文章编号：2095-2163(2020)06-0294-06 中图分类号：TP391 文献标志码：A

基于图分类的中文长文本匹配算法

郭佳乐[1]，卜 巍[2]，邬向前[1]
（1 哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001；2 哈尔滨工业大学 媒体技术与艺术学院，哈尔滨 150001）

摘 要：判断一对文章之间的关系是一项很重要的自然语言处理任务，在新闻系统和搜索引擎等实际服务中有着广泛的应用。然而，相比在信息检索场景中去匹配一对句子或者匹配一个查询-文档对而言，长文章通常具有丰富的语义信息和复杂的逻辑结构，这也使得长文章之间的匹配成为一个相对独立且很有挑战的任务。本文围绕长文章匹配的难点，提出了基于图分类框架的长文本匹配算法，通过将长文本匹配任务等价的转化为图分类任务，使用图表示学习的范式来求解，从而获得长文本匹配的结果。算法模型主要包括基于图表示学习来实现对长文本的建模，基于图注意力神经网络的图节点特征提取，以及图池化等步骤。在两个大型公开数据集上的训练和测试实验结果表明：本文提出的算法可以实现高质量的文本匹配，同时各项评价指标均达到了目前最先进的结果。

关键词：自然语言处理；文本匹配；图注意力神经网络；图池化

Chinese document matching based on graph classification

GUO Jiale[1], BU Wei[2], WU Xiangqian[1]
（1 School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China;
2 School of Media Technology and Art, Harbin Institute of Technology, Harbin 150001, China）

【Abstract】 Identifying the relationship between two documents is an important task in natural language processing area, which has been popularly applied to Internet services such as news recommendation systems or search engines. However, compared with sentence matching or query-doc matching in information retrieval, document matching is more challenging and independent since documents always contained rich semantic information and complicated structure. This paper tries to focus on the challenges to document matching and propose a matching pipeline based on graph classification by transferring the matching task equally to a graph classification problem. The pipeline mainly includes modeling documents pairs based on graph representation learning, graph attention neural network based node feature extraction and graph pooling. Two public datasets are used to verify the performance of our proposed methods, and the results achieve the state-of-the-art.

【Key words】 Natural Language Processing; Text Matching; Graph Attention Network; Graph Pooling

0 引 言

近年来随着移动互联网的蓬勃发展，涌现了大量基于内容分发服务的 App 和相关自媒体平台，它们在日常生活中扮演着越来越重要的角色，改变了人们获取信息的方式和途径。当用户在浏览感兴趣的内容时，这些平台常常会主动为用户推荐相同或者相近话题的其他文章。而如何判断文章之间的主要内容是否关于相同或相近的主题，即判断一对文章之间的关系，其实可以归结为自然语言处理领域中的文本语义匹配任务。

文本匹配任务本质上是判断源文本与目标文本之间的语义相似度（比如：查询-文档的匹配、问题-答案的匹配等），而如何正确地建模文本中蕴含的语义信息对于实现更高质量的文本匹配结果十分重要。目前，在实际建模过程中主要有两个难点：一是词汇和短语本身是有歧义的，存在广泛的指代和省略等问题；二是待匹配的文本篇幅很长，文本中词汇、短语、句子本身的语义受到复杂文章结构很大的影响。

本文针对互联网上的文本信息，譬如新闻文章都具有较长的篇幅以及一定的行文逻辑结构，设计出一种高精度的基于图分类的中文长文本匹配模型，并在 CNSE 和 CNSS 数据集上进行训练和测试。本文的主要贡献总结如下：

（1）提出了基于图分类的长文本匹配算法。该算法将输入的文本对转化为图结构，利用基于注意

基金项目：国家自然科学基金（61672194）；国家重点研究与发展计划（2018YFC0832304）；中国黑龙江省杰出青年科学基金（JC2018021）；国家机器人与系统国家重点实验室项目（SKLRS-2019-KF-14）；中兴通讯产学研合作论坛合作项目。

作者简介：郭佳乐（1994-），男，硕士研究生，主要研究方向：自然语言处理、深度学习；卜 巍（1977-），女，博士，副教授，主要研究方向：数字媒体技术、数字图像处理、医学图像分析等；邬向前（1973-），男，博士，教授，博士生导师，主要研究方向：数字图像处理、模式识别、生物特征识别等。

