**郑州大学毕业设计（论文）**

题目：基于指称表达式所指代物体的视觉问题回答系统设计

指导教师： 田侦 职称：

指导教师(校外)： 职称：

学生姓名： 谢梓聪 学号： 201824100445

专 业： 软件工程

院（系）： 计算机与人工智能学院

完成时间： 2022年5月12日

2022年 5 月 12 日

基于指称表达式所指代物体的视觉问题回答系统设计

摘要

近年来深度学习的出现极大程度地促进了人工智能的发展，人工智能又极大地便利了人类的生活。在众多人工智能应用当中人机交互需求最为广泛，因此同时结合图像文本语音等多模态的人工智能变得越来越重要。文本是反映人类的高层抽象语义信息的载体，图像是反映人类视觉系统的成像，如何同时结合文本和图像模拟和人类的实际交互是近年人工智能的一大难题。

视觉问题回答和指称表达式理解都是近年来视觉-语言人工智能研究领域的主流任务，视觉问题回答是指给定一副图像回答一个给定的与该图像相关的问题，指称表达式理解是指给定一副图像和一个指代某个图像物体的表达式来确定物体在图像的具体位置。

该篇论文旨在对于当前视觉问题回答任务的主流形式较为简单和较难用于实际的问题进行解决，提出一个新的双阶段任务形式，给定一副图片，先通过指称表达式指代一副图片中某个感兴趣的物体，若系统反馈找到了这样的物体，再提出一个对该物体的问题用于系统回答。该任务想法来源于真实的人机交互场景，可为盲人提供帮助，盲人可通过该任务的两次交互了解周围情况，同时该项任务可作为基础任务为后续高级人机交互任务如视觉语言导航提供很大帮助。针对提出的任务，本文使用了大型文本图像预训练模型VL-BERT在这两个下游任务的标准数据集上分别进行微调，再将两个任务的实时预测结合起来，并且做了一个双阶段的交互式的系统用于任务理解和结果可视化。综上可知本文研究成果以及贡献有如下三点：

（1）相比当前视觉问题回答形式提出更细致更易应用于实际人机交互场景的任务：基于指称表达式所指代物体的视觉问题回答。

（2）针对该任务提出使用大型文本图像预训练开源模型VL-BERT分别在指称表达式理解和视觉问题回答任务两个标准数据集上进行微调，在预测阶段将两个任务预测结果结合来完成该任务。

（3）针对提出任务和所用模型使用Vue、ElementPlus和Django开发一个交互式的前后端分离的Web系统用于任务理解和结果可视化。

关键词：视觉问题回答，指称表达式理解，人机交互，VLBERT，视觉语言

**Abstract**

In recent years, the emergence of deep learning has greatly facilitated the development of artificial intelligence, which in turn has greatly facilitated human life. Among the many AI applications, human-computer interaction is the most widely demanded, so it is becoming more and more important to combine multimodal AI at the same time. Text is the carrier of high semantic information reflecting human language, and image is the imaging reflecting human visual system, how to combine both text and image to achieve actual human interaction is a major challenge of AI in recent years.

Both visual question answering and referring expression comprehension are mainstream tasks in the field of visual-linguistic AI research in recent years. Visual question answering refers to answering a given question related to a given image, and referring expression comprehension refers to determining the specific location of an object given a image and an expression referring to an object in the image.

This paper aims to solve the problem that the current mainstream form of visual question answering tasks is simple and difficult to use in practice. A new two-stage task is proposed, in which given an image, an object of interest is first referred by an referring expression, and if such an object is found by the system, then a question about the object is posed for the system to answer. This task is inspired from a real human-computer interaction scenario and can help blind people to understand their surroundings through the two interactions of this task. For the proposed task, we use a large text-image pre-training model, VL-BERT, to fine-tune two downstream tasks, denotational expression understanding and visual question answering benchmark datasets respectively, and then combines real-time predictions for both tasks, and makes a two-stage interactive heuristic system for task understanding and result visualization. All codes and documents of this paper are open source. In summary, the research results and contributions of this paper are as follows:

(1) A more detailed and easy-to-apply task for real human-computer interaction scenarios is proposed: referring expression comprehension based visual question answering.

(2) For this task we propose to use the large text-image pre-training open-source model VL-BERT to fine-tune it on referring expression comprehension and visual question answering benchmark datasets respectively, and then combine the predictions of the two tasks to accomplish the task.

(3) An interactive and heuristic front-end and back-end separated web system for task understanding and result visualization was developed using Vue, ElementPlus and Django for the proposed task and the model used.

**Keywords: Visual Question Answering, Referring Expression Comprehension, Human-Computer Interaction, VLBERT, Vision and Language**

**目录**

[1.引言 6](#_Toc103529885)

[1.1研究背景 6](#_Toc103529886)

[1.2研究内容 6](#_Toc103529887)

[1.3研究意义 7](#_Toc103529888)

[2.相关工作 7](#_Toc103529889)

[2.1指称表达式理解任务研究现状 7](#_Toc103529890)

[2.2视觉问题回答任务研究现状 9](#_Toc103529891)

[2.3其他视觉语言任务研究现状 10](#_Toc103529892)

[2.4图像文本大型预训练及微调 10](#_Toc103529893)

[3.系统设计 11](#_Toc103529894)

[3.1任务设计 11](#_Toc103529895)

[3.2算法设计 12](#_Toc103529896)

[3.2.1预训练模型微调 12](#_Toc103529897)

