基于跨模态对比学习的视觉问答主动学习方法

张北辰[1)] 李 亮[2)] 查正军[3)] 黄庆明[1),2),4)]

[1)](中国科学院大学计算机科学与技术学院 北京 101408)

[2)](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

[3)](中国科学技术大学信息科学技术学院 合肥 230027)

[4)](鹏城实验室 深圳 广东 518055)

摘 要 视觉自动问答技术是一个新兴的多模态学习任务,它联系了图像内容理解和文本语义推理,针对图像和问题给出对应的回答.该技术涉及多种模态交互,对视觉感知和文本语义学习有较高的要求,受到了广泛的关注.然而,视觉自动问答模型的训练对数据集的要求较高.它需要多种多样的问题模式和大量的相似场景不同答案的问题答案标注,以保证模型的鲁棒性和不同模态下的泛化能力.而标注视觉自动问答数据需要花费大量的人力物力,高昂的成本成为制约该领域发展的瓶颈.针对这个问题,本文提出了基于跨模态特征对比学习的视觉问答主动学习方法(CCRL).该方法从尽可能覆盖更多的问题类型和尽可能获取更平衡的问题分布两方面出发,设计了视觉问题匹配评价(VQME)模块和视觉答案不确定度度量(VAUE)模块.视觉问题评价模块使用了互信息和对比预测编码作为自监督学习的约束,学习视觉模态和问题模式的匹配关系.视觉答案不确定性模块引入了标注状态学习模块,自适应地选择匹配的问题模式并学习跨模态问答语义关联,通过答案项的概率分布评估样本不确定度,寻找最有价值的未标注样本进行标注.在实验部分,本文在视觉问答数据集 VQA-v2 上将 CCRL 和其他最新的主动学习算法进行了性能比较,实验结果表明该方法在各个问题模式下均超越之前的方法,该方法对比当前性能最好的主动学习方法在不同的采样率下平均提升了 1.65%的准确率.在仅标注 30%的数据下,该方法可以达到 100%样本标注下性能的 96%;在 40%的标注比例之下,该方法可以达到 100%样本标注下性能的 97%.这说明该方法可以选取出具有高指导价值的样本,节约了标注花费的同时最大化视觉自动问答的模型性能.

关键词 主动学习;跨模态语义推理;对比学习;视觉问答;互信息

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.01730

Contrastive Cross-Modal Representation Learning Based Active Learning for Visual Question Answer

ZHANG Bei-Chen[1)] LI Liang[2)] ZHA Zheng-Jun[3)] HUANG Qing-Ming[1),2),4)]

[1)](*School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing* 101408)

[2)](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[3)](*School of Information Science and Technology, University of Science and Technology of China, Hefei* 230027)

[4)](*Peng Cheng Laboratory, Shenzhen, Guangdonog* 518055)

Abstract Visual question answer (VQA) is a newly developing multi-modal learning task that bridges both the comprehensions of the visual content and the textual question to generate a corresponding answer. It attracts a lot of attention from the community and involves the interaction of different modalities, which requires the capability of image perception and textual semantic

收稿日期:2021-06-15;在线发布日期:2021-11-26.本课题得到科技部科技创新 2030-"新一代人工智能"重大项目(2018AAA0102000)、国家自然科学基金(61732007,61771457,U21B2038)、中国科学院青年创新促进会(20200108)、中央高校基本科研业务费专项资金资助.
张北辰,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为主动学习、半监督机器学习、计算机视觉生成.E-mail:zhangbeichen14@mails.ucas.ac.cn.**李 亮**(通信作者),博士,副研究员,中国计算机学会(CCF)高级会员,主要研究方向为计算机视觉、模式识别、机器学习.E-mail:liang.li@ict.ac.cn.**查正军**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为图像视频分析与检索、多媒体大数据分析、人工智能等.**黄庆明**(通信作者),博士,教授,中国计算机学会(CCF)会士,主要研究领域为多媒体内容分析、模式识别、计算机视觉.E-mail:qmhuang@ucas.ac.cn.

learning. However, the training of VQA has great requirements for the dataset. It requires a wide variety of question patterns and a large number of question answer annotations with different answers for similar scenarios to ensure the robustness of the model and the generalization ability under different modalities. Thus, it is very time-consuming and expensive to label a VQA dataset, which becomes a bottleneck for the development of VQA. In view of these problems, this paper proposes a contrastive cross-modal representation learning based active learning (CCRL) method for VQA. The key idea of CCRL is to cover more question patterns and make the distribution of answers more balanced. It consists of a visual question matching evaluation (VQME) module and a visual answer uncertainty estimation (VAUE) module. The visual question matching evaluation module utilizes mutual information and contrastive predictive coding as the constraints to learn the alignment relationship between visual content and question pattern. The answer uncertainty module introduces the label state learning model. It selects matched question patterns for each image and learn the semantic relationship between cross-modal questions and answers. Then the model estimates the uncertainty of the answer based on the distribution of its probability, by which CCRL can select most informative samples and label them. In the experiment, this work implements the latest active learning algorithms on the VQA task and performs performance evaluation on VQA-v2 dataset. The experimental results demonstrate that CCRL outperforms the previous methods in all question patterns and averagely improves the accuracy by 1.65% compared to the state-of-the-art active learning method. With 30% labeled samples, CCRL achieves 96% of the performance with 100% labeled data. With 40% labeled samples, CCRL achieves 97% of the performance with 100% labeled data. This indicates that CCRL can select instructive and diverse samples, which greatly cuts down the annotation cost and maximizes the VQA performance respectively.

Keywords　active learning; cross-modal semantic reasoning; contrastive learning; visual question answer; mutual information

1　引　言

随着计算机网络技术和硬件计算能力的进步，网络上信息交互日渐频繁，信息类型日益丰富，人们对网络信息也日加依赖. 这导致全球互联网数据的规模剧增，数据的模态趋向多元化，图像、视频、文本等数据形式共存于各种媒体，这推动了跨模态学习任务[1-3]的发展. 近来，跨模态的视觉问答[4-7] (Visual Question Anwser，VQA)任务备受学界和工业界的瞩目，逐渐被应用在社会行为、生物健康、新闻传媒和文娱社交等领域[8-9]，它沟通了计算机视觉和自然语言两种模态，针对图像和问题，能够自动推理对应的答案. 视觉问答作为一个跨模态语义理解的任务，对视觉文本的内容理解有极高的要求，这使得该任务模型的训练依赖大量的、多样的标注样本. 然而，针对视觉问答的跨模态数据标注所需的时间和人力成本远高于普通的单模态任务，数据的采集也更加困难[10]. 同时，为了保证模型的泛化能力和鲁棒性，对问题多样性的需求和答案分布的数据平衡性也提出了更高的要求. 这些问题制约了视觉自动问答技术的发展.

为了解决这些问题，本文引入了主动学习[11]的策略来选择和标注视觉问答数据. 主动学习的目标是选择尽可能少的样本，同时最大化模型的性能. 图 1 展示了针对视觉问答任务的主动学习算法流程. 标注系统只需要对部分有代表性的、有指导价值的图像进行标注，就可以在节省大量标注成本的情况下训练得到一个性能极佳的视觉问答模型. 这其中的关键，在于如何高效的从未标注的数据池中选择出最具代表性的样本，使其具有更加多样的问题模式和平衡的答案分布.

近年来，主动学习已经在诸如图像分类[12-14]和语义分割[15-17]等计算机视觉任务上取得了巨大的

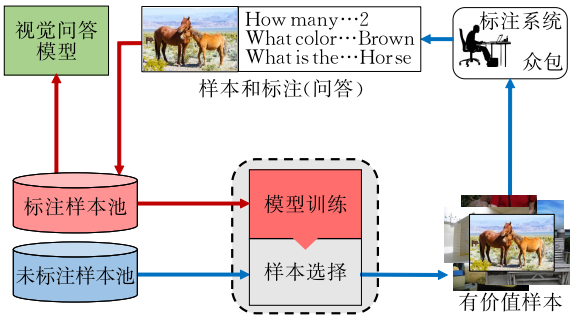

图 1 针对视觉问答任务的主动学习流程(首先使用初始的标注数据对主动学习模型进行训练.训练结束后,根据模型推理选择一个有较高标注价值的未标注样本集合,然后由标注系统进行标注,并转移到标注池.主动学习将重复采样和标注,直到模型性能满足用户的需求,或标注预算全部耗尽)

成功,但是针对视觉问答的主动学习技术尚未被提出.当前的主流主动学习方法可以被归纳为两类.基于数据分布的方法[14,18-19]和基于不确定性的方法[20-22].从数据分布角度,主动学习要求数据具有足够的多样性来代表整个数据集;从不确定性的角度,主动学习要求所选择的数据是模型当前难以理解的,这样才能使得所选数据对模型提升最大.

基于数据分布的方法通过提取数据表示来估计数据在特征空间的分布情况,进而从未标注数据池中选择尽可能覆盖整体数据分布的有价值样本.该方法的出发点在于,如果标注的数据可以代表整体数据集的分布,那么基于这些数据的训练在该数据集上会是充分的,即训练可以覆盖到整体的数据分布.然而,由于学习不同图像模式的难度不同,难度高的类别需要更多的样本,而容易的则需要更少的样本.基于数据分布的方法只考虑了多样性,却忽略了样本类别的泛化难度,这可能会降低采样的价值.另一方面,这些基于单模态的主动学习方法并不适用于视觉问答.视觉问答是一个跨模态的任务,图像和文字问答的信息以不同模态相互纠缠,因此很难用一种统一的数据表征同时刻画出多种模态的语义.传统的主动学习方法致力于分析单模态的数据分布,而忽略了多模态的数据特点,这会导致模型只关注单一模态的语义丰富程度而忽略了视觉问答数据形式下的分布特点.

