文章编号：2095-2163(2023)12-0138-06 中图分类号：TP391.4 文献标志码：A

基于门控机制的联合关系推理视觉问答模型

胡 婷，何利力
（浙江理工大学 计算机科学与技术学院，杭州 310018）

摘 要：与问题相关的视觉对象提取准确度不够，以及视觉对象之间的关系推理能力不足，是现有视觉问答模型视觉推理能力不足的主要原因。针对这两个方面的问题，本文提出一种基于门控机制的联合关系推理视觉问答模型（VARG）。该模型利用视觉注意力机制关注多个与问题相关的区域，通过筛选机制提取与问题最相关的前 N 个区域，并在此基础上建立视觉关系特征进行视觉关系推理，引入门控选择机制，动态的控制视觉特征和视觉关系特征对于答案的贡献，以此提升模型视觉推理能力。经在 VQA V2 数据集上进行实验，证明了模型的有效性。

关键词：视觉问答；注意力机制；门控机制；视觉关系推理

A visual question answering model for joint relational reasoning based on gating mechanisms

HU Ting, HE Lili
(School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: The lack of accuracy in extracting visual objects related to the problem and the lack of reasoning ability of the relationship between visual objects are two main reasons for the lack of reasoning ability of the existing visual question-answering model. To solve these two problems, this paper proposes a gated mechanism-based joint relation reasoning visual question answering model (VARG), which uses the visual attention mechanism to focus on multiple areas related to the problem, extracts the top N areas most relevant to the problem through the screening mechanism, and establishes visual relation features on this basis for visual relation reasoning. The gating selection mechanism is introduced to dynamically control the contribution of visual features and visual relation features to the answers, to improve the visual reasoning ability of the model. Experiments on VQA V2 data set demonstrate the validity of the model.

Key words: visual question answering; attention mechanism; gate control mechanism; visual relational reasoning

0 引 言

视觉问答[1]是近年来兴起的多模态任务，其涉及计算机视觉和自然语言处理等领域。具体来说，以一幅图像和一个自然语言描述的问题作为输入，要求机器基于问题，通过推理图像中的视觉元素来做出回答[2]。目前，在盲人问询、医疗、儿童早教等领域有着广泛应用。

注意力机制通过赋权重的方式，对于视觉区域的重要程度进行划分，数值越高则与问题的关联性越强。Yang[3]等人提出一种堆叠注意力网络，首次将注意力机制应用在视觉问答中；Derson [4]等人提出自上而下和自下而上组合的注意力机制，自下而上的机制关注图像中视觉区域，每一个区域有对应的特征向量，自上而下的机制决定特征权重；Google 提出 Transformer[5]模型，模型中提出了自注意力机制和多头注意力机制；Yu 等[6]在多头注意力机制的基础上提出深度协同注意网络，其由自注意单元和引导注意单元构成，协同的建模模态内和模态间的相互作用。除注意力机制之外，Zhu 等[7]通过加入外部知识库帮助模型理解图像内容；兰红等人[8]提出问题引导的空间关系图推理视觉问答模型，显式的建模场景中对象之间的联系，但模型并未考虑重点视觉特征对于问题的突出贡献；Ma[9]等提出融合注意力机制和关系提取的视觉问答模型，根据问题突出图像视觉区域，并通过不同尺度对视觉域区进行组合，但模型并未考虑图像的空间位置特征对于模型的贡献。

作者简介：胡 婷（1994-），女，硕士研究生，主要研究方向：计算机视觉。
通讯作者：何利力（1966-），男，博士，教授，博士生导师，主要研究方向：数据分析。Email：llhe@ zju.edu.cn

收稿日期：2022-12-24

本文针对视觉问答与问题相关视觉对象的提取准确度不够，以及视觉对象之间的关系推理能力不足等问题，提出了一种基于门控机制的联合关系推理视觉问答模型（VARG）。该模型从视觉特征和视觉关系特征两个角度对问题分别建模，并引入门控选择机制评估两者与问题的相关性，动态控制两种特征对于预测答案的重要程度。

