计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800207

利用全局与局部帧级特征进行基于共享注意力的视频问答

王雷全[1] 候文艳[2] 袁韶祖[1] 赵 欣[2] 林 瑶[2] 吴春雷[1]
1 中国石油大学(华东)计算机科学与技术学院 山东 青岛 266555
2 中国石油大学(华东)海洋与空间信息学院 山东 青岛 266555

摘 要 视频问答是视觉理解领域中非常重要且具有挑战性的任务。目前的视觉问答(VQA)方法主要关注单个静态图片的问答,而现实生活中的数据是立体动态的视频。此外,由于问题的复杂性,视频问答任务必须根据问答问题恰当地处理多种视觉特征才能获得高质量的答案。文中提出了一个通过利用局部和全局帧级别的视觉信息来进行视频问答的多共享注意力网络。具体来说,以不同帧率提取视频帧,并以此提取帧级的全局与局部视觉特征,这两种特征包含了多个帧级别特征,用于对视频时间动态建模,再以共享注意力的形式建模全局与局部视觉特征的相关性,然后结合文本问题来推断答案。在天池视频问答数据集上进行了大量的实验,验证了所提方法的有效性。

关键词:视频问答;共享注意力机制;全局和局部帧级特征
中图法分类号 TP391

Multi-Shared Attention with Global and Local Pathways for Video Question Answering

WANG Lei-quan[1],HOU Wen-yan[2],YUAN Shao-zu[1],ZHAO Xin[2],LIN Yao[2] and WU Chun-lei[1]
1 College of Computer Science and Technology,China University of Petroleum,Qingdao,Shandong 266555,China
2 College of Oceanography and Space Informatics,China University of Petroleum,Qingdao,Shandong 266555,China

Abstract Video question answering is a challenging task of significant importance toward visual understanding. However,current visual question answering (VQA) methods mainly focus on a single static image,which is distinct from the sequential visual data we faced in the real world. In addition,due to the diversity of textual questions,the VideoQA task has to deal with various visual features to obtain the answers. This paper presents a multi-shared attention network by utilizing local and global frame-level visual information for video question answering (VideoQA). Specifically,a two-pathway model is proposed to capture the global and local frame-level features with different frame rates. The two pathways are fused together with the multi-shared attention by sharing the same attention funtion. Extensive experiments are conducted on Tianchi VideoQA dataset to validate the effectiveness of the proposed method.

Keywords Video question answering,Shared attention mechanism,Global and local pathways

1 引言

视频问答的任务是给定一个视频和相关的文本问题,由计算机自动推断出正确答案,它在人机交互、信息检索等领域都有广泛应用。视频问答是一个具有挑战性的领域,它需要用到视频识别、自然语言处理等多个领域的知识。

视觉问答(VOA)主要包括图像问答和视频问答问题的多样性是整个视觉问答(VQA)领域面临的一个常见问题。不同的问题需要不同层次的视觉特征来推断答案。图像问答只涉及静态图像的特征,相比而言,视频问答所面临的情况更加复杂,它必须处理具有冗余信息的帧序列。随着深度学习技术的发展,视频问答取得了长足的进步,但有些问题仍未得到妥善解决。如图1所示,答案不仅可能与帧级的局部视觉特征相关,还可能与全局视觉特征相关,且这些特征分散于多个视频帧中。如何平衡复杂的视觉信息以准确地回答视频问题,值得进一步的研究。

基于注意力的模型在视频描述[1-2]、视觉问答[3]等计算机视觉任务上取得了巨大的成功。目前,大多数基于注意力的视频问答模型都着重于根据问题注意哪个重点区域。为了关注到不同单词的重要性,文献[4]提出了协同注意力机制,以同时学习视觉注意力和问题注意力。然而,由于问题和视频内容的多样性,这些方法忽略了不同层级视觉特征的互补性。多个视觉特征的融合,如级联[5]或沿时间维度的双线性池化[6]可能会导致模型准确性低或计算成本过高[7]。多头注意

到稿日期:2020-08-29 返修日期:2020-09-30 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:科技部重点研发计划(2018YFC1406204),中央高校基本科研业务费专项资金(19CX05003A-11)
This work was supported by the National Key Research and Development Program(2018YFC1406204) and Fundamental Research Funds for the Central Universities(19CX05003A-11).
通信作者:王雷全(richiewlq@gmail.com)

