E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2023)02-0510-12

论文引用格式:Li Y, Wu Q F, Liu J T and Zou J L. 2023. Leading weight-driven re-position relation network for figure question answering. Journal of Image and Graphics,28(02):0510-0521(黎颖,吴清锋,刘佳桐,邹嘉龙. 2023. 引导性权重驱动的图表问答重定位关系网络. 中国图象图形学报,28(02):0510-0521)[DOI:10.11834/jig.211026]

引导性权重驱动的图表问答重定位关系网络

黎颖,吴清锋*,刘佳桐,邹嘉龙
厦门大学信息学院,厦门　361005

摘　要:**目的**　图表问答是计算机视觉多模态学习的一项重要研究任务,传统关系网络(relation network,RN)模型简单的两两配对方法可以包含所有像素之间的关系,因此取得了不错的结果,但此方法不仅包含冗余信息,而且平方式增长的关系配对的特征数量会给后续的推理网络在计算量和参数量上带来很大的负担。针对这个问题,提出了一种基于融合语义特征提取的引导性权重驱动的重定位关系网络模型来改善不足。**方法**　首先通过融合场景任务的低级和高级图像特征来提取更丰富的统计图语义信息,同时提出了一种基于注意力机制的文本编码器,实现融合语义的特征提取,然后对引导性权重进行排序进一步重构图像的位置,从而构建了重定位的关系网络模型。**结果**　在2个数据集上进行实验比较,在FigureQA(an annotated figure dataset for visual reasoning)数据集中,相较于IMG+QUES(image + questions)、RN和ARN(appearance and relation networks),本文方法的整体准确率分别提升了26.4%,8.1%,0.46%,在单一验证集上,相较于LEAF-Net(locate,encode and attend for figure network)和FigureNet,本文方法的准确率提升了2.3%,2.0%;在DVQA(understanding data visualization via question answering)数据集上,对于不使用OCR(optical character recognition)方法,相较于SANDY(san with dynamic encoding model)、ARN和RN,整体准确率分别提升了8.6%,0.12%,2.13%;对于有Oracle版本,相较于SANDY、LEAF-Net和RN,整体准确率分别提升了23.3%,7.09%,4.8%。**结论**　本文算法围绕图表问答任务,在DVQA和FigureQA两个开源数据集上分别提升了准确率。

关键词:计算机视觉;图表问答(FQA);多模态融合;注意力机制;关系网络(RN);深度学习

Leading weight-driven re-position relation network for figure question answering

Li Ying, Wu Qingfeng*, Liu Jiatong, Zou Jialong
School of Informatics, Xiamen University, Xiamen　361005, China

Abstract: **Objective** Figure-based question and answer (Q&A) is focused on learning the basic information representation of data mining in real scenes and provide the basis of judgment for reasoning in terms of the text information of the joint questions. It is widely used for multi-modal learning tasks. Existing methods can be segmented into two categories in common: 1) model tasks are based on neural network framework algorithms directly. The statistical graph is processed by the

收稿日期:2021-11-05;**修回日期**:2021-12-17;**预印本日期**:2021-12-24

*通信作者:吴清锋　qfwu@xmu.edu.cn

基金项目:国家重点研发计划资助(2017YFC1703303);福建省自然科学基金项目(2020J01435,2019J01846);中国福建省对外合作项目(2019I0001)

Supported by: National Key R&D Program of China(2017YFC1703303); Natural Science Foundation of Fujian Province, China(2020J01435, 2019J01846); External Cooperation Project of Fujian Province, China(2019I0001)

convolutional neural network to obtain the feature map of the image information, the question text is encoded by the recurrent neural network to obtain the sentence-level embedding representation vector. The output answer is obtained by the fusion inference model. To capture the overall representation of the fusion of multi-modal feature information, the popular attention mechanism is concerned about the obtained image feature matrix as the input of the text encoder in recent years. However, the interaction between the relationship features in the multi-modal scene has a huge negative impact on the extraction of effective semantic features. 2) A multi-module framework algorithm is used to decompose the task into multiple steps. Different modules are used to obtain the feature information at first, the obtained information is then used as the input of the subsequent modules, and the final output results are obtained through the subsequent algorithm modules. However, this type of method needs to rely on additional annotation information to train individual modules, and the complexity is quite higher. So, we develop a weight-driven re-located relational network model based on fusion semantic feature extraction. **Method** We clarify the whole framework for weight-driven re-located relation network, which consists of three modules in the context of image feature extraction, the attention-based long short-term memory (LSTM) and joint weight-driven re-located relation network. 1) For the image feature extraction module, image feature extraction is implemented via fusing the convolutional layer and the up-sampling layer. To make the extracted image feature information more suitable for the scene task, we design a fusion of convolutional neural network and U-Net network architecture to construct a network model that can extract the semantic meaning of low-level and high-level image features. 2) For the attention-based LSTM module, we joint the problem-based reasoning feature representation in terms of attention mechanism. LSTM can just retain the influence of existing words on unrecognized words. To obtain a better vector representation of the sentence, we can capture different contextual information based on attention mechanism. 3) For the joint leading weight-driven re-located relation network module, we propose a paired matching mechanism, which guides the matching process of relationship features in the relationship network. That is to calculate the inner product of the feature vector of each pixel with the feature vectors of all the pixels, the similarity can be obtained between it and all the points and the pixel can be obtained by averaging in the entire group at the end. However, to resolve the high complexity problem, it ignore the overall relationship balance that can be obtained by the original pairwise pairing method although the relationship features matching pair sequence obtained by the above method. Therefore, our re-located operation is carried out to achieve a balanced effect for overall relationship. 1) Remove the relationship feature of the pixel paired with itself from the obtained relationship feature pair set; 2) swap locations in the relationship feature list of each pixel according to a constant one exchange and this iterative rule; and 3) add the location information of the pixels and the sentence-level embedding. Especially, the generation of relational features is composed of three parts: a) the feature vector of two pixels, b) the coordinated value of the two pixels, and c) the embedding representation of the question text. **Result** The experiment is compared to the 2 datasets with the latest 6 methods. 1) For the FigureQA (an annotated figure dataset for visual reasoning) dataset, compared to IMG + QUES (image + questions), relation networks (RN) and ARN (appearance and relation network), the overall accuracy rate is increased by 26.4%, 8.1%, and 0.46%, respectively. 2) For a single verification set, compared to LEAF-Net (locate, encode and attend for figure network) and FigureNet, the accuracy is increased by 2.3% and 2.0% of each. 3) For the understanding data visualization via question answering (DVQA) dataset, the overall accuracy of the DVQA dataset is increased by 8.6%, 0.12%, and 2.13% compared to SANDY (san with dynamic encoding model), ARN, and RN, and 4) For the Oracle version, compared to SANDY, LEAF-Net and RN, the overall accuracy rate has increased by 23.3%, 7.09%, 4.8%, respectively. **Conclusion** Our model has good results on the two large open source datasets in the statistical graph Q&A beyond baseline model.