1　VARG 模型

视觉问答的输入一般由一张图像和一个自然语言问题组成，视觉问答的任务是根据图像内容生成问题对应的答案。图像和问题属于不同模态，需要通过不同模型分别提取出特征，再通过特征的融合处理，最后预测出可能的答案。本文提出的 VARG 模型结构如图 1 所示，该模型由特征提取、视觉注意力模块、视觉关系推理模块、门控选择机制和答案预测 5 部分组成。

1.1　特征提取

图像和文本是两个完全不同的模态，需要采用不同的模型进行特征提取，对于输入的图像，本文选择基于 resnet−101[10] 网络的 faster−r−cnn[11] 模型进行物体的特征检测，并选择最相关的前 36 个检测区域，每个检测区域可以通过视觉向量 $\boldsymbol{y} = [y_1, y_2, \cdots, y_k]^{\mathrm{T}}$ 和位置向量 $\boldsymbol{B} = [b_1, b_2, \cdots, b_k]^{\mathrm{T}}$ 表示。其中，每个区域的视觉向量 $y_i \in d_y$，每个区域的位置特征 $b_i = [e_i, f_i, w_i, h_i]$。其中 e_i、f_i 是检测框 i 的中心坐标，w_i、h_i 是检测框 i 的宽和高，上述向量将会作为图像特征提取层的输出，输入到视觉注意力模块和视觉关系推理模块。

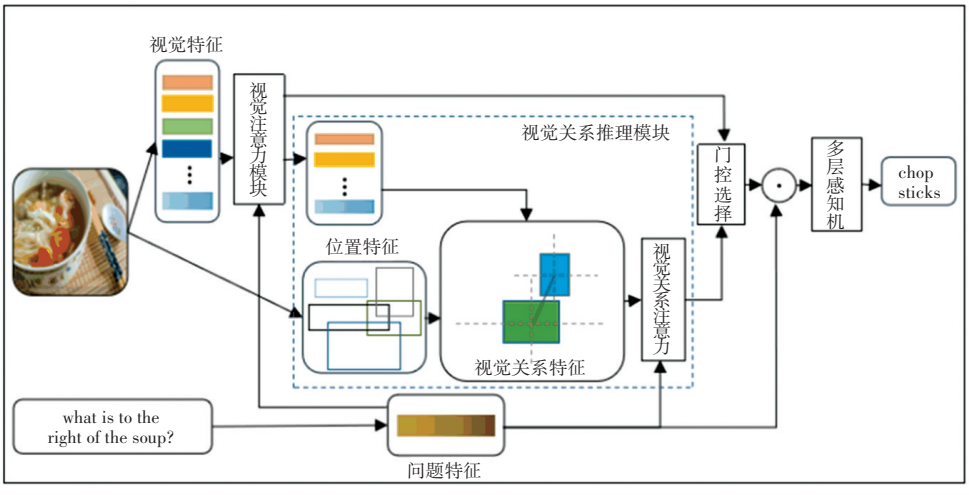

图 1　基于门控机制的联合关系推理视觉问答模型结构图

Fig. 1　Structure diagram of joint relation reasoning visual question answering model based on gating mechanism

对于输入问题，本文用空格与标点将句子分割为单词，从而将问题转化为单词序列，将问题的长度设置为 14 个单词[12]，将多于规定的部分单词剪裁丢弃，少于规定的部分进行补 0 操作，采用预训练的 Glove[13] 方法进行词嵌入，根据文本信息获取具体的词向量，将每个单词转变成 300 维的词向量，将词嵌入序列依次输入长短期记忆网络 LSTM[14] 中，得到一组问题特征向量 $\boldsymbol{x} \in R^{d_x}$。