力通过在不同的表示子空间中使用多个注意力模块来显示出有效性。但是，在多头注意力机制下采用不同特征融合的策略鲜有研究。本文探讨了如何在多头注意机制下聚合多模态帧级别信息，以获得视频问答的视觉信息特征，从而实现快速的视频问答。

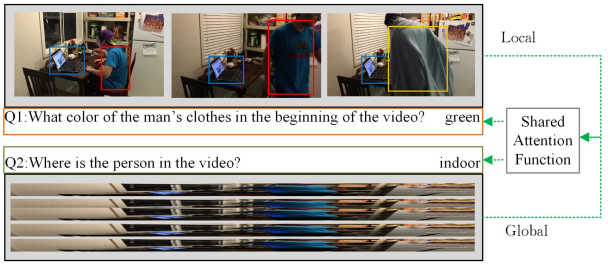


图 1 利用全局与局部帧级特征进行视频问答

Fig. 1 Exploiting global and local visual features for video question answering

基于上述考虑，本文提出用多头共享注意力网络聚合多模态视频特征以进行视频问答。本文首先提取了视频的帧级全局和局部视觉特征：全局特征是在时间序列上的多个帧级别信息；局部特征是通过目标检测方法提取视频帧的物体局部特征。为了更好地融合视频的时空信息，本文通过一个共享注意力权重的多头注意力融合网络对视频时空信息进行动态建模。多头注意力的优势是能够获取不同的表示子空间的信息，从而更好地理解视频内容。同时，每个单头模块中增设了多个注意力模块来缓解过拟合。为了获取这种时空特征的关联性，在每个单头注意力中通过共享相同的注意力来融合全局和局部特征。综上所述，本文的主要贡献如下：

(1)提出了一种用于视频问答的双路模型。这两条路分别用于提取帧级全局特征和局部特征，并利用互补的视觉特征来推断答案。

(2)设计了一种基于双路的多共享注意力机制视频问答模型。通过在同一注意力头中共享相同的注意力，为全局和局部注意力显式地建模相关性。

(3)本文通过实验对该方法进行了实证分析。在 2018 年之江杯全球人工智能大赛中的天池数据集上的实验结果验证了本文方法的有效性。

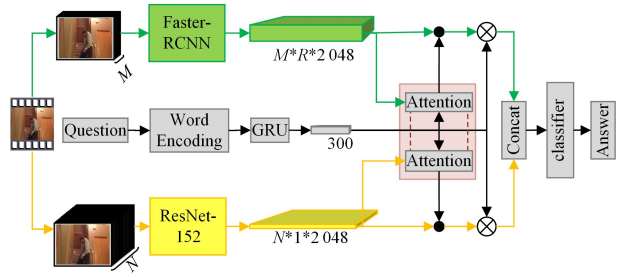
2 相关工作

图像问答旨在针对给定的图像和自然语言问题提供正确的答案。基于深度学习的方法在图像问答中更有优越性，因此将其中的 CNN 或 LSTM 用于视觉或文本特征提取。最近，注意力机制在图像问答中扮演着越来越重要的角色，它为研究关注图像的重点区域提供了指导。SAN(Stacked Attention Network)[8]使用多层注意力机制，通过多次查询图像来定位相关的视觉区域并逐步推断答案。为了同时关注不同单词和图像的重要性，一些工作使用 co-attention[5]来对图像和问题进行加权融合。Xu 等[3]提出了空间记忆网络来估计句子中每个图像块和标记之间的相关性。BUTD attention[9]使用物体检测方法 Faster R-CNN[10]提取了检测级别的特征，在视觉和语言任务之间建立了更紧密的联系。本文将重点放在视频问答任务上，该问题与静态图像问答有许多显著差异。

