面向视觉对话的自适应视觉记忆网络

赵 磊，高联丽*，宋井宽

(电子科技大学计算机科学与工程学院 成都 611731)

【摘要】视觉对话中最具挑战的难点是视觉共指消解问题，该文针对此问题设计了一种自适应视觉记忆网络 (AVMN)。该方法直接将视觉信息存储于外部记忆库，整合了文本和视觉定位过程，进而有效缓解了在这两个过程中所产生的误差。此外在很多场景下，仅依据图片便可对提出的问题进行回答，历史信息反而会导致不必要的误差。因此，模型自适应地读取外部视觉记忆，并融合了残差视觉信息。实验证明，相比于其他方法，该模型在各项指标上均取得了更优的效果。

关 键 词 自适应； 注意力机制； 记忆网络； 视觉对话

中图分类号 TP391 **文献标志码** A doi:10.12178/1001-0548.2021057

Adaptive Visual Memory Network for Visual Dialog

ZHAO Lei, GAO Lianli*, and SONG Jingkuan

(School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731)

Abstract The key challenge in visual dialogs is the problem of visual co-reference resolution. This paper proposes an adaptive visual memory network (AVMN), which applies external memory bank to directly store grounded visual information. The textual and visual positioning processes are integrated so that the possible errors in the two processes are effectively relieved. Moreover, the answers can be produced only based on the question and image in many cases. The historical information somewhat causes unnecessary errors, so we adaptively read the external visual memory. Furthermore, a residual queried image is fused with the attended memory. The experiment indicates that our proposed method outperforms the recent approaches on the evaluation metrics.

Key words adaptive; attention mechanism; memory network; visual dialog

当前，计算机视觉[1]与自然语言处理[2]相结合的跨模态任务获得大量关注，如图像描述生成 (image captioning)[3-4]、视觉问答 (visual question answering)[5-6]等。视觉对话任务是指计算机根据图片、图片描述以及历史对话信息对人所提出的问题进行流畅自然地回答。视觉对话技术可以应用于大量的实际生活场景中，如协助视觉障碍患者完成对周围环境的感知；如升级客服系统，使之智能化地对消费者所提出的问题作答；或让机器人拥有类似于人的交流能力。

视觉对话是一项充满挑战性的任务。其中，视觉共指消解问题是关键的一个研究点，它是指如何找到问题中的代词在图片中的具体目标指代。在视觉对话任务中最常用的数据集 VisDial 中，有近 38% 的问题以及 19% 的答案包含代词，如‘he’‘his’‘it’‘there’‘they’‘that’‘this’等。文献 [7] 通过神经模块网络确定问题中的代词在历史对话中所指代的具体实体，然后从输入的图片完成视觉定位。文献 [8] 提出了适用于视觉对话的双重注意力网络，它通过多头注意力机制学习问题与历史对话信息之间的潜在关联，然后利用自底向上的注意力机制完成视觉上的目标检测。文献 [9] 提出了递归的视觉注意力来对历史对话进行遍历，直至找到高置信度的视觉指代。总结先前的工作，它们都是通过文本定位和视觉定位两个步骤来解决视觉共指消解问题。然而，每一步过程都有可能产生误差，从而导致最终回答的问题精度不足。误差产生的主要原因是问题中的代词在对话历史中所指代的目标依然难以确定。如在历史对话中其指代的目标在比较靠前的轮次，或者存在语义相近，容易混淆的文本目标，这

收稿日期：2021－03－01；修回日期：2021－07－04

作者简介：赵磊 (1991－)，男，博士生，主要从事计算机视觉、自然语言处理方面的研究.

*通信作者：高联丽，E-mail：juana.alian@gmail.com

都容易导致文本定位的误差。而由历史对话中所找到的文本指代完成视觉定位同样容易产生误差。其原因为图像中背景信息比较复杂，如背景中有同目标类似的物体，亦或其背景的颜色特征、纹理特征与目标相近等，容易误检而造成误差。同时先前工作都忽视了在很多情况下，问题的回答不需要利用历史对话，简单的视觉信息可以直接完成作答。

本文将对话过程中已完成定位的视觉信息存储在外部的记忆库中，从而将上述的两个步骤进行整合。在每回答一个问题时，不需要从历史对话中寻找问题中代词具体的指代，而是直接从视觉记忆库中进行读取。通过外部视觉记忆库对文本定位和视觉定位的整合，将先前的两步定位可能产生的误差缩减为对单步视觉记忆读取的误差，理论上单步的误差要小于两步的误差。为了更好地处理视觉信息可直接作答的情形，在读取视觉记忆库的时候，采用了自适应的方式，即动态地学习一个置信度。进一步地，引入视觉残差连接来缓解此问题，从而更好地应对不同的情况。