[3.2.2双阶段任务结合 12](#_Toc103529898)

[3.3界面设计 13](#_Toc103529899)

[3.3.1前端设计 13](#_Toc103529900)

[3.3.2后端设计 14](#_Toc103529901)

[3.4 开发与部署环境 15](#_Toc103529902)

[3.4.1开发环境 15](#_Toc103529903)

[3.4.2部署环境 16](#_Toc103529904)

[4.总结与展望 16](#_Toc103529905)

[4.1总结 16](#_Toc103529906)

[4.2展望 16](#_Toc103529907)

[参考文献 17](#_Toc103529908)

[致谢 19](#_Toc103529909)

# 1.引言

## 1.1研究背景

自从2012年卷积神经网络[1]出现后，深度学习得到了飞速发展，卷积神经网络、循环神经网络两大主流网络分别统治了图像和文本领域相关任务，如卷积神经网络广泛应用于图像分类、目标检测、语义分割和实例分割等任务，循环神经网络广泛应用于机器翻译、命名实体识别、情感分析、机器问答等任务，且深度学习的出现使得人工智能各个应用相比传统方法精度都得到了质的提升。但人工智能的需求并不仅限于单模态的文本或图像领域，如何解决多模态任务如图像和文本相结合的任务是实现真正意义上人机交互的不可或缺的基石。

近五年来视觉语言人工智能研究领域逐步发展，小到如视觉问题回答[2]、指称表达式理解[3]、图像描述等，大到视觉对话[4]、视觉语言导航[5]等，其中视觉问题回答是指给定一个图片对给定问题进行回答，指称表达式理解是指给定一个图片找到给定表达式指代的具体物体，模型使用从最简单的CNN+LSTM相结合到逐步嵌入多种注意力机制、模块化机制和图机制再到近期主流的大型文本图像预训练模型，这些任务的出现和成熟一定程度上促进了人机交互领域的发展，但是多数任务存在形式简单、与实际应用有差距等问题，在现有模型不断提升精度的情况下，需要有更加复杂以及贴合实际人机交互场景的任务提出，从而更准确地评估当前多个模型的精度以及落地难度。

## 1.2研究内容

本文就是基于主流视觉语言任务中的视觉问题回答这一任务，针对其存在的形式简单单一、与实际人机交互场景差距较大的问题进行解决，提出一个双阶段的交互任务。给定一张图片，首先给定一个指定图片中某物体的表达式，由机器代理判断是否存在这样的物体，若存在这样的物体那再次给定一个与该物体相关的问题，并由机器代理回答，这实际上是将原来的指称表达式理解和视觉问题回答任务相结合的一个任务，通过指称表达式理解来达到更细致更实际地问题回答。将视觉问题回答常用标准数据集VQA2.0[6]的任务与本论文提出任务相比，前者问题类别主要分为计数类问题、判断图片场景的描述正确与否问题、颜色问题以及一部分常识推理类问题，后者则更注重于实际人机交互场景的问题，更能将问题聚焦于一个场景下多个物体中的一个具体物体。两者都着重于同时理解图像和文本，但后者相比前者更贴合实际交互场景，能够很好地在多个类似物体中聚焦到感兴趣的物体或区域，再通过一次问答更细致地了解该物体的情况。前者包含面广但真正意义上贴合实际人机交互场景的比例不大，后者旨在解决能快速应用到实际的小部分问题而并未包含其他很多类问题。除了提出该项任务之外，本文还同时提出使用现阶段在指称表达式理解和视觉问题回答性能均表现极好的大型图像文本预训练模型，在这两个下游任务上进行微调后并在预测阶段将预测结果相结合来完成所提出任务。最后为了任务理解和结果可视化，本文同时提出开发一个前后端相分离的Web系统，使用Vue3框架的Composition API和ElementPlus组件框架进行前端开发，使用Vuex做组件间的全局数据交互，使用Vue Router做组件的路径映射，使用TypeScript代替JavaScript，使用Axios向后端发出异步请求，上述提到前端技术均是为了使得项目在开发过程中代码复用性强、可扩展性强以及可阅读性强。后端整体采用Django开发，具体使用Django-Rest-Framework做接口开发，使用CORS-Headers做跨域访问从而实现Vue前端与Django后端的交互。

## 1.3研究意义

本文所研究工作的意义在于在现有视觉语言任务上提出更加贴合实际人机交互场景的任务，使得该任务上的研究成果能够快速在实际人机交互场景中应用。人工智能现阶段的前沿研究成果均存在落地难、与实际应用场景相差较大的问题，因此使得工业界人工智能的应用相比科研界人工智能的研究成果总要落后一阶段。本文在视觉问题回答这一任务上，使用指称表达式理解作为辅助任务来聚焦一个具体物体，从而实现在这个具体物体上的问题回答。该项任务可在后续应用于机器代理辅助盲人了解周围环境以及视觉语言导航等人机交互场景。

