计算机应用研究
Application Research of Computers
ISSN 1001-3695,CN 51-1196/TP

《计算机应用研究》网络首发论文

题目：基于复合图文特征的视觉问答模型研究
作者：邱南，顾玉宛，石林，李宁，庄丽华，徐守坤
DOI：10.19734/j.issn.1001-3695.2020.12.0537
收稿日期：2020-12-15
网络首发日期：2021-04-23
引用格式：邱南，顾玉宛，石林，李宁，庄丽华，徐守坤．基于复合图文特征的视觉问答模型研究[J/OL]．计算机应用研究.
https://doi.org/10.19734/j.issn.1001-3695.2020.12.0537

基于复合图文特征的视觉问答模型研究 *

邱 南，顾玉宛，石 林，李 宁，庄丽华，徐守坤†

(常州大学 计算机与人工智能学院 阿里云大数据学院，江苏 常州 213164)

摘 要：针对当前主流视觉问答(visual question answering，VQA)任务使用区域特征作为图像表示而面临的训练复杂度高、推理速度慢等问题，提出一种基于复合视觉语言的卷积网络(composite visionlinguistic convnet，CVlCN)来对视觉问答任务中的图像进行表征。提出的方法将图像特征和问题语义通过复合学习表示成复合图文特征，然后从空间和通道上计算复合图文特征的注意力分布，以选择性地保留与问题语义相关的视觉信息。在 VQA-v2 数据集上的测试结果表明，提出的方法在视觉问答任务上的准确率有明显的提升，整体准确率达到 64.4%。模型的计算复杂度较低且推理速度更快。

关键词：视觉问答；复合视觉语言特征；区域特征；多模态融合

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2020.12.0537

Research on visual question answering model based on composite graphic features

Qiu Nan, Gu Yuwan, Shi Lin, Li Ning, Zhuang Lihua, Xu Shoukun†

(*School of Computer Science & Artificial Intelligence*, *Aliyun School of Big Data*, *Changzhou University*, *Changzhou Jiangsu* 213164, *China*)

Abstract: In view of the problems of high training complexity and slow inference speed involved by the current mainstream visual question answering task which uses regional features as image representations, this paper proposed a convolutional network (composite visionlinguistic convnet, CVlCN) based on composite visual language to extract the image features in visual question answering tasks. The proposed method represents image features and problem semantics into composite picture-text features through composite learning, and then calculates the attention distribution of composite picture-text features from space and channels to selectively retain visual information related to problem semantics. The experimental results show that, on the VQA-v2 dataset, the test accuracy of the proposed method on the visual question answering task is obviously improved, and the overall accuracy is 64.4%. And the model has low computational complexity and fast inference speed.

Key words: visual question answering; composite visionlinguistic feature; regional feature; multimodal fusion

0 引言

视觉问答(visual question answering，VQA)是一项新颖的多模态学习任务，它已经引起了深度学习、计算机视觉和自然语言处理领域的广泛兴趣。视觉问答模型需结合问题和图像进行理解，并给出合理的回答。即给计算机一张图片和关于这张图片的问题，让计算机像人类一样作出回答。与图像字幕[1, 2]、图文检索[3, 4]等多模态任务相比，视觉问答任务更具有挑战性。因为视觉问答任务中的问题不是预先确定好的，而其他任务中需要回答的问题是固定的。视觉问答也涉及许多其他的计算机视觉子任务，如目标的识别、检测，计数，视觉推理等任务。

在早期基于深度学习的视觉问答模型[5]中，通常使用循环神经网络获取问题中的语义信息，采用卷积神经网络来提取视觉特征，然后通过融合函数对视觉和语义两种模态数据进行融合，最后将融合后的特征送入分类器中预测答案。许多工作展开了大量研究，如 MRN[6]继承深度残差学习的思想提出多模态残差网络来有效地学习视觉和文本信息的联合表示；MCB[7]提出的多模态紧致双线性池化方法去组合多模态特征。这些方法主要对多模态数据的全局特征进行表示学习，在提取图像特征过程中由于缺少文本语义的交互和卷积网络有限的感受野大小，提取的图像网格特征可能会夹杂着其他干扰信息而影响模型的准确回答。

许多工作中通过引进注意力机制来改善视觉问答中的视觉特征，使获取的视觉特征更符合问题语义。SANs[8]提出堆叠注意力网络将一个问题的语义表示作为查询去搜索与回答有关的图像区域；Anderson 等人[9]提出自底向上注意力(bottom-up attention)机制来解决视觉问答中的图像表示问题，这与以往使用卷积神经网络提取的网格特征不同，自底向上注意力使用预先训练好的目标检测器(Faster R-CNN[10])来检测图像中的目标区域，然后将基于区域的视觉特征用作视觉问答任务中的图像表示。MCAN[11]遵循自底向上注意力使用目标区域作为图像表示，提出深度模态共注意力网络来将问题中的关键单词和图像中关键对象联系起来，进而准确的回答问题。但是，使用区域特征往往需要进行复杂的预训练工作，这会使得整体模型的训练复杂度高。在推理过程中，由于存在选择目标区域的步骤，整体网络模型推理速度较慢。因此，许多最新的方法[11, 12]直接在已经提取好的区域特征上

收稿日期：2020-12-15；**修回日期：**2021-02-01 **基金项目：**国家自然科学基金资助项目(61906021)；常州市城市大数据分析与应用技术重点实验室资助项目(CM20193007)

作者简介：邱南(1995-)，男，江苏宿迁人，硕士研究生，主要研究方向为视觉问答、多模态深度学习；顾玉宛(1982-)，女，江苏吴江人，讲师，博士，主要研究方向为图像处理；石林(1979-)，男，江苏常州人，副教授，主要研究方向为大数据分析与应用、数字孪生技术；李宁(1974-)，男，甘肃庆阳人，副教授，主要研究方向为大数据分析与应用、图像识别与自然语言处理；庄丽华(1980-)，女，讲师，硕士，主要研究方向为大数据处理，图像识别；徐守坤(1972-)，男(通信作者)，吉林蛟河人，教授，博士，主要研究方向为深度学习、自然语言处理、数字孪生技术等(flybyron@hotmail.com).

