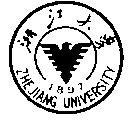
**涉密论文** □ **公开论文** □



**本 科 生 毕 业 论 文（设计）**

**题目 基于深度学习的视觉问答**

**姓名与学号 钱旭峰 3140102491**

**指导教师 邵建**

**年级与专业 14级 计算机科学与技术**

**所在学院 计算机学院**

**提交日期 2018-05-30**

**浙江大学本科生毕业论文（设计）承诺书**

1. 本人郑重地承诺所呈交的毕业论文（设计），是在指导教师的指导下严格按照学校和学院有关规定完成的。

2. 本人在毕业论文（设计）中除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **浙江大学** 或其他教育机构的学位或证书而使用过的材料。

3. 与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

4. 本人承诺在毕业论文（设计）工作过程中没有伪造数据等行为。

5. 若在本毕业论文（设计）中有侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

6. 本人完全了解 **浙江大学** 有权保留并向有关部门或机构送交本论文（设计）的复印件和磁盘，允许本论文（设计）被查阅和借阅。本人授权 **浙江大学** 可以将本论文（设计）的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编本论文（设计）。

作者签名： 导师签名：

签字日期： 年 月 日 签字日期： 年 月 日

致 谢

非常感谢肖俊老师，赵洲老师，王东辉老师，邵建老师在本次研究期间给与的极大鼓励和督促。非常感谢陈隆学长，蔺越檀学长，庞璋阳学长在本次研究期间给予的极大的帮助和支持。每当遇到困难时，老师和学长的支持是推动本次研究最大的动力。没有以上的老师和学长，本次研究绝不可能顺利进行。

摘 要（中文）

视觉问答(VQA)对人工智能的研究具有重大的意义同时极具挑战性，因为他涉及同时处理图像和相关文本，而且对模型的结构和超参的选择有较高的要求。视觉问答模型的重点在于图像特征的提取，文本特征的提取，attention权重的计算和图像特征与文本特征融合的方式。双线性特征融合被证明在融合图像特征与文本特征时有较好的的效果，但是其计算复杂度太高。目前一般的attention机制只是专注于面向图片区域的焦点。本文所介绍的CSF模型引用MFB模型来融合图像特征与文本特征，在达到较好的效果的同时大大减少了计算复杂度，同时结合spatial-wise attention和channel-wise attention来提高attention机制的性能。实验证明，将以上两者结合之后，在基于VQA2.0数据集上，本文的CSF模型相对前人提出的baseline的准确率上有了较大的提升。

**关键词：**视觉问答；特征提取；特征融合；焦点机制

Abstract （英文）

The visual question answering (VQA) has great significance for the research of artificial intelligence and it is very challenging since it involves dealing with images and related texts simultaneously, and has higher requirements on the architecture of the model and the choice of super parameters. The focus of the visual question-answering model is the extraction of image features, the extraction of text features, the calculation of attention weights, and the fusion of image features and text features. Bilinear feature fusion has been proved to have a good performance when merging image features with text features, but its computational complexity is too high. At present, most of attention mechanisms only focuses on the spatial-wise attention. The CSF model introduced in this paper cites the MFB model to fuse image features and text features. It achieves better performance while greatly reducing computational complexity. CSF model also combines spatial-wise attention and channel-wise attention to improve the performance of the attention mechanism. Experiments show that the CSF model has greatly improved the accuracy of the baseline proposed before, based on the VQA2.0 data set.

**Key words**: visual question answering、feature extraction、feature fusion、attention

# 目 录

**第一部分 毕业论文（设计）**

[致 谢 II](#_Toc515005180)

[摘 要（中文） III](#_Toc515005181)

[Abstract （英文） IV](#_Toc515005182)

[目 录 V](#_Toc515005183)

[1 绪论 1](#_Toc515005184)

[1.1 课题背景 1](#_Toc515005185)

[1.2 本文研究目标和内容 2](#_Toc515005186)

[1.3 本文结构安排 2](#_Toc515005187)

[2 视觉问答综述 3](#_Toc515005188)

[2.1 背景介绍 3](#_Toc515005189)

[2.2 VQA数据集 3](#_Toc515005190)

[2.3 视频问答 4](#_Toc515005191)

[2.4 焦点机制 4](#_Toc515005192)

[2.5 数据集偏差问题 5](#_Toc515005193)

[2.6 未知词汇问题 6](#_Toc515005194)

[2.7 模块化方法 6](#_Toc515005195)

[2.8 组成模型 7](#_Toc515005196)

[2.9 本章小结 7](#_Toc515005197)

[3 研究方案 9](#_Toc515005198)

[3.1 数据集 9](#_Toc515005199)

[3.1.1 VQA2.0 9](#_Toc515005200)

[3.1.2 问题预处理 10](#_Toc515005201)

[3.1.3 答案预处理 10](#_Toc515005202)

[3.2 总体模型 11](#_Toc515005203)

[3.2.1 Baseline 12](#_Toc515005204)

[3.2.1.1 图片处理 12](#_Toc515005205)

[3.2.1.2 问题处理 13](#_Toc515005206)

[3.2.1.3 图像特征与文本特征的结合 13](#_Toc515005207)

[3.2.2 Final 14](#_Toc515005208)

[3.3 CSF子模块 15](#_Toc515005209)

[3.3.1 CSF流程 15](#_Toc515005210)

[3.3.2 CSF\_A 16](#_Toc515005211)

[3.3.3 CSF\_B 16](#_Toc515005212)

[3.4 本章小结 17](#_Toc515005213)

[4. 实验结果 18](#_Toc515005214)

[4.1 对照实验结果 18](#_Toc515005215)

[4.2 实验结果分析 19](#_Toc515005216)

[5 本文总结 21](#_Toc515005217)

[参考文献 22](#_Toc515005218)

[作者简历 25](#_Toc515005219)

《浙江大学本科生毕业论文（设计）任务书》

《浙江大学本科生毕业论文（设计）考核表》

**第二部分 文献综述和开题报告**

文献综述和开题报告封面

指导教师对文献综述和开题报告具体内容要求

目录 I

[一、文献综述 1](#_Toc515005220)

[1．背景介绍 1](#_Toc515005221)

[1.1 基本内容 1](#_Toc515005222)

[1.2 愿景 1](#_Toc515005223)

[2．国内外研究现状 2](#_Toc515005224)

[2.1研究方向及进展 2](#_Toc515005225)

[2.1.1 任务定义 2](#_Toc515005226)

[2.1.2 用于训练和评估的VQA数据集 2](#_Toc515005227)

[2.1.3 视频问答 3](#_Toc515005228)

[2.1.4 注意机制 3](#_Toc515005229)

[2.2存在问题 4](#_Toc515005230)

[2.2.1 数据集偏差问题 4](#_Toc515005231)

[2.2.2 未知词汇问题 4](#_Toc515005232)

[3．研究展望 5](#_Toc515005233)

[3.1 模块化方法 5](#_Toc515005234)

[3.2 组成模型 5](#_Toc515005235)

[4．参考文献 6](#_Toc515005236)

[二、开题报告 10](#_Toc515005237)

[1．问题提出的背景 10](#_Toc515005238)

[1.1背景介绍 10](#_Toc515005239)

[1.2本研究的意义和目的 10](#_Toc515005242)

[2．论文的主要内容和技术路线 11](#_Toc515005243)

[2.1主要研究内容 11](#_Toc515005244)

[2.2技术路线 11](#_Toc515005245)

[2.2.1 数据集 11](#_Toc515005246)

[2.2.2 图片处理 12](#_Toc515005247)

[2.2.3 问题处理 13](#_Toc515005248)

[2.2.4 图像特征与文本特征的结合 14](#_Toc515005249)

[2.2.5 CSF层 15](#_Toc515005250)

[2.3可行性分析 15](#_Toc515005251)

[3．研究计划进度安排及预期目标 16](#_Toc515005252)

[3.1进度安排 16](#_Toc515005253)

[3.2预期目标 16](#_Toc515005254)

[4．参考文献 16](#_Toc515005255)

[三、外文翻译 19](#_Toc515005256)

[视觉问答中的多模型双线性池化分解与共同注意学习 19](#_Toc515005257)

[摘要 19](#_Toc515005258)

[1.介绍 19](#_Toc515005259)

[3.多模式因式分解双线性池 20](#_Toc515005260)

[4.VQA网络架构 23](#_Toc515005261)

[4.1 MFB基础模型 23](#_Toc515005262)

[4.2 MFB与Co-Attention 24](#_Toc515005263)

[四、外文原文 25](#_Toc515005264)

[毕业论文（设计）文献综述和开题报告考核 36](#_Toc515005265)

第一部分

**毕业论文（设计）**

# 1 绪论

## 1.1 课题背景

视觉问答（VQA）在计算机视觉和自然语言处理领域受到越来越多的研究人员的关注。 由于深度学习的成功，计算机视觉领域已经得到了巨大的进步，特别是在低级和中级任务上，如图像分割或对象识别。这些进步促进了研究人员对解决视觉与语言相结合的任务和更复杂的高层次推理的信心。VQA是这一趋势的典型例子。VQA构成了深度视觉理解和普通人工智能（AI）的基准测试。 虽然近期VQA领域取得了成功，但它仍然是一个很大的挑战和未解决的任务。

VQA涉及图像和相关文本问题，机器必须确定正确答案。该任务跨越计算机视觉和自然语言处理领域，它需要同时对问题有较深刻的理解并且解析图像的视觉元素。 VQA是评估深层视觉理解的基本任务，本身被视为计算机视觉领域的首要目标。深度视觉理解可以被定义为算法从图像中提取高级信息并基于该信息执行推理的能力。在这方面，VQA是用来评估这种能力的其他任务的替代方案。例子包括图像描述任务[1]，[2]以及近期关于视觉和对话的研究[3]。

研究VQA的另一个动机是它本身的实用性。 能够回答关于图像的问题的系统具有直接的实际应用，例如个人助理，或者在机器人中作为视觉障碍者的辅助系统。但是目前的VQA数据集并不直接处理这个设置，因为问题通常是以非面向对象的方式收集的。 现实的问题可能需要图像中不存在的信息，并涉及罕见的词汇和概念。 相比之下，目前数据集中的大多数问题都是纯粹的视觉问题（例如关于计数或颜色），并集中在常见的概念上。例如，在一个最流行的数据集VQA中，只要1,000个不同的答案可以正确回答90％以上的问题。 (仿宋小四号或12磅，1.5倍行距)

## 1.2 本文研究目标和内容

对于视觉问答（VQA）的研究具有深刻的学术意义和广阔的应用前景。随着计算机视觉（CV），自然语言处理（NLP）技术的不断发展和成熟，计算机视觉将越来越融合自然语言处理，对图片数据的语义化和结构化，可以说是自然语言处理在计算机视觉里的一个首要应用，这两三年紧密结合自然语言处理的视觉任务也越来越多。2014年和2015年大热的基于CNN+RNN的看图说话（Image Captioning）：给任意一张图，系统可以输出语句来描述这幅图里的内容。随后，2015年和2016年视觉问答（VQA）又大热。VQA是看图说话的进阶应用：以前看图说话是给张图，系统输出语句描述，而VQA更强调互动，人们可以基于给定的图片输入问题，识别系统要给出问题的答案。更深层次的讲，计算机视觉（CV），自然语言处理（NLP）两者的未来发展会借助各自的优势齐头并进，融合到General AI的框架之下，将视觉信息和文本信息相结合也是人工智能从简单单一任务迈向复杂的，需要深层次理解的任务的重要也是必要的一步。

视觉问答（VQA）具有广阔的应用前景。包括问答辅助系统，人工智能助手，搜索助手，盲人辅助装置等等。在任何需要同时理解视觉信息和文本信息的任务中，视觉问答（VQA）都能起到巨大的帮助。同时，这也是未来General AI发展的必要基础之一，也正是应为如此视觉问答（VQA）获得了巨大的市场关注。

## 1.3 本文结构安排

本文的第1节将具体介绍VQA的背景，对于VQA研究的目的和具体内容。之后的第2节将介绍目前该领域的研究现状，所用到的方法和存在的问题。第3小姐讲具体描述由本文提出的CSF模型和，包括具体的结构和参数。第4节将陈列一系列实验来对比前人提出的baseline和本文提出的模型的优劣，以及详细说明本文提出的CSF在不同结构，不同参数的情况下的对照试验。在第5节，本文将对CSF模型进行进一步的分析和解释，并指明可以改进之处。最后在第6节将对本研究做一个总结。

# 2 视觉问答综述

## 2.1 背景介绍

由于自然语言处理和计算机视觉方面的发展，计算机已经逐渐向通用人工智能的方向进步，计算机有望在不久的将来同时自动理解图像和文本，并作出进一步的推理和猜测。目前视觉问答是一件极具挑战性的任务，因为他要求深刻的理解图像和文本，并进行复杂的推理，最后推测出最佳的答案。视觉问答可以被看做是图像描述任务和图像文本检索任务的一般化。因此视觉问答是实现通用人工智能的重要的一步。

最近对VQA[4]，[5]的兴趣来源于计算机视觉领域中低级和中级任务的最新进展。这鼓励了对更高级别任务的进一步研究，以及将愿景与其他方式，特别是语言相结合。历史上，计算机视觉与语言的最早集成之一是可追溯到1972年的SHRDLU系统，它允许使用语言指示计算机在模拟的“块状世界”中移动对象。其他尝试创建的会话机器人代理也都是在视觉世界中发展起来的。然而，这些早期的作品往往局限于特定的领域和简单的语言。 深度学习现在已经应用于计算机视觉中几乎所有可以想象的问题，卷积神经网络（CNN）正在接近人类在比如图像分割[6]和物体识别等任务中的表现。 深入学习感知任务的成功推动了对高级任务的热情。VQA尤其体现了人们对实现高级图像理解的信心。

## 2.2 VQA数据集

VQA数据集中的每一个条目中都包含一个三元组，包括了一个图像，一个问题和它的正确答案。一些早期的数据集的产生是半自动的，但现代数据集大多是通过众包手动创建的[4] ，[7]。用真实的答案创建这些问题非常耗时，而今天最大数据集[8]包含了几十万个条目，代表了研究者重大的努力。 这些数据集旨在用于评估和训练VQA系统，两者都需要大量的数据。

现有的数据集主要沿着三个方向变化：1）数据集的大小，即图像和问题中表示的数量和概念的种类多少;2）所需要的推理量，例如，对于对象的检测是否是否需要对多个事实或概念进行推理; 3）输入图像中存在的信息之外的多少信息对于推断答案是必要的，例如，常识或有关主题的特定信息。大多数数据集倾向于视觉层面的问题，并且只需要很少的常识以外的外部知识。这些特征反映了即使是当前最先进的方法，仍然困扰于简单视觉问题。

在目前的数据集中，最著名的几个数据集有VQA-real，Visual genome and visual7W，Zero-shot VQA。其中VQA-real由两部分组成，一部分使用名为VQA-real的自然图像，另一部分使用名为VQA-abstract的剪贴画图像。VQA-real包含123,287个训练条目和81,434个测试条目。人类注释者被鼓励提供有趣和多样的问题，并提供简明扼要的答案（通常为2至3个单词）。数据集允许以开放式和多选式答案形式进行评估，后者为每个问题提供17个额外的（不正确的）候选答案。总体而言，数据集包含614,163个问题。

