

电子科技大学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER THESIS



论文题目 复杂推理的起点：知识库在视觉问答的应用

学科专业 控制科学与工程

学号 201721070835

作者姓名 陈小兵

指导老师 郑文锋 副教授

分类号 _____ 密级 _____

UDC 注¹ _____

学 位 论 文

复杂推理的起点：知识库在视觉问答的应用

(题名和副题名)

陈小兵

(作者姓名)

指导老师

郑文锋 副教授

电子科技大学 成都

(姓名、职称、单位名称)

申请学位级别 硕士 学科专业 控制科学与工程

提交论文日期 _____ 论文答辩日期 _____

学位授予单位和日期 电子科技大学 年 月

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

The Start of Complex Reasoning: The Application of Knowledge Base in Visual Question Answering

**A Master Thesis Submitted to
University of Electronic Science and Technology of China**

Discipline: Control Science and Engineering

Author: Xiaobing Chen

Supervisor: Dr. Wenfeng Zheng

School: School of Automation Engineering

摘要

得益于神经网络带来的机器视觉和自然语言处理的快速发展，视觉问答是人工智能领域近几年兴起的热门研究方向。视觉问答是给定一张图片和一个图像相关的自然语言问题，输出问题答案的人工智能任务。虽然由于图像内容的复杂性和问题的开放性使得任务充满挑战，但其相较于人工智能中零散的任务，更接近通用人工智能，因此视觉问答模型的研究具有高度的研究价值和广阔的应用场景。目前已经提出的模型按照是否引入外源知识库可以分为联合嵌入模型和基于知识库的模型两类，本文将分别对这两种模型进行研究。本文首先对视觉问答的国内外研究状况做了详细的调查研究，对其涉及的模型、数据集、知识库均作出了系统性的分析。针对已有的联合嵌入模型使用静态词向量的缺陷，本文使用动态词向量对其进行改进，并结合 Faster R-CNN 和注意力机制，构建了 None KB-Specific Network(N-KBSN) 模型。为了进一步提高模型的准确性，文章首次提出引入知识库的图嵌入，在 N-KBSN 模型的基础上，构建了 KB-Specific Network(KBSN) 模型。在多个数据集上的实验结果显示，本文改进的动态词向量能够实现提供更好的文本特征，而引入的知识库图嵌入也显著提高了结果的准确率。另外，本文还创建了两个从 DBpedia 提取的、具有丰富语义的知识库 dbv 和 dba，这两个知识库为以后的知识库图嵌入研究提供数据集支持。

关键词：视觉问答，联合嵌入模型，知识库，N-KBSN，KBSN

ABSTRACT

Benefited from the rapid development of machine vision and natural language processing brought by neural networks, Visual Question Answering(VQA) is a popular research direction in the field of artificial intelligence in recent years. VQA is a task that outputs the answer to the question given a picture and a natural language question related to the image. Although the complexity of the image content and the open-ended form of the question make the task full of challenges, it is closer to general artificial intelligence than the other tasks and has high research value and broad applications. The models proposed can be generally divided into two types: joint embedding models and Knowledge Base(KB)-based models according to whether the external knowledge base is involved. In this paper, we will address these two types perspectively. We first overviewed past related researches, including VQA models, datasets and knowledge bases. To improve past joint embedding models, which all used static word vectors, we proposed to use dynamic word vectors, combining Faster R-CNN and attention mechanism to construct the None KB-Specific Network (N-KBSN) model. To further improve the accuracy of the model, we first proposed to introduce the graph embedding of the knowledge base to the N-KBSN model, the KBSN model was constructed. Experimental results on several datasets show that the improved dynamic word vectors can provide better text features, and the introduction of knowledge base graph embedding also significantly improves the accuracy of the results. In addition, we also created two knowledge bases dbv and dba with rich semantics extracted from DBpedia. These two knowledge bases can be the basis for future research on the embedding of knowledge base graphs.

Keywords: Visual Question Answering, Joint Embedding Model, Knowledge Base, N-KBSN, KBSN

目 录

| | |
|----------------------------------|-----------|
| 第一章 绪 论 | 1 |
| 1.1 研究工作的背景与意义 | 1 |
| 1.2 视觉问答的国内外研究状况 | 2 |
| 1.2.1 联合嵌入模型 | 4 |
| 1.2.2 基于外源知识库的视觉问答模型 | 13 |
| 1.3 本文的主要贡献与创新 | 21 |
| 1.4 本论文的结构安排 | 22 |
| 第二章 视觉问答数据集 | 24 |
| 2.1 基于视觉的数据集 | 25 |
| 2.2 基于知识的数据集 | 29 |
| 2.3 实验数据集 | 32 |
| 2.4 本章小结 | 33 |
| 第三章 基于动态词向量的联合嵌入模型 | 34 |
| 3.1 基于 Faster R-CNN 的图像特征化 | 36 |
| 3.2 基于 ELMo 的文本特征化 | 38 |
| 3.3 基于多头注意力机制的特征增强 | 41 |
| 3.4 实验 | 44 |
| 3.4.1 实验设置 | 44 |
| 3.4.2 剔除研究 | 45 |
| 3.4.3 实验结果分析 | 45 |
| 3.5 本章小结 | 45 |
| 第四章 基于知识库图嵌入的视觉问答模型 | 46 |
| 4.1 知识库概述 | 46 |
| 4.2 KBSN 模型 | 54 |
| 4.2.1 知识库子图提取 | 56 |
| 4.2.2 知识库子图嵌入 | 60 |
| 4.2.3 基于图神经网络的图嵌入 | 61 |
| 4.3 实验 | 63 |
| 4.3.1 数据集 | 63 |
| 4.3.2 实验结果分析 | 63 |

目录

| | |
|--------------------------|-----------|
| 4.4 本章小结 | 63 |
| 第五章 全文总结与展望 | 64 |
| 5.1 全文总结 | 64 |
| 5.2 后续工作展望 | 64 |
| 致 谢 | 65 |
| 参考文献 | 66 |
| 攻读硕士学位期间取得的成果 | 75 |

第一章 绪论

1.1 研究工作的背景与意义

视觉问答（VQA）是近几年学界新兴研究的热门方向之一。得益于神经网络架构在自然语言处理和图像识别相关任务的成功应用，学界将研究的视线移向对系统智能要求更高的视觉问答任务。视觉问答任务是一类输入为图像和用自然语言表达的文本问题，输出为基于图像内容理解并且用自然语言方式呈现的答案的计算机视觉任务。简而言之，任务目标是构建一个像人类智能一样的问答系统——能够从给定的图片中，抽象凝结出图中物体的类别、空间关系、活动、场景等高阶信息；并根据问题的不同，针对性得给出合理的答案。

视觉问答主要涉及计算机视觉、自然语言处理、知识表达与推理三个领域。作为一个多学科交叉的领域，想实现高准确率的系统表现，既依托单个分支下理论、算法、应用系统的快速发展，作为其基础设施；同时还对各子系统的结合方式提出了很高的要求。正是由于视觉问答任务需要处理语言和图像两种重要的数据类型，这使得智能体更像人类一般思考和推理。智能体的“视觉系统”能够接收含有深层次信息的图像源；智能体的“神经系统”能解析图像信息和理解语言内涵；智能体的“语言系统”能够遣词造句，输出人类可理解的语言形式。因此视觉问答被认为是人类构建“人工智能完全体”的重要一步^[1-3]。

视觉图灵测试^[3]是一种能够衡量智能体系是否在图像语义理解方面达到人类水平的测试方法，视觉问答任务被认为是智能系统通过视觉图灵测试的关键性技术。除了作为视觉图灵测试的核心部分，视觉问答还有其他具有价值的应用场景。
a) 作为盲人或是有视觉障碍问题的病患的辅助系统，他们可以通过自然语言询问，就能获得细粒度的图像或者视频信息，能极大地帮助其获得场景的语义理解，在互联网和现实场景中均能作为一种便利的“视觉补充”。b) 作为一种扩充人机交互的方式，在人机交互上可以实现多种的便利查询。通过对已有图像的询问，获得更深层次的背景知识，例如，对一副未曾见过的艺术名画询问其作者和作画背景，可以更深入的理解图像背后隐藏的人文和历史知识。通过源图像可以搜索具有相似“特征”的图像，例如，向系统查询一张埃菲尔铁塔的夜景图，将能获得更多具有相关特征的图像素材。同样可以通过图像描述查询到对应或者相似的图像。

总的来说，作为一个跨领域的人工智能任务，视觉问答代表着研究者对未来“通用人工智能”的探索，既能够提供一种跨模态的数据处理方式，又能够向机器理解和解决复杂问题、甚至完成推理的人工智能新阶段迈进。

1.2 视觉问答的国内外研究状况

视觉问答任务广阔的应用场景和对人工智能发展的深远意义驱动着研究者不断深化在视觉问答的问题深度、数据集构建、算法演化等方面的研究。

由于视觉问答任务的最终目的是面向真实的人类交互场景，因此 VQA 模型需要解决开放性问题的挑战，因此对问题类型的研究很重要。按照问题的形式划分，视觉问答的问题可以分为二值否问题^[4-6]、多选问题^[1,5]、开放性问题^[1]。按照问题的内容划分，问题分为识别类和推理类。识别类问题包括物体识别、物体检测、属性分类、计数问题、空间关系判定，此类任务在以往的计算机视觉的研究中已经达到了较高的识别准确率，在某些物体识别和物体检测任务上已经能逼近甚至超越人类水平。推理类问题包括场景识别、常识推理和知识库推理等，这类问题形式多变、层次复杂、需要外源知识、甚至需要多步推理，例如：“图片中有什么东西在伞下？”——需要能准确识别物体的空间位置关系、“图片中的交通路口是否可以通行？”——需要基于常识的推理、“图片中的汽车属于什么品牌？”——需要基于外部专业知识库提供隐藏信息。

除了以上提到的两种问题分类，本文提出了一种全新问题分类标准——按照答案与问题和图像的相关性划分。我们认为：对于不同的视觉问答问题，其答案-图像相关度、答案-文本相关度存在差异，即有的答案更依赖于准确的图像分析，而有的答案却对图像不敏感。而这种与输入信息的相关性差异能帮助我们更好的理解模型决策的内部机制。我们提出以 Q、q、I、i 定性的表示“答案-问题强相关”、“答案-问题弱相关”、“答案-图像强相关”、“答案-图像弱相关”，从而组合得出四种问题类型 QI、Qi、qI、qi，如表1-1所示。

表 1-1 根据答案和源信息的相关性划分出四类任务

| | 答案-问题相关性 | 答案-图像相关性 | 问题类型 |
|-----|----------|----------|------|
| 相关性 | 强 | 强 | QI |
| | 强 | 弱 | Qi |
| | 弱 | 强 | qI |
| | 弱 | 弱 | qi |

QI 类型的问题需要同时结合图像特征和文本特征，这是一种最典型的视觉问答类型。Qi 类型的问题答案可以直接根据问题文本得出，这类图像无关的问题可以退化为文本问答任务。qI 类型的答案可以直接从图像中得出，这类问题退化为

图像识别任务。例如，对于图1-1，问题1：“图片中的狗是什么颜色？”，模型可以通过直接识别出图像中狗的色彩属性得出正确答案，那么问题1就是qI类型。对于同样的图片，问题2：“图片上的狗是不是属于动物？”是图像无关的Qi类型。



图 1-1 对于同一张图片，问题的不同将会决定模型是否需要外源知识或是常识。

qi类型的问题答案与问题文本和图像的相关性都比较弱，这类问题包含两类类型，一类为错误或者偏僻的问题，例如偏僻单词的问题、低频的语法结构；另一类为涉及常识和外源知识的问题，回答这类问题需要额外的知识，甚至还需要多步的推理，例如，对于图1-1，问题3：“图片中的动物是什么颜色？”，模型除了需要正确识别图片中的狗和颜色属性，还需要知道“狗属于动物”的常识，才能在“狗的颜色”——黄色和“草地的颜色”——绿色之间做出正确的预测。目前，对于错误或者偏僻的qi问题研究较少。

明确了VQA的问题类型之后，我们将进一步介绍视觉问答模型的国内外发展状况。

视觉问答任务要求系统能同时正确理解问题文本内容和图像内容，一般而言视觉问答系统包含三个主要模块，a) 文本处理模块：从问题文本中提取特征，使得特征中包含足够多的语义信息。b) 图像处理模块：从图像中提取特征，理解图像中的物体信息、场景信息、活动信息、空间构成信息、颜色信息，将像素信息转化为系统可计算的数值量或者标签。c) 答案生成模块：采用某种方式整合文本特征和图像特征，为系统建立一条高泛化能力、高稳健性、高准确率的答案生成通路。

从视觉问答的处理过程可以看出，算法的核心有三个部分组成：如何提取出高层次的图像特征，例如，物体、属性、场景等；如何挖掘问题文本中的语义信息，以求能深入的理解问题内容，确定答案的形式和内容；如何结合图像特征和

文本特征，得出正确或是最佳答案。

受神经网络在计算机视觉和自然语言处理成功应用的影响，从 2014 至今的视觉问答研究多采用了神经网络模型，使用卷积神经网络 CNN 提取图像特征，使用卷积神经网络 RNN 或者长短期记忆 LSTM 处理文本信息，再通过不同的方式“融合”图像特征和文本特征得出答案。图像特征提取的方法一般使用预处理后的卷积神经网络，例如 VGGNet^[7]、ResNet^[8] 和 GoogLeNet^[9]。问题文本的特征提取则借鉴了自然语言处理中的成果，例如词袋模型（CBOW）^[10]、长短期记忆（LSTM）^[11]、门控复发单位（GRU）^[12-14]。系统输出答案的方式有两种（如图1-2），最常见的方式是将任务视为分类问题，根据候选选项的概率大小，确定答案。第二种方式则直接由系统遣词造句合成答案语句，此类方法多出现在有额外知识库的视觉问答系统中，例如 Attributes-LSTM^[15]、ACK^[16]、Ahab^[17]、Facts-VQA^[18]、Multimodal KB^[19]。

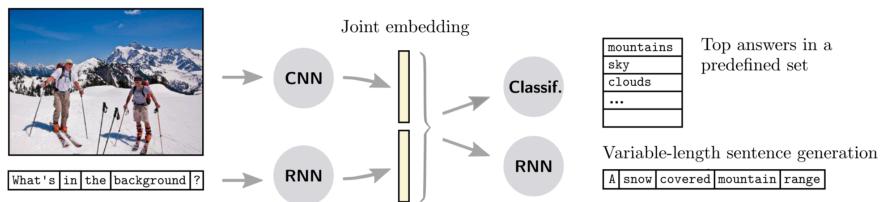


图 1-2 常见的 VQA 方法是将图像和问题文本映射到同一特征空间，再组合融合两者
的形成新的特征向量，特征向量作为分类器或者循环神经网络 RNN（也可能
是长短期记忆 LSTM）的输入，输出得到最终的答案

按照我们提出的基于答案和源信息相关度的划分标准，我们可以将现有的 VQA 模型划分为两个大类：联合嵌入模型和基于知识库的模型。联合嵌入模型是将图像和文本特征融合训练，这类模型对 QI、Qi、qI 三种类型的问题有效的建模，但是无法回答 qI 类型。而基于知识库的模型便是希望引入额外特征从而解决弱问题和图像相关性的问题。其中根据知识库的引入方式的不同，基于知识库的模型又可以分为知识库查询类和知识库嵌入类。

本节将分别介绍三类模型，并比较三种类型的优劣点。

1.2.1 联合嵌入模型

联合嵌入模型先将视觉信息和问题文本信息分别特征化，再通过特征向量串联^[10]、卷积^[20]、逐元素相乘^[1]、逐元素相加^[11] 等池化方法融合图像特征和文本特征，最终得到最优答案。

Malinowski 等人首次提出了应用于真实场景视觉问答任务的联合嵌入模型

Neural-Image-QA^[11]。Neural-Image-QA 是一个由卷积神经网络 CNN 和长短期记忆 LSTM 组成的深度网络，先使用在 ImageNet 预处理过的卷积神经网络 CNN 对图像进行特征提取，得到的特征向量和问题文本一起传输到长短期记忆 LSTM 中，从而生成答案的单词序列。模型在 DAQUAR 数据集上完成训练和测试，对于答案只有一个词语的问题，准确率为 19.43%，对于答案是多个词语的问题，准确率为 17.49%。不同于 Malinowski 的 Neural-Image-QA，Gao 等人认为问题和答案在句法结构上有所不同，因此编码问题的 LSTM 和解码答案的 LSTM 为采用两个独立的网络，使用不同的权重矩阵，结合卷积神经网络 CNN 构成了 mQA 模型^[21]。

Noh 等人认为单单使用相同权重参数的深度卷积神经网络去处理不同的问题，并期待能得到足够准确的答案，这是很困难的^[12]。因此他们提出 DPPnet，在卷积神经网络 CNN 中添加一个动态参数层，动态参数层中的参数会根据问题的不同而改变，这使得每个问题输入都对应一个独特的分类网络。模型由三个部分组成，一个部分作为分类网络的卷积神经网络，第二个部分是参数预测网络，由门控复发单位编码问题序列，再通过一个全连接层输入动态参数，第三个部分是一个哈希函数，将参数预测网络输出的动态参数配置到分类网络中。如图1-3。

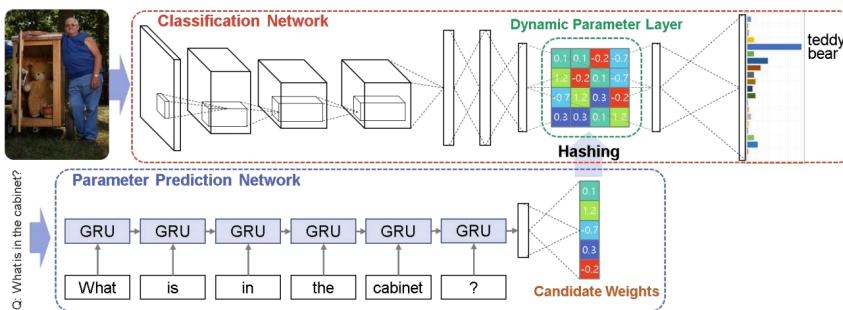


图 1-3 带有动态参数层的卷积网络模型 DPPnet

Zhou 等人同样适用预处理后的卷积神经网络 CNN，但在处理问题文本时选择了比长短期记忆 LSTM 更为简单的词袋模型 BOW，提出了 iBOWIMG 模型^[10]。iBOWIMG 模型受到 BOWIMG^[1] 在 VQA 数据集上优于部分基于长短期记忆 LSTM 模型的启发，在原有基础上将 VGGNet 替换为在图像特征提取表现更优的 GoogLeNet^[9]，将图像特征向量和文本特征向量串联后送入 softmax 层预测问题答案（如图1-4），在 COCO-VQA 数据集上的测试展现出具有竞争力的表现。

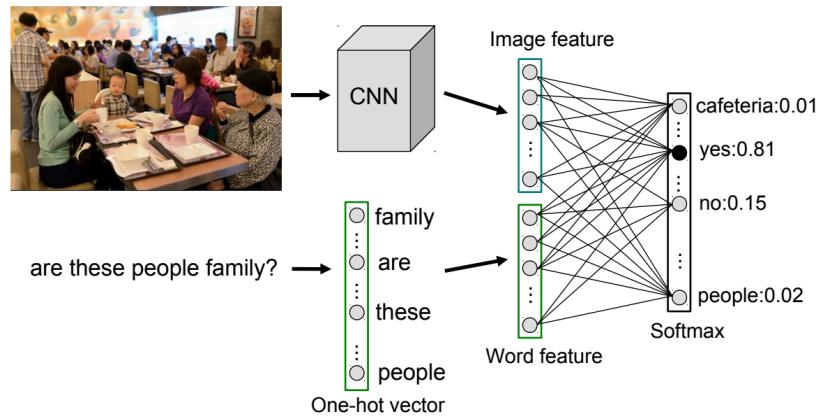


图 1-4 iBOWIMG 使用词袋 BOW 模型作为词特征向量编码器

Lin 等人将卷积神经网络 CNN 不仅应用于编码图像内容，而且也应用于问题文本的提取^[20]。在处理图像特征和文本特征时使用一个多模态的卷积层输出联合特征向量，再使用 softmax 层预测最终的答案。如图1-5。

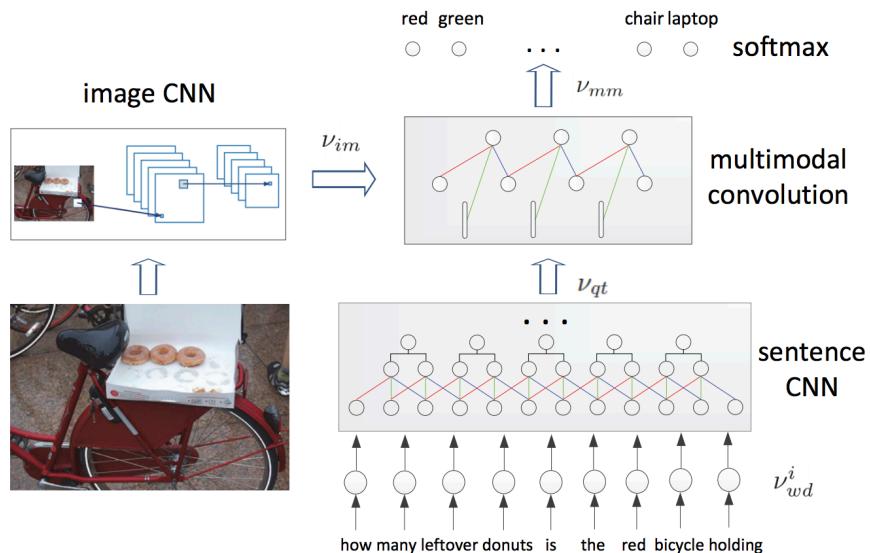


图 1-5 图像特征和问题文本特征提取时均使用 CNN

除了使用不同的方法提取图像和文本特征以外，联合嵌入模型的另一个能够显著改善模型准确率的方向就是实验不用特征向量融合的池化方法。Malinowski 等人通过对不同的特征向量融合方法的比较，可以看出系统的准确率与特征向量融合方法有关，不同方法之间准确率最多能相差 9 个百分点之多^[11]。除了以上提到的 iBOWIMG 采用向量串联的方式,Lin 使用向量卷积的方式外，Antol 等人提出的模型使用逐元素相乘的方法融合两者^[1]，Saito 等人认为不同的特征融合方法各

有特点，会保留或损失不同的特征，为了充分利用不同方法所保留的特征，提出了一种融合逐元素相加和逐元素相乘相结合的模型 DualNet。模型同样利用了使用不同卷积神经网络 CNN 提取的图像特征，例如在真实场景图像采用了 VGG-19^[7]、ResNet-152 和 ResNet-101^[8]。DualNet 对提取出的文本特征和图像特征分别使用逐元素相加和逐元素相乘的方法得到两个不同的联合向量，再将两个的联合向量串联得到最终的合成向量，如图1-6。

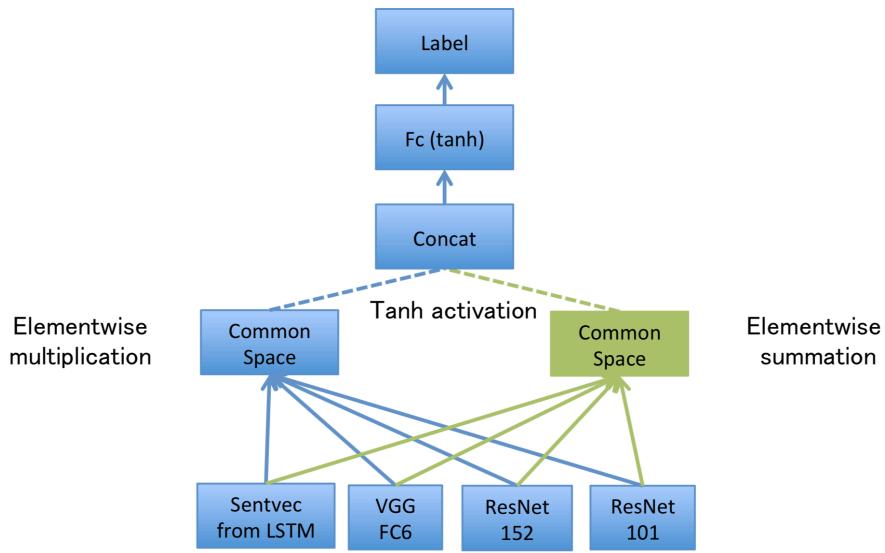


图 1-6 DualNet 针对真实场景图像的模型架构

Fukui 等人认为向量之间的外乘运算中，所有元素之间的互动更加活跃，应该能保留更加丰富的特征信息，因此提出一种更为复杂的多模态紧凑双线性池化方法（MCB）。一般的双线性模型会对两个向量的外乘结果线性化，外乘操作会得到异常高维的向量，例如外乘的两个向量维度均为 2048、输出向量维度为 3000 时，那么训练参数的数量将达到 125 亿个之多，这会导致巨大的计算开销。而提出的多模态紧凑双线性方法能避免直接计算向量外乘，同时保留了大量特征，模型架构如图1-7。