进行训练和评估。此外，问题的语义理解也会影响视觉问答模型回答的准确性。在以往自然语言处理的方法[13]中，RNN、LSTM 等循环神经网络提取的问题语义往往不充分，对图像和问题进行推理时会损失关键的信息，影响整体视觉问答的效果。

综合以上问题，为了提升整体视觉问答模型的推理速度并保持较高的准确率，本文提出了复合视觉语言的卷积网络(composite visionlinguistic ConvNet，CVlCN)架构来对视觉问答任务中的图像进行表征，即将图像特征和问题语义通过复合学习表示成复合视觉语言特征，然后从空间和通道上计算复合图文特征的注意力分布，以选择性地保留与问题语义相关的视觉信息。通过提取不同卷积层的视觉信息，以形成多尺度的特征映射，更有效地回答问题。对于问题表示，本文继承并扩展了 Transformer[13]中的编码器(Encoder)分支，通过多层编码器的堆叠以获取丰富的问题语义，并更好地与视觉信息进行复合表示。整体视觉问答模型的建模过程变得更简单、模型更轻量，且以端到端的形式进行训练。在测试标准集(test-standard)上获得整体准确率 64.4%的较高性能，推理速度比以区域特征作为图像表示的方法快接近 16 倍。

1　相关工作

1.1　视觉问答中的视觉特征

在单模态和双模态任务中，视觉特征在提升性能上扮演着重要的角色。如典型的图像分类任务，视觉特征带来了显著的性能提升。当前主流视觉问答模型的准确率很大程度上依赖于提取的视觉特征，如使用 VGG[14]、ResNet[15]等典型分类网络提取的网格特征，自底向上注意力选取的区域特征。但许多视觉问答任务的工作重心都集中于视觉和语言的多模态融合策略上，对于图像表示则采用主流的区域特征方案。本文的工作从用于分类任务的基础卷积网络出发，构造出提取特定问题语义下视觉信息的卷积网络架构。实验表明，本文提出的 CVlCN 有着更快的推理速度和较高的准确率。

1.2　视觉和语言的多模态复合表示

将视觉和语言编码成视觉语言特征的复合学习，在许多视觉和语言任务中已经引起了广泛的兴趣。如基于文本的图像搜索任务[16]，使用复合表示来学习包括多个原语的特征编码。随着 BERT[17](bidirectional encoder representations from transformers，BERT)的预训练策略在机器翻译领域取得显著地性能提升，许多研究工作将 BERT 的预训练策略扩展到视觉和文本的联合表示中，以此解决视觉问答[18]、图像字幕[2]、图文匹配[3]等任务。以往用于图像分类等单模态任务的卷积网络仅仅学习视觉信息的组成，它们没有明确将图像表示和语言语义以一种复合的方式联系起来。与文献[16]工作类似，本文将视觉特征和语言语义编码成视觉语言特征，使用注意力机制保存复合特征相关性的视觉信息。通过在基础分类网络的特征映射之间注入语言语义，使得整体卷积网络能够动态地保存复合图文特征下的视觉信息。

1.3　注意力机制

注意力机制能够有选择地突出与任务相关的信息，抑制与任务不相关的信息，从而提高神经网络的工作效率。随着 Transformer[13]架构在机器翻译领域取得显著性成果，注意力机制也因此在自然语言处理和计算机视觉领域获得快速发展。对于视觉问答领域，使用注意力机制能够有效地建立视觉特征与问题语义之间多模态的联系，如 Co-Attention[19]，MCAN[11]等多模态学习方法。因此，基于注意力机制的多模态学习已经在视觉问答中广泛应用。本文在问题处理方面继承并扩展了 Transformer 中的编码器(encoder)分支，使得模型能够获取丰富的问题语义信息，并更好地与视觉信息进行复合表示。对于视觉特征，通过从空间和通道联合计算复合图文特征下的权重分布，进而保存不同特征映射下的有用特征，最终使得基础卷积网络可以更好地适应视觉问答任务。

2　主要模型

本文提出一种复合视觉语言的卷积网络(composite visionlinguistic ConvNet，CVlCN)架构来提取复合图文特征下的视觉信息，继承并扩展了 Transformer[13]中的编码器(encoder)分支来表示问题语义。最后通过多模态双线性融合方法融合图像特征与文本语义并送入分类器中预测答案，从而形成完整的视觉问答模型。整体网络架构如图 1 所示。对于问题表示，首先使用 GloVe 词嵌入模型将问题表示成词向量，然后输入到 LSTM 中获取上下文语义，进一步通过 4 层编码器获取问题语义的长范围依赖。对于图像表示，首先使用卷积神经网络提取图像中的视觉特征，然后在不同尺度的特征映射之后嵌入复合视觉语言的注意力模块来获取特定问题语义下的视觉信息。最后，使用多模态双线性融合方法融合图像特征和问题语义并预测出答案。

