

中图法分类号: TP391.41 文献标识码: A 文章编号: 1006-8961(2023)11-3562-13
论文引用格式: Wang F, Shi F Y, Zhao J, Zhang X S and Wang X F. 2023. Answer mask-fused visual question answering model. Journal of Image and Graphics, 28(11):3562-3574(王峰, 石方宇, 赵佳, 张雪松, 王雪枫. 2023. 融合答案掩码的视觉问答模型. 中国图象图形学报, 28(11): 3562-3574)[DOI:10.11834/jig.211137]

融合答案掩码的视觉问答模型

王峰, 石方宇, 赵佳*, 张雪松, 王雪枫

阜阳师范大学计算机与信息工程学院, 阜阳 236041

摘要: 目的 现有的视觉问答模型由于受到语言先验的影响, 预测准确率不高。虽然模型能够根据数据集中问题和答案的统计规律学习到它们之间简单的对应关系, 但无法学习到问题和答案类型之间深层次的对应关系, 容易出现答非所问的现象。为此, 提出了一种使用答案掩码对预测结果中的无关答案进行遮盖的方法, 迫使模型关注问题和答案类型之间的对应关系, 提高模型的预测准确率。**方法** 首先对数据集中的答案进行聚类并为每一类答案生成不同的答案掩码, 然后使用预训练的答案类型识别模型预测问题对应的答案类型, 并根据该模型的预测结果选择相应的答案掩码对基线模型的预测结果进行遮盖, 最终得到正确答案。**结果** 提出的方法使用 UpDn(bottom-up and top-down)、RUBi(reducing unimodal biases)、LMH(learned-mixin +h)和 CSS(counterfactual samples synthesizing) 4种模型作为基线模型, 在3个大型公开数据集上进行实验。在 VQA(visual question answer)-CP v2.0 数据集上的实验结果表明, 本文方法使 UpDn 模型的准确率提高了 2.15%, LMH 模型的准确率提高了 2.29%, 融合本方法的 CSS 模型的准确率达到 60.14%, 较原模型提升了 2.02%, 达到了目前较高的水平。在 VQA v2.0 和 VQA-CP v1.0 数据集上的结果也显示本文方法提高了大多数模型的准确率, 具有良好的泛化性。此外, 在 VQA-CP v2.0 上的消融实验证明了本文方法的有效性。**结论** 提出的方法通过答案掩码对视觉问答模型的预测结果进行遮盖, 减少无关答案对最终结果的影响, 使模型学习到问题和答案类型之间的对应关系, 有效改善了视觉问答模型答非所问的现象, 提高了模型的预测准确率。

关键词: 视觉问答; 语言先验; 答案聚类; 答案掩码; 答案类型识别

Answer mask-fused visual question answering model

Wang Feng, Shi Fangyu, Zhao Jia*, Zhang Xuesong, Wang Xuefeng

School of Computer and Information Engineering, Fuyang Normal University, FuYang 236041, China

Abstract: Objective Visual question answering (VQA) is essential for artificial intelligence (AI) in recent years. Current VQA is concerned of the linkage of natural language processing and computer vision more. Therefore, VQA-related model is required for text and image information processing simultaneously, and the information of these two modes can be fused to infer the answer. Such popular VQA models like VQA v2.0 dataset have been developing in terms of a deep neural network and trained samples. However, these prior language models-based tasks can be simplified to learn the surface relationship for answer questions between questions and answers. The weakness of uneven distribution of answers is still to be

收稿日期: 2021-12-16; 修回日期: 2022-04-21; 预印本日期: 2022-04-28

* 通信作者: 赵佳 zhaojia11b@mails.ucas.ac.cn

基金项目: 国家自然科学基金项目(61906044); 中国博士后科学基金项目(2020M681984); 安徽省高校自然科学研究重点项目(2023AH050406)

Supported by: National Natural Science Foundation of China(61906044); China Postdoctoral Science Foundation(2020M681984); Key Projects of Natural Science Research in Anhui Colleges and Universities(2023AH050406)

challenged for its weak generalization and poor performance in the VQA-CP v2.0 dataset. Specifically, language problems-prior has threatened for its prediction errors of the model and the predicted answer and question are in irrelevance. To optimize this non-linkage and generalization of the model, we develop an answer mask-related method to cover the irrelevant answers for predictable results, which can forge the model to learn the deeper relationship between question and answer. The prediction accuracy of the model can be improved as well. **Method** The prediction results of the baseline model is masked via the answer mask. It is necessary to cluster all candidate answers and fewer answers-involved for each type of answer can be used to preserve accurate classification through more answers-irrelevant coverage of mask-of the prediction results. The answers consist of non-contextual words and phrases. Conventional Word2Vec and Glove is still challenged for its effectiveness of these encoded answers. Such clip is illustrated as the encoder to extract the answer features. And, the k-means algorithm is used to cluster answer-extracted feature vectors. After clustering, original dataset can be modified and the corresponded type is changed to the clustering-after type answer of the dataset, and different answer mask vectors are generated for each answer type. The answer mask vector is structured of 0 and 1. The elements of the vector can be assigned to 1 when the corresponding positions are contained for each answer type, and the others are configured to 0; the impact of irrelevant answers of prediction can be eliminated for final results of the baseline model. We design an answer type recognition model, which uses the questions and answers types for pre-training. Input question-based model can be used to predict the answer type corresponding to the question. The model's accuracy can reflect the quality of clustering work, and its prediction results are the basis for the optioned answers mask types- task. The baseline model is focused on encoding the image and text and depth neural network is linked to fuse the image and text features. The preliminary prediction results can be obtained through the classifier as well. First, corresponding answer mask vector is leaked out in terms of answer type identification model-based prediction results. Then, the multiplied prediction results are generated via the baseline model and the distribution of irrelevant answers are covered in the prediction results of the baseline model. At the end, final results are predicted. The model is trained to learn the corresponding relationship between the types of questions and answers. **Result** We selected out UpDn, RUBi, LMH and CSS as baseline models and experiments are carried out on three large public datasets mentioned below. VQA-CP v2.0 dataset-related experiments can show its potentials for model's accuracy. Three sort of accuracy of the UpDn, LMH and CSS model are improved by 2.15%, 2.29% and 2.02% each. Among them, the higher accuracy of the CSS model is reached to 60.14%. Additionally, our model's accuracy is preserved when VQA v2.0-related accuracy is reduced. The VQA v2.0-based experimental results show that the accuracy of most baseline models are improved further. Among them, the accuracy of the CSS model is optimized by 3.18%. To demonstrate better generalization of our model, comparative experiments are carried out on VQA-CP v1.0 dataset further. The experimental results show that our method is mutual-benefited for most of baseline models, which is sufficient to reflect its potential ability of generalization. Furthermore, ablation experiment on VQA-CP v2.0 shows that the accuracy can be optimized further in terms of the answer mask. **Conclusion** We develop an answer mask-related method to cover irrelevant answers in the model prediction results and the final influence of irrelevant answers can be alleviated. The model is yielded to learn the corresponding relationship between the question and the answer type, and its challenge can be resolved for the question-irrelevant model's prediction answer to a certain extent, and the model's generalization and accuracy can be optimized as well.

