计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200015

视觉问答中问题处理算法研究

徐 胜 祝永新
中国科学院上海高等研究院 上海 201210
中国科学院大学 北京 100049
(xusheng@sari.ac.cn)

摘 要 当前对视觉问答(Visual Question Answering,VQA)建模的研究多种多样,但现有的 VQA 模型有一个共同的缺点:训练和推理较为耗时。研究表明,VQA 模型中文本处理部分主要基于长短期记忆网络(Long Short Term Memory,LSTM),而 VQA 模型的整体性能也受制于文本处理部分的 LSTM 网络,由于 LSTM 网络具有循环的特性,LSTM 网络中复杂的数据流难以有效利用 GPU 的并行计算优势来加速计算。针对以上问题,以优化模型的训练速度为目的,提出了一个新模型 SCMP(Simple Conv1d MaxPool1d)来代替 LSTM 网络处理输入模型的自然语言文本。在 VQA2.0 数据集上的实验结果表明,该模型与现有的模型相比训练速度提高了 10 倍,并且没有对 VQA 模型的精度造成损失。此外,文中提出了一种新颖的方法来对 VQA2.0 数据集中的文本数据进行数据增强。实验结果表明,数据增强可以提高 VQA 模型的精度,同时加速模型收敛,使用增强后的数据训练的模型(SCMP)在验证集上的评估分数为 63.46%,优于目前现存的 VQA 模型。

关键词:视觉问答;自然语言处理;卷积神经网络;长短期记忆网络;词嵌入
中图法分类号 TP391

Study on Question Processing Algorithms in Visual Question Answering

XU Sheng and ZHU Yong-xin
Shanghai Advanced Research Institute,Chinese Academy of Sciences,Shanghai 201210,China
University of Chinese Academy of Science,Beijing 100049,China

Abstract At present,there are various researches on the modeling of Visual Question Answering (VQA) tasks,but existing VQA models have a common drawback,i.e. training and reasoning are time-consuming. Research shows that the text processing part of the VQA model is mainly based on LSTM (Long Short Term Memory) networks,and the overall performance of the VQA model is also limited by the LSTM network used for the text processing. Due to the recurrent nature of the LSTM network,the complex data streams in the LSTM network can hardly take advantages of GPU parallel computing to accelerate. Aiming at the above problems,and for the purpose of optimizing the training speed of the model,a new model named SCMP (Simple Conv1d MaxPool1d) is proposed in this paper to replace the LSTM network to deal with incoming natural language questions. The experimental results on the VQA2.0 dataset show that the training speed of the model is 10 times faster than the existing model,and there is no loss for the accuracy of the VQA model. In addition,this paper proposes a novel method for data augmentation of question datasets in VQA2.0 datasets. Experimental results show that data augmentation can improve model prediction performance and accelerate model convergence. The model trained with enhanced data (SCMP) obtains an evaluation score of 63.46% on the validation set,which is better than the existing VQA model.

Keywords Visual question answering,Natural language processing,CNN,LSTM,Word embedding

1 引言

视觉问答解决的任务如下:如图 1 所示,对于给定的图片和一个基于图片内容的自然语言问题,视觉问答系统能基于图片和问题给出答案。视觉问答因其广泛的应用领域(例如图像检索、智能机器人、帮助失明或视力受损的人[1]等),尤其

到稿日期:2019-12-02 返修日期:2020-03-20 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:国家自然科学基金(U1831118);中国科学院战略性先导科技专项(XDA19000000,XDA19090106);上海市科学技术委员会科研计划项目(18511103502)
This work was supported by the National Natural Science Foundation of China (U1831118),Strategic Leading Science and Technology Project of the Chinese Academy of Sciences (XDA19000000,XDA19090106) and Shanghai Committee Science and Technology Scientific Research Project (18511103502).
通信作者:祝永新(zhuyongxin@sari.ac.cn)