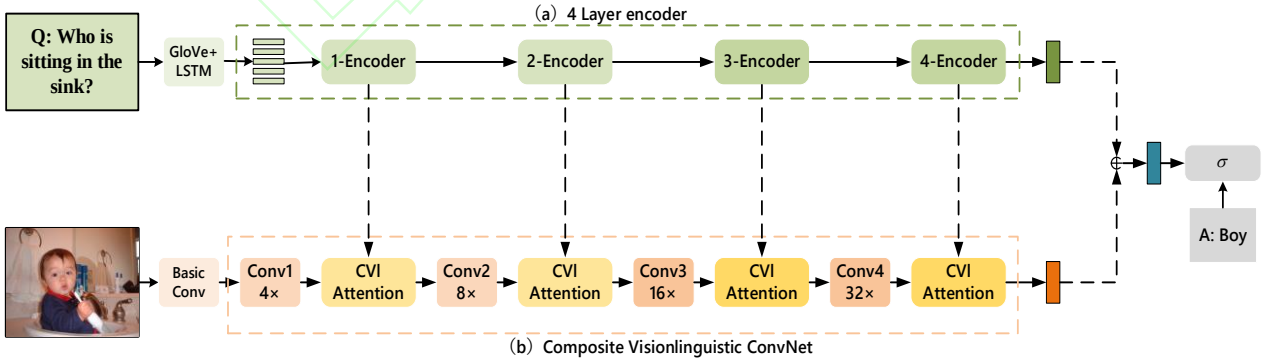


图 1　整体视觉问答网络框架

Fig. 1　Overall visual question answering network framework

2.1　图像的特征层级

为了获取不同尺度大小的视觉信息，本文采用自底向上路径[20]的方式，基于输出特征映射的尺度大小，对卷积神经网络的卷积层进行划分。在整体卷积网络中，许多连续的卷积层能够产生相同尺度的特征映射，本文将这些层划分成一组，从而形成不同尺度大小的特征层级。如图 2 所示，以 ResNet[15]为例，本文将输出尺度大小相同的卷积层组合成一个卷积模块，并把每个卷积模块中最后一个残差块的激活特征映射作为该模块的输出。因此，图 1(b)中的卷积模块可以表示为{basic conv,conv1,conv2,conv3,conv4}，每个卷积模块的输出特征大小相对于原始输出图像分别有{4,4,8,16,32}倍缩放像素。假设输入图像的原始大小为 224×224，则各个卷积模块输出的特征大小分别为{56,56,28,14,7}。通常{basic conv}模块用于提取初始图像的纹理、形状等特征，且实验中

使用在 ImageNet 数据集上的预训练网络，因此对{basic conv}模块采取参数固定的措施，即在模型训练的过程中不进行参数学习。最终，图像表示的特征层级划分为 I={conv1,conv2,conv3,conv4}，$i \in I$，i 为某个特征层级。

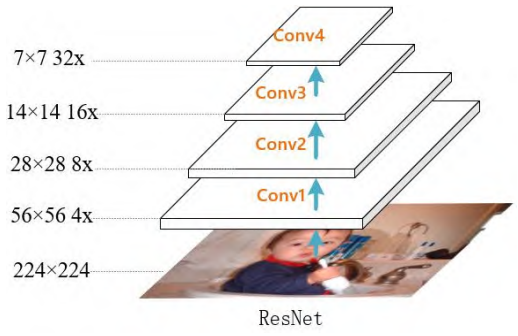


图 2　ResNet 网络的前馈特征层级

Fig. 2　Feedforward feature hierarchy of resnet network

2.2　复合视觉语言的卷积网络架构

为了在图像处理阶段获取特定问题语义下的视觉内容，本文设计了复合视觉语言的卷积网络(composite vision-linguistic ConvNet, CVlCN)架构，即在不同尺度大小的特征映射之间嵌入复合视觉语言的注意力模块，从而提取特定问题语义下的视觉信息。如图 1(b)所示。

2.2.1 复合视觉语言的注意力模块

为了动态学习特定问题语义下的视觉内容，受到文献[16]中的复合 Transformer 模块的启发，本文设计一种适应于不同图像特征映射下的复合视觉语言的注意力模块(Composite Visionlinguistic Attention Module, CVl-Attention)，如图 3 所示。具体地，对第 i 层编码器(在 2.3 小节中详细描述)的输出 $Y^i=[y_1^i;...;y_n^i] \in \mathbb{R}^{n\times d}$ 使用最大池化和多层感知机(MLP)进行特征维度的重塑，从而获得问题特征 $y^i \in \mathbb{R}^{1\times d}$。然后，将问题特征 y^i 和第 i 层特征层级输出的图像特征 $x^i \in \mathbb{R}^{h^i\times w^i\times c^i}$ 进行拼接并复合成视觉语言表示 x_{vl}^i。最后，从空间和通道两个方向分别计算复合视觉语言特征的注意力分布，以选择性地突出相关问题语义下的视觉信息 x^i。具体实现如下：

a)问题与图像的复合过程。为了使问题语义特征与图像特征进行复合表示。首先，对第 i 层编码器的输出 Y^i 进行最大池化，将原先 $n\times d$ 维的特征转换成 $1\times d$ 维特征。紧接着，经过多层感知机(MLP)对语义特征进行线性映射，以重新校准语义信息。然后将校准后的语义信息 y^i 和第 i 层级提取的图像特征 x^i 进行级联，并使用卷积核为 1×1 的卷积层进行多模态复合学习，以获得视觉语言表示 x_{vl}^i。如下式：

$$y^i = \mathrm{MLP}(\mathrm{MaxPool}(Y^i)) \tag{1}$$

$$x_{vl}^i = \mathrm{Convs}(\mathrm{Concat}(x^i, y^i)) \tag{2}$$

其中，$Y_j^i \in \{y_1^i,...,y_j^i,...,y_n^i\}$，$y_j^i$ 为第 i 层编码器输出的第 j 个单词语义。为了能够和 3 维图像进行级联，将 $y^i \in \mathbb{R}^{1\times d}$ 重塑成 $y^i \in \mathbb{R}^{1\times1\times d}$。Concat(·,·) 表示两个模态数据进行级联操作，即问题特征 y^i 从空间上使用广播机制以匹配图像特征 x^i 的形状。Convs(·) 是两层卷积核为 1×1 的卷积操作，其中，$x^i, x_{vl}^i \in \mathbb{R}^{h^i\times w^i\times c^i}$。

b)图像的注意力特征。对于复合表示的视觉语言特征 x_{vl}^i，分别从空间和通道上计算其注意力分布 A_{sp}^i，A_{ch}^i。然后将从视觉语言特征上获取的空间和通道注意力特征进行联合，形成联合的注意力分布 A_{ja}^i，以显著性地提取复合视觉语言下的视觉信息 x_{att}^i。如下式：

