计算机科学
COMPUTER SCIENCE

自监督学习用于3D真实场景问答

李祥, 范志广, 林楠, 曹仰杰, 李学相

引用本文

李祥, 范志广, 林楠, 曹仰杰, 李学相. 自监督学习用于3D真实场景问答[J]. 计算机科学, 2023, 50(9): 220-226.

LI Xiang, FAN Zhiguang, LIN Nan, CAO Yangjie, LI Xuexiang. Self-supervised Learning for 3D Real-scenes Question Answering [J]. Computer Science, 2023, 50(9): 220-226.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900256

自监督学习用于3D真实场景问答

李　祥[1]　范志广[2]　林　楠[1]　曹仰杰[1]　李学相[1]
1 郑州大学网络空间安全学院　郑州 450000
2 中山大学计算机学院　广州 510000
(lixiang.zg@qq.com)

摘　要　近年来,视觉问答逐渐成为计算机视觉领域的研究热点之一。目前大多数研究是围绕2D图像的问答,但2D图像存在由视点改变、遮挡和重投影引入的空间模糊性。现实生活中,人机交互的场景往往是3D的,研究3D问答更具实际应用价值。已有的3D问答算法能感知3D对象以及它们的空间关系,并能回答意义复杂的问题。但是,由点云组成的3D场景和问题属于两种模态的数据,这两种模态数据之间存在明显的差异,难以对齐,两者潜在的相关特征容易被忽略。针对这一问题,提出了一种基于自监督学习的3D真实场景问答方法。该方法首次在3D问答模型中引入对比学习,通过3D跨模态对比学习对齐3D场景和问题,缩小两种模态的异构差距,挖掘两者的相关特征。此外,将深度交互注意力网络用于处理3D场景和问题,对3D场景中的对象和问题中的关键词做充分的交互。在ScanQA数据集上进行的大量实验表明,3DSSQA在EM@1这个主要指标上的准确度达到了24.3%,超过了目前最先进的模型。

关键词:3D问答;自监督学习;对比学习;点云;深度交互注意力
中图法分类号　TP181

Self-supervised Learning for 3D Real-scenes Question Answering

LI Xiang[1],FAN Zhiguang[2],LIN Nan[1],CAO Yangjie[1] and LI Xuexiang[1]
1 School of Cyber Science and Engineering,Zhengzhou University,Zhengzhou 450000,China
2 School of Computer Science and Engineering,Sun Yat-sen University,Guangzhou 510000,China

Abstract　Visual question answering(VQA)has gradually become one of the research hotspots in recent years. Most of the current question-answering research is 2D-image-based,often suffering from spatial ambiguity introduced by viewpoint changing,occlusion,and reprojection. In practice,human-computer interaction scenarios are often three-dimensional,yielding the demand for 3D-scene-based question answering. Existing 3D question answering algorithms have so far been able to perceive 3D objects and their spatial relationships,and can answer complex questions. However,point clouds represented by 3D scenes and the target questions belong to two different modalities,which are extremely difficult to align,leading to their unconspicuous related features are easy to be ignored. Aiming at this problem,this paper proposes a novel learning-based question answering method for realistic 3D scenes,called 3D self-supervised question answering(3DSSQA). Within 3DSSQA,a 3D cross-modal contrastive learning model(3DCMCL) is proposed to first align point-cloud data with question data globally for modality heterogeneity gap reduction,before mining related features between the two. In addition,a deep interactive attention(DIA) network is adapted to align 3D objects with keywords in a more fine-grained granularity,facilitating sufficient interactions between them. Extensive experiments on the ScanQA dataset demonstrate that 3DSSQA achieves an accuracy of 24.3% on the main EM@1 metric,notably surpassing state-of-the-art models.

Keywords　3D question answering,Self-supervised learning,Contrastive learning,Point clouds,Deep interactive attention

1　引言

在3D问答任务中,模型从3D场景中接受视觉信息,并回答3D场景对应的文本问题[1]。这项任务不仅需要具备在3D场景中识别和定位对象的基本感知能力,还应具备理解3D场景并根据问题进行推理的能力[2],图1给出了一个3D问答任务的样本。

与其他的多模态任务类似,传统的视觉问答(Visual

到稿日期:2022-09-28　返修日期:2023-03-28
基金项目:国家自然科学基金面上项目(61972092);郑州市协同创新重大专项(20XTZX06013)
This work was supported by the General Project of the National Natural Science Foundation of China(61972092) and Collaborative Innovation Major Project of Zhengzhou(20XTZX06013).
通信作者:李学相(lxx@zzu.edu.cn)

Question Answering，VQA）主要局限于 2D 空间。在过去几年里，基于图像的视觉问答受到广泛的关注，并且涌现出了各种不同类别的算法，如基于联合嵌入的方法、基于注意力机制的方法、基于外部知识的方法等，这些算法都取得了不错的性能。但是基于 2D 图像的视觉问答难以准确理解 3D 世界，存在一些无法忽视的问题。例如，2D 图像不能准确地表现 3D 场景中物体的相对方向和距离，当一些物体被其他物体遮挡时难以进行识别。在多张图像中识别同一物体也存在困难。

