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 基于知识库的视觉问答技术研究

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

基于知识库的视觉问答技术研究

（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。

Research on Visual Question Answering Technology Based on Knowledge Base

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

Discipline:	**Control Science and Engineering**
Author:	**Xiaobing Chen**
Supervisor:	**Prof. Wenfeng Zheng**
School:	**School of Automation Engineering**

摘　要

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。跨领域的视觉问答接近通用人工智能，有很高的研究价值和广阔的应用场景。按照是否引入外源知识库，现有模型分为联合嵌入模型和基于知识库的模型，这两类模型在视觉问答任务中均有不错的表现。然而主流的联合嵌入模型存在数据集依赖、网络容量小和文本表征能力不足的缺陷。另一方面，通过引入外源知识库，基于知识库的模型克服了联合嵌入模型的网络容量限制，能回答涉及常识或外源知识的推理问题。但其需要通过人工构建知识库查询语句，极大的限制了模型的泛化能力。本文分别改进了联合嵌入模型的文本特征化方法和基于知识库的模型的通用性，主要包括以下内容:

1）引入动态词向量改进联合嵌入模型的文本特征化方法。目前的联合嵌入模型的文本特征化方法仍然使用静态词向量方法，考虑到静态词向量无法有效表征一词多义和一词多用的情况，本文在视觉问答模型中引入动态词向量，并结合 Faster R-CNN 和注意力机制，提出了基于动态词向量的联合嵌入模型（N-KBSN）。实验结果证明动态词向量能实现更好的文本特征表示，进而提高准确率。

2）构建了一个知识库图嵌入模块，以扩展基于知识库的模型的通用性。本文构建的知识库图嵌入模块分别从图像和文本中提取核心实体，并映射为知识库实体，再以核心实体为中心提取出子图，并将子图转换为低维向量，实现子图嵌入。为了实现好的子图嵌入，我们首先从 DBpedia 中提取了两个具有丰富语义的实验知识库：DBV 和 DBA。并基于这两个知识库，选取了一系列知识库嵌入模型进行链路预测实验。实验结果显示，DBV 知识库的实体间具有清晰的对应关系，能实现优异的节点嵌入。并且 TransE 模型能实现很好的知识库嵌入，因此我们以 TransE 为核心构建了知识库图嵌入模块。

3）合并知识库图嵌入模块和 N-KBSN 模型，构建了基于知识库图嵌入的视觉问答模型（KBSN）。在多个数据集上的实验结果证明，知识库图嵌入模块提高了视觉问答的准确率。尤其在面对需要常识或外源知识的复杂问题时，准确率提升明显。

关键词： 视觉问答，联合嵌入模型，知识库，N-KBSN，KBSN

ABSTRACT

Visual Question Answering (VQA) is an artificial intelligence task that outputs an answer to a question given a picture and a related natural language question. Compared with other tasks, VQA is closer to General Artificial Intelligence(GAI). Therefore, the research of VQA model has high research value and promising application scenarios. According to whether the knowledge base is introduced, the existing models are divided into joint embedding models and knowledge base-based models. These two types of models have good performance in VQA tasks. However, the mainstream joint embedding model has the defects of data set dependence, small network capacity and insufficient text representation ability. On the other hand, by introducing an external knowledge base, the knowledge base-based model overcomes the network capacity limitation of the joint embedding model and can answer inference questions involving common sense or external knowledge. However, it needs to construct knowledge base query statements manually, which greatly limits the generalization ability of the model. This paper improves the text representation method of the joint embedding model and the generality of the model based on the knowledge base, mainly including the following:

1) Introduce dynamic word embeddings to improve the text characterization method of the joint embedding model. The current text embedding method of the joint embedding model still uses the static word embedding method. Considering that the static word vector cannot effectively represent the polysemy and multi-word, our paper introduces dynamic word embeddings to the VQA model, combining Faster R-CNN and attention mechanism, proposed a joint embedding model (N-KBSN) based on dynamic word embeddings. The experimental results prove that the dynamic word embedding can achieve better text feature representation, thereby improving accuracy.

2) Construct a knowledge base graph embedding module to extend the versatility of knowledge-based models. The knowledge base graph embedding module constructed in this paper extracts core entities from images and text, and maps them as knowledge base entities, then extracts the sub-graphs closely related to the core entities, and converts the sub-graphs into low-dimensional vectors to realize sub-graph embedding. In order to achieve good subgraph embedding, we first extracted two experimental knowledge bases with rich semantics from DBpedia: DBV and DBA. Based on these two knowledge bases,

a series of knowledge base embedding models are selected to produce link prediction. The results show that there is a clear correspondence between the entities of the DBV, which can achieve excellent node embedding. And the TransE model can achieve a good knowledge base embedding, so we built the knowledge base graph embedding module based on TransE.

3) Merge the knowledge base graph embedding module and the N-KBSN model, and construct a VQA model (KBSN) based on the knowledge base graph embedding. Experimental results on multiple data sets prove that the knowledge base graph embedding module improves the accuracy of VQA. The accuracy improves significantly while processing complex problems that require common sense or external knowledge.

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

目 录

第一章 绪 论

1.1 研究工作的背景与意义

视觉问答（VQA）是近几年学界新兴的研究方向之一。视觉问答是一类输入为图像和用自然语言表达的文本问题，输出自然语言方式答案的人工智能任务。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。通过自然语言询问，使用者能获得细粒度的图像或者视频信息，获得一种便利的“视觉补充”。b) 扩充人机交互方式，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，视觉问答的研究代表着对未来“通用人工智能”的探索，既能够提供一种跨模态的数据处理和融合方式，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

视觉问答任务具有广阔的应用场景和对人工智能发展的深远意义，自 2014 年 VQA 挑战以来，大量的视觉问答模型被提出。按照是否引入外源知识库，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。联合嵌入模型设计特征提取网络分别提取图像和文本特征，进行跨模态的特征融合，最后使用分类器预测答案，一般为端对端的网络。基于知识库的模型则通过引入外源知识库，获得额外特征和知识，试图克服联合嵌入模型数据集依赖等缺陷。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。

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[4]、卷积[5]、逐元素相乘[1]、逐元素相加[6]等方法融合图像特征和文本特征，最后使用分类器预测答案。

Malinowski 等人首次提出了应用于真实场景的联合嵌入模型 Neural-Image-QA[6]。Neural-Image-QA 使用卷积神经网络 CNN 提取图像特征，得到的特征向量和问题文本一起传输到长短期记忆 LSTM 中，生成答案的单词序列。模型在 DAQUAR[7] 数据集上的准确率为 19.43%。这种 CNN+RNN 的基本范式也被后来的研究者大量使用。

在文本特征化的方面，Zhou 等人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[4]，并迁移预训练的 GoogLeNet[8] 提取图像特征，在 COCO-VQA 数据集上的表现良好。Gao 等人认为问题和答案在句法结构上有所不同，因此使用两个独立的 LSTM 网络编码问题和解码答案，并结合卷积神经网络构成了 mQA 模型[9]。Lin 等人既将卷积神经网络 CNN 应用于编码图像内容，也应用于问题文本的特征提取，并且使用一个多模态的卷积层输出联合特征向量，提出了双 CNN 模型[5]。

在图像特征提取方面，除了不同模型使用不同的预训练 CNN 外，Noh 等人认为单一权重配置的深度卷积神经网络无法有效处理不同的问题[10]。他们在卷积神经网络 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。提出的 DPPnet 模型由三个部分组成，作为分类网络的卷积神经网络、由门控复发单位构成的参数预测网络、将参数预测网络输出的动态参数配置到分类网络的哈希函数。

除了使用不同的方法提取图像和文本特征以外，跨模特征融合的方式也被大

量研究。Malinowski 等人通过对不同的特征向量融合方法的比较，证明了系统的准确率与特征向量融合方法有关[6]——不同方法之间准确率最多能相差 9 个百分点。除了以上提到的 iBOWIMG 采用向量拼接的方式，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特征融合方法[1]。Saito 等人认为不同的特征融合方法会保留不同层次的的特征，因此提出了一种逐元素相加和逐元素相乘相结合的模型 DualNet[11]。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，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）[12]。

由于注意力机制已经在大量的深度学习任务中被证明有效，视觉问答模型也广泛使用注意力机制。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）。模型针对“图像问题对”生成对应的注意力映射，对问题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[13]。在 Toronto COCO-QA[14], DAQUAR[7], 和 VQA[1] 三个数据集上的测试结果都实现了最优结果，证明了注意力机制在提高视觉问答任务上的有效性。

Shih 等人使用 CNN 对图片的不同区域编码，根据图像特征和文本特征的点乘结果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。在辨别物体颜色的任务上得到了最优结果[15]。类似的工作还有 Ilievski 等人提出的“聚焦型动态注意力模型“[16]。

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。而 Yang 等人则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[17]。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[18]；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”[19]。

还有研究者希望存储部分训练信息供后续的迭代训练使用，从而在原有 CNN+RNN 结构基础上，引入了动态记忆网络[20–22]。Jiang 等人在 CNN+RNN 的架构上，新增了一个成分记忆模块[20]，用于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”。Kumar 等人为解决文本问答（Text-QA）任务而提出动态记忆网络（DMN）[21]。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、情节记忆模块和答案模块构成，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特征，

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。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。受到动态记忆网络的启发，Xiong 等人在原有网络的基础上改善了输入和记忆模块，使其不仅能处理文本信息外，还能处理图像信息，因此提出了应用到于视觉问答任务动态记忆网络 +（DMN+）[22]。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。该模型在 DAQUAR[7] 和 VQA 数据集[1] 上的测试结果都得到了具有竞争力的表现。

1.2.2 基于外源知识库的视觉问答模型

联合嵌入模型由于其灵活的结构和在通用数据集上的优异表现，成为了视觉问答任务中的主流模型，但是联合嵌入模型存在以下缺陷：

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，因此对于测试集中的全新概念，模型很难得出正确的答案。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，但考虑到目前知识的巨大体量，这种数据集扩充的方式面临巨大的挑战。

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。

为弥补数据集依赖和网络容量小的缺陷，研究者在联合嵌入模型的基础上引入外源知识库，提出了一些基于知识库的视觉问答模型。根据知识库的使用方式，基于知识库的视觉问答模型分为知识库查询类和知识库嵌入类。

知识库查询类的目标是根据图像和文本创建知识库查询语句，通过知识库查询获得答案。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[23] 和 FVQA 模型[24]。

Wang 等人引入 DBpedia 知识库，提出了 Ahab 模型[23]。Ahab 分别使用预训练的 Fast R-CNN[25] 和两个不同的 VGGnet[26] 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。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。在 KB-VQA 数据集上，Ahab 在每种问题类型上的准确率都远高于联合嵌入模型。

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，这极大的限制了模型能处理的问题类型，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[24]。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。在自建的 FVQA 数据集上和多个模型的对比结果显示，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，并且可以反映出整个推理过程，实现了推理过程的去黑盒化。

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，并且使得模型能应用于更大规模的开放型数据集，避免了自建数据集带来的训练和评估问题。为了提高视觉问答系统的问题的灵活性，Wu 等人通过改进常见的 CNN+LSTM 的嵌入模型，提出了基于知识库的通用嵌入模型[27]。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。该模型采用 Toronto COCO-QA[14] 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。

1.2.3 前人工作总结

联合嵌入模型是视觉问答任务的主流模型。大量模型探索了不同的特征提取方法、特征融合方法和注意力机制。总体来看，联合嵌入模型具有较很高的灵活性，并且在通用数据集上的表现也很好，例如在 VQA 挑战中 2015-2019 年的最优模型均是联合嵌入模型，因此其具有很高的研究价值。

注意力机制的引入能有效的提高模型的准确率。在引入注意力机制前，无论对于图像输入还是文本输入，特征提取网络将输入看做一个整体，提取的特征包含问题无关的信息，降低了模型的分类效果。引入注意力机制后，输入的编码方式改变，提取的特征是局部信息的综合，既包含了更多的重要信息，也减少了模型的无关运算，提高了执行效率。

通过引入知识库，基于知识库的模型能够改善联合嵌入模型数据集依赖和网络容量小的问题，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。但其下的两类模型也各有优劣。知识库查询类模型通过人为设置查询语句，实现了在复杂推理任务上远远高于联合嵌入类模型的准确率。但也同样因为查询语句需要人为设计，当问题类型数量剧增时，人工成本骤增，这会大大限制分类数量和分类模型的精度。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。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

总而言之，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。而知识库嵌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 论文主要研究内容

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，本文重点研究了联合嵌入模型和知识库嵌入类模型。在已有工作优秀思想的基础上，考虑到模型现有的问题，本文分别对这两类模型进行了改进，从而提出了两个全新的模型。具体来说，本文的主要研究内容如下：

1. 视觉问答模型中的文本特征化方法的改进。通过对已有的联合嵌入模型的系统性研究，本文提炼出了视觉问答模型的基础架构，并且详细分解了一系列代表模型的结构。虽然目前的联合嵌入模型已经有较好的准确率，但其在文本特征化中仍使用静态词向量。考虑到真实预料中一词多义和一词多成分的情况，本文构建了一个 biLSTM 网络，用于学习场景化的词向量，引入 Elmo 动态词向量，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，并构建了一系列对比模型试验动态词向量和静态词向量对结果的影响。

2. 构建知识库图嵌入模块。为了提高基于知识库的模型的泛化能力，本文使用知识库图嵌入的方式引入知识库。为了得到优秀的知识库节点嵌入，本文进行了知识库嵌入实验。首先对原有的 DBpedia 知识库进行预处理，通过遴选数据子

集、数据清洗、去 URI 化等步骤，构建了 DBA 和 DBV 两个实验知识库。随后使用 TransE 翻译模型在 DBA 和 DBV 两个实验集上进行了链路预测实验，通过对比实验，评估 TransE 模型在知识库嵌入的有效性，并以此构建知识库子图提取模块。

