计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(5):114−119 [doi: 10.15888/j.cnki.csa.007925]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 基于对称注意力机制的视觉问答系统①

路　静, 吴春雷, 王雷全

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 吴春雷, E-mail: wuchunlei@upc.edu.cn

**摘　要**: 近年来, 基于图像视觉特征与问题文本特征融合的视觉问答 (VQA) 引起了研究者们的广泛关注. 现有的大部分模型都是通过聚集图像区域和疑问词对的相似性, 采用注意力机制和密集迭代操作进行细粒度交互和匹配, 忽略了图像区域和问题词的自相关信息. 本文提出了一种基于对称注意力机制的模型架构, 能够有效利用图片和问题之间具有的语义关联, 进而减少整体语义理解上的偏差, 以提高答案预测的准确性. 本文在 VQA2.0 数据集上进行了实验, 实验结果表明基于对称注意力机制的模型与基线模型相比具有明显的优越性.

**关键词**: 视觉问答; 注意力机制; 对称注意力; 卷积神经网络; 特征提取

引用格式: 路静,吴春雷,王雷全.基于对称注意力机制的视觉问答系统.计算机系统应用,2021,30(5):114–119. http://www.c-s-a.org.cn/1003-3254/7925.html

## Visual Question Answering with Symmetrical Attention Mechanism

LU Jing, WU Chun-Lei, WANG Lei-Quan

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

**Abstract**: In recent years, Visual Question Answering (VQA) based on the fusion of image visual features and question text features has attracted wide attention from researchers. Most of the existing models enable fine-grained interaction and matching by the attention mechanism and intensive iterative operations according to the similarity of image regions and question word pairs, thereby ignoring the autocorrelation information of image regions and question words. This paper introduces a model based on a symmetrical attention mechanism. It can effectively reduce the overall semantic deviation by analyzing the semantic association between images and questions, improving the accuracy of answer prediction. Experiments are conducted on the VQA2.0 data set, and results prove that the proposed model based on the symmetric attention mechanism has evident advantages over the baseline model.

**Key words**: Visual Question Answering (VQA); attention mechanism; symmetrical attention; Convolutional Neural Network (CNN); feature extraction

## 1　引言

近年来, 基于视觉和语言的跨模态任务, 如视频场景识别[1]、图文匹配[2]、视觉问答[3] 等, 在学术界和工业界引起了越来越多研究者的兴趣. 其中, 视觉问答 (Visual Question and Answering, VQA) 可以用来测试智能机器对多模态信息的理解和推理能力, 故被认为是一种评估当前机器学习模型实现程度的“视觉图灵测试”. 因此, VQA 越来越受到重视, 它的具体任务是给定一张图片和一个问题, 通过两者的合理融合生成相应的答案. VQA 研究的重点在于如何更加全面的理解视觉内

① 基金项目: 山东省重点研发计划 (2019GGX101015); 中央高校自主创新科研计划 (20CX05018A, 18CX02136A)
Foundation item: Key Research and Development Program of Shandong Province (2019GGX101015); Innovative Research Program of the Central Universities of China (20CX05018A, 18CX02136A)
收稿时间: 2020-09-15; 修改时间: 2020-10-13, 2020-10-28; 采用时间: 2020-11-04; csa 在线出版时间: 2021-04-28

容和自然语言, 如何更精准地提取和表示模态特征, 以及如何更有效地融合跨模态信息. 为了挖掘图像突出区域与问题文本中重要词之间的对应信息, 在 VQA 任务中引入了注意机制. 目前主流的算法是将问题信息与图片信息经过注意力机制生成含有双边信息的特征, 再将其放到答案预测器中生成结果. 但是这种算法只考虑了问题和图像的双边信息, 却忽略了图像信息和问题信息自身的关联性. 因此, 本文提出一种新的模型, 该模型通过利用图片和问题的自关联性和共同注意力信息, 进一步提升答案和图片的契合度. 本文提出的模型在回答问题的准确率上与基线模型相比取得了一定的提升, 这进一步说明了该模型的有效性.

本文中, 创新点可以总结归纳为如下 3 点:

(1) 本文在单模态特征中增强了特征区域间的关联性, 使图像中区域框之间及问题中单词之间的关系更紧密.