Key words: visual question answering (VQA); language priors; answer clustering; answer mask; answer type recognition

0 引言

视觉问答 (visual question answering, VQA) (Antol等, 2015)任务是人工智能领域新兴的一项研究内容,属于自然语言处理和计算机视觉交叉领域,具有广泛的应用前景,逐渐成为热门研究方向。视

觉问答任务要求模型根据给定的问题和与问题相关的图像给出相应答案,这种需要处理跨模态信息的任务对模型有着非常高的要求。与传统的基于文本的问答任务不同,视觉问答需要模型根据图像回答对应的问题。模型不仅需要使用自然语言处理技术和图像处理技术对问题文本和图像进行处理,还需要融合这两种不同模态的特征来推理答案。相较于

文本,图像模态的特征所包含信息更加复杂,模型很难理解图像中各目标物之间的语义关系和空间关系,而且文本和图像模态的特征属于不同的特征空间,在特征融合时也存在一定困难。因此视觉问答是一个复杂且困难的研究方向。

随着深度学习的发展,视觉问答模型从早期简单的特征融合分类模型转变为深度神经网络模型,准确率得到极大提升。视觉问答任务有专门的数据集,用于训练模型并衡量模型的性能。使用较为广泛的数据集 VQA v2.0(Goyal 等,2017),是通过对 VQA v1.0(Antol 等,2015)进行优化得到的。图1展示了 VQA v2.0数据集的部分样本实例。在视觉问答数据集中,答案通常分为“yes/no”、“number”和“other”3种类型,这3种类型的预测准确率也是判断一个模型优劣的重要标准。目前的模型在 VQA v2.0数据集上已经有了较高的预测准确率。

但 Agrawal 等人(2018)发现,当前的视觉问答模型容易利用数据集中答案的分布规律而不是图像与问题的结合来得到正确答案,在回答某一类问题时,预测的答案往往是训练集中此类问题出现次数最多的答案,模型在预测时仅通过问题就能够得到正确答案,说明视觉问答任务中存在严重的语言先验问题,并且模型未将图像信息充分利用,无法满足视觉问答的任务需求。为了解决这个问题,他们对 VQA v2.0数据集进行了重新分割,得到了 VQA-CP v2.0(countering priors in vqa)数据集。VQA-CP v2.0数据集的训练集和测试集中的答案分布不均衡,模型如果仅通过答案分布规律回答问题,在测试集上的预测准确率将会很低,该数据集由于能够衡量模型受到语言先验影响的程度,因此成为目前视觉问答任务最常用的数据集。当前的模型因为受到语言先验的影响,在 VQA-CP v2.0数据集上的表现并不能令人满意。

与现实生活相同,视觉问答数据集中的每一类问题都有与之对应的一类答案。例如,询问图像中物体数量的问题,答案都是“number”类型的答案,询问“是否”的问题,答案都为“yes/no”类型的答案,这种关系是一一对应的。但是由于语言先验等原因,模型往往只通过问题和答案之间的表面关系来回答问题,没有学习到问题和答案之间这种更深层次的对应关系,存在答非所问现象。模型应该回答“yes/

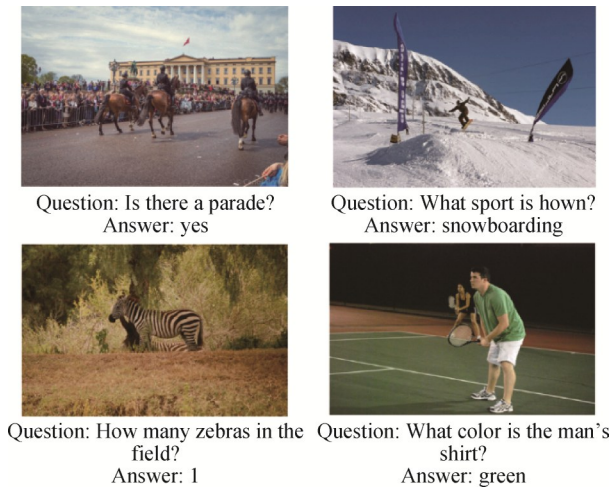


图1 VQA v2.0数据集部分样本实例
Fig.1 Sample examples of VQA v2.0 dataset

no”类型的答案,却回答了“number”类型的答案,如图2所示。

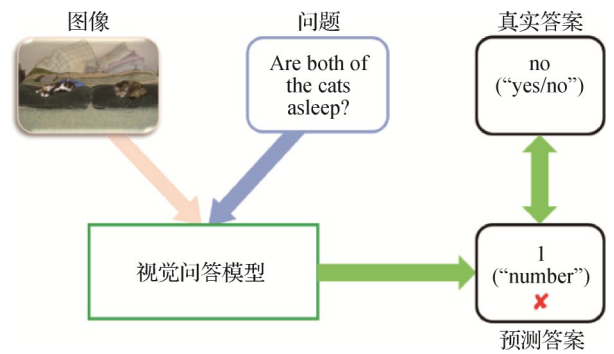


图2 答非所问现象示例
Fig.2 An example of irrelevant answer phenomenon

为了缓解这一现象,本文提出使用答案掩码对模型的预测结果进行遮盖的方法,将预测结果的概率分布限制在问题对应的答案类型中,通过训练使模型学习到问题和答案类型之间的对应关系,提高模型的预测准确率。首先,通过 K-means 算法对数据集中的答案重新聚类,并为每一类答案生成一个由 0、1 组成的答案掩码向量。然后,使用预训练的答案类型识别模型预测输入问题对应的答案类型,并根据该预测结果选择相应的答案掩码。使用基础视觉问答模型得到初步的概率分布向量,并将问题对应的答案掩码与概率分布向量进行融合,遮盖无关答案的概率分布。最后,使用数据集中的样本对模型进行训练,优化基础视觉问答模型的网络参数,缓解答非所问现象,提高模型预测准确率。

1 相关工作

1.1 视觉问答

视觉问答任务自提出以来便引起广泛关注,其流程可以总结为特征提取、特征融合以及答案预测3个步骤(Zhou等,2015)。

视觉问答模型通常使用卷积神经网络(convolutional neural network, CNN)提取图像特征。Simonyan和Zisserman(2014)提出的VGGNet(Visual Geometry Group Network)和Szegedy等人(2015)提出的GoogLeNet是早期视觉问答模型常用的视觉特征提取器,但是由于参数量大、训练速度慢等原因,逐渐被精度更高、训练速度更快的ResNet(residual network)(He等,2016)所取代。目前,视觉问答模型最常用是Faster R-CNN(region-based convolutional neural network)(Ren等,2017)模型。对于问题文本,模型通常使用循环神经网络(recurrent neural network, RNN)提取特征。模型使用GloVe(global vectors for word representation)(Pennington等,2014)获取词向量表示,并通过长短期记忆(long short-term memory, LSTM)或双向循环门控单元(gated recurrent unit, GRU)(Cho等,2014)对词向量进行编码。基于注意力机制的Transformer(Vaswani等,2017)和BERT(bidirectional encoder representation from transformers)(Devlin等,2019)也是目前常用的文本特征提取器。