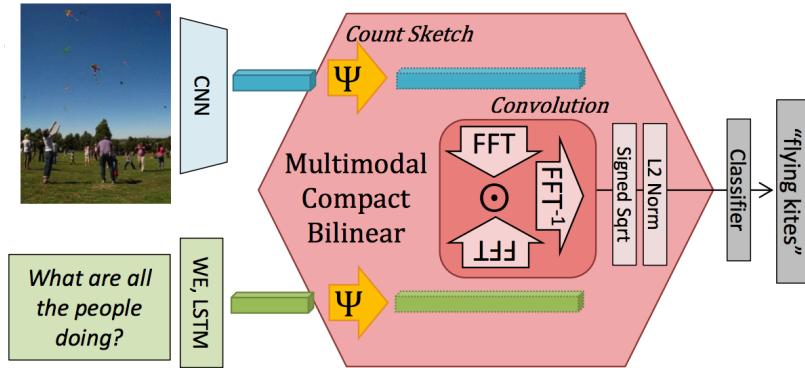


图 1-7 使用多模态紧凑双线性池化融合图像和文本特征

以上介绍的模型从图像特征提取、文本特征提取、特征融合方式上都做出了不同的改进和创新，但这些模型还并未引入注意力机制。由于注意力机制已经在大量的深度学习任务中表现优异，并且，近三年来的视觉问答模型大量引入注意力机制，因此，我们将带有注意力机制的模型特别提出，并加以介绍。

1.2.1.1 注意力机制

人类获取外部视觉信息时，会自动形成一种“像素不均衡”，在同一视野范围内的像素被视觉中枢神经系统根据“关注区域”的远近、相关性特征自动分配不同的分辨率，使得“关注区域”内的像素具有极高的分辨率，而其他的像素仅仅作为视觉信息输入，并不参与大脑的语义处理（如图所示1-8）。因此视觉注意力机制帮助大脑过滤了低相关性的视觉信息，减少了待处理数据的体积，极大地提高了信息处理速率并松弛了大脑负载。



图 1-8 人类视觉系统的“像素不均匀”现象

近几年，受到人类视觉注意力机制的启发，在神经网络中引入注意力机制变得十分热门，在自然语言处理和计算机视觉领域的应用也极大得帮助了原有算法

精度和计算效率的提升。Google Deepmind 团队提出了一种带有注意力机制的循环神经网络 (RNN)，并成功应用于图像分类任务，获得了优于以往卷积神经网络 (CNN) 的基线水平的分类精度^[22]。随后，带有注意力机制的循环神经网络便被广泛应用于自然语言处理和计算机视觉的多个子领域^[23–25]。Bahdanau 等人将注意力机制引入神经机器翻译任务，仍然使用“编码-解码”的翻译模式，但一改以往将源语言文本映射为一个固定长度的向量的编码方式，而是将原语言文本编码为向量序列，解码时将翻译和位置对应因素联合学习，训练向量序列中各向量对翻译词组的不同权重，加和完成翻译结果的推断，得到了以往最优的结果^[23]。Xu 等人受到注意力机制在机器翻译和物体识别任务成功应用的启发，将带有注意力机制的循环神经网络应用于自动生成图像标注，并且在 Flickr9k, Flickr30k 和 MS COCO 三个数据集上均获得了最优的结果^[24]。随后，更多注意力机制的变型或优化研究均在图像标注任务上展开^[26–29]。

相较起图像标注任务，视觉问答任务除了要求系统能理解图片内容，生成语义和句式合理的自然语言文本以外，还需要联合学习问题文本和聚焦与问题相关的图像细节。这些任务特性决定了视觉问答任务可以利用已有较为先进的图像标注任务的框架，同时融合自然语言处理的最新成果。注意力机制在自然语言处理和计算机视觉上的成功应用便成为了视觉问答算法快速发展的基石。

Chen 等人最先将注意力机制引入视觉问答任务，提出了基于注意力机制的可配置卷积神经网络 (ABC-CNN) 用于针对“图像问题对”生成对应的注意力映射，将问题的语义信息和图像区域建立映射，使得答案生成取决于被关注区域，减少无关区域的影响^[30]（模型架构如图1-9）。在 Toronto COCO-QA^[31], DAQUAR^[32], 和 VQA^[1] 三个数据集上的测试结果都提升了最优结果，证明了注意力机制在提高视觉问答任务上的有效性，同时注意力权重图能反应系统的推理过程，为参数的微调提供了依据。

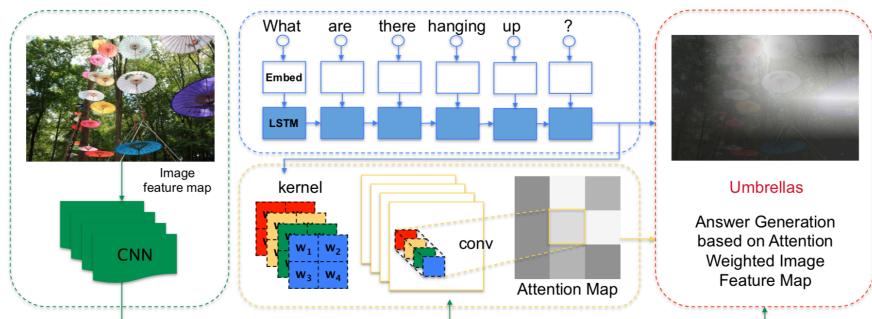


图 1-9 ABC-CNN 使用 CNN 提取图像特征，LSTM 提取问题文本特征，黄色方框内
为用于推测与问题相关的图像区域的注意力机制

Shih 等人使用简单的 word2vec 方法编码“问题-答案”对，使用预处理后的卷积神经网络 CNN 对图片的不同区域编码，将编码后的文本特征向量和图片特征向量映射到同一特征空间，根据特征之间的点乘运算决定每个图像区域的权重，最后结合权重化以后的图像特征和文本特征得出答案。架构如图1-10。在辨别物体颜色的任务上得到了最优结果^[33]。类似的工作还有 Ilievski 等人提出的“聚焦型动态注意力模型”^[34]。

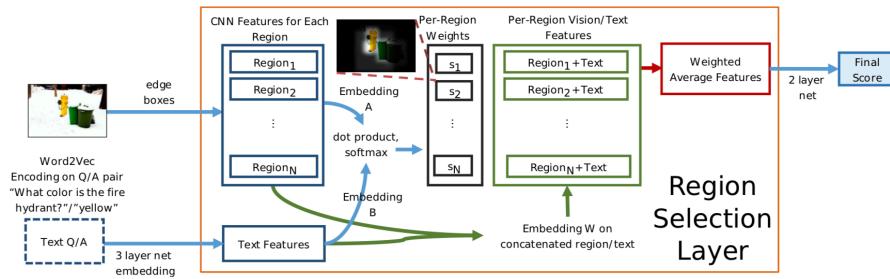


图 1-10 使用图片区域选择层实现注意力机制的架构

包括以上提到的在内，多数注意力机制对问题文本和图像区域特征进行一次运算，直接生成图像注意力权重图。针对这种情况，Yang 等人提出堆栈式注意力网络——使用问题的语义表达对图像进行多次查询，不断缩小答案相关区域，实现更高的精度^[35]。注意力机制在视觉问答上的其他应用还有，同时使用对图像和问题使用注意力机制的联合注意力模型^[36]；不采用图像区域赋值方法，而是过滤掉不相关区域的“自适应硬性注意力网络”^[37]。

对于神经网络训练这类参数密集和计算密集的框架，注意力机制能带来两个重要的改变。一方面，无论对于图像输入还是文本输入，原有的方法都选择将输入看做一个整体，因此映射后的向量需要包含完整的输入信息，对于包含词组过多文本或是场景过于复杂的图像，编码后的向量根本无法区分开输入的局部特征，这使得神经网络的可解释性大大降低。引入注意力机制后，编码方式改变，将输入视为局部信息的综合，保留了文本中单词和图像中像素区域的信息，通过可视化处理，能清晰的看出神经网络的推理过程，增强了系统的可解释性，可以称之为一种“弱化黑盒的处理”。另一方面，注意力机制非常符合人类对于语言和视觉信息的处理方式，这背后的假设是：针对绝大多数任务，只需要从信息源的局部便能获得充分正确的答案。类似于人类，具有注意力机制的智能体应当能获得更高的执行的效率和更高的答案精度。

1.2.1.2 动态记忆网络

无论是在自然语言理解还是图像内容理解，人类在获取单词或者图像像素区域的语义时不会将其与语境割裂来看，通常上下文语境对于准确理解文本和图像信息是非常重要的，因为在语言和图像中存在大量具有歧义特性的内容，例如，在语言中一个单词具有不同的语义，也可能有不同的词性，只有在上下文的语境中才能确定词语的真正含义。记忆力与上下文语境相似，是神经网络在训练过程中存储的“经验”，这种“经验”有助于以后的训练，这种累积经验能创造更准确的答案，基于这样的假设，研究人员为从序列化的输入中获得更准确的输入，而引入了动态记忆网络^[13, 14, 38]。

Jiang 等人在常见的 CNN 解析图像、LSTM 解析问题文本的架构上，新增一个成分记忆模块^[38]，旨在融合每一次训练过程中的局部图像信息和文本信息，并提供给下一次训练使用，从而使网络存储了训练过程的“经验”，这与之后提出的动态记忆网络有同样的思想，模型训练流程如图1-11。

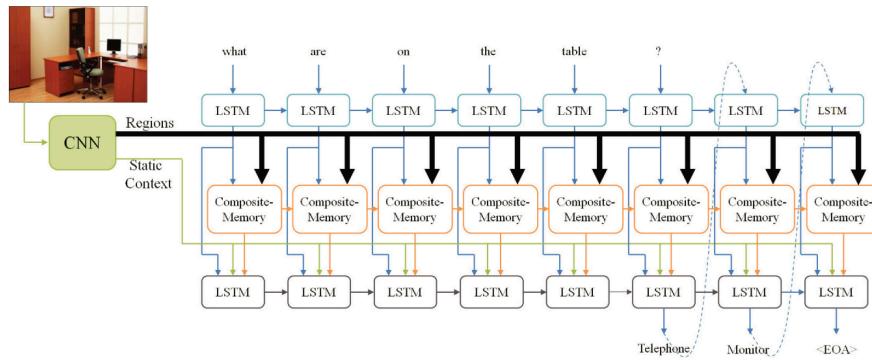


图 1-11 成分记忆模型的训练流程

Kumar 等人为解决文本问答（Text-QA）任务而提出动态记忆网络（DMN）^[13]。动态记忆网络（DMN）是一个用于生成文本问题答案的神经网络框架，它由输入模块、问题模块、情节记忆模块和问题模块构成，输入模块用于编码文本输入；问题模块用于编码文本问题；情节记忆模块接受由输入和问题模块得到的分布式向量，再使用注意力机制选择部分接受到的向量，结合选择后的向量与以往存储的“记忆”生成新的“记忆”向量，并不断迭代；答案模块根据最终的记忆向量生成答案，模型架构如图1-12。

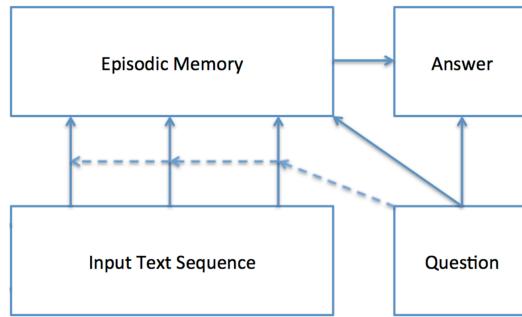
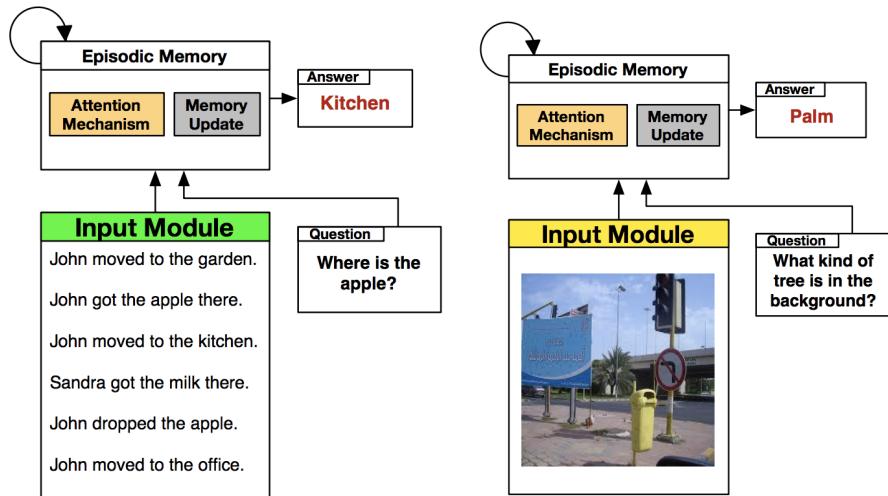


图 1-12 DMN 基础架构

动态记忆网络（DMN）在文本问答、语义分析、词性标注任务上取得了最优的结果，受到其在处理序列化的文本信息上的优异表现的启发，Xiong 等人在原有网络的基础上改善了输入和记忆模块，除了能处理文本信息外，还能处理图像信息，提出应用到视觉问答任务动态记忆网络+（DMN+）^[14]，如图1-13。动态记忆网络+（DMN+）将原有的输入模块中处理文本编码的门控复发单元（GRU）更换为双向门控复发单元（bi-GRU）以得到文本或图像区域更完整的上下文信息；使用基于注意力机制的门控复发单元替换原有的软性注意力机制。更新后的动态记忆网络+（DMN+）在 DAQUAR^[32] 和 VQA 数据集^[1] 上的测试结果都得到了具有竞争力的表现。



(a) 应用于文本问答的动态记忆网络 (b) 应用于视觉问答的动态记忆网络 +
(DMN) 模型架构 (DMN+) 模型架构

图 1-13 DMN+ 与 DMN 架构对比

1.2.2 基于外源知识库的视觉问答模型

视觉问答任务基于图像场景回答问题，图像理解、问题理解和答案生成是实现准确的视觉问答系统的算法核心。图像理解、问题理解和答案生成三者又可以根据人类思考逻辑将其划分为两个逻辑层次，问题理解成为逻辑基点，图像理解和答案生成都根据问题的不同而采用适当的算法策略——注意力机制便是一种借助问题理解而实现计算效率更高的图像解析方法，答案生成中关心的答案类型和答案词组长度也需要依照问题的不同而选择。因此问题的解析过程对于视觉问答算法的准确性和计算成本都有很大的影响。

正如上文提及的，问题可以分为识别和推理两个大类，推理任务中既要求系统能准确识别图像中的对象，往往也会涉及图像中无法获取的先验知识。先验知识包括众所周知但不会显性呈现的常识和面对特定领域需要具备的专业知识，例如，判断路口是否可以通行时，涉及基本交通规则的常识，判断艺术品的作者这类专业问题时，需要借助与该艺术品相关的知识储备。

先验知识对视觉问答系统提出了更高的要求，这也揭露了主流的联合嵌入模型的缺陷：

第一，数据集依赖。联合嵌入模型的答案生成来源于训练集中的问题和答案文本，这意味着训练集中包含的知识和文本内容是整个视觉问答系统的所有知识来源，因此对于测试集中涉及的全新概念或答案，系统根本无法得出正确的答案。不断扩充包含更多先验知识的训练集是提高精度的方式之一，但对于整个世界蕴含的不可计量的知识而言，这种数据集扩充的方式成本巨大。

第二，网络容量小。联合嵌入模型要求网络本身能存储学习到的知识，目前网络的容量相较于需要学习的知识是严重不足的。

第三，黑盒效应明显。对于识别和分类等问题而言，可解释性与高精确度相比，显得不那么重要，但是对于需要明确推理过程的问答系统而言，黑盒的不可解释性会降低提问者对系统的可信度。

对于以上三个联合嵌入模型的缺陷，一种可行的解决方案是将推理过程和知识学习分离，引入外源知识库。可扩展的外源知识库可以解决网络容量的限制问题；知识库中结构化的数据能为推理提供路径，提高系统的可解释性。因此基于知识库的视觉问答模型是并行于联合嵌入模型的另一个重要研究方向。

在基于知识库的视觉问答模型中，知识库的使用方式分为两类。一类为知识库查询类，依照查询知识库查询获得答案的思路，模型提取图片的实体、将实体映射到知识库、转化自然语言为查询语句、查询知识库。代表模型为 Ahab^[17] 和 FVQA 模型^[18]。这些模型依靠精准的查询语句，对于预先设定好的模板问题能实

现优于基线模型的准确率，然而却面临着问题模板设计成本高、数据集难构建、模型泛化能力差等缺点。

另一类为知识库嵌入类，这种方式不用设计复杂的查询语句，而是将知识库的数据转化为额外的特征向量，并联合图像特征和问题特征一起训练。这种方式能省去问题模板和查询语句设计的人工成本，并将模型在更大规模的开放性数据集进行训练。代表为基于知识库的通用嵌入模型^[16]。

本节将简单介绍以上两类基于知识库的模型，并指出其优缺点。

1.2.2.1 知识库查询类

Ahab Wang 等人提出的 Ahab 视觉问答系统利用 DBpedia 作为知识库，实现对需要先验知识的问题的推理应答，即使问题中涉及不包含于图像中的概念^[17]。Ahab 的主要思路为三步，第一步，将图像中的概念链接到知识库中相同的概念，形成从图像到知识库的映射，第二步，将自然语言的文本问题处理为知识库查询语句，实现从自然语言的句法和语义结构变换到相应的查询语句结构，第三步，将知识库的查询结果转换为自然语言表达。利用以上三步，Ahab 可以不通过数据集训练获取知识，而使用自然语言到知识库的两次转化完成问答任务。

具体来说，为了建立图像概念到知识库实体之间的映射，首先检测图像包含的概念，再将提取出的图像概念和知识库实体建立链接。Ahab 分别使用预训练的 Fast R-CNN^[39] 和两个不同的 VGGnet^[7] 从图像中提取物体对象、图像场景和图像属性三种视觉概念。所有提取出的图像信息都使用资源描述框架（RDF）的形式表示，例如，“图像中包含长颈鹿对象”被表示为（图像，包含，对象 1），（对象 1，名称，长颈鹿）。每个视觉概念则被直接链接到具有相同语义的知识库概念，如图所示1-14。

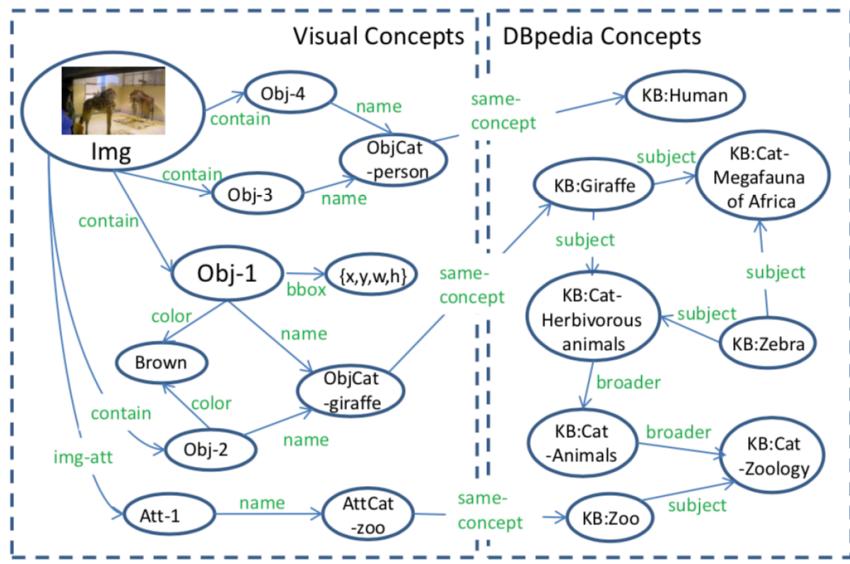


图 1-14 Ahab 中链接图像信息和知识库实体的 RDF 图结构

在问题文本处理方面，Wang 等基于自建的 KB-VQA 数据集——其中的问题需要常识或外源知识，设定了 23 种问题模板，将自然语言问题转化为相应的知识库查询语句，直接从知识库中查询得到答案。Ahab 在 KB-VQA 数据集上的表现如图 1-15。

| Question Type | Accuracy(%) | | | Correctness (Avg.) | | |
|-----------------------|-------------|------|-------|--------------------|------|-------|
| | LSTM | Ours | Human | LSTM | Ours | Human |
| <i>IsThereAny</i> | 64.9 | 86.9 | 93.6 | 3.6 | 4.5 | 4.7 |
| <i>IsImgRelate</i> | 57.0 | 82.2 | 97.1 | 3.3 | 4.2 | 4.9 |
| <i>WhatIs</i> | 26.9 | 66.9 | 94.5 | 2.1 | 3.7 | 4.8 |
| <i>ImgScene</i> | 30.4 | 69.6 | 85.9 | 2.3 | 3.8 | 4.5 |
| <i>ColorOf</i> | 14.6 | 29.8 | 93.2 | 1.7 | 2.5 | 4.7 |
| <i>HowMany</i> | 32.5 | 56.1 | 90.4 | 2.3 | 3.3 | 4.6 |
| <i>ObjAction</i> | 19.7 | 57.1 | 90.5 | 1.8 | 3.5 | 4.7 |
| <i>IsSameThing</i> | 54.9 | 77.5 | 91.5 | 3.2 | 4.2 | 4.6 |
| <i>MostRelObj</i> | 32.1 | 80.4 | 92.9 | 2.3 | 4.2 | 4.6 |
| <i>ListObj</i> | 1.9 | 63.0 | 100 | 1.1 | 3.6 | 4.8 |
| <i>IsTheA</i> | 74.5 | 80.4 | 92.2 | 3.9 | 4.2 | 4.7 |
| <i>SportEquip</i> | 2.1 | 70.8 | 79.2 | 1.2 | 3.9 | 4.2 |
| <i>AnimalClass</i> | 0.0 | 87.0 | 95.7 | 1.0 | 4.5 | 4.8 |
| <i>LocIntro</i> | 2.5 | 67.5 | 95.0 | 1.1 | 3.6 | 4.8 |
| <i>YearIntro</i> | 0.0 | 46.9 | 93.8 | 1.0 | 2.9 | 4.8 |
| <i>FoodIngredient</i> | 0.0 | 58.1 | 74.2 | 1.0 | 3.4 | 4.3 |
| <i>LargestObj</i> | 0.0 | 66.7 | 96.3 | 1.0 | 3.8 | 4.8 |
| <i>AreAllThe</i> | 29.6 | 63.0 | 81.5 | 2.3 | 3.7 | 4.3 |
| <i>CommProp</i> | 0.0 | 76.9 | 76.9 | 1.0 | 4.1 | 4.2 |
| <i>AnimalRelative</i> | 0.0 | 88.2 | 76.5 | 1.1 | 4.4 | 4.1 |
| <i>AnimalSame</i> | 41.2 | 70.6 | 94.1 | 2.6 | 3.8 | 4.8 |
| <i>FirstIntro</i> | 25.0 | 25.0 | 75.0 | 2.0 | 1.5 | 4.1 |
| <i>ListSame Year</i> | 25.0 | 75.0 | 50.0 | 1.8 | 4.2 | 3.0 |
| <i>Overall</i> | 36.2 | 69.6 | 92.0 | 2.5 | 3.8 | 4.7 |

图 1-15 Ahab、联合嵌入模型和人类作答在 23 种问题上的表现。Accuracy 是得分超过 3 的问题数量的比例，Correctness 是某类问题得分的加权平均数。

从图2-8中可以看出，Ahab 在每种问题类型上都优于联合嵌入模型，但离人类的正确率还是有一定差距，尤其在“判断物体颜色”和“比较两个物品的诞生先后”两种问题。但是由于 KB-VQA 在不同问题类型上数量的不均衡和问题样本数过小的缺陷，Ahab 在真实场景中对于推理问题的解决上仍然有待检验。