3. 基于知识库图嵌入的视觉问答模型。合并之前工作中构建的知识库图嵌入模块和 N-KBSN 模型，本文提出了基于知识库图嵌入的视觉问答模型——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。最后通过在 VQA2.0 和 KB-VQA 数据集上的实验评估知识库图嵌入对模型的效果。

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，对视觉问答的国内外研究状况作了归纳，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，视觉问答任务及架构基础。本章首先介绍视觉问答任务的基础知识，包括定义、问题类型和数据集。其中，按照答案与问题和图像的相关性，本章提出了一种新的问题分类标准；按照“是否需要知识”的维度，本章将现有的代表性数据集划分为基于视觉的数据集和基于知识的数据集，并做了概括性描述。本章还提出了视觉问答模型的基础架构，并分项介绍了其特征提取、注意力机制、特征融合、答案生成的方法、原理、公式等内容。

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，本文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。最后构建了一系列对比模型进行视觉问答实验，根据实验结果，详细分析了动态词向量和静态词向量对模型准确率的影响，并解释了 N-KBSN 优异表现的原因，最后和已有模型进行了对比。

第四章，基于知识库图嵌入的视觉问答模型。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。本章提出了一个基于知识库图嵌入的视觉问答模型——KBSN 模型，并详细介绍了本文构建的知识库图嵌入模块。随后通过知识库嵌入实验，验证了知识库嵌入的有效性和模型选取的合理性。

最后本章在 KB-VQA [23] 和 VQA2.0 数据集上训练和测试 KBSN 模型，通过对比其他模型，分析实验结果，证明了知识库图嵌入模块的有效性。

第五章，全文总结与展望。本章回顾了全文的研究内容和研究结论，明确了工作的创新点和贡献，分析了研究中的不足。最后对于存在的问题，提出了改进方法以及潜在的发展方向。

第二章 视觉问答任务及基础架构

虽然视觉问答任务是从 2014 年才被提出的新兴人工智能研究，但是由于其跨学科的特性以及图像处理和自然语言处理领域的快速发展，目前视觉问答模型迭代速度很快，大量的数据集、模型被提出。本章将介绍视觉问答研究中重要的基础知识和模型的基础架构。

2.1 视觉问答任务

视觉问答是向智能系统给出图片和问题，系统返回答案的任务。由于图像内容的复杂性和问题的开放性，视觉问答的研究难度较大，正因如此，其相较于其他子任务更接近通用人工智能。本节详细分析了视觉问答的问题类型，并介绍目前主要的数据集。

2.1.1 问题类型划分

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 模型面临着开放性问题的挑战，问题类型的研究是构建模型前的关键步骤之一。

按照答案的形式划分，视觉问答的问题可以分为二值否问题[28–30]、多选问题[1,29]、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、空间关系判定等，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，在某些物体识别任务上已经能迫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研究者更好的理解模型决策的内部机制。本文提出以 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2-1所示。

表 2-1 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，这类文本强相关的问题更依赖模型对问题文本的解析。qI 类型的答案可以直接从图像中得出，这类问题能很好的评估模型对图像识别的能力。例如，对于图2-1，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?”是图像弱相关的 Qi 类型。

图 2-1 视觉问答示例图

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2-1，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。

2.1.2 视觉问答数据集

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 [7]、COCO-QA [14]、VQA [1]、VQA 2.0 [31]、CLEVR [32]、KB-VQA [23]、FVQA [24]。

以上的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。本文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2-2。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”。

表 2-2 视觉问答数据集的对比

Dataset	#images	#QA pairs	Image source	Knowledge based
DAQUAR [7]	1,449	12,468	NYU-Depth	
COCO-QA [14]	69,172	117,684	COCO	
VQA [1]	204,721	614,163	COCO	
VQA 2.0 [31]	204,721	1,105,904	COCO	
CLEVR [32]	100,000	999,968	Synthetic images	
KB-VQA [23]	700	2402	COCO+ImgNet	√
FVQA [24]	1,906	4,608	COCO	√

2.1.2.1 基于视觉的数据集

DAQUAR DAQUAR 以 NYU-Depth V2 中带有语义分割标注的图片为基础扩展得到 [7]。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 个问答对，测试集中包含 5674 对问答对，问答对由算法生成或是人类志愿者提供，算法生成的问答对根据给定的模板生成。

COCO-QA COCO-QA 包含来自 MS COCO 的 123287 张真实场景图片，问题-答案对是运用算法从 MS COCO 数据集的图像标注中生成，问题划分为物体识别、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频率从 24.98% 下降到 7.30%。COCO-QA 的训练集包含 78736 个问答对，测试集包含 38948 个问答对。

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[7]。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[33] 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，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。

VQA 2.0 针对 VQA 数据集的语言偏见问题，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。针对同样的问题，面对不同的图片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，同样需要能准确识别图片的差异。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。

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。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。

2.1.2.2 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[23]。

KB-VQA 数据集从 MS COCO[33] 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。

为了准确评估系统在需要先验知识的问题的表现，KB-VQA 人工地标注问题回答所需的知识层次，“视觉问题”、“常识问题”和“知识库问题”。其中“视觉问题”表示不需要获取额外的知识便能正确作答的问题，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?”。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。

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[24]。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[33] 和 ImageNet[34] 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[35]、ConceptNet[36] 和 WebChild[37] 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”。提取的知识以资源描述框架（RDF）的形式存储作为“支持事实”。数据集最终包含 4608 个需要先验知识的问题，涉及 3458 条事实。

2.2 视觉问答模型架构

视觉问答任务要求系统能同时正确理解问题文本和图像，从视觉问答的处理过程可以看出，算法的核心由三个部分组成：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，以求能深入地理解问题内容，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，并且模型基本上都由特征提取、注意力机制、特征融合、答案生成四个部分组成。模型使用卷积神经网络 CNN 提取图像特征，使用循环神经网络 RNN 或者长短期记忆 LSTM 处理文本信息，再通过不同的注

意力机制增强特征的表达能力，最后融合特征，使用分类器输出答案，整体架构如图2-2所示。

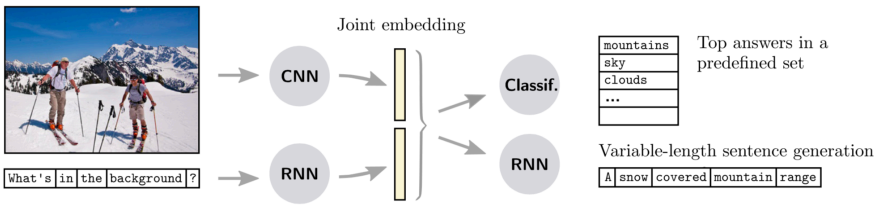


图 2-2 VQA 模型的一般架构

2.2.1 特征提取

特征提取部分一般由图像特征提取和文本特征提取两个部分组成，但在一些引入知识库的模型中，可能会添加一个知识库特征提取部分，用于丰富特征。图像特征提取的方法一般使用预训练的卷积神经网络。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）[4]、长短期记忆（LSTM）[6]、门控复发单位（GRU）[10,21,22]。

2.2.1.1 图像特征提取

在卷积神经网络出现以前，图像的特征提取一般是使用人工设计特征，例如SIFT[38]、HOG[39]。虽然这些特征在特定任务上表现良好，但是其泛化性能较差，这意味着特征提取的成本较高。而卷积神经网络是一种深度学习模型，具有分层特征学习能力以及更好的识别和泛化性能[40]。

卷积神经网络从机器视觉的成功应用开始，成为一众人工智能子领域的研究模型，已经被广泛用于物体检测、姿态检测、自然语言处理、语音识别等领域。并且随着迁移学习的兴起，大量性能优异的预训练卷积神经网络被用于图像特征提取，例如 AlexNet[41]、VGGNet[26]、ResNet[42]、GoogLeNet[8] 等。虽然有大量新的卷积神经网络被提出，但是它们都使用输入层、卷积层、池化层、全连接层和分类层组成的基本架构，如图2-3所示。

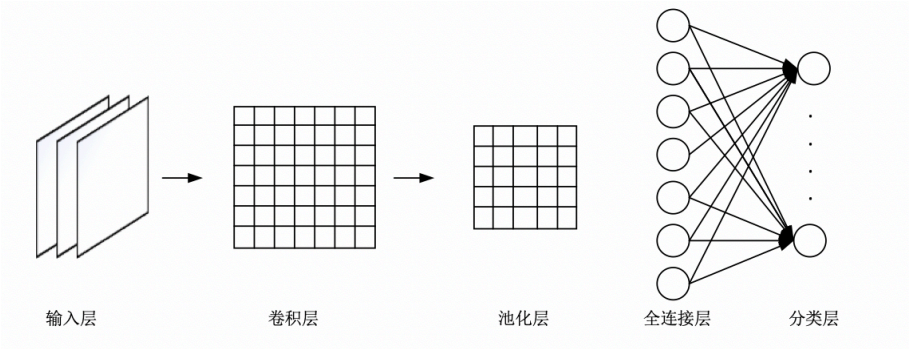


图 2-3 卷积神经网络的基本结构

卷积层是卷积神经网络的核心组成。对于高像素的彩色照片，全连接的神经网络将会产生大量参数，巨量的参数不利于模型的训练。卷积层的节点通过卷积核采样特征图的局部，并对局部的特征值进行卷积操作，再通过平移卷积核在特征图上滑动，从而扫描整张图片。不同的卷积核能够捕获图像的不同层次的特征，底层的卷积核得到的是边缘、颜色、纹理等粗粒度的特征，高层的卷积核则得到抽象的语义信息。给定输入图像 I，H_i 表示第 i 个卷积层的特征图，其中 $H_0 = I$，则相邻卷积层的特征图满足关系：

$$H_i = f(H_{i-1} \otimes W_i + b_i) \tag{2-1}$$

其中，W_i 表示第 i 层卷积核的权重向量，b_i 表示第 i 层卷积核的偏置，$\otimes$ 表示卷积操作，f 为非线性激活函数。常见的激活函数有 Sigmoid, tanh, ReLu, Leaky ReLu, Maxout 等，其中 ReLu 的公式为：

$$f(x) = max(0, x) \tag{2-2}$$

因为其具有计算量小、收敛快等优点，被广泛用作激活函数。

池化层紧跟卷积层，对特征图进行下采样，降低特征维度，一方面能减少网络的计算复杂度，加快收敛，另一方面能提取主要特征。常见的池化方式有平均池化和最大值池化。

通过多个交替的卷积层和池化层后，特征进入全连接层和分类层，对每个类别进行概率预测，得到概率分布 Y(l_i 表示第 i 个标签的类别)。因此整个卷积网络可以被视为接受输入 H_0，在网络参数 W, b 的条件下，得到正确类别 i 的函数，如下式，

$$Y(i) = P(L = l_i | H_0; (W, b)) \tag{2-3}$$

网络的训练目标是最小化损失函数 $L(W,b)$，从而更新模型参数。常见的损失函数均值平方、二值交叉熵、softmax 交叉熵等。

2.2.1.2 文本特征提取

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向量，是自动化处理文本相关的任务的基础。

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[43]，LSTM[44]、潜在语义分析（LSA）[45]，GloVe[46]。

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，但是其并没有使用传统的 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。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。

循环神经网络（RNN）是处理序列化数据最为常用的模型，通过将句子中词特征循环迭代后，得到句子特征。但由于循环神经网络具有遗忘性，存在长时依赖问题，长短期记忆神经网络（LSTM）逐渐替代 RNN。LSTM 由三个门控制，分别是输入门、遗忘门和输出门。输入门控制着网络的输入，遗忘门控制着记忆单元，自动学习需要保存的记忆，输出门控制网络的输出。

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。

以上提及的文本特征化方法被广泛的使用在视觉问答模型中。值得注意的是，

以上方法都是将文本转换为固定的静态词向量，而静态词向量缺乏对上下文的感知，因此不能有效的表征多义词和具有多语法成分的词组。本文的重要改进之一便是引入动态词向量，具体内容将在 N-KBSN 模型架构中介绍。

2.2.2 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。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降低噪音干扰。

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地提升了算法的精度和计算效率。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[47]。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[48–50]。Bahdanau 等人将注意力机制引入神经机器翻译任务，将原语言文本编码为向量序列，解码时将翻译和位置对齐联合学习，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[48]。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，并且在多个数据集上均获得了最优的结果[49]。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[51–54]。

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。因此，在视觉问答模型中，从作用对象来分，注意力机制可以分为图像自注意力机制、文本自注意力机制、引导注意力机制。不同的模型使用的注意力机制细节不同，本文将使用多头注意力机制（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），具体的实现细节详见 N-KBSN 模型。

2.2.3 特征融合

对于视觉问答任务而言，图像特征和文本特征具有异质性，数据来源和特征分布都不同，因此好的特征融合对于模型的准确性具有重要的意义。

多模态的融合方式可以分为协同表示和联合表示，协同表示是指将一种模态特征映射到另一种模态的特征空间，再融合；联合表示则是将不同模态的特征映射到同一特征空间。具体而言，协同表示是将图像特征映射到文本特征空间，再使用各类特种融合方法；联合表示则分别提取图像特征和文本特征，并且将两种特征统一维度，再进行相加或者拼接。联合表示的方式具有很高的灵活性，图像和文本特征的提取分开，并且融合的方式的选择也更为多样，因此目前大多模型均是使用该模式，本文也是使用联合表示，分别提取特征，再融合。

常用的特征融合方法有：向量串联[4]、卷积[5]、逐元素相乘[1]、逐元素相加[6]，另有模型使用更复杂的特征融合方法，例如，动态参数层[10]、多模态紧凑双线性池化方法（MCB）[12]。

2.2.4 答案生成

答案生成是解码融合后的特征，输出答案的部分。在视觉问答模型中，系统输出答案的方式有两种，最常见的方式是将任务视为分类问题，根据候选项的概率大小，确定答案。第二种方式则直接由系统合成答案语句，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[55]、ACK[27]、Ahab[23]、Facts-VQA[24]、Multimodal KB[56]。