1 自适应视觉记忆网络

1.1 数据处理

视觉对话任务中的输入主要包括文本类数据和视觉类数据两种模态数据。其中，文本类数据包括当前轮次所提出的问题$\boldsymbol{q}_t$，历史对话$\boldsymbol{H}_t=(\boldsymbol{C},(\boldsymbol{q}_1,\boldsymbol{a}_1),(\boldsymbol{q}_2,\boldsymbol{a}_2),\cdots,(\boldsymbol{q}_{t-1},\boldsymbol{a}_{t-1}))$，以及候选答案$\boldsymbol{A}$。视觉类数据包括图片$\boldsymbol{I}$，以及视觉记忆库$\boldsymbol{M}_t=(\boldsymbol{m}_0,\boldsymbol{m}_1,\cdots,\boldsymbol{m}_{t-1})$。

本文对文本类数据均利用词嵌入方法将每一个词映射为词向量。随后，映射后的当前问题$\boldsymbol{q}_t$利用自注意力机制得到带权重的词向量$\boldsymbol{q}^a$，用以表示在问题中重要的词语。同时，将映射之后的历史对话和候选答案都输入 LSTM 中，取最后一个隐藏层的状态为其对应特征，分别为$\boldsymbol{H}_t^l=(\boldsymbol{h}_0,\boldsymbol{h}_1,\cdots,\boldsymbol{h}_{t-1})$和$\boldsymbol{A}^\lambda$。

视觉类数据中的图片$\boldsymbol{I}$利用在 Visual Genome 上预训练好的 Faster R-CNN 提取目标级特征$\boldsymbol{V}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n)$。本文将提取的目标数量固定为 36 个。初始的视觉记忆库$\boldsymbol{m}_0$是由图片描述$\boldsymbol{C}$对图片$\boldsymbol{I}$进行软注意力计算所得。

1.2 网络模型

本文所采用的网络框架为编码器–解码器模式。整体框架图如图 1 所示，其中自适应视觉记忆模块是整个网络的重点。它的输入为当前问题的带权重特征$\boldsymbol{q}^a$对图片$\boldsymbol{I}$的特征$\boldsymbol{V}$进行注意力计算所得到的视觉特征$\boldsymbol{V}^q$，具体如下：

$$\boldsymbol{\alpha}^{q\_v}=\mathrm{softmax}\left(f^q(\boldsymbol{q}^a)\circ f^v(\boldsymbol{V})\right) \tag{1}$$

