

# 利用图像描述与知识图谱增强表示的视觉问答

王屹超, 朱慕华, 许晨, 张琰, 王会珍\*, 朱靖波

东北大学计算机科学与工程学院自然语言处理实验室, 沈阳, 中国

{yichaow98, zhumuhua}@gmail.com, {xuchenneu, zhangyanneu}@163.com

{wanghuizhen, zhujingbo}@mail.neu.edu.cn

## 摘要

视觉问答作为多模态任务, 需要深度理解图像和文本问题从而推理出答案。然而在许多情况下, 仅在图像和问题上进行简单推理难以得到正确的答案, 事实上还有其它有效的信息可以被利用, 例如图像描述、外部知识等。针对以上问题, 本文提出了利用图像描述和外部知识增强表示的视觉问答模型。该模型以问题为导向, 基于协同注意力机制分别在图像和其描述上进行编码, 并且利用知识图谱嵌入, 将外部知识编码到模型当中, 丰富了模型的特征表示, 增强模型的推理能力。在OKVQA数据集上的实验结果表明本文方法相比基线系统有1.71%的准确率提升, 与先前工作中的主流模型相比也有1.88%的准确率提升, 证明了本文方法的有效性。

**关键词:** 视觉问答; 多模态融合; 知识图谱; 图像描述

## Exploiting Image Captions and External Knowledge as Representation Enhancement for Visual Question Answering

Yichao Wang, Muhua Zhu, Chen Xu, Yan Zhang, Huizhen Wang\*, Jingbo Zhu

NLP Lab, School of Computer Science and Engineering,

Northeastern University, Shenyang, China

{yichaow98, zhumuhua}@gmail.com, {xuchenneu, zhangyanneu}@163.com

{wanghuizhen, zhujingbo}@mail.neu.edu.cn

## Abstract

Visual question answering (VQA), as a multi-modal task, requires deep understanding of images and questions. However, conducting reasoning simply on images and questions may fail in some cases. In fact there exists other information that we can use for the task, such as image captions and external knowledge base. In this paper we propose a novel approach to incorporating information of image captions and external knowledge into VQA models. To utilize image captions, the approach adopts the co-attention mechanism and encodes image captions with the guidance from the question. The approach incorporates external knowledge by using knowledge graph embedding as initialization of word embeddings. Experimental results on the OKVQA dataset show that the proposed approach achieves an improvement of 1.71% over the baseline system and 1.88% over the best-reported previous system.

**Keywords:** Visual Question Answering, Multimodal Fusion, Knowledge Graph, Image Captioning

\* 通信作者: 王会珍 (wanghuizhen@mail.neu.edu.cn)

## 1 引言

近年来,深度学习技术极大推动了包括自然语言处理和计算机视觉在内的人工智能领域的发展,人们希望机器可以像人一样思考、交流。本文关注一项非常具有挑战性的多模态理解任务:视觉问答 (Antol et al, 2015)。视觉问答要求模型在读懂文本问题的同时对图像内容有很好的理解,并利用两种模态的信息获得正确的答案。<sup>0</sup>因此,视觉问答的发展得益于计算机视觉和自然语言处理两个领域的蓬勃发展。在应用方面,视觉问答在帮助视觉障碍人士了解世界,构建智能问答系统提升人机交互体验等场景下有巨大的应用潜力。

目前领域内的视觉问答模型广泛采用双线性编码的范式 (Kim et al, 2018),其基本思想是对图像和文本分别进行编码,如图像领域的VGG (Simonyan et al, 2014)、ResNet (He et al, 2016),自然处理领域Word2vec (Mikolov et al, 2013)、Glove (Pennington et al, 2014)、BERT (Devlin et al, 2018)等等,然后基于注意力机制学习图像和文本问题之间的隐含对齐特征(Yu et al, 2019; Anderson et al, 2018),将图像特征与文本特征进一步融合从而推理出正确的答案。然而在许多情况下,直接从图像和文本问题学习得到的特征表示很可能存在不足,需要引入额外的信息来增强特征表示,从而得到更好的模型性能。如例1所示,对于“男人在玩什么乐器”这个问题在图像描述中可以找到与答案相关的信息,即“电子琴”、“乐器”等,引入图像描述的信息有助于解答该问题。例2则说明引入外部知识的必要性:在图像描述中没有与问题“图中物体有什么用”的相关信息,但是通过描述可以从外部知识间接获得与答案相关的知识信息。将这些信息与图像和问题相关的知识信息编码到模型当中来丰富模型的特征表示,提升模型的推断能力,有助于模型更准确的生成正确答案。



| 例1   | 例2  |
|--|---|
|  |  |
| 问题:男人在玩什么乐器?   | 问题:图中物体有什么用?  |
| 答案:电子琴   | 答案:灭火   |
| 图像描述:一个男人坐着弹电子琴  | 图像描述:一个消防栓立在地上  |
| 外部知识: (电子琴,属于,乐器)  | 外部知识: (消防栓,用作,灭火)   |