## 2.3 视频问答

除了前面提到的关于图像问答的研究之外，关于视频的VQA也有一些研究工作。Zhu等人[9]使用来自不同领域的现有视频集合，从烹饪场景到电影和网络视频，汇集了100,000多个视频和400,000个问题的数据集。Tapaswi等人 [10]提出了一个名为MovieQA的任务，其中的模型被要求根据整一部电影，标题，剧本和剧情总结来回答问题。

## 2.4 注意力模型

对联合嵌入模型最有效的改进办法之一是使用视觉焦点。人类有能力通过关注图像中的某一个区域而不是一次处理整个场景来快速理解图像。在深度神经网络中模仿人类的注意力已成功应用于机器翻译，阅读理解[11]，文本问答[12]，物体识别[13]和图像描述[14]当中，并且也用于大多数现代VQA模型（例如[15]和[16]）。焦点机制背后的主要思想是让模型专注于图像的某些区域。 该技术涉及到1）使用特定区域的图像特征2）神经网络内的相互作用。VQA模型一般使用CNN来提取描述整个图像的全局特征向量，这可能包含不相关或噪音信息。相反，我们现在为图像的不同区域提取局部特征。些特征是在最后的空间合并之前从预训练的CNN中的早的层中获得的。 网络使用图像区域特征和问题特征来计算每个区域的注意标量。函数是习得的并作为网络的附加层。注意权重可以被解释为给定区域的I相关性，并且图像最终表示为图像区域特征的加权总和。针对给定问题/图像计算的焦点权重可以以“注意映射”的形式可视化，以用于观察VQA模型的内部。 每一个对应于输入图像的特定区域，并且这些值被叠加到图像上,它们被解释为模型赋予每个图像区域的重要性。目前使用焦点机制是非常有效并且常见的做法。



## 2.5 数据集偏差问题

最近有几项研究指出了VQA数据集的一个基本问题[17],[18]。 单纯的文本问题通常会提供强有力的线索，足以使模型来得到正确答案，而不需要考虑输入图像的内容。 这些线索可能很明显。 例如，以“你有看到一个......”开头的问题几乎在10次中可以用9次“是”来提供正确答案。 这种缺陷可能源于答案比例之间的不平衡。例如，以“多少......” 开头的问题常常有“一”或“二”来作为正确答案，但很少是“17”。这个缺陷也可能以更加微妙，并以条件偏差的形式表现出来。 例如，我们可以想象，如果要正确回答“What is the color ...”，若问题包含单词“car”，则很有可能直接用“gray”就可以正确回答这个问题，若问题包含单词“flower”，则很有可能直接用“red”就可以正确回答这个问题。数据集中的图像数据也存在类似但是更加微妙的偏差。偏差是现实世界固有的，VQA模型在某种程度上习得并利用偏差是可取的。然而，现在的方法已经被证明过分依赖数据集的偏差，并且基本上被简化为训练对问题的死记硬背。这对视觉理解的目标是相违背的。VQA模型即没有显示输入图像， 只能从问题中猜测仍然可以达到56％的准确率，而再输入图片的情况下准确率为65％[18]。

## 2.6 未知词汇问题

在真实环境中使用的VQA方法，例如机器人或个人AI助手，必须能在开放，无限制的环境下正常工作。目前的VQA系统的训练模式，即用问题数据集及其答案进行培训，只能涵盖有限对象和概念。尽管VQA数据集的规模在不断扩大，但是仍然没有一套有限的范例能够覆盖现实世界中所有的对象，行为，关系等，因此应该设计理想的VQA系统来解决这个问题。 当前方法的第二个问题是模型总是被激励在数据集上表现良好，但是这会导致模型忽略罕见的单词和概念，而是集中于数据集中最常见的概念。

最近的研究在争论解决一个叫做zero-shot VQA的任务[19]，[20]，其中问题（或提出的多项选择答案）具体涉及在任何训练集问题中都未见过的单词。例如，即使没有“zebra”参与训练集，也可能出现“How many zebras are in the image?”这样的问题。该任务要求模型有强大的泛化能力。例如，一个相关的训练问题“How many giraffes are in the image?”应该被视为一个学习计数的机会，而不是只针对长颈鹿的计数。我们期望VQA最终将需要高层次推理的学习作为原理方法，而不是从有限的例子中进行蛮力学习。

## 2.7 模块化方法

目前大多数VQA模型都使用整体神经网络和端到端训练来学习数据表示，推理过程以及从训练示例中获取背景知识。另外，为了将VQA分成不同的子任务人们探索了模块化方法[21]，[22]。模块化原则允许在某种程度上将子任务互相分离，并且使用中间监督并利用多种类型的训练数据，而不仅仅是用“端到端”问题/答案对。比如，使用预训练的词向量表示是这个原则的一个非常成功的例子。 词向量被预训练来捕捉基于语言的语义相似性，并且以类似的方法，可以从辅助数据中预先训练其他数据表示以获得视觉相似性和其他类型的背景信息。用于VQA的模块化系统还允许在某种程度上从高层推理中去除视觉感知。 例如，Wang等人[23]在一系列计算机视觉算法的基础上提出了一个VQA模型，用于检测视觉元素，例如对象，人员以及它们之间的关系。 因此，VQA模型只需要对图像内容的这种明确的高级表示进行推理。

## 2.8 组成模型

图像和语言的组成性质有助于学习类似的组合模型[24]。该方法旨在解决广义化的挑战，即将学习模型应用于文本和视觉元素组合。 组成模型由Hendricks等人在图像描述的任务[25]提出。Andreas等人[26]第一个提出VQA组合体系结构，并把它称为神经模块网络。在他们的方法中，输入问题被处理来得到回答问题所需的一组操作（即相应的模块组合）。深度神经网络与被训练过的模块组装在一起，每个模块都对应于其中一种操作。 因此，模型网络专门针对每个问题量身定制，并最终应用于图像以推断答案。CLEVR合成图像数据集（主要用于训练组合文本和视觉共同推理）[27]专门被设计用于评估VQA中新组合的泛化性能。它包含各种颜色和材料的形状逼真的图像。数据集还包含注释，用来描述每个问题所需的推理类型。该数据集激发了一系列关于组合模型的研究[26],[28]。 额外的注释通过充当中间标注数据来促进组合模型的训练。这种标注数据对应于为每个问题执行的模块的组合。所有上述研究都展示了合成数据集的独特功能。然而，目前还不清楚如何最好地将它们应用于真实图像，以及如何仅使用端到端来训练它们，即只用最终答案而不用标注数据来训练他们。

## 2.9 本章小结

视觉问答是当下人工智能和深度学习方面最热门的一个方向，他极具挑战性的同时又对通用人工智能发展具有重大意义。当下对于视觉问答的研究具体的方向非常明确但又存在很多问题，对于这些问题人们提出了很多极具创新特色的解决方案，但仍存在可以提升之处。

# 3 研究方案

## 3.1 数据集

### 3.1.1 VQA2.0

本文使用VQA2.0数据集[29]来训练和测试模型，VQA2.0的图像集由来自MS-COCO数据集的约200,000幅图像组成[30]，每个图像3个问题，每个问题10个答案。数据集分为三部分：训练（80k图像和248k问题），评估（40k图像和122k问题）和测试（80k图像和244k问题）。 此外，还有一个名为test-dev的25％测试分割子集。我使用开放式（OE）模式来回答问题，开放式模式要求模型更具图片和问题直接提供答案，而不是从十几个选项中选出对的那个，但是由于开放式（OE）的答案非常难以评估，答案存在歧义性和同义性问题，所以本文首先对答案和问题进行了预处理，使得问题和答案更易于训练和评估。



**图 1 VQA数据集图片举例**

**表 1 问题和答案的预处理举例**

'answers': [['green', 1.0], ['green and yellow', 0.3], ['white', 0.3]],

'ans\_num': [['green', 8], ['green and yellow', 1], ['white', 1]],

'question':['<PAD>','<PAD>','<PAD>','<PAD>','<PAD>','<PAD>','what','color','shirts','are','the','baseball','players','wearing'],

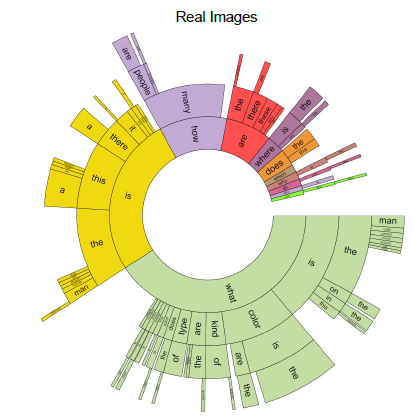
### 3.1.2 问题预处理

每张图片（如**图 1**）对应10个问题，训练集中共有443,757个问题，验证集中共有214,354个问题，本文首先将问题预处理（比如将缩写扩展，将英文数字替换成阿拉伯数字等等），再将问题分为一个个的word。由于每个问题长短不一，为了方便处理本文将问题的标准长度定为14个word，若问题长度超过14则将后面超过的部分舍去，若不足14个word则用<PAD>在前面补全，虽然当长度超过14会导致部分信息丢失，但只有797/443,757 = 0.18%的问题会被截去少量部分，所以对总体的影响不大，问题首先被处理成如**表 1**中的question的样子，之后将问题中的每个word对应成index(int)来方便处理。在计算最终准确率的时候，由模型给出的答案分部选出最佳答案，之后在总准确率上加上该答案对应的分数。

### 3.1.3 答案预处理

每个问题对应10个答案，这10个答案都是由志愿者提供的答案，10个答案中存在相同的答案，如表3.1中举的例子，该问题只有3个答案，但’green’这个答案出现了8次。之后根据每个问题中答案出现的次数根据**公式 1**来计算每个答案的分数。然后计算每个在数据集中的总分，选取总分大于一个阈值的所有答案作为候选答案，本文选取的阈值为16，候选答案总数为3097。 **图 2**是VQA2.0中的问题的前四个word的分布

**公式 1**

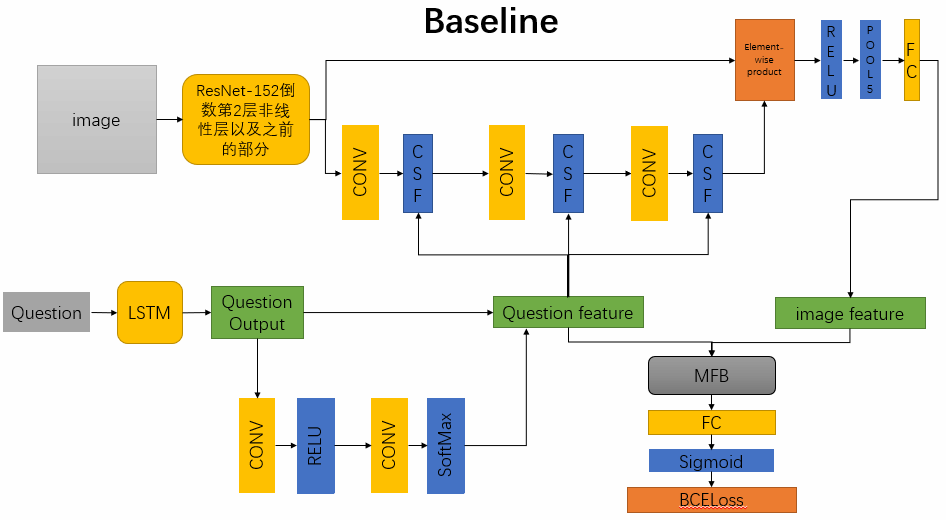


**图 2 VQA2.0数据集中问题类型的大致分布，该图提供了每个问题的前4个单词的分布，样本包括了60K个问题，图中面积越大的部分说明该类型的比重越大。**

## 3.2 总体模型

### 3.2.1 Baseline

#### 3.2.1.1 图片处理



**图 3 总体模型-Baseline，其中的CSF是本文自定义的模型**

如**图 3**所示，这是总体的Baseline模型，首先处理图片，b本文用目前最流行的方法将图片通过卷积神经网络（CNN，这里主要用到了ResNet-152[31]架构），但与其他VQA模型不同的是，本文并不是直接利用pool5层之后的结果，而是用ResNet-152最后3层conv层。

#### 3.2.1.2 问题处理

VQA2.0中含有大量的问题，问题的长短不一，我们在这里取所有问题的前14个单词，将问题中的每个单词用预训练过的glove词向量数据集将单词转化为word embedding表示，若问题不到14个单词，则空出的位置用0向量来表示，之后将这些word embedding通过LSTM[32]来获得最终的question feature ，其中LSTM将会循环14次，我取最后一次循环的hidden state作为question feature 。



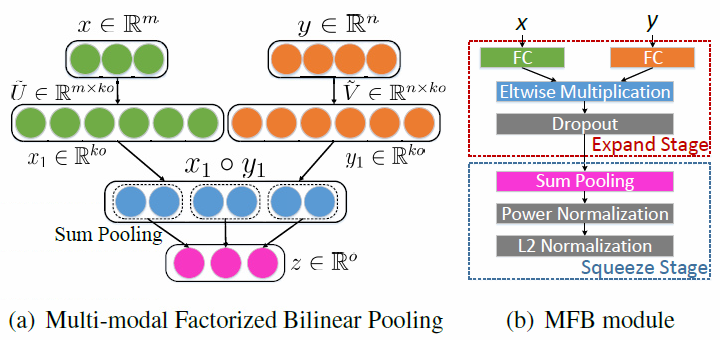
除此之外，除了给image加上attention机制之外，还可以给question加上attention机制，由于人类单单从question就可以确定question中重要的word是那几个，所以本文只用question本身来确定question的attention。获得LSTM在每个阶段的hidden state作为attention处理模块的输入，依次通过一个1维卷积层，RELU非线性层，第二个1维卷积层，最后通过一个SoftMax层得到attention权重，把attention权重和每个阶段的hidden state做加权和之后得到question feature 。



#### 3.2.1.3 图像特征与文本特征的结合

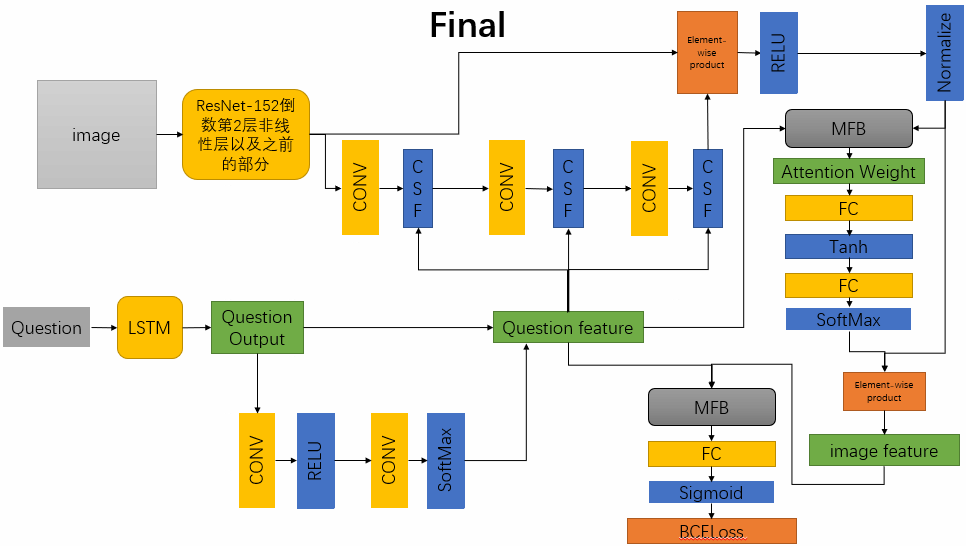
在得到最终的加权image feature 和最终的question feature 之后，我用MFB模块[33]将两者结合起来，其中MFB（如**图 4**）模块用到了bilinear pooling技术，但和一般的bilinear pooling不同，MFB用Factorized Bilinear Pooling来减少参数个数和计算复杂度，从而来大大降低内存消耗量和运行时间，在结合和之后我得到了最终的融合向量，其中k为人为定义的超参，k越大就是复杂度越高但表示能力越强，之后将c通过FC层之后得到d，与其他VQA模型不同的是，我不将d通过softmax模块来得到答案分布，而是通过sigmoid模块来得到最终的答案分布，之后用BCELoss来计算loss.