基于不确定性的方法致力于评估任务模型对未标注数据的不确定程度,其动机是具有高不确定度的数据可以为目标模型带来更大的性能提升.许多主动学习方法在标注信息的监督下训练深度模型,然后基于方差、熵和损失等测量数据不确定性.然而,在主动学习的迭代采样中,样本的标注量在大部分时间内都是不充足的,这导致了该方法无法给出合理的不确定性分数.此外,视觉问答是跨模态任务,其模型输出为针对视觉问题的文本答案,传统的不确定性指标无法准确地在该任务上进行度量.这大大影响了不确定性方法的主动学习采样质量.

此外,视觉问答任务本身也具有很大的挑战.传统的数据集,其数据和标注之间是一对一的关系;然而在视觉问答中,数据为图像,标注分为问题和答案两部分,不但需要标注出最适合图像内容的问题,还要标注问题的答案.因此视觉问答的主动学习不只需要考虑图像内容的丰富程度,还应该考虑图像所关联的问题与答案的多样性.传统的主动学习任务绝大部分是单模态的,它们对跨模态的视觉问答任务只能给出单一模态下数据的标注价值,而不能将不同模态的指标统一量化.同时,因为视觉问答中问题域是共享的,所以要尽量保证答案分布的平衡性,避免视觉问答数据集出现长尾分布等标签不平衡的问题[23],从而降低模型的泛化能力.

基于以上的问题,本文提出了基于跨模态特征对比学习的视觉问答主动学习方法(Contrastive Cross-modal Representation Learning Based Active Learning,CCRL).该方法从问题模式和答案分布两方面入手,主要由两部分组成:视觉问题匹配评价模块(Visual Question Matching Evaluation,VQME)和视觉答案不确定性度量模块(Visual Answer Uncertainty Estimation,VAUE).在视觉问题匹配评价模块内,图像和问题分别基于卷积神经网络和词向量,并通过多头自注意力模块,得到关键区域视觉特征和问题文本序列特征.因为主动学习下只能得到少量的标注,为了充分训练模型的跨模态匹配评价能力,本文在主动学习模型内引入了基于无监督训练的对比学习和互信息约束作为模型更新的策略,将图像特征和不同的问题模式特征随机组合,得到视觉问题匹配对.之后对这些匹配对进行判别,对图像和问题来自于同一分布,即问题可以对应图像视觉语义的,最大化互信息,不匹配的则最小化互信息,同时利用对比预测编码[24-25](Contrastive Predictive Coding,CPC)对比机制约束图像问题的特征空间,以此产生自监督的对抗损失.互信息约束可以对特征提取器进行微调优化,使模型尽可能地提取和视觉问答任务相关的特征,提高特征的可判别性,对比学习则优化了跨模态特征空间,无监督地学习图像的问题匹配关系.此外判别器也在训练中具备了识别图像和问题配对的能力,从而为每一个

未标注图像挑选其适合的问题，得到符合视觉内容的问题模式. 利用已标注的数据训练视觉答案不确定性度量模块，并设计了一个基于问答概率分布的不确定性度量公式对未标注样本进行标注重要性评估. 通过自适应地获取图像所匹配的问题模式并预测其概率分布，该模块可以针对每个未标注样本预测其不确定分数. 此外，模块还分析了已标注数据池的答案标签分布，并以此对标注重要性分数进行规约放缩，进一步保证主动学习采样下标签分布的平衡性. 通过以上两个模块，CCRL 对每个图像都可以获得与之匹配的问题模式的不确定性，之后针对不确定性和答案标签的分布情况，选择问题模式不确定性高，答案分布更加平衡的语义丰富的图像进行问答标注.

在实验部分，本文在通用的视觉问答数据集 VQA-v2[26] 上进行实验，并将最新的、性能最好的主动学习方法在视觉问答任务上实现以进行性能对比. 在多种采样比例之下，CCRL 均取得了最好的效果.

本文第 2 节简要介绍主动学习和视觉问答领域的相关工作；第 3 节描述本文提出的 CCRL 模型架构和理论计算过程；第 4 节提供本文的实验描述、实验结果和分析结论；第 5 节为本文结论和未来工作的概述.

2　相关工作

2.1　主动学习

主动学习在业界和学界已有几十年的研究历史. 传统的主动学习方法可以归纳为以下三种：单个样本查询采样、基于流的选择性采样和基于数据池的采样. 由于互联网的发展和计算能力的提升，大规模数据的获取变得更加容易，模型对标注量的需求也逐年增多，因而近年来的主动学习研究方向主要集中在基于数据池的主动学习采样. 当前，基于数据池的主动学习目前可以分为两类：基于数据分布(distribution-based)的方法[14,18-19] 和基于不确定性(uncertainty-based)的方法[20-22].

基于数据分布的方法旨在通过对最具代表性的样本进行采样来增加已标注数据池的数据多样性. 传统的基于分布的主动学习通过分析训练模型的梯度、误差或预测概率变化来估计数据分布的多样性. Sener 等人[12] 将核心集(Core-set)技术引入主动学习任务. 这种方法评估中间特征进行核心集的采样，在类别较少的数据集上具有良好的性能. 然而，该方法在多类别或高维数据下表现较差. 近年来，一些主动学习方法在模型中引入了对抗学习以充分利用标签的标注信息. Ducoffe 等人提出了 DFAL[27] 方法，使用已标注样本训练特征提取器，而 Sinha 等人提出的 VAAL[22] 和 Wang 等人提出的 DAAL[28] 则通过变分自编码器(Variational Auto-Encoder，VAE)来学习视觉特征和隐空间. 最新的 TA-VAAL[29] 模型则引入基于任务的排序损失，以此将对抗过程的损失进行基于任务的重排序. 上述模型将标注状态映射到二值化的标签，将所有未标注样本的价值视为相等，但是实际上未标注样本的重要性各不相同，某个未标注样本和已标注样本越相似，其标注价值就越低. 此外，视觉问答的问题也会因为同类问题的增多而价值下降，这会降低主动学习采样效率. 此外，近年来许多工作着眼于标签分布的平衡性问题，TA-VAAL[9] 通过探究分类器分类边界，在控制边界区域样本标签平衡的同时，获取对机器学习任务贡献最大的样本. Vab-AL[14] 利用生成模型为未标注样本拟合标签并以此判别标签的置信度，以此控制数据集标签分布的平衡性. 上述最新方法中的表征仅来自于视觉模态，而视觉问答作为跨模态任务，其数据分布于视觉特征空间和文本语义空间，这种单模态的主动学习方法很难对跨模态样本的多样性进行合理评估[10]. 同时视觉问答任务的文本为不定长的序列，标签边界和拟合标签的思路在该跨模态数据下效果不佳.

基于不确定性的方法通过估计样本的不确定程度，选取能够给模型带来更多性能提升的样本进行标注. 传统方法通过计算分类误差[30] 或者高斯过程[31] 的置信度来估计不确定性. 这些方法在特定任务中表现良好，但不能扩展到大规模数据集且采样效果不稳定. Yoo 等人[16] 则提出了一种基于损失学习(Learning Loss，LL)的主动学习模型，但该方法的损失预测精度受任务模块性能影响较大，在采样早期视觉问答模型受制于标注量不足，损失预测的精度不足. Agarwal 等人[19] 提出了基于上下文的主动学习方法，将上下文信息作为监督信息训练不确定性模型进行核心集采样. MOCU[21] 模型则利用梯度更新策略，使不确定度和损失一起迭代更新，获取不确定度估计. UncertainGCN[32] 和 MPART[33] 则引入图的知识，前者构建样本间的分布图结构以评估不确定度，后者结合半监督学习，对未标注样本计

算弱监督标签参与模型更新.但是在跨模态的视觉问答任务中,多种模态的交互导致上下文关系的获取变得困难,对数据的不确定性评估能力也因此大幅下降.同时视觉问答模型的输入输出涉及语言和视觉两种模态,损失预测很难对多种模态数据进行不确定性的评估.

2.2 视觉问答

近几年来,视觉问答任务[4-6]越来越受到学界和业界的关注.该任务涉及视觉数据和文本信息,对跨模态的语义理解有较高的要求.视觉问答最简单的解决方案是对全局特征进行多模态融合[34].然而上述多模态融合方法会导致局部图像信息的丢失,进而失去关键区域的视觉语义信息.因此,近来的方法将视觉注意力机制引入到视觉问答中.Anderson 等人[5]提出了自下而上的注意力机制并从多个尺度进行跨模态融合.MCAN[6]通过协调沟通跨模态注意力查询,学习给定问题的关联图像特征.Bosselut 等人[35]和 Ren 等人[36]使用图神经网络建模问答逻辑关系,前者建模了语义关系,后者着重于优化语义空间挖掘因果关系.Xiong 等人[37]借鉴蒸馏网络的思想,通过教师模型进行自我提问以优化答案推理.Urooj 等人[38]提出了一个基于胶囊网络的弱监督图像问答模型.本文中为了测试主动学习样本挑选的质量,需要使用一个视觉问答模型作为评价模型,而此模型和主动学习采样算法无关.因此本文采用了图像问答领域更有影响力的 MCAN 模型作为任务模型,测试主动学习采样质量.

3 视觉问答主动学习方法

3.1 符号定义和任务介绍

本节约定了符号的使用,并对视觉问答的主动学习任务进行了详细介绍.在主动学习中,给定一个未标注的视觉问答数据集 D 和视觉问答任务模型 Θ.在主动学习初始阶段,数据集 D 中的 M 个样本会被随机选中并通过标注系统获得它们的问题答案标注,它们组成了初始的标注数据池 D_L,其余未标注数据组成了初始的未标注数据池 D_U.主动学习算法的关键在于从未标注池 D_U 中选择最有指导意义的样本进行标注.x_U 表示无标签数据池中的一个图像样本,(x_L, q_L, y_L) 表示有标签数据池中的一个样本,包含图像 x_L,问题 q_L 以及答案 y_L.初始化之后,主动学习模型将迭代地选择固定数量的样本进行标注,并将其从 D_U 转移到 D_L,以更新有标注和未标注数据池.如图 1 所示,这个过程将重复进行,直到任务模型的性能满足用户的要求,或者数据标注的预算用尽.