Ahab 模型最大的缺点是，模型的准确率高度依赖问题模板和查询语句的人为设计。人为设计必将带来的高额成本，并且需要使用特定的数据集进行训练。当问题类型数量剧增时，人工的对每种类型设定对应的算法是不切实际的，因此 Ahab 的扩展性面临挑战。

但相较于主流使用统计方法的联合嵌入模型，Ahab 利用知识库取代知识学习过程的方法在复杂推理任务，尤其是需要运用先验知识的问题上，表现更好。

FVQA Ahab 将问题解析为知识库查询语句时，需要预先确定问题模板，这极大的限制了系统面对多样化问题的能力，因此 Wang 等人改变了问题到查询语句的映射方式提出了 FVQA 模型^[18]。FVQA 模型使用长短期记忆（LSTM）网络训练一个 28 类的查询语句分类器，实现将问题到查询语句的分类过程。通过 LSTM 的分类，文本问题被映射为（REL, AS, VC）的查询类型，其中 VC 表示视觉概念、

REL 表示谓语、AS 表示知识来源。对于所有 28 种查询类型，查询语句都由下面的形式构成：

```
Find ?X, ?Y, subject to
{ (ImgID, Contain, ?X) and (?X, VC-Type, VC) and (?X, REL, ?Y) }
```

其中 ImgID 表示图片的标号，?X 表示在图片 ImgID 中类型为 VC 的视觉概念，?Y 表示在知识库中与?X 通过谓语 REL 链接的概念。再根据 AS 是图片还是知识库的类型，使用不同的方法得到最终的答案。

引入支持向量机（SVM）^[40] 和使用长短期记忆（LSTM）的联合嵌入模型^[15] 作为基线模型：只提供问题的 SVM-Question 和 LSTM-Question、只提供图片的 SVM-Image 和 LSTM-Image 以及同时提供问题文本和图片的 SVM-Question+Image 和 LSTM-Question+Image，各模型使用自建的 FVQA 数据集作为训练和测试集中测试，不同模型的正确率见表1-2。

表 1-2 不同模型在 FVQA 数据集上的测试正确率，Top-1 表示只取得分最高的预测结果，Top-3 和 Top-10 以此类推。灰色数据表示使用与问题对应的完全正确的查询类型时的正确率。

| Method | Overall Acc. (%) | | |
|---------------------|------------------|--------------|--------------|
| | Top-1 | Top-3 | Top-10 |
| SVM-Qusetion | 11.19 | 20.68 | 32.14 |
| SVM-Image | 17.55 | 30.75 | 49.02 |
| SVM-Qusetion+Image | 17.99 | 31.83 | 49.55 |
| LSTM-Question | 10.30 | 18.26 | 31.02 |
| LSTM-Image | 22.69 | 36.21 | 58.59 |
| LSTM-Question+Image | 23.37 | 37.02 | 52.51 |
| gt-QQmaping | 64.23 | 71.58 | 72.74 |
| top-1-gt-QQmaping | 53.63 | 60.70 | 61.59 |
| top-3-gt-QQmaping | 58.19 | 65.89 | 66.83 |

从表1-2中 Top-1 一列可以看出，无论是 SVM-Question+Image 与 SVM-Image 之间的正确率差距还是 LSTM-Question+Image 与 LSTM-Image 的正确率差值都非常小，这说明问题的解析对于 SVM 和 LSTM 这两种模型正确率的提升没有太大的帮助，而两个模型总体的正确率也处于较低的水平，说明统计方法在样本较小

的语料库中很难学习到知识间真正的逻辑关联。而 FVQA 模型使用问题到查询映射模型能从问题文本中提取到关键信息，并能利用关键信息组成有意义的语言结构，再结合额外知识库搜索到正确答案，答案获得的过程反映了推理的过程。gt-QQmaping（灰色背景）使用问题对应的正确查询类型，因此正确率反映了理想状况下 FVQA 模型从查询类型到生成查询语句过程中的误差情况，知识库查询过程的错误率在 30% 左右。top-1-gt-QQmaping 与 gt-QQmaping 之间的差距则代表问题到查询类型的正确率在最终答案的影响，top-3-gt-QQmaping 的准确率高于 top-1-gt-QQmaping 的原因在是因为前者拥有更高的问题到查询类型映射的准确率。

表1-3提供了不同方法在不同答案来源上的正确率，对比表中 Image 和 KB 两列容易看出，答案来源于视觉概念的准确率在所有模型上均远高于知识库来源，这说明表中涉及的三种模型都只能从图像和问题文本中包含的概念中提取答案，一旦答案涉及都额外知识库中的“新”概念，准确率便急剧下降，即使是使用额外知识库的 gt-QQmaping。

表 1-3 不同方法在不同答案来源上的正确率

| Method | Answer-Source | | | | | |
|---------------------|---------------|--------------|--------------|-------------|--------------|--------------|
| | Image | | | KB | | |
| | Top-1 | Top-3 | Top-10 | Top-1 | Top-3 | Top-10 |
| SVM-Qusetion | 12.80 | 24.53 | 36.48 | 0.68 | 2.03 | 3.72 |
| SVM-Image | 19.92 | 34.88 | 55.11 | 2.03 | 3.72 | 9.12 |
| SVM-Qusetion+Image | 20.43 | 36.07 | 55.73 | 2.03 | 4.05 | 9.12 |
| LSTM-Question | 11.71 | 20.49 | 34.21 | 1.01 | 3.72 | 10.14 |
| LSTM-Image | 25.49 | 40.40 | 65.12 | 4.39 | 8.78 | 15.88 |
| LSTM-Question+Image | 26.01 | 41.12 | 58.05 | 6.08 | 10.14 | 16.22 |
| gt-QQmaping | 72.65 | 80.13 | 80.13 | 9.12 | 15.54 | 24.32 |
| top-1-gt-QQmaping | 60.89 | 68.27 | 68.27 | 6.08 | 11.15 | 17.91 |
| top-3-gt-QQmaping | 66.10 | 74.15 | 74.15 | 6.42 | 11.82 | 18.92 |

FVQA 模型提出了一种以句法结构中的谓语为核心的先验知识问题的解答思路，首先从问题中解析出关键的谓语信息，在问题到查询类型模型中，结合谓语、视觉概念和答案来源决定了 28 种不同的查询类型，再使用生成的查询语句搜索基

于 12 种谓语构建的知识库，最终预测答案。“主语-谓语-宾语”的一般句式结构中谓语表示了主语和宾语之间的相互作用，即使在相同的主语和宾语情况下，不同的谓语能表达出截然不同的语义信息，而绝大多数问题也能够直接通过谓语，推断答案的范畴。以谓语为基础的优势有几点，第一，易于问题分类。问题的自然语言表达方式众多，但无论如何改变句式结构，表达相同含义的谓语有限，通过对谓语的语义划分能够划分出问题的不同类型。第二，便于知识库的查询。知识库中的实体之间通过不同的谓语连接，形成错综复杂的知识网络，一个实体有众多连接，但一个谓语只连接两个实体，且往往谓语的两端就是问题的答案。

FVQA 模型的缺陷有三点，第一，分类数量的确定和分类模型的精度。FVQA 模型由于查询语句的生成依赖于查询类型，因此问题到查询类型映射的准确性会直接影响到答案生成的正确率。表1-2是在 FVQA 数据集上进行的，由于数据集中问题的类型只有 28 种，因此 FVQA 模型在问题到查询类型映射模型中使用了 28 类的分类器，但在实际问答环境中问题类型的具体数量远远多于 28 种，且无法预先确定。想训练能应用于实际情景中，FVQA 模型不仅需要数据集的扩充，还需要提高模型本身的分类精度。第二，不能回答以谓语为答案的问题。所有 28 种查询类型都要求能从问题中提取关键谓语，并且所有答案都是物体对象，如果问题询问对象之间的关系，模型则无法从问题文本中获得谓语，不能得到答案。第三，不能很好的处理含有多个动词的复杂推理问题。FVQA 模型的查询语句过于简单，仅仅将一次查询结果作为答案，在面对需要多级推理的问题时，便无法直接得到答案。

1.2.2.2 知识库嵌入类

知识库查询类模型通过将问题转换为特定的知识库查询语句，限制了问题类型。为了提高视觉问答系统的问题的灵活性，Wu 等人又通过改进常见的 CNN+LSTM 的嵌入模型，提出了基于知识库的通用嵌入模型^[16]。模型的基本架构由图像属性提取网络（CNN）、图像描述生成网络、外部知识库查询网络以及答案生成网络（LSTM）构成，模型架构如图1-16。

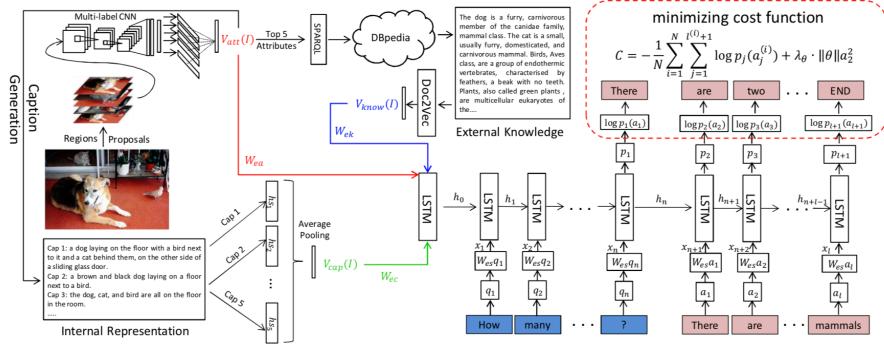


图 1-16 结合外部知识库的通用嵌入模型

图像属性提取网络将图像属性提取问题视为多标签的分类问题，以图像的多个子区域作为输入，输出前五个从 MS COCO 中筛选得到的图像属性 $V_{att}(I)$ ，属性可能为物体名称、动作或者描述特征的形容词。提取出的图像属性分别作为图像描述生成网络和外部知识库查询网络的输入，图像描述生成网络将 [15] 中的高层次的属性表达输入 LSTM 网络生成基于图像属性的描述，再将文本描述转化特征向量 $V_{cap}(I)$ 。外部知识库查询网络首先分别将五个图像属性转化为知识库查询语句，查询到 DBpedia 知识库中相应的对象后，返回其“comment”——“comment”往往包含关于知识库对象最重要的解释信息，如图1-17。Wu 等人使用 Doc2Vec^[41]将其转化为 $V_{know}(I)$ 。最后将 $V_{att}(I)$ 、 $V_{cap}(I)$ 、 $V_{know}(I)$ 以及问题文本作为答案生成网络（LSTM）的输入，训练网络生成答案。



图 1-17 使用‘dog’属性的查询语句以及返回的‘comment’内容

在评估模型准确率时，由于该模型面对开放性问题，因此不同于 Ahab 和 FVQA 只能使用专门设计的数据集，该模型采用 Toronto COCO-QA^[31] 和 VQA^[1] 两个数据集进行评测，分别获得了 69.73% 和 55.96% 的准确率。

相对于知识库查询类的模型，知识库嵌入类的模型无需设计查询语句和构建数据集，因此可以使用通用数据集训练和评价；相对于不引入知识库的联合嵌入模型，基于知识库的模型能提供图像和问题以外的信息，并且具有很高的知识存储容量。

知识库嵌入类的模型的缺点是，由于答案的生成依然依赖于训练集，因此在面对复杂的推理问题时，表现差于知识库查询类的模型。

总结上文提到的所有视觉问答模型，可以看出联合嵌入模型是主流，注意力机制和动态记忆网络的出现也都是为了弥补原有联合嵌入模型的计算效率不高、记忆短缺等问题。虽然引入外源知识库能更为彻底的解决神经网络知识存储不足的问题，实现特征提取和答案推理的分离，但是其下属的两个类型：知识库查询类和知识库嵌入类在多个方面仍然存在优劣。三种模型的比较可以简单描述为以下关系：

解决识别类问题： 知识库嵌入类 > 联合嵌入模型 > 知识库查询类

解决推理类问题： 知识库查询类 > 知识库嵌入类 > 联合嵌入模型

模型迁移能力： 知识库嵌入类 = 知识库查询类 > 联合嵌入模型

联合嵌入模型由于其模块组合的灵活性，因此具有很高的改进空间。而知识库潜入类的模型由于引入了知识库，能引入额外的特征，因此提高了其对推理问题的解决能力，有很好的研究前景，并且能够使用通用数据集训练，可实现端对端的训练。而知识库查询类的模型基于人为设计的查询模板，在推理问题上能实现更好的精度，但是查询模板和数据集的构建成本过高，限制了其进一步的发展。

1.3 本文的主要贡献与创新

虽然视觉问答任务具有广阔的应用前景和有价值的研究意义，但其仍然属于起步阶段，面临着诸多亟待解决的问题。我们认为现有的模型还存在以下三个主要问题。

第一，泛化能力受限于数据集。目前的视觉问答主流模型为联合嵌入模型，其主要思路为联合图像处理模型输出的图像特征和自然语言处理输出的文本特征，并利用神经网络训练联合后的特征，得到答案。和其他基于神经网络的模型一样，联合嵌入模型的泛化性能主要由训练集的大小、内容的多样性等因素影响。然而数据集的收集和整理工作需要消耗大量的的人工成本，因此“数据集偏见”成为目前模型泛化性能的主要瓶颈之一。

第二，结果的可解释性匮乏。目前的视觉问答模型大多仍然属于分类模型，候选答案来源于训练集，模型对候选答案评分，并将得分最高的候选答案作为输出。而分类的标准存在于模型中的大量参数之中，模型得出答案的过程和标准并不明确。匮乏的可解释性在需要多步推理的问题上表现尤为突出。该类问题不同于识别任务，答案的得出是分步进行的，每一步的正确推理都对答案的得出至关重要。而通过分类模型得出的答案无法给出每一步的推理过程。

第三，缺少通用架构。正如以上提到的，联合嵌入模型和基于知识库的视觉问答模型是目前研究的重点方向。基于知识库的视觉问答模型的提出是为了解决训练集的有限性和答案的黑盒性。外源知识库能够扩展模型可搜寻的答案范围，结构化且语义明确的实体之间的关系也可以提供答案的可解释性。然而，目前基于知识库的视觉问答模型根据各自的特点建立独特的数据集，该类自建的数据集从数据量和多样性的角度都不如通用的数据集，例如 VQA2.0^[42]。因此由于使用的数据集不同，目前两个主要的研究方向的模型之间很难建立统一的标准以衡量性能之间的差异性，造成了联合嵌入模型和基于知识库的 VQA 模型的割裂。

为改善第一个问题中提到的数据集限制，我们依照答案和源信息相关性，研究了主要的视觉问答数据集，并且从中选择数据集构成实验数据集。实验数据集需要包含 QI、Qi、qI、qi 全部四种问题类型，从而扩展问题类型的多样性和数据量。

如上文提到的，目前的视觉问答模型可以划分为联合嵌入模型、知识库查询类模型、知识库嵌入类模型，本文重点研究联合嵌入模型和知识库嵌入类模型。针对两类模型，我们分别进行了改进，从而提出了两个全新的模型。具体来说，本文的主要贡献和创新点如下：

1. 对于视觉问答问题，我们根据答案分别与问题和图像之间的依赖性的不同，提出了一个视觉问答类型的划分标准，划分出 QI、Qi、qI、qi 四种问题类型。并使用该标准解释了联合嵌入模型与基于知识库的模型的差异。
2. 针对现有联合嵌入模型仍使用静态词向量的局限，我们提出了一个基于动态词向量的联合嵌入模型——None KB-Specific Network (N-KBSN) 模型。N-KBSN 模型能根据上下文语境动态的计算词向量，从而能有效解析词语的多义性和多成分特性，提高模型的准确性。
3. 区别于其他基于知识库的模型，我们提出将知识库转化为图嵌入——使用低维特征向量表示与问题相关的知识子图，而不是使用查询语言获取知识库中的子节点^[17, 18]。在 N-KBSN 模型的基础上，我们引入知识库图嵌入模块，构建了基于知识库图嵌入的 VQA 模型——KB-Specific Network (KBSN)。知识库的图嵌入由子图提取模块和子图嵌入模块两个主要部分组成。知识库的图嵌入能表达实体之间的结构信息，从而增强特征的表达能力，并且低维的特征向量具有计算便利性，可以实现大规模的训练和预测，消除了人工设计查询语言的复杂性。

1.4 本论文的结构安排

本文的章节结构安排如下：

第一章，绪论。本章节主要介绍了视觉问答任务的研究内容和应用前景，提出了一种新的问题类型的划分标准，还对视觉问答的国内外研究状况作了比较完整的归纳，其中重点介绍了已有的联合嵌入模型和基于知识库的模型，最后阐述和总结本文的研究内容。

第二章，视觉问答数据集。本章对已有的主要的视觉问答数据集进行了概括性的介绍，并且根据数据集的问题是否需要外源知识为标准，分为基于视觉的数据集和基于知识的数据集，最后介绍了本文使用的实验数据集及其选取策略。

第三章，基于动态词向量的联合嵌入模型。本章首先分析了联合嵌入模型在视觉问答任务优异表现的原因，并且分析了现有模型的特点和局限。针对现有模型的局限，我们提出了基于动态词向量的联合嵌入模型——N-KBSN 模型，随后详细介绍了 N-KBSN 模型的问题文本和图像特征提取模块、自注意力和引导注意力模块。最后使用 VQA2.0 数据集^[42] 训练和测试模型，并且通过剔除实验分析了各个模块在模型中的作用，通过和其他最优模型的比较证明了其有效性。

第四章，基于知识库图嵌入的视觉问答模型。KB-Specific Network(KBSN)。本章简要介绍了知识库的发展历史，并且分析了几个重要的知识库各自的特点。随后提出了知识库的图嵌入模块，并在上一章提出的 N-KBSN 模型的基础上，提出了一个基于知识库图嵌入的视觉问答模型——KBSN 模型，最后使用 KB-VQA^[17] 和 FVQA^[18] 数据集训练和测试模型，通过对比其他模型，分析实验结果，证明了 KBSN 模型在回答常识型问题和知识型问题的优越性。

第五章，全文总结与展望。主要总结全文研究内容及结论，并阐述后续工作的主要研究方向。

第二章 视觉问答数据集

从 LeCun 的 MNIST 数据集^[43] 到如今大量的人工智能任务的数据集，优质的数据集已经成为研究工作的重要部分，是监督学习模型的训练基础。优秀的数据集需要具有足够大的容量^[44]、规范友好数据格式、较小的数据偏见等特点。

视觉问答任务是在经历了计算机视觉和自然语言处理任务成功之后，新兴出现的人工智能任务——要求系统能同时理解多模信息，并完成信息整合与推理。自从 2014 年以来，多个高质量的视觉问答数据集被提出：DAQUAR^[32]、COCO-QA^[31]、VQA^[1]、VQA 2.0^[42]、CLEVR^[45]、KB-VQA^[17]、FVQA^[18]。

以上数据集有不同的图像来源，各自的问题对的数量也不同，但是其中最重要的区别在于问题回答是否需要额外知识，例如常识和专业知识。我们将不需要额外知识的数据集称为“基于视觉的数据集”——问题的答案往往来源于图像信息的准确提取，而需要额外知识的数据集称为“基于知识的数据集”——图像信息仅仅作为推理的一环，答案依赖于图像和问题以外的知识。根据以上的划分标准，以上提到的数据集的统计信息如表2-1。

表 2-1 视觉问答数据集的对比，其中 KB-VQA 和 FVQA 属于“基于知识的数据集”，其他均属于“基于视觉的数据集”。

| Dataset | #images | #QA pairs | Image source | Knowledge based |
|-------------------------|---------|-----------|------------------|-----------------|
| DAQUAR ^[32] | 1,449 | 12,468 | NYU-Depth | |
| COCO-QA ^[31] | 69,172 | 117,684 | COCO | |
| VQA ^[1] | 204,721 | 614,163 | COCO | |
| VQA 2.0 ^[42] | 204,721 | 1,105,904 | COCO | |
| CLEVR ^[45] | 100,000 | 999,968 | Synthetic images | |
| KB-VQA ^[17] | 700 | 2402 | COCO+ImgNet | ✓ |
| FVQA ^[18] | 1,906 | 4,608 | COCO | ✓ |

正如表格所示，KB-VQA 和 FVQA 属于“基于知识的数据集”，其他均属于“基于视觉的数据集”。本章我们将从这两个方面概述视觉问答数据集，并分析各自的特点，最后从中挑选出本文使用的实验数据集。

2.1 基于视觉的数据集

DAQUAR DAQUAR 从 NYU-Depth V2 中带有语义分割标注的图片基础上扩展而来^[32]。数据集包含 1449 张图片，图片多为室内场景，这大大地限制了数据集的场景丰富性，是该数据集的一大劣势。数据集由训练集和测试集两部分组成，训练集中包含 6794 对“问题-答案”，测试集中包含 5674 对“问题-答案”，“问题-答案”对由算法生成或是人类志愿者提供，算法生成的“问题-答案”对根据给定的模板生成，详见图2-1。

| | Description | Template | Example |
|-------------------|---------------------|--|--|
| Individual set | counting | How many {object} are in {image_id}? | How many cabinets are in image1? |
| | counting and colors | How many {color} {object} are in {image_id}? | How many gray cabinets are in image1? |
| | room type | Which type of the room depicted in {image_id}? | Which type of the room is depicted in image1? |
| | superlatives | What is the largest {object} in {image_id}? | What is the largest object in image1? |
| | counting and colors | How many {color} {object}? | How many black bags? |
| negations type 1 | negations type 1 | Which images do not have {object}? | Which images do not have sofa? |
| | negations type 2 | Which images are not {room_type}? | Which images are not bedroom? |
| | negations type 3 | Which images have {object} but do not have a {object}? | Which images have desk but do not have a lamp? |

图 2-1 DAQUAR 问题模板

DAQUAR 数据集较小并且问题的类型只有三种：物体识别、色彩识别、计数，并且答案类型多以单词为主，因训练和测试系统复杂问题的推理能力较弱，偏于传统的物体识别任务。

COCO-QA COCO-QA 包含来自 MS COCO 的 123287 张真实场景图片，问题-答案对则是运用算法从 MS COCO 数据集的图像标注中生成的，为了方便生成算法的运用，将问题划分在物体识别、色彩识别、计数、地点查询四种类型。DAQUAR 数据集在实际测试过程中，被发现仅仅通过简单的猜测答案的方式都能获得较高的正确率，这使得高准确率出现了极大的偏差，不能公正的测试系统的“推理”能力。为了克服该缺点，COCO-QA 去除了出现频数极低和极高的一些答案，使得常见答案出现的评率从 24.98% 下降到 7.30%。COCO-QA 的训练集包含 78736 对“问题-答案”，测试集包含 38948 对“问题-答案”，在四个类别中的分布如图2-2。

| CATEGORY | TRAIN | % | TEST | % |
|----------|-------|---------|-------|---------|
| OBJECT | 54992 | 69.84% | 27206 | 69.85% |
| NUMBER | 5885 | 7.47% | 2755 | 7.07% |
| COLOR | 13059 | 16.59% | 6509 | 16.71% |
| LOCATION | 4800 | 6.10% | 2478 | 6.36% |
| TOTAL | 78736 | 100.00% | 38948 | 100.00% |

图 2-2 coco-vqa 中“问题-答案”对的分布情况

VQA VQA 数据集是视觉问答领域发展的一个重要拐点，在此之前的数

据集的问题类型被限制在一些模板之中，这使得数据集不能很好地测试出视觉问答系统在真实语境下的表现，例如，DAQUAR 将答案仅仅限制在 16 种基本颜色和 894 种物体类别中^[32]。VQA 数据集中的问题和答案是无限制、开放式的，且全部由人类产生，同时图片的数量相较 DAQUAR 提高了两个数量级，到达 254731 张，极大的提高了数据集的容量。VQA 数据集不仅包含从 MS COCO^[46] 中提取的 204721 张真实场景的图片，还提供了 50000 张合成的抽象场景图（如图2-3），丰富了数据库场景的多样性，同时为高阶的场景推理和复杂空间推理提供了便利。



图 2-3 VQA 中的真实、抽象场景图像实例

为了实现对复杂推理的训练和测试，VQA 数据集在问题设置上采用了人工的方式，每张图片都有 3 个人类提出的问题。答案则分为开放式和多项选择两种形式，开放式答案由于答案并不唯一，因此难以确定标准答案，因此正确答案的评估方法也引入人工评估机制：对于同一个开放性问题由十个人分别作答，如果有三个及以上的被测者均提供了同一答案，该答案被视为正确答案。多项选择的答案则由四种类型、18 个候选选项组成（如图2-4）：

