DOI：10.13718/j.cnki.xdzk.2024.10.018
原蕾，王科俊. 利用人工智能神经网络体系结构生成视觉问答系统中的自然语言解释 [J]. 西南大学学报(自然科学版)，2024，46(10)：212-221.

利用人工智能神经网络体系结构生成视觉问答系统中的自然语言解释

原蕾[1]，　王科俊[2,3]

1. 郑州工商学院 信息工程学院，郑州 451400；2. 北京理工大学(珠海校区) 信息学院，广东 珠海 519088；
3. 哈尔滨工程大学 智能科学与工程学院，哈尔滨 150001

摘要：模型可解释性长期以来一直是人工智能领域备受关注的问题. 在视觉问答(Visual Question Answering，VQA)系统中，特别需要处理视觉(图像)和语言(问题)之间的协同推理，以产生解释性强且可靠的答案. 然而，现有方法通常集中于单独处理视觉和语言特征，未能捕捉到 VQA 所需的高低级交互关系，也未能提供答案生成过程的解释. 针对以上问题，该研究提出一种创新方法，即基于 Transformer 的可解释路径 VQA 方法. 首先利用 Transformer 编码器层分别提取预训练的卷积神经网络(Convolutional Neural Network，CNN)和领域特定语言模型(Language Model，LM)的视觉和语言特征. 随后，解码器层被嵌入，并对编码特征进行上采样，用于最终的 VQA 预测. 通过在具有挑战性的 VQA-X 数据集和 e-SNLI-VE 数据集上大量的实验验证了该方法的有效性. 实验结果表明：该方法在定性和定量评估方面明显优于其他先进方法，不仅有助于解释 VQA 模型的单幅图像结果，还为理解 VQA 模型的行为提供了有益参考.

关　键　词：可解释；视觉系统；人工智能；神经网络；转换器
中图分类号：TP393　　**文献标志码**：A
文 章 编 号：1673-9868(2024)10-0212-10　　开放科学(资源服务)标识码(OSID)：

Natural Language Explanations in Visual Question Answering Systems Generated Using Artificial Intelligence Neural Network Architectures

YUAN Lei[1]，　WANG Kejun[2,3]

1. *School of Information Engineering，Zhengzhou Technology and Business University，Zhengzhou 451400，China*；
2. *School of Information，Beijing Institute of Technology(Zhuhai Campus)，Zhuhai Guangdong 519088，China*；
3. *College of Intelligent Systems Science and Engineering，Harbin Engineering University，Harbin 150001，China*

收稿日期：2023-12-19
基金项目：教育部产学合作协同育人项目(220600440151815).
作者简介：原蕾，硕士，讲师，主要从事计算机网络研究.

Abstract：The interpretability of models has long been a prominent challenge in the field of artificial intelligence. In Visual Question Answering (VQA) systems, particularly, there is a critical need to facilitate collaborative reasoning between visual (image) and linguistic (question) components in order to generate answers that are both highly interpretable and reliable. However, existing methods often focus on separately handling visual and linguistic features, failing to capture the intricate interplay required for VQA and lacking in providing explanations for the answer generation process. To address these issues, this study explores and introduces an innovative approach, known as Interpretable Transformer-Based Path Visual Question Answering. This method begins by leveraging Transformer encoder layers to separately extract visual and linguistic features from pre-trained Convolutional Neural Network (CNN) and domain-specific language model (LM). Subsequently, decoder layers are embedded to upsample encoded features for the final VQA predictions. Extensive experiments conducted on challenging VQA-X datasets and e-SNLI-VE datasets validate the effectiveness of this approach. Experimental results indicated that the proposed method outperforms other state-of-the-art methods in qualitative and quantitative evaluations. This research not only contributes to elucidating single-image results in VQA models but also provides profound insights into understanding the behavior of VQA models.

Key words：interpretable；vision systems；artificial intelligence；neural networks；converters

在过去的几年里，人工智能(Artificial Intelligence，AI)领域取得了显著进展，尤其是在自然语言处理(Natural Language Processing，NLP)和计算机视觉(Computer Vision，CV)领域[1]．这些领域的迅速发展引领了新一代智能系统崭露头角，其中包括自动化问答系统，它们的出现和发展引起了业界的广泛关注和研究．

视觉问答是人工智能领域的一项新兴课题，该课题结合了计算机视觉和自然语言处理两个学科领域的知识，其任务是把给定的视觉信息(图像)和与视觉信息相关的自然语言问题作为输入，生成的自然语言答案作为输出，即输入图像和与图像相关的文本问题，并输出确定的正确答案．该领域的研究受到了广泛关注，因其具有广泛的应用前景，例如在虚拟助手、医疗诊断、自动驾驶和智能客服等领域[2-5]．然而，尽管 VQA 系统在回答问题方面取得了显著进展，但这些系统在解释决策和答案生成过程方面仍然存在挑战．这些不足是由于 VQA 系统通常被视为黑盒子，用户难以理解为何系统会给出特定的答案所致[6]．这样缺乏解释性不仅限制了 VQA 系统在关键任务中的应用，还降低了用户对系统的信任度和接受度．因此，提高 VQA 系统的可解释性成为当前研究的热点问题之一．