是在人工智能快速发展的今天，被认为是一个 AI 完备问题，可以作为视觉图灵测试[2]。

问题：What animals is this in the picture?
答案：Cat

图 1　视觉问答

Fig. 1　Visual question answering

视觉问答作为一个极具挑战性的研究方向，涉及自然语言处理技术、计算机视觉技术以及视觉和文本的融合技术。视觉问答中文本的处理因对视觉问答模型的性能有重大影响而占据重要位置。因此，视觉问答中文本的处理有着重要的研究意义。

研究表明，VQA 模型中的文本处理部分主要基于 LSTM 网络，而 VQA 模型的整体性能也受制于 LSTM 网络，而基于 LSTM 网络的模型的训练和推理过程是非常耗时的。耗时的训练致使实验周转时间延长，同时也限制了研究人员的迭代速度，由此导致模型所能使用的数据集规模受到制约。最重要的是，耗时的推断导致视觉问答难以部署在实时应用程序中。

针对这些不足，本文考虑设计了一个轻量化的文本处理网络来替换 LSTM 网络。对视觉问答任务中的文本处理过程进行分析可知，视觉问答任务中的问题文本都是简短的单句，根据文献[3]，99.75%的问题文本不超过 14 个单词。因此，可以认为视觉问答文本处理中没有多语句间的复杂语义特征需要提取。考虑到轻量化的网络可能会给模型的精度带来细微的影响，本文对文本部分的数据做增强处理。

本文的主要贡献如下：

(1)提出了一个文本处理模型 SCMP，该模型能加快文本的处理速度，而不会牺牲整个 VQA 模型的准确性。

(2)采用了一种精巧的方法来扩充问题文本数据集，该数据集在几个主流模型上均取得了良好的效果。

本文第 2 节讨论了相关工作；第 3 节介绍了视觉问答任务的内容，讨论了 SCMP 模型与基准模型的优劣；第 4 节介绍了对数据集中的文本部分进行扩增的方法；第 5 节给出了实验过程，并对实验结果进行了分析；最后总结全文。

2　相关工作

视觉问答任务是深度学习领域的一个研究热点，涉及自然语言处理和计算机视觉，以及两者的融合。自然语言处理是视觉问答模型中的一个重要组成部分[1]。

循环神经网络(Recurrent Neural Network，RNNs)一直主导着自然语言处理领域的发展。自然语言文本序列的本质以及文本前后的相互依存关系与 RNNs 尤其是 LSTM 的设计理念相吻合。因此，RNNs 受到了研究人员的普遍欢迎，而视觉问答作为一个涉及自然语言处理的多模态任务也不例外。尽管研究者已经提出许多模型来解决这个多模态任务，如 Vis ＋ LSTM 模型[4]、动态参数预测模型(DPPnet)[5]、堆叠注意力网络(SAN)[6]、分层联合注意网络(CoAtt)[7]、多模压缩双线性池模型(MCB)[8]，以及自下而上和自上而下注意模型(Bottom-Up and Up-Down)[9]，但上述 VQA 模型均基于 LSTM 网络。虽然它们取得了不错的性能，但是 LSTM 网络循环的特性使其难以有效利用 GPU 的并行计算特性进行加速计算[10]。以文献[9]为例，模型使用 VQA2.0 数据集，在两块 K20 GPU 上训练 12 h。因此，有必要设计一个轻量化的文本处理网络来替换普遍存在的 LSTM 网络。

视觉问答中文本处理部分的关键在于词向量的表征[11]以及如何将这些表征有效地组合成句向量的表征。本文提出了 SCMP 模型，其与目前视觉问答中广泛使用的 LSTM 网络的不同点在于：该轻量化设计的模型能充分利用 GPU 的并行加速计算特性。实验结果表明，该模型在训练速度上明显优于现有模型。为了进一步提高模型的准确性，本文提出了一项新颖的数据增强技术来增强训练数据中的问题文本。实验结果表明，该方法能显著提高模型的精度。