正确答案 一个，从被测者回答中取最为常见的作为正确答案

混淆答案 三个，不看图，仅根据问题作答的答案

常见答案 十个，数据集中最出现频数最高的十个答案

随机答案 四个，除去已经列出的选项，随机挑选四个答案

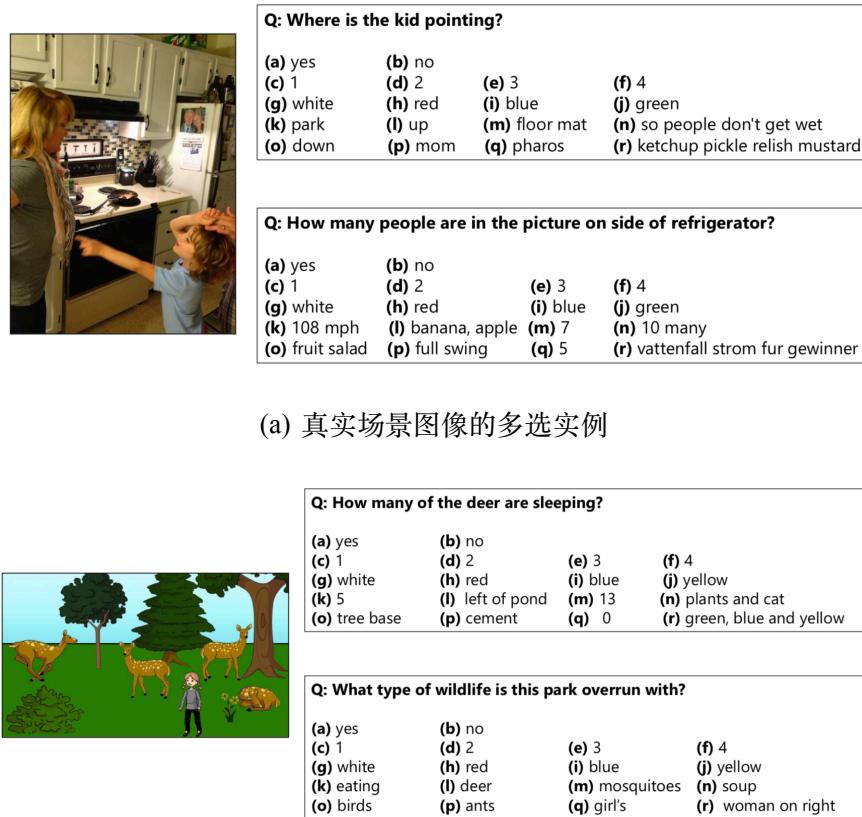


图 2-4 VQA 中的真实和抽象场景图像的多项选择实例

VQA 2.0 VQA 数据集由于其建立了开放性问题和多项选择问题的评测标准，成为众多算法的测试数据集，但 VQA 数据集存在语言偏见问题。具体表现为，即使是完全无视图像的算法也能在 VQA 数据集上得到 49.6% 的准确率^[31]，这意味着在 VQA 数据集的测试环境下，系统对于视觉信息的需求程度远远小于语言信息，这种状况相较于人类对于图像问答任务中的真实体验而言，是严重不符的。例如，答案为“是或否”的问题占所有问题的 38%，并且大约 59% 的二值问题答案都为“是”；询问“什么运动”的问题中有 41% 的答案为“网球”；询问数量的问题中有 39% 的答案为“2”。

针对以上问题，VQA 2.0 通过在原有的 VQA 数据集基础上补充新的“混淆数据”实现数据集对视觉信息的增强。“混淆数据”和原始数据一样由（图像 I，问题 Q，答案 A）的形式组织，不同的是新补充的图像与原有图像相似，但回答同样的问题 Q 却得到不同的答案 A(如图2-5)。针对同样的问题，在不同图片背景下需要得到不同的答案，这要求系统不仅能理解自然语言问题，同样需要关注图片的语义差异，才能得到正确的答案，这种平衡的方法能够筛选掉弱化图像理解的算

法，强化了图像理解在视觉问答任务的重要性。补充后的 VQA 2.0 包含 110 万对“图像-问题”、20 万张关联 1300 百万个问题的真实场景图片，数据量几乎是 VQA 数据集的两倍，成为了开放性问题的新测试标准。

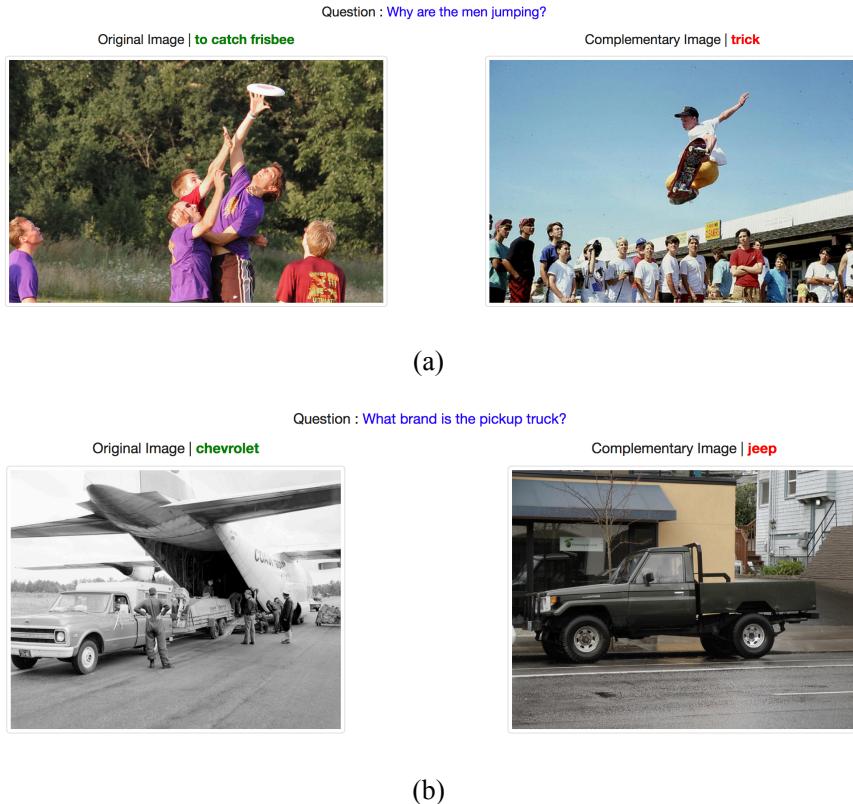


图 2-5 VQA 2.0 中针对同一问题的不同图像和答案实例

CLEVR 为了更加准确地衡量视觉问答系统各个方面的推理能力，Johnson 等人提出了一个结合语言和基本视觉推理诊断数据集 CLEVR。CLEVR 包含 10 万张由空间立方体组成的合成图像、将近 100 万个问题，其中包含 85.3 万个独特的问题。在图像的设置上，CLEVR 为了减小识别难度，关注系统的视觉推理能力，采用了由空间立方体组成的合成图像，并且每张图像均有包含所有物体位置和属性的说明（如图2-6）。CLEVR 的问题也均由程序生成得到，涉及属性识别、计数、比较、逻辑运算等子任务。

为了减少问题的偏见，数据集生成的问题中有 85% 是独特的；为了控制问题的准确性，数据集剔除了有歧义的问题，例如，询问“正方体右边的球体是什么颜色？”时，如果“正方体”右边有多个“球体”，问题便产生歧义，答案变得不唯一，使得评估过程变得复杂和不准确；为了保持问题的复杂性，数据集拒绝了一些看似复杂但实际上限定条件无效的问题，例如，询问“球体前面的圆柱体是否为金属

的?”时,如果场景中仅有一个“圆柱体”,那么问题中的“球体前面”的限定便可以被忽略,这种情况降低了问题的复杂性。

由于 CLEVR 数据集对图像和问题具有完全的掌控,能实现其他数据集难以实现的能力测试,要求系统具有短期的记忆力、注意力机制、组合推理能力。但同样因为其简单的图像场景设置,CLEVR 不能测试出视觉问答系统在常识推理、复杂推理的表现,并且也不能衡量系统在真实场景中的识别能力和稳定性。

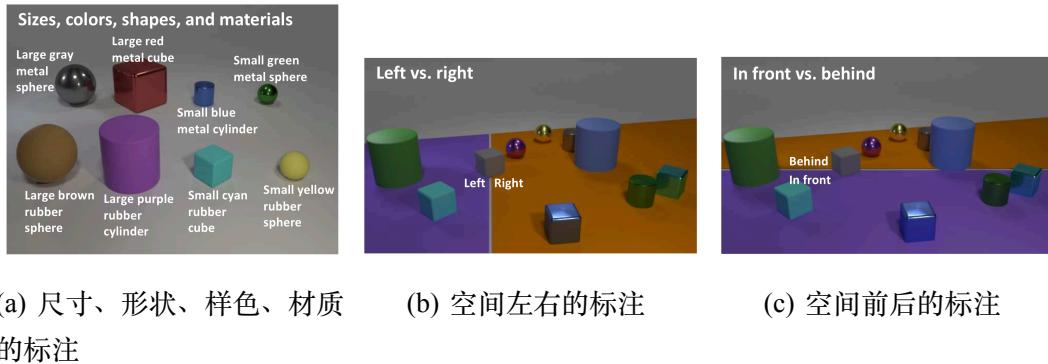


图 2-6 CLEVR 中图像标注

2.2 基于知识的数据集

KB-VQA 真实场景中的开放性问题可能涉及常识或者特定领域知识的先验知识,为了更好的评估 VQA 算法对需要高层次知识问题的准确推理能力,Wang 等人构建了只包含复杂推理问题的数据集 KB-VQA^[17]。

KB-VQA 数据集从 MS COCO^[46] 中挑选出 700 张图片样本,挑选出的图片包含 150 个物体类别和 100 个场景类别。每张图片附带有 3-5 个由人工生成的“问题-答案”对,所有的问题被限定在 23 种问题模板中,例如,“图片中是否存在某种概念?”,“图片中的某个物体被生产于什么地方?”等,详见2-7。

为了准确评估系统在需要先验知识的问题的表现,KB-VQA 人工地赋予每个问题一个表示所需不同知识类型的标签,“视觉问题”、“常识问题”和“知识库问题”,其中“视觉问题”表示仅仅从图片中便可以获得答案的问题,例如,“物体是否存在于图片?”、“列出图片中包含的所有事物?”等,“常识问题”需要结合成人级别的常识和图像内容得出答案,例如,“图片涉及什么场景?”,“知识库问题”则需要某个领域特定的知识才能完成作答,例如,“图中的物品在哪一年被发明?”。23 种问题模板在不同问题标签的分布如图2-8。

| Name | Template | Num. |
|-----------------------|---|------|
| <i>IsThereAny</i> | Is there any <i><concept></i> ? | 419 |
| <i>IsImgRelate</i> | Is the image related to <i><concept></i> ? | 381 |
| <i>WhatIs</i> | What is the <i><obj></i> ? | 275 |
| <i>ImgScene</i> | What scene does this image describe? | 263 |
| <i>ColorOf</i> | What color is the <i><obj></i> ? | 205 |
| <i>HowMany</i> | How many <i><concept></i> in this image? | 157 |
| <i>ObjAction</i> | What is the <i><person/animal></i> doing? | 147 |
| <i>IsSameThing</i> | Are the <i><obj1></i> and the <i><obj2></i> the same thing? | 71 |
| <i>MostRelObj</i> | Which <i><obj></i> is most related to <i><concept></i> ? | 56 |
| <i>ListObj</i> | List objects found in this image. | 54 |
| <i>IsTheA</i> | Is the <i><obj></i> a <i><concept></i> ? | 51 |
| <i>SportEquip</i> | List all equipment I might use to play this sport. | 48 |
| <i>AnimalClass</i> | What is the <i>(taxonomy)</i> of the <i><animal></i> ? | 46 |
| <i>LocIntro</i> | Where was the <i><obj></i> invented? | 40 |
| <i>YearIntro</i> | When was the <i><obj></i> introduced? | 32 |
| <i>FoodIngredient</i> | List the ingredient of the <i><food></i> . | 31 |
| <i>LargestObj</i> | What is the largest/smallest <i><concept></i> ? | 27 |
| <i>AreAllThe</i> | Are all the <i><obj></i> <i><concept></i> ? | 27 |
| <i>CommProp</i> | List the common properties of the <i><obj1></i> and <i><concept/obj2></i> . | 26 |
| <i>AnimalRelative</i> | List the close relatives of the <i><animal></i> . | 17 |
| <i>AnimalSame</i> | Are <i><animal1></i> and <i><animal2></i> in the same <i>(taxonomy)</i> ? | 17 |
| <i>FirstIntro</i> | Which object was introduced earlier, <i><obj1></i> or <i><concept/obj2></i> ? | 8 |
| <i>ListSameYear</i> | List things introduced in the same year as the <i><obj></i> . | 4 |

图 2-7 KB-VQA 中 23 中问题模板及对应的问题数量

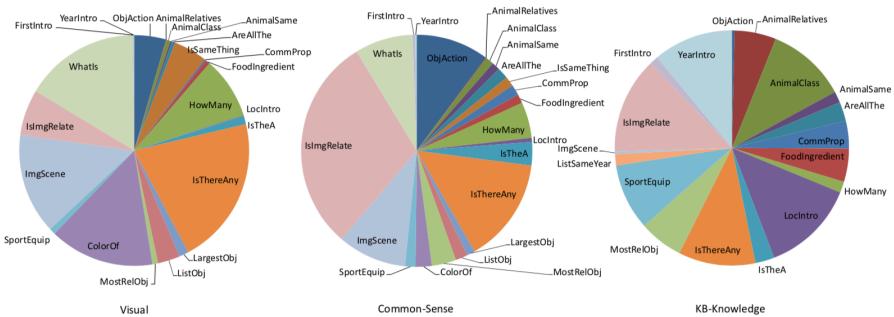


图 2-8 KB-VQA 中 23 中问题模板对应“视觉问题”、“常识问题”和“知识库问题”的分布情况

数据集中的“视觉问题”、“常识问题”和“知识库问题”数量分别是 1256、883 和 263，就图片和问题的数量而言，KB-VQA 数据集相较于 COCO-QA 等数据集是非常小的，而且从图2-7也容易看出，有 16 种问题类型的问题数量都不超过 100 个，甚至有个位数的问题数量，数据集的不均衡和小容量很难准确得评估出系统在细分问题类型上的推理能力。但需要先验知识的问题占比要远高于大型数据集，DAQUAR^[32] 几乎全是“视觉问题”，COCO-VQA^[31] 仅仅包含 5.5% 的问题需要常识，没有问题需要额外的知识库。

KB-VQA 在评估系统的复杂推理能力方面提供了一个解决方案，但数据集的容量、平衡性和多样性方面还需要更多的丰富，并且随着数据集的扩充，自动化和标准化的评估方式也相应的需要完善。

FVQA 为了评估视觉问答系统在需要先验知识的问题上的表现，Wang 等人提出了 FVQA 数据集^[18]。回答 FVQA 中的问题需要额外的知识，但不同于一般的数据集，FVQA 将（图片，问题，答案）的三元组数据扩展为（图片，问题，答案，支持事实）的四元组形式，其中“支持事实”是回答问题所需要的额外知识，使用资源描述框架（RDF）的三元组形式，例如（猫，可以，爬树）。

FVQA 从 MS COCO^[46] 和 ImageNet^[47] 中挑选出 1906 张图片，并对图片预处理，提取出三种类型的视觉概念：物体对象、场景和行为，最终提取出 326 种物体对象、21 种场景和 24 种行为。为了获取与视觉概念相关的知识，FVQA 以 DBpedia^[48]、ConceptNet^[49] 和 WebChild^[50] 为知识源，从三种知识库中与视觉概念相关的所有知识中筛选出包含 12 种常见的谓语的知识，例如，关于分类的知识——“目录属于”、关于地点的知识——“地点所在”、关于大小比较的知识——“体积大于”，详见图2-9。提取的知识以资源描述框架（RDF）的形式存储作为“支持事实”。

| KB | Predicate | #Facts | Examples |
|------------|--|--------|--|
| DBpedia | Category | 35152 | (Wii, Category, VideoGameConsole) |
| ConceptNet | RelatedTo | 79789 | (Horse, RelatedTo, Zebra), (Wine, RelatedTo, Goblet) |
| | AtLocation | 13683 | (Bikini, AtLocation, Beach), (Tap, AtLocation, Bathroom) |
| | IsA | 6011 | (Broccoli, IsA, GreenVegetable) |
| | CapableOf | 5837 | (Monitor, CapableOf, DisplayImages) |
| | UsedFor | 5363 | (Lighthouse, UsedFor, SignalingDanger) |
| | Desires | 3358 | (Dog, Desires, PlayFrisbee), (Bee, Desires, Flower) |
| | HasProperty | 2813 | (Wedding, HasProperty, Romantic) |
| | HasA | 1665 | (Giraffe, HasA, LongTongue), (Cat, HasA, Claw) |
| | PartOf | 762 | (RAM, PartOf, Computer), (Tail, PartOf, Zebra) |
| | CreatedBy | 96 | (Bread, CreatedBy, Flour), (Cheese, CreatedBy, Milk) |
| WebChild | Smaller, Better, Slower, Bigger, Taller, ... | 38576 | (Motorcycle, Smaller, Car), (Apple, Better, VitaminPill), (Train, Slower, Plane), (Watermelon, Bigger, Orange), (Giraffe, Taller, Rhino) |

图 2-9 从三种知识库中提取的知识涉及的 12 种谓语及相应的数量

FVQA 的问题和答案均使用人工的方式收集得到，被试者先选择图片中的一个视觉概念和一个与视觉概念相关的支持事实，再根据视觉概念和支持事实给出问题和答案，答案的来源要么是图片中的视觉概念要么是支持事实中涉及的概念。数据集最终包含 4608 个需要先验知识的问题，涉及 3458 条事实。根据视觉概念的类型，这些问题可以归为物体对象、场景和行为三种类型；根据支持事实的来源，可以归为 DBpedia、ConceptNet 和 WebChild 三种类型；根据答案来源，可以归为图片来源和知识库来源两种类型，不同分类在训练集和测试集的数量分布如图2-10。

| Criterion | Categories | Train | Test | Total |
|----------------|------------|-------|------|-------|
| Visual Concept | Object | 2087 | 1997 | 4084 |
| | Scene | 273 | 220 | 493 |
| | Action | 14 | 17 | 31 |
| Answer-Source | Image | 2014 | 1938 | 3952 |
| | KB | 360 | 296 | 656 |
| KB-Source | DBpedia | 345 | 343 | 688 |
| | ConceptNet | 1881 | 1757 | 3638 |
| | Webchild | 148 | 134 | 282 |
| Total | | 2374 | 2234 | 4608 |

图 2-10 不同分类在训练集和测试集的数量分布

从统计的数据上不难看出，绝大多数问题是针对图像中的物体对象，这与提供的视觉概念中物体对象的高占比有强关联，从知识来源上分析，答案除了能从图像中获得外，还包含 14% 的答案需要从额外知识库中获得，并且问题中不包含“是或否”的二值问题，这降低了系统“猜中正确答案”的情况。

FVQA 和同样包含先验知识的数据集 KB-VQA 两者都能通过查询语言获取知识库中的数据，但不同于 KB-VQA，FVQA 拥更多的图片和问题数量，并且所有问题都需要额外知识。FVQA 增加了 ConceptNet 和 WebChild 作为知识源，提高了知识库的多样性，能回答更多类型的问题，而不用预先设定问题模板。但 FVQA 数据集中几乎所有的答案都是物体对象，且为单个词语，不能训练模型给出对象关系的答案。FVQA 数据集的支持事实多为单一谓语的句子，句式结构简单，如果用做训练集，不能考察模型应对多动词结构问题时的答案正确率。

然而两个数据集都面临着同样的问题：数据量的扩充和问题类型的扩充。两个数据集的问题收集都是通过人工的方式，并且参与者数量有限，因此直接导致了问题数量远低于其他自动化方法生成的数据集。大规模的协同工作和探索更多自动化方法是扩充数据集容量的方向。两个数据集都受到问题类型的限制，KB-VQA 使用预先设定的问题模板，限制了问题的开放程度，FVQA 虽然没有使用预先设定的问题模板，但其筛选的 12 种谓语间接的限制了问题的类型。

2.3 实验数据集

从表格2-1可以看出，在基于视觉的数据集中，VQA2.0 的图像数量、问答对数量远远高于其他数据集，并且由于数据集中添加的“混淆数据”——相似的图像

而不同的问答对，数据集对于模型对于图像的理解提出了更高的要求，比起其他数据集更加均衡。然而和其他基于视觉的数据集一样，VQA2.0 的问题对于外源知识的需求很低，这种需要外源知识或者常识的问题缺失限制了其对基于知识库的 VQA 模型的检验。而基于知识的数据集则要求模型能够解决外源知识引入的问题。

为了衡量模型在识别类和推理类两类问题的表现，我们采用了 VQA2.0 数据集、KB-VQA 和 FVQA 三个数据集，其中 VQA2.0 数据集用于训练 N-KBSN 模型，而 KB-VQA 和 FVQA 则用于训练 KBSN 模型，最后混合三种数据集用于测试模型的表现。

2.4 本章小结

根据数据集中的问题是否涉及常识和外源知识，本章分别介绍了基于视觉和基于知识的数据集，分析总结了各自的优劣点，并构建了实验数据集。

第三章 基于动态词向量的联合嵌入模型

视觉问答来的研究受到机器学习算法在自然语言处理和图像识别等领域成功应用的启发，因此从 2015 年视觉问答任务出现至今，大量的 VQA 模型都使用了联合嵌入模型，使之成为目前视觉问答的主流模型。顾名思义，联合嵌入模型是将任务的源信息——图像和问题文本——表示为向量，再通过特征融合，将不同模态的信息映射到统一的向量空间，最终从联合表征中提取出答案。因为这种架构的模型易于训练，研究者采用不同的图像特征的提取方法、不同的文本特征的提取方法、两种模态的不同融合方法，做了许多尝试。

Antol 等人在 2015 年发布了开放问题的视觉问答数据集 VQA^[1]之后，在数据集的基础上提出了 VQA 挑战。VQA 挑战中涌现了大量视觉问答模型，模型的准确率也逐年升高，图3-1展示了 2015 年-2019 年 VQA 挑战中的最优模型的准确率。通过研究其中表现优异的模型，我们发现几乎所有模型都使用了联合嵌入模型，并且加入注意力机制之后准确率能够进一步提升，例如，四年的冠军模型都是使用了注意力机制的联合嵌入模型，其中 2019 年的冠军模型^[51]能在 VQA2.0 数据集下获得总体 75% 作用的准确率，相较于四年前的模型准确率得到了 20% 的提升，并且距离人类表现也只有 5% 左右的差距。

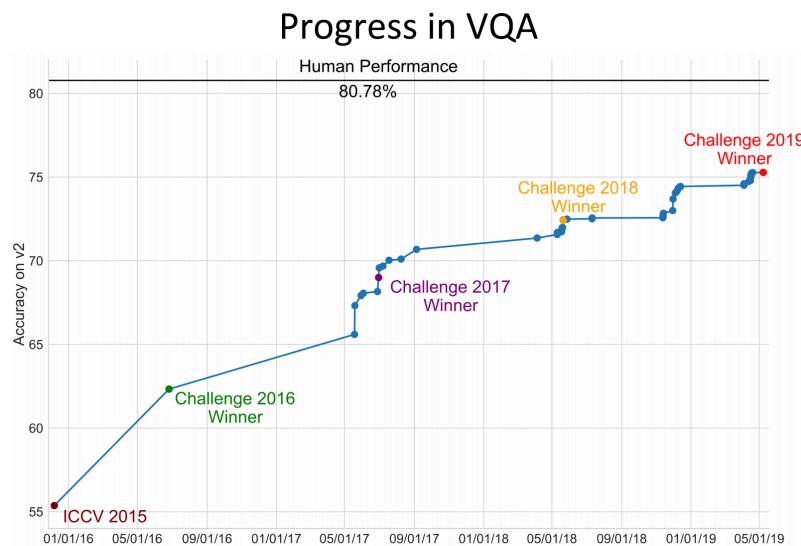


图 3-1 视觉问答模型在 VQA 挑战的准确率曲线，其中四年的优胜模型都是使用联合嵌入模型。

我们认为联合嵌入模型在 VQA 挑战的优异表现有以下几点原因。

- 1) 引入注意力机制。2016 年的优胜模型^[34]提出“动态关注注意力 (FDA)”模

型，目的是根据问题中的关键词动态的对图像的不同区域分配带有权重的注意力，得到图像的全局特征和局部特征的结合。2017 和 2018 年的优胜模型都使用论文^[52] 中的自上而下和自下而上的图像注意力机制，2019 年的优胜模型使用了 Transformer^[53] 的多头注意力机制。注意力机制的引入能够减少无关特征的干扰，提高计算效率，并且一定程度的提高可解释性。

