科学研究与学术论文写作基础

于静

中国科学院信息工程研究所

2022.03.01 @上海大学计算机学院

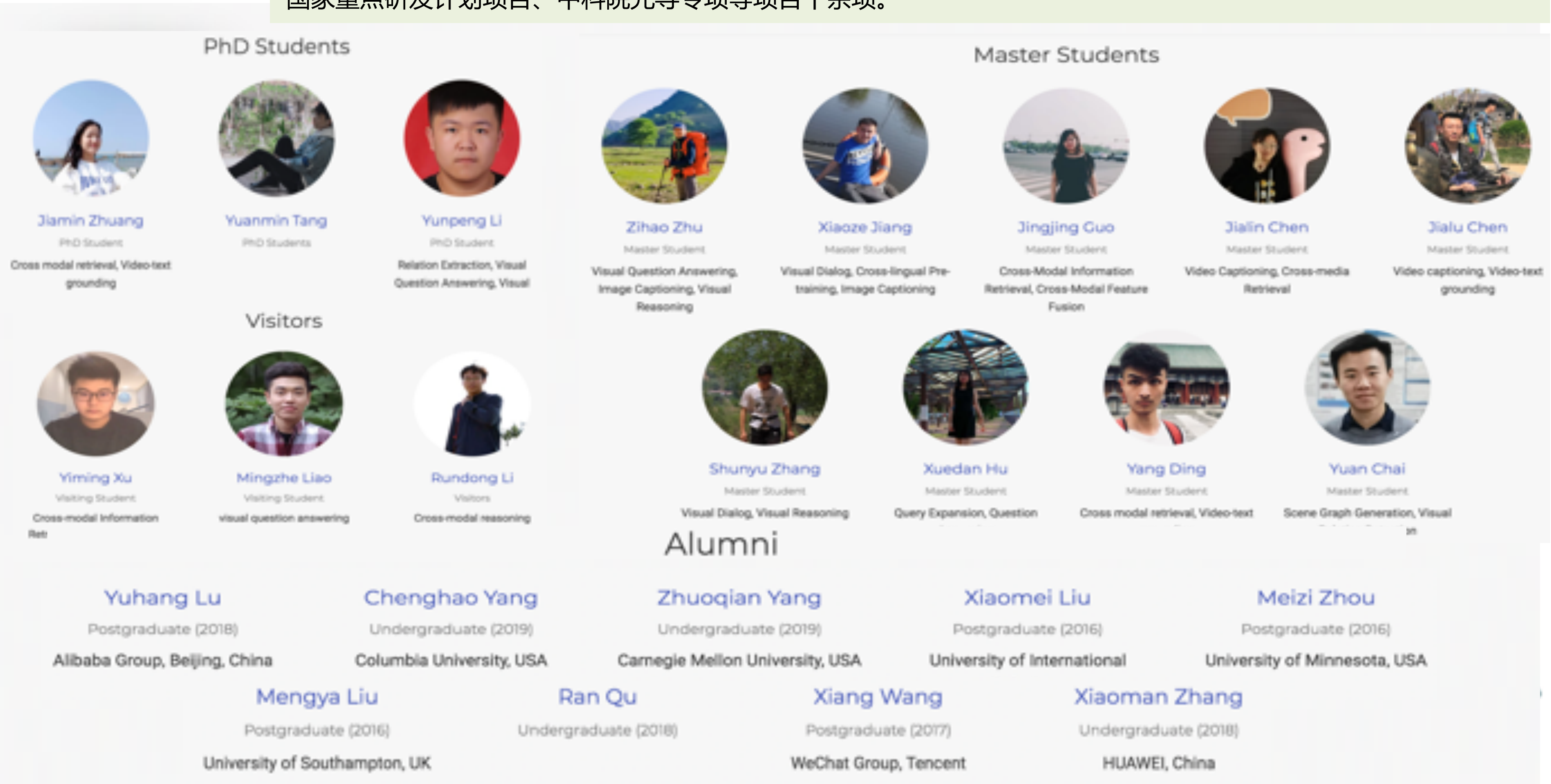
认知启发的跨模态智能研究组（GogModal）

于静，中国科学院信息工程研究所　副研究员

研究组主页：https://mmlab-iie.github.io/

研究方向：跨模态智能，包括视觉问答、视觉对话、跨媒体检索等

在国内外学术会议及期刊发表论文40余篇，包括ICML、ICLR、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。主持与参与国家自然科学基金项目、国家重点研发计划项目、中科院先导专项等项目十余项。



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

Visitors

Yiming Xu
Visiting Student
Cross-modal Information Ret

Mingzhe Liao
Visiting Student
visual question answering

Rundong Li
Visitors
Cross-modal reasoning

Master Students

Zihao Zhu
Master Student
Visual Question Answering, Image Captioning, Visual Reasoning

Xiaoze 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

Shunyu Zhang
Master Student
Visual Dialog, Visual Reasoning

Xuedan Hu
Master Student
Query Expansion, Question

Yang Ding
Master Student
Cross modal retrieval, Video-text

Yuan Chai
Master Student
Scene Graph Generation, Visual

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/ 学术论文之英文规范

4/ 学术论文之日常积累

5/ 本讲总结与心得分享

学术研究
与论文写作

IDEA
CREATIVE
SMART art
IMAGINATION
MIND
idea
VISUAL
design
concept
creative
SOLUTION CONCEPT
image
MIND
imagination
DESIGN
solution
CONCEPTION
art
smart
IMAGE
creative
mind
visual
CREATIVE
INSPIRATION
CONCEPTION
image
creative
idea

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？

CogModal
GROUP

为什么要做学术研究？

方法论

解决问题

问题发现

系统调研

有效方法

实践验证

认知力

科学素养

学术品位

结构思维

严谨逻辑

清晰表达

价值观

长线思维

不急功近利

不患得患失

交流共享

持续积累

CogModal GROUP

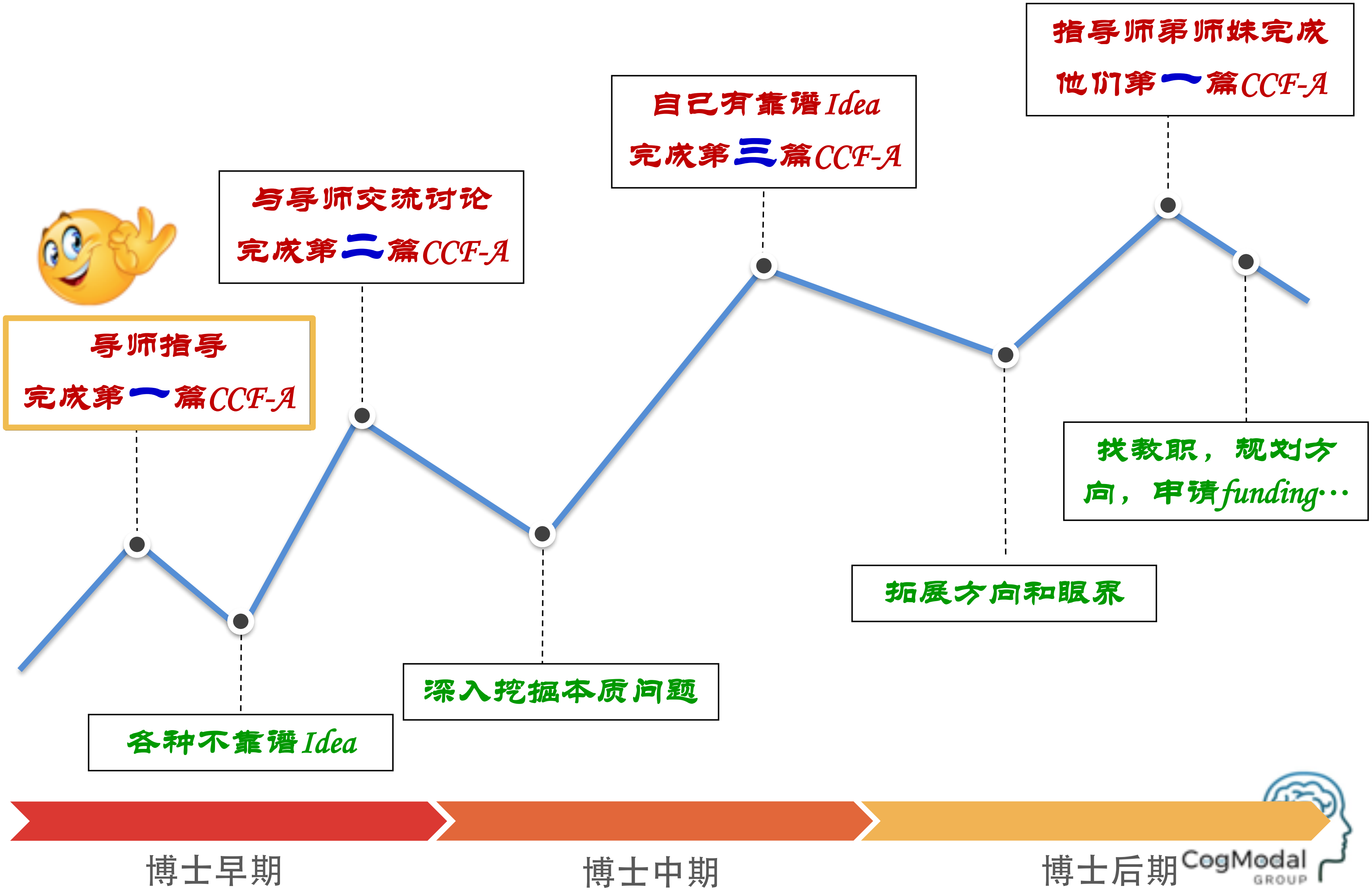
学术研究的基本要求

规律： 依据研究规律，确定研究目标

规则： 依据领域共识，细化研究方法

规范： 依据学术规范，行文准确表达

学术研究修炼之路——四生四世



导师指导
完成第一篇CCF-A

各种不靠谱Idea

与导师交流讨论
完成第二篇CCF-A

深入挖掘本质问题

自己有靠谱Idea
完成第三篇CCF-A

拓展方向和眼界

指导师弟师妹完成
他们第一篇CCF-A

找教职，规划方向，申请funding…

博士早期

博士中期

博士后期

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

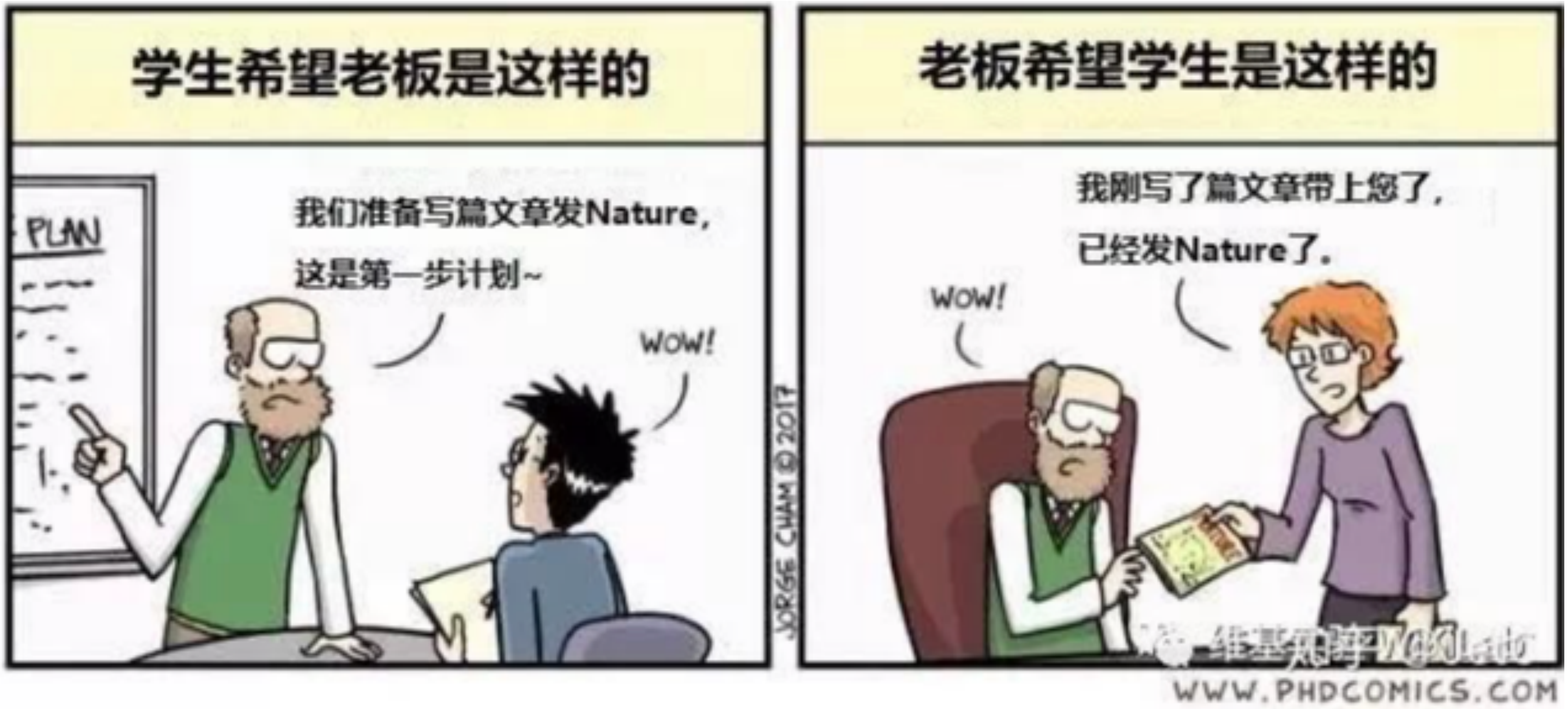
研究领域 → 研究动机 → 问题提出 → 方法设计 → 实验验证 → 总结展望

Topic Motivation Problem Method Experiments Conclusion

论文写作过程

gModal GROUP

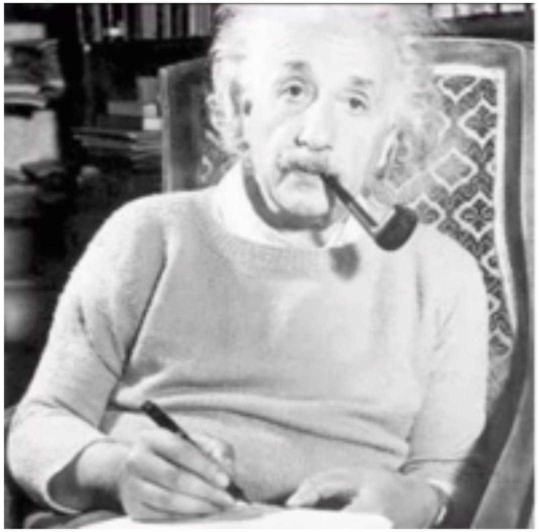
我们的论文出了什么问题？



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



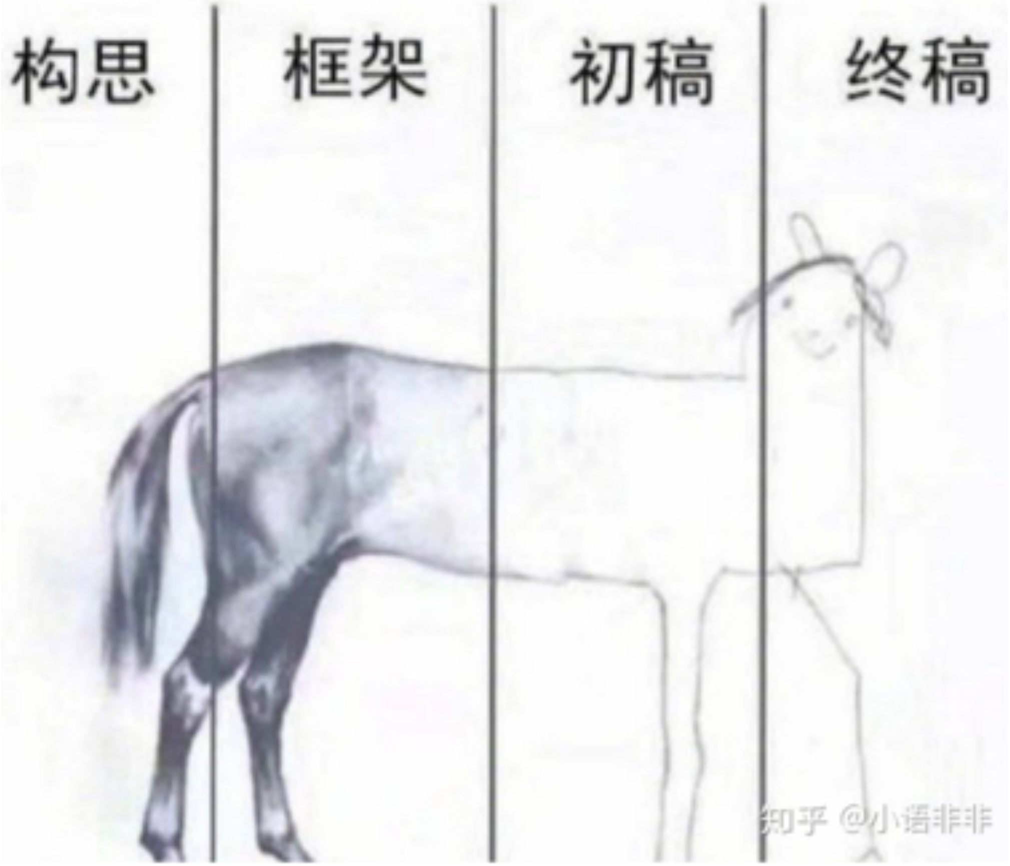
导师眼中的论文

审稿人眼中的论文

论文总是写给自己看，别人看不懂！

CogModal
GROUP

我们的论文出了什么问题？



论文无法准确表达研究内容！

CogModal GROUP

我们的论文出了什么问题？

论文 Story 难以自圆其说！

CogModal GROUP

我们的论文出了什么问题？

我离论文写完就差这么一点点

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

不知道论文如何逐步完善！

我们的论文出了什么问题？

结果： *CCF-A*的研究内容只达到*CCF-C*的论文水平！

CogModal
GROUP

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

CogModal GROUP

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

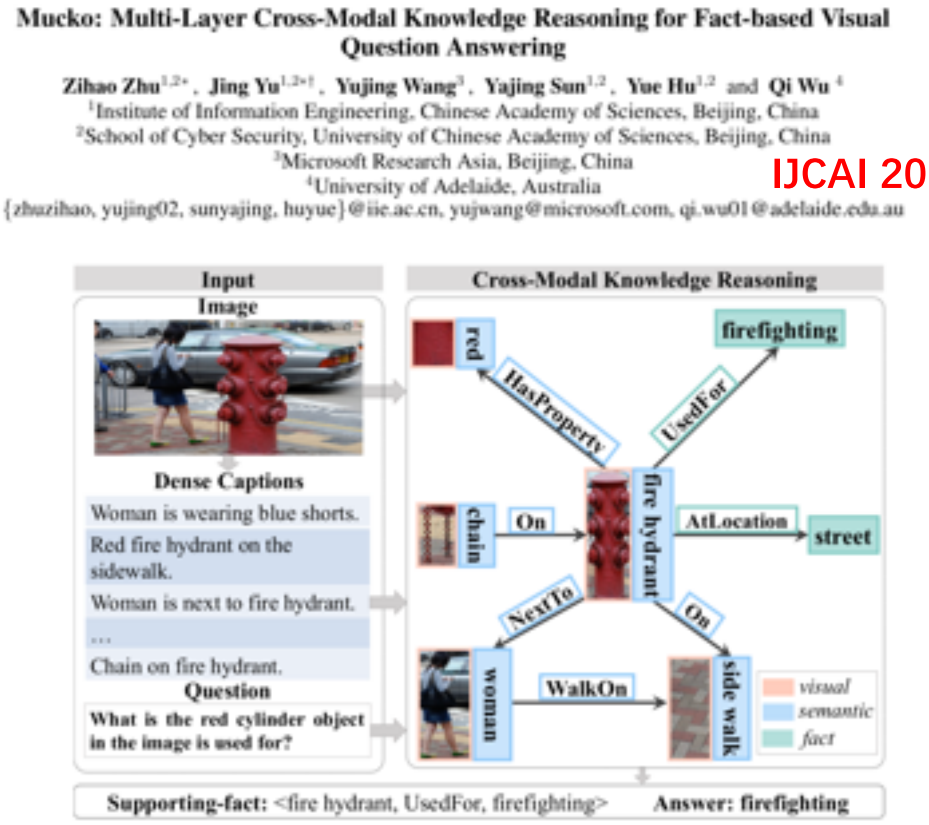
参考《中国计算机学会推荐国际学术会议和期刊目录》

学术论文
之写作思路

IDEA
SMART art CREATIVE
visual
IMAGINATION
MIND concept
VISUAL idea
design
concept
creative idea
SOLUTION CONCEPT
image MIND
imagination
DESIGN
DESIGN mind
solution
CONCEPTION
art
smart
IMAGE
talent creative
smart
mind
mind
art
visual
CREATIVE
INSPIRATION
image
CONCEPTION
creative
DESIGN
idea