随着 3D 传感器和 3D 视觉技术的快速发展，3D 问答成为一种可行的方案，并具有广阔的前景，其广泛应用于现实世界环境中的人机交互[3]、增强现实中的信息查询[4]、虚拟现实和基于语言的自动驾驶汽车导航[5]。由于点云是不规则的且规模较大，因此 3D 问答任务具有更复杂的输入数据，其往往包含大量的对象和复杂的空间关系，在 3D 场景中进行问答具有一定的挑战性。尽管前人在提高场景理解方面做出了巨大努力，但 3D 点云和问题这两种模态的特征表示有明显的区别，存在“语义鸿沟”，难以对齐且难以利用跨模态的信息。此外，现有的方法往往忽视了 3D 场景和问题在特征空间中的潜在关系。在建立数据集时，问题需要与 3D 场景保持一致，而不是随意创建。例如，对于“最靠近门的椅子是什么颜色”，从对应的 3D 场景中可以找到椅子这个对象，也就是说，问题中涉及的关键对象或属性可以在 3D 场景中找到。

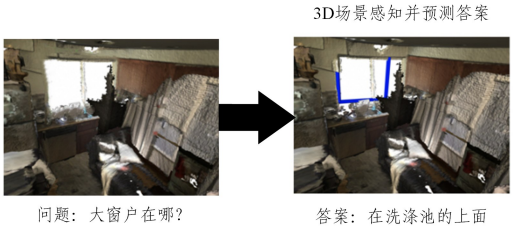


图 1　3D 问答任务的例子

Fig. 1　Sample of 3D question answering

为了解决上述问题，本文提出了一种基于自监督学习的 3D 真实场景问答方法（3D self-Supervised Question Answering，3DSSQA）。自监督学习属于无监督学习范式的一种，它不需要人工标注的类别标签信息，而是利用数据本身提供的监督信息来学习样本数据的特征表达，并用于下游任务。对比学习是自监督学习中的一类重要的方法。通过引入对比学习，模型能够学习到一个共同的低维空间来嵌入 3D 点云特征和问题特征，从而对齐 3D 点云和问题，有利于 3D 点云和问题在堆叠 Transformer 网络中进一步交互。3DSSQA 方法可以有效地感知 3D 场景并定位与问题相关的对象，之后以此为依据来推断出答案。在得到 3D 场景和问题的特征表示之后，将其送到 3D 跨模态对比学习框架中，使 3D 场景和对应问题的互信息（Mutual Information，MI）最大化，以便模型能够更好地理解 3D 场景和问题之间的关联信息。然后，将这两种模态特征输入堆叠 Transformer 网络，在多个堆叠的 DIA 层中建立统一的语义表达，从而回答给定的问题。

本文方法的创新性贡献可以总结如下：

（1）为了探索 3D 场景和问题之间潜在的相关特征，提出了一种基于自监督学习的 3D 真实场景问答方法（3DSSQA）。该方法首次在 3D 问答中使用对比学习，提高了视觉和文本的表示能力，使模型更加关注 3D 场景和问题之间的关系。

（2）为了解决具有挑战性的 3D 问答任务，使用了堆叠 Transformer 网络，这种结构可以在交互过程中减少视觉文本信息的丢失，使两种信息的融合更加充分。

（3）本文方法在 ScanQA 数据集上取得了显著的效果，优于现有的方法，该方法提高了 3D 场景理解的能力，有助于 3D 问答任务的进一步发展。

2　相关工作

视觉问答任务在人工智能领域具有较高的热度。这个任务要求机器根据一个图像和一个对应的自然语言问题提供准确的自然语言答案。早期的视觉问答主要通过联合嵌入的方法进行交互，联合嵌入的方法往往通过简单的机制将两种特征进行整合，如串联、逐元素乘法或逐元素加法等[6-8]。后来出现了基于注意力机制的跨模态交互方法，该方法模仿人类的注意力模式，侧重于学习问题分词和图像区域之间的相互作用，使问答的过程更具有可解释性[9-11]。随着研究的深入，Transformer 结构被引入到视觉问答中，通过 Transformer 的多头注意力，加强了图像和文本特征的细粒度交互，取得了不错的性能[12-14]。近年来，基于预训练的方法获得快速的发展，它们使用预训练的编码器-解码器架构，充分利用大规模视觉文本数据集，显著提升了模型的性能[15-17]。然而，图像和 3D 点云的表示不同，图像是 3D 世界的一个映射，缺少深度这一维度。因此，基于图像的 VQA 方法不能直接迁移到 3D 场景理解。

与基于图像的视觉问答相比，3D 问答是一个新兴的研究方向，现有的工作侧重于使用跨模态的 Transformer 进行点云特征与文本特征的融合。例如，Azuma 等[1]提出了一个 3D 问答的基线模型，被称为 ScanQA。ScanQA 使用基于 Transformer 的编码器层和解码器层将语言信息引导的多个 3D 物体特征以及文本信息融合在一起。Ye 等[18]提出了一种新的 3D 问答框架“3DQA-TR”，它使用 3D-L BERT 将外观、几何和语言问题的多模态信息相互关联，来预测目标答案。传统的 3D 场景理解工作更多地关注单个物体，而忽略了物体之间的关系。为了解决这个问题，Yan 等[2]提出了 Transformer-VQA3D。该模型首先使用一个跨模态 Transformer 来融合问题和物体的特征。然后，通过应用场景图初始化并取场景图的附加边来进行场景图感知注意，获得物体之间的关系并推断出答案。这些算法都在尝试解决 3D 问答任务，但是它们没有利用 3D 场景与问题之间的互信息，而是直接对齐这两种模态的数据，导致模型并没有充分学习到 3D 场景和问题之间统一的语义表达。

对比学习是一种自监督学习方法。简单来说，对比学习指通过比较正负样本对来学习表示，正样本对之间的相似度应尽可能高，而负样本对之间的相似度应尽可能低。对比学习经常被应用在基于图像的视觉问答中。多模态编码器学习图像文本的交互具有挑战性。为了应对这个问题，Li 等[15]提出了 ALBEF 模型，通过图像文本对比学习来对齐图像特征和文本特征，使多模态编码器更容易进行跨模态学习，并使