与图像问答相比，基于视频的问答是一个未被广泛探索的领域，将基于图像问答的几种注意力方法直接应用于视频问答时效果并不理想。截至目前，对视频问答的研究还很少。Yu 等[11]提出了一种语义注意力机制，将视频中检测到的物体与文本结合起来生成答案。Kim 等[12]引入了深度嵌入式记忆网络，利用视频问答的场景和对话的信息流来重构描述。Na 等[13]利用多层 CNN 捕获电影的顺序信息以推断答案。Gao 等[14]利用结构化片段与编码后的文字特征来进行视频时序结构特征的提取，进而生成回答。在这些工作中，没有一个权威的数据集被视频问答领域的大多数研究者接受，不同的方法适用于不同的视频问答数据集，大多数数据集(如 TGIF-QA[15])，要么单调，要么太短。在 2018 年之江杯全球人工智能大赛中使用的天池视频问答数据集是对该任务的补充，其中包含了许多具有丰富场景的长视频。本文考虑到该数据集的特点，提出了一种用于视频问答的将全局和局部特征融合的多共享注意力网络。

3 基于共享注意力的视频问答

本节首先详细描述利用全局与局部帧级特征进行基于共享注意力的视频问答方法。图 2 为该视频问答方法的流程图。由于之江杯视频问答数据集的答案主要是短语和单词，因此我们将视频问答任务视为一个对常见答案的分类问题。然后，详细介绍了本文的视频问答多共享注意力网络。最后，展示了分类器和损失函数。



注：红色框中的内容是一个共享注意力头，用于建模全局和局部视觉特征的相关性，并结合问题进行视频问答

图 2 具有共享注意力的 VQA 模型流程图(电子版为彩色)

Fig. 2 Flowchart of VQA model with shared attention

3.1 特征提取

本文将单条问题 q 表示为一个单词序列 $\{w_1, w_2, \cdots, w_t\}$ 的集合，其中 t 是问题 q 的长度。通过预先训练好的 GloVe 词嵌入[16]将每个单词 w_t 初始化为 y_t。然后，嵌入的单词 $\{y_1, y_2, \cdots, y_t\}$ 将被输入 Bi-GRU[17]中。再将最终隐藏状态 h_t 用作嵌入问题 q_e，经实验验证，h_t 维度设置为 300 时效果较好。

$$h_t = GRU(y_t, [\overrightarrow{h_{t-1}}, \overleftarrow{h_{t-1}}]) \tag{1}$$

视频 v 可表示为全局和局部两种特征。考虑到视频的冗余性，构建全局特征时从 v 采样 N 帧，构建局部特征时从 v 采样 M 帧。在构建全局特征时，本文使用 Resnet-152[18]的最后一个全连接层提取的特征作为视频的时间序列表示，全局路径表示为 $f^g = \{f_1^g, f_2^g, \cdots, f_N^g\}$。在构建局部空间特征时，

本文用 Fast-RCNN[19] 来提取空间中每个物体的表示，每帧保留 R 个物体，局部路径表示为 $f^l=\{f_1^l,f_2^l,\cdots,f_M^l\}$。因此，$f^g$ 和 f^l 的维数分别为 $N*1*2048$ 和 $M*R*2048$。

利用上述符号，可将视频问答问题表述为：给定一个视频 $v=\{f^g,f^l\}$ 和一个问题 q，通过本文提出的多共享注意力网络来自动推断答案 a。

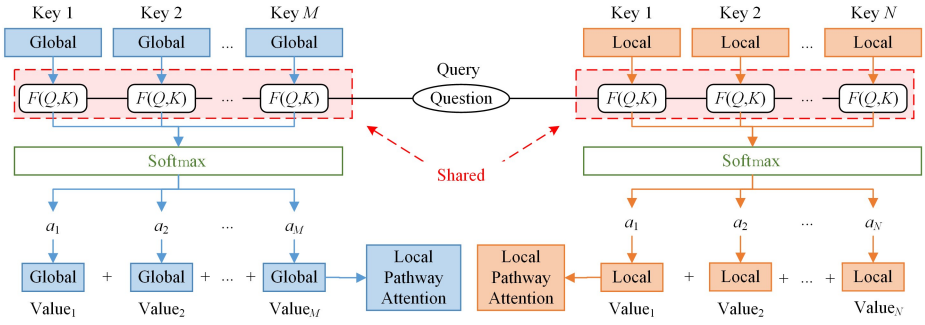
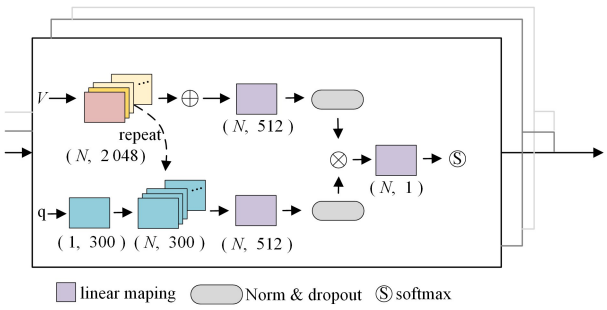