收稿日期：2020-04-24

力机制的图神经网络对节点特征进行抽取，通过基于多层感知机的图分类模块融合全图信息的特征表示向量，进行图分类计算，从而得到长文本匹配的结果，完成长文本匹配任务。

(2)提出了一种基于图注意力机制的图池化算法，增强图节点融合过程中最大化图的可辨别性，有效的提升了图的表示效果，从而获得了更鲁棒且优异的长文本匹配结果。

(3)针对网络模型的输入，设计了多尺度卷积神经网络模块对节点编码，从而提升网络模型输入特征质量，获得更丰富且鲁棒的节点语义表示。同时在模型中引入更多的非线性，增强了模型的拟合能力。

(4)本文提出的中文长文本匹配算法，在 CNSE 和 CNSS 两个公开数据集上超越了先前算法模型的结果，各项指标均达到了目前最好的匹配结果。

1　相关工作

1.1　文本匹配

为了衡量两篇文本之间的相似度，深度学习模型之前主流的模型有向量空间模型(vector space model，VSM)、隐含语义分析(Latent Semantic Analysis，LSA)模型以及引入例如知识库等外部的语义知识资源来辅助计算，但是传统方法和模型仍旧受限于离散的单词表示。基于深度学习的文本语义匹配模型通常将文本编码为分布式特征表示，通过用于度量学习的孪生结构(siamese structure)来学习文本之间的相似度信息[1]。其中有大量的研究工作聚焦于使用深度神经网络来对文本进行编码，但是过去的这些研究主要都集中于短文本之间的匹配，当随着待匹配文本篇幅增加，基于 RNN 的模型在超长序列中传递信息会丢失重要的语义信息，而基于 CNN 的模型也无法充分地表示结构更复杂的长文本的语义信息。

此外，大多数文本语义匹配的工作都忽略了长文本所具有的结构信息，而这种结构信息对于语义匹配是十分重要的，应当加以有效地利用。Liu 等人提出层次性结构(hierarchical structure)来聚焦于句子级别的层次信息。为了更充分地利用文本中的结构信息，Jiang 等人提出基于孪生多深度注意力机制的层次性循环神经网络(SMASH RNN)模型，来处理长文本匹配的任务[2]。Liu 等人基于分治思想将待匹配的长文本转化到多个关键词的匹配，并融合结果，形成整体的匹配得分，同时提出了两个中文长文本匹配数据集[3]。

1.2　图表示学习

传统的图结构用不同的符号命名节点，用邻接矩阵来存储节点之间的关系。但这种表示方法节点之间没有语义关系，且表示形式稀疏，很难应用于深度学习的模型当中。因此后续有很多工作聚焦于将节点特征低维稠密化，比如 Deepwalk 模型，通过随机游走获得当前节点的上下文信息[4]，之后针对大规模网络计算，LINE 模型被提出[5]。Node2vec 模型被提出改进的随机游走策略，可以同时考虑到局部和宏观的信息，并且具有很高的适应性[6]。

1.3　图池化

目前针对图池化算法相关的工作大致可以划分为基于节点选择的图池化算法和基于节点聚类的图池化算法两类。基于节点选择的模型比如 gPool 算法，将节点 Embedding 隐射到一维空间中，根据值的大小选择其中 top k 个节点，再进行图卷积的计算，自适应地选择图全部节点的子集来形成一个新的小图[7]。还有研究者提出 SortPooling 算法通过 Weisfeiler-Lehman(WL)算法 可以对节点进行着色，而节点的颜色可以定义节点之间的次序，通过 1-D 卷积的方法进行卷积运算，从而得到全图的表示，用于图分类任务[8]。由于基于节点选择的图池化算法一定程度上忽略了图的层级结构信息，而这对提升最终图表示的可辨别性有一定的辅助作用，因此有研究者提出了一种端到端的可微可微图池化模块 DiffPool[9]。为了实现图中某一个节点分配到哪一个簇，应该与其他节点的簇分配相互约束，有研究者提出了 StructPool 来高效地学习高层次的图表示，使用条件随机场(conditional random fields，CRF)显示地捕捉了图中不同节点之间的高阶结构关系[10]。

