本科毕业设计(论文)开题报告

专业：人工智能　　班级：10051901

<table>
<tr><td>学号</td><td>2019302973</td><td>姓名</td><td colspan="2">牛远卓</td><td>指导教师</td><td colspan="2">牛凯</td></tr>
<tr><td colspan="2">报告题目</td><td colspan="6">基于深度学习的跨模态图像文本匹配研究</td></tr>
<tr><td colspan="2" rowspan="2">题目来源(划 √)</td><td>实践教学
□</td><td>科教融合
√</td><td>创新创业
□</td><td>产教融合
□</td><td>国际合作
□</td><td>本研贯通
□</td></tr>
<tr><td colspan="6"></td></tr>
<tr><td colspan="2">论文类型(划 √)</td><td colspan="6">工程设计类□　　理论研究/论文类 √　　其 他 □</td></tr>
<tr><td colspan="2">报告日期</td><td colspan="3">2023 年 2 月 19 日</td><td>报告地点</td><td colspan="2">线上报告</td></tr>
</table>

开题报告(不少于 1000 字)

一、选题背景、意义及依据

当前新一轮科技革命和产业革命正在发生变革，这与我国高质量发展形成历史性交汇。“十四五”时期我国经济发展应抢抓这一重要变革机遇，为高质量发展“动力换挡”导入强劲引擎。伴随移动互联网、大数据、超级计算、传感网、脑科学等新理论新技术的驱动，以人工智能技术为代表的新一轮科技革命蓬勃发展，以前所未有的速度和方式改变着经济发展，成为高质量发展的重要引擎。计算机视觉与自然语言处理作为人工智能是两个代表性的研究领域。在上述两个领域中，图像文本匹配任务非常重要。

随着互联网上数据规模的急剧增大，数据类型越来越呈现多样化的特点，用户感兴趣的数据模态不再单一，用户的需求也越来越呈现出从单一模态到跨模态的发展态势[1]。模态是指数据的表达形式，包括文本、图像、视频和音频等。跨模态匹配是至少两种模态的数据之间互相匹配，通常是以一种模态来匹配另一种模态的相关数据[2]。通过找出不同模态数据之间的潜在关联，实现相对准确的交叉匹配。

1976 年，文章[3]提出视觉对言语感知的影响，后被用于视听语音识别（Audio Visual Speech Recognition，AVSR）技术并成为多模态概念的雏形。2019 年，文章[4]将多模态学习主要研究方向分为多模态表示、多模态翻译、多模态对齐、多模态融合和多模态协同感知等。跨模态学习是多模态学习的分支，其充分利用了多模态学习中模态间表示、翻译和对齐等策略。跨模态学习与多模态融合的相似之处在于，二者的数据都来自所有模态，但不同之处在于，前者的数据只在某一模态可用，而后者的数据则用于所有模态。

针对图像和文本之间复杂的语义交互作用，传统的跨模态匹配主要采用统计分析方法，如典型相关性分析方法（Canonical Correlation Analysis，CCA）[5]和跨模态因子分析方法

（Cross-modal Factor Analysis，CFA）[6]，其对实际应用场景中不同模态数据的复杂相关性难以建模。文章[7]研究了多媒体信息中文本和图像的联合建模问题，用典型相关析来学习两个模态间的相关性，然而其学习到的都是线性映射，无法有效建立不同模态数据的高阶相关性。近年来，深度学习（Deep Learning）的兴起为跨模态图文匹配提供了新选择，并逐渐成为该领域的热点和主流[8]。一方面，相比于传统方法，深度网络因其高度非线性结构，更适合对模态内特征和模态间语义关系进行挖掘；另一方面，鉴于小批量训练策略的优势，深度网络能够支持对海量数据的处理[9]。基于深度学习的跨模态图文匹配研究因其良好的性能而倍受关注。

