基于深度学习的图片问答系统设计研究

周远侠 于 津*
(汕头大学工学院计算机科学与技术系 广东 汕头 515000)

摘 要 对 VQA(Visual Question Answering)数据集进行统计分析,得到相应统计特征,在此基础上提出数据预处理方法:仿聚类法。局部修改 VGGNet 提取的图像特征,与使用 LSTM 获取的问题特征连接后通过多层感知器,连接以 K 个可能输出的 softmax 分类器构成模型 LcVMS。经过低频剔除法与仿聚类法预处理后,LcVMS 在数据集上准确率从 43.21% 提高到 44.45%。实验表明,以 LcVMS 模型为系统应答逻辑的图片问答系统能较好地分辨物体、数量、颜色和位置等信息,在一定程度上可媲美幼儿智商,具备一定的实用价值。

关键词 视觉问答 对话系统 自然语言处理 卷积神经网络 循环神经网络

中图分类号 TP387 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2018.12.038

DESIGN OF IMAGE QUESTION AND ANSWER SYSTEM BASED ON DEEP LEARNING

Zhou Yuanxia Yu Jin*
(*Department of Computer Science and Technology, College of Engineering, Shantou University, Shantou 515000, Guangdong, China*)

Abstract In the paper, the statistical analysis of VQA (Visual Question Answering) data set was carried out, and the corresponding statistical characteristics were obtained. On the basis of that, we proposed one kind of data preprocessing method called imitation clustering. We modified the image features extracted by VGNet locally, connected them with the question features obtained by LSTM, and connected them with K possible output softmax classifiers to form our model LcVMS through multi-layer perceptron. After low frequency rejection and imitation clustering pretreatment, the accuracy of LcVMS on data sets increased from 43.21% to 44.45%. LcVMS model was used as the response logic to construct an image question and answer system. The experimental results show that the system can distinguish information such as the object, the quantity, the color and the position. To a certain extent, it has young children's IQ, and has certain practical value.

Keywords Visual question answering Dialogue system Natural language processing Convolutional neural network Recurrent neural network

0 引 言

近年来,包括深度学习在内的机器学习理论有了巨大进展,人类见证了人工智能在众多领域的研究及应用成果。2015 年,学术界提出的自由形式和开放式视觉问答 VQA 任务[1],逐步成为人工智能研究的热门方向。VQA 系统将图像与自由形式和开放式的自然语言表述问题作为输入,产生的自然语言表述答案作为输出。VQA 任务需要具有精准识别、物体检测、活动识别、知识库推理和常识推理等功能的问答系统来完成,而这些功能所涉及的计算机视觉、自然语言处理和知识推理等领域在过去 10 年中取得了显著的进步。

图片问答聊天机器人涉及的领域主要有视觉问答、文本问答和图像处理。

计算机视觉问答的研究始于 2014 年,一开始的研究,其设定和数据集都比较有限[2-3]。例如,文献[2]只考虑答案来自 16 个基本颜色或 894 个对象类别的预先设计的闭合问题。文献[3]考虑从对象、属性、对象之间的关系来构建固定词汇表的模板进而生成的问

收稿日期:2018-05-31。**周远侠**,硕士,主研领域:机器学习,深度学习。**于津**,副教授。

题。相比之下,文献[1]在一年后提出的任务涉及人类的开放式,自由形式的问题和答案,增加了提供正确答案所需的知识的多样性和推理的种类[1]。从数据集上看,后者[1]的数据集(问答对 36 万,图片 12 万)比前两者[2-3](分别为 2 591 和 1 449 图像)大两个数量级,这对在视觉问答这个更为困难和无约束的任务上取得成功很重要。视觉问答的问题是开放式的,然而,了解问题的类型和哪些类型的算法能更好地回答问题也很有必要。为此,后者[1]还分析所提问题的类型和提供的答案类型,通过几种可视化展示了所提问题的惊人的多样性,并探讨问题的信息内容和答案与图像说明的区别。他们设计出一个将处理问题的 LSTM 与处理图像的卷积神经网络相结合以产生答案的模型[1],并以几种使用文本和最先进的视觉特征相结合的方法作为基准,对类似的模型进行了评估。除此以外,其他相关工作有:文献[4]最早设计出一个将处理问题的 LSTM(Long short-term memory)网络与处理图像的卷积神经网络相结合以产生答案的模型。文献[5]引入了 10 k 图像的数据集,并提出了描述场景的特定方面的说明。百度的 Gao 等[6]收集了 COCO 图像的中文问题和答案。文献[7]使用微软的 COCO 数据集说明自动生成对象、计数、颜色、位置四种类型的问题。

从最新(2016 年、2017 年)发表的文献上看,有些学者已经开始尝试将“带有注意力”的模型加入到现有的视觉问答研究中,例如,文献[8-11]使用了基于视觉注意的模型,其中注意机制通常产生突出显示与回答问题相关的图像区域。这几篇文献都专注于识别“往哪里看”或“视觉注意区域”的问题。文献[12]认为,除了建模“往哪里看”或“视觉注意区域”之外,同样重要的是要模拟“听问题的重点”或“提问注意”。因此他们提出了一个新颖的 VQA 的共同关注模式,同时关注问题和图像的“重点区域”。主要通过新颖的一维卷积神经网络(CNN)以分层方式进行改进。然而,在引入了“带有注意力”的机制后,模型的复杂度会提高不少。