单模态编码器能够更好地理解图像和文本的语义。ALBEF 应用对比学习来学习全局视觉信息和文本信息之间的关联，未能考虑输入中的局部信息。为此，Yang 等[19]提出了一种新的框架 TCL。TCL 引入跨模态对齐(CMA)、模态内对比(IMC)和局部 MI 最大化这 3 个对比模块。这 3 个模块的组合不仅使模型能学习跨模态图像文本对齐和模态内部有意义的信息，还能捕获结构化的局部信息。Wang 等[20]提出了一个统一的视觉语言预训练模型 VLMo，该模型联合学习图像-文本对比学习，屏蔽语言模型和图像-文本匹配任务。它利用跨模态对比学习来获得理想的表示，使融合编码器更容易学习多模态交互。基于图像的视觉问答使用对比学习取得了不错的性能。

对比学习使模型更加关注图像和文本之间的关系。与基于图像的视觉问答相比，3D 问答能够避免图像固有的空间模糊性，更容易捕获对象的几何信息和空间关系。但是，3D 点云与问题这两种模态的数据存在明显差异，难以融合。因此，本文将对比学习从基于图像的视觉问答任务迁移到 3D 问答任务中。

3 方法

本章将全面介绍 3DSSQA 方法。该方法的总体结构如图 2 所示。这个模型主要使用 VoteNet 获得 3D 场景中的对象特征，使用 BiLSTM 获得问题特征，使用 3D 跨模态对比学习(3D Cross-Modal Contrastive Learning，3DCMCL)对齐视觉和文本的单模态表示，提高了视觉和文本的表示能力，使模型更加关注 3D 场景和问题之间的关系，之后将它们作为输入送到堆叠 Transformer 网络。3.1 节介绍了输入嵌入，包括视觉表示学习和文本表示学习；3.2 节介绍了 3D 跨模态对比学习；3.3 节介绍了堆叠 Transformer 网络；3.4 节介绍了物体感知和预测答案模块。

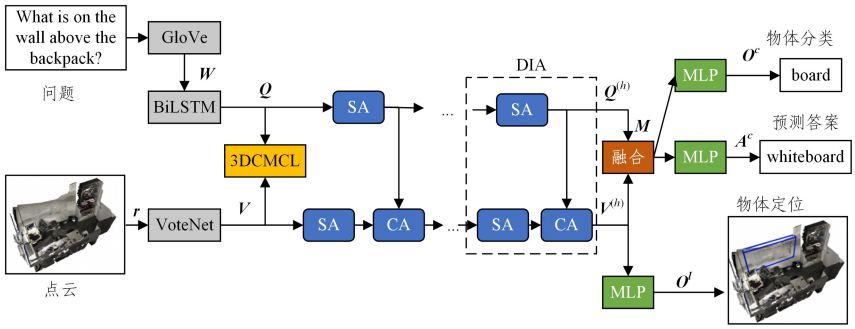


图 2 3DSSQA 的整体框架图

Fig. 2 Overall framework of 3DSSQA

3.1 输入嵌入

3.1.1 视觉表示学习

本文首先对输入的点云进行特征表示，输入的点云大小为 $n\times 3$，n 个点中的每个点都有三维坐标，点云可以表示为 $p\in\mathbb{R}^{n3}$。根据前人对 3D 场景理解的研究[21-22]，本文使用额外的点特征，如点的高度、颜色、法线和多视图图像特征[23]，将 2D 外观特征映射到点云中。这些组合的点云特征可以表示为 $r\in\mathbb{R}^{135}$。其次，直接使用 VoteNet[24]检测 3D 场景中的对象。VoteNet 的基础网络 PointNet++[25]可以通过处理点云得到对象框。最后，使用带有 GELUs 激活函数的非线性层对其进行投影，以获得对象框表示 $\boldsymbol{V}=\{v_1,v_2,\cdots,v_m\}\in R^{md_s}$，其中 m 是对象框的数量，d_s 是对象框的维度。

3.1.2 文本表示学习

为了方便提取问题特征，每个问题被统一成由 n 个单词组成的句子，超过 n 个单词的部分会被删除，少于 n 个单词的句子会用 0 来填充。单词可以表示为 d_h 维的词嵌入 $\boldsymbol{D}=\{w_1,w_2,\cdots,w_n\}R^{n*d_h}$。其中，$n$ 表示每个问题包含的最大单词数，d_h 表示词嵌入的维度。然后，词向量被送到双向长短期记忆网络(BiLSTM)[26]以编码句子嵌入 $\boldsymbol{Q}=\{q_1,q_2,\cdots,q_n\}R^{n\times d_s}$。其中，$d_s$ 是 BiLSTM 中隐藏状态的维度。同样，采用非线性层把问题特征映射到与视觉模态相同的嵌入空间。

3.2 3D 跨模态对比学习

3D 点云和问题属于不同的特征空间，跨模态信息存在明显的差异，难以对齐。本文首次在 3D 问答中使用对比学习来对齐 3D 点云和问题，缩小 3D 点云和问题的异构差距，挖掘两者的相关特征。这样的学习过程有利于 3D 点云和问题的进一步融合，可以实现从 3D 场景到语言理解跨模态的知识共享。