在 VQA 领域，可解释性意味着系统能够清晰地解释其答案生成的依据和过程，使用户能够理解系统的决策逻辑，这不仅包括系统对问题和图像的理解，还包括系统对答案生成路径的解释．譬如在回答关于图像中物体的问题时，一个具有良好可解释性的 VQA 系统应该能够解释为何选择了某个特定的物体作为答案，并提供与这一选择相关的推理过程．可解释性 VQA 系统的重要性不仅仅体现在用户交互和决策支持方面，还涉及到伦理和法律等更为广泛的社会问题．在一些应用中 VQA 系统的决策可能会对人们的生活产生直接影响，因此这些决策必须能够被清晰地解释和追溯．例如在医疗领域，AI 系统用于辅助医生进行疾病诊断，可解释性 AI 系统可以提供关于诊断依据的详细信息，使医生和患者能够理解系统的决策逻辑，并作出明智的治疗决策．在自动驾驶领域，可解释性也是一个重要的问题，自动驾驶车辆需要做出复杂的决策，例如避免碰撞、超车和停车等．如果这些决策不能被解释，必将难以确定责任和安全性．在金融领域，可解释性 AI 系统可以帮助分析师和投资者更好地理解市场趋势和交易建议，有助于制定更明智的

投资策略，并降低金融风险. 一些国家和地区也出台了法规要求AI系统具有可解释性，以确保其决策公正且不受偏见影响[7].

为了提高VQA系统的可解释性和效果，本文提出一种使用Transformer编码器层和解码器层的统一方法，充分利用完整Transformer架构的优势，融合视觉和语言特征为检索到的答案提供解释. 实验结果表明，与一些最先进的方法相比，本文方法可以更准确地生成答案，并且解释更合理、更接近事实. 本文的主要目的和意义是通过引入自然语言解释，提高VQA系统的可解释性和效果，旨在开发一种创新的神经网络体系结构(Transformer)，该体系结构不仅能够回答问题，还能够以自然语言的形式解释答案生成的过程，从而使用户更易理解系统的决策过程. 这种自然语言解释不仅有助于用户理解系统的决策，还可以提供关于答案推理过程的详细信息，使用户能够追踪答案的生成路径.

1 基础概念及研究现状

1.1 视觉问答

VQA任务涉及回答有关图像的问题，需要理解图像内容和文本信息. 在VQA系统中，问题和图像使用文本和视觉特征化技术中的一种或多种分别嵌入，然后使用融合技术(如串联、元素相乘或关注等)将文本和视觉特征向量结合起来. 从融合阶段获得的向量可以使用分类技术进行分类，或者可以将其用于VQA生成问题的答案. 图1显示了整个VQA系统的结构.

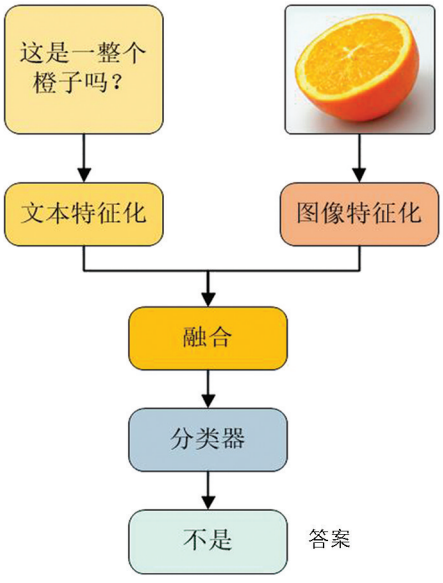

图1 VQA系统结构

Truong等[8]首次提出了视觉问题解答，每当提供一幅图像和一个有关该图像的自然语言问题时，模型应能为该问题提供准确的答案. 他们的系统由2个模块组成：① 多层感知(MLP)神经网络分类器，它负责处理图像和问题的特征；② 长短期记忆(Long Short-Term Memory，LSTM)模块，紧随其后的是Softmax层，用于生成最终的答案. 为了获得答案，首先将VGGnet最后一个全连接层的视觉特征与LSTM编码的问题表示相集成，然后将它们传递给分类器. 王虞等[9]通过迁移学习和跨模式门控方法，采用基于注意力的机制来提高VQA性能；高鸿斌等[10]认识到当人类回答有关图像的问题时，他们会考虑图像中包含的信息和图像之外的背景知识. 为了解决这个问题，他们提出了外部知识VQA，并使用外部知识来回答视觉问题. 该研究关于外部知识重要性的发现强调了参考多种信息来源的重要性；Li等[11]提出了一种模型，该模型使用带有Resnet的Faster R-CNN(Faster Region-based Convolutional Neural Network)来提取图像特征，然后使用onehot词向量对问题进行编码，并由门控循环单元(Gated Recurrent Unit，GRU)生成问题表示；Sharma等[12]提出了视觉问题解答的再关注(Re-Attention for Visual Question Answering)模型，该模型利用答案中包含的有价值信息，再次关注图像以获得更精确的答案.