不同于传统的主动学习任务(图像分类和语义分割等),视觉问答主动学习是一个跨模态的任务,视觉问答的标注由问题和答案两部分组成,因此应用于该任务的主动学习采样需要同时考虑问题标注和答案标注的标注价值.针对此特点,本文提出了基于跨模态特征对比学习的视觉问答主动学习方法(Contrastive Cross-modal Representation Learning Based Active Learning,CCRL).如图 2 和图 3 分别展示了本文 CCRL 方法中的视觉问题匹配评价(Visual Question Matching Evaluation,VQME)模块和视觉答案不确定度度量(Visual Answer Uncertainty Estimation,VAUE)模块.视觉问题匹配评价模块对问题图像进行随机配对,并通过互信息约束和对比学习的方式,对来自同一分布的图像问题配对最大化其互信息(Mutual Information,MI)并拉近其特征空间的距离,对来自不同分布的图像和问题配对,最小化其互信息并扩大其特征空间的距离,从而使得特征生成器可以学习到具有较高相关度的特征,并使判别器具有判别图像问题是否匹配的评价能力.针对未标注图片获得其匹配问题之后,视觉答案不确定度度量模块进行问答预测,并给出答案的概率分布和不确定度分数估计,选择对模型训练最有益的样本进行标注,节约标注成本的同时最大化模型性能.

3.2 视觉问题匹配评价 VQME

视觉问答主动学习的标注由问题和答案两部分组成,且涉及了多种模态,因此如何对问题和答案两部分的数据重要性进行衡量是该任务的难点.基于此,本文提出了视觉问题匹配评价模块 VQME.如图 2 所示,该模块对图像和问题文本分别构建特征空间,并对跨模态特征进行随机的配对,利用互信息最大化和 CPC 对比学习(Contrastive Learning)的机制使 VQME 可以无监督地学习跨模态图像问题特征,并给出图像和问题配对的匹配分数.通过匹配分数,CCRL 模型即可判断未标注样本和哪些问题模式的数据分布更为接近,进而挑选出更具有询问价值的、提升视觉问题模型性能的样本进行问答标注.

对于视觉模态,通过分析可以得知,视觉问答大多涉及的是图像中的关键物体的问答,问答内容几

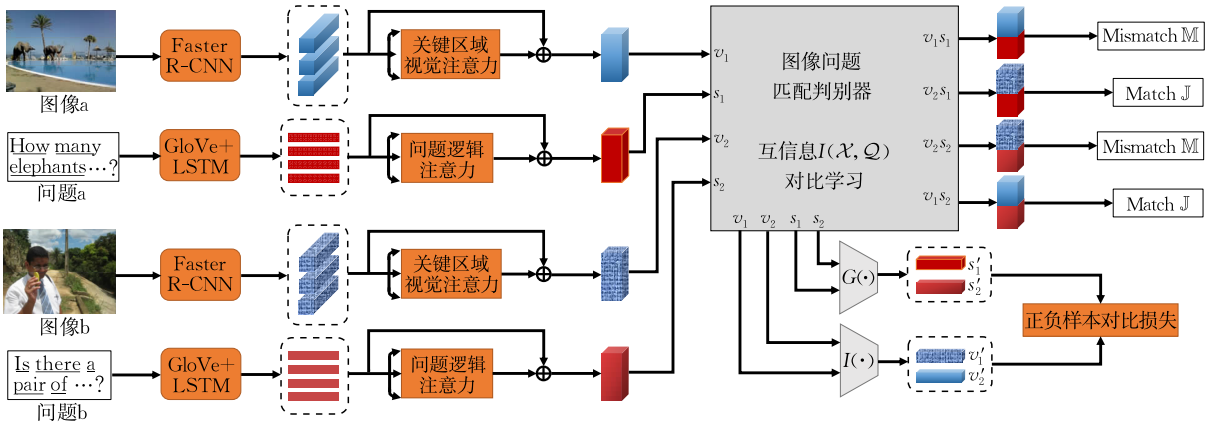

图 2　视觉问题匹配评价模块 *VQME*（图像 a 与问题 a，图像 b 与问题 b 是相互匹配的. 使用基于 *Faster RCNN* 的检测模型对图像检测关键物体区域，并提取关键区域视觉特征. 使用 *GloVe* 和 *LSTM* 模型提取问题文本序列特征. 之后使用图像问题注意力机制，分别对关键区域视觉特征和文本序列特征进行权重学习，重组视觉和文本特征. 之后，对不同的视觉文本特征进行随机配对，使用对比学习的机制和互信息最大化作为约束，训练模型判断图像问题配对的匹配程度）

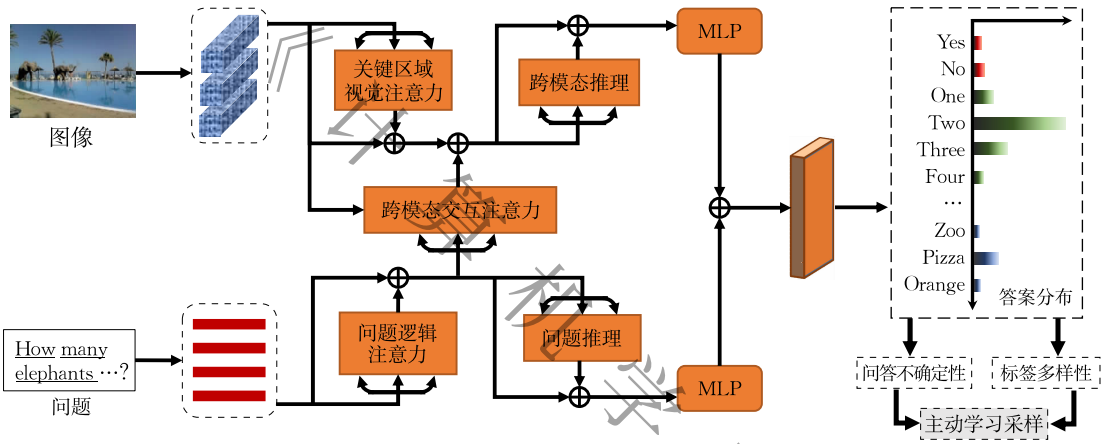

图 3　视觉答案不确定度度量模型 *VAUE*（该模型输入图像和问题，输出答案的概率分布. 根据概率的分布评估模型对于该图像问题的配对不确定度，同时评估标注数据池的答案分布，综合给出样本的标注重要性）

乎不会涉及图像的背景内容[39]. 而传统的卷积神经网络所得到的特征图过于笼统，模型很难捕捉到问答相关的关键信息. 鉴于此，VQME 同其他 VQA 任务[6,35-37]一样，都采用了基于 Faster-RCNN[40] 结构的图像检测模型，将图像内的关键目标物体检测并定位，之后提取关键区域的视觉特征，得到关联图像关键目标的关键区域视觉信息. 在特征提取时，为了保证所有关键目标区域均被提取，本文设计了对关键目标根据排序的算法：首先将关键区域按照置信度排序，之后根据区域重复关系对置信度进行适度缩放，重排序后按照重要性选取前 36 个关键区域，尽可能保证关键目标的全覆盖. 其计算过程如下所示：

$$V=\{v_i\}_{i=1}^{N}=f_{\text{CNN}}(f_{\text{R-RCNN}}(x)) \tag{1}$$

其中 $N=36$ 是关键目标个数，f 为神经网络计算过程，$V=\{v_i\}_{i=1}^{N}$ 是得到的 N 个关键区域视觉特征.

此时，模型得到多个关键区域的视觉信息. 为了增强这些区域信息之间的联系，同时使模型有选择性地对这些区域进行关注，VQME 引入了自注意力机制（self-attention）对关键区域视觉特征计算关键区域视觉注意力. 自注意力计算如下所示：

$$selfAtt(q,K,V)=\text{softmax}\left(\frac{qK}{\sqrt{d}}\right)V \tag{2}$$

其中 q 为查询变量，K 为关键变量，V 为查询数据库，d 为变量维度.

此外，该模型中的自注意力模块引入了多头机制，以此能够帮助模型更加充分地学习图像语义并挖掘各个关键物体之间的逻辑关系等. 多头机制使模型从更多的维度和更多的特征子空间上对视觉特征施加不同的注意力. 基于关键区域提取和注意力集散式（2），关键区域视觉注意力的计算过程如下所示：

$$\begin{aligned}&\boldsymbol{head}_j=\left[selfAtt(v_i\boldsymbol{W}_j^{Q},V\boldsymbol{W}_j^{K},V\boldsymbol{W}_j^{V})\right]_{i=1}^{d},\\&v_{att}=Concat(head_1,head_2,\cdots,head_h)W^{0}\end{aligned} \tag{3}$$

其中自注意力 $\boldsymbol{head}_j$ 为第 j 个注意力向量，$\boldsymbol{W}^Q$，$\boldsymbol{W}^K$，$\boldsymbol{W}^V$ 为第 j 个注意力下的投影计算矩阵. 得到全部的注意力权重后，将其连接成一个向量，经过 W^0 进行单层感知机计算后，得到关键区域视觉注意力特征. 该计算模块建模了多个关键区域视觉特征之间的关联，以此计算了 h 个注意力向量并组成了多头的关键区域视觉注意力.