如图 3 所示，3D 跨模态对比学习(3DCMCL)的目标是使匹配的 3D 场景-问题对尽可能接近，使未匹配的 3D 场景-问题对相互远离。为了测量 3D 场景和问题的匹配程度，需要引入互信息这一概念。互信息(MI)的概念起源于概率论和信息论。它是对两个概率分布或随机变量之间的依赖关系或共享信息数量的评估。互信息的值越大，3D 场景和问题的相关性就越高。由于 3D 场景和问题之间存在很强的相关性，因此本文把 3D 场景与对应的问题定义为正样本对；相反，把 3D 场景和不相关的问题定义为负样本对。

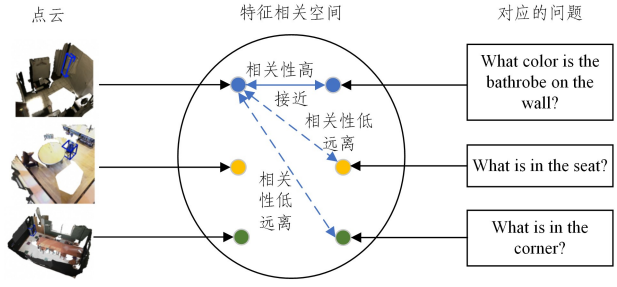


图 3 3D 跨模态对比学习的示意图

Fig. 3 Illustration of 3D cross-modal contrastive learning

3D 跨模态对比学习的主要思想是通过比较正负样本对

之间的互信息来提高 3D 问答模型的学习能力。3D 场景和问题之间的互信息的表达式如式(1)所示：

$$I(\boldsymbol{V},\boldsymbol{Q})=\sum_{\boldsymbol{V},\boldsymbol{Q}}P(\boldsymbol{V},\boldsymbol{Q})\log\frac{P(\boldsymbol{V}|\boldsymbol{Q})}{P(\boldsymbol{V})} \tag{1}$$

其中，$\frac{P(\boldsymbol{V}|\boldsymbol{Q})}{P(\boldsymbol{V})}$表示 3D 场景和问题的相似度，本文使用$S_c(\boldsymbol{V},\boldsymbol{Q})$近似地表示公式$\frac{P(\boldsymbol{V}|\boldsymbol{Q})}{P(\boldsymbol{V})}$，这里$S_c(\boldsymbol{V},\boldsymbol{Q})$关心的是完整的 3D 场景和全局的问题特征。根据 Misra 等[27]所做的工作，本文利用余弦相似性来计算S_c。此时，如果给出的第 j 个 3D 场景的特征V_j和它对应的问题特征$\boldsymbol{Q}_j$，则S_c可以表示为：

$$S_c(\boldsymbol{V},\boldsymbol{Q})=\exp\left(\frac{cosine(\boldsymbol{V}_j,\boldsymbol{Q}_j)}{\tau}\right) \tag{2}$$

其中，τ来源于物理学中的温度系数，是一个超参数。

式(1)处理起来非常困难，要估计高维随机变量的 MI，InfoNCE[28]是一个不错的选择。它是一种分类交叉熵损失，可在一组负样本中识别正样本。InfoNCE 已经被证明是 MI 的下界[28]。3D 场景和问题的互信息 $I(\boldsymbol{V},\boldsymbol{Q})$可以表示为：

$$I(\boldsymbol{V},\boldsymbol{Q})\log U'-L^{\mathrm{NCE}} \tag{3}$$

其中，U'表示负样本的数目，它被认为是一个常量；L^{NCE}表示 InfoNCE 损失。从式(3)可以看出，想要保证互信息 $I(\boldsymbol{V},\boldsymbol{Q})$最大化，只需使互信息 $I(\boldsymbol{V},\boldsymbol{Q})$的下界最大化。此时，最小化$L^{\mathrm{NCE}}$可以使互信息的下界最大化。

给出参与运算的 U 个 3D 场景和对应的 U 个问题。结合式(1)和式(2)，3D 场景到问题的 InfoNCE 损失可以定义为：

$$L^{\mathrm{NCE}}(\boldsymbol{V},\boldsymbol{Q})=-\sum_{i=1}^{U}\log\frac{S_c(\boldsymbol{V}_i,\boldsymbol{Q}_j)}{\sum_{j=1}^{U}S_c(\boldsymbol{V}_i,\boldsymbol{Q}_j)} \tag{4}$$

其中，$\boldsymbol{V}_i$和$\boldsymbol{Q}_j$是正样本对，V_i与剩下的不匹配的 $U-1$ 个问题样本$\boldsymbol{Q}_{\mathrm{neg}}=\{\boldsymbol{Q}_1,\boldsymbol{Q}_2,\cdots,\boldsymbol{Q}_{U-1}\}$组合成负样本对。分子表示正样本对之间的互信息，分母表示 U 对样本的互信息之和，包括正样本对和负样本对。

同理，问题到 3D 场景的 InfoNCE 损失的表达式如式(5)所示：

$$L^{\mathrm{NCE}}(\boldsymbol{Q},\boldsymbol{V})=-\sum_{i=1}^{U}\log\frac{S_c(\boldsymbol{Q}_i,\boldsymbol{V}_j)}{\sum_{j=1}^{U}S_c(\boldsymbol{Q}_i,\boldsymbol{V}_j)} \tag{5}$$

其中，$\boldsymbol{Q}_i$和$\boldsymbol{V}_j$是正样本对，$\boldsymbol{Q}_i$与剩下的不匹配的 $U-1$ 个 3D 场景样本$\boldsymbol{V}_{\mathrm{neg}}=\{\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_{U-1}\}$组合成负样本对。