$$\boldsymbol{V}^q=\boldsymbol{\alpha}^{q\_v}\cdot\boldsymbol{V} \tag{2}$$

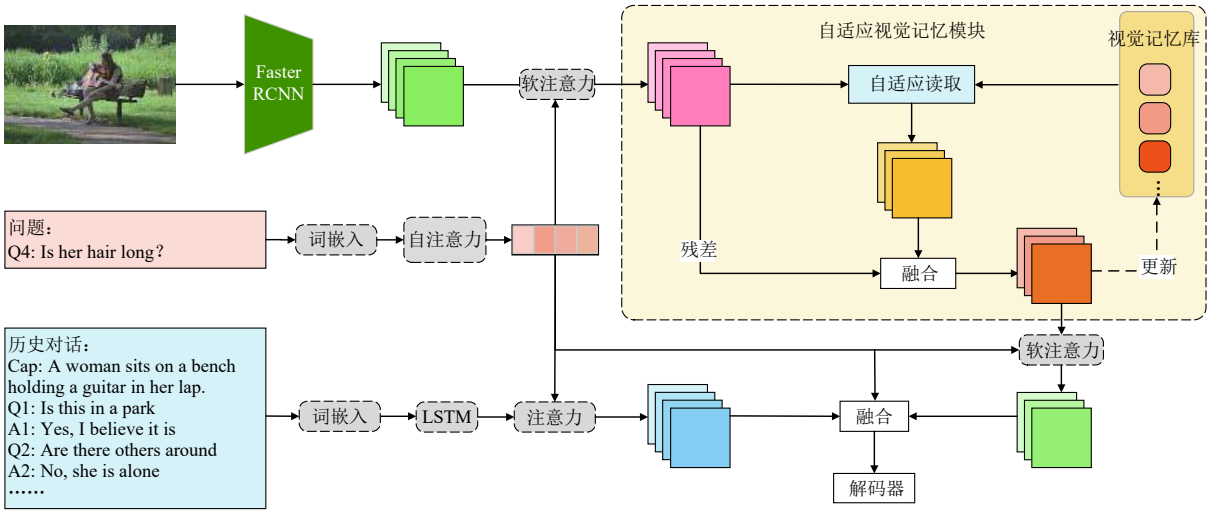

图 1 本文所设计的自适应视觉记忆网络 AVMN 的框架图

式中，f^q和f^v分别表示非线性变换函数；“∘”表示哈达玛积；“·”表示矩阵相乘。

之后，$\boldsymbol{V}^q$输入到自适应视觉记忆模块中读取外部的视觉记忆库以完成初步的目标定位。其详细流程如算法 1 所示。

算法 1 自适应视觉记忆模块数据读写流程：

Function AdaptQuery($\boldsymbol{V}^q$,$\boldsymbol{M}_t$)

$\lambda\leftarrow\sigma(\mathrm{Linear}(\boldsymbol{V}^q))$

Weighted Mem: $\boldsymbol{M}_t^w \leftarrow (1-\lambda)\boldsymbol{M}_t$

$\boldsymbol{M}_t^{v\_a} \leftarrow \text{Soft}-\text{ATT}(\boldsymbol{V}^q, \boldsymbol{M}_t^w)$

$\boldsymbol{m}_t \leftarrow \text{Fusion}(\boldsymbol{V}^q, \boldsymbol{M}_t^{v\_a})$

$\boldsymbol{M}_{t+1} \leftarrow \boldsymbol{M}_t + \boldsymbol{m}_t$　　更新

return $\boldsymbol{m}_t, \boldsymbol{M}_{t+1}$

end function

考虑到在很多情形下问题的回答不需要用到视觉相关的历史信息，直接利用问题便可从图片中定位到目标特征。因此，本文将$\boldsymbol{V}^q$经过线性变换处理后得到的特征输入到 sigmoid 函数中学习一个参数λ，并用此参数得到带有权重的外部视觉记忆信息$\boldsymbol{M}_t^w$。然后利用软注意力机制读取到视觉记忆$\boldsymbol{M}_t^{v_a}$，具体如下：

$$\boldsymbol{\alpha}^{v\_m} = \text{softmax}(f^{v\_a}(\boldsymbol{V}^q)) \tag{3}$$

$$\boldsymbol{M}_t^{v\_a} = \boldsymbol{\alpha}^{v\_m} \cdot \boldsymbol{M}_t^w \tag{4}$$

式中，f^{v_a}表示非线性变换。进一步地，将视觉特征$\boldsymbol{V}^q$与取得的视觉记忆$\boldsymbol{M}_t^{v_a}$做融合，也可以视为对视觉特征$\boldsymbol{V}^q$做残差连接。具体融合方式为：

$$\boldsymbol{\alpha}^{v\_f} = \text{FC}(\text{Norm}(\text{Gate}(\boldsymbol{V}^q))) \tag{5}$$

$$\boldsymbol{\alpha}^{m\_f} = \text{FC}\left(\text{Norm}\left(\text{Gate}\left(\boldsymbol{M}_t^{v\_a}\right)\right)\right) \tag{6}$$

$$\boldsymbol{m}_t = \text{FC}\left(\left[\boldsymbol{\alpha}^{v\_f} \cdot \boldsymbol{V}^q, \boldsymbol{\alpha}^{m\_f} \cdot \boldsymbol{M}_t^{v\_a}\right]\right) \tag{7}$$

式中，FC 均表示全连接线性变换；Norm 和 Gate 分别表示 L2 正则化运算和门函数；[,] 表示向量之间的级联操作。此阶段所读取到的最终特征$\boldsymbol{m}_t$还要被更新到外部记忆库中。

为进一步地提炼所读取出来的视觉特征，使其更专注于所提出的问题，利用经过自注意力计算的问题$\boldsymbol{q}^a$对$\boldsymbol{m}_t$做如下计算：

$$\boldsymbol{V}^f = \sigma(\text{FC}(\boldsymbol{q}^a)) \cdot \boldsymbol{m}_t \tag{8}$$