本文计划用深度学习的方法，将图像和文本基于注意力机制映射到公共空间中计算相关性，进行匹配。并尝试用不同相关性对比其优劣，探索其中的可解释性。

二、国内外研究现状

2014 年，文章[10]将跨模态建模策略分为直接建模和间接建模，前者指通过建立共享层来直接度量不同模态数据间的相关性，后者指通过构建公共表示空间来建立不同场景不同模态间的语义关联。类似地，2015 年，文章[11]将多模态数据间建立关联的策略分为基于共享层与基于公共表示空间的两种关联方法，该文章对跨模态深度学习模型的设计进行了深入分析。2018 年，文章[12]针对模态间内容相似性度量的技术难点，将跨模态匹配分为公共空间学习方法和跨模态相似性度量方法，并对不同跨模态匹配技术进行总结。2018 年，文章[13]将跨模态匹配方法分为基于子空间的方法、基于深度学习的方法、基于哈希变换的方法和基于主题模型的方法，指出当前跨模态匹配面临的主要问题是缺乏对模态内局部数据结构和模态间语义结构关联的研究。同年，文章[14]从信息抽取与表示、跨模态系统建模两个维度评述了基于表示学习的跨模态匹配模型，并总结了特征抽取方面的研究成果。2018 年，文章[15]探索了联合图正则化的跨模态匹配方法。2019 年，文章[16]简要介绍了近年来跨模态特征匹配及优化的研究进展，并对跨模态数据联合分析方法及跨模态特征匹配面临的问题与挑战进行了概述。文章[15-16]对跨模态匹配方法的具体分支进行了梳理，为相关领域的探索提供了新思路。

基于特征表示的方法面向跨模态原始数据，其关注点在于获得更好的输入特征，通过模态特征学习减小模态异构问题。相比于基于特征表示的方法，基于图文匹配的方法更关注于不同模态间的结构关联，此类方法通过研究图像和文本模态间的语义对应关系来增强模态间特征表示的一致性。目前主流的基于图文匹配的方法按照模态间语义结构关联的不同可分为三类：图像-文本局部对齐的方法、跨模态重构的方法和图文联合嵌入的方法。

（1）图像-文本局部对齐的方法

图像-文本局部对齐的方法一般通过学习同一实例不同模态特征之间的关系来推断句子片段与图像区域之间的潜在对齐，进而实现图文匹配。

由于图像-文本局部对齐的方法更关注局部精细的信息，也常用于细粒度的跨模态图文匹配任务。文章[17]针对服装领域提出了 Fashion BERT 模型，相比于感兴趣区域（Region of Interest，RoI）模型，时尚文本倾向于描述更精细的信息。因此，Fashion BERT 模型由 BERT（Bidirectional Encoder Representations from Transformers）模型[18]引申得到。BERT 是一种双向注意力语言模型，作为 Transformer[19]在自然语言处理任务的变体之一，其主要作用是对单模态文本数据进行编码。Fashion BERT 在提取图像表示时将每个图像分割成相同像素的补丁，作为 BERT 模型的序列输入，在匹配时将文本标记和图像补丁序列进行连接。实验表明该方法可以在一定程度上掩盖图像中不相关的信息，减小了检测到无用和重复区域的可能性。

（2）跨模态重构的方法

与图像-文本局部对齐的方法关注局部信息的方式不同，跨模态重构的方法更关注全局信息，此类方法通常利用一种模态信息来重构对应模态，同时保留重建信息，能够增强跨模态特征一致性及语义区分能力。

由于跨模态相关性是高度非线性的，而受限玻尔兹曼机模型（Restricted Boltzmann Machine，RBM）很难直接对这种相关性进行学习。考虑在每个模态的预训练层上训练 RBM 的方法，文章[20]提出不同模态数据共享权重的双模深度自编码器模型，在仅给定数据的情况下进行跨模态重建，从而发现跨模态的相关性。在此研究基础上，文章[21]提出了一种图像文字匹配的方法，引入了结构-内容神经语言（Structure-Content Neural Language Model，SC NLM）模型，SC-NLM 通过编码器学习图像句子联合嵌入，并根据编码器产生的分布式表示，将句子的结构与内容分离。

（3）图文联合嵌入的方法

相比于图像-文本对齐的方法和跨模态重构的方法，图文联合嵌入的方法一般结合了全局和局部信息作为语义特征的嵌入，因此能够学习到更好的特征判别性。此类方法一般通过图像和文本模态数据的联合训练及语义特征的嵌入来学习图像文本的相关性，进而实现图文匹配。