从科学问题是否探究本质说起

	CCF—A	*CCF—C*
科学问题	聚焦领域前沿，探索本质问题	追逐领域热点，解决表面问题
解决方法	专注核心挑战，具普适性方法	盲从热门方法，模型增量修改

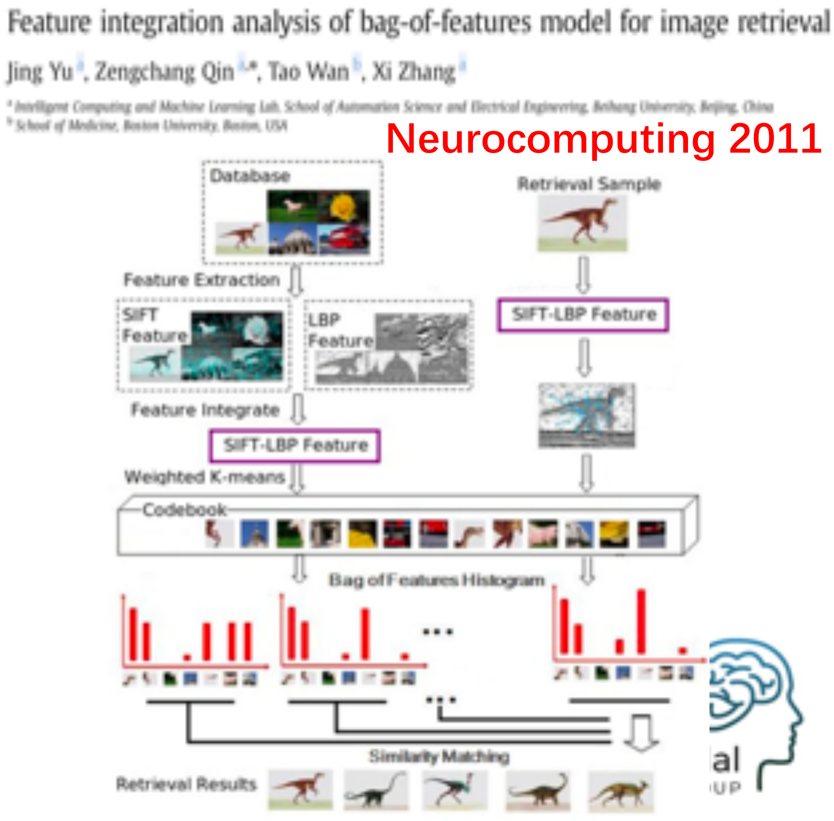


Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering

Zihao Zhu[1,2*], Jing Yu[1,2*†], Yujing Wang[3], Yajing Sun[1,2], Yue Hu[1,2] and Qi Wu[4]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Microsoft Research Asia, Beijing, China
[4]University of Adelaide, Australia
{zhuzihao, yujing02, sunyajing, huyue}@iie.ac.cn, yujwang@microsoft.com, qi.wu01@adelaide.edu.au

IJCAI 2020

Input
Image
Dense Captions
Woman is wearing blue shorts.
Red fire hydrant on the sidewalk.
Woman is next to fire hydrant.
…
Chain on fire hydrant.
Question
What is the red cylinder object in the image is used for?
Cross-Modal Knowledge Reasoning
red
HasProperty
firefighting
UsedFor
chain
On
fire hydrant
AtLocation
street
NextTo
On
woman
WalkOn
side walk
visual
semantic
fact
Supporting-fact: <fire hydrant, UsedFor, firefighting>
Answer: firefighting



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

Jing Yu[a], Zengchang Qin[a,*], Tao Wan[b], Xi Zhang[a]
[a] Intelligent Computing and Machine Learning Lab, School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
[b] School 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：创新点与研究内容的高度凝练
- 摘要：标题的扩充
- 引言：摘要的扩充
- 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
- 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

CogModal GROUP

一篇论文的组成——标题 (Good Cases)

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

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

Evolving Attention with Residual Convolutions (ICLR 2021)

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

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

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！

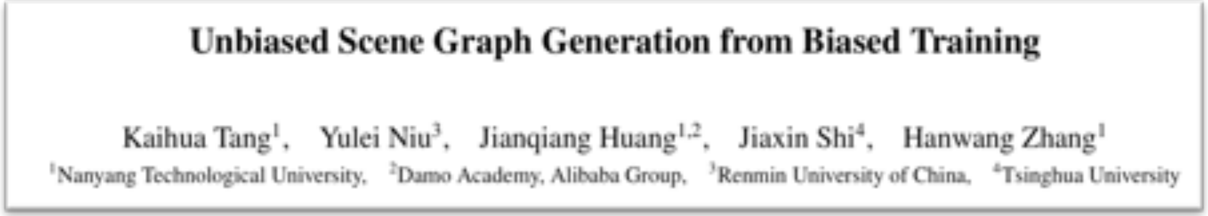
一篇论文的组成——摘要

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

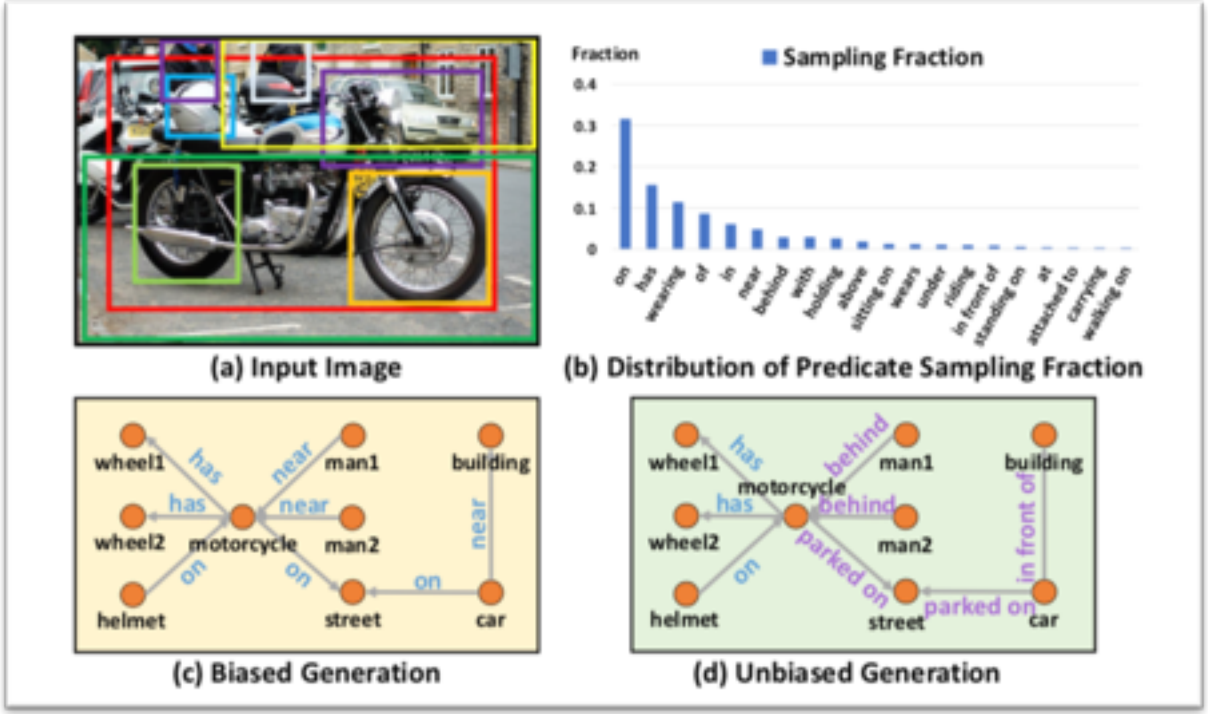
2020 CVPR



Unbiased Scene Graph Generation from Biased Training

Kaihua Tang[1], Yulei Niu[3], Jianqiang Huang[1,2], Jiaxin Shi[4], Hanwang Zhang[1]

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



(a) Input Image

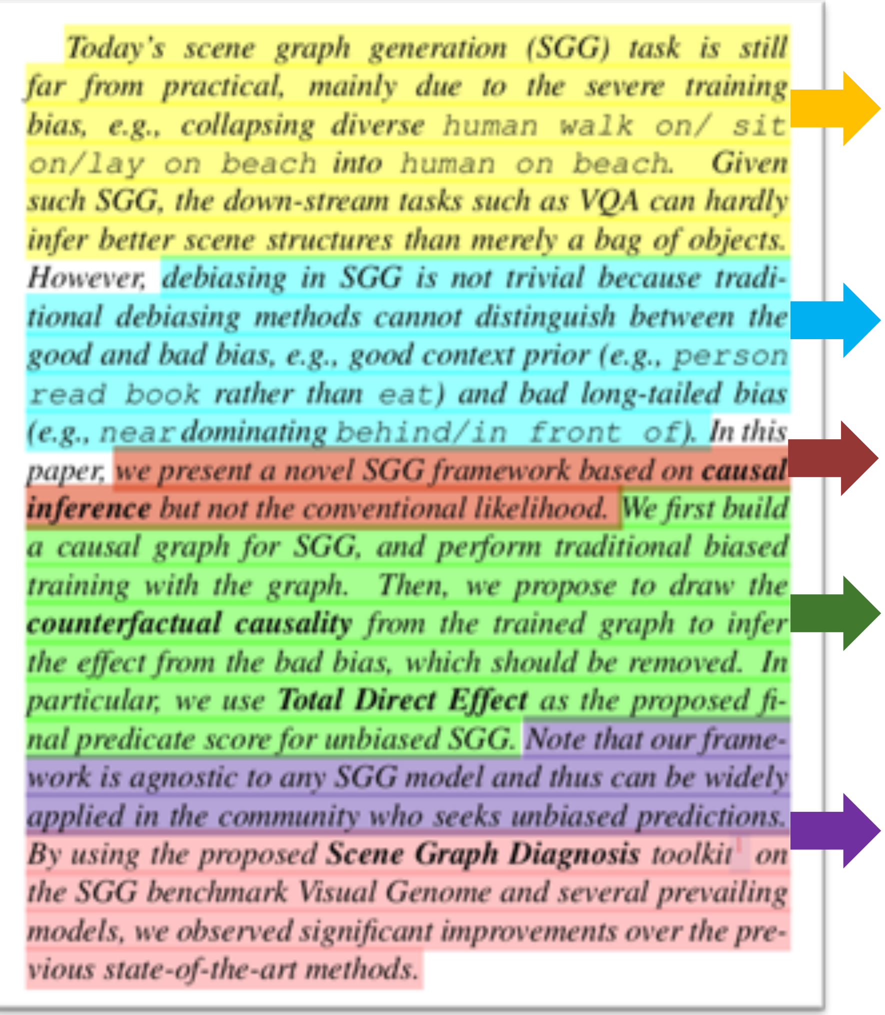
(b) Distribution of Predicate Sampling Fraction

(c) Biased Generation

(d) Unbiased Generation

一篇论文的组成—摘要

Unbiased Scene Graph Generation from Biased Training
2020 CVPR



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*** *toolkit 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

本文的具体方法：

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）

- 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

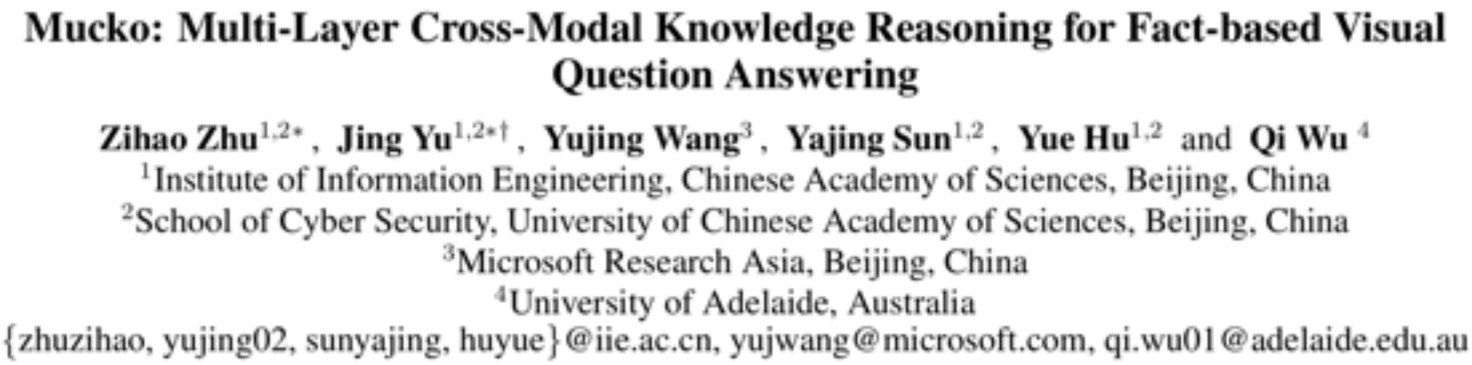
- 问题-方法-实验，相互呼应
  - 问题：有理有据，足够具体
    - 背景阐述聚焦重点
    - 问题提出明确具体
    - 聚焦研究动机，总结现状问题
    - 基于研究动机，概述研究方法
    - 面向领域需求，拔高论文贡献
  - 方法：针对问题设计，每一步设计目标明确
  - 实验：针对方法逐一证明，针对动机逐一分析

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

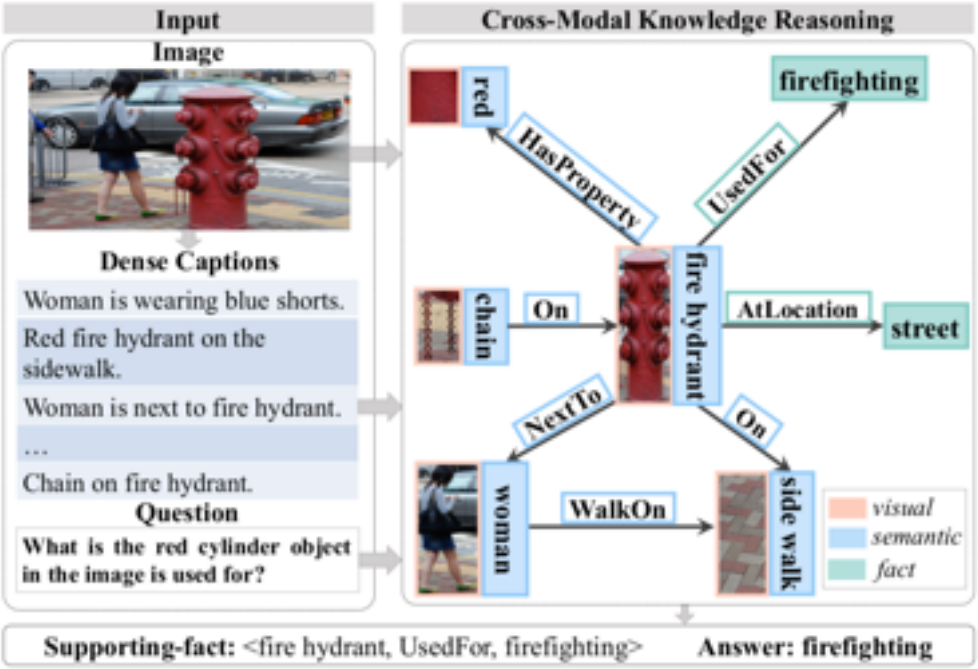
一篇论文的组成——引言

2020 IJCAI



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

Zihao Zhu[1,2*], **Jing Yu**[1,2*†], **Yujing Wang**[3], **Yajing Sun**[1,2], **Yue Hu**[1,2] and **Qi Wu**[4]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Microsoft Research Asia, Beijing, China
[4]University of Adelaide, Australia
{zhuzihao, yujing02, sunyajing, huyue}@iie.ac.cn, yujwang@microsoft.com, qi.wu01@adelaide.edu.au



Woman is wearing blue shorts.

Red fire hydrant on the sidewalk.

Woman is next to fire hydrant.

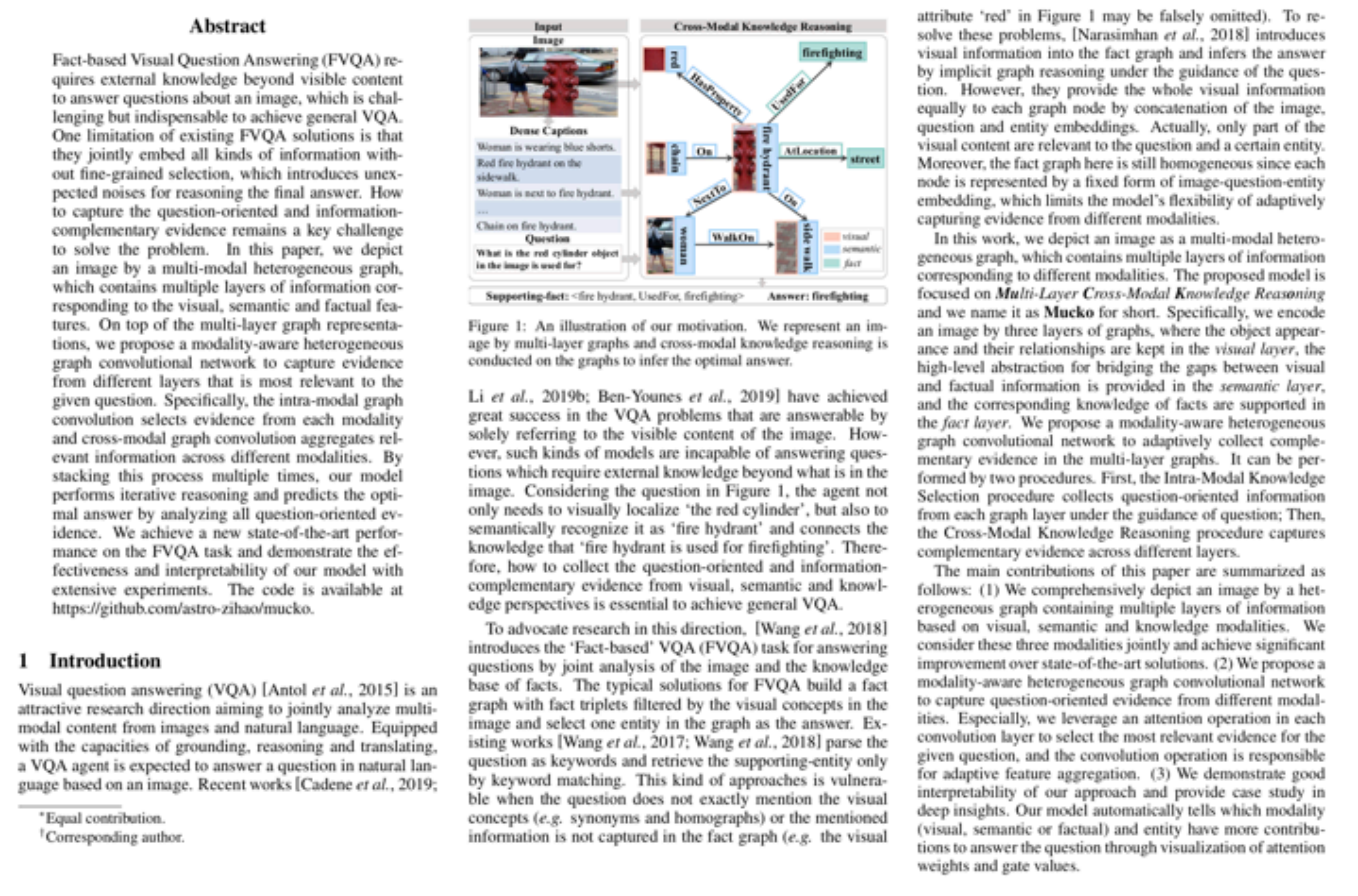
...

Chain on fire hydrant.

What is the red cylinder object in the image is used for?

Supporting-fact: <fire hydrant, UsedFor, firefighting> **Answer: firefighting**

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 multimodal 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; 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.

Input
Image
Dense Captions
Woman is wearing blue shorts.
Red fire hydrant on the sidewalk.
Woman is next to fire hydrant.
....
Chain on fire hydrant.
Question
What is the red cylinder object in the image is used for?
Cross-Modal Knowledge Reasoning
red
HasProperty
UsedFor
firefighting
chain
On
fire hydrant
AtLocation
street
NextTo
On
woman
WalkOn
side walk
visual
semantic
fact
Supporting-fact: <fire hydrant, UsedFor, firefighting>
Answer: firefighting

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.

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 ***Mu**lti-Layer **C**ross-Modal **K**nowledge Reas**o**ning* 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.

[*] Equal contribution.
[†] Corresponding author.