- 2) VQA2.0 数据集的局限性。VQA 挑战以 VQA2.0 为数据集，然而根据我们提出的依照答案和源信息统计相关性的标准（详见表1-1），VQA2.0 中需要常识或者外源知识的 qi 类型仅仅占所有问题的 5.5%^[17]，这意味着回答绝大多数的问题都不需要额外的信息。然而在现实中的开放性问题中，涉及常识或者外源知识的问题广泛存在，因此 VQA2.0 数据集存在局限性，而这种局限性使得模型只需要关注图像和文本，因此联合嵌入模型成为了主要架构。
- 3) 得益于图像识别和自然处理模型的进步。联合嵌入模型具有灵活的组合模式，很容易从将其他任务中表现优异的模型迁移过来形成新的模型。

在本文的研究中，为实现一个通用的 VQA 架构，我们分成两个阶段完成，第一部分是沿袭联合嵌入模型的思路，构建一个基于动态词向量的联合嵌入模型——None KB-Specific Network (N-KBSN) 模型，该模型仅仅使用图像特征和文本特征，不使用外源知识。第二个阶段是在 N-KBSN 模型的基础上，融合知识库的图嵌入，提出 KB-Specific Network (KBSN) 模型。

本章将重点介绍 N-KBSN 模型，并且使用 VQA2.0 数据集训练。N-KBSN 由三个主要部分组成：问题文本和图像特征提取模块、自注意力和引导注意力模块、特征融合和分类器。其中，图像特征提取使用在多目标检测中表现优秀的 Faster R-CNN^[39]，问题文本特征提取使用能够获得上下文信息的 ELMo 模型^[54]，并使用从 Transformer 中借鉴的多头注意力机制^[53] 分别实现图片的自注意力 (V-SA)、问题文本的自注意力 (Q-SA)、由问题引导的对图像的注意力 (Guided Attention, GA)，最后通过特征融合预测答案。N-KBSN 模型的基础架构如图3-2。

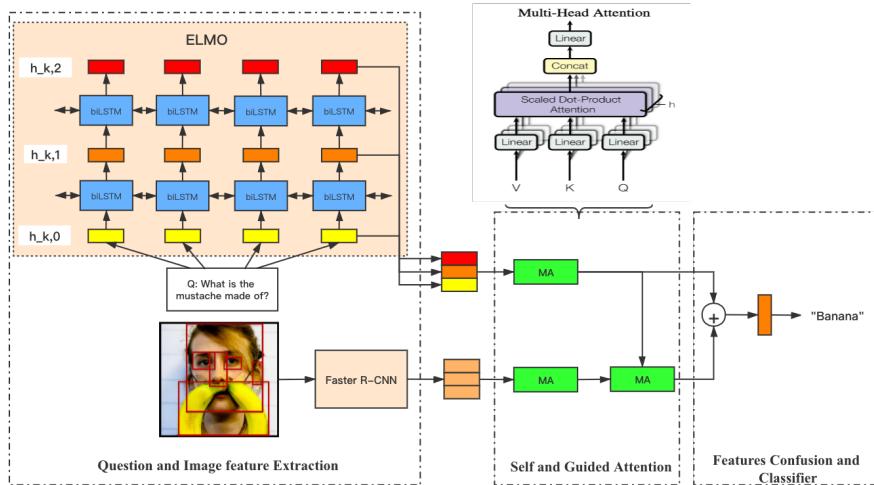


图 3-2 N-KBSN 基础结构

3.1 基于 Faster R-CNN 的图像特征化

目标检测是机器视觉领域的重要应用之一，目标检测的核心任务是准确、快速的从图像中定位出目标并且能识别目标的类别、属性等特性。在 2012 年深度学习正式介入计算机视觉目标检测任务之前，传统的目标检测算法一直是以滑动窗口卷积等较为传统的方式进行区域选择、特征提取和分类回归等步骤，例如可变形的组件模型（DPM）方法^[55] 等。这些传统的目标检测算法大多区域选择的策略效果差、时间复杂度高，并且因为由于是人工提取特征，提取的特征层次较低，致使模型的鲁棒性较差。在深度学习兴起并逐渐成为计算机视觉的核心方法之后，大批优秀的目标检测算法出现，例如 R-CNN^[56]、SPP-Net^[57]、Fast R-CNN^[58]、Faster R-CNN^[39]、Mask R-CNN^[59]、YOLO^[60] 及其后续版本等。以上的模型大致分为两个主要类别，第一类为两级式检测框架，包含一个用于区域提议的预处理步骤，使得整体流程是两级式的，例如一系列的 R-CNN 模型；第二类为单级式检测框架，即无区域提议的框架，这是一种单独提出的方法，不会将检测提议分开，使得整个流程是单级式的，例如 YOLO 系列的模型。由于 Faster R-CNN 在各个目标识别任务的出色表现，本节将省略对单级式检测框架的介绍，并着重介绍本文中图像处理的核心模型 Faster R-CNN。

区别于传统的滑动卷积窗口来判断目标的可能区域，R-CNN 采用选择性搜索的方法来预先提取一些可能包含目标物体的候选区域（region proposal），再使用卷积神经网络提取各个图像区域的特征，再将提取的特征送入 SVM 分类器完成类别识别，最后使用回归器对目标位置进行修正。这种方法显著的提升识别速度，降低了计算成本，也提高了准确率。因为 R-CNN 需要分别对每一个生成的候选区域

进行一次特征提取，存在着大量的重复运算，制约了算法性能。为了减少 R-CNN 的重复计算，研究者提出了 SPP-Net。该算法通过在网络的卷积层和全连接层之间加入空间金字塔池化层（Spatial Pyramid Pooling）来对利用 CNN 进行卷积特征提取之前的候选区域进行裁剪和缩放使 CNN 的输入图像尺寸一致。随后的 Fast R-CNN 借用了 SPP-Net 的空间金字塔池化层，设计了兴趣区域池化（RoI Pooling），将图像中的多个兴趣区域池化成相同大小的特征图，并使用这些特征图同时预测物体类别和框出对象的区域。这种方法解决了输入候选区域尺寸不一致的问题，并且提高了计算速度。但是 Fast R-CNN 在生成生成候选区域的较慢，为了解决这一问题，R-CNN 的作者又提出了 Faster R-CNN。

Faster R-CNN 同样沿袭了先前 R-CNN 和 Fast R-CNN 的两级式检测框架，但是为了解决之前的大量候选框导致的速度慢的问题，Faster R-CNN 设计了一个用于选择和判断候选区域的网络（Region Proposal Network, RPN），该网络将 CNN 处理后的全局图像特征作为输入，输出候选区域，最终的分类器结合全局的图像特征和候选区域预测各个区域的类别。

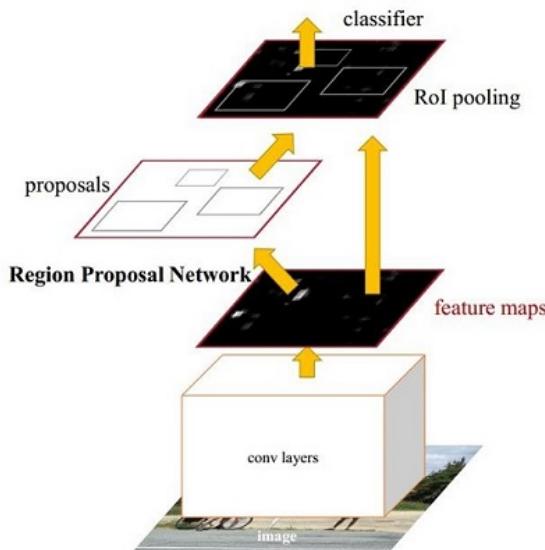


图 3-3 Faster R-CNN 基本结构

如图3-3，Faster R-CNN 可以根据功能的不同将模型分为四个模块：卷积层、区域候选网络（RPN）、兴趣区域池化（RoI Pooling）、分类器。卷积层使用 CNN 及其变型提取图像特征，生成的特征图被共享用于后续 RPN 层和全连接层，这种对图像的处理方式不同于 R-CNN 和 Fast R-CNN。后两者都是先从原始图像中提取候选区域，再分别对候选区域提取特征。RPN 网络用于预测候选区域，该网络在 CNN 输出的图像特征上滑动，在每个空间区域，网络都会预测类别得分，通过

softmax 判断锚点 (anchors) 属于前景 (foreground) 或者背景 (background)，利用候选框回归 (bounding box regression) 修正锚点获得精确的候选区域。兴趣区域池化层收集输入的特征图和候选区域，综合这些信息后提取候选区域特征图，送入后续全连接层判定目标类别。分类器利用候选区域特征图计算区域的类别，同时再次使用候选框回归获得检测框最终的精确位置。

在本文中，我们使用联合在 ImagNet^[61] 上预训练的 ResNet-101 和在 Visual Genome^[4] 上预训练 Faster R-CNN 提取图像特征。给定图像 I ，我们从图像中提取 m 个大小不固定的图像特征， $X = \{x_1, x_2, \dots, x_m\}, x_i \in \mathbb{R}^D$ ，每一个图像特征编码一个图像区域。对卷积层输出的特征图，使用非极大抑制 (non-maximum suppression) 和单元重合 (IoU) 阈值筛选出排名靠前的候选区域，通过设定一个目标检测概率的阈值，我们获得一个动态的被检测对象的数量 $m \in [10, 100]$ ，并且使用零填充使得 $m = 100$ 。对于每个所选区域 i ， x_i 被定义为该区域的特征图的均值池化结果，并将 m 个区域的 x_i 拼接成为最终的图像特征。

3.2 基于 ELMo 的文本特征化

和众多自然语言处理任务一样，在视觉问答任务中如何准确理解问题内容对最终的答案准确率上有着决定性的影响。而自然语言理解中最为基本和核心的便是文本表达，文本表达将自然语言转换为计算机可处理的数字，为自动化处理文本相关的任务建立了基础。在文本表达中，独热向量 (one-hot) 是最早也是最为简单的词向量。但是其稀疏性会带来的“维度灾难”和因简单的编码方式而造成“语义鸿沟”。基于分布式假设——即处于相似上下文的词语具有相似的含义，研究者先后提出了多种使用分布式表示的词向量模型，例如，CBOW，Skip-Gram，word2vec^[62]，潜在语义分析 (LSA)^[63]，GloVe^[64]，ELMo^[54]，BERT^[65] 等。

CBOW 和 Skip-Gram 均是使用神经网络模型训练上下文信息得到词向量。word2vec 也使用了 CBOW 与 Skip-Gram 来训练模型与得到词向量，但是并没有使用传统的 DNN 模型，而是使用霍夫曼树来代替隐藏层和输出层的神经元，提高了计算效率，因此被研究者广泛地使用作为预训练的词向量。但是由于 word2vec 使用滑动窗口来限定上下文信息，因此得到的词向量仅仅使用了局部的语义和语法信息。不同于 word2vec 使用局部语料，潜在语义分析 (LSA) 采用统计计数的方式获得语料的全局信息，其统计预料库中每两个词共同出现的次数构成共现矩阵，并采用了基于奇异值分解 (SVD) 的矩阵分解技术对大矩阵进行降维，得到词向量。然而 LSA 方法中的 SVD 计算量很大，并且共现矩阵仅能表示两个词语同时出现的次数，并不能表示词语之间的远近关系。为了改进 word2vec 的局部预料限制

和 LSA 的计算复杂性，GloVe 使用衰减函数改造 LSA 的共现矩阵，使得词语间的远近关系得以表达。GloVe 还构建了词向量和共现矩阵之间的近似关系，使用梯度下降算法取代了 LSA 中的奇异值分解，大大减少了计算代价，并且得到了远超 LSA 和 word2vec 的性能。

以上提到的所有模型都是通过对语料库的学习得到静态的词向量，即每个单词对应一个确定的实数向量，这种固定向量在处理词汇的多义性上表现不佳。无论是中文词语还是英文单词都广泛得存在一词多义的现象，即同一个词在不同的语境下含义发生变化，例如，在中文中，“他正在算账”和“下回找你算账”中的“算账”由于文化演化而产生了更复杂的引申义，又如英文中的“where is the bank?”和“It is the bank of the river”中的“bank”在第一句中译为“银行”而第二句中译为“河畔”。为解决一词多义的问题，研究者提出了动态词向量，ELMo 和 BERT 便是其中的代表。ELMo 在多个 NLP 任务中均提高了模型的准确率，因此本文将着重介绍并使用 ELMo 模型处理视觉问答任务中的文本，并在后续的处理中结合类似于 BERT 的注意力机制。

ELMo 是一种深度场景化的词表示。其模型深度能够对复杂的词语使用特性——语法和语义特征进行有效建模，而其模型的动态性能根据词语的上下文的不同生成动态向量，进而为解决一词多义提供了可能。ELMo 采用了两个阶段获得词向量，第一个阶段是用大量的文本语料训练一个深度双向语言模型（biLSTM）；第二个阶段从预训练网络中提取对应单词的网络各层的内部状态（internal state），并通过函数转化为词向量。ELMo 模型的结构如图3-4。

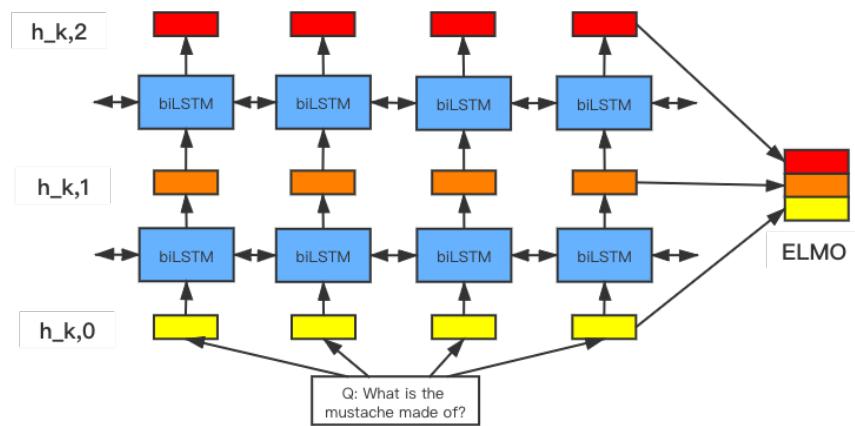


图 3-4 ELMo 使用两层的 biLSTM 层，并对每一层的输出加权得到词嵌入。

语言模型是对语句的概率分布的建模。语言模型分为前向和后向，前向是指已知上文的词语，推理下一个词语的方式，而后向则是已知后文的内容，求解上

一个词语的方式。对于一个具有 N 个单词的句子 $S = (t_1, t_2, \dots, t_N)$ 而言，前向语言模型就是求解以下公式的最大值：

$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^N p(t_k | t_1, t_2, \dots, t_{k-1}) \quad (3-1)$$

其中 $p(t_1, t_2, \dots, t_N)$ 为序列的联合概率, $p(t_k | t_1, t_2, \dots, t_{k-1})$ 表示已知 t_k 的上文 $(t_1, t_2, \dots, t_{k-1})$ 的条件下，求解 t_k 的条件概率。对应的后向语言模型的公式为

$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^N p(t_k | t_{k+1}, t_{k+2}, \dots, t_N) \quad (3-2)$$

ELMo 使用双向 LSTM (biLSTM) 模型作为语言模型的基础。首先将“上下文无关的”初始词向量 y_k^{LM} 输入 L 层的前向 LSTM。在位置 k 上，LSTM 将输出一个“上下文相关”的词表征 $\vec{h}_{k,j}^{LM,j}$ ，其中 $j = 1, \dots, L$ 。最后一层的 LSTM 输出 $\vec{h}_{k,j}^{LM,L}$ 通过一个 softmax 层预测下一个词语的初始词向量 y_{k+1}^{LM} 。后向 LSTM 类似于前向 LSTM 有 L 层并且在 k 位置上得到一个词表征 $\overleftarrow{h}_{k,j}^{LM}$ 。最后通过最大似然的方式训练双向 LSTM 模型，公式如下：

$$\sum_{k=1}^N (\log_p(t_k | t_1, t_2, \dots, t_{k-1}; \Theta_x, \overrightarrow{\Theta_{LSTM}}, \Theta_s) + \log_p(t_k | t_{k+1}, t_{k+2}, \dots, t_N; \Theta_x, \overleftarrow{\Theta_{LSTM}}, \Theta_s)) \quad (3-3)$$

其中， Θ_x 和 Θ_s 分别是初始词向量训练时的两个 softmax 层参数， $\overrightarrow{\Theta_{LSTM}}$ 和 $\overleftarrow{\Theta_{LSTM}}$ 则是双向语言模型的参数。

当完成预训练阶段后，向网络输入一个新句子，句子中每个单词都能得到对应的三种 Embedding: 最底层是初始的词向量 y_k^{LM} ；前向 LSTM 输出的 $\vec{h}_{k,j}^{LM}$ ；后向 LSTM 输出的 $\overleftarrow{h}_{k,j}^{LM}$ 。ELMo 将三种词向量串联，得到

$$R_k = [y_k^{LM}, \vec{h}_{k,j}^{LM}, \overleftarrow{h}_{k,j}^{LM}] \quad j = 1, \dots, L = [\vec{h}_{k,j}^{LM}] \quad j = 0, \dots, L \quad (3-4)$$

其中 $h_{k,0}^{LM}$ 是初始词向量， $h_{k,j}^{LM} = [\vec{h}_{k,j}^{LM}, \overleftarrow{h}_{k,j}^{LM}]$ 是每个 biLSTM 层输出的结果。

最后使用以下公式得到对应单词的具有“上下文信息”的词向量。

$$ELMo_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^L s_j^{task} h_{k,j}^{LM} \quad (3-5)$$

其中是 s_j^{task} 任务相关训练得到的权重参数， γ 是一个任务相关的 scale 参数。

在本文中，我们将句子最大长度裁剪为 14 个，并使用零填充的方式将不足 14 个词的句子补足为 14，即 $n = 14$ 。并将每个单词转化为 50 维的初始词向量，即 $y_k^{LM} \in \mathbb{R}^{50}$ 。假定双向语言模型的层数为 $L = 2$ ，隐层节点数为 H_{dim} ，输出维度为

$output_{dim} \in \mathbb{R}^d$, 则 $ELMo_k^{task} \in \mathbb{R}^{2d}$, 输出的文本特征 $Y \in \mathbb{R}^{n \times 2d}$ 。

3.3 基于多头注意力机制的特征增强

正如绪论中提到的, 注意力机制的引入帮助神经网络提高了预测精度, 并且减少了计算复杂度。视觉问答任务由于需要处理多模态的数据——图像和文本, 比起仅需要处理单模态的数据的任务更需要进行高效的计算。同时, VQA 任务输入的图像和问题文本具有高度的相关性, 因此两种模态的数据之间的交互对于结果的准确性的提升也具有显著的影响。对于以上两个需求, 我们在 N-KBSN 中使用了 Transformer^[53] 的多头注意力机制 (Multi-head Attention, MA) 实现图片的自注意力 (V-SA)、问题文本的自注意力 (Q-SA)、由问题引导的对图像的注意力 (Guided Attention, GA)。

注意力机制本质上是找到一个方式对已有信息分配合适的权重, 并以此提高输出的准确性。我们可以将注意力函数描述成映射查询 (query) 到一些键值对 (key-value pair) 并由此得到输出。假定查询矩阵 $Q = \{q_1, q_2, \dots, q_m\}$, 其中查询向量 $q_i \in \mathbb{R}^{1 \times d_q}$; key 矩阵 $K = \{k_1, k_2, \dots, k_n\}$, 其中 $k_j \in \mathbb{R}^{1 \times d_k}$; value 矩阵 $V = \{v_1, v_2, \dots, v_n\}$, 其中 value 向量 $v_i \in \mathbb{R}^{1 \times d_v}$, 那么注意力特征可以通过对 value 矩阵的加权得到, 权重可以通过查询矩阵和 key 矩阵得到:

$$Attention(Q, K, V) = score(Q, K)V \quad (3-6)$$

其中 $score(Q, K)$ 为计算权重的函数, 有多种计算方式, 本文使用 Transformer 中的缩放点乘法:

$$score(Q, K) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (3-7)$$

其中 q_i 和 k_j 要求具有相同的维度。因此可以得到:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3-8)$$

为了进一步提高注意力特征的表达能力, 引入多头注意力机制。多头注意力机制的实现过程是, 将上式的 Q, K, V 输入到 h 个具有不同权重的线性层, 得到 $(Q_i, K_i, V_i), i = 1, 2, \dots, h$, 再分别计算得到 $Attention(Q_i, K_i, V_i), i = 1, 2, \dots, h$, 最后将 h 个注意力特征拼接并通过一个线性层获得期望维度的注意力特征, 如图3-5。多头注意力机制的公式为:

$$MA(Q, K, V) = [head_1, head_2, \dots, head_h]W \quad (3-9)$$

$$head_i = \text{Attention}(Q_i, K_i, V_i) \quad (3-10)$$

其中 W 为线性层的权重。

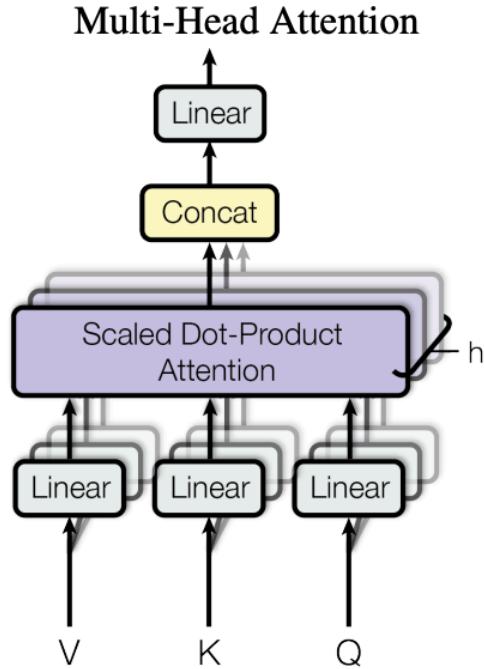


图 3-5 多头注意力的架构。多头注意力特征由 h 个缩放点乘注意力特征拼接得到。

基于以上多头注意力机制的思想，本文分别使用三种注意力特征：图片的自注意力（V-SA）、问题文本的自注意力（Q-SA）、由问题引导的对图像的注意力（GA）。假设文本词向量矩阵为 Y ，图像特征图为 X ，则在计算 V-SA 时， $Q = K = V = X$ ，即输出的图像特征为 $SA = MA(X, X, X)$ ；在计算 Q-SA 时， $Q = K = V = Y$ ，即输出的文本特征为 $SA = MA(Y, Y, Y)$ ；在计算引导注意力特征时， $Q = Y$ 为词向量矩阵， $K = V = X$ 为图像特征矩阵，并且词向量和图像特征向量具有相同的维度，即输出的由问题引导的图像特征为 $GA = MA(Y, X, X)$ 。三种注意力组合的结构构成一个共同注意力模块（MCA），结构如图3-6。共同注意力模块以输入为原始的图像特征和文本特征，输出为经过注意力机制的图像和文本特征。

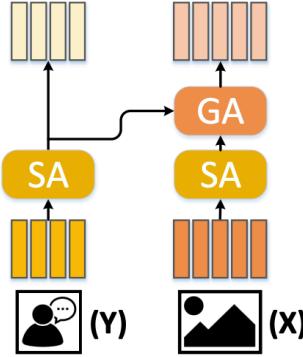


图 3-6 共同注意力模块（MCA）由一个 V-SA、一个 Q-SA 和一个 GA 组成。

为了提高使用深度的注意力机制提取更高层次的特征，MCAN 论文^[51]提出了 Encoder-Decoder 和 Stacking 两种级联 MCA 层的方式，如图3-7。