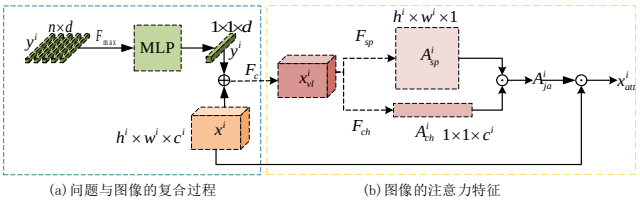
$$A_{sp}^i = \mathrm{sigmoid}(\mathrm{Convs}_{\mathrm{sp}}(\frac{1}{c^i}\sum_j^{c^i} x_{vl}^i(:,:,j))) \tag{3}$$

$$A_{ch}^i = \mathrm{sigmoid}(\mathrm{Convs}_{\mathrm{ch}}(\frac{1}{h^i\times w^i}\sum_j^{h^i}\sum_k^{w^i} x_{vl}^i(j,k,:))) \tag{4}$$

$$A_{ja}^i = A_{sp}^i \odot A_{ch}^i \tag{5}$$

$$x_{att}^i = x^i \odot A_{ja}^i \tag{6}$$

其中 $A_{sp}^i \in \mathbb{R}^{h^i\times w^i\times1}$，$A_{ch}^i \in \mathbb{R}^{1\times1\times c^i}$，$A_{ja}^i \in \mathbb{R}^{h^i\times w^i\times c^i}$，$x_{att}^i \in \mathbb{R}^{h^i\times w^i\times c^i}$；$\mathrm{Convs}_{\mathrm{sp}}(\cdot)$ 和 $\mathrm{Convs}_{\mathrm{ch}}(\cdot)$ 是两层卷积核为 1×1 的卷积操作，用于学习空间和通道上的注意力特征。A_{ja}^i 是由 A_{sp}^i 和 A_{ch}^i 推导的联合注意力矩阵。⊙ 为哈达玛积，进行二元运算。



(a) 问题与图像的复合过程　　(b) 图像的注意力特征

图 3　复合视觉语言的注意力模块

Fig. 3　Composite visionlinguistic attention module

2.2.2 复合视觉语言的卷积网络

通常，不同特征层级输出的特征映射尺度大小会随着输入图像的大小而改变。例如，输入图像像素大小为 224×224，则各个特征层级 I 的输出特征大小分别为{56,28,14,7}。特别地，对于 conv1 层 56×56 的像素大小，若使用文献[21]中的视觉自注意力模块，将三维空间的图像重塑成一维时会造成很大的维数灾难，整体模型的计算复杂度非常高。因此，在不同尺度大小的特征层之间设计保证模型的性能和减缓模型计算复杂度的模块具有一定的挑战性。

本文 CVlCN 方法在各个特征层级 I={conv1,conv2,conv3,conv4}，$i \in I$ (i 为某个特征层级)之后嵌入复合视觉语言的注意力模块，从而对不同尺度大小的特征映射编码视觉信息，包括颜色、纹理等。进一步，经过和问题语义进行复合学习，使得原先提取视觉特征的卷积网络具有一定的任务性，从而由浅到深地获取特定问题语义下的视觉特征。另外，CVlCN 方法从空间和通道两个方向对复合图文特征计算注意力分布，并进行联合以提取特定问题语义下的视觉特征。这不仅加强整体视觉问答模型对问题与视觉之间关系的推理能力，而且采用“软性”信息选择机制降低了大尺度特征映射下的计算复杂度。

2.3　问题表示

为了表示文本语义，本文首先将输入问题以单词的形式做标记，并且最多修剪 14 个单词，这种处理类似于文献[22]。然后使用 GloVe 词嵌入表示模型，将问题中每个单词进一步转换成向量的形式。最终单词序列 $Y \in \mathbb{R}^{n\times d}$ 的大小为 $n\times300$，其中 $n \in [1,14]$ 是问题中单词的数量。正式地，将转换后的词向量输入到 LSTM 中，通过循环神经网络具有的短期记忆能力，使得问题向量特征能够获得短期上下文语义。

为了使拥有短期上下文语义的问题特征获得长范围的依赖，并相应的缓解梯度爆炸或消失的问题，本文引入文献[13]中的编码器-解码器结构。与文献[13]中不同，本文只应用了编码器模块。首先，对输入序列进行“位置编码”，保持有关标记在序列中的相对或绝对位置的信息。然后，将词向量序列输送到编码器风格的 Transformer 模块中，并经过多层堆叠产生最终的语义表示 $Y \in \mathbb{R}^{n\times d}$。图 4 展示了一层编码器的网络结构，其中包含了一个多头自注意力模块和一个小型全连接网络，两者都通过直连边进行恒等映射并依序堆叠起来。图 1(a)描述了 4 层编码器依序堆叠的过程，并且每一层编码器额外向下输出，用于和视觉特征进行信息的交互。

在多头自注意力模块中，首先将单词序列 Y 线性映射到三个不同的特征空间，即查询向量 $Q \in \mathbb{R}^{n\times d}$、键向量 $K \in \mathbb{R}^{n\times d}$ 和值向量 $V \in \mathbb{R}^{n\times d}$。给出一个查询向量 $q \in \mathbb{R}^{1\times d}$，通过缩放点积函数计算查询向量 q 和键向量 K 的相关性，再应用 softmax 函数去获得注意力分布，并在值向量 V 上进行加权求和，最终形成注意力特征 $H \in \mathbb{R}^{1\times d}$。值得注意的，由于输入向量较高的

维度，通常点积模型计算的值会有很大的方差，从而导致 softmax 函数的梯度会比较小。因此，在查询向量 Q 和键向量 K 之间进行点积时除以 $\sqrt{d}$：

$$H=\mathrm{att}(q,K,V)=\mathrm{softmax}(\frac{qK^{\mathrm{T}}}{\sqrt{d}})V \tag{7}$$

