****

**专业型硕士学位论文开题报告书**

|  |  |  |
| --- | --- | --- |
| **姓名** | **：** | **兰又铭** |
| **专业** | **：** | **电子信息** |
| **研究方向** | **：** | **自然语言处理** |
| **录取类别** | **：** | **非定向全日制** |
| **入学时间** | **：** | **2020.09** |
| **校内指导教师** | **：** | **马霄** |
| **校外指导教师** | **：** | **曾江峰** |
| **学院（中心）**  **名称** | **：** | **信息与安全工程学院** |

**中南财经政法大学研究生院**

一、基本情况

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 | 兰又铭 | 所在学院（中心） | 信息与安全工程学院 |
| 专 业 | 计算机技术专硕 | 研究方向 | 自然语言处理 |
| 年 级 | 2020级 | 指导教师 | 马霄 |
| 论文题目 | 基于图文多模态可解释性的视觉问答方法 | | |

二、选题背景、研究意义、文献综述与实践现状

|  |
| --- |
| **（一）选题背景及研究意义**  **1）选题背景：**  随着深度学习技术和计算资源的发展，计算机视觉和自然语言处理领域中的众多基础任务得到了广泛关注，并取得了一定的研究进展，如图像分类、目标检测、问答系统、对话系 统等。随着基础任务的突破以及计算机感知能力的提高，研究者们开始聚焦于计算机视觉和自然语言处理的交叉领域。从研究的角度出发，研究者们希望计算机不仅具有基础的感知与认知能力，而且进一步具备多模态推理能力。从应用的角度考虑，用户希望能够以更加友好的交互方式使用计算机视觉系统。因此，视觉语言领域(Vision and Language)的相关任务在近年来得到了广泛关注。  视觉问答 (Visual Question Answering，VQA)：视觉问答致力于回答一个给定图像的自然语言问题。VQA已经取得了巨大的进步，但是仍然存在许多尚未解决的问题。大多数现存的VQA模型，忽略了一种类型问题：这些问题与图像中的文字相关。为了正确的回答这类问题，需要理解图像中的文字并进行推理。  **2）研究意义：**  视觉问答任务被提出的本意是帮助那些在视觉上有障碍的人士能更好的理解周围的事物和感知他们所处的环境，并且可以帮助他们进行恢复训练，值的肯定的是视觉问答的研究意义与应用领域已经远远超出了人们本来的期许。视觉问答系统不仅可以帮助视觉受损的人群，更可以应用于医学领域、教育领域和娱乐领域。视觉问答系统正在以其特有的研究魅力吸引着无数科研工作者和人工智能商业应用领域的广泛关注，其“看图说话”的独特功能模式或将应用于图像检索、儿童早教、盲人导航、辅助驾驶等工作领域，与其他技术领域相结合从而引领技术革新也是未来发展的趋势之一，视觉问答系统的研究及应用无疑是一个值得探索的领域。  **（二）国内外文献综述**  关于可解释性视觉问答的研究，虽然视觉问答系统才兴起不久，但是对其的研究方法却各有不同。近年来已发表的综述不胜枚举，借鉴文献的归纳方式本文将四个方面介绍视觉问答主要的研究现状，分别为：基于基于预训练的方法、基于图像中文本信息的视觉问答、基于注意力机制的方法、基于注意力机制的方法、视觉与文本任务中多模态交互方法。  **1）视觉问答**  视觉问答己经成为了人工智能领域最活跃的领域之一。VQA是一个复杂的任务，其目的是回答一个与给定改的图像内容相关的自然语言问题。它要求同时对图像和问题内容进行理解，并且还需要找到图像内容与问题语义的关系。 自从提出了一个大规模的VQA数据集（VQALO）, VQA的性能己经取得了很大的进步，并且已经有许多相关的数据集被提出来。VQA1.0的规模是更加庞大，并且 其包含的样本比DAQUAR和COCO-QA数据集中的样本更加接近自然情况下 的问答场景。但是VQA1.0数据集存在很大的偏置问题，一个模型在不看图像信息的情况下能够利用这种偏置作为模型的先验信息来回答问题。收集并提出了 VQA2.0 数据集来弥补VQA1.0数据集中的偏置问题。为了学习独立于 语言的视觉推理能力，一些工作提出了包含抽象的生成图像的VQA数据集（例如, CLEVR, NLVR和FigureQ）。