# 2.相关工作

## 2.1指称表达式理解任务研究现状

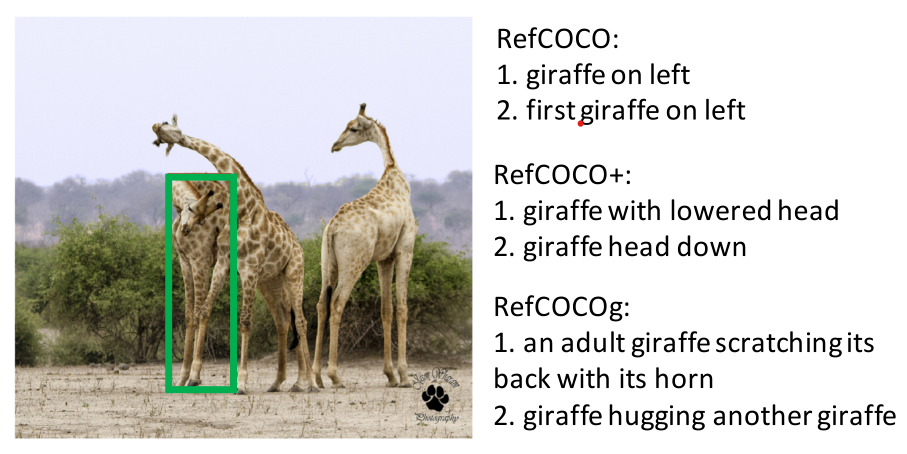
指称表达式是指代某个场景下某一个具体物体的自然语言语句，表达式往往根据场景下某个物体的外观、属性、相对其他物体的位置和关系信息来确切指向一个物体，指称表达式理解是给定一副图片标出指称表达式所指代的具体物体边界框位置，是一个将计算机视觉和自然语言处理相结合的任务，需要同时理解文本和图像来进行预测，具体样例可参阅图2.1。不同于目标检测提前定义好所有目标的类别，指称表达式理解所要指代的物体类别在预测时才能根据表达式确定，因此指称表达式理解比传统计算机视觉问题更加困难。指称表达式理解这一任务从提出以来逐步发展出很多方法与模型[7]，主要分为使用CNN+LSTM将图片和文本特征嵌入到一个特征空间的传统方法[3], [8], [9]、在传统方法上添加模块化机制[10], [11]、图机制[12]、注意力机制[10], [11], [13]的改进方法、使用额外解析器的模型[14]和大型图像文本预训练模型[15], [16]等，具体各方法在该任务常用标准数据集RefCOCO+[8]测试集上表现如表2.1所示，其中表现最好的模型ViLBERT[15]在该任务上的精度已经足够应用于实际当中。

图2.1 指称表达式理解任务举例

表2.1 使用检测候选物体时上述模型在RefCOCO+的IOU比较

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | 图像特征提取模型 | 分类 | testA | testB |
| MMI[3] | vgg16 | CNN/LSTM | 54.03 | 42.81 |
| Visdif[8] | vgg16 | CNN/LSTM | 50.10 | 37.48 |
| Spe+Lis+RI[9] | frcnn-resnet101 | CNN/LSTM | 58.68 | 48.23 |
| CMN[10] | vgg16 | 模块化和注意力机制改进 | 54.32 | 47.76 |
| MattNet[11] | mrcnn-resnet101 | 模块化和注意力机制改进 | 71.62 | 56.02 |
| LGRANs[12] | frcnn-vgg16 | 图机制改进 | 64.0 | 53.4 |
| CM-Att-Erase[13] | frcnn-resnet101 | 注意力机制改进 | 73.65 | 58.03 |
| NMTree[14] | frcnn-vgg16 | 额外解析器改进 | 72.02 | 57.52 |
| ViLBERT[15] | frcnn-resnet101 | 图像文本大型预训练 | 78.52 | 62.61 |
| VLBERT[16] | frcnn-resnet101 | 图像文本大型预训练 | 77.72 | 60.99 |

## 2.2视觉问题回答任务研究现状

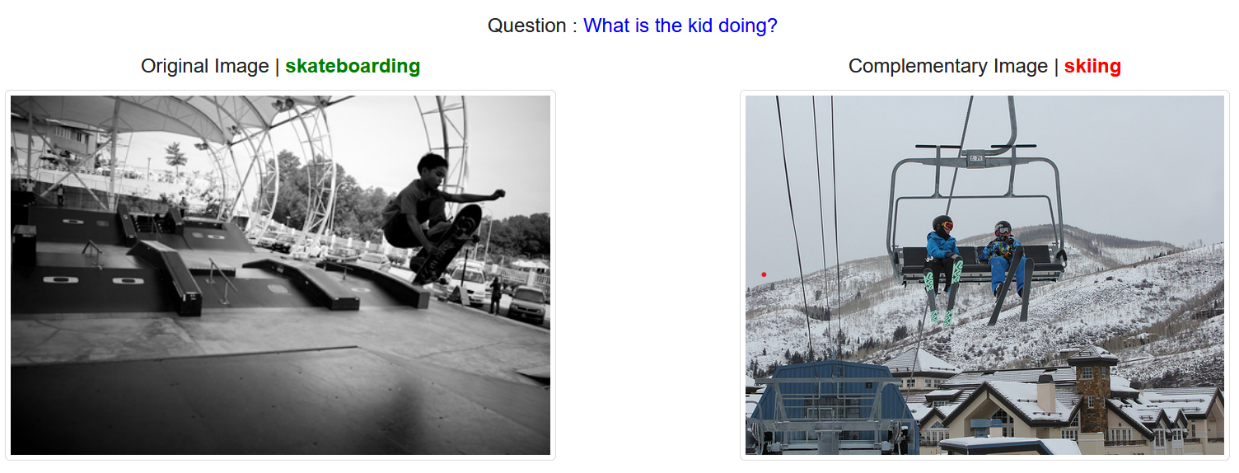
视觉问题回答是计算机视觉和自然语言处理交叉领域的一个任务，是给定一副图片和一个与该图相关的问题，对该问题进行回答的任务，它需要结合图片中的视觉元素和一些常识甚至是知识库来正确回答问题。问题类型主要分为正确与否的二分类问题、多选一选择答案的多分类问题、完形填空式问答以及开放性的生成问题，相比前三者是分类问题即答案集合已经预先定义好，后者是开放性的问题，需要使用符合句法结构的自然语言短语来回答，标准数据集上常将该类问题分为物体类别判定、场景某物体数量计数、物体颜色判定以及物体位置判定等，具体样例可参阅图2.2。总体来看，现阶段视觉问题回答任务中问题语义信息量较少形式较为单一，若应用于实际的话会存在应用场景较少的问题。因此在传统视觉问题回答之上提出了很多更复杂更有应用场景的视觉问答任务如视频问题回答、交互场景下的视觉问答[17]、基于事实[18]和知识库[19]的视觉问答、视觉对话[4]等等。视觉问题回答这一任务自2015年提出以来也发展出了各类模型用于解决不同形式的上述提到的视觉问题回答任务[20]，主要分为使用CNN+LSTM将图片和文本特征嵌入到一个特征空间的传统方法、在传统方法上基于注意力机制、可分解式模型、额外知识库的改进方法以及大型文本图像预训练模型等，具体各方法在该任务常用标准数据集VQA2.0[6]测试集上表现如表2.2所示，虽然在该数据集上表现精度较高，但由于现有数据集均存在数据偏差、语言先验以及问题形式不够丰富、语义结构不够丰富等问题，仍不足够应用到实际人机交互场景当中。