1.2 视觉问答的自然语言解释生成

深度学习模型的可解释性一直是研究人员关注的一个重要领域，而对决策的可视化解释一直是各种研究工作的主题. 本文提出了对问答过程进行合理解释作为增强VQA模型可靠性的解决方案. 然而，传统的解释是生成模型并使用热图来表示图像不同区域对决策过程的影响. 这些方法利用数学原理可视化对神经网络底层逻辑进行分析，对于发展深度学习研究至关重要. 但对于大多数人工智能系统的非专家用户来

说，这些解释是抽象的，甚至是不可理解的. 因此，需要一种能够帮助他们简单判断深度学习给出的判断可靠性的方法.

为了解决这个问题，学者们提出了用自然语言生成人类可理解的 VQA 解释的模型，取代了抽象的热图，这样即便是无专业背景知识的个人也能理解模型的解释. 然而，这些模型使用了与回答问题不同的解释生成模型，使得验证回答模型是否正确理解了解释中的逻辑变得更加困难. 为了解决影响模型生成的解释能够可靠性的局限性，Guo 等[13]使用单一模型生成答案和解释，确保生成的解释能够准确反映模型用于回答问题的逻辑. 但该方法只参考了输入图像，由于缺乏参考信息，模型的性能受到了限制. 目前，基于预训练 Transformer 的大规模模型，如生成预训练 Transformer(GPT)-2[14]模型和对比语言图像预训练(CLIP)[15]模型迅速发展，这些模型提高了理解图像的能力以及使用大量训练数据生成句子的准确性. 尽管这些模型保证了底层逻辑的一致性，但仍然存在许多生成解释难以置信的情况.

2　方法

问题定义：给定图像 X 和相关问题 V；VQA 任务的目标是预测答案 $\hat{G}$. 在数学上，这个问题可以表述为：

$$\hat{G} = f(X, V, \theta) \tag{1}$$

其中，θ 为模型参数，f 为答案预测函数.

架构概述：如图 2 所示，本文模型由 4 个主要部分组成：① 使用多媒体领域 LM (MulELMo)和 BiLSTM 提取问题特征，以捕捉上下文信息；② 使用 ResNet 结合 CNN 进行图像特征提取，捕获低级特征；③ Transformer Encoder 用于融合提取的图像(视觉)和问题(语言)特征，并建立高级全局特征模型；④ Transformer Decoder 用于对编码特征进行上采样，以进行最终预测.

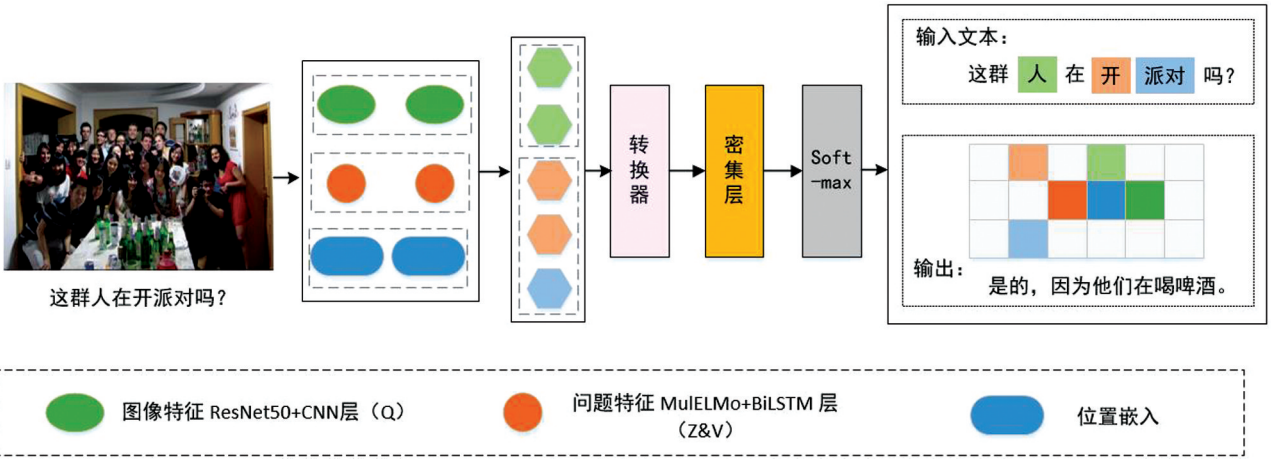

图 2　本文模型的总体架构

2.1　问题特征提取

为了提取问题特征，本文尝试了各种通用和特定领域的 LM(Language Model)，例如 ELMo(Embeddings from Language Model)、BERT(Bidirectional Encoder Representations from Transformer)和 MulBERT(Multimodal BERT)，最终选择表现最好的 MulELMo. 使用 BiLSTM 的 MulELMo(MulELMo 是在多媒体摘要文本上进行训练的)，其具有与 ELMo 相同的网络结构. ELMo 是从 Bi-LM 中学习到的这些特征的特定任务组合，其中所有层都被扁平化为单一向量，如公式(2)所示.