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 multimodal 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

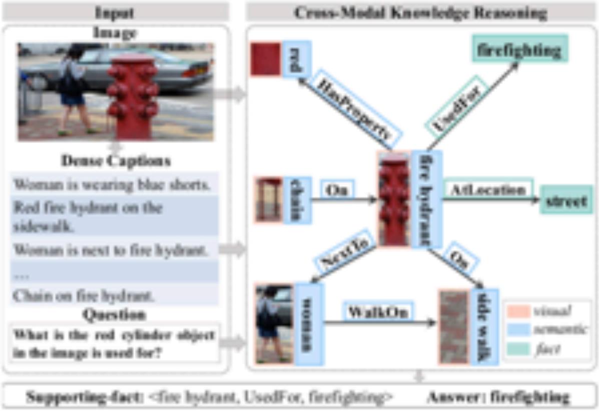

Input
Image
Dense Captions
Woman is wearing blue shorts.
Red fire hydrant on the sidewalk.
Woman is next to fire hydrant.
...
Chain on fire hydrant.
Question
What is the red cylinder object in the image is used for?
Cross-Modal Knowledge Reasoning
red
HasProperty
firefighting
UsedFor
chain
On
fire hydrant
AtLocation
street
NextTo
On
woman
WalkOn
side walk
visual
semantic
fact
Supporting-fact: <fire hydrant, UsedFor, firefighting>
Answer: firefighting

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 r

tion. However, th

equally to each gra

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.

现有方法问题：分类归纳

从具体方法归纳问题

In

gene

corresponding to different modalities. The proposed model is focused on ***M**ulti-Layer **C**ross-Modal **K**nowledge Reas**o**ning* 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 Cro

complem

方法介绍：呼应挑战和问题

The m

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.

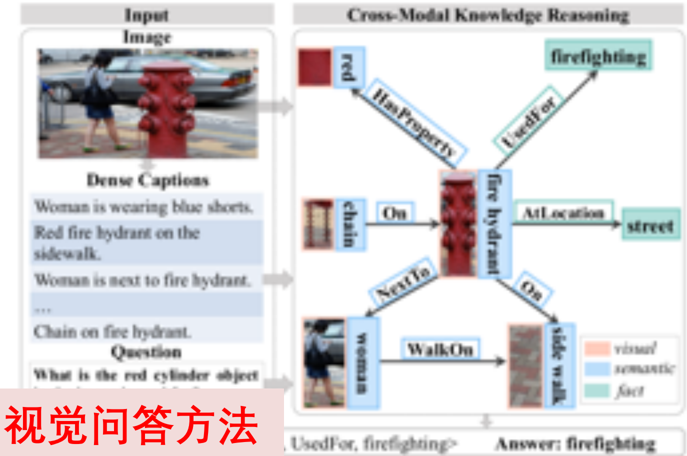
贡献：凝练方法的普适性

一篇论文的组成——相关工作

基本要求

- 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

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 decipt 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.

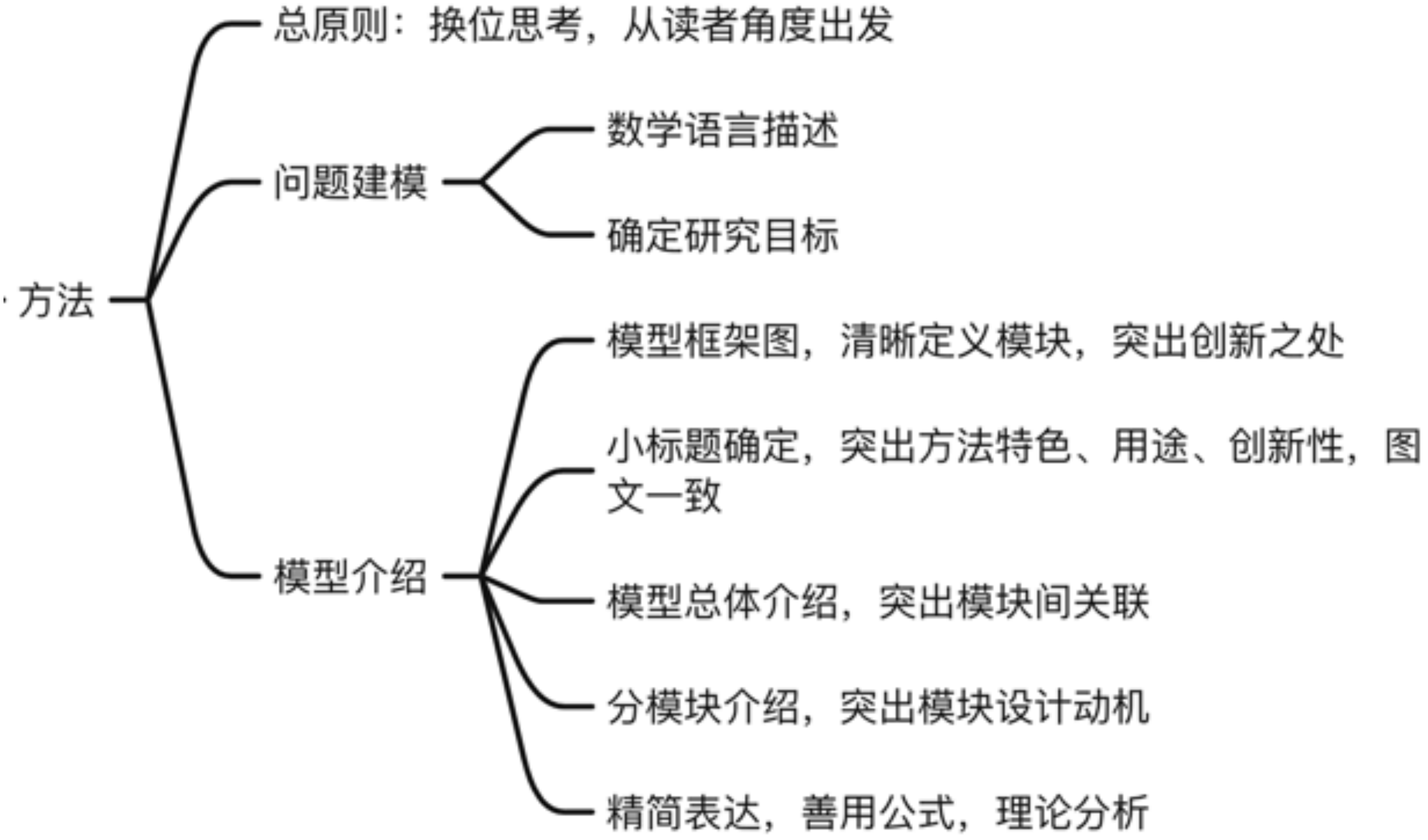
异构图神经网络方法

Heterogeneous Graph Neural Networks. Graph neu- ntum in the last few d 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.

CogModal
GROUP

一篇论文的组成——方法

基本要求



- 方法
  - 总原则：换位思考，从读者角度出发
  - 问题建模
    - 数学语言描述
    - 确定研究目标
  - 模型介绍
    - 模型框架图，清晰定义模块，突出创新之处
    - 小标题确定，突出方法特色、用途、创新性，图文一致
    - 模型总体介绍，突出模块间关联
    - 分模块介绍，突出模块设计动机
    - 精简表达，善用公式，理论分析

CogModal
GROUP

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

CogModal
GROUP

一篇论文的组成——方法

2020 IJCAI

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

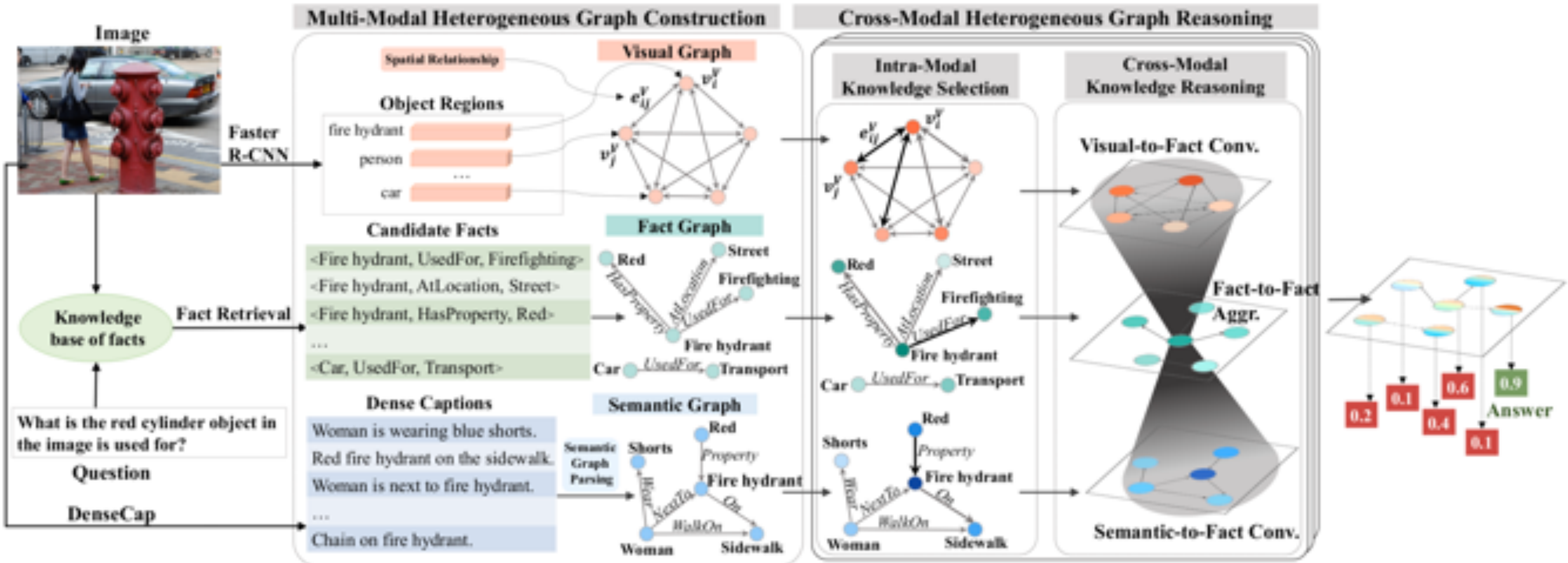
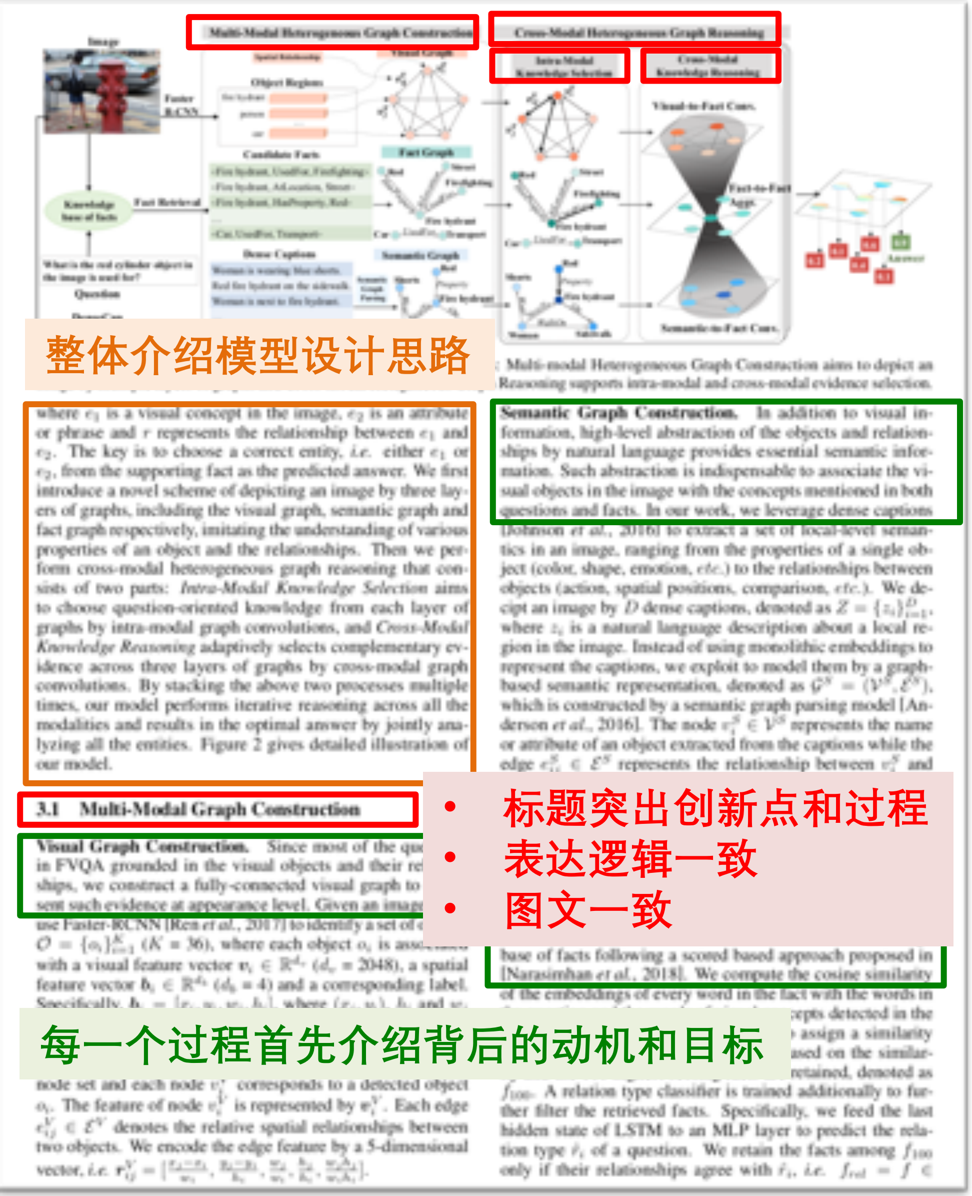


Figure 2: An overview of our model. The model contains two modules: Multi-modal Heterogeneous Graph Construction aims to depict an image by multiple layers of graphs and Cross-modal Hetegeneous Graph Reasoning supports intra-modal and cross-modal evidence selection.

CogModal GROUP

一篇论文的组成——方法



Multi-modal Heterogeneous Graph Construction aims to depict an … Reasoning supports intra-modal and cross-modal evidence selection.

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 que in FVQA grounded in the visual objects and their re ships, we construct a fully-connected visual graph to sent such evidence at appearance level. Given an image use Faster-RCNN [Ren *et al.*, 2017] to identify a set of o $O = \{o_i\}_{i=1}^{K}$ ($K = 36$), where each object o_i is asso 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), h_i and w_i … 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.* $r_{ij}^V = [\frac{x_j - x_i}{w_i}, \frac{y_j - y_i}{h_i}, \frac{w_j}{w_i}, \frac{h_j}{h_i}, \frac{w_j h_j}{w_i h_i}]$.

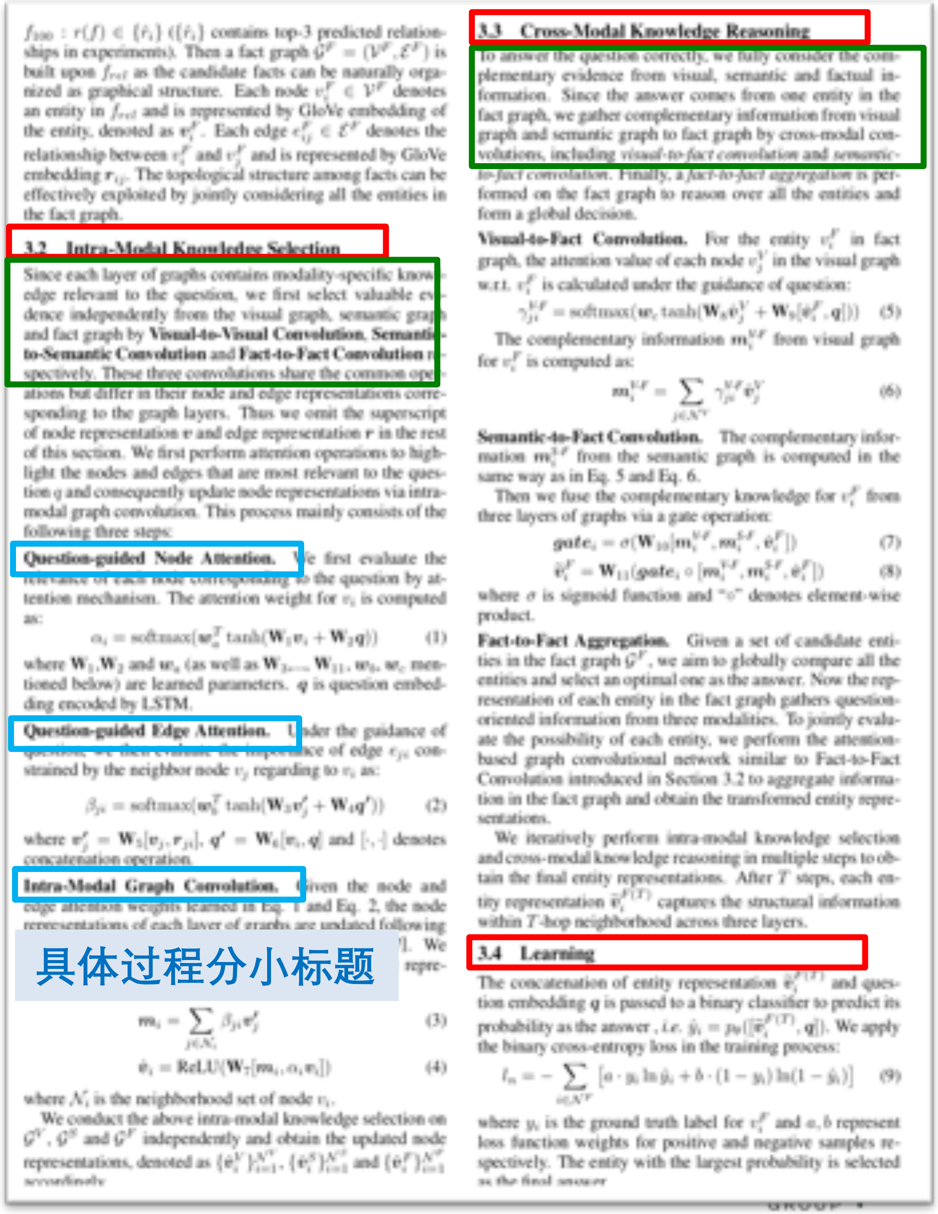
Semantic Graph Construction. In addition to visual information, high-level abstraction of the objects and relationships by natural language provides essential semantic information. Such abstraction is indispensable to associate the visual objects in the image with the concepts mentioned in both questions 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

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 … cepts detected in the … assign a similarity … ased on the similar… retained, 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 to 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.* $f_{rel} = f \in$

整体介绍模型设计思路

- 标题突出创新点和过程
- 表达逻辑一致
- 图文一致

每一个过程首先介绍背后的动机和目标



$f_{100} : r(f) \in \{\hat{r}_i\}$ ($\{\hat{r}_i\}$ 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 $W_3, \ldots, W_{11}, w_b, w_c$ 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 … We … repre-

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

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

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 $\{\hat{v}_i^V\}_{i=1}^{\mathcal{N}^V}$, $\{\hat{v}_i^S\}_{i=1}^{\mathcal{N}^S}$ and $\{\hat{v}_i^F\}_{i=1}^{\mathcal{N}^F}$ 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:

$$\gamma_{ji}^{V\text{-}F} = \text{softmax}(w_c \tanh(W_8 \hat{v}_j^V + W_9[\hat{v}_i^F, q])) \quad (5)$$

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

$$m_i^{V\text{-}F} = \sum_{j \in \mathcal{N}^V} \gamma_{ji}^{V\text{-}F} \hat{v}_j^V \quad (6)$$

Semantic-to-Fact Convolution. The complementary information $m_i^{S\text{-}F}$ 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:

$$gate_i = \sigma(W_{10}[m_i^{V\text{-}F}, m_i^{S\text{-}F}, \hat{v}_i^F]) \quad (7)$$

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

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:

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

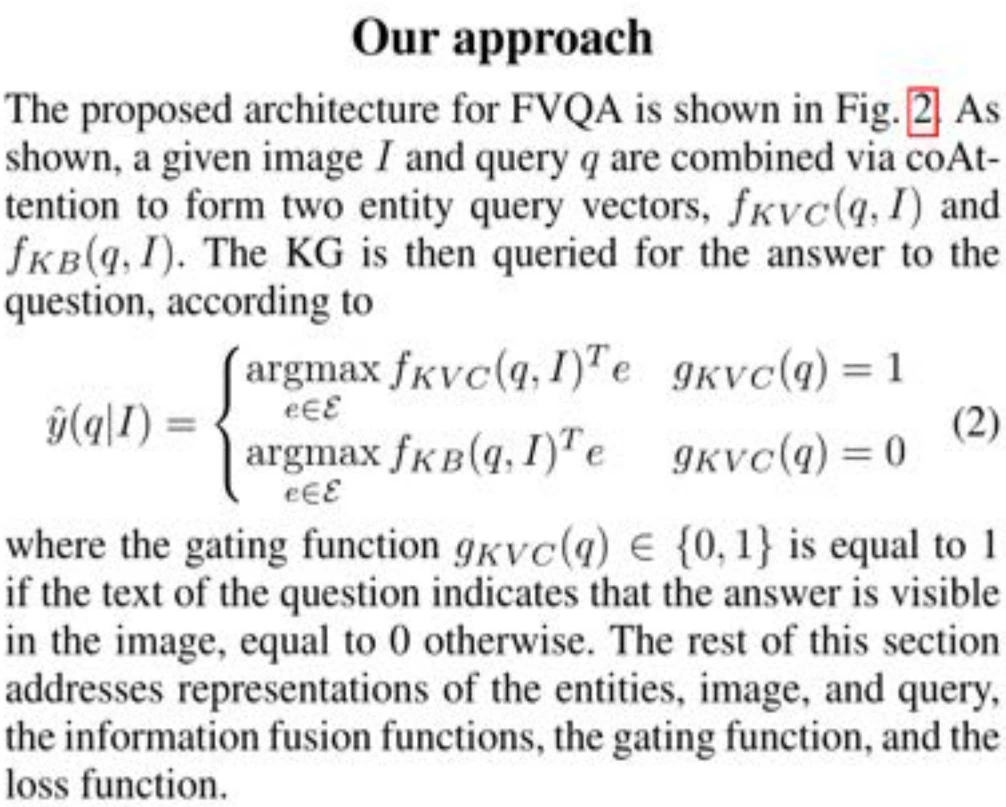
where y_i is the ground truth label for v_i^F and a, b represent loss function weights for positive and negative samples respectively. The entity with the largest probability is selected as the final answer.