基于文本的问答在自然语言处理和文本处理领域是一个很好的研究问题。文献[13]中让机器回答阅读理解多项选择问题,试图解决开放域的机器理解问题;文献[14]合成了文本描述和 QA(问答)对。这些方法为视觉问答的研究提供了灵感。视觉问答的自然基础是图像,即需要理解文本(问题)和视觉(图像)。由于关于图片提出的问题是由人类产生的,因此常识知识和复杂推理也显得很有必要。

图像处理的相关技术为视觉问答提供一定的支持和借鉴,比如图像标记[15-16]与图像说明[17-18]。和视觉问答相比,这些任务虽然需要视觉和语义知识,但是说明通常不具有针对性[18]。相比之下,视觉问答中的问题往往需要详细的有针对性的图像信息,所以和一般的图像标记与图像说明不一样。

我们在对数据集统计分析的基础上提出数据预处理方法仿聚类法,建立合适的 LcVMS 模型,并以此设计出图片问答系统。这个图片问答系统需要精准识别、物体检测、活动识别、知识库推理和常识推理等多种 AI 功能,对人工智能的学术研究有积极意义;从工业角度上看,一个成熟的图片问答系统,能协助视觉障碍用户积极获取视觉信息。因此,以 VQA 这个目标驱动型任务为导向,以深度学习为基础,研究图片问答聊天机器人的系统设计,既有理论研究意义,也有实际应用价值。

1 框架设计

1.1 任务描述

最近几年,计算机视觉 CV(Computer Vision)、自然语言处理 NLP(Natural Language Processing)和知识表示与推理 KR(Knowledge Representation & Reasoning)快速发展,越来越多的学者投入到上述学科的交叉任务研究中。2015 年,Aishwarya 等提出了自由形式和开放式视觉问答 VQA 的任务:给定图像和相应的以自然语言表述的自由形式的、开放式的问题,智能系统响应以自然语言表述的准确的问题答案。这里最需要强调的一点就是,问题和答案都是开放式的,不加任何限制,视觉问题可以选择性地针对图像的不同区域,包括背景细节等。因此,一个理想的系统通常需要具备比生成通用图像标题的系统更详细地理解图像和进行复杂推理的能力。我们也是基于这样的任务目标设计图片问答系统。

1.2 图片特征提取模块

在 2015 年 ICLR(International Conference on Learning Representations)会议上,Karen Simonyan 和 Andrew Zisserman 提出了 VGG16 和 VGG19 两个非常好的深层卷积神经网络 DCNN(Deep Convolutional Neural Network)[19]。我们模型的图片特征提取模块选择的网络是 VGG16。网络所有卷积层有相同的配置,卷积核大小均为 3×3,步长为 1,填充为 1;共有 5 个最大池化(max pooling)层,大小都为 2×2,步长为 2;卷积层的

通道数目,或者称为宽度从 64 开始,每次经过一个最大池化层翻倍,一直到 512 为止;共有三个全连接层,前两层都有 4 096 通道,第三层共 1 000 路及代表 1 000 个标签类别;除了最后全连接的 softmax 层外,其他层都需要使用整流线性单元(ReLU)非线性激活函数。由于我们只用这个 VGG 网络来提取图片的特征,而不是做图像分类任务,因此最后的 softmax 层就要删除,这样修改后的 VGG 网络如图 1 所示。

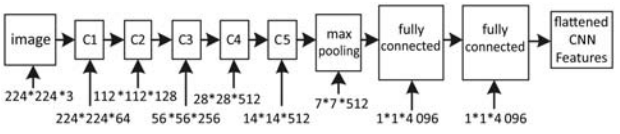

图 1　修改后的 VGG 网络

VGG 网络在最后一个隐藏层加入 L2 正则化,修改后的 VGG 网络输入是一个图片,输出则是一个该图片的"数字化表达形式":为了与后边模型衔接,其后再加一层平整化(flatten)得到一个包含 4 096 个元素的一维的数组,我们以后将这个数组称之为卷积神经网络提取的图片特征,简称 CNN Features。

1.3　问题特征提取模块

由于文字不能直接作为神经网络的输入,这里需要选择一种文字的数字化表达形式。我们选用的 word2vec 词向量表示,是一种适合机器学习、特别是深度学习的输入和表示空间的语言模型。这里可以简单地把 word2vec 当成一种把词语变成向量表示的方法,具体原理不做深入讨论。

对于问题的处理,我们尝试两种不同的方式:

模型 1:普通的词袋模型(Continuous Bag of Words Model):首先利用自然语言处理 NLP(Natural Language Processing)的 word2vec 技术,将问题中的每个单词先转化为一个 300 维的向量(vector),然后将所有词语的向量相加。此处将获得一个 300 维的向量作为 Question 的数字化表达。