在特征融合方面,Fukui等人(2016)提出的MCB(multimodal compact bilinear)模型使用双线性池化的方法进行特征融合;Ben-Younes等人(2017)为了解决双线性模型参数量过多的问题,提出了基于Tucker分解的模型,在保证双线性交互的同时大量减少了模型的参数,提高了特征融合的效率;Yu等人(2017)提出多模态分解双线性池化的方法,引入了注意力机制进行特征联合嵌入。除此之外,基于注意力机制的模型也有很好的表现。Yang等人(2016)提出的SANs(stacked attention networks)模型通过问题对图像进行多次堆叠式的注意力引导,查找图像中与问题相关的区域并推测答案;Lu等人(2016)提出的共同注意模型主张两个模态的特征相互引导,同时关注图像和问题的关键区域;Anderson等人(2018)提出的UpDn(up-down)模型使用自底向

上和自上而下注意力相结合的方式融合特征,是视觉问答任务的经典模型,且具有良好的可扩展性,后来的模型大多都在UpDn的基础上进行改进。

除以上方面,Wu等人(2016)从外部知识库中引入了与图像、问题相关的先验知识对模型进行训练,加深模型对图像的理解。Li等人(2019)提出关系感知图注意模型,利用图关系网络和注意力机制辅助模型理解图像中各目标物之间的语义关系和空间关系,提高了模型的性能。

视觉问答模型特征提取的发展依赖于计算机视觉和自然语言处理技术的进步。在当前条件下,视觉问答模型在特征融合方面还可以进行深入研究,基于外部知识和图关系网络的模型也有广阔的研究前景。

1.2 视觉问答中的语言先验和答非所问

Agrawal等人(2016)对视觉问答模型进行分析,发现大部分模型的泛化性很差。Jabri等人(2016)也发现,模型仅学习答案的分布规律就能取得不错的效果。Kafle等人(2017)对视觉问答任务的未来工作进行分析,指出需要提高模型对图像的关注并平衡数据集的偏差。Agrawal等人(2018)明确指出,视觉问答任务中存在语言先验现象,并对数据集进行了重新划分。由于这种语言先验性,模型不能真正理解图像的内容,不能准确学习到问题和答案类型之间的联系,容易出现答非所问现象。

目前,克服语言先验的方法大致分为两类,一类是平衡数据集,另一类是通过对抗或融合的方式抵消语言先验。Chen等人(2020)提出生成反事实样本的CSS(counterfactual samples synthesizing)模型,通过生成的反事实样本平衡数据集。Liang等人(2020)对CSS模型进行改进,通过对比学习使模型更加关注图像和问题的关键信息。Zhu等人(2020)提出SSL(self-supervised learning)模型,引入自监督学习的思想,生成负样本平衡数据集并利用负样本产生的监督信号对基础模型进行指导,增加模型对图像的关注度。Cadène等人(2019)提出RUBi(reducing unimodal biases)模型,通过一个“仅问题模型”来捕获偏差,并将其预测结果生成的掩码向量与基础模型的预测结果进行融合,动态调整损耗以补偿偏差。Clark等人(2019)提出LMH模型,将基础模型的预测结果与“仅问题模型”的预测结果进行融合,减轻了语言先验。

对于由语言先验引起的答非所问现象,也有学者进行研究。Gokhale 等人(2020)提出 MUTANT 模型,使用经过突变生成的反事实样本进行训练,并计算模型预测的答案类型与真实答案类型的损失,迫使模型学习问题和答案类型之间的关系,减少答非所问现象。Jing 等人(2020)提出的模型将问题分解为多种表示形式,计算问题类型表示和答案之间的相关性得分并映射成(0,1)之间的掩码向量,最后使用掩码向量对答案进行遮盖,进而得到预测结果。Guo 等人(2022)通过一个映射函数将融合后的问题和图像特征映射为掩码向量,并对基础视觉问答模型的预测结果进行遮盖。Agrawal 等人(2018)提出 GVQA (grounded visual question answering) 模型,通过问题类型预测模型对问题的答案类型进行预测,并将预测结果与基础视觉问答模型的预测结果相加。贾荫鹏(2020)将答案类型预测模型的预测结果作为“第三模态”特征,与问题和图像特征融合来预测答案。

这些方法都对缓解答非所问现象有所帮助,但是也存一些问题。MUTANT 模型使用的答案类型种类较少,模型学习到的问题和答案类型之间的关系不够精确;GVQA 等使用答案掩码的模型生成的答案掩码不能消除无关答案对结果的影响;贾荫鹏在重新划分答案类别时使用了大量的人工标注,在融

合“第三模态”特征时也引入了一些噪声,影响实验结果。本文方法与上述方法不同,在保证模型能够正确识别问题对应的答案类型的同时,尽可能地将答案类型划分得更细致,使用0、1组成的答案掩码对无关答案进行遮盖,以消除无关答案对实验结果的影响。

2 本文方法

本文方法的总体框架如图3所示。其中,黑色箭头表示数据传输方向,黄色箭头为梯度反向传播方向,⊗为选择器,⊙为点乘。首先,对数据集中的候选答案进行聚类,并为每个答案类型 C 生成答案掩码 m_c 。然后,通过答案类型识别模型预测问题对应的答案类型,并根据预测结果 $P_{qc}(C|Q)$ 选择相应的答案掩码,使用基础视觉问答模型对输入的图像和问题进行预测,得到初步的预测结果 $f_{vqa}(v, q)$;最后,将基础视觉问答模型的预测结果与答案掩码融合,得到最终的预测答案 P ,计算损失并反向传播梯度,优化基础视觉问答模型。