具体过程分小标题

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

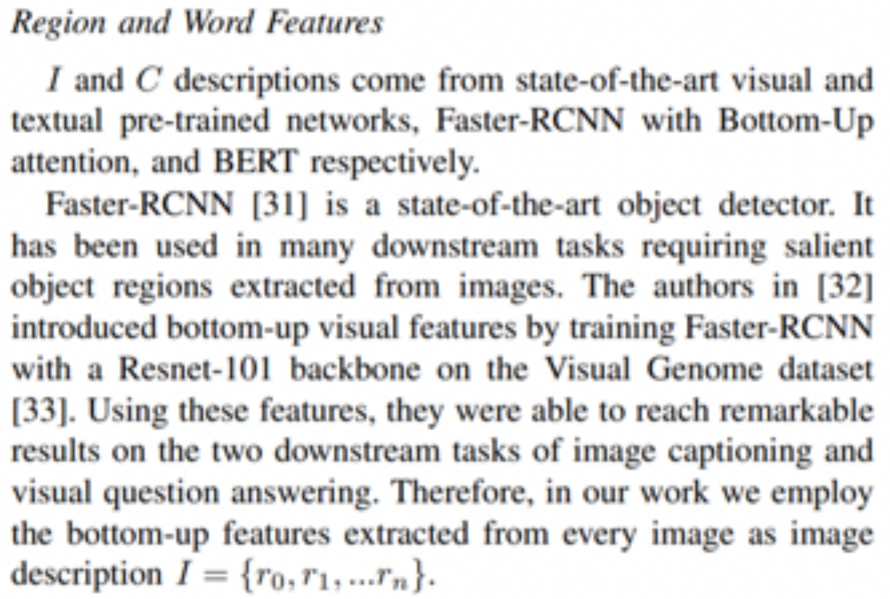
The proposed architecture for FVQA is shown in Fig. 2. As shown, a given image I and query q are combined via coAttention to form two entity query vectors, $f_{KVC}(q, I)$ and $f_{KB}(q, I)$. The KG is then queried for the answer to the question, according to

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

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.

一篇论文的组成——方法(实例分析)

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

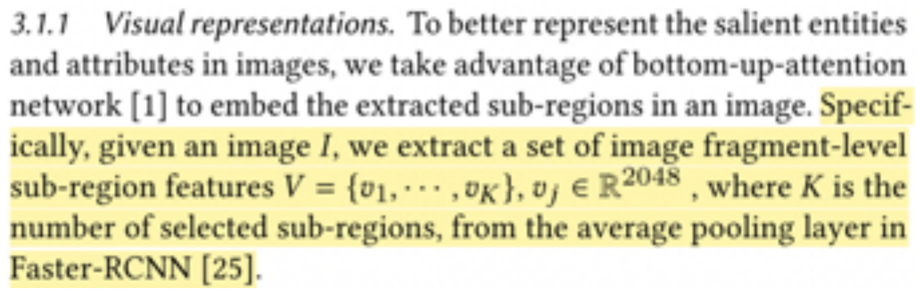


Region and Word Features

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 $I = \{r_0, r_1, ...r_n\}$.

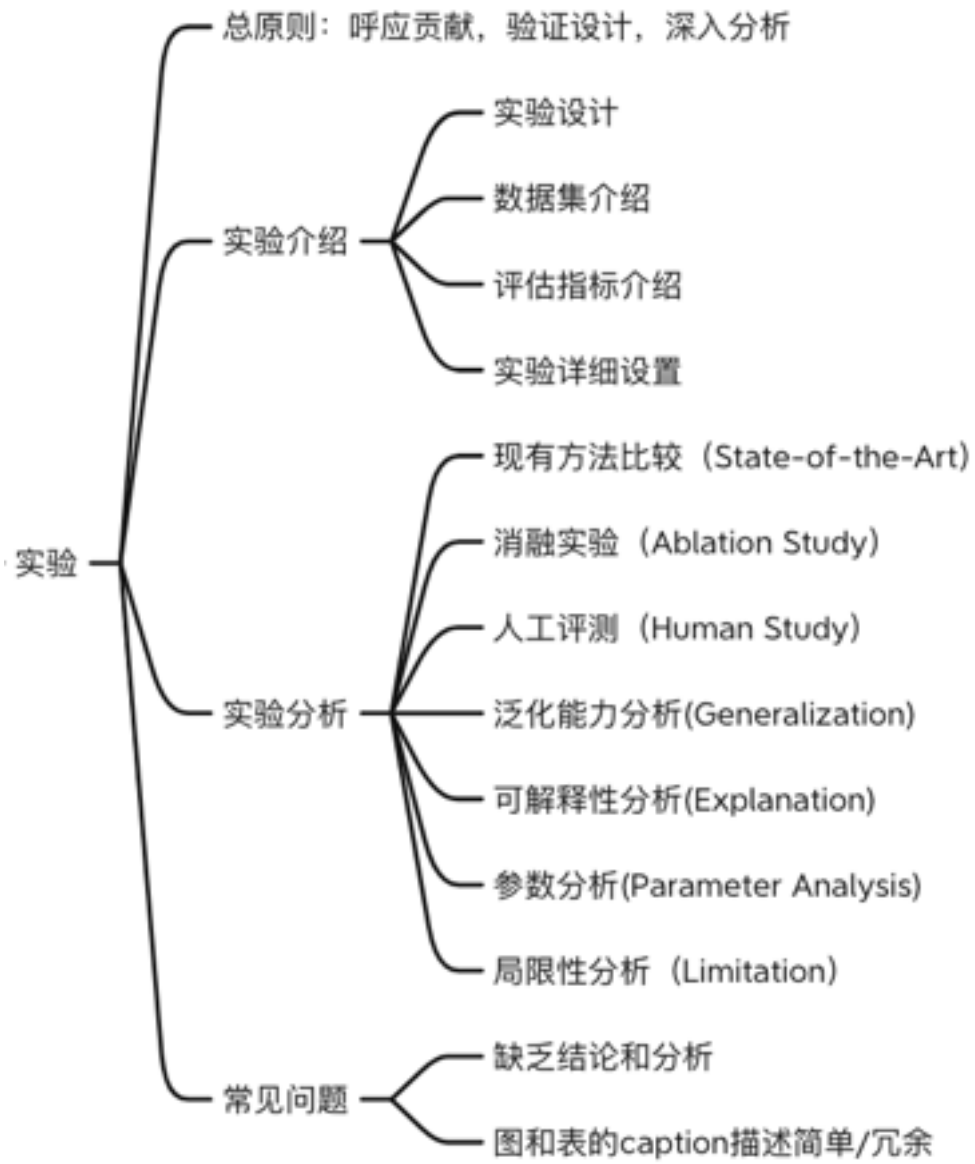
Structured Multi-modal Feature Embedding and Alignment for Image-Sentence Retrieval
2020 ACM MM



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 $V = \{v_1, \cdots, v_K\}, v_j \in \mathbb{R}^{2048}$, where K is the number of selected sub-regions, from the average pooling layer in Faster-RCNN [25].

一篇论文的组成——实验

基本要求



- 实验
  - 总原则：呼应贡献，验证设计，深入分析
  - 实验介绍
    - 实验设计
    - 数据集介绍
    - 评估指标介绍
    - 实验详细设置
  - 实验分析
    - 现有方法比较（State-of-the-Art）
    - 消融实验（Ablation Study）
    - 人工评测（Human Study）
    - 泛化能力分析(Generalization)
    - 可解释性分析(Explanation)
    - 参数分析(Parameter Analysis)
    - 局限性分析（Limitation）
  - 常见问题
    - 缺乏结论和分析
    - 图和表的caption描述简单/冗余

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

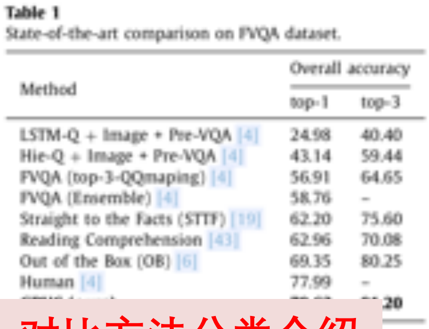


Table 1
State-of-the-art comparison on FVQA dataset.

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-QQmaping) [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	–
[illegible]	[illegible]	**[illegible]20**

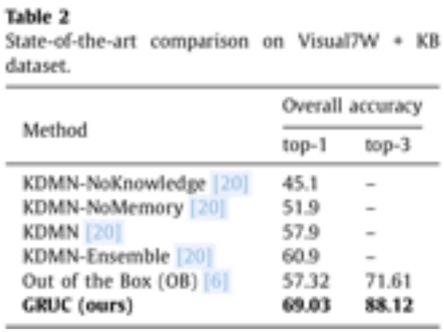


Table 2
State-of-the-art comparison on Visual7W + KB dataset.

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**

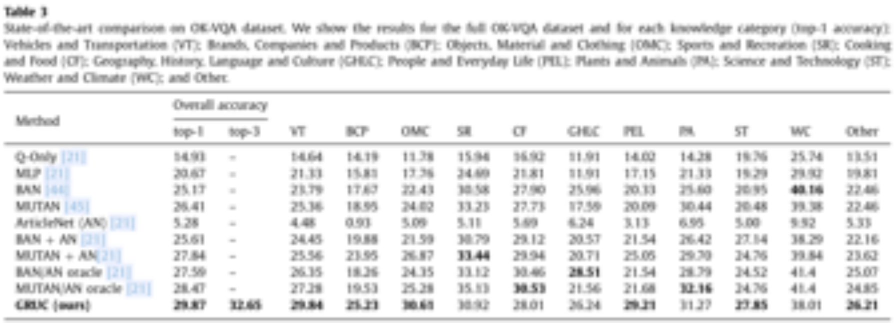
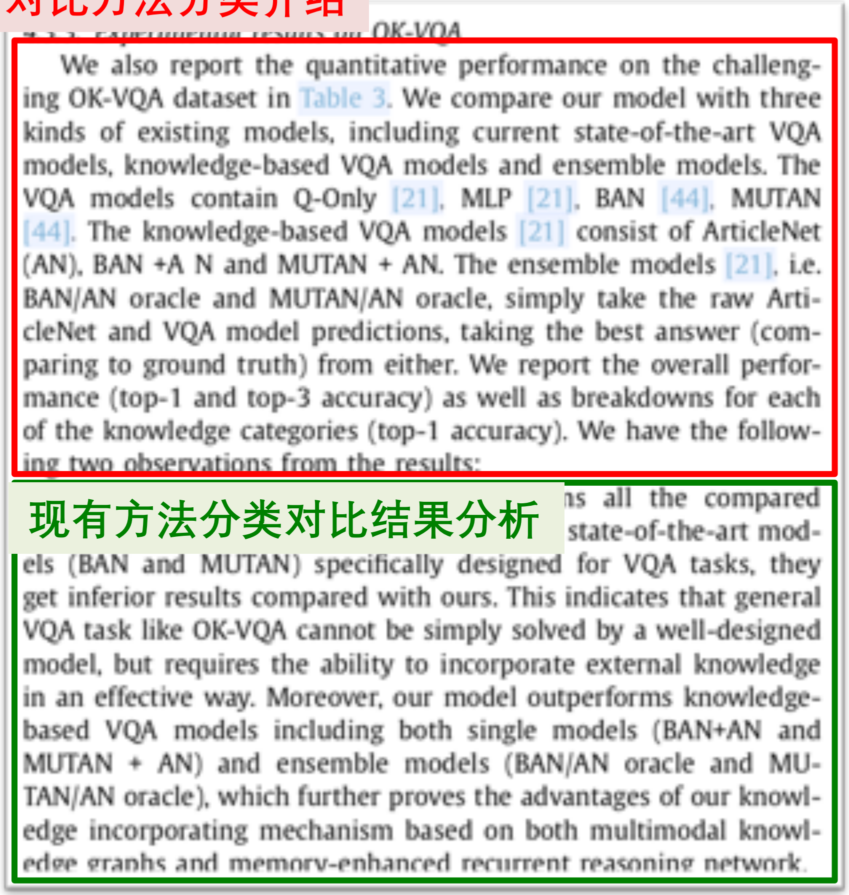


Table 3
State-of-the-art comparison on OK-VQA dataset. We show the results for the full OK-VQA dataset and for each knowledge category (top-1 accuracy): Vehicles and Transportation (VT); Brands, Companies and Products (BCP); Objects, Material and Clothing (OMC); Sports and Recreation (SR); Cooking and 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**

对比方法分类介绍



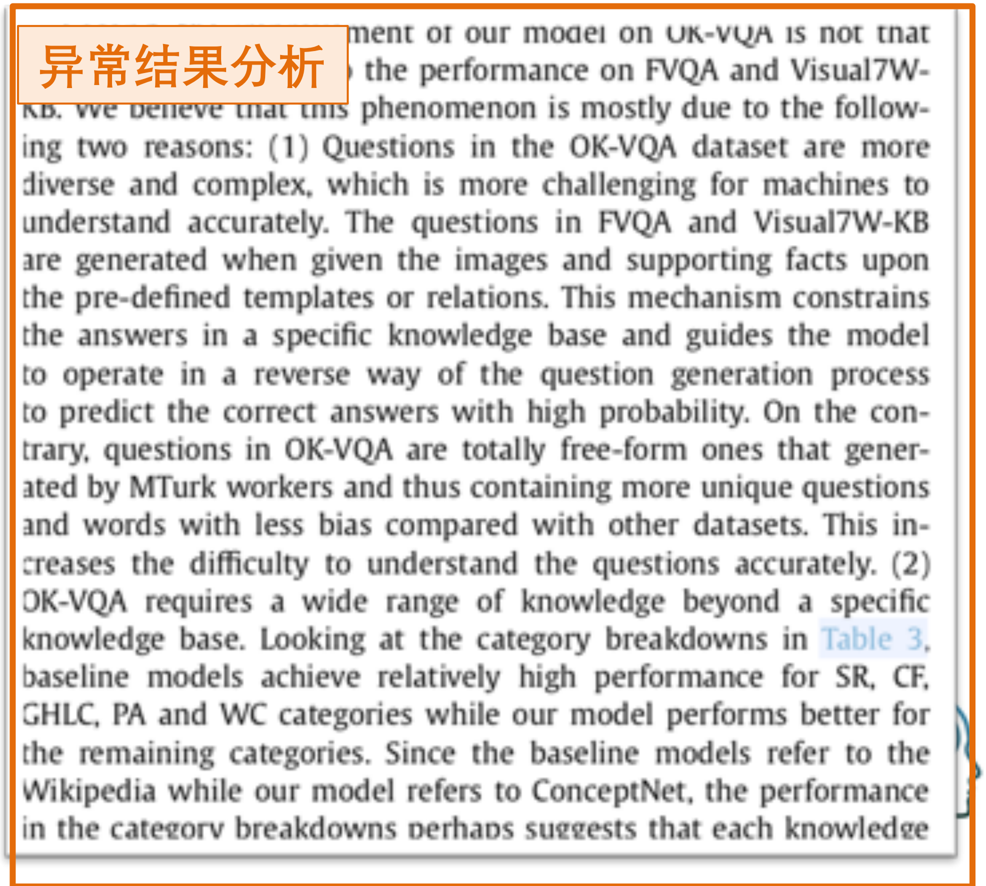
4.3.3. Experimental results on OK-VQA

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 +A N 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

ns 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.

异常结果分析



ment 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

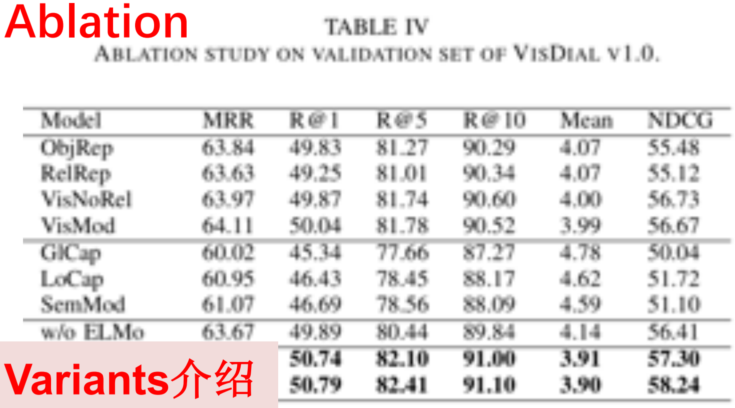
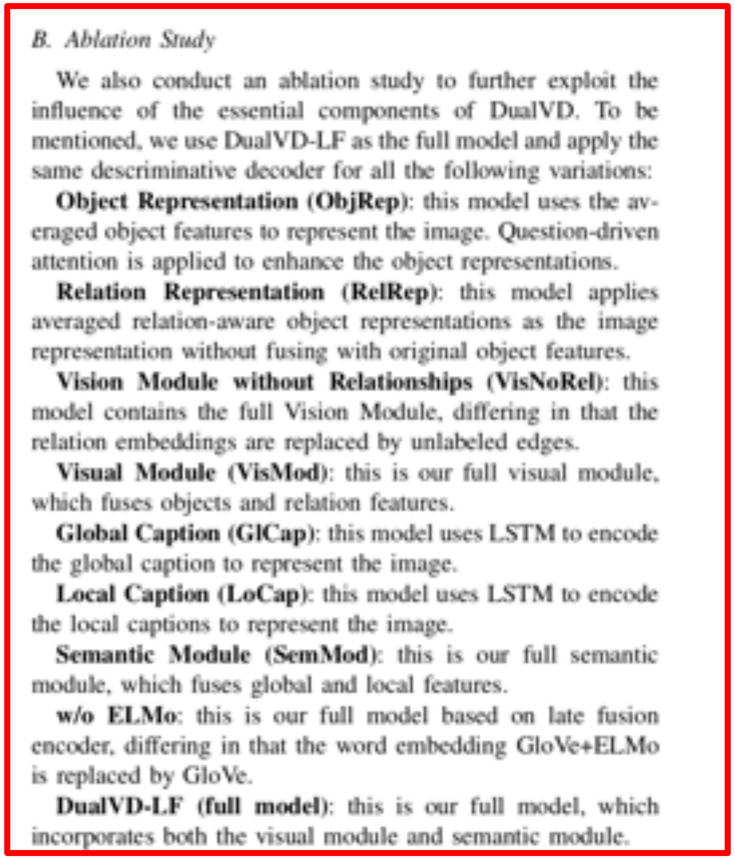


TABLE IV

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 descriminative 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.