模型 2:LSTM 模型:首先利用 word2vec 技术,将问题中的每个单词先转化为一个 300 维的向量(vector),这一步和模型 1 一样。然后按照句子中单词排列的顺序将每个单词依次输入带有一个隐藏层的 LSTM 网络,将网络输出的 512 维向量作为问句的数字化表达。

模型 1 和模型 2 的主要区别在于:模型 1 直接将各词的词向量表示加和,作为整个问句的向量表示,是一种平均化的方法。这种方法完全不考虑词在句子中出现的顺序,类似于"把词扔进一个袋子里",所以叫词袋模型。模型 2 用的 LSTM 模型是 RNN 的一种,词语在问句中的顺序会直接影响输入的顺序,进而影响问句的特征表达。

1.4　训练模块结构

神经网络设计的一个关键是确定结构。结构(architecture),是指网络作为一个整体,包含多少单元,以及这些单元之间是如何连接的。

1989 年,通用近似定理(universal approximation theorem)提出,一个前馈神经网络如果具有线性输出层和至少一层具有任何一种"挤压"性质的激活函数的隐含层,只要给予网络足够数量的隐含单元,就可以任意的精度来近似任何从一个有限维空间到另一个有限维空间的 Borel 可测函数[20]。这里不展开讨论这个定理的具体内容,只引用一个结论:通用近似定理表明,无论需要神经网络学习什么函数,一个足够大的多层感知机 MLP(multilayer perceptron)一定能够表示这个函数。

结构设计除了考虑网络的神经元数量以外,还需要考虑层与层之间如何连接。默认的神经网络层会采用矩阵描述的线性变换,每个输入单元连接到每个输出单元,这就是所谓的全连接。也有某些网络不使用全连接,通常,学术界无法对通用神经网络的结构给出更具体的建议,需要具体问题具体分析。这里只简单阐述模型 1 与模型 2 的结构,如图 2、图 3 所示。

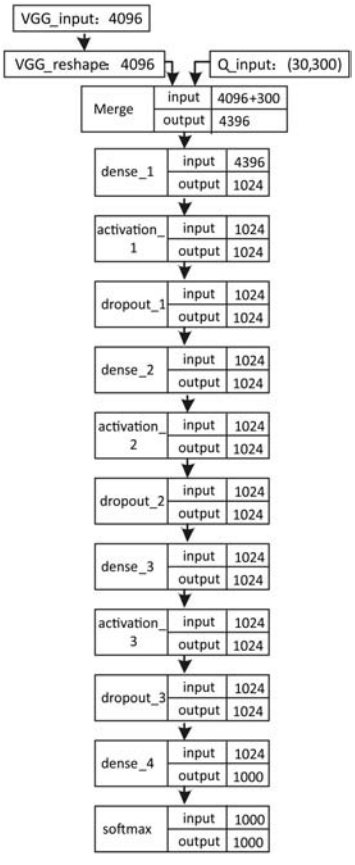

图 2　模型 1 结构图

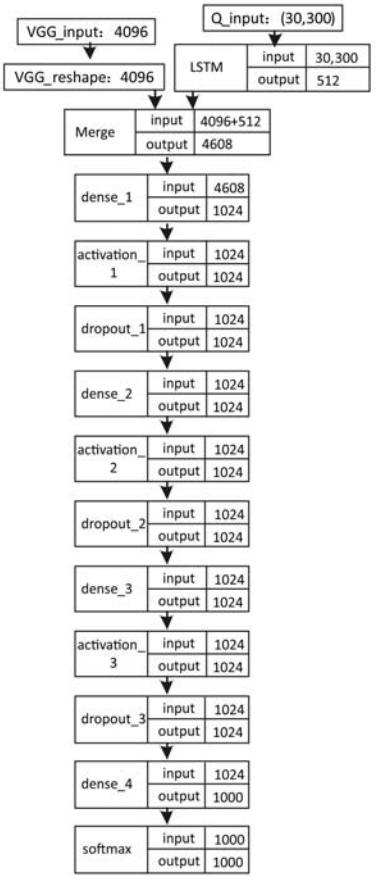

图 3 模型 2 的结构图

1.5 分类模块

输出单元的选择与损失函数的选择紧密相关，通常损失函数使用数据分布和模型分布的交叉熵，而表示输出的形式决定了交叉熵函数的形式。输出层的作用是将隐藏层提供的特征进行变换以完成整个网络的任务，模型的任务是，把“看到一个图片回答开放式问题”转化成“从 K 个最有可能的回答中，找到最符合图片的答案”，这样就把一个开放的 AI 问题，转化成了 multi-class 的分类问题。softmax 函数是一个用来表示一个具有多个可能取值的离散型随机变量分布的函数，常用来作为分类器的输出，因此这里选择 softmax 作为分类器，以获得出现频率最高的 K 个答案的分布。整个模型最后会以交叉熵作为损失函数实现端对端学习。