3　SCMP 模型

本节首先对视觉问答任务进行阐述，然后对本文提出的 SCMP 模型进行了详细介绍和分析。

3.1　视觉问答任务

本文研究的视觉问答任务定义为：如图 2 所示，对于输入的图片 I_i 和一个基于给定图片 I_i 的自然语言问题 q_{ij}，视觉问答模型将会融合 I_i 和 q_{ij} 的特征计算出一个自然语言答案 a_{ij}。

$$a_{ij}=\arg\max_{\theta} p_0(a_{ij}\mid q_{ij},I_i;\theta),a\in\Omega \tag{1}$$

其中，Ω 是出现频次最高的前 K 个待分类候选答案组成的集合，θ 是视觉问答模型中待学习的参数。

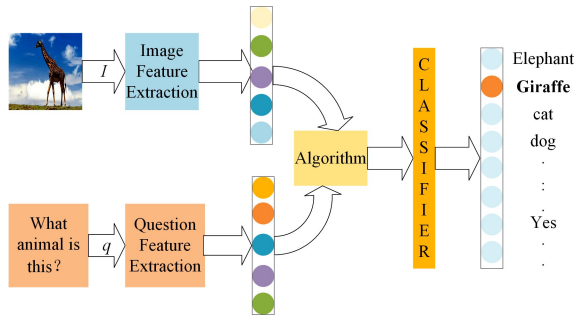


图 2　视觉问答模型的简化框架

Fig. 2　Simplified framework for visual question answering model

3.2　SCMP 模型理论

对于 VQA 模型而言，问题通常用一个简短的句子或者短语来描述查询对象的重要信息。具有局部感受野和共享权重的卷积核能够获取连续单词之间丰富的语义特征[12]，因此本文提出的 SCMP 模型也是基于卷积神经网络延展而来的。

首先 SCMP 模型的输入是一个问题文本，表示为 $Q=[w_1,w_2,\cdots,w_L]$，其中 w_l 代表问题序列中的第 l 个单词，L 代表问题的最大长度。然后，通过词向量模型(见式(2))将每个单词映射到向量空间中，并且把所有的词向量按单词顺序拼接形成一个新的矩阵来表示问题 Q。因此问题文本 Q 可以

被描述为式(3)。式(3)中，‖表示按照问题的单词顺序进行拼接操作，$\boldsymbol{V}_{1:L}$可以视作一个特殊的文本特征矩阵。

$$v_l = f_{Word\text{-}Embedding}(w_l) \tag{2}$$

$$\boldsymbol{V}_{1:L} = [v_1 \| v_2 \| \cdots \| v_L] \tag{3}$$

采用一维卷积来提取短语的抽象语义特征。采用窗口大小为 s 的卷积核来对第 l 个单词进行卷积，卷积后的结果如式(4)所示：

$$v_l^s = \varphi(W^s v_{l:l+s-1} + b^s) \tag{4}$$

其中，$\varphi(\cdot)$是一个非线性激活函数，W^s 和 b^s 分别是卷积核和偏置项。

经过卷积操作后，新的问题文本特征可以表示为：

$$Q^s = [v_1^s \| v_2^s \| \cdots \| v_{L+1-s}^s] \tag{5}$$

接着，对提取的文本特征进行大小为 s 的最大池化操作，计算结果为：

$$\tilde{Q}^s = \max_l(v_l^s) = \max_l[v_1^s, v_2^s, \cdots, v_{L+1-s}^s] \tag{6}$$

通过卷积和最大池化操作，可以很好地提取连续单词之间的局部语义特征。最后将提取完成的语义特征输入到全连接层，得到包含语义特征的单维文本特征向量 $\boldsymbol{q}$。基于上述步骤，使用 SCMP 实现了从问题文本 Q 到问题文本特征向量 $\boldsymbol{q}$ 的映射。