表 1. 视觉问答示例: 引入图像描述和外部知识的必要性

为了引入图像描述信息,我们首先利用图像描述生成技术生成图像的“显式”描述,即图像描述,再基于注意力机制以问题为导向生成图像描述的特征表示,从而让模型可以更加充分地学习图像和文本的对齐信息。为了使模型能更好的回答类似例2中的问题,我们引入外部知识库提供额外的问答知识。使用知识库作为答案的来源是利用外部知识库的常用方法 (Wang et al, 2017; Narasimhan et al, 2018; Zhu et al, 2020),但是这种方法依赖于知识库对答案实体的覆盖程度,另外检索所消耗的计算成本也很高。因此本文利用知识图谱嵌入 (Malaviya et al, 2020)对文本问题与图像描述进行编码,从而实现将外部知识引入到问答系统的目标。在基于双线性编码的基准系统上,本文通过从图像中额外提取出来的图像描述和基于ConceptNet知识库 (Liu et al, 2004)学习得到的知识图谱嵌入,为系统增强图像语义信息的同时也融入外部常识信息,提升了模型的推理能力从而更准确的生成答案。在OKVQA (Marino et al, 2019)数据集上进行了广泛的实验,发现本文问答系统的答案准确率与基线系统相比有1.71%的性能提升,

©2021 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

<sup>0</sup>根据视觉信息的来源不同可以把视觉问答分为图像问答和视频问答两个不同的子任务,本文工作关注前者。

与先前工作中所提的主流模型相比也有1.88%的显著提升。本文的主要贡献点可以总结如下：

- (1) 本文采用注意力机制将图像描述信息融入到视觉问答模型当中，增强图像与问题之间的对齐信息。
- (2) 本文采用知识图谱嵌入对问答和图像描述中的文本信息进行编码，从而实现将外部知识融入到系统当中，提升模型的推理能力。
- (3) 通过OKVQA数据集对本文所提出的模型进行有效性验证，在答案准确率方面，模型效果与基线模型相比有显著提升。

## 2 相关工作

视觉问答 (Antol et al, 2015)任务提出以来，受到了自然语言处理领域和计算机视觉领域研究者的广泛关注。目前主流的视觉问答模型采用双线性编码的范式 (Kim et al, 2018)，即图像与文本问题对应各自领域编码方法。为了学习图像和文本之间的内在联系，注意力机制被用来捕捉图像和自然语言之间的隐含对齐，取得了一定的效果。与直接将整张图像进行编码的方法不同，Anderson et al. (2018)通过“自下而上”的方法确定图像中包含的物体区域框，利用文本问题与每个区域框相似度计算得出注意力权重，最后基于注意力机制模型学习到图像区域与文本问题之间的隐含对齐。该方法一经提出更引起了视觉问答领域研究人员大量的关注，后续的工作都在注意力基础上做了相应的改进 (Yu et al, 2019; Guo et al, 2020)。本文在捕获图像和问题的“隐式”关系的同时，将图像描述作为“显式”的语义信息并且基于注意力机制学习图像描述与问题之间的“显式”对齐，增强模型对图像与文本的编码能力。

视觉问答的另一个重要研究方向在于如何将知识库信息引入到问答系统中。Wang et al. (2017)、Narasimhan et al. (2018)等工作将知识库引进视觉问答模型，利用知识库中的实体拓展答案的候选集合。具体来说，Wang et al. (2017)所采用的数据集在原本的图像-问题-答案三元组基础上额外分配一个事实知识来支撑答案的推理。视觉问答模型需要通过在知识库中选择对应的知识实体作为答案。Narasimhan et al. (2018)、Zhu et al. (2020)等则是利用在图像提取到的物体、场景、动作等信息在知识库中检索最相关的知识，并推理出答案。基于知识库的视觉问答优点在于模型不再局限于固定的答案集合，而是利用知识库动态提供答案候选。但这种基于检索的方式使模型性能依赖于识别物体、场景等子模块的效果，并且检索知识占据大部分运行时间。本文通过知识图谱嵌入 (Malaviya et al, 2020)引入知识库信息，在运行效率得到保证的同时为模型融入常识知识，提升模型推理能力。

## 3 本文方法

本节首先介绍视觉问答任务的形式化定义，然后介绍本文提出的视觉问答模型。如图1所示，该模型主要由表示层、多模态注意力层和输出层三部分所组成。

### 3.1 问题定义

在本文设定的视觉问答任务中，问答系统的输入包括预先给定的图像 $I$ 和文本问题 $Q$ ，以及根据图像自动生成的图像描述 $C$ 和额外引入的外部知识图谱 $G$ （这里将外部知识信息表示为 $E = \{G, C\}$ ）。视觉问答系统的目标是从答案集合 $A$ 中得到满足下列公式的最优答案 $\hat{a}$ ：

$$\hat{a} = \arg \max_{a \in A} p_{\theta}(a|I, Q, E) \quad (1)$$

其中 $p_{\theta}(a|I, Q, E)$ 表示在给定 $I, Q, E$ 的前提下，生成答案 $a \in A$ 的条件概率， $\theta$ 为可训练参数。

### 3.2 模型结构

本文使用Yu et al. (2019)作为基线模型，该模型采用双线性编码范式，基于协同注意力机制使模型学习文本问题与图像之间的隐式对齐。在基线模型的基础上，本文所提的模型框架如图1所示，与基准系统相比本文方法将额外生成的图像描述作为输入，利用协同注意力学习文本问题与图像描述的对齐信息，丰富模型表示，此外还引入知识图谱嵌入对文本进行编码，为模型融入额外的知识信息。



