涉密论文 □　公开论文 □

浙江大学

本 科 生 毕 业 论 文

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

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

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）

第二部分 文献综述和开题报告

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。

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。目前，视觉问答模型性能提升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这 4 个方面。本文主要针对 attention 权重的计算和图像特征与文本特征融合这两个方面，以及其他细节方面的地方相对于前人的模型做出了改进。本文的主要工作在于本文使用 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。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。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)。

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 小节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 [15][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y^I，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$\mathrm{i} = 1 \ldots \mathrm{m}$提取局部特征$\{y_i^I\}$。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$a_i = f^{att}(y_i^I, y^Q)$。函数$f^{att}(.)$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$y^I = \sum_i a_i y_i^I$。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a_i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，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

$$\text{Accuracy(a)} = \min\left\{\frac{Count(a)}{3}, 1\right\}$$

公式 1

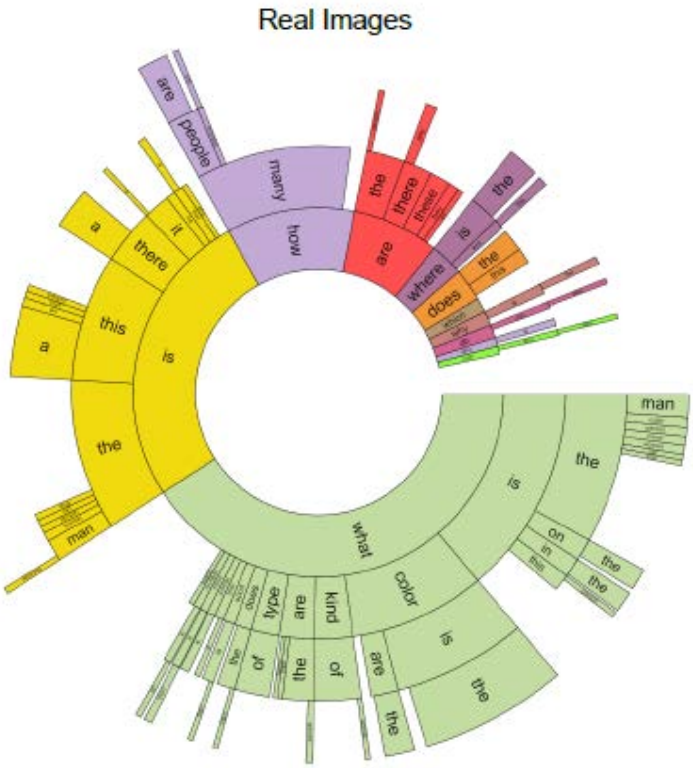

图 2 VQA2.0 数据集中问题类型的大致分布，该图提供了每个问题的前 4 个单词的分布，样本包括了 60K 个问题，图中面积越大的部分说明该类型的比重越大。

3.2 总体模型

3.2.1 Baseline

3.2.1.1 图片处理

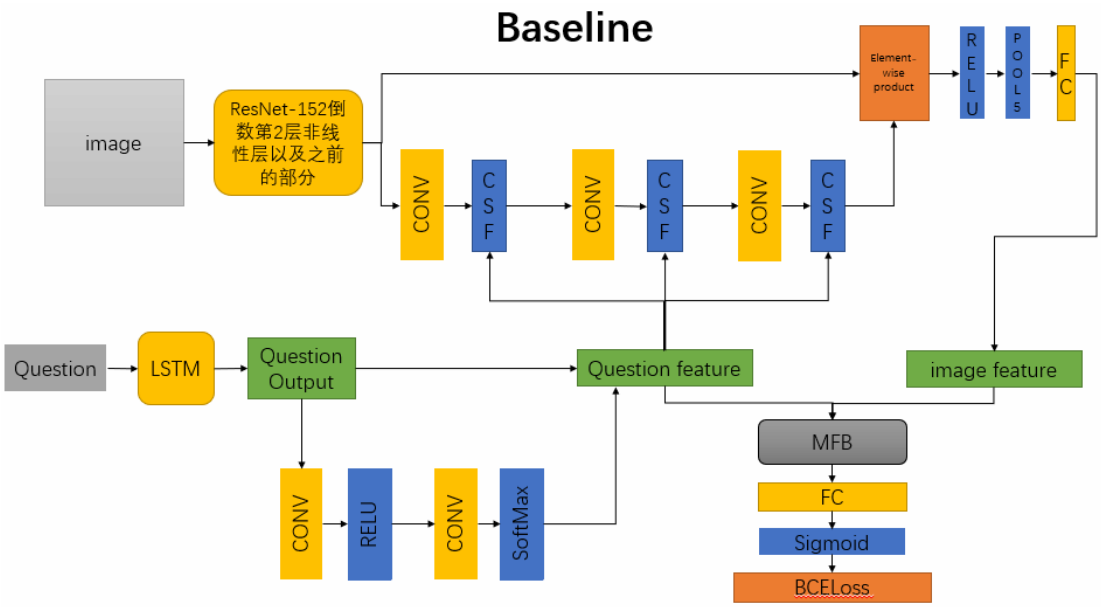

图 3 总体模型-Baseline，其中的 CSF 是本文自定义的模型

如**图 3** 所示，这是总体的 Baseline 模型，首先处理图片，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 b_{que}，其中 LSTM 将会循环 14 次，我取最后一次循环的 hidden state 作为 question feature

b_{que}。

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 b_{que}。

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 a_{img}和最终的 question feature b_{que}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a_{img}和b_{que}之后我得到了最终的融合向量$\mathrm{c} = \mathrm{SumPooling}\left(\tilde{U}^T a_{img} \circ \tilde{V}^T b_{que}, k\right)$，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.

