Computer Science and Application 计算机科学与应用, 2019, 9(12), 2403-2410
Published Online December 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.912267

Hans汉斯

Improve Image Question and Answer Accuracy by Using Text Feature Enhancement and Attention Mechanism

Zou Jiang, Murong Jiang*, Chunna Zhao, Yaqun Huang

School of Information and Engineering, Yunnan University, Kunming Yunnan
Email: *jiangmr@ynu.edu.cn

Received: Dec. 5th, 2019; accepted: Dec. 18th, 2019; published: Dec. 25th, 2019

Abstract

Image question and answer is one of the main directions for the successful application of deep learning in the field of computer vision. It has been widely used in artificial intelligence, natural language processing, image recognition and so on. The accuracy of the image question and answer is not only related to the design of the feature fusion module in the image question answering system, but also related to the degree of matching between the image feature and the semantic level of the question feature. In this paper, the text features and visual features of the image are first combined as the enhanced features of the image. Then, the text features are extracted from the question, and then the attention mechanism is added. The enhanced features of the image and the text features of the question are merged, and make answer prediction for fusion features. The experimental results show that the proposed method can solve the problem of mismatch between image features and text features, and improve the accuracy of the image question answering system.

Keywords

Text Feature of the Image, Image Q & A, Attention Mechanism, Feature Enhancement

利用文本特征增强与注意力机制提高图像问答准确率

江　邹，蒋慕蓉*，赵春娜，黄亚群

云南大学信息学院，云南　昆明
Email: *jiangmr@ynu.edu.cn

收稿日期：2019年12月5日；录用日期：2019年12月18日；发布日期：2019年12月25日

*通讯作者。

文章引用: 江邹, 蒋慕蓉, 赵春娜, 黄亚群. 利用文本特征增强与注意力机制提高图像问答准确率[J]. 计算机科学与应用, 2019, 9(12): 2403-2410. DOI: 10.12677/csa.2019.912267

摘 要

图像问答是深度学习在计算机视觉领域成功应用的主要方向之一，在人工智能、自然语言处理、图像识别等方面有着广泛应用。图像问答的准确率不仅与图像问答系统中特征融合模块的设计有关，而且与图像特征与问题特征语义层次匹配程度有关。本文首先将图像的文本特征和视觉特征融合后作为图像增强特征，之后对问题提取文本特征，再加入注意力机制，将图像增强特征与问题文本特征进行特征融合，对融合特征做出答案预测。实验结果表明，本文方法可以解决图像特征与文本特征层次不匹配的问题，提高图像问答系统的准确率。

关键词

图像文本特征，图像问答，注意力机制，特征增强

1. 引言

图像问答是指给定一张图像和一个用自然语言描述的问题，计算机能自主根据图像内容做出相应回答的过程，它是深度学习在计算机视觉领域成功应用的主要研究方向之一。随着人工智能、自然语言处理、深度学习、图像识别等技术的发展，图像问答在汽车导航、盲人识路、机器人系统等领域有广泛应用[1]。

图像问答的实现方式主要采用 CNN-RNN 框架[2] [3] [4]。其中，CNN 为卷积神经网络，主要用于图像特征提取，RNN 为循环神经网络，主要对问题文本特征的提取。由于 CNN 使用了全局图像特征来表示输入图像，会导致一些无关或噪音信息输入到问答模块，对生成答案造成干扰，因此，将注意力机制引入到 CNN-RNN 框架中并与图像问答相结合的方法已成为图像问答系统的主流方法[5]。Li 等人[6]提出基于属性和描述的图像问答并引入注意力机制，将任务拆分为解释和推理两个步骤，首先理解图像的内容，然后根据理解对答案进行推理。Yuan 等人[7]提出基于图像全局－局部特征以及注意力机制的图像文本描述算法，充分利用了图像的全局和局部特征。Liu 等人[8]提出构建联合多图像特征的 Global-Local Fusion 模型来做信息增广，采用混阶注意力模型来提取与问题相关的局部特征信息。Yu 等人[9]提出基于图注意力网络的视觉问答，将注意力机制先后用于图像的一元表达和二元表达上，把图像建模成一个图模型，图注意力模型就是在图像的图结构表达上进行推理。Lin 等人[10]提出多级注意力机制视觉问答模型，基于注意力机制的算法，利用问题的多重文本粒度来融合各种特征。这些方法都是使用 CNN 提取图像视觉特征与问题文本特征直接进行融合，再加入注意力机制生成每个图像区域的权重，视觉特征的不完整以及权重的选取都会导致图像特征与问题特征语义层次的不匹配，影响图像问答的结果。