Key words: computer vision; figure question answering; multimodal fusion; attention mechanism; relation network (RN); deep learning

0 引 言

图表问答(figure question answering,FQA)是近年来提出的计算机视觉中的多模态任务,是用于处理图表图像与文本的多模态融合问题。图表问答与视觉问答(Chaudhry 等,2020)和视觉推理类似,都是同时输入一个图表图像与一个文本问题,并通过

一个算法模型得到最终的输出答案。但是,图表问答与视觉问答和视觉推理任务的区别很大,并不能简单看成同一类问题。具体区别如下:在输入的图像数据方面,3 个任务用到的图像内容与结构有很大的不同(Wu 和 Nakayama,2020),视觉问答使用自然图像作为输入图像,图像结构非常复杂,没有明显的结构特征,视觉推理任务主要是关于图像中物体的属性信息内容进行推理,具有更明显的结构特性(Antol 等,2015),而图表问答使用各类图表图像作为输入图像,有着显著结构化的特点,重点关注图表内部元素间的对比关系,如大小、比例和折线趋势等(Teney 等,2018);在输入的文本问题方面,在视觉问答与视觉推理任务上问题文本的单词有固定的语义,而图表问答任务中的问题文本的语义会根据图表中的文本信息而变化(Kafle 等,2020)。因此,理论上来说,在视觉问答任务或视觉推理任务中加入额外的语料库辅助文本问题的嵌入表示的做法并不适用于图表问答(Brill 等,2002)。

在图表问答中,目前已有的方法大致分为两类。一类是为图表问答任务直接进行建模的神经网络算法框架(Andreas 等,2016),算法通过使用卷积神经网络处理图表图像来获得图像特征信息的特征图,再通过使用循环神经网络处理问题文本获得句子级嵌入表示向量,最后使用融合推理模型(Pal 等,2012)得到最终的输出答案。其中很多算法工作会引入注意力机制,并将获得的图像特征矩阵作为问题文本编码器的输入,从而捕捉多模态特征信息融合的整体表征(Zhou 等,2015)。然而,多模态场景中关系特征之间的交互作用对有效语义特征的提取具有巨大的负影响。

另一类算法是将图表问答任务进行多步骤多模块分解(Miller 等,2020),算法使用不同的模块来获取图像的特征信息,如图像颜色等,再将获取的信息特征作为后续模块的输入,并通过后续的算法模块(Zhang 等,2017)处理得到最终的问题答案。然而,这类方法需要额外的标注信息来训练各个独立的模块,而且算法的复杂度往往过高(Gibson,2010)。

本文提出一种解决图表问答问题的新方法。首先在图像处理部分,提出了融合低级图像特征和高级图像特征的方法来提取更丰富的语义信息,在文本问题处理部分,引入注意力机制来获得更好的句子向量表示,并在关系推理部分提出一种基于引导性权重的配对匹配机制,然后通过对于图像内部关系以及图像和文本问题之间的交叉关系进行图像的位置重构,融合推断图像内部的关系以及图像与文本问题的相互关系,最终构建引导性权重驱动的重定位关系网络(leading weight-driven re-position relation network,LWR-RN)模型。

本文贡献如下:

1)提出了融合低层图像特征和高层图像特征的方法来获得更丰富的图表语义信息,从而提高模型在图表问答任务中的能力。

2)在文本编码器中引入注意力机制捕获不同的上下文信息,来获得更好的句子向量表示。

3)提出了基于引导性权重的配对匹配机制和重定位关系网络,通过融合推断图像内部关系以及图像与文本问题间的相互关系,获得更好的关系特征表示。

4)实验结果证明了本文方法在公共数据集上的有效性和鲁棒性。

1 相关工作

1.1 图表问答

图表问答 FQA 作为新兴的多模态任务,近年来已有许多关于图表问答的开源数据集公开,如 FigureQA(an annotated figure dataset for visual reasoning)数据集(Kahou 等,2017)和 DVQA(understanding data visualization via question answering)数据集(Kafle 等,2018)。故本文选择在这两个公开数据集上验证所提网络模型的有效性。

图表问答数据集公开的同时也发布了几个基线模型。IMG + QUES(image + questions)(Pal 等,2012)是两个开源数据集的基线模型,也是计算机视觉中视觉问答等多模态任务的经典模型框架。用来获得图像特征的图像编码器使用了多层卷积神经网络,用来获得文本嵌入表示的文本编码器则使用长短期记忆网络(long short-term memory, LSTM),最后通过多层全连接神经网络进行推理得到最终输出答案。

关系网络 RN(relation networks)(Santoro 等,2017)是应用于图表问答的另一个基线模型。RN 最早是由 Google 公司的 DeepMind 团队为了改善视觉推理任务提出的算法模型,并在 CLEVR(composi-

tional language and elementary visual reasoning diagnostics)数据集上展示了强大的效果(Johnson 等,2017),随后发现它在图表问答上的效果大大超过了其他基线方法(Zhang 等,2017)。其算法的核心思想为将通过卷积神经网络处理得到的特征图上的每个像素点作为一个对象,将每两个像素的特征向量配对在一起,形成一个包含所有像素之间关系的关系对特征(Mertens 等,2004),再使用一个多层感知器来推理每两个像素之间的关系,最后从这些两两关系的整体关系表示来推理最终答案(Echihabi 和 Marcu,2013)。其简单的两两配对方法可以包含所有像素之间的关系,从而使模型在两个任务上都取得了良好的效果(Miller 等,2020)。但两两配对方法不仅包含大量冗余信息,而且关系对的特征数量会进行平方式增长,这给后续的推理网络在计算量和参数量等方面带来了很大的负担。

研究者们也提出了许多新颖的算法来提升图表问答任务的准确率。Kafle 等人(2018)提出的 SANDY (san with dynamic encoding model)模型采用多次迭代的注意力机制实现对图像关键区域的捕捉,并与提出的结构相配合。FigureNet 模型(Reddy 等,2019)提出了一个多模块算法框架来解决图表问答问题。这个模块通过大量的标注数据进行预训练,用于提取特征和元素的对应值,例如图像颜色,然后得到特征向量,并拼接问题的嵌入表示。最后,通过多层全连接神经网络进行推理。但受模型设计的限制,FigureNet 模型只能应用于柱形图和饼图。ARN 模型(Zou 等,2020)是关系网络框架算法。它首先通过多个识别模块识别图像的元素、字符和结构等信息,然后通过得到的信息将其构建成表格的形式,最后通过表格问答模型得到答案。LEAF-Net (locate, encode and attend for figure network) 模型(Chaudhry 等,2020)使用了许多开源的预训练模型。首先,通过光学字符识别(optical character recognition,OCR)来识别图像中的字符信息,然后将其定位到嵌入问题中,同时通过预训练的 ResNet-152 得到图像特征图,最后通过空间注意力机制将特征图作为隐藏层信息添加到文本编码器 LSTM 中,来获得句子级表示。