3.3 SCMP 模型和 LSTM 网络的比较

本文提出的模型是在视觉问答挑战赛冠军模型[9]的基础上改进而来的，冠军模型如图 3 所示，本文将以此模型为基准模型(以下简称 Benchmark Model)。本文的模型与 Benchmark Model 的不同之处在于文本特征的提取方式不同。

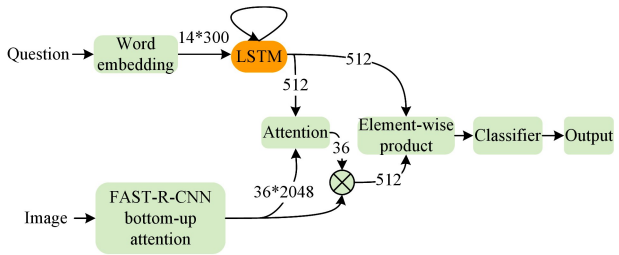


图 3 视觉问答基准模型

Fig. 3 Visual question answering benchmark model

对于 Benchmark Model 来说，其文本的编码是通过 LSTM 或 GRU 实现的，而本文提出的模型的文本编码过程如图 4 所示。

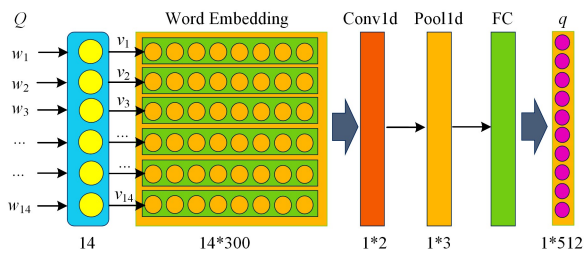


图 4 SCMP 框架图

Fig. 4 Overview of SCMP structure

首先将文本输入 Q 采用 NLTK 3.3(Natural Language Toolkit)分词工具分词为 $w_1, w_2, w_3, \cdots, w_{14}$ 单词序列，下标 14 表示一个句子序列的单词个数，也即句子序列的长度。本文将所有的文本序列都修剪成固定长度以提升处理效率[3]。

随后采用 GloVe(Global Vectors for Word Representation)[13]模型对分词后的单词做词的向量化表示。$\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \cdots, \boldsymbol{v}_{14}, \boldsymbol{v}_i \in R^k$ 分别代表以上单词 w_i ($0 \leqslant i \leqslant 14$)做词嵌入后的词向量。$k$ 表示向量的空间维度，此处是 300D。

向量化表示后，对形成的矩阵沿句子序列的方向进行卷积，卷积核的大小设为 1×2，卷积核的窗口覆盖句子内 2 个连续的单词，滑动的滤波器窗口应用于这个句子以生成相应的特征映射。然后，在特征映射的基础上使用最大池化(1×3)来提取最显著的语义特征。最后，在全连接层之后产生该句子向量的最终表征 q。

该方法的合理性在于：在视觉问答任务中，图像已经提供了足够的信息，视觉问答模型只需要从问题中提取一些关键信息来补全视觉信息以匹配出答案。因此，卷积运算考虑句子内一些连续的单词，滤波器应用于这些单词窗口以生成相应的语义特征。在提取的语义特征之上使用最大池化来抽象更高层的显著语义特征[14]。从 Q 到 q 的神经网络映射：$Q \rightarrow \boldsymbol{q}$，旨在将自然语言词向量融合表征为定长的自然语言句向量 $\boldsymbol{q}$。

对于 Benchmark Model，LSTM 网络(见图 5)的三重门(输入门 i_t、输出门 O_t 和遗忘门 f_t)设计使其在处理长期依赖的文本数据上有一定的优势，但由式(7)－式(11)可知，其循环递归的特性和复杂的数据流导致其只能串行计算，难以采用 GPU 进行有效的并行加速计算[10]。式(7)－式(11)中，σ 为 logistic sigmoid 激活函数，tanh 是双曲正切激活函数。