针对模态特征的不一致性导致的跨模态迁移困难的问题，文章[22]使用弱对齐的数据来

学习具有强对齐的跨模态表示，在共享层使用多层感知器将文本信息映射到与视觉模态相同维度的表示空间中。该模型同时用到了微调和统计正则化的方法，可以在训练数据没有明确对齐的情况下跨模态检测相同的概念，具有良好的匹配性能。为了寻找公共表示空间来直接比较不同模态的样本，文章[23]提出了深度监督跨模态匹配（Deep Supervised Cross-Modal Retrieval，DSCMR）方法，通过最小化样本在标签空间和公共表示空间中的判别损失来监督模型学习判别特征，以保持不同类别语义样本间的区分度，并使用权重共享策略来消除多媒体数据在公共表示空间中的跨模态差异。相比以往的方法，DSCMR 的学习策略可充分利用成对标签信息和分类信息，有效学习了异构数据的公共表示。

三、课题研究目标、研究内容、研究方法及关键技术

1．研究目标

基于深度学习方法，拟采用卷积神经网络完成训练图像的特征提取，并通过设计创新性的网络结构，实现更高质量的特征学习，获得更具判别力的图像特征，最终利用递归神经网络生成图像文本匹配。

2．研究内容和方法

(1)前期准备：阅读 5-7 篇相关领域的科研论文，了解图像识别相关的基本知识，掌握卷积神经网络和对齐技术的基本原理并学会使用深度学习相关工具，例如 PyTorch 等。

(2)搭建多模态双向递归神经网络结构：基于 Deep CNN 和注意力机制搭建卷积神经网络算法，实现高层图像特征提取，并针对搭建的网络结构进行改进。

(3)搭建双向递归神经网络：基于 Bi-LSTM 搭建递归神经网络算法，并根据卷积神经网络提取的高层语义特征与图像特征匹配，并针对搭建的网络结构进行改进。

(4)性能评估：通过在数据集 MSCOCO 和 Flickr30k 进行实验，采用 Recall@1 指标以及 Recall@5 指标进行匹配准确率评估。

3．关键技术

本课题主要采用卷积神经网络，循环神经网络 LSTM，跨模态匹配等关键技术开展研究工作。

四、论文所遇到的困难和问题、拟采取的解决措施及预期达到的目标

1．遇到的困难和问题：

(1)对于参考论文所用技术及算法的学习仍不完善，对于深度学习领域研究的实践经验较匮乏。

(2)研究所涉及的知识领域较广，没有实验经验，无法在实践之前就对各个模型的预期效果进行评估。

(3)难以将多种不同的方法与技术进行有效融合，进而建立一个能够获得更高性能的模型。

2．解决措施及预期目标：

(1)尽快学习相关知识结构，详细阅读并理解不同参考文章中的模型、方法、实践步骤。

(2)迅速了解并掌握 PyTorch 等代码编写工具，提升代码编写能力，对各篇论文中的不同模型进行复现并对预期效果进行评估。

(3)整合各个模型的优缺点，预期能够构建一个相对传统方法有更好表现得深度学习模型，并达到性能上的优化和提升。

五、论文进度安排

2022 年 12 月 1 日-2022 年 12 月 15 日，文章调研

2022 年 12 月 16 日-2022 年 12 月 31 日，数据获取及预处理

2023 年 1 月 1 日-2023 年 1 月 31 日，开题准备

2023 年 2 月 1 日-2023 年 2 月 28 日，深度学习模型搭建及调试

2023 年 3 月 1 日-2023 年 3 月 31 日，深度学习模型搭建及调试

2023 年 4 月 1 日-2023 年 4 月 15 日，匹配性能评估，模型改进

2023 年 4 月 16 日-2023 年 4 月 30 日，参数优化

2023 年 5 月 1 日-2023 年 5 月 31 日，结果整理，论文写作

2023 年 6 月 1 日-2023 年 6 月 15 日，论文写作，答辩准备

六、参考文章

[1] 冯 霞 胡志毅. 跨模态检索研究进展综述[J].计算机科学，2021, 48 (8): 13-23.

[2] Kunpeng Li, Yulun Zhang，et al. Visual Semantic Reasoning for Image-Text Matching. ICCV, 2019.

[3] Mcgurk H，Macdonald H. Hearing lips and seeing voices[J]. Nature，1976，264(5588): 746-748.

[4] Baltrusaitis T，Ahuja C，Morency L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2019，41(2): 423-443.