2.3 本章小结

本章从视觉问答任务的定义、问题类型和数据集三个方面介绍该领域的基本知识，其中，本章提出了一种按照答案与问题和图像相关性的问题分类标准。本章还介绍了视觉问答模型的一般架构，并从图像特征提取、文本特征提取、注意力机制、特征融合、答案生成五个方面展开说明了目前的实践情况以及原理和公式。

第三章 基于动态词向量的联合嵌入模型

从 2015 年视觉问答任务出现至今，大量的视觉问答模型模型都属于联合嵌入模型，使之成为目前视觉问答的的主流模型。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，研究者大量尝试了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。

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。研究其中表现优异的模型容易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得到进一步提升，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[57] 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。

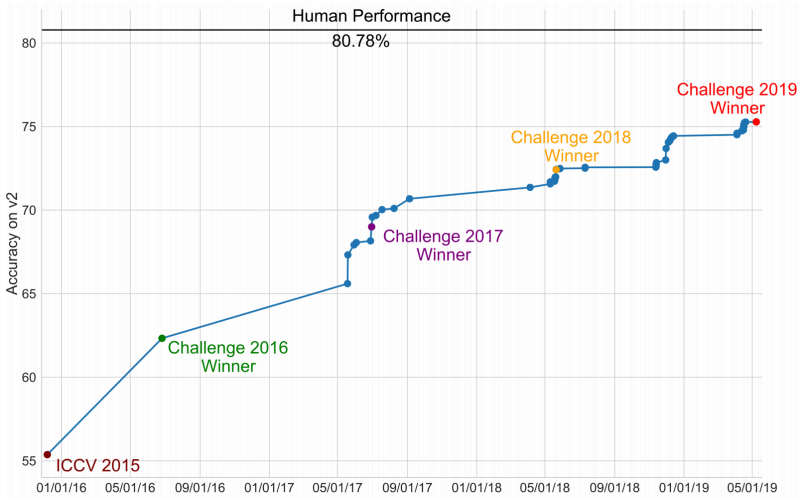


图 3-1 VQA 挑战中模型的准确率曲线

本文分析联合嵌入模型在 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[16] 提出“动态关注注意力（FDA）”模型，目的是根据问题中的关键词，对图像的不同区域动态地分配权重，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 [58] 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[59]

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，然而根据本文提出的依照答案和源信息相关性的标准（详见表2-1），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%[23]，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。

按照图像特征化方法、文本特征化方法、特征融合方法、VQA2.0 的准确性、是否使用静态词向量，具有代表性的联合嵌入模型的总结为表3-1所示。

表 3-1 代表性联合嵌入模型的比较

模型	图像特征化	文本特征化	特征融合	VQA 准确率 (%)	静态词向量
LSTM Q+I [1]	VGGnet	LSTM	逐项点乘	54.1	是
iBOWIMG [4]	GoogleNet	词袋模型	串联	55.9	是
DPPNet [10]	VGGnet	GRU	动态参数层	57.4	是
D-CNN [5]	CNN	CNN	CNN	58.4	是
MCB [12]	RestNet	LSTM	MCB	64.2	是
2017-winer [60]	Faster R-CNN	Glove+GRU	逐项点乘	69.87	是
2018-winer [58]	Faster R-CNN	Glove+GRU	逐项点乘	72.27	是
2019-winer [57]	Faster R-CNN	Glove+LSTM	MLP	75.26	是

如上表所示，现有的联合嵌入模型采用了不同的图像特征化、文本特征化、特征融合方法的组合，但是所有现有的模型的文本特征化均使用静态词向量。静态词向量使用一个语料库作为数据集，训练得到每个词语的分布式表示，这种表示方法的优点在于词语的向量预先训练得到，因此应用于不同的下游任务时，无需再训练，提高了计算效率。然而在真实的语言环境中，同一词语在不同的语境中表示不同的含义，也可能作为不同的语法成分，而这些差异并不能被静态词向量有效表示，因此可能出现语义和语法的偏差。

为了解决静态词向量的问题，本章构建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。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 [25]，问题文本特征提取使用能够获得上下文信息的 ELMo 模型 [61]，并使用从 Transformer 中借鉴的多头注意力机制 [59] 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。

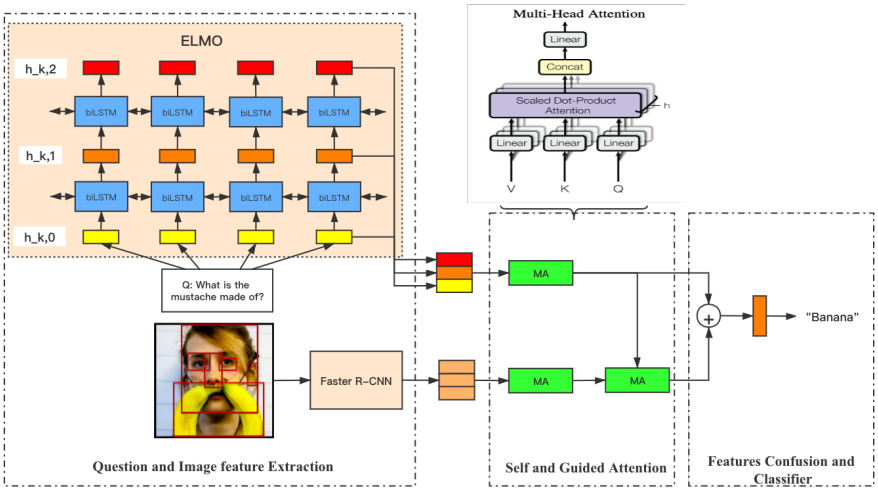


图 3-2 N-KBSN 基础结构

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。传统的目标检测算法通过滑动窗口的方式选取区域选择、提取特征提并分类，例如可变形的组件模型（DPM）方法 [62] 等。这些传统的目标检测算法大多区域选择的效果差、时间复杂度高。并且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。由于深度学习在视觉问答任务的优异表现，大批优秀的目标检测算法出现，例如 R-CNN [63]、SPP-Net [64]、Fast R-CNN [65]、Faster R-CNN [25]、Mask R-CNN [66]、YOLO [67] 及其后续版本等。以上的模型大致分为两个主要类别，第一类为两阶段检测模型，由候选区域识别和区域特征提取组成，例如 R-CNN 系列模型；第二类为单阶段检测模型，使用端对端的训练，不添加候选区域识别的网络，例如 YOLO 系列模型。由于 Faster R-CNN 在各个目标识别任务的出色表现，本节选择 Faster R-CNN 进行图像特征提取。

不同于传统目标检测模型使用的滑动卷积窗口，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，并送入 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。但由于候选区域的提取是相互独立的，因此可能存在像素重叠，使得 R-CNN 会对同一区域重复提取特征。

为了减少 R-CNN 的重复计算，研究者提出了 SPP-Net。该算法使用空间金字塔池化层（Spatial Pyramid Pooling）裁剪和缩放候选区域，使得图像大小一致，再输入到卷积层进行特征提取。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 的两阶段检测框架，并构建了一个筛选候选区域的网络（Region Proposal Network, RPN），用于控制候选区域的数量。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。最后使用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。

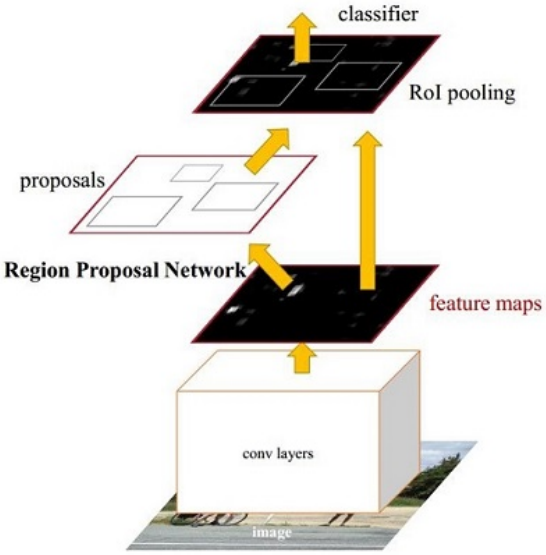


图 3-3 Faster R-CNN 基本结构

如图3-3，根据功能的不同，Faster R-CNN 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，兴趣区域池化层结合图像的特征图和候选区域，得到区域特征图，再使用全连接层和分类器，预测每个区域的类别，同时利用候选框回归得到对象的检测框。

N-KBSN 模型使用联合在 ImagNet[68] 上预训练的 ResNet-101 和在 Visual Genome[28] 上预训练 Faster R-CNN 提取图像特征。给定图像 I，模型从图像中提取 m 个大小不固定的图像特征 $X = \{x_1, x_2, ..., x_m\}, x_i \in \mathbb{R}^D$，每一个图像特征编码一个图像区域。每个图像区域的特征维度为 2048。对卷积层输出的特征图，模型使用非极大抑制（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，网络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。因此，每一张输入的图像将会被转化为一个 100×2048 的图像特征，供后续的注意力模块使用。

3.2 基于 ELMo 的文本特征化

如表3-1所示，以往的视觉问答模型的文本特征化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，因此本文将引入 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 模型是一种能感知上下文的词向量生成模型，其模型深度能够有效建模词语复杂的语义和语法，能根据词语的上下文生成动态向量，进而为解决一词多义和一词多用提供了可能。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。

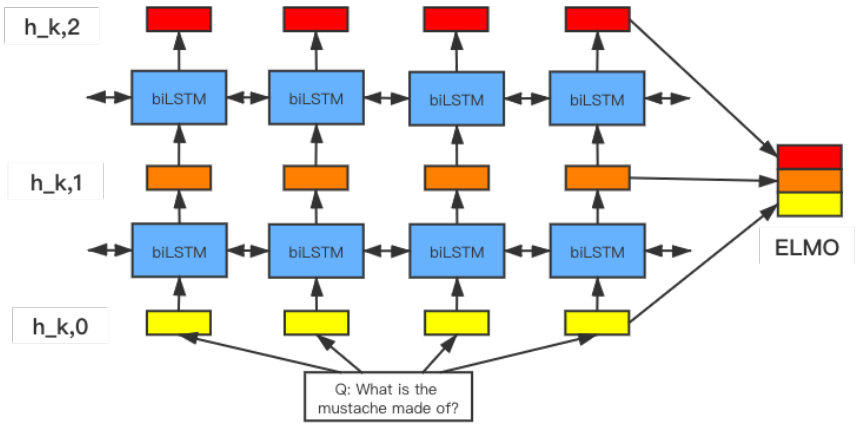


图 3-4 ELMo 模型架构

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 $S=(t_1,t_2,...,t_N)$ 而言，前向语言模型就是求解以下公式的最大值：

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

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

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

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 $j=1,...,L$。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：

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

其中，Θ_x 和 Θ_s 分别是训练阶段时的两个 softmax 层的参数，$\overrightarrow{\Theta_{LSTM}}$ 和 $\overleftarrow{\Theta_{LSTM}}$ 是 biLSTM 的参数。

完成预训练的模型对输入句子的每个单词输出三种 Embedding: 最底层是初始

的词向量 y_k^{LM}；前向 LSTM 输出的 $\overrightarrow{h}_{k,j}^{LM}$；后向 LSTM 输出的 $\overleftarrow{h}_{k,j}^{LM}$。ELMo 将三种词向量串联，得到

$$R_k = [y_k^{LM}, \overrightarrow{h}_{k,j}^{LM}, \overleftarrow{h}_{k,j}^{LM} | j = 1, ..., L] = [h_{k,j}^{LM} | j = 0, ..., L] \tag{3-4}$$

其中 $h_{k,0}^{LM}$ 是初始词向量，$h_{k,j}^{LM} = [\overrightarrow{h}_{k,j}^{LM}; \overleftarrow{h}_{k,j}^{LM}]$ 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动态词向量。

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

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。

模型首先将句子最大长度裁剪为 14，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，N-KBSN 中使用了 Transformer [59] 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。注意力函数可以被描述成映射查询（query）到一些键值对（key-value pair）并由此得到输出。假定查询矩阵 $Q = \{q_1, q_2, ..., q_m\}$，其中查询向量 $q_i \in \mathbb{R}^{1 \times d_q}$；key 矩阵 $K = \{k_1, k_2, ..., k_n\}$，其中 $k_j \in \mathbb{R}^{1 \times d_k}$；value 矩阵 $V = \{v_1, v_2, ..., v_n\}$，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：

$$Attention(Q, K, V) = score(Q, K)V \tag{3-6}$$

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：

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

其中 q_i 和 k_j 要求具有相同的维度。因此可以得到：

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

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 $(Q_i,K_i,V_i), i = 1,2,...,h$，再分别计算得到 $Attention(Q_i,K_i,V_i), i = 1,2,...,h$，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：

$$MA(Q,K,V) = [head_1, head_2, ..., head_h]W \tag{3-9}$$

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

其中 W 为线性层的权重。

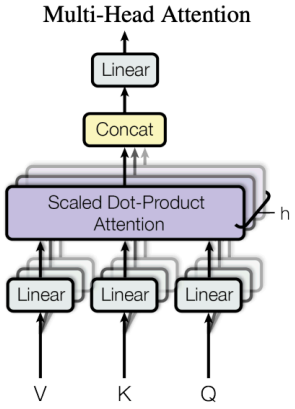


图 3-5 多头注意力的架构

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。

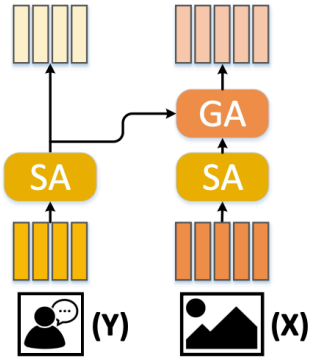
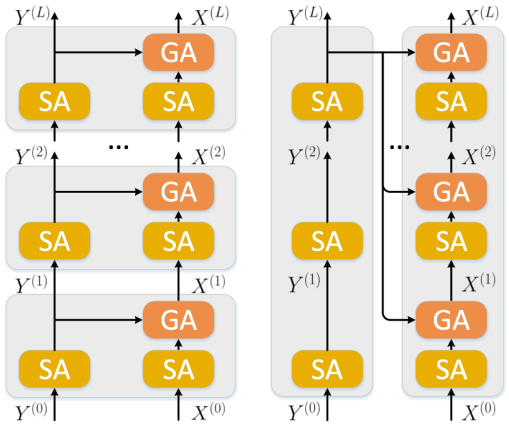


