



CCF 学生领航计划

第一期：英文学术论文写作指南

于静
中科院信息工程研究所

2022.06.22@CCF SPP

认知启发的跨模态智能研究组 (GogModal)



于静，中科院信息工程研究所，副研究员， yujing02@iie.ac.cn

近年研究：跨模态智能，包括视觉问答、视觉对话、跨媒体检索等

CCF YOCSEF学术委员会委员，CCF 多媒体技术专委会专业委员。在国内外学术会议及期刊发表论文50余篇，包括ICML、CVPR、AAAI、IJCAI、ACM MM、TIP、TMM、PR等，担任ACL、CVPR、ICCV、AAAI、IJCAI、TMM、PR等国际会议和期刊审稿人。主持与参与国家重点研发计划项目、中科院先导专项等项目。讲授中国科学院大学博士课程《基于深度学习的自然语言处理》

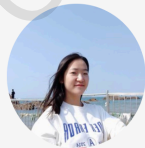
研究组主页



知乎专栏



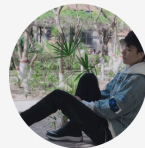
PhD Students



Jiamin Zhuang

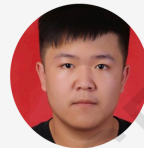
PhD Student

Cross modal retrieval, Video-text grounding



Yuanmin Tang

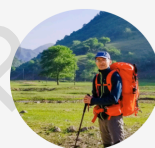
PhD Students



Yunpeng Li

PhD Student

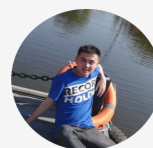
Relation Extraction, Visual Question Answering, Visual



Zihao Zhu

Master Student

Visual Question Answering, Image Captioning, Visual Reasoning



Xiaozhe Jiang

Master Student

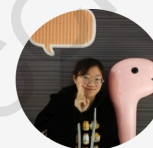
Visual Dialog, Cross-lingual Pre-training, Image Captioning



Jingjing Guo

Master Student

Cross-Modal Information Retrieval, Cross-Modal Feature Fusion



Jialin Chen

Master Student

Video Captioning, Cross-media Retrieval



Jialu Chen

Master Student

Video captioning, Video-text grounding

Visitors



Yiming Xu

Visiting Student

Cross-modal Information Retrieval, Multi-style Image



Mingzhe Liao

Visiting Student

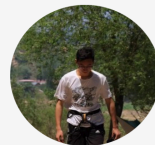
visual question answering



Rundong Li

Visitors

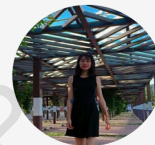
Cross-modal reasoning



Shunyu Zhang

Master Student

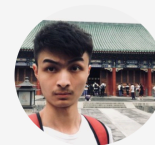
Visual Dialog, Visual Reasoning



Xuedan Hu

Master Student

Query Expansion, Question Answering



Yang Ding

Master Student

Cross modal retrieval, Video-text grounding



Yuan Chai

Master Student

Scene Graph Generation, Visual Relation Detection

Alumni

Yuhang Lu

Postgraduate (2018)

Alibaba Group, Beijing, China

Chenghao Yang

Undergraduate (2019)

Columbia University, USA

Zhuoqian Yang

Undergraduate (2019)

Carnegie Mellon University, USA

Xiaomei Liu

Postgraduate (2016)

University of International

Meizi Zhou

Postgraduate (2016)

University of Minnesota, USA

Mengya Liu

Postgraduate (2016)

University of Southampton, UK

Ran Qu

Undergraduate (2018)

Xiang Wang

Postgraduate (2017)

WeChat Group, Tencent

Xiaoman Zhang

Undergraduate (2018)

HUAWEI, China

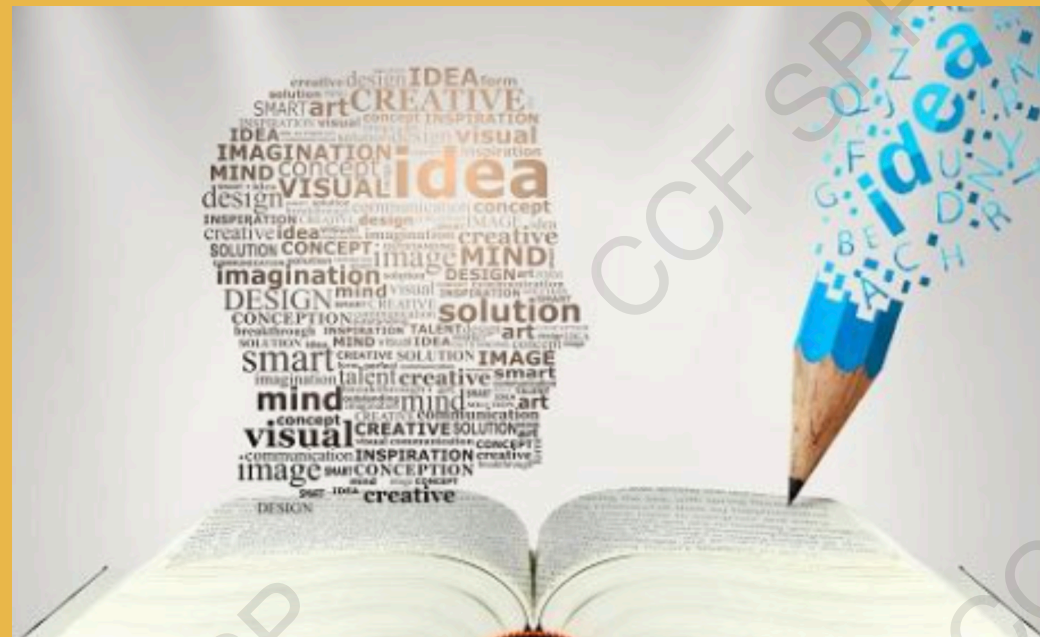
主要内容

- 1/ 学术研究^与论文写作
- 2/ 学术论文之写作思路
- 3/ 学术论文之图表制作
- 4/ 学术论文之英文规范
- 5/ 学术论文之日常积累
- 6/ 本讲总结与心得分享

主要内容

- 1/ 学术研究^与论文写作（第一期）
- 2/ 学术论文之写作思路（第一期）
- 3/ 学术论文之图表制作（第二期）
- 4/ 学术论文之英文规范（第二期）
- 5/ 学术论文之日常积累（第二期）
- 6/ 本讲总结与心得分享（第二期）

学术研究 与论文写作



什么是学术研究？

技术——走野路



科研——逛景区



学术——攀珠峰



为什么要做学术研究？



你的梦想什么

我要翻身!



没钱，没有梦想，
没有目标，浑浑噩噩



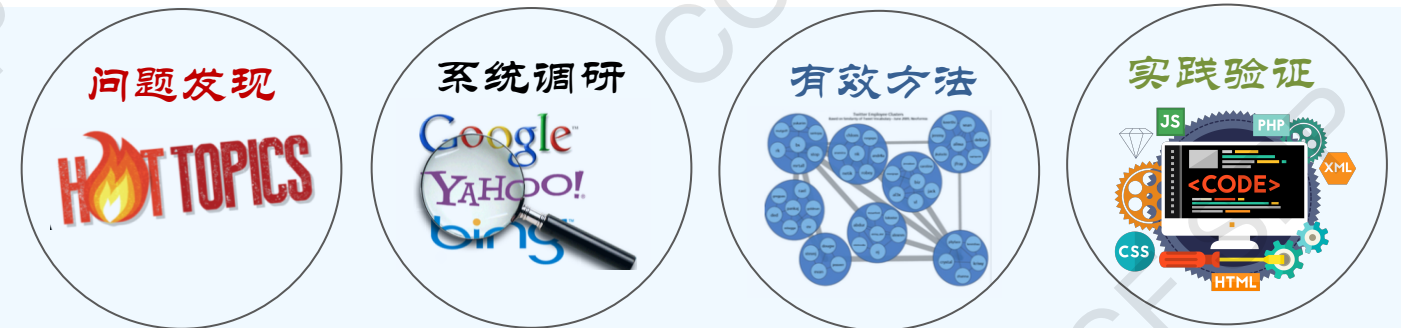
人丑就要多读书



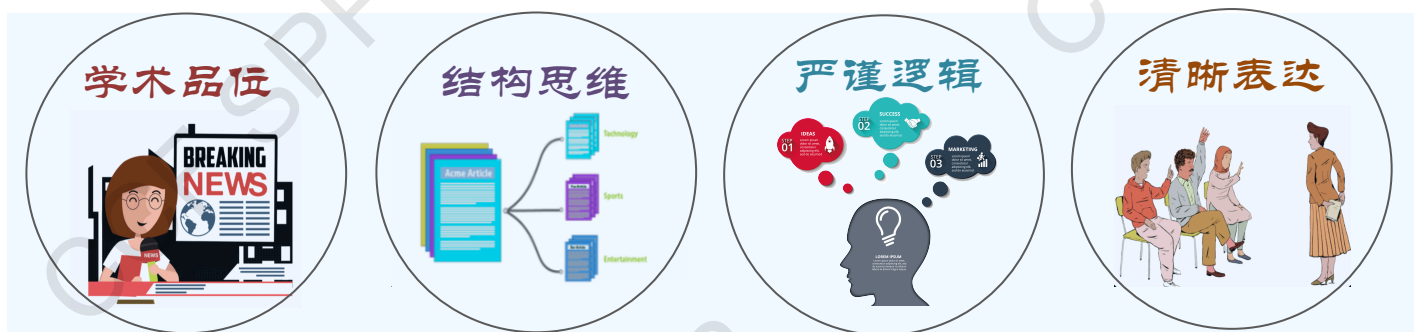
卡里有钱就行了

为什么要做学术研究？

方法论
解决问题



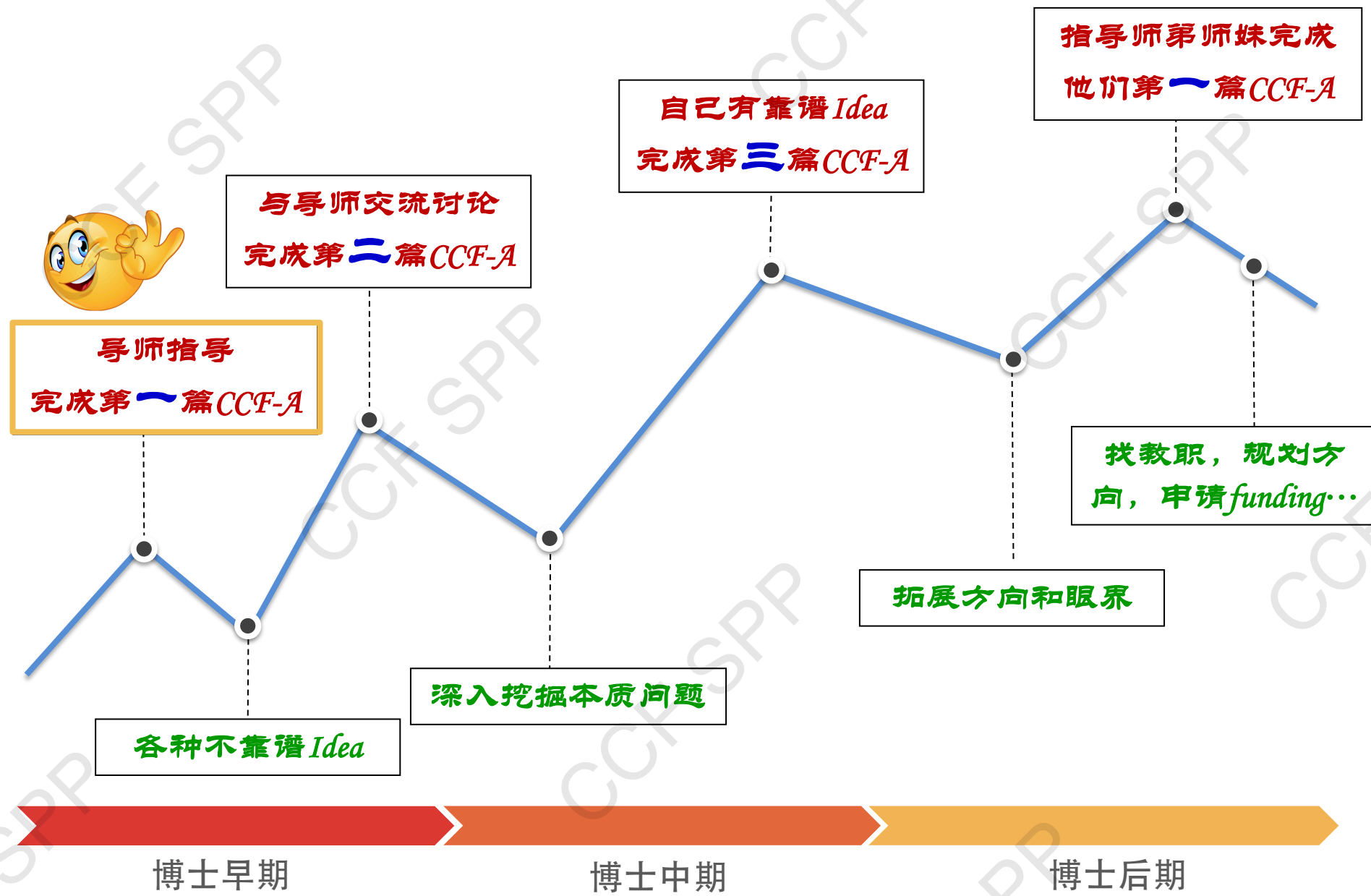
认知力
科学素养



价值观
长线思维



学术研究修炼之路——四生四世



学术研究修炼之路——第一篇论文的诞生

确定 *Topic*:

- 导师指导
- 自力更生
- 找人请教

大量阅读文献:

- *Idea List*
- *Math List*
- *English List*

确定 *Method*:

- *Novel enough?*
- *Good enough?*
- *Feasible enough?*

实验分析:

- *Comprehensive*
- *Strength/Weakness*
- *How/When/Why?*

学术研究过程

研究领域 → 研究动机 → 问题提出 → 方法设计 → 实验验证 → 总结展望
Topic *Motivation* *Problem* *Method* *Experiments* *Conclusion*

论文写作过程



我们的论文出了什么问题？

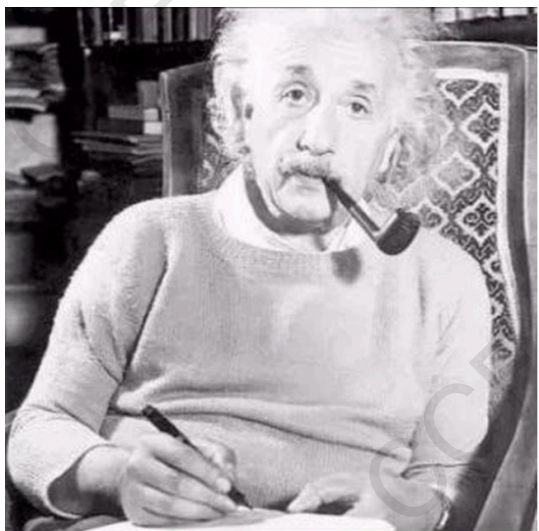


credit: 图片来自互联网

学生和导师对论文写作的认识不一致！

我们的论文出了什么问题？

我眼中的论文



credit: 图片来自互联网

导师眼中的论文

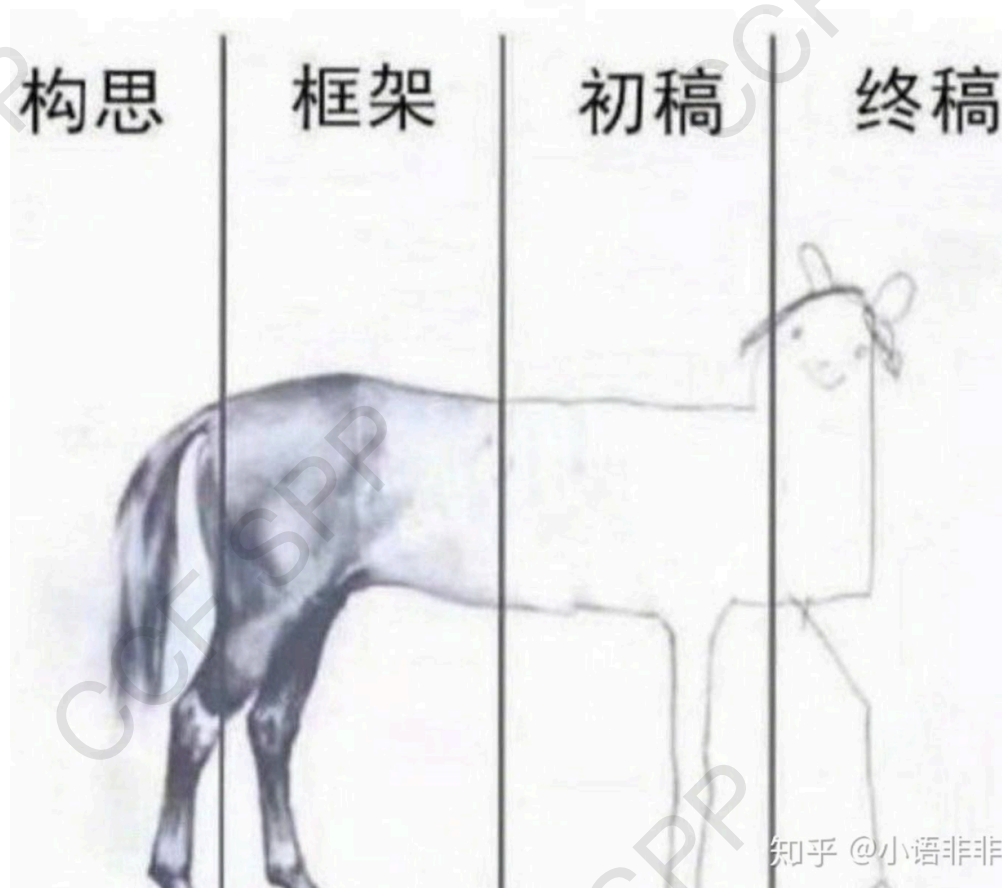


审稿人眼中的论文



论文总是写给自己看，别人看不懂！

我们的论文出了什么问题？



credit: 图片来自互联网

论文无法准确表达研究内容！

我们的论文出了什么问题？



credit: 图片来自互联网

论文 *Story* 难以自圆其说！

我们的论文出了什么问题？

我离论文写完就差这么**一点点**



credit: 图片来自互联网

写作拖延，无法赶上 *deadline*!