为了使注意力模块关注输入信息的不同部分，多头自注意力模块利用多个查询 $Q=\{q_1,q_2,...,q_t\}$ 来平行地计算从输入信息中选取的多组数据，每组对应于一个缩放点积注意力函数。注意力特征 H 由此可得：

$$H=\{\mathrm{att}(q_1,K,V)\oplus...\oplus\mathrm{att}(q_t,K,V)\} \tag{8}$$

其中 $\oplus$ 表示向量拼接。

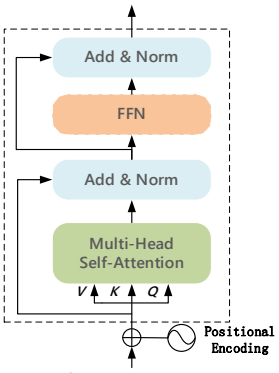


图 4　单层 Encoder 网络结构

Fig. 4　Single-layer Encoder network structure

2.4　多模态双线性融合方法和分类器

对经过多层卷积和注意力机制提取的图像特征 $X\in\mathbb{R}^{h^4\times w^4\times c^4}$ 使用全局平均池化，然后重塑形状输送到线性映射中。其中，$h^4\times w^4\times c^4$ 为特征层级 I 中第 4 层级的输出形状。而经过多层编码器获得的文本特征 $Y\in\mathbb{R}^{n\times d}$，其已经包含关于问题的丰富语义信息。因此，使用式(1)将问题特征的维度转换成 $\hat{y}\in\mathbb{R}^{1\times d}$。最后使用多模态双线性融合函数计算最终融合特征 f。公式如下：

$$\hat{x}=\mathrm{Flat}(\mathrm{GlobalPool}(X)) \tag{9}$$

$$f=\mathrm{LayerNorm}(W_x^T\hat{x}+W_y^T\hat{y}) \tag{10}$$

其中，GlobalPool(·) 是全局平均池化操作，Flat(·) 是将池化后的图像特征 $X\in\mathbb{R}^{1\times1\times c^4}$ 重塑成 $\hat{x}\in\mathbb{R}^{c^4}$，$W_x\in\mathbb{R}^{c^4\times d_f}$，$W_y\in\mathbb{R}^{h\times d_f}$，$f\in\mathbb{R}^{d_f}$，通常 c^4 和 h 维度相同。LayerNorm 是层归一化操作，用于稳定网络训练。

最后，采用文献[11，22]的分类处理，将融合特征 f 映射成分类向量 $v\in\mathbb{R}^m$，然后应用 Sigmoid 函数进行对数几率回归。其中，m 是训练集中回答最频繁的答案。最终使用二值交叉熵(binary cross-entropy)作为损失函数来训练 m 类分类器。

3　实验结果及分析

本文采用大型 VQA 基准数据集 VQA-v2 进行实验来评估本文方法的有效性。本章首先详细介绍实验过程中采用的训练策略和预防过拟合的方法。其次，从损失值分布、准确率、计算复杂度、推理速度等角度对模型进行实验对比。最后，通过消融实验和可视化操作来评估复合图文特征对视觉问答模型的影响。

3.1　VQA-v2 数据集

VQA 数据集主要由现实图像与抽象卡通图像组成。它包含人类标记的问题和答案，其中图像基本来自于 MS COCO 数据集。每个图像有 3 个问题并且每个问题有 10 个回答。数据集是划分成训练集、验证集、测试集，分别有 80k 图像和 444k 问题回答对，40k 图像和 214k 问题回答对，80k 图像和 448k 问题回答对。此外，有两个在线测试子集：测试开发集(*test-dev*)和测试标准集(*test-standard*)用于在线评估模型的性能。评估结果包括(Yes/No，计数，其他)准确率和整体的准确率类别。

3.2　参数设置

本文的实验均部署在 Linux(Ubuntu 16.04)操作系统上，使用 PyTorch 深度学习框架进行模型构建，在两块 GPU(TITAN-XP)上进行分布式并行训练模型。实验中模型的超参数设置如下。将输入图像大小重塑为 224×224×3，问题特征和融合特征的维度分别为 1024，2048。在问题编码器中，多头自注意力的潜在维度是设置成 1024，多头数量设置成 8，因此每一头的维度为 128。使用文献[22]中的策略将答案词汇的大小设置成 3129。

本文采用 Adam[21]优化器优化网络模型，Adam 优化器中的超参数设置为：$\beta_1=0.9,\beta_2=0.98$。初始学习率 η=0.0001，并采用学习率预热[23]的策略。假设在前 m 数据周期进行预热操作，然后每一个周期 $j(1\le j\le m)$ 的学习率计算为 jn/m，本文 m 设置成 4。整体的训练周期设置成 25，在周期 15 和 20 时分别进行学习率衰减，衰减率为 0.2。输入图像批次大小设置成 128。本文将原先划分的训练集和验证集用作训练，使用测试开发集(test-dev)和测试标准集(test-standard)进行模型的评估。

为了防止模型过拟合，本文采用数据增强操作。即对训练集图像采用随机水平翻转、-10 与 10 度范围内的随机旋转、-45 到 45 度范围内的随机仿射等。对这一组数据增强操作，本文采用随机应用的策略，即从操作列表中，随机选取部分操作来进行应用。这种综合的数据增强方案使得模型具有抗过拟合的能力。

3.3　实验对比与分析

实验使用 VQA-v2 数据集进行训练，对模型的评估准则采用标准的 VQA 准确率，即“总体”、“Yes\No”、“计数”、“其他”等四个准确率类别。

本文在 CVlCN 架构中使用 ResNet-101 作为主干网络。由图 5 可知，在模型的训练过程中，模型的整体损失稳定下降。在 1-4 周期内进行学习率预热，使得模型不会因初始参数的差异性带来模型准确率的不稳定。模型分别在周期为 13、19 时进行学习率的衰减，从第 19 个周期开始，损失下降平缓，模型逐渐趋于收敛。本文将 19 到 25 周期内的模型提交到 VQA-v2 在线测试集 *test-dev* 和 *test-standard* 上，获得本文方法在四个准确率类别上的测试结果。最终在周期为 23 的模型上测试出最优的准确率，在线测试结果如图 6 所示。在 *test-dev* 测试集上，四个准确率类别“总体”、“Yes/No”、“计数”“其他”分别为 64.09%、80.57%、44.91%、54.37%。在 *test-standard* 测试集上，四个类别准确率分别为 64.39%、80.95%、44.97%、54.42%。因此，本文提出的 CVlCN 架构的视觉问答模型能够获得较高的准确率。