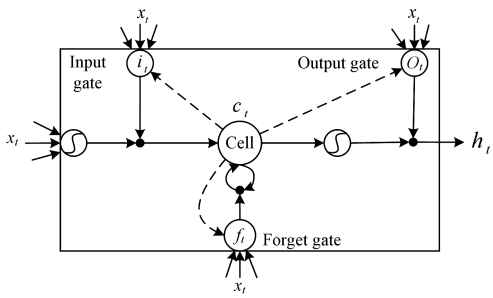


图 5 LSTM 网络图

Fig. 5 LSTM Network

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{7}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{8}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{9}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{10}$$

$$h_t = o_t \tanh(c_t) \tag{11}$$

此外，该基准模型在对文本编码的过程中，问题被分词并映射为词向量后，依次送入 LSTM 网络，这将大大限制整个视觉问答模型的训练速度。

相比被广泛采用的 LSTM 模型，本文提出的模型能在几乎没有精度损失的情况下减少操作数。模型中假设输入层节点数为 N_i，隐藏层节点数为 N_h，输出层节点数为 N_o，操作数为 N_{ops}。

由图 3 和图 5 可知，两层 LSTM 网络的操作数可以近似估算为[15]：

$$N_{ops}=2[(N_i+N_h)\cdot N_h\cdot 4\cdot 2+N_h\cdot(8+4+4)]$$
$$\approx 8.6\times 10^6 \tag{12}$$

由图 4 可知，本文提出的 SCMP 模型的操作数可估算为：

$$N'_{ops}=(N_i-1)\cdot 300\cdot 3+(N_i-3)\cdot 300$$
$$\approx 1.5\times 10^6 \tag{13}$$

通过以上分析可知，相比 LSTM 网络，本文提出的 SCMP 模型将导致操作数大幅减少。此外，本文还没有把 LSTM 网络中计算激活函数所必需的大量取幂和指数运算考虑进来。因此可以断定，本模型的轻便性和数据流的清晰简便性更占优势。

4　对数据集中文本部分进行扩增的方法

本文采用 Back Translation 方法对 VQA2.0 数据集中的文本问题数据集进行了扩增。

该方法通过将原始问题从英语翻译成另一种语言，然后再翻译成英语来扩增问题，该过程使训练问题的数量加倍，从而丰富了数据集的样本空间。如图 6 所示，该方法使用两个翻译引擎：一个是从英文到中文，另一个是从中文到英文，以获得问题的同义转述[16]。

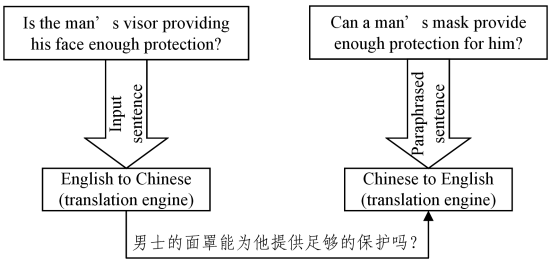


图 6　用中文作为演示语言来解释数据增强过程

Fig. 6　Explaining data augmentation with Chinese as a demonstration language

5　实验与分析

本节将进行两组实验，分别研究文本数据增强和文本处理模型 SCMP 的表现。实验选用 VQA2.0 数据集，所有实验均在 2 台 Tesla K20 GPU 上完成。

5.1　数据增强实验结果分析

表 1 列出了数据增强对模型性能的影响。实验过程中参考文献[7-9]中的最佳实验性能设置。

表 1　在 VQA2.0 上的测试集准确率

Table 1　Test Accuracy on VQA2.0

Method	Test Accuracy/%
CoAtt-Model[7]	59.22
MCB-Model[8]	62.27
Benchmark Model[9]	70.34
CoAtt-Model+ Data Augmentation	60.19
MCB-Model + Data Augmentation	63.12
Benchmark Model+ Data Augmentation	70.67