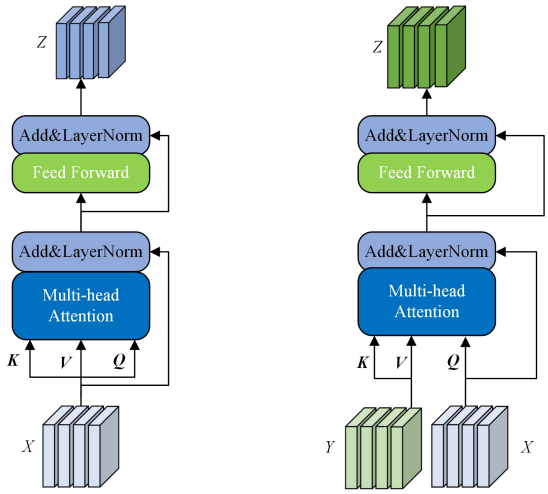
综合$L^{\mathrm{NCE}}(\boldsymbol{V},\boldsymbol{Q})$和$L^{\mathrm{NCE}}(\boldsymbol{Q},\boldsymbol{V})$这两个损失，最终的 3D 跨模态对比损失可以定义为：

$$\mathscr{L}_{vqc}=\frac{1}{2}(L^{\mathrm{NCE}}(\boldsymbol{V},\boldsymbol{Q})+L^{\mathrm{NCE}}(\boldsymbol{Q},\boldsymbol{V})) \tag{6}$$

具体来说，通过最小化$\mathscr{L}_{vqc}$把匹配的 3D 场景和问题之间的互信息最大化，把未匹配的 3D 场景和问题之间的互信息最小化。3D 跨模态对比损失迫使视觉特征和文本特征在嵌入空间中更好地对齐。

3.3　堆叠 Transformer 网络

在视觉表示学习中，点云被送到 VoteNet 进行物体识别，这进行的是局部运算。对于问题来说，BiLSTM 通过全局运算提取句子的特征。此外，3D 点云和问题属于两个不同的模态，因此，来自 VoteNet 的点云特征与来自 BiLSTM 的问题特征有着不同的分布。为了融合视觉特征和文本特征，本文使用堆叠 Transformer 网络。堆叠 Transformer 网络由多个级联的深度交互注意力层(DIA)组成，每一个 DIA 层将处理后的特征表示向下一个 DIA 层传递，逐步细化参与的点云特征和问题特征，将 3D 场景中的对象与问题语义联系起来。这种结构的优势是输入特征的数量等于输出特征的数量，实例没有减少。如图 2 所示，DIA 层是由单模态注意力(SA)单元和跨模态注意力(CA)单元组合而成，首先对 3D 点云和问题这两种模态内的关系进行建模，然后利用跨模态注意力单元对每个对象$\boldsymbol{v}_m\in\boldsymbol{V}$和每个问题$\boldsymbol{q}_n\in\boldsymbol{Q}$之间的密集交互进行建模。SA 单元和 CA 单元的框架如图 4 所示。SA 单元由一个多头注意力层和一个前馈层组成。一般情况下，SA 单元的输入是问题特征或者 3D 对象特征。类似地，CA 单元主要包含多头注意力层和前馈层，它的输入往往是视觉和文本这两种模态的特征。



(a)单模态注意力(SA)　　(b)跨模态注意力(CA)

图 4　两个注意力单元的框架

Fig. 4　Framework of two attention units

具体来说，问题特征 $\boldsymbol{Q}$ 被送到 SA 得到具有深度语义信息的问题表示 $SA(\boldsymbol{Q})$，点云特征 V 经过 SA 获得各个对象之间关联的 $SA(\boldsymbol{V})$，使用 SA 的输出作为 CA 中多头注意力的键和值，在 CA 中输出带有问题信息的 3D 对象表示 $CA(\boldsymbol{V})$。假设第 1 层的 DIA 的输入特征表示为$\boldsymbol{V}^0$和$\boldsymbol{Q}^0$，输出特征表示为 $\boldsymbol{V}^1$和$\boldsymbol{Q}^1$。以此类推，第 h 层 DIA 的输入特征表示为$\boldsymbol{V}^{(h-1)}$和$\boldsymbol{Q}^{(h-1)}$，输出特征可以表示为 $\boldsymbol{V}^h$和$\boldsymbol{Q}^h$。从堆叠 Transformer 网络输出的图像特征$\boldsymbol{V}^{(h)}=\{x_1,x_2,\cdots,x_m\}\in\mathbb{R}^{m\times d}$和问题特征$\boldsymbol{Q}^{(h)}=\{y_1,y_2,\cdots,y_n\}\in\mathbb{R}^{n\times d}$交给多模态融合网络进行聚合。首先，使用两层 MLP(FC(d)-GELU-Dropout(0.1)-FC(1))用于处理 $\boldsymbol{V}^{(h)}$和$\boldsymbol{Q}^{(h)}$得到注意力权重。然后，在每个模态中加入注意力权重。详细过程见式(7)和式(8)。

$$\boldsymbol{V}'=\sum_{k=1}^{m}\mathrm{Softmax}(MLP(\boldsymbol{V}^{(h)}))x_k \tag{7}$$

$$\boldsymbol{Q}'=\sum_{k=1}^{n}\mathrm{Softmax}(MLP(\boldsymbol{Q}^{(h)}))y_k \tag{8}$$

最后，为了稳定地训练，参与特征$\boldsymbol{V}'$和$\boldsymbol{Q}'$被送到 LayerNorm，运算过程如式(9)所示：

$$\boldsymbol{M}=LayerNorm(\boldsymbol{W}_1^{\mathrm{T}}V'+\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{Q}') \tag{9}$$

其中，$\boldsymbol{W}_1$，$\boldsymbol{W}_2 \in \mathbb{R}^{d \times d_m}$ 是两个线性投影矩阵，d_m 是融合特征的维度。

