathbf{\Omega}}=\{(\mathbf{g}_i,\mathbf{y}_i)\}_{i=1}^{n}$。因此，模型可以在$\hat{\mathbf{\Omega}}$上训练一个预测模型。具体来说，本文将预测模型 $f_{\Pi}(\mathbf{g})$， $\mathbf{\Pi}=\{\mathbf{W}_p\}$形式化为完全连接的线性层。对于每个 g_i，对应的预测 y_i 公式为：

$$\hat{\mathbf{y}}_i=\text{sigmoid}(\mathbf{W}_p\mathbf{g}_i) \quad \text{.....(3.10)}$$

其中 $\mathbf{W}_p$ 表示预测模型的参数向量。

最后，本文将每个 $\hat{y}_i$ 和 y_i 之间的损失函数指定为常用的交叉熵函数，给出以下关于参数 {Π, Φ, Θ} 的目标函数为：

$$\mathcal{L}_{CE}(\mathbf{\Pi},\mathbf{\Phi},\mathbf{\Theta})=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{l}\mathbf{y}_{ij}\log\hat{\mathbf{y}}_{ij}+(1-\mathbf{y}_{ij})\log(1-\hat{\mathbf{y}}_{ij}) \quad \text{.....(3.11)}$$

为了进一步完善模型以获得更好的泛化能力，本文继续在训练集上对模型进行自训练。自训练方法可以为每篇论文选择多个高置信度的伪标签，有效地解决了之前的模型只把论文和它所发表的场所相关联的问题。一般自训练目标函数可以用 Kullback-Leibler （KL） 散度表示：

$$\mathcal{L}_{ST}(\mathbf{\Pi},\mathbf{\Phi},\mathbf{\Theta})=\mathbf{KL}[Q\,\|\,P]=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{l}q_{ij}\log\frac{q_{ij}}{\hat{\mathbf{y}}_{ij}} \quad \text{.....(3.12)}$$

其中，Q 是目标分布，P 是模型当前的预测分布。自训练的核心概念是迭代地利用预测 P 获得目标 Q 作为伪标签来改进模型。

常用的获取目标分布 Q 的方式有硬标签和软标签。硬标签是将超过提前设置的阈值的高置信度预测转换为 1。软标签[61]首先将预测 $\hat{y}_{ij}$ 提高到二次方锐化预测，然后通过频率归一化进行处理。这样做的优势是可以增强高置信度的预测并同时减弱低置信度的预测。在实践中，实验发现软标签方法比硬标签方法效果更好。一种可能的原因是硬标签丢失了太多原始目标分布的信息，容易产生错误传播。相反，软标签的目标分布是为每个实例计算的，不需要设置置信度阈值。模型的上述目标函数可以通过基于梯度的方法进行解决，并且随机优化算法也能够有效地处理大规模训练数据集。

3.6 本章小结

本章首先阐述了针对论文发表场所预测这一文本分类的实际应用场景现有的模型的缺陷。然后为了解决这些缺陷，本章详细介绍了模型的重要组成部分，包括摘要图构建，基于注意力机制的消息传递，基于注意力机制的图读出过程和模型预测层。VPALG 首先构建了三种形式的摘要图用于将序列文本结构化，接着利用注意力消息传递衡量不同邻居消息的重要性，同时利用节点级别的注意力衡量论文摘要文本中不同单词的重要性。在最后模型预测过程中还采用了自训练的方法获得训练数据高置信度的伪标签以进一步提升模型的表现。

第 4 章 论文发表场所预测算法的实验结果与分析

4.1 实验说明

4.1.1 实验数据集

为了验证提出的模型 VPALG 的有效性，本文收集了三个数据集进行评估，分别名为 PubMed、Computer Science Journals（CSJ）、Computer Science Conferences（CSC）。这些数据集的统计数据被展示在表 4.1 中。数据集的详细介绍如下：

（1）PubMed：该数据集是从 PubMed[1]收集的，PubMed 是一个大型生物医学出版物数字图书馆。本文采用了 2007-2016 年发表在 PubMed 上的论文，剔除了发表期刊标有“Predecessor”、“No New Content”和“Now Select”的异常论文。PubMed 数据集是不平衡的，其中摘要数量最多和最少的期刊分别包含大约 10,000 和 100 篇摘要。

（2）CSJ：该数据集收集了 2007-2016 年 JMLR、TOIS、TPAMI 等 28 种计算机科学领域领先期刊的论文摘要。该数据集相对均衡。

（3）CSC：该数据集收集了 2007 年至 2016 年 ICML、KDD 和 NeurIPS 等计算机科学领域 37 个主要会议的论文摘要集合。该数据集相对均衡。

表 4 .1 基准数据集的统计数据

数据集	摘要数目	类别数目	单词平均数目	单词总数
PubMed	837782	1130	203.27	323194
CSJ	72800	28	156.94	37255
CSC	73744	37	153.60	37747

