

分类号	<u>TP391</u>	密级	<u>公开</u>
UDC	<u>004.8</u>	学位论文编号	<u>D-10617-308-(2020)-03008</u>

重庆邮电大学硕士学位论文

中文题目	基于多重注意力机制和特征融合算法的 视觉问答系统研究
英文题目	Research on Visual Question Answering based on Multiple Attention Mechanism and Feature Fusion Algorithm
学 号	S170301006
姓 名	周思桐
学位类别	工学硕士
学科专业	控制科学与工程
指导教师	蔡林沁 教授
完成日期	2020 年 4 月 23 日

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的
研究成果。尽我所知,除了文中特别加以标注和致谢的地方外,论文中不包含他人已
经发表或撰写过的研究成果,也不包含为获得 重庆邮电大学 或其他单位的学
位或证书而使用过的材料。与我一同工作的人员对本文研究做出的贡献均已
在论文中作了明确的说明并致以谢意。

作者签名: 周思桐

日期: 2020 年 7 月 1 日

学位论文版权使用授权书

本人完全了解 重庆邮电大学 有权保留、使用学位论文纸质版和电子版
的规定,即学校有权向国家有关部门或机构送交论文,允许论文被查阅和借阅等。
本人授权 重庆邮电大学 可以公布本学位论文的全部或部分内容,可编入有
关数据库或信息系统进行检索、分析或评价,可以采用影印、缩印、扫描或拷贝等
复制手段保存、汇编本学位论文。

(注:保密的学位论文在解密后适用本授权书。)

作者签名: 周思桐

日期: 2020 年 7 月 1 日

导师签名: 蒋林元

日期: 2020 年 7 月 2 日

摘要

视觉问答任务是一个结合计算机视觉研究与自然语言处理两个领域的前沿方向。视觉问答系统可根据问题语义,从与问题相匹配的图像中找寻有用信息对问题进行答案预测。视觉问答任务模型包含图像特征处理、文本特征处理、多模态特征融合和答案预测四个模块,其中图像特征处理和文本特征处理均属于特征提取的范畴。在当前的视觉问答研究中,如何进行特征提取、多模态特征融合以及注意力机制的改进一直都是研究的难点问题,故而本文将针对这三个问题展开探索与研究:

1. 基于 Faster-RCNN 目标检测算法的图像预处理模型。本文利用 Faster-RCNN 与 Resnet101 相结合的方式处理图像信息, Faster-RCNN 用于识别属于类的对象实例,并使用边界框对它们进行定位,进而 Resnet101 模型对 VQA v2 数据集进行预处理,提取 2048 维图像特征向量,图像特征信息则以矩阵向量的文件形式参与到视觉问答模型的训练中。

2. 基于多模态特征融合的视觉问答模型研究。为了解决跨模态特征融合的问题,基于 1 的工作基础,本文采用预训练好的词向量工具和长短时记忆网络对文本特征进行表征,形成一个 2048 维的特征向量来表示问题。然后将 2048 维图像特征向量和 2048 维的问题特征向量输入多模态分解双线性池化特征融合算法模块中,生成融合特征。最后预测答案模块,以 SoftMax 为分类器进行答案预测输出。通过在 VQA v2 数据集上的实验结果证明了本文构建的视觉问答模型的合理性和科学性。

3. 基于多模态特征融合的多重注意力机制的视觉问答模型研究。为了加强模型语义信息和更准确的抓取图片特征信息,本文在基于 2 的工作基础上加入自注意力机制、引导注意力机制和多头注意力机制,构成基于多重注意力机制的视觉问答模型,旨在更好的捕捉图片及文本之间的相关语义信息,缩短多模态特征融合的鸿沟。实验结果表明多重注意力机制与多模态分解双线性池化的特征融合算法相结合的视觉问答模型具有较高的准确率,且优于先进模型。

关键词: 视觉问答, 目标检测算法, 多模态特征融合, 多重注意力机制

Abstract

The visual question answering task is a frontier direction that combines computer vision research and natural language processing. The visual question answering system can find useful information from the images matching the question according to the semantics of the question to predict the answer to the question. The visual question answering task model includes four modules: image feature processing, text feature processing, multi-modal feature fusion and answer prediction. Among them, image feature processing and text feature processing belong to the category of feature extraction. In the current visual question answering research, how to perform feature extraction, multi-modal feature fusion and improvement of attention mechanism have always been the difficult problems of research, so this article will explore and study these three problems:

1. Image preprocessing model based on Faster-RCNN target detection algorithm. In this thesis, Faster-RCNN and Resnet101 are combined to process image information. Faster-RCNN is used to identify object instances belonging to the class and use bounding boxes to locate them. The Resnet101 model preprocesses the VQA v2 data set and extracts 2048 Dimensional image feature vectors and image feature information participate in the training of visual question answering models in the form of matrix vector files.

2. Research on visual question answering model based on multi-modal feature fusion. In order to solve the problem of cross-modal feature fusion, based on the working basis of 1, this thesis uses pre-trained word vector tools and long-term and short-term memory networks to characterize the text features, forming a 2048-dimensional feature vector to represent the problem. Then the 2048-dimensional image feature vector and the 2048-dimensional problem feature vector are input into the multimodal decomposition bilinear pooling feature fusion algorithm module to generate fusion features. Finally, the answer prediction module uses SoftMax as the classifier for answer prediction output. The experimental results on the VQA v2 data set prove that the visual question answering model constructed in this thesis is reasonable and scientific.

3. Research on visual question answering model based on multiple attention mechanism of multi-modal feature fusion. In order to strengthen the semantic information of the model and capture more accurate image feature information, this thesis adds a self-

attention mechanism, a guided attention mechanism and a multi-head attention mechanism on the basis of the work based on 2, to form a visual question answering model based on the multiple attention mechanism. It aims to better capture the relevant semantic information between pictures and text, and shorten the gap of multi-modal feature fusion. The experimental results show that the visual question answering model combined with the multi-attention mechanism and the multi-modal decomposition bilinear pooling feature fusion algorithm has higher accuracy and is superior to the advanced model.

Keywords: visual question answering, target detection algorithm, multi-modal feature fusion, multiple attention mechanism

目录

第 1 章 绪论	1
1.1 课题研究背景及意义	1
1.2 国内外研究现状	3
1.2.1 基于特征融合的网络模型	4
1.2.2 基于注意力机制的网络模型	5
1.2.3 基于外部知识库的网络模型	6
1.2.4 各模型的表现小结	7
1.3 研究难点	8
1.4 课题研究内容与章节安排	9
1.4.1 课题研究内容	9
1.4.2 课题章节安排	10
第 2 章 视觉问答系统的理论基础	12
2.1 视觉问答的基本框架	12
2.1.1 图像编码模块	13
2.1.2 问题编码模块	14
2.1.3 特征融合模块	14
2.1.4 答案预测模块	15
2.2 数据集	16
2.2.1 数据集简介	16
2.2.2 相关数据集详解	17
2.3 本章小结	20
第 3 章 基于目标检测算法的图像预处理模型	21
3.1 图像预处理模型	21
3.2 残差神经网络 ResNet	22
3.2.1 残差网络	22
3.2.2 ResNet101 的网络结构	23
3.3 局部特征识别与提取	24

3.3.1 目标检测算法	24
3.3.2 Faster-RCNN 模型	25
3.4 模型实验设置与训练	28
3.4.1 模型的实验设置	28
3.4.2 模型训练步骤	29
3.4.3 模型训练结果	30
3.5 本章小结	33
第 4 章 基于多模态特征融合的视觉问答	34
4.1 视觉问答整体架构	34
4.2 文本信息处理模块	35
4.2.1 GloVe 模型	35
4.2.2 长短时记忆网络	37
4.3 多模态特征融合模块	39
4.3.1 双线性池化模型	40
4.3.2 MLB 融合模型	41
4.3.3 MFB 融合模型	42
4.4 实验结果及分析	43
4.4.1 VQA 数据集	44
4.4.2 评价指标	45
4.4.3 实验设置	46
4.4.4 实验结果分析	47
4.5 本章小结	49
第 5 章 基于多模态特征融合的多重注意力机制的视觉问答	51
5.1 模型整体架构	51
5.2 注意力机制	52
5.2.1 注意力机制的发展趋势	52
5.2.2 Attention 机制的原理与计算流程	53
5.3 相关注意力机制介绍	55
5.3.1 Self-Attention	55

5.3.2 Multi-Head Attention	56
5.4 基于多重注意力机制的视觉问答模型.....	57
5.4.1 多重注意力机制网络	57
5.4.2 多模态融合与输出预测	59
5.5 实验结果及分析	59
5.5.1 数据集	59
5.5.2 实验设置	61
5.5.3 实验结果	61
5.6 本章小结	65
第 6 章 总结与展望	66
6.1 工作总结	66
6.2 未来的研究方向	67
参考文献	69
致谢	75
作者攻读硕士期间从事的科研工作及取得的研究成果	77

第1章 绪论

1.1 课题研究背景及意义

随着科学的进步和研究团队的逐步扩大,人工智能的发展呈迅猛之势加速突破各领域的研究瓶颈,应用驱动和技术创新已然成为推进研究进展的最新趋势,人工智能已成为社会各界追捧的热门研究方向,在人工智能的众多研究方向中,自然语言处理和计算机视觉占据了相当重要的位置。

回顾自然语言处理的发展历史,自然语言处理技术也从最初基于规则、统计概率的方法逐步向基于深度学习的方法发展。特别是到2000年以后,自然语言处理技术如雨后春笋般蓬勃发展,其中包括简单的神经语言模型、共享参数的多任务学习模型、流行甚久的词嵌入模型、应用最为广泛的神经网络模型、注意力机制模型以及预训练语言模型,在这些模型的加持下,问答系统(Question Answering System, QA)^[1]从图灵测试^[2]发展到如今的成熟应用,自然语言技术功不可没,见证了各式各样问答系统的诞生。问答系统的本质是信息检索的一种表现形式,其能够根据用户提出的自然语言文本问题进行简洁明了且准确的回答,这样的信息检索方式取代了传统的信息检索模式,很大程度的提高了信息检索的效率。自此问答系统的研究便成为了最有研究意义的方向之一,同时也奠定了问答系统在自然语言处理和人工智能领域的重要性。

计算机视觉主要是解决机器“如何看”的问题,在人类世界中眼睛则可以实现这一功能,但是计算机必须借助摄像机和电脑才能得以实现,在“看”的过程中摄像机代替眼睛完成看的动作,电脑则代替大脑实现目标识别、目标跟踪、目标测量等一系列的运算,因此计算机视觉也可以理解为是利用图像处理技术对图像进行预处理和特征提取。基于深度学习的计算机视觉已经在该领域取得了长足的进步,随着5G时代低时延、超高速的到来定会为计算机视觉的进步提供新一轮的契机。

自然语言处理技术的日益成熟使得问答系统的种类呈多样化发展趋势,与此同时计算机视觉作为人工智能领域的另一个举足轻重的研究方向也不甘落后,计算机视觉的良好发展趋势为以后视觉问答系统的研究奠定了坚实的基础。问答系统领域的百花齐放和日益成熟,再考虑到人类对获取图像信息多样性的需求,催生

了新的问答模式——视觉问答系统(Visual Question Answering, VQA)^[3], 视觉问答系统是问答系统的一个特殊存在, 相比传统的问答系统多了一个图片作为输入, 而计算机视觉中的“看”刚好可以巧妙的解决图面输入的问题。视觉问答系统作为跨模态多元化发展的代表之一, 为跨模态的技术发展增添了很多机遇和挑战。

正是因为深度学习技术的不断发展与创新以及人类对未知领域的探索精神和浓厚兴趣, 视觉问答系统的发展形势一片大好。视觉问答系统不管是对技术层面的要求还是科学领域的基础理论研究都有极其严谨, 这无疑会吸引着各类科研爱好者前仆后继的进行研究与探索, 并且视觉问答任务作为一个全新的发展方向还有很大的提升空间可以进行探究。视觉问答究其本质是根据视觉信息回答相关的自然语言问题, 换言之便是可以让计算机实现“看图说话”的功能, 这不仅需要计算机能准确获取自然语言信息, 还必须学会在给定的图像中抓取需要的图片特征信息, 将两种不同模态下的特征进行融合从而在答案集中筛选预测出合适的答案。该模式不是单纯的文本交互问答, 而是采用“文本+图像”的双模态交互方式, 打破了以往的文本对文本的交互方式, 为问答系统打开了新思路。

视觉问答系统是最近几年才从传统问答中衍生出来的新方向, 虽然一开始的基线模型并不能很好的完成问答任务, 但是在深度学习技术日益发展今天, 自然语言处理技术和计算机视觉的目标检测算法也在逐步完善, 基于两个领域下是技术支持视觉问答系统的准确率已经有了质的进步。毋庸置疑的是视觉问答系统为传统的问答系统拓宽了一个新的研究方向, 同时也促进着深度学习技术向着多模态方向发展。是在未来几年中, 视觉问答的准确率将会越来越高, 性能也会有所提升。

视觉问答任务被提出的本意是帮助那些在视觉上有障碍的人士能更好的理解周围的事物和感知他们所处的环境, 并且可以帮助他们进行恢复训练, 值的肯定的是视觉问答的研究意义与应用领域已经远远超出了人们本来的期许。视觉问答系统不仅可以帮助视觉受损的人群, 更可以应用于医学领域、教育领域和娱乐领域。视觉问答系统正在以其特有的研究魅力吸引着无数科研工作者和人工智能商业应用领域的广泛关注, 其“看图说话”的独特功能模式或将应用于图像检索、儿童早教、盲人导航、辅助驾驶等工作领域, 与其他技术领域相结合从而引领技术革新也是未来发展的趋势之一, 视觉问答系统的研究及应用无疑是一个值得探索的领域。

1.2 国内外研究现状

自从2014年M. Malinowski和M. Fritz提出了“开放世界”的概念^[3],便打开了探索视觉问答世界的大门,成为人工智能史上关于视觉问答的第一次勇敢的尝试。他们在文中主要提出一种将文本语义的编码解析模型和贝叶斯框架的图像切割模型相结合的方法用于回答与图像相关的问题,该方法可以解决真实场景下包含的较为复杂的自然语言问题,例如其可以回答有关计数、判断、对象等类型的问题,并且还建立了第一个关于视觉问答任务的基线标准。此后,视觉问答系统一直是国内外的热门研究方向,吸引着无数科研工作者和科研机构投身于此。

国际的研究机构如卡耐基梅隆大学、加州大学伯克利分校、斯坦福大学等,都在视觉问答这一跨模态任务中取得了一定的成就和研究成果。斯坦福大学人工智能实验室主任李飞飞教授提出了“视觉基因组”(visual genome)计划^[4],将跨模态的视觉问答作为主要研究内容。微软研究院所研究的项目“语境中的公共对象”的重要任务之一便是探索视觉问答的图像信息和文本信息。

2015年A. Agrawal等人^[5]提出了一个相对完整的视觉问答模型,针对问题文本采用双层的长短时记忆网络,针对图像的特征信息则采用VGGNet,然后采用简单的特征融合方式将两个特征融合进而预测答案。同时,他们还提出了迄今为止最大的视觉问答数据集VQA, VQA发展到现在已经有两个版本的数据集, VQA v1主要为动画场景的图像, VQA v2数据集主要为真实场景的图像。

国内也有越来越多的人致力于视觉问答研究,高校组织如中国科技大学、浙江大学、哈尔滨工业大学、北京邮电大学、电子科技大学、吉林大学等,都纷纷加入视觉问答的研究队伍中来,近三年来发表的硕博论文从2017年的3篇增长到现在的20篇,足以证明视觉问答具备的研究意义。

以此同时,2015年百度深度学习实验室的M. Ren等人^[6]提出了一种全新的框架,他们希望视觉问答模型能够简单有效,且对图像信息产生一个全局概念,因此与M. Malinowski等人的做法相异,他们将文本信息喂入长短时记忆网络前,先将图像特征信息作为第一个单词,然后进行输入,实验结果证实这一策略的正确性。值得注意的是M. Ren等人还提供了一个新的评估视觉问答任务的数据集COCO-QA,该数据集也来源于Microsoft COCO图片数据集,不过他们为每张图片提供了

标签。现在 COCO-QA 已经成为很多评估视觉问答模型的数据集。一般的视觉问答模型在处理文本信息时都会采用循环神经网络，但 Ma Lin 等人^[7]针对视觉问答框架特点做出了新的尝试，他们的图片和文本均采用卷积神经网络进行特征提取工作，图像特征是用 VGGNet 模型进行提取的。虽然该模型在 COCO-QA 数据集上的表现没有很好，但是也为视觉问答提供了另一个研究思路。

虽然视觉问答系统才兴起不久，但是对其的研究方法却各有不同。近年来已发表的综述不胜枚举，借鉴文献[8],[9],[10],[11]的归纳方式本文将从三个方面介绍视觉问答主要的研究现状，分别为：基于特征融合的网络模型、基于注意力机制的网络模型和基于外部知识库的网络模型。

1.2.1 基于特征融合的网络模型

基于特征融合的网络模型是目前视觉问答领域最为常见的一种研究方法。因为视觉问答任务的本质便是要根据问题中提取的语义信息与图片中提取的图像特征相结合，然后预测出正确答案。由于文本特征与图像特征分属于两个不同的模态，如何将二者更完美的结合起来得到准确的特征信息便成为视觉问答的研究难点之一。基于特征融合的网络模型之所以成功还得归功于自然语言处理和计算机视觉中的深度学习技术的进步。

2015 年 M. Malinowski 等人^[12]提出了“Neural-Image-QA”的模型，该模型主要是靠长短时记忆网络(LSTM)完成特征融合工作的。作者用卷积神经网络完成了图像特征提取工作，其次用具有用长短期记忆细胞实现的递归神经网络处理输入的问题，然后将文本特征和图像特征统一馈送到称短时记忆网络的编码器中实现特征融合。其实这一过程用“特征联合嵌入”来描述更为合适，因为在进入网络之前两个模态的特征并未做其他处理。随后在 2016 年 A. Fukui 等人^[13]提出一种双线性池化的模型来解决多模态特征融合的问题，该模型被称为“多模式紧凑双线性池化”(Multimodal Compact Bilinear Pooling, MCB)。作者将两个模态的特征投影到高维的傅里叶空间中得到两个表征不同信息的向量进行乘法，然后进行卷积。J. H. Kim 等人^[14]使用多模式残差学习框架(Multimodal Residual Networks, MRN)来让模型共同学习文本和图像的特征信息，其主要的研究思路是借鉴深度残差学习的理念，利用元素模型的残差学习对联合残差映射使用元素智能乘法。最后，他们的实

验结果表明 MRN 模适用于特征融合的任务, 并且在 VQA 数据集上实现了当时的最好结果。

Yu Zhou 等人^[15]为了能更好的捕获多模态信息进行特征融合, 于 2017 年开发一种多模态分解双线性池(Multi-modal Factorized Bilinear Pooling, MFB)方法, 该方法有效的融合了多模态特征, 使得视觉问答系统性能超越了其他双线性模型。但 MFB 模型中的高维信息与较高的计算复杂度在提升模型准确率的同时也阻碍了其的适用性。因此两年之后 Yu Zhou 等人通过级联多个 MFB 块, 提出了一种广义多模态因式分解高阶池化的方法(Multi-modal Factorized High-order pooling approach, MFH)^[16], 与之前的 MFB 模型相比, MFH 降低了计算复杂度并且能够更全面的捕获文本与图像特征之间的相关性, 从而回答更多的图像相关问题。

特征融合的方法很符合属于跨模态任务的视觉问答模型的需求, 是目前大部分视觉问答系统主要的研究方法之一。特征融合的方法也从最开始的直接嵌入、朴素相加、点乘的方式逐渐演变成双线性池化的方式, 这大大提升了特征融合的有效性。MFB 和 MFH 作为新型的特征融合改进方式不仅提高了视觉问答的准确率, 而且为特征融合提供了更多的选择。

1.2.2 基于注意力机制的网络模型

基于注意力机制的网络模型是视觉问答模型中最为重要的一类模型, 其主要是根据空间注意(spatial attention)重点关注的图像中具有语义信息的区域, 给不同的区域赋予不同的权重信息进而参与神经网络的训练。

2015 年 Chen Kan 等人^[17]提出一种基于注意力机制的卷积神经网络的模型, 其核心思路是根据问题文本中所包含的语义信息来搜索图片中的图像特征信息, 故生成“问题引导的注意力图”。具体操作是将图像特征映射与可配置的卷积内核进行卷积以此来实现搜索, 同时将问题文本中的语义信息嵌入转换到视觉映射空间的卷积内核中, 则视觉映射空间中便包含了由问题文本所确定的图像特征信息。

2016 年 Yang Zichao 等人^[18]提出“堆叠注意网络”(SAN)迭代地关注视觉信息推断答案。Lu Jiasen 等人^[19]提出了一个“层次共同关注模型”(HieCoAtt), 共同关注图像信息和问题文本信息, HieCoAtt 对称地处理图像特征和问题特征。A. Fukui

等人也尝试了将注意机制结合到的 MCB 模型中并使视觉问答系统得到了一定的提升。

在 2018 年中,视觉问答中的注意力机制受到了更多的关注,其中表现最为亮眼的是 Shi Yang 等人^[19]提出的问题类型引导注意力模型(QTA),他们是在视觉问答中的 MCB 模型基础上加入问题引导的注意力机制,系统的整体准确率有了 3%的提升,此外作者还使用了残差网络和和 R-CNN 的模型处理图片数据集提取图像特征。M. Malinowski 等人^[21]提出了一种通过引导硬注意力机制的视觉问答模型,即根据特征向量的大小来选择特征向量的子集进行进一步的处理。Li Mengfei 等人^[22]提出一个文本引导的双分支注意力网络(TDAN)的视觉问答方法。不同的注意力产生的权重信息不同,因此 TDAN 模型设置了两个分支,分别进行答案预测,然后再通过问题引导的双分支结构产生相应的权重,然后将两个分支中的关键层合为一个,并产生最终的输出。模型在两个权威数据集上均进行了评估,结果显示 VQA2.0 数据集中,总体准确率从 61.89%提升到 63.94%,COCO-QA 数据集中则从 62.5%提升到了 63.98%。Peng Liang 等人^[23]提出了一种新的逐个区域注意网络,该网络不是传统的整体区域网络,而是由多个区域注意网络构成,可以实现同时定位多个目标区域的功能,从而增强语义信息和图像特征之间的一致性。

从上述的研究近况可以看出,由于注意力机制的有效运用视觉问答模型的准确率总体均有提升,但究其本质可以发现,注意力机制的运用只是加强了重要信息的语义特征并将其馈送进入训练网络,不管模型中在何处加入注意力机制或者是加入任意种类的注意力,都无法改变模型缺乏解释性和推理性的问题。因此,如何赋予视觉问答系统的推理性和可解释性仍然是一个悬而未决的难题。

1.2.3 基于外部知识库的网络模型

视觉问答系统虽然是根据所给定的图片进行问题回答,但是在实际应用过程中肯定会涉及到一些非视觉信息的语义问题,如回答“图中的天气好吗?”“图中有几只牛头梗呢?”这样类型的问题,必然需要模型具备一定的推理能力和特定域的知识储备,不然是无法获取“牛头梗”真正的所指代的动物是什么。在处理此类问题时,上述三种类型的视觉问答模型便捉襟见肘了,因此又催生了另外一个类型的视觉问答模型,基于外部知识库的网络模型。比较常见且规模较大的外部知识库

有 DBpedia^[24]、Freebase^[25]、YAGO^[26]、OpenIE^[27]、NELL^[28]、WebChild^[29] 和 ConceptNet^[30]。

2015 年 Wu Qi 等人^[31]提出了一个叫做“Ahab”的视觉问答框架，该网络框架的图像特征同其他模型一样均采用 CNN 进行提取，然后利用知识库 DBpedia 中表示相似含义的语义节点与之关联，最后通过查询的方式获得最终答案。该模式的主要缺点是只可以回答有限类型的问题。2016 年 Wu Qi 等人^[32]在之前的工作基础上又提出了一种名为 FVQA 的改进模型，同上述工作一样先给定一张图片，用 CNN 提取特征信息。与之不同的是提取的特征信息与 DBpedia 版本中的简短描述检索与之相关的外部知识嵌入到固定大小向量中，然后该向量会被馈送到 LSTM 的网络中，用于并生成答案。另外，作者在此基础上又增加了两个知识库分别是 ConceptNet 和 WebChild。