2 实验分析

2.1 数据集统计分析

2.1.1 数据来源

用于模型训练和分析的数据，来自 VQA（Visual Question Answering）组织公布的数据集，具体下载地址在 http://www.visualqa.org/vqa_v1_download.html。我们使用 V1 版本的数据集，这个数据集包括：微软的 COCO（MSCOCO）数据集，123 287 张图片，其中训练集图片 82 783 张，测试集图片 40 504 张；图片对应的 Question 共 369 861 个，其中训练集 248 349 个，测试集 121 512 个；每个 Question 对应的 Answer 有 10 个，经过数据预处理后[1]，训练集和测试集的数量与 Question 相同（一个 Question 可能对应多个 Answer，这不会影响数据集个数），即训练集 248 349 个，测试集 121 512 个。

2.1.2 统计与分析

1）Question 部分的统计：

（1）设阈值 word_count_threshold = 1 000，出现次数超过 1 000 次的单词，认为是高频词；其余为低频词。其中前 20 个高频词出现次数如表 1 所示。高低频词个数、个数比例、出现次数、出现次数占所有词出现次数比例如表 2 所示。

表 1 前 20 个高频词出现次数

单词	次数	单词	次数	单词	次数
?	320 161	a	45 681	man	18 392
the	225 976	on	41 629	does	14 668
is	200 545	how	40 158	people	13 492
what	118 203	many	38 230	picture	12 518
are	76 624	color	37 322	s	11 779
this	64 512	of	37 023	to	11 758
in	49 209	there	29 182		

表 2 高低频词个数与词频统计

	个数	个数比例	出现次数总和	出现次数比例
总词数	14 770	100%	2 284 620	100%
高频词	208	1.41%	1 890 568	82.75%
低频词	14 562	98.59%	394 052	17.25%

（2）对每个 Question，在前 k 个词出现的前提下第 $k+1$ 个词出现的比例进行统计，其中 XXX 为低频词标记；统计结果如表 3 所示。

表 3 Question 前 k 个词出现的前提下第 $k+1$ 个词出现的比例

What：117327/320029（36.66%）	is：38090/117327（32.46%）	the：22268/38090（58.46%）
		on：3215/38090（8.44%）
		in：2917/38090（7.66%）
		this：2495/38090（6.55%）
		XXX

续表 3

What: 117327/320029 (36.66%)	color: 33251/117327 (28.34%)	is:25770/33251 (77.50%)
		are:5620/33251 (16.90%)
		XXX
	kind: 6828/117327 (5.82%)	of:6816/6828 (99.82%)
		XXX
	are: 6089/117327 (5.19%)	the:4055/6089 (66.60%)
		XXX
	type: 4746/117327 (4.05%)	of:4724/4746 (99.54%)
		XXX
	sport: 2765/117327 (2.36%)	is:2140/2765 (77.40%)
		XXX
	room: 1781/117327 (1.52%)	is:1503/1781 (84.39%)
		XXX
	animal: 1726/117327 (1.47%)	is:1574/1726 (91.19%)
		XXX
	does: 1690/117327 (1.44%)	the:4055/6089 (66.60%)
		XXX
	number: 1115/117327 (0.95%)	XXX
	game: 1035/117327 (0.88%)	XXX
	XXX	XXX
Is: 94094/320029 (29.40%)	the: 36732/94094 (39.04%)	man:4329/36732 (11.79%)
		woman:1750/36732 (4.76%)
		person:1389/36732 (3.78%)
		cat:1375/36732 (3.74%)
		dog:1250/36732 (3.40%)
		XXX
	this: 29398/94094 (31.24%)	a:12875/29398 (43.80%)
		an:1582/29398 (5.38%)
		person:1429/29398 (4.86%)
		man:1318/29398 (4.48%)
		XXX

续表 3

Is: 94094/320029 (29.40%)	there: 13109/94094 (13.93%)	a:7829/13109 (59.72%)
		XXX
	it:6308/94094 (6.70%)	XXX
	that:2017/94094 (2.14%)	a:1297/2017 (64.30%)
	he:1992/94094 (2.12%)	XXX
	she:1050/94094 (1.12%)	XXX
	XXX	XXX
How: 39911/320029 (12.47%)	many:37804/39911 (94.72%)	people:6124/37804 (16.20%)
		animals:1164/37804 (3.08%)
		XXX
	XXX	XXX
Are: 27322/320029 (8.54%)	the:9067/27322 (33.19%)	XXX
	there:6829/27322 (24.99%)	any:2213/6829 (32.41%)
		XXX
	these:4911/27322 (17.97%)	XXX
	they:2439/27322 (8.93%)	XXX
	all:1299/27322 (4.75%)	XXX
	XXX	XXX
Where: 6998/320029 (2.19%)	is:4619/6998 (66.00%)	the:3704/4619 (80.19%)
		XXX
	are:1595/6998 (22.79%)	the:1240/1595 (77.74%)
		XXX
	XXX	XXX
Does: 10243/320029 (3.20%)	is:4619/6998 (66.00%)	the:3704/4619 (80.19%)
		XXX
	are:1595/6998 (22.79%)	the:1240/1595 (77.74%)
		XXX
	XXX	XXX
Do: 4098/320029 (1.28%)	you:1478/4098 (36.07%)	XXX
	XXX	XXX
	the:1059/4098 (25.84%)	XXX
	XXX	XXX
Which: 3513/320029 (1.10%)	XXX	