表 1 中所有模型均在 VQA2.0 train-val 上进行训练，并在 VQA2.0 test-standard 上进行了测试。实验结果表明，使用增强的 VQA2.0 数据集进行训练的模型性能优于没有增强的 VQA2.0 数据集训练的模型，这也充分说明了提出的文本数据增强技术的有效性和普适性。

5.2　SCMP 模型实验结果分析

表 2 所列的实验结果表明，采用 SCMP 模型来代替 Benchmark Model 中的 LSTM(GRU)使训练速度提高了 10 倍左右，从而使精度损失较小。本实验只是在验证集上显示得分以供进一步分析，因为 test-standard 每天可以向 VQA 评估服务器提交 1 次，并且总共不超过 10 次。总结文献[17-19]发现，验证结果与测试结果呈较好的相关性。

由表 2 也可看出，基于 Benchmark Model 进行的一些改进，即用简单的一维卷积池操作(SCMP)替换复杂的 LSTM 以对问题文本进行编码，可以提高 VQA 模型的训练速度。

表 2　在 VQA2.0 上的验证集准确率和训练时间

Table 2　Validation accuracy and training time on VQA2.0

Method	Val Accuracy/%	Training Time/h
Benchmark Model (LSTM)	63.12	8～12
Ours (SCMP)	63.10	0.67～1

表 3 所列的实验结果表明，通过数据增强训练的 SCMP 模型的性能优于 Benchmark Model 模型。此外，与 NeurIPS 2018 中的“UpDn+DoE”相比，“SCMP+Data Augmentation”的性能更好。

表 3　在 VQA2.0 上的验证集准确率

Table 3　Validation Results on VQA2.0

Method	Val Accuracy/%
Benchmark Model	63.12
UpDn+DoE[20]	63.43
Ours (SCMP)+Data Augmentation	63.46

5.3　数据增强加速模型收敛

图 7 给出了训练过程中的损失曲线，可知用带有数据增强的数据训练模型时损失曲线迅速下降，采用数据增强的模型能在 4～6 个 epoch 达到没有数据增强的模型训练 19～21 个 epoch 才能达到的水平。该曲线表明数据增强可以加速训练模型的收敛。

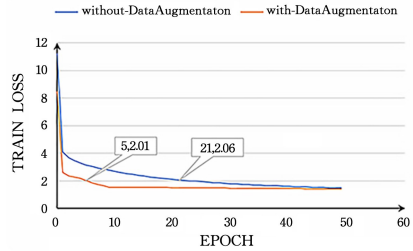


图 7　SCMP 模型在有无数据扩增下的训练损失曲线

Fig. 7　Train loss of ours model with and without data augmentation

5.4　小结

传统的视觉问答模型多使用 LSTM 来完成问题中词向量到句向量的映射。但是由于 LSTM 网络循环递归的本质，GPU 很难对其实现有效的加速[10]。针对上述问题，本文设计了一个仅使用卷积和池化操作的轻量化网路。而对于卷积和池化操作，可以充分利用 GPU 的并行计算特性来加速。为了对提出的网络进行评估，本文进行了相关实验。实验结果表明，所提模型在速度上有很大的优势。

结束语　本文提出了一种快速轻便的 SCMP 模型，用于

替代视觉问答模型中编码问题文本的 LSTM。为弥补 LSTM 难以充分利用 GPU 并行计算特性来加速的缺点，所提出的模型完全是前馈的，它由卷积、池化和线性层组成，可采用 GPU 的并行计算来加速。此外，本文提出了一种新颖的数据增强方法，即先把英文翻译为中文，再将中文翻译为英文来回作为增强 VQA2.0 数据集的方法，实验结果表明其可以显著提升模型的精度。