M. Narasimhan 等人^[33]开发了一种将问题-图像对和事实同时联合嵌入到一个可以允许有效搜索进行回答的模型。首先从知识库中筛选出事实类型，然后使用 LSTM 从问题中预测事实关系类型，对馈送到网络的两个向量之间的点积对检索到的事实进行排序，然后返回最高级别的事实来回答查询。与现有的技术相比，该模型将问题和事实从外部知识库中过滤出来并嵌入到知识检索中，FVQA 数据集上的事实预测准确率为 64.50%。

上述的方法中都在各自使用的数据集上进行了评估，并验证了使用外部知识库可以提高视觉问答模型的平均准确率。但是基于外部知识库的网络模型只能获取训练集中已经存在的先验知识，知识库也不能包含人类世界所有的答案，显然人造数据集的速度也远远跟不上知识增长的速度，因此基于外部知识库的网络模型也不是解决视觉问答任务的长久之计。

1.2.4 各模型的表现小结

本文将上述的三种视觉问答网络模型均是近年来相对热门且合理的研究方法，为视觉问答的发展奠定了厚实的基础，提供了可继续深入研究的方向。在本节，选取出一些在以上三类视觉问答模型中具有代表性的网络框架模型，然后简单统计了各模型的准确率表现，可以看出模型之间在统一数据及上的表现有着很小的差距，但也有各自的优势。因为同一个模型在不同的数据集上的表现有所差异，所以

本文最主要统计在 VQA 数据集上的模型。具体数据统计如表 1.1 所示。

表 1.1 部分模型在 VQA 数据集上的准确率统计

Model	test-dev				test-std
	Y/N	Number	Other	All	All
MCB ^[13]	80.81	35.91	46.43	59.40	57.39
NMN ^[34]	81.20	38.00	44.00	58.60	58.70
SAN ^[18]	79.30	36.60	46.10	58.70	58.90
MRN ^[14]	80.81	35.91	46.30	59.40	57.39
DNMN ^[35]	81.10	38.60	45.40	59.40	59.40
HieCoAtt ^[19]	79.70	38.70	51.70	61.80	62.10
MCB+Att ^[13]	82.20	37.70	54.80	64.20	—
DMN ^[36]	83.00	39.10	53.90	64.30	64.20

根据表 1.1 的数据可以看出视觉问答任务在 VQA 数据集上的表现不佳，有很大的提升空间，特别是针对“Number”类型的问题，答案准确率还没有达到 50% 的准确率，这给本文后面的研究指明了方向。

1.3 研究难点

视觉问答系统作为一个被发掘不久的衍生方向，存在很多固有难点和未知的挑战，是一项十分具有挑战性的任务。视觉问答系统一方面要解决图像特征的提取问题，在图像中可能会存在一些噪声，如物体相互遮挡、光照变化、模糊不清、背景混乱等问题，除此以外如何准确定位目标物体及获取图片信息也是难点之一。另一方面还要解决自然语言理解问题，自然语言中混乱的语言表达模式、丰富的语气词均会干扰文本意义信息的理解和提取。最后，由于抽象提取出来的视觉特征和文本特征分属于不同模态，如何进行跨模态特征融合，消除两个模态中的间隙也是视觉问答面临的又一难点。

本文对视觉问答的难点进行深入研究探讨，整理得出视觉问答任务的诸多难点关键集中在以下几点：

1. 到目前为止，还没有专门针对视觉问答任务的统一专属通用模型，图片和文本就分属于两个不同的模块，进而关于视觉问答任务的解决方式只是分模块处理，换言之便是分解视觉问答的任务，不同的模块对应相应的训练模型，每个模块各司其职。

2. 为了让视觉问答任务能更贴近真实自然环境的回答，故而视觉问答的相关

数据集包含了很多类别的图片，数据量相当庞大，在模型训练过程中必将耗费大量的时间。

3. 传统的特征融合模型在特征融合过程中会丢失一些重要语义信息，满足不了视觉问答模型的跨模态融合要求，需要发掘更多适合的融合模型。

4. 一部分的视觉问答模型虽然采用了共同注意力机制，但是只用于其中某一模态便会忽略视觉语义信息和文本语义信息之间的相互影响，对特征融合模块的进行带来一定的阻力。

1.4 课题研究内容与章节安排

1.4.1 课题研究内容

本文主要关注基于多重注意力机制和特征融合算法的视觉问答系统研究。以深度学习的神经网络为算法支撑，基于多模态分解双线性池化模型与多重注意力机制，构建一个完整且适合处理图像信息与文本信息的多模态视觉问答系统，实现架构基于多重注意力机制和特征融合算法的视觉问答模型。本文主要工作包括以下三点：

1. 基于 Faster-RCNN 目标检测算法的图像预处理模型。针对视觉问答系统在训练过程中所投入的时间成本过大，文中提出的图像预处理方式，通过结合 ResNet101 网络利用 Faster R-CNN 模型对视觉问答数据集中的图片数据集进行预处理，实现细粒度的图像特征提取，便于工作 2 中进行端到端的模型训练，同时也节省了模型训练的时间。提取出来的图像特征以文件的形式作为训练数据直接喂入视觉问答模型中，为以后统一视觉问答模型奠定基础。

2. 基于多模态特征融合的视觉问答系统。传统的 VQA 基线模型其高维表示和高计算复杂度可能严重限制了它们在实际应用中的适用性。本文提出了一个新的视觉问答系统框架，对于多模态特征融合模块，采用多模态分解双线性池化的方法，有效地结合了多模态特征，使 VQA 的性能优于其它双线性池化模型。首先问题编码模块是将自然语句的问题文本利用 GloVe 词向量模型进行预训练，然后用 LSTM 网络抽取问题语义特征信息。其次特征融合模块是将图像和问题特征变换到同一个特征空间，通过多模态分解双线性池化特征融合算法将图片特征和文本

特征相融合。最后答案预测模块以 SoftMax 为分类器进行答案预测输出。

3. 构建基于多模态特征融合的多重注意力机制的视觉问答系统。本文提出基于多重注意力机制的视觉问答系统,旨在通过多层多种注意力机制更好的捕获语义信息,从而进一步解决不同模态之间的语义鸿沟问题。首先,本文从卷积神经网络的高层语义中生成图像的语义属性,选择与问题相关的属性作为语义注意力。其次,本文从卷积神经网络中提取有空间信息的中间层输出,然后通过一个长短时记忆网络将目标区域编码为上下文已知的视觉特征,本文进一步通过多层感知机定位与答案相关的区域作为视觉注意力。最后,本文使用 SoftMax 层联合优化语义注意力,视觉注意力和问题表达,预测答案。并且本文通过实验讨论不同的注意力机制对视觉问答系统的影响。在一个完整的系统架构中,将 MFB 模型与多重注意力机制相结合,为视觉问答任务提供了一个相对统一的系统,进一步提高了系统的准确率,实验结果显示本文的模型表现优于现有的大多数视觉问答模型。

1.4.2 课题章节安排

本文主要研究了面向虚拟学习环境的智能问答,共分为六章,具体章节安排如下:

第1章,绪论。首先交代了论文研究课题的研究背景及意义,然后分析了近年来的视觉问答模型并将它们的研究方法归纳整理为三类,分别对这三类视觉问答模型的国内外研究现状进行阐述与比较,并指出本研究课题目前存在的一些关键问题与对应的解决方法;最后对本文的主要研究工作进行了内容概括与章节安排。

第2章,视觉问答系统的理论基础。本章首先介绍了对视觉问答系统框架,并且对其图片处理、文本处理、特征融合、答案预测这四个模块的训练模型和研究方法进行了细致的阐述;然后对现有的视觉问答数据集的现状和具体情况进行了详细的调查,同时统计了一些较为典型的模型,并比较了它们在各大数据集中的表现;最后分别介绍了视觉问答系统中所涉及到的算法理论,其中包括处理图片信息的卷积神经网络、处理文本信息的递归神经网络及在深度学习领域比较常见的捕获细粒度特征信息的注意力机制网络等。

第3章,本章提出了一种基于 Faster-RCNN 目标检测算法的图像预处理模型。针对视觉问答系统在训练过程中所投入的时间成本过大,提出的图像预处理方式,

通过结合残差网络利用目标检测算法模型对视觉问答数据集中的图片数据集进行预处理,实现细粒度的特征提取,便于后续的端到端的模型训练,节省模型训练时间。

第 4 章,本章提出了一种基于多模态特征融合的视觉问答模型。针对跨模态特征融合,利用多模态分解双线性池化模型把已经提取好的图像信息特征与文本特征进行融合、预测答案,最后在实验中加入基因组数据集作为扩充数据集,从而进一步提高模型准确率。本文在 VQA v2 的数据集上分别进行实验,逐步验证了图像处理模块的 Faster R-CNN 模型、文本处理模块的 GloVe 模型和 LSTM 网络、特征融合模块的融合方式、外部数据集的扩充等方面在视觉问答模型的准确率上均有一定的提高。

第 5 章,本章提出了一种基于多模态特征融合的多重注意力机制的视觉问答模型框架。为了加强模型语义信息和更准确的抓取图片特征信息,本章主要是在第四章基于多模态特征融合的视觉问答系统的基础上加入自注意力机制、引导注意力机制、多头注意力机制等多重注意力机制,旨在更好的捕获图片及文本之间的高语义信息,从而缩短多模态特征融合的鸿沟。

第 6 章,总结与展望。分析并总结本文关键研究内容,指出相关不足之处,对未来工作做进一步展望。

第2章 视觉问答系统的理论基础

视觉问答是人工智能领域中一个全新的研究方向,随着科研力量的逐步壮大,跨模态任务的提出与解决已是大势所趋,特别是针对视觉信息与文本信息的跨模态融合吸引了大量的科研工作者的目光。如何利用现有技术对视觉问答系统进行研究?如何在现有技术上创新?都是困扰广大研究人员的问题。在本章中,首先以总分的形式来介绍视觉问答系统的基本框架构成,包括介绍视觉问答的研究思路与算法基础,然后介绍了现在比较主流的七个视觉问答数据集的基本情况和来源,为后面的实验研究奠定理论基础,加深读者对视觉问答理论研究的理解。

2.1 视觉问答的基本框架

视觉问答系统一共可以分为四个核心模块:图像编码模块、问题编码模块、特征融合模块和答案预测模块,这是一个完整的视觉问答系统所必需的四个模块。一个简单的视觉问答模型应该能基本实现“看图说话”的功能,即据用户给定的图片,针对用户提出的问题,经过一系列的深度处理技术能够合理的回答用户所问问题。若把视觉问答的处理过程抽象化,则:

模型的输入:“图片+文本”——像素形式的图片 I , 英文形式的问句问题 Q 。

模型的输出: 文本形式的预测答案 A 。

模型处理过程: 图像编码模块一般采用卷积神经网络进行操作将图片 I 中的特征信息提取出来形成具有一定语义信息的特征向量 v ; 问题编码模块一般采用循环神经网络对问句 Q 的语义特征信息进行编码形成高语义特征向量 q 。但所提取到的特征向量 v 和 q 原本就分属于不同的模态,故而它们二者之间存在较大的语义结构鸿沟,为了缩短它们之间的语义鸿沟,需要对做多模态特征融合的操作减少差距。故而特征融合模块主要负责将由不同方式提取的图像信息特征和文本信息特征之间建立语义上的映射关系,从而得到融合的多模态特征 m 。答案预测模块则是在多模态特征 m 的基础上采用多层感知机和 SoftMax 函数来得到候选答案分布 \tilde{a} 。

以上即是视觉问答系统在处理四个模块时的基本操作过程,由于视觉问答系统刚起步几年,至今未找到统一处理的模型,因此仍处于各模块分而治之的阶段,

下一节将对各模块的处理过程进行详细的介绍。其系统概要图如图 2.1 视觉问答系统总概图所示。

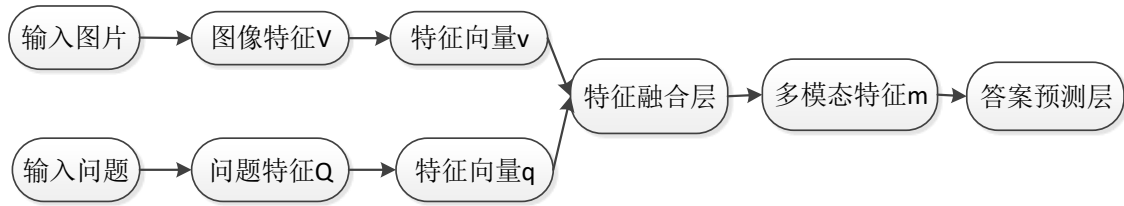


图 2.1 视觉问答系统总概图

2.1.1 图像编码模块

视觉问答系统中的图像编码模块常常采用深度卷积神经网络(Convolution Neural Network, CNN)进行图像特征提取,本文将输入的图片定义为 $I (I \in R^{W \times H \times 3})$, 提取后的图像特征为图像特征向量 v 。该过程如公式(2.1)所示:

$$v = CNN_{fc}(I) \quad (2.1)$$

v 的维度可在实验中进行设定,可设置为 1024、2048 或 4096 维,在本文中的后续实验中 v 的维度均设定为 2048 维。

CNN 的种类可根据不同的任务需求进行选择,常用的特征提取模型有 VGGNet^[37]、GoogLeNet^[38]以及 ResNet^[39],其中 ResNet 在 2015 年表现尤为突出,是当年各大图像赛事的识别模型冠军,本文在第三章图像预处理部分便采用的 ResNet 来提取图像特征。一般情况下,用于图像特征提取的 CNN 模型会在 ImageNet 上进行预训练,预训练好的 CNN 模型具备更有效的图像理解能力,然后再参与到视觉问答的图像特征提取中达到事半功倍的效果。

虽然经过预训练的 CNN 模型有相对良好的性能,但其能识别的物体种类却很单一,相对于视觉问答任务中所涉及的图片均来源于复杂世界真实场景,且包含真实场中各式各样的物体,对图像的理解能力的要求极高,这样的 CNN 模型便存在很大的劣势。为了解决这一问题,人们开始把目光转向注意力机制,利用注意力机制能增强局部特征的优势来弥补 CNN 模型在提取图片的全局特征时丢失的局部特征。

在视觉问答任务中,提出的问题往往是根据所给图片的内容进行提问,所以网络模型在提取图像特征时只需要能捕获到局部信息便可预测答案。为了最大程度的提取到局部图像特征 $V^h = [v_1^h, \dots, v_k^h]$ 提高网络模型的执行效率,可利用 Attention

机制对问题语义信息和局部图像特征 V^h 进行权重筛选, 得到针对所述问题相关的图像特征向量 v 。

$$v = \text{Attention}(V^h, q) \quad (2.2)$$

2.1.2 问题编码模块

在目前的研究中, 视觉问答仅是处理英文形式的文本信息, 由于公开数据集的限制, 视觉任务从提出至今还未涉及中文形式的问题回答, 因此在问题编码模块主要以英文文本的处理方式进行研究。故将问题文本进行英文分词得到 w_1, \dots, w_N , 其中 w_i 是经过分词后的第 i 英文单词的分词形式, 则该模块的输入问题为 $Q = \{w_1, \dots, w_N\}$, N 为问题的长度, 即分词后单词的个数。在问题编码模块中, 首先要将每个英文单词表示为词向量, 便于模型提取语义信息, 然后在具有 d 维的语料库或词典 D 中进行预训练, 将单词转化成 d 维的 one-hot 向量, 被进行编码后的问题为 $Q^o = \{w_1^o, \dots, w_n^o\} \in R^{d \times N}$, 最后每个单词的 one-hot 向量会被嵌入到词向量空间。

$$Q^D = W_D Q^o \quad (2.3)$$

其中 W_D 的第 i 列是 D 中第 i 个单词对应的词向量表达。在训练时 W_D 既可以做随机初始化也可以利用其他预训练的词向量, 如 GloVe^[40]、Skip-gram 及 CBOW^[41]等。使用预训练的词向量进行初始化往往能得到更好的效果。

问题文本经过语料库或词典的预训练流程后所获得的问题词向量 Q^D 还需进一步编码成问题特征向量 q , 文献[42], [43]的简单做法则是把问题词向量分组后做组内平均, 然后将各组均向量进行拼接; 或者是将所有问题词向量平均或相加。但是朴素拼接和相加并不能很好的获取语义信息, 最好最有效的方式是采用循环神经网络(Recurrent Neural Network, RNN)进行词向量编码, 这不仅能准确编码词义, 刚能将词间语义信息进行表征与捕获。

2.1.3 特征融合模块

特征融合模块主要根据图像特征向量 v 和问题特征向量 q 进行操作, 从而得到多模态特征向量 m , 用公式可表示为

$$m = \text{Fusion}(v, q) \quad (2.4)$$

融合的方式有很多种,简单的特征融合方式为直接拼接 $m=[v;q]$ 或者将同维的元素一一对应进行加法操作 $m=(W_v v+b_v)+(W_q q+b_q)$,这两种融合方式都过于朴素,忽略了两个模态下的特征信息存在差异的情况,或加剧向量维数或造成语义误差从而影响答案预测模块。经过一段时间的科研探索,前文提到的 A. Fukui 等人^[13]发现双线性池化的方法可以很好的融合多模态特征,给后面的研究提供了有效的模型。

若令图像特征向量 $v \in R^m$ 问题特征向量 $q \in R^n$ 作为双线性池化的输入,多模态特征向量 $m \in R^o$ 作为输出,以多模态低秩双线性池化模型 MLB^[44]为例可得如公式(2.5)的操作:

$$m_i = v^T W_i q + b_i \quad (2.5)$$

其中 W_i 为参数矩阵,可进一步进行矩阵分解,以减少权重矩阵 W_i 的秩,使其具有较少的正则化参数。

权重矩阵可改写为:

$$W_i = U_i V_i^T \quad (2.6)$$

$$m_i = v^T (U_i V_i^T)_i q + b_i \quad (2.7)$$

其中 $U_i \in \mathbb{R}^{N \times d}$ 和 $V_i \in \mathbb{R}^{M \times d}$, 这限制了 W_i 的等级最多为 $d < \min(N, M)$ 。

由于在第四章中会着重介绍模型 MLB 的计算分解步骤,故在此不再详细介绍。

2.1.4 答案预测模块

经过融合操作的多模态特征 m 将作为答案预测模块的输入进行答案预测。自然语言处理中答案预测模块通常被归结为分类问题,本节将讨论视觉问答任务中如何在开放式的配置下进行答案预测,首先,在所给数据集中确定候选答案组成集合 A , 然后集合 A 中每一个候选答案都可以当成一个类分布在答案候选集中,最后以概率分布情况预测正确答案,概率大者为正确答案。

$$A = \max_{\tilde{a} \in A} P(\tilde{a} | I, Q) \quad (2.8)$$

在答案预测中一般选取多层感知机 MLP 和 SoftMax 函数实现这一过程:

$$\tilde{a} = \text{Soft max}(MLP(m)) \quad (2.9)$$

其中, $\tilde{a} = P(\tilde{a}|I, Q)$ 。网络训练可采用交叉熵:

$$L(\tilde{a}, t) = -\sum_{i=1}^{|A|} t_i \log \tilde{a}_i \quad (2.10)$$

其中, t 是正确答案所对应的 One-Hot 特征向量。

上述方法有一个前提是要求每一个问题都需对应一个唯一正确的答案, 但是并不是所有的问题都满足此条件, 很多问题的答案并不只有一个, 这时答案预测就不能被划分为简单的分类问题, 可以就实际情况将其定义为二分类问题来处理, 实现二分类问题可以将 Sigmoid 函数用于替换公式(2.9)中的 SoftMax 函数, 则误差函数为:

$$L(\tilde{a}, t) = -\sum_{i=1}^{|A|} t_i \log(a_i) + (1-t_i) \log(1-a_i) \quad (2.11)$$

其中, $t_i \in [0, 1]$, t_i 表示第 i 个答案和问题的匹配程度。上述误差函数中一个问题允许对应多个答案, 即使不在候选集中的正确答案也是被允许用来参与到样本集的训练, 从而可以提供更为丰富的训练样本。

2.2 数据集

2.2.1 数据集简介

从 2014 年开始, 视觉问答的数据集种类发展的越来越多, 但是用于评估视觉问答任务的主流数据集屈指可数, 其中包括 DAQUAR^[3], COCO-QA, VQA 数据集^[45], FM-IQA^[46], Visual7W^[47]和 Visual Genome^[48]和 CLEVR 数据集^[49], 其中 VQA 数据集、COCO-QA 数据集占据了模型评估的主导地位, CLEVR 数据集作为后起之秀也逐渐吸引着越来越多的视觉问答模型在此之上进行训练与评估。

本节提到的视觉问答任务的主要数据集除 DAQUAR 外, 所有的图片均来自于微软开发的 MS-COCO 数据集, 其包含 328000 张图片, 每个图片平均设置 5 个字幕描述, 图片内容涵盖 91 个常见的对象种类, 共有 200 多万个标记实例。此外, 除 MS-COCO 的数据集图像外, Visual7W 和 Visual Genome 还使用了 Flickr100M 中的图像。

本小节主要介绍上述七种数据集, 并详细的分析了每个数据集的构成比例、使

用原则、应用情况、评价标准以及优劣势。

如表 2.1 所示是本文介绍的 7 个数据集的数据统计情况。

表 2.1 视觉问答数据集统计数据

Dataset	Image	Question	Image source	Task type	Eval.server
DAQUAR	1449	12468	NYU-Depth v2	OE	NO
COCO-QA	123287	117686	MS-COCO	MC	NO
VQA-real	204721	614163	MS-COCO	OE	YES
VQA-abstract	50000	150000	MS-COCO	MC	YES
FM-IQA	123287	250560	MS-COCO	OE&MC	NO
Visual Genome	108249	1700000	MS-COCO	OE	NO
CLEVR	100000	999968	—	—	NO

2.2.2 相关数据集详解

(1) DAQUAR

DAQUAR(The DATaset for QUestion Answering on Real-world Images)是第一个发布的视觉问答数据集，用于回答真实场景下室内的图像相关问题。它来源于 NYU-Depth V2 数据集，主要有 1449 张图片，问答对一共有 12468 个，其中图像训练集为 795 张、测试集为 654 张；问答对训练集有 6794 个、测试集有 5674 个。该数据集还提供另外一个版本的配置，称为 DAQUAR-37，顾名思义其数据集中含 37 个对象类别，在 DAQUAR-37 中训练问答对仅有 3825 个，测试问答对为 297。

虽然 DAQUAR 是视觉问答任务中的第一个数据集，但是它的数据量无法用于视觉问答模型的训练，更别提用其来评估模型的性能。另一方面 DAQUAR 数据集包含的是室内场景下的图片，这也极大的限制了其的应用与推广，而且室内场景的图片还会受到光照、物体种类繁多、空间杂乱等因素的影响，从而使大部分问题无法给出答案，就连人类在这个数据集上的测试标准也只达到 50.2% 的准确率。

(2) COCO-QA

COCO-QA 数据集中共包含 123287 张图片，均来自于 MS-COCO 数据集，其中训练集 72738 张、测试集 3894 张，每张图片配置一个问答对，每个答案为一个单词。该数据集中的问答对是根据 MS-COCO 数据集中的图像标题利用自然语言处理技术自动生成，例如，图片标题为一个女孩在打排球，自动生成的问题可能是“图片中的男孩在玩什么”，答案为“打排球”。

在 COCO-QA 数据集中用于回答问题的答案均为一个单词,并且数据集中只包含了 435 个不一样的答案,故而在数据集上的评价略为简单,不适应于视觉问答模型的评估。此外,该数据集中用于处理问答对的自然语言处理算法存在误差,成为了该数据集的一大的劣势,作者为了方便处理语言信息,把较长的语句分解开,使得算法在处理语法和从句并不能很好的获取语义信息。数据集中的问题类型只有 4 种,即对象类型、颜色类型、计数类型、位置类型,所占比例分别为 69.84%、16.59%、7.47%、6.10%,且这四种问题在回答时只需要模型捕捉到图片的局部特征便可进行回答,这样大大降低了数据集的适用性和用于评估模型的广泛性。