1.2 U-Net

U-Net 是一种基于全卷积网络(fully convolution network,FCN)的语义分割网络,最初用于分割医学图像。U-Net 网络结构类似于 FCN 网络结构,只有卷积层和池化层,没有全连接层。与 FCN 网络不同的是,U-Net 网络模型的上采样和下采样使用了相同数量的卷积运算,并且使用 skip connection 来连接下采样层和上采样层,使得下采样层的图像特征信息可以直接传输到上采样层,对图像像素的定位更加准确。在效率方面,U-Net 只需要进行一次训练,而 FCN 需要 3 块进行训练才能实现更准确的 FCN-8 s 结构(Ronneberger 等,2015),所以 U-Net 网络的效率也高于 FCN 网络。

U-Net 在计算机视觉领域的各个方向也取得了不错的效果。Ranjan 等人(2020)使用 MRA(multi-resolution analysis)网络结合 U-Net 的复合结构对遥感图像进行区域分割。Shopovska 等人(2018)在对多光谱图像去除马赛克的任务中使用了 U-Net 网络,也取得了不错的效果。He 等人(2020)结合注意力地图和 U-Net 网络对低光照度图像实现图像亮度增强、去噪和色彩恢复。Goel 等人(2021)在灰度图像转换为彩色图像的任务中设计了 U-Net 模型框架,进行自动化着色。杨佳林等人(2021)利用 U-Net 网络提取遥感图像中的道路。在图像特征提取时,本文将使用 U-Net 网络来获得更加有效的图像特征。

1.3 注意力机制

注意力机制借鉴人类视觉的注意力选择机制,通过一定的计算方法来捕捉需要关注的关键信息,而忽略其他信息。随着计算机领域的发展,注意力机制在各个方向均取得了不错的效果。Google 团队 Vaswani 等人(2017)提出了在循环神经网络(recurrent neural network,RNN)模型中引入注意力机制进行图像分类。Luong 等人(2015)提出了使用不同 attention 架构的集成模型来改进机器翻译任务。而 Transformer 算法框架中更是提出了各种注意力机制的巧妙应用,可广泛应用于各个领域(Fukui 等,2019)。闫茹玉和刘学亮(2020)结合新颖的注意力机制模型增强对图片特征的提取准确率,从而提高视觉问答效果。在图表问答算法中,许多研究者都应用到了注意力机制,如 Kafle 等人(2018)采用多次迭代的注意力机制实现对图像关键区域的捕捉。

注意力机制的核心思路是将当前应用场景的特征向量看做由〈Key,Value〉数据特征对构成,并将目标中的某个特征向量标记为 Query,通过计算

Query 与各个特征对之间的相关性，得到 Key 与 Query 的对应关系，然后对 Value 按照一定的计算方式进行加权求和，再通过使用如 Softmax 函数的计算方式将权重值映射到 0～1 的范围进行归一化，使得最终的概率分布总和为 1，最后与对应的 Value 特征向量进行加权求和，即可得到 Key 的 attention 值。该过程的计算式为

$$\alpha_i = f_{\text{softmax}}(s(\boldsymbol{X}_i, \boldsymbol{q})) \tag{1}$$

式中，$\boldsymbol{q}$ 代表 Query 特征向量，$\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N]$ 表示 N 个输入信息。通常会令 Key = Value = $\boldsymbol{X}$，则 α_i 为得到的注意力分布。$s(\boldsymbol{X}_i, \boldsymbol{q})$ 为注意力打分机制，表示在给定任务相关的查询向量 $\boldsymbol{q}$ 时，第 i 个输入向量 $\boldsymbol{X}_i$ 受注意的程度。3 种常见的注意力打分机制分别为点积模型、双线性模型以及加性模型，计算式为

$$s(\boldsymbol{X}_i, \boldsymbol{q}) = \begin{cases} \boldsymbol{b}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{q} \\ \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{W}\boldsymbol{q} \\ \boldsymbol{v}^{\mathrm{T}}\tanh(\boldsymbol{W}\boldsymbol{X}_i + \boldsymbol{U}\boldsymbol{q}) \end{cases} \tag{2}$$

式中，$\boldsymbol{W}$，$\boldsymbol{U}$，$\boldsymbol{v}$ 均为可学习的参数权重。在本文模型中，注意力机制是序列到序列模型的基础，将注意力机制集成到经典的 LSTM 网络中，以提取问答句中长期相关词之间的关系。

2 方 法

本文提出一种用于图表问答的新型关系网络模型，命名为引导性权重驱动的重定位关系网络。首先通过融合卷积层和 U-Net 网络来进行低层图像特征和高层图像特征的提取，同时基于注意力机制的推理特征表示实现融合语义的特征提取，然后根据多级图像特征对引导性权重进行排序，并进一步重新排列每个图像特征的位置，来建立一个关系网络，完成特征配对。

2.1 融合语义特征提取

传统的卷积神经网络（convolutional neural networks，CNN）模型可以通过多层卷积运算来提取高层图像特征，但随着网络的加深会造成对低层图像特征的遗忘。而在图表问答中，低层和高层的图像特征语义都起着重要的作用。因此，本文尝试通过加深卷积神经网络的层次来获取更高层次的图像特征，同时融合较低层次的图像特征，使提取的图像特征信息更适合场景任务。

本文通过融合 CNN 和 U-Net 网络结构，来构建一个能够同时提取低层和高层图像特征语义的网络模型，如图 1 所示。图 1 中所有特征图的通道数均为 64。首先通过多层卷积和平均池化层来获得高阶图像特征语义，当卷积到 1×1 特征图时，再通过反卷积操作进行上采样还原大小。同时，将上采样过程中的特征映射（即高级图像特征语义）与卷积过程中相同大小的特征映射（即低级图像特征语义）进行融合，从而融合低层和高层图像特征成为更丰富的图像语义特征图。整个网络结构可分为下采样和上采样两个过程，其中下采样过程包括卷积模块和池化模块，上采样过程包括反卷积模块和融合模块。