与 Encode-decode 结构相比，堆叠 Transformer 网络每经过一个 SA 都要进行视觉信息和文本信息的交互，减少了初始语义信息的丢失，融合的特征表示包含了更加丰富的信息。

3.4 物体感知和预测答案模块

物体感知和预测答案模块对得出答案起到关键作用，需要将融合特征表示映射到不同的空间，从而完成各类任务。物体感知和预测答案模块主要包括物体定位模块、物体分类模块和预测答案模块。

3.4.1 物体定位模块

物体定位模块的目标是依据问题内容在 3D 场景中定位目标对象。3D 点云往往是不规则的且规模较大。因此，有效地定位关键物体对预测答案至关重要。本文把经过堆叠 Transformer 网络之后的视觉特征 $\boldsymbol{V}^{(H)} \in \mathbb{R}^{m \times d}$ 送入两层 MLP 得到物体定位置信度 $\boldsymbol{O}^l \in \mathbb{R}^n$，从而确定这 n 个对象框与问题相关的可能性。参考前人的工作[21]，该模型使用交叉熵损失来训练这个模块。

3.4.2 物体分类模块

物体分类模块的作用是预测与问题有关联的对象名称。3D 定位任务与这个模块类似，在 3D 定位任务中对 3D 场景的描述含有对象名称，而 3D 场景问答中大多数问题不包含对象名称。这里给出 18 个 ScanNet 基本类别（如“电视”“摇椅”“沙发椅”）。为了预测 18 个类别标签，融合特征 M 被送到两层 MLP，映射到与类别标签相同的空间，经过 softmax 计算可以得到物体分类可能性分数 $\boldsymbol{O}^c \in \mathbb{R}^{18}$。其中，最高的可能性得分对应的类别标签即为预测结果。最后，使用交叉熵损失来训练这一模块。

3.4.3 预测答案模块

与基于图像的视觉问答相似，预测答案模块需要将融合特征表示映射到低维答案特征空间。具体来说，多模态融合特征表示 $\boldsymbol{M}$ 被送到 MLP 并经过 softmax 处理得到向量 $\boldsymbol{A}^c \in \mathbb{R}^e$。向量 $\boldsymbol{A}^c$ 表示 e 个候选答案的可能性得分。这里使用二元交叉熵（BCE）损失函数计算最终分数来训练预测答案模块。

在模型优化的过程中，为了最大化视觉问答任务得出答案的准确性，使用了一个复合损失函数来训练整个模型。损失模块总共包含 5 个部分，分别是物体定位模块的物体定位损失 L_{ol}、物体分类模块的物体分类损失 L_{oc}、3D 目标检测的损失 L_{od}、3D 跨模态对比损失 L_{vqc} 和预测答案模块的答案分类损失 L_{ac}。为了统一训练，本文将这些损失进行简单的线性组合，总损失 L_{total} 的计算方式如式(10)所示：

$$L_{\text{total}} = L_{ol} + L_{oc} + L_{od} + L_{vqc} + L_{ac} \tag{10}$$

4 实验

在本节中，在 ScanQA 数据集上将 3DSSQA 方法与最新的方法进行比较，并进行消融实验以验证 3DSSQA 方法中每个模块的有效性。

4.1 数据集与评价指标

本文方法在 ScanQA 数据集上进行训练和评估。ScanQA 数据集是 3D 问答比赛的官方数据集，建立在 ScanNet 数据集的基础上，包含来自 ScanNet 数据集的 800 个室内场景以及对应的 41 363 个问题和 58 191 个答案[1]。它是使用多个短语创建的，包括自动问答对生成、问题过滤、问题编辑和答案收集。该数据集不仅包含问答对，还包含 3D 对象定位注释。

为了评估 3D 问答的性能，本文使用 EM@1 和 EM@10 作为主要评价指标。其中 EM@N 表示前 N 个候选答案中匹配到正确答案的百分比。因为一些问题的答案可以用多种短语或者句子表述，所以本文使用了经常用于图像描述的句子评估指标。比如用 BLEU[29]，ROUGE-L[30]，METEOR[31]，CIDEr[32] 和 SPICE[33] 指标来评估鲁棒的答案匹配。

4.2 实现细节

本文方法对 3D 场景进行数据增强处理，增加训练样本的数量，提升模型的泛化能力。具体来说，本文在 $-5°\sim5°$ 的范围中围绕 3 个坐标轴以任意角度随机旋转 3D 点云。另外，在所有方向上随机平移点云，平移的距离不超过 0.5m。3DSSQA 方法使用点云的几何信息、预处理的多视图图像特征和法线信息进行训练。在训练的过程中，该模型使用了 Adam[34]，批量大小为 16，初始学习率为 5×10^{-4}，超参数 τ 设置为 0.2。参考 Encode-decode 结构，本文的堆叠 Transformer 网络使用了 6 层 DIA。模型进行了 30 轮训练，一直到收敛为止。在 15 轮之后，每轮的学习率降低了 20%。为了减轻模型对其训练数据的拟合，本文将权重衰减因子设置为 1×10^{-5}。所有实验均在 PyTorch 上使用单个 V100 GPU 实现。

