基于注意力机制的视觉问答任务研究

白姣姣 柯显信 曹 斌
(上海大学机电工程与自动化学院 上海 200444)

摘 要 提出一种基于注意力机制的视觉问答系统,通过匹配图像中与问题相关的区域来回答基于图像的问题。通过拼接的方式将问题特征与图像特征映射到一个共享空间,再通过非线性层、线性层以及 Softmax 层来得到注意力权重。该方法将视觉问答任务视为一个多分类任务,将数据集中出现频率最高的 1 000 个答案作为候选答案。利用预训练的 VGG16 模型提取图片特征,利用 LSTM 网络提取问题特征,采用 VQA 数据集进行训练和测试。

关键词 视觉问答 注意力机制 LSTM 人机交互

中图分类号 TP391 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2020.10.023

VISUAL QUESTION ANSWERING TASK BASED ON ATTENTION MECHANISM

Bai Jiaojiao Ke Xianxin Cao Bin
(*School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China*)

Abstract This paper proposes a visual question answering system based on attention mechanism to answers image-based questions by matching areas of the image that are relevant to the problem. It mapped the question features and visual features into a shared space by concatenation, and obtained the attention weight through the nonlinear layer, the linear layer and the Softmax layer. The visual question answering task was regarded as a multi-category task, and 1 000 answers with the highest frequency in the data set were regarded as candidate answers. The image feature was extracted by the pre-trained VGG16 model, the problem feature was extracted by LSTM network, and the VQA data set was used for training and testing.

Keywords Visual question answering Attention mechanism LSTM Human-computer interaction

0 引 言

近几年来人工智能飞速发展,智能机器人的功能也因此越来越强大,具备图像识别、语音识别、情感识别和对话处理等多种功能。对智能机器人视觉问答任务的研究可以帮助盲人这一弱势群体。视觉问答系统可以回答盲人的询问,帮助盲人了解周围环境等。盲人用户可以直接通过语音进行提问,经过一系列处理后返回对应答案,实现盲人辅助。

视觉问答涉及多方面的人工智能技术,如:细粒度识别(这位女士是黄种人吗?)、物体识别(图中的水果是苹果嘛?)、行为识别(这位男士在干什么呢?)和对问题所包含文本的理解(自然语言处理)等。自由式和开放式的视觉问答任务首次出现于 2015 年,虽然出现时间较晚,但经过几年的发展已成为研究的热点。视觉问答任务涉及图像处理、自然语言处理等多个领域,虽然目前的图像处理技术已经可以很好地进行图像检测与识别,但还不能较好地理解图片内容。而视觉问答任务旨在解决这一问题,将图像及基于图像的问题输入模型,经过处理后输出该问题对应的答案。

目前,科研工作者们已经在视觉问答领域做了很多工作。文献[1-3]采用无注意力机制的深度学习模型来处理视觉问答任务,其中文献[1]使用预训练的卷积神经网络编码图像,使用循环神经网络编码问题,将图片与问题特征融合后传入全连接层,最后传入

收稿日期:2019-06-28。国家自然科学基金项目(61273325)。白姣姣,硕士生,主研领域:智能机器人。柯显信,副教授。曹斌,硕士生。

Softmax 层; 文献 [3] 使用三个不同的卷积神经网络分别编码图像、问题,以及二者融合后的特征,其中编码图像的卷积神经网络与 VGG 模型的架构相同。文献 [4-7] 则采用基于注意力机制的深度学习模型来处理视觉问答任务,该方法可以赋予图片不同区域的特征不同的重要性,突出图片中与问题相关的部分。其中: 文献 [5] 为了获得更细粒度的视觉信息,反复计算图像的注意力权重; 文献 [6] 既计算图像的注意力权重,也计算问题的注意力权重; 文献 [7] 将视觉问答任务视为多类别分类问题而不是多选一。文献 [8-10] 借助外部知识库中的信息来处理视觉问答,通过知识库可以使模型能够像人类一样具有“常识”,例如,在回答“图片中有多少种花?”时,模型首先要知道花的种类有哪些,这种方法极大地提高了模型的泛化能力。

1 视觉问答系统

本文提出的视觉问答系统的实现包括语音识别、语音合成、图像识别、视觉问答模型四个子模块。在交互过程中,交互对象声音由麦克风录制,录制的音频由语音识别模块转化为文本; 图像由摄像头捕捉,由图片识别模块提取其特征; 视觉问答模型首先提取文本特征,然后将其与图片特征融合,融合后的特征输入神经网络,经过处理后生成相应回答并合成语音。