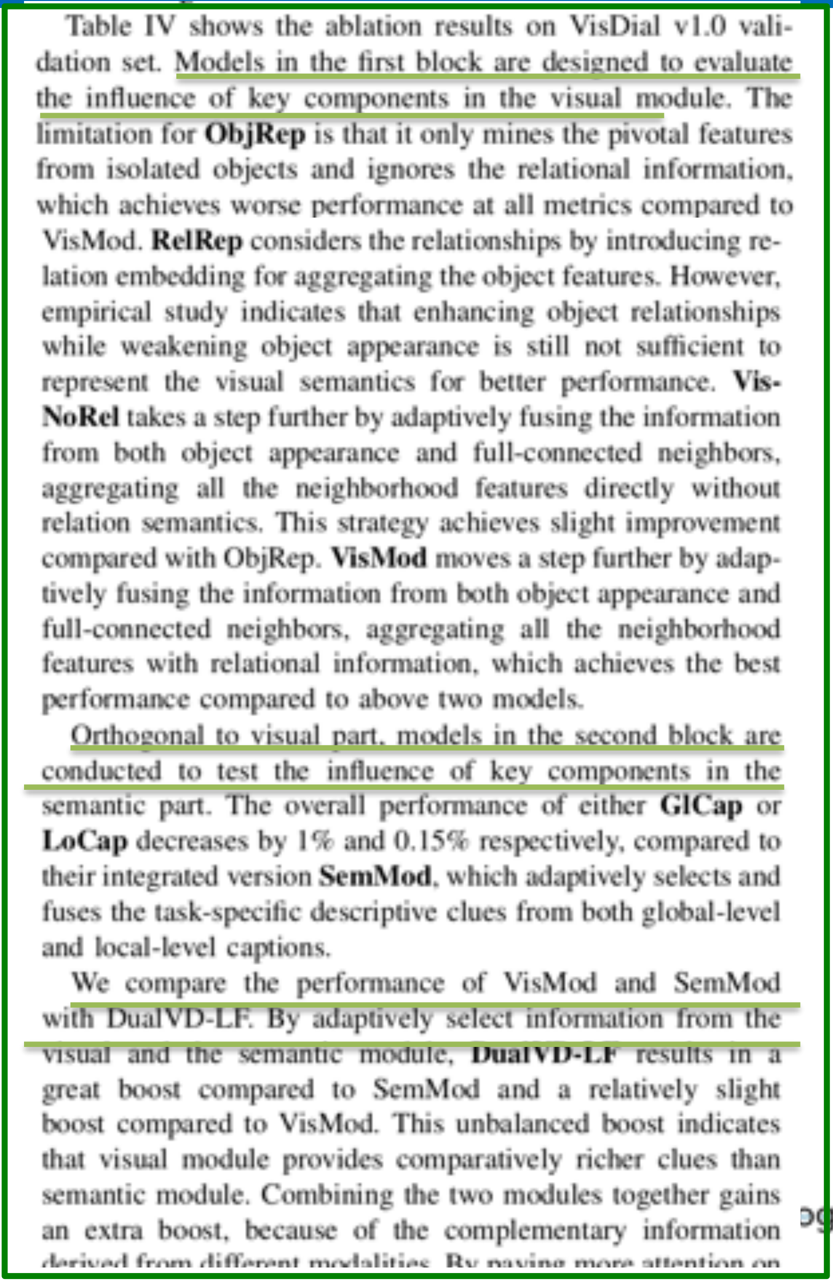


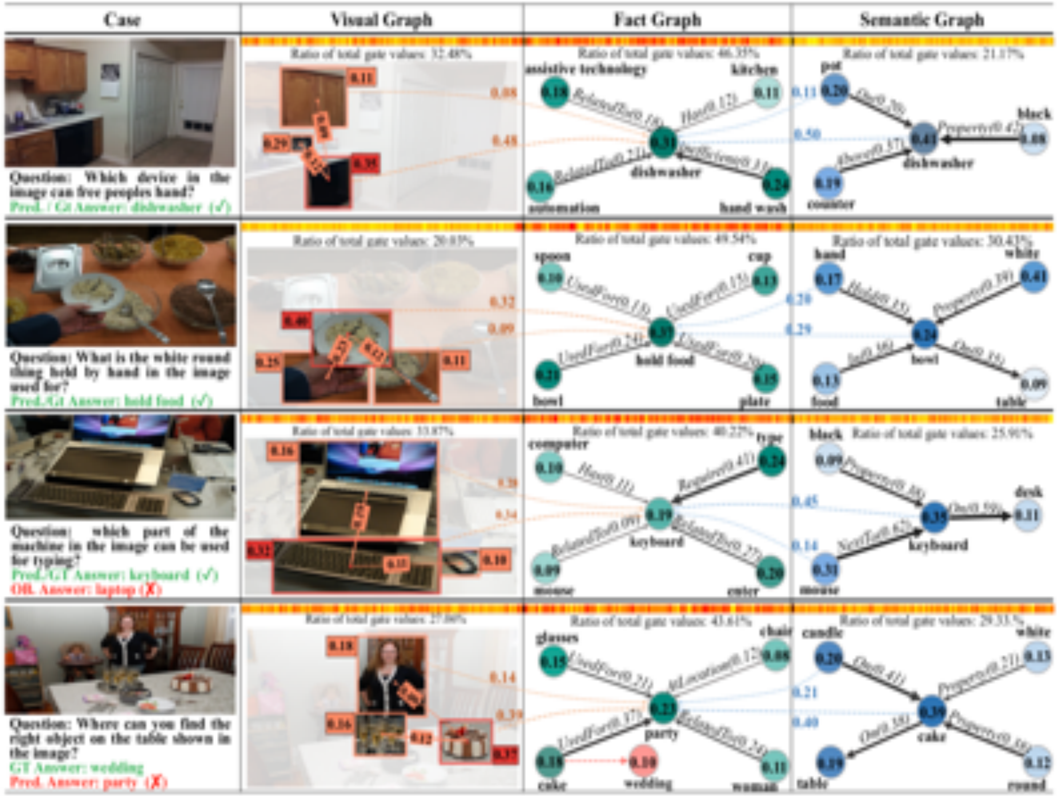
Table IV shows the ablation results on VisDial v1.0 validation set. Models in the first block are designed to evaluate the influence of key components in the visual module. The limitation for **ObjRep** is that it only mines the pivotal features from isolated objects and ignores the relational information, 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.

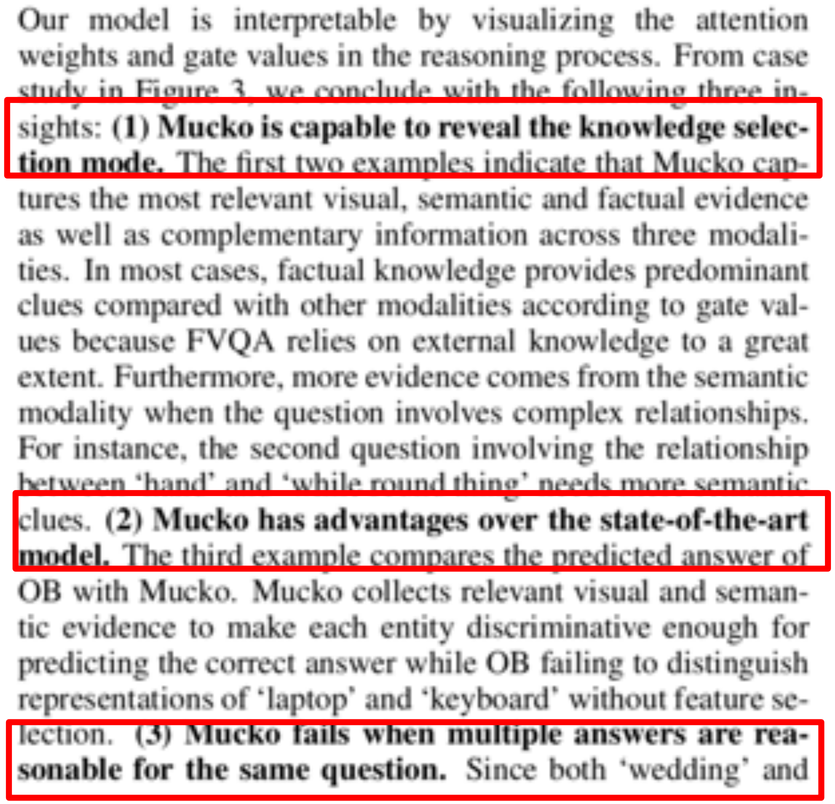
We compare the performance of VisMod and SemMod with DualVD-LF. By adaptively select information from the visual and the semantic module, **DualVD-LF** results in a great boost compared to SemMod and a relatively slight boost compared to VisMod. This unbalanced boost indicates that visual module provides 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



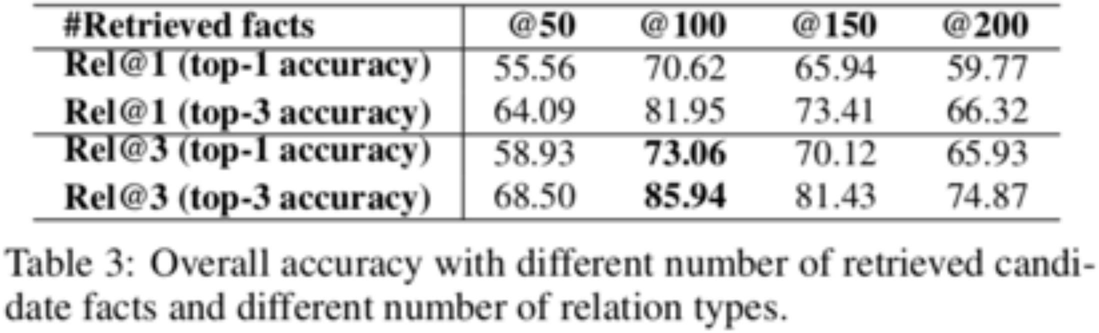
Case
Visual Graph
Fact Graph
Semantic Graph
Question: Which device in the image can free peoples hand?
Pred. / Gt Answer: dishwasher (√)
Question: What is the white round thing held by hand in the image used for?
Pred./Gt Answer: hold food (√)
Question: which part of the machine in the image can be used for typing?
Pred./GT Answer: keyboard (√)
OB. Answer: laptop (✗)
Question: Where can you find the right object on the table shown in the image?
GT Answer: wedding
Pred. Answer: party (✗)
dishwasher
kitchen
hand wash
automation
counter
spoon
cup
hold food
bowl
plate
food
table
computer
keyboard
mouse
enter
desk
glasses
chair
party
cake
wedding
woman
candle
round



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

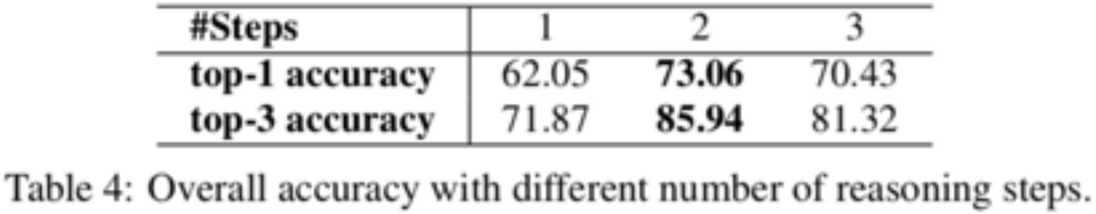
一篇论文的组成——实验

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.

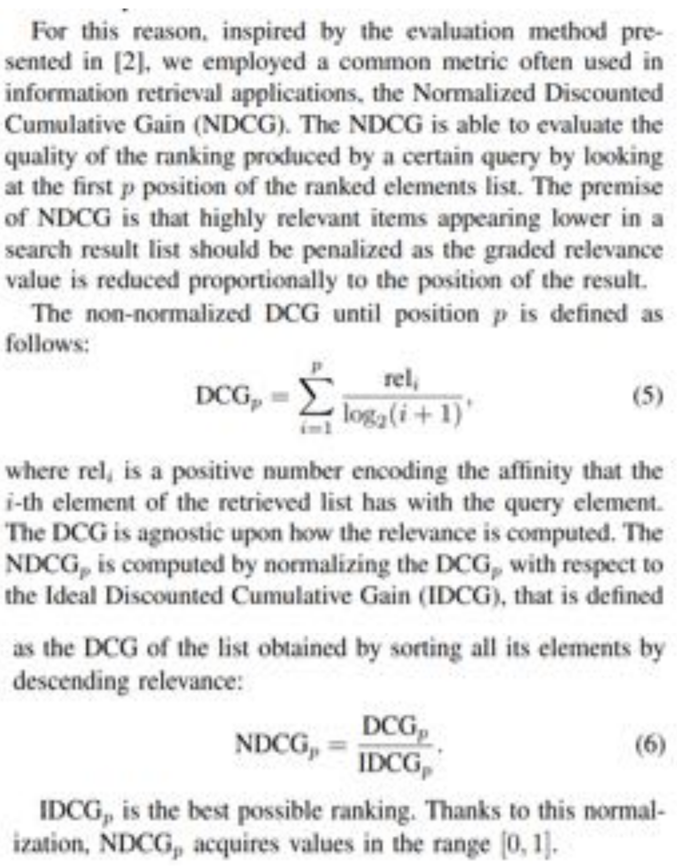
其他需要验证的实验

- 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:

$$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{\mathrm{rel}_i}{\log_2(i+1)}, \qquad (5)$$

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:

$$\mathrm{NDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG}_p}. \qquad (6)$$

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

SOTA实验

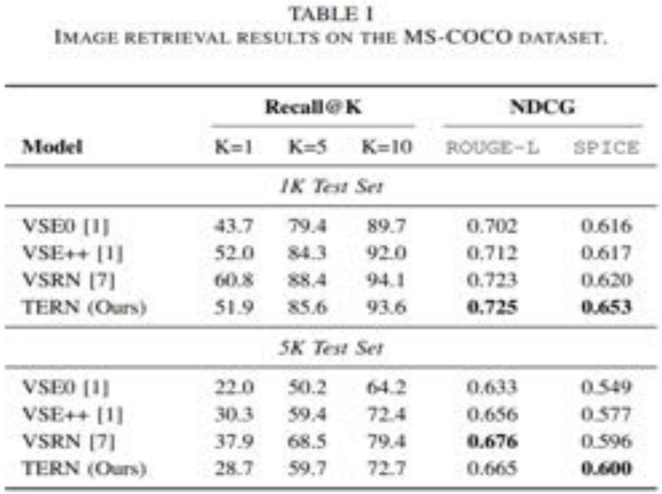


TABLE I
IMAGE RETRIEVAL RESULTS ON THE MS-COCO DATASET.

	Recall@K			NDCG	
Model	K=1	K=5	K=10	ROUGE-L	SPICE
		1K Test Set			
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**
		5K Test Set			
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

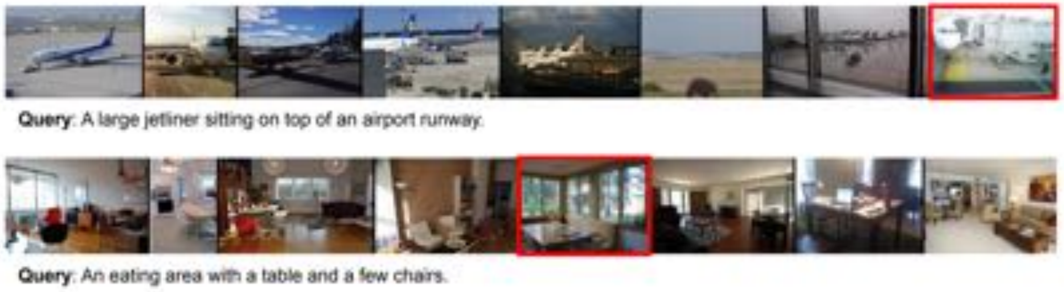

Query: A large jetliner sitting on top of an airport runway.
Query: An eating area with a table and a few chairs.

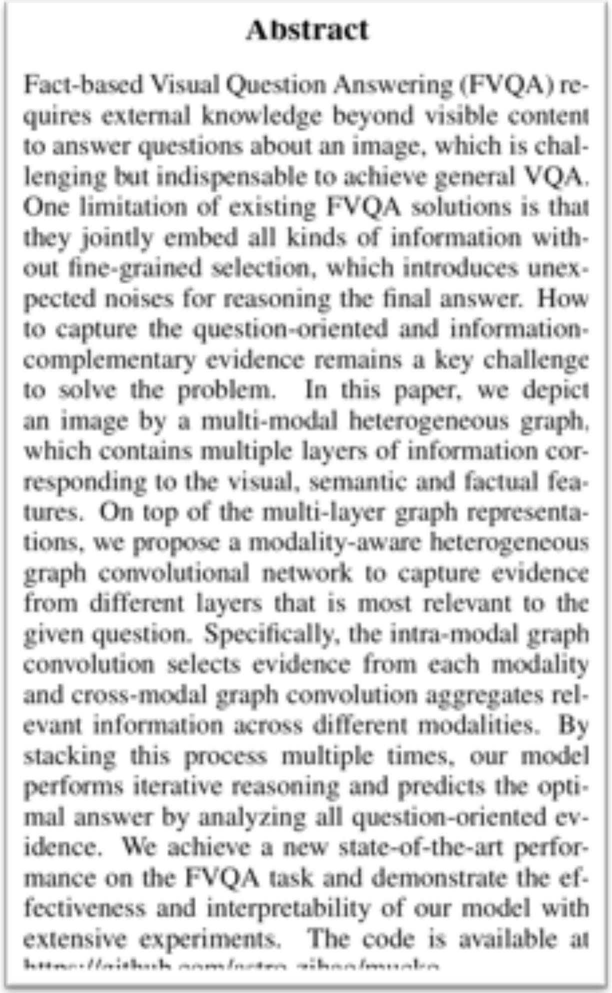
Fig. 4. Example of image retrieval results for a couple of query captions. The red marked images represent the MS-COCO ground truths, and they are not necessarily the best results in these scenarios. In fact, in the very first positions, we find non-matching yet relevant images. These are common examples 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
- 介绍未来工作

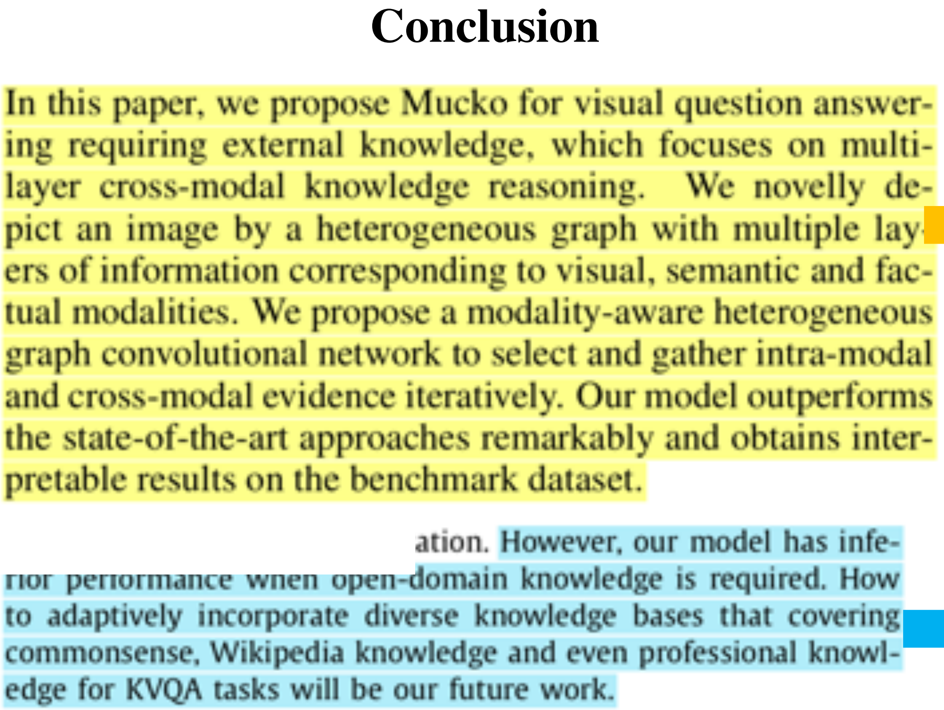
一篇论文的组成——结论



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/astro-zihao/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.

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.