2.1 答案掩码生成

由于数据集中的答案类型划分不够细致,使用原有的答案类型生成的答案掩码无法高效地过滤无关答案,模型无法学习到问题和答案类型之间更细

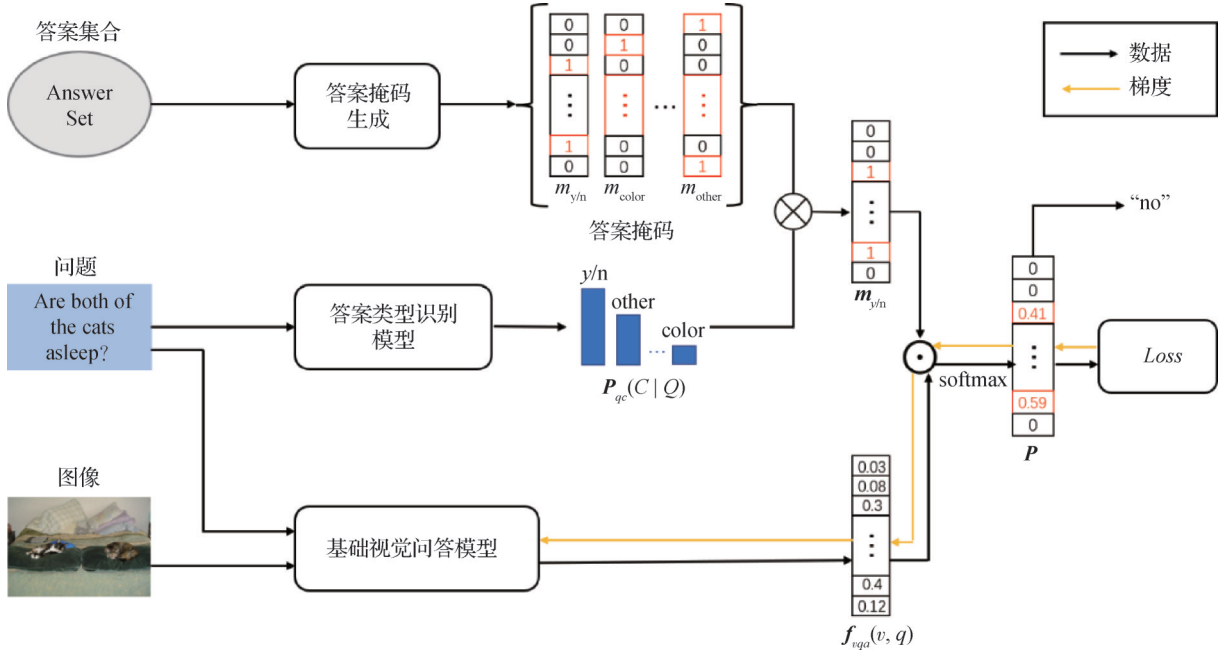


图3 总体框架
Fig. 3 Overall framework

致的对应关系。为了解决这个问题,首先对数据集中的答案进行聚类,由于数据集中的答案数量较多,大部分模型将出现次数超过9次的答案作为候选答案,因此仅对这些候选答案进行聚类。不同数据集的候选答案数量不同,其中VQA v2.0中的候选答案有2 410个,VQA-CP v1.0有1 691个,VQA-CP v2.0有2 274个。图4展示了VQA-CP v2.0中的部分候选答案及出现次数。可以看出,数据集中各答案的数量不同,由单词和短语组成,包含颜色、数字、物体等类别。

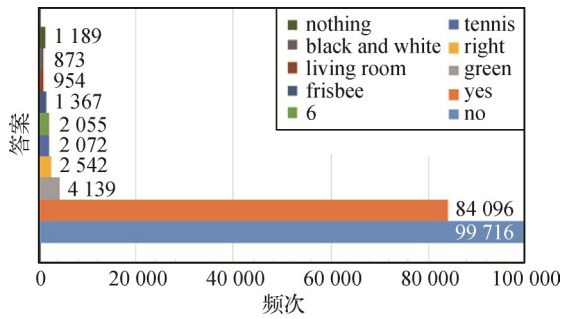


图4 VQA-CP v2.0数据集部分候选答案及出现频次

Fig. 4 Partial candidate answers and frequency of VQA-CP v2.0 dataset

2.1.1 答案特征提取

对答案进行聚类首先需要得到答案的向量表示。数据集中的答案包含单词和短语,并且它们之间不存在上下文关系,传统的 Word2Vec (word to vector) (Mikolov 等, 2013) 和 Glove 不能提取答案特征。由于需要对答案进行聚类,原数据集中的答案类型也不能作为真实标签训练 LSTM 提取答案特征。采用 CLIP (contrastive language-image pre-training) (Radford 等, 2021) 作为编码器提取答案特征。CLIP 使用 4 亿个图像、文本对来训练图像编码器和文本编码器,能够将图像映射到文本描述对应的类别中,预训练的文本编码器提取的文本特征带有丰富的跨模态信息。视觉问答任务中问题对应的答案也来源于图像,使用 CLIP 提取数据集中的答案特征有利于下一步的聚类工作。

2.1.2 答案聚类

使用 K-means 算法将答案聚为 k 类。给定答案向量集 $A = \{a_1, a_2, \dots, a_m\}$, 首先在 A 中随机选择 k 个质心向量 $U = \{\mu_1, \mu_2, \dots, \mu_k\}$, 然后计算第 i 个答案向量 a_i 和第 j 个质心向量 μ_j 的距离 d_{ij} , 将 a_i 划分至与

其距离最小的答案簇 C_j 中, 在答案簇中重新选择新的质心向量 μ'_j 。具体为

$$d_{ij} = \sqrt{(a_i - \mu_j)^2} \quad (1)$$

$$C_j = C_j \cup a_i \quad (2)$$

$$\mu'_j = \frac{1}{|C_j|} \sum_{a \in C_j} a \quad (3)$$

经过多次迭代, 得到最终的 k 个答案簇, 最后根据该聚类结果对数据集中的答案类型进行修改。

2.1.3 答案掩码生成

本文方法使用答案掩码对预测结果中的无关答案进行遮盖, 为每一类答案生成一个由 0、1 组成的答案掩码向量, 用 0 遮盖无关答案的概率分布。

在将答案重新分类后, 为 k 类答案生成 k 个不同的答案掩码。在视觉问答任务中, 同一个模型在加载数据时生成的答案标签是固定的, 答案标签的每个元素都对应一个答案。对于类别 C_j , 首先定义一个向量 $s = \{0\}^{n \times 1}$, n 为候选答案个数。然后确定 C_j 中的所有答案在答案标签中的位置, 并将 s 中对应位置的元素置为 1, 即得到答案掩码 $m_j = \{0, 1\}^{n \times 1}$ 。

2.2 答案类型识别模型

得到答案掩码后, 还需要选择使用哪一类答案掩码对预测结果进行遮盖。为此, 设计了一个答案类型识别模型, 识别问题对应的答案类型, 如图 5 所示。答案类型识别模型主要有两个作用。1) 检验聚类结果是否满足要求。根据问题和答案类型之间的对应关系, 可以通过答案类型识别模型的预测准确率来检验聚类结果是否满足需求。如果模型准确率高, 说明问题和答案类型的关联性强、聚类结果好; 反之则说明聚类结果不能满足要求, 需要重新进行聚类。2) 根据预测结果选择答案掩码。使用答案类型识别模型对问题对应的答案类型进行预测, 并根据预测该结果选择相应的答案掩码。

本文将问题识别答案类型看作一个多分类任务, 设计了一个分类模型来加以实现。给定 N 组样本 $H = \{Q_i, C_i\}_{i=1}^N$, 其中 Q_i, C_i 为第 i 组样本的问题和答案类别, 目标是学习一个映射函数 $f_{qc}: Q \rightarrow C^{k \times 1}$, 通过该映射函数可以由问题得到答案类型的预测结果。

对于输入的问题, 首先通过 GloVe 模型得到词嵌入, 然后通过 GRU 对词嵌入向量进行编码, 获得

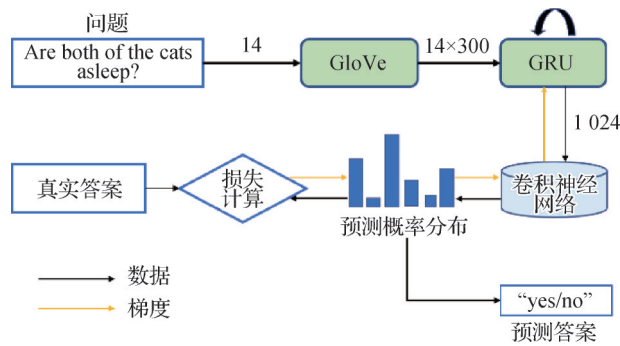


图5 答案类型识别模型整体结构

Fig. 5 Overall structure of answer type identification model

问题特征向量 q , 最后通过分类器对答案类型进行预测。具体为

$$P_{qc} = P_{qc}(C|Q) = \text{softmax}(f_{qc}(q)) \quad (4)$$

得到预测结果后, 计算结果与真实标签的损失 L_{qc} , 并根据该损失优化模型参数。具体为

$$L_{qc} = -\frac{1}{N} \sum_{i=1}^N [g_{c_i} \log_2(P_{qc}) + (1 - g_{c_i}) \log_2(1 - P_{qc})] \quad (5)$$