(3) VQA 数据集

从 2014 年 VQA 数据集被提出至今,已经发展到两个版本的数据集了,包括 VQA v1.0 和 VQA v2.0。本节中主要介绍 VQA v1.0 的数据内容,VQA v2.0 的内容将会在本文的第四章实验数据及部分进行详细说明。VQA v1.0 版本的数据集分为来源于 MS-COCO 真实场景的图片数据集 VQA-real 和剪贴画场景图片即抽象场景图片数据集 VQA-abstract,该数据集由人工采集问题和答案,并将问题类型分为“yes/no”、“number”、“other”、“all”。

在 VQA-real 数据集中每种图片都有属于自己的图片 ID,文本数据集便是根据图片的 ID 信息与图片建立对应关系,便于从中抽取答案。v1.0 版本中 1 张图片对应 3 个问题,每个问题对应 10 个答案,所设问题的个数共为 614163 个,其中 248349 个用于培训,121512 个用于验证,244302 个用于测试。每个问题对应的 10 个答案均有属于自己的注释器,设置 10 个答案的目的是作为评估度量。

VQA-abstract 数据集和 VQA-real 在格式上相同,主要区别在于 VQA-abstract 中的图片以卡通动画场景为主。构成卡通动画场景的是 20 种人类卡通形象、30 种动物形象和 100 多种物体形象,其中人类卡通形象还存在性别、年龄大小,种族类别的差异。VQA-abstract 包括 50000 张合成图片,旨在通过合成场来平衡数据集使其更加具有多样性,同样按照 1:3:10 的原则设置一张图片对应三个问题,每个问题对应 10 个答案。

VQA-real 数据集和 VQA-abstract 数据集都具备开放式与多种选择式两种评估方式。由于 VQA-real 包含真实场景的信息内容,有充足的图片可供训练评估,文本数据集除了问答对还有问答对补充数据集,该数据集的多样性符合视觉问答模

型的训练,更适用于评估模型,因此受到广泛关注与应用。

(4) FM-IQA

FM-IQA(Freestyle Multilingual Image Question Answering)数据集使用的 123287 张图像同样是来自于 MS-COCO 数据集,该数据集还有 250560 个问答对。一开始 FM-IQA 数据集是中文形式的,后来为了提高数据集的应用性,便将其进行人工翻译成为英文形式的数据集。与其他数据集不同的是其允许答案是一个完整的句子,而不是简短的单词形式的答案,换言之,只要不超出图片的内容范围,注释者便可以不受约束的提问任何类型的问题。这无疑在原有数据集中增添了更为多样化的问题形式,极大程度的增加了模型进行自动评估的难度,因此一般情况下不建议采用此数据集进行评估。

(5) Visual Genome 和 Visual7W

Visual Genome 数据集是 2016 年斯坦福大学的李飞飞组在“视觉基因组”项目中发布的图片语义理解数据集,是视觉问答数据集中较大规模的数据集之一,他们发布 VG 数据集是希望可以推动图像中的高语义研究。而后 VG 数据集已经成为了视觉关系检测的标准数据集。VG 数据集的构建由四个部分组成,分别为区域描述(Region Description)、区域图(Region Graph)、场景图(Scene Graph)和问题/答案对(QA),即图片根据其内容被划分为一个一个的区域,每个区域都有对应的自然语言描述的标签;每个区域中的对象、属性、关系又被提取出来,从而构成一个局部的场景图;把整张图片的局部场景图进行合并便可得到该图片的全局场景图;同时每张图片都配置了多对 QA,在 VG 数据集中一共包含 170 万个问题/答案对。该数据集特殊的场景结构化注释不仅将视觉信息巧妙的联系对应,而且还囊括了视觉特征的属性与关系,很好的诠释了图像语义,便于视觉任务的图像信息提取,因此深受视觉问答研究人员的青睐。

VG 数据集图片内容包含了场景、动物、人类、运动等图片数据集应该具备的基准,同时, VG 数据集由六种类型的问题组成:“What”、“Where”、“When”、“Who”、“How”和“Why”,并将这些问题/答案对设置成两种模式,其一基于区域场景问答模式,其二是自由形式的问答模式。前者在设置时注释器需要提供基于区域场景的图像特征问题/答案对;后者在设置注释器时则需要提供 8 个 QA 对并显示图片。截止目前, VG 数据集是视觉问答任务中最大的数据集,但由于作者在

创建数据集时没有对其设置评估标准，故而本文在实验部分将其作为了额外的补充数据集。

Visual7w 数据集一共包含 47300 张图片均来自于 VG 数据集，是包含附加注释的 Visual Genome 的子集，该图片也是来源于 COCO 数据集。相比于 VG 数据集中六个类型的问题，Visual7w 数据集还多增加了“Which”，共 7 个“W”类型的问题，这也是其数据集名字的缘由。用户提问时必须以这七个类型中的任意一个疑问词开始提问，而且问题内容必须与图像有关。

(6) CLEVR

CLEVR(Compositional Language and Elementary Visual Reasoning)是一种用于组合语言和初级视觉推理的诊断数据集，和上述数据集不同，它主要是为解决视觉问答任务中缺乏推理问题的难点而构建的，它不仅具有详细的注释，而且还可以用于测试视觉问答系统的推理性。

CLEVR 数据集中包括训练问题 699989 个，验证问题 149991 个，测试问题 149988 个，该数据集的图片内容基本上由一些简单的颜色各异、形状不同的几何体构成，但是文本数据集中的问题却是一些较为复杂的逻辑推理问题，诸如“图中是否有相同数量的金属球？”、“有多少红色的圆柱体”此类的问题。

CLEVR 中的问题测试了视觉推理的各个方面，包括计数、比较、属性标识、空间关系、逻辑运算。CLEVR 中的每个问题都用自然语言和功能程序表示。功能程序表示可以精确确定回答每个问题所需的推理能力。特别地，CLEVR 中的图片和问题由模拟引擎生成，在 CLEVR 官方网站作者提供了该模拟引擎的数据集生成代码，可供用户随时随地为 CLEVR 中的图像生成新的问题。

2.3 本章小结

本章主要介绍了视觉问答系统所涉及到的关键技术及理论知识。首先介绍了对视觉问答系统框架，并且对其图片处理、文本处理、特征融合、答案预测这四个模块的训练模型和研究方法进行了细致的阐述；然后对现有的视觉问答数据集的现状和具体情况进行了详细的调查和阐述，同时对每个数据集的可否用于评估模型的原因进行了分析，并粗略的统计了典型模型在各大数据集集中的表现。

第3章 基于目标检测算法的图像预处理模型

在视觉问答任务中的四个模块中，图像编码模块中的卷积神经网络或是其他网络模型能否高效的提取到图像特征并且进行准确的语义表征尤为重要。在视觉问答系统中要实现“看图说话”的功能，必先让模型学会“看图”，只有准确获取到图片的特征信息才能更好地回答文本问题。因此如何训练一个完善的目标识别算法模型对其进行的表征，便是本章主要研究的内容。本章主要探索了基于 Faster-RCNN 目标检测算法的图像预处理模型，经过图片预处理过程以后将图片中的图像信息封装成压缩文件的形式参与到视觉问答模型的训练中，这样的处理形式不仅可以节约模型训练时间，还可以高效准确的提取图像特征，为第四、第五章研究视觉问答模型奠定了坚实的基础。

3.1 图像预处理模型

在之前的研究中，总是同时将图片数据集和文本数据集进行训练，每次调试模型都要将数据集中的图片重复处理，图片数据集的数据量巨大，加之图片处理所需时间较长，与文本数据集的训练相比则需要花费大量的时间成本；另外因为提取图像特征的模型单一，提取出来的特征并不能很好的表征图像。为了解决上述两个问题，本文提出了一种基于目标检测算法的图像预处理模型，该模型主要包含的算法有残差网络 Resnet101 和目标检测算法 Faster-RCNN。

1. 残差网络：用于提取图片中的全局图像特征。
2. 目标检测算法 Faster-RCNN：用于识别抽取图像的局部特征。

采用 Faster-RCNN 与 Resnet101 相结合的方式处理图像信息，不仅可以弥补传统 CNN 处理图片时只能提取图像的全局特征的劣势，还可以让图像信息形成细粒度的图像特征，加强图像的语义表征。同时本文的图像“预处理”的意思是将图片数据集单独处理后形成图像特征文件，该文件有每张图片的全部信息，其中包括目标对象信息、位置关系、逻辑关系等，并且这些图片特征可以直接参与到视觉问答的模型训练中去。将图片编码模块分离出来好处多多，不用重复进行图片的训练还可以节省训练时间。

3.2 残差神经网络 ResNet

3.2.1 残差网络

残差网络在深度学习的发展历程中占据了相当重要的位置，残差神经网络 (Residual Neural Network, ResNet)^[50]由微软研究院提出，并在 2015 年 ImageNet 大赛中获得冠军，现在残差网络的模型深度已经发展到 152 层了。作者何凯明在设计残差网络之初，灵感源于如何解决深度模型的退化问题。

通常情况下，深度模型的网络层数越深，则网络模型的学习能力就会相对的变强，但实际情况却与之相反，当模型的深度逐渐加深时，模型的准确率接近饱和以后便开始下降。研究后发现，该情况的发生不是因为模型过拟合了，而是因为当深度模型超过它适度的层数以后会引起训练误差，这就是模型退化问题。为了解决该问题的产生，何凯明提出了残差学习的模型，将前一层的输出通过恒等映射传到后一层，这样的操作既可以保证信息完整性，又可以简化模型学习的难度。

一个堆积层结构的残差学习单元，令其输入为 x ，与之对应的学习特征为 $H(x)$ ，则它可以学习到的残差为 $F(x) = H(x) - x$ ，原始的学习特征是 $H(x)$ 。之所以这样是因为残差学习相比原始特征直接学习更容易。当残差为 $F(x) = 0$ 时，此时堆积层仅仅做了恒等映射，至少网络性能不会下降，实际上残差不会为 0，这也会使得堆积层在输入特征基础上学习到新的特征，从而拥有更好的性能。残差学习单元的结构如图 3.1 所示。

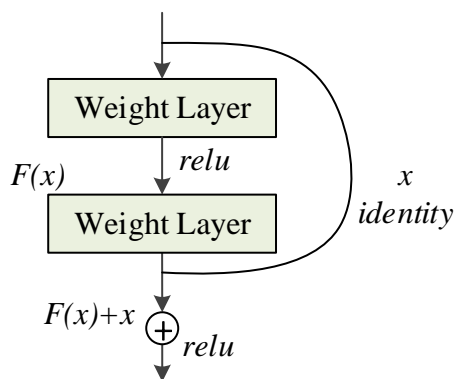


图 3.1 残差学习单元

直观上看，因为残差一般比较小，那么学习难度就小很多，残差学习单元需要学习的内容就少，这就是为何残差学习较为容易的原因，但也可以从数学角度对这

个问题进行分析, 可得残差单元表示为:

$$\begin{aligned} y_l &= h(x_l) + F(x_l, W_l) \\ x_{l+1} &= f(y_l) \end{aligned} \quad (3.1)$$

其中, x_L 和 x_{L+1} 分别表示第 L 个残差单元的输入和输出, 注意每个残差单元一般包含多层结构。 F 是残差函数, 表示学习到的残差, 而 $h(X_L) = X_L$ 表示恒等映射, f 是 *ReLU* 激活函数。基于上式, 求得从浅层 $l-1$ 到深层 $L-1$ 的学习特征。

$$x_L = x_l + \sum_{l=1}^{L-1} F(x_l, W_l) \quad (3.2)$$

残差网络有以下三大优势:

1. 简化学习过程, 同时又增强梯度传播。

普通的网络是让其学习原始的信号, 然后残差网络则是让其学习信号的差值, 这就是差异所在, 这样的学习效果更有效, 而且还简化了学习过程。

2. 打破网络模型的不对称性。

残差网络之所以达到很深的网络层数而没有出现模型退化问题, 在于残差网络设置了跳层连接, 这样的设置可以增强信息的梯度流动, 从而层数加深也不影响其训练效果。在训练过程中权重矩阵一直是一个高维度矩阵高维, 但其实这个矩阵中好多维度里面是没有包含信息的, 这样的网络就不具备强大的表征能力, 这是由于网络的对称性决定的, 但是残差网络的跳层连接巧妙的打破了这种网络对称性。

3. 提高网络泛化能力。

从文献[51]中可以得知, 于 VGGNet 模型而言, 如果删除网络模型中的任意一层必会造成模型性能奔溃, 与之相反的是, 当深度残差网络训练完一个深层网络后, 测试时随机删减掉某一网络层并不会影响网络, 从而使网络性能出现大的退化, 其实深层的残差网络相当于不同深度的浅层神经网络的堆叠, 即使删除某一层并不会产生较大影响, 故而残差网络有较强的网络泛化能力。

3.2.2 ResNet101 的网络结构

由残差网络的优化历程、网络优势、参数量等方面的考量, 本文在图片预处理模型中采用 ResNet101 网络对图片数据集进行预处理。下表 3.1 为 ResNet101 的卷积层参数与输出维度。

表 3.1 resnet101 的卷积层、输出维度

卷积层名	输出维度	卷积层参数
res1	256×256	$7 \times 7, 64$ $3 \times 3, \text{max pool}$
res2_x	128×128	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
res3_x	64×64	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
res4_x	32×32	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$
res5_x	16×16	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$

其中 $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$ 表示卷积核为 $1 \times 1, 64$ 通道、 $3 \times 3, 64$ 通道、 $1 \times 1, 256$ 通

道的卷积层分别有 3 个，共有九层卷积层。四个卷积块加上步长为 2 的 7×7 的卷积和 3×3 的最大池化层共同构成 101 层可学习参数的卷积层。

3.3 局部特征识别与提取

3.3.1 目标检测算法

基于深度学习的目标检测算法已经随着技术的发展日益成熟，目前较为出名的算法有 R-CNN 系列算法，YOLO 算法、SSD 算法等，不同任务根据各算法特点做出不同选择。R-CNN 是由 R. Girshick 发表于 2014 年的 CVPR 上的网络模型，首次成功地将深度学习的方法用在目标检测上^[52]。2015 年 R. Girshick 等人^[53]提出了 Fast-RCNN 网络解决候选框中的重复计算问题，并且优化了损失函数，在 Pascal VOC2007 得到 0.67mAP。次年 R. Girshick 等人继续提出改进后的 Faster-RCNN 网络^[54]，通过引入 RPN(Region Proposal Network)网络将前景分类、框回归过程在共享特征图上同时进行，该过程只需 10ms 左右，不仅提高了模型预测速度，而且还在 VOC2007 上取得 0.73 的 mAP 的成绩。J. Redmon 在 2016 年提出的 YOLO^[55]可

以一次到位地进行目标的识别与位置的回归,成为首个真正意义上的 end-to-end 目标检测模型,大幅提升整个模型的速度,但定位精度有所下降。

3.3.2 Faster-RCNN 模型

Faster-RCNN 模型在目标检测领域有很高的准确率,并且具有很快的识别速度,因此本文在提取图像特征时采用了 Faster-RCNN。该模型的网络组成架构如图 3.2 所示。首先采用卷积神经网络提取图像特征,然后利用区域提案网络(RPN)生成目标边界框,通过对目标边界框的不断精修,得到最优的目标定位,然后映射出目标区域图像特征,用于表示图像信息。

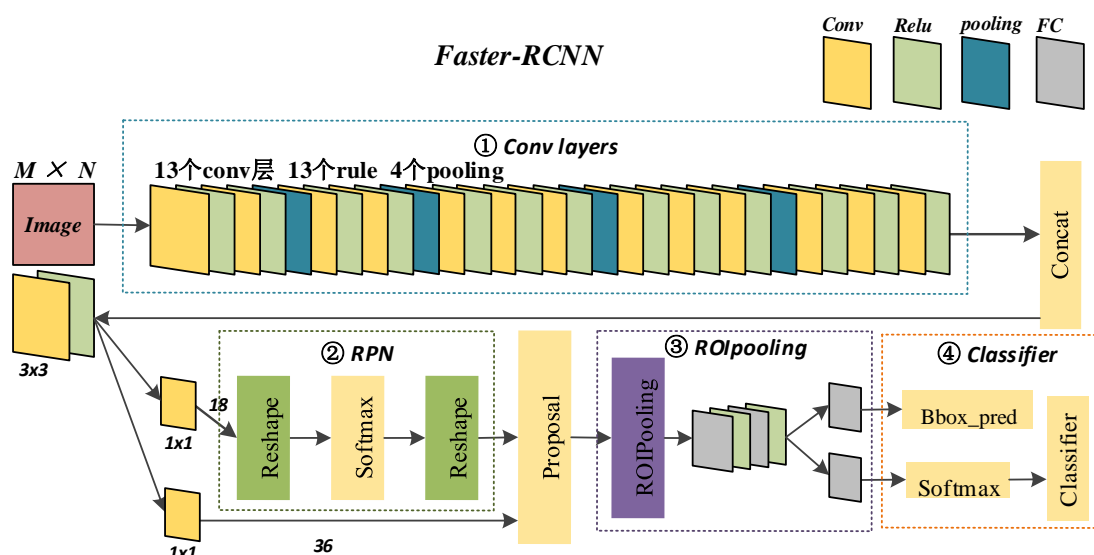


图 3.2 Faster-RCNN 模型

一般情况下, Faster-RCNN 的模型算法可分为三步:

Step1: 原始图像特征提取。

在图像分类模型中,深度残差网络比 VGGNet、GoogLeNet 在提升准确率上有很大的优势,因此在选择何种卷积神经网络处理图片特征时本文优先选择 ResNet101 作为提取图像特征的网络,虽然 ResNet152 有 152 层的残差网络,效果更好,但 ResNet101 已经能满足实验所需。故而本文使用在 ImageNet 数据集上预训练的 ResNet101 来进行基础图像特征的提取。以 448×448 的彩色图像作为模型的输入,利用卷积神经网络生成多个图像特征图。如图 3.2 所示,图像提取出多个特征图,为目标分类、属性分类和位置回归工作做准备。

Step2: 训练区域提案网络(RPN),定位目标区域。

模型首先生成多个边框，然后利用非极大抑制方法(NMS)与交并比(IOU)方法选择合适边框，并对选择的边框不断进行回归与位置精修。区域提案网络工作流程如图 3.3 所示。

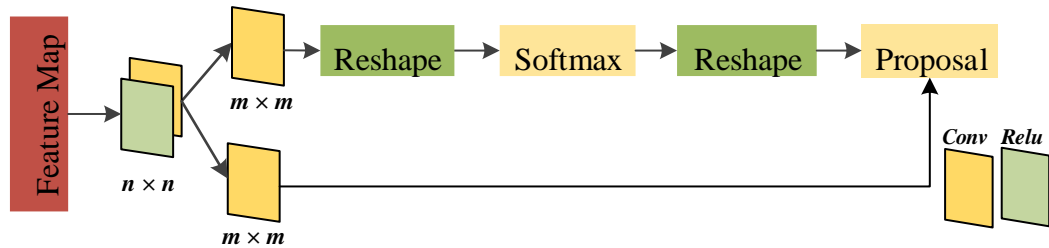


图 3.3 PRN 工作流程示意图

从图 3.3 中可以看出区域提案网络有两条工作路线，图像特征分别有两种处理方式，其一图像特征再次卷积操作重新调整大小，然后经分类函数对边框进行分类，分成带目标边框（图像前景）和不带目标边框（图像背景）。其二为了使模型获得更加准确的提案框，图像特征会再次通过卷积操作，计算边框回归的偏移量，删除不符合条件的提案框，边框或大或小或存在噪音等均被删除。经过上述操作，进而使得区域提案网络完成目标定位功能。

Step3: ROI 池化，固定边长输出。

全连接层需要固定大小的输入，因此要将卷积特征大小相异的候选框进行统一然后送入全连接层。ROI 池化的思想来自于 SPPNet^[56]，即将大小不同的输入，从水平与竖直方向均分成 n 等份，然后对每一份卷积特征进行最大池化 *max pooling* 处理。即使大小不同的区域处理过后，输出的结果均为 $n \times n$ ，从而实现固定边长输出。在本文中池化后边框大小定为 14×14 。

为了更好训练网络，使模型达到更好的效果，本文对模型输出和损失函数进行改进，本研究中目标检测网络损失计算如图 3.4 所示，为了模型能更好的识别出目标属性，获得质量更高的图像特征，所以在目标分类和边界框回归的基础上，增加目标属性的分类。固定维度的特征向量被用于三种操作，*cls_score* 层用于目标的分类操作，*bbox_predict* 层通过不断训练使得候选区域位置更加准确，*attr_score* 层用于目标属性的分类操作。

从图 3.4 中可以看出，这部分有三个代价函数，一是目标区域中物体分类的代价函数，二是边界框位置的代价函数，三是目标属性分类代价函数，它们对应的损失公式如公式(3.3)、(3.4)和(3.5)所示。

$$L_{cls} = -\log p_u \quad (3.3)$$

$$L_{loc} = \sum_{i=1}^4 g(t_i^u - v_i) \quad (3.4)$$

$$L_{attr} = -\log p_a \quad (3.5)$$

目标分类代价函数中，是由一个概率值 p_u 决定的，它代表的是图像中真实分类。

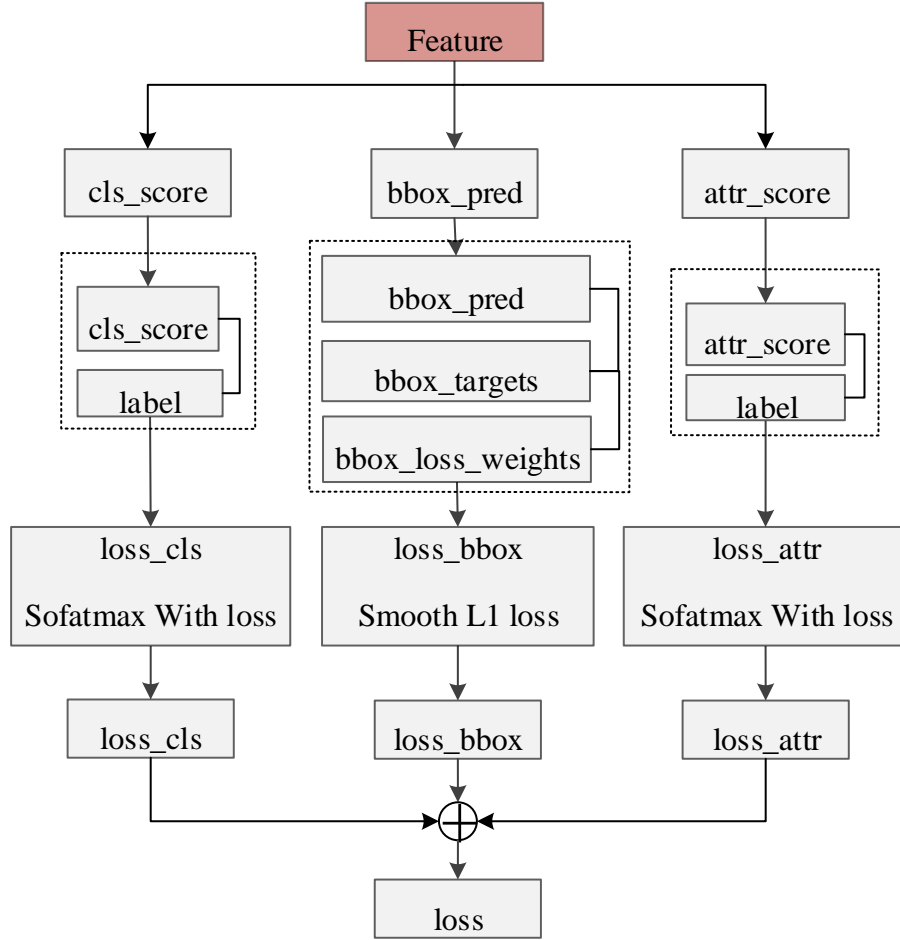


图 3.4 Faster-RCNN 损失计算图

在边界框位置代价函数中，函数 $g(x)$ 是一个损失函数，该函数公式见公式(3.6)， t^u 是一个预测的缩放参数，与真实分类相对应，而参数 v 表示的是真实的缩放参数，二者只差作为 $g(x)$ 函数的参数进行位置回归函数的损失计算。

$$g(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & |x| \geq 1 \end{cases} \quad (3.6)$$