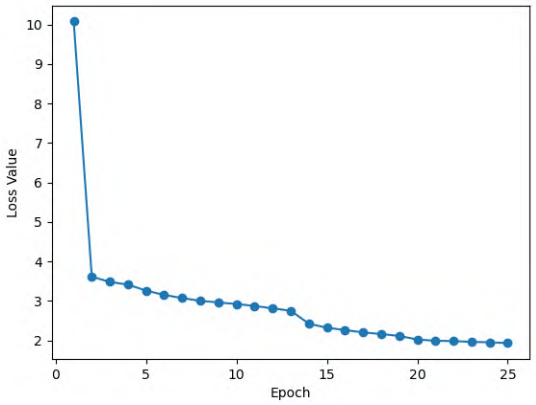


图 5　随周期变化的训练损失变化曲线图

Fig. 5　Variation curve of training loss with period

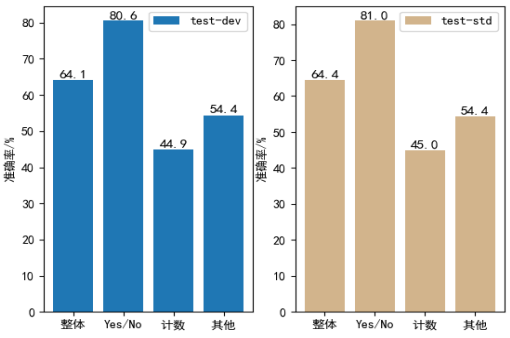


图 6　本文方法在测试集 test-dev 和 test-standard 上的结果

Fig. 6　The results of this method on the test sets test-dev and test-standard

本文选取使用区域特征作为图像表示的代表性方法作为实验对比对象，如 Bottom-Up[9]、文献[24]等方法。另外，选取多模态融合策略和其他策略的代表性方法，如 vqa machine[25]、MCB[7]、MRN[6]、SAN[8]等方法。表 1 列出了在 VQA 在线评测数据集 *test-standard* 上四个准确率类别(“整体”，“Yes/No”，“计数”，“其他”)的对比结果。值得注意的是，“▵”一列以优秀的区域特征 Bottom-Up 方法作为基线，计算其他方法“整体”准确率类别的差距。由实验结果可见，与使用区域特征方法的模型相比，本文提出的 CVlCN 架构的视觉问答模型，在准确率方面取得较好的效果，“整体”准确率为 64.4%，超过文献[24]方法，与 Bottom-Up 方法的单个模型具有可比性，并且在“计数”类别上比 Bottom-Up 方法高 1.1%。实验也和其他类型的方法进行对比，本文提出的算法模型比堆叠注意力机制的 SAN 算法和多模态残差网络 MRN 的“整体”准确率分别高出 9.1%、7%。相比于多模态双线性融合的 MCB 方法和 vqa machine 方法，本文的模型在“整体”准确率类别上分别高出 2.1%、1.4%。因此，本文提出的 CVlCN 架构在视觉问答任务中取得了较高的准确率。

表 1　不同 VQA 模型在 *test-standard* 测试集上的准确率对比

Tab. 1　Comparison of the accuracy of different VQA models on the test-standard test set

模型	Yes/No	计数	其他	整体	▵
Bottom-Up	**82.2%**	43.9%	**56.2%**	**65.7%**	-
SAN	69.8%	35.7%	47.2%	55.3%	-10.4%
MRN	75.7%	36.2%	46.3%	57.4%	-8.3%
MCB	78.8%	38.3%	53.3%	62.3%	-3.4%
vqa machine	79.8%	40.9%	53.4%	63.0%	-2.7%
文献[24]方法	80.9%	44.3%	54.0%	64.0%	-1.7%
本文 CVlCN	81.0%	**45.0%**	54.4%	64.4%	**-1.3%**

如表 1 所示，本文提出的模型在准确率上与 Bottom-Up 的单个模型具有可比性，并且在“计数”问题类型上超过了 Bottom-Up。对同样使用区域特征的文献[24]方法，本文方法的四个准确率类别都比文献[24]还要高。为了衡量本文 CVlCN 方法比使用区域特征的方法在训练复杂度和推理速度上的有效性，本文选取提出区域特征的 Bottom-Up 方法和使用区域特征获得优秀性能的 MCAN 方法进行 FLOPs 和推理速度的对比。整体实验在一块 GPU(TITAN-XP)上进行，图像大小都设置成 224×224 并使用 ResNet-101 作为主干网络，Bottom-Up 和 MCAN 都使用同样配置的 Faster R-CNN 来提取区域特征。如表 2 所示，本文 CVlCN 方法的 FLOPs 是 54.2GFLOPs，Bottom-Up 方法和 MCAN 方法分别是 189.7GFLOPs 和 255.1GFLOPs，本文整体模型的计算复杂度要比使用区域特征的方法低。另外，通过测试模型的推理速度可以发现，本文 CVlCN 的推理速度比使用区域特征的 Bottom-Up 快接近 16 倍。(在这里，推理速度是使用一张图片和一个问题进行测试。)进一步，本文通过图 7 将整体 Bottom-Up 和本文 CVlCN 视觉问答模型各阶段的推理速度进行可视化。在同样设备环境下，由于 Bottom-Up 模型要从图像中选择区域特征，因此需要消耗大量时间。并且，使用区域特征进行端到端训练往往需要进行目标位置的微调，整体模型的训练复杂度变得很高且额外的位置标注很难获得。事实上，许多最新的方法[11,12]直接在已经提取好的区域特征上进行训练和评估。而本文提出的整体视觉问答模型只需构建 CVlCN 架构去表示图像特征，不需要进行图像区域选择的过程。因此，相比于使用区域特征作为图像表示的视觉问答模型，如 Bottom-Up 和 MCAN，本文所提出的方法在准确率上相对较低，但本文使用基于复合图文特征的卷积网络来表征视觉问答任务中的图像信息，没有区域选取的步骤，这使得整体视觉问答模型以端到端的形式进行训练，模型训练复杂度较低、推理速度更快，整体建模过程更简单。