图 3　共享注意力示意图

Fig. 3　Shared attention diagram

计算注意力权重时，首先通过线性映射方式将经过词嵌入的问题 q_e、全局特征 $f^g=\{f_1^g,f_2^g,\cdots,f_N^g\}$ 和局部特征 $f^l=\{f_1^l,f_2^l,\cdots,f_M^l\}$ 映射到相应的空间中，然后计算特征相似度。为了简单起见，式(2)和式(3)简化了该过程，细节如图 4 所示。此外，受文献[20]的启发，本文算法利用类似的多头注意力对输出的注意力权重取平均，该过程可被视为注意力融合。为了避免过度拟合，本文在网络中使用了 dropout 和 batch normalization[21] 的结构。



注：V 可以用全局特征 f^g 或局部特征 f^l 代替

图 4　注意力单元的示意图

Fig. 4　Illustration of the attention unit

以上计算过程可以表述为：

$$a_i=\mathrm{avg}(\sum_k^n[f_i^g,q_e]W_a^k),i=1,2,\cdots,N \tag{2}$$

$$b_j=\mathrm{avg}(\sum_k^m[f_j^l,q_e]W_b^k),j=1,2,\cdots,M*R \tag{3}$$

其中，W_b 和 W_a 是学习的参数，k 是多头单元的数量。本文计算相似度时使用了“concat”，其他形式，如点乘、点加也是适用的。特别地，在相同的 k 下全局特征和局部特征的权重 W_a^k 和 W_b^k 是相同的，即共享相同的注意力可将全局和局部特征连接在一起。一方面，这样可以增强全局特征和局部特征的联系，另一方面，共享注意力机制也可以看作是一种数据增强。接下来，本文用 softmax 函数得到了作为注意力权重的 α_i 和 β_j。

3.2　多头共享注意力网络

考虑到问答问题的多样性，本文设计了一个提取全局特征和局部特征的并行架构，以了解视频中的帧或图像中需要注意的区域。不同于 Stacked Attention[8] 或 co-attention[5]，本文提出采用共享注意力来建模局部和全局视觉特征集合之间的相关性，如图 3 所示。

$$\alpha_i=\frac{\exp(a_i)}{\sum_{d=1}^{N}\exp(a_d)},i=1,2,\cdots,N \tag{4}$$

$$\beta_j=\frac{\exp(b_j)}{\sum_{d=1}^{M*R}\exp(b_d)},j=1,2,\cdots,M*R \tag{5}$$

然后，通过线性函数 F 和 G 将 f^g 和 f^l 的加权和问题 q_e 点乘在一起得到问题和视觉特征联合嵌入的 x。

$$x=F(q_e)\odot G([\sum_i^N\alpha_i f_i^g,\sum_j^{M*R}\beta_j f_j^l]) \tag{6}$$

最后，将问题和视觉特征 x 传递给答案分类器，该分类器由两层 MLP 和 softmax 层组成，则推测的答案类别可表示为：

$$\hat{p}=\mathrm{softmax}((xW_1+b)W_2+b_2) \tag{7}$$

其中，W_1 和 W_2，b_1 和 b_2 分别是两层 MLP 的权重与偏置。在优化时，模型采用交叉熵作为损失函数进行学习。

$$L=-\sum(p\log(\hat{p})-(1-p)\log(1-\hat{p})) \tag{8}$$

4　实验和分析

4.1　数据集

在 2018 年之江杯全球人工智能大赛发布的天池视频问答数据集上，我们对本文模型进行了评估。该数据集包括 10524 个视频和 52620 个问答对，每个问题有 1～3 个答案。与其他视频问答数据集相比，天池视频问答数据集包含了大量具有丰富场景的长视频，包括电影、动漫、相机拍摄的视频等。其中，8083 个视频用于训练，1641 个视频用于验证，800 个视频用于测试。