提出了基于外部事实知识的VQA数据集, 探索需要外部知识来辅助视觉问答任务。  **2）基于预训练的方法**  预训练是计算机视觉和自然语言领域的重要技术.随着迁移学习、自监督学习、无监督学习等领域的发展，利用外源数据集进行有效的预训练成为了计算机视觉与自然语言处理领域的重要研究问题与技术手段。在大规模视觉数据集和文本数据集上，对基础模型进行预训练并将其迁移到下游任务上，模型的性能可以得到显著提升。早期，视觉问答模型的视觉分支采用VGG或 ResNet结构，在大规模图像数据集ImageNe上进行预训练.文本分支通常采用 GloVe对词表示进行初始化。视觉分支的预训练经历了从网格特征到区域特征，再回到网格特征的探索过程。Anderson等提出使用目标检测的模型作为视觉分支的基础模型,借助 VisualGenome数据集上的区域标注、类别标注和属性标注等对目标检测模型进行预训练。这一视觉特征在视觉问答、看图说话等一系列视觉语言任务上取得了明显的突破，在后续的研究方法中得到了广泛应用。最近Jiang等发现，通过对模型稍加修改，可以使网格特征得到比区域特征更好的性能，且计算速度得到了大幅提高。Jiang等指出，预训练数据集 VisualGenome中的目标及属性标注和图像的分辨率对预训练具有重要作用。早期的视觉对话方法中,视觉模块与语言模块独立进行预训练，因此在预训练过程中，视觉语言之间的联系没有得到很好的捕捉。近年来，研究人员考虑到基于Transformer结构的预训练模型在自然语言处理领域取得了突破性的性能提升，进而将其应用到视觉语言领域。其中，具有代表性的工作包括 ViLBERT,VLBERT,LXMERT,UNITER, OSCAR,in,VisualBERT等，上述方法大多采用了 BERT的结构。预训练任务包括掩码语言建模 (Masked Language Modeling)和跨模态匹配 (Crossmodality Matching)等。预训练数据集的选择包括 ConceptualCaptions, SBUcaptions,MSCOCO,VisualGenome等。借助在大规模数据集上进行跨模态的自监督预训练任务，上述方法在包括视觉问答在内的多个下游任务中取得了最佳性能。  **3）基于图像中文本信息的视觉问答**  最近，有几个工作己经提出基于图像中文本信息的视觉问答任务（称作TextVQA 任务）。LORKAS是最近提出的解决TextVQA任务的工作。LoRRA使用了 一个外部的OCR系统来识别图像中的文本信息，并通过在OCR tokens上运用注意力方法, 并比较注意力权重与分类器答案列表的答案分数，如果注意力权重更大一些则复制 图像上的文  本信息作为问题的答案。LoRRA也是用了问题指导的图像区域注意力方 法来识别图像中最重要的视觉信息。然而，LoRRA仅仅包含了有限的模态间的交互 数量（问题与图像，问题与OCRtokens）。这些现存的方法的性能严重的被多模态交 互方法限制。这一章提出的MLCI模型就是用来解决这个限制。   1. **基于注意力机制的方法**   注意力机制是视觉问答任务中的主流技术之一。基于视觉问答与对话综述意力机制的方法通过对问题或图像进行注意力加权，增强视觉与语言之间的交互，来对问题和图像的主体信息进行准确捕捉。早期，最具代表性的方法为基于问题的图像注意力，包括 SMem和SAN.其中，SMem 计算每个单词的词向量与图像区域特征的相关性，得到基于问题的图像注意力。SAN使用 CNN 或 LSTM 计算整个问题的特征表示，并将注意力模块进行堆叠以实现多步迭代。上述方法的注意力机制均是单向的，即先计算语言特征，再计算视觉区域关于语言特征的注意力。为了更好地进行视觉语言交互，后续工作提出使用协同注意力的模式。例 如，分层注意力模型 (Hierarchical Question Image Attention,HieCoAtt)并行地计算关于问题的图像注意力特征与关于图像的问题注意力特征，从而得到最终的视觉表示与文本表示。考虑到协同注意力机制在浅层网络中的成功应用，模块协同注意力网络(ModularCo AttentionNetworks,MCAN)进一步将协同注意力拓展到深层模型中。具体而言，MCAN 借鉴了 Transformer模型的结构，利用自注意力单元对单词之间和区域之间的交互关系进行建模，并引导注意力单元对单词与区域之间的关系进行建模。