为了更好训练参数，损失函数 $g(x)$ 使用的是 $L1$ 函数，由公式可以看出，该函数容错率较高，能够在训练过程中减少梯度爆炸的情况出现。在目标属性分类代价

函数中，函数值由概率值 p_a 决定。最后，Faster-RCNN 算法总损失是三者的加权和，如果分类为背景，则不考虑 IOC 代价函数，利用联合损失函数使得模型更容易收敛。入控制分类损失和回归损失的平衡，实验中分别把 λ 设为 0.1、0.5、0.75 和 1，研究发现 $\lambda=1$ 时效果最好。

$$L = \begin{cases} L_{cls} + \lambda L_{loc} & u \text{ is foreground} \\ L_{cls} & u \text{ is background} \end{cases} \quad (3.7)$$

根据多任务损失定义，可以推出 Faster-RCNN 的损失函数：

$$\begin{aligned} L\{(p_i), (u_i)\} = & \frac{1}{N_{cls}} \sum_i L_{cls}\{p_i, p_i^*\} + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}\{t_i, t_i^*\} \\ & + \frac{1}{N_{attr}} \sum_i L_{attr}\{a_i, a_i^*\} \end{aligned} \quad (3.8)$$

其中 p_i 为锚点(anchor)预测为目标的概率，真实答案标签可表示为：

$$p_i^* = \begin{cases} 0 & \text{negative label} \\ 1 & \text{positive label} \end{cases} \quad (3.9)$$

在公式(3.10)中， $L_{cls}\{p_i, p_i^*\}$ 是一个对数损失函数，代表目标区域与非目标区域的对数损失。

$$L_{cls}\{p_i, p_i^*\} = -\log[p_i^* p_i + (1 - p_i^*)(1 - p_i)] \quad (3.10)$$

$L_{reg}\{t_i, t_i^*\}$ 是回归损失， $t_i = \{t_x, t_y, t_w, t_h\}$ 是一个向量，表示预测的边界框的 4 个参数化坐标；其中变量 t_i^* 是真实答案的坐标，用公式(3.11)计算，这里的 R 是 *smooth L1* 函数。

$$L_{reg}(t_i, t_i^*) = R(t_i - t_i^*) \quad (3.11)$$

在属性预测损失函数部分， a_i 是模型预测出的目标物体属性， a_i^* 是目标属性真实分类。

3.4 模型实验设置与训练

3.4.1 模型的实验设置

1. 硬件配置

任何具有 12GB 或更大内存的 NVIDIA GPU 都可以训练 Faster-RCNN 和

ResNet-101。

2. 环境配置

Caffe、Python 3.6、PyTorch 0.4.1、cuda 9.0、cuDNN 7.0.4。

Caffe 是一款内部提供模板框架的开源软件平台框架，可以在 GPU 和 CPU 之间自由切换，具有最新的库版本，在视觉、语音、多媒体等领域的学术研究项目中均能看到 Caffe 的身影。Caffe 不仅是基于 C++/CUDA 的架构，而且还支持 MATLAB 接口、Python 接口和命令行模式。其优点有：

(1)上手快：以文本形式给出模型的解析。Caffe 的还会教用户如何定义模型、什么是模型的最优设置，并且提供其预训练权重。对于初识 Caffe 的人，极易理解与应用，入门级快。

(2)速度快：不管层数再深的模型都可以在上面运行，并且能够允许海量数据同时并行运算，例如测试 AlexNet 模型，若将 Caffe 配合 cuDNN 一起使用，处理一张图片仅仅只需 1.17ms，比其他的 TensorFlow 的处理速度快多了。

(3)模块化：平台架构具有很强的可拓展性，在原有项目中模块可扩展到新的任务与设置上。并且可以根据所需模型在 Caffe 提供的各层类型中自行定义。

3.4.2 模型训练步骤

1. 构建 Cython 模块
2. 建立 Caffe 和 pycaffe
3. 下载预先训练的模型，并将其置于 data\faster_rcnn_models。
4. 运行 tools/demo.ipynb 以在演示图像上显示对象和属性检测。
5. 运行 tools/generate_tsv.py 以将边界框特征提取到制表符分隔值(tsv)文件。
6. 重新创建每个图像具有 10 到 100 个特征的预训练特征文件，参数设置为 MIN_BOXES=10 和 MAX_BOXES=100。
7. 下载视觉基因组数据集。提取所有的 JSON 文件，以及图像目 VG_100K，并和 VG_100K_2 到文件夹 VGdata 中。
8. 视觉基因组数据集以 pascal voc 格式为每个图像生成 xml 文件。该脚本将提取最重要的 2500/1000/500 个对象/属性/关系，并对视觉基因组数据进行基本清理。但是请注意，本文的训练代码实际上仅使用 xml 文件中注释的子集，即仅基于

中找到的经过手工过滤的 vocab, 仅包含 1600 个对象类和 400 个属性 20 个关系, 文件位置为 data/genome/1600-400-20。关系标签可以包含在数据层中, 但当前不使用, 为特征融合模型扩充数据集做准备。

9. 经过 Faster-RCNN 训练所提取的图片特征信息以 npz 的文件形式进行存储, 在后面的模型训练中, 视觉问答的网络模型会直接读取 npz 文件, 不再需要同步处理提取图像特征从而占用内存和消耗大量的时间, 很大程度的提高了训练效率, 便于第四章、第五章的视觉问答模型进行训练。

3.4.3 模型训练结果

1. 训练说明

本文增加模型对于目标物体属性的分类, 并没有增强 Faster-RCNN 模型在目标检测评价指标 mAP 上的表现, 只是增加模型对于目标属性的识别能力, 因为本研究目的是为了获取到更高质量的图像特征, 为图像标注算法和视觉问答算法打基础, 并没有改进 Faster-RCNN 算法的准确性。在本工作中, 将训练好的目标检测网络作用在图像标注数据集中的图像数据集上, 将图像目标特征以 npz 文件保存, 为图像标注研究和视觉问答研究做基础。

存储图片特征的 npz 文件是由函数 np.savez() 输出的是一个扩展名为 npz 的压缩文件, np.savez() 函数的第一个参数是文件名, 其后的参数都是需要保存的数组, 使用 np.savez() 函数可以将多个数组保存到同一个文件中。传递数组时可以使用关键字参数为数组命名, 非关键字参数传递的数组会自动起名为 arr_0、arr_1、……、arr_n, 它包含多个与保存的数组对应的 .npy 文件 (由 save() 函数保存), 文件名对应数组名。读取 npz 文件时使用 np.load() 函数, 返回的是一个类似于字典的对象, 因此可以通过数组名作为关键字对多个数组进行访问。

以图片 ID 为 COCO_train2014_000000006809.jpg 的图片为例, 经过图片预处理模型以后图片特征便被封装为 COCO_train2014_000000006809.jpg.npz 的压缩文件, 其中包含有 x、image_w、image_h、bbox、num_bbox 四个参数的信息。x 代表图片特征向量信息, image_w 为图片的宽, image_h 为图片的高, bbox 为图片中每个对象的框信息, num_bbox 为图片中的框个数, npz 文件中具体的参数解析示例如图 3.5 所示。

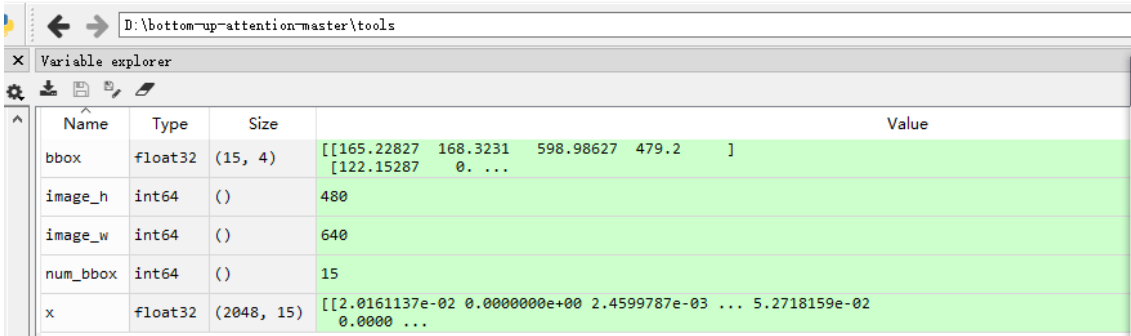
```
IPython console
Console 1/A

In [29]:
In [29]: runfile('G:/visualization of attention/6-1-npz.py', wdir='G:/visualization of attention')

In [30]: image_data6809.files
Out[30]: ['x', 'image_w', 'bbox', 'num_bbox', 'image_h']

In [31]: x = image_data6809['x']
In [32]: bbox = image_data6809['bbox']
In [33]: num_bbox = image_data6809['num_bbox']
In [34]: image_w = image_data6809['image_w']
In [35]: image_h = image_data6809['image_h']
```

图 3.5 npz 文件中所包含的参数示例



Name	Type	Size	Value
bbox	float32	(15, 4)	[[165.22827 168.3231 598.98627 479.2 122.15287 0. ...
image_h	int64	()	480
image_w	int64	()	640
num_bbox	int64	()	15
x	float32	(2048, 15)	[[2.0161137e-02 0.0000000e+00 2.4599787e-03 ... 5.2718159e-02 0.0000 ...

图 3.6 参数类型及大小示例

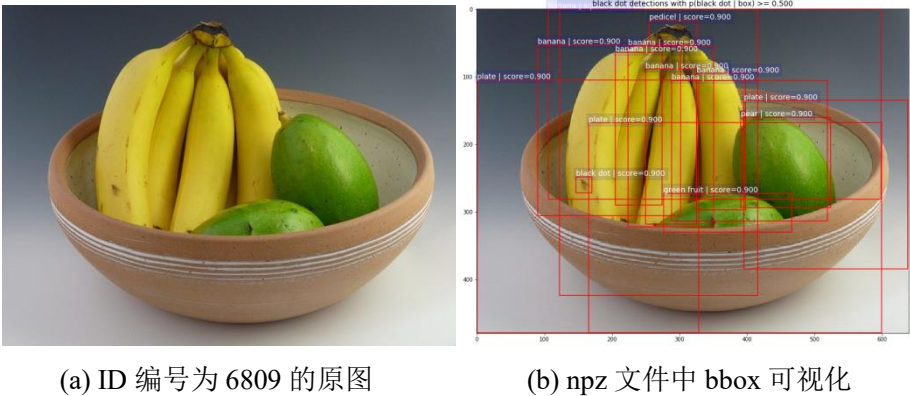
由图 3.6 可以看出，图片特征向量 X 中包含 15 个 2048 维的特征向量 x；该图片的宽 image_w 为 640；高 image_h 为 480；图片中一共有 15 个 bbox 为 15 行 4 列的矩阵，每一行代表一个 bbox 对象框，每个 bbox 有两个确定的坐标共四个参数形成一个矩形框；图片中的框个数 num_bbox 为 15。不管是在 VQA 数据集还是 CLEVR 数据集中，本文在进行图片预处理时,num_bbox 的取值大小设置为[10,100]，即每张图片会根据自己图中的对象复杂程度产生 10-100 个不等的矩阵框，以此实现“自适应”的图片特征需求。

如图 3.7 所示为图片 ID 为 COCO_train2014_000000006809.jpg.npz 中 bbox 中包含的矩阵向量示例，该矩阵示例包含了该图片中每一个图像特征的位置、大小信息，该图片 bbox 的大小为（15,4）15 行 4 列的数组矩阵。



图 3.7 bbox 中包含的矩阵向量示例

为了进一步将图片 ID 为 COCO_train2014_000000006809.jpg 提取特征以后形成的 npz 文件更加便于理解，本文加入了原图与预处理图片后的效果对比情况，如图 3.8 所示，可以很清晰的看出图片中产生了 15 个 bbox，每个 bbox 对应一个属性标签。



(a) ID 编号为 6809 的原图 (b) npz 文件中 bbox 可视化

图 3.8 特征提取后的可视化示例

2. 训练结果可视化

运行 tools/demo.py 以可视化训练数据和预测。如图 3.8 可以看出 Faster-RCNN

模型在处理图片时,针对不同图片提取的目标信息各有不同,充分体现了“自适应”的实验目的,更细致更智能的将图片中每一个特征信息用窗口框起来,这些特征窗口则以矩阵形式的文件封装起来。

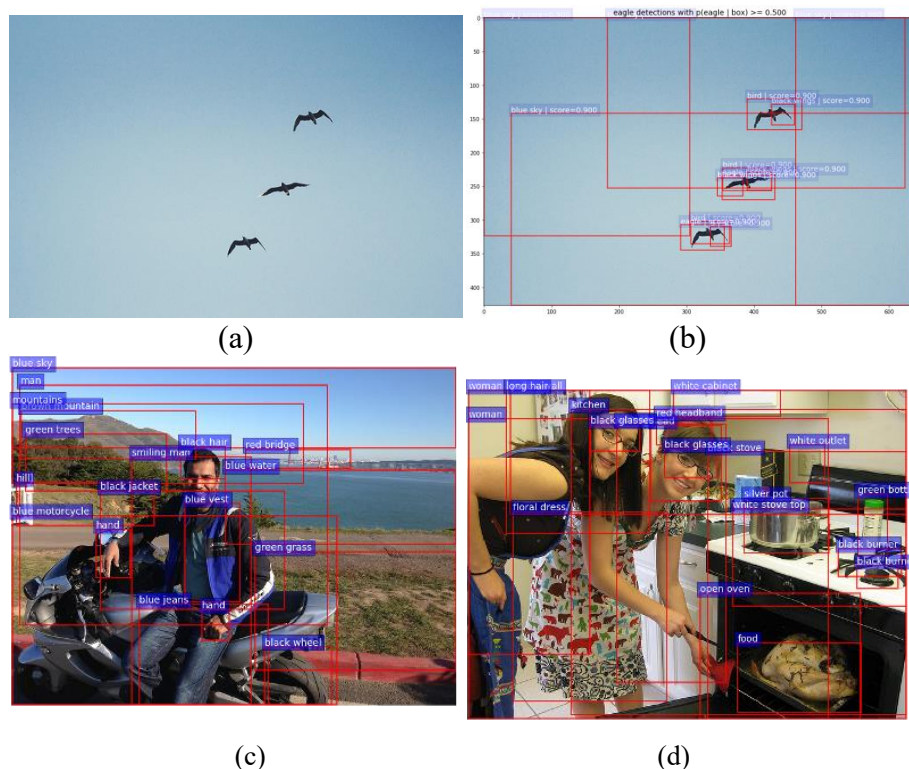


图 3.8 图片数据集可视化示例

图 3.8 中的(a)为数据集的原图, (b)、(c)、(d)为可视化示例。此图片只是用于展示 Faster-RCNN 预处理模型后的图片可视化, 每一个对象包围框都有一个对应的属性类, 但是在后续的模型训练研究中本文只是用它的特征向量并非预测标签。

3.5 本章小结

本章主要提出了一种基于目标检测算法的图像预处理模型, 采用 Faster-RCNN 与 Resnet101 相结合的方式处理图像信息。首先利用残差网络 Resnet101 提取图片中的全局图像特征, 然后根据目标检测算法 Faster-RCNN 来识别抽取图像的局部特征。本实验中并没有用 mPA 来评测目标检测模型的准确性, 因为本章的主要目的是得到图像特征编码后的 npz 文件, 用于参与第四、五章的视觉问答模型训练。

第4章 基于多模态特征融合的视觉问答

一般的视觉问答基线模型在四个模块中的处理均采用比较简单且为端到端的网络模型进行训练，一方面这样的训练方法在面对视觉问答任务时常常会耗费大量的训练、调参、模型消融的时间，同时也会占据大量的硬件装备存储空间；另一方面在处理两个不同模态的特征信息上存在一定的困难，传统的特征拼接方式会导致特征信息流失和语义理解差异。所以针对上述两方面：如何准确提取图片特征，提高训练的效率，节省研究时间？如何将图片特征与文本特征更好的融合在一起清晰的表达图片信息与文本语义？便是本章着重解决的问题。本章主要的研究内容是采用先进的卷积神经网络和残差网络对图片进行特征提取，然后利用单词的全局向量训练文本信息，其次对不同的特征融合方式逐一训练旨在找到更适合的融合模型，最后在 VQA v2 数据集中进行了测试与评估验证了多模态融合算法的有效性。

4.1 视觉问答整体架构

一个完整的视觉问答框架包括你：图像特征编码模块、文本特征编码模块、多模态特征融合模块、答案预测模块。本节主要简述了每个模块运用的方法模型及其主要算法步骤。

1. 图像特征编码模块：采用 Faster-RCNN 与 Resnet101 相结合的方式处理图像信息，提取图像特征向量，形成自适应的预训练特征文件，该文件内存储的是已经提取出的图像目标特征，其中包含图片中目标对象信息、位置关系、逻辑关系等，便于后续与文本特征进行特征融合训练。

2. 文本特征编码模块：将自然语句的问题文本编码成 *one-hot* 词向量特征，结合 300 维的 GloVe 词向量模型捕捉单词的语义特征，再利用 LSTM 网络对文本特征进行编码从而抽取问题语义特征信息。

3. 多模态特征融合模块：将已经识别出的图像目标特征与经处理后的问题特征联合嵌入到同一个特征空间，通过多模态分解双线性池化(MFB)特征融合算法将图片特征和文本特征相融合。

4. 答案预测模块：将答案预测作为分类问题，将 SoftMax 函数作为分类器在候选答案集中进行答案预测输出。

如图 4.1 为视觉问答的整体框架图。

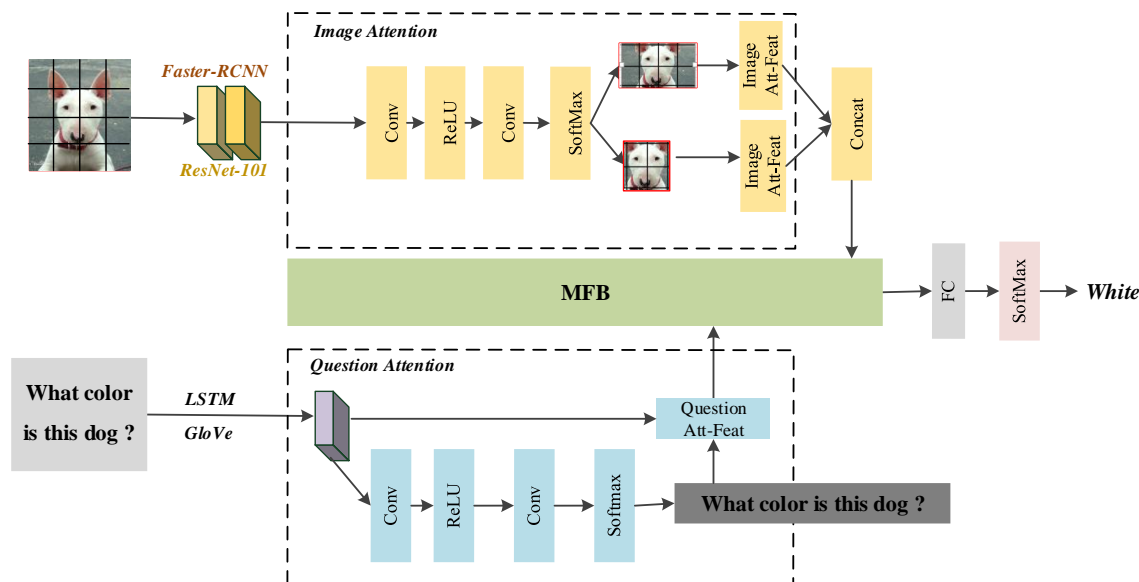


图 4.1 视觉问答整体框架

4.2 文本信息处理模块

在本节中主要介绍两种文本处理方式，其一是词向量模型 GloVe，其二是长短期记忆网络。首先利用 GloVe 词向量模型在训练过程中使用全局信息更加准确的捕获单词的语义信息，然后使用 LSTM 网络对文本特征进行编码。

4.2.1 GloVe 模型

训练词向量的方法有很多种，当前运用比较广泛的是局部上下文窗口的方法和基于全局矩阵分解的方法，例如 T. Mikolov 等人在 2013 年提出来的词袋模型与 skip-gram 模型^[41]就是基于上下文窗口的方法，LSA(Latent Semantic Analysis)是基于全局矩阵分解的方法。上述两种方法各有优缺点，CBOW 和 skip-gram 模型采用局部上下文窗口方法虽可以进行词汇类比，但是不能很好地利用全局词汇共现信息；LSA 虽可以统计全局词汇信息进行利用，但是词汇类比表现不佳。

为了弥补上述两种方法的不足，2014 年 R. Nithyanand 等人^[40]提出了一种基于全局词频统计的词向量表征工具 GloVe(Global Vectors for Word Representation)，其

主要是把文本中的每一个单词表征为一个实数向量,该向量可以将单词之间的语义信息如类比性、相似性等进行表征,这种方法成功的将全局词汇共现信息进行统计从而达到提升表征效果的目的。以下便对 GloVe 模型的原理进行说明。

令词汇的共现矩阵为 X , 则 X_{ij} 表示在第 i 个词汇上下文中出现次数的总和, 用 $P_{ij} = P(j/i) = X_{ij}/X_i$ 表示第 j 个词汇在第 i 个词汇上下文中出现的概率。

GloVe 的作者猜想是否可以通过训练词向量,使得词向量在函数 F 的计算之后得到 $P_{ij} = P(j/i) = X_{ij}/X_i$ 的比值, 具体如下:

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}} \quad (4.1)$$

其中, w_i, w_j, \tilde{w}_k 为词汇 i, j, k 对应的词向量, 它们的维度均表示为 d , F 在这里暂时为未知函数由后文推导得出, P_{ik}/P_{jk} 可由语料库计算而来。词向量均属于同一线性向量空间, 故而可对 w_i, w_j 进行差分操作, 将公式(4.1)转变为:

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}} \quad (4.2)$$

由于公式(4.2)中 $w_i - w_j, \tilde{w}_k$ 是维度为 d 的词向量, P_{ik}/P_{jk} 是一个标量, 可用向量的内积解决二者之间的差异, 因此, 可进行以下操作:

$$F\left\{(w_i - w_j)^T \tilde{w}_k\right\} = F\left(w_i^T w_k - w_j^T w_k\right) = \frac{P_{ik}}{P_{jk}} \quad (4.3)$$

由公式(4.3)的构造提示, 作者联想到用指数计算, 故而将函数 F 指定为指数函数, 则有:

$$\exp\left(w_i^T w_k - w_j^T w_k\right) = \frac{\exp(w_i^T w_k)}{\exp(w_j^T w_k)} = \frac{P_{ik}}{P_{jk}} \quad (4.4)$$

同时, 要确保公式(4.4)中两边的分子分母对应相等, 即有:

$$\exp(w_i^T w_k) = P_{ik}, \exp(w_j^T w_k) = P_{jk} \quad (4.5)$$

然后, 将语料库中的所有词汇进行转化, 考察 $\exp(w_i^T w_k) = P_{ik} = X_{ik}/X_i$, 即

$$w_i^T w_k = \log\left(\frac{X_{ik}}{X_i}\right) = \log X_{ik} - \log X_i \quad (4.6)$$

因为公式(4.6)中的 $w_i^T w_k$ 具有对称性, 即调换 i 和 k 的值对结果无影响, 因此为了确

保等式右侧同样具有对称性，在参数中引入两个偏置项，即

$$w_i^T w_k = \log X_{ik} - b_i - b_k \quad (4.7)$$

其中， $\log X_{ik}$ 已经被纳入在 b_i 的计算中。

为使公式(4.7)等式两边尽量相等，此时模型转为学习词向量的表示，目标函数则可定为二者的平方差：

$$J = \sum_{i,k=1}^V \left(w_i^T \tilde{w}_k + b_i + b_k - \log X_{ik} \right)^2 \quad (4.8)$$