**图 4 MFB模块的流程图，MFB将两种不同表示的特征向量结合起来，MBF具体可分为expand阶段和squeeze阶段。**

### 3.2.2 Final

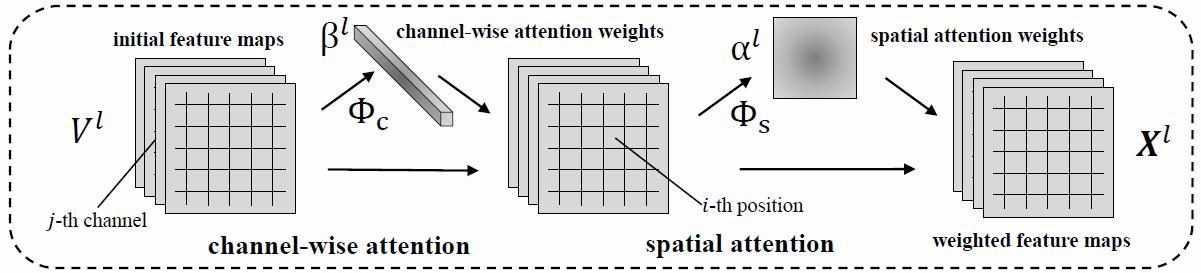


**图 5 总体模型：Final， Final相对于Baseline来说准确率有较大的提升**

Final模型（如**图 5**）相对于Baseline虽然复杂度更高，但在准确率方面有较大的提升。在结构上，两者唯一的不同在于Final的ResNet152[31]前面部分的输出在与CSF模块加权相加之后得到的image feature tensor不会直接通过线性层来转化为image feature vector，而是在normalize之后与question feature通过MFB融合来计算image feature的attention weight，之后在于normalize之后的image feature tensor加权和之后得到最终的image feature。

## 3.3 CSF子模块

### 3.3.1 CSF流程



**图 6 CSF子模块的流程，文本创建了两种CSF模块，分别为CSF\_A和CSF\_B，两者差别仅在于计算attention权重的函数不同，其流程都与上图一样，本文进行了相关对照试验并对使用不同模块的总体模型的准确率进行了比较**

在模型根据图片和问题得出答案的时候，并不是图片所有的区域都和问题相关，也不是所有的区域都能够对于得出答案提供有帮助的信息，所以如果能够对于那些重要的区域给与更高的关注，比如那些和问题的语义相关的区域，那么相信模型能够有更好的表现，所以在CSF模块（如**图 6**）中本文对spatial-wise进行了权重计算。与一般attention权重计算不同的是CSF不仅对spatial-wise进行了权重计算，还对channel-wise进行了权重计算。对channel-wise进行了权重计算相当于对某些channel进行了特别的关注，而每一个channel对应一个卷积核，而每个卷积核又对应一种样式，所以相当于对于卷积层得到的结果中的样式进行了有选择的关注。

CSF模块分为两步，首先对channel-wise进行权重计算，之后对spatial-wise进行了权重计算。文本创建了两种CSF模块，分别为CSF\_A和CSF\_B，两者差别仅在于计算attention权重的函数不同，其流程都与上图一样，本文进行了相关对照试验并对使用不同模块的总体模型的准确率进行了比较（详见实验部分）。

### 3.3.2 CSF\_A

CSF\_A模块的流程如图（6）所示，本文引用了前人的模块[34]，但是本文对其进行了改善。首先计算channel-wise权重，将由上一层得到的tensor 重塑成，其中，代表了第i个channel的feature，之后本文对每个做mean pooling，得到向量，是的均值。之后根据**公式 2**将将v和由LSTM得到的question feature h相结合，得到了channel-wise attention权重，并与上一层得到的tensor做加权和得到tensor 。

**其中**。

**公式 2**

之后计算spatial-wise权重，将由上一层得到的tensor 重塑成，其中，代表了第i块区域的特征向量。之后本文用**公式 3**将V和由LSTM得到的question feature h相结合，得到了spatial-wise attention权重，并与上一层得到的tensor做加权和得到CSF\_A模块的最终结果。

**其中**

**公式 3**

### 3.3.3 CSF\_B

CSF\_B的流程和CSF\_A一样（如**图 6**），唯一不同之处在于将question feature 和临时image feature相结合的函数不同，在CSF\_B中本文用到了MFB模块来结合question feature 和临时image feature。



首先我们得到第l-1层conv层的image feature ，首先计算channel-wise attention map ，，与在前人论文中提到的channel-wise attention map计算不同，我们先将通过形变转化为,其中代表feature map中的第i个channel，之后对U中的每一个u做meanpooling，得到，之后v和参数做外积得到image matrix ，其中k为人为定义的的超参，之后将T中的每一行与一同作为MFB[33]的输入得到channel-wise attention map ，以channel-wise attention map 作为权重与image feature 相乘得到新的image feature 。



之后计算spatial attention map ，，将通过形变转化为，其中代表feature map中的第i个区域，之后将V中的每一列与一同作为MFB的输入得到spatial attention map , 以spatial attention map 作为权重与image feature 相乘得到新的image feature 。



## 3.4 本章小结

本文用CSF+MFB的创新形式来解决基于VQA2.0的视觉问答挑战

本文所做的研究与一般视觉问答不同之处在于：

1. 本文使用open-ended模式，答案的准确率采用分数累积，而不是一般的多项选择
2. 本文采用CSF模块（包括CSF\_A和CSF\_B）不仅对spatial-wise进行了权重计算，还对channel-wise进行了权重计算。
3. 本文采用MFB模块和ResNet152 FC层之前的tensor来结合LSTM的输出来计算每个区域的权重，而不是直接把image feature和question feature结合
4. 本文采用SigMoid来计算最后的分布，而不是一般的softmax(实验部分会有对比两者的差异)

# 4. 实验结果

## 4.1 对照实验结果

本次实验用python实现，以pytorch为框架，默认batch size=10，初始学习率为0.0007，默认运行25个epoch，实验代码详见<https://github.com/AllenAnthony>。表 2为Baseline模型在不同超参下的准确率结果，**表 3**为Final模型在不同超参下的准确率结果。

**表 2 Final模型的对照试验，其中fine-tuning表示在加入CSF模块后，保持CSF模块中间的Conv层在与训练的参数初始化之后，在训练时继续微调，其中的decay表示再训练的时候当准确率不在提升时，将learning\_rate设为decay \* learning\_rate**

|  |  |
| --- | --- |
| 实验说明 | 准确率（%） |
| Baseline | 53.71 |
| Baseline + question attention | 48.20 |
| Baseline + Sigmoid | 54.15 |
| Baseline + fine-tuning | 53.46 |
| Baseline + 1 layer CSF\_B | 51.65 |
| Baseline + 1 layer CSF\_A | 53.72 |
| Baseline + 2 layer CSF\_A | 55.50 |
| Baseline + 3 layer CSF\_A | 52.29 |
| Baseline + 2 layer CSF\_A + fine-tuning | 55.64 |
| Baseline + 2 layer CSF\_A + decay=0.1 | 53.75 |
| Baseline + 2 layer CSF\_A + decay=0.5 | 53.89 |

**表 3 Final模型的对照试验，其中fine-tuning表示在加入CSF模块后，保持CSF模块中间的Conv层在与训练的参数初始化之后，在训练时继续微调，其中的decay表示再训练的时候当准确率不在提升时，将learning\_rate设为decay \* learning\_rate**

|  |  |
| --- | --- |
| 实验说明 | 准确率（%） |
| Final | 55.18 |
| Final + question attention | 48.20 |
| Final + fine-tuning | 55.50 |
| Final + 1 layer CSF\_A | 55.44 |
| Final + 2 layer CSF\_A | 55.48 |
| Final + 3 layer CSF\_A | 55.41 |
| **Final + 2 layer CSF\_A + fine-tuning + Sigmoid + decay=0.5** | **58.34** |

## 4.2 实验结果分析

question attention：基于Baseline模型，在加入question attention之后，准确率反而降低，与预期结果相反。

Sigmoid：在Baseline上，相对于softmax，使用Sigmoid使准确率提高了0.44%，提升效果较为理想。由于本次研究的准确率计算在于分数的累加，所以Sigmoid效果优于softmax与预期的结果相同。

fine-tuning：在Final上，fine-tuning使准确率提升了0.42%，在预训练的基础上，使Conv层在训练的时候继续微调参数，使模型准确率提高，与预期的相符。

CSF\_A：2层CSF\_A在Baseline上使准确率提升了1.79%，有了较大的提升。本次实验进行了CSF\_A层数为1,2,3的实验，当CSF\_A层数为2的时候效果最佳，与预期相同，有了较大的提升，说明channel-wise attention + spatial-wise attention的效果出众。

CSF\_B：CSF\_B在大大加大计算复杂度的同时并没有提升模型的表现。虽然模型的训练时间大大增加，但是模型的准确率却并没有提高，与CSF\_A相反，也与预期的相反。

dey：在准确率下降的时候降低学习率使模型的准确率有了微小的提升，实验证明dey在0.5左右的时候有较好的效果。在梯度下降的时候，为了使准确率收敛，防止准确率在达到一定高度的时候来回震荡，使其无法收敛，本文在准确率下降的时候使学习率以指数下降，是模型最终收敛到一个较为稳定且相对较高的准确率。

最后在做了大量实验时候，得到的最高准确率为58.34，相比Baseline的53.71有了4.63%的巨大提升

# 5 本文总结

在经过大量试验之后，本文得到了较好的结果58.34%，虽然相对于目前世界上对于视觉问答的研究的最高的准确率62%还有较大的差距(这里不包括ensemble模型等用来刷分的技巧)，但是任然相对于Baseline的53.71%有了较大的提高。

其中的MFB在Baseline中都有用到，不仅较好的将question feature和image feature相结合，而且相对于普通的bilinear pooling大大减少了计算的复杂度。而本文的CSF模型则同时利用channel-wise attention和spatial-wise attention，效果出众。

参考文献

1. H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, and J. Platt, “From captions to visual concepts and back,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2015, pp. 1473–1482.
2. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2014, pp. 3156–3164.
3. A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, “Visual dialog,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2017.
4. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D.Parikh, “VQA: Visual question answering,” in Proc. IEEE Int. Conf. Computer Vision, 2015, pp. 2425–2433.
5. Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, “Visual question answering: a survey of methods and data sets,” Computer Vision and Image Understanding, to be published.
6. G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” in Proc. Conf. Computer Vision and Pattern Recognition (CVPR), July 2017.
7. Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, “Visual7W: Grounded question answering in images,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 4995–5004.
8. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” arXiv Preprint, arXiv:1602.07332, 2016.
9. L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, “Uncovering temporal context for video question and answering,” arXiv Preprint, arXiv:1511.04670, 2015.
10. M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler,“Movieqa: Understanding stories in movies through question-answering,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 4631–4640.
11. M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, “Bidirectional attention flow for machine comprehension,” arXiv Preprint, arXiv:1611.01603, 2016.
12. C. Xiong, V. Zhong, and R. Socher, “Dynamic coattention networks for question answering,” arXiv Preprint, arXiv:1611.01604, 2016.
13. P. Sermanet, A. Frome, and E. Real, “Attention for fine-grained categorization,”arXiv Preprint, arXiv:1412.7054, 2014.
14. K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: neural image caption generation with visual attention,” in Proc. Int. Conf. Machine Learning, 2015, pp. 2048–2057.
15. J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image coattention for visual question answering,” in Proc. Advances Neural Information Processing Systems, 2016, pp. 289–297.
16. Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked attention networks for image question answering,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 21–29.
17. Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in visual question answering,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR), 2017.
18. A. Jabri, A. Joulin, and L. van der Maaten, “Revisiting visual question answering baselines,” in Proc. European Conf. Computer Vision (ECCV) 2016, pp. 727–739.
19. S. K. Ramakrishnan, A. Pal, G. Sharma, and A. Mittal, “An empirical evaluation of visual question answering for novel objects,” arXiv Preprint, arXiv:1704.02516, 2017.
20. D. Teney and A. van den Hengel, “Zero-shot visual question answering,” arXiv Preprint, arXiv: 1611.05546. 2016.
21. P. Wang, Q. Wu, C. Shen, and A. v d. Hengel, “The VQA-machine: Learning how to use existing vision algorithms to answer new questions,” arXiv Preprint, arXiv:1612.05386, 2016.
22. Q. Wu, C. Shen, A. v d. Hengel, P. Wang, and A. Dick, “Image captioning and visual question answering based on attributes and their related external knowledge,”arXiv Preprint, arXiv:1603.02814, 2016.
23. A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” in Proc. Conf. Empirical Methods Natural Language Processing (EMNLP), 2016, pp. 457–468.
24. Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik, “Learning to generalize to new compositions in image understanding,” arXiv Preprint, arXiv:1608.07639, 2016.
25. L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. J. Mooney, K. Saenko, and T. Darrell, “Deep compositional captioning: Describing novel object categories without paired training data,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2015, pp. 1–10.
26. R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, “Learning to reason: End-to-end module networks for visual question answering,” arXiv Preprint, arXiv:1704.05526, 2017.
27. J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick, “CLEVR: A diagnostic data set for compositional language and elementary visual reasoning,” in Proc. Conf. Computer Vision and Pattern Recognition (CVPR), 2017.
28. J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, F. Li, C. L. Zitnick, and R. B. Girshick, “Inferring and executing programs for visual reasoning,” CoRR, 2017. [Online]. Available: http://arxiv.org/abs/1705.03633
29. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In International Conference on Computer Vision (ICCV), pages 2425–2433, 2015. 1, 2, 3, 5, 6, 7, 8
30. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision (ECCV), pages 740–755, 2014. 5
31. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015. 4
32. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In International Conference on Computer Vision (ICCV), pages 2425–2433, 2015. 1, 2, 3, 5, 6, 7, 8
33. Zhou Yu†, Jun Yu†, Jianping Fan‡, Dacheng Tao. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. In International Conference on Computer Vision (ICCV).
34. Long Chen1 Hanwang Zhang2 Jun Xiao1 Liqiang Nie3 Jian Shao1 Wei Liu4 Tat-Seng Chua5. SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. arXiv:1611.05594v2 [cs.CV] 12 Apr 2017