MCAN进一步对上述两个注意力单元进行模块化组合，最终得到级联的模块协同注意力网络，双线性注意力网络(BilinearAttentionNetworks,BAN)则将低秩双线性池化(Low rankBilinearPooling)应用到注意力计算上，以增强视觉与文本的注意力表示。   1. **视觉与文本任务中多模态交互方法**   早期解决视觉文本任务的方法经常通过以一个模态为条件学习另一 个模态上的注意力的方法（即跨模态注意力）来建模跨模态交互，例如以问题为引导 的图像注意力。受到自然语言领域中自注意力架构El成功实践的启发，最近的几个工作提出了模态内部的交互方法，也就是一个模态内部的注意力。然而现存的大多数方法是为两个输入模态定制的。很少有方法适合更多的输入模态。为了弥补这个问题，这一章为多模态（超过三个模态）输入设计了有效的跨模态和模态内部的交互方法。   1. **研究主题实践现状分析**   自从 2014 年 M. Malinowski 和 M. Fritz 提出了“开放世界”的概念，便打开了探索视觉问答世界的大门，成为人工智能史上关于视觉问答的第一次勇敢的尝试。他们在文中主要提出一种将文本语义的编码解析模型和贝叶斯框架的图像切割模型相结合的方法用于回答与图像相关的问题，该方法可以解决真实场景下包含的较为复杂的自然语言问题，例如其可以回答有关计数、判断、对象等类型的问题，并且还建立了第一个关于视觉问答任务的基线标准。此后，视觉问答系统一直是国内外的热门研究方向，吸引着无数科研工作者和科研机构投身于此。国际的研究机构如卡耐基梅隆大学、加州大学伯克利分校、斯坦福大学等，都 在视觉问答这一跨模态任务中取的了一定的成就和研究成果。斯坦福大学人工智能实验室主任李飞飞教授提出了“视觉基因组”(visual genome)计划，将跨模态的视觉问答作为主要研究内容。微软研究院所研究的项目“语境中的公共对象”的重要任务之一便是探索视觉问答的图像信息和文本信息。2015 年 A. Agrawal 等人提出了一个相对完整的视觉问答模型，针对问题文本采用双层的长短时记忆网络，针对图像的特征信息则采用 VGGNet，然后采用简单的特征融合方式将两个特征融合进而预测答案。同时，他们还提出了迄今为止最大的视觉问答数据集 VQA，VQA 发展到现在已经有两个版本的数据集，VQA v1主要为动画场景的图像，VQA v2 数据集主要为真实场景的图像。国内也有越来越多的人致力于视觉问答研究，高校组织如中国科技大学、浙江大学、哈尔滨工业大学、北京邮电大学、电子科技大学、吉林大学等，都纷纷加入视觉问答的研究队伍中来，近三年来发表的硕博论文从2017年的3篇增长到现在的20篇，足以证明视觉问答具备的研究意义。以此同时，2015 年百度深度学习实验室的M. Ren等人提出了一种全新的框架，他们希望视觉问答模型能够简单有效，且对图像信息产生一个全局概念，因此与M. Malinowski等人的做法相异，他们将文本信息喂入长短时记忆网络前，先将图像特征信息作为第一个单词，然后进行输入，实验结果证实这一策略的正确性。值得注意的是M. Ren等人还提供了一个新的评估视觉问答任务的数据集COCO QA，该数据集也来源于Microsoft COCO图片数据集，不过他们为每张图片提供了标签。现在 COCO-QA 已经成为很多评估视觉问答模型的数据集。一般的视觉问答模型在处理文本信息时都会采用循环神经网络，但 Ma Lin 等人针对视觉问答框架特点做出了新的尝试，他们的图片和文本均采用卷积神经网络进行特征提取工作，图像特征是用 VGGNet模型进行提取的。虽然该模型在 COCO-QA 数据集上的表现没有很好，但是也为视觉问答提供了另一个研究思路。虽然视觉问答系统才兴起不久，但是对其的研究方法却各有不同。近年来已发表的综述不胜枚举。坎姆布鲁等人通过在一个带有注释人类解释的语料库上训练一个语言模型，训练自然语言推理任务的一般自然语言解释，提出了一个常识推理的两阶段框架，首先训练一个自然语言解释模型，然后进一步训练一个以生成的解释作为附加信息的预测模 型。这些方法在预测性能和可解释性方面都取得了良好的性能。然而，需要大量带有人类解释的标记例子，这是昂贵的，有时不可能获得。  本文提出一种改进的基于EM算法的可解释性视觉问答模型结构，可以更好的适应当前可解释性视觉问答的特点。 |