系统将视觉问答任务视为一个多分类任务进行处理,其流程如图 1 所示。

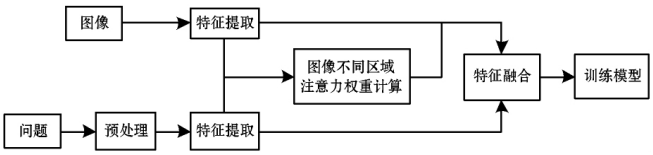


图 1 视觉问答任务流程

1.1 图像特征提取

卷积神经网络凭借其突出的性能在图像处理领域获得了举足轻重的地位,它于 1998 年被 Lecun 等首次提出,被称为 LeNet。该网络有 3 个卷积层,2 个下采样层和 1 个全连接层,卷积层的卷积核大小均为 5 × 5,该模型在识别手写字符和打印字符的任务中取得了优秀的效果。2012 年,Alex 等提出了 AlexNet 模型,该模型是卷积神经网络的突破模型,在“ILSVRC”图像分类的比赛中获得了冠军,证明了通过增加网络的深度可以提高网络的性能。该模型包括 5 个卷积层和 3 个全连接层,卷积层的第一层卷积核大小为 11 × 11,步长为 4,第二卷积核大小为 5 × 5,步长为 1,剩余卷积层的卷积核大小都为 3 × 3,步长为 1。在 2014 年“ILSVR”挑战赛中赢得了定位任务冠军和分类任务亚军的 VGG 卷积网络模型进一步加深了网络的结构,由于其优异的性能被人们广泛运用于各种图像处理任务。然而,神经网络层数的增加使得网络的训练变得困难,出现层数越大错误率越高的退化现象。何凯明等于 2015 年提出了 ResNet 残差神经网络,该模型成功地解决了退化问题。随着卷积神经网络的不断发展,人们又相继提出了 R-CNN、Fast R-CNN、Faster R-CNN 等区域卷积神经网络来更好地处理目标检测的问题。

本文模型使用预训练的 VGG16 卷积网络进行图像特征提取,使用 VGG16 模型最后一层池化层提取得到的特征作为图像特征。VGG16 的模型结构如图 2 所示,该模型包含 13 个卷积层,5 个最大池化层和 2 个全连接层,卷积核的大小均为 3 × 3,最后一层池化层输出的特征向量为(7,7,512)。

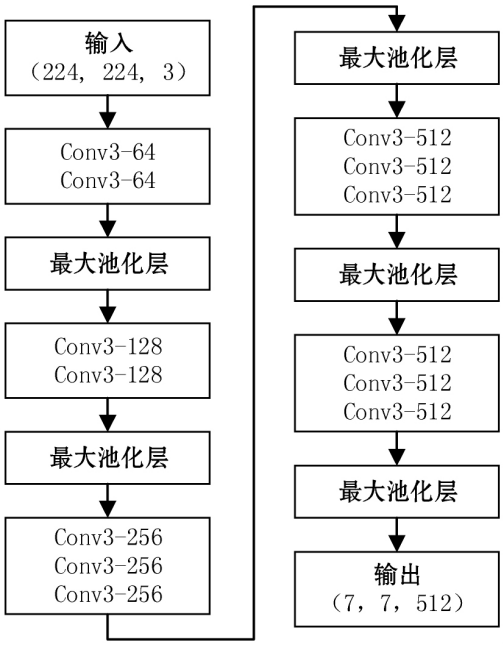


图 2 VGG16 模型

1.2 问答对特征提取

首先处理答案,统计数据集中答案出现的频率,选取出现频率最高的 1 000 个答案作为标签,并通过独热编码(One-Hot Encoding)。

接着处理问题,在对问题进行预处理之前,首先判断与之对应的答案是否被编码,舍弃答案未被编码的问答对。然后,采用 NLTK 工具包对问题进行预处理。通过正则表达式对问题进行分词处理,匹配句中的单词并去掉标点符号。英文分词比中文分词容易实现,可以直接基于空格分词。在对问题进行分词处理后,句子中仍然存在“the”“that”“this”等出现频率相当高的词,这类词的存在对问题的理解并没有太大的作用,还会增加数据量。NLTK 工具包中提供了一个英文停用词表,通过删除在问题中出现的该表中的词来实现