对于文本模态，VQME 通过 GloVe[41] 模型对问题进行词向量化，之后使用长短期记忆网络[42](Long-Short Term Memory，LSTM)，对文本数据进行序列化迭代处理，提取更好的问题文本表征. 其计算过程如下：

$$S=\{s_i\}_{i=1}^{\mathrm{LEN}},$$
$$s_i=\mathrm{LSTM}(s_{i-1},\mathrm{GloVe}(w_i)) \tag{4}$$

在每一个时刻 i，LSTM 接受上一时刻的输出 s_{i-1} 以及第 i 个位置的单词 GloVe 向量.

获得 LSTM 文本特征之后，模型问题文本序列产生了语义理解，但是无法整体获取问题文本的语义逻辑. 为了加强模型对问题文本的逻辑理解力，为后续的跨模态对比学习和图像问题匹配提供语义可分的特征空间，本文也采用了多头原理计算了问题逻辑注意力权重，计算如下所示：

$$\boldsymbol{head}_j=[selfAtt(s_i\boldsymbol{W}_j^Q,S\boldsymbol{W}_j^K,S\boldsymbol{W}_j^V)]_{i=1}^d,$$
$$v_{att}=Concat(head_1,head_2,\cdots,head_h)W^0 \tag{5}$$

除关键视觉特征替换为 GloVe 词向量序列，计算方式和关键区域视觉注意力部分基本一致.

获得视觉特征和问题文本特征后，VQME 需要评估图像和问题的匹配程度. 通常此类问题采用余弦相似度(Cosine)或者欧氏距离，计算特征的相似度. 但是这种方法在该任务下存在很多问题. 首先受限于标注量，数据集无法提供足够的标注信息训练一个特征对齐的深度网络；其次常规的跨模态特征对齐技术，其效果不足以支撑主动学习选取高质量样本.

因此，本文在该任务中引入了互信息最大化和对比学习机制，利用正负例样本的对抗，最大化样本和特征的互信息，同时无监督地训练匹配评价判别模型. 互信息的定义如下：

$$\mathcal{I}(X,Z)=H(X)-H(X|Z)$$
$$=\int_x\int_z p(z|x)\tilde{p}(x)\log\frac{p(x,z)}{p(x)p(z)}\mathrm{d}z\mathrm{d}x \tag{6}$$

其中 X 和 Z 为输入集合和隐变量集合，H 是信息熵公式：$H=\sum_i p_i\times\log p_i$. $\tilde{p}(x)$ 为 X 的先验分布.

互信息衡量了两个随机变量之间相互依赖程度的度量，它可以看成是一个随机变量中包含的关于另一个随机变量的信息量. 基于此，VQME 通过最大化互信息来对特征提取部分进行优化.

为了设计一个端到端的模型，本文将互信息最大化和视觉问题匹配评价合二为一，利用了对比学习的机制同时约束这两个条件，使特征生成部分和匹配评价部分在对抗过程中共同提升. 为了约束互信息，本文引入了 JS 散度[43](Jensen-Shannon Divergence，JSD)以评估互信息. JSD 定义如下：

$$D_{\mathrm{JS}}(P\|Q)=\frac{1}{2}D_{\mathrm{KL}}\left(P\middle\|\frac{P+Q}{2}\right)+\frac{1}{2}D_{\mathrm{KL}}\left(Q\middle\|\frac{P+Q}{2}\right),$$
$$D_{\mathrm{KL}}(P|Q)=\int_i P(i)\ln\frac{P(i)}{Q(i)} \tag{7}$$

其中 D_{KL} 为KL散度[43](Kullback-Leibler Divergence，KLD). 引入 JSD 后，互信息的评估可以被表示为如下形式：

$$\mathcal{I}(X,Z)=D_{\mathrm{JS}}(p(z|x)\tilde{p}(x)\|p(z)\tilde{p}(x)) \tag{8}$$

可以发现，KL 散度是无上界的，而 JS 散度则存在明确的上确界 $\frac{\log 2}{2}$. 这使得互信息的优化更加可靠稳定. 之后，借鉴 Nowozin 等人[44]的 f 散度函数共轭推导，可以得到基于 JSD 的互信息估计公式如下：

$$D(P|Q)=\mathrm{E}_{x\sim P}[\log\sigma(T(x))]+\mathrm{E}_{x\sim Q}[\log(1-T(x))],$$

其中 σ 为判别器函数，T 将高维向量转换到实数域.

基于式(8)，可以得到互信息最大化的目标函数如下：

$$L_{\mathrm{MI}}=\mathbb{E}_{(v,s)\sim p(s|v)\tilde{p}(x)}[\log\sigma(T(v,s))]-$$
$$\mathbb{E}_{(v,s)\sim p(s)\tilde{p}(x)}[\log(1-\sigma(T(v,s)))] \tag{9}$$

其中 v 表示视觉特征，s 表示文本特征. 对于判别器 σ，在训练过程中的对比损失会更新判别器权重，促使判别器可以更好地分辨图像和问题是否是相互匹配的. 交替的对抗训练可以有效地提高模型的鲁棒性，增强模型对两种不同模态数据的语义理解能力，从而给出一个精确的匹配程度得分.

为了进一步提升特征提取的质量，本文借鉴了 CPC[24](Contrastive Predictive Coding)损失和 SimCLR[45]工作的特征投影机制，引入了对比学习学习图像问题的匹配关系. 因为视觉特征 v 和文本特征 s 不在同一特征空间内，本文分别设计了视觉投影器 $I(\cdot)$ 和文本投影器 $G(\cdot)$，分别将视觉和文本特征投影到相同的特征空间. 之后利用对比学习的机制，约束匹配的特征距离更近，不匹配的特征距离更远. 对于一个 batch 为 B 的图像问题组合，跨模

态对比学习损失函数如下：

$$l_{\mathrm{CL}}=-\log\frac{\sum_{i=1}^{B}\exp(v_i'\cdot s_i'/\tau)}{\sum_{i\neq j}\exp(v_i'\cdot s_j'/\tau)} \tag{10}$$

其中 $v_i'=I(v_i)$，$s_i'=G(s_i)$ 为投影后的视觉和文本特征，下标相同表示匹配，不同则不匹配，τ 为温度超参数，调节对比损失. 对比学习损失函数中的分子项可以约束模型拉近匹配特征的相似度，而分母项则约束模型拉远匹配特征的相似度. 通过两项对比，为特征嵌入关键语义信息，优化跨模态特征空间分布，使其更加可分.

根据式(9)和式(10)，本文设计了如图 2 所示模型结构. 在获取视觉特征 v 和问题文本特征 s 之后，模型对不同的 v 和 s 进行随机组合，如果 v 和 s 来自同一个样本，即问题和图像内容是匹配的，则视为正例，如果来自不同样本，即问题和图像是不相干的，则视为负例. 之后，在互信息最大化部分，模型评估图像问题配对的互信息大小，并根据是否匹配的监督信息优化判别器和特征提取部分. 在对比学习部分，对跨模态的特征分别进行投影，利用对比学习的机制拉近匹配特征的距离，拉远不匹配特征的距离，从而在正负例学习中提升特征的可判别性，构建一个基于对比学习机制的匹配评价模型.

对于视觉和文本的特征生成器，Faster-RCNN 和 GloVe 模型均使用了预训练的权重. 在对比学习过程中，对比损失依旧会对特征生成器进行微调. 正负例的学习过程可以优化生成的特征空间，在无监督的对比下使特征具有相当的可分性.

在对比学习中，特征学习器可以类比于 GAN 的生成模块，而互信息最大化和 CPC 对比学习模块可类比于判别模块. 不同于生成对抗模型，对比学习中生成模块和判别模块的损失是相同的，即对比损失使得生成的特征更加具有判别性，同时使得判别模块更加精准，两者的训练目标相同. 因此对于对比学习，其生成模块(特征学习器)和判别模块的梯度更新方向是一致的，两者的收敛性基本上同步. 基于该收敛特性，本文在实验中以判别器的收敛性和准确率来判断对比学习特征提取的收敛程度，在判别器收敛平稳、判别准确率不再下降后停止 VQME 模块的训练.

3.3　视觉答案不确定度度量 VAUE

训练充分的视觉问题匹配评价模块 VQME，可以将所有的未标注图像与各种问题模式进行匹配，得到匹配分数. 匹配分数高的问题，则被认为是更适合该图像的问题. 通过统计已有数据池内的标注问题和不同问题类型的问答准确率，可以得出当前标注数据池下，模型对哪些问题具有较低的识别解答能力，进而主动学习可以去选择这些困难问题对应的图像进行标注，最大程度提升模型性能. 但是，仅仅通过问题和图像的匹配关系选择样本，对于主动学习来说仍旧是不充分的. 视觉问答问题的特殊性在于标注分为问题和答案两部分，VQME 提高了问题的多样性，答案标签分布的不平衡问题仍然存在. 在现实应用中，随机构建的数据集极有可能存在数据标签的长尾分布问题，而问答标注中答案的分布不平衡，也会导致标注训练集对视觉问答模型的训练效果降低，影响其回答尾端问题的性能.

针对问答的答案标注部分，本文提出了视觉答案不确定度度量(VAUE)模块，结构如图 3 所示. 本文引入自注意力和问答注意力的结构对视觉特征和问题特征进行分别处理. 其中，关键区域视觉注意力和问题逻辑注意力的计算过程同式(3)和式(5)，对关键区域视觉特征和文本序列特征进行多头自注意力权重的学习，组合得到新的注意力特征.

分别学习不同模态特征后，为了解决跨模态数据语义不对齐的问题，VAUE 在视觉特征内引入了问题文本，在视觉问答处理中间过程进行了跨模态的语义交互. 利用问题文本特征计算注意力权重并使用权重对视觉特征进行重组. 此计算过程如下所示：

$$\begin{aligned}&\boldsymbol{head}_j=\left[selfAtt(v_i\boldsymbol{W}_j^{Q},S\boldsymbol{W}_j^{K},S\boldsymbol{W}_j^{V})\right]_{i=1}^{d},\\&v_{att}=Concat(head_1,head_2,\cdots,head_h)W^{0}\end{aligned} \tag{11}$$

上式使用视觉模态的注意力特征作为查询值，和文本问题模态的注意力特征做注意力计算，通过跨模态的交互利用文本问答信息指导视觉特征的注意力分布. 这种跨模态的数据交互使 VAUE 能够更好地评估图像问答的概率分布.