作者简历

姓名：钱旭峰 性别：男 民族：汉 出生年月：1994-11-23 籍贯：浙江省宁波市余姚市

2014.09-2018.07 浙江大学攻读学士学位

获奖情况：无

参加项目：校园SRTP项目

发表的学术论文：无

**本科生毕业论文（设计）任务书**

**一、题目：基于深度学习的视觉问答**

**二、指导教师对毕业论文（设计）的进度安排及任务要求：**

**起讫日期 20 年 月 日 至 20 年 月 日**

**指导教师**（**签名） 职称**

**三、系或研究所审核意见:**

**负责人**（**签名）**

**年 月 日**

**毕 业 论 文（设计） 考 核**

**一、指导教师对毕业论文（设计）的评语：**

**指导教师(签名）**

**年 月 日**

**二、答辩小组对毕业论文（设计）的答辩评语及总评成绩：**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **成绩**  **比例** | **文献综述**  **占（10%）** | **开题报告**  **占（15%）** | **外文翻译**  **占（5%）** | **毕业论文（设计）质量及答辩**  **占（70%）** | **总评成绩** |
| **分值** |  |  |  |  |  |

**答辩小组负责人（签名）**

**年 月 日**



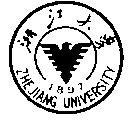
第二部分

**文献综述和开题报告**



**本 科 生 毕 业 论 文**

**文献综述和开题报告**



**学生姓名 钱旭峰**

**学生学号 3140102491**

**指导教师**  邵健

**年级与专业**  计算机科学与技术1404

**所在学院**  计算机学院

**一、题目：基于深度学习的视觉问答**

**二、指导教师对文献综述和开题报告的具体要求：**

**指导教师（签名）**

**年 月 日**

**目 录**

[**一、文献综述 1**](#_Toc510550644)

[**1.1 基本内容 1**](#_Toc510550645)

[**1.2 愿景 1**](#_Toc510550646)

[**2．国内外研究现状 2**](#_Toc510550647)

[**2.1研究方向及进展** 2](#_Toc510550648)

[2.1.1 任务定义 2](#_Toc510550649)

[2.1.2 用于训练和评估的VQA数据集 2](#_Toc510550650)

[2.1.4 注意机制 3](#_Toc510550651)

[**2.2存在问题** 4](#_Toc510550652)

[2.2.1 数据集偏差问题 4](#_Toc510550653)

[2.2.2 未知词汇问题 4](#_Toc510550654)

[**3．研究展望** 5](#_Toc510550655)

[3.1 模块化方法 5](#_Toc510550656)

[3.2 组成模型 5](#_Toc510550657)

[**4．参考文献** 6](#_Toc510550658)

[**二、开题报告 10**](#_Toc510550659)

[**1．问题提出的背景 10**](#_Toc510550660)

[1.1背景介绍 10](#_Toc510550661)

[1.2本研究的意义和目的 10](#_Toc510550664)

[**2．论文的主要内容和技术路线** 11](#_Toc510550665)

[2.1主要研究内容 11](#_Toc510550666)

[2.2技术路线 11](#_Toc510550667)

[2.2.1 数据集 11](#_Toc510550668)

[2.2.2 图片处理 12](#_Toc510550669)

[2.2.3 问题处理 13](#_Toc510550670)

[2.2.4 图像特征与文本特征的结合 14](#_Toc510550671)

[2.2.5 CSF层 15](#_Toc510550672)

[2.3可行性分析 15](#_Toc510550673)

[**3．研究计划进度安排及预期目标** 16](#_Toc510550674)

[3.1进度安排 16](#_Toc510550675)

[3.2预期目标 16](#_Toc510550676)

[**4．参考文献** 16](#_Toc510550677)

[**三、外文翻译** 19](#_Toc510550678)

[视觉问答中的多模型双线性池化分解与共同注意学习 19](#_Toc510550679)

[摘要 19](#_Toc510550680)

[1.介绍 19](#_Toc510550681)

[3.多模式因式分解双线性池 20](#_Toc510550682)

[4.VQA网络架构 23](#_Toc510550683)

[4.1 MFB基础模型 23](#_Toc510550684)

[4.2 MFB与Co-Attention 23](#_Toc510550685)

[**四、外文原文** 25](#_Toc510550686)

[毕业论文（设计）文献综述和开题报告考核 36](#_Toc510550687)

# 一、文献综述

## 1．背景介绍

### 1.1 基本内容

视觉问答（VQA）在计算机视觉和自然语言处理领域受到越来越多的研究人员的关注。 由于深度学习的成功，计算机视觉领域已经得到了巨大的进步，特别是在低级和中级任务上，如图像分割或对象识别。这些进步促进了研究人员对解决视觉与语言相结合的任务和更复杂的高层次推理的信心。VQA是这一趋势的典型例子。VQA构成了深度视觉理解和普通人工智能（AI）的基准测试。 虽然近期VQA领域取得了成功，但它仍然是一个很大的挑战和未解决的任务。

VQA涉及图像和相关文本问题，机器必须确定正确答案。该任务跨越计算机视觉和自然语言处理领域，它需要同时对问题有较深刻的理解并且解析图像的视觉元素。 VQA是评估深层视觉理解的基本任务，本身被视为计算机视觉领域的首要目标。深度视觉理解可以被定义为算法从图像中提取高级信息并基于该信息执行推理的能力。在这方面，VQA是用来评估这种能力的其他任务的替代方案。例子包括图像描述任务[1]，[2]以及近期关于视觉和对话的研究[3]。

研究VQA的另一个动机是它本身的实用性。 能够回答关于图像的问题的系统具有直接的实际应用，例如个人助理，或者在机器人中作为视觉障碍者的辅助系统。但是请注意，目前的VQA数据集并不直接处理这个设置，因为问题通常是以非面向对象的方式收集的。 现实的问题可能需要图像中不存在的信息，并涉及罕见的词汇和概念。 相比之下，目前数据集中的大多数问题都是纯粹的视觉问题（例如关于计数或颜色），并集中在常见的概念上。例如，在一个最流行的数据集VQA中，只要1,000个不同的答案可以正确回答90％以上的问题。

### 1.2 愿景

最近对VQA[4]，[5]的兴趣来源于计算机视觉领域中低级和中级任务的最新进展。这鼓励了对更高级别任务的进一步研究，以及将愿景与其他方式，特别是语言相结合。历史上，计算机视觉与语言的最早集成之一是可追溯到1972年的SHRDLU系统，它允许使用语言指示计算机在模拟的“块状世界”中移动对象。其他尝试创建的会话机器人代理也都是在视觉世界中发展起来的。然而，这些早期的作品往往局限于特定的领域和简单的语言。 深度学习现在已经应用于计算机视觉中几乎所有可以想象的问题，卷积神经网络（CNN）正在接近人类在比如图像分割[6]和物体识别等任务中的表现。 深入学习感知任务的成功推动了对高级任务的热情。VQA尤其体现了人们对实现高级图像理解的信心。

## 2．国内外研究现状

### 2.1研究方向及进展

#### 2.1.1 任务定义

VQA的一个实例由一个图像和以纯文本形式给出的相关问题组成。 机器的任务是确定正确的答案，在当前数据集中答案通常是几个字或一个短语。 通常考虑两种实用的变体，一种开放式，另一种为多项选择[4]，[7]。 在后者中，提出了一组候选答案。这使得对生成的答案的评估比在开放式设置中更容易，其中机器的输出与标准答案（即人类提供的答案）之间的比较面临同义词和释义的问题。

#### 2.1.2 用于训练和评估的VQA数据集

关于查为VQA研究专门编制的数据集。 这些数据集每一个条目中都包含一个三元组，包括了一个图像，一个问题和它的正确答案。一些早期的数据集的产生是半自动的，但现代数据集大多是通过众包手动创建的[4] ，[8]。用真实的答案创建这些问题非常耗时，而今天最大数据集[8]包含了几十万个条目，代表了研究者重大的努力。 这些数据集旨在用于评估和训练VQA系统，两者都需要大量的数据。

现有的数据集主要沿着三个方向变化：1）数据集的大小，即图像和问题中表示的数量和概念的种类多少;2）所需要的推理量，例如，对于对象的检测是否是否需要对多个事实或概念进行推理; 3）输入图像中存在的信息之外的多少信息对于推断答案是必要的，例如，常识或有关主题的特定信息。大多数数据集倾向于视觉层面的问题，并且只需要很少的常识以外的外部知识。这些特征反映了即使是当前最先进的方法，仍然困扰于简单视觉问题。

在目前的数据集中，最著名的几个数据集有VQA-real，Visual genome and visual7W，Zero-shot VQA。其中VQA-real由两部分组成，一部分使用名为VQA-real的自然图像，另一部分使用名为VQA-abstract的剪贴画图像。VQA-real包含123,287个训练条目和81,434个测试条目。人类注释者被鼓励提供有趣和多样的问题，并提供简明扼要的答案（通常为2至3个单词）。数据集允许以开放式和多选式答案形式进行评估，后者为每个问题提供17个额外的（不正确的）候选答案。总体而言，数据集包含614,163个问题。

#### 2.1.3 视频问答

除了前面提到的关于图像问答的研究之外，关于视频的VQA也有一些研究工作。Zhu等人[9]使用来自不同领域的现有视频集合，从烹饪场景到电影和网络视频，汇集了100,000多个视频和400,000个问题的数据集。Tapaswi等人 [10]提出了一个名为MovieQA的任务，其中的模型被要求根据整一部电影，标题，剧本和剧情总结来回答问题。

#### 2.1.4 注意机制

对联合嵌入模型最有效的改进办法之一是使用视觉注意。人类有能力通过关注图像中的某一个区域而不是一次处理整个场景来快速理解图像。在深度神经网络中模仿人类的注意力已成功应用于机器翻译，阅读理解[11]，文本问答[12]，物体识别[13]和图像描述[14]当中，并且也用于大多数现代VQA模型（例如[15]和[16]）。关注机制背后的主要思想是让模型专注于图像的某些区域。 该技术涉及到1）使用特定区域的图像特征2）神经网络内的相互作用。VQA模型一般使用CNN来提取描述整个图像的全局特征向量，这可能包含不相关或噪音信息。相反，我们现在为图像的不同区域提取局部特征。这些特征是在最后的空间合并之前从预训练的CNN中的较早的层中获得的。 网络使用图像区域特征和问题特征来计算每个区域的注意标量。函数是习得的并作为网络的附加层。注意权重可以被解释为给定区域的I相关性，并且图像最终表示为图像区域特征的加权总和。针对给定问题/图像计算的注意权重可以以“注意映射”的形式可视化，以用于观察VQA模型的内部。 每一个对应于输入图像的特定区域，并且这些值被叠加到图像上,它们被解释为模型赋予每个图像区域的重要性。目前使用注意机制是非常有效并且常见的做法。



### 2.2存在问题

#### 2.2.1 数据集偏差问题

最近有几项研究指出了VQA数据集的一个基本问题[17],[18]。 单纯的文本问题通常会提供强有力的线索，足以使模型来得到正确答案，而不需要考虑输入图像的内容。 这些线索可能很明显。 例如，以“你有看到一个......”开头的问题几乎在10次中可以用9次“是”来提供正确答案。 这种缺陷可能源于答案比例之间的不平衡。例如，以“多少......” 开头的问题常常有“一”或“二”来作为正确答案，但很少是“17”。这个缺陷也可能以更加微妙，并以条件偏差的形式表现出来。 例如，我们可以想象，如果要正确回答“What is the color ...”，若问题包含单词“car”，则很有可能直接用“gray”就可以正确回答这个问题，若问题包含单词“flower”，则很有可能直接用“red”就可以正确回答这个问题。数据集中的图像数据也存在类似但是更加微妙的偏差。偏差是现实世界固有的，VQA模型在某种程度上习得并利用偏差是可取的。然而，现在的方法已经被证明过分依赖数据集的偏差，并且基本上被简化为训练对问题的死记硬背。这对视觉理解的目标是相违背的。VQA模型即没有显示输入图像， 只能从问题中猜测仍然可以达到56％的准确率，而再输入图片的情况下准确率为65％[18]。

#### 2.2.2 未知词汇问题

在真实环境中使用的VQA方法，例如机器人或个人AI助手，必须能在开放，无限制的环境下正常工作。目前的VQA系统的训练模式，即用问题数据集及其答案进行培训，只能涵盖有限对象和概念。尽管VQA数据集的规模在不断扩大，但是仍然没有一套有限的范例能够覆盖现实世界中所有的对象，行为，关系等，因此应该设计理想的VQA系统来解决这个问题。 当前方法的第二个问题是模型总是被激励在数据集上表现良好，但是这会导致模型忽略罕见的单词和概念，而是集中于数据集中最常见的概念。

最近的研究在争论解决一个叫做zero-shot VQA的任务[19]，[20]，其中问题（或提出的多项选择答案）具体涉及在任何训练集问题中都未见过的单词。例如，即使没有“zebra”参与训练集，也可能出现“How many zebras are in the image?”这样的问题。该任务要求模型有强大的泛化能力。例如，一个相关的训练问题“How many giraffes are in the image?”应该被视为一个学习计数的机会，而不是只针对长颈鹿的计数。我们期望VQA最终将需要高层次推理的学习作为原理方法，而不是从有限的例子中进行蛮力学习。

## 3．研究展望

### 3.1 模块化方法

目前大多数VQA模型都使用整体神经网络和端到端训练来学习数据表示，推理过程以及从训练示例中获取背景知识。另外，为了将VQA分成不同的子任务人们探索了模块化方法[21]，[22]。模块化原则允许在某种程度上将子任务互相分离，并且使用中间监督并利用多种类型的训练数据，而不仅仅是用“端到端”问题/答案对。比如，使用预训练的词向量表示是这个原则的一个非常成功的例子。 词向量被预训练来捕捉基于语言的语义相似性，并且以类似的方法，可以从辅助数据中预先训练其他数据表示以获得视觉相似性和其他类型的背景信息。用于VQA的模块化系统还允许在某种程度上从高层推理中去除视觉感知。 例如，Wang等人[23]在一系列计算机视觉算法的基础上提出了一个VQA模型，用于检测视觉元素，例如对象，人员以及它们之间的关系。 因此，VQA模型只需要对图像内容的这种明确的高级表示进行推理。

### 3.2 组成模型

图像和语言的组成性质有助于学习类似的组合模型[24]。该方法旨在解决广义化的挑战，即将学习模型应用于文本和视觉元素组合。 组成模型由Hendricks等人在图像描述的任务[25]提出。Andreas等人[26]第一个提出VQA组合体系结构，并把它称为神经模块网络。在他们的方法中，输入问题被处理来得到回答问题所需的一组操作（即相应的模块组合）。深度神经网络与被训练过的模块组装在一起，每个模块都对应于其中一种操作。 因此，模型网络专门针对每个问题量身定制，并最终应用于图像以推断答案。CLEVR合成图像数据集（主要用于训练组合文本和视觉共同推理）[27]专门被设计用于评估VQA中新组合的泛化性能。它包含各种颜色和材料的形状逼真的图像。数据集还包含注释，用来描述每个问题所需的推理类型。该数据集激发了一系列关于组合模型的研究[26],[28]。 额外的注释通过充当中间标注数据来促进组合模型的训练。这种标注数据对应于为每个问题执行的模块的组合。所有上述研究都展示了合成数据集的独特功能。然而，目前还不清楚如何最好地将它们应用于真实图像，以及如何仅使用端到端来训练它们，即只用最终答案而不用标注数据来训练他们。

## 4．参考文献

[1] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, and J. Platt, “From captions to visual concepts and back,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2015, pp. 1473–1482.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2014, pp. 3156–3164.