总结工作
体现效果

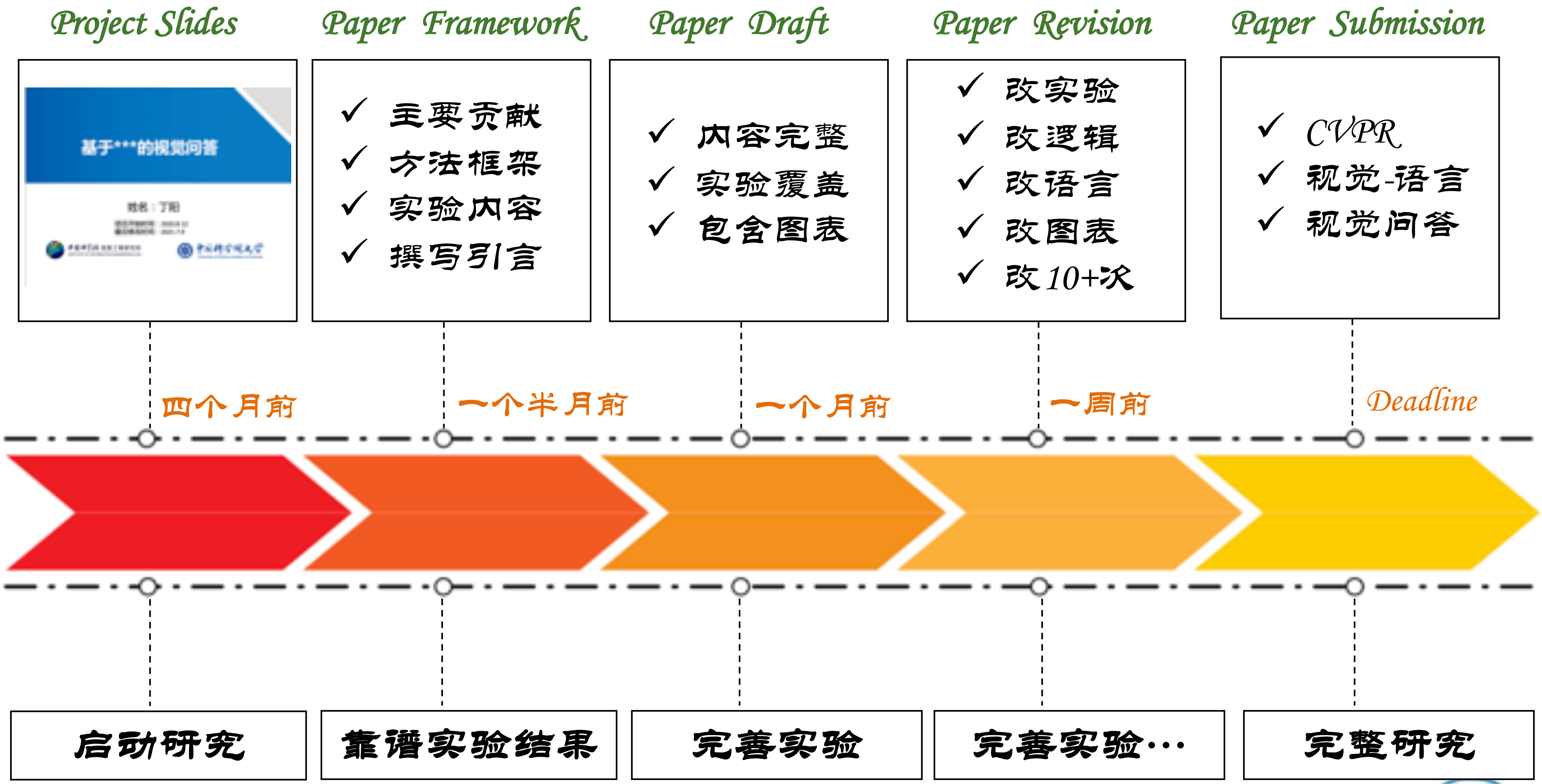
说明局限
指明方向

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



Project Slides

基于***的视觉问答

姓名：丁阳

四个月前

启动研究

Paper Framework

- ✓ 主要贡献
- ✓ 方法框架
- ✓ 实验内容
- ✓ 撰写引言

一个半月前

靠谱实验结果

Paper Draft

- ✓ 内容完整
- ✓ 实验覆盖
- ✓ 包含图表

一个月前

完善实验

Paper Revision

- ✓ 改实验
- ✓ 改逻辑
- ✓ 改语言
- ✓ 改图表
- ✓ 改10+次

一周前

完善实验…

Paper Submission

- ✓ CVPR
- ✓ 视觉-语言
- ✓ 视觉问答

Deadline

完整研究

会议论文如何扩展为期刊论文？

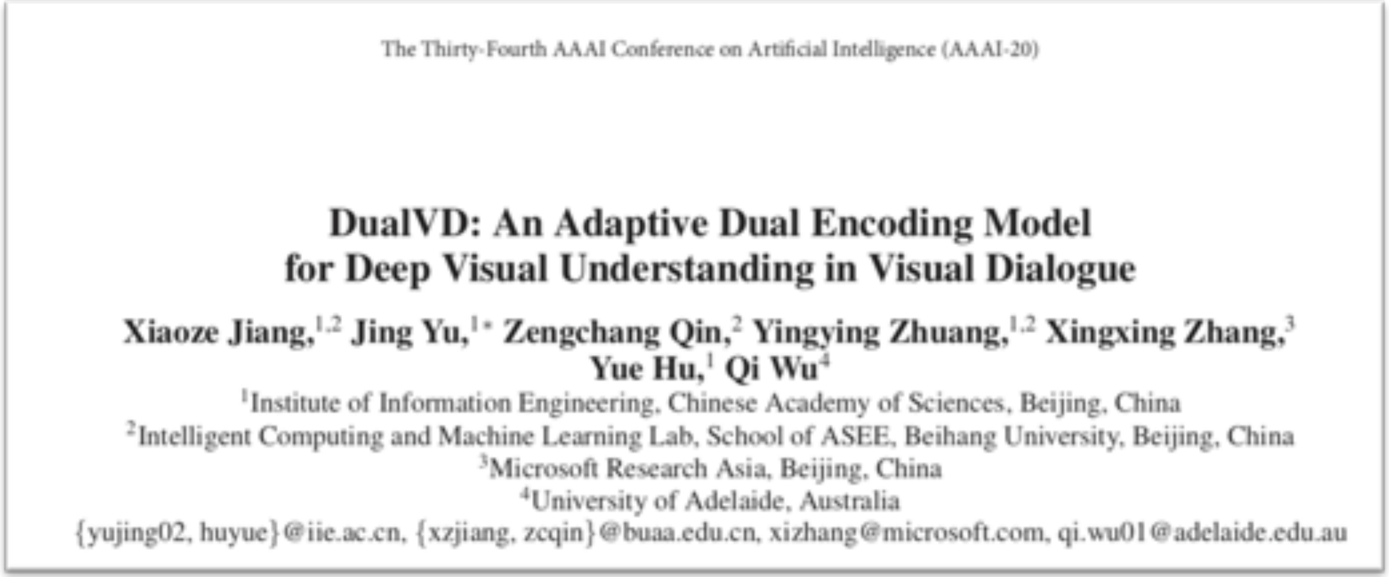
基本思路

- 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

CogModal
GROUP

看一个例子

会议论文-*Title*

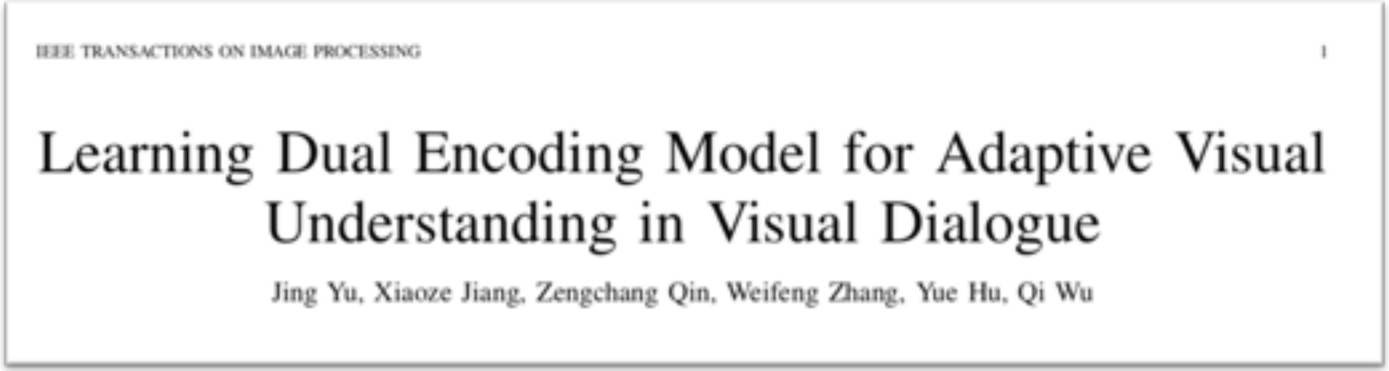


The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

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 Wu[4]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]Intelligent Computing and Machine Learning Lab, School of ASEE, Beihang University, Beijing, China
[3]Microsoft Research Asia, Beijing, China
[4]University of Adelaide, Australia
{yujing02, huyue}@iie.ac.cn, {xzjiang, zcqin}@buaa.edu.cn, xizhang@microsoft.com, qi.wu01@adelaide.edu.au

期刊论文-*Title*



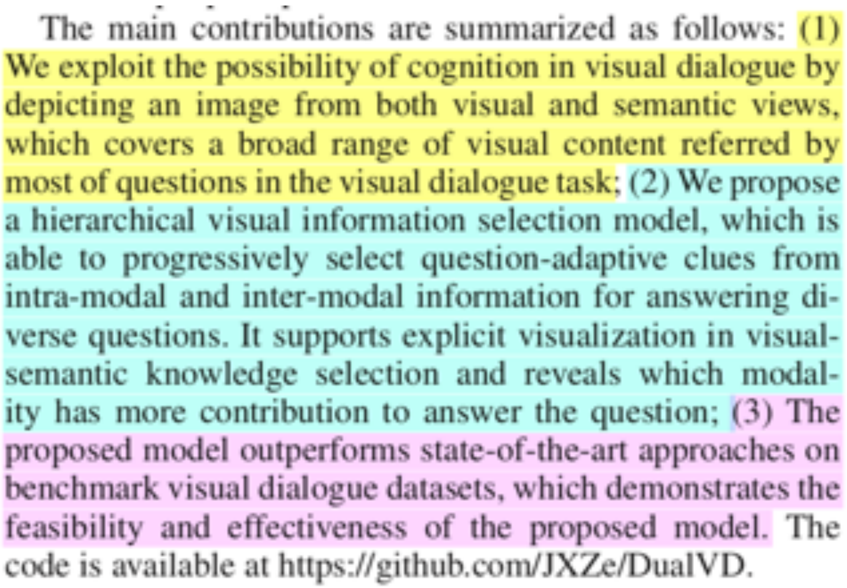
IEEE TRANSACTIONS ON IMAGE PROCESSING 1

Learning Dual Encoding Model for Adaptive Visual Understanding in Visual Dialogue

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

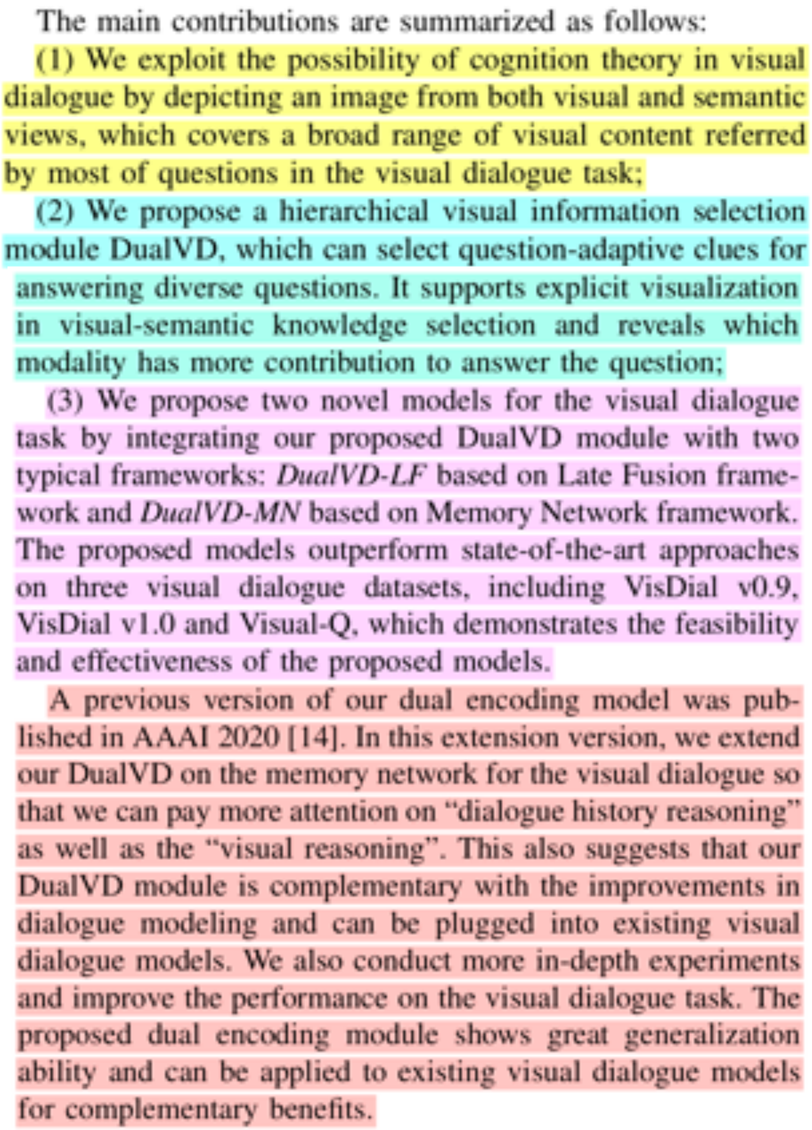
看一个例子

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;

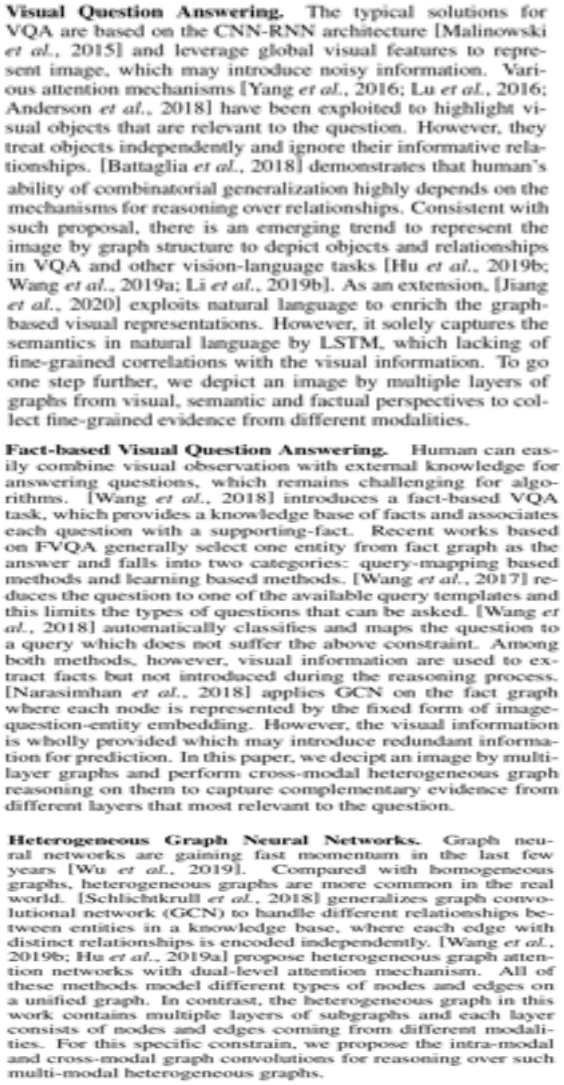
(2) We propose a hierarchical visual information selection module DualVD, which can select question-adaptive clues for answering diverse questions. It supports explicit visualization in visual-semantic knowledge selection and reveals which modality has more contribution to answer the question;

(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.

看一个例子

会议论文-Related Work

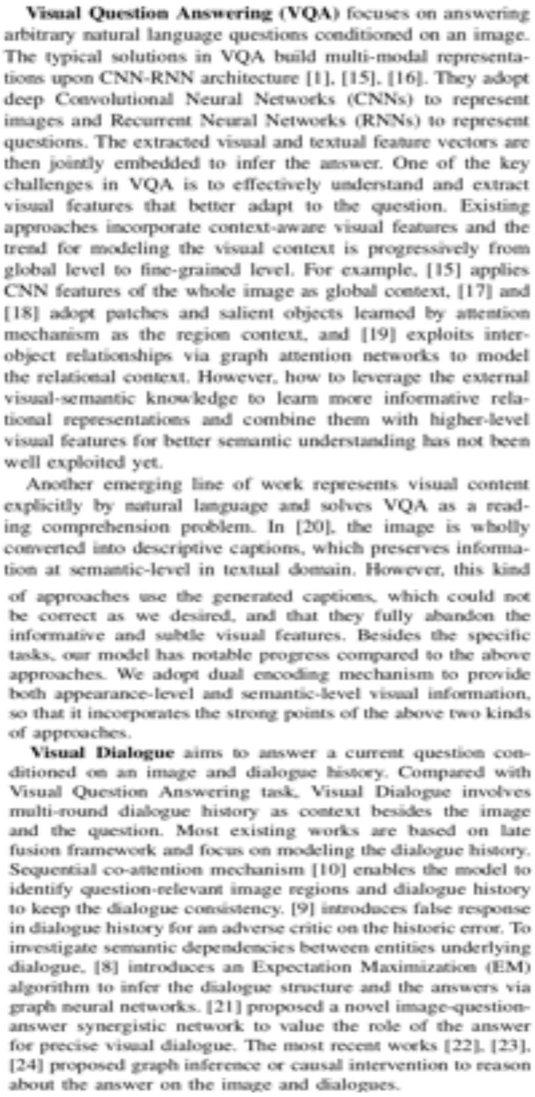


Visual Question Answering. The typical solutions for VQA are based on the CNN-RNN architecture [Malinowski *et al.*, 2015] and leverage global visual features to represent image, which may introduce noisy information. Various attention mechanisms [Yang *et al.*, 2016; Lu *et al.*, 2016; Anderson *et 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 decipt 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.

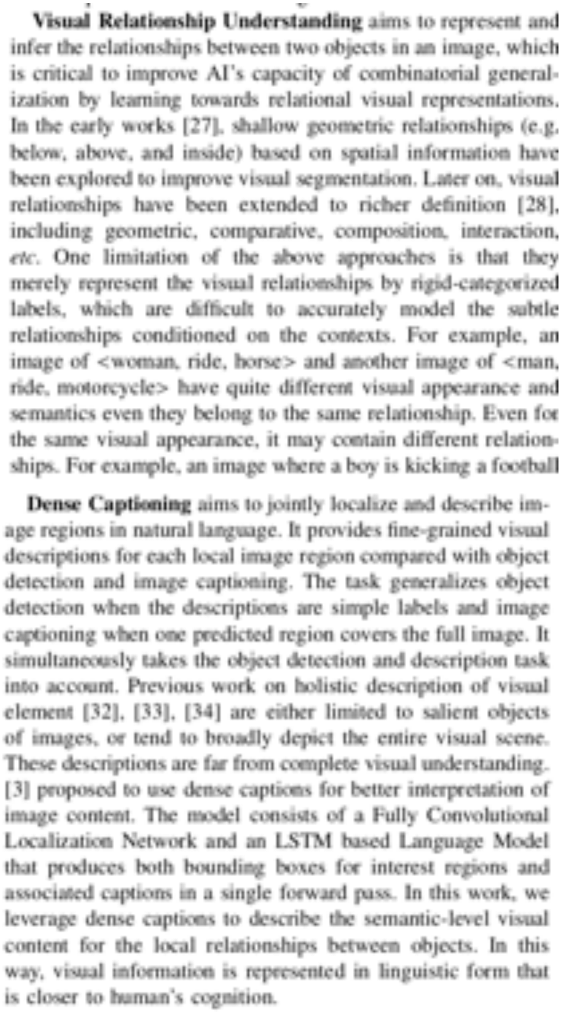
期刊论文-Related Work



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.



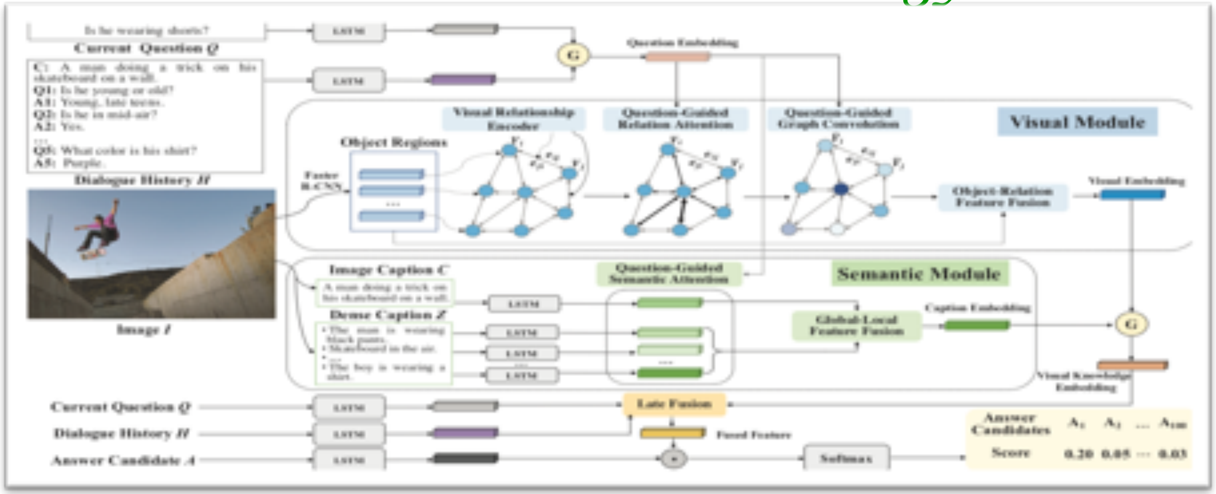
Visual Relationship Understanding aims to represent and infer the relationships between two objects in an image, which is critical to improve AI's capacity of combinatorial generalization by learning towards relational visual representations. In the early works [27], shallow geometric relationships (e.g. 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.