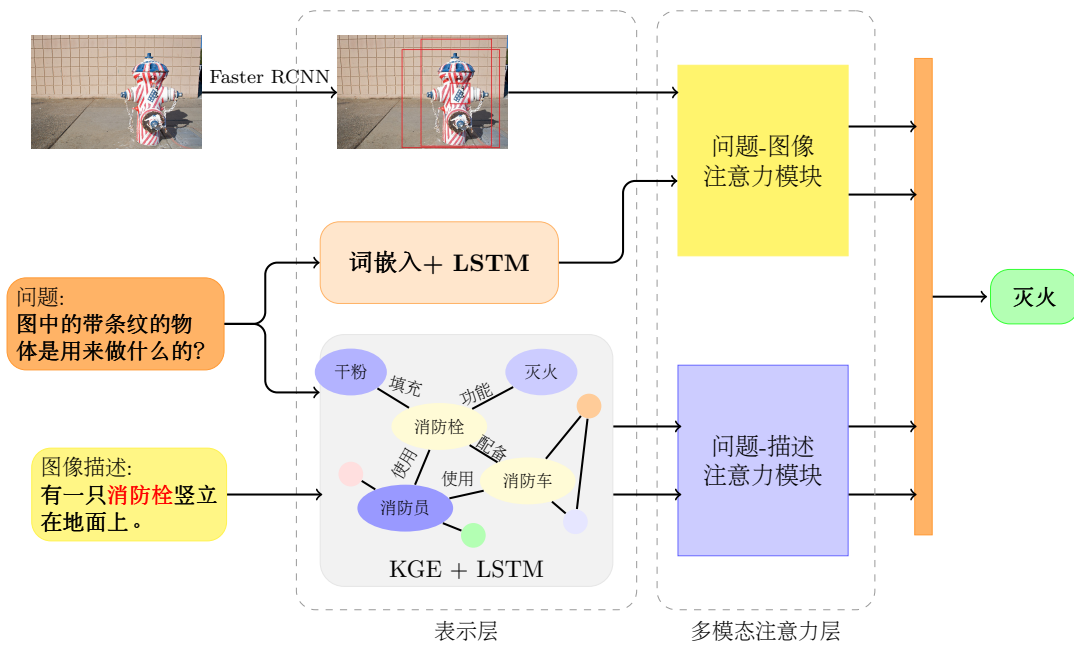


图 1. 视觉问答系统的整体框图

### 3.2.1 表示层

如图1(左虚线框)所示，模型的表示层包括图像、文本问题、图像描述，分别对应模型的三种输入。表示层的主要目的是将模型的输入映射到一定维度的向量特征，这些特征再被应用于模型训练中。下面将对三类输入表示方法分别进行论述。在具体介绍这三种表示之前，我们首先介绍本文采用的知识图谱以及图谱嵌入 (KGE, Knowledge Graph Embedding) 表示学习方法。

#### 知识图谱嵌入表示学习

我们使用ConceptNet (Liu et al, 2004)作为外部知识来源，ConceptNet是一个多语言的知识库，表示单词或者短语实体之间的常识知识 (包含维基百科、开放常识、游戏 (Singh et al, 2002)等)。本文从ConceptNet知识库中抽取与问答数据集语言相关的150万个节点，并且利用 (Malaviya et al, 2020)中的表示学习方法将节点所对应的单词转化成维度为 $d_k$ 的特征表示。

#### 图像表示

本文对图像表示采用“自下而上”的区域编码方法 (Anderson et al, 2018): 利用Visual Genome数据集 (Krishna et al, 2017)预训练得到Faster R-CNN模型 (Ren et al, 2015)，基于该模型提取图像中的区域对象。最终一张图像可以表示为一个特征向量 $V \in \mathbb{R}^{n_v \times d_v}$ 和区域位置坐标 $B \in \mathbb{R}^{n_b \times d_b}$ :

$$V, B = \text{FasterRCNN}(I) \tag{2}$$

其中 $V = \{v_i | i = 1, 2, \dots, n_v\}$ ,  $B = \{b_i | i = 1, 2, \dots, n_b\}$ , 且 $v_i \in \mathbb{R}^{d_v}$ ,  $b_i \in \mathbb{R}^{d_b}$ 。  $n_v = n_b$ 表示区域框数量， $d_v$ 和 $d_b$ 表示区域框特征维度与位置坐标维度。

#### 文本问题表示

文本问题的表示分拆为两个并行的视角，每个视角由两个阶段组成。视角1利用词嵌入Glove (Pennington et al, 2014)将文本转换成特征向量，然后利用长短时记忆网络 (Long Short-Term Memory, LSTM) 进一步生成包含上下文信息的特征向量；另一视角则采用知识图谱嵌入 (Zhu et al, 2020)初始化词向量，再利用长短时记忆网络进行表征学习。问题 $Q = \{q_i | i = 1, 2, \dots, n_q\}$ 经过上述转换过程得到的特征向量 $X_q \in \mathbb{R}^{n_q \times d_x}$ 形式化表示如下:

$$X_{qq} = \text{LSTM}(\text{Glove}(Q)) \tag{3}$$

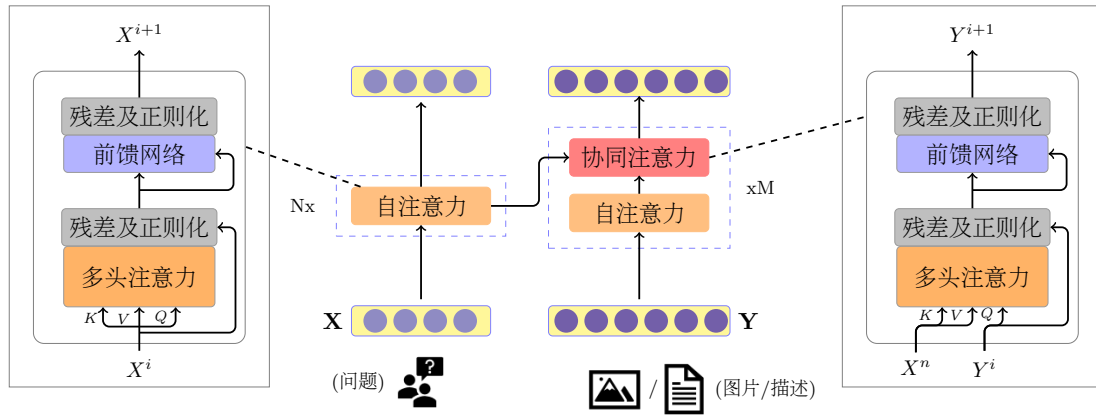


图 2. 多模态注意力模块

$$X_{qk} = LSTM(KGE(Q)) \quad (4)$$

其中  $X_{qg} \in \mathbf{R}^{n_q \times d_g}$  对应采用Glove词嵌入得到的特征表示,  $X_{qk} \in \mathbf{R}^{n_q \times d_k}$  表示采用知识图谱嵌入编码得到的特征表示,  $n_q$  表示问题的长度,  $d_g$  和  $d_k$  表示Glove词嵌入和知识图谱嵌入的维度。

### 图像描述表示

首先利用图像描述生成模型 (luo et al, 2018) 生成图像描述  $C \in \{c_i | i = 1, 2, \dots, n_c\}$ , 这里  $n_c$  表示图像描述的长度。与问题的表示方法不同, 本文对图像描述的表示只采用知识图谱嵌入将文本映射成特征向量, 然后经过长短期记忆网络产生图像描述的特征表示  $X_{ck} \in \mathbf{R}^{n_c \times d_k}$ :

$$X_{ck} = LSTM(KGE(C)) \quad (5)$$

### 3.2.2 多模态注意力层

多模态注意力层 (图1右虚线框) 包含问题-图像注意力模块、问题-描述注意力模块, 通过注意力机制学习同模态或者不同模态表示之间的交互信息。多模态注意力模块实现细节如图2所示, 问题-图像注意力模块与问题-描述注意力模块具有相同的实现方式, 区别只在于将图像表示替换为图像描述表示, 因此本文只以问题-图像注意力模块为例进行说明。该模块采用类似编码器-解码器的结构形式 (Yu et al, 2019), 首先左侧通过N层自注意力机制对文本问题进行编码, 学习文本问题的自注意力特征, 使模型对问题有一定的理解; 另一侧图像先经过自注意力编码, 学习图像自身的特征表示, 该特征表示与经过N层自注意力编码的文本问题特征表示作为协同注意力的输入计算得到在文本问题指导下的图像的多模态特征表示。经过M层解码器之后, 最终整个模块的输出为: (编码器端得到的) 文本问题的自注意力特征表示和 (解码器端得到的) 问题导向下的图像特征表示; 将图像替换为图像描述, 文本问题与图像描述作为问题-描述注意力模块的输入, 经过上述的计算过程, 问题-描述注意力模块的最终输出为: (编码器端得到的) 文本问题的自注意力特征表示和 (解码器端得到的) 问题导向的图像描述特征表示。

**自注意力模块** 包含多头注意力层、正则化与残差链接层、前向层。以文本问题自注意计算为例, 如图2输入的文本特征或视觉特征  $X \in \mathbf{R}^{n_q \times d_x}$  经过矩阵映射之后得到相应的查询矩阵  $Q \in \mathbf{R}^{n_q \times d_{query}}$ , 键值矩阵  $K \in \mathbf{R}^{n_q \times d_{key}}$  和实值矩阵  $V \in \mathbf{R}^{n_q \times d_{value}}$ , 这里  $d_{query} = d_{key} = d_{value} = d$ 。注意力层采用缩放点积运算, 其计算方法如下所示:

$$Q = XW^Q, K = XW^K, V = XW^V \quad (6)$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (7)$$

为了进一步提高表示能力，可以采用**多头注意力机制** (Vaswani et al, 2017)。多头注意力包含 $h$ 个注意力运算，每个注意力运算对应了缩放点积运算，将运算结果拼接成为多头注意力层的输出表示：

$$MAttention(Q, K, V) = Concat(Attention_1, Attention_2, \dots, Attention_h)W^o, \quad (8)$$