2　基于图分类的中文长文本匹配算法

2.1　算法概况

本文提出基于图分类的中文长文本匹配算法，即将长文本匹配任务等价地转化为图分类的任务，在图分类任务的模式下求解问题，得到原任务的结果。如图 1 所示，首先将待匹配的一对长文本转化为图的结构，然后通过图表示学习来提取节点的特征，最后通过融合全图节点的表示来获得图的表示，并对图的表示进行图分类，获得的图分类结果等价于长文本匹配的结果。

2.2　文本对的图结构表示算法

给定一个文本对包含文章 D_A 和 D_B，对两篇文章分别通过 TextRank 算法来抽取出文章中的命名

实体和关键词作为图的顶点并合并,把文章中的所有句子通过与每个关键词计算 TF-IDF 相关度来对其划分,使得每个句子都只隶属于一个顶点。这样每个顶点均包含了一个句子子集 $S(v)$,其中 $S(v) = \{S_A(v), S_B(v)\}$。得到的文本对图结构表示的可视化结果,如图 2 所示。

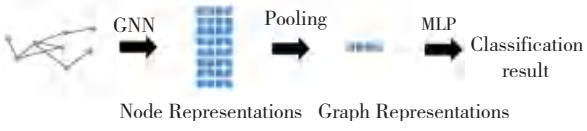

图 1 基于图分类的长文本匹配算法框架

Fig. 1 The pipeline of the proposed document matching algorithm based on graph classification

然后把关键词所包含的句子集合中的文本内容使用深度神经网络(Deep Neural Network,DNN)和词项相似度(Term Similarity,TS)两种编码模式向量化。其中 DNN 编码器使用卷积神经网络来对文本进行编码,即把 $S(v) = \{S_A(v), S_B(v)\}$ 的 $S_A(v)$、$S_B(v)$ 分别输入到 *CNN* 模型中得到向量 $c_A(v)$、$c_B(v)$。对其计算逐元素差的绝对值和逐元素乘,最后把两个向量进行拼接,即固定长度的输出向量 $m_{AB}(v)$;对于 *TS* 编码器,每个节点 v 用 5 种常见的词项相似度度量方式来计算 $S_A(v)$ 和 $S_B(v)$ 的相似度,包括 *TF - IDF* 余弦相似度、*TF* 余弦相似度、*BM*25 余弦相似度、*Jaccard* 相似度和 *Ochiai* 相似度。通过把 5 种相似度向量进行拼接,得到针对每个节点的固定长度向量 $m'_{AB}(v)$。$m_{AB}(v)$ 与 $m'_{AB}(v)$ 拼接,得到节点 v 的语义向量。

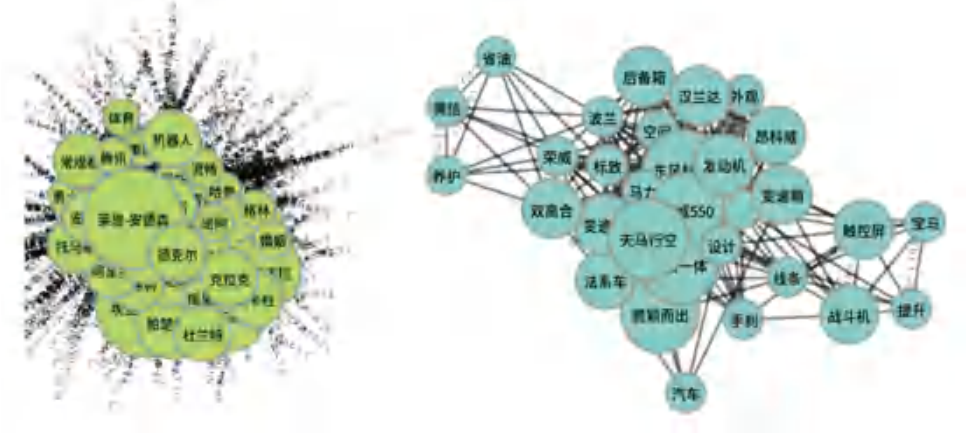

图 2 文本对图结构表示的可视化结果

Fig. 2 Visualization of the graph structure of the documents pairs

2.3 基于注意力机制图神经网络的节点特征抽取

近年来很多图表示学习的工作使用常规的图卷积神经网络,获得了很好的结果,但是其在聚合节点的一阶邻居信息时,对于每个邻居的信息给予的权重是相同的,并没有考虑该信息的价值贡献大小。而在信息聚合时,对于不同邻居的信息根据其价值来给以不同大小的权重,实现了选择性的信息聚合,从而有利于更有效地对节点特征进行变换。