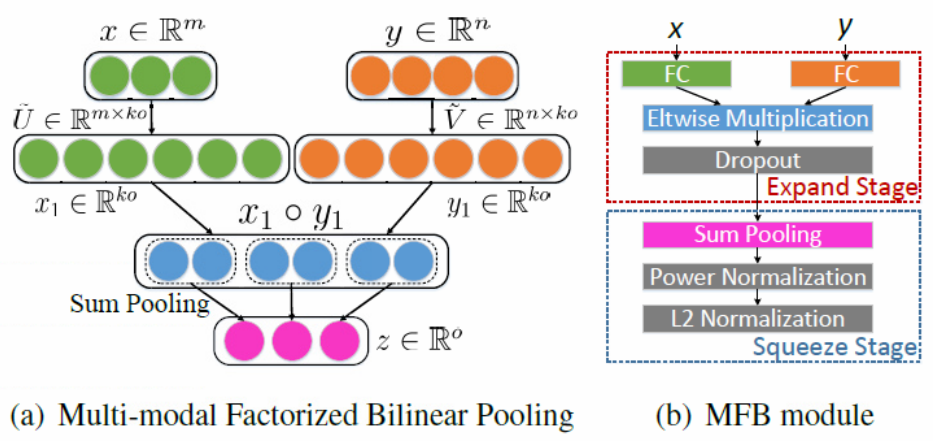

图 4 MFB 模块的流程图，MFB 将两种不同表示的特征向量结合起来，MBF 具体可分为

expand 阶段和 squeeze 阶段。

3.2.2 Final

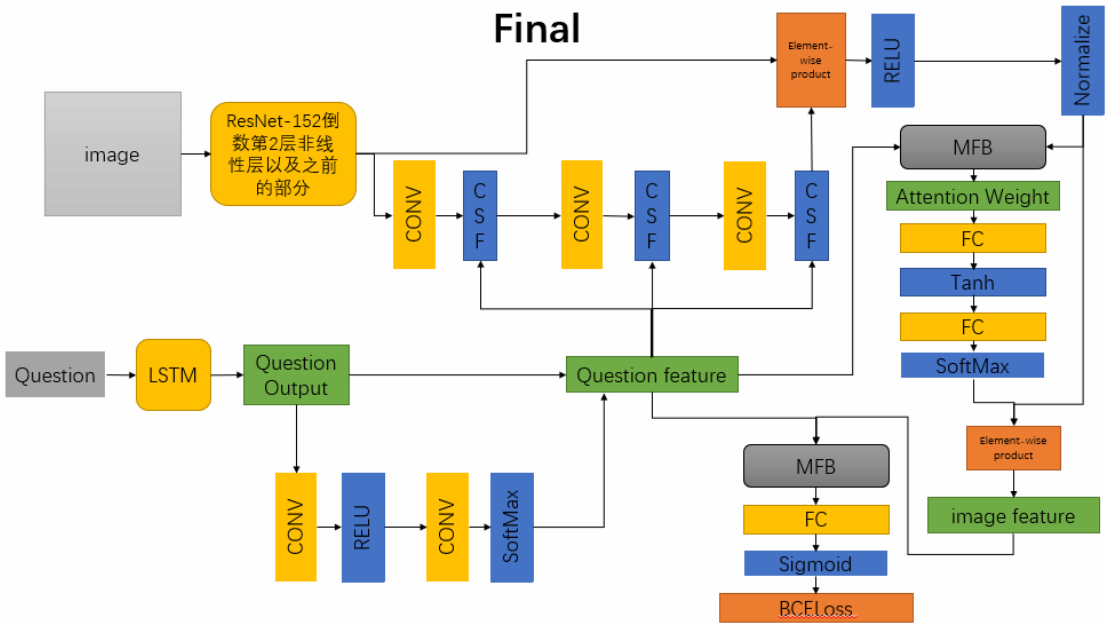

图 5 总体模型：Final， Final 相对于 Baseline 来说准确率有较大的提升

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

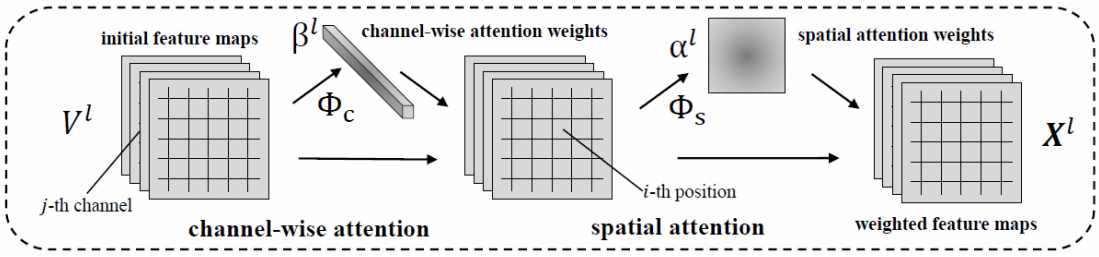

图 6 CSF 子模块的流程，文本创建了两种 CSF 模块，分别为 CSF_A 和 CSF_B，两者差别仅在于计算 attention 权重的函数不同，其流程都与上图一样，本文进行了相关对照试验并对使用不同模块的总体模型的准确率进行了比较

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 $V \in R^{C*H*M}$ 重塑成$U = [u_1, u_2, \dots, u_C]$，其中$u_i \in R^{W*H}$，代表了第 i 个 channel 的 feature，之后本文对每个u_i做 mean pooling，得到向量$v = [v_0, v_1, \dots, v_C], v \in R^C$，$v_i$是$u_i$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 $V \in R^{C*H*M}$。

$$b = \tanh\big((W_c \otimes v + b_c) + W_{hc}h\big)$$

$$\beta = \text{softmax}(W_{k2}b + b_{k2})$$

其中$W_c \in R^k, W_{\text{hc}} \in R^{k*d}, W_{k2} \in R^k, b_c \in R^k, b_{k2} \in R^1$。

公式 2

之后计算 spatial-wise 权重，将由上一层得到的 tensor $V \in R^{C*H*M}$重塑成$V = [v_1, v_2, \dots, v_m]$，其中$v_i \in R^c, m = H * M$，$v_i$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。

$$a = \tanh\big((W_s V + b_s) + W_{hs}h\big)$$

$$\alpha = \text{softmax}(W_{k1}a + b_{k1})$$