式中，σ表示 sigmoid 函数。同时将历史对话作为答案生成的补充信息。同样利用注意力机制使历史对话中的有效信息集中到相关问题上，具体为：

$$\boldsymbol{\alpha}^{q\_h} = \text{softmax}\left(\text{FC}\left(f^h\left(\boldsymbol{H}_t^l\right) \circ f^{q\_h}(\boldsymbol{q}^a)\right)\right) \tag{9}$$

$$\boldsymbol{H}^q = \sum_{r=0}^{t-1} \boldsymbol{\alpha}^{q\_h} \boldsymbol{h}_r \tag{10}$$

最终将当前问题特征、外部记忆库所读出来的视觉特征及历史对话特征进行融合，具体方式为：

$$\boldsymbol{F} = f^{q\_h\_v}\left(\left[\boldsymbol{q}^a, \boldsymbol{H}^q, \boldsymbol{V}^f\right]\right) \tag{11}$$

式中，$f^{q_h_v}$为线性变换；[,] 表示级联操作；$\boldsymbol{F}$则是融合之后的特征，也是整个框架中编码器的输出。它之后被输入到解码器中，用以给候选的 100 个答案进行排序。

本文中解码器采用多任务学习机制，即判别式和生成式的融合。其中，判别式解码器是通过计算每个候选答案的特征与编码器输出的融合特征之间的点乘相似度，用 softmax 函数获得候选答案的后验概率。并通过对交叉熵损失函数的最小化来训练模型。生成式编码器是用 LSTM 语言模型来直接生成答案，并通过对数似然损失函数完成训练。本文将两者损失函数相加，完成对最终模型的训练。

2　实验结果及分析

2.1　数据集

本文所有实验都在数据集 VisDial 1.0[10] 上进行。该数据集采集于 Amazon Mechanical Turk 数据采集平台。其中，训练集的图片均来自于 COCO 2014 数据集，共包含大约 12.3 万张图片。验证集和测试集的图片则采集于 Flickr 数据集，分别包含 2 000 和 8 000 张图片。训练集和验证集中，每张图片对应 10 轮问答，测试机则仅有一轮问答。每个问题都包含有 100 个候选答案。

2.2　评价指标

实验中所采用的评价指标共 4 类，包括：平均排序 (mean)、平均排序倒数 (mean reciprocal rank, MRR)、召回率 (recall@)、归一化折现累计收益 (normalized discounted cumulative gain, NDCG)。

平均排序用于表示人工标注的正确答案在所有候选答案排序中的平均排名。平均排序倒数是指将所有正确答案的排名取倒数，并做平均化处理。召回率表示在所有候选答案的排序中人工标注的正确答案位于前 k 所占的比例，本文将 k 设置为 1、5 和 10。归一化折现累计收益则是考虑到候选答案中可能存在多个正确答案的情形，它旨在处罚那些正确但又排名较低的答案。

2.3　实验设置

本文所设计的模型主要基于 PyTorch1.0 实现。模型在数据集上共训练 15 个周期，批大小设为 32，初始学习率设为 0.001，经历一个热身周期，并在第 10 个周期后降至 0.000 1。训练优化器选用 Adam。

2.4 定量及定性实验

为验证本文所设计模型的有效性，将此模型和近年来效果最优的算法进行对比。对比方法包括：

1) VGNN[11]：利用图神经网络将视觉对话模拟为基于局部观测节点的图模型推导。每轮对话被视为图节点，对应的回答表示为图中缺失的一个值。

2) CorefNMN[7]：利用模块神经网络完成字词级别的目标定位。

3) DVAN[12]：以双重视觉注意力网络来解决视觉对话中的跨模态语义相关性。充分地挖掘了局部视觉信息和全局视觉信息，并利用 3 个阶段的注意力获取来生成最终的答案。

4) FGA[13]：针对视觉对话的因子图注意力方法，可以有效地整合多种不同模态的数据。

5) RVA[9]：用于遍历历史对话信息的递归注意力机制。

6) DualVD[14]：自适应的双重编码模型。学习更丰富的、全面的视觉特征用以回答多样的问题。

表 1 为本文所提出的算法 AVMN 与上述方法在 VisDial 1.0 测试集上的实验结果在平均排序倒数 (MRR)、召回率 (recall@k)、平均排序 (mean)、归一化折现累计收益 (NDCG) 各项指标上的对比。其中，AVMN*表示解码器为判别式的，AVMN 表示解码器采用多任务学习方式，在训练的时候加入了生成式损失函数。