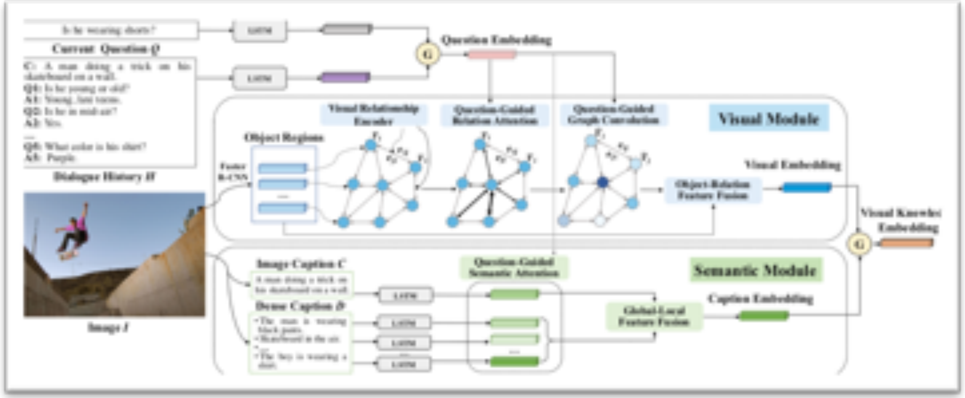
CogModal GROUP

看一个例子

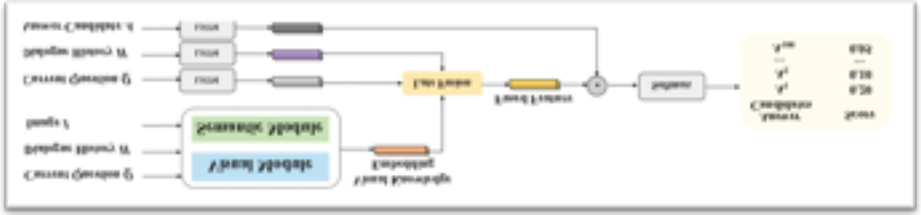
会议论文-*Methodology*

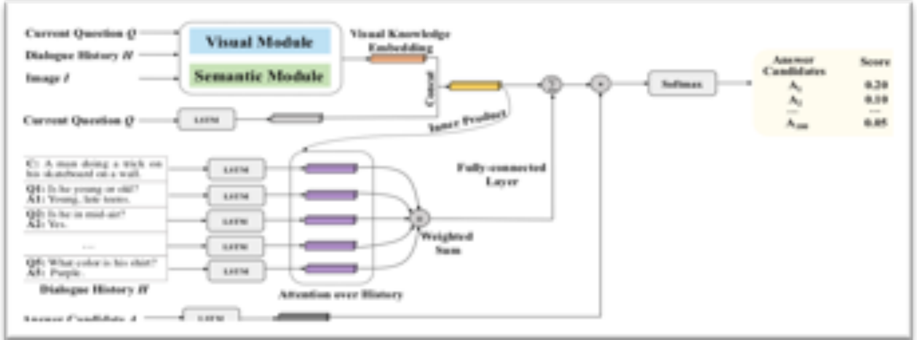


期刊论文-*Methodology*



Is he wearing shorts?
Current Question Q
Question Embedding
Visual Relationship Encoder
Question-Guided Relation Attention
Question-Guided Graph Convolution
Visual Module
Object Regions
Faster R-CNN
Object-Relation Feature Fusion
Visual Embedding
Visual Knowledge Embedding
Dialogue History H
Image I
Image Caption C
Dense Caption Z
Question-Guided Semantic Attention
Semantic Module
Global-Local Feature Fusion
Caption Embedding
LSTM
G





看一个例子

期刊论文-*Experiments*

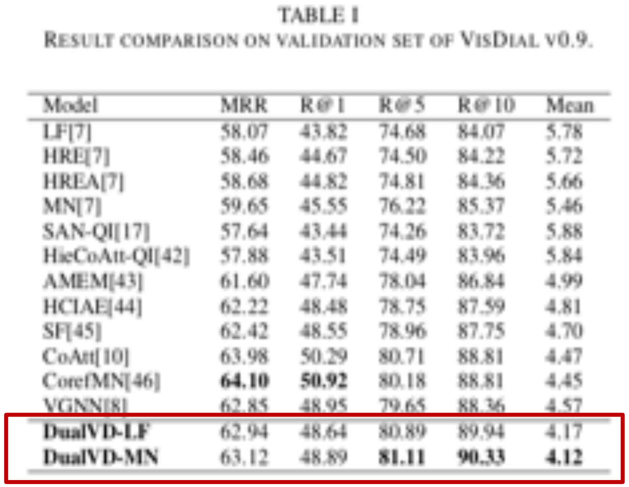


TABLE I
RESULT COMPARISON ON VALIDATION SET OF VISDIAL V0.9.

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
HieCoAtt-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**

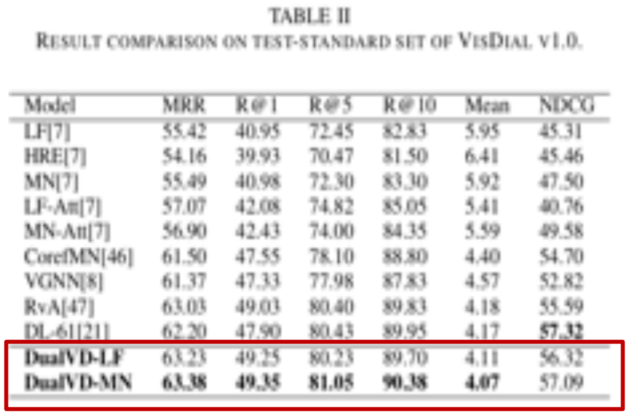


TABLE II
RESULT COMPARISON ON TEST-STANDARD SET OF VISDIAL V1.0.

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

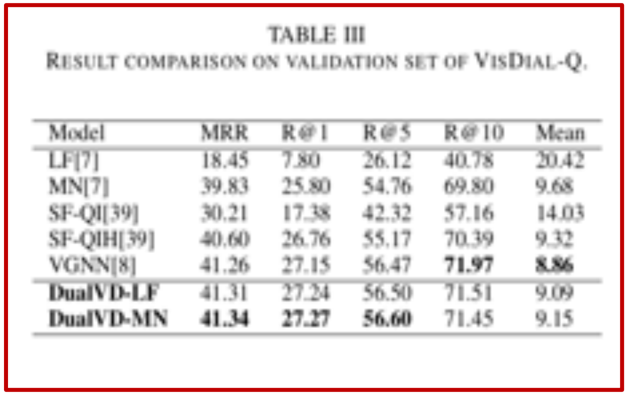


TABLE III
RESULT COMPARISON ON VALIDATION SET OF VISDIAL-Q.

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

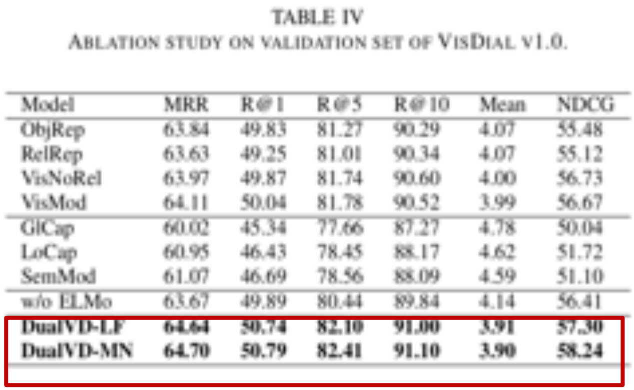


TABLE IV
ABLATION STUDY ON VALIDATION SET OF VISDIAL V1.0.

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**

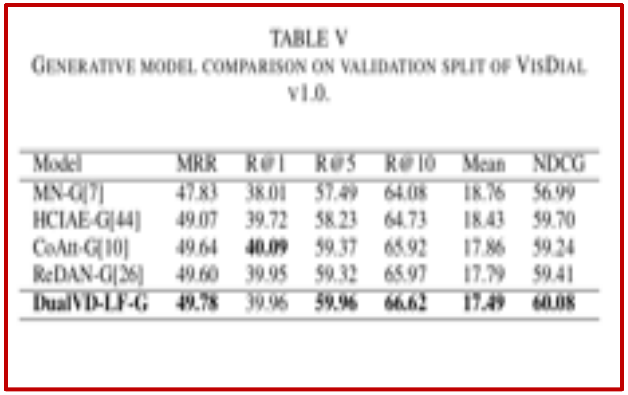


TABLE V
GENERATIVE MODEL COMPARISON ON VALIDATION SPLIT OF VISDIAL V1.0.

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**

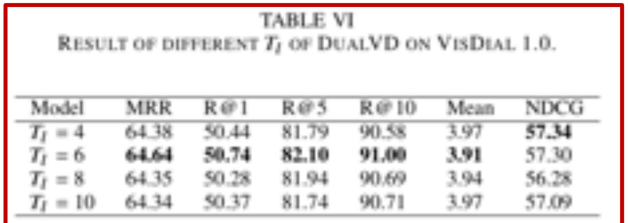


TABLE VI
RESULT OF DIFFERENT T_I OF DUALVD ON VISDIAL 1.0.

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

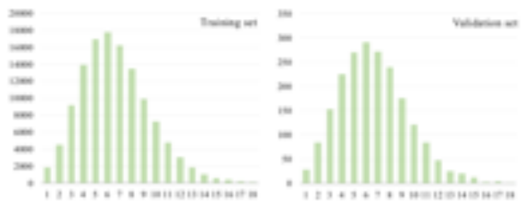


Fig. 10. Analysis of dense caption distribution on "train"and "val" splits of VisDial v1.0. x-axis represents the number of dense captions while y-axis represents the number of images generating corresponding number of captions.

看一个例子

期刊论文-*Experiments*

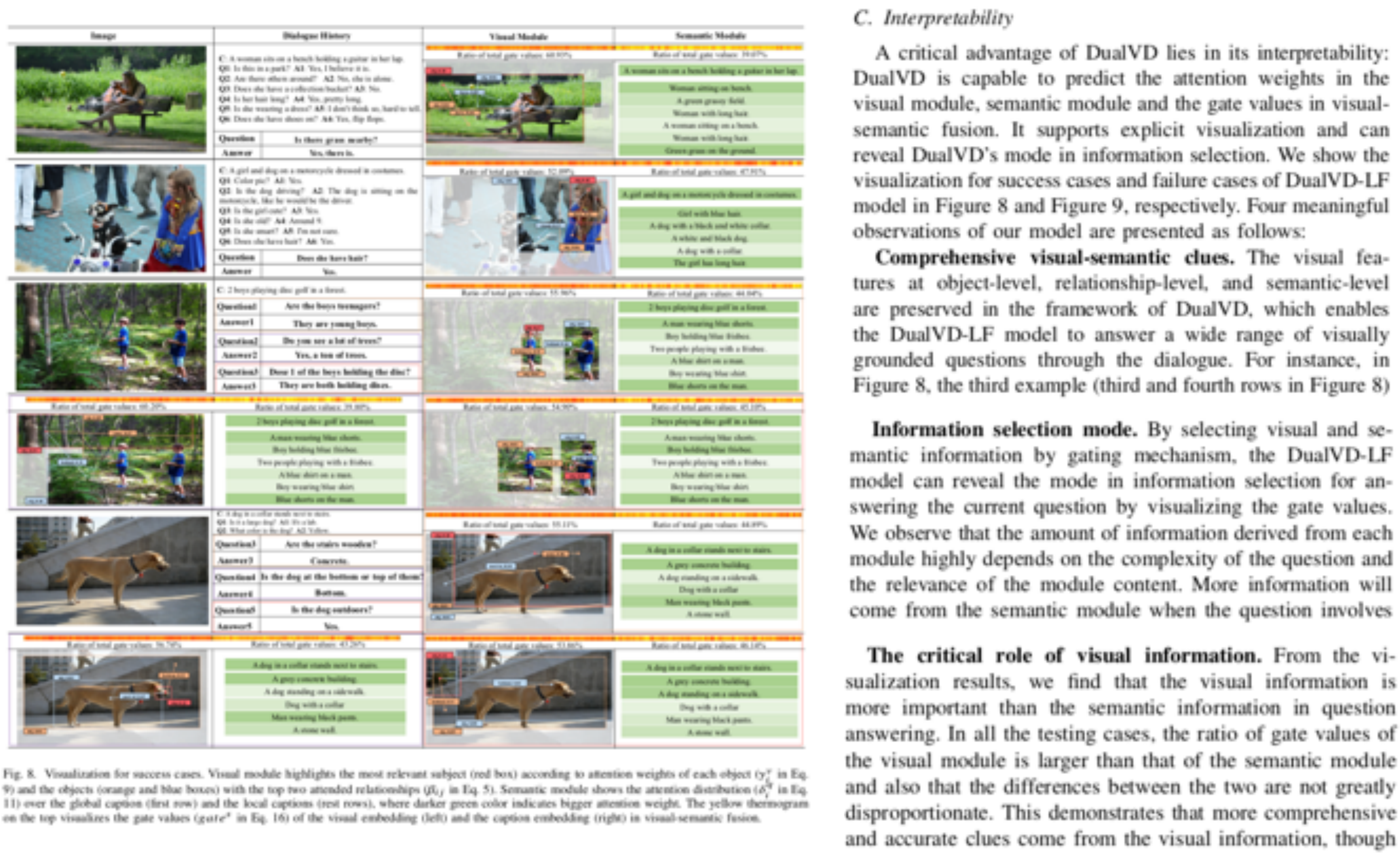


Image
Dialogue History
Visual Module
Semantic Module
C: A woman sits on a bench holding a guitar in her lap.
Q1: Is this in a park? A1: Yes, I believe it is.
Q2: Are there others around? A2: No, she is alone.
Q3: Does she have a collection bucket? A3: No.
Q4: Is her hair long? A4: Yes, pretty long.
Q5: 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
Is there grass nearby?
Answer
Yes, there is.
C: A girl and dog on a motorcycle dressed in costumes.
Q1: Color pic? 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
Does she have hair?
Answer
Yes.
C: 2 boys playing disc golf in a forest.
Question1
Are the boys teenagers?
Answer1
They are young boys.
Question2
Do you see a lot of trees?
Answer2
Yes, a ton of trees.
Question3
Does 1 of the boys holding the disc?
Answer3
They are both holding discs.
Question3
Are the stairs wooden?
Answer3
Concrete.
Question4
Is the dog at the bottom or top of them?
Answer4
Bottom.
Question5
Is the dog outdoors?
Answer5
Yes.
Ratio of total gate values
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.
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.
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.
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_i^T 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_i^q 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

CogModal GROUP

看一个例子

期刊论文-*Experiments*

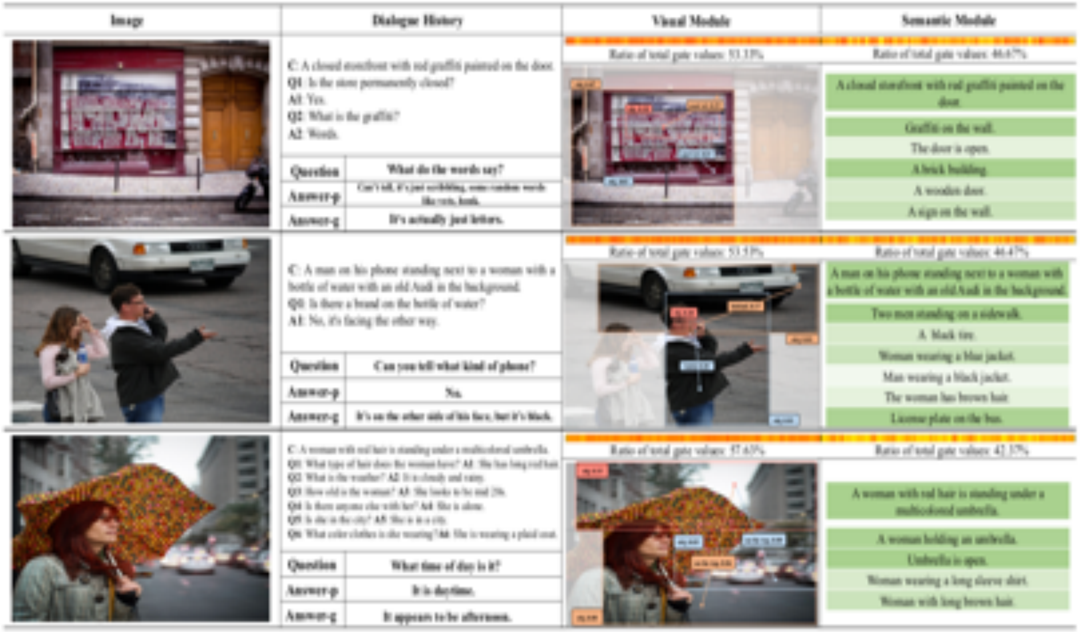


Image	Dialogue History	Visual Module	Semantic Module
	C: A closed storefront with red graffiti painted on the door. Q1: Is the store permanently closed? A1: Yes. Q2: What is the graffiti? A2: Words. Question: What do the words say? Answer-p: Can't tell, it's just scribbling, some random words like vets, honk. Answer-g: It's actually just letters.	Ratio of total gate values: 53.33%	Ratio of total gate values: 46.67% A closed storefront with red graffiti painted on the door. Graffiti on the wall. The door is open. A brick building. A wooden door. A sign on the wall.
	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.	Ratio of total gate values: 53.53%	Ratio of total gate values: 46.47% A man on his phone standing next to a woman with a bottle of water with an old Audi in the background. Two men standing on a sidewalk. A black tire. Woman wearing a blue jacket. Man wearing a black jacket. The woman has brown hair. License plate on the bus.
	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.	Ratio of total gate values: 57.63%	Ratio of total gate values: 42.37% A woman with red hair is standing under a multicolored umbrella. A woman holding an umbrella. Umbrella is open. Woman wearing a long sleeve shirt. Woman with long brown hair.

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.

学术论文
之英文规范

NETWORK
SHARING
SOCIAL MEDIA
VIDEO
EMAIL
SMS
CHAT
PICTURE
INTERNET
SERVER FARM
GROWTH
IDEA
YES
STATISTICS
LAST YEAR
THIS YEAR
QUALITY
STRATEGY
VISION
SUCCESS
INNOVATION
VISION
RESEARCH
35%
150%
TIME
MONEY
BUSINESS
QUALITY

简洁的表达

施一公

论文只是一个载体，是为了向同行们宣告你的科研发现，是科学领域交流的重要工具。所以，在科研论文写作时，一定要谨记于心的就是：用最简单的话表达最明白的意思！

Q：那如何才能最简单的表达自己的意思呢？
A：删掉后不影响句子表达的部分，全部删掉！

It is impossible for us to accomplish the transformation of the whole society overnight

It is impossible for us to transform the whole society overnight

No Chinglish !

- Chinglish会让人读起来奇奇怪怪

We compare our single-model results with previous best published results on VQA/GQA test-standard sets and NLVR2 public test set, and did one or two times.

解决方法：

1.多多阅读native speaker的文献，学习用词方法和表达习惯

2.以英语的思维来写文章，英文中没有的表达就不使用

3.搜集经典的Chinglish表达，遇到类似的表达及时更改

We compare our single-model results with pre-vious best published results on VQA/GQA test-standard sets and NLVR2public test set, and tried a couple of times.

No Chinglish !

- 地道English，更加简洁

As we know, the channel gain varies much more slower than the channel phase, and we have thousands of ways to prove it.

解决方法：

1.忘记中文中的经典短语，尤其是附加的修饰性短语

2.英文表达注重逻辑，一句话表达一个观点

3.善用副词表达观点，而不是短语

删除不地道的表达：

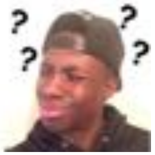
As we know, the channel gain varies much more slower than the channel phase.

使用副词替换短语：

Obviously, the channel gain varies much more slower than the channel phase.

逻辑严谨——不使用未定义术语

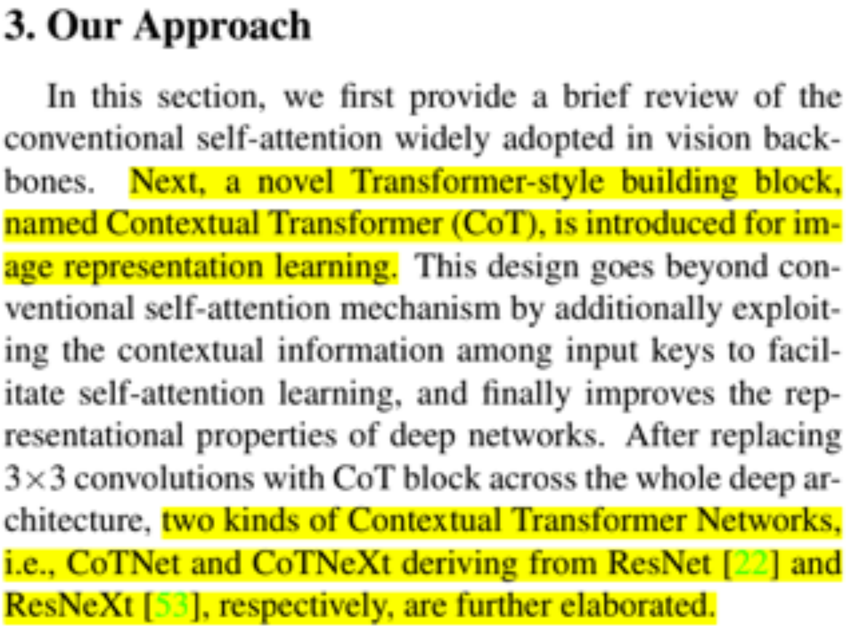
假设审稿人突然在论文中读到这么一句话：



Now, we could use CoTNet model to predict the result of that dataset.