我们的论文出了什么问题？

当**论文**修改到第五稿的状态



credit: 图片来自互联网

不知道论文如何逐步完善！

我们的论文出了什么问题？



发不了SCI不要紧

还可以来发知音，故事会啊

credit: 图片来自互联网 知乎 @语非

结果：CCF-A的研究内容只达到CCF-C的论文水平！

我们的论文出了什么问题——小结

学生和导师对论文写作的认识不一致！

论文总是写给自己看，别人看不懂！

论文无法准确表达研究内容！

论文 *Story* 难以自圆其说！

不知道论文如何逐步完善！

写作拖延，无法赶上 *deadline*！

...

写作思路

英文规范

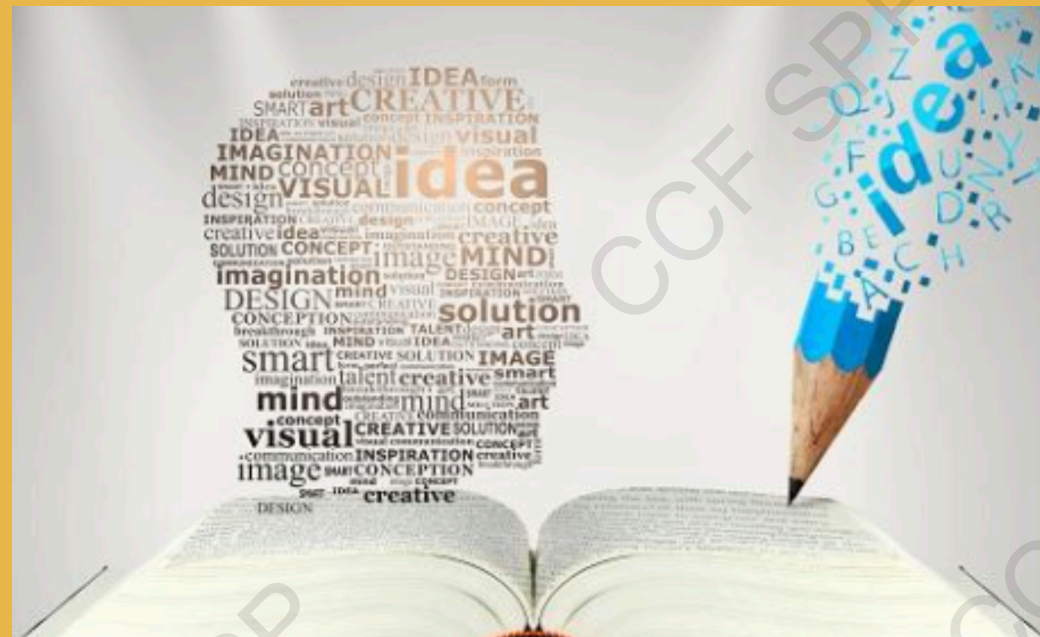
日常积累

我们的目标——AI领域主要会议及期刊

研究方向	主要会议	主要期刊
计算机视觉/ 多媒体	CVPR, ICCV, ECCV, ACM MM, ICASSP, ICMR, ICME 中国多媒体大会(ChinaMM) 中国模式识别与计算机视觉大会(PRCV)	TIP, IJCV, TMM, PR, TCSVT, TOMCCAP, CVIU 《计算机学报》《软件学报》 《计算机研究与发展》《中国图象图形学报》
自然语言处理	ACL, EMNLP, NAACL, COLING, CoNLL, AACL 全国计算语言学大会 (CCL) 全国知识图谱与语义计算大会 (CCKS) CCF国际自然语言处理与中文计算会议 (NLPCC) 全国机器翻译研讨会 (CWMT)	TACL, TASLP, TALLIP, Computer Speech and Language
机器学习/ 人工智能	NeurIPS, ICML, ICLR, IJCAI, AAAI	TPAMI, TNNLS
数据挖掘	SIGMOD, SIGKDD, VLDB, ICDE, SIGIR, CIKM, WWW, WSDM	TKDE, VLDBJ, TKDD

参考《中国计算机学会推荐国际学术会议和期刊目录》

学术论文 之写作思路



从科学问题是否探究本质说起

CCF—A

科学问题 聚焦领域前沿，探索本质问题

解决方法 专注核心挑战，具普适性方法

Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering

Zihao Zhu^{1,2*}, Jing Yu^{1,2*†}, Yujing Wang³, Yajing Sun^{1,2}, Yue Hu^{1,2} and Qi Wu⁴

¹Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

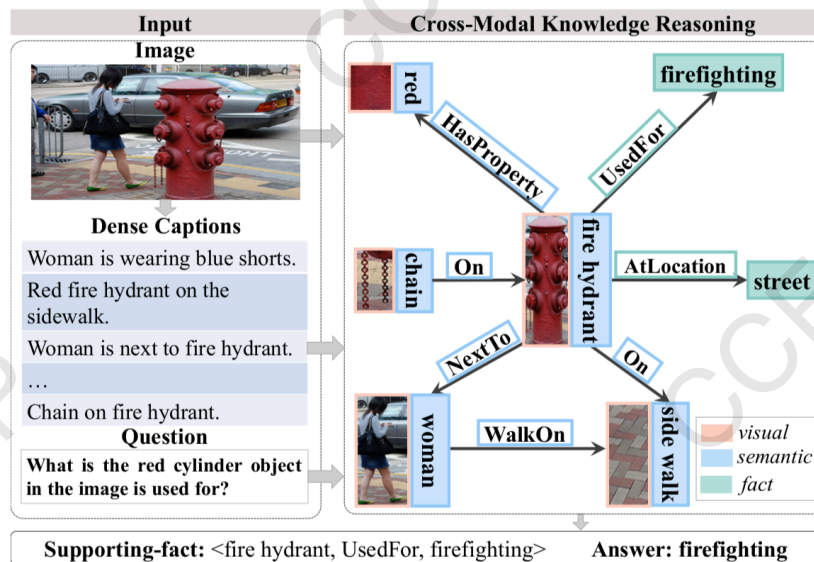
²School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

³Microsoft Research Asia, Beijing, China

⁴University of Adelaide, Australia

IJCAI 2020

{zhuzihao, yujing02, sunyajing, huyue}@iie.ac.cn, yujwang@microsoft.com, qi.wu01@adelaide.edu.au



CCF—C

追逐领域热点，解决表面问题

盲从热门方法，模型增量修改

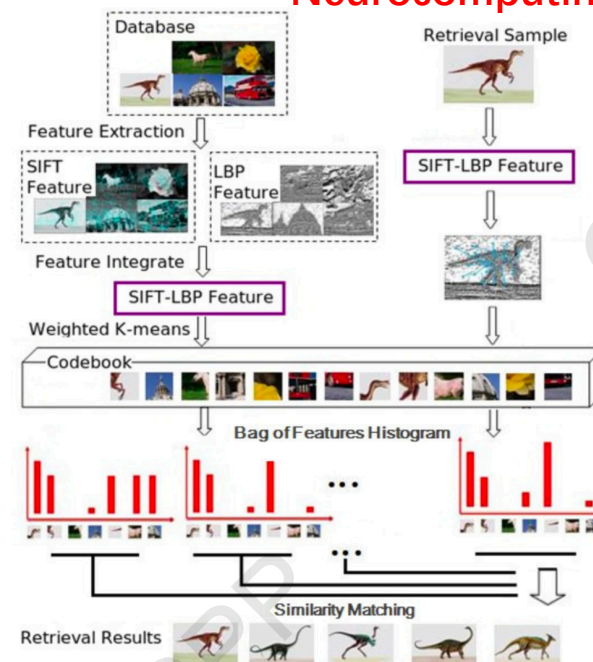
Feature integration analysis of bag-of-features model for image retrieval

Jing Yu^a, Zengchang Qin^{a,*}, Tao Wan^b, Xi Zhang^a

^aIntelligent Computing and Machine Learning Lab, School of Automation Science and Electrical Engineering, Beihang University, Beijing, China

^bSchool of Medicine, Boston University, Boston, USA

Neurocomputing 2011



从科学问题是否探究本质说起

CCF—A

- 问题-方法-实验，相互呼应

- 问题：有理有据，足够具体
- 方法：针对问题设计，每一步设计目标明确
- 实验：针对方法逐一证明，针对动机逐一分析

CCF—C

- 问题-方法-实验，各为其说

- 问题：大家都在研究，所以我研究
- 方法：step1->step2->step3
- 实验：达到了SOTA，缺乏分析

一篇论文的组成

- ◆ **标题：**核心问题与创新点的高度凝练
- ◆ **摘要：**动机、亮点与效果的概要介绍
- ◆ **引言：**动机、现状、方法和贡献介绍
- ◆ **相关工作：**引言现有工作的详细介绍
- ◆ **研究方法：**引言模型部分的详细介绍
- ◆ **实验分析：**引言模型效果的详细介绍
- ◆ **总结展望：**经过实验验证后给的结论
- ◆ **致谢：**对作者以外人员/机构的感谢
- ◆ **参考文献：**正文出现的按照格式引用

一篇论文的组成——标题

基本要求（不超过15个单词）

☀ **英文形式规范**

☀ **语言精炼简洁**

☀ **范围大小适当**

好标题

☀ **反应核心问题**

☀ **突出技术创新**

☀ **保护知识产权**

☀ **易于记忆传播**

一篇论文的组成——标题 (Bad Cases)

From shallow to deeper: compositional reasoning
over graphs for visual question answering

太宽泛!

PERT: **adaPtive** Evidence-driven Reasoning **neTwork** for Machine Reading
Comprehension with Unanswerable Questions

不规范!

Understanding like humans: multimodal representation for the visual
information in visual dialog

没依据!

Graph Neural Networks for Image-Text Matching

没创新!

A **Plug-and-Play novel Tree Loss Function** for Unbiased Scene Graph
Generation based on **Upgraded Transformer framework**

太冗余!

KBGN: Knowledge-Bridge Graph Network for Adaptive Vision-Text Reasoning
in Visual Dialogue.

难记忆!

一篇论文的组成——标题 (My Own)

MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering (CVPR 2022)

ET-BERT: A Contextualized Datagram Representation with Pre-training Transformers for Encrypted Traffic Classification (WWW 2022)

CogTree: Cognition Tree Loss for Unbiased Scene Graph Generation (IJCAI 2021)

DualDV: An Adaptive Dual Encoding Model for Visual Dialogue (AAAI 2020)

Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering (IJCAI 2020)

DAM: Deliberation, Abandon and Memory Networks for Generating Detailed and Non-repetitive Responses in Visual Dialogue (IJCAI 2020)

一篇论文的组成——标题 (CV & NLP & ML)

Zero-Shot Text-to-Image Generation
(DALL·E, arxiv 2021)

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (ICCV 2021)

Sketch, Ground, and Refine: Top-Down Dense Video Captioning (CVPR 2021)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
(NAACL 2019)

Knowledgeable Prompt-tuning : Incorporating Knowledge into Prompt Verbalizer for
Text Classification (ACL 2022)

Semi-Supervised Classification with Graph Convolutional Networks (ICLR 2017)

A Simple Framework for Contrastive Learning of Visual Representations
(SimCLR, 2020 ICML)

一篇论文的组成——摘要

基本要求

- ☀ 标题的扩充、引言的概括
- ☀ 涵盖动机、亮点、效果等
- ☀ 200词左右
- ☀ 逻辑清晰

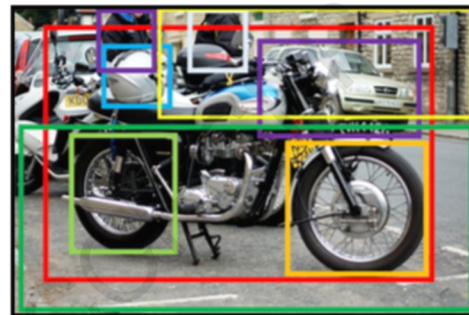
一篇论文的组成——摘要

CVPR 2020

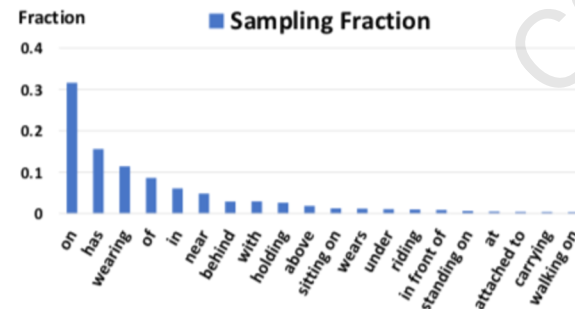
Unbiased Scene Graph Generation from Biased Training

Kaihua Tang¹, Yulei Niu³, Jianqiang Huang^{1,2}, Jiaxin Shi⁴, Hanwang Zhang¹

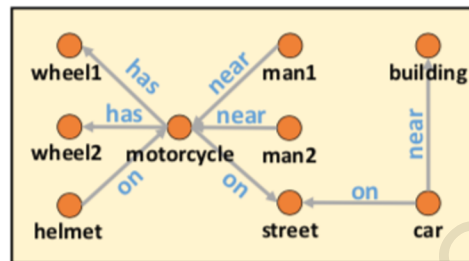
¹Nanyang Technological University, ²Damo Academy, Alibaba Group, ³Renmin University of China, ⁴Tsinghua University



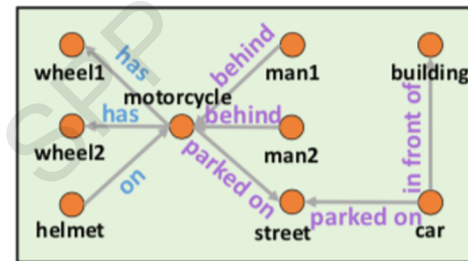
(a) Input Image



(b) Distribution of Predicate Sampling Fraction



(c) Biased Generation



(d) Unbiased Generation

一篇论文的组成—摘要

Unbiased Scene Graph Generation from Biased Training CVPR 2020

Today's scene graph generation (SGG) task is still far from practical, mainly due to the severe training bias, e.g., collapsing diverse human walk on/ sit on/lay on beach into human on beach. Given such SGG, the down-stream tasks such as VQA can hardly infer better scene structures than merely a bag of objects. However, debiasing in SGG is not trivial because traditional debiasing methods cannot distinguish between the good and bad bias, e.g., good context prior (e.g., person read book rather than eat) and bad long-tailed bias (e.g., near dominating behind/in front of). In this paper, we present a novel SGG framework based on **causal inference** but not the conventional likelihood. We first build a causal graph for SGG, and perform traditional biased training with the graph. Then, we propose to draw the **counterfactual causality** from the trained graph to infer the effect from the bad bias, which should be removed. In particular, we use **Total Direct Effect** as the proposed final predicate score for unbiased SGG. Note that our framework is agnostic to any SGG model and thus can be widely applied in the community who seeks unbiased predictions. By using the proposed **Scene Graph Diagnosis** toolkit¹ on the SGG benchmark Visual Genome and several prevailing models, we observed significant improvements over the previous state-of-the-art methods.