假设在下采样过程中经过一次卷积和池化操作之后得到的特征图为 $\boldsymbol{C}_i$，$i \in [1, n]$，n 为下采样的

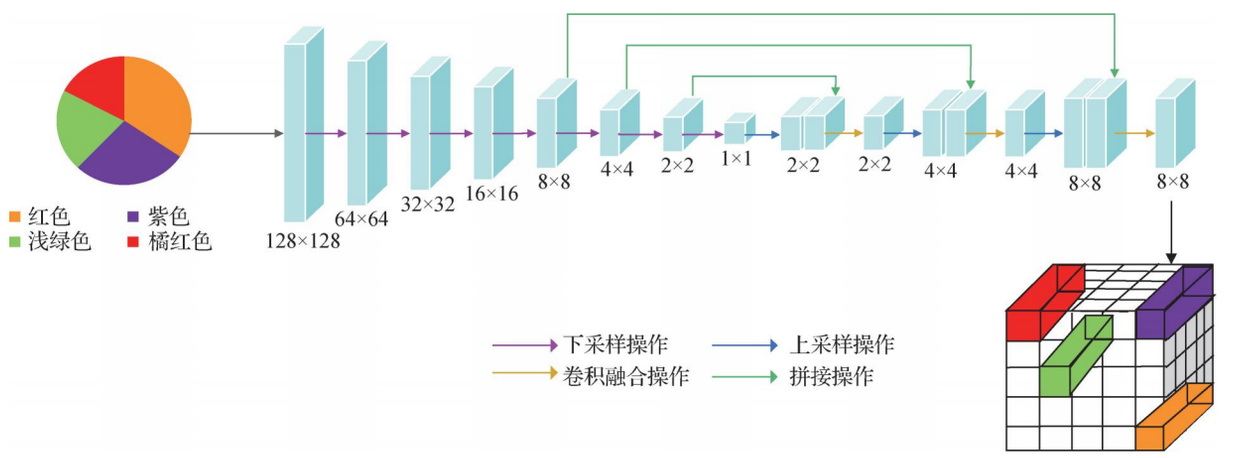

图 1 融合语义特征提取流水线

Fig. 1 The pipeline of fusion semantic feature extraction

次数,那么下采样过程可以描述为

$$\boldsymbol{C}_i = AugPool(Conv(\boldsymbol{C}_{i-1})),\ \boldsymbol{C}_0 = \boldsymbol{I} \tag{3}$$

式中,$\boldsymbol{C}_0$ 是原始的输入图像 $\boldsymbol{I}$;$Conv$ 代表卷积操作,每个卷积层都由 64 个卷积核组成,尺寸为 3×3,步长为 1;$AvgPool$ 则代表尺寸为 2×2 并且步长为 2 的平均池化操作。每个特征图 $\boldsymbol{C}_i \in \mathbf{R}^{\frac{w}{2}\times\frac{h}{2}\times k}$ 的大小是前一个特征图 $\boldsymbol{C}_i \in \mathbf{R}^{w\times h\times k}$ 尺寸的一半,特征通道数保持不变。

上采样过程中则是通过反卷积模块和融合模块依次反复交替的形式,即先经过一个反卷积模块得到尺寸放大一倍的特征图后,再通过融合模块对得到的特征图与下采样过程中相同尺寸的特征图进行特征融合。融合模块包含两个操作,首先将输入的两个相同尺寸的特征图在特征通道维度上进行拼接操作,然后经过一层卷积层进行融合操作。融合模块能够将两个相同尺寸的特征图通过融合之后得到同输入一样尺寸的特征图。假定反卷积后的特征图为 $\boldsymbol{D}_i, i \in [1,n]$。则上采样过程描述为

$$\boldsymbol{D}_i = f_{\mathrm{DeConv}}(\boldsymbol{F}_{i-1}), \boldsymbol{F}_0 = \boldsymbol{C}_n \tag{4}$$

$$\boldsymbol{F}_i = f_{\mathrm{Concat}}(\boldsymbol{C}_{n-i}, \boldsymbol{D}_i) \tag{5}$$

式中,f_{DeConv} 表示反卷积操作,每个反卷积层包含 64 个卷积核,尺寸为 3×3,步长为 2;f_{Concat} 表示拼接操作,即将 $\boldsymbol{C}_{n-i}$ 和 $\boldsymbol{D}_i$ 拼接起来得到 $\boldsymbol{F}_i$。反卷积模块完成后得到两倍大小的特征图,再将当前层反卷积模块的输出结果和下采样过程中相同尺寸的输出特征图作为当前层融合模块的输入,得到融合特征后的输出结果再传入下一层的反卷积模块。需注意,反卷积模块第 1 层的输入是下采样过程中的最终输出结果,并且上采样过程中全程特征通道数保持不变。其中,反卷积模块的作用是在进行卷积运算的同时放大特征图的大小来获取高阶的图像特征,而融合模块不仅能融合高层次和低层次的图像特征,而且能降低特征的维数。使用多个反卷积模块和融合模块交替的方式是为了在不同大小的特征图上进行高层和低层图像特征的融合。

2.2　基于注意力机制的文本编码器

在文本编码器算法模块,由于图表问答中的文本问题具有句子篇幅较短且句式相对固定的特点,故大多数算法模型采用长短期记忆模型(LSTM)(Gers 等,2000)来获得文本问题的嵌入式表示,但普通的 LSTM 只能保留现有单词对未知单词的影响。故本文通过引入注意机制,来捕获不同的上下文信息,从而获得更好的句子级向量表示。本文使用 AT-LSTM(attention-based LSTM)模型(Wang 等,2016)作为文本编码器的算法模型,其隐藏层的神经元数量设置为 256。具体来说,可以将问题文本表示为 $\boldsymbol{Q} = [e_1, e_2, \cdots, e_T]$,其中 e_t 表示句子中的单词。首先,需要通过嵌入矩阵将所有单词嵌入向量空间。定义向量的长度为 64,则单词的向量可以表示为 $\boldsymbol{x}_t = \boldsymbol{E}_{e_t,*} \in \mathbf{R}^{64}$。接着在每一步的操作中,将当前的单词嵌入表示向量输入 LSTM 神经元来获得隐藏状态的更新。

定义 $\boldsymbol{H} \in \mathbf{R}^{d\times N}$ 为由隐藏向量组成的矩阵。$\boldsymbol{H} = [h_1, h_2, \cdots, h_N]$ 是 LSTM 生成的隐藏层向量表示,其中 d 是隐藏层的大小,N 是给定句子的长度。进一步地,$\boldsymbol{v}_a$ 代表给定方面信息的嵌入表示,$\boldsymbol{e}_N$ 则是一个 1 维向量。注意力权重向量 $\boldsymbol{\alpha}$ 和一个加权的隐藏表示 $\boldsymbol{r}$ 的计算式为

$$\boldsymbol{M} = \tanh\begin{bmatrix}\boldsymbol{W}_h\boldsymbol{H} \\ \boldsymbol{W}_v\boldsymbol{v}_a \otimes \boldsymbol{e}_N\end{bmatrix} \tag{6}$$

$$\boldsymbol{\alpha} = f_{\mathrm{softmax}}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{M}) \tag{7}$$

$$\boldsymbol{r} = \boldsymbol{H}\alpha^{\mathrm{T}} \tag{8}$$

式中,投影参数 $\boldsymbol{M} \in \mathbf{R}^{(d+da)\times N}$,$\boldsymbol{\alpha} \in \mathbf{R}^N$,$\boldsymbol{r} \in \mathbf{R}^d$,$\boldsymbol{W}_h \in \mathbf{R}^{d\times d}$,$\boldsymbol{W}_v \in \mathbf{R}^{d_a\times d_a}$,$\boldsymbol{w} \in \mathbf{R}^{d+d_a}$,$\boldsymbol{W}_v$,$\boldsymbol{W}_h$ 和 $\boldsymbol{w}$ 均为可学习的参数权重。$\boldsymbol{\alpha}$ 是一个注意力权重组成的向量,$\boldsymbol{r}$ 是具有给定方面的句子的加权表示。最终的文本表示为

$$\boldsymbol{h}^* = \tanh(\boldsymbol{W}_p\boldsymbol{r} + \boldsymbol{W}_x\boldsymbol{h}_N) \tag{9}$$