[3] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and

D. Batra, “Visual dialog,” in Proc. IEEE Conf. Computer Vision and Pattern

Recognition, 2017.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D.Parikh, “VQA: Visual question answering,” in Proc. IEEE Int. Conf. Computer Vision, 2015, pp. 2425–2433.

[5] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, “Visual

question answering: a survey of methods and data sets,” Computer Vision and Image Understanding, to be published.

[6] G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” in Proc. Conf. Computer

Vision and Pattern Recognition (CVPR), July 2017.

[7] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, “Visual7W: Grounded question answering in images,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 4995–5004.

[8] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” arXiv Preprint, arXiv:1602.07332, 2016.

[9] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, “Uncovering temporal context

for video question and answering,” arXiv Preprint, arXiv:1511.04670, 2015.

[10] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler,“Movieqa: Understanding stories in movies through question-answering,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 4631–4640.

[11] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, “Bidirectional attention

flow for machine comprehension,” arXiv Preprint, arXiv:1611.01603, 2016.

[12] C. Xiong, V. Zhong, and R. Socher, “Dynamic coattention networks for question answering,” arXiv Preprint, arXiv:1611.01604, 2016.

[13] P. Sermanet, A. Frome, and E. Real, “Attention for fine-grained categorization,”arXiv Preprint, arXiv:1412.7054, 2014.

[14] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio,

“Show, attend and tell: neural image caption generation with visual attention,” in Proc. Int. Conf. Machine Learning, 2015, pp. 2048–2057.

[15] J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image coattention for visual question answering,” in Proc. Advances Neural Information

Processing Systems, 2016, pp. 289–297.

[16] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked attention networks for image question answering,” in Proc. IEEE Conf. Computer Vision and Pattern

Recognition, 2016, pp. 21–29.

[17] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V

in VQA matter: Elevating the role of image understanding in visual question

answering,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR), 2017.

[18] A. Jabri, A. Joulin, and L. van der Maaten, “Revisiting visual question

answering baselines,” in Proc. European Conf. Computer Vision (ECCV) 2016,

pp. 727–739.

[19] S. K. Ramakrishnan, A. Pal, G. Sharma, and A. Mittal, “An empirical evaluation

of visual question answering for novel objects,” arXiv Preprint, arXiv:1704.02516, 2017.

[20] D. Teney and A. van den Hengel, “Zero-shot visual question answering,” arXiv Preprint, arXiv: 1611.05546. 2016.

[21] P. Wang, Q. Wu, C. Shen, and A. v d. Hengel, “The VQA-machine: Learning how to use existing vision algorithms to answer new questions,” arXiv Preprint,

arXiv:1612.05386, 2016.

[22] Q. Wu, C. Shen, A. v d. Hengel, P. Wang, and A. Dick, “Image captioning and

visual question answering based on attributes and their related external knowledge,”arXiv Preprint, arXiv:1603.02814, 2016.

[23] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach,

“Multimodal compact bilinear pooling for visual question answering and visual

grounding,” in Proc. Conf. Empirical Methods Natural Language Processing

(EMNLP), 2016, pp. 457–468.

[24] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik, “Learning to

generalize to new compositions in image understanding,” arXiv Preprint, arXiv:1608.07639, 2016.

[25] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. J. Mooney, K. Saenko,

and T. Darrell, “Deep compositional captioning: Describing novel object categories without paired training data,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2015, pp. 1–10.

[26] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, “Learning to reason:

End-to-end module networks for visual question answering,” arXiv Preprint,

arXiv:1704.05526, 2017.

[27] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R.

B. Girshick, “CLEVR: A diagnostic data set for compositional language and elementary visual reasoning,” in Proc. Conf. Computer Vision and Pattern

Recognition (CVPR), 2017.

[28] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, F. Li, C. L. Zitnick,

and R. B. Girshick, “Inferring and executing programs for visual reasoning,”

CoRR, 2017. [Online]. Available: http://arxiv.org/abs/1705.03633

# 二、开题报告

## 1．问题提出的背景

### 1.1背景介绍

## 最近几年，随着人工智能的研究再次兴起，特别是在计算机视觉（CV），自然语言处理（NLP）和知识表示与推理（KR）结合文本与图片或者视频方向的研究逐渐火热[1],[2],[3],[4],[5],[6]，人们对于视觉问答（VQA）的研究也迅速兴起。像将图片和文本相结合的任务，比如图片描述，视觉问答等，是人工智能从简单单一任务迈向复杂的，需要深层次理解的任务的重要也是必要的一步。VQA模型主要完成的任务是通过被给与的图片和问题，需要给出相应的答案，其中视觉信息一般由深度神经网络提取，而文本信息一般用词向量表示并用循环神经网络来提取，其中的难点在与我们需要同时对图片和问题有一个深层次的理解，并且将对图片的抽象理解和对问题的抽象理解相结合来得到答案。在对VQA的研究中人们发现了很多问题，也对此提供了许多新兴的技术和解决方案，其中最重要的技术集中在对数据集的研究和对模型的改进上，特别是在模型的研究上，人们提出了各种各样的想法。

## 在对于模型的改进中，人们提供了许多极具创兴的技术，包括注意机制[7],[8]，预训练模型[9]，记忆增强神经网络[10]，运行时检索附加信息[11],[12]，组成模型[13],[14]等等，其中最重要的两个焦点在于如何使用注意机制和如何将视觉特征和文本特征相结合。对于如何使用注意机制，前人提出了非常有效，极具参考价值的内容，注意机制一般把抽象文本信息与视觉信息结合通过神经网络来提取，而对于文本信息与视觉信息相结合，前人也提出了许多很好的模型，包括MCB[15],MLB[16],MFB[17]等。我的工作也将着重这两点对VQA模型作进一步的改进。

### 1.2本研究的意义和目的

对于视觉问答（VQA）的研究具有深刻的学术意义和广阔的应用前景。随着计算机视觉（CV），自然语言处理（NLP）技术的不断发展和成熟，计算机视觉将越来越融合自然语言处理，对图片数据的语义化和结构化，可以说是自然语言处理在计算机视觉里的一个首要应用，这两三年紧密结合自然语言处理的视觉任务也越来越多。2014年和2015年大热的基于CNN+RNN的看图说话（Image Captioning）：给任意一张图，系统可以输出语句来描述这幅图里的内容。随后，2015年和2016年视觉问答（VQA）又大热。VQA是看图说话的进阶应用：以前看图说话是给张图，系统输出语句描述，而VQA更强调互动，人们可以基于给定的图片输入问题，识别系统要给出问题的答案。更深层次的讲，计算机视觉（CV），自然语言处理（NLP）两者的未来发展会借助各自的优势齐头并进，融合到General AI的框架之下，将视觉信息和文本信息相结合也是人工智能从简单单一任务迈向复杂的，需要深层次理解的任务的重要也是必要的一步。

视觉问答（VQA）具有广阔的应用前景。包括问答辅助系统，人工智能助手，搜索助手，盲人辅助装置等等。在任何需要同时理解视觉信息和文本信息的任务中，视觉问答（VQA）都能起到巨大的帮助。同时，这也是未来General AI发展的必要基础之一，也正是应为如此视觉问答（VQA）获得了巨大的市场关注。

## 2．论文的主要内容和技术路线

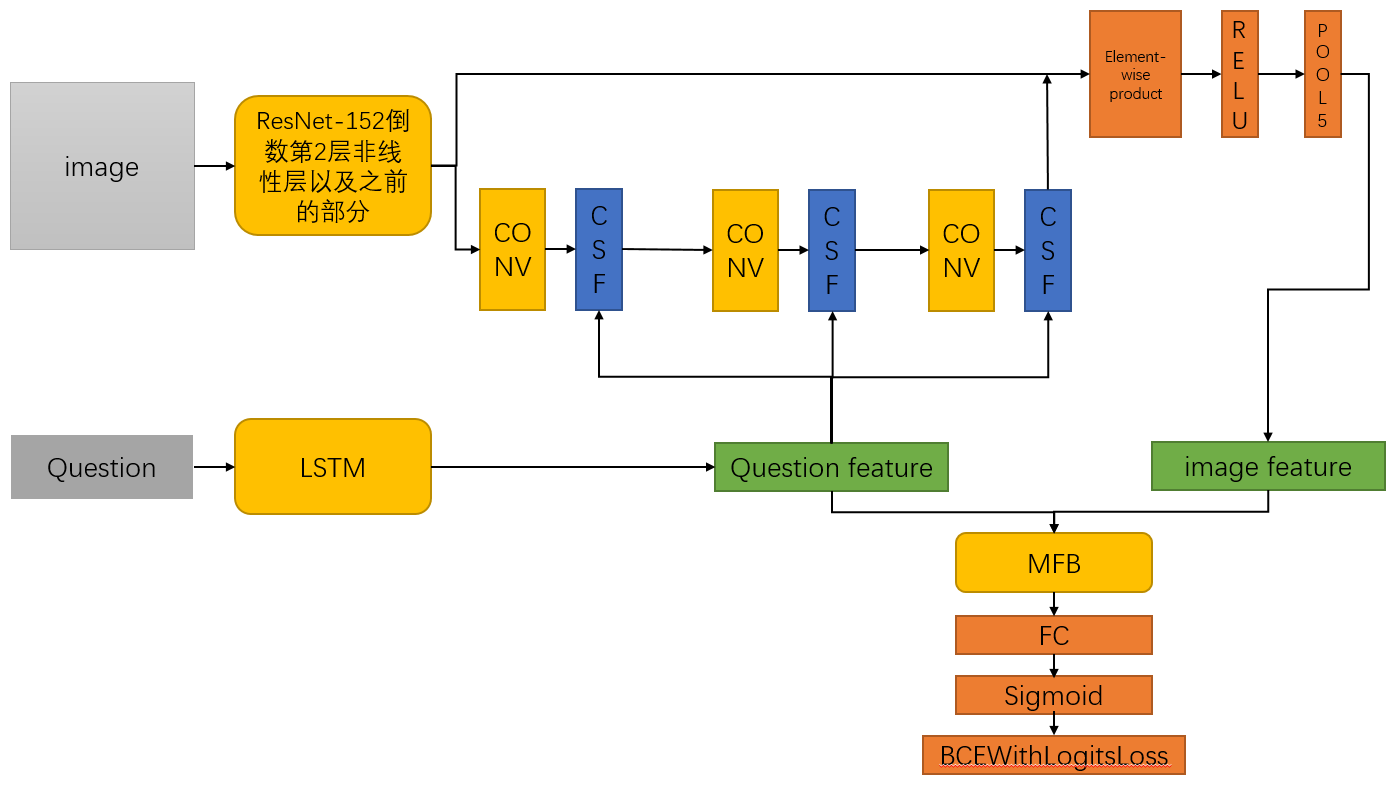
### 2.1主要研究内容

我的主要研究内容在于将注意机制模块进行改进，同时将视觉抽象特征与文本抽象特征用更好的方法相结合，并将以上两者同时应用到视觉问答（VQA）中。其中我主要用到了SCA-CNN[18]中的将Spatial Attention和Channel-wise Attention在卷积神经网络中相结合的技术，不仅在Spatial维度上加入Attention，并在Channel维度上加入Attention，并重复使用多次来得到最终的加入attention的image feature，同时我将其中视觉抽象特征与文本抽象特征用更好的方法（MFB）相结合，改进了其中的原始的将视觉抽象特征与文本抽象特征经过线性映射后直接相加的步骤，采用更为合理的MFB模块将两者结合，希望以此来提高VQA模型的准确性。

### 2.2技术路线

### 2.2.1 数据集

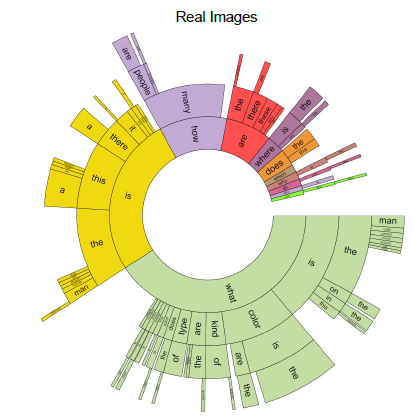
我使用VQA2.0数据集[21]来训练和测试模型，VQA2.0的图像集由来自MS-COCO数据集的约200,000幅图像组成[22]，每个图像3个问题，每个问题10个答案。数据集分为三部分：训练（80k图像和248k问题），评估（40k图像和122k问题）和测试（80k图像和244k问题）。 此外，还有一个名为test-dev的25％测试分割子集。我使用开放式（OE）模式来回答问题，开放式模式要求模型更具图片和问题直接提供答案，而不是从十几个选项中选出对的那个，但是由于开放式（OE）的答案非常难以评估，答案存在歧义性和同义性问题，所以我在这里选出答案集中出现最频繁的3196个答案作为答案的候选集，我的模型将从这3196个答案中选出最合理的那个。



图（1）总体模型

#### 2.2.2 图片处理

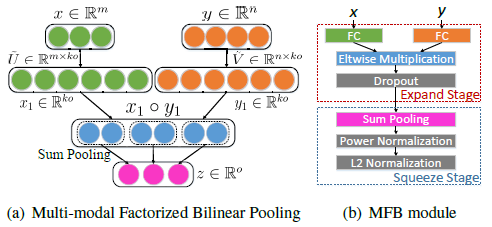
如图（1）所示，这是我总体的模型，首先处理图片，我用目前最流行的方法将图片通过卷积神经网络（CNN，这里主要用到了ResNet-152[19]架构），但与其他VQA模型不同的是，我并不是直接利用pool5层之后的结果，而是首先用原来的ResNet-152最后3层conv层之前的部分处理图片得到初步的image feature ，之后将通过原ResNet-152倒数第三层conv层，之后将得到的feature再通过我自己定义的第一个CSF层（具体实现详见CSF部分），得到，之后又将上一层得到的通过原ResNet-152倒数第二层conv层，之后将得到的feature再通过我自己定义的第二个CSF层，得到，之后又将上一层得到的通过原ResNet-152倒数第一层conv层，之后将得到的feature再通过我自己定义的第三个CSF层，得到，就是得到的最终的加上了attention的image feature。



图（2）VQA2.0数据集中问题类型的大致分布，该图提供了每个问题的前4个单词的分布，样本包括了60K个问题，图中面积越大的部分说明该类型的比重越大。

#### 2.2.3 问题处理

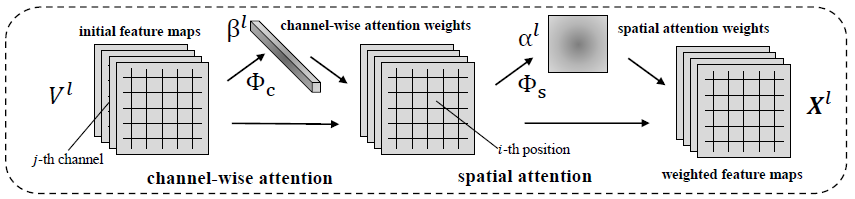
VQA2.0中含有大量的问题，问题的长短不一，我们在这里取所有问题的前14个单词（由于问题长度超出14个单词的问题极少，所以那些超出14个单词的问题在这里直接被截断），将问题中的每个单词用预训练过的glove词向量数据集将单词转化为word embedding表示，若问题不到14个单词，则空出的位置用0向量来表示，之后将这些word embedding通过LSTM[20]来获得最终的question feature ，其中LSTM将会循环14次，我取最后一次循环的hidden state作为question feature 。



图（3）MFB模块的流程图，MFB将两种不同表示的特征向量结合起来，MBF具体可分为expand阶段和squeeze阶段。

#### 2.2.4 图像特征与文本特征的结合

在得到最终的加权image feature 和最终的question feature 之后，我用MFB模块将两者结合起来，其中MFB模块用到了bilinear pooling技术，但和一般的bilinear pooling不同，MFB用Factorized Bilinear Pooling来减少参数个数和计算复杂度，从而来大大降低内存消耗量和运行时间，在结合和之后我得到了最终的融合向量，其中k为人为定义的超参，k越大就是复杂度越高但表示能力越强，之后将c通过FC层之后得到d，与其他VQA模型不同的是，我不将d通过softmax模块来得到答案分布，而是通过sigmoid模块来得到最终的答案分布，之后用BCEWithLogitsLoss来计算loss.