SGG存在的挑战：数据偏置

SGG方法的问题：未区分偏置的好坏

本文方法的思路：因果推断

本文具体的亮点：

构建因果图-->反事实推断-->优化目标

本文方法的优势：

- 模型无关，广泛适用
- 模型效果达到新SOTA

一篇论文的组成——引言

基本要求（摘要的扩展，最难的部分**）**

☀ **研究背景与挑战**

☀ **提出问题与原因**

☀ **相关工作与不足**

☀ **本文研究思路**

☀ **本文主要贡献**

一篇论文的组成——引言

CCF—A

- 问题-方法-实验，相互呼应

- 问题：有理有据，足够具体

- 背景阐述聚焦重点

- 问题提出明确具体

- 聚焦研究动机，总结现状问题

- 基于研究动机，概述研究方法

- 面向领域需求，拔高论文贡献

- 方法：针对问题设计，每一步设计目标明确

- 实验：针对方法逐一证明，针对动机逐一分析

CCF—C

- 问题-方法-实验，各为其说

- 问题：大家都在研究，所以我研究

- 方法：step1->step2->step3

- 实验：达到了SOTA，缺乏分析

一篇论文的组成——引言

IJCAI 2020

Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering

Zihao Zhu^{1,2*}, Jing Yu^{1,2*†}, Yujing Wang³, Yajing Sun^{1,2}, Yue Hu^{1,2} and Qi Wu⁴

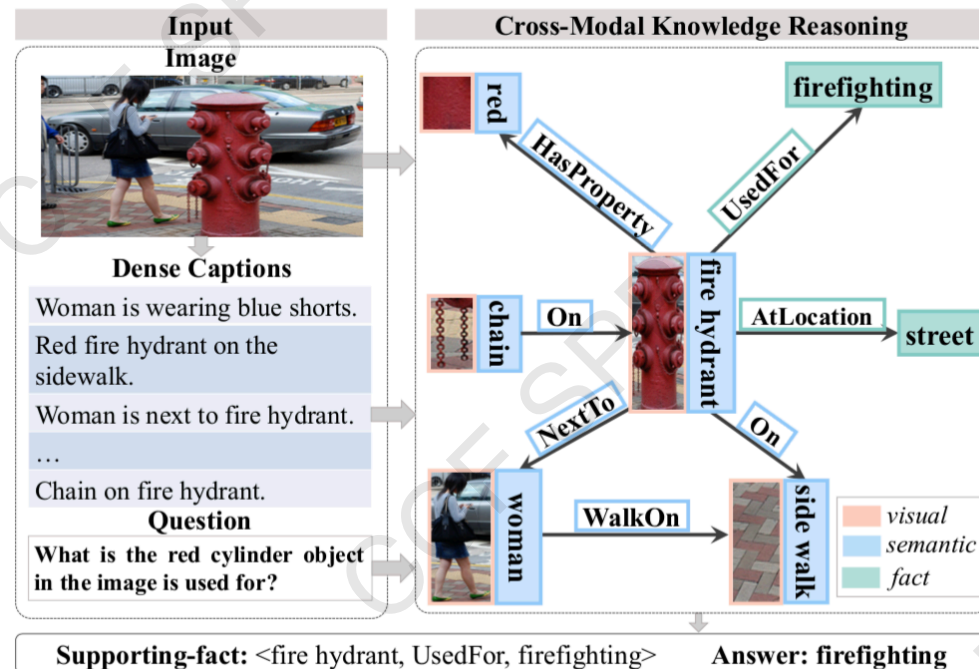
¹Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

²School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

³Microsoft Research Asia, Beijing, China

⁴University of Adelaide, Australia

{zhuzihao, yujing02, sunyajing, huyue}@iie.ac.cn, yujwang@microsoft.com, qi.wu01@adelaide.edu.au



一篇论文的组成——引言

Abstract

Fact-based Visual Question Answering (FVQA) requires external knowledge beyond visible content to answer questions about an image, which is challenging but indispensable to achieve general VQA. One limitation of existing FVQA solutions is that they jointly embed all kinds of information without fine-grained selection, which introduces unexpected noises for reasoning the final answer. How to capture the question-oriented and information-complementary evidence remains a key challenge to solve the problem. In this paper, we depict an image by a multi-modal heterogeneous graph, which contains multiple layers of information corresponding to the visual, semantic and factual features. On top of the multi-layer graph representations, we propose a modality-aware heterogeneous graph convolutional network to capture evidence from different layers that is most relevant to the given question. Specifically, the intra-modal graph convolution selects evidence from each modality and cross-modal graph convolution aggregates relevant information across different modalities. By stacking this process multiple times, our model performs iterative reasoning and predicts the optimal answer by analyzing all question-oriented evidence. We achieve a new state-of-the-art performance on the FVQA task and demonstrate the effectiveness and interpretability of our model with extensive experiments. The code is available at <https://github.com/astro-zihao/mucko>.

1 Introduction

Visual question answering (VQA) [Antol *et al.*, 2015] is an attractive research direction aiming to jointly analyze multi-modal content from images and natural language. Equipped with the capacities of grounding, reasoning and translating, a VQA agent is expected to answer a question in natural language based on an image. Recent works [Cadene *et al.*, 2019;

*Equal contribution.

†Corresponding author.

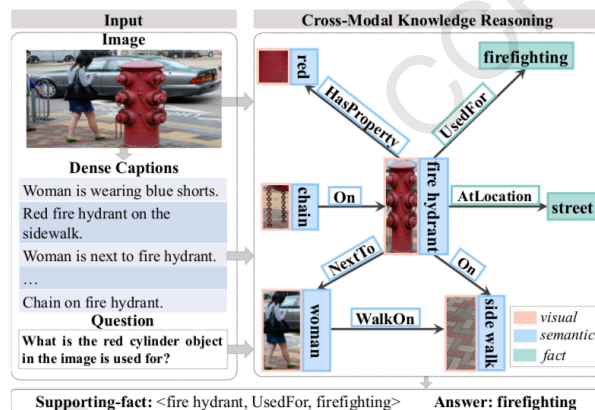


Figure 1: An illustration of our motivation. We represent an image by multi-layer graphs and cross-modal knowledge reasoning is conducted on the graphs to infer the optimal answer.

Li *et al.*, 2019b; Ben-Younes *et al.*, 2019] have achieved great success in the VQA problems that are answerable by solely referring to the visible content of the image. However, such kinds of models are incapable of answering questions which require external knowledge beyond what is in the image. Considering the question in Figure 1, the agent not only needs to visually localize ‘the red cylinder’, but also to semantically recognize it as ‘fire hydrant’ and connects the knowledge that ‘fire hydrant is used for firefighting’. Therefore, how to collect the question-oriented and information-complementary evidence from visual, semantic and knowledge perspectives is essential to achieve general VQA.

To advocate research in this direction, [Wang *et al.*, 2018] introduces the ‘Fact-based’ VQA (FVQA) task for answering questions by joint analysis of the image and the knowledge base of facts. The typical solutions for FVQA build a fact graph with fact triplets filtered by the visual concepts in the image and select one entity in the graph as the answer. Existing works [Wang *et al.*, 2017; Wang *et al.*, 2018] parse the question as keywords and retrieve the supporting-entity only by keyword matching. This kind of approaches is vulnerable when the question does not exactly mention the visual concepts (e.g. synonyms and homographs) or the mentioned information is not captured in the fact graph (e.g. the visual

attribute ‘red’ in Figure 1 may be falsely omitted). To resolve these problems, [Narasimhan *et al.*, 2018] introduces visual information into the fact graph and infers the answer by implicit graph reasoning under the guidance of the question. However, they provide the whole visual information equally to each graph node by concatenation of the image, question and entity embeddings. Actually, only part of the visual content are relevant to the question and a certain entity. Moreover, the fact graph here is still homogeneous since each node is represented by a fixed form of image-question-entity embedding, which limits the model’s flexibility of adaptively capturing evidence from different modalities.

In this work, we depict an image as a multi-modal heterogeneous graph, which contains multiple layers of information corresponding to different modalities. The proposed model is focused on **Multi-Layer Cross-Modal Knowledge Reasoning** and we name it as **Mucko** for short. Specifically, we encode an image by three layers of graphs, where the object appearance and their relationships are kept in the *visual layer*, the high-level abstraction for bridging the gaps between visual and factual information is provided in the *semantic layer*, and the corresponding knowledge of facts are supported in the *fact layer*. We propose a modality-aware heterogeneous graph convolutional network to adaptively collect complementary evidence in the multi-layer graphs. It can be performed by two procedures. First, the Intra-Modal Knowledge Selection procedure collects question-oriented information from each graph layer under the guidance of question; Then, the Cross-Modal Knowledge Reasoning procedure captures complementary evidence across different layers.

The main contributions of this paper are summarized as follows: (1) We comprehensively depict an image by a heterogeneous graph containing multiple layers of information based on visual, semantic and knowledge modalities. We consider these three modalities jointly and achieve significant improvement over state-of-the-art solutions. (2) We propose a modality-aware heterogeneous graph convolutional network to capture question-oriented evidence from different modalities. Especially, we leverage an attention operation in each convolution layer to select the most relevant evidence for the given question, and the convolution operation is responsible for adaptive feature aggregation. (3) We demonstrate good interpretability of our approach and provide case study in deep insights. Our model automatically tells which modality (visual, semantic or factual) and entity have more contributions to answer the question through visualization of attention weights and gate values.

一篇论文的组成——引言

Abstract

Fact-based Visual Question Answering (FVQA) requires external knowledge beyond visible content to answer questions about an image, which is challenging but indispensable to achieve general VQA. One limitation of existing FVQA solutions is that they jointly embed all kinds of information without fine-grained selection, which introduces unexpected noises for reasoning the final answer. How to capture the question-oriented and information-complementary evidence remains a key challenge to solve the problem. In this paper, we depict an image by a multi-modal heterogeneous graph, which contains multiple layers of information corresponding to the visual, semantic and factual features. On top of the multi-layer graph representations, we propose a modality-aware heterogeneous graph convolutional network to capture evidence from different layers that is most relevant to the given question. Specifically, the intra-modal graph convolution selects evidence from each modality and cross-modal graph convolution aggregates relevant information across different modalities. By stacking this process multiple times, our model performs iterative reasoning and predicts the optimal answer by analyzing all question-oriented evidence. We achieve a new state-of-the-art performance on the FVQA task and demonstrate the effectiveness and interpretability of our model with extensive experiments. The code is available at <https://github.com/astro-zihao/mucko>.

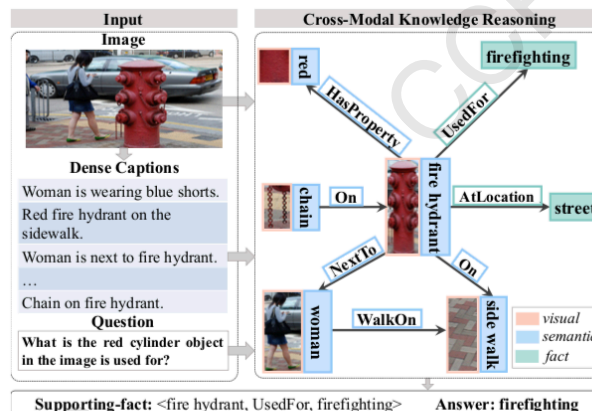


Figure 1: An illustration of our motivation. We represent an image by multi-layer graphs and cross-modal knowledge reasoning is conducted on the graphs to infer the optimal answer.

Li *et al.*, 2019b; Ben-Younes *et al.*, 2019] have achieved great success in the VQA problems that are answerable by solely referring to the visible content of the image. However, such kinds of models are incapable of answering questions which require external knowledge beyond what is in the image. Considering the question in Figure 1, the agent not only needs to visually localize ‘the red cylinder’, but also to semantically recognize it as ‘fire hydrant’ and connects the knowledge that ‘fire hydrant is used for firefighting’. Therefore, how to collect the question-oriented and information-complementary evidence from visual, semantic and knowledge perspectives is essential to achieve general VQA.

To advocate research in this direction, [Wang *et al.*, 2018] introduces the ‘Fact-based’ VQA (FVQA) task for answering questions by joint analysis of the image and the knowledge base of facts. The typical solutions for FVQA build a fact graph with fact triplets filtered by the visual concepts in the image and select one entity in the graph as the answer. Existing works [Wang *et al.*, 2017; Wang *et al.*, 2018] parse the question as keywords and retrieve the supporting-entity only by keyword matching. This kind of approaches is vulnerable when the question does not exactly mention the visual concepts (*e.g.* synonyms and homographs) or the mentioned information is not captured in the fact graph (*e.g.* the visual

attribute ‘red’ in Figure 1 may be falsely omitted). To resolve these problems, [Narasimhan *et al.*, 2018] introduces visual information into the fact graph and infers the answer by implicit graph reasoning. However, the information is equally to each graph node by concatenation of the image, question and entity embeddings. Actually, only part of the visual content are relevant to the question and a certain entity. Moreover, the fact graph here is still homogeneous since each node is represented by a fixed form of image-question-entity embedding, which limits the model’s flexibility of adaptively capturing evidence from different modalities.

In this paper, we propose a modality-aware heterogeneous graph convolutional network to adaptively collect complementary evidence in the multi-layer graphs. It can be performed by two procedures. First, the Intra-Modal Knowledge Selection procedure collects question-oriented information from each graph layer under the guidance of question; Then, the Cross-Modal Knowledge Reasoning procedure aggregates the information across different modalities. The proposed model is focused on **Multi-Layer Cross-Modal Knowledge Reasoning** and we name it as **Mucko** for short. Specifically, we encode an image by three layers of graphs, where the object appearance and their relationships are kept in the *visual layer*, the high-level abstraction for bridging the gaps between visual and factual information is provided in the *semantic layer*, and the corresponding knowledge of facts are supported in the *fact layer*. We propose a modality-aware heterogeneous graph convolutional network to adaptively collect complementary evidence in the multi-layer graphs. It can be performed by two procedures. First, the Intra-Modal Knowledge Selection procedure collects question-oriented information from each graph layer under the guidance of question; Then, the Cross-Modal Knowledge Reasoning procedure aggregates the information across different modalities.

The main contributions of this paper are as follows: (1) We comprehensively depict an image by a heterogeneous graph containing multiple layers of information based on visual, semantic and knowledge modalities. We consider these three modalities jointly and achieve significant improvement over state-of-the-art solutions. (2) We propose a modality-aware heterogeneous graph convolutional network to capture question-oriented evidence from different modalities. Especially, we leverage an attention operation in each convolution layer to select the most relevant evidence for the given question, and the convolution operation is responsible for adaptive feature aggregation. (3) We demonstrate good interpretability of our approach and provide case study in deep insights. Our model automatically tells which modality (visual, semantic or factual) and entity have more contributions to answer the question through visualization of attention weights and gate values.

从具体方法归纳问题

可解决的小问题

方法介绍：呼应挑战和问题

如何解决问题

贡献：凝练方法的普适性

提出的新问题？
解决的新视角？
实现的新框架？
实现的新方法？
达到的新SOTA？
具备的新能力？

现有方法问题：分类归纳

准确、专业评价

有主要实验结果后可以动笔！

1 Introduction

大背景

Visual question answering (VQA) [Antol *et al.*, 2015] is an attractive research direction aiming to jointly analyze multi-modal content from images and natural language. Equipped with the capacities of grounding, reasoning and translating, a VQA agent is expected to answer a question in natural language based on an image. Recent works [Cadene *et al.*, 2019;

背景：VQA需要知识

挑战：引入互补知识

一篇论文的组成——相关工作

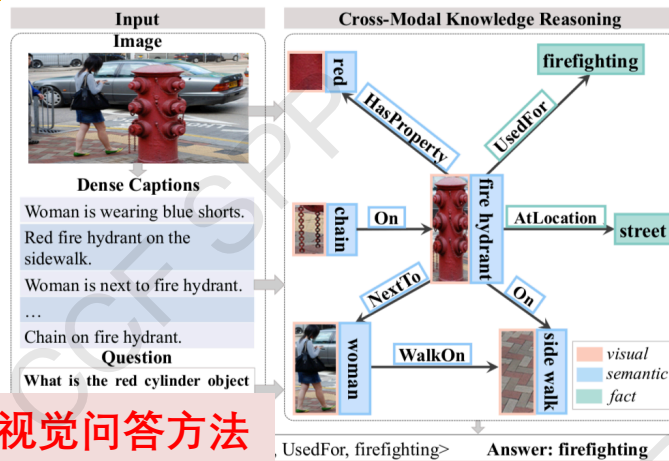
基本要求

- ☀ 包括理解本文的所有主题
- ☀ 包括问题相关的所有工作
- ☀ 从不同维度划分主题
- ☀ 同一主题方法归类
- ☀ 总结问题
- ☀ 引出本研究的区别和贡献

不要所有看过的论文！

不要罗列写上的论文！

一篇论文的组成——相关工作



视觉问答方法

Visual Question Answering. The typical solutions for VQA are based on the CNN-RNN architecture [Malinowski *et al.*, 2015] and leverage global visual features to represent image, which may introduce noisy information. Various attention mechanisms [Yang *et al.*, 2016; Lu *et al.*, 2016; Anderson *et al.*, 2018] have been exploited to highlight visual objects that are relevant to the question. However, they treat objects independently and ignore their informative relationships. [Battaglia *et al.*, 2018] demonstrates that human's ability of combinatorial generalization highly depends on the mechanisms for reasoning over relationships. Consistent with such proposal, there is an emerging trend to represent the image by graph structure to depict objects and relationships in VQA and other vision-language tasks [Hu *et al.*, 2019b; Wang *et al.*, 2019a; Li *et al.*, 2019b]. As an extension, [Jiang *et al.*, 2020] exploits natural language to enrich the graph-based visual representations. However, it solely captures the semantics in natural language by LSTM, which lacking of fine-grained correlations with the visual information. To go one step further, we depict an image by multiple layers of graphs from visual, semantic and factual perspectives to collect fine-grained evidence from different modalities.