式中,$\boldsymbol{h}^* \in \mathbf{R}^d$,$\boldsymbol{W}_p$、$\boldsymbol{W}_x$ 是投影的参数矩阵。

2.3　引导性权重驱动的重定位关系网络

本文网络由图像特征提取模块、基于注意力机制的 LSTM 模块和用于图表问答的重定位关系网络 3 个部分组成,如图 2 所示。

2.3.1　基于引导性权重的配对匹配机制

在关系网络中,两两配对机制是将每个对象视为等价对象。但特征图中的每个像素与整体之间的关系必然不同,换句话说,如果将特征图上的每个像素点视为一个对象,则每个对象与整个群体之间的关系必然不是等同,整个群体中每个对象的重要性也不一定相等。

为了衡量两个向量之间的关系,提出基于引导性权重的配对匹配机制,用于指导关系网络中关系

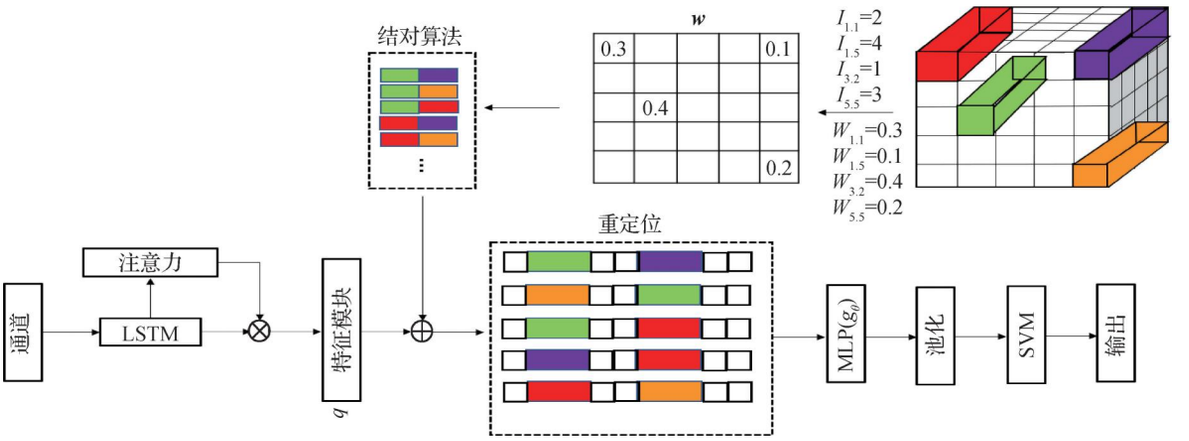

图 2　引导性权重驱动的重定位关系网络框架

Fig. 2　The framework of leading weight-driven re-position relation network (LWR-RN)

特征的配对过程。即计算每个像素点的特征向量与所有像素点的特征向量的内积,得到它与所有点之间的相似性度量,然后通过平均得到整个组中的像素。假定 $\boldsymbol{F}=\{\boldsymbol{f}_{1,1},\boldsymbol{f}_{1,2},\cdots,\boldsymbol{f}_{n,n}\}$ 表示特征图上所有行列像素的特征向量,$i,j\in[1,n]$。那么所有特征向量的引导性权重可以描述为

$$W_{i,j}=\frac{\exp\left(\sum_{p=1}^{n}\sum_{q=1}^{n}\boldsymbol{f}_{p,q}^{\mathrm{T}}\boldsymbol{f}_{i,j}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\exp\left(\sum_{p=1}^{n}\sum_{q=1}^{n}\boldsymbol{f}_{p,q}^{\mathrm{T}}\boldsymbol{f}_{i,j}\right)} \tag{10}$$

式中,$i,j,p,q\in[1,n]$。将权重从大到小排序,得到权重向量 $\boldsymbol{W}=[W_{1,1},\cdots,W_{i,j}]$,$W_{i,j}$ 的索引用符号 $I_{i,j}$ 表示。对图像内部元素进行排序,使具有较高引导性权重的像素可以承担更多的配对任务,这不仅可以大大减少生成的关系特征的数量,而且可以使整体元素之间的关系更加紧密。然后,进一步利用引导性权重对图像的语义特征进行配对,每个像素与自身和引导性权重低于其自身的像素配对。通过这样的配对过程,可以丢弃接近原始配对产生的总数的一半的冗余关系特征表达式,同时允许具有较大引导性权重的像素在整体中承担更多对关系表达。最后得到像素点对的集合 $\boldsymbol{FF}=\{f_{1,1}f_{1,2},\cdots,f_{i,j}f_{p,q}\}$,集合中的每个元素满足 $W_{i,j}>W_{p,q}$。

2.3.2　重定位关系网络

尽管通过上述方法获得的关系特征配对序列解决了高复杂性问题,但它不再具有通过原始配对方法获得的整体关系平衡,通过在通道的维度上拼接形成关系特征的配对,具有高引导性权重的像素位于最前面的结果将为模型创建一种无形的学习指导关系。因此,进一步执行重新定位操作,以使整体关系达到平衡。

首先在获得的关系特征对的集合中移除与其自身配对的像素的关系特征,然后在每个像素点的关系特征列表中按照一个不变一个交换以此反复的规则进行交换位置的操作。这个过程描述为

$$f_{i,j}f_{p,q}=\begin{cases}f_{i,j}f_{p,q} & (I_{p,q}-I_{i,j})\bmod 2=1\\ f_{p,q}f_{i,j} & (I_{p,q}-I_{i,j})\bmod 2=0\end{cases} \tag{11}$$

因此,可以把像素特征向量和像素的关系看成一个节点和一个边,这种关系必须是有方向的。通过方向关系,可以传递不同元素之间的关系,这样就形成了关系图结构。然后,将像素的位置信息和文本的句子级嵌入表示添加到关系特征中。最终的关系特征由 3 部分组成:两个像素的特征向量、两个像素的坐标值和问题文本的嵌入表示。