$$ELMo_t^{task} = E(W_t; \Theta^{task}) = \gamma^{task} \sum_{y=0}^{L} s_y^{task} b_{(b,y)}^{LW} \tag{2}$$

其中，s^{task} 为串联多个层表示的 $softmax$ 归一化权重；γ^{task} 为优化和缩放的超参数；Θ^{task} 为特定任务的参数. L 为网络层数，W_t 为单词 t 的词嵌入，$b^{LW}_{(b,y)}$ 为第 y 层 BiLSTM 的隐藏状态.

使用预训练的 MulELMo 来提取给定问题 V 的上下文特征，如公式(3)所示. MulELMo 在各种多媒体文本挖掘任务中很大程度上优于 ELMo 和之前最先进的方法 $\boldsymbol{I}_V$.

$$\boldsymbol{I}_V = MulELMo(V) \tag{3}$$

利用 $MulELMo$，一个 1 024 维向量 $\boldsymbol{I}_V$ 被送入 BiLSTM 层，然后对来自两个方向的信息建模，并连接前向 $\overrightarrow{b_x}$ 和后向 $\overleftarrow{b_x}$. ($\overrightarrow{b_x}$ 和 $\overleftarrow{b_x}$ 分别为前向和后向 LSTM 在时间 x 的隐藏表示，在给定时间 x 的输出隐藏表示为 b_x，如公式(4)所示.

$$b_x = [\overrightarrow{b_x} \parallel \overleftarrow{b_x}] \tag{4}$$

其中，$\parallel$ 为连接运算符，BiLSTM 处理后的问题特征 $\boldsymbol{I}_V \in \boldsymbol{R}^{l\times d}$，$l$ 为问题长度，d 为每个词的向量大小. 在转发到 BiLSTM 层之前，$\boldsymbol{I}_V$ 被填充以匹配最大问题长度 $l_{\max}$ 到 $\boldsymbol{I}^{pad}_V$.

$$\boldsymbol{I}^l_V = BiLSTM(\boldsymbol{I}^{pad}_V) \tag{5}$$

其中，填充到最大问题长度的问题向量 $\boldsymbol{I}^{pad}_V \in \boldsymbol{R}^{l\max\times d}$，问题特征 $\boldsymbol{I}^l_V \in \boldsymbol{R}^{l\max\times 512}$；$\boldsymbol{I}^l_V$ 将经过一个密集层、一个位置编码层和一个丢失层，并输出一个最终问题特征矩阵.

2.2 图像特征提取

ResNet(Residual Network)与 CNN(Convolutional Neural Network)：使用预训练的 VGG19、InceptNet、DenseNet 和 ResNet50 进行提取图像特征比较(图 3)，最终选择性能最佳的 ResNet50 作为本文图像预训练的神经网络结构.

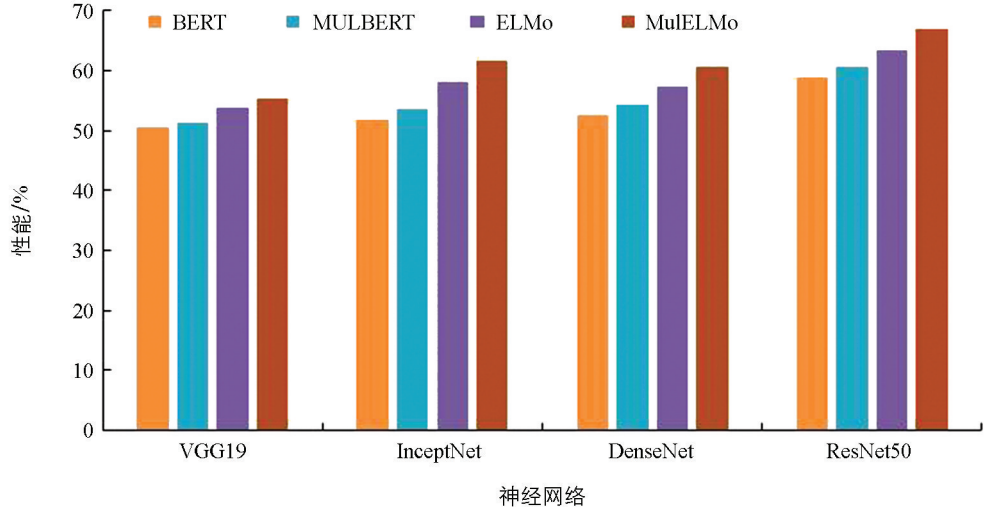

图 3 不同网络结构的性能

本文重塑了图像 X，使其与 ResNet50 的形状(224，224，3)相匹配，不需要 ResNet50 作为分类器，而是作为特征提取器，因此去掉了最后 3 个全连接层，只保留最后一个平均池化层的输出作为图像特征 $\boldsymbol{I}_X$.