此外,朴素图卷积需要在建图时完成邻接矩阵的构建,而在实际应用当中这一步需要做大量的预处理计算,同时在进行图卷积运算时,将邻接矩阵加载到内存中或者 GPU 显存中时非常耗时,不利于算法在工业界大规模图上的实际应用,因此本文使用基于自注意力机制的图节点特征抽取算法(Graph Attention Network,GAT),来代替朴素图卷积方法。这样可以省略文本建图时必须构建邻接矩阵的过程,同时在训练测试加载数据时大大提升处理效率。基于注意力机制图神经网络模型结构如图 3 所示。自注意力机制在计算相似度得分时,使用公式(1)计算:

$$\alpha_{ij} = \frac{\exp\left(\mathrm{LeakyReLU}(a^T[Wh_i \parallel Wh_j])\right)}{\sum_{k\in N_i}\exp\left(\mathrm{LeakyReLU}(a^T[Wh_i \parallel Wh_k])\right)}. \tag{1}$$

其中,a 和 W 是训练参数,$[a \parallel b]$ 代表拼接(concatenate)操作。α_{ij} 代表节点 i 与节点 j 之间的相似度,h_i 代表 i 节点的属性特征。LeakyReLU 为激活函数,增强模型的非线性表达能力。

此外,为了增加模型的容量以及模型训练过程的稳定性,引入了多头机制(Multi - head mechanism)。多头机制可以将特征向量映射到不同的子空间,通过聚合多个子空间的映射结果,来得到更好的特征表示。通过堆叠多层的 GAT 层,可以抽取到更有效的节点特征表示,在相邻的 GAT 层之间同样引入非线性激活函数 LeakyReLU 来增加模型的非线性,使得训练得到的模型更鲁棒。

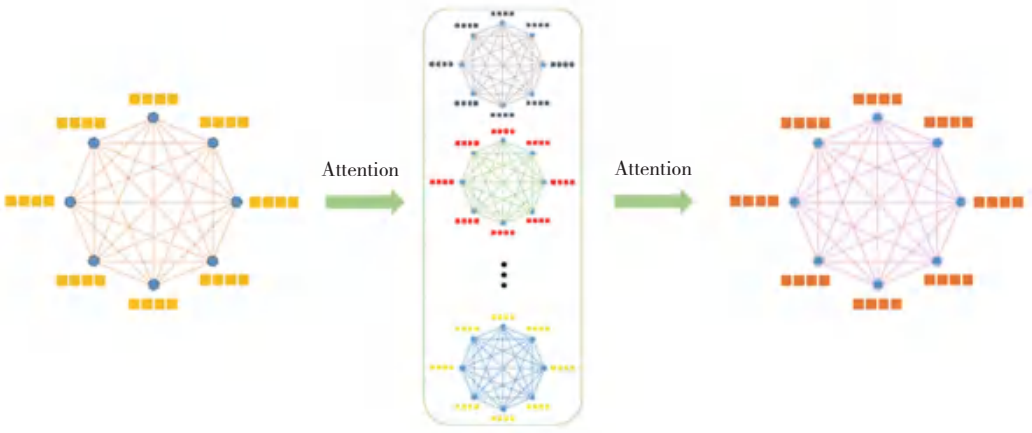

图 3　基于注意力机制图神经网络

Fig. 3　Self attention based graph neural network model

2.4　基于图注意力机制的图池化算法

学习图数据的高层语义表示对于图分类任务十分重要，除了引入图卷积操作来对图数据进行处理，类比图像文本领域常见的池化操作，如何对图结构数据引入图池化操作也是一个很重要的研究方向，目前最普遍的融合算法策略是对所有节点的特征向量形成的矩阵在节点数量维度上做求和(Sum Pool)或者平均池化(Mean Pool)/最大池化(Max Pool)等操作，从而得到一个与节点特征维度相同的向量，并用这个向量作为全图的最终表示，送入到图分类模块进行图的分类。但是，这些朴素的融合方式都可能丢失掉重要特征的信息，同时在融合时没有考虑到图节点之间的语义信息交互的关系，以及完全丢失了图所具有的拓扑结构对最终图表示向量的贡献，因此获得的全图表示的可分辨性较弱，在一定程度上限制了图分类模型性能进一步提升的潜力。