(2) 本文提出对称注意力机制的图像问答模型, 该模型可以将图像与问题文本之间的双边信息以及图像区域与问题词的自相关统一在一起, 实现了较全面的语义理解与融合.

(3) 在 VQA2.0 上通过大量的实验对新模型进行了验证. 新模型准确率比 DCA[4] 的模型提高了 1.22%, 表明了该方法的有效性.

## 2 相关工作

### 2.1 视觉问答

像图文匹配一样, 视觉问答在人工智能领域作为一种综合计算机视觉和自然语言理解的任务被大家重视. 与其他视觉任务 (行为识别[5], 目标检测[6], 图像描述[7]) 不同, 除了视觉语言基础信息外, 许多视觉问答示例还需要问题或图像中未包含的其他信息, 例如关于世界的背景常识. 问题的答案可以分为以下几种: 是/否、多选择、计数和开放式的单词/短语 (关于什么、在哪里、谁、…). VQA 在大多数研究中被划分成一个分类问题, 图像和问题作为输入, 答案作为输出类别. 目前, 视觉问答的解决方案都是使用卷积神经网络 (Convolutional Neural Networks, CNN) 对图像进行建模, 使用循环神经网络 (Recurrent Neural Networks, RNN) 或长短期记忆网络 (Long Short Term Memory networks, LSTM) 对于问题特征进行建模.

### 2.2 注意力机制

注意力机制建立在人脑视觉注意机制的基础上, 在用于视觉问答之前, 已经被用于机器翻译[8], 图像描述[9] 等任务. 在视觉问答系统中, 注意力机制与神经网络[10] 结合被用来选择与问题信息最相关的图像区域. Yang 等[11] 构建了一个叠加注意网络, 以连续的方式生成图像上的多个注意图, 目的是进行多个推理步骤. 在 EGCS[12] 中, 作者使用 multi-hop 图像关注机制,目的是捕获问题中的细粒度信息. Shih[13] 应用现有的区域提议算法来生成对象区域, 并选择最相关的区域来预测答案. Xiong 等[14] 提出了一种基于注意力的门控循环单位 (Gated Recurrent Unit, GRU), 以促进答案的预测. 除了视觉注意力, 目前 Nguyen 等[4] 已经提出了一种具有问题注意力的共同注意力机制. 同样, 本文将共同注意力机制应用于图像区域和问题关键词, 但是与文献 [4] 不同的是, 本文提出的模型考虑了问题信息和图像信息的自身关联性, 独立对待句子中的每个单词和图像中每个区域框.

## 3 视觉问答方法模块介绍

### 3.1 基于 LSTM 和 RCNN 的特征构造

由于问题词具有顺序性, 故使用双向 LSTM 对其编码. 具体来说, 一个包含 $N$ 个单词的问题首先被转换成一个 *GloVe* 向量序列$\left\{el_1^{\mathrm{T}},\cdots,el_n^{\mathrm{T}}\right\}$, 然后将其输入到具有残差连接的单层双向 LSTM (*Bi-LSTM*) 中, 过程可由如下公式表示:

$$\overrightarrow{t_n} = BiLSTM\left(\overrightarrow{t_{n-1}}, el_n^{\mathrm{T}}\right) \tag{1}$$

$$\overleftarrow{t_n} = BiLSTM\left(\overleftarrow{t_{n+1}}, el_n^{\mathrm{T}}\right) \tag{2}$$

创建一个矩阵$T=[t_1,\cdots,t_N]\in\mathbb{R}^{d\times N}$, 其中$t_N=\left[\overrightarrow{t_n}^{\mathrm{T}},\overleftarrow{t_n}^{\mathrm{T}}\right]^{\mathrm{T}}(n=1,\cdots,N)$. 同时, 为了获取输入图像的表示, 将 Bi-LSTM 中最后隐藏状态$h_T=\left[\overrightarrow{t_N}^{\mathrm{T}},\overleftarrow{t_1}^{\mathrm{T}}\right]^{\mathrm{T}}$保存. 这里的 Bi-LSTM 网络参数采用随机初始化.