图2.2 视觉问题回答任务举例（VQA2.0）

表2.2 各视觉问题回答模型在VQA2.0上的精度比较

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 分类 | test-dev | test-std |
| BUTD[26] | 注意力机制改进 | 65.32 | 65.67 |
| ViLBERT[16] | 大型预训练 | 70.55 | 70.92 |
| VisualBERT[27] | 大型预训练 | 70.80 | 71.00 |
| LXMERT[28] | 大型预训练 | 72.42 | 72.54 |
| VL-BERT[15] | 大型预训练 | 71.19 | 72.22 |

## 2.3其他视觉语言任务研究现状

随着计算机视觉和自然语言处理在近年来得到了很大发展，以及多模态数据在自然界普遍存在，因此视觉语言相结合多模态方向的人工智能研究在近年来发展迅猛且仍在持续增长，除了本文提到的视觉语言问答和指称表达理解之外，还有其他很多正在持续研究的视觉语言任务，但是视觉语言各项任务都会面临图像和文本两个模态间存在语义鸿沟的挑战[20]：图像是高维的且比纯文本含有更多噪声；图像缺乏语言所含有的句法结构和语法规则且没有自然语言中可直接做句法解析和正则表达式匹配的工具包；图像捕捉真实世界的丰富性而自然语言已经达到高水平的抽象。其他相关视觉语言任务诸如视觉常识推理[21]、图像描述、视觉语言导航[5]等任务，其中视觉常识推理是指给定一个有关某个图像的问题，机器必须正确回答且提供一个合理的理由证明自己的答案；图像描述是指给定一个图片给出有关该图片内容的自然语言描述；视觉语言导航是指给定一条导航指令让机器代理在3D室内全景场景下进行导航。这些任务在提出以来都获得了很大的提升，但是与视觉问题回答一样，视觉语言任务的关键是要解决图片和文本的特征融合，而图片和文本之间的语义鸿沟一定程度上限制了各类视觉语言模型的性能，因此这些任务往往存在形式上的限制如视觉常识推理局限为四选一的选择问题，这些局限表明了当前大多数视觉语言模型还无法像人类一样完成高难度的视觉语言理解，因此视觉语言还处在一个需要持续得到发展的阶段。

## 2.4图像文本大型预训练及微调

自Transformer[22]提出完全使用自注意力机制代替RNN来完成自然语言处理任务以来，使用Transformer做文本大型预训练便大放光彩[23]，在通用型数据集上做大型自监督预训练使得模型学得某个领域的泛化特征，并在若干个该领域下游任务标准数据集上做微调便可在各个下游具体任务上达到极佳的性能。这种使用Transformer做大型预训练很快也被应用到图像上[24]，并代替传统CNN架构来完成泛化特征提取并迅速应用到各下游任务上。随着计算机视觉和自然语言处理领域Transformer的大放异彩，这种大型预训练的方法很快被应用到视觉语言领域[15], [16]，通过在大量图像文本对上做大型预训练，然后在若干个下游任务上如视觉问题回答、指称表达理解等标准数据集上做微调即可达到很好的效果，本文使用VLBERT[16]作为模型即是利用了大型预训练的优点，使用一个模型微调又同时在两个任务上都达到很高的性能，从而更快速的完成所提出的任务。

# 3.系统设计

## 3.1任务设计

因本文上述所提到的当前视觉问题回答存在问题形式局限、问题语义结构不够丰富、问题不够精确等问题，现有模型较难快速应用到有实际需求的人机交互场景中，故本文提出一个新型双阶段的交互式问答任务：基于指称表达式所指代物体的视觉问题回答，即先通过指称表达式理解这一辅助任务聚焦在感兴趣的物体上，再通过具体的视觉问答进一步了解该物体的情况。具体是先通过明显的外观和属性描述、相对其他明显物体的位置和关系描述来明确指代某个物体，后通过询问该物体详细外观信息来了解该问题更多信息。