本文为了在省略邻接矩阵的情况下依然能够得到鲁棒的结果，提出了基于图注意力机制的图池化算法(Graph Attention Pooling, GATPool)来完成图节点的融合过程，基于图注意力机制的图池化模型算法模型如图 4 所示。算法把输入的节点特征矩阵(N 个节点)通过一层图注意力网络来得到一个 $N \times N$ 的相似度矩阵(attention matrix) $\boldsymbol{A}$ 来衡量图中任意两个节点之间的相似度，如公式(2)所示，同时相似度矩阵可以作为一种既融合了节点的特征表示又显示地建模了图拓扑结构的模块。然后使用公式(3)对相似度矩阵按列进行 softmax 归一化后按行求和，公式(4)依据每个节点与其他节点的交互计算，得到了每个节点在全图中的重要性得分 $\boldsymbol{Z}$。通过选择值最大的 K 个节点作为图的代表 $\boldsymbol{X}'$，通过公式(7)完成节点的选择后，通过平均池化(Mean Pooling)得到全图的特征表示 $\boldsymbol{X}_{out}$。

$$\boldsymbol{A} = \mathrm{GAT}(\boldsymbol{X}), \tag{2}$$

$$\boldsymbol{Z} = Sum(Softmax(\boldsymbol{A})), \tag{3}$$

$$\mathrm{idx} = \mathrm{top-rank}(\boldsymbol{Z}, \lceil KN \rceil), \tag{4}$$

$$\boldsymbol{Z}_{mask} = \boldsymbol{Z}_{\mathrm{idx}}, \tag{5}$$

$$\boldsymbol{X}' = \boldsymbol{X}_{\mathrm{idx}}, \tag{6}$$

$$\boldsymbol{X}_{\mathrm{out}} = Mean(\boldsymbol{X}' \odot \boldsymbol{Z}_{mask}). \tag{7}$$

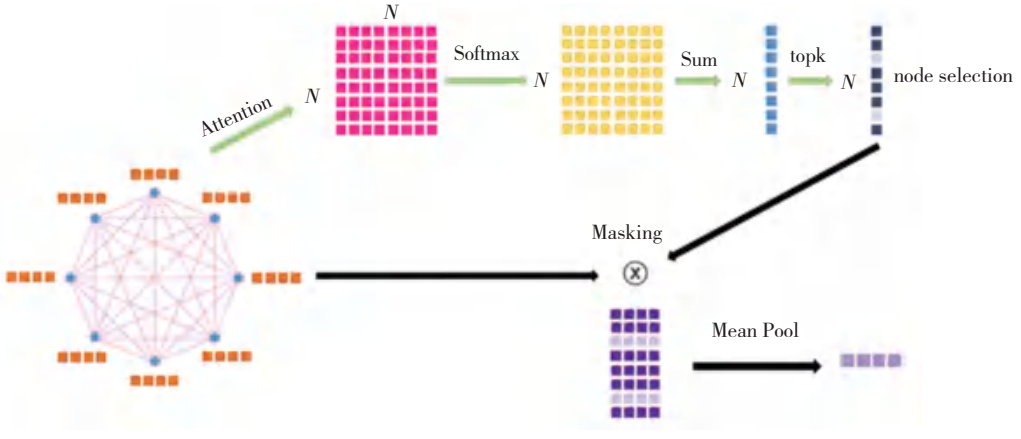

图 4　基于图注意力机制的图池化模型

Fig. 4　Graph Attention Pool, GATPool

2.5 多尺度卷积神经网络模块

为了充分地挖掘和利用输入信息,本文提出使用多尺度卷积神经网络模块(记为 Inception)来对每个节点所包含的文本信息 S 进行编码,从而得到节点的表示向量 $\boldsymbol{X}$, 如公式(8)。通过组合三种大小的卷积核形成层次性的模型结构,可以有效地捕获输入文本中在不同语义层次和空间的语义表示,通过最大池化(max pooling)操作对特征进行筛选,最后对不同卷积核的输出结果进行拼接(concatenate),得到输入文本的编码向量 $\boldsymbol{X}$。