针对图像特征与问题特征语义层次不匹配的问题，本文分别对图像进行视觉特征提取和图像文本特征提取并将两种特征合并后作为图像的增强特征，然后再对问题提取问题文本特征，采用 MCB 的融合方式进行特征融合，最后模型对融合特征做出答案预测。在问题的特征提取中采用了长短期记忆网络 LSTM，在图像的特征提取中采用 VGG-16 模型和 Neural 模型相结合的方法，提取图像的视觉特征和文本特征，并将两种特征融合作为图像的增强特征与问题文本特征进行 MCB 融合后进行答案预测，提高图像问答系统的准确率。

2. 本文方法描述

图像问答是指用一幅图像和一个与该图像内容相关的问题作为输入，要求系统最终输出该问题的正确答案。本文实现步骤描述如下：将图像传入 Neural 模型生成图像的文本特征，同时传入改进后的 VGG-16 模型生成图像的视觉特征，引入注意力机制并将两者特征融合作为图像的增强特征，然后再将问题的文本特征提取，将两种特征采用 MCB 方式融合并依次通过全连接层和 Softmax 层分类后，得到最终的结果。模型架构如图 1 所示：

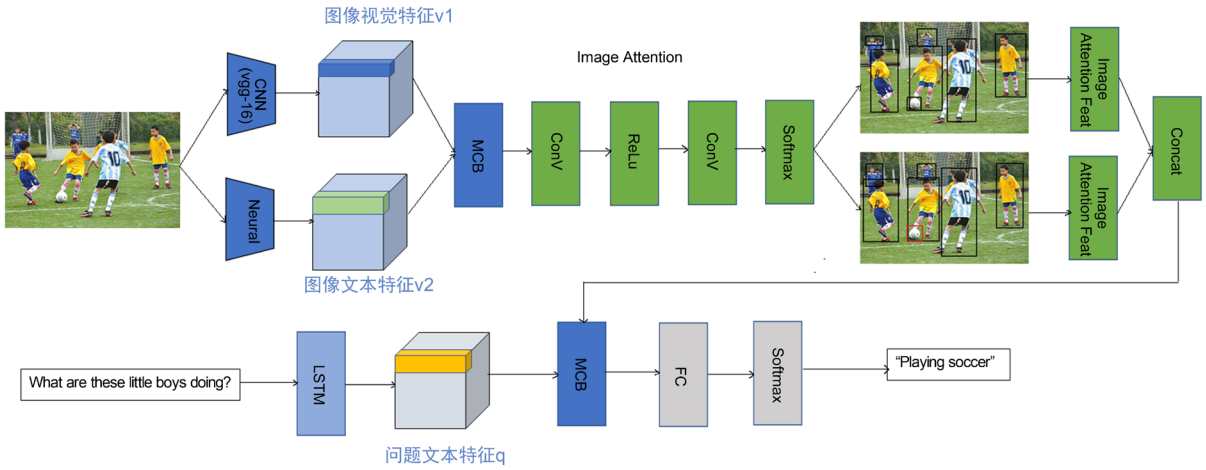

Figure 1. Model architecture
图 1. 模型架构

2.1. 提取图像增强特征

提取图像特征是指将一张以像素形式的图像输入，输出为具有高层语义信息的特征向量 v。作为特征提取器的卷积神经网络都是在 ImageNet 图像识别任务中提出的标准模型，常见的有 VGG-16、VGG-19、GoogLe-Net 以及 ResNet，借助这些 CNN 模型能够间接的利用 ImageNet 上的大量训练数据对图像进行更好的特征提取。本文采用 VGG-16 预训练模型，去掉模型的最后两层，从全连接层提取具有 4096 维的特征，作为图像的视觉特征 $v1$。