本文的亮点在于：

(1)开创性地对视觉问答中的文本数据集进行了数据增强。

(2)深入分析了视觉问答中的文本处理过程，能根据自然语言任务的实际情况合理建模，不刻意追求模型的复杂度，在速度与精度中找到一个较好的平衡点。

本文的不足之处在于：

(1)无法对数据增强的有效性从理论上做出充分阐述。

(2)对于本文提出的文本处理模型 SCMP，没有在精度上进行进一步的优化。

因此，我们在后续研究中将对除 VQA2.0 数据集外的其他视觉问答数据集进行数据增强，以验证本文提出的数据扩增方法的有效性；另外我们将在精度和速度方面进一步优化本文提出的文本处理模型 SCMP。

参 考 文 献

[1] AGRAWAL A, LU J, ANTOL S, et al. VQA: Visual Question Answering[J]. International Journal of Computer Vision, 2017, 123(1): 4-31.

[2] DESTA M T, CHEN L, KORNUTA T. Object-based reasoning in VQA[C]// 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 1814-1823.

[3] TENEY D, ANDERSON P, HE X, et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge [C]// Computer Vision and Pattern Recognition. 2018: 4223-4232.

[4] REN M, KIROS R, ZEMEL R. Exploring models and data for image question answering[C]// Advances in Neural Information Processing Systems. 2015: 2953-2961.

[5] NOH H, HONGSUCK SEO P, HAN B. Image question answering using convolutional neural network with dynamic parameter prediction[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 30-38.

[6] 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.

[7] LU J, YANG J, BATRA D, et al. Hierarchical co-attention for visual question answering[J]. Advances in Neural Information Processing Systems, 2016: 289-297.

[8] 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.

[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] CHANG A X, MARTINI B, CULURCIELLO E, et al. Recurrent Neural Networks Hardware Implementation on FPGA[J]. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2015, 5(1): 401-409.

[11] ZHANG L U, SHEN C L, LI S S. Emotion classification algorithm based on emotion-specific word vectors [J]. Computer Science, 2019, 46(S1): 93-97.

[12] MA L, LU Z, LI H, et al. Learning to answer questions from image using convolutional neural network[C]// National Conference on Artificial Intelligence. 2016: 3567-3573.

[13] PENNINGTON J, SOCHER R, MANNING C D, et al. Glove: Global Vectors for Word Representation[C]// Empirical Methods in Natural Language Processing. 2014: 1532-1543.

[14] SHEN D, WANG G, WANG W, et al. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms[J]. arXiv: 1805.09843, 2018.

[15] SUN Z, ZHU Y, ZHENG Y, et al. FPGA Acceleration of LSTM Based on Data for Test Flight[C]// 2018 IEEE International Conference on Smart Cloud (Smart Cloud). 2018: 1-6.

[16] YU A W, DOHAN D, LUONG M T, et al. Qanet: Combining local convolution with global self-attention for reading comprehension[J]. arXiv: 1804.09541, 2018.

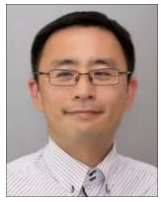
[17] GOYAL Y, KHOT T, SUMMERS-STAY D, et al. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6904-6913.

[18] AGRAWAL A, KEMBHAVI A, BATRA D, et al. C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset[J]. arXiv: 1704.08243, 2017.

[19] SINGH J, YING V, NUTKIEWICZ A. Attention on attention: Architectures for visual question answering (vqa)[J]. arXiv: 1803.07724, 2018.

[20] RAMAKRISHNAN S, AGRAWAL A, LEE S. Overcoming language priors in visual question answering with adversarial regularization[C]// Advances in Neural Information Processing Systems. 2018: 1541-1551.

XU Sheng, born in 1993, postgraduate. His main research interests include deep learning and natural language processing.



ZHU Yong-xin, born in 1969, Ph.D, researcher, is a member of China Computer Federation. His main research interests include computer system architecture, system-level chip design, big data, and artificial intelligence.