这里 $W^o \in \mathbf{R}^{(d \times h) \times d}$ 是可训练参数。多头注意力层的输出 $f \in \mathbf{R}^{n_q \times d}$ 再经过**残差链接与层正则化**以防止梯度消失和加速模型收敛：

$$f = LayerNorm(X + MAttention(Q, K, V)) \quad (9)$$

经过前向层后得到自注意力模块的最终输出为：

$$Z = LayerNorm(f + FFN(f)) \quad (10)$$

**协同注意力模块**与自注意力模块不同的之处在于查询矩阵，键值矩阵和实质矩阵的生成方式。如图2所示， $Y \in \mathbf{R}^{m \times d_y}$ 作为指导特征与 $X \in \mathbf{R}^{n \times d_x}$ 生成对应的矩阵：

$$Q = XW_Q, K = YW_K, V = YW_V. \quad (11)$$

之后的运算与自注意力模块相同，在协同注意力模块中 $X$ 文本问题特征， $Y$ 为图像或图像描述特征，因为这里的 $X$ 和 $Y$ 可以是不用维度的特征，所以能够进行多个模态特征之间的注意力学习。

### 3.2.3 输出层

上述模型得到了四部分的输出，分别是：(1)问题引导的图像特征 $V \in \mathbf{n}_v \times \mathbf{d}_v$ ；(2)问题自注意力特征 $X_q \in \mathbf{n}_q \times \mathbf{d}_q$ ；(3)问题引导并融入知识图谱表示的图像描述特征 $x_c \in \mathbf{n}_c \times \mathbf{d}_k$ ；(4)融入知识图谱表示的问题自注意力特征 $X_k \in \mathbf{n}_q \times \mathbf{d}_k$ 。将这四部分特征通过线性层映射到统一维度，并且利用加和的融合方式生成最终的向量，然后将该特征向量送入到与答案集合长度相同的分类器当中，分类得到预测结果。

## 4 实验

### 4.1 实验数据

我们希望实验所用的数据集在获取答案时需要借助额外知识信息。而先前大多数的数据集都对额外知识施加一些约束，使答案的推理预测相对简单，例如KB-VQA数据集 (Wang et al, 2015)使用模板生成的方式生成答案；FVQA数据集 (Wang et al, 2017)规定只能利用该数据集给定的知识库，不能具备很好的通用性。因此为了验证方法的有效性，本文选用OKVQA数据集 (Marino et al, 2019)。

OKVQA数据集包含14,031张图像和14,055个自然语言（英文）问题，文本问题的平均长度为6.8个单词，答案的平均长度为2.0个单词。我们把问题集合划分训练集和验证集，数量分别为9009和5046。另外该数据集答案分为以下11个类别：车辆和运输工具（VT）；品牌、公司和产品（BCP）；物品、材料和服装（OMC）；体育运动和娱乐活动（SR）；烹饪和事物（CF）；地理、历史、语言和文化（GHLC）；人与日常生活（PEL）；动植物（PA）；科学技术（ST）；天气和气候（WC）以及Other类。

### 4.2 实验设置

**表示层：**针对输入的图像，本文采用Anderson et al. (2018)的方法，基于FasterRCNN (Ren et al, 2015)从图像中提取区域框特征，设定最大区域框数量 $n_v = 100$ ，每个图像特征维度 $d_v = 2048$ 。图像描述的生成则采用Luo et al. (2018)的方法，本文实验中保留得分最高的2个图像描述结果。对文本问题的表示利用Glove (Pennington et al, 2014)将每个词映射成300维的词嵌入，另外还利用知识图谱嵌入 (Malaviya et al, 2020)将问题和图像描述中的词映射到 $d_k = 1024$ 维的向量空间。文本问题和图像描述最大长度为 $n_q = 20, n_c = 14$ 。随后利用LSTM网络将文本向量维度即问题向量、图像描述向量和知识图谱向量均映射到 $d_q = d_k = 2048$ 。

| 超参数名称         | 取值   |
|---------------|------|
| Epoch         | 20   |
| Batch size    | 32   |
| Dropout       | 0.1  |
| Warm up       | 0.1  |
| Learning rate | 5e-5 |
| Loss function | BCE  |
| Optimizer     | Adam |

表 2. 本文实验中所涉及的超参数设置

**多模态注意力层：**编码器和解码器层数分别为 $N=6$ 以及 $M=6$ 。多头注意力机制中，设置 $head = 8$ 。在计算注意力时查询矩阵、键值矩阵和实值矩阵的隐藏层维度 $d = 1024$ 。

**输出层：**视觉问答任务中答案来自于训练集出现的答案集合，本文实验答案来自训练集中出现次数为3次及以上的答案集合，答案集合的大小为 $d_{ans} = 2255$ 。实验中涉及的超参数如表2所示，本试验采用二分类交叉熵(BCE, Binary Cross Entropy)损失函数，优化算法为Adam。

本文使用视觉问答挑战赛(Aishwarya et al, 2017)中提出的评价标准来评估我们提出的模型有效性：

$$Acc(ans) = \min\left(1, \frac{\#\{humans\ provided\ ans\}}{3}\right) \quad (12)$$