1.2　视觉注意力模块

注意力机制在视觉问答领域应用已久，并取得了良好的效果。本文通过软注意力机制构成视觉注意力模块，使模型聚焦于与问题相关的视觉区域上。公式表示如下。

$$u = RELU(\boldsymbol{W}_y y_i + b_y) \tag{1}$$

$$t = RELU(\boldsymbol{W}_x x + b_x) \tag{2}$$

$$f = \boldsymbol{W}(u + t) \tag{3}$$

$$\omega = \mathrm{softmax}(f) \tag{4}$$

所有视觉区域的注意力特征矩阵为

$$\boldsymbol{Y} = \sum_{i=1}^{K} \omega_i \boldsymbol{y}_i \tag{5}$$

其中，$\boldsymbol{W}_y$、$\boldsymbol{W}_x$、$\boldsymbol{W}$ 是参数矩阵；b_y、b_x 为偏置量；y_i 为视觉特征输入；x 为问题特征输入。

1.3　视觉关系推理模块

不同视觉对象存在着不同的空间位置关系，不同问题所关注的视觉关系也不同。基于此，本文构建视觉关系推理模块，其主要由视觉特征筛选、视觉关系特征构建和视觉关系注意力 3 部分组成。其中，视觉特征筛选是在视觉注意力模块的基础上，基

于注意力权重筛选出信息度最高的 N 个区域;视觉关系特征构建用于构造每个视觉区域间的位置关系;视觉关系注意力用于推断与问题最相关的视觉关系特征。

1.3.1 视觉特征筛选

通过视觉注意力模块,所有图像特征区域均获得注意力权重,将加权后的视觉注意力特征通过注意力机制选取相关度最高的 n 个视觉特征区域,注意力权重 ω 用于过滤掉小于 $\max(n)$ 的视觉区域。通过筛选的视觉区域 v^n 将用于后续视觉关系注意力,此机制能够有效降低计算复杂度,提高模型效率,增强后续模型的执行效率。相关表达公式如下:

$$\omega_{\max(n)} = \max(\omega[n]) \tag{6}$$

$$\begin{cases} w^i = 1, \omega_i \geqslant \omega_{\max(n)} \\ w^i = 0, \omega_i < \omega_{\max(n)} \end{cases} \tag{7}$$

$$Y^n = wY \tag{8}$$

其中,$\omega_{\max(n)}$ 表示第 n 高的注意力权重数值,只有权重值大于等于 $\omega_{\max(n)}$ 的视觉区域可以通过筛选,用于构建视觉关系特征。

1.3.2 视觉关系特征构建

视觉区域之间的位置坐标需要进行一定的数学运算,用以描述视觉区域之间的空间位置关系。将位置特征 b_i、b_j 进行数学运算来表示关系因子 α_{ij},α_{ij} 表示区域 j 相对于区域 i 的位置关系。将关系因子 α_{ij} 与视觉特征 y_i 级联,表示为视觉关系特征 $\tilde{y}_{ij}$。具体公式如下:

$$\tilde{y}_{ij} = \boldsymbol{W}_{ij}[\alpha_{ij}, y_i] \tag{9}$$

$$\alpha_{ij} = [\frac{e_j - e_i}{w_i}, \frac{f_j - f_i}{h_i}, \frac{w_j}{w_i}, \frac{h_j}{h_i}, \frac{w_j h_j}{w_i h_i}] \tag{10}$$

其中,$\boldsymbol{W}_{ij}$ 为转换矩阵;α_{ij} 为关系因子;$\tilde{y}_{ij}$ 为视觉关系特征。

1.3.3 视觉关系注意力

不同问题关注的视觉区域间关系不同,为了充分挖掘问题与视觉关系间的相关性,本文参考 Yu 等[6]设计的引导注意单元用于问题特征与视觉关系特征的融合,如图 2 所示。其中,多头注意力机制[5]是在缩放点积注意力机制基础上的改进。通过拼接 g 个并行的缩放点积注意力,将拼接结果作为多头注意力的输出,通过位置前馈层获得最终结果。且多头注意力层和前馈层的输出都应用归一化操作和残差连接,加速了模型的收敛。缩放点积基本公式为

$$R_{att}(Q, K, V) = \text{softmax}(\frac{Q K^{\mathrm{T}}}{\sqrt{d_k}}) V \tag{11}$$