VAUE 分别对图像和问题进行特征提取并组合特征，之后使用 softmax 分类器预测问题的答案，计算过程如下：

$$p(y|x,q)=\mathrm{softmax}(concat(v,s)) \tag{12}$$

VAUE 的答案分类器部分，在每轮主动学习标注结束后，都会使用最新的标注数据池重新标注一遍，以保证得到的答案概率分布 p 和主动学习采样是基于已有的标注信息得到的. 在每轮采样前，VAUE 使用

标注数据池的数据进行训练，目标函数如下所示：

$$L_{\mathrm{VAUE}}=\mathrm{E}[\log[p(y_L|x_L,q_L)]],\ (x_L,y_L,q_L)\in D_L \tag{13}$$

得到每个答案项的概率分布之后，VAUE 会分析图像问题配对的不确定度. 不确定度高说明模型对该图像下的问题理解较差，进而说明标注此样本能够使得视觉问答模型的性能得到较大的提升. 传统的主动学习方法大多使用方差(variance)和交叉熵(cross entropy)等[31]指标作为不确定度的度量，但是这些指标都存在无法准确度量分类置信度的问题. 基于此，本文使用了一种新的不确定分数作为配对不确定度的评估公式，该公式如下所示：

$$f_{\mathrm{uct}}(x,q)=1-\frac{HC(p(y|x,q))}{Var(p(y|x,q))}\times p_{\max}(y) \tag{14}$$

其中 Var 为方差，$p_{\max}(y)$ 为预测概率中最大的概率. HC(Highest Concentration)如下所示：

$$HC(\boldsymbol{p}(y|x,q))=\frac{1}{C}\left(\left(\frac{1}{C}-P(y_{\max}|x)\right)^2+(C-1)\left(\frac{1}{C}-\frac{1-P(y_{\max}|x)}{C-1}\right)^2\right) \tag{15}$$

其中 C 为答案项的个数. HC 衡量了当前概率分布的最大集中度. 对于概率向量 $\boldsymbol{p}(y|x,q)$，如果概率的分布越集中，则其他的概率项被选中的可能就越低，那么分类器就有更高的置信度选择集中程度高的概率项作为预测值，那么此次预测的配对不确定度就会越低. 由式(15)，可以得到本文提出的不确定式(14)具有三条数学性质：(1) $f_{\mathrm{uct}}\in[0,1]$；(2) f_{uct} 和最大概率成负相关；(3) f_{uct} 和概率向量的集中程度成负相关. 上述性质保证了不确定度归约到[0,1]区间，方便进行比较. 同时性质(2)和(3)保证了不确定度的衡量同时考虑了最大概率项和其他概率项的整体分布.

通过上述公式，VAUE 评估了问答模型对当前图像问题配对的不确定度. 此外，VAUE 还考虑了答案分布的多样性，对于标注数据池中频繁出现的答案项，对其不确定度进行适当的缩放，以保证最后得到的标注数据集，是一个尽可能平衡的数据集，防止出现长尾分布. 基于此，VAUE 会分别统计标注数据池内三类问题(是否、数量和其他)的答案分布情况，之后将每项答案占比作为系数规约到不确定性分数上，从而得到最终的主动学习标注重要性. 由于每一张图像可能存在多种合适的问题，因此标注重要性会综合每张图像所有的匹配问题，给出总的衡量分数. 计算方式如下所示：

$$score(x)=\sum_i f_{\mathrm{uct}}(x,q_i)\times rate(y_{\max}) \tag{16}$$

其中 i 为匹配图像 x 的第 i 个问题，$rate(y_{\max})$ 衡量了概率最大项 $y_{\max}$ 在已标注数据池内的同类问题中所占的比重. 通过同类已标注问答的数量，平衡数据集的标签分布.

通过式(16)的计算，VAUE 对未标注样本进行了不确定度评估，同时根据标注数据池内已有的答案标注，对不确定度进行了数值调整，从而促使模型更倾向于选择能够平衡标签分布的样本进行问答标注，在保证主动学习采样的高效性的同时，又防止最后的标注数据集出现长尾分布等标签不平衡的问题.

3.4 整体算法流程

根据 VQME 和 VAUE，本文构建了 CCRL 的视觉问答主动学习方法. 总体算法流程如算法 1 所示，在训练阶段，通过随机组合得到的视觉注意力特征和问题文本注意力特征，构造正例和负例，通过互信息最大化和 CPC 对比学习联合训练 VQME，并利用标注数据池的数据训练 VAUE 的视觉问答部

算法 1. CCRL 算法.

输入：VQME 模块 Θ_1，VAUE 模块 Θ_2，所有数据 D，标注预算 B，每轮标注量 I，训练迭代次数 $epoch$，学习率 α_1 和 α_2

输出：数据量为 B 的标注数据池

随机选择 I 个未标注数据进行标注，初始化标注数据池 D_L、未标注数据池 $D_U=D-D_L$
WHILE $size(D_L)<B$ DO
  FOR $e=1$ to $epoch$ DO
    计算 L_{VQME} 和 L_{VAUE}
    更新 VQME：$\Theta_1'=\Theta_1-\alpha_1\nabla L_{\mathrm{VQME}}$
    更新 VAUE：$\Theta_2'=\Theta_2-\alpha_2\nabla L_{\mathrm{VAUE}}$
  END FOR
  统计 D_L 内答案分布
  使用 VQME 计算所有未标注图像 x_U 匹配的问题模式$\{q_i|\mathrm{VQME}(x_U,q_i)>0.5\}$
  为所有匹配的图像问题对(x_U,q_i)，计算不确定度分数 $f_{\mathrm{uct}}(x_U,q_i)$
  综合答案分布，计算每个未标注样本的标注分数：
  $score(x_U)=f_{\mathrm{uct}}(x_U,q_i)*rate(y_{\max})$
  选取分数前 I 的样本：
  $U=\underset{x\in D_U}{\arg\max}(score(x))$
  $D_L=D_L\cup U$
  $D_U=D_U-U$
ENDWHILE
RETURN 最终的标注数据 D_U

分.之后 VQME 对所有未标注的图像和问题模式进行匹配评价,将匹配分数较高的图像问题组输入 VQUE,通过答案的概率分布评估模型对此图像问题的不确定度,之后 VQUE 统计了标注数据池内的答案分布,根据答案的多样性调整不确定度,得到每张图像的标注分数.此时,具有多种问题模式、问答分布更加平衡的图像往往具有较高的标注分数,因此模型对分数较高的样本进行问答标注.最终经过多轮标注,CCRL 仅用有限标注工作量获得的数据集,能够更好地协助视觉问答模型训练.

4　实验与分析

4.1　实验设置

为了验证 CCRL 模型在视觉问答主动学习任务上的性能,本文在大规模的视觉问答数据集上设计了主动学习性能试验,将 CCRL 和其他最新的、性能最优的主动学习方法进行比较,并进一步设计了多种实验分析视觉问答下的主动学习机制和各类主动学习方法的优缺点.

在实际应用中,主动学习模型选择具有标注价值的未标注的样本,使用标注系统进行标注.为了在实验中充分模拟主动学习的实际应用场景,本文使用了一个大规模的视觉问答数据集作为数据池并初始化为标注池和未标注池.未标注池中的数据,其标注信息(问题和答案)对于模型是不可见的,只有被选中为标注样本后,标注信息才可以使用.

4.1.1　视觉问答数据集 VQA-v2

VQA-v2[26]是目前最常用的视觉问答数据集,其中的图像数据全部来自 MS-COCO 数据集[46],每张图像都包含一个或多个关键目标物体,具有丰富的视觉信息和语义关系.其中的问题答案均为人工多轮标注.实验部分将 VQA-v2 划分为三部分:训练集(80 000 张图像和 444 000 个问题答案)、验证集(40 000 张图像和 214 000 个问题答案)、测试集(80 000 张图像和 448 000 个问题答案).在主动学习环境下,训练集内所有的问题答案都是不可见的.实验结果最终包含三类问题类型的预测准确率(数字问题,是否问题,其他问题)以及总的准确率,实验结果由 VQA Challenge 2021[47]竞赛主办方官方网站提交后计算得到.此外,测试集有两个子集版本,分别为 Test-dev 和 Test-standard,受限于在线测试,Test-standard 测试集仅提供总准确率.

4.1.2　主动学习设置

在本文的主动学习实验部分,VQA-v2 的训练集作为主动学习的采样池,初始化时从训练集选择 $M=2\%$ 的数据进行标注,放入数据池作为初始化的标注数据池,其余数据放入未标注数据池.在每轮主动学习迭代时,随机选择 $I=2\%$ 的数据进行标注,标注预算 $B=20\%$,一共标注 9 轮.为了保证性能比较的公平性和可靠性,测试性能的视觉问答模型均采用 MCAN[6]模型,使用同一个随机初始化的标注数据池作为采样的开始,且进行 5 次实验取平均值.测试时使用主动学习选出的数据训练 MCAN 模型,测试模型的问答准确率作为主动学习性能依据.在展示不同模型主动学习性能时,我们使用随机采样得到 2%的初始数据池,且所有方法均使用该初始数据池,此时 VQA 模型具有相同的性能,所以在性能图中只展示第一轮采样(4%)到最后一轮(20%)的数据.在 2%初始化时,Test-dev 上的准确率为 48.69%,Test-std 上为 49.03%.