图 2 为模型经过第一层处理后得到的卷积和池化特征图。图 3 为模型经过提取具有 4096 维特征的特征图。

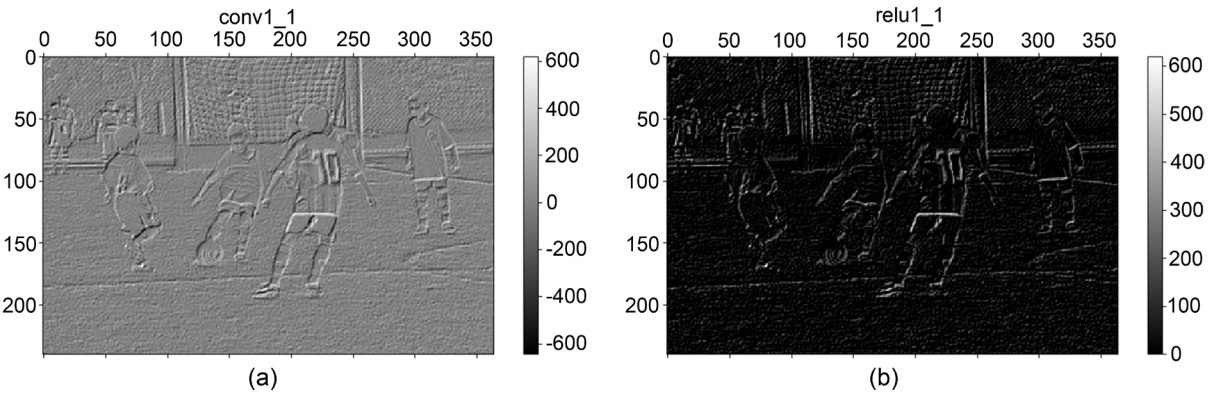

Figure 2. (a) First layer convolution map; (b) First layer pooling feature map
图 2. (a) 第一层卷积特征图；(b) 第一层池化特征图

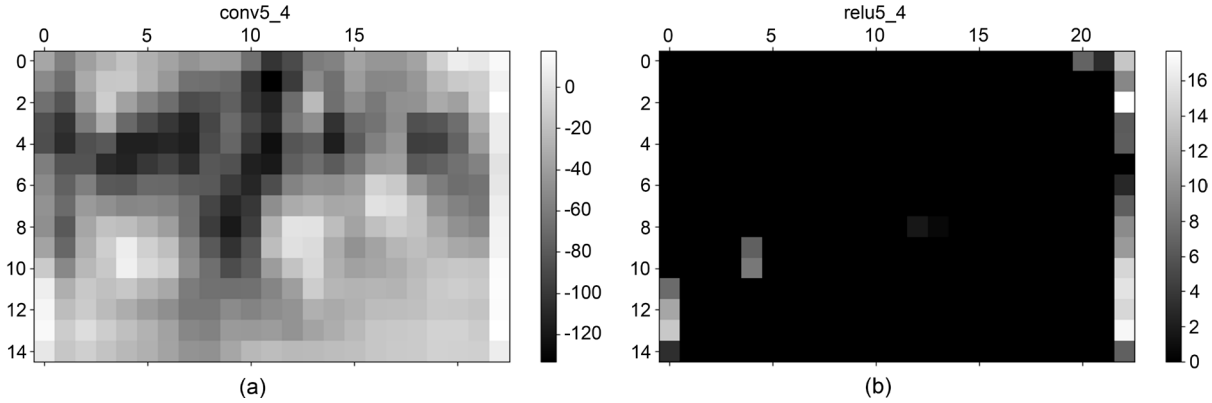

Figure 3. (a) Convolutional feature map; (b) Pooled feature map
图 3. (a) 卷积后的特征图；(b) 池化后的特征图