去停用词处理。接着,进行词型还原的处理,即将英文单词的复数或第三人称单数形式还原成单词原型,该步骤同样利用 NLTK 工具包实现。最后,将问题中的单词统一恢复为小写。问题预处理变化如表 1 所示。

表 1　问题预处理

原问题	What are these photos taken looking through?
分词	['What', 'are', 'these', 'photo', 'taken', 'looking', 'through']
去停用词	['What', 'photos', 'taken', 'looking']
词型还原	['What', 'photo', 'taken', 'looking']
小写还原	['what', 'photo', 'taken', 'looking']

问题通过上述预处理,降低特征维度,减少没有用的信息,增强模型的泛化能力,同时也可以避免模型过拟合。经过预处理后,问题的最大长度是 13 个单词,采用尾部对齐的方式,将不足 13 个单词的句子后面补 0,并统计每个句子的实际长度。采用预训练的 Glove 模型进行词嵌入,该模型的向量维度是 300 维。Glove 模型是通过构造一个共现矩阵来学习,共现矩阵主要是计算一个单词在上下文中出现的频率。

最后将问题词向量经过一个单元数为 512 的 LSTM 神经网络,得到问题的特征向量为(1,512)。

LSTM 神经网络是一种特殊的循环神经网络模型,循环神经网络能够在记忆单元中储存之前的信息,可以很好地处理序列问题,因此在自然语言处理领域得到了广泛的应用。LSTM 模型可以学习长期依赖的信息,解决了梯度爆炸的问题,其结构如图 3 所示。

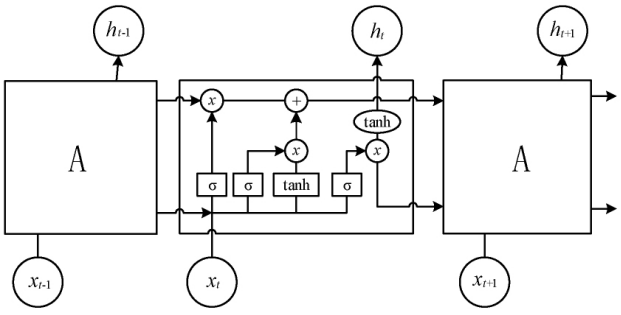


图 3　LSTM 结构示意图

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3}$$

式中:f_t、i_t、o_t 分别表示 t 时刻遗忘门、输入门和输出门的状态;h_{t-1} 代表前一时刻的记忆;x_t 为当前时刻的输入;h_t 为当前时刻的输出;W、b 分别代表权重和偏置,为神经网络的可训练参数。

1.3　图像注意力计算

深度学习中的注意力机制(Attention),实现了将更多的注意力资源投入某一区域的功能,与人类的注意力机制类似。人类可以通过快速浏览整幅图像,获得需要重点关注的区域,也就是人们常说的注意力焦点,然后将注意力集中在这一区域,以获取需要重点关注的目标的信息,而忽略其他无用的信息。注意力机制可以帮助人类从大量的信息中筛选出高价值的信息,是人类在长期进化中形成的一种生存机制。从本质上讲,深度学习中的注意力机制和人类的注意力机制类似,也是为了从繁多的信息中筛选出对当前任务更有用的信息。目前,注意力机制已被广泛应用于自然语言处理、图像识别、语音识别等各种领域,并取得了良好的效果。

本文所使用的图像特征的注意力计算方法如图 4 所示。将图像特征与问题特征拼接后经过一个非线性层,之后经过一个线性层和 Softmax 层,从而得到一幅图像不同区域的注意力权重。

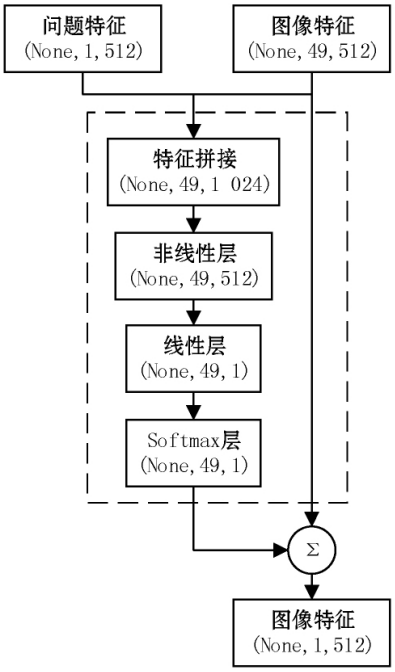


图 4　图像注意力计算示意图

如图 4 所示,在拼接图像与问题特征时,将问题的特征向量拼接在图像每一个区域的特征向量之后。最后经过 Softmax 层得到图像各区域的注意力权重后,进行如下运算:

$$\hat{\boldsymbol{v}} = \sum_{i=1}^{K} \alpha_i \boldsymbol{v}_i \tag{4}$$

式中:α 代表各区域注意力权重;v 代表图像各区域的特征向量;K 值为 49,是图像被划分的区域数。

非线性层的计算公式如下:

$$\tilde{\boldsymbol{y}} = \tanh(W\boldsymbol{x} + b) \tag{5}$$

$$\boldsymbol{g} = \sigma(W'\boldsymbol{x} + b') \tag{6}$$

$$y = \tilde{y} \cdot g \tag{7}$$

式中: x 代表图像与问题特征拼接后的向量; y 代表计算得到的特征向量; "·"代表逐元素相乘。

1.4 特征融合

在得到了图片特征和问题特征后,需要对二者进行特征融合。通过计算注意力之后得到的图片特征向量为 512 维,通过 LSTM 神经网络得到的问题向量也为 512 维。将二者进行点乘来实现特征融合,同时也将二者按列进行拼接实现特征融合对比。

1.5 训 练

将特征融合之后的向量通过两层全连接层和一层 Softmax 层。搭建的视觉问答模型结构如图 5 所示。模型使用交叉熵作为损失函数,全连接层单元数分别为 1 024 和 1 000,使用 RMSprop 作为优化函数,batch 的数目为 200,epoch 的数目为 20。

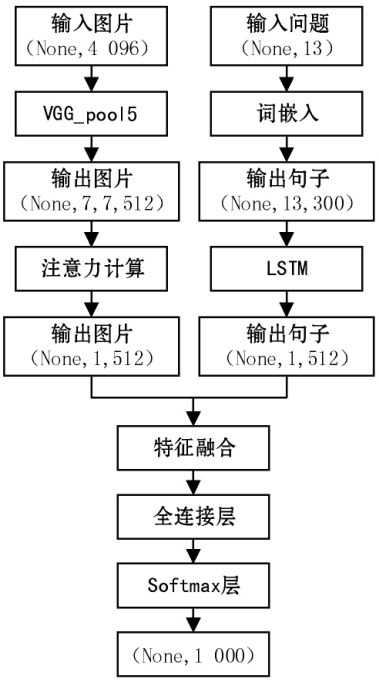


图 5 模型结构图

2 实 验

2.1 数据集

视觉问答任务发展至今,已有很多的数据集,包括 DQAUAR、COCO-QA、FM-IQA、Visual Genome 和 VQA 等。DQAUAR 数据集中的部分图片质量较低,且图片的内容较为单一,只包含室内场景,限制了问题的多样性,但它所有答案的个数不超过 1 000 个。COCO-QA 数据集的主要问题是问答对由自然语言处理模型根据图片标题自动生成,答案的个数也不超过 1 000 个。FM-IQA 数据集中有的问题的回答是一个句子,这给统计答案频率增加了难度。Visual Genome 数据集中频率最高的 1 000 个答案仅占所有答案的 65%,而 VQA 数据集中频率最高的 1 000 个答案约占答案总数的 82.7%。

综上,本文模型采用 VQA 数据集进行训练与测试,该数据集中包含了 82 783 幅训练图片、81 434 幅测试图片以及 40 505 幅验证图片,而且图片涉及了多种多样的场景。每幅图片对应 3 个问答对,问题的种类繁多,如 'what is this' 'what color' 'is this' 'does the' 'how many' 等。其中被编码的 1 000 个答案对应训练集中的 387 976 句问题,测试集中的 186 937 句问题。对于问题中未出现在预训练的"Glove"中的词,编码为 0。VQA 数据集的举例如图 6 所示,数据集中的问题与答案如表 2 所示。

图 6 VQA 数据集举例(穿着橘色上衣的人)

表 2 数据集中针对图 6 的问题

问题	答案
What is this photo taken looking through?	net
What position is this man playing?	pitcher
What color is the players shirt?	orange

2.2 实验结果分析

针对第 1 章节中介绍的模型,使用 VQA 训练集进行训练,并与下列无注意力机制模型的准确率进行对比:

1) 模型一: 图片特征由 VGG16 最后一层全连接层提取得到,特征向量为(None,1,4 096),将其经过全连接层得到特征向量为(None,1,300);问题经过预处理和词嵌入后得到的特征向量为(None,13,300)。将图像作为问题的最后一个单词,实现特征融合后得到特征向量为(None,14,300),之后分别传入单元数为 1 024 的 LSTM 和 BiLSTM 神经网络,最后传入 Softmax 得到分类结果。

2) 模型二: 只使用问题特征,只通过 LSTM 神经网络,其余参数设置与模型一中一致。

3) 模型三: 图片特征由 VGG16 最后一层全连接层提取得到,特征向量为(None,1,4 096),其经过全连

接层得到特征向量为(None,1,1 024) ; 问题经过预处理和词嵌入后得到的特征向量为(None,13,300) ,其通过 LSTM 神经网络之后得到特征向量为(None,1,1 024) 。二者按行融合后得到特征向量(None,2,1 024) ,之后通过单元数为 512 的 LSTM 神经网络,最后再通过 Softmax 层实现分类。

4) 模型四: 将文献 [4] 中计算注意力权重的方式应用到本文模型当中,即二者除注意力权重计算方法外,其余参数与方法均一致。

不同模型在验证集上得到的准确率结果如表 3 所示。

表 3　实验结果

模型	测试集准确率
模型一(LSTM)	0.401 4
模型一(BiLSTM)	0.410 2
模型二	0.357 1
模型三	0.407 3
模型四	0.440 1
注意力模型(点乘融合)	0.441 8
注意力模型(拼接融合)	0.424 4

可以看出,基于注意力机制的模型的准确率高于其他模型,而使用点乘融合方式的注意力模型准确率高于使用拼接融合方式的注意力模型,与模型四不相上下,但本文模型相对需要较少的计算。所有模型的准确率都高于模型二,这说明模型在训练过程中确实使用了图像信息。

为了验证模型在不同类型问题上的准确率,从测试集中提取出 53 692 对关于“是非”的问答,23 192 对关于“数量”的问答,以及 19 962 对关于“颜色”的问答,结果如表 4 所示。

表 4　分类实验结果

模型	询问颜色	询问数量	是非问题
模型一(LSTM)	0.334 1	0.267 2	0.542 5
模型一(BiLSTM)	0.341 4	0.245 4	0.549 1
模型二	0.284 7	0.238 2	0.501 8
模型三	0.320 5	0.257 7	0.531 9
模型四	0.374 4	0.272 1	0.554 6
注意力模型(点乘融合)	0.371 2	0.269 6	0.567 8
注意力模型(拼接融合)	0.360 4	0.299 7	0.550 0

可以看出,所有模型均是在是非问题上的准确率最高,因为是非问题只有“Yes”“No”两种答案,这也是模型二在是非问题上准确率在 50% 左右的原因,而其他模型有图片信息作为输入,准确率均高于模型二。注意力模型在关于数量问题上的准确率最低,说明模型不能很好地完成数数任务。

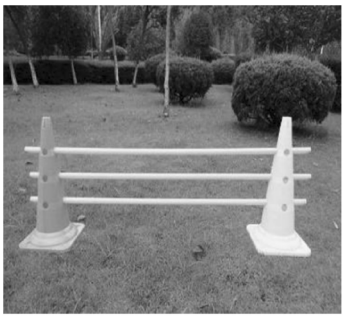
将图片和问题在上述六个模型上作对比,对于同一个问题,不同模型预测的答案中排名前三的结果如图 7 所示。

模型	What color is the signal?
模型一(LSTM)	green(0.220 6) , silver(0.188 3) , red(0.186 8)
模型二(BiLSTM)	green(0.356 4) , red(0.228 1) , red and white (0.004 4)
模型二	white(0.517 8) , red(0.166 2) , green(0.162 4)
模型三	green(0.166 7) , white(0.070 2) , red and white (0.059 7)
模型四	green(0.382 4) , red(0.136 4) , silver(0.011 9)
注意力模型(点乘融合)	green(0.380 0) , red(0.023 4) , red and white (0.004 9)
注意力模型(拼接融合)	Green(0.381 1) , red(0.240 8) , red and white (0.175 4)

(a) 询问颜色(绿色交通灯,灯杆红白相间,背景有雾)