图（4）CSF层的大致流程，首先计算channel-wise attention weight，得到加权image feature之后在计算spatial attention weight，之后再次得到加权image feature，其中权重的计算需要用到question feature



#### 2.2.5 CSF层

添加在CNN倒数后3层conv层之后的CSF层用来将question feature 和临时image feature相结合来计算attention map，其中包括了channel-wise attention map和spatial attention map，首先我们得到第l-1层conv层的image feature ，首先计算channel-wise attention map ，，与在前人论文中提到的channel-wise attention map计算不同，我们先将通过形变转化为,其中代表feature map中的第i个channel，之后对U中的每一个u做meanpooling，得到,之后v和参数做外积得到image matrix ，其中k为人为定义的的超参，之后将T中的每一行与一同作为MFB的输入得到channel-wise attention map ，以channel-wise attention map 作为权重与image feature 相乘得到新的image feature ，之后计算spatial attention map ，，将通过形变转化为,其中，代表feature map中的第i个区域，之后将V中的每一列与一同作为MFB的输入得到spatial attention map , 以spatial attention map 作为权重与image feature相乘得到新的image feature ，在通过下一层conv层之后，得到image feature 如此循环3次来得到最终的image feature 。



### 2.3可行性分析

技术可行性：CNN和LSTM都是比较成熟的算法，能够提取有效的图片特征和文本特征，MFB已被证明在融合不同数据特征上效果出色且效率较高

经济可行性：前期开发工作基本在自己电脑和实验室服务器上完成，开支较小。

法律可行性：无不良企图，以为公众服务为目的。

开发方案的选择：从自身条件出发选择平时使用比较多也比较广泛的python和pytorch作为开发工具。设计的模型也在理论上能够用开发工具实现。

## 3．研究计划进度安排及预期目标

### 3.1进度安排

|  |  |  |
| --- | --- | --- |
| 序号 | 时间 | 内容 |
| 1 | -4月5日 | 开题报告，外文翻译，系统分析 |
| 2 | 4月6日-5月1日 | 完成模型设计，并初步实现，得到较高的准确率 |
| 3 | 5月2日- 5月4日 | 检查模型，中期检查 |
| 4 | 5月5日-5月15日 | 调整模型，进一步提高准确率 |
| 5 | 5月16日-5月31日 | 撰写毕业论文，报告 |
| 6 | 6月1日-6月4日 | 成果验收 |
| 7 | 6月5日-6月12日 | 论文答辩 |
| 8 | 6月13日-6月15日 | 毕业设计总结、文档整理 |

### 3.2预期目标

实现上述模型，并希望在某些地方做出更多的改进来进一步提高模型的准确率和计算速度。

## 4．参考文献

[1] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Doll´ar,

J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig.

From Captions to Visual Concepts and Back. In CVPR, 2015. 1, 2

[2] X. Chen and C. L. Zitnick. Mind’s Eye: A Recurrent Visual Representation

for Image Caption Generation. In CVPR, 2015. 1, 2

[3] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan,

K. Saenko, and T. Darrell. Long-term Recurrent Convolutional

Networks for Visual Recognition and Description. In CVPR,2015. 1, 2

[4] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying Visual-

Semantic Embeddings with Multimodal Neural Language Models.TACL, 2015. 1, 2

[5] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for

Generating Image Descriptions. In CVPR, 2015. 1, 2

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A

Neural Image Caption Generator. In CVPR, 2015. 1, 2

[7] J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image coattention for visual question answering,” in Proc. Advances Neural Information Processing Systems, 2016, pp. 289–297.

[8] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked attention networks for image question answering,” in Proc. IEEE Conf. Computer Vision and Pattern

Recognition, 2016, pp. 21–29.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of

word representations in vector space,” arXiv Preprint, arXiv:1301.3781, 2013.

[10] C. Xiong, S. Merity, and R. Socher, “Dynamic memory networks for visual and textual question answering,” in Proc. Int. Conf. Machine Learning, 2016, pp. 2397–2406.

[11] P. Wang, Q. Wu, C. Shen, A. v d. Hengel, and A. Dick, “Explicit knowledge-based reasoning for visual question answering,” arXiv Preprint, arXiv:1511.02570, 2015.

[12] Q. Wu, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel, “Ask me anything: Freeform visual question answering based on knowledge from external sources,” in Proc.IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 4622–4630.

[13] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik, “Learning to generalize to new compositions in image understanding,” arXiv Preprint,

arXiv:1608.07639, 2016.

[14] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. J. Mooney, K. Saenko,

and T. Darrell, “Deep compositional captioning: Describing novel object categories without paired training data,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2015, pp. 1–10.

[15] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847, 2016.

[16] J.-H. Kim, K. W. On, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325, 2016.

[17] Zhou Yu†, Jun Yu†\_, Jianping Fan‡, Dacheng Tao§ Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering in ICCV, 2015.

[18] Long Chen1 Hanwang Zhang2 Jun Xiao1\_ Liqiang Nie3 Jian Shao1 Wei Liu4 Tat-Seng Chua5 SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning arXiv:1611.05594v2 [cs.CV] 12 Apr 2017

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015. 4

[20] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In International Conference on Computer Vision (ICCV), pages 2425–2433, 2015. 1, 2, 3, 5, 6, 7, 8

[21] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In International Conference on Computer Vision (ICCV), pages 2425–2433, 2015. 1, 2, 3, 5, 6, 7, 8

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision (ECCV), pages 740–755, 2014. 5

# 三、外文翻译

## 视觉问答中的多模型双线性池化分解与共同注意学习

# 摘要

视觉问答（VQA）具有挑战性，因为它需要同时理解图像的视觉内容和问题的文本内容。 用细粒度的方式表达图像和问题的方法以及融合这些多模型特征的方法在模型性能中起关键作用。 已经证明基于双线性池化的模型优于传统的VQA线性模型，但其高维表示和高计算复杂性严重限制了它们在实践中的适用性。 对于多模型特征融合，我们在这里开发了一种多模型双线性池化分解（MFB）方法，高效并有效地组合多模型特征，与其他双线性池化方法相比，我们的模型具有更高的性能。对于细粒度的图像和问题表示，我们使用端到端的深度网络架构来开发“共同注意”机制，共同学习图像注意和问题注意。 在新的网络体系结构中将所提出的MFB方法与共同注意学习相结合，为VQA提供了一个统一的模型。 我们的实验结果表明，具有共同注意模型的单个MFB在真实世界VQA数据集上实现了最先进的性能表现。

## 1.介绍

由于计算机视觉和自然语言处理方面的最新进展， 计算机预计将能够在不久的将来自动理解图像和自然语言的语义。 这样的进步同时激发了新的研究课题，如图像文本检索[35, 37]，图像字幕[5, 34]和视觉问答[3, 19]。

与图像文本检索和图像字幕相比（只需要底层算法为给定图像搜索或生成自由形式的文本描述），视觉问答（VQA）是一项更具挑战性的任务，需要细致的理解图像和问题的语义并且支持复杂推理以正确推测最佳匹配答案。 在某些方面，VQA可以被视为图像字幕和图像文本检索的扩展。 因此，建立有效的VQA算法，可以实现与人类相近的表现，这是实现人工智能的重要一步。

现有的VQA通常有三个阶段：（1）将图像表示为视觉特征和将问题表示为文本特征; （2）结合这些多模型特征以获得融合的图像问题特征;（3）使用集成图像问题特征来学习多级分类器并预测最佳匹配答案。深度神经网络（DNNs）有效且灵活，许多现有方法在一个DNN模型中对三个阶段进行建模，并通过反向传播以端到端方式训练模型。在三个阶段中，特征表示和多模型特征融合特别影响VQA性能。

对于多模型特征融合，大多数现有的方法只是使用线性模型进行多模型特征融合（例如拼接或逐元素相加），以将图像的视觉特征与问题的文本特征相结合[39, 18]。 由于多模型特征分布可能变化很大，因此通过这种线性模型获得的综合图像问题表示可能不足以充分表达图像的视觉特征与问题的文本特征之间的复杂关系。 与线性池化不同，双线性池化[28] 最近已经被用于集成不同的CNN特征来用于细粒度图像识别[17]。 然而，高维的输出特征和大量的模型参数可能会严重限制双线性池化的适用性。 Fukui等人提出了多模型紧凑双线性（MCB）池化模型，使用Tensor Sketch算法有效地同时减少参数和计算时间[6]。 使用MCB模型，该团队为VQA提出了网络架构，并赢得了2016年VQA挑战。然而，MCB模型依赖高维输出特性来保证强大的性能，因为内存使用量巨大，这可能会限制其适用性。为了克服这个问题，Kim等人提出了基于两个特征向量Hadamard 积的多模低秩双线性（MLB）池化模型[12]。 由于MLB生成较低维度的输出特征和较少参数的模型，因此与MCB竞争非常激烈。 然而，MLB收敛速度较慢，对学习的超参数敏感。 为了解决这些问题，我们在这里开发多模型分解双线性池化（MFB）方法，该方法具有MLB紧凑输出特性和MCB强大表现能力的双重优势。

关于特征表示，直接使用全局特征进行图像表示可能会引入与给定问题无关的噪音信息。 因此，视觉注意机制 [34] 就自然而然得被添加到VQA任务中，自适应学习给定问题的相关图像区域。 建模视觉注意会显着提高性能[6]。 然而，大多数现有的方法只是在不考虑问题注意的情况下模拟图像注意，尽管问题注意也非常重要，因为以自然语言解释的问题也可能包含可被视为噪音的俗语。 因此，基于我们的MFB方法，我们使用共同学习模块为VQA任务设计了深度网络架构，共同学习图像注意和问题注意。

总之，这项研究的主要贡献如下：首先，我们开发了一种简单而有效的多模型分解双线性池化（MFB）方法，将图像的视觉特征与问题的文本特征相融合。 MFB 明显优于现有的多模型双线性池化方法，如MCB [6]和MLB [12]。 其次，基于MFB模块，设计了一种共同学习体系结构，共同学习图像和问题的注意。 我们采用共同注意模型的MFB方法实现了VQA数据集的目前最高性能。 我们还进行详细和广泛的实验，以说明为什么我们的MFB 方法是有效的。 我们的实验结果证明归一化技术在双线性模型中非常重要。

## 3.多模式因式分解双线性池

给定不同形式的两个特征向量，例如图像的视觉特征和问题的文本特征，最简单的多模态双线性模型定义如下：



其中是投影矩阵，是双线性模型的输出。 因为偏倚隐含在W中，因此这里省略了它。为了获得o维输出z，我们需要学习。 尽管双线性汇集可以有效地捕捉特征维度之间的相互作用，但它也引入了大量的参数，导致高计算成本和过度拟合风险。受到单模型数据矩阵分解技巧的启发[15, 25]，投影矩阵在等式(2)可以被分解为两个低秩矩阵：



其中k是被分解的矩阵和的因子或潜在维数。是Hadmard乘积或两个向量的逐元素乘法，是一个全1向量。



为了通过公式Eq.(3)获得输出特征，要学习的权重是相应的两个三阶张量和。 不失一般性的，我们可以用简单的重塑操作重新定义U和V作为2-D矩阵和。相应地，Eq.(3)可以重写为如下：



其中函数SumPooling(x,k)意味着使用尺寸为k的一维非重叠窗口在x上进行相加池化操作。我们将这个模型命名为Multimodal Factorized Bilinear pooling(MFB)。

MFB的详细程序如图[1(a)](#_bookmark7)所示。该方法可以通过组合一些常用的层，如全连接层，逐元素乘和池化层来轻松实现。 此外，为了防止过度拟合，在逐元素乘法层之后添加一个dropout层。由于引入了逐元素乘法，因此输出神经元的大小可能会有显著的变化，并且该模型可能会收敛到令人不满的局部最优。 因此，类似[6]， 在MFB输出之后加上功率归一化()和L2归一化层()。整个MFB模块的流程图如图1(b)所示。



与MLB的关系。Eq.(4)表明Eq.(1)中的MLB是MFB当k=1的特例，其对应于rank-1因式分解。形象地说，MFB可以分解为两个阶段(见图1(b)):首先，将来自不同模块的特征扩展到高维空间，然后与逐元素乘法相结合。 之后，执行相加池化层以及归一化层以将高维特征压缩为紧凑的输出特征，而MLB直接将特征投影到低维输出空间并执行逐元素乘法。因此，在输出特征具有相同维度的情况下，MFB的表示能力比MLB更强大。