图 3-6 共同注意力模块结构

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[57]提出了 Encoder-Decoder 和 Stacking 两种级联 MCA 层的方式，如图3-7。其中，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。



(a) Stacking (b) Encoder-Decoder

图 3-7 两种 MCA 层的级联方式

根据文章给出的两种级联方式在多个任务的表现情况[57]，N-KBSN 模型使用 Encoder-Decoder 的级联方式，假定 SA^1，SA^2，...，SA^L 表示不同层的自注意力，GA^1，GA^2，...，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，

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

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

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 $X^{(L)} = [x_1^{(L)}, ..., x_m^{(L)}] \in \mathbb{R}^{m \times d}$ 和文本特征 $Y^{(L)} = [y_1^{(L)}, ..., y_n^{(L)}] \in \mathbb{R}^{n \times d}$，模型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。

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

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

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

$Y^{(L)}$ 的计算方式类似。特征融合的公式如下。

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

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。模型使用交叉熵更新参数。

3.4 实验

本节将构建一系列有不同网络结构或者参数设置的视觉问答模型，并使用通用的开放型问答数据集 VQA2.0 [31] 训练和评估模型。实验的目的是为了对比动态词向量和静态词向量对结果准确率的影响，并通过大量实验找到超参数最优的 N-KBSN 模型。整体代码使用 Python 实现，以 pytorch 为机器学习平台，并使用带有 32G 内存和 GPU 的计算机训练模型。

3.4.1 实验设置

1）数据集。本实验使用 VQA2.0 数据集训练模型。数据集被划分为 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，本文还使用了从 Visual Genome 中提取出 49 万个问答对，用于增强训练集。

2）评估方式。为了实现对开放性问题的训练和测试，VQA2.0 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。答案全为开放式问题，开放式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。因此文本使用正确率作为评价参数，包括总体正确率和子项正确率。子项正确率根据答案类型分为是否、计数和其他。

3）固定模块的参数设置。本节的实验目的是对比不同的词向量嵌入策略对于模型的准确率的影响，因此模型的其他部分应保持相同的参数设置。具体来说，对于图像特征提取模块，Faster R-CNN 的候选图像区域数 $m = 100$，单个图像区域的特征维度 $x_i = 2048$，因此单张图像特征 $X \in \mathbb{R}^{100\times2048}$。自注意力和引导注意力模块中使用的多头注意力隐层维度 $d = 512$，头数 $h = 8$，即每个头的隐层维度 $d_h = d/h = 64$，MCA 层数 $L = 6$。我们从所有答案中筛选出出现次数大于 8 的单词或词组，构建得到大小为 $N = 3129$ 答案词典，即分类的类别数为 3129。

激活函数使用 ReLU。Adam 优化器参数为 $\beta_1 = 0.9$、$\beta_2 = 0.98$，学习率为 $min(2.5te^{-5}, 1e^{-4})$，其中 t 为训练的 epoch 数，从第 10 代开始，每过两代学习率衰减为当前的 1/5。批样本数 $batch = 32$，训练代数 $epoch = 13$。

3.4.2 模型选择和对比

在固定模型的其他部分不变的情况下，本实验将构建使用不同的文本特征化方法的模型，以评估动态和静态词向量对于结果准确率的影响。实验选取了静态词向量中具有代表性、并且被广泛应用的预训练的 word2vec 和 Glove 词向量，并且级联一个单层的 LSTM 网络，将其特征维度转换为 512 维，便于后续和图像特征的融合。其中 word2vec 词向量使用的是 word2vec 模型在包含 1000 亿个单词和词组的 Google News 上训练得到的词向量。word2vec 词向量是由表示 300 万个单词和词组的 300 维向量组成。Glove 词向量是从 Wikipedia 和 Twitter 等语料库中训练得到，包含 200 万个 300 维的向量。为了使用静态词向量，首先输入样本的问题文本被裁剪为长度为 14 的单词序列，再使用查找表得到每个单词的静态向量。对于数据集中不存在预训练词向量的单词，其词向量初始化为零向量。

不同于静态向量的配置方式，为获得动态词向量，首先将预训练的深度双向 LSTM 网络 (biLSTM) 嵌入到模型中，并通过训练得到 Elmo 词向量的权重参数和尺度参数。为了进一步探究最佳的 Elmo 参数，本实验选用了三种不同参数的预训

练 ELMO 模型，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3-2。如表所示，三种不同参数的 Elmo 模型主要差异在于模型深度和词向量的维度，理论上来说，更深的网络深度具有更大的容量，并且高维的词向量能包含更多的语义信息。同样将问题文本裁剪为 14 个单词序列，并将整个单词序列作为输入，通过两层的 biLSTM 网络，得到包含上下文语境的 Elmo 词向量，再通过一个 LSTM 网络将 Elmo 词向量统一转化为 512 维，融合特征 $z \in \mathbb{R}^{512}$。

表 3-2 三种 ELMO 的参数配置

Model	Parameters (M)	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

几种词向量的统计信息如表3-3。其中，由于 Elmo 模型使用字符级别的编码，因此即使对于语料库中不存在的单词，仍然可以获得其初始词向量，进而得到其 Elmo 词向量，因此 Elmo 词向量的数量理论上无限。为了方便实验结果的分析，配置了不同文本特征的模型分别表示为：$baseline(w2v)$, $baseline(glove)$, $N-KBSN(s)$, $N-KBSN(m)$, $N-KBSN(l)$。

表 3-3 word2vec, Glove, 三种 elmo 模型的统计信息

名称	预训练语料库（大小）	词向量维度	词向量数量
word2vec	Google News（1000 亿单词）	300	300 万
Glove	Wikipedia 2014 + Gigaword 5（60 亿单词）	300	40 万
$ELMO_s$	WMT 2011（8 亿单词）	256	/
$ELMO_m$		512	
$ELMO_l$		1024	

3.4.3 实验结果及分析

3.4.3.1 VQA 实验结果分析

本次实验使用 VQA2.0 数据集训练和评估六个模型，分别是 $baseline(random)$, $baseline(w2v)$, $baseline(glove)$, $N-KBSN(s)$, $N-KBSN(m)$, $N-KBSN(l)$。其中

baseline(*random*) 的词向量使用随机初始化，并级联 LSTM 网络。同时，模型对比也加入了往年在 VQA 挑战中取得优胜的模型，2017 – *winner*(*glove*) 和 2019 – *winner*(*glove*)。各个模型在 val 验证集上的实验结果如表3-4。

表 3-4 使用不同文本特征化的模型在 val 数据集的结果

模型	总体正确率	是否	计数	其他
baseline(*random*)	62.34	78.77	41.92	55.27
2017 – *winner*(*glove*)	63.22	80.07	42.87	55.81
baseline(*w2v*)	64.37	81.89	44.51	56.31
baseline(*glove*)	66.73	84.56	49.52	57.72
2019 – *winner*(*glove*)	67.22	84.80	49.30	58.60
N – *KBSN*(*s*)	67.27	84.76	49.31	58.73
N – *KBSN*(*m*)	67.55	85.03	49.62	59.01
N – *KBSN*(*l*)	**67.72**	**85.22**	**49.63**	**59.20**

如表所示，从答案类型单项的准确率方面看，所有模型的实验结果都表现为，是否 > 其他 > 计数，且任意两个单项的准确率差值在不同模型的结果中保持稳定，这说明这种单项的准确率差异与模型无关，来源于问题本身和数据集的特性，例如，答案为“是否”的样本的随机猜测的期望准确率为 50%，而“计数”类型的随机猜测准确率很低，两种答案类型本身的难度差异导致了“是否”类型的准确率始终远远高于“计数”类型。

前五个模型均使用静态词向量作为文本特征，*baseline*(*random*) 由于使用随机初始化的文本特征，词向量能包含的语义和语法信息相较于预训练词向量更少，因此其各项准确率均为最低。2017 – *winner*(*glove*) 由于使用更为简单的注意力机制，因此准确率低于本文的 baseline 模型。对比 *baseline*(*w2v*) 和 *baseline*(*glove*) 的实验结果可以看出，即使 word2vec 词向量的语料库接近 20 倍于 glove 词向量的语料库，基于 glove 的模型还是呈现出全面的领先，在其他部分相同的情况下。分析其原因，正如论文 [46] 所说，glove 词向量使用了共现矩阵，相较于 word2vec 只使用局部的上下文信息，引入了语料库的全局信息，提高了表征能力，因此，即使使用了更大的语料库训练得到的 word2vec 词向量表征能力依然差于 glove。

最后三个模型是本文提出的 N-KBSN 模型，从结果可以看出，对比 *baseline*(*glove*)，各项的准确率均有显著提升，这证明了动态词向量确实能一定程度的提高模型的文本表示能力，进而提高整体的结果准确率。并且三个 N-KBSN 模型的准确率均

高于 $2019-winner(glove)$。而对比三种不同参数的 elmo 模型不难发现，随着模型深度和特征维度的提高，虽然整体的准确率均有提升，但提升幅度逐渐减小。具体来说，$N-KBSN(l)$ 的 elmo 模型参数量是 $N-KBSN(m)$ 的三倍多，但是准确率并无明显提升。本文选用 $N-KBSN(l)$ 作为参考模型，供之后的基于知识库图嵌入的 KBSN 模型使用。

如表3-4所呈现出的结果，使用了 Elmo 动态向量的 N-KBSN 模型的实验结果全面优于使用 Glove 静态词向量的 $baseline(glove)$。为了进一步探究其原因，本节以 $N-KBSN(m)$ 和 $baseline(glove)$ 作为实验模型，进行了定量分析和定性分析。

3.4.3.2 定量分析

为了定量分析 $N-KBSN(m)$ 和 $baseline(glove)$ 的差异，首先将 VQA2.0 的 train 随机采样，分别构成大小为原始大小 10%, 30%, 50%, 70%, 90%, 100% 的训练子集，并按照之前实验相同的参数配置，训练两个模型。两个模型在 val 验证集上的总体准确率如表3-5所示，图示见3-8。

表 3-5 $N-KBSN(m)$ 和 $baseline(glove)$ 在不同训练子集的实验结果统计

准确率	$baseline(glove)$	$N-KBSN(m)$	差值
训练子集（10%）	54.65	**55.03**	0.38
训练子集（30%）	60.30	**60.81**	0.51
训练子集（50%）	62.87	**63.40**	0.53
训练子集（70%）	64.12	**64.72**	0.60
训练子集（90%）	65.98	**66.69**	0.71
训练子集（100%）	66.73	**67.55**	0.82

图 3-8 $N-KBSN(m)$ 和 $baseline(glove)$ 在不同训练子集的实验结果统计图

根据上述图和表格可以看出，随着训练数据的增加，各自的准确率均单调上升，并且上升的速度逐渐减小，使得准确率趋于稳定，这说明增大训练数据量能够帮助提升准确率，但是提升会趋于饱和。值得注意的是，仅仅使用 10% 的训练数据，也能得到较好的准确率。

$N-KBSN(m)$ 的准确率始终高于 $baseline(glove)$，这说明动态词向量的确能提升模型的准确率。随着数据量的增大，两个模型的准确率差值逐渐增大，这是因为动态词向量能有效表征多义词的表现，更大的训练数据意味着更有可能包含词语在不同语境的使用，静态词向量则无法有效处理这种情况。

3.4.3.3 定性分析

为了定性的分析 $N-KBSN(m)$ 和 $baseline(glove)$ 在验证集结果的差异，本节将两个模型在验证集的预测结果进行对比分析，以下使用 $Res(elmo)$ 和 $Res(glove)$ 分别代表其结果集。$Res(elmo)$ 和 $Res(glove)$ 的大小均为验证集中问题的数量: 214354，其中，两个模型给出不同答案的问题数为 52990，约占总数的 1/5。在给出的不同答案中，$Res(elmo)$ 答对且 $Res(glove)$ 答错的占比为 27.8%，$Res(glove)$ 答对且 $Res(elmo)$ 答错的占比为 26.4%，结果都错的比例为 45.8%，结果都对的比例为 0%，如表3-6。

表 3-6 $Res(elmo)$ 和 $Res(glove)$ 在验证集的答案的差异

不同答案的问题数（比例）	52,990 (24.7%)
仅 $Res(elmo)$ 答对数（比例）	14,711 (27.8%)
仅 $Res(glove)$ 答对数（比例）	13,991 (26.4%)
都答错数（比例）	24288 (45.8%)
都答对数（比例）	0 (0%)

从总体的答案差异看，$N-KBSN(m)$ 答对的数量略高于 $baseline(glove)$，这也符合验证集上的整体结果的表现。

为了定性分析 $N-KBSN(m)$ 优于 $baseline(glove)$ 的可能原因，我们从答案中挑选了一些样本，如图3-9。图 (a)$Res(elmo)$ 正确识别出“浴池窗帘”的颜色，而 $Res(glove)$ 的答案则为”白色“，可能的原因是模型错误的识别了”shower” 而不是”shower curtain” 的颜色。在图 (b) 中类似，$Res(glove)$ 计数对象为”people” 而 $Res(elmo)$ 则能正确的计数”elderly people” 词组。图 (c) 则体现了 $Res(elmo)$ 能更好的识别长句中的词组而不仅仅是识别单词的能力。