续表 3

Can:2551/320029(0.80%)	you:1430/2551(56.06%)	
	XXX	
Why:1752/320029(0.55%)	is:1006/1752(57.42%)	
	XXX	
Who:1718/320029(0.54%)	XXX	
Has:1556/320029(0.49%)	XXX	
Was:1366/320029(0.43%)	XXX	
Would:1297/320029(0.41%)	XXX	
Could:1273/320029(0.40%)	XXX	
Did:1246/320029(0.39%)	XXX	
XXX	XXX	

2) Question 部分的分析:

(1) 由表 1 - 表 2 可以看出,出现次数大于 1 000 次的高频词只有 208 个,所占比例非常小,只有 1.44%,超过 98% 的单词出现的次数不足 1 000;高频词出现次数占所有词出现次数的比例为 82.75%,即出现频率超过 80% 的单词不到 1.5%。

(2) 由表 3 可以看出,Question 的分布非常不均匀,比例非常小的一部分 Question 问的次数非常多,而大部分 Question 出现的频率都比较低。

3) Answer 部分的统计:

(1) 对 Answer 的词频进行统计,出现最多的前 20 个 Answer 及其次数如表 4 所示。

表 4 出现最多的前 20 个 Answer 及其次数

单词	次数	单词	次数	单词	次数
yes	86 619	blue	4 974	tennis	1 663
no	54 664	4	3 808	baseball	1 524
2	11 941	green	3 714	right	1 516
1	6 991	black	3 436	orange	1 484
white	6 756	yellow	2 785	6	1 406
3	6 488	brown	2 526	left	1 390
red	5 318	5	2 196		

(2) 选取出现次数最多的 1 000 个 Answer,统计对应的 Question 数量,为 320 029 个,占 Question 总数量的 320 029/369 861 = 86.53%。

4) Answer 部分的分析:

出现次数前 5 个 Answer 占所有 Answer 出现次数的 (86 619 + 54 664 + 11 941 + 6 991 + 6 756)/369 861 = 45.14%;出现次数最多的 1 000 个 Answer 占所有 Answer 出现次数的 86.53%:可见 Answer 的分布是十分不均衡的,而且分布很像一个长尾分布,但是只需要 1 000 个 Answer 就可以覆盖超过 85% 的 Question。

2.2 数据预处理

通过前一章的数据分析,我们提出一个数据预处理方法:仿聚类法。

首先统计 Question 里出现的单词词频,得到“高频词”和“低频词”;然后模仿聚类和分类算法的思想,以词频为密度、以问句起始词为类目对问句进行“聚类”和“分类”,在每个类里又使用相同的方式,以词频为密度、以问句第二个词为类目再次进行归类,以此类推,最后会将样本从原空间映射到新空间,直接合并低频样本。

例如:一个经过处理后的样本,低频词汇被 xxx 替代,样本会变成:question = ['why','does','the','player','have','one','xxx','xxx','up','?'],这是从改变样本着手提高模型准确率的方法,简称为“仿聚类法”。

2.3 模型搭建

模型搭建如下:

1) CcVMS 模型:

- 图像通道:VGGNet 去除最后的 softmax 层,将输出的 4 096 维矩阵展开成 4 096 维向量,简称 CNN Features;VGGNet 参数被固定为 ImageNet 分类而学习的参数,并且在图像通道中不作调整。
- 问题通道:普通的词袋模型:首先利用自然语言处理 NLP(Natural Language Processing)的 word2vec 技术,将问题中的每个单词先转化为一个 300 维的向量,然后直接将各词的词向量表示加和,作为整个问句的向量表示。此处将获得一个 300 维的向量作为问句的数字化表达。
- MLP 层:将图像通道的 4 096 维向量和问题通道的 300 维向量拼接得到 4 396 维向量,作为多层前馈神经网络(MLP)的输入。多层前馈神经网络包括三个全连接层,每个全连接层包含一个全连接级、一个激活函数级和一个 Dropout 级,然后连接一个全连接级,将输出的 1 024 维转化为 1 000 维,最后作为 softmax 分类器的输入。
- 输出层:softmax 分类器,输出维度同样是 1 000。

整个模型以交叉熵作为损失函数实现端对端学习。

2）LcVMS 模型:

• 图像通道:VGGNet 去除最后的 softmax 层,将输出的 4 096 维矩阵展开成 4 096 维向量,简称 CNN Features;VGGNet 参数被固定为 ImageNet 分类而学习的参数,并且在图像通道中不作调整。

• 问题通道:同样使用 word2vec 技术,将问题中的每个单词先转化为一个 300 维的向量,然后按照句子中单词排列的顺序将每个单词依次输入带有一个隐藏层的 LSTM 网络,将网络输出的 512 维向量作为问句的数字化表达。