但是该目标函数会对全部的共现词汇使用相同的权重信息，进而作者在此基础上做出了修正，方法是根据词汇共现统计信息修改目标函数中的权重，具体操作如下：

$$J = \sum_{i,k=1}^V f(X_{ik}) \left(w_i^T \tilde{w}_k + b_i + b_k - \log X_{ik} \right)^2 \quad (4.9)$$

其中， V 表示词汇的数量。

同时，权重函数 f 必须满足以下三个要求：

- (1) $f(0)=0$ ，当词汇共现的次数为 0 时，此时对应的权重应该为 0。
- (2) $f(x)$ 为非减函数，这样才能保证当词汇共现的次数越大时，其权重不会出现下降的情况。
- (3) 针对出现频率过多的词汇，函数 $f(x)$ 给它们设置的权重信息要稍微小一点，从而避免过度加权的情况发生。

为满足上述三点特性，作者制定了相应的权重函数：

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} \quad (4.10)$$

其中， x_{\max} 设定为 100，且 $\alpha=3/4$ 时效果较好，由作者在实验中得出此结论。

Glove 模型集合了局部窗口上下文方法和全局词汇信息共现统的优势，是这两种方法的综合体现，与基于全局矩阵分解的方法相比，GloVe 模型更不需要统计共现次数为 0 的单词，最大限度的减少了运算量与数据存储空间。

4.2.2 长短时记忆网络

循环神经网络通常被用于处理文本信息，但是由于它的结构相对简单，只能够处理较短的文本问题，面对较长序列编码时往往会丢失一些重要的语义信息。研究

工作者为了解决循环神经网络的这一梯度消失问题, S. Hochreiter 与 J. Sutskever 等人^[57]提出一种改进版的 RNN 网络, 即长短时记忆网络(LSTM), 在原网络中的隐藏状态 h 的基础上加入一个状态单元 c , 如图 4.2 所示。

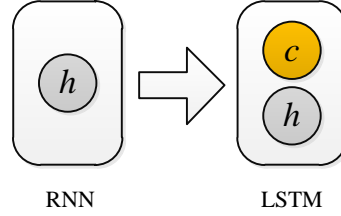


图 4.2 RNN 演变为 LSTM 的过程

与传统 RNN 相比, LSTM 在网络中多加入的状态单元 c , 其实就是增加了三个控制器, 也成为“门”, 分别是遗忘门、输入门、输出门。

将输入的文本序列设为 $x = (x_1, x_2, \dots, x_n)$, n 为文本序列的长度, 如图 4.3 所示, 从左往右, 遗忘门控制上一时刻的单元状态 c_{t-1} ; 输入门控制当前时刻的网络输入 x_t ; 输出门控制当前时刻的单元状态 c_t 。状态单元 c 其本质为全连接层, 输入输出均为向量, 输出为 $[0,1]$ 的实数向量。假设 W 为门的权重向量, b 为偏置项, 则一个最基础的门可表示为:

$$g(x) = \sigma(Wx + b) \quad (4.11)$$

由此类推, 长短时记忆网络的门单元表示及输出如下所示:

$$\text{遗忘门:} \quad f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \quad (4.12)$$

$$\text{输入门:} \quad i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (4.13)$$

$$\text{输出门:} \quad o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (4.14)$$

$$\text{单元状态:} \quad c_t = f_t * c_{t-1} + i_t * \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (4.15)$$

$$\text{单元输出:} \quad h_t = o_t * \tanh(c_t) \quad (4.16)$$

其中, f_t , i_t , o_t 为三个门对应的控制状态, σ 为 *sigmoid* 激活函数, x_t 表示语句中的第 t 个单词向量, W 和 b 分别表示三个门所对应的权重向量和偏置项, 如 W_{fx} 和 b_f 表示遗忘门的权重矩阵与偏置量。

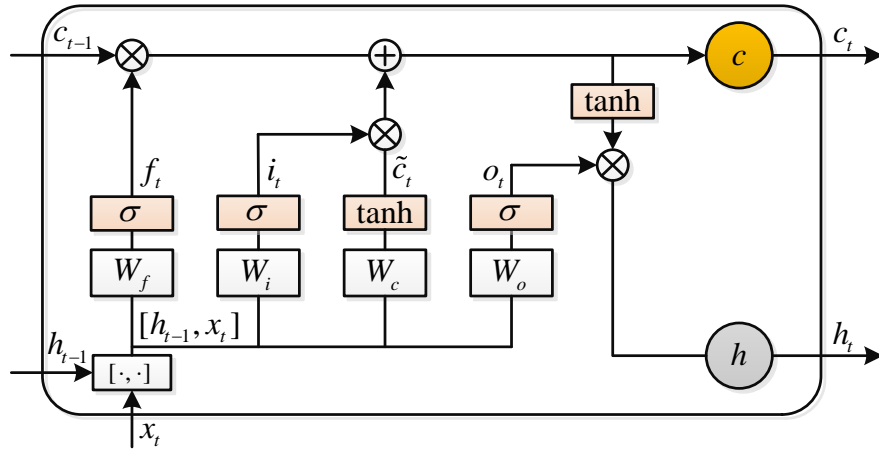


图 4.3 LSTM 单元组成结构

本文在问题编码模块采用预训练的 GloVe 词向量模型来表示问题，该模型可利用词嵌入矩阵将高维度空间的张量转化为低维度的空间向量，从而计算空间中两个向量的距离，它们之间的距离越短，则证明两个向量越相似。然后采用 LSTM 网络对这些问题进行编码处理，上述单元状态 c 中的输入和输出分别为 $x_t = W_D * q_n$ 、 $h_t = LSTM(x_t)$ 。

此外，并不是所有文本语料库中问题 Q 都和 GloVe 模型中的词向量表示成一对对应关系，因此 GloVe 训练模型中没有向量表示的单词可以通过取均值的方式进行处理。

4.3 多模态特征融合模块

多模式特征融合在 VQA 中扮演着重要而有趣的角色。获得图像和问题特征后，级联或逐元素求和最常用于多模式特征融合。由于两个特征集在不同模态下的分布（即图像和图像的视觉特征问题的文字特征）可能会有很大差异，融合特征的代表能力可能是不足，限制了最终的预测性能。A. Fukui 等人^[13]首先介绍了双线性模型来解决 VQA 中的多模式特征融合问题。在与上述方法相反，他们提出多模式紧凑型双线性池(MCB)，使用两个特征向量的外积来产生二次扩展的高维特征。为了降低计算成本，他们使用了基于样本的近似方法来利用该属性两个向量的投影可以表示为它们的卷积。MCB 模型优于简单模型融合方法并表现出卓越的性能在 VQA 数据集上。不过，MCB 通常需要高尺寸特征以确保强大的性能，由于 GPU 内存的限制，可能会严重限制其适用性。为了克服这个问题，J. H. Kim 等人^[44]提

出了多模式低秩双线性池(MLB)方法基于两个特征向量的 **Hadamard** 积（即图像特征和问题特征）。

与线性模型相比，双线性模型提供了丰富的表示。它们已被应用于各种视觉任务，如对象识别、分割和视觉问答。最先进的性能利用扩展的表示。然而，双线性表示往往是高维的，限制了对计算复杂的适用性。

4.3.1 双线性池化模型

很多多模态任务，比如 VQA、视觉定位等，都需要融合两个模态的特征。特征融合即输入两个模态的特征向量，输出融合后的向量。最常用的方法是拼接(concatenation)、按位乘(element-wise product)、按位加(element-wise sum)。MCB 的作者认为这些简单的操作效果不如外积(outer product)，不足以建模两个模态间的复杂关系。但外积计算存在复杂度过高的问题。 n 维的向量，外积计算得到 n^2 的向量。于是 MCB 被提出，MCB 将外积的结果映射到低维空间中，并且不需要显式计算外积。

双线性(Bilinear)就是向量外积的计算。双线性池化(Bilinear Pooling)是对双线性融合后的特征进行池化。L. Tsungyu 等人^[58]的做法是首先对卷积得到的 feature map 的每个位置的特征向量进行向量外积计算，再对所有位置外积计算的结果进行 sum pooling 得到特征向量 x 。 x 经过 *signed square root* 和 *L2 normalization* 得到最后的特征。但是双线性特征的维度是极高的，紧凑双线性池化(Compact Bilinear Pooling, CBP)是对双线性池化的一种降低维度的近似^[59]。

双线性池化的操作可以表示为：

$$B(x) = \sum_{s \in S} x_s x_s^T \quad (4.17)$$

在线性核的情况下有：

$$\begin{aligned} \langle B(x), B(y) \rangle &= \left\langle \sum_{s \in S} x_s x_s^T, \sum_{u \in U} y_u y_u^T \right\rangle \\ &= \sum_{s \in S} \sum_{u \in U} \langle x_s x_s^T, y_u y_u^T \rangle \\ &= \sum_{s \in S} \sum_{u \in U} \langle x_s, y_u \rangle^2 \end{aligned} \quad (4.18)$$

因为：

$$\begin{aligned}
\langle B(x), B(y) \rangle &= \sum_{s \in S} \sum_{u \in U} \langle x_s, y_u \rangle^2 \\
&\approx \sum_{s \in S} \sum_{u \in U} \langle \phi(x), \phi(y) \rangle \\
&\equiv \langle C(x), C(y) \rangle
\end{aligned} \tag{4.19}$$

对多项式核的进行低维近似的映射函数 Φ ，可以用来做对双线性池化的压缩。Tensor Sketching^[60]是一种近似多项式核的算法，可以用 Tensor Sketching 进行压缩。

4.3.2 MLB 融合模型

双线性模型使用线性变换的二次展开，考虑每一对特征。

$$f_i = \sum_{j=1}^N \sum_{k=1}^M w_{ijk} x_i y_k + b_i = x^T W_i y + b_i \tag{4.20}$$

其中 x 和 y 是输入向量， $W_i \in \mathbb{R}^{N \times M}$ 是输出 f_i 的权重矩阵， b_i 是输出 f_i 的偏置。注意，参数的数目是 $L \times (N \times M + 1)$ ，包括偏置向量 b 其中 L 是输出特征的数目。

P. Hamed 等人^[61]建议采用低秩双线性方法，以减少权重矩阵 W_i 的秩，使其具有较少的正则化参数。他们将权重矩阵改写为 $W_i = U_i V_i^T$ 其中 $U_i \in \mathbb{R}^{N \times d}$ 和 $V_i \in \mathbb{R}^{M \times d}$ ，这限制了 W_i 的等级最多为 $d < \min(N, M)$ 。

基于这个想法， f_i 可以重写如下：

$$f_i = x^T W_i y + b_i = x^T U_i V_i^T y + b_i = 1^T (U_i^T x \circ V_i^T y) + b_i \tag{4.21}$$

其中 $1 \in \mathbb{R}^d$ 表示一列向量， \circ 表示 Hadamard 乘积。然而，对于一个特征向量 f ，我们需要两个三阶张量， U 和 V ，其元素是 $\{f_i\}$ ，为了将权重张量的阶数减少 1，我们用 $P \in \mathbb{R}^{d \times c}$ 替换 1 和 $b \in \mathbb{R}^c$ 替换 b_i ，然后，重新定义为 $U \in \mathbb{R}^{N \times d}$ 和 $V \in \mathbb{R}^{M \times d}$ 得到投影特征向量 $f \in \mathbb{R}^c$ 。然后，可得到：

$$f = P^T (U^T x \circ V^T y) + b \tag{4.22}$$

其中 d 和 c 是决定联合嵌入维数和低秩双线性模型输出维数的超参数。

公式(4.22)中的一个低秩双线性模型可以使用两个线性映射来实现，而不存在嵌入两个输入向量的偏差，Hadamard 乘积可以学习乘法中的联合表示方法，以及具有偏置的线性映射，将联合表示投影到给定输出维数的输出向量中。然后，将该结构作为深度神经网络的池化方法。现在，我们讨论了基于该模型的低秩双线性池

的可能变化，该模型受神经网络研究的启发。

在公式(4.22)中线性投影 U 和 V ，可以有自己的偏置向量。因此，每个输入向量 x 和 y 的线性模型被整合成一种加性形式，称为全模型。

$$\begin{aligned} f &= P^T \left[(U^T x + b_x) \circ (V^T y + b_y) \right] + b \\ &= P^T \left\{ U^T x \circ V^T y + \left[\text{diag}(b_y) U^T \right] x + \left[\text{diag}(b_x) V^T \right] y \right\} + \left[b + P^T (b_x \circ b_y) \right] \quad (4.23) \\ &= P^T (U^T x \circ V^T y + U'^T x + V'^T y) + b' \end{aligned}$$

其中， $U'^T = \text{diag}(b_y) U^T$ ， $V'^T = \text{diag}(b_x) V^T$ ， $b' = b + P^T (b_x \circ b_y)$ 。

4.3.3 MFB 融合模型

给定两个不同模式下的特征向量，例如图像的视觉特征 $x \in \mathbb{R}^m$ 和问题的文本特征 $y \in \mathbb{R}^n$ ，定义最简单的多模态双线性模型如下：

$$z_i = x^T W_i y \quad (4.24)$$

其中 $W_i \in \mathbb{R}^{m \times n}$ 是投影矩阵， $z_i \in \mathbb{R}$ 是双线性模型的输出。这里省略了偏置项，因为它隐含在 W 中。要获得一个 c 维输出 z ，我们需要学习 $W = [W_1, \dots, W_o] \in \mathbb{R}^{m \times n \times o}$ 。虽然双线性池可以有效地捕获特征维度之间的成对相互作用，但它也引入了大量的特征参数可能导致高计算成本和过度拟合的风险。受单模态数据矩阵分解技巧的启发，公式(4.24)中的投影矩阵 W_i 可以分解作为两个低秩矩阵：

$$\begin{aligned} z_i &= x^T U_i V_i^T y = \sum_{d=1}^k x^T u_d v_d^T y \\ &= 1^T (U_i^T x \circ V_i^T y) \end{aligned} \quad (4.25)$$

其中 k 是因式分解矩阵 $U_i = [u_1, \dots, u_k] \in \mathbb{R}^{m \times k}$ 和 $V_i = [v_1, \dots, v_k] \in \mathbb{R}^{n \times k}$ 的因子或潜在维数， \circ 是 Hadamard 乘积或两个向量的元素乘法， $1 \in \mathbb{R}^k$ 是一个全一向量。

通过公式(4.24)获得输出特征 $z \in \mathbb{R}^o$ ，需要相应地学习的权重是两个三阶张量 $U = [U_1, \dots, U_o] \in \mathbb{R}^{m \times k \times o}$ 和 $V = [V_1, \dots, V_o] \in \mathbb{R}^{n \times k \times o}$ 。在不失去通用性的情况下，我们可以将 U 和 V 重新表述为二维矩阵 $\tilde{U} \in \mathbb{R}^{m \times ko}$ 和 $\tilde{V} \in \mathbb{R}^{n \times ko}$ 与简单的重塑操作。因此，公式(4.24)可以改写如下：

$$z = \text{SumPooling}(\tilde{U}^T x \circ \tilde{V}^T y, k) \quad (4.26)$$

其中, 函数 $SumPooling(x, k)$ 意味着使用具有大小 k 的一维非重叠窗口在 x 上执行和池。这就是多模态分解双线性池(MFB)模型。

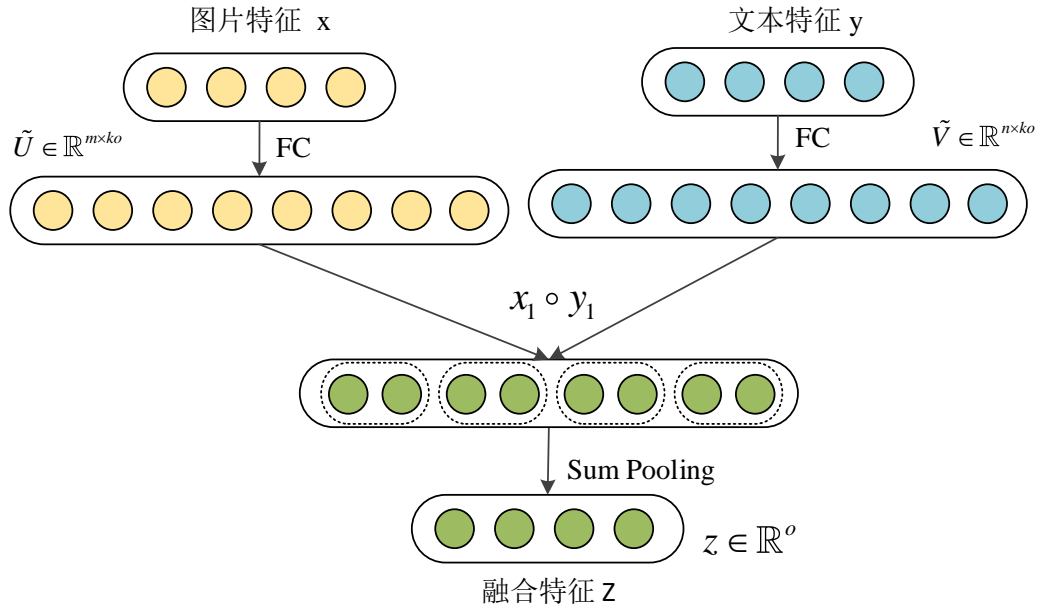


图 4.4 多模态分解双线性池(MFB)

该方法可以很容易地实现, 结合一些常用的层, 如完全连接, 元素乘法和池化层。此外, 为了防止过度拟合, 在元素级乘法层之后添加了一个 **dropout** 层。由于引入了元素乘法, 输出神经元的大小可能会有很大的变化, 并且模型可能收敛到不令人满意的局部最小值。因此, 类似于文献^[58], 在 **MFB** 输出后附加幂归一化 ($z \leftarrow sign(z)|z|^{0.5}$) 和 ℓ_2 归一化 ($z \leftarrow z/\|z\|$) 层。公式(4.26)表明 **MLB** 在公式(4.23)中是具有 $k=1$ 的 **MFB** 的一个特例, 它对应于秩为 1 因式分解。**MFB** 可以分解为两个阶段: 首先, 将来自不同模态的特征扩展到高维空间, 然后与逐元素乘法集成。此后, 执行求和合并后再进行归一化层, 以将高维特征压缩为紧凑的输出特征, 而 **MLB** 将特征直接投影到低维输出空间并执行逐元素乘法。因此, 在输出特征尺寸相同的情况下, **MFB** 的表示能力要比 **MLB** 强大。

4.4 实验结果及分析

为了验证图像预处理的方式适用于本文提出的视觉问答系统, 同时能在多种特征融合方式中找到最优的特征融合模型, 进一步将跨模态特征完美融合, 在本节中, 首先介绍了视觉问答数据集 **VQA v2** 及其相关的评价指标和模型评估方式, 然

后详细地介绍了实验过程中各个模型所涉及到的实验设置，最后将本文提出的视觉问答模型与近年来优秀的模型做比较。从实验结果可以看出本文所提出的视觉问答框架比基线模型和其他一般的视觉问答模型具有更加优秀的成绩。

4.4.1 VQA 数据集

数据集 VQA v2.0 是 VQA v1.0 的第二个版本，图片主要是由微软提出的 MS-COCO 数据集构成，共计 204721 张图组成，其中每张图片设置 3~100 个不等的问题，每个问题对应 10 个答案（这 10 个答案中有些是相互重复的答案），文本数据集中还包括训练互补对作为补充文本数据。



图 4.5 COCO_train2014_000000578958.jpg

如图 4.5 是图片 ID 为 COCO_train2014_000000578958.jpg 的图片内容，与之对应的文本数据集分别是问题数据集和答案数据集，该图片中只含三个问题，每个问题有相应的问题 ID，每个问题对应不同的问题类型且包含 10 个答案，同时每个答案也有属于自己的编号。如下字段为文本数据集中截取出来的部分文本，格式保留了数据集中的文本格式，其中该三个问题所对应的 10 个答案均相同，故而答案部分只写了其中一个答案表示。

Image id: 578958

Question:

"question_id": 578958000: "Are the sheep babies?"

"question_id": 578958001 : "What plant is growing beneath the sheep's feet?"

" question_id": 578958002 : "How many sheep are grazing?"

Answer:

"question_id":578958000,"multiple_choice_answer":"yes","question_type":"are the","answer_type":"other"

"question_id":578958001,"multiple_choice_answer":"grass",
"question_type":"what","answer_type":"number"

"question_id":578958002,"multiple_choice_answer":"6","question_type":"how many","answer_type":"yes/no"

VQA 数据集中共有训练集（图像 82783 和问题 443757）、验证集（图像 40504 和问题 214354）和测试集（图像 81434 和问题 447793），共计 204721 张图片、1105904 个问题、6581110 个答案。表 4.1 为 VQA v2 数据集的统计情况。

表 4.1 VQA v2 数据集的统计情况

数据集	图片	问题	答案	互补对
训练集	82783	443757	4437570	200394
验证集	40504	214354	2143540	95144
测试集	81434	447793	—	—
总计	204721	1105904	6581110	295538

4.4.2 评价指标

深度学习的模型中，评价指标是衡量一个模型性能是否稳定，是否具有良好鲁棒性的关键。在视觉问答任务中，VQA v2.0 数据集为自己量身定制了一种新的评价指标，并且将其称为“正确率”。在该数据集中，每个问题均配置了 10 个答案，这些答案是由不同的注释者给出的，故而工作人员便将这 10 个答案和模型给出的预测答案进行比对。

$$Acc(a) = h(a, T) = \min \left\{ \frac{\sum_i 1\{a = T_i\}}{3}, 1 \right\} \quad (4.27)$$

其中 T 为不同注释者给出的答案组成的集合。

若以上述指标的评估标准，可认为只要模型预测出来的答案和标注的答案（3 个及 3 个以上）相吻合，则认为答案预测准确。即模型预测答案与人工标注的答案一致，判定为模型完全正确；相反则判定模型部分正确。每种类型问题的得分均值便为该种问题的“正确率”。这种评价标准能够客观准确的评价模型，并保证视觉问答任务评估结果的可靠性和有效性。

4.4.3 实验设置

1. 硬件要求

本工作使用 windows10 操作系统的服务器，该服务器的容量参数为：显存容量 24GB、内存容量 16GB、机械硬盘容量 4TB、固态硬盘容量 256GB，硬件装备满足实验要求并且为后续研究提供了方便快捷的实验环境。GPU 为专业级显卡，不仅可以提高模型的训练速度，而且还可以节约时间成本。

2. 软件要求

Python 3.6、PyTorch 0.4.1、cuda 9.0、cuDNN 7.0.4。

3. 实验设置

(1) 实验一：融合模型+图像预处理

该实验的主要目的是将两种比较先进的特征融合模型与经图像预处理后的图片数据集直接通过端到端的模式进行训练。图片预处理过程在第三章内已经完成，因此实验中的参数设置和实验二相同。

(2) 实验二：最佳融合模型+其他网络模型

该实验模块主要是研究 MFB 模型与其他网络模型结合训练的结果是否能提高视觉问答任务的准确率，一共设置了 4 组实验。

MFB+GloVe：融合特征向量的维度设置为 1024，采用 300 维的 GloVe。

MFB+LSTM：融合特征向量的维度设置为 1024，LSTM 的隐层节点数量设置为 512，采用双层的 LSTM 网络。

MFB+GloVe+LSTM：同前两个模型的设置。

MFB+GloVe+LSTM +VG：增加了视觉基因组数据集。

每组实验的 Epoch 设置为 20，每个 Epoch 平均消耗 800~1100 秒，共耗时 6~9 小时，根据加载模型的不同，训练时间也会相应变化。Batch_size 为 64，每个 batch 为 6933，实验一共迭代了 443712 次，模型损失值接近于 0 并且达到稳定状态，说明训练成功。Adam 求解器设置为 $\beta_1 = 0.9$ ， $\beta_2 = 0.99$ 。基础学习率设置为 0.0007，每 4 万次迭代衰减一次，指数速率为 0.5。在每个 LSTM 层(dropout=0.3)和 MFB 模型之后使用 dropout 层。