式中, g_{c_i} 表示第 i 个样本的真实答案类型标签。

本文方法需要根据问题选择正确的答案掩码, 因此要求答案类型识别模型有较高的预测准确率。为此, 使用预训练好的答案类型识别模型与基础模型进行融合, 并且在训练过程中仅对基础视觉问答模型进行优化, 答案类型识别模型不参与反向传播过程。

2.3 模型融合

给定 N 组样本 $D = \{I_i, Q_i, a_i\}_{i=1}^N$, 其中 I_i, Q_i, a_i 为第 i 组样本的图像、问题和答案, 视觉问答模型的目的是学习一个映射函数 $f_{vqa}: I \times Q \rightarrow A^{n \times 1}$, 通过该函数可以生成 n 个候选答案的概率分布。

本文方法具有良好的可移植性, 为了验证该方法的效果, 将其与具有代表性 UpDn、RUBi、LMH 和 CSS 模型作为基础模型进行融合。

UpDn 模型采用 GloVe 对问题文本进行词嵌入, 并用 GRU 提取问题特征向量 q 。对于图像, UpDn 采用基于自底向上注意力的 Faster R-CNN 提取特征 v 。得到问题和图像特征后, UpDn 利用问题特征引导图像特征生成自上而下的注意力, 关注图像中的关键区域, 得到答案的概率分布向量 P_{vqa} , 具体为

$$P_{vqa} = P_{vqa}(A|I, Q) = \text{softmax}(f_{vqa}(v, q)) \quad (6)$$

预测结果与真实答案标签 ga_i 的交叉熵损失 L_{UpDn} 为

$$L_{UpDn} = -\frac{1}{N} \sum_{i=1}^N [ga_i \log_2(P_{vqa}) + (1 - ga_i) \log_2(1 - P_{vqa})] \quad (7)$$

通过式 (7) 得到 L_{UpDn} 后, 计算梯度并反向传播, 调整网络参数。

RUBi 在 UpDn 的基础上增加了一个仅问题模型, 该模型学习映射函数 $f_{qa}: Q \rightarrow A^{n \times 1}$, 根据输入的问题预测相应的答案。仅问题模型可具体表示为

$$P_{qa} = P_{qa}(A|Q) = \text{softmax}(f_{qa}(q)) \quad (8)$$

RUBi 将 UpDn 和仅问题模型融合得到最终的预测结果。具体为

$$P_{QM} = f_{vqa}(v, q) \odot \sigma(f_{qa}(q)) \quad (9)$$

式中, $\odot$ 表示点乘含义, σ 表示 sigmoid 激活函数。

计算该结果与答案标签的交叉熵损失 L_{QM} , 并计算仅问题模型与答案标签的交叉熵损失 L_{Q0} , 两者相加作得到 L_{RUBi} 。

LMH 也引入仅问题模型减轻语言先验问题。与 RUBi 的策略不同, LMH 引入的仅问题模型使用问题类型(问题的疑问词)作为输入进行预测, 用于捕获数据集中的偏差。LMH 将 UpDn 和仅问题模型融合得到预测结果, 具体为

$$P_{LMH} = \text{softmax}(\log_2(p) + g(x) \log_2(b)) \quad (10)$$

式中, $p = f_{vqa}(v, q)$, $b = f_{qa}(q)$, $g(x)$ 是一个可学习的函数, 函数表示为 $\text{softplus}(\omega \cdot h_i)$, ω 为可学习的参数向量, $(\cdot)$ 为实际点乘运算, h_i 为第 i 个样本在模型最后一个隐藏层的状态。

计算交叉熵损失 L_{LM} , 为了防止 $g(x) = 0$ 时模型无法学习到偏差, 在损失中加入惩罚项, 得到最终的损失 L_{LMH} , 具体为

$$R = \lambda H(\text{softmax}(g(x) \log_2(b))) \quad (11)$$

式中, λ 为超参数, $H(z) = -\sum_j z_j \log_2(z_j)$ 。

CSS 使用反事实样本训练模型来减轻语言先验。首先计算问题中各单词或图像的各个区域对最终答案的贡献度, 然后根据贡献度确定问题的关键词或图像的关键区域并进行遮盖, 最后为其分配错误答案, 得到反事实样本。CSS 模型以 UpDn 为基础, 在训练过程中使用 LMH 的方法计算损失并优化模型。

将本文方法与基础模型进行融合。以 UpDn 为例, 对于输入的第 i 组样本, 答案类型识别模型预测该组样本的问题对应的答案类型为 C_j , 因此选择答案掩码 m_j 对 UpDn 模型的预测结果进行遮盖, 得到最终预测结果 P_i 。具体为

$$P_i = \text{softmax}(m_j \odot f_{\text{vqa}}(v_i, q_i)) \quad (12)$$

式中, v_i 和 q_i 为第 i 组样本中的图像和问题特征向量。

最后, 计算预测结果与真实答案标签的交叉熵损失 $Loss$, 并根据该损失调整基础模型的网络参数。

$$Loss = -\frac{1}{N} \sum_{i=1}^N [g a_i \log_2(P_i) + (1 - g a_i) \log_2(1 - P_i)] \quad (13)$$

3 实验方案

3.1 数据集

本文使用视觉问答领域中的3个大型公开数据集对融合模型进行训练和评估。

1) VQA v2.0 数据集。VQA v2.0 分为训练集(train)、验证集(val)和测试集(test)。该数据集的图像都来自 MS COCO (microsoft common objects in context) 数据集 (Lin 等, 2014)。训练集包含 82 783 幅图像、443 757 个问题和 4 437 570 个答案; 验证集包含 40 504 幅图像、214 354 个问题和 2 143 540 个答案; 测试集包含 81 434 幅图像和 447 793 个问题。VQA v2.0 在 VQA v1.0 的基础上进行了优化, 为每幅图像增加了 1 个问题和对应的答案, 对数据集进行了扩充。对于每个问题, 都有两幅相似的图像与之对应, 并且由这两幅图像得到的答案也互不相同。通过增加样本的方式平衡数据集, 迫使模型关注图像来回答问题, 在一定程度上减少了语言先验性。

2) VQA-CP v1.0 数据集。由于 VQA v1.0 训练集和测试集的答案分布相似, 对于同一类问题, 模型不需要关注图像, 只需要回答训练集上该类问题出现次数最多的答案就能够在测试集上取得不错的效果, 存在较强的语言先验性。VQA-CP v1.0 对其进行重新分割, 将训练集和测试集打乱, 通过贪心算法使训练集和测试集的答案分布尽可能不同, 模型无法利用语言先验性在验证集上获得好的效果。

3) VQA-CP v2.0 数据集。虽然 VQA v2.0 通过增加样本的方式减轻了语言先验性, 但是其训练集和验证集上的答案分布仍然相似。例如询问数量的

问题, 训练集和验证集出现最多的答案都为“1”, 不管图像中有多少物体, 模型只需要回答“1”就能在验证集上取得不错的结果, 语言先验依然严重。经过重新分割, VQA-CP v2.0 改善了这种现象, 图6为数据集中较为常见的“How many”和“What color is the”类型的问题在训练集和测试集上的答案分布情况, 由于答案数量较多, 仅展示了出现次数最多的前5个答案及其占比情况。从图6可以看出, VQA-CP v2.0 训练集和测试集答案分布差异较大, 能够更好地衡量模型受语言先验影响的程度, 是目前最常用的视觉问答数据集。本文使用模型在 VQA-CP v2.0 数据集上的准确率作为衡量模型性能的主要指标。