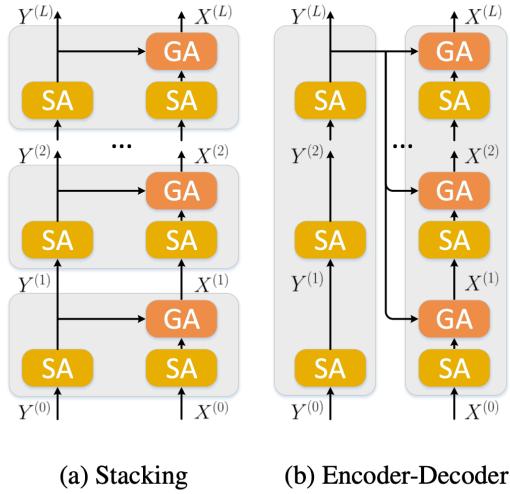


图 3-7 两种 MCA 层的级联方式，Stacking 将上一层的输出直接作为下一层的输入，Encoder-Decoder 将最后一层的问题自注意力特征作为每一层图像的查询矩阵。

根据文章给出的两种级联方式在多个任务的表现情况^[51]，在本文中，我们使用 Encoder-Decoder 的级联方式，假定 SA^1, SA^2, \dots, SA^L 表示不同层的自注意力， GA^1, GA^2, \dots, GA^L 表示不同层的引导注意力， $X^{(k)}$ 和 $Y^{(k)}$ 分别表示第 k 层输出的图像特征和文本特征。因此第 k 层 Encoder-Decoder 级联的注意力模块的公式为，

$$Y^{(k)} = SA^{(k)}(Y^{(k-1)}) \quad (3-11)$$

$$X^{(k)} = GA^{(k)}(Y^{(L)}, SA^{(k)}(X^{(k-1)})) \quad (3-12)$$

其中图像特征 $X^{(0)} = X$, $Y^{(0)} = Y$ 。

在获得经过多层注意力的图像特征 $X^{(L)} = [x_1^{(L)}, \dots, x_m^{(L)}] \in \mathbb{R}^{m \times d}$ 和文本特征 $Y^{(L)} = [y_1^{(L)}, \dots, y_n^{(L)}] \in \mathbb{R}^{n \times d}$, 我们对所有分量权重求和, 进一步得到最终的图像特征 x 和文本特征 y 。以图像特征为例, 公式如下。

$$\alpha = \text{softmax}(MLP(X^{(L)})) \quad (3-13)$$

$$x = \sum_{i=1}^m \alpha_i x_i^{(L)} \quad (3-14)$$

其中 $\alpha = [\alpha_1, \dots, \alpha_m]$ 是图像特征分量的权重。

$Y^{(L)}$ 的计算方式类似。我们使用一下公式融合两种特征。

$$z = \text{LayerNorm}(W_x^T x + W_y^T y) \quad (3-15)$$

其中 $W_x, W_y \in \mathbb{R}^{d \times d_z}$ 是线性映射矩阵, d_z 是融合后的特征向量的维度。最后我们使用 softmax 函数计算融合特征在 N 个类别的答案, N 为训练集中出现频率最高的答案。最后我们使用交叉熵更新模型参数。

3.4 实验

为了确定不同的参数配置对于 N-KBSN 模型在视觉问答任务上的表现, 我们使用通用的开放型问答数据集 VQA2.0^[42] 训练模型, 并和目前的最优模型性能进行比较分析。

3.4.1 实验设置

数据集 我们使用 VQA2.0 数据集训练模型。按照 VQA 挑战中的划分标准将数据集分为 train/val/test 三个数据子集, 它们分别包含 8 万图像 +44.4 万问答对、4 万图像 +21.4 万问答对、8 万图像 +44.8 万问答对。答案包含“是否”、“数量”和“其他”三种类型, 图片均为从 MS-COCO 数据集中提取的真实场景。此外, 根据同时存在于 VQA2.0 和 Visual Genome 中的图片, 我们还使用了从 Visual Genome 中提取出 49 万个问答对, 用于增强训练集。

超参数设置 在文本中, 输入的图像特征 x_i 的维度为 2048, 因此 $X \in \mathbb{R}^{100 \times 2048}$ 。为衡量不同文本特征的性能表现, 我们使用三种 ELMO 的参数配置, 分别是 $ELMO_s/ELMO_m/ELMO_l$, 它们的参数量、LSTM 的隐层大小、输出大小、 $ELMo_k^{task}$ 大小见表3-1。我们使用一个线性层将 elmo 词向量统一转化为 512 维, 融合特征 $z \in \mathbb{R}^{1024}$ 。

表 3-1 三种 ELMO 的参数配置

| Model | Parameters (Millions) | LSTM Size | Output Size | ELMO Size |
|--------------|----------------------------------|------------------|--------------------|------------------|
| $ELMO_s$ | 13.6 | 1024 | 128 | 256 |
| $ELMO_m$ | 28.0 | 2048 | 256 | 512 |
| $ELMO_l$ | 93.6 | 4096 | 512 | 1024 |

多头注意力中的隐层维度为 512，头数 $h = 8$ ，因此每个头的隐层维度 $d_h = d/h = 64$ 。根据^[66]的建议，我们将答案词典大小设为 $N = 3129$ 。MCA 层数 L 设为 6；优化器 Adam 参数为 $\beta_1 = 0.9$ 、 $\beta_2 = 0.98$ ；学习率为 $\min(2.5te^{-5}, 1e^{-4})$ ， t 为训练的 epoch 数；batch 大小为 64。

3.4.2 剔除研究

3.4.3 实验结果分析

3.5 本章小结

第四章 基于知识库图嵌入的视觉问答模型

4.1 知识库概述

人类智能体通过学习和实践不断获取知识与经验，并能将习得的知识存储在记忆系统中，面对相关问题时能准确、快速地调用相关的知识和经验，完成识别和推理过程，成功解决问题。人工智能系统的终极目标便是能像人类一般快速、准确地解决未知问题，甚至超越人类的物理极限，实现范围更广、更艰深的任务解决。人类在真实世界中的学习是不断将非结构化的信息重构为结构化的知识的过程，知识库（KB）是一种包含常识和描述真实世界的事实的知识集，在不同的应用情景中有不同的内部结构。

知识库最早被应用于人工智能中的专家系统^[67]，专家系统是一种建立在知识库基础上，使用推理方法完成复杂推理过程，最终实现与人类专家同水平的决策能力的计算机系统，被广泛应用于医学诊断、分子结构推理、自然语言理解等领域。专家系统面向的专家任务需要特定领域的知识，这也使得知识库成为专家系统的核心之一。针对不同领域的任务构建知识的表达方式是困难的，因为专家知识可能是不精确的，同时要从知识库中获取答案的过程依赖于人工的制定复杂的规则，知识库精度和人力成本等因素制约了专家系统在更多领域的应用。

知识库也被应用于在自然语言处理的任务，例如机器翻译和文本问答。知识库中的本体包含某个领域中的各种概念和概念间的关系，本体在机器翻译中可作为知识源^[68]。语言学中的多义词在不同的语境中被解释为不同的含义，人类能根据上下文语境的不同选择出最恰当的词语，但对于机器翻译系统便是一大难题。当机器翻译系统能够获得足够多的本体作为知识源时，能较好地解决多义词的解释问题，从而得到更加准确的翻译结果^[69]。

文本问答系统在早期作为专家系统的交互界面，在之后的发展中逐渐独立出来成为自然语言处理的一个分支，文本问答系统根据给出的文本问题，从文本知识库中提取答案，此时的文本知识库往往是文本组成的文档，还未使用资源描述框架（RDF）的结构化数据。大多数文本问答系统都采用相对标准的结构：根据问题文本建立查询、利用信息提取方法（IR）确定可能包含答案的文章位置、进一步确定答案所在的片段，这种架构下不使用任何与答案相关的额外知识^[70]。Hermjakob 等人提出了将手写规则和概念本体相结合的问答系统——Webclopedia^[70]。Webclopedia 由对输入问题进行句法和语义的解析的问题解析模块、用于文档查询的查询模块、用于获得与答案相关的文档的信息提取模块、片

段解析模块、答案匹配模块和答案生成模块构成。系统在多个模块中使用了知识库提高精确度，在问题解析过程中使用了语言知识库——由 30000 个节点的概念层级、140 个问题/答案类型和词库组成，帮助系统确定问题的句法结构；在查询模块中使用了 WordNet^[71] 扩展与问题关键词关联的信息；在答案匹配模块中也使用了常识和事实知识库，系统的架构如图4-1。

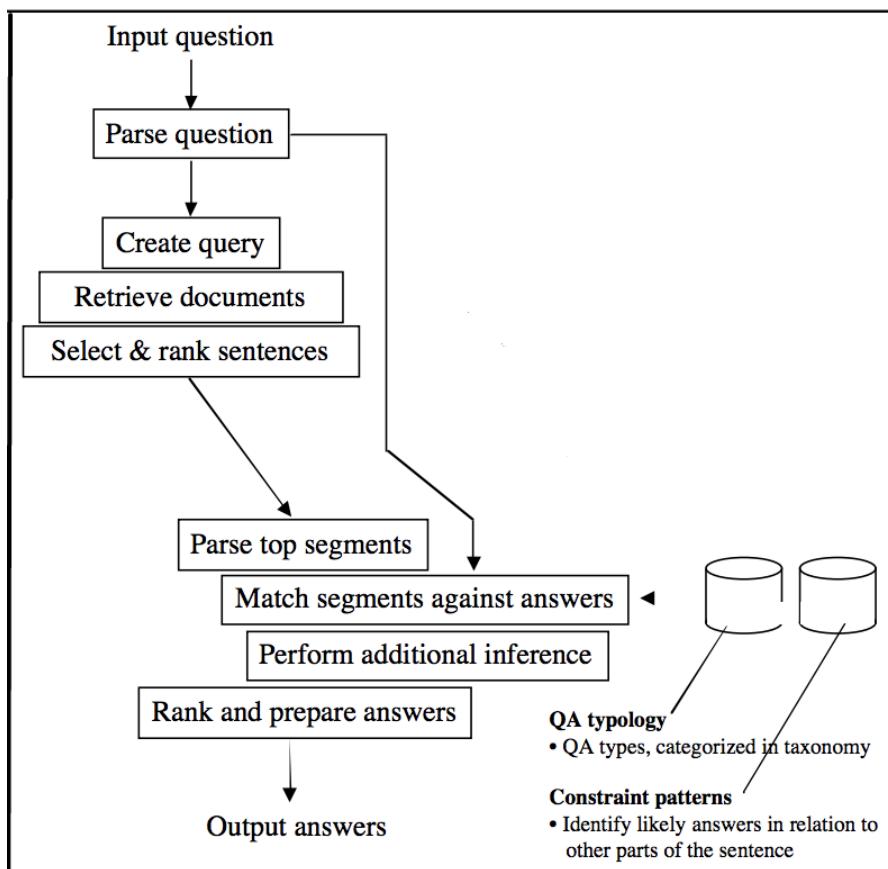


图 4-1 Webclopedia 系统架构

应用信息提取技术（IR）的问答系统有一个非常明显的缺点——只能根据问题确定答案相关的文章或者段落，不能给出更为直接的答案。为解决这种缺陷，研究人员探索了更多的方法。

Burke 等人一改通常的从文章中提取答案的方式，先将被频繁问到的问题（FAQ）以“问题-答案”对的形式存储为知识库，再从新问题中寻找与知识库匹配程度最高的“问题-答案”对，进而获得答案^[72]。在此方法中最核心的步骤是对新旧问题之间的匹配，为了使匹配的问题之间的语义相似度最大，系统还使用了 WordNet^[71] 的语义知识，WordNet 能提供词语和其同义词集合、同义词集合之间的关系，因此能避免一些匹配过程中的歧义错误，提高匹配的准确度。这里以“匹

配”为核心思想的算法最大的障碍是常见问题集的容量、深度和广度问题，因此通常对于范围较小的场景而言，才能实现较好的匹配准确度。Rinaldi 等人提出一个专门针对技术领域的基于知识的问答系统 ExtrAns^[73]。ExtrAns 以技术手册为知识库，将问题文本和知识库都转化为一种称为“最小逻辑形式”（MLF）的语义表达，并通过逻辑证明提取出答案。

随着资源描述框架（RDF）在构建知识库的兴起，知识库也由原来的文档形式转化为冗余更小、可扩展性更强、易用性更强的结构化数据库。面对由于互联网技术的普及带来的海量网页、文章、超文本、图片等多种模态的资源，研究者们对信息的整合进行了探索^[74-78]，语义网和相关技术的出现促进了大尺度知识库的发展，出现了 DBpedia^[48]、OpenIE^[79]、Yago^[80]、Freebase^[81]、Wikidata^[82] 等多种含有常识和特定领域知识的知识库，这些配置灵活、结构统一且语义丰富的外源知识库也促进了基于知识库的视觉问答方法的兴起。

Yago 知识库通常由人工和自动化提取两种方式构建得到，对比这两种不同构建方式，自动化提取的知识库往往质量较低，容易包含错误信息，而人工构建的知识库能满足较高的精度要求，但由于人工构建的成本较高，因此此类知识库有数据容量受限、构建周期长、内容老化快等缺陷。

Suchanek 等人结合 Wikipedia 文章的广博性和 WordNet 优秀的语义分类，提出了自动化生成本体的知识库 YAGO^[80]。Wikipedia 的文章对某个话题或概念进行详细的多角度说明，同时大多数文章都归属于一个或者多个类别，类别页面既包含了大量实体和概念，可以作为知识库中的本体，同时类别页面也隐含着概念之间的平行关系和所属关系，这能提供一定的结构关系。YAGO 利用 Wikipedia 目录页面提取出其中的实体和实体之间的关系，同时结合 WordNet 中概念的清晰层次关系，实现了 97% 的准确率。初始版本中涉及 90 万个实体和 500 万个实体之间的关系。

YAGO 被设计为可扩展的知识库，能够结合特定领域的知识源或是从网络上提取得到的信息构建领域相关的知识库，因此之后的研究者也在此基础上进行了多种的扩展。YAGO2 在 YAGO 基础上引入 GeoNames——包含超过 700 万个地点信息，在“实体-关联”的表示方法中加入了时间和空间维度，不仅能丰富事实的准确性，还能反应出实体在时空层面的变化^[83]。YAGO3 构建了一个多语言的知识库^[84]。

DBpedia Wikipedia 是由非盈利组织维基媒体基金会（Wikimedia Foundation）构建的世界上最大的多语言的开放性网络百科全书，其通过文章的形式对词条进行多方面的介绍，文章中包含大量的结构化信息，例如文字、信息框模板、分类

信息、图片、地理坐标信息、超链接等，这些多模态的信息能丰富知识的多样性，并且建立知识的关联。但作为网络应用，Wikipedia 的搜索能力和其他网络应用一样，只能满足关键词的搜索，这种状况大大的降低了知识之间的关联和价值，同时因为其作为大规模协同性内容编辑平台，文章内容也难以避免的出现数据矛盾、不一致的分类和错误。

Auer 等人为了充分挖掘 Wikipedia 中已有的人类知识，并构建知识结构，提出了 DBpedia 知识库^[48]。Wikipedia 为实现统一的文章风格，因此在文章编辑中镶嵌了一些信息框模板，如图4-2。DBpedia 利用信息框提取算法检测信息框模板，

| <pre> {{infobox City Korea full_name=Busan Metropolitan City image=[[Image:Haeundaebeachbusan.jpg 250px Haeundae Beach, Busan]] rr=Busan Gwangyeoksi mr=Pusan Kwangyōksi hangul=부산 광역시 hanja=釜山廣域市 short_name=Busan (Pusan; 부산; 釜山) population=3,635,389 ... area=763.46 km² government=[[Metropolitan cities of South Korea Metropolitan City]] divisions=15 wards (Gu), 1 county (Gun) region=[[Yeongnam]] dialect=[[Gyeongsang Dialect Gyeongsang]] map=[[Image:Busan map.png Map of South Korea highlighting the city]] }} </pre> | <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="2" style="text-align: center;">Busan Metropolitan City</th> </tr> </thead> <tbody> <tr> <td colspan="2" style="text-align: center;"></td> </tr> <tr> <th colspan="2" style="text-align: center;">Korean name</th> </tr> <tr> <td style="padding: 2px;">Revised Romanization</td> <td style="padding: 2px;">Busan Gwangyeoksi</td> </tr> <tr> <td style="padding: 2px;">McCune-Reischauer</td> <td style="padding: 2px;">Pusan Kwangyōksi</td> </tr> <tr> <td style="padding: 2px;">Hangul</td> <td style="padding: 2px;">부산 광역시</td> </tr> <tr> <td style="padding: 2px;">Hanja</td> <td style="padding: 2px;">釜山廣域市</td> </tr> <tr> <td style="padding: 2px;">Short name</td> <td style="padding: 2px;">Busan (Pusan; 부산; 釜山)</td> </tr> </tbody> </table> | Busan Metropolitan City | |  | | Korean name | | Revised Romanization | Busan Gwangyeoksi | McCune-Reischauer | Pusan Kwangyōksi | Hangul | 부산 광역시 | Hanja | 釜山廣域市 | Short name | Busan (Pusan; 부산; 釜山) |
|--|--|-------------------------|--|--|--|-------------|--|----------------------|-------------------|-------------------|------------------|--------|--------|-------|-------|------------|-----------------------|
| Busan Metropolitan City | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | |
| Korean name | | | | | | | | | | | | | | | | | |
| Revised Romanization | Busan Gwangyeoksi | | | | | | | | | | | | | | | | |
| McCune-Reischauer | Pusan Kwangyōksi | | | | | | | | | | | | | | | | |
| Hangul | 부산 광역시 | | | | | | | | | | | | | | | | |
| Hanja | 釜山廣域市 | | | | | | | | | | | | | | | | |
| Short name | Busan (Pusan; 부산; 釜山) | | | | | | | | | | | | | | | | |

图 4-2 Wikipedia 的信息框模板和加载效果

并且提取出关键的信息，再将信息转化为资源描述框架（RDF）的三元组结构，从而将 Wikipedia 的文章内容转化为机器可读的结构化信息。最初版本的 DBpedia 知识库包含关于 195 万实体的信息，实体内容包括人物、地点、音乐专辑和电影，除了实体外还包含 65.7 万个图片链接、160 万个外部网页链接、18 万个其他资源描述框架（RDF）数据库、20.7 万个 Wikipedia 目录和 7.5 万个 YAGO 类别^[80]。随着开放社区的数据丰富，2016 年推出的版本中已经包含 6600 万实体，实体的类型扩充了视频、游戏、组织、物种和疾病^[85]。资源描述框架的三元组数据量也从 1 亿增长到 130 亿之多。

为了增强 DBpedia 的数据易用性，Auer 等人提供了三种数据获取方式：链接数据、SPARQL 协议和可下载的 RDF 文件。链接数据通过 HTTP 协议获取发布与互联网上的 RDF 数据，提供给语义网络浏览器、语义网路爬虫和语义网络查询客户端访问^[86]。SPARQL 是专门针对资源描述框架的查询语言，通过 SPARQL 终端

向`http://dbpedia.org/sparql`发送查询指令，DBpedia 知识库会返回相应的查询结果。可下载的 RDF 文件包含序列化的 RDF 三元组数据，DBpedia 将整个数据库按照数据的类型分为众多子数据集，例如，文章目录集、目录标签集、地理坐标集、图像集等。

知识库的内容多样性、易用性和大体量为 DBpedia 应用提供了良好的基础设施，因此一些自然语言问答和交互的应用都选择建立在 DBpedia 丰富的知识之上。NLI-GO DBpedia 是一个针对通用自然语言交互的应用程序，程序可以接受自然语言问题，并通过 SPARQL 查询 DBpedia 知识库，给出答案，实际上这就是基于 DBpedia 的文本问答系统^[87]，类似的还有款基于 DBpedia 的聊天机器人——DBpedia Chatbot。许多基于知识库的视觉问答研究也选择了数据更加准确的 DBpedia^[16-18]。

OpenIE 应用于构建知识库的信息提取技术（IR）往往需要人为构建大量手写规则，并选择合适的语料库，当已有的提取模型面对全新领域的语料库时，需要重新编写提取规则或者标注数据，这种系统在面对快速迭代和具有丰富多样性的互联网数据时，便会遇到自动化程度低、语料库异质性和效率问题。

为了节省信息提取过程的自动化程度，并能大范围应用于不同领域，Banko 等人提出了一种能自主学习不同语料库的信息提取模型——开放信息提取技术（Open IE）^[79]。Open IE 以语料库为输入，通过内部算法对语料库中的语句进行一次遍历，最终提取出语句中蕴含的（实体，关系，实体）三元组数据，在整个过程中不需要人工参与，因此可以应用于不同领域知识库的构建。

Banko 等人还提出了一种应用高扩展性 Open IE 模型的系统 TEXTRUNNER。TEXTRUNNER 由自监督学习器、单通道提取器、基于冗余的评估器三个主要模块构成。自监督学习器以小的语料样本作为训练集，首先使用语句解析器从样本中粗略地提取出（实体，关系，实体）的三元组数据，再对提取出的内容进行标注，标注为“可信”和“不可信”两种标签，将带有标签的数据作为朴素贝叶斯分类器的训练样本。提取器遍历整个语料库，提取出所有可能的三元组数据。对于同一个句子，提取器能生成一个或多个三元组数据，这些数据将被送入学习器训练得到的分类器中，保留所有分在“可信”类别的数据。在得到所有提取出的知识后，评估器融合相同的数据，计算不同的数据的数量。基于以上统计，评估器对每一个三元组数据分配一个用于判断知识正确性的概率值，其中的假设是，如果从多个语句中提取出相同的知识，那么该知识拥有较高的可信度。

在实验阶段，TEXTRUNNER 从包含 1.3 亿个句子的 900 万个网页中提取出 6000 万个三元组数据，平均每个句子提取出 2.2 个关系数据。通过数据过滤、随机

抽取、人工判定等方式，作者对提取数据的完整性和正确性进行了概率评估，过滤后的数据包含 1130 万个三元组数据，其中 780 万的数据被评估为“格式正确”且概率标签在 0.8 以上，80.4%“格式正确”的数据通过人工评估被认定为正确的，从实体间的关系看，“格式正确”的数据中反映抽象事实的占 86%，其中 77.2% 是正确的；反映具体事实的占 14%，其中 88.1% 是正确的，如图所示 4-3。

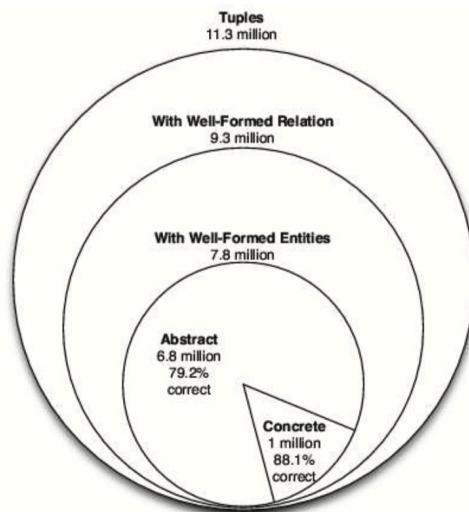


图 4-3 TEXTRUNNER 在实验环境下知识提取的正确率

Wu 等人在 TEXTRUNNER 的基础上提出了 WOE 开放信息提取系统^[88]。WOE 改进了自监督学习方式用于构建提取器，TEXTRUNNER 在提取过程中使用解析器直接从语料库中提取（实体，关系，实体）的三元组数据，而 WOE 则先从 Wikipedia 的信息框中提取“属性-值”对，再使用匹配器从文章中找到包含文章主语和“属性-值”对的句子作为语料库中的训练数据。随后测试了两种解析方法的提取器：WOE-parse 和 WOE-pos，WOE-pos 使用和 TEXTRUNNER 类似的解析方法，根据简单的词性标签，从语料库中的句子解析出（实体，关系，实体）的数据，WOE-parse 则选择更复杂的依赖解析树，希望能再复杂长句的解析中得到更好的精确度。

开放信息提取系统会对每个输出的三元组数据给定一个置信度，如果给定一个置信度的下限，高置信度的数据被保留，低置信度的数据被过滤，此时可以通过精确度和召回率测试系统的性能。精确度是指保留的数据中正确的数据所占比例，能反映整体精确度的平均水平。召回率是指保留的数据中正确的数据占所有正确数据的比例，能反映正确的数据在不同置信度的分布情况。

实验分析显示，因为使用了更友好的训练数据，WOE-pos 在精确度上更优于 TEXTRUNNER，而 WOE-parse 在解析树的帮助下实现了最好的性能，特别是在