[5] Harold H. Relations between two sets of variates[J]. Biometrika，1936，28(3): 321-377.

[6] Li D G，Dimitrova N，Li M K，et al. Multimedia content processing through cross-modal association[C]//ACM International Conference on Multimedia，2003.

[7] Rasiwasia N，Pereira J C，Coviello E，et al. A new approach to cross- modal multimedia retrieval[C]//International Conference on Firenze，2010.

[8] Ji Z Y，Yao W N，Wei W，et al. Deep multi-level semantic Hashing for cross-modal retrieval[J]. IEEE Access，2019，7: 23667-23674.

[9] Wang C，Yang H J，Meinel C. Deep semantic mapping for cross-modal retrieval[C]// International Conference on Tools with Artificial Intelligence，2015.

[10] Feng F X，Wang X J，LI R F. Cross-modal retrieval with correspondence autoencoder[C]// ACM International Conference on Multimedia，2014.

[11] 冯方向. 基于深度学习的跨模态检索研究[D]. 北京：北京邮电大学，2015.

[12] Peng Y，Huang X，Zhao Y. An overview of cross-media retrieval: concepts，methodologies，benchmarks，and challenges[J]. IEEE Transactions on Circuits and Systems for Video Technology，2018，28(9): 2372-2385.

[13] 欧卫华，刘彬，周永辉，等. 跨模态检索研究综述[J]. 贵州师范大学学报(自然科学版)，2018，36(2)：114-120.

[14] 李志义，黄子风，许晓绵. 基于表示学习的跨模态检索模型与特征抽取研究综述[J]. 情报学报，2018，37(4):422-435.

[15] Ayyavaraiah M，Venkateswarlu B. Joint graph regularization based semantic analysis for cross-media retrieval: a systematic review[J]. International Journal of Engineering & Technology，2018，7: 257-261.

[16] Ayyavaraiah M，Venkateswarlu B. Cross media feature retrieval and optimization: a contemporary review of research scope，challenges and objectives[C]//International Conference on Computational Vision and Bio Inspired Computing，2019.

[17] Gao D H，Jin L B，Chen B，et al. Fashion BERT: text and image matching with adaptive loss for cross-modal retrieval [C]//International ACM SIGIR Conference on Research and Development in Information Retrieval，2020.

[18] Devlin J，Chang M W，Lee K，et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810. 04805，2018.

[19] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[C]//Annual Conference on Neural Information Processing Systems，2017.

[20] Ngiam J，Khosla A，Kim M，et al. Multimodal deep learning[C]// International Conference on Machine Learning，2011.

[21] Kiros R，Salakhutdinov R，Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models[J]. arXiv:1411. 2539，2014.

[22] Castrejón L，Aytar Y，Vondrick C，et al. Learning aligned cross-modal representations from weakly aligned data[C]// IEEE Conference on Computer Vision and Pattern Recognition，2016.
[23] Zhen L L，Hu P，Wang X，et al. Deep Supervised Cross-modal Retrieval[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition，2019.

指导教师意见：

同意开题。

签名：

2023 年 02 月 20 日

<table>
<tr><td colspan="2">开题评议小组成员：</td></tr>
<tr><td colspan="2">开题评议小组意见：(包括对论文的选题、难度、进度、工作量、论文形式意见)：
1. 论文选题： ☐ 有理论意义; ☐有实用价值;☐ 有理论意义与实用价值;
☐ 意义不大。
2. 论文的难度： ☐ 偏高; ☐ 适当;☐ 偏低。
3. 论文的工作量： ☐ 偏大; ☐ 适当;☐ 偏小。
4. 进度： ☐ 可行;☐ 不可行;
5. 学生开题报告中反映出的综合能力和表达能力： ☐ 好; ☐ 较好;一般;☐ 较差。
6. 论文形式意见：☐ 可行;☐ 不可行;
7. 对论文选题报告的总体评价： ☐ 好; ☐ 较好;☐ 一般;☐ 较差。
(在相应的方块内作记号“√”)</td></tr>
<tr><td>评 议
结 论</td><td>是否同意论文选题报告：☐ 同意;☐ 需重做
(在相应的方块内作记号“√”)

评议小组组长签名： 年 月 日</td></tr>
</table>