三、研究方案

|  |
| --- |
| **（一）研究目标、研究思路、研究内容和拟解决的实际问题**  本研究的目标是实现面向金融实体识别模型的改进，提高开放环境下系统的准确率、召回率和F1分数值。并将其应用于金融领域实体识别之中，实现在公开数据集上准确识别命名实体。  研究思路：  本文的主要内容包括以下几部分：  首先是绪论。该部分主要介绍选题背景和研究意义，国内外研究现状，主要研究内容、研究思路和研究方法，以及创新点。  第一章 相关理论基础。本部分主要介绍与本研究相关的理论和方法。   1. 可解释性视觉问答数据集的构建。可分为模型的训练集和测试集两部分。   第四章 可解释性视觉问答实验与分析。  最后部分是结论与建议。  **（二）拟采取的研究方法及可行性分析**  **1）研究方法**  本课题拟采用定性与定量分析相结合、主观理论分析与客观实验分析相结合的分析方法。具体包括以下几种研究方法：  （1）文献研究法  文献研究法是根据研究目的，透过调查文献来获得资料，从而全面地、正确地了解掌握所要研究问题的一种方法。文献研究法被广泛用于各种学科研究中。其作用有：①能了解有关问题的历史和现状，帮确定研究课题。②能构成关于研究对象的一般印象，有助于观察和访问。③能得到现实资料的比较资料，有助于了解事物的全貌。在本研究中，通过广泛查阅研究国内外相关文献和研究资料，了解各类型命名实体识别技术的发展现状，从已有的模型之中总结出优点与存在的缺陷，并在其基础上加以完善和改进，整理相关文献，利用理论基础确定研究内容及思路。  （2）实验法  实验法是通过主动变革、挖制研究对象来发现与确认事物间的因果联系的一种科研方法。其主要特点是：第一、主动变革性。观察与调查都是在不干预研究对象的前提下去认识研究对象，发现其中的问题。而实验却要求主动操纵实验条件，人为地改变对象的存在方式、变化过程，使它服从于科学认识的需要。第二，控制性。科学实验要求根据研究的需要。借助各种方法技术，减少或消除各种可能影响科学的无关因素的干扰，在简化、纯化的状态下认识研究对象。第三，因果性。实验是发现确认事物之间的因果联系的有效工具和必要途径。在本研究中具体表现为：通过一整套设定好的实验条件与参数，在同一实验环境下使用同一公开数据集来检测不同模型的预测精度，以形成对比证明新模型的优越性。  （3）BERT  BERT的全称为Bidirectional Encoder Representation from Transformers，是一个预训练的语言表征模型。于2018年由Devlin等提出，并在11项自然语言处理任务上取得优异成绩。该模型采用表义能力较强的双向 Transformer 网络结构来预训练语言模型。Transformer是一种基于 Attention 机制叠加而成的深度网络，不但可以捕捉长距离特征，且具有较好的并行计算能力。  （4）EM算法  EM(Expectation Maximization)算法是在数据不完备的情况下求解最大似然或最大后验估计问题的常用方法。它由Dempster等提出，是一种经典的统计算法，当某一数据模型丢失了某些数据时，EM算法利用当前模型的不完整数据通过反复计算，对缺失数据获得最大的后验概率估计，从而提高模型性能。EM算法的基本原理可以描述如下：假设完全数据集，其中为己观测数据，为缺失数据集，那么在模型参数口上关于的联合密度函数为：。它把问题求解分为两步：E-step和M-step，E-step用于对完全数据集的似然函数求条件期望，M-step用于对求出的期望值求最大化。   1. UniLM   UniLM模型既可以针对自然语言理解任务进行微调，也能用于自然语言生成任务。该模型使用了三种类型的语言建模任务进行预训练：单向模型、双向模型、序列到序列预测模型。其统一建模的实现使用了共享的 Transformer 网络，并且还使用了特定的自注意掩码来控制预测条件所处的上下文。在 GLUE 基准以及 SQuAD 2.0 和 CoQA 问答任务上，UniLM 优于 BERT。   1. **可行性分析：**   关于本研究的可行性，学生将从数据获取，理论支撑，软硬件支持和自身能力四个方面进行论证：   1. **本研究的特色与创新之处** |