召回率上。

Fader 等人在分析 TEXTRUNNER 和 WOE 的结果之后发现，不连贯提取和无信息提取两种错误频繁出现。不连贯提取是指被提取的关系语句由多词组成，但语义不连贯而无意义。无信息提取是指提取内容忽略了句子的关键信息，例如，“父亲对母亲做出承诺”，系统返回无信息的（父亲，做出，承诺）而不是（父亲，做出承诺对，母亲）。以上的两种错误都是由系统不能提取出具有完整句法结构的关系语句造成的，Fader 等人在 Open IE 系统中引入了一定的句法限制，提出了 REVERB 开放信息提取系统^[89]。30% 的 REVERB 提取数据的概率标签在 0.8 或更高，相较起 TEXTRUNNER 的 0.13%，在精确度上实现了越阶式的增长，不连贯提取和无信息提取的错误率也大幅减少。

Freebase Bollacker 等人试图结合一般数据库的扩展和 Wikipedia 等百科全书的多样性，提出了 Freebase 数据库^[81]。Freebase 和其他常用的知识库相同，使用资源描述框架的三元组形式结构化真实世界的知识，但同时继承了网络百科全书的开放和协同的思想，所有的内容创造和维护都由社区成员协作完成。Freebase 存储的元组数据超过 1 亿 2500 万条，超过 4000 种类型和 7000 种属性，允许使用查询语言通过 HTTP 协议获取数据。

2014 年，Google 宣布关停 Freebase 并将数据迁移至 Wikidata。

Wikidata Wikidata 是为了更高效地开放使用和管理 Wikipedia 文章中数据而提出的协同知识库^[82]。由于 Wikidata 的出发点是希望通过大规模协同的方式构建知识库，因此 Wikidata 的数据具有开放性、多版本共存、多语言、易用性和持续更新的特性。Wikidata 向所有用户提供数据扩展和编辑的权限；Wikidata 为保证模糊数据的存疑性，相互之间有冲突的数据被同时展示；考虑到数字、日期、坐标等语言无关的数据内容，Wikidata 与 Wikipedia 相同设计为多语言版本；Wikidata 数据被组织成 Json、RDF 的形式发布于网络，通过网络服务能够轻松获取数据；社区成员的持续更新能保持 Wikidata 的时效性。

Wikidata 数据的基本单元被称为项目（Item），每个项目包含名称标签、“Q+ 数字”的项目编码、描述、别名、和声明。声明中包含一系列属性和相应的值，用于详细描述项目的特点，项目页面如图4-4。

label: Douglas Adams (Q42)

description: English writer and humorist
Douglas Noël Adams | Douglas Noel Adams
[In more languages](#)

property: educated at

rank: 1

statement group: Statements

value: St. John's College

qualifiers: end time (1974), academic major (English literature), academic degree (Bachelor of Arts), start time (1971)

opened references: 2 references

reference URL: http://www.bnfb.com/people/731/000023662/
original language of work: English
retrieved: 7 December 2013
publisher: NNDB
title: Douglas Adams (English)

collapsed reference: Brentwood School

value: Brentwood School

qualifiers: end time (1970), start time (1969)

add reference: + add reference

add statement: + add (statement)

图 4-4 wikidata 项目页面

项目之间通过有向无环图的方式构成，节点代表项目，有向线段代表项目之间的关系，如图4-5。截止到 2018 年，Wikidata 已拥有超过 5000 万个项目。

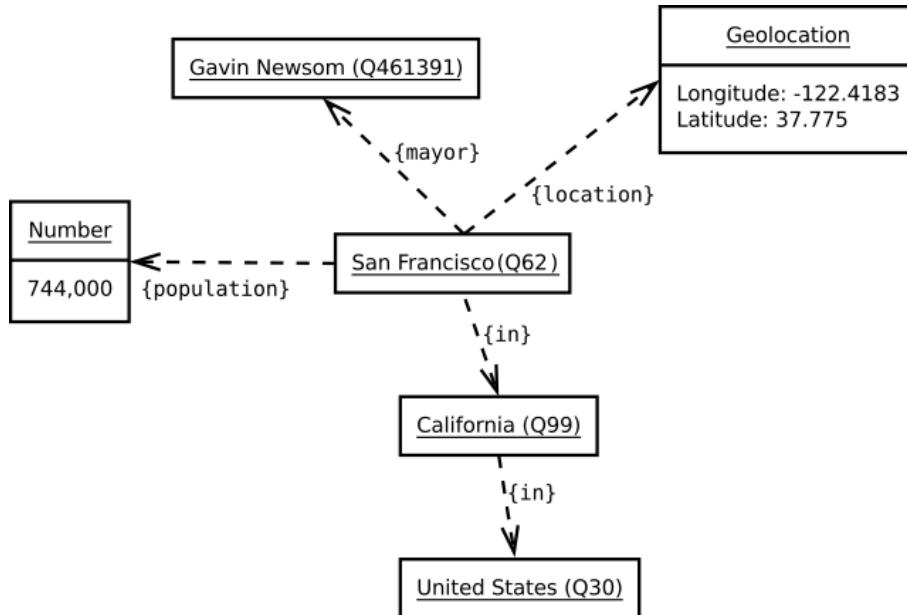


图 4-5 wikidata 项目之间的有向无环图结构

Wikidata 于 2012 年提出，相较起以往的知识库，开放性更强，限制也更少。对比 YAGO 和 DBpedia，Wikidata 不是从 Wikipedia 的目录或者信息框中提取信息，相反 Wikidata 被社区成员独立构建，并为 Wikipedia 作为知识源，数据被链接到

Wikipedia 文章中。对比 Freebase 将对象按类型划分的方式，Wikidata 支持对所有对象赋予任意属性。

4.2 KBSN 模型

使用 RDF 的数据表达方式，实体和实体之间通过属性建立了联系，这些有丰富语义的实体之间相互联系，构成了知识库。通过可视化的方式，实体作为节点，属性或者实体关系为边，知识库可以以图的形式呈现，因此知识库也被称为知识图谱。

正如绪论中提到的，知识库因其丰富的知识存储量、多样化的知识内容、复杂的知识关联、结构化的数据存储方式，可以作为问答系统或者其他信息检索任务的重要基础。目前使用知识图谱的主流方式是通过 SPARQL 等结构化查询语言对知识库中的内容进行精准的检索和提取，这种方式人为地建立查询规则、设计相应的知识库存储。

在基于知识库的视觉问答模型中，知识库的使用方式大致分为两种。一种方式为知识库查询类，依照主流的知识库查询的思路，模型提取图片的实体、将实体映射到知识库、转化自然语言为查询语句、查询知识库^[17,18]。这些模型依靠精准的查询语句，对于预先设定好的模板问题能实现优于基线模型的准确率，然而却面临着问题模板设计成本高、数据集难构建、模型泛化能力差等缺点。

另一种方式为联合嵌入类，这种方式不用设计复杂的查询语句，而是将知识库的文本信息转化为额外的特征向量，并联合图像特征和问题特征一起训练。这种方式能省去问题模板和查询语句设计的人工成本，并将模型在更大规模的开放性数据集进行训练。然而，此前的模型却仅仅使用知识库中单个节点的文本信息，例如论文^[16]根据从图像中预测的属性生成 DBpedia 查询，得到相关属性的“comment”本文内容，再将这种成段的文本信息转换为固定的特征向量，作为由知识库提供的额外特征与其他两种模态的特征融合。在这种方式中，模型虽然引入了额外的特征，试图提高表征能力，但是这种额外特征仅仅局限于单个节点，因此必然损失了节点互联形成的结构关系，而这种结构关系正是知识库的核心——通过多种关系连接而组织起来的具有丰富语义表征能力的实体网络。

为了利用知识库中关联数据的结构信息，我们在 N-KBSN 模型的基础上，引入使用图嵌入表示的知识库，提出了 KBSN 模型。KBSN 模型使用了 N-KBSN 模型的问题文本和图像特征提取模块、自注意力和引导注意力模块，而在特征融合时引入了知识库的图嵌入。

知识库的图嵌入是 KBSN 模型有别于其他基于知识库的视觉问答模型的创新

之处，其背后的思路为：先从图像和问题文本中识别出核心概念，将核心概念映射为知识库中的核心实体，通过剔除核心实体外无关的实体和链接形成以核心实体为中心的子图，再将各个子图转换为图嵌入，最后子图嵌入融合为图嵌入，以此作为额外特征。

按照以上的思路，知识库的图嵌入由子图提取模块和子图嵌入模块两个主要部分组成。子图提取模块的作用是完成从图像和问题文本到知识库的映射。具体来说，子图提取模块包含“图像-知识库映射”和“文本-知识库映射”。“图像-知识库映射”使用 Faster R-CNN 预测得到图像中包含的物体，再使用 SPARQL 查询知识库，得到图像相关的核心实体，“文本-知识库映射”使用 DBpedia Spotlight^[90] 模型识别、整合问题文本，得到问题相关的核心实体。需要指出的是，在此，我们不是使用完整的 DBpedia 知识库，而是根据问答这种任务类型的特点，挑选出特定的数据子集构成实验知识库。

子图嵌入模块是将提取得到的子图映射为图嵌入。具体来说，我们首先使用实验知识库训练 TransE 模型，得到实验知识库的嵌入表示。然而将子图提取模块输出的子图的节点和边都映射为向量，最后经过图神经网络获得子图的嵌入表示。

最后将子图的嵌入表示、图像特征、文本特征三者融合，分类得到答案，KBSN 的基础架构如图4-6。

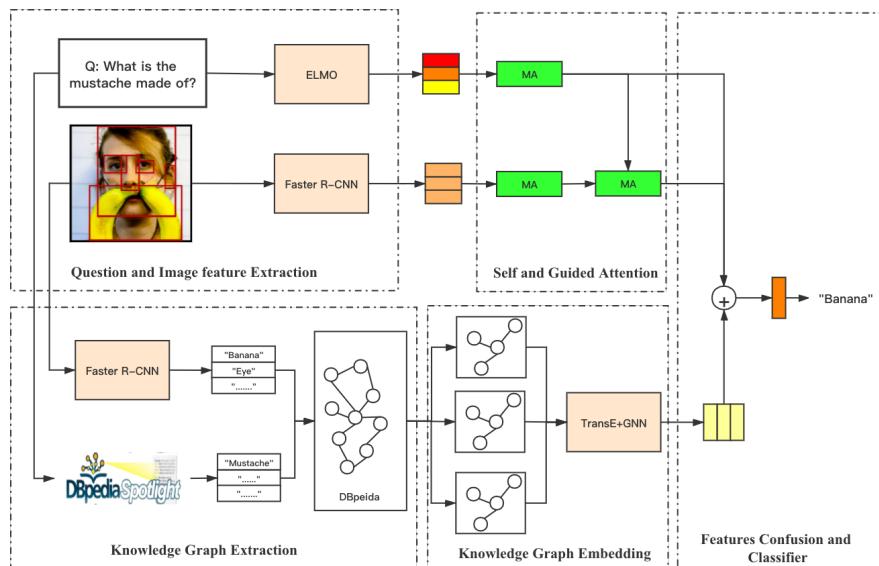


图 4-6 KBSN 的基础架构

4.2.1 知识库子图提取

对于基于知识库的视觉问答任务而言，准确的实现自然语言和图像中涉及的实体到知识库实体的映射是至关重要的，一方面能够大大地减少知识库中无关信息的噪声干扰，提高精确度，另一方面准确的映射能够极大的减少计算冗余，提高运行速度。

在知识库子图提取的准确性和计算效率综合的考量下，我们首先以 DBpedia 为基础，收集了部分子数据集组成实验知识库，再使用在 N-KBSN 模型中相同的 Faster R-CNN 从图像中识别出关键实体，再设计 SPARQL 查询语句从实验知识库中提取出图像相关的子图。另一方面，对于问题文本中的核心实体，我们直接使用 DBpedia Spotlight 完成从文本到 DBpedia 节点的映射，进而提取出问题相关的子图。

需要注意的是，针对图像的物体识别是使用的 N-KBSN 的 Faster R-CNN，但是在 N-KBSN 中，是将所有区域的图像特征融合作为图像特征，而此处则加上分类层，使用 softmax 预测并输出各个区域的类别信息。由于使用的 Faster R-CNN 已经在之前的章节详细介绍了，本节将省略面向图像的子图提取，重点介绍面向文本的子图提取。

4.2.1.1 知识库收集和分析

在本文中，因为 DBpedia 丰富的实体及其属性，并且相对规范和统一的数据内容，我们使用 DBpedia 作为提供额外特征的知识库。然而由于 DBpedia 从众包的 wikipedia 提取得到，因此知识库中除去我们关心的和实体语义高度相关的属性外，还保留着一些基于语义网思想的属性，例如用于连接其他知识库实体的外链、参考引用的外链、主页地址、图片链接、未经处理的 infobox 属性等，除此之外，完整数据集中还包含包括英语在内的多语言版本，而以上这些信息对于回答开放性问题帮助很小，因此在本文的研究中，我们甄选了 DBpedia 数据集中“语义高度相关”的子数据集构成实验知识库，具体的数据集和其简要描述如表4-1，知识库的参数统计如表4-2。

表 4-1 实验知识库包含的 DBpedia 数据子集及其描述

| DBpedia 数据集 | 描述 |
|--------------------------|-----------------------------|
| instance_types_en | 连接实体及其类型 |
| labels_en | 实体标签 |
| mappingbased_literals_en | 连接 object 为 literal 的高质量的谓语 |
| mappingbased_objects_en | 连接 object 为对象的高质量的谓语 |
| persondata_en | 和 person 相关的信息，例如出生日期等 |

表 4-2 实验知识库的参数统计

| 统计参数 | 数量 |
|------|------------|
| 三元组 | 59,998,758 |
| 类 | 426 |
| 实体 | 5,377,081 |
| 主语 | 14,556,042 |
| 属性 | 1,377 |
| 宾语 | 19,495,719 |

正如表4-1所示，实验知识库只包含了英语版本的知识库，但值得注意的是，本文提出的模型结构同样适用于其他语言类型，扩展到多语言的应用只需要将数据集和知识库替换为指定语言版本即可。实验知识库包含完整的类别信息、完整的实体标签和高质量的属性信息，足够应对绝大部分的常规问题，例如 VQA2.0 数据集中涉及的对象和属性都存在于实验知识库中。

我们使用 virtuoso opensource 将上述数据集加载成一个命名图 (<http://dbpedia.org>)，并使用本地服务器提供 SPARQL 应用交互接口，以便后续提取子图。

4.2.1.2 面向文本的子图提取

和基于知识库的问答任务相似，基于知识库的视觉问答任务中的问题文本中并不是每个词语对于答案的得出都起着同等重要的作用。例如对于问题 “Is this book writen by Ernest Miller Hemingway”，人类回答者可以忽略句式中的谓语“is”、代词“this”，而将句子缩减为 (book, writen by, Ernest Miller Hemingway)。这种去除了辅助句法和语法结构的词语而得到的缩减形式便能够反映问题的

关键信息，而其中的”book” 和”Ernest Miller Hemingway” 这类名词在知识库中，被称之为命名实体 (named entity)，在 DBpedia 中是以类似于 *DBpedia : book* 和 *DBpedia : Ernest_Miller_Hemingway* 这种 URI 的节点形式存在。

对于这些存在于文本中的命名实体的提取便是本小节中面向文本的子图提取的关键步骤之一。而在命名实体的提取中，消除歧义是非常重要的。同一个单词在不同的语境下表达不同的意思，如果不能根据语境正确地判断出单词的特定语义，那么句义的理解就可能偏移，甚至意思完全无法理解。在文本-知识库映射中则体现为，同一个单词在不同语境下对应不同的 DBpedia 资源，例如”Washington”可以同时对应 *DBpedia : George_Washington* 和 *DBpedia : Washington,_D.C.*，前者指向“乔治-华盛顿”，一个人，而后者则指向“华盛顿特区”，一个地方，两者的含义千差万别。

为了实现较为准确的命名实体识别，我们使用 DBpedia Spotlight 模型^[91] 实现文本-知识库映射。包括“人物”、“地点”、“组织”这种常见的类别，DBpedia Spotlight 能够实现 272 类 DBpedia 资源的识别，因此能够很好的识别绝大部分问题中涉及的实体。我们还可以通过针对数据集的特点使用针对性的配置，进一步提高实体的识别准确率。

DBpedia Spotlight 模型主要由三个阶段实现，短语识别阶段从输入的自然语言句子中提取出可能存在 DBpedia 资源的短语；候选实体筛选阶段将前一阶段得到的一系列短语映射到 DBpedia 资源，形成候选实体列表；消除歧义阶段根据短语的上下文语境，从候选实体列表中挑选出最佳的 DBpedia 资源，完成从文本-知识库映射。

短语识别阶段首先通过字符匹配算法从句子中提取词典中包含的短语，再对每个短语自动标注词性，并且去除词性为动词、形容词、副词、介词，剩下的短语作为候选短语。

候选实体筛选阶段根据 DBpedia 的 Disambiguation 数据集——包含和特定短语容易混淆的所有其他短语，囊括每一个候选短语的歧义形式的 DBpedia 资源，例如对于候选短语”Washington”，*DBpedia : George_Washington* 和 *DBpedia : Washington,_D.C.* 都被加入候选实体列表，以便下一阶段的使用。这一阶段实现了由短语到 DBpedia 资源的映射，并且为了提高结果的准确性，在这一阶段只进行最小化的筛选，尽量多的包含候选实体。

消除歧义阶段使用生成概率模型^[92]，根据短语的上下文信息，计算短语和实体匹配的概率，再依照概率阈值得到短语匹配的 DBpedia 资源，其中短语也称为“实体指称”。假定短语 s ，上下文 c ，每个实体 e 和短语匹配的概率可以根据以下公

式得到,

$$P(e, s, c) = P(e)P(s|e)P(c|e) \quad (4-1)$$

其中, $P(e)$ 表示实体出现的概率, $P(s|e)$ 表示以短语 s 指代实体 e 的概率, 因为多种不同的短语可以指代同一个 DBpedia 资源, 例如短语”Washington”和”George_Washington”都可以指代 DBpedia : George_Washington, $P(c|e)$ 表示实体在特定语境出现的概率。通过最大似然概率, 得到最匹配的实体 e , 即

$$e = \operatorname{argmax} P(e, s, c) \quad (4-2)$$

假定一个包含 M 个实体指称的 wikipedia 数据集, $P(e)$ 可以使用以下公式计算,

$$P(e) = \frac{\operatorname{count}(e)}{|M|} \quad (4-3)$$

其中 $\operatorname{count}(e)$ 表示指向实体 e 的实体指称的数量。 $P(s|e)$ 的公式为,

$$P(s|e) = \frac{\operatorname{count}(e, s)}{\operatorname{count}(e)} \quad (4-4)$$

对于短语 s , 它的上下文 c 可以使用一个单词窗口来框定, 在本文中, 我们设定窗口大小为 50。假定上下文 c 包含 n 个单词 $t_1 t_2 \dots t_n$, 那么 $P(c|e)$ 的公式为,

$$P(c|e) = P_e(t_1)P_e(t_2)\dots P_e(t_n) \quad (4-5)$$

其中 $P_e(t)$ 表示单词 t 出现在实体 e 的上下文的概率, 计算公式为,

$$P_e(t) = \lambda P_{e-ML}(t) + (1 - \lambda)P_{LM}(t) \quad (4-6)$$

$$P_{e-ML}(t) = \frac{\operatorname{count}_e(t)}{\sum_t \operatorname{count}_e(t)} \quad (4-7)$$

其中 $P_{e-ML}(t)$ 是 $P_e(t)$ 的最大概率, $P_{LM}(t)$ 是在 wikipedia 数据集上计算得到的通用语言模型。

为了防止短语都连接到“空实体”, 同样我们需要计算“空实体”的得分 $P(NIL, s, c)$, 使用以下公式分别计算 $P(NIL)$ 、 $P(s|NIL)$ 和 $P(c|NIL)$,

$$P(NIL) = \frac{1}{|M|} \quad (4-8)$$

$$P(s|NIL) = \prod_{t \in S} P_{LM}(t) \quad (4-9)$$

$$P(c|NIL) = \prod_{t \in C} P_{LM}(t) \quad (4-10)$$

而所有得分小于 $P(NIL, s, c)$ 的实体都会被剔除。

在计算得到实体的得分之后，根据得分的高低排序便可以得到最匹配的 DBpedia 资源，完成文本-知识库映射。

4.2.2 知识库子图嵌入

针对本文提出的 KBSN 模型，我们将从问题文本中提取得到的 DBpedia 实体视为核心节点，核心节点从词性的角度看，绝大多数都为名词，从句义的整体来看代表整个句子的核心概念，例如问题“Is there snow on the mountains?”中，我们提取出实体 *DBpedia : Snow*，并且提取出以 *Snow* 为核心节点的子图。子图中包含大量语义高度相关的属性能作为丰富概念的不同语义层次，例如其属性 *Subject* 为 *Category : Snow*——表示其分类，属性 *seeAlso* 为 *Blizzard*——表示其同义概念。然而图结构的知识子图并不能很好的计算处理，因此我们使用分布式表示将知识子图中的实体和关系转化为低维向量。这样做的优点有以下几点：

- 1) 计算的便利性，向量化的节点能够方便的衡量节点的差异和相似度，显著提升计算效率。
- 2) 实现多模信息的融合，KBSN 模型中涉及图像特征、文本特征和知识子图特征三种不同模态的数据结构。知识子图的嵌入能够很好得融合入另外两种特征，这种统一的特征表达方式能够也是适应目前的计算框架——以多维向量为基础的计算方式。
- 3) 便于知识库的扩展，文本使用 DBpedia 为主要的知识库，然而对于其他主流的知识库，如 Freebase，WordNet 等，使用的实体和属性名称不尽相同，这会限制模型迁移。而使用分布式表示能够将不同的知识来源映射到同一个语义空间，从而建立统一的表示空间，实现不同知识库的相互适应，提高模型的扩展能力。

文本将使用 TransE^[93] 模型为基础，实现对子图的嵌入表示。TransE 模型的思路来源于词向量中呈现出的词向量聚集和向量空间的平移不变性。具体来说，在词嵌入空间中具有相似语义的词表示呈现出聚集情况，例如向量 $e(German)$ 和 $e(France)$ 等国家名称距离接近；平移不变性表现为 $e(king) - e(queen) \approx e(man) - e(woman)$ 。前者说明有效的嵌入能够表征词的语义相似性，后者说明向

量空间中存在一些固定关系能够连接不同的词嵌入。而在知识库中实体之间是通过显性的关系连接构成一个三元组，这种显性的关系也许能帮助找到一个好的图嵌入方式，使得向量空间中存在和显性关系暗合的隐藏关系，而这种隐藏关系在 TransE 中被称为“翻译”。假定 E 为实体的集合， R 为关系的集合，训练集为 $S = \{(h, r, t)\}$ ，其中三元组 (h, r, t) 中 h 表示“头实体”， r 表示“关系”， t 表示“尾实体”，它们的嵌入向量分别用 l_h 、 l_r 、 l_t 表示。TransE 希望得到的向量存在以下关系，

$$l_h + l_r \approx l_t \quad (4-11)$$

公式可以看做向量 l_h 经过关系 r 翻译后得到了 l_t 。

为了学习到符合以上公式的向量，模型使用 $d(h + r, t)$ 计算两个向量的差异度，函数 d 使用 L1 或者 L2 距离计算公式。模型的思路为如果对一个正确存在的三元组的 h 或者 t 替换成其他的实体，那么新的差异度 $d(h^n, r, t^n)$ 数值应该尽量大，以体现新三元组的错误性。因此 TransE 使用以下损失函数，

$$Loss = \sum_{(h, r, t) \in S} \sum_{(h^n, r, t^n) \in S^n} |\gamma + d(h + r, t) - d(h^n, r, t^n)| \quad (4-12)$$

其中， γ 为正确的三元组和错误三元组差异度之间的距离超参数。

$$S^n = (h^n, r, t) | h^n \in E \cup (h^n, l, t) | t^n \in E \quad (4-13)$$

S^n 表示替换了头实体或者尾实体的三元组的集合。

比起以往的模型，TransE 参数较少，计算复杂度低，但能够直接建立实体和关系的复杂语义联系，并且在大规模的知识库上依然有较好的表现，因此文本将使用 TransE 将知识库子图中的实体和关系转化为向量表示。

本文 TransE 的参数设置为，向量维度 $k = 50$ ，随机梯度下降的学习率 $\lambda = 0.01$ ， $\gamma = 1$ ， $d()$ 使用 L2 距离公式。