为了验证本文提出的 3DSSQA 方法，将其与以下基线方法进行比较：RandomImage+MCAN[1]，VoteNet+MCAN[1]，ScanRefer + MCAN (pipeline)[1]，ScanQA[1]，ScanRefer + MCAN(e2e)[1]。RandomImage+MCAN 是一个 2D 问答模型，与之进行比较的目的是展示 3D 问答模型具有一定的优越性。该 2D 问答模型使用了预训练的 MCAN[12]，MCAN 基于 Transformer 结构并依靠编码器和解码器完成跨模态信息的交互。因为 2D 问答模型不能直接处理 3D 场景，因此本文在 ScanNet 数据集上运行 2D 视觉问答模型。ScanNet 数据集的图像来自与问答对相关联的目标对象周围的图像。每个问题使用了 3 张图像。VoteNet+MCAN 检测 3D 空间中的对象，并把它们送到 MCAN 中。该方法没有在 3D 空间中对目标对象进行定位。ScanRefer[21] 是一种 3D 对象定位方法，用于将给定的语言描述定位到 3D 空间中的相应目标对象。ScanRefer+MCAN(pipeline) 方法是分两阶段进行的。第一阶段，ScanRefer 使用 VoteNet 识别房间中的对象，然后从候选对象中选择与语言描述相关的对象。第二阶段，将 ScanRefer 定位的对象送到 2D 视觉问答模型 MCAN 中。ScanRefer+MCAN(e2e) 是一种端到端的方法，该方法在学习 3D 定位的同时也进行问答，直接根据对象框特征和问题内容预测答案。

ScanQA 方法使用了编码器和解码器结构，不仅检测 3D 空间中的对象，还要预测对象的类别并对其定位。

4.3 实验结果和分析

表 1 列出了本文提出的 3DSSQA 方法和基线模型的比较结果。为了看起来更直观，每一列中最好的结果都被加粗。

从表 1 可以看出，3DSSQA 的所有评价指标都高于 RandomImage＋MCAN，这验证了 3D 问答模型的性能显著优于 2D 视觉问答模型。这是因为 3D 数据编码对象之间的真实形状属性和空间关系，不会在 2D 图像中因视点改变、遮挡和重投影而引入歧义。在 EM@1 这个主要指标上，本文模型 3DSSQA 比 VoteNet＋MCAN 高 4.59%，验证了物体定位模块和物体分类模块有利于提升 3D 问答模型的性能；3DSSQA 超过 ScanRefer＋MCAN(pipeline)6.78%，超过 ScanRefer＋MCAN(e2e)3.74%，这表明本文方法在解决 3D 问答任务方面是有效的，可以有效地进行 3D 点云特征和问题特征的融合。在表 1 中，3DSSQA 几乎在所有指标上都优于最新的基线 ScanQA，这证明了堆叠 Transformer 网络联合对比学习框架是有效的，3DSSQA 方法在细粒度交互中能细致地理解 3D 场景和问题语义，生成令人满意的答案。

表 1　在 ScanQA 数据集上与已有算法性能比较

Table1　Performance comparison with previous works on ScanQA

Model	EM@1	EM@10	BLEU-1	ROUGE	METEOR	CIDEr	SPICE
RandomImage＋MCAN[1]	22.31	53.11	26.66	31.27	12.13	60.37	9.05
VoteNet＋MCAN[1]	19.71	50.76	29.46	30.97	12.07	58.23	10.44
ScanRefer＋MCAN(pipeline)[1]	17.52	49.92	19.17	24.40	9.38	44.25	6.24
ScanRefer＋MCAN(e2e)[1]	20.56	52.35	27.85	30.68	11.92	57.36	10.58
ScanQA[1]	23.45	**56.51**	31.56	34.34	13.55	67.29	11.99
3DSSQA	**24.30**	55.69	**32.67**	**35.30**	**13.93**	**69.02**	**12.58**

4.4　消融实验

在本节中，本文方法在 ScanQA 数据集上进行消融实验，目的是验证其中每个部分的有效性。这些实验主要通过控制使用或者不使用其中一些关键模块来进行，如表 2 所列。对于 3D 跨模态对比学习来说，与原始模型相比，不加入对比学习的模型在各项指标上都有所下降。这验证了 3D 跨模态对比学习的有效性，它能够缩小两种模态异构的差距，有利于两种模态特征的进一步融合。下面主要看 EM@1 这个指标。没有物体定位模块的 3DSSQA 比默认设置的 3DSSQA 下降了 1.61%，体现了物体定位模块的重要性。物体定位模块能够找到对象框，提供位置信息，协助机器给出准确答案，没有物体分类模块的 3DSSQA 在各项指标上普遍低于原始模型，这验证了物体分类模块的显著作用。物体分类模块提供 3D 场景中的对象类别信息，有效地提升了模型理解 3D 场景的能力。

表 2　在 ScanQA 数据集上的消融研究

Table 2　Ablation studies on ScanQA

预测答案	物体定位	物体分类	对比学习	EM@1	EM@10	BLEU-1	ROUGE	METEOR	CIDEr	SPICE
√	√	√	√	**24.30**	**55.69**	**32.67**	**35.30**	**13.93**	**69.02**	**12.58**
√	√		√	23.67	55.65	30.86	33.94	13.38	66.44	11.17
√		√	√	22.69	55.41	30.43	33.29	13.26	64.89	10.33
√			√	20.40	50.66	26.15	29.64	11.56	56.84	8.39
√	√	√		24.28	55.06	31.58	35.01	13.67	67.86	10.90

结束语　本文提出了一种基于自监督学习的三维真实场景问答方法(3DSSQA)来完成 3D 空间理解任务。3DSSQA 利用 3D 跨模态对比学习使 3D 场景和对应问题的互信息最大化，缩小两种模态的异构差距，从而挖掘两者之间潜在的相关特征。然后，通过堆叠 Transformer 网络融合 3D 场景信息和问题信息，有效地减少了原始信息的丢失。实验结果表明，本文的方法在 ScanQA 数据集上优于基线模型，取得了不错的性能，提升了模型在 3D 空间中感知和推理答案的能力。

参 考 文 献

[1] AZUMA D,MIYANISHI T,KURITA S,et al. ScanQA:3D Question Answering for Spatial Scene Understanding[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:19129-19139.