4.2　实验细节

在实验中，Adam 优化器的权重衰减率被设置为 0.001，文本问题词嵌入的维度为 300，问题与视觉特征的联合嵌入维数为 1024。实验证明，*batch size* 设置为 16 时具有最佳效果。在注意力网络中，激活函数采用了 Leaky ReLU，并将 *drouput* 值设置为 0.3。GRU 中的所有权重均采用均匀分布初始化，其他权重均采用正态分布初始化。全局特征包含 19

帧，局部特征包含 6 帧，Fast-RCNN 的边界框数量 R 取值为 36。在构建答案集时，本文抽取了最常见的 1 357 个答案进行分类。本文的评价指标是准确率。

4.3 不同方法的比较

为了验证算法的有效性，将本文方法与几种典型的视频问答方法进行了比较，包括 Soft-attention[22]，Gtanh-attention[9]，Co-attention[4]和 Bilinear-attention[23]，结果如表 1 所列。这些方法应用了多种注意力机制来融合视觉特征。可以看到，本文方法取得了最好的结果。尽管 Co-attention[4]也在视觉特征层面上增强了联系，但本文方法的准确率比其高出了 0.75%，这也证明了本文方法的有效性。其原因之一是多共享注意力强化了全局和局部信息的联系。

表 1 视频问答不同注意力方法的比较

Table 1 Comparison of different attention approaches on videoQA

Methods	*Accuracy*/%
Soft-attention	53.17
Co-attention	56.25
Gtanh-attention	55.96
Bilinear-attention	56.18
Multi-shared attention (ours)	57.00

本文除利用了多头策略外，共享注意力机制有效地捕捉到了全局和局部信息的相关性也是性能提升的一个关键。一般来说，全局和局部特征的注意力结合优于全局或局部特征的单一注意力。它表明了由于问题类型的不同，全局特征和局部特征包含的视频信息具有一定互补性。因此，使用共享注意力比仅使用全局特征或者局部特征上的注意力有更好的效果。在每个“单头”的注意力单元中都采用这种共享机制来将全局特征和局部特征联系起来。从数据的角度来看，共享注意力也可以看作是一种提高泛化性能的数据增强方式。

4.4 多共享注意力的比较

表 2 列出了多共享注意力的性能。从表 2 可以看出，在注意力机制中应用多头策略十分有利于实验结果的提升，这是因为多头注意力通过在不同表示子空间中使用多个注意力单元来提升网络的性能。特别地，当多头注意力单元 k 的数目设置为 3 时，多共享注意力网络达到了最佳性能。

表 2 具有不同数目的注意头的多共享注意力的性能

Table 2 Performances of multi-shared attention with different number of attention heads

Approach	multi-head	*Accuracy*/%
attention on global	3	39.65
attention on local	3	53.41
respective attention	3	54.32
shared attention	2	55.97
	3	57.00
	4	56.74

4.5 不同问题类型的比较

为了更好地验证文本方法的性能，图 5 将不同类型问题回答的准确性做了可视化比较。当问题类型为“where”和“color”时，本文提出的带有全局和局部特征的多共享注意力表现得很好。相反，本文方法在“doing”类型的问题回答上表现最差。其主要原因是本文为了追求问答效率，没有在视频问答方法中使用 3D 卷积或光流等运动特征，这是今后要解决的主要问题。图 6 给出了本文算法的一些例子。从图 6 也可以看出，本文算法在问题为帧级全局或局部视觉特征相关的问题上表现良好（见图 6(a)与图 6(b)），不过在回答动态时序特征相关的问题时则表现较差（见图 6(c)）。