• MLP 层:将问题通道的 300 维向量作为 LSTM 的输入,得到 512 维向量的输出,和图片通道的 4 096 维向量拼接得到 4 068 维向量,作为多层前馈神经网络的输入。这里采取拼接而不是点乘或者其他融合,是为了能保持相对原始的信息。多层前馈神经网络的结构设计和模型 1 相同,只是第一个全连接层的输入维度不一样。

• 输出层:softmax 分类器,输出维度同样是 1 000。整个模型以交叉熵作为损失函数实现端对端学习。

2.4　结果分析

衡量不同模型性能的指标为准确率,表 5 为实验对比结果。

表 5　不同模型仿聚类法处理前后准确率对比

数据预处理方式	CcVMS 模型	LcVMS 模型
不处理	30.97%	43.21%
仿聚类法	31.59%	44.45%

由于原始数据集的不同 Answer 数目超过 10 000 个,这里我们选取出现次数最多的 $K=1\ 000$ 个作为最后分类输出;大部分 Answer 的出现频率极低,如果 K 太大,低频样本训练不充分,不容易分对,而且可能对高频样本产生干扰,导致高频样本分错;去除这部分低频样本,效果会更好。前边数据集分析提到,Answer 的分布是十分不均衡的,而且分布很像一个长尾分布,但是只需要 1 000 个 Answer 就可以覆盖超过 85% 的 Question,因此 softmax 的输出设计为 1 000 维是合理的。

仿聚类法在样本端对低频样本的输入进行改变,让某些低频样本合并;仿聚类法可以认为是对样本映射到另一个空间,高频的样本从原空间映射到新空间,基本保持不变,而低频样本会进行合并,多个低频样本会映射到新空间的同一个位置;映射到新空间同一个位置的这些低频样本的类标往往不一样,它们的类标最后可能变成一个高频样本的类标,也有可能维持原状;对于前者,这些低频样本会合并到新空间中相近的高频样本;后者则会直接被剔除。

另外,在经过仿聚类法处理的 LcVMS 模型中,真正率 TPR(True Positive Rate)或称为灵敏度(sensitivity)为 57.38%。

如前文所述,出现次数最多的 1 000 个 Answer 占所有 Answer 出现次数的 86.53%,准确率最高的模型选取了 $K=1\ 000$,覆盖了 86.53% 的问题答案,另外还有 13.47% 的问题是没有答案的,也就是说,这些低频问题的无论选择 1 000 个答案里边的哪一个,都不是正确的。

真正率的计算公式是:

$$\text{真正率}=\frac{\text{正样本预测结果数}}{\text{正样本实际数}} \tag{1}$$

真正率计算的是,在属于出现次数最多的 $K=1\ 000$ 个 Answer 里边的样本中,分类正确的比例。

3　图片问答系统

通过对数据集进行分析,搭建模型,我们训练出了 LcVMS 模型,在测试集上准确率达到 44.5%。这里将以模型 LcVMS 为系统应答逻辑构建了图片问答系统。

3.1　系统设计

系统分为输入端、应答逻辑、存储器后端和输出端四大模块,这样满足功能模块化,并且架构清晰,模块之间解耦;每个模块可以选择不同的部件,这些部件都是可插拔的,可以随时更换。具体如图 4 所示。

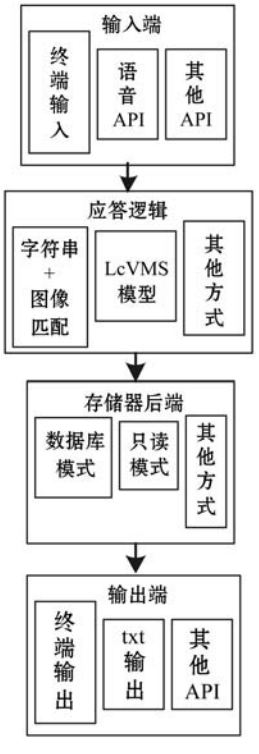

图 4　系统模块设计

输入端:输入端可以使用合适的 API,比如网页前端上传图片,或者语音 API 输入 Question;这里我们使用简单的终端输入,输入包括图片名在内的图片文件本地存放路径与人工提问的 Question,系统会读取图片与 Question。

应答逻辑:应答逻辑功能同样可以设计不同的部件来承担。

方案 1 使用字符串与图片的模糊匹配:当输入图片时,使用修改的 VGGNet 提取 CNN Features,并与数据集所有图片提取出来的 CNN Features 作比较,用 4 096 维的向量距离计算匹配度;当输入 Question 时,数据集的所有 Question 用编辑距离(Edit Distance)计算匹配度;图片和 Question 计算出来的匹配度与设定的匹配度阈值作比较,大于阈值,使用最接近的图片和 Question 对应的回答,小于匹配度,则返回一个"安全的回答",比如"I don't know"。

方案 2 使用 2.3 节搭建的 LcVMS 模型,输入的图片使用修改的 VGGNet 提取 CNN Features;输入的 Question 由 word2vec 转化为数字化表达,并通过 LSTM 提取序列特征;两个特征拼接后通过 MLP,最后由 softmax 返回 Answer。