为获取图像的文本特征。本文采用 Neural 模型来生成图像对应的文本描述。输入一幅图像，采用随机采样得到第一个单词，然后将其输入到长短期记忆网络 LSTM 中得到第二个单词，一直重复操作直到结束符号或者达到预先设定的句子最大长度，这样就得到了图像的文本描述，再利用最大化后验概率对该文本描述进行验证。经过反复学习，最终输出描述图像内容的一句话，作为该图像的文本特征 $v2$。

图 4 为利用 Neural 模型生成的图像文本特征样例。

Figure 4. Text examples of images
图 4. 图像文本样例

将图像视觉特征 $v1$ 与图像文本特征 $v2$ 进行 MCB 融合，即可得到图像特征 v。即利用两个特征的外积来计算：

$$v_1 \otimes v_2 = v_1 v_2^{\mathrm{T}} \tag{1}$$

将图像视觉特征 $v1$ 与图像文本特征 $v2$ 进行外积运算后线性变换 W，得到隐含表达 z，

$$z = W\left[v_1 \otimes v_2\right] \tag{2}$$

再将表达 z 经过卷积/FFT 得到融合后的结果即为图像增强特征 v。

2.2. 提取问题文本特征

提取问题文本特征是指输入一个以英文问句形式输入的问题 $Q = \{w_1, \cdots, w_N\}$，(其中 w_i 是将问题分词后得到的英文单词，N 为问题的长度)，输出为问题的特征向量 q。

首先需要将英文单词表示为词向量。假设 V 是语料库确定的词典，则每个单词首先将会被转换为 $|V|$

维的 one-hot 向量。经过 one-hot 编码后的问题为 $Q^o=\{w_1^o,\cdots,w_N^o\}$，随后每个单词的 one-hot 向量将被嵌入到词向量空间中：

$$Q^e=W_eQ^o \tag{3}$$

其中 W_e 的第 i 列是 V 中第 i 个单词对应的词向量表达。

然后，利用 LSTM 对词向量 Q^e 进行编码，得到一个固定长度的特征向量 q。

图 5 为问题文本特征提取的流程图。

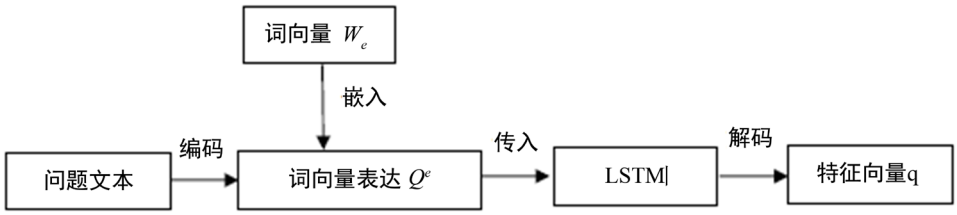

Figure 5. Question feature extraction flow chart
图 5. 问题特征提取流程图

2.3. 将注意力机制加入到图像特征提取中

注意力机制就是从一系列局部特征 $X=[x_1,\cdots,x_n]$ 中寻找与引导条件 g 最相关的部分。加入注意力机制操作可分为两步：第一步是为每个局部特征 x_i 产生一个权重 a_i 用于表明该局部特征与引导特征 g 之间的相关性，然后利用 $S_i=a_i/\sum_j a_j$ 计算出每个 x_i 对应 softmax 值，来表示每个 x_i 获得关注的概率大小；第二步是求所有局部特征的加权和，得到的向量 $\tilde{x}$ 代表根据引导条件 g 在输入特征 X 上的关注结果。

图 6 为在图像特征提取中引入注意力机制得到的焦点图。

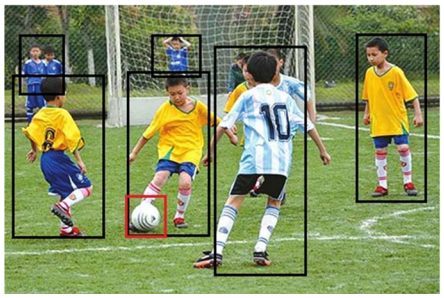

Figure 6. Focus map of interest in the visual attention mechanism
图 6. 视觉注意力机制感兴趣的焦点图