4.1.3　作为对比的主动学习方法

本实验使用随机选取标注方法作为主动学习基线方法.为了展示充分展示 CCRL 的优越性能,本文选取了最新的、性能最优的主动学习方法和 CCRL 进行比较.其中基于不确定度的方法有 LL[16]、SRAAL[13]和 MOCU[21],基于样本多样性的方法有 CDAL[19]、Vab-AL[14]和 TA-VAAL[9].这些方法在实验中都使用其开源的代码,或按照其论文叙述进行复现.

4.2　实验性能比较

4.2.1　Test-dev 实验结果

图 4 展示了在 Test-dev 测试集上各个主动学习方法在视觉问答任务中的性能.通过观察可以发现,首先,CCRL 的方法是大幅领先于随机采样基线方法,对比最好的主动学习方法,准确率平均提升了 1.61%.这显著的性能提升,说明本文提出的基于跨模态对比学习的 CCRL 方法在视觉问答主动学习任务上具有十分优越的性能.其次,通过分析其他方法的性能曲线,可以发现基于不确定度的方法(LL、SRAAL、MOCU),其性能略高于基于样本分布的方法(CDAL、Vab-AL 和 TA-VAAL).该现象基本符合本文在引言部分对于视觉问答主动学习任务的分析.因为视觉问答任务是跨模态的任务,其样本为视觉模态而问答标注为自然语言文本模态,这就导致了过去传统的基于样本分布的方法,很难在视觉

问答数据上给出一个充分的多样性评估，跨模态的语义表达难以在单一模态的框架下进行学习和挖掘.而相对的，CCRL 方法通过其 VAUE 模块对问题和答案两者的数据分布进行了建模，解决了多种模态相互纠缠下的多样性评估问题，避免了上述提到的跨模态样本分布的痛点，帮助模型更好地理解了视觉问答数据的语义，从而选取出更具有标注价值的样本.最后，CCRL 在仅标注 20%的样本时，即可达到 65.01%的准确率，对比使用全部样本标注下达到的 70.04%的准确率，可以发现 CCRL 使用 20%的标注量达成了全部标注量训练下 92.82%的性能，对标注量的节省十分显著.

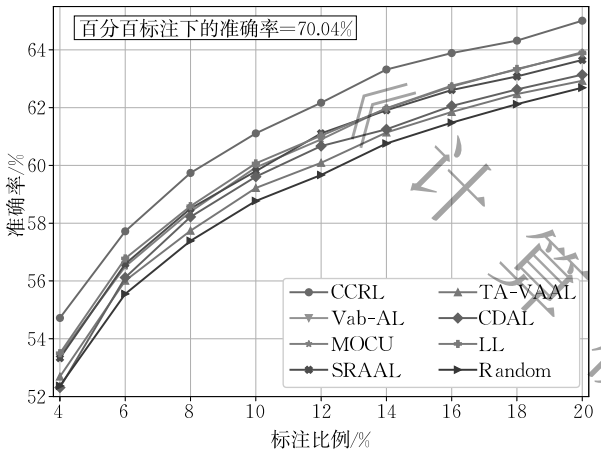

图 4 在 Test-dev 上的主动学习性能比较(2%初始化采样时，性能均为 48.69%)

4.2.2 Test-dev 三类问答实验结果

为了更加深入地对模型进行分析，图 5～图 7 分别展示了数字类、是否类、其他类三类问题的准确率.是否类问题是比较简单的问答类型，少量数据训练即可达到较高准确率，而其他类问题的模式则更为复杂，包含了方位、颜色、形状等问题形式，对问答标注量需求较高.分析实验结果可以发现，首先，CCRL 在三类问题下性能均达到最佳，几乎没有出现问答性能的偏向.因为 CCRL 中的问题答案匹配评估机制和 VAUE 模块在标注分数计算中引入了问题模式多样性的度量，CCRL 能够正确地评估三类问题模式的标注价值，合理分配有限的标注预算，将有限的标注量使用在对性能提升最大的样本.其次，当前性能最好的 LL 和 SRAAL 方法，可以发现 LL 在其他类和数字类问题下都取得了优越的效果，但是是否类问题下性能却下降严重，而 SRAAL 则仅在是否类问题表现较好.因为 LL 通过损失评估得到不确定性，而是否类问题性能初期就相对较

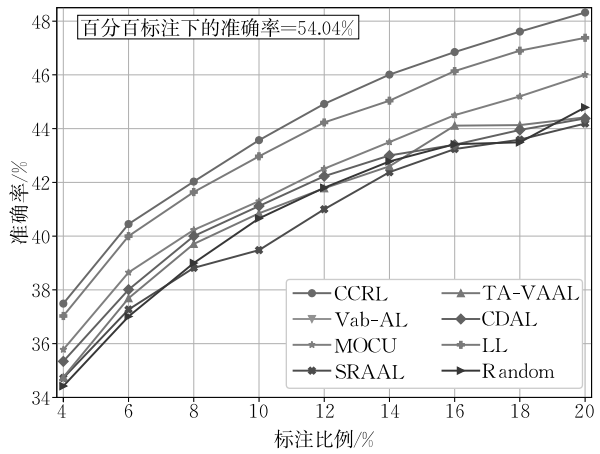

图 5 在 Test-dev 上的数字类问题的主动学习性能比较(2%初始化采样时，性能均为 31.60%)

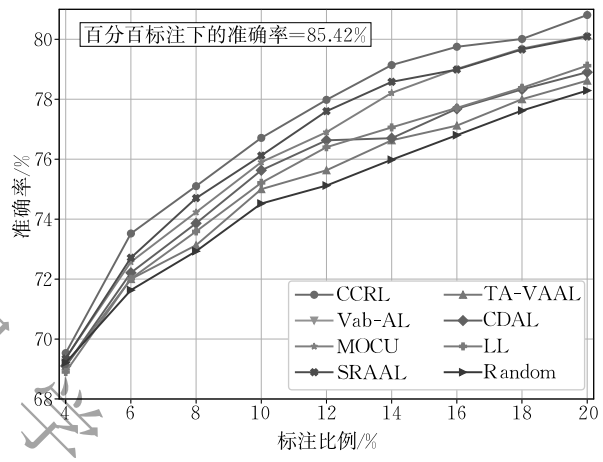

图 6 在 Test-dev 上的是否类问题的主动学习性能比较(2%初始化采样时，性能均为 66.34%)

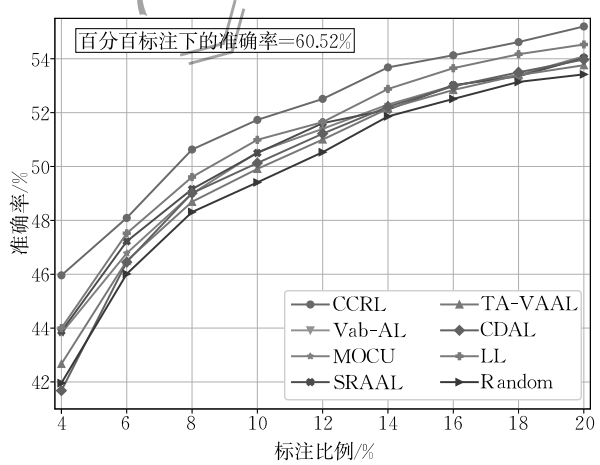

图 7 在 Test-dev 上的其他类问题的主动学习性能比较(2%初始化采样时，性能均为 37.46%)

高的是否类问题，其问答损失往往较低，这使得 LL 对该类问题的选取过少，出现了采样的不平衡导致了性能的下降.而 SRAAL 方法受数据集的偏置影响，对数目较多的是否类问题产生了采样冗余，导致

其他两类问题下表现不佳.第三,对比基于标签平衡的 Vab-AL 方法,本文的 CCRL 全面超过该方法.这也说明 CCRL 能够更好地评估多模态数据下的标签分布,针对更复杂的标签形式具有更优越的样本评估能力.

4.2.3　Test-standard 实验结果

Test-standard 是一个更全面的 VQA-v2 测试集,其上的性能测试如图 8 所示.首先,可以发现 CCRL 依旧取得了最高的性能,平均领先其他最优方法 1.65%的准确率.在 Test-standard 这个更加丰富全面的测试集下的表现,进一步验证了 CCRL 主动学习采样的高效性,这得益于模型对跨模态问题和图像的匹配技术.其次,分析曲线可以看出,CCRL 方法和随机主动学习采样方法的曲线更加平滑,这说明 CCRL 方法得到的采样,和随机采样一样都具有极高的稳定性,很少出现不稳定的标注价值评估.这种稳定评估的特性,也使得 CCRL 方法更加适用于实际应用场景.最后 CCRL 利用 20%的标注量达到了 64.9%的准确率,是百分百数据标注下性能的 92.3%,这能够最大化性能的同时极大地节约标注量.