$$\boldsymbol{I}_X = ResNet50(X) \tag{6}$$

其中，$ResNet50()$是预训练 ResNet50 模型的特征提取器. $\boldsymbol{I}_X$ 被输送到另一个核大小为 3 的二维 CNN 层，即 ReLu 的激活函数，并转发到一个密集层以缩小通道. 重塑和扁平化是为了尽可能多地保留信息，并输出由 $\boldsymbol{I}^l_X$ 表示的图像特征矩阵，如公式(7)所示. 这种结构在图像特征矩阵 $\boldsymbol{I}^l_X$ 的第一个维度与问题特征矩阵 $\boldsymbol{I}^l_V$ 的第一个维度相匹配的同时，保留了尽可能多的信息.

$$\boldsymbol{I}^l_X = Convolution2D(ResNet50(\boldsymbol{I}_X)) \tag{7}$$

其中，$\boldsymbol{I}_X^l \in \boldsymbol{R}^{7\times7\times512}$，$\boldsymbol{I}_X \in \boldsymbol{R}^{l\max\times512}$. Convolution2D 表示二维卷积操作，$ResNet50(\boldsymbol{I}_X)$是 ResNet50 模型的输出，即图像 X 的特征表示.

2.3　转换器

Transformer 由编码器—解码器结构组成，如图 4 所示.

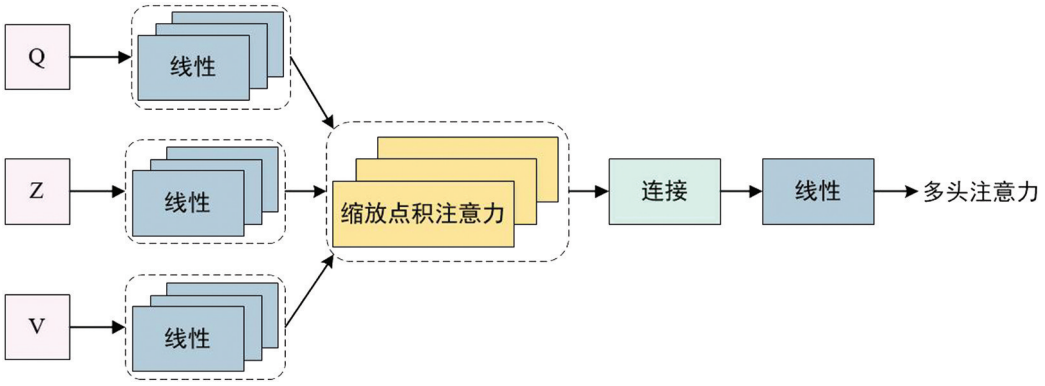

图 4　Transformer 结构图

每个编码器层都由多头自注意力和前馈神经网络组成. 与编码器一样，解码器也有 3 个子层，其中两个子层与编码器类似(多头自注意力和前馈)，而第 3 个子层对编码器的输出进行多头注意. 输入矢量首先被转换成 3 个不同的向量：值向量 $\boldsymbol{v}$、键向量 $\boldsymbol{z}$ 和查询向量 $\boldsymbol{q}$，然后将不同输入的向量组合成 3 个不同的矩阵，即 $\boldsymbol{Q}$、$\boldsymbol{Z}$ 和 $\boldsymbol{V}$.

$$Att(\boldsymbol{V},\boldsymbol{Z},\boldsymbol{Q}) = softmax\left(\frac{\boldsymbol{V}\cdot\boldsymbol{Z}^{\mathrm{T}}}{\sqrt{d_z}}\right)\cdot\boldsymbol{Q} \tag{8}$$

其中，值矩阵 $\boldsymbol{V}\in\boldsymbol{R}^{L\times d}$，键矩阵 $\boldsymbol{Z}\in\boldsymbol{R}^{L\times d}$，查询矩阵 $\boldsymbol{Q}\in\boldsymbol{R}^{L\times d}$，$\boldsymbol{Z}^{\mathrm{T}}$ 为矩阵 $\boldsymbol{Z}$ 的转置矩阵，L 为序列长度，d 为特征深度，d_z 为键矩阵 $\boldsymbol{Z}$ 的维度，$\sqrt{d_z}$ 为其平方根(即缩放因子，用于缩放点积，防止经过 softmax 函数后的梯度消失). 查询被传递给组件，组件搜索最相似的键，并返回与该键相关的值. 两个矩阵乘法以及一个 softmax 函数有助于加快该过程. softmax($\boldsymbol{V}\cdot\boldsymbol{Z}^{\mathrm{T}}$)会生成一个概率分布，其峰值位于相关查询关键字的位置. 这可以作为一个伪掩码，通过将其与 $\boldsymbol{V}$ 进行矩阵相乘，可以得到网络首先需要关注的集中值.