该任务启发于实际人机交互场景，想象一个盲人在一个室内或室外场景下，他需要机器代理先使用自然语言描述一下所看到的场景，这其中会包含一些物体的外观属性和一些物体间的相对位置和关系信息，因此这里选择使用明显的外观属性和相对位置和关系信息来指代，因为现有图像描述模型结果往往还不足够精确到很详细的属性，然后盲人为了了解周围情况需要对各个场景物体做更详细的了解，此时便需要先指代感兴趣的物体再详细地问该物体的外观信息。又比如在智能家居应用中，服务机器人要辅助行动不便老人看一些场景的实时情况，则老人可根据自己对场景的熟悉指代机器到某一场景下查看并询问该物体的详细情况。

该任务相比于当前视觉问题回答常用基准数据集[6], [25]的任务，之所以更贴合实际应用场景，是因为它能在一个场景多个不同类物体或同类不同外观物体之间进行区分后指代感兴趣的物体进行视觉问答，它能够以不含歧义的方式精确指代感兴趣的物体，而不是仅仅宽泛的问整个场景的一些正确与否问题来了解整个场景，启发于人类生活中的场景问答习惯，采用多次做一个双阶段先指代某物体再具体询问的任务来使其更贴合实际人机交互场景。

## 3.2算法设计

3.2.1预训练模型微调

为解决上述所提出任务，需要同时可以完成指称表达式理解和视觉问题回答的模型用来做实时预测，在前期调研视觉问题回答[20]、指称表达理解综述[7]后，选择使用现阶段表现性能极佳且能同时应用到这两个下游任务的大型文本图像预训练模型VL-BERT模型[16]在这两个下游任务基准数据集上进行微调。

考虑到VLBERT原模型是使用16G V100进行分布式多卡预训练和微调的，且安培架构的GPU（30系列、A40、A4000等）需使用CUDA11.x及以上版本的CUDA，而VLBERT需要使用CUDA9.0做一些模块编译，因此最后选择租用AutoDL平台的32G GPU一卡24小时，分别在指称表达理解常用基准数据集RefCOCO+[8]和视觉问题回答常用基准数据集VQA2.0[6]上做轮次为5的训练，微调时的训练参数参照VLBERT，在32G GPU上，可将batch size做适应性的修改，分别将两个下游任务batch size扩大两倍，视觉问题回答微调阶段batch size使用128，采用Adam优化器，学习率为，，，权重衰减为，学习率在最开始2000步内采用学习率线性衰减进行学习率预热。指称表达理解微调阶段batch size使用8，累计梯度步数为2，即总batch size为16，学习率为，，，权重衰减为，学习率在最开始梯度下降500步内采用学习率线性衰减进行学习率预热。具体微调结果与原论文对比请参照表3.1。

表3.1 VLBERT在指称表达理解和视觉问答上微调结果与[16]精度对比

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 任务 | batch size | 微调时间 | 优化器 | 学习率 |  |  | VLBERT精度 | 本文微调精度 |
| VQA | 128 | 5h | Adam |  | 0.9 | 0.999 | 71.16 | 70.93 |
| REC | 16 | 20h | Adam |  | 0.9 | 0.999 | 77.72/60.99 | 76.44/60.63 |

3.2.2双阶段任务结合

完成上述提出任务需要将两个下游任务微调后的模型在预测阶段结合起来，首先通过指称表达理解指代出感兴趣物体的边界框，然后将边界框部分的图像特征提取出来，和二阶段提出的问题一起输入到微调后的视觉问题回答模型，并输出该问题的答案。具体需要找到VLBERT开源代码中的REC和VQA数据集定义部分，首先需要将Dataset类内部的batch size参数均改为1，因为在实际使用时每次只需要有一个样本输入到模型中，其次添加一个脚本用于指称表达理解得出检测框之后提出检测框的特征并和问题一同输入到视觉问题回答模型中，后续需要将该部分代码添加到后端请求用于实时返回模型预测结果。