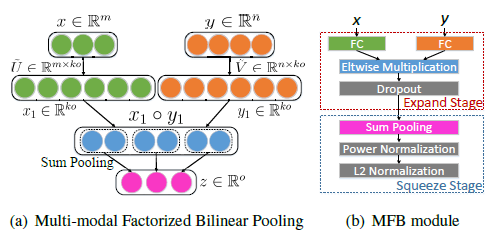


图1.多模型分解双线性池化的流程图和MFB模块的完整设计。

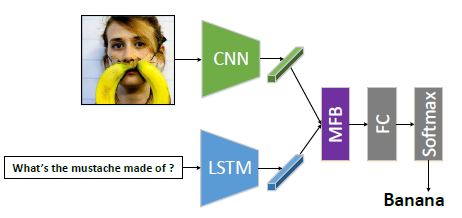


图2. VQA的MFB基础网络体系结构

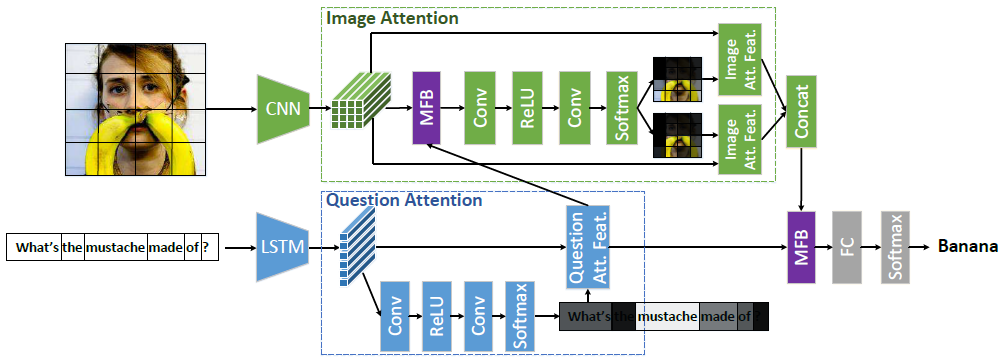


图3.具有Co-Attention网络体系结构的MFB用于VQA。 与MFB基础网络不同，图像和问题首先分别表示为细粒度特征。然后，在框架中同时建模问题注意和图像注意模块，以提供更准确的答案预测。

## 4.VQA网络架构

VQA的目标是回答关于图像的问题。模型的输入包含图像和关于图像的相应问题。我们的模型提取图像和问题表示，使用图[1(b)](#_bookmark8)中的MFB模块集成多模块特征，将每个单独的答案视为一个类，并执行多分类任务以预测正确的答案。在本节中，介绍了两种网络体系结构。 首先是带有一个MFB模块的MFB基础模型，该基础模型用于对使用不同超参的模型进行对照分析，以便与其他基准模型进行比较。 第二个网络引入Co-Attention学习，同时学习图像注意和问题注意，以更好地捕捉图像和问题之间的细粒度关联，这使该模型成为一个更具表达能力的模型。

## 4.1 MFB基础模型

如同 [6]，我们使用152层ResNet模型提取图像特征[7]，并预先在ImageNet数据集上进行训练。 图像被调整大小为448x448，并且用2048-D pool5层的特征（用L2归一化）用于图像表示。问题首先被标记为单词，然后进一步转化为最大长度为T的one-hot特征向量。然后，one-hot特征向量通过一个embedding层并进入一个带有1024个隐藏单元的双层LSTM网络[8]。每个LSTM层为每个字输出一个1024-D特征。 如同[6]，我们从每个LSTM网络中提取最后一个词的输出特征，并串联从两个LSTM网络中获得的特征以形成用于问题表示的2048-D特征向量。 为了预测答案，我们简单地使用前N个最频繁的答案作为N个类别，因为它们遵循长尾分布。

提取的图像和问题特征被送到MFB模块以生成融合特征z。最后，z被送到用KL散度损失的N分类器。因此，ResNet以外的所有权重（由于GPU内存限制）都以端对端方式共同优化。整个网络架构如图2所示.