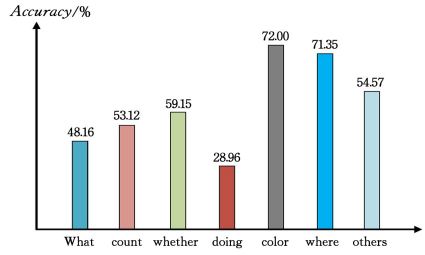
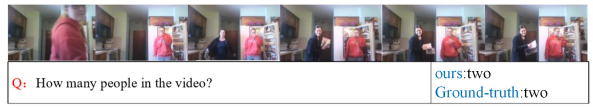


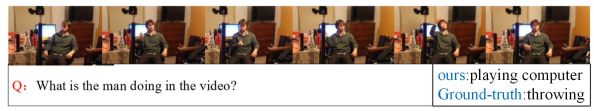
图 5 回答不同类型问题的表现

Fig. 5 Performances on answering the different types of questions

(a)



(b)



(c)

图 6 生成答案的示例

Fig. 6 Examples of generated answers

结束语 本文提出了一个同时利用局部和全局特征进行视频问答的网络，利用多头共享注意力机制，将全局和局部的特征关联起来，并结合问题来关注视频中的重点区域。在天池数据集上的实验结果表明了该方法的有效性。在未来的工作中，我们将考虑在保证回答问题速度的前提下，使用动态的特征来改善“What is the man doing?”这一类问题的回答质量。

参 考 文 献

[1] WU C, WEI Y, CHU X, et al. Hierarchical attention-based multimodal fusion for video captioning[J]. Neurocomputing, 2018, 315: 362-370.

[2] XU Z L, DONG H W. Video Question Answering Scheme Based on Prior MASK Attention Mechanism[J]. Computer Engineering, 2021, 47(2): 52-59.

[3] XU H, SAENKO K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering[C]// European Conference on Computer Vision. Cham: Springer, 2016: 451-466.

[4] XIONG C, ZHONG V, SOCHER R. Dynamic coattention networks for question answering[J]. arXiv: 1611.01604, 2016.

[5] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C]// Advances in

Neural Information Processing Systems. 2016:289-297.
[6] FUKUI A,PARK D H,YANG D,et al. Multimodal compact bilinear pooling for visual question answering and visual grounding [J]. arXiv:1606.01847,2016.
[7] KIM K M,CHOI S H,KIM J H,et al. Multimodal dual attention memory for video story question answering[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018:673-688.
[8] 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.
[9] ANDERSON P,HE X,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] KRISHNA R,ZHU Y,GROTH O,et al. Visual genome:Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision,2017,123(1):32-73.
[11] YU Y,KO H,CHOI J,et al. End-to-end concept word detection for video captioning,retrieval,and question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:3165-3173.
[12] KIM K M,HEO M O,CHOI S H,et al. Deepstory:Video story qa by deep embedded memory networks[J]. arXiv:1707.00836,2017.
[13] NA S,LEE S,KIM J,et al. A read-write memory network for movie story understanding[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:677-685.
[14] GAO L,ZENG P,SONG J,et al. Structured two-stream attention network for video question answering[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:6391-6398.
[15] JANG Y,SONG Y,YU Y,et al. Tgif-qa:Toward spatio-temporal reasoning in visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:2758-2766.
[16] PENNINGTON J,SOCHER R,MANNING C D. Glove:Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014:1532-1543.
[17] CHO K,VAN MERRIËNBOER B,GULCEHRE C,et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv:1406.1078,2014.
[18] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.
[19] REN S,HE K,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.
[20] JABRI A,JOULIN A,LAURENS V D M. Revisiting visual question answering baselines[C]// European Conference on Computer Vision. Cham:Springer,2016:727-739.
[21] IOFFE S,SZEGEDY C. Batch Normalization:Accelerating Deep Network Training by Reducing Internal Covariate Shift [J]. arXiv:1502.03167,2015.
[22] XU K,BA J,KIROS R,et al. Show,attend and tell:Neural image caption generation with visual attention[C]// International Conference on Machine Learning. 2015:2048-2057.
[23] KIM J H,JUN J,ZHANG B T. Bilinear attention networks [C]// Advances in Neural Information Processing Systems. 2018:1564-1574.

WANG Lei-quan, born in 1981, Ph.D, senior experimenter, is a member of China Computer Federation. His main research interests include cross media analysis and action recognition.