2.4. 将图像特征和问题特征融合

将图像特征 v 和问题特征 q 进行 MCB 融合，得到融合后的特征 m。

2.5. 答案预测

首先根据训练集确定一些比较常见的答案组成候选答案集 A，然后将每个候选答案看作一个类，预测正确答案在候选答案集 A 上的概率分布，最后取概率最大的候选答案 $\tilde{a}$ 作为预测结果，计算公式如(4)所示。

$$\tilde{a}=\max\{P(a\mid I,Q)\} \tag{4}$$

其中，预测答案 a 一般采用多层感知机 MLP 加 Softmax 函数实现。

3. 实验结果分析

3.1. 评价指标

对应每一个问题，假设系统输出的答案为 a，对于该答案是否正确，本文采用的是 VQA 官网[11]给出的评价标准：

$$Acc(a)=\min\left(1,\frac{\sum_1^k \mathbb{I}(a_i=a)}{3}\right)$$

在上面公式中，$a_1,a_2,\cdots,a_k$ 是每个问题的正确标注答案的集合，k 一般取值为 10，预测答案只要与 3 个及以上的标注答案一致，即被认为是正确的。

3.2. 实验数据集

本文采用的数据集为 VQA2.0 [11]，其中训练集有 82,783 张图像，443,757 个问题；验证集有 40,504 张图像，214,354 个问题；测试集有 81,434 张图像，447,793 个问题。在 VQA2.0 中，每一张图像平均有 5 个问题，每个问题有 10 个预选答案。图像、问题和答案之间存在着一一对应的关系。

3.3. 实验结果与分析

图 7 为本文实验的部分结果。

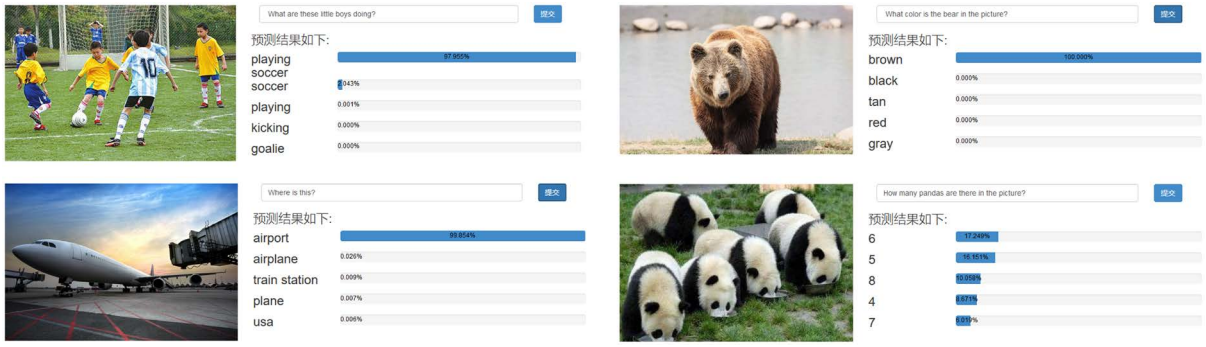

Figure 7. Image question and answer example
图 7. 图像问答实例

为了验证本文方法的有效性，选取 VQA 官网上的几个相关模型与本文模型在 VQA2.0 数据集上进行对比，结果如表 1 所示。

Table 1. Results of each model on the dataset
表 1. 各模型在数据集上的结果

model	yes/no	number	other	overall
Prior [11]	61.26	0.40	1.26	26.13
Language-only [11]	67.88	30.59	29.13	46.21
d-LSTM + n-I [11]	72.23	36.20	39.68	56.27
Our-1	78.82	38.56	51.26	59.86
Our-2	83.37	44.29	55.39	63.45