4.4.4 实验结果分析

在本节中，主要分析两个实验的结果。实验一先确定与图片预处理模型相匹配的特征融合模型，实验二对最优特征融合模型进行消融实验，验证文本处理模块模型的合理性有助于提升视觉问答任务的性能。从实验结果得以证实本文提出的基于特征融合模型的视觉问答系统有良好的表现力。

1. 实验一：融合模型+图像预处理

设置本实验的目的是寻找与 Faster R-CNN 目标检测网络相匹配的最优融合模型。下表为各融合模型在 VQA v2 数据集的训练情况对比。

表 4.2 各融合模型在 VQA v2 数据集的训练情况

序号	模型	Accuracy%			
		Y/N	Number	Other	All
1	MCB ^[13]	81.2	35.1	49.3	60.8
2	MLB ^[44]	84.02	37.9	54.77	65.07
3	MFB ^[15]	82.5	38.3	55.2	64.6
4	MLB+Image Preprocessng	82.38	44.52	56.69	64.75
5	MFB+Image Preprocessng	83.54	46.03	57.7	65.87

注*：表 4.2 中 1-3 行的数据均来自于近年的参考文献。

如表 4.2 各融合模型在 VQA v2 数据集的训练情况，第 4、5 行是本文对图片数据集进行预处理，提取每张图片的自适应特征以后，再结合特征融合模型在 VQA v2 数据集上做验证实验，以下是对表格内容的分析总结：

(1) 比较第 1、2、3 行的模型可以看出多模态低秩双线性池化模型 MLB 和多模态分解双线性池化模型 MFB 在四个问题类型的准确率上表现亮眼且远远高于多模态紧凑型双线性池化模型 MCB，故后续操作中没有与 MCB 进行对比实验。

(2) 模型“MLB+Image Preprocessng”相较于 MLB 模型，只有在“Number”和“Other”这两种问题类型的准确率上有明显提升，分别提高了 6.62%和 1.92%，但是在“Y/N”和“All”类型的准确率反而降低了。

(3) 从模型“MFB+Image Preprocessng”与 MFB 的这两组数据的比较可以看出图像预处理模块与 MFB 模型配合度极高，对视觉问答系统的性能有很大的提升，相较于原模型四个类型的问题准确率均有上升，特别是针对“Number”类型的问题提高了 7.73%的准确率，结合效果明显优于“MLB+Image Preprocessng”模型。

由实验一的结果可以看出本文提出的图片预处理模块与 MFB 模型相结合后,视觉问答模型提升的准确率最高,综上可得,MFB 更适合用于本实验中图片预处理的模块,进行端到端的视觉问答模型训练。本文将采用 MFB 作为我们实验二中的特征融合模型。

2. 实验二: 最佳融合模型+其他网络模型

以实验一的训练结果为基础,在该节实验中,本文进行了四次消融实验,主要目的是完善本文提出来的视觉问答系统模型,其次是证明加载词向量模型 GloVe 的准确性,然后将长短时记忆网络应用在视觉问答模型中提升系统的性能,最后增加额外的视觉基因组数据集 Visual Genome(VG)作为训练补充。

各基线模型及本文模型的最终实验结果如表 4.3 所示。

表 4.3 本文模型及其他模型在 VQA v2 数据集上的实验结果对比

序号	模型	Test-dev			
		Yes/No	Number	Other	All
1	VAQ-team ^[45]	80.5	36.8	43.1	57.8
2	MRN ^[14]	82.3	38.4	49.3	61.7
3	NMN ^[34]	81.2	38.0	44.0	58.6
4	SAN ^[18]	79.3	36.6	46.1	58.7
5	MCB ^[13]	82.2	37.7	54.9	64.2
6	DAN ^[62]	83.0	39.1	53.9	64.3
7	MUTAN ^[63]	83.6	39.4	54.4	64.7
8	MLB ^[44]	84.1	38.2	54.9	65.1
9	MFB ^[15]	84.0	39.8	56.2	65.9
10	MLBP ^[64]	83.9	39.7	56.5	65.9
11	MFB(ours)	83.54	46.03	57.7	65.87
12	MFB+GloVe(ours)	83.52	46.13	57.74	65.90
13	MFB+LSTM(ours)	83.49	46.19	57.72	65.93
14	MFB+GloVe+LSTM(ours)	83.99	46.31	57.75	66.12
15	MFB+GloVe+LSTM +VG(ours)	84.03	46.57	57.78	66.17

注*: 表 4.3 中 1-10 行的数据均来自于近 1-3 年的参考文献。

在表 4.3 本文模型及其他模型在 VQA v2 数据集上的实验结果对比中,第 13 至 17 行是本文进行多次实验后上获得的最终数据结果,以下是对表格内容的详细分析与总结:

(1) 将表 4.3 中的第 11 行单一特征融合模型 MFB 与第 1 至 8 行中的各类模

型相比, MFB 模型的优势相当明显, 比 2018 年最具竞争力的 MUTAN 模型在 All accuracy 上高出了 1.17%, 在“other”类型的问题中准确率高出 3.3%, 在“number”类型的问题上提高了 7.63%的准确率, 计数方面取的这样的进步不仅是得益于 MFB 的优势, 而且也验证了实验一中图片特征预处理方法的有效性。

(2) 表中第 12 行的模型是在特征融合模型选为 MFB 的基础上, 再加上 300 维的 GloVe 词向量模型针对文本数据集进行训练。相比于单一的 MFB 模型, 增加了 GloVe 词向量模型以后, 视觉问答模型在 All accuracy 上 0.03%的提高, “number”类型的问题准确率提高了 0.1%。GloVe 模型作为基于全局词频统计的词向量工具, 在捕捉单词的语义特征上有卓越的成绩, 也可以用于视觉问答任务。同时进行消融实验的还有表中的第 13 行, 相比于 11 行的模型, “MFB+LSTM”模型在“other”类型的问题上提高了 0.26%的准确率, 其他指标的准确率也有微弱的上升。不管是 GloVe 还是 LSTM, 均适用于视觉任务中的文本处理模块。

(3) 表中的第 14 行与第 11、12、13 行进行比较, 可以很明显的看出, 模型的各项准确率均比前三个模型的高。与单一的融合模型相比, 其中“MFB+GloVe+LSTM”在“Yes/No”的问题回答上提高了 0.45%的准确率, 在“number”类型的问题上提高了 0.27%。文本处理模块有了 GloVe 和 LSTM 的联合训练有助于模型更好的理解文本语义, 产生了不错的处理效果, 进一步提升了视觉问答模型各个指标的准确率。

(4) 在考虑了特征融合模块、图像预处理模块、和文本处理模块以后, 本文在实验中增加了视觉基因组数据集 Visual Genome。从实验结果可以看出, 相比于“MFB+GloVe+LSTM”模型, 增加了 VG 数据集的模型针对上述四个指标的准确率分别提高了 0.04%、0.26%、0.03%、0.05%。准确率的提高不仅验证了扩大数据集的策略是正确的, 而且也为后续的研究提供了可行的思路。“MFB+GloVe+LSTM+VG”的模型性能也超越了表中所提及的所有视觉问答模型, 验证了本文的研究方法的科学性。

4.5 本章小结

本章提出了一种基于多模态特征融合的视觉问答模型框架。首先结合 ResNet101 网络利用 Faster R-CNN 模型对视觉问答数据集中的图片数据集进行预

处理,实现细粒度的特征提取便于后续的端到端的模型训练,节省模型训练时间;然后使用 GloVe 模型对文本数据集进行共现矩阵的统计,据此获取文本单词的向量表示及其空间中的线性结构,并结合 LSTM 网络对文本信息进行表征;其次寻求自然的多模态特征融合方式对图片特征与文本特征相结合达到预测答案的效果;最后在实验中加入补充数据集 Visual Genome 进一步提高模型准确率。本文在 VQA v2 的数据集上分别进行实验,逐步验证图像处理模块的 Faster R-CNN 模型、文本处理模块的 GloVe 模型和 LSTM 网络、特征融合模块的融合方式、外部数据集的扩充等方面在提高视觉问答模型的准确率上有何种程度的影响。模型在类型为“yes/no”问题上的答案预测准确率较高,但是在类型为“number”问题的答案预测效果表现一般,若是想进一步提高视觉问答系统网络模型的准确率,可以从完善视觉问答数据集、加强模型语义信息理解、准确提取图片特征信息等方面入手。关于加强模型语义信息和图片特征信息的实验内容,将在第五章进行详细的阐述。

第 5 章 基于多模态特征融合的多重注意力机制的视觉问答

本章为了探讨注意力机制在视觉问答系统中是否可以缩短多模态特征之间的语义鸿沟问题,提出一种基于多模态特征融合的多重注意力机制的视觉问答模型,旨在利用多重注意力机制根据文本特征信息捕获图像的局部信息特征,从而达到准确理解自然语言的目的,并且能在图像特征集中获取所需的目标特征。本章首先概述了基于多模态特征融合的多重注意力机制的视觉问答方法的总体框架,然后介绍了图像特征与问题语义特征的深度抽取模型,随后阐述了各个注意力机制的模型,并详细的介绍了多重注意力机制网络的实现细节,最后给出了相关实验结果和实验分析。

5.1 模型整体架构

本文针对视觉问答每个模块的不同需求,提出基于多模态特征融合的多重注意力机制的视觉问答模型,该模型的算法结构如图 5.1 所示。

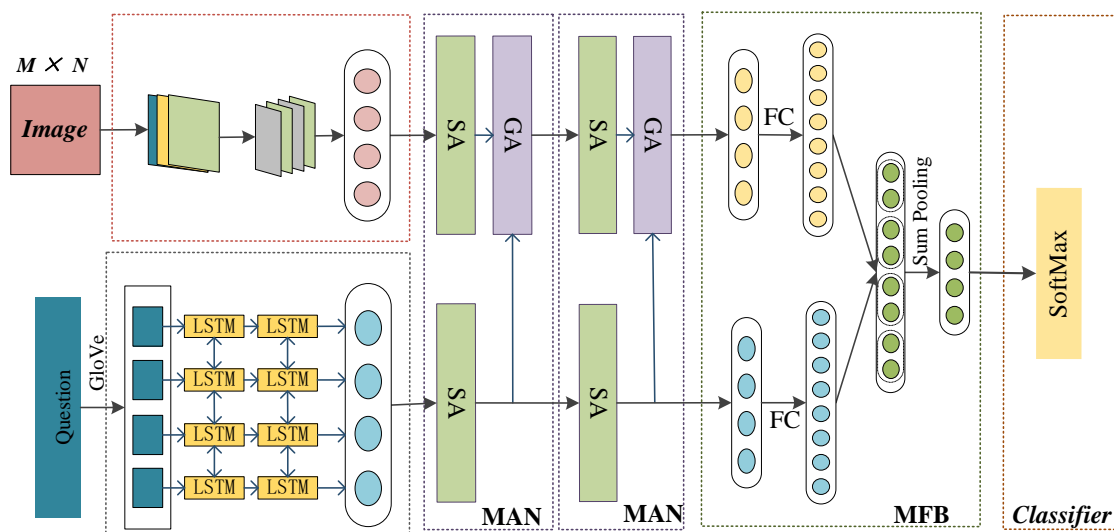


图 5.1 基于多模态特征融合的多重注意力机制的视觉问答模型

1. 图片预处理模块: 首先利用 ResNet101 网络对图片数据集进行预处理,提取图片的全局特征图,然后结合目标识别算法 Faster R-CNN 模型进行局部图像特征抽取,一张图片经过全局-局部的特征提取后从而形成多个自上而下的自适应特征便于进行融合操作。

2. 文本处理模块：利用预训练后的 300 维 GloVe 词向量工具对文本对进行向量化，然后使用长短时记忆网络对这些向量进行语义编码。

3. 特征融合模块：经过上述两个模块处理后的不同模态特征，图像特征和文本特征在多模态分解双线性池化的模型下进行特征融合。

4. 注意力机制层：使用自注意力机制、引导注意力机制和多头注意力机制形成多层注意力网络，逐层加入验证注意力机制在图像处理和文本处理中的有效性，针对不同的特征采用不同的注意力层。

5. 答案预测模块：使用双层神经网络以及 SoftMax 函数预测答案。

5.2 注意力机制

5.2.1 注意力机制的发展趋势

随着深度学习的继续深入和各种交叉学科的兴起，注意力机制已经不局限于最初提出的视觉图像领域了，它逐渐渗透到其他学科，如图像分类、目标识别、自然语言处理、语音交互等各种领域。例如，2014 年 Google mind 团队利用注意力机制结合 RNN 模型用于图像分类^[65]研究。同年，研究机器翻译任务的 D. Bahdanau 等人^[66]首次将注意力机制运用到自然语言处理中，实现了同时进行翻译与对齐任务的功能。随后在各种网络模型中均可以看到注意力机制的身影，注意力机制也在此期间得到了很好的发展，衍生出好多种变体可供不同的模型选择。图 5.2 所示为 Attention 研究进展的时间线，简单梳理了注意力机制的发展历程和应用领域。

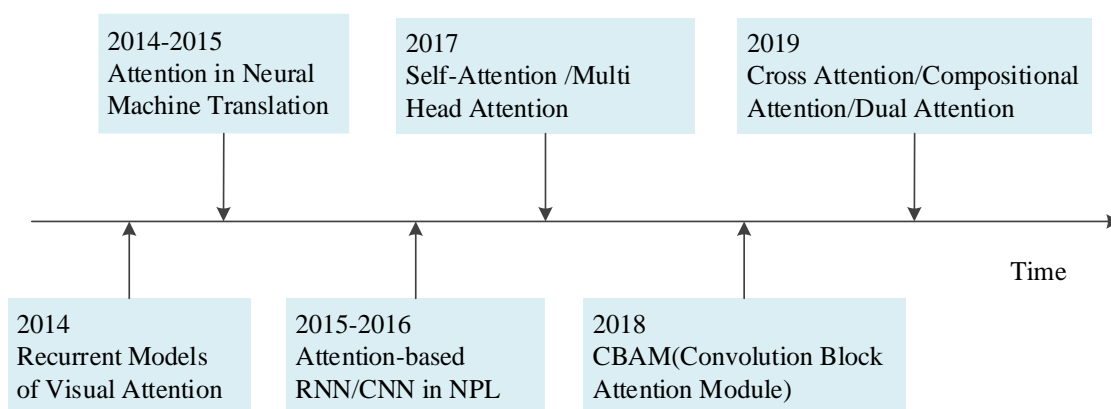


图 5.2 Attention 研究进展的时间线

2018 年 W. Sanghyun 等人^[67]提出了卷积块注意力模块(Convolutional Block

Attention Module, CBAN), 这是一种用于前馈卷积神经网络的简单而有效的注意力模块, 该模块会沿着两个独立的维度(通道和空间)依次推断注意力图, 然后将注意力图与输入特征图相乘以进行自适应特征细化。

2019 年 Fu Jun 等人^[62]提出了一种双重注意网络(DANet)可将本地功能与其全局依赖项进行自适应集成, 特别地他们在扩张的顶部附加了两种类型的注意力模块 FCN, 分别对空间和通道维度中的语义相互依赖性进行建模。同年, Yi Tay 等人^[68]提出了一种新的注意力, 综合利用了两种相似度 Pairwise Affinity 和 Distance Dissimilarity, 采用了两种计算方法和两种不同的激活函数进行复合构造相似度, 并在多个任务/数据集上实现了最先进的性能。

5.2.2 Attention 机制的原理与计算流程

1. Attention 机制的原理

关于 Attention 机制的原理可以将其抽象成键值对的映射关系, 例如把一个序列最为一个查询(Query), 则序列中的元素则组成一系列的键-值对(Key-Value), 如图 5.3 所示。

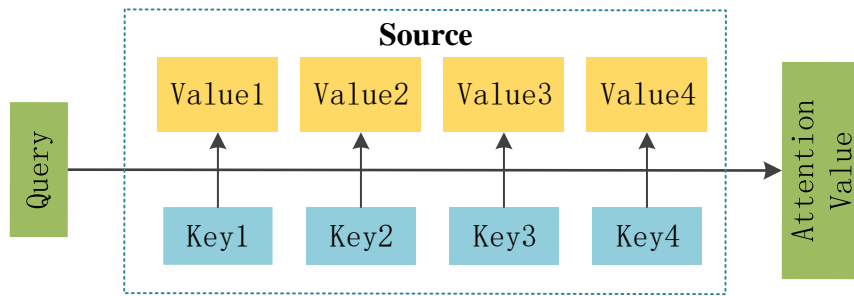


图 5.3 Attention 机制本质思想

有图 5.3 的映射关系图可以看出, Attention 机制的本质思想是通过计算 Key 和 Query 之间的相似性和关联性来获得 Value 的权重系数, 进而对 Value 的权重系数进行加权求和操作。即一系列 Key-Value 组成 Source 中的元素,

将其过程用数学公式抽象表示则为:

$$Attention(Query, Source) = \sum_{i=1}^{L_x} Similarity(Query, Key_i) * Value_i \quad (5.1)$$

其中, $L_x = \|Source\|$ 代表 Source 的长度。

就此 Attention 机制便可以对序列的重要信息进行筛选聚焦, 从而忽略掉不重

要的干扰信息，Value 的信息越重要，分配的权重值越大。

2. Attention 机制的计算流程

Attention 机制的计算流程主要分为三步：

Step1: 利用拼接、点积、Cosine 相似性和感知机 MLP 等计算相似度的函数，计算 Query 与每一个 Key 的相似度然后获得权重；

Step2: 可利用 SoftMax 函数对权重进行归一化处理；

Step3: 将处理后的权重信息与对应的键值 Value 加权求和即可得到 Attention。

在目前的 NLP 研究中，Value 和 Key 通常指同一个，即 Value=Key。如图 5.4 所示为 Attention 机制的计算流程图。

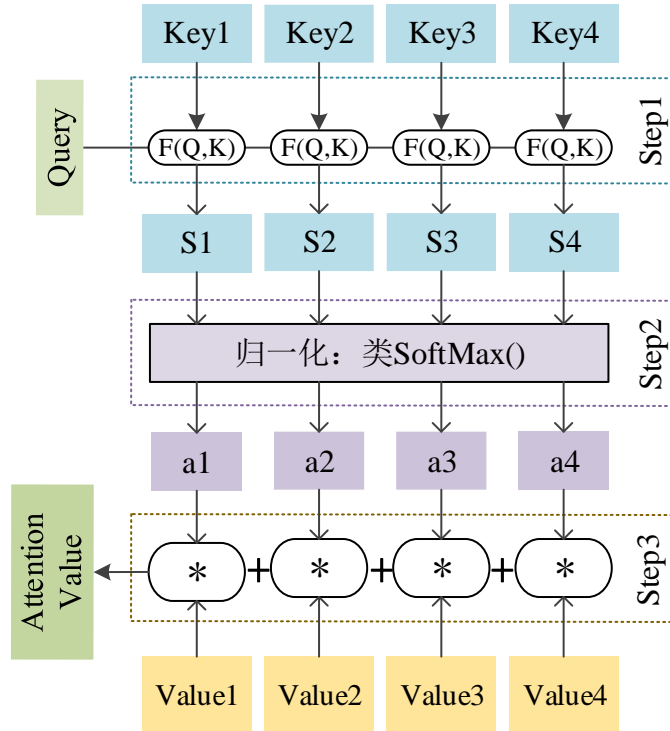


图 5.4 Attention 机制的计算流程图

针对 Step1，根据 Query 和第 i 个 Key_i ，利用不同的计算机制如向量点积和向量 Cosine 相似性，或者是引入神经网络进行计算，具体的求解方式的公式见(5.2)、(5.2)、(5.2)。

$$\text{点积:} \quad \text{Similarity}(\text{Query}, \text{Key}_i) = \text{Query} \cdot \text{Key}_i \quad (5.2)$$

$$\text{Cosine 相似性:} \quad \text{Similarity}(\text{Query}, \text{Key}_i) = \frac{\text{Query} \cdot \text{Key}_i}{\|\text{Query}\| \|\text{Key}_i\|} \quad (5.3)$$

$$\text{MLP 网络:} \quad \text{Similarity}(\text{Query}, \text{Key}_i) = \text{MLP}(\text{Query} \cdot \text{Key}_i) \quad (5.4)$$

针对 Step2, 由于 Step1 中选用的计算方法不同, 因此需要将 Step1 中的得分利用 SoftMax 函数进行归一化, 把这些得分数值换算为元素权重之和为 1 的概率分布, 在此过程中由于 SoftMax 函数的参与可以将重要权重信息突显出来, 计算公式为:

$$a_i = \text{Soft max}(Sim_i) = \frac{e^{Sim_i}}{\sum_{j=1}^{L_x} e^{Sim_j}} \quad (5.5)$$

针对 Step3, 在 Step2 的基础上求出 a_i 对应的 $Value_i$ 的权重系数, 进一步加权求和, 求出对应 Query 的 Attention 数值:

$$\text{Attention}(\text{Query}, \text{Source}) = \sum_{i=1}^{L_x} a_i * Value_i \quad (5.6)$$

综上便是 Attention 机制的计算过程解析, 目前大部分的 Attention 机制均满足以上三个步骤的计算思路与方法。

5.3 相关注意力机制介绍

注意力模型最近几年在深度学习各个领域被广泛使用, 无论是图像处理、语音识别还是自然语言处理的各种不同类型的任务中, 都很容易遇到注意力模型的身影。所以, 了解注意力机制的工作原理对于关注深度学习技术发展的技术人员来说有很大的必要。除了理解图像的视觉内容外, VQA 还要求充分理解自然语言问题的语义。因此, 有必要学习他同时对问题进行文本关注, 同时对图像进行视觉关注。

5.3.1 Self-Attention

于 2017 年, Google 的机器翻译团队发表的论文《Attention is all you need》里面提到自注意力(Self-Attention)机制^[69]可以用来进行学习文本表示, 从此 Self-Attention 机制受到了各界的关注, 在自然语言处理领域的探索也多了起来, 逐渐向各个领域延伸。Self-Attention 也被称作内部 Attention(Intra Attention), 它在自动文本摘要、文本继承和阅读理解等任务中均有出色的表现, 抽取句子的相关信息时, 它一般不会受其他额外信息的干扰, 只关注自注意力本身。

计算过程与普通的 Attention 机制相同, 但计算的对象从 Query 和 Value 变成了 Target 和 Source。Self-Attention 的优势是在计算过程中可以通过一个计算步骤

将句子中任意两个单词的直接联系起来, 因此可以缩短远距离特征之间的距离, 便于特征之间的高效利用, 同时 Self-Attention 可以增加计算的并行性, 这都是 Self-Attention 备受青睐的原因。

5.3.2 Multi-Head Attention

放缩点积注意力(Scaled Dot-Product Attention)的输入由维度 d_{key} 的查询键以及维度 d_{value} 的值组成, 结构如图 5.5(a)所示。在计算中 d_{key} 和 d_{value} 通常被设置为相同的数字 d 。用所有键来计算查询的点积分别除以 \sqrt{d} , 并应用 *SoftMax* 函数来获得值上的注意权重。给定查询 $q \in \mathbb{R}^{1 \times d}$, n 个键值对 (封装在一个关键矩阵 $K \in \mathbb{R}^{n \times d}$ 和一个值矩阵 $V \in \mathbb{R}^{n \times d}$), 通过对从 Q 和 K 中学习到的注意力的所有值 V 的加权求和, 得到所有注意力的特征 $f \in \mathbb{R}^{1 \times d}$:

$$f = A(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (5.7)$$

Query, Key, Value 首先进过一个线性变换, 然后输入到放缩点积 Attention, 注意这里要做 h 次, 其实也就是所谓的多头, 每一次算一个头。