从表 1 可看出，本文所提出的 AVMN 即使在没有加入生成式损失函数的情况下已经在各项指标上全面超过了各对比方法。在采用多任务学习方式后，实验结果又获得了可观的提升，进一步和对比方法拉开了一定差距。具体地，完整的 AVMN 在平均倒数排序 MRR 上的结果比所有对比方法中最优的方法 DualVD 提升了 0.6%，在召回率 R@1 上比效果最佳的 FGA 提升了 0.59%，在同样代表精确性的平均排序上取得了 4.03 的结果。在保证答案的精确度的同时，它在归一化折现累计收益 NDCG 上也取得了 56.92 的结果，相比相关的最优方法取得了 0.7% 的提升。FGA 在 R@5 上的结果比 AVMN 略高，但是它利用因子图将多种类型数据进行交互，所取得的提升建立在代价较大的计算上。以上实验结果证明了 AVMN 的先进性。

表 1 本文算法与其他算法的结果对比

算法	MRR	R@1	R@5	R@10	mean	NDCG
VGNN	61.37	47.33	77.98	87.83	4.57	52.82
CorefNMN	61.50	47.55	78.10	88.80	4.40	54.72
DVAN	62.58	48.90	79.35	89.03	4.36	54.70
FGA	63.70	49.58	**80.98**	88.55	4.51	52.10
RVA	63.03	49.03	80.40	89.83	4.18	55.59
DualVD	63.23	49.25	80.23	89.70	4.11	56.32
AVMN*	63.79	50.1	80.95	89.97	4.08	56.46
AVMN	**63.83**	**50.17**	80.90	**90.17**	**4.03**	**56.92**

AVMN 在 VisDial 1.0 上的定性实验结果如图 2 所示。其中 Baseline 代表没有加入自适应视觉记忆模块的基准模型，GT 代表人工标注的正确答案，Predict 代表 AVMN 预测的答案。从图中前两个示例可以看出，AVMN 所生成的答案相较基准模型更为准确，和 GT 一致。同时，它也可以对不存在代词的问题进行准确的回答，如后两个示例所示。

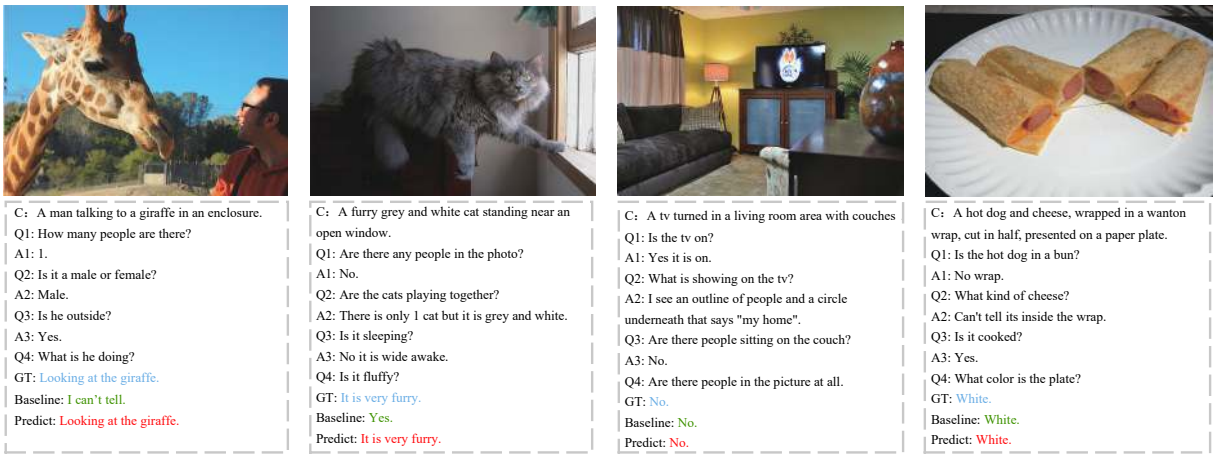

图 2 本文所设计的自适应视觉记忆网络 AVMN 在 VisDial 1.0 数据集上的定性结果

2.5 消融实验

在此实验部分，设计针对本文所提出的算法 AVMN 中主要组成部分在 VISDial 1.0 验证集上的消融实验。实验中主要设置了两个算法的变体：1) 没有使用记忆库的原始模型；2) 仅使用了记忆库，但没有采用自适应读取的模型。表 2 为消融实