其中“yes/no”、“number”以及“other”分别对应三种不同答案类型下模型预测答案的正确率，“overall”则是在对应数据集上的总体表现。Prior 表示用训练集上最常见的答案来回答测试集上的问题，Language-only 是仅利用问题对答案去进行预测，采用的是单个 LSTM 架构，d-LSTM + n-I 是一个基础视觉问答模型，Our-1 是未进行图像文本特征提取的模型，Our-2 是把图像文本特征和视觉特征融合作为最终图像特征的模型。从表 1 中可以看出，答案为 yes/no 类型的问题，本文模型准确率在 80%左右，比前面所提到的模型准确率高出 10%左右；答案为 number 类型的问题，本文模型准确率在 40%左右，比前面所提到的模型准确率高出 8%左右；答案为 other 类型的问题，本文模型准确率在 55%左右，比前面所提到的模型准确率高出 15%左右；总体来说本文模型准确率比前面所提到的模型准确率高出 10%左右。Our-1 模型是未提取图像文本特征直接采用图像的视觉特征作为图像特征的模型，与模型 Our-2 相比，各种类型的问题准确率均高出 5%左右，由此可见，本文所提出的模型，利用图像文本特征和视觉特征融合后作为图像的增强特征，图像问答的准确率较高。

4. 总结

本文对图像分别进行视觉特征提取和文本特征提取，并将两种特征融合为最终的图像特征，目的是从多个角度来真正的理解图像中的内容；再加入注意力机制，可以在推理答案的过程中更加关注图像与问题对应部分的信息，而不是将图像和问题进行粗略的处理，进一步加强了图像局部特征与问题特征的相关性，提高了图像问答系统预测答案的准确度。

从实验结果来看，虽然本文使用的模型能达到不错的效果，仍然存在一定不足：

1) 本文方法对于“yes/no”问题的回答效果相较于其它问题来说虽然表现最好，但离一个成熟的图像问答系统来说，回答此类问题的准确率还需要进一步提高。

2) 本文方法对于“number”问题的回答效果较差，这是由于所用的神经网络对于少数或者边界比较模糊或者有遮挡物的物体识别准确率不高导致的，需要对神经网络模型做进一步改进。

基金项目

国家自然科学基金(61862062)和云南省高校科技创新团队支持计划项目(IRTSTYN)。

参考文献

[1] 张天. 用于图像问答的深层注意力网络结构研究[D]: [硕士学位论文]. 云南: 云南大学, 2017.

[2] Malinowski, M. and Fritz, M. (2014) Multi-World Approach to Question Answering about Real-World Scenes Based on Uncertain Input. *Proceedings of the Advances in Neural Information Processing Systems*, Montreal, 8-13 December 2014, 1682-1690.

[3] Gao, H., Mao, J., Zhou, J., *et al*. (2015) Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question. *Proceedings of the Advances in Neural Information Processing Systems*, Cornell University, Ithaca, New York, 2 November 2015, 2296-2304.

[4] Malinowski, M., Rohrbach, M. and Fritz, M. (2015) Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images. *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 1-9. https://doi.org/10.1109/ICCV.2015.9

[5] Wu, Q., Shen, C., Liu, L., *et al*. (2016) What Value Do Explicit High Level Concepts Have in Vision to Language Problems? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 203-212. https://doi.org/10.1109/CVPR.2016.29

[6] 李庆. 基于深度神经网络和注意力机制的图像问答研究[D]: [硕士学位论文]. 合肥: 中国科学技术大学, 2018.

[7] 袁爱红. 图像内容的语义描述与理解[D]: [博士学位论文]. 陕西: 中国科学院大学, 2018.

[8] 刘瑾莱. 基于深层神经网络推理的图像问答技术研究和应用[D]: [硕士学位论文]. 北京: 北京邮电大学, 2019.

[9] 于东飞. 基于注意力机制与高层语义的视觉问答研究[D]: [博士学位论文]. 合肥: 中国科学技术大学, 2019.
[10] 林靖豪. 用于视频问答的多级注意力循环神经网络算法研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2018.
[11] https://visualqa.org/.