## 3.3界面设计

3.3.1前端设计

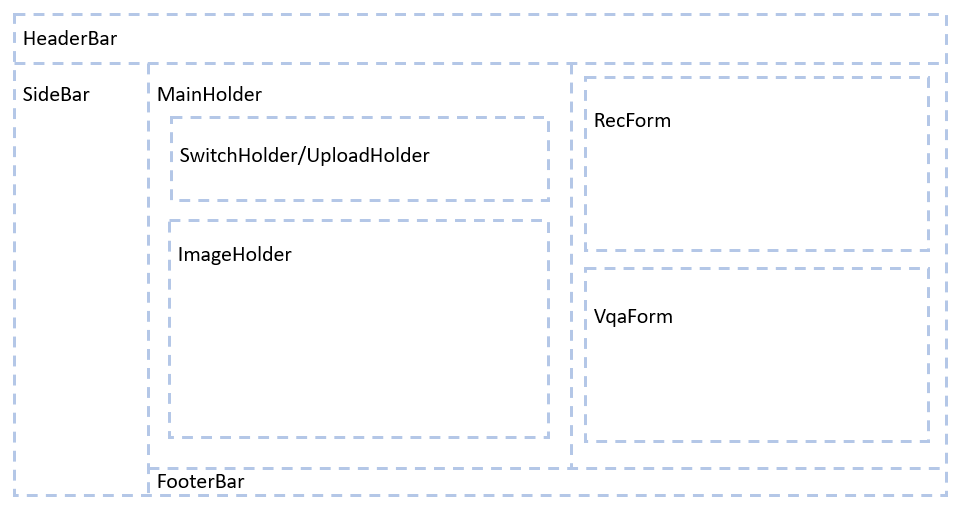
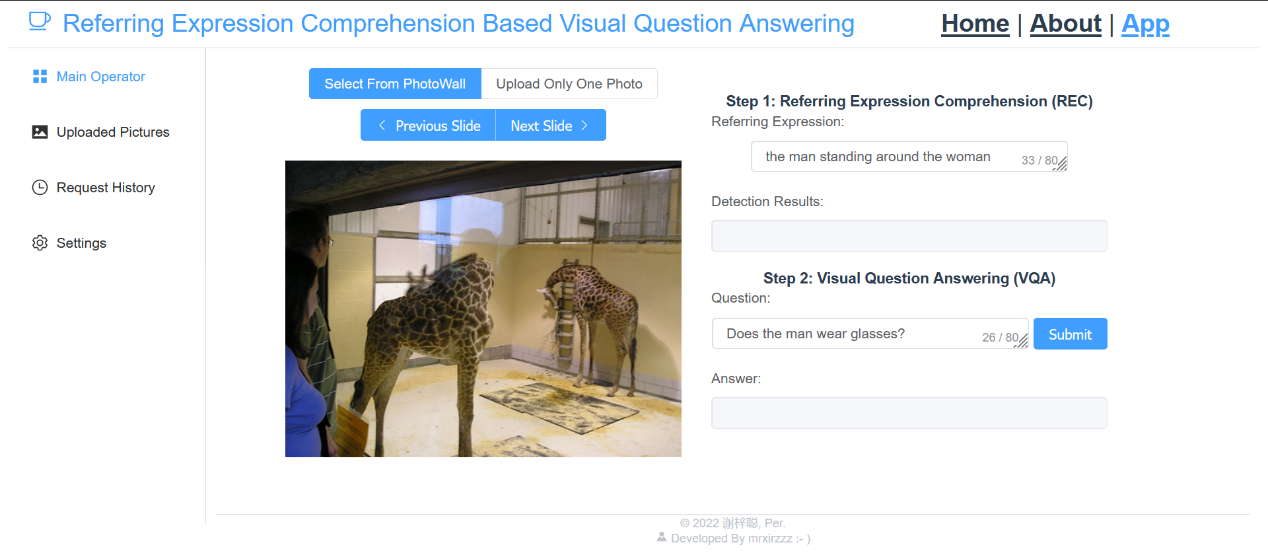
前端设计使用Vue3开发，基于Vue3开发了两类组件，分别是内容填充组件和布局组件，其中内容填充组件包含图像填充组件、上传文件组件、切换图片组件、指称表达理解任务表单提交组件、视觉问题回答任务表单提交组件、照片墙组件、提交历史查询组件，布局组件包含顶部标题栏、左侧标签页栏、中部内容栏和底部版权栏，具体各个布局组件如何布局和内容组件如何填充可见图3.3所示，整体界面如图3.4所示。使用Vue所提供的组件思想去开发整个网页的模块化布局和内容组件有利于提升代码可复用性、可阅读性和可扩展性。使用Vuex去做各个组件间的数据共享读写库，使用Vue Router去做某个路径下哪些内容组件填充在中部布局组件中，具体映射情况可参见表3.2所示。

图3.3 组件间布局和填充关系

表3.2 Vue Router中组件和路径之间的映射关系

|  |  |
| --- | --- |
| 包含组件 | URL |
| ImageHolder、SwitchHolder/UploadHolder、RecForm、VqaForm | /app/Main/ |
| RequestHistory | /app/Request/ |
| RequestHolder | /app/Request/{:request\_id}/ |
| PhotoWall | /app/PhotoWall/ |
| SettingPanel | /app/Setting/ |

图3.4 整体操作界面

Vue3整体采用TypeScript代替JavaScript进行开发，使用SFC（Single File Component）和Composition API进行开发，使用Axios代替Ajax向后端发出Promise形式的Restful请求，一定程度上提高了代码简洁性和可阅读性。前端组件库方面使用基于Vue3开发的Element Plus组件库，一定程度上美化了界面设计，且由于Element Plus也是采用模块化的设计思想设计的各个常用业务组件，因此同时提高了代码可复用性和可扩展性。总体来看，前端使用Vue3结合Element Plus的方式去进行开发，并使用很多开发模式、开发框架和开发包去提高前端代码的可复用性、可扩展性、可阅读性、鲁棒性和简洁性。

3.3.2后端设计

由于上述模型部署时需要使用Python环境，因此考虑使用Python方面Web开发比较成熟的框架Django来做接口开发，同时结合使用Django-Rest-Framework去做Django代码的简化并可快速开发出对各个数据库表的增删改查的可浏览式接口，同时提高代码的简洁性、可复用性和可扩展性。

数据库方面由于设计数据库实体表较少，因此直接使用Python内置的SQLite数据去做数据库引擎，具体实体表包含用于图像上传的IMG表、用于提交指称表达理解任务请求的REC表、用于进一步提交视觉问题回答任务请求的VQA表，其中VQA和REC间存在一对多的联系，即每个VQA实体需要依托于一个REC实体存在，一个REC实体可以有多个对应的VQA实体。数据库关系模式设计如下表3.3-3.5所示。