方案 1 需要事先将所有数据集的图片特征提取出来并保存,当输入一个新图片时,提取的新图片特征必须与所有数据集的图片特征进行一次匹配度计算,同时输入的 Question 也必须与数据集的每个 Question 计算一次编辑距离,然后计算匹配度。数据集每个图片与 Question 的匹配度相加后,还要找出最符合的那一个图片和 Question,返回相应的 Answer。如果数据集每个图片与 Question 的匹配度相加后都没有超过阈值,那么应答逻辑返回一个"I don't know"。很明显,第一种方案的计算量非常大,返回 Answer 的速度非常慢。并且实验表明,准确率比较低。这是因为高频问题匹配度高的样本,往往也是某一类 Question 对应 Answer 概率最大的样本。比如,当 Question 是"How many"一类的问题时,由于数据集样本中"2"这个 Answer 出现频率大于其他数字,这就导致 Answer 为"2"的样本匹配度高的概率最大,因此每次提问"How many"这类问题时应答逻辑都会回答"2"。对于低频的问题,由于数据集中所有样本的匹配度都小于阈值,这会使应答逻辑返回"I don't know"这样一个没有任何意义的 Answer。

方案 2,LcVMS 模型提取输入的图片特征与 Question 序列特征后,只需要通过神经网络的计算就可以返回 Answer。由于 LcVMS 模型的全连接层少,结构也不复杂,因此应答逻辑的响应比较快,并且有较高的准确率。这里我们选用 LcVMS 作为应答逻辑的部件。当然,假如以后有更合适的模型,我们可以替换这个部件,不影响其他模块。

存储器后端:存储器后端也可以选用不同的部件。比如,当希望系统有学习能力时,我们可以选用数据库模式,每次输入图片与 Question,应答逻辑返回 Answer 后,我们人工给系统一个反馈,告诉系统 Answer 是否正确,如果不正确,人工输入一个正确的 Answer。数据库模式可以把图片、Question 与正确的 Answer 一起记录下来,扩展原来的数据集。当应答逻辑选用字符串与图片模糊匹配的方式时,也需要数据库模式来存储数据集的样本,才能计算匹配度。我们搭建的模型 LcVMS 是预先训练好的,所以已经不需要数据集,因此我们这里只选用最简单的只读模式作为存储器后端,即存储器后端不做任何操作。

输出端:输出端我们选用简单的终端输出,即把应答逻辑产生的 Answer 通过终端输出;如果有合适的 API,比如语音合成 API 等,以后也可以替换。

这样一个完整的图片问答系统,就搭建完成了,具体模块与选用部件如表 6 所示。

表 6 具体模块与选用部件

模块	输入端	应答逻辑	存储器后端	输出端
选用部件	终端输入	LcVMS 模型	只读模式	终端输出

3.2 系统展示评述

下面对图片问答系统的智能程度进行展示与评估,输入的图片是随机从互联网下载的图片。测试结果如图 5 - 图 9 所示,"Ask a question"后的字体为人工输入的提问,方括号内的字符是系统的输出,详细 QA 对正确与否如表 7 所示。

图 5 一个穿红衣白裤的男人在玩飞盘

图 6　两个男人在踢足球

图 7　一张桌子上边放着书和台灯

图 8　一个穿白背心的女人在打网球

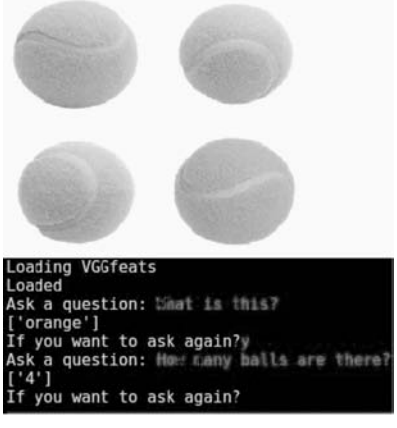

图 9　四个网球

表 7　测试 QA 对

图片	Question	Answer	是否正确
图 5	What is he doing?	frisbee	正确
	What color is his clothing?	white	大致正确
	Is he playing football?	no	正确
图 6	What are they doing?	soccer	正确
	How many men are there?	2	正确
图 7	Where is the lamp?	on desk	正确
图 8	What is she doing?	tennis	正确
	What color is her shirt?	white	正确
图 9	What is this?	orange	错误
	How many balls are there?	4	正确

从以上 5 个简单测试可以看出,图片问答系统具有一定智能,在一定程度上达到了幼儿智商。首先,问答系统可以识别出应该回答什么:是回答“Yes”或者“No”,还是回答数量、颜色、位置;其次,问答系统回答的准确率看起来要高于 44.45%,这是因为上边人工提问的这些 Question 没有特别古怪或者特别难的问题,只是询问物体、数量、颜色或者位置,而数据集里边的 Question,会更多样化一些,详情可以参照 2.1 节的数据统计与分析。