每个问题对应了10个给定的答案（可相同或不同），上式表明，每个预测答案所得到的分数为该预测答案匹配到的真实答案数量除3，然后与1取最小值。

### 4.3 主实验结果

在OKVQA数据集上的主实验结果如表3所示。本文使用Yu et al. (2019)作为基线模型，并在基线系统的基础上实现了本文所提方法。此外为了更好地体现本文方法的效果，我们在表3中汇总了在OKVQA数据集上开展的相关工作，包括BAN (Kim et al, 2018)、Mucko (Zhu et al, 2020)以及MUTAN (Marino et al, 2019)。从实验结果我们可以看出，本文实现的基线系统要显著优于先前工作，具体地说，基线系统与最好的先前工作Mucko的性能基本可比，甚至要略优于Mucko。而本文所提方法与基线相比，则是获得准确率有1.71%的增长，同时在细粒度类别上，相比于基线模型绝大部分类别都有准确率上的提升，特别是科学技术（ST），天气和气候（WC），物品、材料和服装（OMC）分别有7.86%、4.97%和4.30%的性能提升。另外我们也可以看到本文方法获得了在OKVQA数据集上的最好性能。在细分类别方面，模型也在车辆和运输工具（VT）；物品、材料和服装（OMC）；烹饪和事物（CF）；人与日常生活（PEL）；科学技术（ST）；天气和气候（WC）和Other等大多数类别上为最好的结果。

| Model    | ALL          | VT           | BCP          | OMC          | SR           | CF           | GHLC         | PEL          | PA           | ST           | WC           | Other        |
|----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| BAN      | 25.17        | 23.79        | 17.67        | 22.43        | 30.58        | 27.90        | 25.96        | 20.33        | 25.60        | 20.95        | 40.16        | 22.46        |
| MUTAN    | 26.41        | 25.36        | 18.95        | 24.02        | 33.23        | 27.73        | 17.59        | 20.09        | 30.44        | 20.48        | 39.38        | 22.46        |
| B+AN     | 25.61        | 24.45        | 19.88        | 21.59        | 30.79        | 29.12        | 20.57        | 21.54        | 26.42        | 27.14        | 38.29        | 22.16        |
| M+AN     | 27.84        | 25.56        | <b>23.95</b> | 26.87        | 33.44        | 29.94        | 20.71        | 25.05        | 29.70        | 24.76        | 39.84        | 23.62        |
| B+oracle | 27.59        | 26.35        | 18.26        | 24.35        | 33.12        | 30.46        | <b>28.51</b> | 21.54        | 28.79        | 24.52        | 41.40        | 25.07        |
| M+oracle | 28.47        | 27.28        | 19.53        | 25.28        | 35.13        | 30.53        | 21.56        | 21.68        | <b>32.16</b> | 24.76        | 41.40        | 24.85        |
| Mucko    | 29.02        | -            | -            | -            | -            | -            | -            | -            | -            | -            | -            | -            |
| 基线系统     | 29.19        | 26.35        | 22.79        | 25.09        | <b>40.00</b> | 30.78        | 26.1         | 24.63        | 30.26        | 24.76        | 37.67        | 26.79        |
| 本文方法     | <b>30.90</b> | <b>28.31</b> | 21.86        | <b>29.39</b> | 39.37        | <b>34.04</b> | 27.52        | <b>25.51</b> | 29.79        | <b>32.62</b> | <b>42.64</b> | <b>29.34</b> |

表 3. 主试验结果



| Model | ALL   | VT    | BCP   | OMC   | SR    | CF    | GHLC  | PEL   | PA    | ST    | WC    | Other |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 本文方法  | 30.90 | 28.31 | 21.86 | 29.39 | 39.37 | 34.04 | 27.52 | 25.51 | 29.79 | 32.62 | 42.64 | 29.34 |
| -KGE  | 30.28 | 28.4  | 21.63 | 25.79 | 39.93 | 33.42 | 22.55 | 23.18 | 30.74 | 24.29 | 42.48 | 30.08 |
| -Cap  | 29.30 | 26.67 | 23.72 | 25.51 | 39.96 | 30.81 | 24.26 | 24.21 | 30.37 | 24.76 | 39.53 | 26.85 |

表 4. 消融实验

## 5 实验分析

### 5.1 消融实验对比

表4展现了消融实验的结果，从中可以看到在去掉知识图谱表示和图像描述之后模型的性能都所下降，这也证明融入图谱知识和图像描述的有效性。

具体而言，在仅去掉知识图谱表示的情况下，虽然模型总体性能仅下降0.68%，但是细粒度类别上例如：物品、材料和服装（OMC）；地理、历史、语言和文化（GHLC）；科学技术（ST）类别上分别下降了3.60%，4.97%和8.33%，这说明一些特殊的领域，知识信息起着相对更为重要的作用，相反一些领域例如车辆和运输工具（VT）和体育运动和娱乐活动（SR）仅凭借图像信息也可以很好的推理出答案。另外，在仅去掉图像描述信息之后，模型的整体准确率下降较为明显，在细粒度类别方面基本或多或少性能都有下降。体现了图像描述对于模型提供的有利信息更具有普遍性。另一种角度来说，图像描述是图像的自然语言表示，这种显式的表示与自然语言问题之间的交互相对容易，可以弥补一些跨模态特征之间的“语义鸿沟”。

### 5.2 词嵌入的影响

为了分析采用知识图谱嵌入（KGE）的作用，我们分别利用Glove词嵌入和随机初始化同维度词嵌入替换，并与采用知识图谱嵌入的结果进行对比。分析结果如下表所示：

| 模型           | 整体准确率 |
|--------------|-------|
| 本文方法(KGE)    | 30.90 |
| 本文方法(Glove)  | 30.28 |
| 本文方法(random) | 25.87 |