后端接口开发主要基于Django-Rest-Framework提供的三层架构来写，即分models、serializers、viewsets三层来写，models.py包含数据库实体表的各字段定义，serializers.py包含数据库增删改查的序列化操作，viewsets.py包含Serializer序列化后的数据以及需要渲染到的HTML模板，这里主要需要对Serializer得到的序列化数据做一些过滤处理等操作而不需要模板渲染。最后在urls中去做viewsets和访问url之间的映射即可。

表3.3 IMG表物理模式设计

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 注释 |
| id | integer | 主键 |
| img | varchar(100) | 非空 |

表3.4 REC表物理模式设计

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 注释 |
| id | integer | 主键 |
| referring\_expression | varchar(255) | 非空 |
| result | varchar(255) | 允许空 |
| result\_image | varchar(255) | 允许空 |
| img\_id | integer | 外键至IMG |

表3.5 VQA表物理模式设计

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 注释 |
| id | integer | 主键 |
| question | varchar(255) | 非空 |
| answer | varchar(255) | 允许空 |
| rec\_id | integer | 外键至REC |

同时为了使Vue开发的8080前端应用和Django开发的8000后端应用能够进行数据交互即前端向后端发送请求后端向前端相应的过程，还需要使用CORS-Headers实现跨域访问，从而完成整个系统的实现。

## 3.4 开发与部署环境

3.4.1开发环境

开发环境前端和后端均是基于VSCode进行开发，前端基于Node.js构建JavaScript环境、基于npm淘宝源做包管理工具，需安装vue3、elementplus、vuex、vuerouter、axios和typescript；后端基于Conda建立虚拟Python3.8环境，使用pip清华源安装django、django-rest-framework、cors-headers以及VLBERT模型依赖包。

3.4.2部署环境

部署后端由于需要租用AutoDL平台GPU服务器，但租用平台的服务器实例只允许开发固定一个端口6006，因此不能在该GPU服务器上同时占用前端Vue应用的8080和后端Django应用的8000，故考虑使用AutoDL平台GPU服务器部署后端Django应用，使用阿里云免费服务器部署前端Vue应用。后端服务器系统为Ubuntu20.04，其他环境为Cuda9.0，cuDNN7.6.5，gcc4.9.1。前端服务器环境为Ubuntu20.04，其他环境为Node.js16.15.0，npm8.5.5。

# 4.总结与展望

## 4.1总结

第一，本文相比传统视觉问题回答任务提出了一个新的能更快速应用于实际应用场景的双阶段视觉问题回答任务，首先通过指称表达式指代图片中的一个具体物体，若存在这样的物体则再提出与该物体相关的问题让系统回答。该任务能在一个场景下众多物体中首先允许系统聚焦在感兴趣的物体上，然后再通过问答了解这个物体更详细的信息，符合实际人机交互的场景，能更快速用于实际中。

第二，本文通过VLBERT大型图像文本预训练模型分别在指称表达理解和视觉问题回答这两个下游任务上进行微调，并在实时预测时先通过指称表达理解找到物体，得到边界框后将该物体视觉特征和问题输入到视觉问题回答模型用于最后输出，使用相对其他物体位置关系以及联系或简单的外观属性来定位目标，再通过问答详细了解具体的其他外观属性。

第三，通过Vue和Django开发前后端分离的Web系统用于任务理解和结果可视化，在前后端均使用相关技术来提高代码可复用性、可扩展性、鲁棒性和可阅读性，并通过前后端分离方式来部署系统。

## 4.2展望

提出上述任务后，使用上述方法的准确率较低，一部分误差来源于一阶段指称表达理解边界框的偏差，另一部分来源于二阶段视觉问答的答案误差，同时由于使用大型预训练模型，会导致预测耗时长要求GPU性能强。对于问题一，考虑使用更强的视觉特征提取模型来提取图像和候选框的特征。对于问题二，考虑针对该任务在已有数据集标注上做一些适应性修改形成新提出任务上的数据集并在该数据集上微调而不是在VQA2.0上微调，并提出对应的评价指标用于更好的模型收敛。对于问题三，一方面考虑在中低性能GPU设备上使用模型压缩和模型蒸馏来降低模型所耗显存和提高模型推理速度，另一方面考虑降低图片分辨率和文本嵌入向量的长度从而降低预测阶段并行进入模型的数据显存。

除此之外，该任务并未包含所有类型的视觉问答，只是提出一类可以快速用于实际人工智能场景的任务，且该任务只是作为整个场景交互的一部分，若要真正完成人机交互，还需要和其他很多视觉语言任务相结合，如在辅助盲人视觉语言导航过程中，可能需要先做场景描述，描述后盲人通过指称表达指代某具体场景物体，便可后续详细了解该物体，导航至该物体或对该物体执行某个操作。总体而言，该任务在一定程度上对逐步实现真正意义上人机交互具有启发意义，通过后续在该任务上做改进工作，对实现自主型人工智能机器代理具有现实意义。

# 参考文献

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘ImageNet Classiﬁcation with Deep Convolutional Neural Networks’, p. 9, 2012.

[2] A. Agrawal *et al.*, ‘VQA: Visual Question Answering’, arXiv, arXiv:1505.00468, Oct. 2016. doi: 10.48550/arXiv.1505.00468.