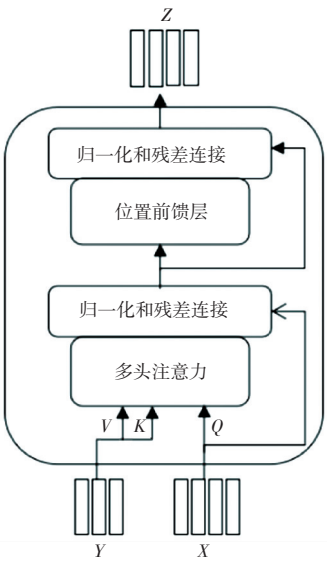

图 2 引导注意力单元

Fig. 2 Guided attention module

本文将训练矩阵 $\boldsymbol{W}_Q$ 与问题特征 x 相乘得到查询 Q,训练矩阵 $\boldsymbol{W}_K$、$\boldsymbol{W}_V$ 与视觉关系特征 $\tilde{y}_{ij}$ 相乘得到 K、V。将上述矩阵通过缩放点积注意力公式,求得每个视觉关系特征 $\tilde{y}_{ij}$ 的注意力特征。

$$R_{att}(\tilde{y}_{ij}) = \text{softmax}(\frac{(\boldsymbol{W}_Q x)(\boldsymbol{W}_K \tilde{y}_{ij})^{\mathrm{T}}}{\sqrt{d_k}})(\boldsymbol{W}_V \tilde{y}_{ij}) \tag{12}$$

多头注意力公式如下:

$$z = \text{MultiHead}(\tilde{y}_{ij}) = Concat(head_1, \cdots, head_g)\ W^{\circ} \tag{13}$$

$$head_t = R_{att}(\tilde{y}_{ij}^t) \tag{14}$$

LayerNorm( )表示进行层归一化。

$$z^1 = \text{LayerNorm}(x + z) \tag{15}$$

$$z^2 = \text{LayerNorm}(z^1 + \text{FFN}(z^1)) \tag{16}$$

FFN( )为位置前馈网络,由两个全连接层组成,公式如下:

$$\text{FFN}(z^1) = \max(0, z^1 W_1 + b_1) W_2 + b_2 \tag{17}$$

1.4 门控选择机制

图像信息由视觉特征和视觉关系特征共同组成,当根据问题给出回答时应根据需要,动态采纳视觉特征和视觉关系特征的比例。应对不需要位置关系推理的问题(如:识别和计数问题),则更依赖于视觉注意力模块,而其他需要依据位置关系回答的问题,则更多依赖视觉关系特征。基于此,本文引入门控机制[15]动态控制二者的输出,将上述两种不同特征映射至同一空间:

$$W^{\text{gate}} = \text{sigmoid}(\boldsymbol{W}_g[\boldsymbol{Y}, Z]) + b_g \tag{18}$$

其中,$\boldsymbol{W}_g$ 是权重矩阵。通过 W^{gate} 融合视觉特

征和视觉关系特征公式如下：

$$M = \boldsymbol{W}_t(\boldsymbol{W}^{\text{gate}} \circ [\boldsymbol{Y}, Z]) + b_t \tag{19}$$

其中，M 是总体视觉表示，$\boldsymbol{W}_t$ 是权重矩阵。

1.5　答案预测

将总体视觉特征 M 和问题特征 x 融合：

$$A = M \circ W_x x \tag{20}$$

将上述结果通过多层感知机（MLP），再执行归一化操作，进行答案预测，上述公式如下：

$$P(a_n) = \text{softmax}(\text{MLP}(A))_n \tag{21}$$

式中：W_x 是训练参数，$P(a_n)$ 是候选答案 a_n 的概率，$a_n \in R^N$（N 代表候选答案的数目），其中选择概率值最大的答案作为最终预测答案。选用二元交叉熵（BCE）作为损失函数对模型进行优化。