表 5. 不同词嵌入的结果对比

从结果可以看到，将知识图谱嵌入替换为Glove词嵌入后，模型性能略有下降(0.62%)。而当随机初始化同维度的词嵌入后，模型的效果有5.03%的明显下降。模型中所有文本使用Glove编码后，编码的信息较单一，导致模型性能略有下降，而随机初始化向量没有经过大规模语料训练，很难有较好的表示能力，使模型性能下降明显。

### 5.3 图像描述数量的影响

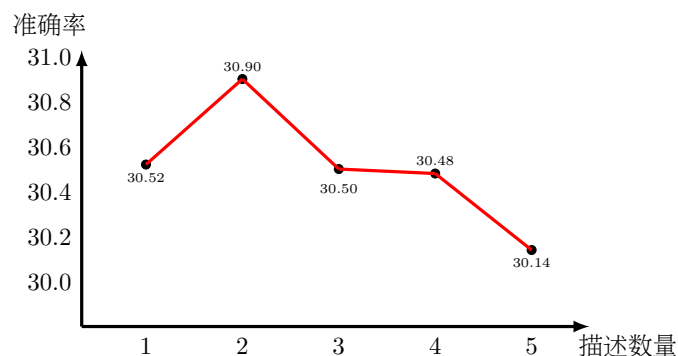


图 3. 不同图像描述数量的对比结果



图像的描述不是唯一的，图像描述生成可以为一张图像产生多个描述。这里我们比较图像描述数量对模型的影响。实验中，使用类似Bert(Devlin et al, 2018)对多句数据的处理方式，将同个图像的不同描述直接进行拼接，图像描述之间利用‘< SEP >’做间隔。图像描述的数量与模型整体准确率的关系如图3所示，从图上结果可以看到从图像生成的描述信息所表达的意思是相近的，因此加入过多的描述信息会引入额外的噪音；仅加入top1的图像描述信息引入的有用信息可能相对有限。经过上述的实验验证， $n = 2$ 时模型的效果最佳。

### 5.4 错误分析


|   |  |   |
|---|--|---|
|    |    |    |
| Q:What type of plane this?  | Q:What type of event is happening?   | Q:What kind of gear do you need to preform this activity?                             |
| C:A plane flying in a cloudy blue sky.  | C:A table that has a bunch of cakes on it.   | C:A person riding skis down a snow covered slope.                                     |
| A:Jet<br>P:Jet  | A:Party<br>P:Party   | A:skis<br>P:skis  |
|  |  |  |
| Q:What type of bike is on the ground?   | Q:What kind of juice?  | Q:What is the small white outdoor house like building called?                         |
| C:A group of people riding motorcycles in a parking lot.                            | C:A plate of food on a dining table.   | C:A park area with green grass and trees.   |
| A:Bmx<br>P:Motorcycles  | A:Orange<br>P:Orange Juice   | A:Gazebo<br>P:Garden  |

表 6. 错误分析

我们对错误案例做了人工分析，如表6所示，其中Q表示给定的问题，C是利用Luo et al.(2018)生成的图像描述，A代表真实答案，P代表预测答案。一方面，在预测正确的案例中可以观察到，类似于“Jet”、“Party”这种不能由图像和问题直接推导出的答案（即基准系统预测错误）得到了正确的预测，知识图谱嵌入的应用起到了一定的效果，图像描述中出现的“skis”关键词信息也帮助模型预测“skis”答案；另一方面，在预测错误的案例中，不难发现图像描述在一些情况下起到的不一定是积极的作用，例如在图像描述中有“Motorcycles”关键词，而问题问到的有关“bike”的内容，这样就会误导系统的判断。任务的评判也存在一定的误差，例如真实答案为“Orange”而预测答案“Orange Juice”。根据所问的问题“‘What kind of juice?’”，系统的回答人为判定是正确的，但是评价标准的原因造成误判。另外，类似“Garden”这种图片中的细粒度场景很难被模型捕捉到，也是造成模型预测错的主要原因之一。综上所述，可能导致模型推断错误的原因总结如下：

- (1) 问题指导下的细粒度场景识别错误。模型没有正确的关注在当前问题下应该关注的图像区域。

- (2) 评价方式导致的错误: 系统所给出的答案, 人工判断是正确的, 然而基于自动化评价方法的原因, 只要预测答案与真实答案没有完全匹配即判为错误。
- (3) 知识库在细粒度类别下覆盖不完全。虽然知识图谱的使用使某些类别正确率明显上升, 但是个别类别下的覆盖率不够, 甚至被模型当做噪声处理。

## 6 总结与展望

本文基于注意力机制利用图像描述增强图像与文本间的对齐表示, 同时利用知识图谱嵌入为模型融入外部知识, 提高模型推理能力。本文的方法在OKVQA数据集进行了有效性验证, 与前人工作相比准确率有了1.88%的显著提升。后续消融试验表明为模型融入图像描述信息和图谱外部知识都促进了模型性能的提升。

未来的工作中我们将关注于如何让模型更准确的捕捉图像的细粒度区域以及如何通过额外知识为不同的细粒度类别提供“专门”的知识信息。另外, 优化视觉问答任务的评价方法, 使评价方法更符合人工的评判标准, 也是比较有意义的方向。