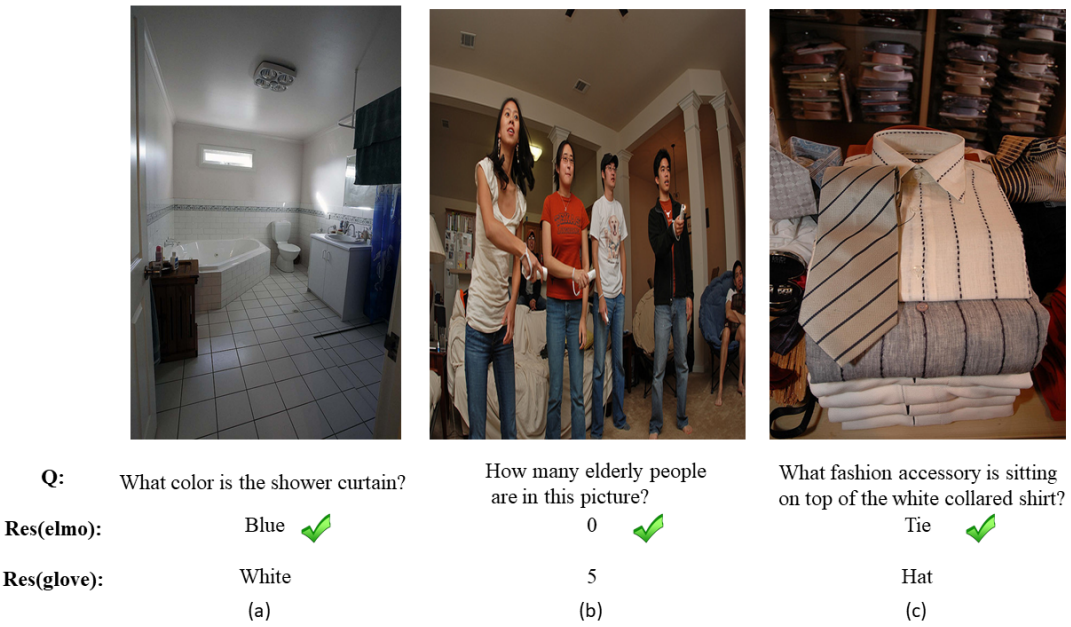


图 3-9 验证集的样本示例

3.4.3.4 和已有模型的比较

如表3-4所示，$N-KBSN(l)$ 在验证集上表现最佳，该模型和 $2017-winner(glove)$ 和 $2019-winner(glove)$ 在 Test-dev 和 Test-std 测试集上的结果如表3-7。测试集上的结果显示，本文提出的 $N-KBSN(l)$ 在各项指标上均优于其他两个模型，该结果和验证集上的结果表现一致，这证明了本文改进的文本特征化方式能够提高预测的准确率。

表 3-7 $N-KBSN(l)$ 模型和其他模型在测试集上的结果对比

模型	Test-dev				Test-std
	All	Y/N	Num	Other	All
$2017-winner(glove)$	65.32	81.82	44.21	56.05	65.67
$2019-winner(glove)$	70.63	86.82	53.26	60.72	70.90
$N-KBSN(l)$	**71.14**	**87.13**	**54.05**	**60.90**	**71.23**

3.5 本章小结

本章的主要内容是构建了一个基于动态词向量的联合嵌入模型（N-KBSN 模型），并通过对比实验证明了动态词向量能够提高视觉问答模型的性能。

本章首先介绍了视觉问答挑战的发展情况，进一步解释了联合嵌入模型主流

地位和优异表现的原因。通过对比已有代表性的联合嵌入模型，本章分析了其潜在的改进方向，从而提出了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。随后详细介绍了 N-KBSN 模型的关键部分：基于 Faster R-CNN 的图像特征化、基于 ELMo 的文本特征化和基于多头注意力机制的特征增强。

实验部分使用通用的 VQA2.0 数据集训练和验证模型。为了探究动态词向量和静态词向量对实验结果的影响，固定其他模块不变，实验使用不同的本文特征化方法构建了多个模型。实验结果证明动态词向量的引入提高了整体结果的准确率，本文构建的 N-KBSN 实现了较高的预测准确率。随后，本节进一步对 $N-KBSN(m)$ 和 $baseline(glove)$ 进行了定性和定量分析，分析结果解释了基于动态词向量的模型整体优于静态模型的原因，也证明了引入动态词向量的有效性。

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[69]。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[70]。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[71]。

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[72]。

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[73]。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[74] 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[75]。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[76–80]，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[35]、YAGO[81]、Freebase[82]、Wikidata[83] 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。

（1）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[81]。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[84]。YAGO3 构建了一个多语言的知识库[85]。

（2）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[35]。DBpedia 利用信息框提取算法检测信息框模板，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[81]。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[86]。资源描述框架的三元组数据量也从 1 亿增长到 130 亿之多。

DBpedia 有很好的数据易用性，有三种数据获取方式：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[87]。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[88]，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[23,24,27]。

（3）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[82]。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。

（4）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[83]。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 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。

正如上文中提到的，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[23,24]。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[27]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，在 N-KBSN 模型的基础上，本章构建了一个知识库的图嵌入模块，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。KBSN 的基础架构如图4-1。

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。

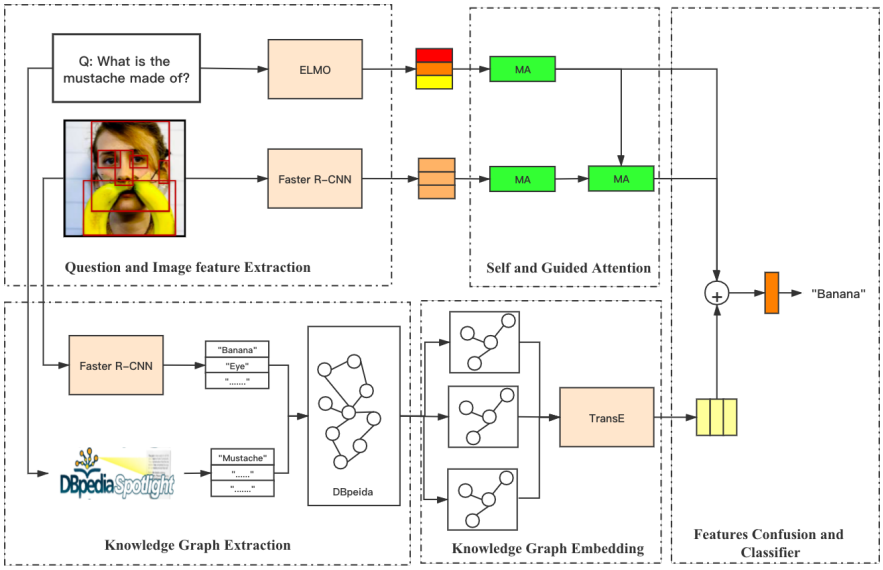


图 4-1 KBSN 的基础架构

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，通过查找表，得到图像相关的核心实体；“文本-知识库映射”使用 DBpedia Spotlight [89] 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，本文不是使用完整的 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，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，然后将子图提取模块输出的子图的节点和边都映射为向量，最后使用向量融合方法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，

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，本文首先以 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。另一方面，对于问题文本中的核心实体，模型直接使用 DBpedia Spotlight 完成从文本到 DBpedia 节点的映射。在获得图像和文本的核心实体后，模型使用贪婪的子图构建方法，提取出所有以核心实体为主语的三元组，构成知识子图。

需要注意的是，针对图像的物体识别共用的 N-KBSN 模型中的 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 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，同一个单词子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，KBSN 模型使用 DBpedia Spotlight 模型 [90] 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。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[91]，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，

$$P(e,s,c) = P(e)P(s|e)P(c|e) \tag{4-1}$$

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

$$e = argmaxP(e,s,c) \tag{4-2}$$

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，

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

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

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

对于短语 s，它的上下文 c 可以使用一个单词窗口来框定，窗口大小设为 50。

假定上下文 c 包含 n 个单词 $t_1t_2...t_n$，那么 $P(c|e)$ 的公式为，

$$P(c|e) = P_e(t_1)P_e(t_2)...P_e(t_n) \tag{4-5}$$

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，

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

$$P_{e-ML}(t) = \frac{count_e(t)}{\sum_t count_e(t)} \tag{4-7}$$

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 $P(c|NIL)$，而所有得分小于 $P(NIL,s,c)$ 的实体都会被剔除。

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

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

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

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，得到核心实体，完成文本-知识库映射。随后，模块从实验知识库中提取出所有以核心实体为主语的三元组，构建出知识子图，完成面向文本的子图提取。

4.2.2 知识库子图嵌入

在 KBSN 模型中，从问题文本和图像中提取得到的 DBpedia 实体被视为核心节点。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，模型识别出核心实体 *DBpedia*：*Snow*，并且提取出以 Snow 为中心的子图。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，因此本文提出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。

多个模型在链路预测任务的实现显示，具有更小参数量的 TransE 模型能有效的建立实体之间的复杂语义，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。具体的实验设置和结果分析详见本章的“知识库嵌入实验”。

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，

$$l_h + l_r \approx l_t \tag{4-11}$$

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，

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

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。

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

S^n 表示替换了头实体或者尾实体的三元组的集合。

4.3 知识库嵌入实验

和词嵌入一样，不同的嵌入方式对节点和节点拓扑结构的特征表示能力有差异，而这种差异又会显著影响后续应用，本文中体现为面向视觉问答任务引入的特征的有效性。因此知识库的嵌入方法需要谨慎选择，并且需要证明其有效性。

本节首先对 DBpedia 进行知识库预处理，得到 DBV 和 DBA 两个实验知识库。随后选取了几个在知识库嵌入表现优异的模型进行对比实验，使用知识图谱中的链路预测任务评估知识库嵌入效果。链路预测是通过已知的网络节点及网络结构等信息，预测网络中任意两个节点之间产生连接的可能性。在本文中，链路预测分为训练和测试两个阶段，在训练阶段，模型将高维的知识图谱中的节点和边映射为低维向量，在测试阶段，给定三元组中的头实体和关系（或尾实体和关系）预测尾实体（或头实体），并使用相似性测度衡量模型的差异。

4.3.1 知识库预处理

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，文文使用 DBpedia 作为提供额外特征的知识库。如图4-1所展示的基础架构，本文会将知识库的实体和关系转化为低维嵌入表示，再根据问题和图像提取出嵌入的知识库子图，并作为额外特征与图像特征和文本特征融合。为实现知识库的分布式表示，需要需要对原有的 DBpedia 知识库进行预处理，分别是遴选数据子集、数据清洗、去 URI 化、创建实验知识库。

（1）遴选数据子集

由于 DBpedia 从众包的 wikipedia 提取得到，完整的 DBpedia 知识库中除了和实体语义高度相关的属性外，还保留着一些语义无关的属性，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，因此本文首先从 DBpedia 中遴选出“包含丰

富语义”的子数据集，具体的数据集及其简要描述如表4-1。

表 4-1 遴选的 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

正如表4-1所示，本文只选取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。遴选的知识库子集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。

（2）数据清洗

在选定知识库子集后，本文使用 virtuoso opensource 将知识库子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，并分析了其数据内容，如表4-2所示。

表 4-2 知识库子集的参数统计

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

如表所示，知识库子集中三元组数量接近 6000 万，数据量非常庞大。一方面，庞大的数据量能覆盖更加广泛的知识内容，这蕴含着提高视觉问答准确性的更大的可能性，但另一方面，知识库的所有实体和关系都需要嵌入到同一个特征空间，数据量越大，特征向量计算难度越大，特征化的效果难以保证。因此本文进一步

研究了数据内容的特征，进一步清洗数据。

表4-3展示了一条知识库中的三元组数据，和实例相同，在 DBpedia 中，以“http://dbpedia.org/resource/ResourceName”的 URI 形式存在的节点被称为“资源”。“资源”是 DBpedia 的核心节点，可能出现在三元组的主语或者宾语。本文的知识库提取模块也是分别从图像和文本中提取 DBpedia“资源”，因此“资源”是本文首要关注的节点。

表 4-3 知识库子集的三元组实例

主语	<http://dbpedia.org/resource/Snow>
谓语	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
宾语	<http://www.w3.org/2002/07/owl#Thing>

为了方便后续的数据清洗，本文首先提取了所有三元组的中包含的 DBpedia 资源，建立了一个资源序列表。资源序列表中包含知识库中接近 1500 万个 DBpedia 资源。本文使用 SPARQL 批量查询了各个资源的谓语数量，得到如表4-4所示的统计信息。

表 4-4 DBpedia 资源的谓语数量

谓语数量范围	谓语数 =0	谓语数 =1	谓语数 >1
[0, 37]	391,904	9,190,505	5,365,537

如表4-4所示，数据集中 DBpedia 资源的谓语数量最小为 0，最大为 37。谓语数量为 0 的资源有将近 40 万，这类资源在三元组中充当宾语。谓语数量为 1 的资源有 900 万左右，所有这些资源的谓语都是“<http://www.w3.org/2000/01/rdf-schema#label>”，宾语都是各自的英文名称，除此以外知识库中不包含任何其他的属性。以上两种资源都对于本次研究没有意义，因为这些资源在知识库中没有其他有价值的属性，因此不能为模型提供额外的信息。所以本文剔除了包含以上两种 DBpedia 资源的三元组，保留了所有剩余的三元组。

（3）去 URI 化

正如表4-3所展示的，DBpedia 的实体多是以 URI 的形式表示的，这种表示方式的优势是 DBpedia 资源直接对应一个网络资源，可以通过浏览器直接访问相应节点，方便构建和其他数字资源的关系。然而在本文的研究中，并不会使用到这种特性，反而因为表示网站地址的前缀和节点的语义并无联系，冗余的表示方法

不仅增大了文本存储的空间，也降低了本文匹配的效率。

因此我们研究了实体和属性的表示方法，将所有节点去 URI 化，去除后的结果实例如表4-5。

表 4-5 节点去 URI 化的实例

	去 URI 前	去 URI 后
主语	<http://dbpedia.org/resource/Snow>	Snow
谓语	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	22-rdf-syntax-ns#type
宾语	<http://www.w3.org/2002/07/owl#Thing>	owl#Thing