四、论文大纲（需列举至三级标题）

|  |
| --- |
| 绪论  一、研究意义与背景  二、国内外文献综述  三、研究思路与方法  第一章 相关理论基础  第一节 视觉问答常用方法介绍  一、基于预训练的方法  二、基于图像中文本信息的视觉问答  三、基于注意力机制的方法  四、视觉与文本任务中多模态交互方法  第二节 文本表示方法  一、词嵌入  二、Word2Vec模型  第二章 可解释性视觉问答数据集的构建  第一节 模型训练集  一、数据来源  二、数据预处理  三、词向量方法  第二节 模型测试集  一、数据来源  二、数据描述  第三章 基于图文多模态可解释性的视觉问答方法  第一节 模型输入层  一、BERT介绍  二、ELV-EST介绍  三、Transformer模型与注意力机制  四、预训练过程  第二节 模型特征层  一、UniLM神经网络模型  二、  三、  第三节 模型输出层  一、注意力机制  第四章 可解释性的视觉问答实验与分析  第一节 实验准备工作  一、实验评估标准  二、实验数据与环境  第二节 实验设定  一、对比模型设定  二、对比参数设定  第三节 实验过程  一、模型参数实验对比与分析  二、不同模型间的实验对比与分析  三、 实验小结  结论及展望  一、研究结论  二、对策建议  三、研究展望  参考文献  附录  致谢 |

五、主要参考文献书目

|  |
| --- |
| [1] Rau L F. Extracting Company Names from Text[C]. Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications,1991: 29-32. |

六、论文写作计划、进行方式和采取主要措施

|  |
| --- |
| 第一阶段 开题准备阶段：  1.2021.03-2021.05：阅读大量文献，确定选题方向，并与指导老师交流讨论，确定选题，并认真撰写开题报告。  2.2021.05-2021.06.11：写出开题报告，制作开题PPT，与指导老师充分沟通，做好开题报告答辩准备  3.2021.06.12：开题报告论证答辩  第二阶段 论文研究阶段  1.2021.06.12-2021.07：收集相关数据来形成训练数据集。阅读金融领域命名实体识别的相关文献，对已有的经典方法进行深入了解，并在总结常用的实体命名方法的创新的切入点与仍存在不足，在此基础上提出改进和创新的想法，为后文的模型的细节改进做铺垫。  2.2021.07-2021.09：对收集到的数据进行清洗和处理，生成词向量，构建初步的命名实体模型，并使用测试数据对其进行训练。  3.2021.09-2021.10：在前文得到的初步模型基础上，对参数做估计、检验及分析以及进一步的调试。与此同时完成测试数据集的选取，将优化过的模型同其他已有方案进行评价指标对比。根据评价指标对比结果，撰写论文报告得出结论以及相关的建议。  第三阶段 论文撰写与修改阶段  1.2020.11 结合前面所做的研究成果，整理完善研究内容，开始撰写完整的学位论文，同时与指导老师交流沟通，不断的修改完善论文。  2.2020.12 提交论文初稿，准备论文预答辩。 |

七、指导教师意见

|  |
| --- |
| 导师签名：  年 月 日 |

八、开题报告会组成人员及意见记录

|  |  |  |
| --- | --- | --- |
| 姓 名 | 职 称 | 工作单位 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| 开题报告记录人签名：  年 月 日 | | |

九、开题报告评价

|  |
| --- |
| 1.综合评价意见。  2.是否通过  （1）通过：按照答辩专家意见修改完善。 [ ]  （2）不予通过：按照答辩专家意见修改后，重新开题。 [ ]  3.答辩成绩  （1）优秀[ ] （2）良好[ ] （3）合格[ ] （4）不合格[ ]  （请在括号内打“√”选择）  4.成员签名： **、 、 、**  、  年 月 日 |