3.2 实验相关设置

实验在 Linux Ubuntu 20.04.1 操作系统上进行, GPU 型号为 GeForce RTX 2080 Ti, 深度学习框架为 PyTorch 1.7.1。实验代码均采用 Python 语言编写, 版本为 Python 3.9.5 和 Python 2.7.5。

本文方法采用的基础视觉问答模型的代码均来自官方, 实验中未对模型的超参数进行修改。设置聚类类别数 $k = 6$, $\text{batch_size} = 512$, 训练次数 $\text{epoch} = 50$ 。

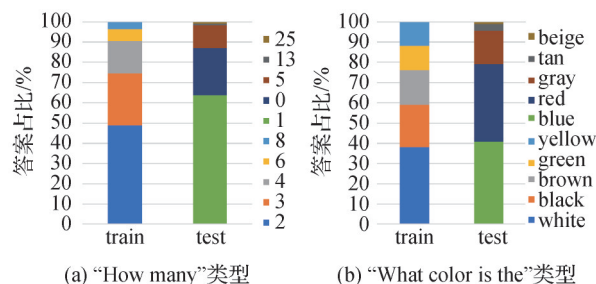


图6 不同类型的问题答案分布情况

Fig. 6 The distribution of different types of questions and answers ((a) "How many" type; (b) "What color is the" type)

4 实验结果与分析

实验主要在 VQA-CP v2.0 数据集上进行, 以模型在 VQA-CP v2.0 测试集上的总体准确率以及在“yes/no”、“number”、“other”3种问题上的准确率作为衡量标准。本文虽然对数据集的答案类别进行了重新划分, 但是为了保证实验结果的公平性和权威性, 仍然按照原始数据集的评分标准进行

评分。

本文方法选择 UpDn、RUBi、LMH 和 CSS 基础模型进行融合实验,为了直观地展示本文方法的有效性和可移植性,将融合后的模型与基础模型在 VQA-CP v2.0 及 VQA v2.0 上的准确率进行比较,如表 1 所示。

表 1 本文方法融合不同基础模型在 VQA-CP v2.0 测试集和 VQA v2.0 验证集上的准确率
Table 1 The accuracies of our method with different basic models on VQA-CP v2.0 test set and VQA v2.0 val set

模型	VQA-CP v2.0 测试集					VQA v2.0 验证集				
	All	Y/N	Num	Other	$\Delta \uparrow$	All	Y/N	Num	Other	$\Delta \uparrow$
UpDn*	39.96	42.83	12.45	46.00		63.44	81.24	41.94	55.61	
UpDn+本文	<u>42.11</u>	<u>54.87</u>	<u>13.07</u>	43.39	+2.15	63.81	81.52	42.79	55.91	+0.37
RUBi*	46.56	68.48	11.52	44.69		60.42	76.98	40.07	53.23	
RUBi+本文	<u>47.25</u>	<u>72.71</u>	<u>11.69</u>	43.66	+0.69	<u>60.75</u>	<u>78.25</u>	40.05	52.93	+0.33
LMH*	53.78	71.46	46.84	46.42		61.55	77.64	40.56	54.89	
LMH+本文	<u>56.07</u>	<u>77.82</u>	46.62	<u>47.26</u>	+2.29	61.41	<u>80.06</u>	38.45	53.30	-0.14
CSS*	58.12	82.22	53.45	46.78		57.12	66.05	39.73	54.94	
CSS+本文	60.14	87.39	52.41	47.98	+2.02	<u>60.30</u>	<u>77.72</u>	38.39	52.86	+3.18

注:加粗字体表示各列最优结果,下划线字体表示较原模型准确率有提高的结果, $\Delta \uparrow$ 表示准确率提升量,*表示重新实验的结果。

由表 1 可以看出,融合了本文方法的 4 种模型在 VQA-CP v2.0 数据集上的准确率都有提高,其中 UpDn 和 LMH 模型分别提高了 2.15% 和 2.29%,CSS 模型提高了 2.02%,达到 60.14%,是目前视觉问答模型在该数据集上的较高水平。RUBi、LMH 和 CSS 都基于 UpDn 模型进行改进,这些模型虽然在 VQA-CP v2.0 数据集上的提高了准确率,但在 VQA v2.0 上的效果并不如原模型。而本文方法不仅提高了模型在 VQA-CP v2.0 数据集上准确率,同时也提高了大部分模型在 VQA v2.0 上的准确率,说明本文方法有更好的泛化性。

为进一步验证本文方法的泛化性,在 VQA-CP v1.0 数据集上进行扩展实验,结果如表 2 所示。可以看出,本文方法使大部分模型的准确率都有所提高。结合在 VQA-CP v2.0 和 VQA v2.0 上的表现,表明本文方法具有良好的泛化性。

表 3 展示了本文方法与一些视觉问答经典模型和表现较好的模型进行比较的结果。其中包括堆叠式注意力模型 SANs(stacked attention networks)、双线性池化模型 MCB(multimodal compact bilinear)、同样关注答案类型的 GVQA(grounded visual question answering)模型、基于自监督学习的 SSL(self

表 2 本文方法融合不同基础模型在 VQA-CP v1.0 测试集上的准确率

Table 2 The accuracies of our method with different basic models on VQA-CP v1.0 test set

模型	VQA-CP v1.0 测试集				
	All	Y/N	Num	Other	$\Delta \uparrow$
UpDn*	38.75	44.14	12.69	43.85	
UpDn+本文	38.63	43.47	12.60	<u>44.26</u>	-0.12
RUBi*	53.55	81.01	12.93	42.57	
RUBi+本文	<u>53.85</u>	80.68	12.11	<u>43.95</u>	+0.3
LMH*	55.08	76.69	24.56	45.87	
LMH+本文	<u>56.80</u>	<u>80.73</u>	<u>30.46</u>	43.58	+1.72
CSS*	60.08	85.93	33.00	45.26	
CSS+本文	60.73	87.34	32.43	<u>45.66</u>	+0.65

注:加粗字体表示各列最优结果,下划线字体表示较原模型准确率有提高的结果, $\Delta \uparrow$ 表示准确率提升量,*表示重新实验的结果。

supervised learning)模型、视觉问答经典模型 UpDn、目前在 VQA-CP v2.0 上表现较好的 RUBi、LMH 和 CSS 模型以及基于 CSS 模型的 CSS+LC 模型。由于

目前表现最好的 MUTANT 模型在生成突变样本时进行了大量的人工标注,并且使用了预训练的 LXMERT (learning cross-modality encoder representations from transformers)模型 (Tan 和 Bansal, 2019), 参数量巨大,未与其进行比较。

从表 3 可以看出,融合本文方法后的 CSS 模型在 VQA-CP v2.0 数据集上的有较好的表现,准确率优于其他模型。

表 3 不同模型在 VQA-CP v2.0 测试集上的准确率
Table 3 The accuracies of different models on the VQA-CP v2.0 test set