让 $\boldsymbol{p}$ 表示关系特征对,$\boldsymbol{f}_{i,j}$ 表示特征图上第 i 行第 j 列像素点的特征向量,则关系特征对的组成可以表示为 $\boldsymbol{P}_{(i,j),(u,v)}=[h^{*},\boldsymbol{f}_{i,j},i,j,\boldsymbol{f}_{u,v},u,v]$。在由引导性权重驱动的配对匹配机制生成的关系特征中,每个关系特征包含两个像素的内容和位置信息,以及问题文本的信息。使用共享参数多层感知器(multi-layer perceptron, MLP)对问题知识先验条件下的每两个像素进行关系推理,得到问题知识先验条件下每两个像素的关系特征向量后,使用平均池化方法聚合关系特征向量以获得最终的特征表示。最后,将全局表示特征向量通过一个支持向量机(support vector machine, SVM),得到该类别对应的每幅图像的最终概率分布 J。该过程表示为

$$J = f_{\mathrm{SVM}}(f_{\mathrm{AvgPool}}(g_\theta(\boldsymbol{P}_{(i,j),(u,v)}))) \tag{12}$$

式中，$\boldsymbol{P}_{(i,j),(u,v)}$ 表示从特征图生成的关系特征，g_θ 表示共享参数 θ 的 MLP，f_{AvgPool} 表示平均池化操作，f_{SVM} 表示经过 SVM 分类器得到最终的概率分布。

3　实　验

3.1　数据集

3.1.1　FigureQA 数据集

FigureQA 数据集(Kahou 等,2017)是微软公司在 2017 年提出的用于图表问答任务的大型开源数据集,分为 5 个独立的包,分别为 1 个训练集、2 个验证集和 2 个测试集。不同的验证集和测试集会使用两种不同的颜色集来生成图表图像,一个颜色集与训练集相同,另一个颜色集与训练集不同,这样避免了颜色对模型的影响。训练集包含 100 000 幅图表图像和 1 327 368 个文本问题,验证集 1 包含 20 000 幅图表图像和 265 106 个文本问题,验证集 2 包含 20 000 幅图表图像和 265 798 个文本问题。然而两个测试集并不开源,评估结果需要发送电子邮件给作者来获得结果。故与其他工作类似,本文直接在验证集上进行算法模型效果的测试。FigureQA 数据集中的图表类型包括垂直柱形图、水平柱形图、折线图、散点图和饼图 5 类,文本问题由 15 种问题类型组成,包括有关各类图表的结构、细节和推理问题。

3.1.2　DVQA 数据集

DVQA 数据集(Kafle 等,2018)是 Kafle 等人与 Adobe 研究实验室合作提出的大型开源图表问答数据集,包含 1 个训练集和 2 个测试集(Test-Familiar 测试集和 Test-Novel 测试集)。训练集包括 200 000 幅图表图像和 2 325 316 个文本问题,Test-Familiar 测试集包括 50 000 幅图表图像和 580 557 个文本问题,Test-Novel 测试集包括 50 000 幅图表图像和 581 321 个文本问题。DVQA 数据集的问题类型分为 3 类:第 1 类是关于结构理解的问题,即对图表图像整体结构的理解;第 2 类是关于数据检索的问题;第 3 类是关于推理的问题,即对图像中元素之间相关关系的推理问题。

3.2　实验设置

本文提出的模型训练主要分为两步:首先是训练重定位关系网络,然后是提取训练好的网络的全局表示的特征向量送到 SVM 分类器得到最终的分类结果。重定位关系网络训练时,使用的损失函数是二分类交叉熵损失函数,使用 Adam 优化器,学习率初始化设置为 2.5E-4,动量参数设置为 0.9,批次设置为 80。当训练 SVM 分类器时,使用的损失函数就是合页损失函数(hinge loss function)。

3.3　实验结果

为了验证本文提出的引导性权重驱动的重定位关系网络对图表问答任务的效果,与近年来在开源数据集 FigureQA 和 DVQA 上的图表问答方面的其他先前工作进行比较。此外,DVQA 数据集上的验证模型还包含两个版本:无动态字典的方法(No OCR)和有 Oracle 版本的方法。表 1 和表 2 的准确率分别显示了在 FigureQA 和 DVQA 两个数据集上的实验结果的比较。对于 FigureQA 数据集,本文模型远远超过了基线模型 IMG + QUES,整体准确率提高了 26.4%;相较于基线模型 RN,整体准确率提高了 8.1%;相较于性能第 2 的模型 ARN,整体准确率提高了 0.46%。在单一验证集上,相较于 LEAF-Net、FigureNet,准确率提升了 2.3%,2%,且 FigureNet 受网络结构的限制,只能应用于柱形图和饼图。在 DVQA 数据集上,对于不使用 OCR 方法,相较于 SANDY、ARN 和 RN,整体准确率分别提升了 8.6%,0.12%,2.13%;对于有 Oracle 版本,相较于 SANDY、LEAF-Net 和 RN,整体准确率分别提升了 23.3%,7.09%,4.8%。

通过实验结果可以看出,本文模型在图表问答的两个大型开源数据集上都能取得良好的效果,得到的结果的准确率明显优于一些现有算法模型,并

表 1　不同算法在 FigureQA 数据集上的实验结果准确率对比

Table 1　Results for the FigureQA dataset of different algorithms

/%

模型	Validation 1 验证集	Validation 2 验证集
IMG + QUES	59.41	57.14
RN	78.21	74.98
LEAF-Net	-	81.15
ARN	85.48	82.95
FigureNet	83.90	-
LWR-RN	**85.91**	**83.43**

注:加粗字体表示各列最优结果,"-"表示无数据。

表 2　不同算法在 DVQA 数据集上的实验结果准确率对比

Table 2　Results for the DVQA dataset for our LWR-RN algorithm compared to existing algorithms

/%

模型	Test-Familiar	Test-Novel
IMG + QUES	32.01	32.01
SANDY（No OCR）	36.02	36.14
ARN(No OCR)	44.50	44.51
RN（No OCR）	42.49	42.50
LWR-RN（No OCR）	**44.76**	**44.49**
SANDY（Oracle）	56.48	56.62
LEAF-Net（Oracle）	72.72	72.89
RN（Oracle）	74.95	75.18
LWR-RN（Oracle）	**79.83**	**79.96**

注：加粗字体为分别为不使用 OCR 和有 Oracle 版本方法的最优值。

远远超过了基线模型。

3.4　消融实验

为了验证本文模型中关系配对机制的每个步骤的重要性，本文设计了一个详细的消融实验，该实验在 FigureQA 数据集的两个验证集上进行。实验包括 5 个模型，具体描述如下：

模型 1：没有基于引导性权重的配对匹配机制（步骤 1））。

模型 2：没有重新定位关系网络（步骤 2））。

模型 3：没有基于引导性权重的配对匹配机制和重新定位关系网络（没有步骤 1）和步骤 2））。

模型 4：完整的引导性权重驱动的重定位关系网络。

模型 5：去除低引导性权重值的像素点的模型。即在完整模型的重定位时，不仅去除冗余的关系特征，还去除了整体数量一半的低引导性权重值的像素点。

表 3 详细说明了每种图表类型的消融实验在不同模型的实验结果。显然，缺少步骤 1）和步骤 2）的模型 3 在每种类型的图表问答中表现不佳，尤其是在直线图和散点图中。举例而言，在散点图类型上与完整模型有较大的差距，准确率相差近 10%。分别缺少步骤 1）和步骤 2）的两个模型，每种图表的准确率也都低于完整模型。可以看出，步骤 1）和步骤 2）对整体效果都有很大影响。