2　实验

2.1　数据集及评价指标

本文实验在 VQA V2 数据集上进行，VQA V2 由训练集、验证集和测试集组成。其中，训练集包含图片82 783张和443 757个问题，验证集包含 40 504 张图片和214 354个问题，测试集包含81 434张图片和447 793个问题。问题分为 3 种类型：是否、数字和其他。一张图片大致有 5 个问题，每个问题包含 10 个人工标注答案。

对于开放式任务，由于每个问题包含了 10 个人工标注的答案，即答案不唯一，因此本文采用 Antol[16] 提出的评估机制对答案进行评估，其公式如下：

$$Accuracy(a) = \min\{\frac{\#\text{humans provided that answer}}{3}, 1\} \tag{22}$$

公式表明，只有当预测答案在 10 个人工标注答案中占 3 个以上，才认为完全正确。

2.2　实验细节

本文使用 pytorch 作为深度学习框架，使用 2 个 GTX 1080ti 作为硬件平台。本文选择在训练集中出现 8 次以上的答案形成答案候选集，故在测试集中产生 3 129 个待选答案的答案集。本文视觉特征维度为 2 048 维，LSTM 模块的隐藏层设置为 512，视觉对象过滤机制的过滤参数 $n \in \{4,6,8,10,12\}$，多头注意力头数设置为 3，使用训练批次大小为 128 的 Adam 优化器训练模型。模型训练 50 个周期，学习率采用预热策略，第一个周期学习率设置为 0.001，之后每个周期学习率增加 0.001，直至 10 个周期，此后每 2 个周期学习率衰减一次，衰减率 0.5。为防止梯度爆炸，选取阈值为 0.25 进行梯度裁剪，并在每个全连接层后采用 dropout 策略，dropout 率设置为 0.5。

2.2.1　参数比较

为了说明基于注意力数值的视觉过滤参数 n 对于实验的影响，本文设置视觉对象过滤机制的参数 $n \in \{4,6,8,10,12\}$，在 VQA V2 测试集进行实验，实验结果如图 3 所示。结果说明，当 n 较小时，通过筛选的视觉区域信息不足，无法生成准确答案；随着 n 值增大，虽然模型性能有所增强，但过多的视觉特征对于答案不仅没有增益效果反而造成模型精度下降。这可能是因为过多的视觉区域进入位置关系模型所致，无关信息的增加会削弱有效区域间的关系信息提取。由于图片中与问题相关的视觉区域个数有限，在一定程度减少视觉区域的个数后，模型的训练更加集中于与问题有关的部分，而过滤掉了无关部分，降低了训练强度，提升了模型的性能。经过多次实验，本文选取 $n = 8$。

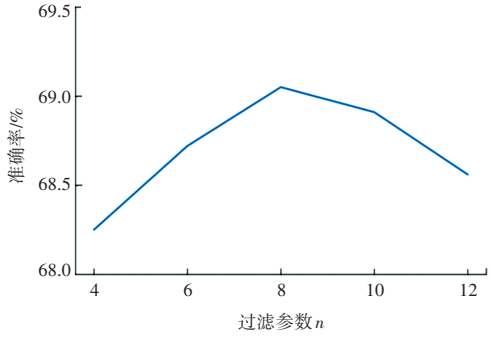

图 3　不同过滤参数 n 下模型的总准确率

Fig. 3　Total accuracy of the model under different filtering parameters

2.2.2　消融实验

VARG 模型由多个模块构成，为了验证不同模块的有效性，进行消融实验。其中统一设定视觉对象过滤参数为 8，在 VQA V2 验证集上进行消融实验。实验结果见表 1。

本文消融实验的对比模型如下：

（1）V-Att（基线模型）：该模型在图推理网络模块仅使用视觉注意力模块输出的特征经过多层感知机进行答案预测，以此为基线模型；

（2）V-Att-Rel-Con（拼接模型）：该模型在图推理模块移除自适应门控，将通过视觉注意力模块的特征和通过位置关系注意力模块的特征进行拼接后，通过多层感知机进行答案预测；