模型	VQA-CP v2.0 测试集				/%
	All	Y/N	Num	Other	
SANs	24.96	38.35	11.14	21.74	
GVQA	31.30	57.99	13.68	22.14	
MCB	36.33	41.01	11.96	40.57	
UpDn*	39.96	42.83	12.45	46.00	
RUBi*	46.56	68.48	11.52	44.69	
LMH*	53.78	71.46	46.84	46.42	
SSL	57.59	86.53	29.87	41.42	
CSS*	58.12	82.22	53.45	46.78	
CSS+LC	59.18	86.99	49.89	47.16	
CSS+本文	60.14	87.39	52.41	47.98	

注:加粗字体表示各列最优结果,*表示重新实验的结果。

本文方法在多个模型上都有很好表现,提高了模型性能。为了验证融合本文方法在提高模型精度中的作用,在 VQA-CP v2.0 测试集上进行消融实验,结果如表 4 所示。其中,Q_CSS 代表仅遮盖问题的关键词生成反事实样本,V_CSS 代表仅遮盖图像的关键区域生成反事实样本。可以看出,将本文方法与生成反事实样本的 LMH 模型融合,可以提高模型的准确率。反事实样本和 LMH 起到平衡数据集的作用,本文方法促使模型学习问题和答案类型之间的关系,从而提高了模型的性能。

在答案聚类过程中,需指定答案类别数 k ,即将答案划分为多少类。超参数 k 对模型的准确率有一定影响。 k 值越大,每个答案类型包含的答案数量越少,答案掩码能够遮盖更多的无关答案,但答案类

表 4 本文方法在 CSS 模型上的消融实验

Table 4 Ablation experiment of this method on CSS model

模型	VQA-CP v2.0 测试集上的准确率/%				
	All	Y/N	Num	Other	$\Delta \uparrow$
+CSS	58.12	82.22	53.45	46.78	
+CSS+本文	60.14	87.39	52.41	47.98	+2.02
+Q_CSS	56.66	80.82	45.83	46.98	
+Q_CSS+本文	59.48	87.29	51.20	47.17	+2.82
+V_CSS	58.23	80.53	52.48	48.13	
+V_CSS+本文	58.94	83.27	55.63	47.10	+0.71

注:加粗字体表示各列最优结果,下划线字体为较原模型准确率有提高的结果, $\Delta \uparrow$ 为准确率提升量。

型识别模型的准确率会降低;反之,答案掩码的作用会减小,但答案类型识别模型的准确率会提高。为此进行了超参数实验,结果如图 7 所示。可以看出,随着 k 的增大,问题类型识别模型的准确率不断降低,融合后的 CSS 模型的准确率大体上也呈现降低趋势。在 $k = 6$ 时,答案掩码的覆盖范围和问题类型识别模型的准确率达到平衡,模型的准确率最高,因此取 $k = 6$ 。将 VQA-CP v2.0 数据集中的答案聚类

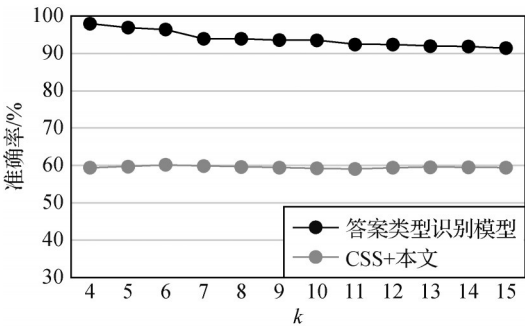


图 7 k 取值对模型准确率的影响
Fig. 7 The influence of k value on model accuracy

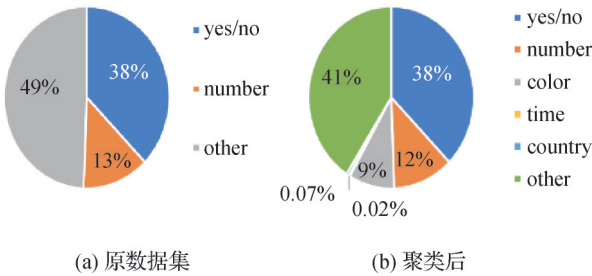


图 8 聚类前后各答案类型分布图
Fig. 8 Distribution of answer types before and after clustering
((a) original dataset; (b) after clustering)

为6类,分别为 yes/no、number、color、time、country 和 other。图8展示了聚类前后各种答案类型在数据集上的分布情况。可以看出,聚类后数据集中的答案类别更多,每个类别所占比重较原数据集也有所不同,生成的答案掩码更加准确,模型能够学习到问题和答案类型之间更细致的关系。

从结果来看,本文方法在总体上提升了模型的预测准确率,提高了模型的性能。对于答非所问现象,对 CSS、LMH 模型以及本文方法在 VQA-CP v2.0 上的预测结果进行了对比分析,本文方法改善了模型答非所问现象,达到了预期效果。图9展示了部分实例的模型预测结果对比。

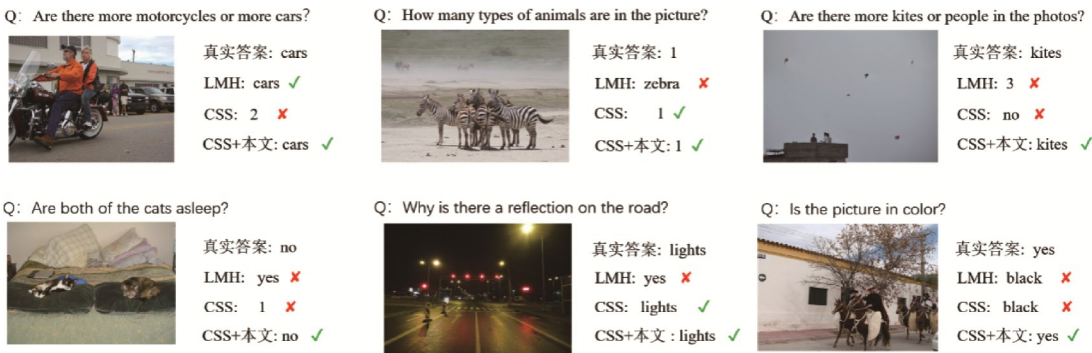


图9 部分实例的模型预测结果对比

Fig. 9 Comparison of model prediction results of partial examples

但是,本文方法对模型准确率的提升程度有限,尤其是在 VQA v2.0 和 VQA-CP v1.0 数据集上,这说明本文方法的泛化性不够强,模型不能准确学习到问题和答案之间更深层次的关系。究其原因,是答案聚类效果不够好、答案类型识别模型准确率不够高所致。下一步的工作重心将放在提高聚类效果及答案类型识别模型准确率上。在提高聚类效果方面,可以使用其他方法获取答案特征。例如,采用视觉问答模型对每个候选答案的预测概率作为候选答案的特征进行聚类。还可以借鉴 Si 等人(2021)提出的答案重排序方法,将视觉问答模型预测结果中概率最高的部分候选答案归为一类,并根据问题类型对这些类别进行合并,得到最终结果。在优化答案类型识别模型方面,可以调整模型的网络结构,提高模型的预测准确率。除此之外,还可以选择性能更好的 LXMERT 等模型作为基础模型与本文方法进行融合。从理论上来说,本文方法适用于各种问答