表 2　计算复杂度和推理速度的对比

Tab. 2　Comparison of computational complexity and inference speed

图像表示	输入图像大小	整体准确率 (test-std)	GFLOPs	推理速度 (ms/image)
Bottom-Up	224×224	65.7%	189.7	634
MCAN	224×224	**70.9%**	255.1	721
本文 CVlCN	224×224	64.4%	**54.2**	**38**

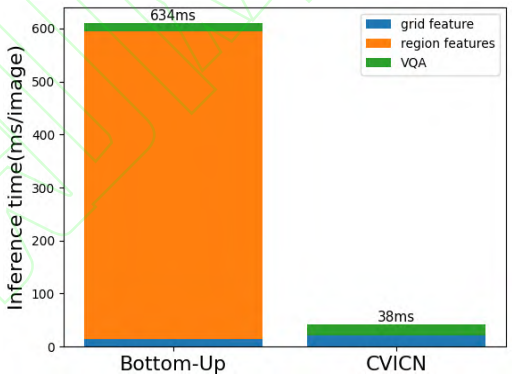


图 7　整体视觉问答模型的推理速度对比

Fig. 7　Comparison of inference speed of overall visual question answering model

为了更好地衡量 CVlCN 架构对整体视觉问答模型的影响，本文在 VQA-v2 数据集上做了消融实验，并在 test-dev 测试集上对实验结果进行比较。如表 3 所示，本文首先设置基线任务 base，即仅使用 ResNet-101 作为主干网络来提取图像特征，没有复合视觉语言的注意力模块参与。然后，按第 2.1 小节对卷积网络划分出不同特征映射大小的特征层级{conv1,conv2,conv3,conv4}，并逐层添加复合视觉语言的注意力模块进行整体视觉问答模型的训练。例如，conv1+conv2 表明在 conv1 和 conv2 特征层级之后都嵌入复合视觉语言的注意力模块。由表 3 可知，在 conv1 之后嵌入复合视觉语言的注意力模块的模型比没有嵌入的 base 任务准确率更高，四个准确率类别(“整体”，“Yes/No”，“计数”，“其他”)比 base 模型分别高出 4.8%、5.4%、4.67%、4.2%，以此验证了复合视觉语言的注意力模块对整体视觉问答模型的有效性。进一步，在不同特征层级之后嵌入复合视觉语言的注意力模块，整体视觉问答模型的准确率稳步提升。最后，在形成本文 CVlCN 架构时，准确率达到最优并趋于饱和。因此，通过逐步在特征层级之后嵌入复合视觉语言模块验证其有效性，本文的 CVlCN 架构能够提取特定问题语义下的细粒度视觉信息。

本文通过将 CVlCN 学习到的特征权重以热点图的方式附加到原始图像中，对复合视觉语言下的视觉特征进行可视化，并进一步解释了学习的视觉特征对整体视觉问答模型回答正确答案的影响。如图 8 所示，所有可视化的图片都来源

于基准数据集 VQA-v2 中 COCO 数据集图像。在第一行正确回答样本中，复合视觉语言下学习的图像特征更关注于问题中要求的区域信息，如左边的卡车卖什么？模型根据问题中关键的‘卡车’单词集中在图像中‘卡车’相关区域。因此，通过学习复合视觉语言下的图像表示可以很好的回答问题，并进一步说明了提取出问题中关键单词及图像中相关区域是重要的。在第二行回答错误样本中，模型基本关注关键的单词和相关的区域，但仍然给出了错误的答案。在定位出关键信息失效的情况下，可能需要设计出专门的模块。如最后数相框的示例中，似乎将座椅也数成相框。通过这些示例的展示，对未来进一步的改进具有指导意义。

表 3　在 *test-dev* 测试集上 CVlCN 架构影响的消融研究

Tab. 3　Ablation of influence of cvlcn architecture on test-dev test set

模型	Yes/No	计数	其他	整体
base	70.1%	38.6%	48.8%	56.4%
conv1	75.5%	43.27%	53.0%	61.2%
conv1+conv2	77.3%	43.9%	53.5%	62.2%
conv1+conv2+conv3	80.2%	44.2%	54.4%	64.0%
本文 CVlCN 方法	**80.6%**	**44.9%**	**54.4%**	**64.1%**