$$\boldsymbol{X} = Inception(\boldsymbol{S}). \tag{8}$$

如图 5 所示,本文设计的多尺度卷积神经网络使用了大小分别为 1、2、3 三种尺寸的卷积核,最左侧的 $K=1$ 的卷积核数量为 32,中间和右侧 $K=1$ 的卷积核数量为 16, $K=2$ 和 3 的卷积核数量均为 32。尺寸为 1 的卷积核主要对输入的特征维度进行变换,使得模型可以在不降低感受野大小的前提下减少模型的参数量,降低模型过拟合的风险。尺寸为 2、3 的卷积核用来捕获文本的 2-gram、3-gram 特征,通过这种局部信息和模式的提取,来增强文本的表示效果。

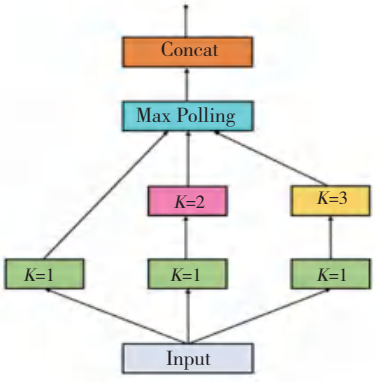

图 5 多尺度卷积神经网络模块

Fig. 5 Multi-scale CNN module

2.6 图分类模块和损失函数

文本匹配任务是一个二分类任务,通过将文本对表示为图结构,则将文本匹配任务等价转化为图的二分类任务。在融合图节点特征得到整个图的向量表示后,通过多层感知机(Multi-Layer Perceptron, MLP)可以实现分类过程。多层感知器的结构如图 6 所示。

两层线性层的隐藏单元数分别为 32 和 16。将 MLP 的输出结果通过与二分类标签进行交叉熵(Cross Entropy)计算作为损失函数。

模型网络的目标函数使用二元交叉熵损失函数(Binary Cross Entropy Loss,记为 BCELoss),计算公式(9)为:

$$Loss = -w_n[y_n \cdot \log x_n + (1-y_n) \cdot \log(1-x_n)]. \tag{9}$$

其中, y_n 是标签值, x_n 是模型网络输出的值, w_n 是该类别的权重值。

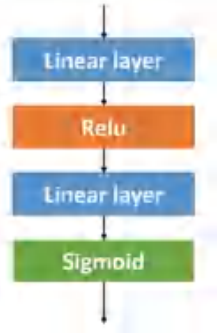

图 6 多层感知机(MLP)结构图

Fig. 6 The scheme of the Multi-layer perceptron(MLP) module

3 实验

3.1 数据集介绍

模型分别在 CNSE(Chinese News Same Event dataset)和 CNSS(Chinese News Same Story dataset)两个数据集上进行训练和测试。数据集的划分比例均保持为训练集:验证集:测试集=6∶2∶2,同时确保了不同集合之间不存在数据泄露,具体划分详情如表 1 所示。

表 1 两个数据集的划分详情

Tab. 1 Description of two evaluation datasets

数据集	正样本	负样本	训练集	验证集	测试集
CNSE	12 865	16 198	17 438	5 813	5 812
CNSS	16 887	16 616	20 102	6 701	6 700

CNSE 数据集中的文章来源于主流中文新闻平台上的长文报道,包括了开放领域丰富多彩的新闻话题。一共包含 29 063 个新闻文章对,由人工根据一对新闻文章是否报道同一件新闻事件来进行标注。

CNSS 数据集中的文章来源于主流中文新闻平台上的长文报道,包括了开放领域丰富多彩的新闻话题。一共包含 33 503 个新闻文章对,由人工根据一对新闻文章是否报道同一系列新闻热点来进行标注。

两个数据集中的所有文章的平均词数是 734,最大词数是 21 791。同时在构造负样本对时兼顾了两篇文章 TF-IDF 的相似度高于一定的阈值,增加了负样本对本身的质量和模型识别判断的难度。

3.2　实验评价指标

实验的评测标准采用通用的二分类评价标准，分别为准确率（*Accuracy*）和 *F*1 值（*F*1 *scores*），分别由公式（10）和（11）进行计算。