[3] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, ‘Generation and comprehension of unambiguous object descriptions’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 11–20.

[4] A. Das *et al.*, ‘Visual dialog’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 326–335.

[5] P. Anderson *et al.*, ‘Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments’, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018, pp. 3674–3683. doi: 10.1109/CVPR.2018.00387.

[6] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, ‘Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering’, arXiv, arXiv:1612.00837, May 2017. doi: 10.48550/arXiv.1612.00837.

[7] Y. Qiao, C. Deng, and Q. Wu, ‘Referring Expression Comprehension: A Survey of Methods and Datasets’, *IEEE Trans. Multimed.*, vol. 23, pp. 4426–4440, 2021, doi: 10.1109/TMM.2020.3042066.

[8] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, ‘Modeling Context in Referring Expressions’, *ArXiv160800272 Cs*, Aug. 2016, Accessed: May 02, 2022. [Online]. Available: http://arxiv.org/abs/1608.00272

[9] L. Yu, H. Tan, M. Bansal, and T. L. Berg, ‘A joint speaker-listener-reinforcer model for referring expressions’, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7282–7290.

[10] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, ‘Modeling relationships in referential expressions with compositional modular networks’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1115–1124.

[11] L. Yu *et al.*, ‘MAttNet: Modular Attention Network for Referring Expression Comprehension’, *ArXiv180108186 Cs*, Mar. 2018, Accessed: Oct. 12, 2021. [Online]. Available: http://arxiv.org/abs/1801.08186

[12] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. van den Hengel, ‘Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks’, 2019, pp. 1960–1968. Accessed: Mar. 29, 2022. [Online]. Available: https://openaccess.thecvf.com/content\_CVPR\_2019/html/Wang\_Neighbourhood\_Watch\_Referring\_Expression\_Comprehension\_via\_Language-Guided\_Graph\_Attention\_Networks\_CVPR\_2019\_paper.html

[13] X. Liu, Z. Wang, J. Shao, X. Wang, and H. Li, ‘Improving referring expression grounding with cross-modal attention-guided erasing’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1950–1959.

[14] D. Liu, H. Zhang, F. Wu, and Z.-J. Zha, ‘Learning to assemble neural module tree networks for visual grounding’, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4673–4682.

[15] J. Lu, D. Batra, D. Parikh, and S. Lee, ‘Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks’, *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[16] W. Su *et al.*, ‘Vl-bert: Pre-training of generic visual-linguistic representations’, *ArXiv Prepr. ArXiv190808530*, 2019.

[17] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, ‘IQA: Visual Question Answering in Interactive Environments’, 2018, pp. 4089–4098. Accessed: Mar. 23, 2022. [Online]. Available: https://openaccess.thecvf.com/content\_cvpr\_2018/html/Gordon\_IQA\_Visual\_Question\_CVPR\_2018\_paper.html

[18] P. Wang, Q. Wu, C. Shen, A. Dick, and A. Van Den Hengel, ‘Fvqa: Fact-based visual question answering’, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2413–2427, 2017.

[19] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. Dick, ‘Explicit knowledge-based reasoning for visual question answering’, *ArXiv Prepr. ArXiv151102570*, 2015.

[20] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, ‘Visual question answering: A survey of methods and datasets’, *Comput. Vis. Image Underst.*, vol. 163, pp. 21–40, Oct. 2017, doi: 10.1016/j.cviu.2017.05.001.

[21] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, ‘From recognition to cognition: Visual commonsense reasoning’, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6720–6731.

[22] A. Vaswani *et al.*, ‘Attention Is All You Need’, *ArXiv170603762 Cs*, Dec. 2017, Accessed: Aug. 25, 2021. [Online]. Available: http://arxiv.org/abs/1706.03762

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, *ArXiv181004805 Cs*, May 2019, Accessed: Nov. 24, 2021. [Online]. Available: http://arxiv.org/abs/1810.04805

[24] A. Dosovitskiy *et al.*, ‘An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale’, *ArXiv201011929 Cs*, Jun. 2021, Accessed: Aug. 25, 2021. [Online]. Available: http://arxiv.org/abs/2010.11929

[25] R. Krishna *et al.*, ‘Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations’, *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017, doi: 10.1007/s11263-016-0981-7.

[26] P. Anderson *et al.*, ‘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*, 2018, pp. 6077–6086.

[27] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, ‘VisualBERT: A Simple and Performant Baseline for Vision and Language’, *ArXiv190803557 Cs*, Aug. 2019, Accessed: Apr. 29, 2022. [Online]. Available: http://arxiv.org/abs/1908.03557

[28] H. Tan and M. Bansal, ‘Lxmert: Learning cross-modality encoder representations from transformers’, *ArXiv Prepr. ArXiv190807490*, 2019.

# 致谢

大学四年以来，感谢各位谆谆教导的老师，使我从中学得各项专业技能，并习得一定程度的专业学习能力，从而对毕业设计各项技术无论熟悉与否都能在较短时间内逐步掌握并把工作做好，其中非常感激田侦老师在我毕业设计阶段悉心教导，对每个阶段的工作存在的问题耐心指导并提出合理建议，对最终论文定稿做了很大帮助，也非常感谢郑州大学以及计算机与人工智能学院平台给我们带来的无限机会，使得我在各项竞赛和创新项目中逐步培养了一些创新和科研意识，提高了遇到问题解决难题的动手能力。最后，再次由衷感谢在我大学四年来对我学业有所帮助的老师和同学们。