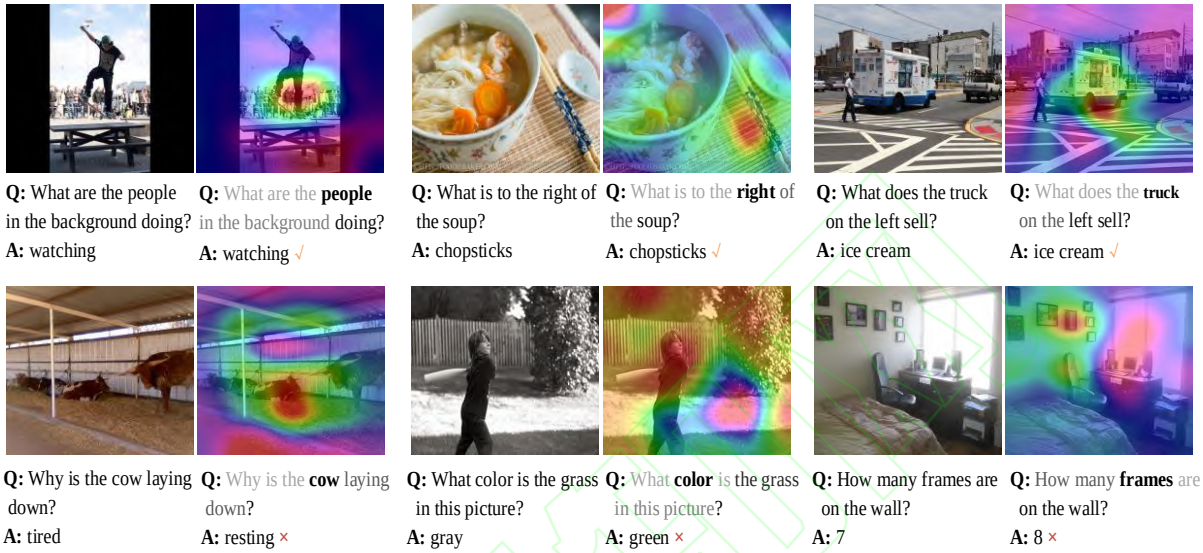


图 8　可视化复合视觉语言下学习的视觉特征

Fig. 8　Visualize the visual features of learning under compound visual language

4　结束语

本文提出一种复合视觉语言的卷积网络(composite visionlinguistic ConvNet，CVlCN)架构来提取复合视觉语言下的图像特征，继承并扩展了 Transformer 中的编码器(encoder)分支来表示问题语义，最后通过多模态双线性融合方法融合视觉信息与文本语义，并送入分类器中预测答案，从而形成完整的视觉问答模型。本文抛弃了主流方法中使用目标检测器从图像中识别出的目标区域作为图像特征，使用不同深度卷积层提取的特征映射作为图像表示。本文使用公共数据集 VQA-v2 进行模型的训练，并与使用区域特征作为图像表示的方法进行对比。实验结果表明，本文提出的方法降低了模型的计算复杂度，大幅度地提高了整体视觉问答模型的推理速度，并且取得较高的准确率，使得整体视觉问答模型的建模过程更简单、模型更轻量，且以端到端的形式进行训练。但模型对复杂场景的推理能力较差，无法给出合理的答案，有待进一步提升。因此，在接下来的视觉问答研究中，针对复杂场景中关系推理能力差等问题，将对问题语义和图像特征在融合阶段进行多模态的关系推理研究。

参考文献：

[1] Huang Lun, Wang Wenmin, Chen Jie, *et al.* Attention on attention for image captioning [C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 4634-4643.

[2] Rennie S J, Marcheret E, Mroueh Y, *et al.* Self-critical sequence training for image captioning [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7008-7024.

[3] Zheng Zhedong, Zheng Liang, Garrett M, *et al.* Dual-Path Convolutional Image-Text Embeddings with Instance Loss [J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) , 2020, 16 (2): 1-23.

[4] Li Xiujun, Yin Xi, Li Chunyuan, *et al.* Oscar: Object-semantics aligned pre-training for vision-language tasks [C]// European Conference on Computer Vision. Springer, Cham, 2020: 121-137.

[5] 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.

[6] 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.

[7] Fukui A, Park D H, Yang D, *et al.* Multimodal compact bilinear pooling for visual question answering and visual grounding [C]/ Proc of the Conference on Empirical Methods in Natural Language Processing. 2016: 457-468. http://doi. org/10. 18653/v1/D16-1044.

[8] Yang Zichao, He Xiaodong, Gao Jianfeng, *et al.* Stacked attention networks for image question answering [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 21-29.

[9] Anderson P, He Xiaodong, Buehler C, *et al.* Bottom-up and top-down attention for image captioning and visual question answering [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6077-6086.

[10] Ren Shaoqing, He Kaiming, Girshick R, *et al.* Faster r-cnn: Towards real-time object detection with region proposal networks [C]// Advances in neural information processing systems. 2015: 91-99.

[11] Yu Zhou, Yu Jun, Cui Yuhao, *et al.* Deep modular co-attention networks for visual question answering [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 6281-6290.

[12] Wang Tan, Huang Jianqiang, Zhang Hanwang, *et al.* Visual commonsense r-cnn [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10760-10770.

[13] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need [C]// Advances in neural information processing systems. 2017: 5998-6008.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]// International Conference on Learning Representations. 2015.

[15] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep residual learning for image recognition [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[16] Chen Yanbei, Gong Shaogang, Bazzani L. Image Search with Text Feedback by Visiolinguistic Attention Learning [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3001-3011.

[17] Devlin J, Chang Ming Wei, Lee K, *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding [C]/ Proc of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171-4186. https://doi. org/10. 18653/v1/N19-1423.

[18] 吝博强，田文洪. 基于层次注意力机制的高效视觉问答模型 [J/OL]. 计算机应用研究: 1-6 [2020-12-08]. https://doi. org/10. 19734/j. issn. 1001-3695. 2019. 11. 0688. (Lin Boqiang, Tian Wenhong. Efficient image question answering model based on layered attention mechanism [J/OL]. Application Research of Computers: 1-6 [2020-12-08]. https://doi. org/10. 19734/j. issn. 1001-3695. 2019. 11. 0688.)

[19] Nguyen D K, Okatani T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6087-6096.

[20] Lin Tsung Yi, Dollár P, Girshick R, *et al.* Feature pyramid networks for object detection [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[21] Bello I, Zoph B, Vaswani A, *et al.* Attention augmented convolutional networks [C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 3286-3295.

[22] Teney D, Anderson P, He X, *et al.* Tips and tricks for visual question answering: Learnings from the 2017 challenge [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4223-4232.

[23] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv: 1412. 6980, 2014.

[24] 闫茹玉，刘学亮. 结合自底向上注意力机制和记忆网络的视觉问答模型 [J]. 中国图象图形学报, 2020, 025 (005): 993-1006. (Yan Ruyu, Liu Xueliang. A visual question answering model combining bottom-up attention mechanism and memory network [J]. Journal of Image and Graphics, 2020, 025 (005): 993-1006.)

[25] Wang Peng, Wu Qi, Shen Chunhua, *et al.* The vqa-machine: Learning how to use existing vision algorithms to answer new questions [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1173-1182.