其中$W_s \in R^{k*C}, W_{hs} \in R^{k*d}, W_{k1} \in R^k, b_s \in R^k, b_{k1} \in R^1$

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 b_{que}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 b_{que}和临时 image feature。

首先我们得到第 l-1 层 conv 层的 image feature V^l，首先计算 channel-wise attention map $\beta^l, \beta^l = \varphi_s(b_{que}, V^l)$，与在前人论文中提到的 channel-wise attention map 计算不同，我们先将$V^l \in R^{W*H*C}$通过形变转化为$\mathrm{U} = [u_1, u_2, \dots, u_C]$，其中$u_i \in R^{W*H}$代表 feature map 中的第 i 个 channel，之后对 U 中的每一个 u 做 meanpooling，得到$\mathrm{v} = [v_1, v_2, \dots, v_C], v \in R^C$，之后 v 和参数$W_c \in R^k$做外积得到 image matrix $\mathrm{T} = W_C \otimes v, T \in R^{k*C}$，其中 k 为人为定义的的超参，之后将 T 中的每一行与b_{que}一同作为 MFB[33]的输入得到 channel-wise attention map β^l，以 channel-wise attention map β^l作为权重与 image feature V^l相乘得到新的 image feature V^{l1}。

之后计算spatial attention map α^l，$\alpha^l = \varphi_c(b_{que}, V^{l1})$，将$V^{l1}$通过形变转化为 $\mathrm{V} = [v_1, v_2, \dots, v_m], m = W * H, v_i \in R^C$，其中$v_i$代表feature map中的第i个区域，之后将V中的每一列与b_{que}一同作为MFB的输入得到spatial attention map α^l，以 spatial attention map α^l作为权重与image feature V^{l1}相乘得到新的image feature X^l。

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