（4）创建实验知识库

为研究知识库大小对模型的影响，本文创建了两个实验数据库：DBV 和 DBA。DBA 知识库是经过以上的处理步骤得到的完整知识库；DBV 知识库是根据特定规则从 DBA 知识库提取得到的，规则为三元组均以 VQA2.0 数据集的问题所包含的 DBpedia 资源为主语。两个知识库的统计信息如表4-6。

表 4-6 DBV 和 DBA 知识库的统计信息

统计量	DBV	DBA
三元组数	98,201	52,247,554
主语数	7,863	5,365,537
谓语数	1,377	1,377

如表4-6所示，DBV 和 DBA 知识库均包含了 1377 个谓语，这和数据清洗前的知识库相同。数据清洗后，DBA 知识库的三元组数比初始的知识库少了 700 万，占比为 13%，这明显得减少了数据冗余，说明数据清洗的必要性。DBV 知识库包含了 7800 个主语，9 万个三元组，虽然其数量远小于 DBA 知识库，但是该知识库的三元组的主语与 VQA2.0 数据集中的问题高度相关，本文构建该知识库去测试其在视觉问答任务的表现。

4.3.2 模型选择和对比

本文选取了几个在知识库嵌入表现优异的模型作为测试模型，包括结构化嵌入（SE）[92]、语义匹配能量函数（SME）[93]、TransE 模型[94]。TransE 模型已经在前文中详细介绍了，其主要思想为将知识库中的实体和关系都嵌入为一个低维

向量，对于一个存在的三元组 (h,r,t)，$d(l_h+l_r,l_t)$ 应该更小，相比于不在同一个三元组的两个实体而言。类似的，SE 仍然使用向量的距离度量实体是否在同一个三元组。不过在 SE 中，关系被嵌入为两个关系矩阵 $R_k^h \in \mathbb{R}^{k\times k}, R_k^t \in \mathbb{R}^{k\times k}$，并且使用 $d(R_k^h h, R_k^t t)$ 度量两个实体的距离。SME 先将知识库的实体和关系转化为向量，再使用神经网络去训练一个语义匹配能量函数 $\epsilon((h,r,t))$，使得处于同一个三元组的两个实体的能量最低，从而实现链路预测。SME 又根据使用的函数不同分为 SME(LINEAR) 和 SME(BILINEAR)。

虽然以上的模型背后的思想都是以关系为基准度量两个实体之间的距离，从而实现头实体和尾实体的匹配，进而完成链路预测。但是其实现方法的不同导致模型的参数数量不同，表4-7对比了四种模型的参数量和在 FB15K、DBV、DBA 上的具体参数量。

表 4-7 不同知识库嵌入模型的理论参数量和实际参数量（百万）

模型	理论参数量	FB15K(M)	DBV(M)	DBA(M)
SE	$O(n_e k+2n_r k^2)$	7.47	7.14	434.93
SME(LINEAR)	$O(n_e k+n_r k+4k^2)$	0.82	2.82	428.54
SME(BILINEAR)	$O(n_e k+n_r k+2k^3)$	1.06	3.06	428.78
TransE	$O(n_e k+n_r k)$	**0.81**	**2.81**	**428.53**

其中 n_e, n_r 分别表示数据集中实体和关系的数量，k 表示嵌入的维度，在计算具体值时，设定 $k=50$。从表中可以看出，随着数据集中的实体和关系数量的增加，所有模型的参数量都逐步增加，其中 SE 模型对关系数量相较其他三个模型更为敏感。总体来看，TransE 具有参数较少的优点，在任意大小的数据集上，其参数量均为最少。SME(LINEAR)、SME(BILINEAR) 和 TransE 的参数量接近，其差距主要取决于向量的维度，TransE 更容易实现更高维度的嵌入。

4.3.3 实验设置

1）数据集。为了测试模型在不同体量的知识库上的表现，本文使用三个数据量差异明显的数据集——FB15K、DBV 和 DBA。FB15K 从 Freebase 中筛选和提取得到的，数据包含 592213 条三元组，涉及 14951 个实体和 1345 种关系。DBV 和 DBA 是本文提出的从 DBpedia 提取得到的两个知识库数据集，其中 DBV 数据集中的三元组数据的主语提取自 VQA2.0 数据集的问题文本，是本文专门设计的面向视觉问答任务的数据集，其包含 98201 条三元组数据，涉及 55341 个实体和 874

种关系。DBA 包含 5200 万条三元组数据，涉及 856 万个实体和 1293 种关系。三个数据集的统计信息如表4-8所示。

表 4-8 FB15K、DBV 和 DBA 数据集的统计对比

统计量	FB15K	DBV	DBA
三元组数	592,213	98,201	52,247,554
实体数	14,951	55,341	8,569,361
关系数	1,345	874	1,293
训练集三元组数	483,142	80,200	41,798,043
验证集三元组数	50,000	9,000	5,224,755
测试集三元组数	59,071	9,001	5,224,756

从数据量的角度分析，DBA 数据集的三元组数和实体数均高出其他两个数据集两个数量级，而这样大体量的数据集一方面更接近真实场景的知识量，另一方面要求模型具有很好的尺度变化能力。DBV 数据集的实体数是 FB15K 的 3.7 倍，但是数据量仅为 FB15K 的 1/7，这说明 DBV 数据集中的单个头实体对应的平均尾实体数少于 FB15K，这种特性理论上能实现更易区分的特征向量。

数据集按照 8 : 1 : 1 的比例被划分为训练集、验证集合测试集，并且保证了测试集中的节点均在训练集或验证集中出现。

2）实验评估。为了评估模型的表示学习的效果，实验使用排名相关的评估标准。具体来说，对于每一个测试三元组，首先先将头实体替换成数据集中其他的实体，组成新的三元组，并通过模型计算得到其差异值或者能量。基于模型的假设，正确的三元组应该具有更低的能量或者差异值，因此对每一个新构成的三元组的差异值或者能量按照从低到高的升序排序，进而得到正确的实体的排名。再对测试三元组进行替换尾实体的操作，并再次计算正确实体的排名，重复直到所有测试三元组均完成了头实体和尾实体的替换。最后将所有正确实体的排名取均值，得到的 Mean Rank 作为一个评估指标。另一个评估标准为正确实体排名前十的比例——hits@10。

3）参数设置。本节实验的四个模型均使用梯度下降算法更新参数，选取学习率的范围 $\lambda \in \{0.001, 0.01, 0.1\}$，特征向量的维度 $k \in \{20, 50\}$，向量的距离计算使用 L_1 或者 L_2 距离计算公式，TransE 中使用的距离超参数 $\gamma \in \{0.5, 1, 2\}$。训练代数最多为 1000 代，并且根据验证集的 Mean Rank 保存最优模型的参数和状态。

其中 TransE 针对 FB15K、DBV、DBA 数据集上的最佳参数配置如表4-9所示。

表 4-9 TransE 对 FB15K、DBV 和 DBA 数据集的最佳参数

参数	FB15K	DBV	DBA
特征向量维度 k	50	50	50
学习率 λ	0.01	0.01	0.01
距离参数 γ	0.5	1	1
距离计算式 d	L_2	L_1	L_1

4.3.4 实验结果及分析

实验结果如表4-10所示。从每个数据集的结果来看，TransE 的 Mean Rank 和 hits@10 均优于其他三个模型。具体来说，在 FB15K 数据集上，TransE 的 Mean Rank 最低，且优于其他三个模型 80 左右，而在 hits@10 的准确率上更是高出 10% 左右，大幅领先于其他三个模型。这一整体的领先也体现在 DBV 数据集上，并且随着数据量的增加，TransE 的优势继续扩大。

表 4-10 四个模型在链路预测任务的实验结果

数据集	FB15K		DBV		DBA	
评估标准	Mean Rank	hits@10(%)	Mean Rank	hits@10(%)	Mean Rank	hits@10(%)
SE	273	28.8	1043	48.5	604836	35.4
SME(LINEAR)	274	30.7	1057	51.4	618265	37.2
SME(BILINEAR)	284	31.3	1150	53.1	634921	42.6
TransE	**195**	**41.2**	**817**	**67.6**	**567420**	**54.6**

根据不同数据集的结果可以得知，随着数据集的体量的增加，待嵌入的节点数增加，模型的 Mean Rank 呈现下降趋势，其中 SE、SME(LINEAR) 和 SME(BILINEAR) 的升高尤其明显，TransE 在面对大数据集时，性能下降平缓，这是由于其模型的简洁性带来的扩展性能的提升。值得注意的是，四个模型在 DBA 的实验结果均明显差于 DBV，并且其 Mean Rank 非常高，大于 50 万，而且模型在 DBA 数据集上的单次训练时间是 DBV 的数百倍。这说明 DBA 知识库虽然包含有更多的数据，理论上能实现更精准的嵌入，但是碍于算力和模型的精度，其并没有很高的实用性。

但是对比 DBV 和 FB15K 上 hits@10 指标，结果的趋势和 Mean Rank 并不相同。DBV 数据集待嵌入的节点数是 FB15K 的 3.7 倍，但是可用于训练的三元组仅

仅是 FB15K 的 1/7，但是其 hits@10 准确率却更高，TransE 在 DBV 上的准确率比在 FB15K 上高出 26.4%，同样的情况发生在其他三个模型上。

为了进一步理解模型在 Mean Rank 和 hits@10 两种指标上呈现出的不同趋势，本文将 FB15K 和 DBV 三元组中的关系类型分成了四种类型：一对一、一对多、多对一、多对多。一对一的关系类型的头实体和尾实体一一对应，一对多的关系的每一个头实体对应多个尾实体，多对一的关系的尾实体对应多个头实体，多对多的关系则是多个头实体和多个尾实体对应。具体来说，对于关系 r，在数据集中，累加其每个尾实体对应的头实体数得到总的头实体数，再除以尾实体个数，得到平均头实体数，同理可以得到平均尾实体数。如果平均头实体数小于 1.5，且平均头实体数小于 1.5，则关系划分为一对一，同理可得其他三种关系，如表4-11所示。

表 4-11 四种关系类型的划分标准

关系类型	平均头实体数	平均尾实体数
一对一	< 1.5	< 1.5
一对多	< 1.5	>= 1.5
多对一	>= 1.5	< 1.5
多对多	>= 1.5	>= 1.5

根据上表的划分标准，本文分析了 FB15K 和 DBV 中关系类型的分布，如表4-12所示。FB15K 数据集中四种类型分布较为均匀，而 DBV 数据集中一对一的关系占比超过 50%，为主要类型。一对一的关系所在的三元组中，头实体和尾实体的对应关系相对清晰，可能更利于模型训练更易区分的特征表示。

表 4-12 FB15K 和 DBV 中关系类型的分布

占比 (%)	FB15K	DBV
一对一	26.2	50.2
一对多	22.7	11.1
多对一	28.3	26.8
多对多	22.8	11.9

为了分析不同关系类型和准确率的关系，本文分析了 TransE 模型在 FB15K 和 DBV 数据集上，对于四种关系类型的 hits@10 的准确率，如表4-13所示。

表 4-13 TransE 在 FB15K 和 DBV 数据集的结果对比

任务	预测头实体 hits@10(%)				预测尾实体 hits@10(%)			
关系类型	一对一	一对多	多对一	多对多	一对一	一对多	多对一	多对多
FB15K	43.7	65.7	18.2	**47.2**	43.7	19.7	66.7	50.0
DBV	**81.8**	**81.7**	**32.8**	47.0	**83.2**	**75.9**	**74.6**	**58.8**

如上表所示，无论是预测头实体还是尾实体，TransE 在 DBV 数据集上，基本上对于所有的关系类型的 hits@10 均高于其在 FB15K 的表现，其中对于一对一的关系的 hits@10 的准确率两倍于 FB15K 的表现，其他类型也有明显的提升，这显示了 TransE 在 DBV 整体 hits@10 指标上优于 FB15K 的原因。

四种关系类型的准确率差异显示，无论是在哪个数据集，TransE 在预测实体数为一的一边时，具有更高的准确率，例如，在预测头实体任务中，“一对一”和“一对多”的准确率高于其他各种类型，而在预测尾实体任务中，“一对一”和“多对一”的准确率高于“一对多”和“多对多”的类型。这说明 TransE 模型更适合处理“一一映射“的数据。

TransE 对 DBV 数据集中一对一的关系，预测准确率达到 81.8% 和 83.2%，表现出极高的预测准确率。高准确率证明了 TransE 模型的节点嵌入的有效性。对于一对一的关系类型，头实体和尾实体存在一一对应关系，而无关节点则应当充分远离。高 hits@10 说明模型能够有效区分正确的匹配和错误的匹配，实现无关节点的距离最远。此外，TransE 之所以在 DBV 上有更优的表现，这说明 DBV 中三元组的实体之间对应性更强，更便于节点的嵌入。表4-12印证了该结论。

结论　SE、SME(LINEAR)、SME(BILINEAR) 和 TransE 在链路预测上的实验证明，TransE 模型在参数更少的情况下，具有更好的节点嵌入表现，且面对不同体量的知识库时，其在尺度扩展的能力也是最优，因此非常适合用于大知识库的节点嵌入。

通过对 TransE 在 FB15K 和 DBV 数据集上更细粒度的实验，结果显示 TransE 模型能够非常有效的编码节点，对于存在一一对应关系的实体，其预测准确率尤其高。此外，DBV 数据集中的实体之间的对应性更强。在训练数据较少并且待嵌入节点数更多的情况下，同样的模型能实现更高的链路预测的准确率，即更优的节点嵌入表示。因此 DBV 数据集优于 FB15K。