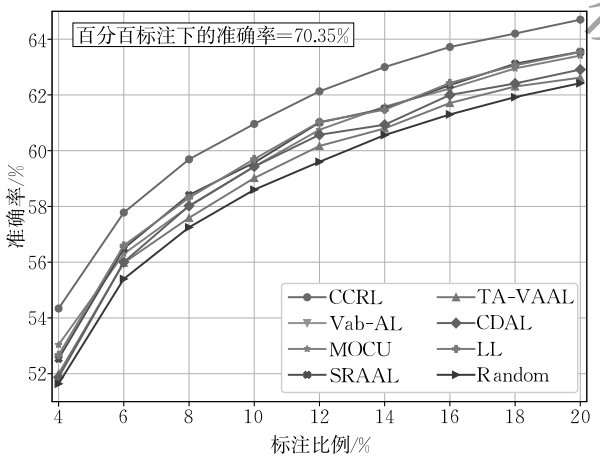

图 8　在 Test-std 上的主动学习性能比较

4.2.4　高标注预算下的主动学习性能

从图 4 和图 8 可以发现,模型在 20%标注量时,视觉问答模型性能的提升幅度依旧很大,说明此时标注量仍是相对不饱和的状态.为了进一步探索饱和状态下的主动学习性能,本节对主动学习实验的采样进行了更多轮迭代,在表 1 中展示了标注预算为 30%和 40%时视觉问答模型的性能.首先,CCRL 在高标注预算下性能仍然超过了其他的主动学习方法.CCRL 在标注量为 30%时,和百分百标注下的训练性能只差 3.25%(Test-dev)和 3.62%(Test-standard),达到了百分百数据标注下训练性能的 95.51%(Test-dev)和 94.85%(Test-standard);在标注量为 40%时,和百分百数据标注下的训练性能只差 2.44%(Test-dev)和 2.63%(Test-standard),达到了百分百标注下训练性能的 96.52%(Test-dev)和 96.26%(Test-standard).这说明本文提出的 CCRL 方法可以通过跨模态的对比学习技术选取出具有更丰富问答语义的样本,极大地削减跨模态视觉问答的数据标注工作量,同时最大化视觉问答模型的性能.其次,通过不同标注量下的性能增长,可以发现 CCRL 方法在 40%采样时,已经达到相对饱和状态,未标注数据池中具有标注价值的样本绝大部分已经被标注,所以此时模型的性能提升变得缓慢.这也符合主动学习的性能变化曲线.相对于随机采样,主动学习的性能曲线应当是上凸的,在标注为 0 和标注为 100%时所有方法曲线汇合,而性能差距则体现在曲线的上凸程度,在标注量较少的情况下尽可能得到高的性能.CCRL 方法的性能变化,相比其他的主动学习方法,更早地进入相对饱和阶段,这也说明 CCRL 方法的样本选取更加高效.

表 1　高标注预算下的主动学习性能　(单位:%)

主动学习方法	Test-dev		Test-standard	
	标注 30%	标注 40%	标注 30%	标注 40%
CCRL	**66.79**	**67.60**	**66.73**	**67.72**
TA-VAAL	65.03	65.95	65.11	66.02
Vab-AL	65.18	66.37	65.27	66.33
CDAL	65.12	66.02	65.40	66.32
MOCU	65.52	66.74	65.61	66.64
LL	65.48	66.45	65.49	66.50
SRAAL	65.67	66.69	65.83	66.74

4.3　消融实验

为了更加细致地验证 CCRL 方法中不同模块的影响,本文设计了消融实验,验证这些关键技术对模型性能提升的贡献.消融实验分为以下几组:(1)CCRL.原始方法;(2)CCRL-A.删除分析标注池答案标签分布部分的 CCRL,仅使用不确定度度量决定未标注样本标注分数;(3)CCRL-B.删除互信息最大化和对比学习部分的 CCRL,使用原始的相似度选取匹配问题;(4)CCRL-AB.作为对照组,同时进行(2)和(3)的删除.

消融实验结果如图 9 所示.首先分析标注池答案标签分布部分,通过 CCRL 和 CCRL-A 的比较,可以发现删除答案标签分布的规约后,主动学习性能平均下降了 0.82%.而对比 CCRL-B 和 CCRL-AB,CCRL-B 性能在多次采样下平均性能超过 CCRL-AB.答案标签分布的规约通过优化标注数据池的标签分布,防止出现长尾分布等标签不平衡现

象,增加了视觉问答的标注多样性,消融实验的结果进一步说明了该模块能够大幅度提升主动学习采样效率,对于 CCRL 是不可缺少的部分.其次分析对比学习部分,通过 CCRL 和 CCRL-B 的比较,可以发现删除对比学习部分后,性能下降十分明显,远超删除答案标签分布规约对模型的影响,准确率平均下降了 1.60%,两倍于 CCRL-A,这说明跨模态对比学习部分作为模型的核心,删除后会严重降低主动学习采样质量.同时 CCRL-A 的性能也远超 CCRL-AB,差距为 1.32%,这也说明本文 CCRL 中互信息最大化和 CPC 对比学习的结合,优化了模型对于跨模态数据的语义理解泛化能力,利用正负例学习极大地提高了视觉问答主动学习的性能.通过消融实验,验证了答案标签分布规约和对比学习两部分都对高质量主动学习采样具有极大的贡献.此外,CCRL 对比 CCRL-AB 在各个采样比例下的准确率提高了 2.13%,这显著的性能提升也进一步验证了本文主要贡献点的有效性.

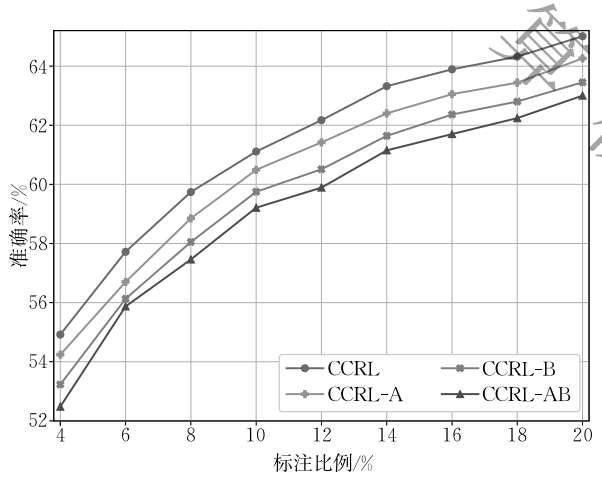

图 9 消融实验性能比较(所有方法使用相同的初始标注池(2%),此时所有方法性能相同,为 49.03%)

4.4 运行时间分析

为了进一步分析跨模态图像问答下的主动学习时间消耗,本文设计了主动学习算法运行时间的比较实验.在实验中,本文对多种算法,使用单张 NVIDIA RTX 3080 显卡,重复 10 次进行 2%数据的采样,并取运行时间的均值,实验结果如表 2 所示.首先,MOCU 和 LL 所需时间最短,因为两者基于样本不确定性,只需要计算得到不确定性并排序即可.而本文的 CCRL 算法的运行时间仅次于 MOCU 和 LL,CCRL 借助于预训练的 Faster-RCNN 视觉特征,采样所需时间开销较小,且问答匹配在采样阶段不需要进行对比学习正负例构建,直接比较匹配分数即可,大大提高了计算效率.此外可以发现 CDAL 和 SRAAL 耗时较多,前者受制于逐样本递进式的算法,无法并行运算,后者则因为变分自编码器和重标注过程增加了运算时间.

表 2 主动学习算法运行时间 (单位:s)

AL 算法	2%样本采样时间	FPS
CCRL	70.79	33.90
TA-VAAL	119.77	20.04
Vab-AL	93.71	25.61
CDAL	176.50	13.60
MOCU	49.81	48.18
LL	42.52	56.44
SRAAL	190.85	12.58

4.5 跨模态特征对比

为了深入探究本文所提出的 VQME 模块特征学习能力,比较不同特征算法在没有答案标签的主动学习环境下的跨模态数据表征质量.比较的特征有本文方法(CCRL)、预训练的 Faster-RCNN+GloVe(FR+G)、MS-COCO 图像描述预训练特征(Bottom-up)、ImageNet 预训练的 Resnet+GloVe(Resnet+G)、纯语言特征(Language).本文参考 MoCo、SimCLR 等论文的特征质量对比方法,设计了如下实验:本文首先在无监督环境下,利用 VQA-v2 数据集预训练了多种特征提取模型,并将所有样本的特征提取保存.之后设计一个单层的 VQA 分类器并将上述无监督特征作为输入,使用答案标签进行监督,并只对最后一层分类器进行梯度反传和微调.根据不同特征得到的分类器性能衡量特征质量.

实验结果如表 3 所示,首先可以发现,本文的 CCRL 对比学习特征能够训练得到 59.52%的整体准确率,对比其他性能最优的 FR+G 特征,领先 5.30%的整体准确率.这说明本文基于互信息最大和 CPC 对比学习的特征学习模块能够更好地提取跨模态样本的关键信息.此外,可以发现 CCRL 特征性能在每个问题类别下,问答性能均超过其他方法,平均性能提升可达 7%,且没有出现明显的偏向.相对的,Bottom-up 方法受制于 caption 任务在数字类方法表现不佳,Faster-RCNN 和 Resnet 则受制于图像特征与文本特征过于独立,在语义繁杂的其它类问题下性能较差.这也进一步表明了,本文的 CCRL 特征能够充分地学习各种问答类型,在一定程度上抵消问题类型的偏置,在各种问答下都取得较好的泛化能力.

表 3　无监督特征在线性分类器下的 VQA 性能(%)，整体性能和各类问题的性能

特征模型	整体性能	是否类	数字类	其他类
CCRL	**59.52**	**76.79**	**38.75**	**49.52**
FR+G	54.22	73.46	35.18	41.83
Bottom-up	53.11	70.40	31.01	45.90
Resnet+G	50.31	67.62	32.12	41.96
Language	44.26	67.01	31.55	27.37

5　结　论

本文首次将主动学习应用到视觉问答的任务标注中，提出了一种基于跨模态特征对比学习的视觉问答主动学习方法(CCRL). 该方法内嵌了视觉问题匹配评价(VQME)模块和视觉答案不确定度度量(VAUE)模块，从覆盖尽可能多的问题类型和获取更平衡的问答分布两方面入手，选取最具标注价值、问答内容最丰富的样本进行问答标注. 视觉问题匹配评价模块利用多头自注意力机制获取视觉特征和问题文本特征，之后使用互信息最大化和 CPC 对比学习机制，一方面约束了特征生成器使其尽可能保留问答关联性高的关键信息，另一方面通过构建正负例训练了能够评价图像问题匹配得分的匹配评价模块. 视觉答案不确定度度量模块，则使用 VQME 获取未标注样本的匹配问题后，测试每个图像问题配对在视觉问答模型下的不确定度，同时根据答案预测概率和标注池的答案分布，得到每个样本的标注得分，从而保证 CCRL 构建一个标签平衡、语义多样的视觉问答数据池.