5 结 论

本文提出一种通过答案掩码对基础视觉问答模型预测结果进行遮盖的方法,减少无关答案对最终结果的影响,通过训练使模型学习到问题和答案类型之间的对应关系。在 VQA-CP v2.0 数据集上的实验结果表明,本文方法能够提高模型的预测准确率。同时,在 VQA v2.0 和 VQA-CP v1.0 数据集上的实验证明了本文方法具有较好的泛化性,在 VQA-CP v2.0 数据集上的消融实验证明了本文方法的有效性。本文方法能够改善答非所问的现象,提高了模型的性能。

任务,可以考虑将本文方法应用于其他问答任务中。

参考文献(References)

Agrawal A, Batra D and Parikh D. 2016. Analyzing the behavior of visual question answering models//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA: Association for Computational Linguistics: 1-13 [DOI: 10.18653/v1/D16-1203]

Agrawal A, Batra D, Parikh D and Kembhavi A. 2018. Don't just assume; look and answer: overcoming priors for visual question answering//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4971-4980 [DOI: 10.1109/CVPR.2018.00522]

Anderson P, He X D, Buehler C, Teney D, Johnson M, Gould S and Zhang L. 2018. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6077-6086 [DOI: 10.1109/CVPR.2018.00636]

- 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]
- Ben-Younes H, Cadene R, Cord M and Thome N. 2017. MUTAN: multimodal tucker fusion for visual question answering//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2631-2639 [DOI: 10.1109/ICCV.2017.285]
- Cadène R, Dancette C, Ben-younes H, Cord M and Parikh D. 2019. RUBi: reducing unimodal biases for visual question answering//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: [s.n.]
- Chen L, Yan X, Xiao J, Zhang H W, Pu S L and Zhuang Y T. 2020. Counterfactual samples synthesizing for robust visual question answering//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10797-10806 [DOI: 10.1109/CVPR42600.2020.01081]
- Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H and Bengio Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics: 1724-1734 [DOI: 10.3115/v1/D14-1179]
- Clark C, Yatskar M and Zettlemoyer L. 2019. Don't take the easy way out: ensemble based methods for avoiding known dataset biases//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics: 1-14 [DOI: 10.18653/v1/D19-1418]
- Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, USA: Association for Computational Linguistics: 1-16 [DOI: 10.18653/v1/N19-1423]
- Fukui A, Park D H, Yang D, Rohrbach A, Darrell T and Rohrbach M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding//Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA: Association for Computational Linguistics: 457-468 [DOI: 10.18653/v1/D16-1044]
- Gokhale T, Banerjee P, Baral C and Yang Y Z. 2020. Mutant: a training paradigm for out-of-distribution generalization in visual Question answering//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. [s.l.]: Association for Computational Linguistics: #63 [DOI: 10.18653/v1/2020.emnlp-main.63]
- Goyal Y, Khot T, Summers-Stay D, Batra D and Parikh D. 2017. Making the V in VQA matter: elevating the role of image understanding in visual question answering//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6325-6334 [DOI: 10.1109/CVPR.2017.670]
- Guo YY, Nie LQ, Cheng ZY and Tian Q. 2022. Loss re-scaling VQA: revisiting the language prior problem from a class-imbalance view. IEEE Transactions on Image Processing, 31: 227-238 [DOI: 10.1109/TIP.2021.3128322]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- Jabri A, Joulin A and van der Maaten L. 2016. Revisiting visual question answering baselines//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer International Publishing: 727-739 [DOI: 10.1007/978-3-319-46484-8_44]
- Jia Y P. 2020. Visual Question Answering Model Based on Answer Type Prediction. Harbin: Harbin Institute of Technology (贾荫鹏. 2020. 基于答案类型预测的视觉问答模型. 哈尔滨: 哈尔滨工业大学)
- Jing C C, Wu Y W, Zhang X X, Jia Y D and Wu Q. 2020. Overcoming language priors in vqa via decomposed linguistic representations. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 11181-11188 [DOI: 10.1609/aaai.v34i07.6776]
- Kafle K and Kanan C. 2017. Visual question answering: datasets, algorithms, and future challenges. Computer Vision and Image Understanding, 163: 3-20 [DOI: 10.1016/j.cviu.2017.06.005]
- Li L J, Gan Z, Cheng Y and Liu J J. 2019. Relation-aware graph attention network for visual question answering//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 10312-10321 [DOI: 10.1109/ICCV.2019.01041]
- Liang Z J, Jiang W T, Hu H F and Zhu J Y. 2020. Learning to contrast the counterfactual samples for robust visual question answering//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. [s.l.]: Association for Computational Linguistics: 3285-3292 [DOI: 10.18653/v1/2020.emnlp-main.265]
- Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48].
- Lu J S, Yang J W, Batra D and Parikh D. 2016. Hierarchical question-image co-attention for visual question answering//Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc: 289-297
- Mikolov T, Chen K, Corrado G and Dean J. 2013. Efficient estimation of word representations in vector space//Proceedings of the 1st International Conference on Learning Representations. Scottsdale,

- USA: [s.n.]
- Pennington J, Socher R and Manning C. 2014. Glove: global vectors for word representation//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics: 1532-1543 [DOI: 10.3115/v1/D14-1162].
- Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sasstry G, Askell A, Mishkin P and Clark J. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR
- Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]
- Si Q Y, Lin Z, Zheng M Y, Fu P and Wang W P. 2021. Check it again: progressive visual question answering via visual entailment//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. [s.l.]: Association for Computational Linguistics: 4101-4110 [DOI: 10.18653/v1/2021.acl-long.317]
- Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s.n.]
- Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]
- Tan H and Bansal M. 2019. LXMERT: Learning cross-modality encoder representations from transformers//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics: 1-12 [DOI: 10.18653/v1/D19-1514]
- 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: 1-15
- Wu Q, Wang P, Shen C H, Dick A and van den Hengel A. 2016. Ask me anything: free-form visual question answering based on knowledge from external sources//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4622-4630 [DOI: 10.1109/CVPR.2016.500]
- Yang Z C, He X D, Gao J F, Deng L and Smola A. 2016. Stacked attention networks for image question answering//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 21-29 [DOI: 10.1109/CVPR.2016.10]
- Yu D F, Fu J L, Mei T and Rui Y. 2017. Multi-level attention networks for visual question answering//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4187-4195 [DOI: 10.1109/CVPR.2017.446]
- Zhou B L, Tian Y D, Sukhbaatar S, Szlam A and Fergus R. 2015. Simple baseline for visual question answering [EB/OL]. [2021-11-19]. <https://arxiv.org/pdf/1512.02167.pdf>
- Zhu X, Mao Z D, Liu C X, Zhang P, Wang B and Zhang Y D. 2020. Overcoming language priors with self-supervised learning for visual question answering//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, Japan: IJCAI.org: 1-7

作者简介

王峰,男,教授,主要研究方向为图像处理和模式识别。

E-mail: wffynu@fynu.edu.cn

石方宇,男,硕士研究生,主要研究方向为视觉问答。

E-mail: 1448571037@qq.com

赵佳,通信作者,男,副教授,主要研究方向为计算机视觉。

E-mail: zhaojia11b@mails.ucas.ac.cn

张雪松,男,硕士研究生,主要研究方向为视觉问答。

E-mail: 1443040527@qq.com

王雪枫,女,硕士研究生,主要研究方向为视觉问答。

E-mail: 906479081@qq.com