另一方面，虽然 DBA 数据集相较于 DBV 拥有更多的数据和实体，但实验结果证明，巨大的数据量成为了知识库嵌入训练的障碍，四个实验模型在 DBV 数据

集上的 hits@10 也远高于 DBA。结果说明 DBA 对于目前的模型而言，并没有太大的实用价值，因此此后的视觉问答任务实验仅仅使用构建的 DBV 知识库。

4.4 视觉问答任务实验

本节实验将使用使用多个数据集训练和评估本文构建的基于知识库图嵌入的视觉问答模型（KBSN）。为了验证知识库图嵌入对结果的影响，实验还构建了一系列对比模型，包括上一章提出的 $N-KBSN(l)$。整体代码使用 Python 实现，以 pytorch 为机器学习平台，并使用带有 32G 内存和 GPU 的计算机训练模型。

4.4.1 实验设置

（1）参数设置

本章提出的 KBSN 模型是在上一章提出的 N-KBSN 模型的基础上，引入了知识库子图提取模块和子图嵌入模块后构建而成。因此图像和文本提取模块、注意力机制、特征融合和分类器的参数配置都和 $N-KBSN(l)$ 相同。其中 Elmo 模型的双向长短期网络（biLSTM）隐藏层维度为 4096，单层输出维度为 512，Elmo 动态词向量维度为 1024。

知识库子图提取模块使用了和图像特征提取模块中相同的 Faster R-CNN 网络。不同之处在于，图像特征提取仅仅使用兴趣区域的特征，而知识库子图提取模块添加了两个全连接层和 softmax 层用于预测图像中的物体，其中，单张图像被检测对象数量 $m \in [5,10]$ 分类器的大小为 1600。Spotlight 模型的单词窗口大小设为 50。

（2）数据集

由于 KBSN 模型引入的外源知识库能一定程度增强模型对先验问题的感知，因此实验除了使用 VQA2.0 数据集[31] 外，还使用了 KB-VQA 数据集[23]。KB-VQA 包含 700 张图片，150 个物体类别和 100 个场景类别，总共包含 2741 个问答对。问题类型包含“常识问题”、“视觉问题”和“知识库”问题，它们各自的数量为 1256、883、263。依照第三章的数据集划分，VQA2.0 被划分为 train/val/test，比例为 2:1:2。由于 KB-VQA 中的图片全部包含于 VQA2.0 中，并且考虑到其数据量较小，因此 KB-VQA 数据集仅分为 train/test，比例为 1:1。

（3）知识库

本节将使用上文构建的提取自 DBpedia 的知识库：DBV。DBV 中三元组的主语均提取自 VQA2.0 数据集中的问题，包含 98201 条三元组。

（4）评估方式

基于 VQA2.0 数据集的实验的评估方式详见第三章实验评估部分。KB-VQA 的评估方式也使用人工标注，具体来说，测试者评估答案的正确性，并且给出五个分数，1：完全错误；2：些许错误；3：及格；4：正确；5：完全正确。得分大于三分的答案被判定为正确，否则为错误。

实验结果均使用正确率为指标，其中 VQA2.0 的实验结果包括总体正确率、是否问题正确率、计数问题正确率和其他问题正确率，KB-VQA 实验结果包括总体正确率、视觉问题正确率、常识问题正确率和知识库问题正确率。

4.4.2 实验结果分析

4.4.2.1 VQA 实验结果分析

本小节使用 VQA2.0 数据集进行实验，参与实验的模型有 $N-KBSN(l)$、$KBSN(spotlight)$、$KBSN(rcnn)$、$KBSN$。其中，$N-KBSN(l)$ 为前文构建的基于动态词的联合嵌入模型，$KBSN(spotlight)$ 和 $KBSN(rcnn)$ 分别为在 $N-KBSN(l)$ 基础上添加文本子图提取和图像子图提取的对照模型，$KBSN$ 为在 $N-KBSN(l)$ 基础上同时引入图像子图提取和文本子图提取的视觉问答模型。各个模型在 val 验证集上的实验结果如表4-14，结果包括总体正确率以及三种答案类型：是否、计数、其他的单项正确率。

表 4-14 使用不同子图提取模块的模型在 val 数据集的结果

模型	总体正确率	是否	计数	其他
$N-KBSN$	67.72	85.22	49.63	59.20
$KBSN(spotlight)$	67.76	85.26	49.64	59.25
$KBSN(rcnn)$	67.95	85.53	50.04	59.31
$KBSN$	**68.04**	**85.61**	**50.23**	**59.40**

如表所示，本章提出的基于知识库图嵌入的视觉问答模型 $KBSN$ 及其两个变体在各项结果的准确率上均高于 $N-KBSN$ 模型，但不同变体的提升幅度有着明显差异。具体来说，仅加入文本子图提取的 $KBSN(spotlight)$ 相较于 $N-KBSN$ 仅仅只有 0.05 个百分比的提升，而加入图像子图提取的 $KBSN(rcnn)$ 则有将近 2 个百分点的提高，这种差异也同时体现在三个子项的准确率上。可能的原因有两点：第一，VQA2.0 中的问题长度普遍较短，因此能提取得到的实体数量较少。第二，图像相较于问题文本具有更丰富的语义，图像子图提取能够获得更多和答案紧密相关的信息，因此对于最终结果的提升作用更大。

计数问题对图像识别的准确率的较高要求，因此计数问题能体现模型对图像中对象的感知能力。在计数问题的准确率上，*KBSN*(*rcnn*) 相比于 $N-KBSN$ 提升了 0.41%，而 *KBSN*(*spotlight*) 仅提升了 0.01%，这说明图像子图提取模块的确增强了模型对图像的解析能力。

KBSN 因为同时添加了两个模块，结果是四个模型中最优的，但是准确率的提升并不明显。

4.4.2.2 KB-VQA 实验结果分析

相较于 VQA2.0 数据集，KB-VQA 数据集的体量更小，但是其问题中除了包含“视觉问题”外，还包含需要复杂推理或者外源知识的“常识问题”和“知识库问题”，因此对模型的推理能力要求极高。本小节仍然使用 $N-KBSN(l)$、*KBSN*(*spotlight*)、*KBSN*(*rcnn*)、*KBSN* 四个模型进行对比实验。

首先将四个模型在 VQA2.0 训练集上完成预训练，再使用 KB-VQA 的训练集进行迁移训练，最后使用 KB-VQA 的测试集评估模型。实验结果中引用了 Ahab 模型的结果，Ahab 模型专门针对 KB-VQA 数据集的问题，人工构建了知识库查询语句，属于知识库查询类模型。实验模型在 KB-VQA 测试集上的实验结果如表4-15，结果的直方统计图如图4-2。如图表所示，$N-KBSN$ 在所有模型中表现最差，由于该模型没有引入任何知识库的嵌入，其推理信息仅仅来源于训练集中的图像和文本，因此其在视觉问题的正确率远远高于常识问题。知识库问题表现极差，正确率仅有 17.4%。

表 4-15 实验模型在 KB-VQA 测试集上的结果

模型	总体正确率	视觉问题	常识问题	知识库问题
$N-KBSN(l)$	37.5	47.8	33.5	17.4
KBSN(*spotlight*)	42.8	49.4	37.3	30.3
KBSN(*rcnn*)	47.3	53.0	42.3	37.1
KBSN	53.7	56.7	50.5	46.2
Ahab	**69.6**	**67.2**	**70.1**	**75.0**

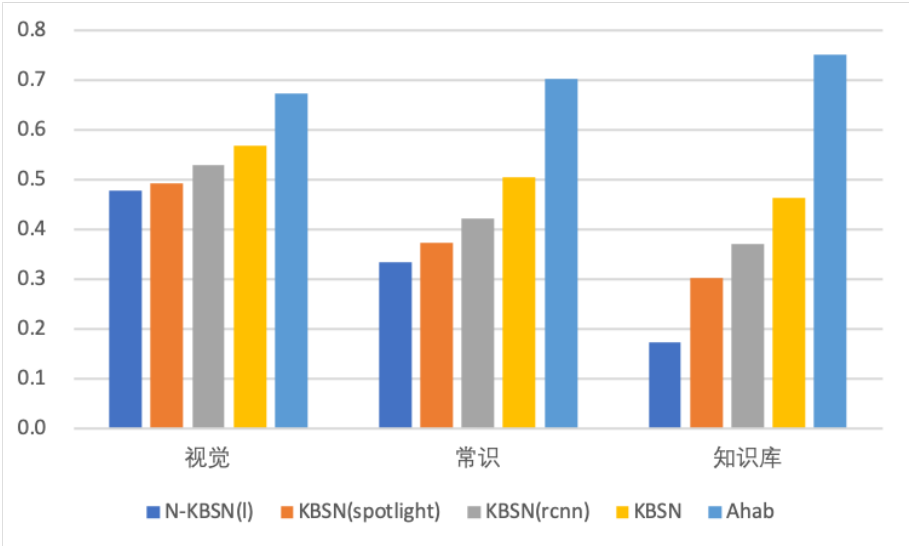


图 4-2 不同模型在不同问题类型的准确率

从知识库问题的正确率来看，*KBSN*(*spotlight*) 和 *KBSN*(*rcnn*) 由于都部分引入了知识库中的节点，所以正确率相较 $N-KBSN$ 大幅提高，这说明无论是文本子图嵌入还是图像子图嵌入均能一定程度的提高模型对知识库问题的解决能力。从视觉问题的准确率看，前四个模型的准确率差距较小，这说明作为基础模型的 $N-KBSN$ 擅长于解决视觉问题，这和其在 VQA2.0 数据集上的优秀表现有关。所有模型在常识问题的准确率均介于视觉问题和知识库问题之间，这说明三类问题的难度梯度非常明显，表现为知识库问题 > 常识问题 > 视觉问题。

本文构建的 *KBSN* 相较于前三个模型，准确率全面领先，这说明文本子图提取和图像子图提取是互补的，每新增一个模块都能提升准确率。

Ahab 在各项内容的准确率上均呈现出大幅的领先，这是因为其针对 KB-VQA 数据集中的每一种问题类型均设计了相应的 SPAQL 查询语句，使其能够精准的查询得到知识库的节点。实验结果也证明了，在推理能力上，知识库查询类模型 (*Ahab*) > 知识库嵌入类模型 (*KBSN*) > 联合嵌入模型 ($N-KBSN(l)$)。

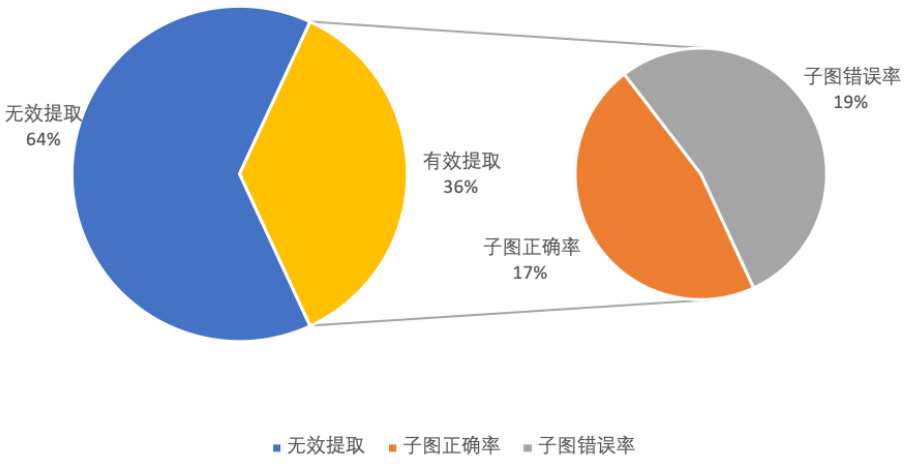
4.4.2.3 子图提取模块分析

根据以上两个在 VQA2.0 和 KB-VQA 数据集的实验可以看出，引入了知识库图嵌入的 KBSN 模型相较于 N-KBSN 模型的准确率提升在不同数据集上的效果差距明显。具体来说，在 VQA2.0 数据集上，KBSN 的总体正确率仅提升了 0.42%，而在 KB-VQA 数据集上，提升了 16.2%。为了探究其准确率提升的差异性，本节探究了子图提取模块在两个数据集上的实际表现。

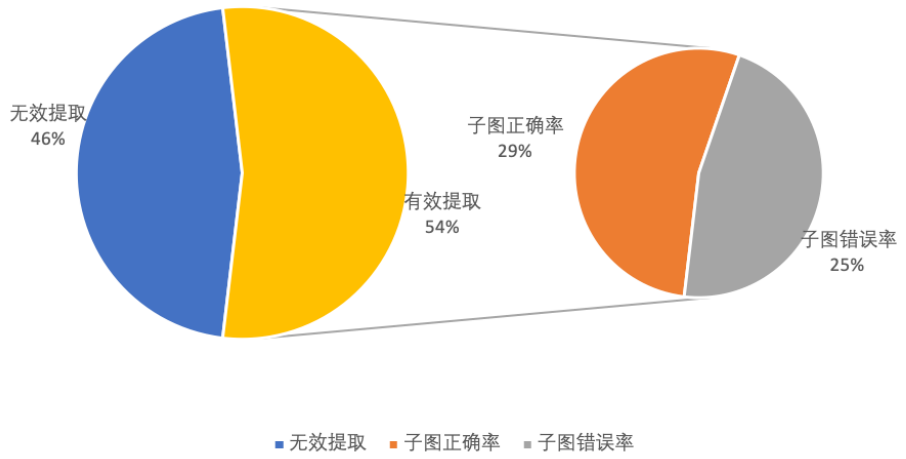
为了衡量子图提取模块的作用，本文提出了子图参与率 *jr* 和子图准确率 *jc* 两

个指标。具体来说，子图参与率是指知识库子图提取模块有效的从问题文本中提取出了核心实体比例。值得注意的是，定义里并不考虑从图像中提取子图，这是因为模型能从所有样本的图像中识别物体，从而构成子图，但是问题文本可能由于太简单而不包含 DBpedia 实体。子图准确率是指针对提取出子图的问题，模型的答案准确率。