4.2.3 基于图神经网络的图嵌入

现实中存在大量可以被转化为图结构的数据，例如化学分子结构、交通网络、知识关联甚至图像，从图结构中挖掘数据关联是图的研究的一个重要领域。在图神经网络被提出以前，传统机器学习处理图的方式主要是通过将图结构转化成形式更简单的数据形式，例如向量^[94]。这种图结构简单化的压缩方式损失了图的拓扑结构信息——压缩后的向量不含有节点之间的连接关系，因此缺乏表征能力。为了解决这一问题，Scarselli 等人提出了图神经网络（GNN）^[95]。受循环神经网络和马尔科夫链在图结构数据上的应用，GNN 统一了两者的优势，使用信息传递

机制，不断更新一系列对应图节点的单元，直到节点状态达到稳定的平衡，最后基于这些节点输出结果。由于这种架构能够处理更为广泛的图，例如有向图、无向图、有环图、无环图，成为了近年来新兴的基于统计的图研究方法。

受到卷积网络在计算机视觉领域所获巨大成功的激励，近来出现了很多为图数据重新定义卷积概念的方法。这些方法属于图卷积网络（GCN）的范畴。Bruna^[96] 等人于 2013 年提出了关于图卷积网络的第一项重要研究，他们基于谱图论（spectral graph theory）开发了一种图卷积的变体，这种方法直接使用图的拓扑结构，根据图的邻居信息进行信息收集。但是由于基于频谱的模型的计算成本随着图的大小而急剧增加，因此对于大图的计算效率较低，另外这种模型只能使用静态的图，面对动态更新的图时，需要执行全新的计算，因此模型的适应性不好。而基于空间的图卷积网络可以解决上述问题，因此也成为了现在主流的 GCN 方法。这些方法遵循循环递归邻域聚合（或者消息传递）的模式，其中每个节点聚合其相邻节点的特征向量用于更新当前节点的特征向量，在多轮聚合迭代后，这种聚合了邻居节点信息的特征向量被用来表示该节点。再根据任务的需要决定输出节点层级的特征（node-level）或者图层级的特征（graph-level）。

在一般的图中，节点可以表示不同的实体，但是连接节点的边没有区别，都只表示为一种连接关系。然而，知识图谱的边具有语义信息，表示实体之间的特定关系，且不同的边可能具有巨大差异的语义，因此除了节点需要表征为向量，边也需要。假设 $G = (V, E)$ 表示一个图， $X(v_i)$ 表示节点 i 的节点向量， $v_i \in V$ ， $X(e_{i,j})$ 表示节点 v_i, v_j 之间的边向量， $e_{i,j} \in E$ 。GNN 利用图结构和节点特征 $X(v_i)$ 来学习一个节点的表征向量 $h(v_i)$ ，或者整个图的表征向量 $h(G)$ 。遵循领域聚合策略，我们通过聚合它的邻近节点的表征向量来迭代更新节点的表征向量，在第 k 层，

$$a_{v_i}^{(k)} = \text{AGGREGATE}^{(k)}(\text{Tr}(h_{e_j}^{(k-1)}, X(e_{i,j}))), v_j \in N(v_i) \quad (4-14)$$

其中 v_j 为节点 v_i 的邻接节点， $\text{Tr}(h_{e_j}^{(k-1)}, X(e_{i,j}))$ 将上一层的节点的表征向量联合边向量进行融合， $\text{AGGREGATE}()$ 为该层的聚合向量。而 k 层的节点表征向量由下式得到，

$$h_{v_i}^{(k)} = \text{COMBINE}(h_{v_i}^{(k-1)}, a_{v_i}^{(k)}) \quad (4-15)$$

其中，我们初始化 $h_{v_i}^{(0)} = X(v_i)$ 。

不同的 $\text{AGGREGATE}()$ 和 $\text{COMBINE}()$ 能组合出不同的用于聚合的体系结构，在本文中，我们使用 GCN^[97] 中的方式，将 AGGREGATE 和 COMBINE 步骤

成在一体如下：

$$h_{v_i}^{(k)} = \text{ReLU}(W * \text{MEAN}\{\text{Tr}(h_{v_j}^{(k-1)}, X(e_{i,j})), h_{v_i}^{(k-1)}\}), v_j \in N(v_i) \quad (4-16)$$

其中 $\text{MEAN}\{\}$ 为 element-wise 的均值池化。

并且为了让知识库中的节点向量能和文本处理模块保持语义的一致性，我们使用 elmo 初始化节点和边特征，即，

$$X(v_i) = \text{elmo}(v_i), X(e_{i,j}) = \text{elmo}(e_{i,j}) \quad (4-17)$$

4.3 实验

4.3.1 数据集

4.3.2 实验结果分析

4.4 本章小结

第五章 全文总结与展望

5.1 全文总结

5.2 后续工作展望

致 谢

在研究生生涯即将结束之际，回看过去三年个人在学术和生活上的成长得失，内心即充满了对各位师友提供的慷慨帮助和精神支持的无限感激，又满怀着对未来的殷切期盼和希望各自安好的美好祝福。因此在论文的结束，希望借此小段略表心中的感激和热切。

首先我要感谢研究生导师郑文锋副教授相识五年以来的一路支持和批判。最初的相识是富有戏剧性的桥段，也必将影响我终身。在课上，我被郑老师对于科学问题和社会问题的独特视角所吸引，在课下，多次的讨论也均引人深思。也正是在多次的交互中，我的思维角度和视野逐渐开阔，并遵循着实证的思路开始思想重塑。研究生阶段的学术探索是在宽松的环境中展开的，长期的学术讨论也帮助我建立起了问题选取、方法定位、实验实施、论文撰写等科学研究的基本方法，本文的研究也离不开老师在选题和实验阶段的帮助，在此特别感谢。

其次，在科学研究思路和内容呈现形式上的提高，我也必须感谢实验室的其他老师，杨波副教授、刘珊副教授和李晓璐博士。每一位老师都从不同的侧面向我传达着作为一个研究者应有的态度和行为模式，这些彰显着他们价值取向的行为也帮助我建立起自我价值。对于一个即将迎接更多科研挑战和生活不确定性的年轻人而言，那些言传身教都难能可贵，尤感敬意。

再者，同实验室其他小伙伴的存在也是研究生生活的一抹亮色，大家长时间的陪伴、定期的聚会、相互的鼓励支持以及每个人独特的人格魅力都是我这三年快乐和幸福的重要来源，也必将成为未来可供追忆的幸福时候。感谢大家这一路的相伴，祝福每一位都能够在自己的人生中安稳而幸福，感谢石天一、张洁勤、肖烨、王爽、王杨、尹超、苗旺、陈阳、徐聪聪。

最后，家人、爱人、挚友的一路相伴和支持给予我无限的力量和勇气。愿亲情、友情、爱情天长地久。

参考文献

- [1] S. Antol, A. Agrawal, J. Lu, et al. Vqa: Visual question answering[C]. Proceedings of the IEEE international conference on computer vision, 2015, 2425-2433
- [2] M. Malinowski, M. Fritz. Towards a visual turing challenge[J]. arXiv preprint arXiv:1410.8027, 2014,
- [3] D. Geman, S. Geman, N. Hallonquist, et al. Visual turing test for computer vision systems[J]. Proceedings of the National Academy of Sciences, 2015, 201422953
- [4] R. Krishna, Y. Zhu, O. Groth, et al. Visual genome: Connecting language and vision using crowd-sourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73
- [5] Y. Zhu, O. Groth, M. Bernstein, et al. Visual7w: Grounded question answering in images[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 4995-5004
- [6] J. Andreas, M. Rohrbach, T. Darrell, et al. Deep compositional question answering with neural module networks. arxiv preprint[J]. arXiv preprint arXiv:1511.02799, 2015, 2:
- [7] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014,
- [8] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 770-778
- [9] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015,
- [10] B. Zhou, Y. Tian, S. Sukhbaatar, et al. Simple baseline for visual question answering[J]. arXiv preprint arXiv:1512.02167, 2015,
- [11] M. Malinowski, M. Rohrbach, M. Fritz. Ask your neurons: A neural-based approach to answering questions about images[C]. Proceedings of the IEEE international conference on computer vision, 2015, 1-9
- [12] H. Noh, P. Hongseok Seo, B. Han. Image question answering using convolutional neural network with dynamic parameter prediction[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 30-38

- [13] A. Kumar, O. Irsoy, P. Ondruska, et al. Ask me anything: Dynamic memory networks for natural language processing[C]. International Conference on Machine Learning, 2016, 1378-1387
- [14] C. Xiong, S. Merity, R. Socher. Dynamic memory networks for visual and textual question answering[C]. International conference on machine learning, 2016, 2397-2406
- [15] Q. Wu, C. Shen, L. Liu, et al. What value do explicit high level concepts have in vision to language problems?[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 203-212
- [16] Q. Wu, P. Wang, C. Shen, et al. Ask me anything: Free-form visual question answering based on knowledge from external sources[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 4622-4630
- [17] P. Wang, Q. Wu, C. Shen, et al. Explicit knowledge-based reasoning for visual question answering[J]. arXiv preprint arXiv:1511.02570, 2015,
- [18] P. Wang, Q. Wu, C. Shen, et al. Fvqa: Fact-based visual question answering[J]. IEEE transactions on pattern analysis and machine intelligence, 2017,
- [19] Y. Zhu, C. Zhang, C. Ré, et al. Building a large-scale multimodal knowledge base system for answering visual queries[J]. arXiv preprint arXiv:1507.05670, 2015,
- [20] L. Ma, Z. Lu, H. Li. Learning to answer questions from image using convolutional neural network.[C]. AAAI, 2016, 16
- [21] H. Gao, J. Mao, J. Zhou, et al. Are you talking to a machine? dataset and methods for multilingual image question[M]. Curran Associates, Inc., 2015, 2296-2304
- [22] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention[C]. Advances in neural information processing systems, 2014, 2204-2212
- [23] D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate[J]. arXiv, 2014, 11: arXiv-1409
- [24] K. Xu, J. Ba, R. Kiros, et al. Show, attend and tell: Neural image caption generation with visual attention[C]. International conference on machine learning, 2015, 2048-2057
- [25] J. K. Chorowski, D. Bahdanau, D. Serdyuk, et al. Attention-based models for speech recognition[M]. Curran Associates, Inc., 2015, 577-585
- [26] K. Cho, A. Courville, Y. Bengio. Describing multimedia content using attention-based encoder-decoder networks[J]. IEEE Transactions on Multimedia, 2015, 17(11): 1875-1886
- [27] J. Wu, S. Xie, X. Shi, et al. Global-local feature attention network with reranking strategy for image caption generation[C]. CCF Chinese Conference on Computer Vision, 2017, 157-167

- [28] L. Li, S. Tang, L. Deng, et al. Image caption with global-local attention.[C]. AAAI, 2017, 4133-4139
- [29] J. Lu, C. Xiong, D. Parikh, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 2
- [30] K. Chen, J. Wang, L.-C. Chen, et al. Abc-cnn: An attention based convolutional neural network for visual question answering[J]. arXiv preprint arXiv:1511.05960, 2015,
- [31] M. Ren, R. Kiros, R. Zemel. Exploring models and data for image question answering[C]. Advances in neural information processing systems, 2015, 2953-2961
- [32] M. Malinowski, M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input[C]. Advances in neural information processing systems, 2014, 1682-1690
- [33] K. J. Shih, S. Singh, D. Hoiem. Where to look: Focus regions for visual question answering[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 4613-4621
- [34] I. Ilievski, S. Yan, J. Feng. A focused dynamic attention model for visual question answering[J]. arXiv preprint arXiv:1604.01485, 2016,
- [35] Z. Yang, X. He, J. 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
- [36] J. Lu, J. Yang, D. Batra, et al. Hierarchical question-image co-attention for visual question answering[M]. Curran Associates, Inc., 2016, 289-297
- [37] M. Malinowski, C. Doersch, A. Santoro, et al. Learning visual question answering by bootstrapping hard attention[J]. arXiv preprint arXiv:1808.00300, 2018,
- [38] A. Jiang, F. Wang, F. Porikli, et al. Compositional memory for visual question answering[J]. arXiv preprint arXiv:1511.05676, 2015,
- [39] S. Ren, K. He, R. Girshick, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. Advances in neural information processing systems, 2015, 91-99
- [40] C.-C. Chang, C.-J. Lin. Libsvm: a library for support vector machines[J]. ACM transactions on intelligent systems and technology (TIST), 2011, 2(3): 27
- [41] Q. Le, T. Mikolov. Distributed representations of sentences and documents[C]. International Conference on Machine Learning, 2014, 1188-1196
- [42] Y. Goyal, T. Khot, D. Summers-Stay, et al. Making the v in vqa matter: Elevating the role of image understanding in visual question answering[C]. CVPR, 2017, 3

- [43] Y. LeCun. The mnist database of handwritten digits[J]. <http://yann.lecun.com/exdb/mnist/>, 1998,
- [44] C. Sun, A. Shrivastava, S. Singh, et al. Revisiting unreasonable effectiveness of data in deep learning era[C]. Computer Vision (ICCV), 2017 IEEE International Conference on, 2017, 843-852
- [45] J. Johnson, B. Hariharan, L. van der Maaten, et al. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning[C]. Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, 2017, 1988-1997
- [46] T.-Y. Lin, M. Maire, S. Belongie, et al. Microsoft coco: Common objects in context[C]. European conference on computer vision, 2014, 740-755
- [47] J. Deng, W. Dong, R. Socher, et al. Imagenet: A large-scale hierarchical image database[C]. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, 248-255
- [48] S. Auer, C. Bizer, G. Kobilarov, et al. Dbpedia: A nucleus for a web of open data[M]. Springer, 2007, 722-735
- [49] H. Liu, P. Singh. Conceptnet—a practical commonsense reasoning tool-kit[J]. BT technology journal, 2004, 22(4): 211-226
- [50] N. Tandon, G. De Melo, F. Suchanek, et al. Webchild: Harvesting and organizing commonsense knowledge from the web[C]. Proceedings of the 7th ACM international conference on Web search and data mining, 2014, 523-532
- [51] Z. Yu, J. Yu, Y. Cui, et al. Deep modular co-attention networks for visual question answering[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019,
- [52] P. Anderson, X. He, C. Buehler, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]. CVPR, 2018, 6
- [53] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need[M]. Curran Associates, Inc., 2017, 5998-6008
- [54] M. E. Peters, M. Neumann, M. Iyyer, et al. Deep contextualized word representations[C]. Proc. of NAACL, 2018,
- [55] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, et al. Object detection with discriminatively trained part-based models[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 32(9): 1627-1645

- [56] R. Girshick, J. Donahue, T. Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, 580-587
- [57] K. He, X. Zhang, S. Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916
- [58] R. Girshick. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2015, 1440-1448
- [59] K. He, G. Gkioxari, P. Dollár, et al. Mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2017, 2961-2969
- [60] J. Redmon, S. Divvala, R. Girshick, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 779-788
- [61] O. Russakovsky, J. Deng, H. Su, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252
- [62] T. Mikolov, I. Sutskever, K. Chen, et al. Distributed representations of words and phrases and their compositionality[C]. Advances in neural information processing systems, 2013, 3111-3119
- [63] T. K. Landauer, P. W. Foltz, D. Laham. An introduction to latent semantic analysis[J]. Discourse processes, 1998, 25(2-3): 259-284
- [64] J. Pennington, R. Socher, C. D. Manning. Glove: Global vectors for word representation[C]. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, 1532-1543
- [65] J. Devlin, M. Chang, K. Lee, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. CoRR, 2018, abs/1810.04805:
- [66] D. Teney, P. Anderson, X. He, et al. Tips and tricks for visual question answering: Learnings from the 2017 challenge[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 4223-4232
- [67] R. Akerkar, P. Sajja. Knowledge-based systems[M]. Jones & Bartlett Publishers, 2010
- [68] S. Nirenburg, J. Carbonell, M. Tomita, et al. Machine translation: A knowledge-based approach[M]. Morgan Kaufmann Publishers Inc., 1994
- [69] K. Knight. Building a large ontology for machine translation[C]. Proceedings of the workshop on Human Language Technology, 1993, 185-190

- [70] U. Hermjakob, E. H. Hovy, C.-Y. Lin. Knowledge-based question answering[C]. Proceedings of the Sixth World Multiconference on Systems, Cybernetics, and Informatics (SCI-2002), 2000,
- [71] G. A. Miller. Wordnet: a lexical database for english[J]. Communications of the ACM, 1995, 38(11): 39-41
- [72] R. D. Burke, K. J. Hammond, V. Kulyukin, et al. Question answering from frequently asked question files: Experiences with the faq finder system[J]. AI magazine, 1997, 18(2): 57
- [73] F. Rinaldi, J. Dowdall, M. Hess, et al. Towards answer extraction: An application to technical domains[C]. European Conference on Artificial Intelligence (15th: 2002), 2002, 26
- [74] J. M. Smith, P. A. Bernstein, U. Dayal, et al. Multibase: integrating heterogeneous distributed database systems[C]. Proceedings of the May 4-7, 1981, national computer conference, 1981, 487-499
- [75] G. Wiederhold. Intelligent integration of information[C]. ACM SIGMOD Record, 1993, 434-437
- [76] V. S. Subrahmanian. Amalgamating knowledge bases[J]. ACM Transactions on Database Systems (TODS), 1994, 19(2): 291-331
- [77] D. W. Embley, D. M. Campbell, R. D. Smith, et al. Ontology-based extraction and structuring of information from data-rich unstructured documents[C]. Proceedings of the seventh international conference on Information and knowledge management, 1998, 52-59
- [78] H. Alani, S. Kim, D. E. Millard, et al. Automatic ontology-based knowledge extraction from web documents[J]. IEEE Intelligent Systems, 2003, 18(1): 14-21
- [79] M. Banko, M. J. Cafarella, S. Soderland, et al. Open information extraction from the web.[C]. IJCAI, 2007, 2670-2676
- [80] F. M. Suchanek, G. Kasneci, G. Weikum. Yago: a core of semantic knowledge[C]. Proceedings of the 16th international conference on World Wide Web, 2007, 697-706
- [81] K. Bollacker, C. Evans, P. Paritosh, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]. Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008, 1247-1250
- [82] D. Vrandečić, M. Krötzsch. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85
- [83] J. Hoffart, F. M. Suchanek, K. Berberich, et al. Yago2: A spatially and temporally enhanced knowledge base from wikipedia[J]. Artificial Intelligence, 2013, 194: 28-61
- [84] F. Mahdisoltani, J. Biega, F. M. Suchanek. Yago3: A knowledge base from multilingual wikipedias[C]. CIDR, 2013,

- [85] Dbpedia version 2016-10, 2016. <https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10>.
- [86] T. Berners-Lee. Linked data, 2016. <http://www.w3.org/DesignIssues/LinkedData.html>.
- [87] P. van Bergen. Nli-go dbpedia demo, 2018. <https://wiki.dbpedia.org/projects/nli-go-dbpedia-demo>.
- [88] F. Wu, D. S. Weld. Open information extraction using wikipedia[C]. Proceedings of the 48th annual meeting of the association for computational linguistics, 2010, 118-127
- [89] A. Fader, S. Soderland, O. Etzioni. Identifying relations for open information extraction[C]. Proceedings of the conference on empirical methods in natural language processing, 2011, 1535-1545
- [90] J. Daiber, M. Jakob, C. Hokamp, et al. Improving efficiency and accuracy in multilingual entity extraction[C]. Proceedings of the 9th International Conference on Semantic Systems (I-Semantics), 2013,
- [91] P. N. Mendes, M. Jakob, A. García-Silva, et al. Dbpedia spotlight: shedding light on the web of documents[C]. Proceedings of the 7th international conference on semantic systems, 2011, 1-8
- [92] X. Han, L. Sun. A generative entity-mention model for linking entities with knowledge base[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011, 945-954
- [93] A. Bordes, N. Usunier, A. Garcia-Duran, et al. Translating embeddings for modeling multi-relational data[C]. Advances in neural information processing systems, 2013, 2787-2795
- [94] S. Haykin, N. Network. A comprehensive foundation[J]. Neural networks, 2004, 2(2004): 41
- [95] F. Scarselli, M. Gori, A. C. Tsoi, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80
- [96] J. Bruna, W. Zaremba, A. Szlam, et al. Spectral networks and locally connected networks on graphs[J]. arXiv preprint arXiv:1312.6203, 2013,
- [97] T. N. Kipf, M. Welling. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016,
- [98] S. A. Hasan, Y. Ling, O. Farri, et al. Overview of the imageclef 2018 medical domain visual question answering task[C]. CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org< <http://ceurws.org>>, Avignon, France (September 10-14 2018), 2018,
- [99] M. Oquab, L. Bottou, I. Laptev, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, 1717-1724

- [100] K. Kafle, C. Kanan. Answer-type prediction for visual question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 4976-4984
- [101] D. Teney, L. Liu, A. van den Hengel. Graph-structured representations for visual question answering[J]. arXiv preprint, 2017,
- [102] P. Lu, L. Ji, W. Zhang, et al. R-vqa: Learning visual relation facts with semantic attention for visual question answering[J]. arXiv preprint arXiv:1805.09701, 2018,
- [103] D. Yu, J. Fu, T. Mei, et al. Multi-level attention networks for visual question answering[C]. Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, 2017, 4187-4195
- [104] A. Agrawal, D. Batra, D. Parikh, et al. Don't just assume; look and answer: Overcoming priors for visual question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 4971-4980
- [105] A. Kembhavi, M. J. Seo, D. Schwenk, et al. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension.[C]. CVPR, 2017, 3
- [106] M. Tapaswi, Y. Zhu, R. Stiefelhagen, et al. Movieqa: Understanding stories in movies through question-answering[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 4631-4640
- [107] Y. Zhu, J. J. Lim, L. Fei-Fei. Knowledge acquisition for visual question answering via iterative querying[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017,
- [108] Q. Wu, C. Shen, P. Wang, et al. Image captioning and visual question answering based on attributes and external knowledge[J]. IEEE transactions on pattern analysis and machine intelligence, 2017,
- [109] Z.-c. Wang, Z.-g. Wang, J.-z. Li, et al. Knowledge extraction from chinese wiki encyclopedias[J]. Journal of Zhejiang University SCIENCE C, 2012, 13(4): 268-280
- [110] K. Kafle, C. Kanan. Visual question answering: Datasets, algorithms, and future challenges[J]. Computer Vision and Image Understanding, 2017, 163: 3-20
- [111] M. Hodosh, P. Young, J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics[J]. Journal of Artificial Intelligence Research, 2013, 47: 853-899
- [112] P. Young, A. Lai, M. Hodosh, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78
- [113] H. Larochelle, G. E. Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine[M]. Curran Associates, Inc., 2010, 1243-1251

- [114] M. Denil, L. Bazzani, H. Larochelle, et al. Learning where to attend with deep architectures for image tracking[J]. Neural computation, 2012, 24(8): 2151-2184
- [115] P. Rajpurkar, J. Zhang, K. Lopyrev, et al. Squad: 100,000+ questions for machine comprehension of text[J]. arXiv preprint arXiv:1606.05250, 2016,
- [116] S. K. Ramakrishnan, A. Pal, G. Sharma, et al. An empirical evaluation of visual question answering for novel objects[J]. arXiv preprint arXiv:1704.02516, 2017,
- [117] K. Saito, A. Shin, Y. Ushiku, et al. Dualnet: Domain-invariant network for visual question answering[C]. Multimedia and Expo (ICME), 2017 IEEE International Conference on, 2017, 829-834
- [118] O. Lassila, R. Swick. Resource description framework (rdf) model and syntax specification[J]. W3C (MIT, INRIA, Keio), 1997, 1-39
- [119] M. Minsky. A framework for representing knowledge[J]. 1974,
- [120] R. C. Schank, R. P. Abelson. Scripts, plans, goals, and understanding: An inquiry into human knowledge structures[M]. Psychology Press, 2013
- [121] B. Zhou, A. Lapedriza, J. Xiao, et al. Learning deep features for scene recognition using places database[C]. Advances in neural information processing systems, 2014, 487-495
- [122] M. Ren, R. Kiros, R. Zemel. Image question answering: A visual semantic embedding model and a new dataset[J]. Proc. Advances in Neural Inf. Process. Syst, 2015, 1(2): 5

攻读硕士学位期间取得的成果

- [1] X. Chen, L. Yin, Y. Fan, et al. Temporal evolution characteristics of pm2. 5 concentration based on continuous wavelet transform[J]. Science of The Total Environment, 2020, 699: 134244
- [2] X. Ni, L. Yin, X. Chen, et al. Semantic representation for visual reasoning[C]. MATEC Web of Conferences, 2019,