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{10}$$

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}. \tag{11}$$

3.3　实验结果与分析

由表 2 中的实验结果可以看出，与先前的 7 种有代表性的文本匹配算法相比，本文提出的算法在 CNSE 和 CNSS 两个数据集上的准确率和 F1 值等各项指标均达到了目前最好的结果。充分说明本文提出的基于图分类的中文长文本匹配算法通过结合注意力机制图神经网络、图注意力池化算法、多尺度卷积神经网络模块等模型可以有效地提升文本匹配任务中的表现。

表 2　基于图分类的中文长文本匹配结果

Tab. 2　Results on low-level feature enhanced model

数据集	CNSE		CNSS	
对比方法	Acc	F1	Acc	F1
C-DSSM[1]	60.17	48.57	52.96	56.75
MatchPyramid[1]	66.36	54.01	62.52	64.56
BM25[1]	69.63	66.60	67.77	70.40
LDA[1]	63.81	62.44	62.98	69.11
SimNet[1]	71.05	69.26	70.78	74.50
BERT fine-tuning[1]	81.30	79.20	86.64	87.08
CIG[1]	84.64	82.75	89.77	90.07
本方法	**86.07**	**84.26**	**90.84**	**91.03**

表 3 中得到了本文提出的模型消融实验的结果，可以看出本文提出的算法需要各模块之间紧密有效的配合，去除其中的某一些模块会不可避免的导致模型在 CNSE 和 CNSS 两个数据集上的准确率和 F1 值等各项指标不同程度的下降。其中，基于自注意力机制的图神经网络可以有效地对节点特征进行抽取，基于图注意力的图池化算法有利于获得更具可辨别性的图全局特征表示，而多尺度卷积神经网络模块可以为网络模型的输入提供语义丰富且鲁棒的节点特征表示。

表 3　模型消融实验结果

Tab. 3　Ablation study of the model

数据集	CNSE		CNSS	
对比方法	Acc	F1	Acc	F1
本方法	**86.07**	**84.26**	**90.84**	**91.03**
-图池化	85.70	83.94	90.59	90.98
-多尺度 CNN	85.83	84.03	90.56	90.93
-图池化-多尺度 CNN	85.61	83.93	90.47	90.75

4　结束语

本文针对中文长文本匹配任务提出了基于图分类框架的长文本匹配算法，通过将长文本匹配任务等价的转化为图分类任务，使用图表示学习的范式来求解，从而获得长文本匹配的结果。同时，从增强图节点特征表示抽取建模，强化图节点融合池化过程，保持图可辨别性，增强模型网络的输入特征、表示等方面来提升基于图表示学习模型的文本匹配算法的性能表现。通过本文提出的各模块之间紧密有效地配合，本文提出的中文长文本匹配算法，在两个公开数据集上进行了充分的实验验证，实验结果表明：各项指标均达到了目前最好的结果，证明了本文方法的有效性和优越性。

参考文献

[1] BROMLEY J, GUYON I, LECUN Y, et al. Signature verification using a" siamese" time delay neural network[J]//Advances in neural information processing systems. 1994: 737-744.

[2] JIANG J Y, ZHANG M, LI C, et al. Semantic text matching for long-form documents[C]//The World Wide Web Conference. 2019: 795-806.

[3] LIU B, NIU D, WEI H, et al. Matching article pairs with graphical decomposition and convolutions[J]. arXiv preprint arXiv:1802.07459, 2018.

[4] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 701-710.

[5] TANG J, QU M, WANG M, et al. Line: Large-scale information network embedding[C]//Proceedings of the 24th international conference on world wide web. 2015: 1067-1077.

[6] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016: 855-864.

[7] GAO H, JI S. Graph u-nets[J]. arXiv preprint arXiv: 1905.05178, 2019.

[8] ZHANG M, CUI Z, NEUMANN M, et al. An end-to-end deep learning architecture for graph classification[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[9] YING Z, YOU J, MORRIS C, et al. Hierarchical graph representation learning with differentiable pooling[C]//Advances in neural information processing systems. 2018: 4800-4810.

[10] YUAN H, JI S. StructPool: Structured graph pooling via conditional random fields[C]//International Conference on Learning Representations. 2019.