基于知识的视觉问答方法

Fact-based Visual Question Answering. Human can easily combine visual observation with external knowledge for answering questions, which remains challenging for algorithms. [Wang *et al.*, 2018] introduces a fact-based VQA task, which provides a knowledge base of facts and associates each question with a supporting-fact. Recent works based on FVQA generally select one entity from fact graph as the answer and falls into two categories: query-mapping based methods and learning based methods. [Wang *et al.*, 2017] reduces the question to one of the available query templates and this limits the types of questions that can be asked. [Wang *et al.*, 2018] automatically classifies and maps the question to a query which does not suffer the above constraint. Among both methods, however, visual information are used to extract facts but not introduced during the reasoning process. [Narasimhan *et al.*, 2018] applies GCN on the fact graph where each node is represented by the fixed form of image-question-entity embedding. However, the visual information is wholly provided which may introduce redundant information for prediction. In this paper, we depict an image by multi-layer graphs and perform cross-modal heterogeneous graph reasoning on them to capture complementary evidence from different layers that most relevant to the question.

Heterogeneous Graph Neural Networks.

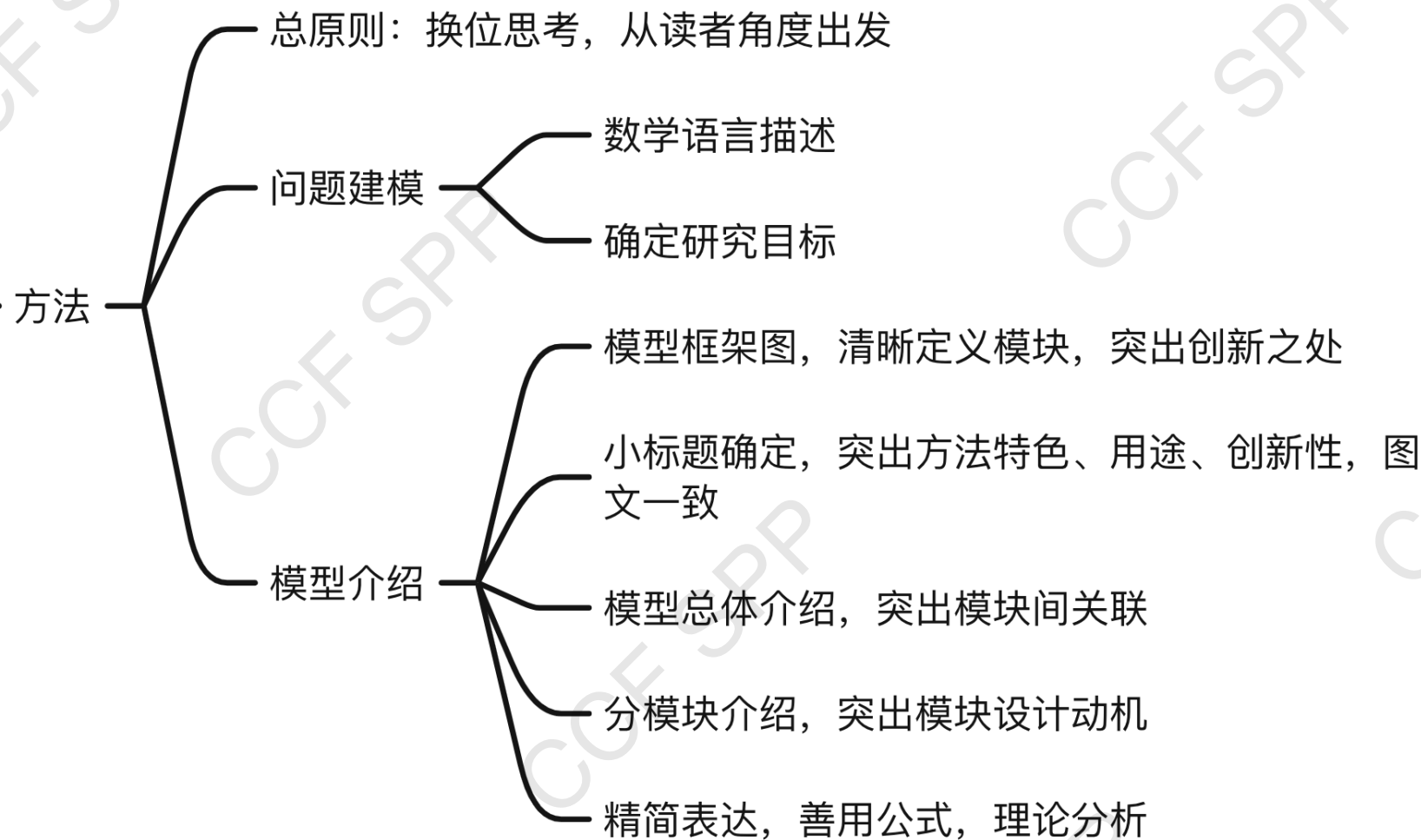
异构图神经网络方法

Graph neural network in the last few years has been combined with homogeneous graphs, heterogeneous graphs are more common in the real world. [Schlichtkrull *et al.*, 2018] generalizes graph convolutional network (GCN) to handle different relationships between entities in a knowledge base, where each edge with distinct relationships is encoded independently. [Wang *et al.*, 2019b; Hu *et al.*, 2019a] propose heterogeneous graph attention networks with dual-level attention mechanism. All of these methods model different types of nodes and edges on a unified graph. In contrast, the heterogeneous graph in this work contains multiple layers of subgraphs and each layer consists of nodes and edges coming from different modalities. For this specific constrain, we propose the intra-modal and cross-modal graph convolutions for reasoning over such multi-modal heterogeneous graphs.

方法写完后可动笔!

一篇论文的组成——方法

基本要求（最容易部分，可以先写）



一篇论文的组成——方法

CCF—A

- 问题-方法-实验，相互呼应

- 动机：有理有据，足够具体
- 方法：针对问题设计，每一步设计目标明确
 - 根据重点，重新组织方法介绍思路
 - 标题和图突出创新性和重点，相互呼应
 - 每一步方法设计都有理可依
- 实验：针对方法逐一证明，针对动机逐一分析

CCF—C

- 问题-方法-实验，各为其说

- 动机：大家都在研究，所以我研究
- 方法：step1->step2->step3
- 实验：达到了SOTA，缺乏分析

一篇论文的组成——方法

IJCAI 2020

Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering

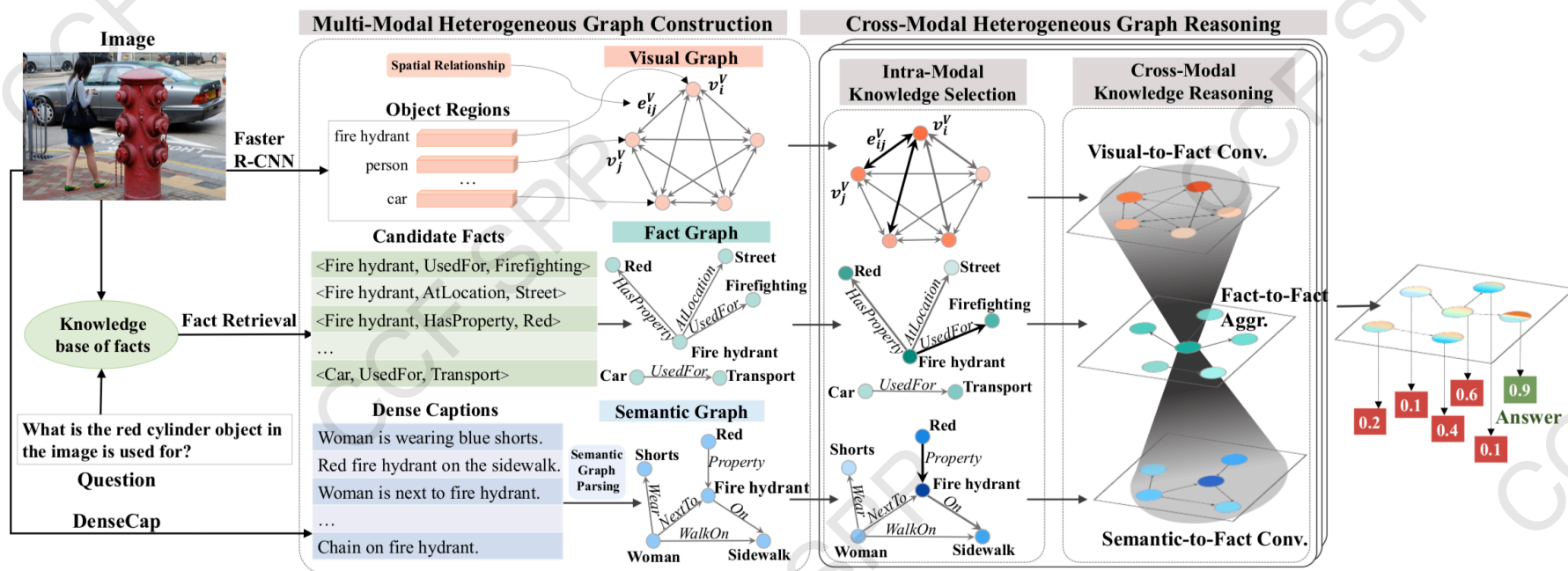
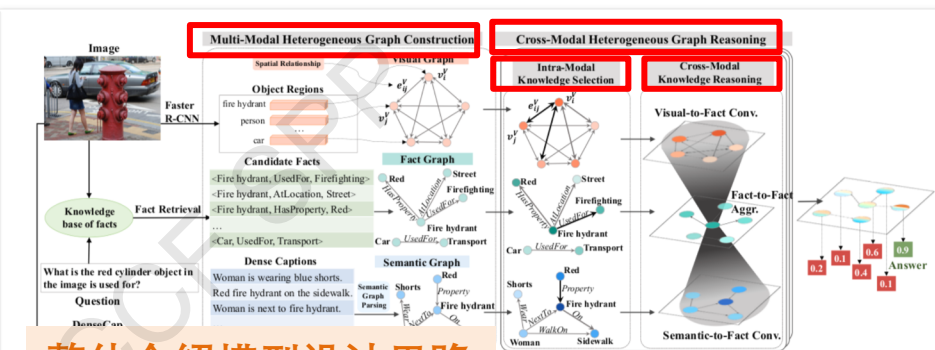


Figure 2: An overview of our model. The model contains two modules: Multi-modal Heterogeneous Graph Construction aims to depict an image by multiple layers of graphs and Cross-modal Heterogeneous Graph Reasoning supports intra-modal and cross-modal evidence selection.

一篇论文的组成——方法



整体介绍模型设计思路

where e_1 is a visual concept in the image, e_2 is an attribute or phrase and r represents the relationship between e_1 and e_2 . The key is to choose a correct entity, i.e. either e_1 or e_2 , from the supporting fact as the predicted answer. We first introduce a novel scheme of depicting an image by three layers of graphs, including the visual graph, semantic graph and fact graph respectively, imitating the understanding of various properties of an object and the relationships. Then we perform cross-modal heterogeneous graph reasoning that consists of two parts: *Intra-Modal Knowledge Selection* aims to choose question-oriented knowledge from each layer of graphs by intra-modal graph convolutions, and *Cross-Modal Knowledge Reasoning* adaptively selects complementary evidence across three layers of graphs by cross-modal graph convolutions. By stacking the above two processes multiple times, our model performs iterative reasoning across all the modalities and results in the optimal answer by jointly analyzing all the entities. Figure 2 gives detailed illustration of our model.

3.1 Multi-Modal Graph Construction

Visual Graph Construction. Since most of the questions in FVQA grounded in the visual objects and their relationships, we construct a fully-connected visual graph to present such evidence at appearance level. Given an image, we use Faster-RCNN [Ren et al., 2017] to identify a set of objects $\mathcal{O} = \{o_i\}_{i=1}^K$ ($K = 36$), where each object o_i is associated with a visual feature vector $v_i \in \mathbb{R}^{d_v}$ ($d_v = 2048$), a spatial feature vector $b_i \in \mathbb{R}^{d_b}$ ($d_b = 4$) and a corresponding label. Specifically, $b_i = [x_i, y_i, w_i, h_i]$ where (x_i, y_i) is the bounding box of o_i and w_i, h_i are the width and height of o_i .

每一个过程首先介绍背后的动机和目标

node set and each node v_i^V corresponds to a detected object o_i . The feature of node v_i^V is represented by v_i^V . Each edge $e_{ij}^V \in \mathcal{E}^V$ denotes the relative spatial relationships between two objects. We encode the edge feature by a 5-dimensional vector, i.e. $r_{ij}^V = [\frac{x_j - x_i}{w_i}, \frac{y_j - y_i}{h_i}, \frac{w_j}{w_i}, \frac{h_j}{h_i}, \frac{w_j h_j}{w_i h_i}]$.

Multi-modal Heterogeneous Graph Construction aims to depict an image by three layers of graphs, including the visual graph, semantic graph and fact graph respectively, imitating the understanding of various properties of an object and the relationships.

Semantic Graph Construction. In addition to visual information, high-level abstraction of the objects and relationships by natural language provides essential semantic information. Such abstraction is indispensable to associate the visual objects in the image with the concepts mentioned in both questions and facts. In our work, we leverage dense captions [Johnson et al., 2016] to extract a set of local-level semantics in an image, ranging from the properties of a single object (color, shape, emotion, etc.) to the relationships between objects (action, spatial positions, comparison, etc.). We depict an image by D dense captions, denoted as $Z = \{z_i\}_{i=1}^D$, where z_i is a natural language description about a local region in the image. Instead of using monolithic embeddings to represent the captions, we exploit to model them by a graph-based semantic representation, denoted as $\mathcal{G}^S = (\mathcal{V}^S, \mathcal{E}^S)$, which is constructed by a semantic graph parsing model [Anderson et al., 2016]. The node $v_i^S \in \mathcal{V}^S$ represents the name or attribute of an object extracted from the captions while the edge $e_{ij}^S \in \mathcal{E}^S$ represents the relationship between v_i^S and v_j^S .

- 标题突出创新点和过程
- 表达逻辑一致
- 图文一致

base of facts following a scored based approach proposed in [Narasimhan et al., 2018]. We compute the cosine similarity of the embeddings of every word in the fact with the words in the question.

cepts detected in the question. We assign a similarity score to each concept, denoted as f_{100} . A relation type classifier is trained additionally to further filter the retrieved facts. Specifically, we feed the last hidden state of LSTM in an MLP layer to predict the relation type $\hat{r}_i$ of a question. We retain the facts among f_{100} only if their relationships agree with $\hat{r}_i$, i.e. $f_{rel} = f$ if $f_{rel} : r(f) \in \{\hat{r}_i\}$ ($\{\hat{r}_i\}$ contains top-3 predicted relationships in experiments). Then a fact graph $\mathcal{G}^F = (\mathcal{V}^F, \mathcal{E}^F)$ is built upon f_{rel} as the candidate facts can be naturally organized as graphical structure. Each node $v_i^F \in \mathcal{V}^F$ denotes an entity in f_{rel} and is represented by GloVe embedding of the entity, denoted as v_i^F . Each edge $e_{ij}^F \in \mathcal{E}^F$ denotes the relationship between v_i^F and v_j^F and is represented by GloVe embedding r_{ij} . The topological structure among facts can be effectively exploited by jointly considering all the entities in the fact graph.

3.2 Intra-Modal Knowledge Selection

Since each layer of graphs contains modality-specific knowledge relevant to the question, we first select valuable evidence independently from the visual graph, semantic graph and fact graph by **Visual-to-Visual Convolution**, **Semantic-to-Semantic Convolution** and **Fact-to-Fact Convolution** respectively. These three convolutions share the common operations but differ in their node and edge representations corresponding to the graph layers. Thus we omit the superscript of node representation v and edge representation r in the rest of this section. We first perform attention operations to highlight the nodes and edges that are most relevant to the question q and consequently update node representations via intra-modal graph convolution. This process mainly consists of the following three steps:

Question-guided Node Attention. We first evaluate the relevance of each node corresponding to the question by attention mechanism. The attention weight for v_i is computed as:

$$\alpha_i = \text{softmax}(w_a^T \tanh(W_1 v_i + W_2 q)) \quad (1)$$

where W_1, W_2 and w_a (as well as $W_3, \dots, W_5, w_b, w_c$, mentioned below) are learned parameters. q is question embedding encoded by LSTM.