模型	How many people are there?
模型一(LSTM)	1(0.312 2) , 2(0.183 1) , 3(0.151 7)
模型一(BiLSTM)	4(0.209 6) , 1(0.184 5) , 0(0.144 2)
模型二	1(0.356 6) , 2(0.227 6) , 4(0.117 3)
模型三	4(0.240 3) , 3(0.186 8) , 0(0.184 8)
模型四	1(0.343 8) , 4(0.145 2) , 2(0.024 6)
注意力模型(点乘融合)	4(0.354 2) , 1(0.063 4) , 0(0.007 4)
注意力模型(拼接融合)	4(0.349 4) , 3(0.221 3) , 2(0.141 9)

(b) 询问数量



模型	Is there somethingon the ground?
模型一(LSTM)	Yes(0.845),no(0.155)
模型一(BiLSTM)	Yes(0.867),no(0.133)
模型二	Yes(0.593),no(0.506)
模型三	Yes(0.804),no(0.054)
模型四	Yes(0.926),no(0.074)
注意力模型(点乘融合)	Yes(0.900),no(0.106)
注意力模型(拼接融合)	Yes(0.903),no(0.086)

(c) 询问是非

图7 模型测试对比

正如图7所示,基于注意力机制的视觉问答模型可以初步辅助盲人解决颜色及是非问题,但在数数方面做得不好。

3 结 语

为了实现多功能的人机交互,本文提出基于注意力机制的视觉问答系统,通过采集问题及周围环境信息,再通过基于注意力的深度学习模型的处理,得到回答并合成语音。实验表明,本文模型在一定程度上完成视觉问答任务,能帮助盲人解决某些场景中的问题,有助于盲人理解周围环境。然而,模型也有很多不足之处,比如特征融合方式、模型准确率等。今后将致力于图像特征与问题特征融合方式的研究、注意力权重计算方法的研究,以及基于知识图谱的模型研究,以提高模型的准确率,更好地完成视觉问答任务,从而实现盲人辅助的功能。

参 考 文 献

[1] Ren M, Kiros R, Zemel R. Exploring models and data for image question answering [C]//Advances in Neural Information Processing Systems, 2015.

[2] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: a neural-based approach to answering questions about images [C]//2015 IEEE International Conference on Computer Vision(ICCV), 2015.

[3] Ma L, Lu Z, Li H. Learning to answer questions from image using convolutional neural network [C]//Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[4] Shih K J, Singh S, Hoiem D. Where to look: Focus regions for visual question answering [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[5] Yang Z, He X, Gao J, et al. Stacked attention networks for image question answering [C]//2016IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[6] 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.

[7] Teney D, Anderson P, He X, et al. Tips and tricks for visual question answering: learnings from the 2017 challenge [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[8] Wu Q, Wang P, Shen C, et al. Ask me anything: Free-form visual question answering based on knowledge from external sources [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[9] Wang P, Wu Q, Shen C, et al. Fvqa: Fact-Based visual question answering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(10): 2413 - 2427.

[10] Wu Q, Shen C, Wang P, et al. Image captioning and visual question answering based on attributes and external knowledge [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1367 - 1381.

(上接第132页)

[13] LibJpeg-turbo [CP/OL]. [2019-07-03]. https://www.libjpeg-turbo.org/.

[14] Engineer Ambitiously™. 应用程序设计模式: 生产者/消费者 [EB/OL]. [2019-07-03]. http://www.ni.com/tutorial/3023/en/.

[15] Unity. Texture2D.LoadImage [EB/OL]. [2019-07-03]. https://docs.unity3d.com/530/Documentation/ScriptReference/Texture2D.LoadImage.html.

[16] Unity. Texture2D.LoadRawTextureData [EB/OL]. [2019-07-03]. https://docs.unity3d.com/ScriptReference/Texture2D.LoadRawTextureData.html.

[17] Unity. Low-level native plug-in interface [EB/OL]. (2017-05-16) [2019-07-03]. https://docs.unity3d.com/Manual/NativePluginInterface.html.

[18] Noguera J M, Segura R J, Ogayar C J. A scalable architecture for 3D map navigation on mobile devices [J]. Personal and Ubiquitous Computing, 2013, 17(7): 1487 - 1502.

[19] Chen Y J, Hung C Y, Chien S Y, et al. Distributed rendering: Interaction delay reduction in remote rendering with client-end GPU-accelerated scene warping technique [C]// IEEE International Conference on Multimedia and Expo Workshops. IEEE, 2017.