同理, 遵循类似的过程对答案进行编码. 将包含 $M$ 个单词的答案转换为$\left\{el_1^{\mathrm{AN}},\cdots,el_{\mathrm{m}}^{\mathrm{AN}}\right\}$, 然后输入到 Bi-LSTM, 产生的隐藏层状态$\overrightarrow{an_M}$和$\overleftarrow{an_M}$. 本文将答案表示为$h_{AN}=\left[\overrightarrow{an_M}^{\mathrm{T}},\overleftarrow{an_1}^{\mathrm{T}}\right]^{\mathrm{T}}$.

对于图像, 将其大小调整为 448×448, 再输入到预训练好的 ResNet-152 网络中提取图像特征. 同时, 将 ResNet-152 的 res5c 层的输出作为对应于 14×14 空间分布区域的图像特征, 用$I=[i_1,\cdots,i_k]\in\mathbb{R}^{d\times K}$表示, 其

中 $K$=14×14 是区域总数, $i_k$表示第 $k$ 个特征向量, ResNet-152 的维度是 2048.

### 3.2 对称注意力模型

图 1 是本文所提出的对称注意力模型. 在 3.1 节中已经得到图像特征 $I$ 和问题特征 $T$: 现将他们输入到对称注意力模型中, 经过模型的训练最终生成相应的模态特征.

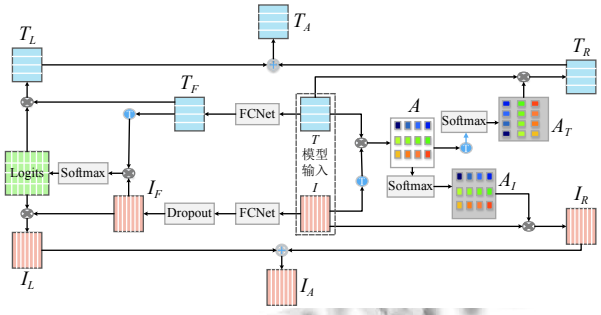


图 1 对称注意力模型架构

模型的右部分是经典的共同注意力架构[4], 首先对于给定的图像特征 $I$ 和问题特征 $T$, 通过叉乘运算得到注意力矩阵 $A$, 再将 $A$ 通过双层 Softmax 生成关于问题的注意力矩阵$A_T$和关于图像的注意力矩阵$A_I$,最后再分别与图像特征和问题特征相乘得到含有双边信息的图像特征$I_R$和问题特征$T_R$. 这个过程可用下面的 5 个公式表示:

$$A = T \times W_R \times I^{\mathrm{T}} \tag{3}$$

$$A_I = softmax(A) \tag{4}$$

$$A_T = softmax\left(A^{\mathrm{T}}\right) \tag{5}$$

$$T_R = T \times A_T \tag{6}$$

$$I_R = I \times A_I \tag{7}$$

其中, 式 (3) 的$W_R \in \mathbb{R}^{N\times K}$表示权重矩阵, 注意力矩阵的维度均为$d \times d$.

模型的左部分增强了单模态特征中特征区域间的关联性. 将问题特征输入到一个全连接层, 得到单词对相关联的问题特征$T_F$, 将图像特征输入到全连接网络以及 dropout 得到图像区域框相关联的图像特征$I_F$. 这个过程可用下面两个公式表示:

$$T_F = FCNet(T) \tag{8}$$

$$I_F = Dropout(FCNet(I)) \tag{9}$$

其中, $FCNet()$是全连接网络, 用 dropout 可以让网络去学习鲁棒性更强的特征 , 这些特征在其它的神经元的随机子集中也存在. 经过以上训练增强了图像信息和问题信息自身的关联性.

再将图像特征$I_F$, 问题特征$T_F$通过叉乘和 Softmax 运算得到权重分布矩阵$logits$, 该矩阵的维度与注意力矩阵 $A$ 一致, 包含了两种特征的融合信息. 最后, 权重分布矩阵分别与图像特征和问题特征相乘得到含有双边信息的图像特征$I_L$和问题特征$T_L$. 这个过程可用下面的 3 个公式表示:

$$logits = softmax\left(T_F^{\mathrm{T}} \times W_L \times I_F\right) \tag{10}$$

$$T_L = logits \times T_F \tag{11}$$

$$I_L = logits \times I_F \tag{12}$$

其中, 式 (10) 中的$W_L \in \mathbb{R}^{N\times K}$表示权重矩阵. 最后, 对于图像特征, 将左部分生成的图像特征$I_L$和右部分生成的图像特征$I_R$融合; 对于问题特征, 进行同样的处理. 用公式表示如下:

$$T_A = T_R \oplus T_F \tag{13}$$

$$I_A = I_R \oplus I_F \tag{14}$$

其中, $\oplus$表示, 两个特征的联合操作. 图像特征$I_A$和问题特征$T_A$是对称注意力模型的输出, 其维度与输入特征的维度一致.