2.3.1　转换器编码器

本文仅使用 Transformer 编码器层来融合前面步骤中提取的图像(视觉)和问题(语言)特征. 在原始 Transformer 编码器中，编码器的输入($\boldsymbol{Q}$，$\boldsymbol{V}$，$\boldsymbol{Z}$)是一串单词，在这里用图像和问题特征对其进行了修改和替换.

在第一层编码器中，图像特征矩阵 $\boldsymbol{I}_X^l$ 用作 $\boldsymbol{Q}$，问题特征矩阵 $\boldsymbol{I}_V^l$ 作为 $\boldsymbol{V}$ 和 $\boldsymbol{Z}$ 的输入，并进行位置编码. 在第二层编码器中，再次使用 $\boldsymbol{I}_X^l$ 作为输入 $\boldsymbol{Q}$，并将第一层编码器中的输出转发给第二层编码器的 $\boldsymbol{V}$ 和 $\boldsymbol{Q}$. 这里输入的 $\boldsymbol{Q}$，$\boldsymbol{V}$ 和 $\boldsymbol{Z}$ 的处理方式与原始 Transformer 编码器层相同.

2.3.2　转换器解码器

本文使用与原始 Transformer 相同的 Transformer 解码层对编码特征进行上采样，然后再次使用两个解码器层进行最终预测. 首先，〈start〉标记的热向量经过可训练嵌入层，将位置编码输入解码器层，$softmax$ 函数将给出每个单击向量的概率分布. 解码器将选取概率最高的单击向量，并将相应的词汇添加到答案中. 解码过程一直持续到解码器生成〈end〉标记的单击向量为止. 在这里，解码器层的工作机制与原始 Transformer 解码器层的工作机制相同.

3 实证结果与分析

3.1 实验设置

本文实验使用VQA-X数据集和e-SNLI-VE数据集. VQA-X数据集是VQA-v2数据集的扩展. VQA-X数据集包含每个答案的解释，是一个结合了视觉和语言的多模态数据集. VQA-X数据集由超过28 000张图像的33 000个问答(QA)对组成，所有这些都来自COCO2014数据集. 具体来讲，使用COCO2014训练集中的24 876张图像作为本文实验的训练集，其中包含29 459个QA对. 为了创建验证集和测试集，本文以3∶4的比例对COCO2014验证集进行分区，分别产生大约1 500个QA对和2 000个QA对. e-SNLI-VE数据集由来自Flickr30k数据集的30 000多张图像组成，每个图像都有几个相应的QA对. 训练集和测试集分别有401 717个QA对和14 740个QA对. 本文实验数据集如图5所示，每张图片都有一个问题及其解释. “A1-A10”表示第n个答案，“E1-E10”表示第n个解释.

图5 本文实验使用的数据集示例

本文选择视觉和语言编码器的主要原因是它们在跨模态提取特征和匹配特征方面的效率. 传统的视觉模型通常是为图像分类和分割等特定任务而设计的，而本文提出的模型可以同时使用视觉和语言编码器来处理图像和文本匹配的多模态任务.

要检索对回答问题和生成解释有用的外部知识，需要一个包含各种类型知识信息的知识库. 维基百科因其结构化的数据收集而成为各种应用的宝贵资源. 本文的问题主要集中在日常生活、体育和动物等主题上，实验从维基百科中选择了最近提出的一个子集作为外部知识集，该子集遵循KAT(Knowledge Augmented Transformer)的方法. 本文所有实验都使用相同的外部知识集；在外部知识检索方面，使用Faiss计算向量之间的余弦相似度来完成检索.

对于实验中的图像和语言多层感知器，本文设置了两个隐藏层，第一层有512个节点，第二层有4 096个节点. 所有模型在训练过程中均训练了30个epoch，以确保模型收敛. 所有图像的大小均调整为224×224像素，并进行随机翻转以防止过度拟合. 学习率最初设置为2×10^{-5}，并逐渐减少直至1×10^{-5}，批量大小固定为32. 所有实验的参数设置与已发表的对比模型源代码一致，未做任何调整，以确保对比实验的公平性. 本文模型使用具有AMD(Advanced Micro Devices)，EPYC 7713P 64核/128线程处理器，512 GB内存和NVIDIA CorporationGA100显卡，总训练时间为9 h.