Question-guided Edge Attention. Under the guidance of question, we then evaluate the importance of edge e_{ji} constrained by the neighbor node v_j regarding to v_i as:

$$\beta_{ji} = \text{softmax}(w_b^T \tanh(W_3 v_j' + W_4 q')) \quad (2)$$

where $v_j' = W_5[v_j, r_{ji}]$, $q' = W_6[v_i, q]$ and $[\cdot, \cdot]$ denotes concatenation operation.

Intra-Modal Graph Convolution. Given the node and edge attention weights learned in Eq. 1 and Eq. 2, the node representations of each layer of graphs are updated following

具体过程分小标题

$$m_i = \sum_{j \in \mathcal{N}_i} \beta_{ji} v_j' \quad (3)$$

$$\tilde{v}_i = \text{ReLU}(W_7[m_i, \alpha_i v_i]) \quad (4)$$

where $\mathcal{N}_i$ is the neighborhood set of node v_i .

We conduct the above intra-modal knowledge selection on $\mathcal{G}^V$, $\mathcal{G}^S$ and $\mathcal{G}^F$ independently and obtain the updated node representations, denoted as $\{\tilde{v}_i^V\}_{i=1}^{N^V}$, $\{\tilde{v}_i^S\}_{i=1}^{N^S}$ and $\{\tilde{v}_i^F\}_{i=1}^{N^F}$ accordingly.

3.3 Cross-Modal Knowledge Reasoning

To answer the question correctly, we fully consider the complementary evidence from visual, semantic and factual information. Since the answer comes from one entity in the fact graph, we gather complementary information from visual graph and semantic graph to fact graph by cross-modal convolutions, including **visual-to-fact convolution** and **semantic-to-fact convolution**. Finally, a **fact-to-fact aggregation** is performed on the fact graph to reason over all the entities and form a global decision.

Visual-to-Fact Convolution. For the entity v_i^F in fact graph, the attention value of each node v_j^V in the visual graph w.r.t. v_i^F is calculated under the guidance of question:

$$\gamma_{ji}^{V \rightarrow F} = \text{softmax}(w_c \tanh(W_8 v_j^V + W_9 [v_i^F, q])) \quad (5)$$

The complementary information $m_i^{V \rightarrow F}$ from visual graph for v_i^F is computed as:

$$m_i^{V \rightarrow F} = \sum_{j \in \mathcal{N}^V} \gamma_{ji}^{V \rightarrow F} v_j^V \quad (6)$$

Semantic-to-Fact Convolution. The complementary information $m_i^{S \rightarrow F}$ from the semantic graph is computed in the same way as in Eq. 5 and Eq. 6.

Then we fuse the complementary knowledge for v_i^F from three layers of graphs via a gate operation:

$$\text{gate}_i = \sigma(W_{10}[m_i^{V \rightarrow F}, m_i^{S \rightarrow F}, \tilde{v}_i^F]) \quad (7)$$

$$\tilde{v}_i^F = W_{11}[\text{gate}_i \circ [m_i^{V \rightarrow F}, m_i^{S \rightarrow F}, \tilde{v}_i^F]] \quad (8)$$

where σ is sigmoid function and " $\circ$ " denotes element-wise product.

Fact-to-Fact Aggregation. Given a set of candidate entities in the fact graph $\mathcal{G}^F$, we aim to globally compare all the entities and select an optimal one as the answer. Now the representation of each entity in the fact graph gathers question-oriented information from three modalities. To jointly evaluate the possibility of each entity, we perform the attention-based graph convolutional network similar to Fact-to-Fact Convolution introduced in Section 3.2 to aggregate information in the fact graph and obtain the transformed entity representations.

We iteratively perform intra-modal knowledge selection and cross-modal knowledge reasoning in multiple steps to obtain the final entity representations. After T steps, each entity representation $\tilde{v}_i^{F(T)}$ captures the structural information within T -hop neighborhood across three layers.

3.4 Learning

The concatenation of entity representation $\tilde{v}_i^{F(T)}$ and question embedding q is passed to a binary classifier to predict its probability as the answer, i.e. $\hat{y}_i = p_\theta([\tilde{v}_i^{F(T)}, q])$. We apply the binary cross-entropy loss in the training process:

$$l_n = - \sum_{i \in \mathcal{N}^F} [a \cdot y_i \ln \hat{y}_i + b \cdot (1 - y_i) \ln (1 - \hat{y}_i)] \quad (9)$$

where y_i is the ground truth label for v_i^F and a, b represent loss function weights for positive and negative samples respectively. The entity with the largest probability is selected as the final answer.

注意：
动机？
其他方式？

一篇论文的组成——方法(实例分析)

Seeing Is Knowing! Fact-based Visual Question Answering

Using Knowledge Graph Embeddings

arxiv

Our approach

The proposed architecture for FVQA is shown in Fig. 2. As shown, a given image I and query q are combined via coAttention to form two entity query vectors, $f_{KVC}(q, I)$ and $f_{KB}(q, I)$. The KG is then queried for the answer to the question, according to

$$\hat{y}(q|I) = \begin{cases} \operatorname{argmax}_{e \in \mathcal{E}} f_{KVC}(q, I)^T e & g_{KVC}(q) = 1 \\ \operatorname{argmax}_{e \in \mathcal{E}} f_{KB}(q, I)^T e & g_{KVC}(q) = 0 \end{cases} \quad (2)$$

where the gating function $g_{KVC}(q) \in \{0, 1\}$ is equal to 1 if the text of the question indicates that the answer is visible in the image, equal to 0 otherwise. The rest of this section addresses representations of the entities, image, and query, the information fusion functions, the gating function, and the loss function.

缺乏动机!

一篇论文的组成——方法(实例分析)

Transformer Reasoning Network for Image-Text Matching and Retrieval 2020 ICPR

Region and Word Features

I and C descriptions come from state-of-the-art visual and textual pre-trained networks, Faster-RCNN with Bottom-Up attention, and BERT respectively.

Faster-RCNN [31] is a state-of-the-art object detector. It has been used in many downstream tasks requiring salient object regions extracted from images. The authors in [32] introduced bottom-up visual features by training Faster-RCNN with a Resnet-101 backbone on the Visual Genome dataset [33]. Using these features, they were able to reach remarkable results on the two downstream tasks of image captioning and visual question answering. Therefore, in our work we employ the bottom-up features extracted from every image as image description $I = \{r_0, r_1, \dots, r_n\}$.

背景知识介绍太多!

Structured Multi-modal Feature Embedding and Alignment for Image-Sentence Retrieval 2020 ACM MM

3.1.1 Visual representations. To better represent the salient entities and attributes in images, we take advantage of bottom-up-attention network [1] to embed the extracted sub-regions in an image. Specifically, given an image I , we extract a set of image fragment-level sub-region features $V = \{v_1, \dots, v_K\}$, $v_j \in \mathbb{R}^{2048}$, where K is the number of selected sub-regions, from the average pooling layer in Faster-RCNN [25].

Good Case!

一篇论文的组成——实验

基本要求

☀️ **一致**：支撑理论/方法、动机

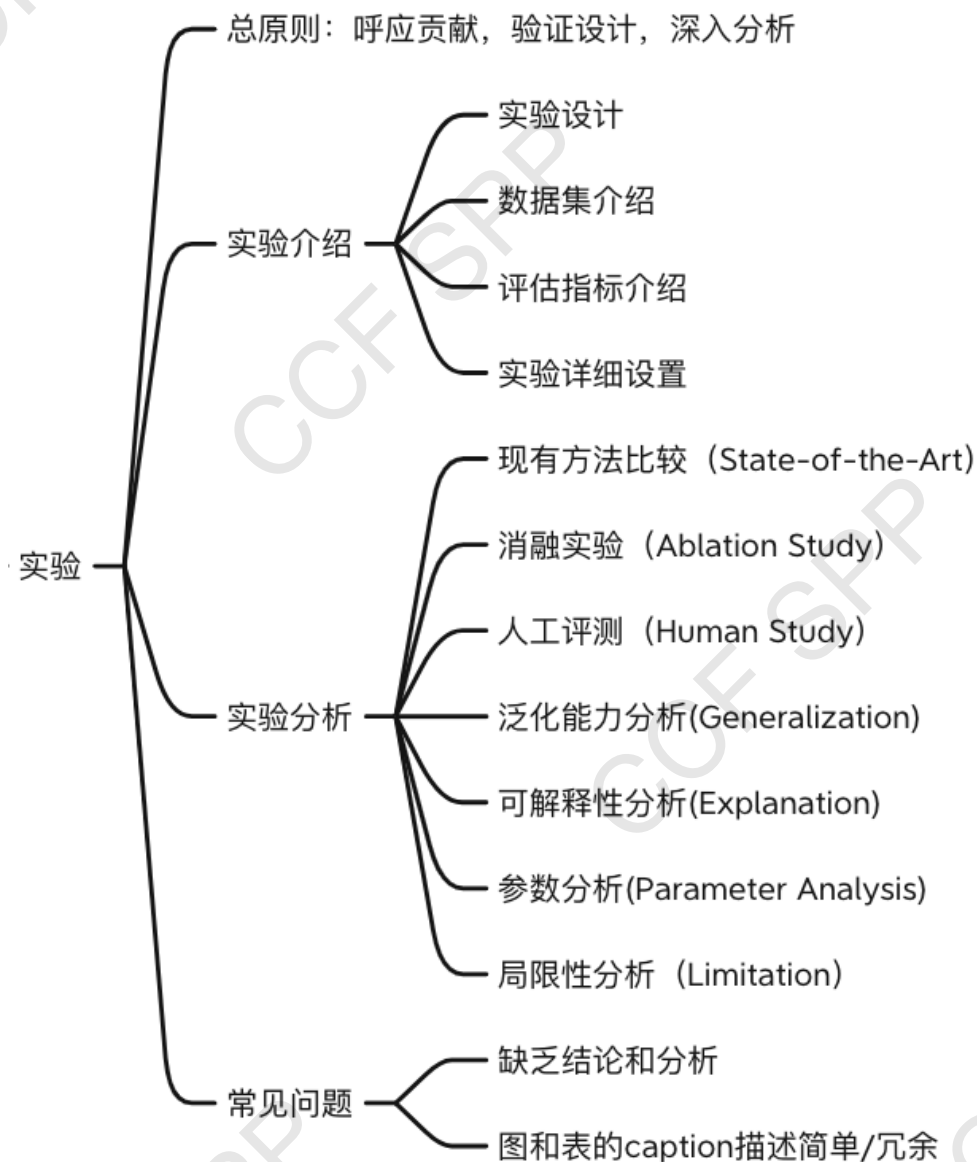
☀️ **核心**：提供重要实验结果

☀️ **诚实**：不只展示最佳个例

☀️ **分析**：给出结果的合理解释

☀️ **局限**：给出方法的能力边界

学习优秀论文的实验写法！



一篇论文的组成——实验

CCF—A

- 问题-方法-实验，相互呼应

- 问题：有理有据，足够具体
- 方法：针对问题设计，每一步设计目标明确
- 实验：针对方法逐一证明，针对动机逐一分析

- 结论先行
- 事实支撑
- 层次分明

CCF—C

- 问题-方法-实验，各为其说

- 问题：大家都在研究，所以我研究
- 方法：step1->step2->step3
- 实验：达到了SOTA，缺乏分析

一篇论文的组成——实验

Table 1

State-of-the-art comparison on FVQA dataset.

Method	Overall accuracy	
	top-1	top-3
LSTM-Q + Image + Pre-VQA [4]	24.98	40.40
Hie-Q + Image + Pre-VQA [4]	43.14	59.44
FVQA (top-3-QQmapping) [4]	56.91	64.65
FVQA (Ensemble) [4]	58.76	-
Straight to the Facts (STTF) [19]	62.20	75.60
Reading Comprehension [43]	62.96	70.08
Out of the Box (OB) [6]	69.35	80.25
Human [4]	77.99	-
GRUC (ours)	69.03	88.12

对比方法分类介绍

Table 3: State-of-the-art comparison on OK-VQA dataset.

We also report the quantitative performance on the challenging OK-VQA dataset in Table 3. We compare our model with three kinds of existing models, including current state-of-the-art VQA models, knowledge-based VQA models and ensemble models. The VQA models contain Q-Only [21], MLP [21], BAN [44], MUTAN [44]. The knowledge-based VQA models [21] consist of ArticleNet (AN), BAN + AN and MUTAN + AN. The ensemble models [21], i.e. BAN/AN oracle and MUTAN/AN oracle, simply take the raw ArticleNet and VQA model predictions, taking the best answer (comparing to ground truth) from either. We report the overall performance (top-1 and top-3 accuracy) as well as breakdowns for each of the knowledge categories (top-1 accuracy). We have the following two observations from the results:

现有方法分类对比结果分析

As all the compared state-of-the-art models (BAN and MUTAN) specifically designed for VQA tasks, they get inferior results compared with ours. This indicates that general VQA task like OK-VQA cannot be simply solved by a well-designed model, but requires the ability to incorporate external knowledge in an effective way. Moreover, our model outperforms knowledge-based VQA models including both single models (BAN+AN and MUTAN + AN) and ensemble models (BAN/AN oracle and MUTAN/AN oracle), which further proves the advantages of our knowledge incorporating mechanism based on both multimodal knowledge graphs and memory-enhanced recurrent reasoning network.

Table 2

State-of-the-art comparison on Visual7W + KB dataset.

Method	Overall accuracy	
	top-1	top-3
KDMN-NoKnowledge [20]	45.1	-
KDMN-NoMemory [20]	51.9	-
KDMN [20]	57.9	-
KDMN-Ensemble [20]	60.9	-
Out of the Box (OB) [6]	57.32	71.61
GRUC (ours)	69.03	88.12

Table 3

State-of-the-art comparison on OK-VQA dataset. We show the results for the full OK-VQA dataset and for each knowledge category (top-1 accuracy): Vehicles and Transportation (VT); Brands, Companies and Products (BCP); Objects, Material and Clothing (OMC); Sports and Recreation (SR); Cooking and Food (CF); Geography, History, Language and Culture (GHLC); People and Everyday Life (PEL); Plants and Animals (PA); Science and Technology (ST); Weather and Climate (WC); and Other.

Method	Overall accuracy		VT	BCP	OMC	SR	CF	GHLC	PEL	PA	ST	WC	Other
	top-1	top-3											
Q-Only [21]	14.93	-	14.64	14.19	11.78	15.94	16.92	11.91	14.02	14.28	19.76	25.74	13.51
MLP [21]	20.67	-	21.33	15.81	17.76	24.69	21.81	11.91	17.15	21.33	19.29	29.92	19.81
BAN [44]	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 [45]	26.41	-	25.36	18.95	24.02	33.23	27.73	17.59	20.09	30.44	20.48	39.38	22.46
ArticleNet (AN) [21]	5.28	-	4.48	0.93	5.09	5.11	5.69	6.24	3.13	6.95	5.00	9.92	5.33
BAN + AN [21]	25.61	-	24.45	19.88	21.59	30.79	29.12	20.57	21.54	26.42	27.14	38.29	22.16
MUTAN + AN [21]	27.84	-	25.56	23.95	26.87	33.44	29.94	20.71	25.05	29.70	24.76	39.84	23.62
BAN/AN oracle [21]	27.59	-	26.35	18.26	24.35	33.12	30.46	28.51	21.54	28.79	24.52	41.4	25.07
MUTAN/AN oracle [21]	28.47	-	27.28	19.53	25.28	35.13	30.53	21.56	21.68	32.16	24.76	41.4	24.85
GRUC (ours)	29.87	32.65	29.84	25.23	30.61	30.92	28.01	26.24	29.21	31.27	27.85	38.01	26.21

异常结果分析

One of our model on OK-VQA is not that the performance on FVQA and Visual7W-KB, we believe that this phenomenon is mostly due to the following two reasons: (1) Questions in the OK-VQA dataset are more diverse and complex, which is more challenging for machines to understand accurately. The questions in FVQA and Visual7W-KB are generated when given the images and supporting facts upon the pre-defined templates or relations. This mechanism constrains the answers in a specific knowledge base and guides the model to operate in a reverse way of the question generation process to predict the correct answers with high probability. On the contrary, questions in OK-VQA are totally free-form ones that generated by MTurk workers and thus containing more unique questions and words with less bias compared with other datasets. This increases the difficulty to understand the questions accurately. (2) OK-VQA requires a wide range of knowledge beyond a specific knowledge base. Looking at the category breakdowns in Table 3, baseline models achieve relatively high performance for SR, CF, GHLC, PA and WC categories while our model performs better for the remaining categories. Since the baseline models refer to the Wikipedia while our model refers to ConceptNet, the performance in the category breakdowns perhaps suggests that each knowledge

一篇论文的组成——实验

不同维度对比分析

Ablation

TABLE IV

ABLATION STUDY ON VALIDATION SET OF VISDIAL V1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
ObjRep	63.84	49.83	81.27	90.29	4.07	55.48
RelRep	63.63	49.25	81.01	90.34	4.07	55.12
VisNoRel	63.97	49.87	81.74	90.60	4.00	56.73
VisMod	64.11	50.04	81.78	90.52	3.99	56.67
GlCap	60.02	45.34	77.66	87.27	4.78	50.04
LoCap	60.95	46.43	78.45	88.17	4.62	51.72
SemMod	61.07	46.69	78.56	88.09	4.59	51.10
w/o ELMo	63.67	49.89	80.44	89.84	4.14	56.41
	50.74	82.10	91.00	3.91	57.30	
	50.79	82.41	91.10	3.90	58.24	

Variants介绍

B. Ablation Study

We also conduct an ablation study to further exploit the influence of the essential components of DualVD. To be mentioned, we use DualVD-LF as the full model and apply the same discriminative decoder for all the following variations:

Object Representation (ObjRep): this model uses the averaged object features to represent the image. Question-driven attention is applied to enhance the object representations.