### 3.3 新模型整体架构

图 2 是本文提出的新模型整体架构, 这里所用答案预测层是目前比较常用的多层感知器 (MultiLayer Perceptron, MLP) 神经网络分类器, 它有 2 个隐藏层和 1000 个隐藏单元 (dropout 为 0.5), 每一层都有 tanh 非线性函数. 首先, 对输入的图像和问题分别提取特征, 图像特征和问题特征作为对称注意力模型的输入, 然后生成包含双边信息的特征, 再作为答案预测分类器的输入, 最终选出得分高的答案. 这个过程可以通过以下公式表示:

$$T_A, I_A = att(T, I) \tag{15}$$

$$Ans = answer(T_A, I_A) \tag{16}$$

其中, $att()$表示对称注意力模型算法, $answer()$是答案预测分类器.

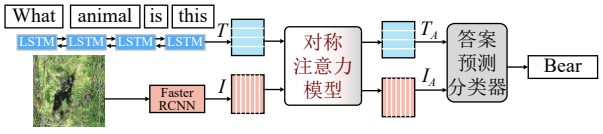


图 2 模型整体架构

### 3.4 总结

和现有的方法相比, 本文摒弃了只采用一个注意力特征矩阵来融合特征的方法或者将两模态特征简单连接的方法. 本模型通过已有的 Faster RCNN 和 Bi-LSTM 的方法构建了图像和问题的特征提取网络, 采用注意架构将两模态信息巧妙融合在一起, 既实现了对双边信息的理解, 又增强了图像区域特征的关联性和问题单词对之间的关联.

## 4 实验

### 4.1 数据集和实验细节

本文使用最流行的数据集 VQA 2.0[15] 来进行实验. VQA[16](也称为 VQA 1.0) 包含来自 MS COCO 数据集[17] 的 204 721 张图像上的人工注释的问答对. 预先将数据集分为 train、val 和 test(或 test-standard) 3 个部分, 它们分别由 248 349 个问题、121 512 个问题和 244 302 个问题组成. 所有的问题都被分为 3 种类型: 是/否、计数和其他, 每个问题都有 10 个自由回答的答案. VQA 2.0 是 VQA 1.0 的更新数据集, 与 VQA 1.0 相比, 它包含的样本更多 (train 数据集有 443 757 个, val 数据集有 214 354 个, test 数据集有 447 793 个), 在语言方面更加均衡. 本文使用具有挑战性的开放式任务的 VQA2.0 数据集评估提出的模型.

与其他工作一样, 本文选择出现 8 次以上的正确答案作为候选答案集. 根据之前的研究, 本文在 train+val 分支上训练模型, 并在 test-standard 和 test-dev 进行测试.

### 4.2 实验细节

本文所有实验均基于 PyTorch 框架, 并在装有一个 Nvidia Tesla P100 GPU 的计算机上进行实验. 使用的优化器的参数是$\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.99$. 在训练过程中, 使 VQA2.0 的学习率 ($\alpha$) 以 0.5 的速率每 7 个迭代下降一次. 所有的模型都在 VQA 2.0 上分别训练了 24 个迭代. 为了防止过拟合, 将每个全连接层的 dropout 设为 0.3、LSTM dropout 设为 0.1. 批次大小为 400, 隐藏层大小为 1024.