验的结果展示。值得注意的是，此实验部分中所有模型的解码器是判别式的。

表 2　针对算法主要模块的消融实验结果

记忆	自适应	MRR	R@1	R@5	R@10	mean	NDCG
		64.40	50.60	81.58	90.27	4.03	56.41
√		64.82	51.21	81.63	90.15	4.01	57.43
√	√	**64.89**	**51.27**	**81.92**	**90.52**	**3.99**	**57.59**

表 2 中第一行是原始模型的实验结果。记忆代表 AVMN 中使用的记忆库。从数据可看出，原始模型相比完整模型的实验结果表现较差。第二行为加入记忆库后模型的实验结果。它在平均排序倒数 MRR 和归一化折现累计收益 NDCG 上提升明显，尤其在 NDCG 上，提升幅度超过 1%。其原因是视觉记忆相比之前的方法缩减了定位步骤，其中间误差减少，准确性以及相关性随之提升。第三行是加入对记忆库自适应读取后完整模型的实验结果。相较于进加入记忆库后的模型，它主要在召回率 R@5 和 R@10 上取得了较大的提升。其原因是自适应读取的加入使得本不需要历史信息的问题得到了更精确的回答。

3　结 束 语

本文设计了一种为解决视觉对话中视觉共指消解的自适应视觉记忆网络 AVMN。先前的方法为缓解指代模糊的问题，基本都是分两步，先从历史对话中找到代词的具体指代，然后再从图片中定位到视觉目标。视觉记忆网络直接将对话历史中已完成定位的视觉信息存储到外部的记忆模块中。这种方式将两步缩减为一步，减少在文本定位和视觉定位两步过程中所产生的误差。同时在面临仅需要图片便能回答的问题，加入了对外部视觉记忆的自适应读取，以及初始图片的残差连接。在视觉对话领域最流行的数据集 VisDial 上的实验结果证明了本文所设计模型相较于其他优秀算法的先进性。消融实验验证了视觉记忆网络内对最终结果的影响，更进一步地证明了它的有效性。

参　考　文　献

[1] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV: IEEE, 2016: 770-778.

[2] 林奕欧, 雷航, 李晓瑜, 等. 自然语言处理中的深度学习: 方法及应用[J]. 电子科技大学学报, 2017, 46(6): 913-919. LIN Y O, LEI H, LI X Y, et al. Deep learning in NLP: Methods and application[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(6): 913-919.

[3] TAKMAZ E, PEZZELLE S, BEINBORN L, et al. Generating image descriptions via sequential cross-modal alignment guided by human gaze[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. [S.l.]: ACL, 2020: 4664-4677.

[4] ZHOU Y E, WANG M, LIU D Q, et al. More grounded image captioning by distilling image-text matching model[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA: IEEE, 2020: 4776-4785.

[5] LI X P, SONG J K, GAO L L, et al. Beyond RNNs: Positional self-attention with co-Attention for video question answering[C]//The 31st Innovative Applications of Artificial Intelligence Conference. Honolulu, Hawaii: AAAI, 2019: 8658-8665.

[6] LE T M, LE V, VENKATESH S, et al. Hierarchical Conditional relation networks for video question answering[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA: IEEE, 2020: 9969-9978.

[7] KOTTUR S, MOURA J, PARIKH D, et al. Visual coreference resolution in visual dialog using neural module networks[C]//The 15th European Conference on Computer Vision. Munich: Springer, 2018: 160-178.

[8] KANG G, LIM J, ZHANG B. Dual attention networks for visual reference resolution in visual dialog[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: ACL, 2019: 2024-2033.

[9] NIU Y L, ZHANG H W, ZHANG M L, et al. Recursive visual attention in visual dialog[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA: IEEE, 2019: 6679-6688.

[10] DAS A, KOTTUR S, GUPTA K, et al. Visual dialog[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI: IEEE, 2017: 1080-1089.

[11] ZHENG Z L, WANG W G, QI S Y, et al. Reasoning visual dialogs with structural and partial observations[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA: IEEE, 2019: 6669-6678.

[12] GUO D, WANG H, WANG M. Dual visual attention network for visual dialog[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2019: 4989-4995.

[13] SCHWARTZ I, YU S, HAZAN T, et al. Factor graph attention[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA: IEEE, 2019: 2039-2048.

[14] JIANG X Z, YU J, QIN Z C, et al. DualVD: An adaptive dual encoding model for deep visual understanding in visual dialogue[C]//The 32nd Innovative Applications of Artificial Intelligence Conference. New York, NY: AAAI, 2020: 11125-11132.

编　辑　蒋　晓