实验数据来自于 *KBSN* 在两个数据集上的结果，结果详见表4-16，双饼图见图4-3，左边的饼状图表示子图参与率，右边饼状图表示子图准确率。



(a) VQA2.0



(b) KB-VQA

图 4-3 子图提取模块在两个数据集的表现

表 4-16 知识子图提取模块在 VQA2.0 和 KB-VQA 数据集上的表现

指标	VQA2.0	KB-VQA
子图参与率 *jr*	35.6	54.3
子图准确率 *jc*	46.5	53.4

由图表可知，在 KB-VQA 数据集上，子图提取模块无论是参与率还是正确率都高于 VQA2.0，这解释了 KBSN 与 N-KBSN 之间准确率之差在不同数据集上表现不同的原因。此外，子图提取模块在 KB-VQA 的子图参与率远高于在 VQA2.0 上的表现，这也说明 KB-VQA 的问题更加多样，且更复杂，因此能从中识别出更多的 DBpedia 实体。

4.5 本章小结

本章的主要内容是构建了一个基于知识库图嵌入的视觉问答模型（KBSN），并使用 VQA2.0 和 KB-VQA 两个差异明显的数据集训练和评估模型，实验结果证明了本文提出的知识库的图嵌入模块能提升模型的准确率。特别是在面对需要常识或者知识库的问题时，准确率提升更明显。

本章首先概括性地介绍了目前主流的知识库，随后详细介绍了本文提出的 KBSN 模型。其中，重点介绍了知识库子图提取和知识库子图嵌入的模型基础及架构。

为了获得好的知识库嵌入，本章进行了知识库嵌入实验。基于 DBpedia 知识库，本章通过一系列知识库预处理方法，构建了两个实验知识库：DBV 和 DBA，并使用了包括 TransE 在内的四个知识库嵌入模型在实验知识库上进行链路预测。实验结果证明了 TransE 模型能实现更好的知识库嵌入，并且本文构建的 DBV 知识库中的实体对应性很强，其知识库嵌入效果很好。

基于知识库嵌入的实验结果，本章使用 KBSN 模型在 VQA2.0 和 KB-VQA 数据集上进行视觉问答实验。实验结果显示，相较于 N-KBSN，引入知识库嵌入的 KBSN 在两个数据集的准确率均有所提升，并且在 KB-VQA 上的提升非常明显。子图提取模块分析显示，相比于 VQA2.0，子图提取模块在 KB-VQA 数据集上呈现出更高的子图参与率和子图正确率，这解释了 KBSN 在该数据集上的优异结果。

第五章 全文总结及展望

5.1 全文总结

本文的研究内容是视觉问答模型的构建，其中重点研究了联合嵌入模型和知识库嵌入模型两类。针对已有联合嵌入模型在文本特征化方面的缺陷，本文构建了基于动态词向量的联合嵌入模型（N-KBSN）。实验结果证明了动态词向量的引入能显著的提高视觉问答模型的准确率。进一步，为解决联合嵌入模型固有的网络容量小等缺陷，在 N-KBSN 的基础上，本文构建了知识库图嵌入模块，提出了 KBSN 模型。实验证明知识库图嵌入模型能一定程度地提高预测准确率，在面对复杂问题时效果尤其好。

（1）首先，本文介绍了视觉问答任务和模型架构。本文对视觉问答任务中的问题类型和数据集进行了分类，根据问题回答是否需要常识或外源知识，划分出基于视觉的数据集和基于知识的数据集，并在之后的实验中使用了这两类数据集。视觉问答模型架构一般包括特征提取、注意力机制、特征融合、答案生成等模块。针对已有模型的问题，文本首先改进了文本特征提取部分，之后又引入了知识库图嵌入模块。

（2）本文改进了联合嵌入模型的文本特征化方法，构建了 N-KBSN 模型。现有模型均使用“静态词向量 +LSTM”完成文本特征提取，这种结构无法有效表征一词多义和一词多成分的情况。为了提高文本的表征能力，N-KBSN 模型使用了一个预训练的双层 biLSTM 模型，以获得能感知上下文的动态词向量。在使用 VQA2.0 的视觉问答实验中，本文构建和测试了一系列基于静态词向量和使用不同参数配置的动态词向量模型。实验结果显示，在模型其他部分相同的情况下，基于动态词向量的模型准确率均明显高于静态词向量模型，并且 biLSTM 的网络越深，准确率提升效果越好。本文还使用定量分析和定性分析解释了 N-KBSN 优秀性能的原因。

（3）在 N-KBSN 模型基础上，本文添加了知识图嵌入模块，构建了 KBSN 模型。知识图嵌入模块由知识库子图提取和知识库子图嵌入两部分组成，其中子图提取部分使用 Faster R-CNN 和 spotlight 模型分别从图像和文本中识别关键实体，并从知识库中提取核心实体相关联的子图。为构建子图嵌入模块，本文进行了知识库嵌入实验。实验使用的数据集是本文从 DBpedia 知识库提取的两个实验知识库：DBV 和 DBA。实验结果证明 TransE 模型在 DBV 知识库上的嵌入效果最好，因此本文最终使用 TransE 模型作为 KBSN 的知识库子图嵌入模块。本文使用 VQA2.0

和 KB-VQA 两个数据集训练和测试了 KBSN 模型，实验结果证明知识库图嵌入模块能提高模型的准确率，并且在面对需要常识或者外源知识的问题时，其能显著提高模型的准确率。

5.2 后续工作展望

本文改进的两个视觉问答模型均在视觉问答数据集上取得较好的准确率，但并没有达到当前最佳，仍有一定的改进和提升空间。

（1）图像特征化方法

由于图像特征化方法并不是本文的研究内容，因此本文直接使用预训练的 Faster R-CNN 网络提取图像特征，参数的设计也是迁移了 2017 年 VQA 挑战的冠军模型[60]，仅仅在训练中微调参数。如果要进一步提升模型的准确率，一方面，可以进行网络搜索获得最佳的超参数组合，另一方面，可以使用效果更好的图像识别模型如 Mask R-CNN。

（2）子图提取方法的改进

本文使用的知识库子图提取的基本思路是：先从图像和问题文本中提取核心实体，再从知识库中提取出以核心实体为中心的子图。这种方法仅仅关注了三元组数据的主语，能减少搜索量。但是贪婪的提取方式必然会引入噪音，从而影响性能。因此以后的工作可以改进子图提取方法，例如建立图像核心实体和文本核心实体的关联，裁剪子图无关连接，考虑核心实体在谓语和宾语的情况。

（3）知识库的建立

本文基于 DBpedia 构建了两个知识库：DBV 和 DBA。根据知识库嵌入实验可知，虽然 DBA 有更大的数据量，但模型在其上的链路预测准确率并不高，这说明了 DBA 知识库很难得到优秀的嵌入表示。但即使是准确率最高的 DBV 知识库，TransE 的准确率也不及 70%，这显然成为了后续知识库嵌入的瓶颈。因此后续的工作一方面可以建立语义对应性更强的知识库，另一方面可以试验更好的知识库嵌入方法，例如图神经网络。

（4）基于知识的数据集的丰富

本文使用了 KB-VQA 数据集，其中包含需要常识或外源知识的问题。从 KBSN 在 VQA2.0 和 KB-VQA 的实验结果可以看出，基于知识的数据集对模型推理和引用外源知识的要求更高，能更全面的衡量模型的推理能力。但是目前基于知识的数据集普遍面临的问题有数据集的标准不统一、数据量相对较小。因此增加数据量和丰富数据多样性是未来的研究方向之一。

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确立起自我价值。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，祝愿答辩组和评阅老师阖家幸福。

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] B. Zhou, Y. Tian, S. Sukhbaatar, et al. Simple baseline for visual question answering[J]. arXiv preprint arXiv:1512.02167, 2015,

[5] L. Ma, Z. Lu, H. Li. Learning to answer questions from image using convolutional neural network.[C]. AAAI, 2016, 16

[6] 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

[7] 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

[8] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015,

[9] 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

[10] H. Noh, P. Hongsuck 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

[11] 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

[12] A. Fukui, D. H. Park, D. Yang, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[J]. arXiv preprint arXiv:1606.01847, 2016,

[13] 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,

[14] M. Ren, R. Kiros, R. Zemel. Exploring models and data for image question answering[C]. Advances in neural information processing systems, 2015, 2953-2961

[15] 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

[16] I. Ilievski, S. Yan, J. Feng. A focused dynamic attention model for visual question answering[J]. arXiv preprint arXiv:1604.01485, 2016,

[17] 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

[18] J. Lu, J. Yang, D. Batra, et al. Hierarchical question-image co-attention for visual question answering[M]. Curran Associates, Inc., 2016, 289-297

[19] M. Malinowski, C. Doersch, A. Santoro, et al. Learning visual question answering by bootstrapping hard attention[J]. arXiv preprint arXiv:1808.00300, 2018,

[20] A. Jiang, F. Wang, F. Porikli, et al. Compositional memory for visual question answering[J]. arXiv preprint arXiv:1511.05676, 2015,

[21] 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

[22] C. Xiong, S. Merity, R. Socher. Dynamic memory networks for visual and textual question answering[C]. International conference on machine learning, 2016, 2397-2406

[23] P. Wang, Q. Wu, C. Shen, et al. Explicit knowledge-based reasoning for visual question answering[J]. arXiv preprint arXiv:1511.02570, 2015,

[24] P. Wang, Q. Wu, C. Shen, et al. Fvqa: Fact-based visual question answering[J]. IEEE transactions on pattern analysis and machine intelligence, 2017,

[25] 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

[26] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014,

[27] 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

[28] R. Krishna, Y. Zhu, O. Groth, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73

[29] 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

[30] 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:

[31] 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

[32] 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

[33] T.-Y. Lin, M. Maire, S. Belongie, et al. Microsoft coco: Common objects in context[C]. European conference on computer vision, 2014, 740-755

[34] 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

[35] S. Auer, C. Bizer, G. Kobilarov, et al. Dbpedia: A nucleus for a web of open data[M]. Springer, 2007, 722-735

[36] H. Liu, P. Singh. Conceptnet—a practical commonsense reasoning tool-kit[J]. BT technology journal, 2004, 22(4): 211-226

[37] 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

[38] D. G. Lowe. Object recognition from local scale-invariant features[C]. Proceedings of the seventh IEEE international conference on computer vision, 1999, 1150-1157

[39] N. Dalal, B. Triggs. Histograms of oriented gradients for human detection[C]. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 2005, 886-893

[40] M. D. Zeiler, R. Fergus. Visualizing and understanding convolutional networks[C]. European conference on computer vision, 2014, 818-833

[41] A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems, 2012, 1097-1105

[42] 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

[43] 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

[44] S. Hochreiter, J. Schmidhuber. Long short-term memory[J]. Neural Comput., 1997, 9(8): 1735–1780

[45] T. K. Landauer, P. W. Foltz, D. Laham. An introduction to latent semantic analysis[J]. Discourse processes, 1998, 25(2-3): 259-284

[46] 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

[47] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention[C]. Advances in neural information processing systems, 2014, 2204-2212

[48] D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate[J]. arXiv, 2014, 11: arXiv-1409

[49] 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

[50] J. K. Chorowski, D. Bahdanau, D. Serdyuk, et al. Attention-based models for speech recognition[M]. Curran Associates, Inc., 2015, 577-585

[51] 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

[52] 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

[53] L. Li, S. Tang, L. Deng, et al. Image caption with global-local attention.[C]. AAAI, 2017, 4133-4139

[54] 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

[55] 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

[56] 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,

[57] 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,

[58] 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

[59] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need[M]. Curran Associates, Inc., 2017, 5998-6008

[60] 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

[61] M. E. Peters, M. Neumann, M. Iyyer, et al. Deep contextualized word representations[C]. Proc. of NAACL, 2018,

[62] 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

[63] 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

[64] 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

[65] R. Girshick. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2015, 1440-1448

[66] K. He, G. Gkioxari, P. Dollár, et al. Mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2017, 2961-2969

[67] 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

[68] 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

[69] R. Akerkar, P. Sajja. Knowledge-based systems[M]. Jones & Bartlett Publishers, 2010

[70] S. Nirenburg, J. Carbonell, M. Tomita, et al. Machine translation: A knowledge-based approach[M]. Morgan Kaufmann Publishers Inc., 1994

[71] K. Knight. Building a large ontology for machine translation[C]. Proceedings of the workshop on Human Language Technology, 1993, 185-190

[72] 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,

[73] 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

[74] G. A. Miller. Wordnet: a lexical database for english[J]. Communications of the ACM, 1995, 38(11): 39-41

[75] 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

[76] 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

[77] G. Wiederhold. Intelligent integration of information[C]. ACM SIGMOD Record, 1993, 434-437

[78] V. S. Subrahmanian. Amalgamating knowledge bases[J]. ACM Transactions on Database Systems (TODS), 1994, 19(2): 291-331

[79] 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

[80] 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

[81] 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

[82] 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

[83] D. Vrandečić, M. Krötzsch. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85

[84] 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

[85] F. Mahdisoltani, J. Biega, F. M. Suchanek. Yago3: A knowledge base from multilingual wikipedias[C]. CIDR, 2013,

[86] Dbpedia version 2016-10, 2016. https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10.

[87] T. Berners-Lee. Linked data, 2016. http://www.w3.org/DesignIssues/LinkedData.html.

[88] P. van Bergen. Nli-go dbpedia demo, 2018. https://wiki.dbpedia.org/projects/nli-go-dbpedia-demo.

[89] 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,

[90] 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

[91] 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

[92] A. Bordes, X. Glorot, J. Weston, et al. A semantic matching energy function for learning with multi-relational data[J]. Machine Learning, 2014, 94(2): 233-259

[93] A. Bordes, J. Weston, R. Collobert, et al. Learning structured embeddings of knowledge bases[C]. Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011,

[94] 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

攻读硕士学位期间取得的成果

[1] 郑文锋, 杨波, 陈小兵. 一种基于图像识别的知识库构建方法 [P]. 中国, 发明专利, 201911309368.5, 2019 年 12 月 18 日

[2] 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

[3] X. Ni, L. Yin, X. Chen, et al. Semantic representation for visual reasoning[C]. MATEC Web of Conferences, 2019,