### 4.3 实验分析

本文采用准确率来评测模型的质量和训练情况. 在图 3 中, 绘制了准确率随迭代次数变化而变化的折线图, 可以看出本文提出的方法模型的准确率折线快速收敛, 不断提高. 再与图 4 的直方图结合观察, 可以看出随着迭代次数的增加, 模型的准确率也在不断提升, 迭代次数为 24 时效果最佳, 最高可以达到 66.34%. 通过图 3, 不难看出在超过 24 个迭代的时候, 由于模型出现了过拟合现象, 模型的准确率会有小幅度的下降. 在表 1 中, 列举了其他模型 (VQA team[15], MCB[18], MF-SIG-T3[19], Adelaide[20], DCA[4]) 和本文提出的模型在 VQA2.0 测试数据集上的准确率, 通过对比发现, 本文建立的模型具有较好的结果.

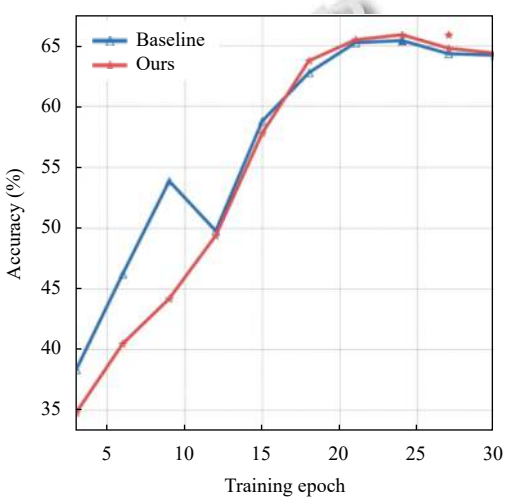


图 3 Overall 准确率损失变化

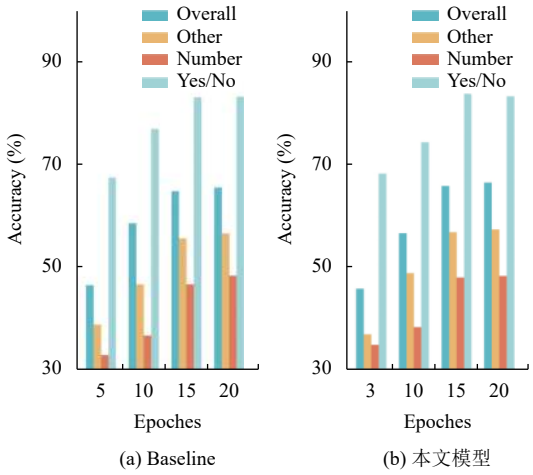


图 4 准确率柱状图

由表 1 可以看出, 本文提出的模型准确率优于基线模型. 新模型的 Overall 问答准确率比 baseline 在 Test-dev 数据集上提升了 1.22%, Other 问答准确率提升了 0.72%, Number 问答准确率提升了 0.9%, Yes/No 问答准确率提升了 0.38%. 这些数据证明: 本文提出的模型可以在较少的训练迭代次数下收敛, 基于对称注

意力机制的模型有效的提升了视觉问答的质量. 相比于传统的特征融合等方法, 对称注意力模型可以通过融合不同模态的信息, 增强问题信息和图像信息的自身关联性来大幅度提升答案分类的准确率.

表 1 与其他方法的实验结果比较 (%)

| Model | Test-dev | | | | Test-standard | | | |
|---|---|---|---|---|---|---|---|---|
| | Overall | Other | Number | Yes/No | Overall | Other | Number | Yes/No |
| VQA team-Prior[15] | – | – | – | – | 25.98 | 1.17 | 0.36 | 61.20 |
| VQA team-Language only[15] | – | – | – | – | 44.26 | 27.37 | 31.55 | 67.01 |
| VQA team-LSTM+CNN[15] | – | – | – | – | 47.22 | 41.83 | 35.18 | 73.46 |
| MCB[18]reported in VQA[15] | – | – | – | – | 62.27 | 53.36 | 38.28 | 78.82 |
| MF-SIG-T3[19] | 64.73 | 55.55 | 42.99 | 81.29 | – | – | – | – |
| Adelaide Model[20] | 62.07 | 52.62 | 39.46 | 79.20 | 62.27 | 52.59 | 39.77 | 79.32 |
| Adelaide +Detector[20] | 65.32 | 56.05 | 44.21 | 81.82 | 65.67 | 56.26 | 43.90 | 82.20 |
| DCA[4] | 65.12 | 56.10 | 47.32 | 83.18 | 66.08 | 56.33 | 47.12 | 83.48 |
| **对称注意力模型** | **66.34** | **56.82** | **48.22** | **83.64** | **66.46** | **57.21** | **48.15** | **83.21** |