在大规模公开数据集 VQA-v2 上本文对 CCRL 和其他最新最优的主动学习方法进行了详尽的实验，并充分分析了实验结果. 实验验证了本文提出的 CCRL 具有优越的性能，在跨模态的视觉问答任务上取得了最好的性能. 同时 CCRL 仅使用少量的标注即可实现接近百分百标注数据集的性能，这说明 CCRL 可以大幅度地节省视觉问答标注工作量的同时，最大化模型的性能. 可以预见，随着跨模态任务和视觉问答应用的普及，未来会有越来越多的科研工作聚焦于跨模态数据集的选取和构建，希望本文可以为相关工作提供一个充分的基准性能和完善的验证实验设置，对后续的相关研究有所启发.

参　考　文　献

[1] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 652-663

[2] Li Liang, Yan Chenggang Clarence, Chen Xing, et al. Distributed image understanding with semantic dictionary and semantic expansion. Neurocomputing, 2016, 174: 384-392

[3] Li Liang, Jiang Shuqiang, Huang Qingming. Learning hierarchical semantic description via mixed-norm regularization for image understanding. IEEE Transactions on Multimedia, 2012, 14(5): 1401-1413

[4] Antol S, Agrawal A, Lu Jiasen, et al. VQA: Visual question answering//Proceedings of the International Conference on Computer Vision. Santiago, Chile, 2014: 1682-1690

[5] Anderson P, He Xiaodong, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6077-6086

[6] Yu Zhou, Yu Jun, Cui Yuhao, et al. Deep modular co-attention networks for visual question answering//Proceedings of the Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6281-6290

[7] Xiao Junbin, Shang Xindi, Yao Angela, Chua Tat-Seng. NExT-QA: Next phase of question-answering to explaining temporal actions//Proceedings of the Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 9777-9786

[8] Li Liang, Wang Shuhui, Jiang Shuqiang, Huang Qingming. Attentive recurrent neural network for weak-supervised multi-label image classification//Proceedings of the International Conference on Multimedia. Seoul, Korea, 2018: 1092-1100

[9] Zhou Baohang, Cai Xiangrui, Zhang Ying, et al. MTAAL: Multi-task adversarial active learning for medical named entity recognition and normalization//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021: 14586-14593

[10] Zhang Beichen, Li Liang, Su Li, et al. Structural semantic adversarial active learning for image captioning//Proceedings of the International Conference on Multimedia. Seattle, USA, 2020: 1112-1121

[11] Cohn D A, Ghahramani Z, Jordan M I. Active learning with statistical models. Journal of Artificial Intelligence Research, 1996, 4: 129-145

[12] Sener O, Savarese S. Active learning for convolutional neural networks: A core-set approach//Proceedings of the International Conference on Learning Representations. Vancouver, 2018: 1-13

[13] Zhang Beichen, Li Liang, Yang Shijie, et al. State-relabeling adversarial active learning//Proceedings of the Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8753-8762

[14] Choi J, Yi K M, Kim J, et al. Vab-AL: Incorporating class imbalance and difficulty with variational bayes for active learning//Proceedings of the Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 6749-6758

[15] Nilsson D, Pirinen A, Gärtner E, Sminchisescu C. Embodied visual active learning for semantic segmentation//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021: 2373-2383

[16] Yoo D, Kweon I S. Learning loss for active learning//Proceedings of the Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 93-102

[17] Cai Lile, Xu Xun, Liew Junhao, Foo Chuan Sheng. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs//Proceedings of the Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 10988-10997

[18] Huang Sheng-Jun, Jin Rong, Zhou Zhi-Hua. Active learning by querying informative and representative examples. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(10): 1936-1949

[19] Agarwal S, Arora H, Anand S, Arora C. Contextual diversity for active learning//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 137-153

[20] Zhang Lijun, Chen Chun, Bu Jiajun, et al. Active learning based on locally linear reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 33(10): 2026-2038

[21] Zhao Guang, Dougherty E R, Yoon Byung-Jun, et al. Uncertainty-aware active learning for optimal Bayesian classifier//Proceedings of the Conference on Learning Representations. Austria, 2021: 9-19

[22] Sinha S, Ebrahimi S, Darrell T. Variational adversarial active learning//Proceedings of the International Conference on Computer Vision. Seoul, Korea, 2019: 5971-5980

[23] Bai Lu, Guo Jia-Feng, Cao Lei, Cheng Xue-Qi. Long tail query recommendation based on query intent. Chinese Journal of Computers, 2013, 36(3): 636-642(in Chinese)
(白露, 郭嘉丰, 曹雷, 程学旗. 基于查询意图的长尾查询推荐. 计算机学报, 2013, 36(3): 636-642)

[24] van den Oord A, Li Yazhe, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv: 1807.03748, 2018

[25] He Kaiming, Fan Haoqi, Wu Yuxin, et al. Momentum contrast for unsupervised visual representation learning//Proceedings of the International Conference on Multimedia. Seattle, USA, 2020: 9726-9735

[26] Goyal Y, Khot T, Agrawal A, et al. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. International Journal of Computer Vision, 2019, 127(4): 398-414

[27] Ducoffe M, Precioso F. Adversarial active learning for deep networks: A margin based approach. arXiv preprint arXiv: 1802.09841, 2018

[28] Wang Shuo, Li Yuexiang, Ma Kai, et al. Dual adversarial network for deep active learning//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 680-696

[29] Kim K-Y, Park D, Kim K, Chun S Y. Task-aware variational adversarial active learning//Proceedings of the Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 8166-8175

[30] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation. representing model uncertainty in deep learning//Proceedings of the International Conference on Machine Learning. New York City, USA, 2016: 1050-1059

[31] Beluch W H, Genewein T, Nürnberger A, Köhler J M. The power of ensembles for active learning in image classification//Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9368-9377

[32] Caramalau R, Bhattarai B, Kim T-K. Sequential graph convolutional network for active learning//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021: 9583-9592

[33] Kim T, Hwang I, Lee H, et al. Message passing adaptive resonance theory for online active semi-supervised learning//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 5519-5529

[34] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A deep learning approach to visual question answering. International Journal of Computer Vision, 2017, 125(1): 110-135

[35] Bosselut A, Le Bras R, Choi Y. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021: 4923-4931

[36] Ren Hongyu, Dai Hanjun, Dai Bo, et al. LEGO: Latent execution-guided reasoning for multi-hop question answering on knowledge graphs//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 8959-8970

[37] Xiong Peixi, Wu Ying. TA-student VQA: Multi-agents training by self-questioning//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2020: 10062-10072

[38] Urooj A, Kuehne H, Duarte K, et al. Weakly-supervised grounded visual question answering using capsules//Proceedings of the Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 8465-8474

[39] Liu Xuejing, Li Liang, Wang Shuhui, et al. Adaptive reconstruction network for weakly supervised referring expression grounding//Proceedings of the International Conference on Computer Vision. Seoul, Korea, 2019: 2611-2620

[40] Ren Shaoqing, He Kaiming, Girshick R B, Sun Jian. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

[41] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1532-1543

[42] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[43] Rao C R, Nayak T K. Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. IEEE Transactions on Information Theory, 1985, 31(5): 589-593

[44] Nowozin S, Cseke B, Tomioka R. f-GAN: Training generative neural samplers using variational divergence minimization// Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 271-279

[45] Chen Ting, Kornblith S, Norouzi M, Hinton G E. A simple framework for contrastive learning of visual representations// Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2020: 1597-1607

[46] Lin T-Y, Maire M, Belongie S J, et al. Microsoft COCO: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755

[47] VQA Challenge 2021. https://visualqa.org/challenge.html

ZHANG Bei-Chen, Ph. D. candidate. His research interests include active learning, semi-supervised machine learning, and generative task.

LI Liang, Ph. D., associate professor. His research interests include active learning, semi-supervised machine learning, and visual generation.

ZHA Zheng-Jun, Ph. D., professor. His research interests include image/video analysis and retrieval, multi-media big data, artificial intelligence.

HUANG Qing-Ming, Ph. D., professor. His research interests include multi-media analysis, pattern recognition and computer vision.

Background

Visual question answer (VQA) that bridges vision and language has gained broad interest from both the computer vision and natural language processing communities. As a cognitive-level task, a good VQA model requires a large nunmber of images with corresponding questions and answers to achieve generalization, which is very costly. This greatly limits the ability of VQA and its application in real. To solve this problem, we design the cross-modal active learning (AL) approach for VQA. The ultimate goal of AL is to enable the best model performance with a limited label budget. AL methods are grouped into a distribution-based method and uncertainty-based method. Although previous methods have made remarkable advances for single-modal tasks, they are not suitable for cross-modal tasks. AL for VQA requires the semantic comprehension of different modalities and accurate measurement for cross-modal informativeness, which is the main challenge for this task.

To solve these problems, this paper proposes a contrastive cross-modal representation learning based active learning (CCRL) method. It consists of a visual question matching evaluation (VQME) module and a visual answer uncertainty estimation (VAUE) module. The VQME adapts CPC contrastive learning and mutual information maximization for image and question matching. Then the VAUE designs a cross-modal acquisition function to evaluate the labeling score based on both VQA uncertainty and answer diversity, by which our CCRL can cover more question patterns and make the distribution of answers more balanced. Experiments and ablation on VQA-v2 show that CCRL outperforms other latest state-of-the-art AL methods with a large margin and it achieves very competitive performance with only a little part of annotations, which verifies its outstanding effectiveness on cross-modal data.

Our research group has already worked on the topic of active learning and cross-modal. Before this work, we have done some relevant studies. These methods proposed by our group have been published in top-tier conferences and journals, such as TPAMI, TIP, TMM, CVPR, ACM MM, etc.

This research was supported by the National Science and Technology Innovation 2030-Major Project of China (2018AAA0102000), and the National Natural Science Foundation of China (61732007, 61771457, U21B2038).