Relation Representation (RelRep): this model applies averaged relation-aware object representations as the image representation without fusing with original object features.

Vision Module without Relationships (VisNoRel): this model contains the full Vision Module, differing in that the relation embeddings are replaced by unlabeled edges.

Visual Module (VisMod): this is our full visual module, which fuses objects and relation features.

Global Caption (GlCap): this model uses LSTM to encode the global caption to represent the image.

Local Caption (LoCap): this model uses LSTM to encode the local captions to represent the image.

Semantic Module (SemMod): this is our full semantic module, which fuses global and local features.

w/o ELMo: this is our full model based on late fusion encoder, differing in that the word embedding GloVe+ELMo is replaced by GloVe.

DualVD-LF (full model): this is our full model, which incorporates both the visual module and semantic module.

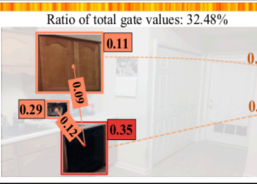
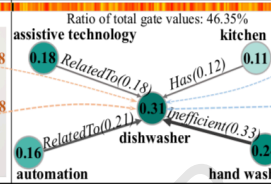
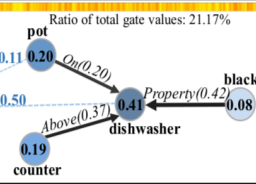
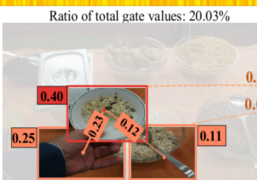
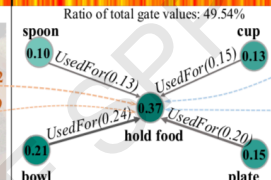
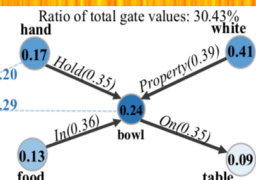
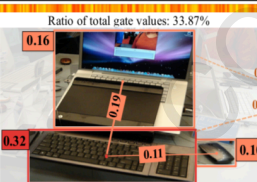
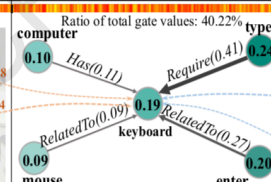
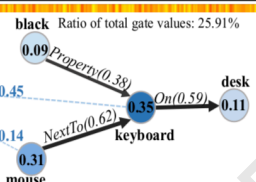
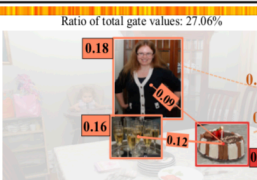
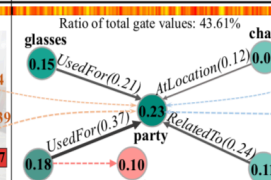
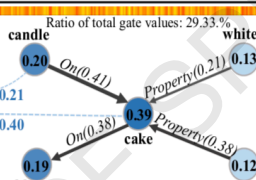
Table IV shows the ablation results on VisDial v1.0 validation set. Models in the first block are designed to evaluate the influence of key components in the visual module. The limitation for **ObjRep** is that it only mines the pivotal features from isolated objects and ignores the relational information, which achieves worse performance at all metrics compared to VisMod. **RelRep** considers the relationships by introducing relation embedding for aggregating the object features. However, empirical study indicates that enhancing object relationships while weakening object appearance is still not sufficient to represent the visual semantics for better performance. **VisNoRel** takes a step further by adaptively fusing the information from both object appearance and full-connected neighbors, aggregating all the neighborhood features directly without relation semantics. This strategy achieves slight improvement compared with ObjRep. **VisMod** moves a step further by adaptively fusing the information from both object appearance and full-connected neighbors, aggregating all the neighborhood features with relational information, which achieves the best performance compared to above two models.

Orthogonal to visual part, models in the second block are conducted to test the influence of key components in the semantic part. The overall performance of either **GlCap** or **LoCap** decreases by 1% and 0.15% respectively, compared to their integrated version **SemMod**, which adaptively selects and fuses the task-specific descriptive clues from both global-level and local-level captions.

We compare the performance of VisMod and SemMod with DualVD-LF. By adaptively select information from the visual and the semantic module, **DualVD-LF** results in a great boost compared to SemMod and a relatively slight boost compared to VisMod. This unbalanced boost indicates that visual module provides comparatively richer clues than semantic module. Combining the two modules together gains an extra boost, because of the complementary information derived from different modalities. By paying more attention on

一篇论文的组成——实验

Interpretation

Case	Visual Graph	Fact Graph	Semantic Graph
 Question: Which device in the image can free peoples hand? Pred./Gt Answer: dishwasher (✓)	 Ratio of total gate values: 32.48%	 Ratio of total gate values: 46.35%	 Ratio of total gate values: 21.17%
 Question: What is the white round thing held by hand in the image used for? Pred./Gt Answer: hold food (✓)	 Ratio of total gate values: 20.03%	 Ratio of total gate values: 49.54%	 Ratio of total gate values: 30.43%
 Question: which part of the machine in the image can be used for typing? Pred./GT Answer: keyboard (✓) OB Answer: laptop (X)	 Ratio of total gate values: 33.87%	 Ratio of total gate values: 40.22%	 Ratio of total gate values: 25.91%
 Question: Where can you find the right object on the table shown in the image? GT Answer: wedding Pred. Answer: party (X)	 Ratio of total gate values: 27.06%	 Ratio of total gate values: 43.61%	 Ratio of total gate values: 29.33%

Our model is interpretable by visualizing the attention weights and gate values in the reasoning process. From case study in Figure 3, we conclude with the following three insights: (1) **Mucko is capable to reveal the knowledge selection mode.** The first two examples indicate that Mucko captures the most relevant visual, semantic and factual evidence as well as complementary information across three modalities. In most cases, factual knowledge provides predominant clues compared with other modalities according to gate values because FVQA relies on external knowledge to a great extent. Furthermore, more evidence comes from the semantic modality when the question involves complex relationships. For instance, the second question involving the relationship between ‘hand’ and ‘while round thing’ needs more semantic clues. (2) **Mucko has advantages over the state-of-the-art model.** The third example compares the predicted answer of OB with Mucko. Mucko collects relevant visual and semantic evidence to make each entity discriminative enough for predicting the correct answer while OB failing to distinguish representations of ‘laptop’ and ‘keyboard’ without feature selection. (3) **Mucko fails when multiple answers are reasonable for the same question.** Since both ‘wedding’ and

一篇论文的组成——实验

Parameter Analysis

#Retrieved facts	@50	@100	@150	@200
Rel@1 (top-1 accuracy)	55.56	70.62	65.94	59.77
Rel@1 (top-3 accuracy)	64.09	81.95	73.41	66.32
Rel@3 (top-1 accuracy)	58.93	73.06	70.12	65.93
Rel@3 (top-3 accuracy)	68.50	85.94	81.43	74.87

Table 3: Overall accuracy with different number of retrieved candidate facts and different number of relation types.

#Steps	1	2	3
top-1 accuracy	62.05	73.06	70.43
top-3 accuracy	71.87	85.94	81.32

Table 4: Overall accuracy with different number of reasoning steps.

其他需要验证的实验

- 泛化能力（更多不同任务）
- 局部模块效果验证
- 存在的局限性
- ...

一篇论文的组成——实验(实例分析)

Transformer Reasoning Network for Image-Text Matching and Retrieval 2020 ICPR

评价指标描述

For this reason, inspired by the evaluation method presented in [2], we employed a common metric often used in information retrieval applications, the Normalized Discounted Cumulative Gain (NDCG). The NDCG is able to evaluate the quality of the ranking produced by a certain query by looking at the first p position of the ranked elements list. The premise of NDCG is that highly relevant items appearing lower in a search result list should be penalized as the graded relevance value is reduced proportionally to the position of the result.

The non-normalized DCG until position p is defined as follows:

$$DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)}, \quad (5)$$

where rel_i is a positive number encoding the affinity that the i -th element of the retrieved list has with the query element. The DCG is agnostic upon how the relevance is computed. The $NDCG_p$ is computed by normalizing the DCG_p with respect to the Ideal Discounted Cumulative Gain (IDCG), that is defined

as the DCG of the list obtained by sorting all its elements by descending relevance:

$$NDCG_p = \frac{DCG_p}{IDCG_p}. \quad (6)$$

$IDCG_p$ is the best possible ranking. Thanks to this normalization, $NDCG_p$ acquires values in the range $[0, 1]$.

SOTA实验

TABLE I
IMAGE RETRIEVAL RESULTS ON THE MS-COCO DATASET.

Model	Recall@K			NDCG	
	K=1	K=5	K=10	ROUGE-L	SPICE
<i>1K Test Set</i>					
VSE0 [1]	43.7	79.4	89.7	0.702	0.616
VSE++ [1]	52.0	84.3	92.0	0.712	0.617
VSRN [7]	60.8	88.4	94.1	0.723	0.620
TERN (Ours)	51.9	85.6	93.6	0.725	0.653
<i>5K Test Set</i>					
VSE0 [1]	22.0	50.2	64.2	0.633	0.549
VSE++ [1]	30.3	59.4	72.4	0.656	0.577
VSRN [7]	37.9	68.5	79.4	0.676	0.596
TERN (Ours)	28.7	59.7	72.7	0.665	0.600

可视化实验



Query: A large jetliner sitting on top of an airport runway.



Query: An eating area with a table and a few chairs.

Fig. 4. Example of image retrieval results for a couple of query captions. The red marked images represent the MS-COCO ground truths, and they are not necessarily the best results in these scenarios. In fact, in the very first positions, we find non-matching yet relevant images. These are common examples where NDCG really succeed over the Recall@K metric.

一篇论文的组成——结论

基本要求

- ☀ 和摘要的区别：经过验证后的结论
- ☀ 论文主要发现总结
- ☀ 介绍小问题的结论
- ☀ 避免过度夸大
- ☀ 内容尽量简洁
- ☀ 介绍未来工作

一篇论文的组成——结论

Abstract

Fact-based Visual Question Answering (FVQA) requires external knowledge beyond visible content to answer questions about an image, which is challenging but indispensable to achieve general VQA. One limitation of existing FVQA solutions is that they jointly embed all kinds of information without fine-grained selection, which introduces unexpected noises for reasoning the final answer. How to capture the question-oriented and information-complementary evidence remains a key challenge to solve the problem. In this paper, we depict an image by a multi-modal heterogeneous graph, which contains multiple layers of information corresponding to the visual, semantic and factual features. On top of the multi-layer graph representations, we propose a modality-aware heterogeneous graph convolutional network to capture evidence from different layers that is most relevant to the given question. Specifically, the intra-modal graph convolution selects evidence from each modality and cross-modal graph convolution aggregates relevant information across different modalities. By stacking this process multiple times, our model performs iterative reasoning and predicts the optimal answer by analyzing all question-oriented evidence. We achieve a new state-of-the-art performance on the FVQA task and demonstrate the effectiveness and interpretability of our model with extensive experiments. The code is available at <https://github.com/lestra-zibao/mucko>.

说明动机
陈述方法

Conclusion

In this paper, we propose Mucko for visual question answering requiring external knowledge, which focuses on multi-layer cross-modal knowledge reasoning. We novelly depict an image by a heterogeneous graph with multiple layers of information corresponding to visual, semantic and factual modalities. We propose a modality-aware heterogeneous graph convolutional network to select and gather intra-modal and cross-modal evidence iteratively. Our model outperforms the state-of-the-art approaches remarkably and obtains interpretable results on the benchmark dataset.

总结工作
体现效果

ation. However, our model has inferior performance when open-domain knowledge is required. How to adaptively incorporate diverse knowledge bases that covering commonsense, Wikipedia knowledge and even professional knowledge for KVQA tasks will be our future work.

说明局限
指明方向

一篇论文的组成——致谢

基本要求

- ☀ 帮助这篇论文的人员、机构、项目资助
- ☀ 审稿人
- ☀ 提供建议的其他科研人员
- ☀ 非*co-author*

一篇论文的组成——参考文献

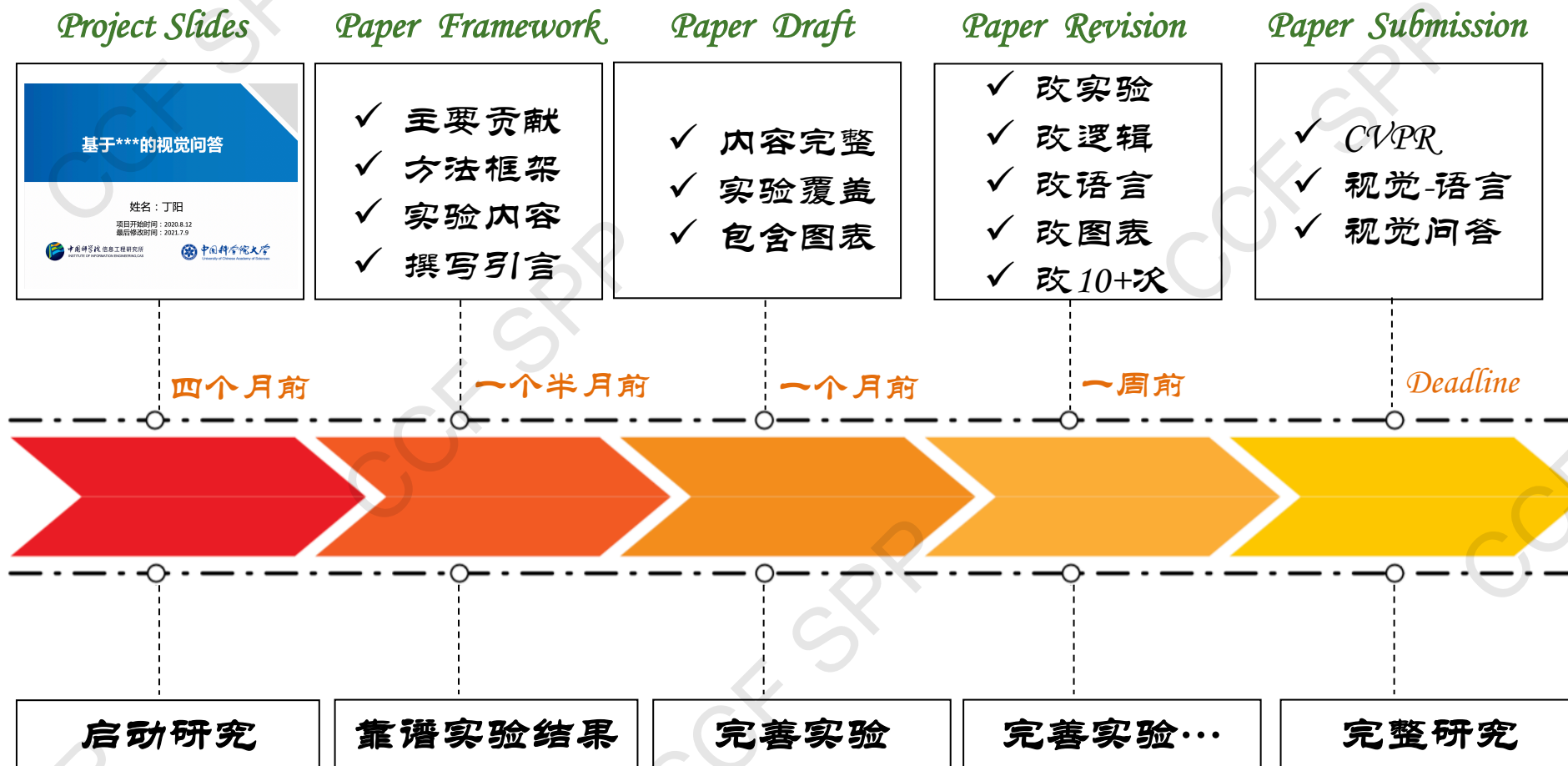
基本要求

☀ 不遗漏，查全

☀ 按照会议/期刊既定格式

☀ 常见错误：大小写、全称/缩写、漏写、名字错拼

论文写作及修改过程



会议论文如何扩展为期刊论文？

基本思路

- ☀ 会议论文已发表
- ☀ 论文题目区别于会议论文
- ☀ 说明和会议论文工作的区别
- ☀ 改进方法
 - 👉 更加通用的框架
 - 👉 更加优化的模型
 - 👉 迁移到其他任务
- ☀ 扩展实验
 - 👉 增加数据集、对比方法、对比任务
 - 👉 更全面、深入分析
- ☀ 完善描述
 - 👉 背景、问题、模型、实验更加详细

看一个例子

会议论文-Title

The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

DualVD: An Adaptive Dual Encoding Model for Deep Visual Understanding in Visual Dialogue

**Xiaoze Jiang,^{1,2} Jing Yu,^{1*} Zengchang Qin,² Yingying Zhuang,^{1,2} Xingxing Zhang,³
Yue Hu,¹ Qi Wu⁴**

¹Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

²Intelligent Computing and Machine Learning Lab, School of ASEE, Beihang University, Beijing, China

³Microsoft Research Asia, Beijing, China

⁴University of Adelaide, Australia

{yujing02, huyue}@iie.ac.cn, {xzjiang, zcquin}@buaa.edu.cn, xizhang@microsoft.com, qi.wu01@adelaide.edu.au

期刊论文-Title

IEEE TRANSACTIONS ON IMAGE PROCESSING

1

Learning Dual Encoding Model for Adaptive Visual Understanding in Visual Dialogue

Jing Yu, Xiaoze Jiang, Zengchang Qin, Weifeng Zhang, Yue Hu, Qi Wu

会议论文-*Contribution*

The main contributions are summarized as follows: (1) We exploit the possibility of cognition in visual dialogue by depicting an image from both visual and semantic views, which covers a broad range of visual content referred by most of questions in the visual dialogue task; (2) We propose a hierarchical visual information selection model, which is able to progressively select question-adaptive clues from intra-modal and inter-modal information for answering diverse questions. It supports explicit visualization in visual-semantic knowledge selection and reveals which modality has more contribution to answer the question; (3) The proposed model outperforms state-of-the-art approaches on benchmark visual dialogue datasets, which demonstrates the feasibility and effectiveness of the proposed model. The code is available at <https://github.com/JXZe/DualVD>.