（3）VARG（视觉问答模型）：本文提出的模型。

实验结果表明，增加区域视觉关系推理后，模型性能得到较大提升。说明对于复杂问题只关注视觉

区域而不考虑位置关系是不够的,视觉特征与图像位置关系相结合,可以更好的理解图像内容。同时可以看出,自适应门控机制相比于拼接有更好地性能,门控机制可以动态控制视觉和位置关系特征方面对于答案的贡献,提升总体性能。上述实验表明,本文提出的 VARG 模型中各个模块均发挥了非常重要的作用。

表 1 消融实验结果

Table 1 Ablation results

模型	准确率/%
V-Att(基线模型)	62.58
V-Att-Rel-Con	63.65
VARG(本文模型)	64.46

2.2.3 定性分析

为分析视觉关系推理模块对于模型预测能力的帮助,展示了 VARG 模型和 V-Att 模型对于不同问题的预测,对比结果如图 4 所示。通过比较可以看出,VARG 模型对于位置类问题,(图 4(a))预测的准确度更高;对于综合类问题(图 4(b)),VARG 也给出了更准确的答案,这说明通过视觉过滤,模型能够更好地关注与问题相关的区域,从而做出更准确地预测。

Q: How many people are riding on the elephant?
Bottom-Up: 3 ×
VARG: 2 √

Q: Is the person on the right wearing a hat?
Bottom-Up: Yes ×
VARG: No √

(a) 位置类问题 (b) 综合类问题

图 4 VARG 和 V-Att 模型结果比较图示

Fig. 4 Comparison of VALG and V-Att model results

2.2.4 模型总体性能比较

不同模型在 VQA V2 测试集上的性能对比结果见表 2。可以看出,本文提出的模型在大部分指标上优于目前较为先进的方法。MCB[18] 提出一种压缩双线性池化方法,将双线性多模态融合时产生的高维度进行压缩,进而优化模型性能;Bottom-up[4] 是 2017 年 VQA Challenge 的冠军;QG-SRGR[8] 将视觉对象之间的空间关系属性结构化建模;DCN[19] 提出一种密集连接的共同注意力机制;VSDC[20] 提出了一种视觉语义对偶信道网络,并在视觉通道引入二元及多元推理;本文提出的 VARG 模型通过视觉注意力以及区域视觉位置关系进行联合推导,在模型的准确度上具有一定的竞争性。

表 2 不同模型实验结果对比

Table 2 Comparison of experimental results of different models

模型	开发测试集				标准测试集			
	总体/%	是/否/%	数字/%	其他/%	总体/%	是/否/%	数字/%	其他/%
LSTM+CNN[17]	-	-	-	-	54.22	73.47	35.18	41.83
MCB[18]	-	-	-	-	62.27	78.82	38.28	53.36
Bottom-up[4]	65.32	81.82	44.21	57.10	65.67	82.20	43.90	56.26
QG-SRGR[8]	66.98	82.82	47.68	56.62	67.34	83.27	47.35	58.04
DCN[19]	66.87	83.51	46.61	57.26	67.04	83.85	47.19	56.95
VSDC[20]	68.55	83.79	48.16	59.31	68.67	84.21	48.57	59.64
VARG(本文模型)	69.09	84.62	48.73	59.36	69.31	84.94	49.43	59.72

3 结束语

本文提出的 VARG 模型,是推理视觉对象以及对象间关系的有效模型,模型通过视觉注意力模块聚焦在高相关度的视觉对象上,通过过滤机制让与问题最相关的视觉对象进入视觉关系推理模块,通过视觉关系注意力输出带权重的视觉关系特征,将得到的视觉特征和视觉关系特征输入到自适应门控中,自适应地选择合适的特征,最后通过多层感知机完成对于候选答案的预测。为了验证本文模型的性能,对模型进行了消融实验,验证了模型各部分的有效性,并与当前效果较好的一些模型进行了对比,实验结果表明,本文模型具有一定的竞争力。

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] 白林亭,文鹏程,李亚晖. 基于深度学习的视觉问答技术研究[J]. 航空计算技术,2018,48(5):334-338.