从表 3 中模型 5 与模型 4 的比较中可以发现，即使直接丢弃一半的低权重值的像素特征，它仍然可以具有相当高的准确率。但是直接抛弃一半图像像素点特征的方式，会在不同的统计图类型上达到两极分化的情况。在饼图这样结构简单且图像信息密集的统计图上，去除一半的低引导性权重值像素点的模型 5 表现甚至优于完整的模型，这是由于饼状图的大量图像信息聚集于高引导性权重值的像素点上，而直接抛弃低引导性权重值的像素点能够让后续的过程不受其影响。但是在折线图这样图像信息分散的统计图类型上，简单地舍弃低引导性权重值的像素点会让模型丢失一部分有效的图像信息，因此表现得很差。

表 3　LWR-RN 在 FigureQA 数据集上的消融测试结果准确率

Table 3　LWR-RN ablation studies on FigureQA dataset

/%

模型	垂直柱形图	水平柱形图	饼图	折线图	散点图
模型 1	86.77	84.52	77.55	76.06	74.16
模型 2	88.93	86.16	78.68	73.63	74.42
模型 3	86.05	84.56	75.96	69.14	71.88
模型 4	**90.45**	**89.56**	80.29	**79.13**	**78.28**
模型 5	89.68	87.44	**82.30**	72.67	76.37

注：加粗字体表示各列最优结果。

表 4 详细列出了消融实验模型在每类问题上的实验结果。通过比较，缺少步骤 1）的模型 1 在前 6 类问题（即柱形图和饼图）的表现都比缺少步骤 2）的模型 2 差，可以看出，步骤 1）和步骤 2）对于不同类型的图表问答推理都起着重要作用。同时缺少步骤 1）和步骤 2）的模型 3 在几乎所有问题类型中都表现最差，只有在两种问题类型中，“Is X the smoothest?”和“ Is X the roughest?”的最高准确率在误差范围内与完整模型基本相同。这是因为在折线图对于整体趋势提问的问题中，图像信息几乎存在于特征图的每个像素中，信息特征的分布相对均匀，从而降低了缺乏引导性权重值的计算带来的影响。同时，这也是在这两类问题上模型 1 的准确率很高，而在移除低引导性权重值像素点的模型 5 中性能较差的原因。虽然移除低引导性权重值的像素点的模型 5 丢弃了一半的低引导性权重值的像素点，但由于引导性权重驱动的配对匹配机制，它仍然可以很

表 4　LWR-RN 对 FigureQA 数据集每类问题的消融实验准确率

Table 4　LWR-RN ablation studies for each unique question in FigureQA

/%

问题类型	模型 1	模型 2	模型 3	模型 4	模型 5
Is X the minimum?	83.34	84.06	81.62	**87.82**	86.36
Is X the maximum?	88.28	89.14	86.94	91.32	**91.76**
Is X the low median?	72.82	75.17	72.44	77.39	**78.21**
Is X the high median?	73.85	75.98	72.72	**78.37**	78.15
Is X less than Y? (bar, pie)	89.92	91.75	89.45	**92.90**	92.23
Is X greater than Y? (bar, pie)	89.65	91.45	90.05	**92.97**	92.30
Does X have the minimum area under the curve?	82.51	82.14	72.28	**86.10**	80.67
Does X have the maximum area under the curve?	83.93	83.54	77.65	**89.46**	86.07
Does X have the lowest value?	77.09	75.57	72.14	**81.05**	75.43
Does X have the highest value?	81.93	81.18	75.10	**87.54**	83.01
Is X less than Y? (line)	75.34	74.26	70.38	**78.89**	74.01
Is X greater than Y? (line)	74.61	75.46	70.93	**79.29**	74.34
Does X intersect Y?	74.87	73.66	70.65	**80.08**	74.95
Is X the smoothest?	62.48	60.29	**62.93**	62.91	61.45
Is X the roughest?	63.08	60.84	**63.58**	63.54	61.99

注：加粗字体表示各行最优结果。

好地利用高引导性权重值的像素点的特征向量。像“Does X intersect Y?”这样的特征问题仍然可以表现得很好。值得注意的是，模型 5 在柱状图上的问答有着很高的准确率，在“Is X the maximum?”和“Is X the low median?”这两种问题类型中，精确度略高于完整的模型。

因此，通过对结果数据更详细的分析，可以发现引导性权重驱动的重定位关系网络中的每一步都发挥了不可替代的作用。无论缺少哪一步，整体性能都会下降，完整模型在各种类型的图表问答中都可以达到整体的最优效果。

4　结　论

针对现有关系网络存在的多模态融合语义提取缺失以及高复杂度的问题，提出了引导性权重驱动的重定位关系网络模型，并详细描述了算法的模型框架。此外，还进行了全面的实验对比和分析，包括有无引导性权重驱动的配对匹配机制等各种消融实验，并与当前前沿的算法模型进行了对比实验，取得了不错的实验结果。

下一步研究将主观定义的引导性权重公式转换为可学习的引导性权重参数矩阵，并将引导性权重集成到后续的关系聚合步骤中。其次，在构建了以引导性权重为导向的图元素结构后，本文只使用一个简单的共享多层感知器对关系特征进行推理，未来可以研究采用图卷积神经网络对图元素结构之间的关系进行关系特征推理。

参考文献(References)

Andreas J, Rohrbach M, Darrell T and Klein D. 2016. Learning to compose neural networks for question answering//Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: Association for Computational Linguistics: 1545-1554 [DOI: 10.18653/v1/N16-1181]

Antol S, Agrawal A, Lu J S, Mitchell M, Batra D, Zitnick C L and Parikh D. 2015. VQA: visual question answering//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2425-2433 [DOI: 10.1109/ICCV.2015.279]

Brill E, Dumais S and Banko M. 2002. An analysis of the AskMSR question-answering system//Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). Philadelphia, USA: Association for Computational Linguistics: 257-264 [DOI: 10.3115/1118693.1118726]

Chaudhry R, Shekhar S, Gupta U, Maneriker P, Bansal P and Joshi A. 2020. LEAF-QA: locate, encode and attend for figure question answering//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Snowmass, USA: IEEE: 3501-3510 [DOI: 10.1109/WACV45572.2020.9093269]

Echihabi A and Marcu D. 2003. A noisy-channel approach to question answering//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Sapporo, Japan: Association for Computational Linguistics: 16-23 [DOI: 10.3115/1075096.1075099]