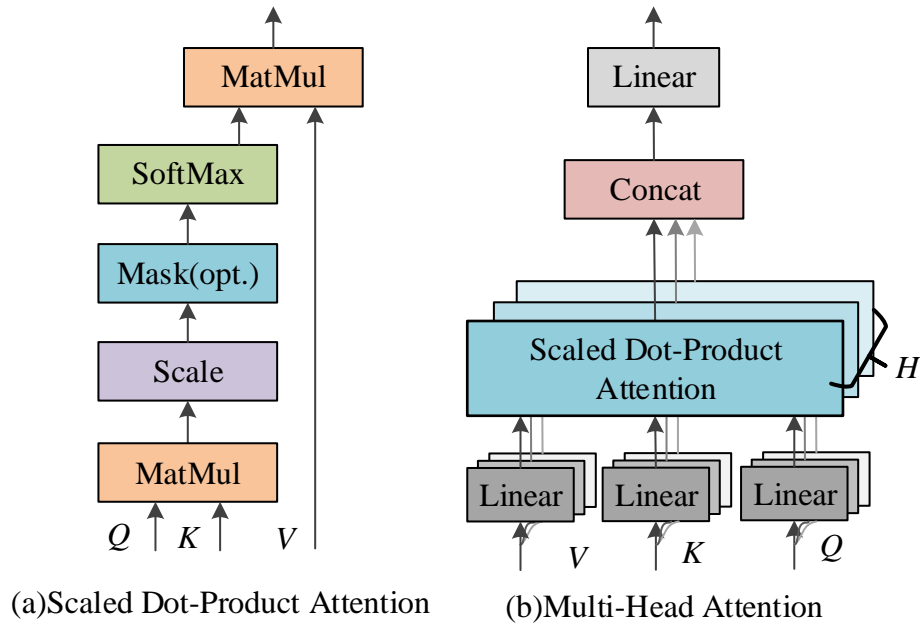


图 5.5 (a)放缩点积注意力 (b)多头注意力机制

而且每次 Q , K , V 进行线性变换的参数 W 是不一样的。然后将 h 次的放缩点积 Attention 结果进行拼接, 再进行一次线性变换得到的值作为多头 Attention 的结

果，多头注意力机制如图 5.5(b)所示。

为了进一步提高注意力特征的表示能力，引入了多头注意，该注意由 h 平行的“heads”组成。每个头对应于独立缩放的点积注意力功能。则输出特征 f 为：

$$f = MA(q, K, V) = [head_1, head_2, \dots, head_h]W^o \quad (5.8)$$

$$head_j = A(qW_j^Q, KW_j^K, VW_j^V) \quad (5.9)$$

其中 $W_j^Q, W_j^K, W_j^V \in \mathbb{R}^{d \times d_h}$ 是第 j 个头的投影矩阵， $W^o \in \mathbb{R}^{h \times d_h \times d}$ 。 d_h 是每个头部输出特征的维数。为了防止多头注意力模型变得太大，通常设 $d_h = d/h$ 。

在多头注意的顶部建立了两个注意单元来处理 VQA 的多模态输入特性，即自注意(SA)单元和引导注意(GA)单元。SA 单元由多头注意层和点向前馈层组成。以一组输入特征 $X = [x_1; \dots; x_m] \in \mathbb{R}^{m \times d_x}$ ，多头注意学习对配对样本 $\langle x_i, x_j \rangle$ 在 X 内的关系，并通过对 X 中所有实例的加权求和来输出特征 $Z \in \mathbb{R}^{m \times d}$ 。前馈层采用多头注意层的输出特征，并通过具有 *ReLU* 激活和 *Dropout* (*FC(4d)-ReLU-Dropout(0.1)-FC(d)*) 的两个完全连接的层进一步变换它们。此外，残差连接和层归一化被应用于两层的输出，以促进优化。GA 单元包含两组输入特征 $X \in \mathbb{R}^{m \times d_x}$ 和 $Y = [y_1; \dots; y_n] \in \mathbb{R}^{n \times d_y}$ ，其中 X 引导 Y 的注意力学习。 X 和 Y 的形状是灵活的，因此它们可以用来表示不同模态的特征（问题特征和图像特征）。GA 单元分别模拟了 X 和 Y 之间 $\langle x_i, y_j \rangle$ 的每个配对样本之间的成对关系。

5.4 基于多重注意力机制的视觉问答模型

5.4.1 多重注意力机制网络

借鉴 Yu Zhou 等人提出的模块化共同注意力网络(Modular Co-Attention Network, MCAN)模型^[70]的做法，本文提出一种基于多重注意力机制网络(Multiple Attention Mechanism Network, MAN)的视觉问答模型。该模型将多头注意力机制中的图像特征输入 X 令为 V 和问题特征输入 Y 令为 Q ，通过将输入特征传递到由深度级联的 L 个 MA 层组成的多层注意模型（由 $MA^{(1)}, MA^{(2)}, \dots, MA^{(L)}$ 表示）来执行多层注意学习。将 $MA^{(l)}$ 的输入特征分别表示为 $V^{(l-1)}$ 和 $Q^{(l-1)}$ ，它们的输出特征分

别由 $V^{(l)}$ 和 $Q^{(l)}$ 表示进一步以递归方式馈入 $MA^{(l+1)}$ 作为其输入。

$$\{V^{(l)}, Q^{(l)}\} = MA^{(l)}(\{V^{(l-1)}, Q^{(l-1)}\}) \quad (5.10)$$

对于 $MA^{(1)}$ ，分别设置其输入特征 $V^{(0)} = V$ 和 $Q^{(0)} = Q$ 。

问题特征向量模块本文采用 Self-Attention 设置 SA 单元，图像特征向量模块采用 Self-Attention 和 Guided-Attention 设置 SA 单元和 SGA 单元，形成多重注意力机制网络 MAN，一个 $SA(Q)$ - $SGA(V, Q)$ 层便形成一个层数 $L=1$ 的 MAN 网络（本文中主要研究 $L=1, 2, 3$ 的实验情况），多重注意力机制网络 MAN 的网络结构（当 $L=2$ 时）图如图 5.5 所示。

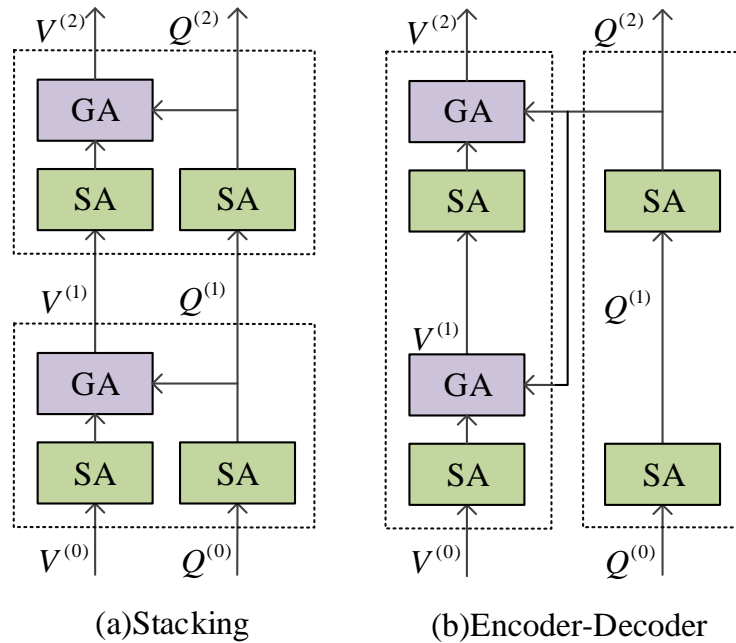


图 5.6 多重注意力机制网络

如图 5.5 左图(a)所示为两层 MA 网络深度叠加，输出 $V^{(L)}$ 和 $Q^{(L)}$ 作为最终的图像和问题特征。如图 5.5 右图(b)所示为编码-解码器，将每个 $MA^{(l)}$ 中 GA 单元的输入特征 $Q^{(L)}$ 替换为问题特征从最后一个 MA 层中提取 $Q^{(L)}$ 。编码-解码器策略可以理解编码器来学习具有 L 叠加 SA 单元的 $Q^{(L)}$ 和使用 $Q^{(L)}$ 到 L 的解码器的问题特征获得具有堆叠 SGA 单元的出席图像特征 $Q^{(L)}$ 。这两种深模型的大小相同， L 相同。作为 $L=1$ 的特例，这两种模型严格地等价于每一种情况。

5.4.2 多模态融合与输出预测

在多重注意力机制学习阶段后, 输出图像特征 $V^{(L)} = [v_1^{(L)}; \dots; v_m^{(L)}] \in \mathbb{R}^{m \times d}$, 问题特征 $Q^{(L)} = [q_1^{(L)}; \dots; q_n^{(L)}] \in \mathbb{R}^{n \times d}$ 已经包含丰富的问题文本和图像区域的注意力权重的相关信息。以 $V^{(L)}$ 为例, 注意力特征的获取如下:

$$\begin{aligned} \alpha &= \text{soft max}(MLP(V^{(L)})) \\ \tilde{v} &= \sum_{i=1}^m \alpha_i v_i^{(L)} \end{aligned} \quad (5.11)$$

其中 $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_m] \in \mathbb{R}^m$ 是学习的注意力权重。通过类比, 我们可以利用独立的注意力减少模型得到 $Q^{(L)}$ 的注意力特征 \tilde{q} 。

使用计算出的 \tilde{v} 和 \tilde{q} , 线性多模态融合函数如下:

$$m = \text{LayerNorm}(W_v^T \tilde{v} + W_q^T \tilde{q}) \quad (5.12)$$

其中 $W_v, W_q \in \mathbb{R}^{d \times d_m}$ 是两个线性投影矩阵。 d_m 是融合特征的共同维数。*LayerNorm* 在这里用来稳定训练。

融合特征 m 投影到向量 $s \in \mathbb{R}^N$ 中 N 后跟一个 *Sigmoid* 函数, 其中 N 是数字训练集中最常见的答案。借鉴文献^[71]中的方法, 使用二进制交叉熵(BCE)作为损失函数在融合特征 m 的顶部训练 N 向分类器。

5.5 实验结果及分析

在本节的实验部分, 为评估模型的性能, 本文在最大的 VQA 基准数据集 VQA-v2 上进行实验。因为不同的注意力机制和特征融合模型可能会影响视觉问答系统的最终性能, 因此进行了广泛的定量和定性消融研究, 以探讨多重注意力机制表现良好的原因。最后, 本文提出的最优模型在两个大型公开数据集中与目前最先进的模型作比较分析。

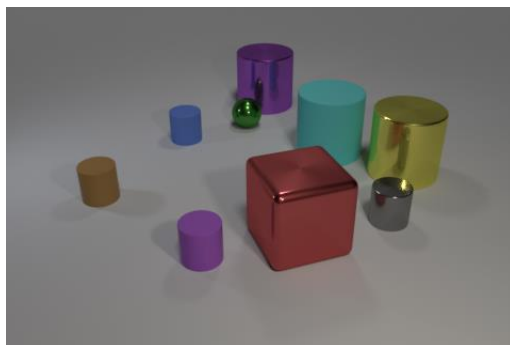
5.5.1 数据集

因为本章是在第三章的基础上进行实验研究验证, 所以同第四章一致使用 VQA v2 数据集来评估模型。VQA v2 数据集的图片来源于 COCO 数据集, 是研究

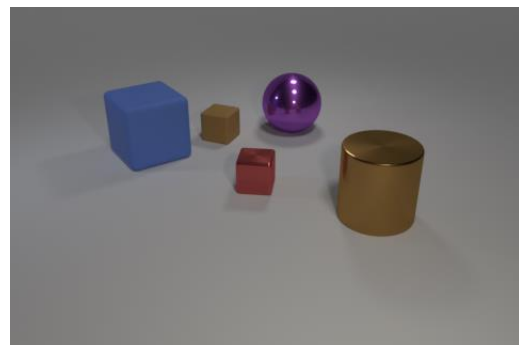
VQA v2 问题中使用的规模最大、最复杂的数据集。VQA v2 数据集包含 248395 个训练问题，121512 个验证集问题，以及 244302 个测试集问题。每张图片对应多个人工标注的 3 个问题，以及每个问题对应 10 个人工标注的答案。VQA v2 数据集类型包括“**Yes/No**”、“**Number**”和“**Other**”三种类型的问题需要回答。其中“**Yes/No**”表示答案是“yes/no”，“**Number**”表示答案是一个数字。“**Other**”涉及到的问题是疑问词类型的，比如“**What**”、“**Who**”、“**Where**”和“**How**”等等。我们使用的训练集是 VQA 训练集加上 VQA 验证集的一半。为了评估提出的模型性能，我们需要在 **test-dev** 和 **test-std** 两个数据集上做模型评估。另外，在评估模型性能时，采用和第三章内容相一致的评估标准，即精确度标准表示在人工标注的 10 个答案中，如果至少有 3 个答案是相同的，那么这个答案就被认为是正确的。

此外，本文在该章节的实验中还增加了 CLEVR 数据集的验证实验，进一步验证本文提出的模型具有良好的稳定性和鲁棒性。CLEVR 数据集中包括训练集图片 70000 张、问题 699989 个；验证集图片 15000 张、问题 149991 个；测试集图片 15000 张、问题 149988 个，每张图片平均对应 10 个问题，每个问题有且只有一个答案且为正确答案。

该数据集的图片如图 5.7 所示，其内容基本上由一些简单的颜色各异、形状不同的几何体构成，但是文本数据集中的问题却是一些较为复杂的逻辑推理问题，并且问题的 ID 是按顺序编码。与其他数据集不同，它主要是为解决视觉问答任务中缺乏推理问题的难点而构建的，是一种用于组合语言和初级视觉推理的诊断数据集，可以用于测试视觉问答系统的推理性。为了便于实验分析，本章在实验部分针对该数据集的不同种类的问题均计算出其准确率以供比较，同时将训练得到的数据进行可视化研究，对模型的稳定性作进一步分析。



(a) CLEVR_train_037336.png



(b) CLEVR_train_037347.png

图 5.7 CLEVR 数据集示例

5.5.2 实验设置

1. 硬件要求：本工作使用 windows10 操作系统的服务器，该服务器的容量参数为显存容量 24GB、内存容量 16GB、机械硬盘容量 4TB、固态硬盘容量 256GB，硬件装备满足实验要求并且为后续研究提供了方便快捷的实验环境

2. 软件要求：Python 3.6、PyTorch 0.4.1、cuda 9.0、cuDNN 7.0.4。

Pytorch 是一个可提供各种常用网络模型的开源机器学习应用平台框架，如 CNN、GRU 和 LSTM，并且可以加载一些预训练好的模型对图像进行处理，节省耗费在搭建软件环境的时间，同时也简化了模型搭建难度。

3. 实验设置

图像特征模块：为了实现图像特征的自适应表达，本文从 ResNet101 最后一个卷积层提取图像的视觉特征，对于大小为 448×448 的图片，可以得到一个大小为 $14 \times 14 \times 2048$ 的图像特征图，然后将提取图像局部特征的 Faster R-CNN 模型中的 RP 参数设置为 100，得到一个维度为 100×2048 的图像特征。

文本特征模块：问题长度 N 设置为 26，同时采用预训练好的语言模型 GloVe 对单词编码。LSTM 的隐层节点数量设置为 512，采用双层的 LSTM 网络。

融合模块：融合特征向量的维度设置为 1024。

注意力模块：按照^[69]中的建议，多头注意力中的潜在维数 d 为 512， $head$ 头数 h 设置为 8，每个 $head$ 的潜在维数为 $d_h = d/h = 64$ 。采用 SGD 训练网络，学习率为 0.001，衰减率为 0.9，权值衰减参数为 $10e-8$ ，dropout 设定为 0.5，batch size 为 64，模型在 epoch=16 时取得最好的成绩。

5.5.3 实验结果

在本节中，主要分析多种注意力机制多视觉问答任务的影响，结合第三章的特征融合模型进行注意力机制的消融实验。实验结果可以看出本文提出的基于多重注意力机制的视觉问答模型准确率优于无注意力机制的模型。本节共在两个大型数据集中进行实验验证。

1. 在 VQA 数据集上的实验结果分析

下表 5.1 为本文模型与其他先进模型在 VQA v2 数据集的实验结果对比。

表 5.1 本文模型及其他先进模型在 VQA-v2 数据集上的实验结果对比

序号	模型	Test-dev				Test-standard			
		Y/N	Num.	Other	All	Y/N	Num.	Other	All
1	NMN ^[34]	81.2	38.0	44.0	58.0	81.2	37.7	44.0	58.2
2	HieCoAtt ^[19]	79.7	38.7	51.7	61.8	-	-	-	62.1
3	MCB+Att ^[13]	82.2	37.7	54.8	64.2	-	-	-	-
4	DAN ^[62]	83.0	39.1	53.9	64.3	82.8	38.1	54.0	64.2
5	MLB+Att ^[44]	84.1	38.2	54.9	65.1	84.0	37.9	54.8	65.1
6	MFB+Att ^[15]	84.0	39.8	56.2	65.9	83.3	38.9	56.3	65.8
7	MFH+Att ^[16]	85.0	39.7	57.4	66.8	85.0	39.5	57.4	66.9
8	MFB	84.03	46.57	57.78	66.17	84.02	46.59	57.75	66.16
9	MFB+MAN(L=1)	84.42	49.17	58.14	67.01	84.41	49.15	58.16	67.05
10	MFB+MAN(L=2)	84.89	49.31	58.45	67.18	84.86	49.32	58.46	67.19
11	MFB+MAN(L=3)	84.68	49.29	58.43	67.11	-	-	-	-
12	Human	95.8	83.4	72.7	83.3	-	-	-	-

注*: 表 5.1 中 1-7 行的数据均来自于近年的参考文献。

以下是对表 5.1 的分析与总结:

(1) 表中 1 至 7 行的模型都是近几年中视觉问答模型的佼佼者, 本文提出的“MFB+MAN(L=2)”模型在各项准确率上都超过了以往的先进视觉问答模型, 取得了相当不错的进步, 并有力地证明了注意力机制在细粒度特征和高语义特征表征有正向促进作用。

(2) 将表中的第 9 行与第 8 行相比较, 此处第 8 行中所指的 MFB 模型是第四章训练的最优模型, 便以此为基线模型分别在网络中加入自注意力单元和引导注意力单元形成层数为 1 的 MAN 网络。MFB 模型加入注意力机制网络以后对提高视觉问答任务的准确率有所帮助, 证明引入注意力机制可以有效的提升模型的泛化能力。

(3) 当 MAN 的层数设为 1 时, 模型“MFB+MAN”相比于模型“MFB”, 模型“MFB+MAN”将“yes/no”、“number”、“other”、“all”四个类型的问题准确率分别提高了 0.39%、2.6%、0.36%、1.01%, 增长率明显高于单一注意力机制的模型, 由此可见多种注意力机制的联合运用更有效的提升了模型的性能。

(4) 当 MAN 的层数设为 2 时, 模型“MFB+MAN”达到了本文中最高准确率; 充分证实了本文提出的基于多重注意力机制和多模态特征融合算法的视觉问答系统的优良性, 和在研究视觉问答任务时思路的科学性、方法的合理性。但是

当模型 MAN 的层数设为 3 时,模型准确率不升反降,意味着单纯的叠加 MAN 网络并不能很好的增加系统性能及准确率。根据文献[72]的研究结论,他们在文中对多个 Head 进行了分析,发现多个 Head 的作用有大多数是冗余的,很多可以被砍掉,并且还多个 Head 进行分类。据此,由于本文中采用了对头注意力机制,单纯的加层可能加剧了 Head 的冗余作用,因此导致模型准确率下降。

(5)表中第 12 行是人类针对视觉问答任务所能回答正确的准确率。不论是从整体正确率还是单一问题的回答,人类的准确率远高于任何视觉问答模型,这足以看出人工智能在视觉理解与自然理解之间难以跨越的鸿沟,倘若机器能更好的理解图片,更智能的理解语义,那视觉问答任务媲美于人类水平便指日可待了。

2. 在 CLEVR 数据集上的实验结果分析

为了进一步验证本文提出的 MAN 模型在视觉问答任务中有良好的稳定性和鲁棒性,本章节中特意增设了在 CLEVR 数据集上的实验对比结果,通过表 5.2 的实验结果分析比较,可以看出 MAN 模型在视觉问答任务中表现稳定且成绩突出。

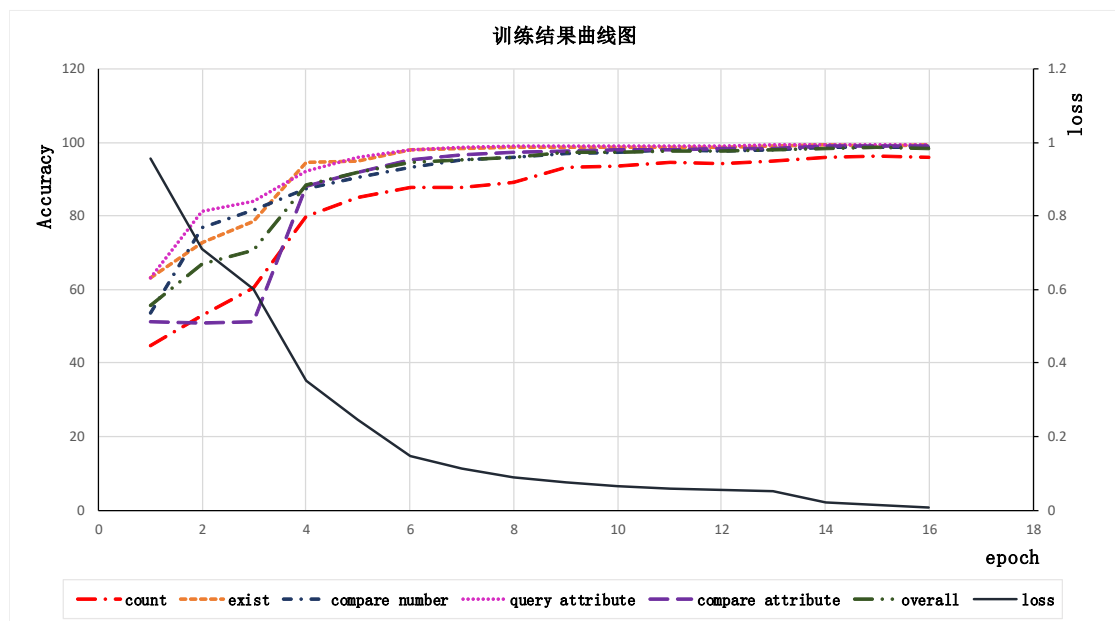


图 5.8 CLEVR 数据集的训练结果曲线图

如图 5.8 CLEVR 数据集的训练结果曲线图所示,本文提出的 MAN 模型 epoch 设置为 16 模型达到最佳效果,当 epoch=16 时损失函数 loss 趋于 0,模型的 batch size 设置为 64,模型一共迭代了 13280 次达到稳定状态,且各个评价指标也趋于平缓状态,证明模型训练很成功。“exist”和“query attribute”两种类型的问题下准确率表现很好,证明模型具有良好的稳定性,“count”类型的问题下准确率曲线

低于其他评价指标，证明模型在准确率上还有一定的提升空间。

下表 5.2 为 MAN 模型与其他先进模型在 CLEVR 数据集的实验结果对比。

表 5.2 本文模型及其他先进模型在 CLEVR 数据集上的实验结果对比

序号	模型	Accuracy%					
		Count	Exist	Compare Number	Compare Attribute	Query Attribute	Overall
1	CNN+LSTM+SAN ^[73]	59.7	77.9	75.1	70.8	80.9	92.6
2	Dependency Tree ^[74]	81.4	94.2	81.6	97.1	90.5	89.3
3	CNN+GRU+FiLM ^[75]	94.5	99.2	93.8	99.0	99.2	97.6
4	MFB+MAN(L=1)	94.60	98.96	97.56	99.26	98.40	97.83
5	MFB+MAN(L=2)	95.82	99.27	98.75	99.0	99.50	98.45
6	Human ^[73]	86.7	96.6	86.4	96.0	95.0	92.6

注*: 表 5.2 中 1-3 行的数据均来自于 1-2 年的参考文献。

以下是对表 5.2 的分析与总结：

CLEVR 数据集是一个组合语言和初级视觉推理的诊断数据集，其问题内容丰富并且可以用于测试视觉问答系统的推理性研究，也可以用于评估模型的性能。本实验在上一节的实验中进行递进研究，所以层数设置为 L=1,2。

(1) 比较表中 4、5 行的数据结果，可以明显的看出本文提出的 MAN 模型在 L=2 时表现优于 L=1 的情况设置，Overall 的准确率提高了 0.62%，Count 提高了 1.22%，Exist 提高了 0.31%，Compare Attribute 提高了 1.19%，Query Attribute 提高了 1.1%，只有 Compare Number 类型的问题下降了 0.26%。证明了适当增加多重注意力机制的层数对模型具有良好的正向提升效果，对模型的准确率和稳定性均有助益。