[3] 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.
[4] 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.
[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017：6000-6010.
[6] YU Z, YU J, CUI Y, et al. Deep modular co-attention networks for visual question answering [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach：IEEE, 2019：6281 - 6290.
[7] ZHU Z, YU J, WANG Y, et al. Mucko：multi-layer cross-modal knowledge reasoning for fact-based visual question answering[J]. arXiv preprint arXiv：2006.09073, 2020.
[8] 兰红，张蒲芬. 问题引导的空间关系图推理视觉问答模型[J]. 中国图象图形学报，2022，27(7)：2274-2286.
[9] MA Y, LU T, WU Y. Multi-scale relational reasoning with regional attention for visual question answering[C]//Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021：5642-5649.
[10] 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.
[11] REN S, HE K, GIRSHICK R, et al. Faster R-CNN：towards real time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6)：1137-1149.
[12] TENEY D, ANDERSON P, HE X, et al. Tips and tricks for visual question answering：learnings from the 2017 challenge [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE, 2018：4223-4232.
[13] 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. Doha, Qatar：ACL, 2014：1532-1543.
[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8)：1735-1780
[15] PEI H, CHEN Q, WANG J, et al. Visual relational reasoning for image caption [C]//Proceedings of the 2020 International Joint Conference on Neural Networks. Glasgow, UK：IEEE, 2020：1-8.
[16] ANTOL S, AGRAWAL A, LU J, et al. Vqa：visual question answering [C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile：IEEE, 2015：2425-2433.
[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] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding [J]. arXiv preprint arXiv：1606.01847, 2016.
[19] NGUYEN D K, OKATANI T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：6087-6096.
[20] WANG X, CHEN Q, HU T, et al. Visual-semantic dual channel network for visual question answering [C]//Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021：1-10.

(上接第 137 页)

参考文献

[1] 彭红星，徐慧明，刘华鼐. 基于改进 ShuffleNet V2 的轻量化农作物害虫识别模型[J]. 农业工程学报，2022，38(11)：161.
[2] 徐硕. 融合姿态信息的步态识别算法研究[D]. 合肥：安徽大学，2021.
[3] 周晨怡. 基于融合算法和深度学习的多模态生物特征识别研究[D]. 广州：南方医科大学，2020.
[4] CYGERT S, CZY ŻEWSKI A. Toward robust pedestrian detection with data augmentation[J]. IEEE Access, 2020, 8：136674.
[5] MA N, ZHANG X, ZHENG H T, et al. Shufflenet v2：Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). IEEE, 2018：116.
[6] ZHANG X, ZHOU X, LIN M, et al. Shufflenet：An extremely efficient convolutional neural network for mobile devices[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：6848.
[7] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston. 2015：1.
[8] RAMACHANDRAN P, ZOPH B, LE Q V. Searching for activation functions[J]. arXiv preprint arXiv：1710.05941, 2017.
[9] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets：Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv：1704.04861, 2017.
[10] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2：Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：4510.
[11] 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.
[12] WANG M, LU S, ZHU D, et al. A high-speed and low-complexity architecture for softmax function in deep learning [C]//Proceedings of 2018 IEEE ASIA Pacific Conference on Circuits and Systems (APCCAS). IEEE, 2018：223-226.
[13] HU H. Decay of correlations for piecewise smooth maps with indifferent fixed points [J]. Ergodic Theory and Dynamical Systems, 2004, 24(2)：495.
[14] KINGMA D P, BA J. Adam：A method for stochastic optimization [J]. arXiv preprint arXiv：1412.6980, 2014.
[15] WEN J, LAI Z, WONG W K, et al. Optimal feature selection for robust classification via l2, 1-norms regularization [C]// Porceedings of 2014 22nd International Conference on Pattern Recognition. IEEE, 2014：517.
[16] PAN S J, TSANG I W, KWOK J T, et al. Domain adaptation via transfer component analysis [J]. IEEE Transactions on Neural Networks, 2011, 22(2)：199.