Fukui H, Hirakawa T, Yamashita T and Fujiyoshi H. 2019. Attention branch network: learning of attention mechanism for visual explanation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 10697-10706 [DOI: 10.1109/CVPR.2019.01096]

Gers F A, Schmidhuber J and Cummins F. 2000. Learning to forget: continual prediction with LSTM. Neural Computation, 12(10): 2451-2471 [DOI: 10.1162/089976600300015015]

Gibson R F. 2010. A review of recent research on mechanics of multifunctional composite materials and structures. Composite Structures, 92(12): 2793-2810 [DOI: 10.1016/j.compstruct.2010.05.003]

Goel D, Jain S, Vishwakarma D K and Bansal A. 2021. Automatic image colorization using U-Net//Proceedings of the 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). Kharagpur, India: IEEE: 1-7 [DOI: 10.1109/ICCCNT51525.2021.9580001]

He W J, Liu Y Y, Feng J F, Zhang W W, Gu G H and Chen Q. 2020. Low-light image enhancement combined with attention map and U-Net network//Proceedings of the 3rd IEEE International Conference on Information Systems and Computer Aided Education (ICISCAE). Dalian, China: IEEE: 397-401 [DOI: 10.1109/ICISCAE51034.2020.9236828]

Johnson J, Hariharan B, Van Der Maaten L, Li F F, Zitnick C L and Girshick R. 2017. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1988-1997 [DOI: 10.1109/CVPR.2017.215]

Kafle K, Price B, Cohen S and Kanan C. 2018. DVQA: understanding data visualizations via question answering//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5648-5656 [DOI: 10.1109/CVPR.2018.00592]

Kafle K, Shrestha R, Price B, Cohen S and Kanan C. 2020. Answering questions about data visualizations using efficient bimodal fusion//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Snowmass, USA: IEEE: 1487-1496 [DOI: 10.1109/WACV45572.2020.9093494]

Kahou S E, Michalski V, Atkinson A, Kádár Á, Trischler A and Bengio Y. 2017. FigureQA: an annotated figure dataset for visual reasoning//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Luong T, Pham H and Manning C D. 2015. Effective approaches to attention-based neural machine translation//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics: 1412-1421 [DOI: 10.18653/v1/D15-1166]

Mertens K C, Verbeke L P C, Westra T and De Wulf R R. 2004. Sub-pixel mapping and sub-pixel sharpening using neural network predicted wavelet coefficients. Remote Sensing of Environment, 91(2): 225-236 [DOI: 10.1016/j.rse.2004.03.003]

Miller J, Krauth K, Recht B and Schmidt L. 2020. The effect of natural distribution shift on question answering models//Proceedings of the 37th International Conference on Machine Learning [EB/OL]. [2021-10-22]. https://arxiv.org/pdf/2004.14444v1.pdf

Pal A, Chang S and Konstan J A. 2012. Evolution of experts in question answering communities//Proceedings of the 6th International AAAI Conference on Weblogs and Social Media. Dublin, Ireland: AAAI: 274-281

Ranjan P, Patil S and Ansari R A. 2020. U-Net based MRA framework for segmentation of remotely sensed images//Proceedings of 2020 International Conference on Artificial Intelligence and Signal Processing (AISP). Amaravati, India: IEEE: 1-4 [DOI: 10.1109/AISP48273.2020.9073131]

Reddy R, Ramesh R, Deshpande A and Khapra M M. 2019. FigureNet: a deep learning model for question-answering on scientific plots//Proceedings of 2019 International Joint Conference on Neural Networks (IJCNN). Budapest, Hungary: IEEE: 1-8 [DOI: 10.1109/IJCNN.2019.8851830]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Santoro A, Raposo D, Barrett D G T, Malinowski M, Pascanu R, Battaglia P and Lillicrap T. 2017. A simple neural network module for relational reasoning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 4937-4983

Shopovska I, Jovanov L and Philips W. 2018. RGB-NIR demosaicing using deep residual U-Net//The 26th Telecommunications Forum (TELFOR). Belgrade, Serbia: IEEE: 1-4 [DOI: 10.1109/TELFOR.2018.8611 819]

Teney D, Anderson P, He X D and Van Den Hengel A. 2018. Tips and tricks for visual question answering: learnings from the 2017 challenge//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4223-4232 [DOI: 10.1109/CVPR.2018.00444]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang Y Q, Huang M L, Zhu X Y and Zhao L. 2016. Attention-based LSTM for aspect-level sentiment classification//Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA: Association for Computational Linguistics: 606-615 [DOI: 10.18653/v1/D16-1058]

Wu Y X and Nakayama H. 2020. Graph-based heuristic search for module selection procedure in neural module network//Proceedings of the 15th Asian Conference on Computer Vision on Computer Vision. Kyoto, Japan: Springer: 560-575 [DOI: 10.1007/978-3-030-69535-4_34]

Yan R Y and Liu X L. 2020. Visual question answering model based on bottom-up attention and memory network. Journal of Image and Graphics, 25(5): 993-1006 (闫茹玉, 刘学亮. 2020. 结合自底向上注意力机制和记忆网络的视觉问答模型. 中国图象图形学报, 25(5): 993-1006) [DOI: 10.11834/jig.190366]

Yang J L, Guo X J and Chen Z H. 2021. Road extraction method from remote sensing images based on improved U-Net network. Journal of Image and Graphics, 26(12): 3005-3014 (杨佳林, 郭学俊, 陈泽华. 2021. 改进 U-Net 型网络的遥感图像道路提取. 中国图象图形学报, 26(12): 3005-3014) [DOI: 10.11834/jig.200579]

Zhang J B, Zhu X D, Chen Q, Dai L R, Wei S and Jiang H. 2017. Exploring question understanding and adaptation in neural-network-based question answering [EB/OL]. [2021-10-22]. https://arxiv.org/pdf/1703.04617.pdf

Zhou B L, Tian Y D, Sukhbaatar S, Szlam A and Fergus R. 2015. Simple baseline for visual question answering [EB/OL]. [2021-10-22]. https://arxiv.org/pdf/1512.02167.pdf

Zou J L, Wu G L, Xue T F and Wu Q F. 2020. An affinity-driven relation network for figure question answering//Proceedings of 2020 IEEE International Conference on Multimedia and Expo (ICME). London, UK: IEEE: 1-6 [DOI: 10.1109/ICME46284.2020.9102911]

作者简介

黎颖,女,硕士研究生,主要研究方向为图表问答。
E-mail: 24320191152542@stu.xmu.edu.cn
吴清锋,通信作者,男,教授,主要研究方向为大数据与云计算、数字媒体技术、智能信息处理技术、计算机图形学和生物信息学。E-mail:qfwu@xmu.edu.cn
刘佳桐,女,硕士研究生,主要研究方向为机器学习。
E-mail: liujiatong@stu.xmu.edu.cn
邹嘉龙,男,硕士研究生,主要研究方向为图表问答。
E-mail:jlzou@stu.xmu.edu.cn