期刊论文-*Contribution*

The main contributions are summarized as follows:

(1) We exploit the possibility of cognition theory in visual dialogue by depicting an image from both visual and semantic views, which covers a broad range of visual content referred by most of questions in the visual dialogue task;

(2) We propose a hierarchical visual information selection module DualVD, which can select question-adaptive clues for answering diverse questions. It supports explicit visualization in visual-semantic knowledge selection and reveals which modality has more contribution to answer the question;

(3) We propose two novel models for the visual dialogue task by integrating our proposed DualVD module with two typical frameworks: *DualVD-LF* based on Late Fusion framework and *DualVD-MN* based on Memory Network framework. The proposed models outperform state-of-the-art approaches on three visual dialogue datasets, including VisDial v0.9, VisDial v1.0 and Visual-Q, which demonstrates the feasibility and effectiveness of the proposed models.

A previous version of our dual encoding model was published in AAAI 2020 [14]. In this extension version, we extend our DualVD on the memory network for the visual dialogue so that we can pay more attention on “dialogue history reasoning” as well as the “visual reasoning”. This also suggests that our DualVD module is complementary with the improvements in dialogue modeling and can be plugged into existing visual dialogue models. We also conduct more in-depth experiments and improve the performance on the visual dialogue task. The proposed dual encoding module shows great generalization ability and can be applied to existing visual dialogue models for complementary benefits.

会议论文-Related Work

Visual Question Answering. The typical solutions for VQA are based on the CNN-RNN architecture [Malinowski *et al.*, 2015] and leverage global visual features to represent image, which may introduce noisy information. Various attention mechanisms [Yang *et al.*, 2016; Lu *et al.*, 2016; Anderson *et al.*, 2018] have been exploited to highlight visual objects that are relevant to the question. However, they treat objects independently and ignore their informative relationships. [Battaglia *et al.*, 2018] demonstrates that human's ability of combinatorial generalization highly depends on the mechanisms for reasoning over relationships. Consistent with such proposal, there is an emerging trend to represent the image by graph structure to depict objects and relationships in VQA and other vision-language tasks [Hu *et al.*, 2019b; Wang *et al.*, 2019a; Li *et al.*, 2019b]. As an extension, [Jiang *et al.*, 2020] exploits natural language to enrich the graph-based visual representations. However, it solely captures the semantics in natural language by LSTM, which lacking of fine-grained correlations with the visual information. To go one step further, we depict an image by multiple layers of graphs from visual, semantic and factual perspectives to collect fine-grained evidence from different modalities.

Fact-based Visual Question Answering. Human can easily combine visual observation with external knowledge for answering questions, which remains challenging for algorithms. [Wang *et al.*, 2018] introduces a fact-based VQA task, which provides a knowledge base of facts and associates each question with a supporting-fact. Recent works based on FVQA generally select one entity from fact graph as the answer and falls into two categories: query-mapping based methods and learning based methods. [Wang *et al.*, 2017] reduces the question to one of the available query templates and this limits the types of questions that can be asked. [Wang *et al.*, 2018] automatically classifies and maps the question to a query which does not suffer the above constraint. Among both methods, however, visual information are used to extract facts but not introduced during the reasoning process. [Narasimhan *et al.*, 2018] applies GCN on the fact graph where each node is represented by the fixed form of image-question-entity embedding. However, the visual information is wholly provided which may introduce redundant information for prediction. In this paper, we depict an image by multi-layer graphs and perform cross-modal heterogeneous graph reasoning on them to capture complementary evidence from different layers that most relevant to the question.

Heterogeneous Graph Neural Networks. Graph neural networks are gaining fast momentum in the last few years [Wu *et al.*, 2019]. Compared with homogeneous graphs, heterogeneous graphs are more common in the real world. [Schlichtkrull *et al.*, 2018] generalizes graph convolutional network (GCN) to handle different relationships between entities in a knowledge base, where each edge with distinct relationships is encoded independently. [Wang *et al.*, 2019b; Hu *et al.*, 2019a] propose heterogeneous graph attention networks with dual-level attention mechanism. All of these methods model different types of nodes and edges on a unified graph. In contrast, the heterogeneous graph in this work contains multiple layers of subgraphs and each layer consists of nodes and edges coming from different modalities. For this specific constrain, we propose the intra-modal and cross-modal graph convolutions for reasoning over such multi-modal heterogeneous graphs.

期刊论文-Related Work

Visual Question Answering (VQA) focuses on answering arbitrary natural language questions conditioned on an image. The typical solutions in VQA build multi-modal representations upon CNN-RNN architecture [1], [15], [16]. They adopt deep Convolutional Neural Networks (CNNs) to represent images and Recurrent Neural Networks (RNNs) to represent questions. The extracted visual and textual feature vectors are then jointly embedded to infer the answer. One of the key challenges in VQA is to effectively understand and extract visual features that better adapt to the question. Existing approaches incorporate context-aware visual features and the trend for modeling the visual context is progressively from global level to fine-grained level. For example, [15] applies CNN features of the whole image as global context, [17] and [18] adopt patches and salient objects learned by attention mechanism as the region context, and [19] exploits inter-object relationships via graph attention networks to model the relational context. However, how to leverage the external visual-semantic knowledge to learn more informative relational representations and combine them with higher-level visual features for better semantic understanding has not been well exploited yet.

Another emerging line of work represents visual content explicitly by natural language and solves VQA as a reading comprehension problem. In [20], the image is wholly converted into descriptive captions, which preserves information at semantic-level in textual domain. However, this kind of approaches use the generated captions, which could not be correct as we desired, and that they fully abandon the informative and subtle visual features. Besides the specific tasks, our model has notable progress compared to the above approaches. We adopt dual encoding mechanism to provide both appearance-level and semantic-level visual information, so that it incorporates the strong points of the above two kinds of approaches.

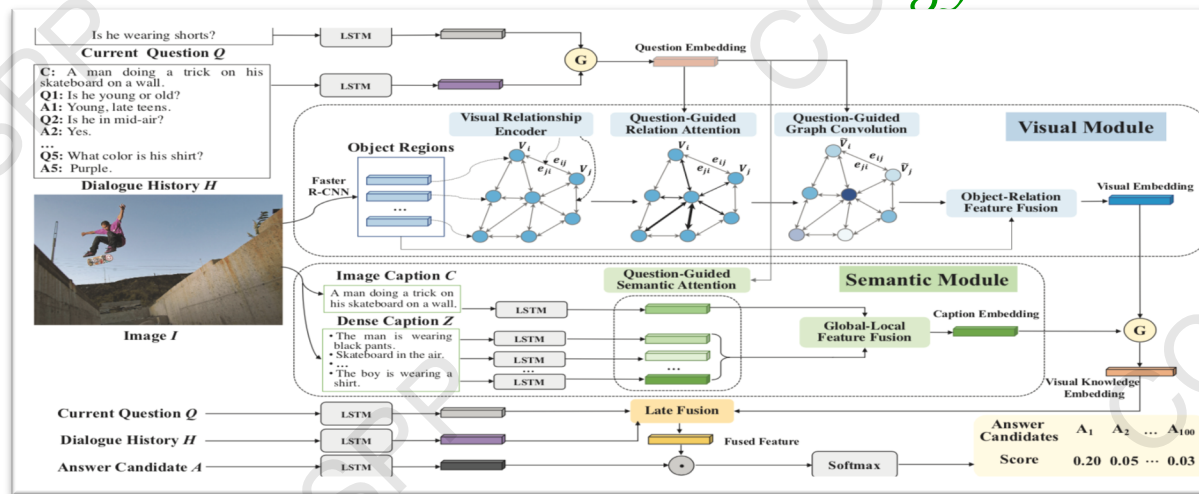
Visual Dialogue aims to answer a current question conditioned on an image and dialogue history. Compared with Visual Question Answering task, Visual Dialogue involves multi-round dialogue history as context besides the image and the question. Most existing works are based on late fusion framework and focus on modeling the dialogue history. Sequential co-attention mechanism [10] enables the model to identify question-relevant image regions and dialogue history to keep the dialogue consistency. [9] introduces false response in dialogue history for an adverse critic on the historic error. To investigate semantic dependencies between entities underlying dialogue, [8] introduces an Expectation Maximization (EM) algorithm to infer the dialogue structure and the answers via graph neural networks. [21] proposed a novel image-question-answer synergistic network to value the role of the answer for precise visual dialogue. The most recent works [22], [23], [24] proposed graph inference or causal intervention to reason about the answer on the image and dialogues.

Visual Relationship Understanding aims to represent and infer the relationships between two objects in an image, which is critical to improve AI's capacity of combinatorial generalization by learning towards relational visual representations. In the early works [27], shallow geometric relationships (e.g. below, above, and inside) based on spatial information have been explored to improve visual segmentation. Later on, visual relationships have been extended to richer definition [28], including geometric, comparative, composition, interaction, *etc.* One limitation of the above approaches is that they merely represent the visual relationships by rigid-categorized labels, which are difficult to accurately model the subtle relationships conditioned on the contexts. For example, an image of <woman, ride, horse> and another image of <man, ride, motorcycle> have quite different visual appearance and semantics even they belong to the same relationship. Even for the same visual appearance, it may contain different relationships. For example, an image where a boy is kicking a football

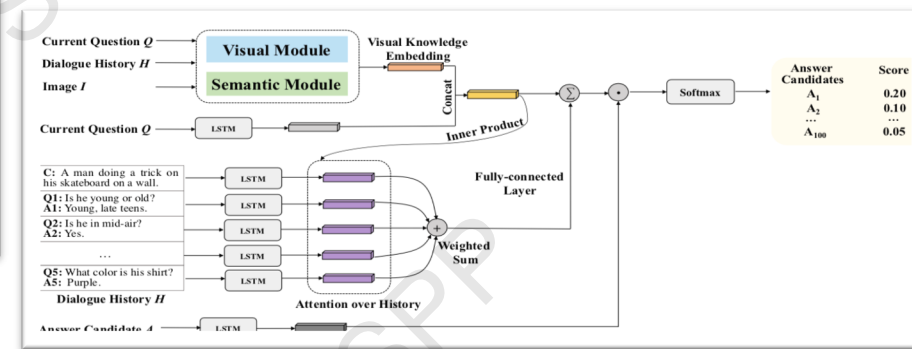
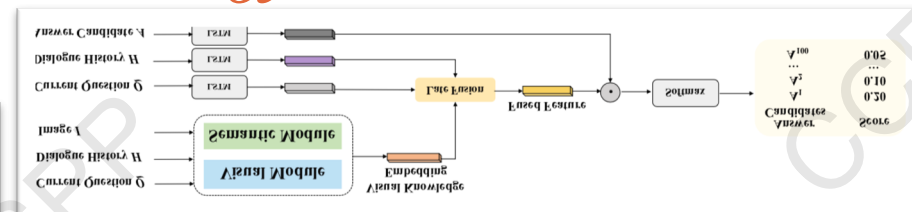
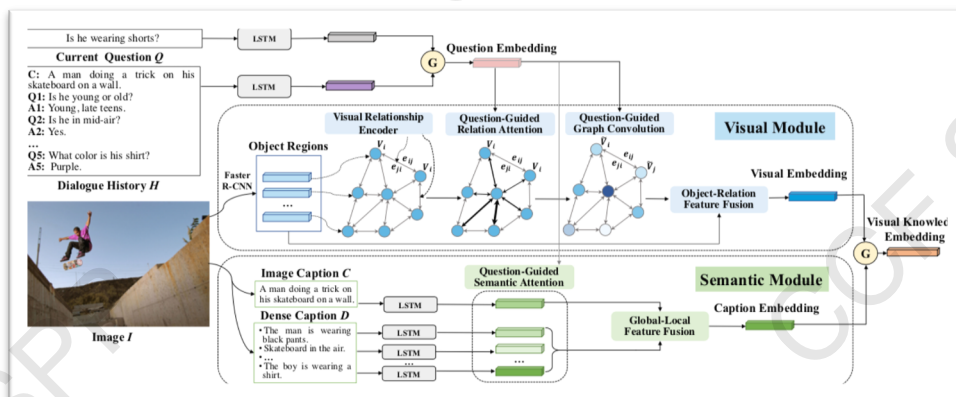
Dense Captioning aims to jointly localize and describe image regions in natural language. It provides fine-grained visual descriptions for each local image region compared with object detection and image captioning. The task generalizes object detection when the descriptions are simple labels and image captioning when one predicted region covers the full image. It simultaneously takes the object detection and description task into account. Previous work on holistic description of visual element [32], [33], [34] are either limited to salient objects of images, or tend to broadly depict the entire visual scene. These descriptions are far from complete visual understanding. [3] proposed to use dense captions for better interpretation of image content. The model consists of a Fully Convolutional Localization Network and an LSTM based Language Model that produces both bounding boxes for interest regions and associated captions in a single forward pass. In this work, we leverage dense captions to describe the semantic-level visual content for the local relationships between objects. In this way, visual information is represented in linguistic form that is closer to human's cognition.

看一个例子

会议论文-Methodology



期刊论文-Methodology



看一个例子

期刊论文-Experiments

TABLE I
RESULT COMPARISON ON VALIDATION SET OF VisDial v0.9.

Model	MRR	R@1	R@5	R@10	Mean
LF[7]	58.07	43.82	74.68	84.07	5.78
HRE[7]	58.46	44.67	74.50	84.22	5.72
HREA[7]	58.68	44.82	74.81	84.36	5.66
MN[7]	59.65	45.55	76.22	85.37	5.46
SAN-QI[17]	57.64	43.44	74.26	83.72	5.88
HicCoAtt-QI[42]	57.88	43.51	74.49	83.96	5.84
AMEM[43]	61.60	47.74	78.04	86.84	4.99
HCIAE[44]	62.22	48.48	78.75	87.59	4.81
SF[45]	62.42	48.55	78.96	87.75	4.70
CoAtt[10]	63.98	50.29	80.71	88.81	4.47
CorefMN[46]	64.10	50.92	80.18	88.81	4.45
VGNN[8]	62.85	48.95	79.65	88.36	4.57
DualVD-LF	62.94	48.64	80.89	89.94	4.17
DualVD-MN	63.12	48.89	81.11	90.33	4.12

TABLE IV
ABLATION STUDY ON VALIDATION SET OF VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
ObjRep	63.84	49.83	81.27	90.29	4.07	55.48
RelRep	63.63	49.25	81.01	90.34	4.07	55.12
VisNoRel	63.97	49.87	81.74	90.60	4.00	56.73
VisMod	64.11	50.04	81.78	90.52	3.99	56.67
GlCap	60.02	45.34	77.66	87.27	4.78	50.04
LoCap	60.95	46.43	78.45	88.17	4.62	51.72
SemMod	61.07	46.69	78.56	88.09	4.59	51.10
w/o ELMo	63.67	49.89	80.44	89.84	4.14	56.41
DualVD-LF	64.64	50.74	82.10	91.00	3.91	57.30
DualVD-MN	64.70	50.79	82.41	91.10	3.90	58.24

TABLE II
RESULT COMPARISON ON TEST-STANDARD SET OF VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF[7]	55.42	40.95	72.45	82.83	5.95	45.31
HRE[7]	54.16	39.93	70.47	81.50	6.41	45.46
MN[7]	55.49	40.98	72.30	83.30	5.92	47.50
LF-Att[7]	57.07	42.08	74.82	85.05	5.41	40.76
MN-Att[7]	56.90	42.43	74.00	84.35	5.59	49.58
CorefMN[46]	61.50	47.55	78.10	88.80	4.40	54.70
VGNN[8]	61.37	47.33	77.98	87.83	4.57	52.82
RvA[47]	63.03	49.03	80.40	89.83	4.18	55.59
DL-61[21]	62.20	47.90	80.43	89.95	4.17	57.32
DualVD-LF	63.23	49.25	80.23	89.70	4.11	56.32
DualVD-MN	63.38	49.35	81.05	90.38	4.07	57.09

TABLE V
GENERATIVE MODEL COMPARISON ON VALIDATION SPLIT OF VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
MN-G[7]	47.83	38.01	57.49	64.08	18.76	56.99
HCIAE-G[44]	49.07	39.72	58.23	64.73	18.43	59.70
CoAtt-G[10]	49.64	40.09	59.37	65.92	17.86	59.24
ReDAN-G[26]	49.60	39.95	59.32	65.97	17.79	59.41
DualVD-LF-G	49.78	39.96	59.96	66.62	17.49	60.08

TABLE III
RESULT COMPARISON ON VALIDATION SET OF VisDial-Q.