另外可以看出,系统的回答都很简短,一般都是一个词,少数会用到两个词,这是因为训练集里边的 Answer 非常简短。对于图片问答,Question 一般都是具体询问图片的某一区域,而不是整个图片的所有信息,因此 Answer 可以用简短的 1 到 3 个词回答出来。

4　结　语

我们在对 VQA 数据集进行统计分析的基础上,提出仿聚类法的数据预处理方法,建立合适的 LcVMS 模型。LcVMS 模型充分考虑模型训练与响应的时间,尽可能提高模型的特征提取和分类速度,更适合作为后台快速响应智能对话。与前人只考虑模型准确率相比,我们兼顾模型与系统,以 LcVMS 为应答逻辑设计了可应用的图片问答系统。我们随机从互联网下载图片,与人工提出的 Question,一起作为图片问答系统的输入,获取 Answer,从应用实验角度来评估图片问答系统的智能程度。实验结果表明,图片问答系统能较好地分辨物体、数量、颜色和位置等信息,具有媲美幼儿的智商,具备一定的实用价值。

参　考　文　献

[1] Antol S, Agrawal A, Lu J, et al. VQA: Visual question an-

swering[C]//IEEE International Conference on Computer Vision. IEEE, 2017:2425-2433.
[2] Malinowski M, Fritz M. A multi-world approach to question answering about real-world scenes based on uncertain input [C]//Proceedings of the Advances in neural information processing systems, 2014.
[3] Geman D, Geman S, Hallonquist N, et al. Visual turing test for computer vision systems[J]. Proceedings of the National Academy of Sciences, 2015, 112(12): 3618-3623.
[4] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A neural-based approach to answering questions about images [C]//Proceedings of the 2015 IEEE International Conference on Computer Vision(ICCV). IEEE Computer Society,2015.
[5] Yu L, Park E, Berg A C, et al. Visual madlibs: Fill in the blank description generation and question answering[C]// IEEE International Conference on Computer Vision. IEEE, 2016:2461-2469.
[6] Gao H, Mao J, Zhou J, et al. Are you talking to a machine? dataset and methods for multilingual image question[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 2. Cambridge: MIT Press, 2015.
[7] Ren M, Kiros R, Zemel R S. Exploring models and data for image question answering[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 2. Cambridge: MIT Press,2015:2953-2961.
[8] 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 (CVPR). IEEE, 2016:4613-4621.
[9] Xiong C, Merity S, Socher R. Dynamic memory networks for visual and textual question answering[C]//Proceedings of the International Conference on Machine Learning, 2016.
[10] Xu H, Saenko K. Ask, Attend and Answer: Exploring question-guided spatial attention for visual question answering [C]//Proceedings of the European Conference on Computer Vision, 2016: 451-466.
[11] Yang Z, He X, Gao J, et al. Stacked Attention Networks for Image Question Answering[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:21-29.
[12] Lu J, Yang J, Batra D, et al. Hierarchical question-image co-attention for visual question answering[C]//Proceedings of the Advances in Neural Information Processing Systems, 2016.
[13] Richardson M, Burges C J, Renshaw E. Mctest: A challenge dataset for the open-domain machine comprehension of text[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013.
[14] Weston J, Bordes A, Chopra S, et al. Towards ai-complete question answering: A set of prerequisite toy tasks[EB]. eprint arXiv:1502.05698, 2015.
[15] Jia D, Berg A C, Li F F. Hierarchical semantic indexing for large scale image retrieval[C]//Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2011:785-792.
[16] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 1. Curran Associates Inc. 2012:1097-1105.
[17] Fang H, Gupta S, Iandola F, et al. From captions to visual concepts and back[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.
[18] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the Computer Vision and Pattern Recognition(CVPR). IEEE, 2015.
[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB]. eprint arXiv:1409.1556, 2014.
[20] Goodfellow I, Bengio Y, Courville A, et al. Deep Learning [M]. Cambridge: MIT Press, 2016.

(上接第 139 页)

参 考 文 献

[1] 李道亮,杨昊. 农业物联网技术研究进展与发展趋势分析[J]. 农业机械学报,2018,49(1):1-20.
[2] 张光河. 物联网概论[M]. 北京:人民邮电出版社,2014:5-50.
[3] 郭雷风,钱学梁,陈桂鹏,等. 农业物联网应用现状及未来展望[J]. 农业科技展望,2015(9):42-46.
[4] 段杰,王茂励,唐勇伟,等. 基于农业物联网的草莓大棚信息监测系统分析[J]. 现代农业科技,2018(8):288-291.
[5] 樊艳英,张自敏,陈冠萍,等. 基于物联网的精准农业玉米长势监测分析系统研究[J]. 农机化研究,2018,40(8):223-227.
[6] 金文,姚凯学. 基于物联网的农业墒情监测系统的设计与实现[J]. 计算机应用与软件,2018,35(3):84-88,204.
[7] 高百惠,徐红亮. 基于物联网的农业生产监控系统设计[J]. 农机化研究,2018,40(2):207-211.
[8] 黄海松,秦志远,张慧,等. 基于农业物联网的农作物生长监测数据融合研究[J]. 江苏农业科学,2017,45(21):241-243,251.