3.2 评估指标

将本文所提出的模型与其他3种最先进的自然语言模型(文献[16]、文献[17]和文献[18])进行定量和

定性评估，其中文献[16]方法不参考图像信息直接回答问题，并生成相应的解释，由单独的问答和解释生成模型. 文献[17]方法可作为评估可解释视觉语言任务的基准，它引入了统一的评估框架，有独立的问答模型和解释生成模型. 文献[18]是一种本质上既紧凑又通用，并且忠实于底层数据的语言模型，该模型引入大规模视觉和语言来生成解释并取得了良好的性能.

本文使用通用的语言建模评价指标，即 BLEU(Bilingual Evaluation Understudy)、METEOR(Metric for Evaluation of Translation with Explicit Ordering)、ROUGE(Recall-Oriented Understudy for Gisting Evaluation)、SPICE(Semantic Propositional Image Caption Evaluation)和 CIDEr(Consensus-based Image Description Evaluation)，来评价生成的解释. 同时，使用准确性来评估生成的答案. 然后，使用公开项目计算所有语言指标得分.

3.3　实验结果

3.3.1　定量分析

如表 1 和表 2 所示，VQA-X 和 e-SNLI-VE 数据集上的语言度量结果证明，本文模型通过 Transformer 捕获全局关系，表现优于其他 3 种对比模型. 这表明转换器中的缩放点积注意力模块生成的注意力图像，突出显示了负责每个生成文本标记的图像区域，从而提高了生成答案的正确性和生成解释的合理性.

表 1　本文模型与其他 3 种对比模型在 VQA-X 数据集上的性能比较　%

模型	解释					回答
	BLEU	METEOR	ROUGE	SPICE	CIDEr	准确率
文献[16]	59.1	21.5	48.3	18.8	92.7	64.5
文献[17]	63.2	22.4	51.6	16.7	103.6	68.2
文献[18]	58.7	19.5	45.7	19.3	89.5	63.6
本文	65.6	24.2	52.5	21.4	108.3	70.8

表 2　本文模型与其他 3 种对比模型在 e-SNLI-VE 数据集上的性能比较　%

模型	解释					回答
	BLEU	METEOR	ROUGE	SPICE	CIDEr	准确率
文献[16]	48.9	19.4	38.2	33.9	87.6	68.7
文献[17]	53.1	20.3	41.5	31.8	98.5	72.4
文献[18]	48.6	17.4	35.6	34.4	84.4	70.5
本文	55.5	22.1	42.4	36.5	103.2	75.3

本文还在 VQA-X 数据集上进行了几次消融实验，以分别评估缺少图像和缺少文本信息对模型的影响，如表 3 所示.

表 3　VQA-X 数据集上特征融合的消融结果　%

方法	解释					回答
	BLEU	METEOR	ROUGE	SPICE	CIDEr	准确率
Transformer(图像＋文本)	65.6	24.2	52.5	21.4	108.3	70.8
文本功能	56.8	19.7	48.1	17.6	97.6	64.7
图像功能	58.5	21.4	50.3	18.7	104.2	66.5

由表 3 可知，当在整个任务中仅引入图像功能时，所有指标上的性能都会下降. 在整个任务中使用纯文本功能，模型在所有指标上的性能也有所下降，但其结果比仅使用图像功能好一些，表明单独引入文本参考信息的模型优于仅引入图像参考的模型. 当同时引入所有图像和文本内容时，可以获得最佳性能. 实

验结果证明本文模型中使用了 Transformer 架构的优势，同时融合了视觉和语言特征的优越性.

3.3.2 定性分析

为了评估本文模型与其他对比模型的使用情况，本文进行了可解释的定性分析，结果如图 6 所示. 样本中的每个图像都附有使用本文模型生成的结果. 括号中标出了正确答案，并用蓝色部分标记了生成过程中有价值的信息. 从图 6 中可以清楚地观察到，与其他先进的模型相比，本文模型的解释更接近图像并且包含更多细节. 实验结果表明，本文基于 Transformer 的模型可以为相关单词分配更多权重，并解释检索到答案的原因，从而有效地提高模型的可信度.