4.1.2 基线模型

由于本文所研究的问题与文档排序和文本分类都有关系，因此在实验中，

[1] https://www.ncbi.nlm.nih.gov/pmc/

本文选择了三种基线方法进行比较，包括文档排序方法、传统的论文发表场地预测方法和基于深度学习的文本分类方法。基线的细节描述如下。

文档排序方法：对于文档排序方法，被用来比较的基线模型将输入摘要视为查询词以从训练的数据找到相关文档，其标签用作查询标签。（1）QL[62]：查询似然（query likelihood，QL）模型是基于 Dirichlet 平滑的良好性能语言模型之一；（2）BM25[63]：一种基于查询词的对一组文档进行排序的经典的检索模型。

传统的场所推荐方法：（1）VRS [44]：一种考虑不同论文发表场所写作风格的方法；（2）TMVR[9] ：一种基于无监督学习主题建模的主题匹配方法；（3）PRS[45]：一种使用卡方统计和 softmax 回归预测论文发表场所的分类方法。

基于深度学习的方法：（1）fastText[64]：一种通过平均 n-gram 嵌入提取文本特征的方法。在这里，本文使用分层损失并将n设置为 3；（2）ELMo[65]：一个通过 Bi-LSTM 训练的深度上下文的词嵌入模型，最终的词向量是不同网络层的词向量的线性组合。（3）BERT [39]：一个基于双向 Transformers 架构的预训练语言模型，采用预训练-微调范式适应特定下游任务。在这里，本文使用 bert-base 模型变体。（4）Bi-LSTM[66]：LSTM 的双向版本，同时利用上下文信息以提取文本中有用的特征。（5）Pubmender[46]：一个利用卷积神经网络提取摘要特征的论文发表场所推荐模型。（6）MPAD[67]：一个基于层次文本图的挖掘不同粒度文本信息的深度图网络模型。对于所有提及的基线方法，本文参考原始论文仔细调整参数，最终报告最佳结果。此外，这些基线是在没有自训练的情况下进行训练。对于所有需要词嵌入的方法，本文统一使用预训练的 GloVe 词嵌入。

4.1.3 评估指标

为了评估各种模型的预测性能，本文采用了三种流行的评估指标，包括正确率（Accuracy@K，Acc@K）、平均倒数排名（Mean Recirocal Rank ，MRR）和 F1 分数（F1-score）。此外，本文还采用一个指标优化精度（Optimized

Precision，OP）来衡量模型在不平衡的 PubMed 数据集中的性能。让$\{(\hat{\mathbf{x}}_i,\hat{\mathbf{y}}_i)\}_{i=1}^{\hat{n}}$表示测试数据集，相关指标的详细描述如下：

Acc@K：该指标衡量测试摘要的真实发表场所被 top-K 预测结果覆盖的比例。它可以被形式化表述为：

$$\text{Acc@K}=\frac{\sum_{i=1}^{\hat{n}}\mathbb{I}\left(\hat{\mathbf{y}}_i\in\Delta_K(i)\right)}{\hat{n}} \quad (4.1)$$

其中$\mathbb{I}(\cdot)$是指示函数；$\Delta_K(i)$表示测试摘要$\hat{\mathbf{x}}_i$的 top-K 预测结果。在实验中，K 的取值为 1、3、5。

MRR：该指标衡量了预测结果中真实论文发表场所的平均排名。它可以被形式化为：

$$\text{MRR}=\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}\frac{1}{\text{Rank}(i)} \quad (4.2)$$

其中 Rank(i)表示$\hat{\mathbf{y}}_i$在测试论文摘要x_i的预测结果中的排名。

F1-score：这是一种被广泛使用的分类指标，可以平衡预测结果的精度和召回率。一般来说，该指标可以被表述为：

$$\text{Precision}=\frac{\sum_{i=1}^{l}TP_i}{\sum_{i=1}^{l}(TP_i+FP_i)},\quad \text{Recall}=\frac{\sum_{i=1}^{l}TP_i}{\sum_{i=1}^{l}(TP_i+FN_i)} \quad (4.3)$$

$$\text{F1-score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4.4)$$

其中TP、FP、TN和FN分别表示真阳性、假阳性、真阴性和假阴性的数量。

OP：为了衡量模型在不平衡数据集中的表现，受之前工作的启发[68]，本文采用了一种新的指标，它同时考虑了所有类的正确率和召回率。该指标可以被表述为：

$$\text{OP}=\text{Acc@1}-\frac{\sum_{i=1}^{l-1}\sum_{j=i+1}^{l}|\text{Recall}_i-\text{Recall}_j|}{l\sum_{k=1}^{l}\text{Recall}_k} \quad (4.5)$$