Model	MRR	R@1	R@5	R@10	Mean
LF[7]	18.45	7.80	26.12	40.78	20.42
MN[7]	39.83	25.80	54.76	69.80	9.68
SF-QI[39]	30.21	17.38	42.32	57.16	14.03
SF-QIH[39]	40.60	26.76	55.17	70.39	9.32
VGNN[8]	41.26	27.15	56.47	71.97	8.86
DualVD-LF	41.31	27.24	56.50	71.51	9.09
DualVD-MN	41.34	27.27	56.60	71.45	9.15

TABLE VI
RESULT OF DIFFERENT T_l OF DUALVD ON VisDial 1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
$T_l = 4$	64.38	50.44	81.79	90.58	3.97	57.34
$T_l = 6$	64.64	50.74	82.10	91.00	3.91	57.30
$T_l = 8$	64.35	50.28	81.94	90.69	3.94	56.28
$T_l = 10$	64.34	50.37	81.74	90.71	3.97	57.09

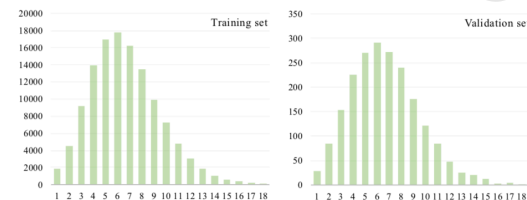


Fig. 10. Analysis of dense caption distribution on “train” and “val” splits of VisDial v1.0. x-axis represents the number of dense captions while y-axis represents the number of images generating corresponding number of captions.

期刊论文-Experiments

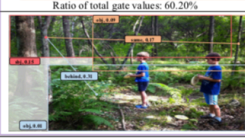
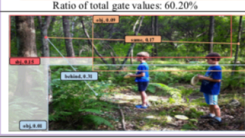
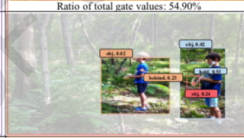
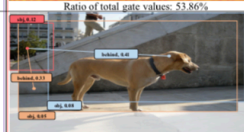
Image	Dialogue History	Visual Module	Semantic Module	
	C: A woman sits on a bench holding a guitar in her lap. Q1: Is this in a park? A1: Yes, I believe it is. Q2: Are there others around? A2: No, she is alone. Q3: Does she have a collection bucket? A3: No. Q4: Is her hair long? A4: Yes, pretty long. Q5: Is she wearing a dress? A5: I don't think so, hard to tell. Q6: Does she have shoes on? A6: Yes, flip flops. Question Answer Is there grass nearby? Yes, there is.	Ratio of total gate values: 60.93% 	Ratio of total gate values: 39.07% A woman sits on a bench holding a guitar in her lap. Woman sitting on bench. A green grassy field. Woman with long hair. A woman sitting on a bench. Woman with long hair. Green grass on the ground.	
	C: A girl and dog on a motorcycle dressed in costumes. Q1: Color pie? A1: Yes. Q2: Is the dog driving? A2: The dog is sitting on the motorcycle, like he would be the driver. Q3: Is the girl cute? A3: Yes. Q4: Is she old? A4: Around 9. Q5: Is she smart? A5: I'm not sure. Q6: Does she have hair? A6: Yes. Question Answer Does she have hair? Yes.	Ratio of total gate values: 52.09% 	Ratio of total gate values: 47.91% A girl and dog on a motorcycle dressed in costumes. Girl with blue hair. A dog with a black and white collar. A white and black dog. A dog with a collar. The girl has long hair.	
	C: 2 boys playing disc golf in a forest. Question1 Answer1 Are the boys teenagers? They are young boys. Question2 Answer2 Do you see a lot of trees? Yes, a ton of trees. Question3 Answer3 Dose 1 of the boys holding the disc? They are both holding discs.	Ratio of total gate values: 55.96% 	Ratio of total gate values: 44.04% 2 boys playing disc golf in a forest. A man wearing blue shorts. Boy holding blue frisbee. Two people playing with a frisbee. A blue shirt on a man. Boy wearing blue shirt. Blue shorts on the man.	
	Ratio of total gate values: 60.20% 	Ratio of total gate values: 39.80% 2 boys playing disc golf in a forest. A man wearing blue shorts. Boy holding blue frisbee. Two people playing with a frisbee. A blue shirt on a man. Boy wearing blue shirt. Blue shorts on the man.	Ratio of total gate values: 54.90% 	Ratio of total gate values: 45.10% 2 boys playing disc golf in a forest. A man wearing blue shorts. Boy holding blue frisbee. Two people playing with a frisbee. A blue shirt on a man. Boy wearing blue shirt. Blue shorts on the man.
	C: A dog in a collar stands next to stairs. Q1: Is it a large dog? A1: It's a lab. Q2: What color is the dog? A2: Yellow. Question3 Answer3 Are the stairs wooden? Concrete. Question4 Answer4 Is the dog at the bottom or top of them? Bottom. Question5 Answer5 Is the dog outdoors? Yes.	Ratio of total gate values: 55.11% 	Ratio of total gate values: 44.89% A dog in a collar stands next to stairs. A grey concrete building. A dog standing on a sidewalk. Dog with a collar. Man wearing black pants. A stone wall.	
	Ratio of total gate values: 56.74% 	Ratio of total gate values: 43.26% A dog in a collar stands next to stairs. A grey concrete building. A dog standing on a sidewalk. Dog with a collar. Man wearing black pants. A stone wall.	Ratio of total gate values: 53.86% 	Ratio of total gate values: 46.14% A dog in a collar stands next to stairs. A grey concrete building. A dog standing on a sidewalk. Dog with a collar. Man wearing black pants. A stone wall.

Fig. 8. Visualization for success cases. Visual module highlights the most relevant subject (red box) according to attention weights of each object (γ_{ij}^v in Eq. 9) and the objects (orange and blue boxes) with the top two attended relationships (β_{ij} in Eq. 5). Semantic module shows the attention distribution (δ_{ij}^s in Eq. 11) over the global caption (first row) and the local captions (rest rows), where darker green color indicates bigger attention weight. The yellow thermogram on the top visualizes the gate values ($gate^s$ in Eq. 16) of the visual embedding (left) and the caption embedding (right) in visual-semantic fusion.

C. Interpretability

A critical advantage of DualVD lies in its interpretability: DualVD is capable to predict the attention weights in the visual module, semantic module and the gate values in visual-semantic fusion. It supports explicit visualization and can reveal DualVD's mode in information selection. We show the visualization for success cases and failure cases of DualVD-LF model in Figure 8 and Figure 9, respectively. Four meaningful observations of our model are presented as follows:

Comprehensive visual-semantic clues. The visual features at object-level, relationship-level, and semantic-level are preserved in the framework of DualVD, which enables the DualVD-LF model to answer a wide range of visually grounded questions through the dialogue. For instance, in Figure 8, the third example (third and fourth rows in Figure 8)

Information selection mode. By selecting visual and semantic information by gating mechanism, the DualVD-LF model can reveal the mode in information selection for answering the current question by visualizing the gate values. We observe that the amount of information derived from each module highly depends on the complexity of the question and the relevance of the module content. More information will come from the semantic module when the question involves

The critical role of visual information. From the visualization results, we find that the visual information is more important than the semantic information in question answering. In all the testing cases, the ratio of gate values of the visual module is larger than that of the semantic module and also that the differences between the two are not greatly disproportionate. This demonstrates that more comprehensive and accurate clues come from the visual information, though

看一个例子

期刊论文-Experiments

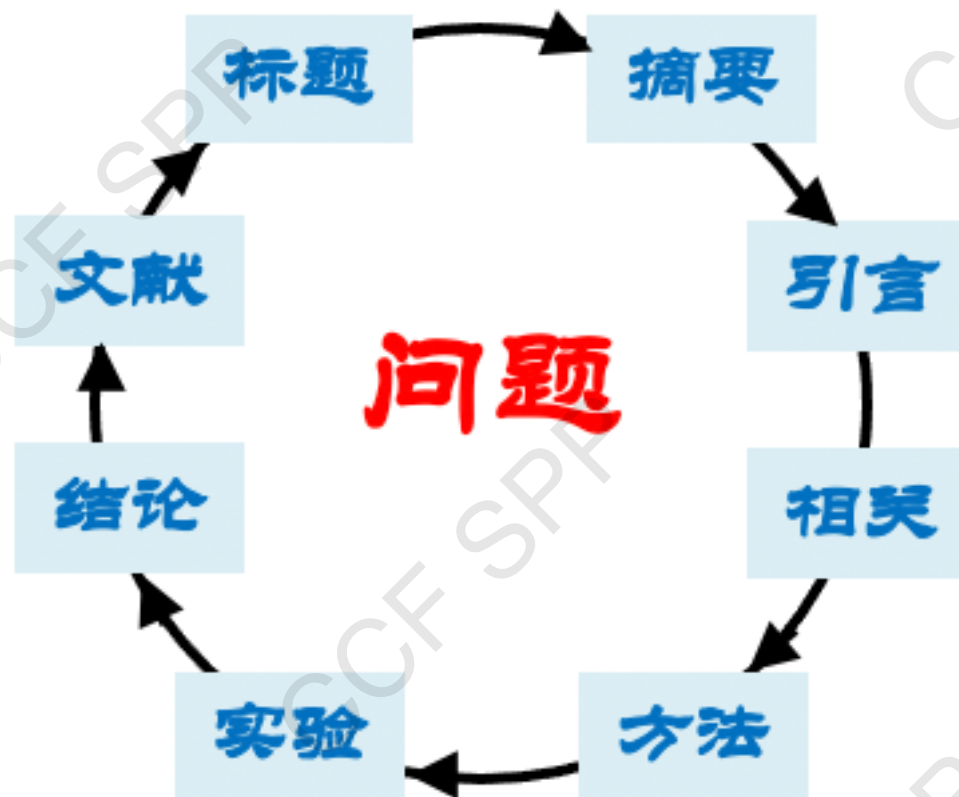
Image	Dialogue History	Visual Module	Semantic Module
	C: A closed storefront with red graffiti painted on the door. Q1: Is the store permanently closed? A1: Yes. Q2: What is the graffiti? A2: Words.		
	Question	What do the words say?	
	Answer-p	Can't tell, it's just scribbling, some random words like vote, bank.	
	Answer-g	It's actually just letters.	
	C: A man on his phone standing next to a woman with a bottle of water with an old Audi in the background. Q1: Is there a brand on the bottle of water? A1: No, it's facing the other way.		
	Question	Can you tell what kind of phone?	
	Answer-p	No.	
	Answer-g	It's on the other side of his face, but it's black.	
	C: A woman with red hair is standing under a multicolored umbrella. Q1: What type of hair does the woman have? A1: She has long red hair. Q2: What is the weather? A2: It is cloudy and rainy. Q3: How old is the woman? A3: She looks to be mid 20s. Q4: Is there anyone else with her? A4: She is alone. Q5: Is she in the city? A5: She is in a city. Q6: What color clothes is she wearing? A6: She is wearing a plaid coat.		
	Question	What time of day is it?	
	Answer-p	It is daytime.	
	Answer-g	It appears to be afternoon.	

Failure case analysis. Since the visualization results explicitly show the visual and semantic clues the model focused on, it provides us evidence to analysis the reasons for the failure cases. In Figure 9, we show three failure cases from three typical failure modes. We can see that our model can accurately attend to the visual and the semantic information that relevant to the questions in all the cases. In the first example, the ground-truth answer is similar to the predicted answer. This mode is so-called *one-to-many problem* in the dialogue system. A second failure mode is due to the incomplete visual information in the image. In the second example, the phone is obscured by the face, which provides limited clues to predict an exact answer for the question “Can you tell what kind of phone?”. The third mode is due to the lack of visual clues for more accurate answer. In the third example, our model accurately focuses on the “sky” and predicts “It is daytime”. However, the image doesn’t provide extra evidence to infer more precise time, i.e. “afternoon” in the ground-truth answer.

Take Home Message

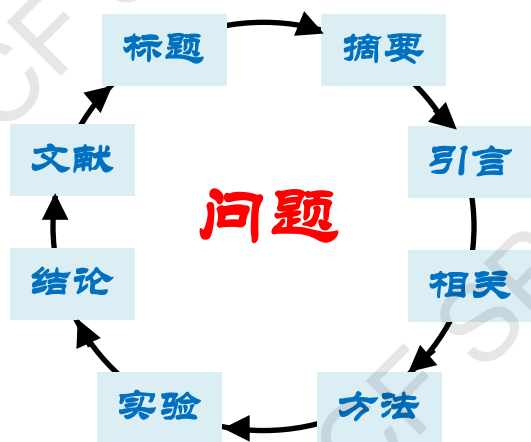


写作思路

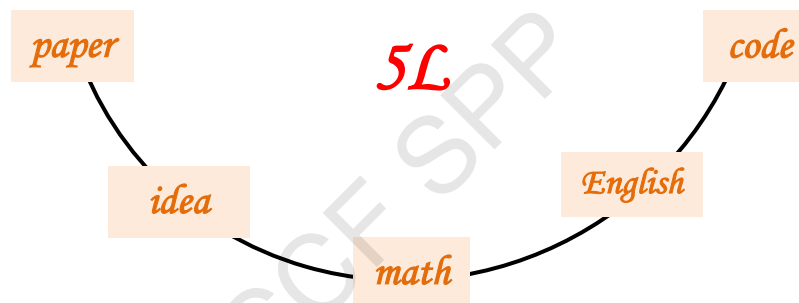
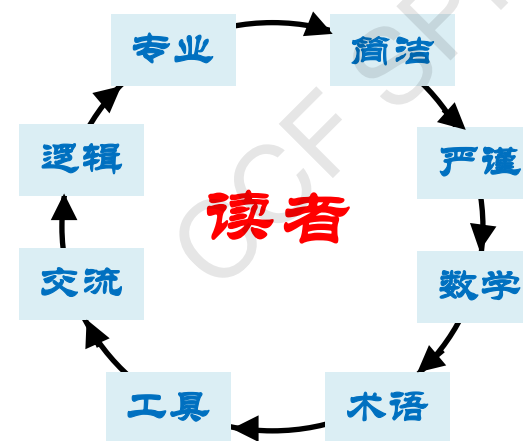


Take Home Message

☀ 写作思路



☀ 英文规范



☀ 日常积累

CCF SPP 第二期《学术论文写作规范与日常积累》，敬请期待！

参考书籍

- ☀ Booth, Wayne C., et al. *The craft of research*. University of Chicago press, 2003.
- ☀ Petre, Marian, and Gordon Rugg. *The Unwritten Rules Of Phd Research*. McGraw-Hill Education (UK), 2010.
- ☀ Macrina, Francis L. *Scientific Integrity*. ASM Press, 2005.

参考课程

- ☀ [董彬，北京大学，北京国际数学研究中心，《研究型学习》](#)

**感谢董彬、沈华伟、吴琦、罗训、崔鹏、王昌栋、龙锦益、陈姝宇等
对本报告内容的指导和启发！**



*As long as I am dreaming, believing and doing,
I can go anywhere and achieve anything!*

于静

邮箱: yujing02@iie.ac.cn

研究组主页: <https://mmlab-iie.github.io/>

知乎专栏: https://www.zhihu.com/column/c_1284803871596797952

研究组主页



知乎专栏



中国科学院 信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING, CAS



中国科学院大学
University of Chinese Academy of Sciences