## 5 结论与展望

本文提出了一种对称注意力机制的图像问答模型, 并在 VQA2.0 数据集上取得优异的成绩. 该算法的亮点在于使用全连接网络来挖掘图像区域之间的相关性, 联合基于共同注意力机制生成的双边信息特征, 达到更加精准的分类效果. 和 DCA 相比, 本文考虑了图像和问题的全面语义理解和融合, 较好地利用了自相关信息. 在未来的工作中, 将进一步探索视觉 (短视频) 问答系统和知识图谱对于答案分类的影响.

## 参考文献


1 袁韶祖, 王雷全, 吴春雷. 基于多粒度视频信息和注意力机制的视频场景识别. 计算机系统应用, 2020, 29(5): 252–256. [doi: 10.15888/j.cnki.csa.007410]

2 Cha M, Gwon YL, Kung HT. Adversarial learning of semantic relevance in text to image synthesis. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 3272–3279. [doi: 10.1609/aaai.v33i01.33013272]

3 Gao P, Jiang Z, You H, *et al*. Dynamic fusion with intra- and inter-modality attention flow for visual question answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. 2019. 6632–6641. [doi: 10.1109/CVPR.2019.00680]

4 Nguyen DK, Okatani T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 6087–6096. [doi: 10.1109/CVPR.2018.00637]

5 Mohammadi S, Majelan SG, Shokouhi SB. Ensembles of deep neural networks for action recognition in still images. Proceedings of 2019 9th International Conference on Computer and Knowledge Engineering. Mashhad, Iran. 2019. 315–318. [doi: 10.1109/ICCKE48569.2019.8965014]

6 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–587. [doi: 10.1109/CVPR.2014.81]

7 Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 4566–4575. [doi: 10.1109/CVPR.2015.7299087]

8 He HH. The parallel corpus for information extraction based on natural language processing and machine translation. Expert Systems, 2019, 36(5): e12349. [doi: 10.1111/exsy.12349]

9 Ge HW, Yan ZH, Zhang K, *et al*. Exploring overall contextual information for image captioning in human-like cognitive style. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea. 2019. 1754–1763.

10 Andreas J, Rohrbach M, Darrell T, *et al*. Neural module networks. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 39–48. [doi: 10.1109/CVPR.2016.12]

11 Yang ZC, He XD, Gao JF, *et al*. Stacked attention networks for image question answering. Proceedings of 2016IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 21–29. [doi: 10.1109/CVPR.2016.

10]

12 Xu HJ, Saenko K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. Proceedings of 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 451–466. [doi: 10.1007/978-3-319-46478-7_28]

13 Shih KJ, Singh S, Hoiem D. Where to look: Focus regions for visual question answering. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4613–4621. [doi: 10.1109/CVPR.2016.499]

14 Xiong CM, Merity S, Socher R. Dynamic memory networks for visual and textual question answering. Proceedings of the 33rd International Conference on Machine Learning. New York City, NY, USA. 2016. 2397–2406.

15 Agrawal A, Lu JS, Antol S, *et al*. VQA: Visual question answering. International Journal of Computer Vision, 2017, 123(1): 4–31. [doi: 10.1007/s11263-016-0966-6]

16 Goyal Y, Khot T, Summers-Stay D, *et al*. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6904–6913. [doi: 10.1109/CVPR.2017.670]

17 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of European Conference on Computer Vision. Zurich, Switzerland. 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

18 Fukui A, Park DH, Yang D, *et al*. Multimodal compact bilinear pooling for visual question answering and visual grounding. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA. 2016. 457–468. [doi: 10.18653/v1/D16-1044]

19 Chen Z, Zhao YP, Huang SY, *et al*. Structured attentions for visual question answering. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 1291–1300. [doi: 10.1109/iccv.2017.145]

20 Teney D, Anderson P, He XD, *et al*. Tips and tricks for visual question answering: Learnings from the 2017 challenge. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 4223–4232. [doi: 10.1109/CVPR.2018.00444]