### 4.2 MFB与Co-Attention

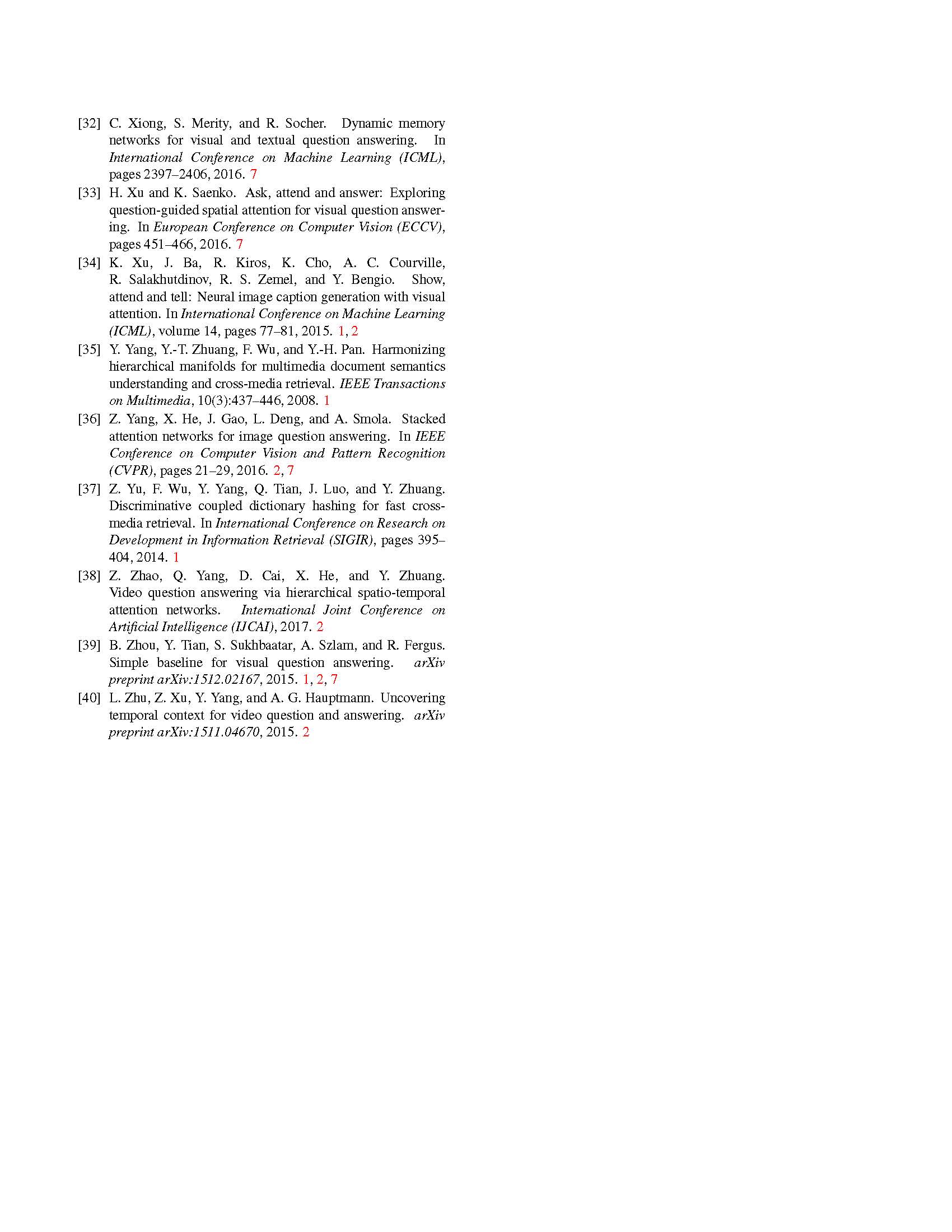
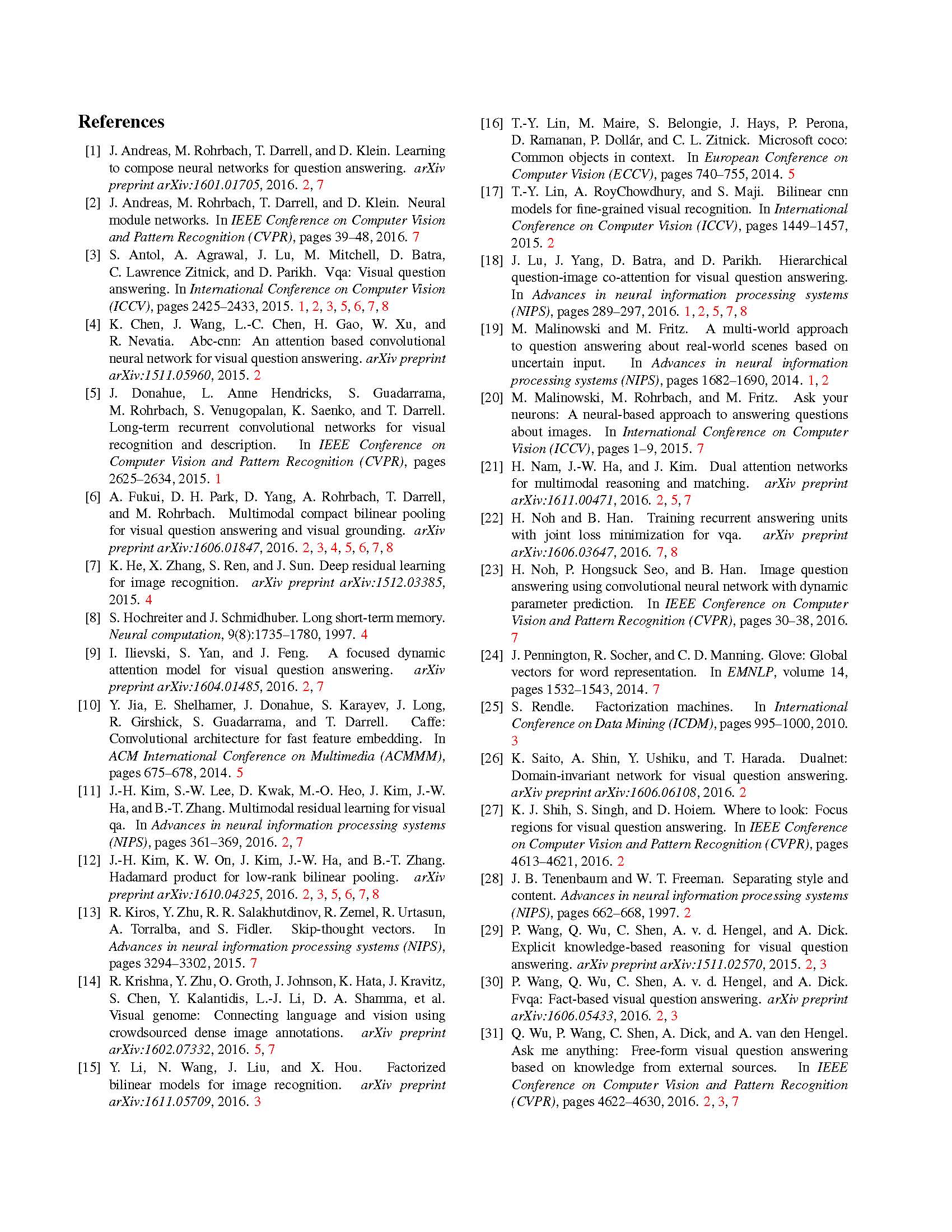
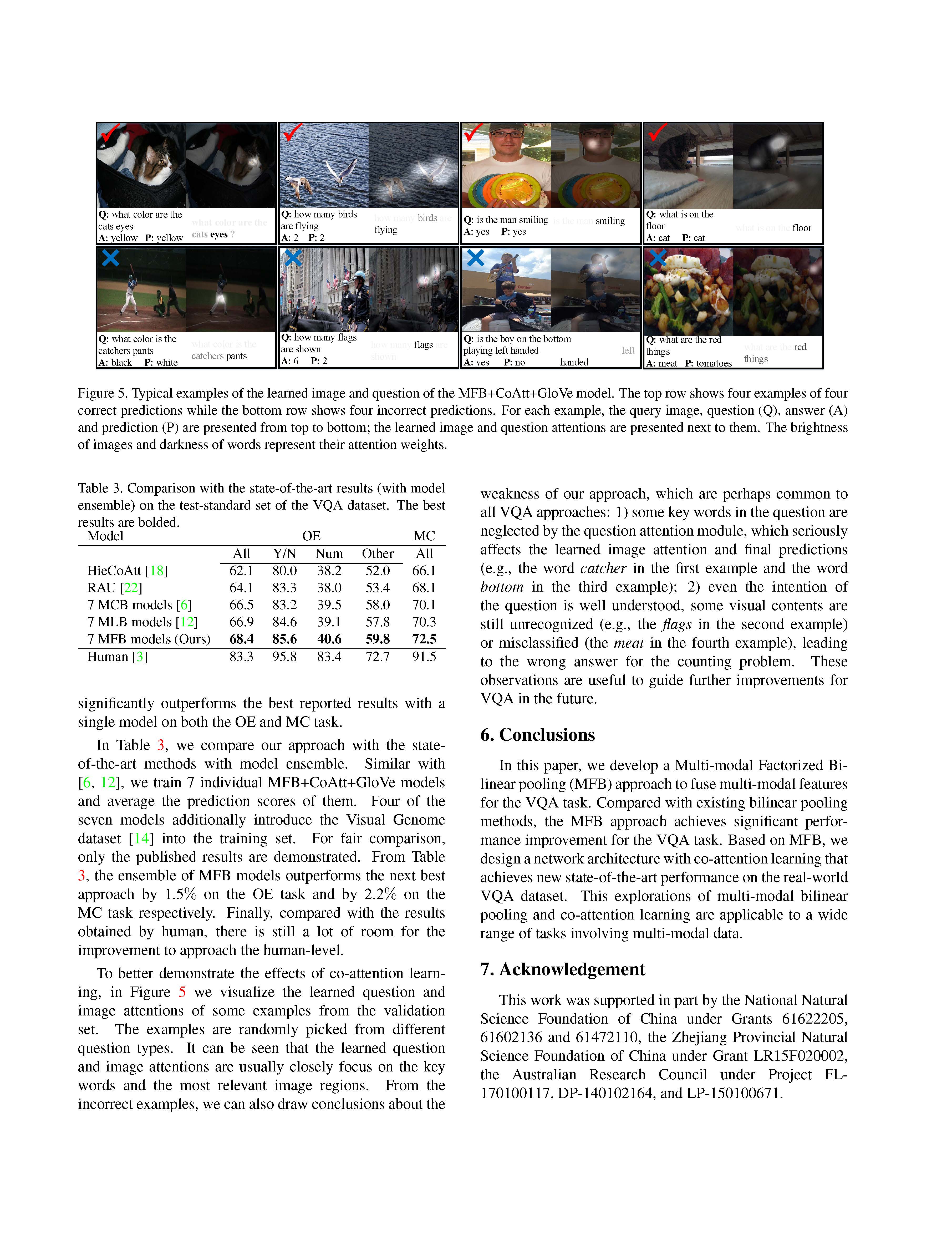
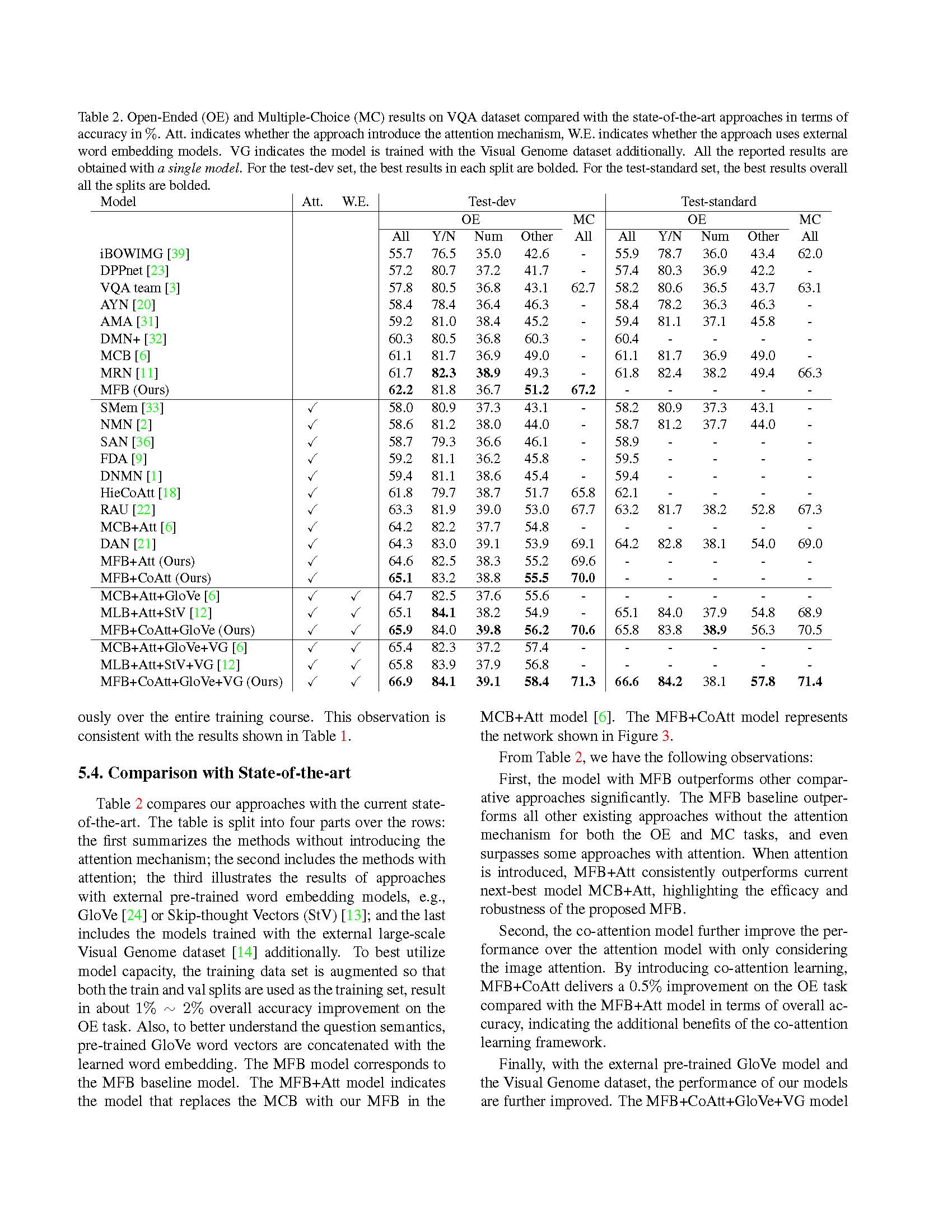
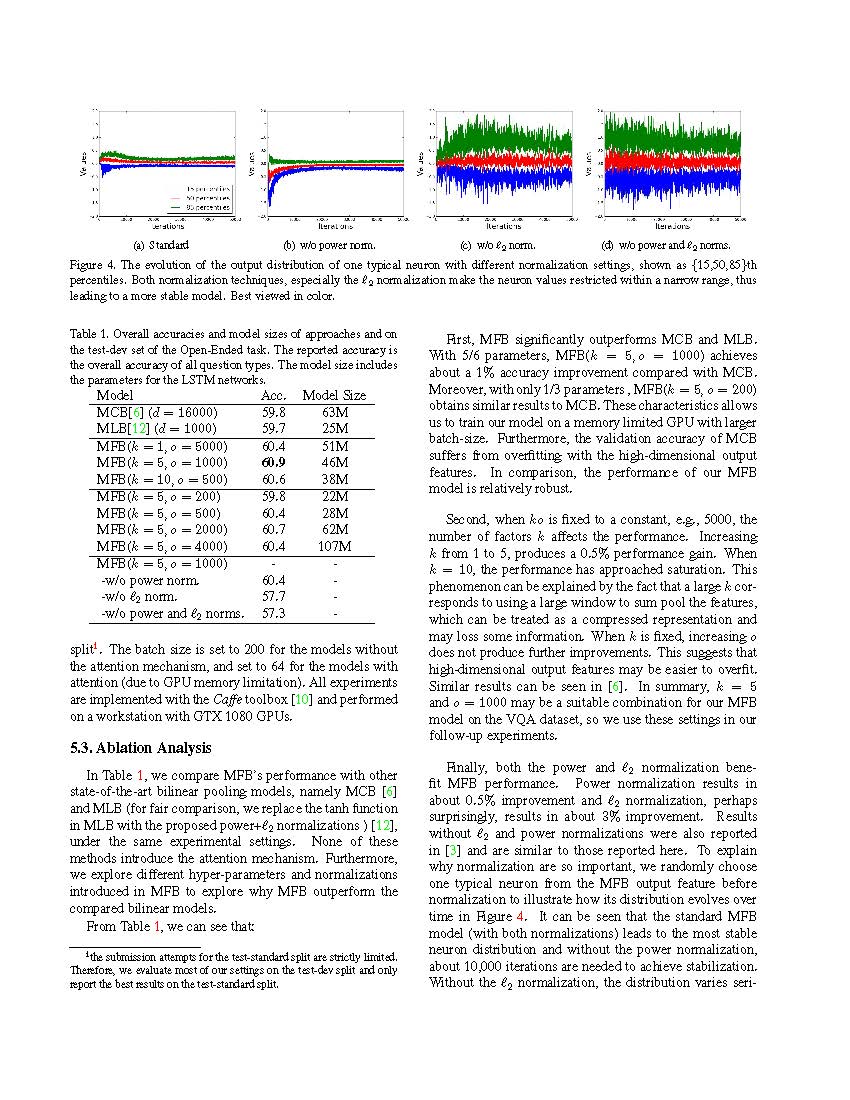
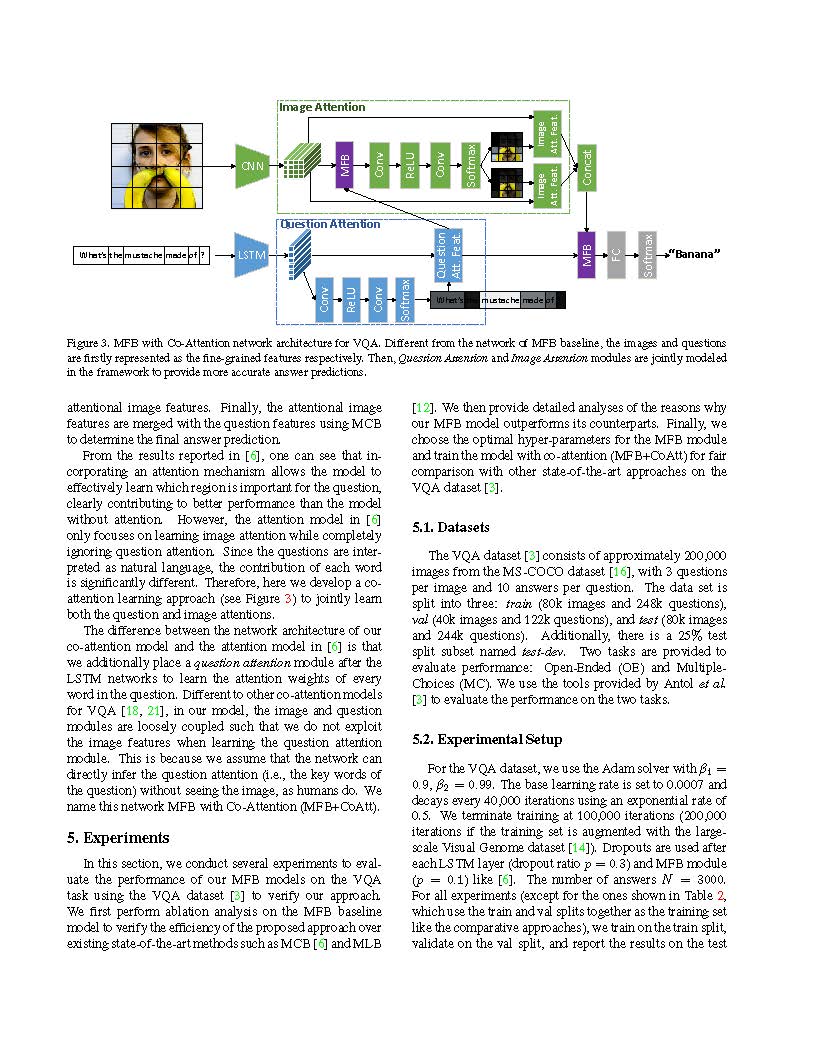
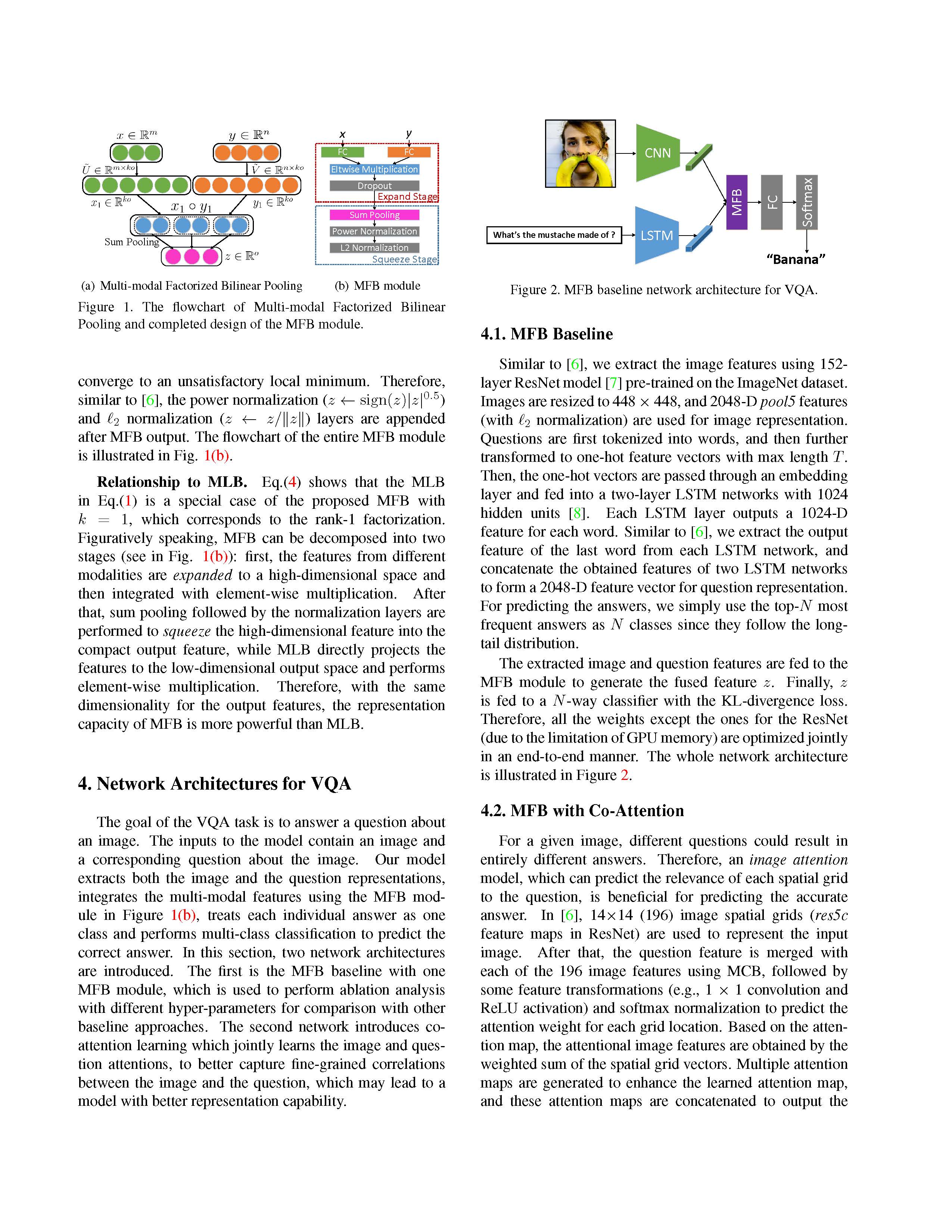
对于给定的图像，不同的问题可能导致完全不同的答案。 因此，可以预测每个图像网格与问题的相关性的图像注意模型对于预测准确答案是有益的。在[6]中，

14x14（196）图像网格（ResNet中的res5c特征图） 用于表示输入图像。之后，使用MCB将问题特征与196个图像特征中的每一个融合，随后进行一些特征变换（例如，1x1卷积和ReLU激活）和softmax归一化以预测每个网格位置的注意权重。 基于注意权重图，注意图像特征通过图像网格矢量的加权和获得。生成多个注意权重图以增强习得的注意权重图效果，并将这些注意权重图连接起来来输出注意图像特征。最后，用MCB来融合注意图像特征与问题特征来进行最终答案预测。

从[6]中我们可以看到，结合注意机制可以让模型有效地了解哪个区域对于问题是重要的，相对于没有注意机制的模型有明显的性能提高。然而，注意模型在[6]只专注于学习图像的注意，同时完全忽视问题的注意。由于问题被解释为自然语言，因此每个单词的贡献明显不同。 因此，我们在这里开发了一种共同注意的学习方法（参见图3）来共同学习问题和图像的注意。

我们的共同注意模型的网络结构和在[6]中的注意模型的结构的不同之处在于我们另外在LSTM网络之后放置一个问题注意模块来学习问题中每个单词的注意权重。 与VQA的其他共同注意模型不同[18, 21]的是，在我们的模型中，图像和问题模块是松耦合的，因此我们在学习问题注意模块时不会利用图像特征。这是因为我们认为网络可以不看到图像而直接推断出问题的注意（即问题的关键词），就像人类那样。 我们将这个网络命名为MFB with Co- Attention（MFB + CoAtt）。

# **四、外文原文**



# 毕业论文（设计）文献综述和开题报告考核

**导师对开题报告、外文翻译和文献综述的评语及成绩评定：**

|  |  |  |  |
| --- | --- | --- | --- |
| **成绩比例** | **文献综述**  **（10%）** | **开题报告**  **（15%）** | **外文翻译**  **（5%）** |
| **分 值** |  |  |  |

**导师签名**

**年 月 日**

**学院盲审专家对开题报告、外文翻译和文献综述的评语及成绩评定：**

|  |  |  |  |
| --- | --- | --- | --- |
| **成绩比例** | **文献综述**  **（10%）** | **开题报告**  **（15%）** | **外文翻译**  **（5%）** |
| **分 值** |  |  |  |

**开题报告审核负责人（签名/签章）**

**年 月 日**