[2] YAN X,YUAN Z,DU Y,et al. CLEVR3D:Compositional Language and Elementary Visual Reasoning for Question Answering in 3D Real-World Scenes[J]. arXiv:2112.11691,2021.

[3] WANG H,GUO B,ZENG Y,et al. Enabling Harmonious Human-Machine Interaction with Visual-Context Augmented Dialogue System:A Review[J]. arXiv:2207.00782,2022.

[4] KIM K,BILLINGHURST M,BRUDER G,et al. Revisiting trends in augmented reality research:A review of the 2nd decade of ISMAR(2008－2017)[J]. IEEE transactions on visualization and computer graphics,2018,24(11):2947-2962.

[5] MITTAL V. Attngrounder:Talking to cars with attention[C]// European Conference on Computer Vision. Cham:Springer,2020:62-73.

[6] MALINOWSKI M,ROHRBACH M,FRITZ M. Ask your neurons:A neural-based approach to answering questions about images[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:1-9.

[7] GAO H,MAO J,ZHOU J,et al. Are you talking to a machine? dataset and methods for multilingual image question[C]// Advances in Neural Information Processing Systems. 2015:2296-2304.

[8] KIM J H,LEE S W,KWAK D,et al. Multimodal residual learning for visual qa[C]// Advances in Neural Information Processing Systems. 2016:361-369.

[9] SHIH K J,SINGH S,HOIEM D. Where to look:Focus regions for visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:4613-4621.

[10] KAZEMI V,ELQURSH A. Show,ask,attend,and answer:A

strong baseline for visual question answering[J]. arXiv:1704.03162,2017.
[11] YANG Z,HE X,GAO J,et al. Stacked attention networks for image question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:21-29.
[12] YU Z,YU J,CUI Y,et al. Deep modular co-attention networks for visual question answering[C]// Proceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition. 2019:6281-6290.
[13] RAHMAN T,CHOU S H,SIGAL L,et al. An improved attention for visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:1653-1662.
[14] ZHOU Y,REN T,ZHU C,et al. Trar:Routing the attention spans in transformer for visual question answering[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:2074-2084.
[15] LI J,SELVARAJU R,GOTMARE A,et al. Align before fuse:Vision and language representation learning with momentum distillation[J]. Advances in Neural Information Processing Systems,2021,34:9694-9705.
[16] ZENG Y,ZHANG X,LI H. Multi-Grained Vision Language Pre-Training:Aligning Texts with Visual Concepts[J]. arXiv:2111.08276,2021.
[17] WANG P,YANG A,MEN R,et al. OFA:Unifying Architectures,Tasks,and Modalities Through a Simple Sequence-to-Sequence Learning Framework[C]// International Conference on Machine Learning. PMLR,2022:23318-23340.
[18] YE S,CHEN D,HAN S,et al. 3D Question Answering[J]. arXiv:2112.08359,2021.
[19] YANG J,DUAN J,TRAN S,et al. Vision-Language Pre-Training with Triple Contrastive Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:15671-15680.
[20] WANG W,BAO H,DONG L,et al. Vlmo:Unified vision-language pre-training with mixture-of-modality-experts[J]. arXiv:2111.02358,2021.
[21] CHEN D Z,CHANG A X,NIEßNER M. Scanrefer:3d object localization in rgb-d scans using natural language[C]// European Conference on Computer Vision. Cham:Springer,2020:202-221.
[22] CHEN Z,GHOLAMI A,NIEßNER M,et al. Scan2cap:Context-aware dense captioning in rgb-d scans[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:3193-3203.
[23] DAI A,NIEßNER M. 3dmv:Joint 3d-multi-view prediction for 3d semantic scene segmentation[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018:452-468.
[24] QI C R,LITANY O,HE K,et al. Deep hough voting for 3d object detection in point clouds[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:9277-9286.
[25] QI C R,YI L,SU H,et al. Pointnet++:Deep hierarchical feature learning on point sets in a metric space[J]. Advances in Neural Information Processing Systems,2017,30:5099-5108.
[26] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.
[27] MISRA I,MAATEN L. Self-supervised learning of pretext-invariant representations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:6707-6717.
[28] OORD A,LI Y,VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv:1807.03748,2018.
[29] PAPINENI K,ROUKOS S,WARD T,et al. Bleu:a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002:311-318.
[30] LIN C Y. Rouge:A package for automatic evaluation of summaries[C]// Text Summarization Branches Out. 2004:74-81.
[31] BANERJEE S,LAVIE A. METEOR:An automatic metric for MT evaluation with improved correlation with human judgments[C]// Proceedings of the ACL Workshop On Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005:65-72.
[32] VEDANTAM R,LAWRENCE ZITNICK C,PARIKH D. Cider:Consensus-based image description evaluation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:4566-4575.
[33] ANDERSON P,FERNANDO B,JOHNSON M,et al. Spice:Semantic propositional image caption evaluation[C]// European Conferenceon Computer Vision. Cham:Springer,2016:382-398.
[34] KINGMA D P,BA J. Adam:A method for stochastic optimization[J]. arXiv:1412.6980,2014.

LI Xiang, born in 1997, postgraduate. His main research interests include visual question answering and so on.

LI Xuexiang, born in 1965, professor, master supervisor. His main research interests include high performance computing and cloud computing.

(责任编辑:杨雪敏)