## 参考文献

- Antol, Stanislaw and Agrawal, Aishwarya and Lu, Jiasen and Mitchell, Margaret and Batra, Dhruv and Zitnick, C Lawrence and Parikh, Devi. 2015. *Vqa: Visual question answering*, International Journal of Computer Vision, 123(1), 4-31.
- Anderson, Peter and He, Xiaodong and Buehler, Chris and Teney, Damien and Johnson, Mark and Gould, Stephen and Zhang, Lei. 2018. *Bottom-up and top-down attention for image captioning and visual question answering.*, In Proceedings of the IEEE conference on computer vision and pattern recognition, 6077-6086.
- Yu, Zhou and Yu, Jun and Cui, Yuhao and Tao, Dacheng and Tian, Qi. 2019. *Deep modular co-attention networks for visual question answering.*, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6281-6290.
- Wang, Peng and Wu, Qi and Shen, Chunhua and Dick, Anthony and Van Den Hengel, Anton. 2017. *Fvqa: Fact-based visual question answering.*, IEEE transactions on pattern analysis and machine intelligence, 40(10), 2413-2427.
- Narasimhan, Medhini and Lazebnik, Svetlana and Schwing, Alexander G. 2018. *Out of the box: Reasoning with graph convolution nets for factual visual question answering.*, arXiv preprint arXiv:1811.00538.
- Narasimhan, Medhini and Schwing, Alexander G. 2018. *Straight to the facts: Learning knowledge base retrieval for factual visual question answering*, Proceedings of the European conference on computer vision (ECCV), 451-468.
- Zhu, Zihao and Yu, Jing and Wang, Yujing and Sun, Yajing and Hu, Yue and Wu, Qi. 2020. *Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering*, arXiv preprint arXiv:2006.09073.
- Malaviya, Chaitanya and Bhagavatula, Chandra and Bosselut, Antoine and Choi, Yejin. 2020. *Common-sense knowledge base completion with structural and semantic context*, Proceedings of the AAAI Conference on Artificial Intelligence, 2925-2933.
- Liu, Hugo and Singh, Push. 2004. *ConceptNet—a practical commonsense reasoning tool-kit*, volume 22. Springer, 211-226.
- Marino, Kenneth and Rastegari, Mohammad and Farhadi, Ali and Mottaghi, Roozbeh. 2019. *Ok-vqa: A visual question answering benchmark requiring external knowledge*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3195-3204.
- Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, Jian. 2015. *Faster r-cnn: Towards real-time object detection with region proposal networks*, arXiv preprint arXiv:1506.01497
- Jiang, Huaizu and Misra, Ishan and Rohrbach, Marcus and Learned-Miller, Erik and Chen, Xinlei. 2020. *In defense of grid features for visual question answering*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10267-10276.

- Guo, Wenya and Zhang, Ying and Wu, Xiaoping and Yang, Jufeng and Cai, Xiangrui and Yuan, Xiaojie. 2020. *Re-Attention for Visual Question Answering*, volume 34. Proceedings of the AAAI Conference on Artificial Intelligence, 91–98.
- Kim, Jin-Hwa and Jun, Jaehyun and Zhang, Byoung-Tak. 2018. *Bilinear attention networks*, arXiv preprint arXiv:1805.07932.
- Simonyan, Karen and Zisserman, Andrew. 2014. *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556.
- He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian. 2016. *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778.
- Pennington, Jeffrey and Socher, Richard and Manning, Christopher D. 2014. *Glove: Global vectors for word representation*, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532–1543.
- Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey. 2013. *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781.
- Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805.
- Krishna, Ranjay and Zhu, Yuke and Groth, Oliver and Johnson, Justin and Hata, Kenji and Kravitz, Joshua and Chen, Stephanie and Kalantidis, Yannis and Li, Li-Jia and Shamma, David A and others. 2017. *Visual genome: Connecting language and vision using crowdsourced dense image annotations*, volume 123. International journal of computer vision, 32–73.
- Teney, Damien and Anderson, Peter and He, Xiaodong and Van Den Hengel, Anton. 2018. *Tips and tricks for visual question answering: Learnings from the 2017 challenge*, Proceedings of the IEEE conference on computer vision and pattern recognition, 4223–4232.
- Singh, Push and others. 2002. *The public acquisition of commonsense knowledge*, Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access.
- Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Lukasz and Polosukhin, Illia. 2017. *Attention is all you need*, arXiv preprint arXiv:1706.03762.
- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. *Vqa: Visual question answering*, Int. J. Comput. Vision, 123(1):4–31.
- Marino, Kenneth and Rastegari, Mohammad and Farhadi, Ali and Mottaghi, Roozbeh. 2019. *Ok-vqa: A visual question answering benchmark requiring external knowledge*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3195–3204.
- Wang, Peng and Wu, Qi and Shen, Chunhua and Hengel, Anton van den and Dick, Anthony. 2015. *Explicit knowledge-based reasoning for visual question answering*, arXiv preprint arXiv:1511.02570.
- Luo, Ruotian and Price, Brian and Cohen, Scott and Shakhnarovich, Gregory. 2018. *Discriminability objective for training descriptive captions*, arXiv preprint arXiv:1803.04376.
- Zhu, Z. and Yu, J. and Wang, Y. and Sun, Y. and Wu, Q. 2020. *Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering*, Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence IJCAI-PRICAI-20.