该指标首先计算每个成对类别的绝对召回距离，然后对其进行归一化，从

而衡量模型预测所有类别的准确率和召回率最高分的能力。本文只报告 PubMed 数据集中 OP 的结果，因为其他数据集相对平衡。

对于所有四个指标，越高的分数代表越好的表现。

4.1.4 实验设置

实验环境：该工作的实验环境汇总如下。（1）操作系统：Ubuntu 16.04.6；（2）显卡：Nvidia Titan XP 12G（3）内存：512G（4）Python：3.7.2（5）PyTorch：1.6.0

实现细节：对于每个数据集，本文随机抽取 10% 的摘要作为验证集，并对剩余的摘要进行 5 折交叉验证实验。本文将每个摘要的最大长度设置为 300，因为观察到大多数摘要包含的单词少于 300 个。对包含超过 300 个单词的摘要采用截断操作。另外，将出现次数少于 10 次的生词替换为“UNK”。 VPALG 的其他实现细节明确如下：层数T，滑动窗口大小S和嵌入维度的大小分别设置为 2、2 和 200；批量大小设置为 128； Adam[69]方法用于稳定优化，其中两个目标函数的学习率都设置为 10−3，损失权重设置为 10−4。对于预测层，在线性层之后应用随机丢弃技术，根据经验将其比率设置为 0.5。此外，在自训练阶段执行提前停止策略，如果当前预测与连续 5 次更新的目标分布一致，则优化停止。

4.2 实验结果

4.2.1 模型表现

本文将所提出的三种方法 VPALG$_{bi}$、VPALG$_{ub}$ 和 VPALG$_{uf}$ 在三个数据集上与选择的基线方法进行了比较。 Acc@K、MRR、F1-score 和 OP 的结果如表 4.1、表 4.2 和表 4.3 所示。从表中的结果本文发现所提出的方法在所有三个数据集中的 Acc@K、MRR、F1-score 和 OP 方面均获得最高分，这表明 VPALG 模型在论文发表场所预测方面的有效性。特别是从指标 OP 可以清楚地看出，所提出的模型可以更好地处理不平衡的数据。本文认为 VPALG 实现最先进性

能的原因有两个。首先是提出的模型能同时捕获全局和局部语义信息，以及论文摘要的网络拓扑结构。与其他模型相比，该模型能够有效地提取关于不同发表场所的精确特征。第二个是在消息传递和图读出阶段引入了双重注意力机制，使模型能够聚合更多邻居的关键信息，让词节点学习到内容丰富的表示，并选择更重要的词来获得图嵌入。因为其他模型忽略了论文摘要的拓扑结构信息，并且未使用注意力机制获得更精准的摘要表征，所以取得了次优的结果。

在三个基于 VPALG 的模型中，本文发现 VPALG$_{ub}$ 和 VPALG$_{uf}$ 总是比 VPALG$_{bi}$ 表现更好。这种现象是预料之中的，虽然构建双向图可以为更新节点嵌入带来很多信息，但也容易出现节点过度平滑问题，即每个节点的特征非常相似以至于无法区分。然而，单向图的构造可以显著缓解这个问题，实验结果也验证了所提出的假设。同时本文还发现了其他有趣的实验现象。与传统的发表场所预测方法相比，文档排名方法（即 QL 和 BM25）具有竞争力。然而，基于深度学习的方法明显优于其他类别的基线。这表明深度学习架构在学习表征方面的巨大优越性。此外，基于图神经网络的模型（即 MPAD 和 VPALG）表现最好，证明了图神经网络在学习文本特征方面的表达能力。

表 4.1 各种模型在 PubMed 数据集中关于 Acc@K、 MRR、F1-score 和 OP 的实验结果，其中 K = 1, 3, 5。

Model	PubMed					
	Acc@K			MRR	F1-score	OP
	1	3	5			
QL	0.273	0.369	0.412	0.255	0.239	0.242
BM25	0.305	0.472	0.526	0.381	0.296	0.275
VRS	0.123	0.225	0.293	0.112	0.107	0.131
TMVR	0.193	0.266	0.332	0.153	0.221	0.193
PRS	0.292	0.454	0.515	0.333	0.320	0.261
fastText	0.399	0.586	0.664	0.413	0.407	0.342
ELMo	0.463	0.642	0.745	0.513	0.456	0.393
BERT	0.474	0.668	0.750	0.523	0.477	0.421
Bi-LSTM	0.511	0.700	0.757	0.539	0.499	0.429
Pubmender	0.497	0.705	0.780	0.579	0.501	0.426
MPAD	0.521	0.714	0.783	0.643	0.531	0.452
VPALG$_{bi}$	0.557	0.728	0.787	0.663	0.544	0.505
VPALG$_{ub}$	**0.579**	0.746	**0.808**	**0.681**	0.567	**0.533**
VPALG$_{uf}$	0.578	**0.748**	0.807	0.679	**0.574**	0.521