(2) 与近两年的参考文献[73],[74],[75]所提出的网络模型相比，本文提出的基于多重注意力机制和特征融合算法的视觉问答模型在各个评价参数指标中具有显著的优势，和表中第 3 行的模型“CNN+GRU+FiLM”相比，Overall 的准确率提高了 0.85%，Count 提高了 1.32%，Exist 提高了 0.07%，Compare Number 提高了 4.95%，Query Attribute 提高了 0.3%，在 Compare Number 类型的问题回答上有很大的提升，证明多重注意力机制的加入使得模型在理解图像信息中有良好的表现力。

(3) 与表中第 6 行的 Human 的回答准确率相比，本文提出的模型具有良好的判断能力和稳定性，每种类型的问题均优于 Human 的准确率，证明该模型的合理性和高效性。

5.6 本章小结

为了加强模型语义信息和更准确的抓取图片特征信息,本章提出了一种基于多模态特征融合的多重注意力机制的视觉问答模型框架。本章内容主要是在第四章基于多模态特征融合的视觉问答系统的基础上加入自注意力机制、引导注意力机制、多头注意力机制等多重注意力机制,旨在更好的捕捉图片及文本之间的相关语义信息,缩短多模态特征融合的鸿沟。其中图片处理部分和文本处理模块延续第四章的处理方式。从实验结果分析,模型在类型为“number”和“other”的问题上的答案预测准确率提升最多,针对多数VQA方法只采用低层或者中层的图像特征结合问题特征进行分类学习的问题,引入了自注意机制与问题引导的最为相关的高层视觉概念,这种方法很好的解决了图像特征与问题特征抽象程度不一样的问题,使得融合的特征更具有判别性,同时采用了多重注意机制使得图像区域特征、视觉概念特征与问题语义之间的联系更加紧密,减小了视觉与自然语言之间的融合间隙,并在两个公开的大型数据集中进行实验,结果充分地体现出多重注意力机制网络在增强语义特征方面有极大的优势。

第6章 总结与展望

视觉问答系统是图片处理与自然语言处两个方向的结合体,其中既包计算机视觉也包括的智能问答,是人工智能领域最热门的研究方向之一。为了满足计算机科学的发展,探寻出一个完整的视觉问答系统,提高其性能并将其成功运用到相应的领域是研究人员迫在眉睫的任务之一。虽然视觉问答系统发展到今天已经取得了一定的进步,但是在模型训练上如何降低训练的时间成本以及如何将两种模态下的特征进行良好的融合从而保证其高语义性都是研究人员关注的主要问题。本文提出了一种基于特征融合的视觉问答算法框架模型,该模型首先将图片预处理提取出图片中的语义信息,通过将图像特征与文本特征进行有效的融合,使得融合后的特征具有高语义的表征能力。同时本文基于第四章的融合模型在第五章的实验中加入了多层注意力机制作为改进方式,提出一种基于特征融合算法的多层注意力机制的视觉问答模型。在实验部分,本文提出的模型均在视觉问答领域最大、最权威的公开数据集 VQA v2 中进行了实验,实验结果表明本文提出的模型具有较高的稳定性且准确率高于大多数模型。

6.1 工作总结

针对视觉问答系统研究中的难点和热点,本文主要对以下内容进行了相关的研究分析和实验探究:

第一,首先交代了论文研究课题的研究背景及意义,其次将近年来视觉问答的研究方法分为四类,分别对这四类视觉问答模型的国内外研究现状进行阐述与比较,然后介绍了相关内容模块的关键模型及方法理论,最后就视觉问答系统中待解决问题提出了本文的解决思路和模型搭建方式。

第二,针对视觉问答系统在训练过程中所投入的时间成本过大,文中提出基于 Faster-RCNN 目标检测算法的图像预处理方式,通过结合 ResNet101 网络利用 Faster R-CNN 模型对视觉问答数据集中的图片数据集进行预处理,实现细粒度的特征提取,便于后续的端到端的模型训练,节省模型训练时间。

第三,针对跨模态特征融合,使用 GloVe 模型对文本数据集进行共现矩阵的

统计, 据此获取文本单词的向量表示及其空间中的线性结构, 并结合 LSTM 网络对文本信息进行表征, 然后利用多模态分解双线性池化模型把已经提取好的图像信息特征与文本特征进行融合、预测答案, 最后在实验中加入了 Visual Genome 数据集以达到扩充数据集的目的, 从而提高模型准确率。本文在 VQA v2 的数据集上分别进行实验, 逐步验证了图像处理模块的 Faster R-CNN 模型、文本处理模块的 GloVe 模型和 LSTM 网络、特征融合模块的融合方式、外部数据集的扩充等方面在视觉问答模型的准确率上均有一定的提高。

第四, 针对增强模型语义信息和图片特征信息, 本文提出了一种基于多模态特征融合的多重注意力机制的视觉问答模型框架。为了加强模型语义信息和更准确的抓取图片特征信息, 主要是在第四章基于多模态特征融合的视觉问答系统的基础上加入自注意力机制、引导注意力机制、多头注意力机制等多重注意力机制, 旨在更好的捕捉图片及文本之间的相关语义信息, 缩短多模态特征融合的鸿沟。

6.2 未来的研究方向

视觉问答系统的研究方向是人工智能世上一次新的尝试, 研究人员大胆探索未知领域, 为我们提供了新颖的研究方向, 成功的将图像特征与文本特征进行融合, 该研究课题从提出到今天的发展已经有六年了, 虽比不上智能问答系统悠久的学术实践历史, 但视觉问答作为智能问答系统的衍生方向, 在学术研究中也取得了不错的成绩, 从一开始模型不到百分之五十的准确率到现在接近百分之七十; 从一开始简单的特征朴素拼接到现在的多模态分解双线性池化; 从一开始单一的注意力机制应用到现在的多层多种注意力机制融合, 视觉问答系统的进步有目共睹。在本文中基于上述的研究思路对视觉问答的模型框架进行搭建实现了跨模态特征融合与多层注意力机制的应用, 在实验部分取得了一定的实验成果, 但是就当前的视觉问答系统和人类判断标准之间还存在相当大的差距, 要想设计出一种更加智能且具有常识理解、常识推理的视觉问答系统可能还需要一段时间的探究与摸索, 以下是本文建议在未来工作中改进视觉问答系统的几点策略:

第一, 优化完善特征融合算法模型。不论是最初的特征拼接, 还是略有突破的双线性池化模型, 还是继续优化发展的多模态分解双线性池化, 都是将两个模态下的特征用向量形式进行表征, 这样多个多维参数在权重信息计算取舍中必会存在

一定的误差,从而影响融合后的特征信息。所以,不断的优化融合算法是提高模型准确率的必经之路。

第二,提高视觉问答系统的可解释性。现有的视觉问答模型大多是基于端到端的训练模式,其中的训练过程往往缺乏可解释性,不利于视觉问答模型在训练中优化算法。

第三,平衡数据集或增加数据集。在当前的视觉问答研究中有大量的开源数据集可供研究选择,但数据来源大多是 COCO 数据集,所以视觉问答模型的基准往往集中于简单的计数或是对象检测,当出现需要回答常识性推理问题时,模型是回答不了的,所以在完善视觉问答系统时,可以构造更加完善的数据集,也可以增设其他知识库,或是形成特定主题的数据集,促进 VQA 的定性分析,不仅可以提高其性能,也能使系统更加具有针对性。

第四,拓展视觉问答系统的应用场景。在目前的研究进展中还未出现完全依附视觉问答研究的应用,在未来的研究中,将视觉问答与虚拟现实相结合,应该是一个不错的选择。同时可将视觉问答系统进行可视化设计,为拓展视觉问答系统的应用提供可行的思路。

最后,随着科技日新月异的发展,人们对信新事物的探究会越来越感兴趣,本文始终坚信在人工智能领域,视觉问答系统的进步会引领新浪潮,再有 5G 通信技术的加持,自然语言处理、计算机视觉、虚拟现实这三大技术领域的融合与应用指日可待。

参考文献

- [1] 毛先领, 李晓明. 问答系统研究综述[J]. 计算机科学与探索, 2012, 6(3): 193-207.
- [2] Turing A. M. Computing Machinery and Intelligence[J]. Mind, 1950, 59(236): 433-460.
- [3] Malinowski M., Fritz M. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input[J]. Neural Information Processing Systems, 2014: 1682-1690.
- [4] Hermann K. M., Kocisky T., Grefenstette E., et al. Teaching Machines to Read and Comprehend[C]//Proceedings of the Advances in Neural Information Processing Systems, 2015: 1693-1701.
- [5] Agrawal A., Batra D., Parikh D. Analyzing the Behavior of Visual Question Answering Models[J]. Empirical Methods in Natural Language Processing, 2016: 1955-1960.
- [6] Ren M., Kiros R., Zemel R. S., et al. Exploring Models and Data for Image Question Answering[J]. Neural Information Processing Systems, 2015: 2953-2961.
- [7] Lin Ma, Zhengdong Lu, Hang Li. Learning to Answer Questions From Image Using Convolutional Neural Network[J]. Computer Science, 2016: 3567-3573.
- [8] Kushal K., Christopher K. Visual question answering: Datasets, algorithms, and future challenges[J]. Computer Vision and Image Understanding, 2017: 3-20.
- [9] Qi Wu, Teney D., Peng Wang, et al. Visual Question Answering: A Survey of Methods and Datasets[J]. Computer Vision & Image Understanding, 2017: 21-40.
- [10] 俞俊, 汪亮, 余宙. 视觉问答技术研究[J]. 计算机研究与发展, 2018, 55(09): 122-134.
- [11] Manmadhan, S., Kooor, B.C. Visual question answering: a state-of-the-art review[J]. Artif Intell Rev, 2020. <https://doi.org/10.1007/s10462-020-09832-7>.
- [12] Malinowski M., Rohrbach M., Fritz M. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images[J]. Computer Science, 2015: 1-9.
- [13] Fukui A., Park D. H., Yang D., et al. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding[J]. Empirical Methods in Natural Language Processing, 2016: 457-468.
- [14] Kim J. H., Lee S. W., Kwak D. H., et al. Multimodal Residual Learning for Visual

- QA[J]. Advances in Neural Information Processing Systems, 2016.
- [15] Zhou Yu, Jun Yu, Jianping Fan, et al. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering[C]//The IEEE International Conference on Computer Vision (ICCV), 2017: 1821-1830.
- [16] Zhou Yu, Jun Yu, Chenchao Xiang, et al. Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering[J]. Neural Networks and Learning Systems, 2018: 5947-5959.
- [17] Kan Chen, Jiang Wang, Liang-Chieh Chen, et al. ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering[J]. Computer Science, 2015.
- [18] Zichao Yang, Xiaodong He, Jianfeng Gao, et al. Stacked attention networks for image question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 21-29.
- [19] Jiasen Lu, Jianwei Yang, Dhruv B., et al. Hierarchical Co-Attention for Visual Question Answering[J]. Computer Vision and Pattern Recognition, 2016: 289-297.
- [20] Yang Shi, Furlanello T., Sheng Zha, et al. Question Type Guided Attention in Visual Question Answering[C]//Computer Vision and Pattern Recognition (CVPR), 2018: 158-175.
- [21] Malinowski M., Doersch C., Santoro A., et al. Learning Visual Question Answering by Bootstrapping Hard Attention[C]//The European Conference on Computer Vision (ECCV), 2018: 3-20.
- [22] Mengfei Li, Gu Li, Yi Ji., et al. Text-Guided Dual-Branch Attention Network for Visual Question Answering//[C]Pacific Rim Conference on Multimedia, 2018: 750-760.
- [23] Liang Peng, Yang Yang, Yi Bin, et al. Word-to-region attention network for visual question answering[J]. Multimedia Tools and Applications, 2018: 3843-3858.
- [24] Auer S., Bizer C., Kobilarov G., et al. DBpedia: A Nucleus for a Web of Open Data[J]. The Semantic Web, 2007: 722-735.
- [25] Bollacker K., Evans C., Paritosh P., et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM Sigmod International Conference, 2008: 1247-1250.
- [26] Hoffart J., Suchanek F M, Berberich K., et al. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence, 2013, 194: 28-

- 61.
- [27] Oren E., Michele B., Stephen S., et al. Open Information Extraction for the Web[J]. Communications of the ACM, 2008: 68-74.
- [28] Carlson A., Betteridge J., Kisiel B., et al. Toward an Architecture for Never-Ending Language Learning[C]//Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010: 11-15.
- [29] Tandon N., Melo G. D., Weikum G. Acquiring comparative commonsense knowledge from the web[C]//Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, 2014.
- [30] Liu H., Singh P. ConceptNet—A Practical Commonsense Reasoning Tool-Kit[J]. Bt Technology Journal, 2004: 211-226.
- [31] Qi Wu, Peng Wang, Chunhua Shen, et al. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources[C]//Computer Vision and Pattern Recognition (CVPR), 2016: 4622-4630.
- [32] Wang P, Wu Q, Shen C, et al. FVQA: Fact-based Visual Question Answering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017:1-1.
- [33] Narasimhan M., Schwing A. G., Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering[C]//The European Conference on Computer Vision (ECCV), 2018: 451-468.
- [34] Andreas J., Rohrbach M., Darrell T., et al. Learning to Compose Neural Networks for Question Answering[J]. Computation and Language, arXiv preprint arXiv:1601.01705, 2016.
- [35] Kumar A., Irsoy O., Ondruska P., et al. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing[C]//Proceedings of the 33 rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume48.
- [36] Caiming Xiong, Merity S., Socher R., Dynamic Memory Networks for Visual and Textual Question Answering[C]//Proceedings of the 33 rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume48.
- [37] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition[J].arXiv preprint arXiv:1409.1556,2014.
- [38] Arora S., Bhaskara A., Ge R., et al. Provable bounds for learning some deep representations[C]//International Conference on Machine Learning, 2014: 584-592.

- [39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015:1904-1916.
- [40] Nithyanand R., Cai X., Johnson R., et al. Glove: A Bespoke Website Fingerprinting Defense[C]//Workshop on Privacy in the Electronic Society, 2014: 131-134.
- [41] Mikolov T., Chen K., Corrado G., et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.
- [42] Shih K J, Singh S, Hoiem D. Where to look: Focus regions for visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4613-4621.
- [43] Zhou B., Tian Y., Sukhbaatar S., et al. Simple Baseline for Visual Question Answering[J]. arXiv: Computer Vision and Pattern Recognition, 2015.
- [44] Kim J. H., On K. W., Lim W., et al. Hadamard Product for Low-rank Bilinear Pooling[J]. preprint arXiv:1610.04325,2016.
- [45] Antol S., Agrawal A., Lu J., et al. VQA: Visual Question Answering[C]//International Conference on Computer Vision (ICCV), 2015: 2425-2433.
- [46] Gao H., Mao J., Zhou J., et al. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering[J]. Computer Science, 2015: 2296-2304.
- [47] Zhu Y., Groth O., Bernstein M. S., et al. Visual7W: Grounded Question Answering in Images[J]. Computer Vision and Pattern Recognition, 2016: 4995-5004.
- [48] Krishna R., Zhu Y., Groth O., et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.
- [49] Johnson J., Hariharan B., Laurens V. D. M., et al. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning[J]. Computer Vision and Pattern Recognition, 2017: 1988-1997.
- [50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep residual learning for image recognition[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.
- [51] Veit A., Wilber M. J., Belongie S. Residual networks behave like ensembles of relatively shallow networks[C]//Advances in Neural Information Processing Systems. 2016: 550-558.
- [52] R. Girshick, J. Donahue, T. Darrell, et al. Rich feature hierarchies for accurate object

- detection and semantic segmentation[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014: 580-587.
- [53] R. Girshick. Fast r-cnn[C]//The IEEE International Conference on Computer Vision (ICCV), 2015: 1440-1448.
- [54] S. Ren, K. He, R. Girshick, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Curran Associates, 2015: 91-99.
- [55] J. Redmon, S. Divvala, R. Girshick, et al. You only look once: Unified, real-time object detection[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 779-788.
- [56] W. Liu, D. Anguelov, D. Erhan, et al. Ssd: Single shot multibox detector[J]. Computer Vision , 2016: 21-37.
- [57] Hochreiter S., Schmidhuber J., Long short-term memory[J]. Neural Computation, 1997: 1735-1780.
- [58] Tsungyu L., Aruni R., Subhransu M., Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1449-1457.
- [59] Gao Y, Beijbom O, Zhang N, et al. Compact bilinear pooling[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 317-326.
- [60] Pham N, Pagh R. Fast and scalable polynomial kernels via explicit feature maps[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013: 239-247.
- [61] Hamed P., Deva R., and Charless C. Fowlkes. Bilinear classifiers for visual recognition[J]. In Advances in Neural Information Processing Systems, 2009: 1482-1490.
- [62] Jun Fu, Jing Liu, Haijie Tian, et al. Dual Attention Network for Scene Segmentation[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 3146-3154.
- [63] Hedi B., Remi C., Matthieu C., et al. MUTAN: Multimodal Tucker Fusion for Visual Question Answering[C]//The IEEE International Conference on Computer Vision (ICCV), 2017: 2612-2620.
- [64] Mingrui Lao, Yanming Guo, Hui Wang, et al. Multimodal Local Perception Bilinear Pooling for Visual Question Answering[J]. IEEE Access, 2018: 57923-57932.
- [65] Mnih V., Heess N., Graves A., et al. Recurrent Models of Visual Attention[J].

- Advances in neural information processing systems, 2014.
- [66] Bahdanau D., Cho K., Bengio Y., et al. Neural Machine Translation by Jointly Learning to Align and Translate[J]. arXiv: Computation and Language, preprint arXiv:1409.0473,2014.
- [67] Sanghyun W., Jongchan P., Joonyoung L., et al. CBAM: Convolutional Block Attention Module[C]//European Conference on Computer Vision, 2018: 3-19.
- [68] Yi Tay, Anh T. L., Aston Zhang, et al. Compositional De-Attention Networks[C]//Neural Information Processing Systems,2019: 6132-6142.
- [69] Ashish V., Noam S., Niki P., et al. Attention is all you need[J].In Advances in Neural Information Processing Systems, 2017: 6000-6010.
- [70] Yu Zhou , Yu Jun , Cui Y , et al. Deep Modular Co-Attention Networks for Visual Question Answering[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 6281-6290.
- [71] Teney D., Anderson P., He X., et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 4223-4232.
- [72] Voita E., Talbot D., Moiseev F., et al. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting[J].Computer Science, 2019: arXiv:1905.09418.
- [73] Johnson J., Hariharan B., Maaten L., et al. Inferring and Executing Programs for Visual Reasoning[C]//The IEEE International Conference on Computer Vision (ICCV), 2017: 2989-2998.
- [74] Cao Qingxing, Liang Xiaodan, Li Bailing, et al. Visual Question Reasoning on General Dependency Tree[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 7249-7257.
- [75] Perez E., Strub F., Vries H., et al. FiLM: Visual Reasoning with a General Conditioning Layer[C]//Thirty-Second AAAI Conference on Artificial Intelligence, 2018: 3942-3951.

致谢

北来南去几时休，人在光阴似箭流。漫漫人生路，我以为时间足够多的、足够长，可以让我看尽重邮每一处花草树木，读遍数图的万卷书，登遍红高粱的每坎梯，吃遍六个食堂的每道菜，殊不知，七年光阴任我蹉跎，在这七年里我都没有做完这些事。本科四年成为追忆，而研究生三年也将成为过去时，临近毕业，更多的是不舍和感恩。感谢母校相伴七年，感谢任课老师倾其相授，感谢导师悉心栽培，感谢亲朋好友鼎力支持，也感谢自己坚持不懈。

首先，十分感谢我的导师蔡林沁教授。初识蔡老师是我本科四年级时去找蔡老师解答习题。早有耳闻蔡老师为师亲善、和蔼，便大着胆子去了，果然待我道明来意后，蔡老师便接过习题，然后细心的询问情况并一一帮我解答。现在想来记忆犹新，后来有幸成为蔡老师研究生团的一员之后才晓得老师平时很忙，当时花了挺长时间为我答疑却丝毫没有催促，感激之情溢于言表。同时，在这三年的学术生涯里面，蔡老师对我的帮助也是极大的。不管是开题的方向选择，或是中期的报告、PPT检查，还是毕业的论文撰写，蔡老师一向秉承学术严格、学术创新的准则要求。不管是传统的机器学习方法还是新颖的深度学习理论，老师都了如指掌，并为我们的研究学习提供了准确的学习思路和理论基础。在日常的学习和生活中蔡老师以身作则、因材施教、待人和善，人性化的实验室管理氛围让我们每一个学生在研究学术之余还能学习一些自己热爱的技能。谢谢老师的辛苦栽培，真心感谢！

然后，我要感谢信息物理系统团队的虞继敏教授、屈洪春教授以及唐晓铭等老师，感谢您们在开题和中期答辩中所提出的宝贵建议和改进思路。也感谢在研究生课堂里，兢兢业业的老师教授们，是你们的专业与风趣为我们的学习营造了良好的学习氛围和学习兴趣。

其次，我还要特别感谢我的师兄徐宏博、周锴、颜勋、丁和恩，师姐刘晓琳、陈富丽，是你们在我学术毫无进展时给予我新思路，是你们在我学术停滞不前时给予我鞭策与鼓励，你们以自身为榜样给予我力量，也为我的研究生生涯增添了丰富多彩的乐趣。感谢我的同门胡雅心、姜娇、陈思维、隆涛和董伟，感谢你们在这三年里的帮助和相伴，无论是清晨的寒冷，还是夜晚的宁静，只要有你们在，实验室

总是充满暖意与欢乐。在我遇到困难的时候，你们总是会化身学术大佬帮我排忧解难。感谢我的师弟董建功、廖忠淑、刘程鹏、易文渊、潘锐、李皓，师妹魏敏、代宇涵、黄宇婷、陈柯佳带来的欢乐、帮助与支持。感谢实验室其他的小伙伴王显豪、杨藺、鲁冲、刘芷倩、王斐、钱洁、晏川又，因为有你们的陪伴和帮助，所以在学习、生活、求职各方面我才能事半功倍。感谢我的小伙伴刘佩，本科三年室友、研究生三年饭友，整整六年的光景，与你同行，何其荣幸。

再者，我必须倍加感谢我的至亲以及我的亲朋好友，若不是你们的理解和支持，仅凭我一人之力难以走到今天。是你们在我迷茫时为我指引方向，是你们在我懒惰时理性劝诫，是你们在我孤独时舍时相伴，是你们在我拮据时出手相助，是你们在我焦虑时从容解忧，有你们，未来无畏。

最后，向在百忙之中审查评阅论文的各位专家，以及参加答辩工作的各位教授老师，送上最衷心的感谢与最诚挚的敬意！

作者攻读硕士期间从事的科研工作及取得的研究成果

参与科研项目：

- [1] 基于体感的虚拟环境情感识别与自然交互理论与方法研究（cstc2015jcyjA40009），重庆市基础与前沿研究计划.
- [2] 基于碰撞预警算法的辅助驾驶系统研究（CYS19273），重庆市科研创新项目,2019.05-2020.5.

发表及完成论文、专利：

- [1] Linqin Cai, **SiTong Zhou**, Xun Yan, RongDi Yuan. A Stacked BiLSTM Neural Network Based on Coattention Mechanism for Question Answering[J], Computational Intelligence and Neuroscience,2019.
- [2] 蔡林沁, **周思桐**, 颜勋, 廖忠淑.一种基于堆叠 Bi-LSTM 网络和协同注意力的虚拟学习环境智能问答方法[P]. 专利号：201910036927.3
- [3] 蔡林沁, **周思桐**, 董建功, 曹世洲, 牟志豪.一种基于碰撞预警算法的辅助驾驶系统[P]. 专利号：201910561949.1

获奖：

- [1] “华为杯”第十五届中国研究生数学建模竞赛，国家级一等奖，2018.12.
- [2] 2019微软“创新杯”中国赛区重庆区域赛，省部级二等奖，2018.12.