图 6 本文方法与其他先进方法的定性比较

4 结语

随着人工智能的快速发展，视觉问答系统作为自然语言处理和计算机视觉交叉领域的热点问题，已引起广泛关注. 在 VQA 系统中，特别是在需要用于关键任务或与人类用户互动的情况下，视觉问答系统的可解释性变得至关重要. 然而，以前的研究仅考虑输入图像，造成信息不足的状况，从而导致错误的答案和令人难以信服的解释. 为此，本文提出一种新颖且可解释生成路径的视觉问答方法. 该方法利用 Transformer 编码器和解码器层来嵌入 VQA 任务的视觉和语言特征；然后将低级图像特征嵌入到特定领域上下文信息中；最后利用这些信息来回答问题. 本文模型利用 CNN 优势在低层提取图像特征，并利用特定领域语言模型提取特定领域的上下文信息，通过 Transformer 捕获高层的全局依赖关系. 在两个流行的基准数据集(VQA-X 和 e-SNLI-VE)上体现了本文模型的先进性能，并通过大量实验证明了本文模型的有效性和可解释性. 本文不仅有助于用户理解系统决策，还可以提供关于答案推理过程的详细信息，使用户能够追踪答案的生成路径. 此外，本文还为深度学习和自然语言处理领域提供了一个创新方法，将注意力机制、LSTM 和 CNN 相结合，以处理跨模态信息. 这一方法可以在其他领域的问题中得到应用，为多模态数据处理和可解释 AI 的研究提供了新的思路. 然而，尽管该方法提供了对 VQA 模型行为的解释，但对于复杂问题和答案仍存在局限性，特别是在面对抽象问题或需要长期推理的问题时，解释的复杂性可能会增加，需要更深入的解释机制. 未来的研究会致力于解决更复杂问题的解释，将涉及到更多的推理机制、对话式 VQA 或需要更多步骤推理的问题. 此外，研究人员还将探索该方法在不同领域的应用，如医疗、法律、教育等，以评估模型的域适应性.

参考文献：

[1] 高楠，彭鼎原，傅俊英，等. 基于专利 IPC 分类与文本信息的前沿技术演进分析——以人工智能领域为例 [J]. 情报理论与实践，2020，43(4)：123-129.

[2] 朱翌，李秀. 医学图像描述综述：编码、解码及最新进展 [J]. 中国图象图形学报，2023，28(7)：1990-2010.

[3] 叶仕俊，张鹏程，吉顺慧，等. 人工智能软件系统的非功能属性及其质量保障方法综述 [J]. 软件学报，2023，34(1)：103-129.

[4] GUO Z H，HAN D Z. Sparse Co-Attention Visual Question Answering Networks Based on Thresholds [J]. Applied Intelligence，2023，53(1)：586-600.

[5] CONG F Z，XU S B，GUO L，et al. Anomaly Matters：an Anomaly-Oriented Model for Medical Visual Question Answering [J]. IEEE Transactions on Medical Imaging，2022，41(11)：3385-3397.

[6] 秦志金，赵荧荧，李凡，等. 多模态语义通信研究综述 [J]. 通信学报，2023，44(5)：28-41.

[7] 朱明婷，徐崇利. 人工智能伦理的国际软法之治：现状、挑战与对策 [J]. 中国科学院院刊，2023，38(7)：1037-1049.

[8] TRUONG L X，PHAM V Q，VAN NGUYEN K. Transformer-Based Approaches for Multilingual Visual Question Answering [J]. International Journal of Asian Language Processing，2022，32(4)：1-18.

[9] 王虞，孙海春. 视觉问答技术研究综述 [J]. 计算机科学与探索，2023，17(7)：1487-1505.

[10] 高鸿斌，毛金莹，王会勇. K-VQA：一种知识图谱辅助下的视觉问答方法 [J]. 河北科技大学学报，2020，41(4)：315-326.

[11] LI H Y，HAN D Z. Multimodal Encoders and Decoders with Gate Attention for Visual Question Answering [J]. Computer Science and Information Systems，2021，18(3)：1023-1040.

[12] SHARMA H，SRIVASTAVA S. Visual Question Answering Model Based on the Fusion of Multimodal Features by a Two-Way Co-Attention Mechanism [J]. The Imaging Science Journal，2021，69(1-4)：177-189.

[13] GUO Z H，HAN D Z. Sparse Co-Attention Visual Question Answering Networks Based on Thresholds [J]. Applied Intelligence，2023，53(1)：586-600.

[14] ZHU H，TOGO R，OGAWA T，et al. Multimodal Natural Language Explanation Generation for Visual Question Answering Based on Multiple Reference Data [J]. Electronics，2023，12(10)：1-19.

[15] BAZIY，RAHHALM M A，BASHMALL，et al. Vision-LanguageModel for Visual Question AnsweringinMedicalImagery [J]. Bioengineering，2023，10(3)：1-17.

[16] XUX，WANGT，YANGY，et al. RadialGraph Convolutional Network for Visual Question Generation [J]. IEEE Transactions on Neural Networks and Learning Systems，2021，32(4)：1654-1667.

[17] CAO F Q，LUO S W，NUNEZ F，et al. SceneGATE：Scene-Graph Based Co-Attention Networks for Text Visual Question Answering [J]. Robotics，2023，12(4)：1-18.

[18] 刘传. 基于门控图卷积网络和协同注意力的视觉问答 [J]. 计算机与数字工程，2023，51(4)：860-865.

责任编辑　夏娟