清晰明了定义：

- 方法一：有清晰的段落结构， 在术语使用前定义解释



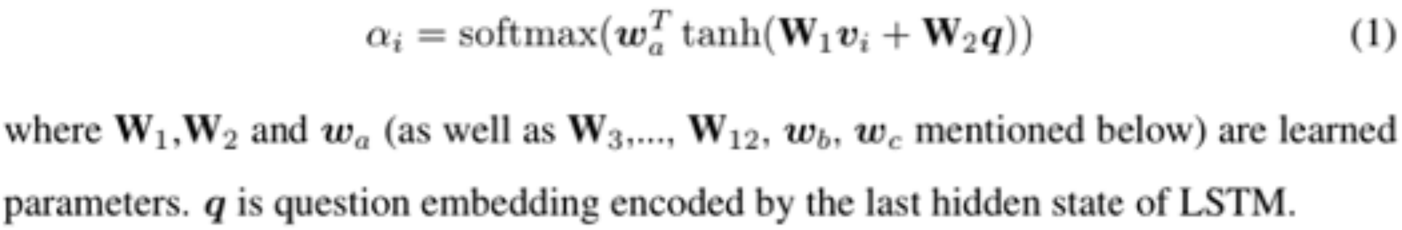
3. Our Approach

In this section, we first provide a brief review of the conventional self-attention widely adopted in vision backbones. Next, a novel Transformer-style building block, named Contextual Transformer (CoT), is introduced for image representation learning. This design goes beyond conventional self-attention mechanism by additionally exploiting the contextual information among input keys to facilitate self-attention learning, and finally improves the representational properties of deep networks. After replacing 3×3 convolutions with CoT block across the whole deep architecture, two kinds of Contextual Transformer Networks, i.e., CoTNet and CoTNeXt deriving from ResNet [22] and ResNeXt [53], respectively, are further elaborated.

逻辑严谨——不使用未定义术语

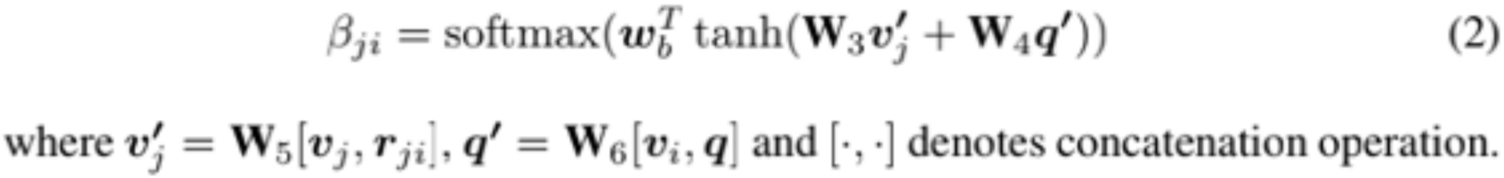
清晰明了的定义：

- 方法二：给出公式后集中对定义及公式中的符号进行解释



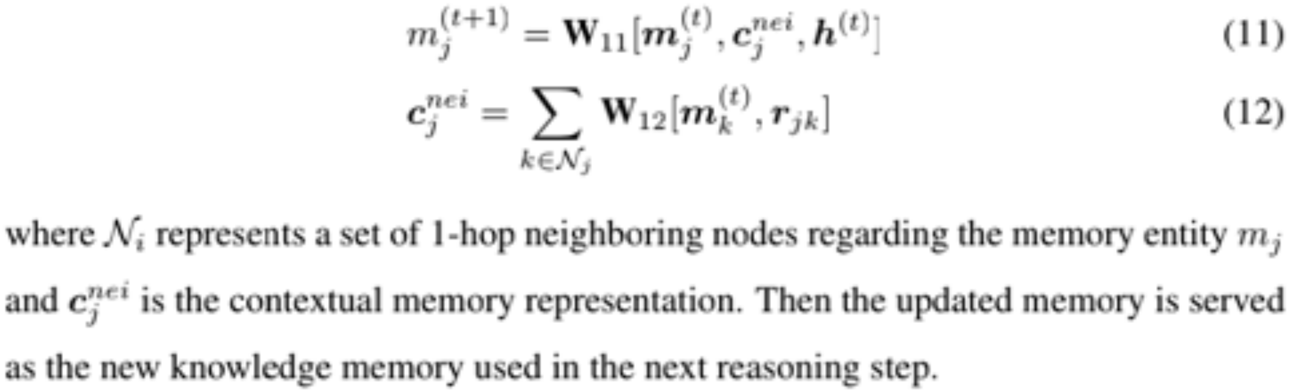
$$\alpha_i = \mathrm{softmax}(\boldsymbol{w}_a^T \tanh(\mathbf{W}_1 \boldsymbol{v}_i + \mathbf{W}_2 \boldsymbol{q})) \quad (1)$$

where $\mathbf{W}_1$,$\mathbf{W}_2$ and $\boldsymbol{w}_a$ (as well as $\mathbf{W}_3$,..., $\mathbf{W}_{12}$, $\boldsymbol{w}_b$, $\boldsymbol{w}_c$ mentioned below) are learned parameters. $\boldsymbol{q}$ is question embedding encoded by the last hidden state of LSTM.



$$\beta_{ji} = \mathrm{softmax}(\boldsymbol{w}_b^T \tanh(\mathbf{W}_3 \boldsymbol{v}'_j + \mathbf{W}_4 \boldsymbol{q}')) \quad (2)$$

where $\boldsymbol{v}'_j = \mathbf{W}_5[\boldsymbol{v}_j, \boldsymbol{r}_{ji}]$, $\boldsymbol{q}' = \mathbf{W}_6[\boldsymbol{v}_i, \boldsymbol{q}]$ and $[\cdot, \cdot]$ denotes concatenation operation.



$$m_j^{(t+1)} = \mathbf{W}_{11}[\boldsymbol{m}_j^{(t)}, \boldsymbol{c}_j^{nei}, \boldsymbol{h}^{(t)}] \quad (11)$$

$$\boldsymbol{c}_j^{nei} = \sum_{k \in \mathcal{N}_j} \mathbf{W}_{12}[\boldsymbol{m}_k^{(t)}, \boldsymbol{r}_{jk}] \quad (12)$$

where $\mathcal{N}_i$ represents a set of 1-hop neighboring nodes regarding the memory entity m_j and $\boldsymbol{c}_j^{nei}$ is the contextual memory representation. Then the updated memory is served as the new knowledge memory used in the next reasoning step.

CogModal GROUP

逻辑严谨——不反复唠叨

- 同一个观点不要反复阐述

经典问题：怕别人看不懂自己“图灵奖”级别的研究，反复解释自己的模型

Abstract:
XXXXXXX each per_x0002_formed by a Graph-based Read, Update, and Control (GRUC) module that conducts parallel reasoning over both visual and semantic information and XXX can XXX. XXXXXXXX

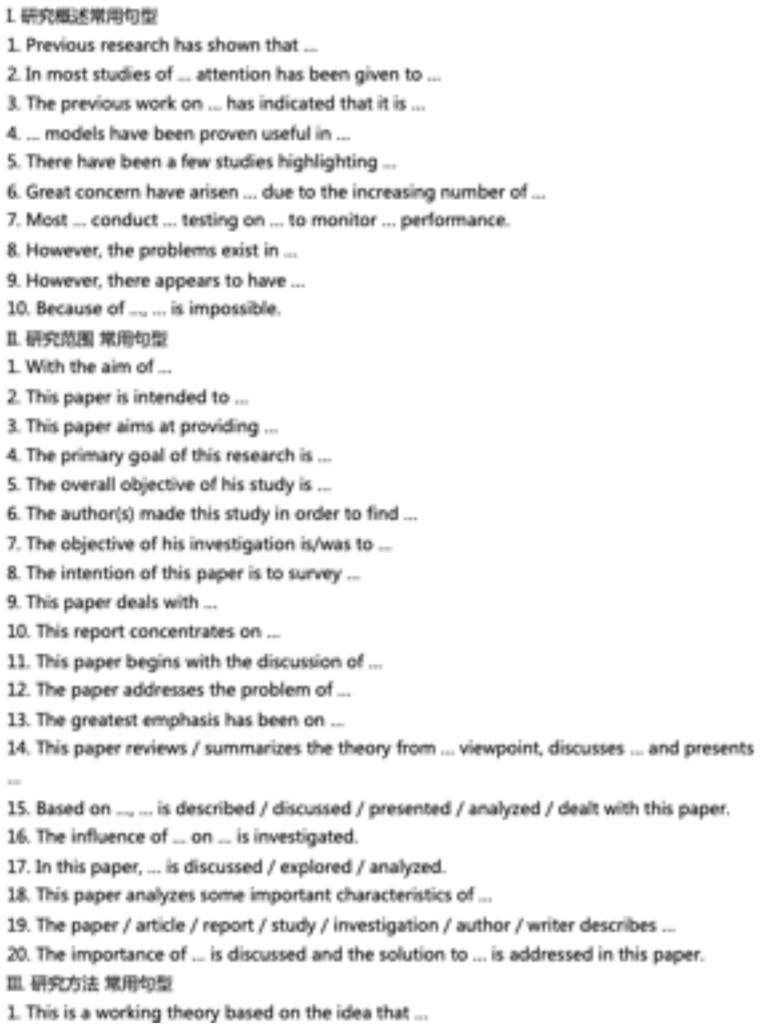
Related Wordk:
XXXXXXXX , But our GRUC can reasoning reasoning over both visual and semantic graph that can XXX. XXXXXXXXXX

Methodology:
XXXXXXXXX, our GRUC includes two modality visual and semantic, which can XXX SO GOOD!. XXXXXXXXXX

不要一直提及自己模型的优点和结构。可以在Abstract中提一句模型解决的问题，在Methodology中提及模型的组成方式，在Experience中根据效果图说明模型的优势，这样就不会造成大量的重复。

润色表达

· 常见表达方式



I. 研究概述常用句型

1. Previous research has shown that ...
2. In most studies of ... attention has been given to ...
3. The previous work on ... has indicated that it is ...
4. ... models have been proven useful in ...
5. There have been a few studies highlighting ...
6. Great concern have arisen ... due to the increasing number of ...
7. Most ... conduct ... testing on ... to monitor ... performance.
8. However, the problems exist in ...
9. However, there appears to have ...
10. Because of ..., ... is impossible.

II. 研究范围 常用句型

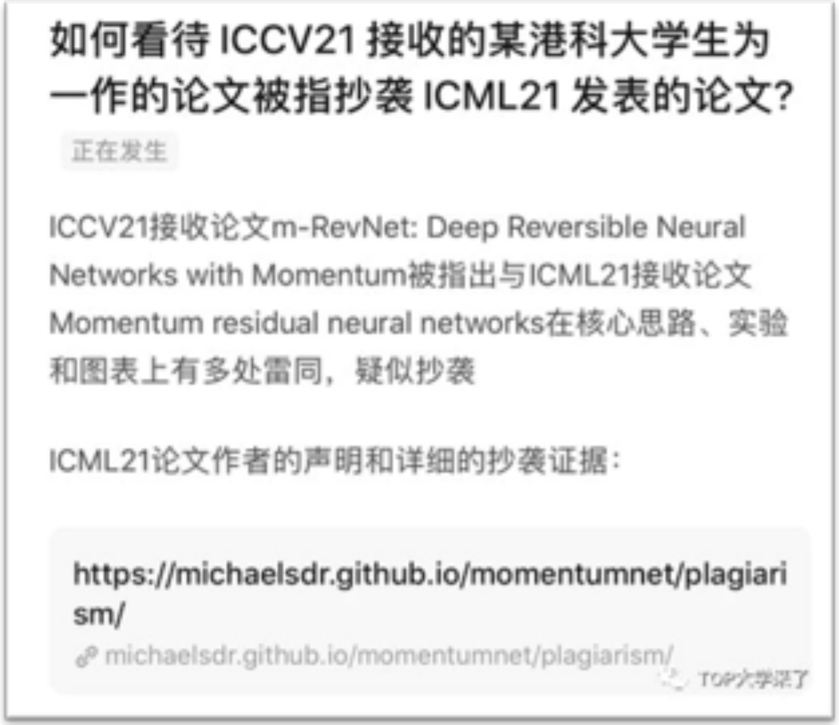
1. With the aim of ...
2. This paper is intended to ...
3. This paper aims at providing ...
4. The primary goal of this research is ...
5. The overall objective of his study is ...
6. The author(s) made this study in order to find ...
7. The objective of his investigation is/was to ...
8. The intention of this paper is to survey ...
9. This paper deals with ...
10. This report concentrates on ...
11. This paper begins with the discussion of ...
12. The paper addresses the problem of ...
13. The greatest emphasis has been on ...
14. This paper reviews / summarizes the theory from ... viewpoint, discusses ... and presents ...
15. Based on ..., ... is described / discussed / presented / analyzed / dealt with this paper.
16. The influence of ... on ... is investigated.
17. In this paper, ... is discussed / explored / analyzed.
18. This paper analyzes some important characteristics of ...
19. The paper / article / report / study / investigation / author / writer describes ...
20. The importance of ... is discussed and the solution to ... is addressed in this paper.

III. 研究方法 常用句型

1. This is a working theory based on the idea that ...

严守学术道德，

切记抄袭！



如何看待 ICCV21 接收的某港科大学生为一作的论文被指抄袭 ICML21 发表的论文?

正在发生

ICCV21接收论文m-RevNet: Deep Reversible Neural Networks with Momentum被指出与ICML21接收论文Momentum residual neural networks在核心思路、实验和图表上有多处雷同，疑似抄袭

ICML21论文作者的声明和详细的抄袭证据：

https://michaelsdr.github.io/momentumnet/plagiarism/

michaelsdr.github.io/momentumnet/plagiarism/

数学规范

- LaTex公式和常用符号规范

1. $\sqrt[3]{\frac{x^3+y^3}{2}}$ VS ((x^3 + y+3)/2)^(1/3)

2. 在VQA问题中，通常$\mathcal{W}$表示待训练参数、Q表示问题、$\mathcal{A}$ 表示答案

   如果一篇论文用 $\mathcal{A}$ 表示问题、Q 表示答案，审稿人很难理解

术语规范

- 机器学习领域学术术语标准（节选）：

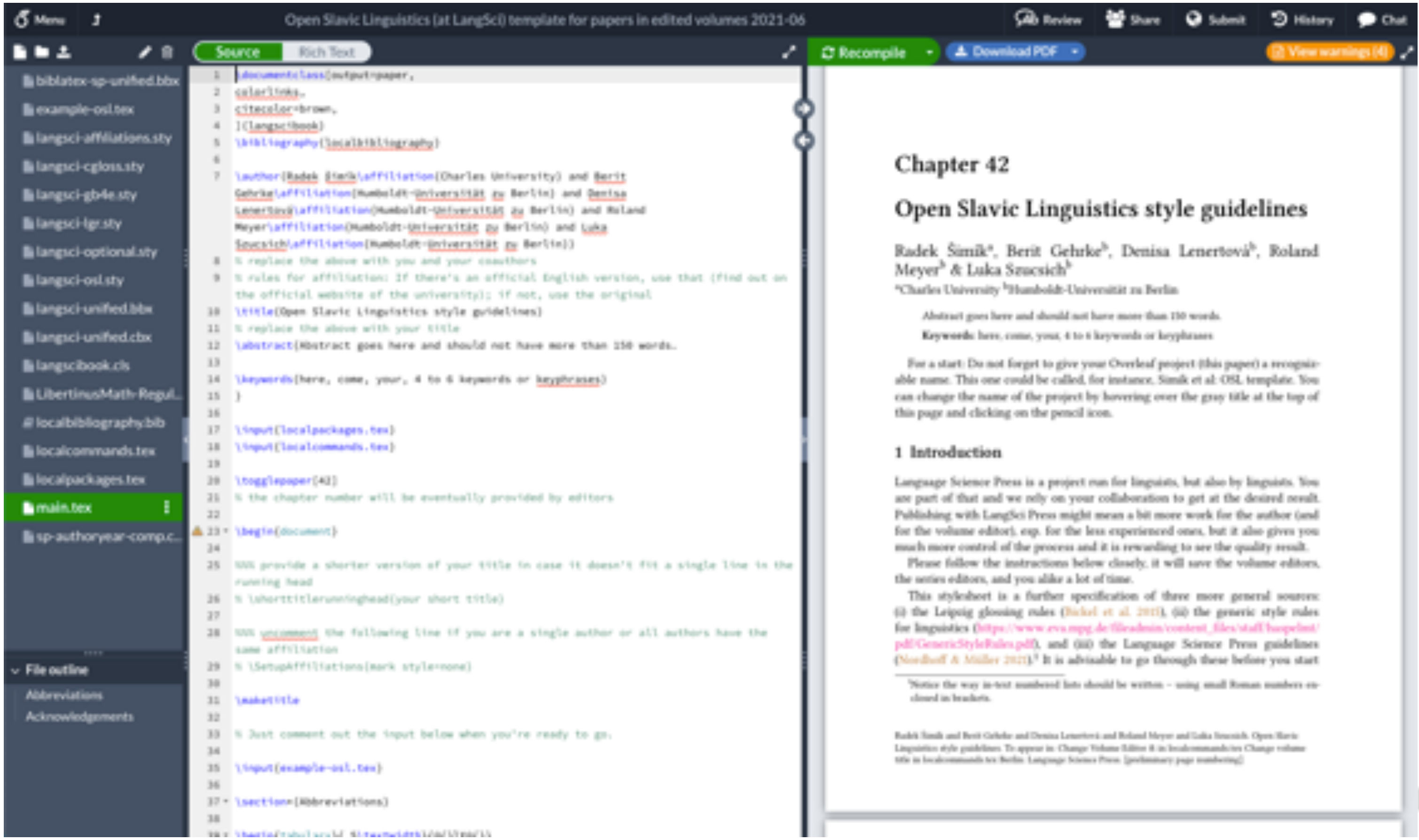
英文表示	缩写	中文含义
Adaptive Moment Estimation Algorithm	Adam	Adam算法
Artificial Neural Network	ANN	人工神经网络
Autoencoder	AE	自编码器
Back Propagation	BP	反向传播
Long Short Term Memory	LSTM	长短期记忆
Maximum Likelihood Estimation	MLE	极大似然估计
Memory Network	MN	记忆网络
Multi-Layer Perceptron	MLP	多层感知机
Adaptive Bitrate Algorithm	ABR	自适应比特率算法
Agent	-	智能体
Attention Mechanism	-	注意力机制

英文表示	缩写	中文含义
Boltzmann Machine	—	玻尔兹曼机
Classification And Regression Tree	CART	分类与回归树
Classifier	-	分类器
Confidence	-	置信度
Context Window	-	上下文窗口
Convolutional Kernel	-	卷积核
Convolutional Neural Network	CNN	卷积神经网络
Cross Entropy	-	交叉熵
Feedforward Neural Network	FNN	前馈神经网络
Gated Recurrent Unit	GRU	门控循环单元
Generative Adversarial Network	GAN	生成对抗网络

More ： https://github.com/jiqizhixin/Artificial-Intelligence-Terminology-Database/blob/master/section/Machine%20Learning.md

写作工具

- LaTex
  - 拥有海量模版，上手简单
  - 通过算法确定间距，格式准确
  - *IEEE*，*ACM*等标准的写作软件



```
\documentclass[output=paper,
colorlinks,
citecolor=brown,
]{langscibook}
\bibliography{localbibliography}

\author{Radek Šimík\affiliation{Charles University} and Berit
Gehrke\affiliation{Humboldt-Universität zu Berlin} and Denisa
Lenertová\affiliation{Humboldt-Universität zu Berlin} and Roland
Meyer\affiliation{Humboldt-Universität zu Berlin} and Luka
Szucsich\affiliation{Humboldt-Universität zu Berlin}}
% replace the above with you and your coauthors
% rules for affiliation: If there's an official English version, use that (find out on
the official website of the university); if not, use the original
\title{Open Slavic Linguistics style guidelines}
% replace the above with your title
\abstract{Abstract goes here and should not have more than 150 words.

\keywords{here, come, your, 4 to 6 keywords or keyphrases}
}

\input{localpackages.tex}
\input{localcommands.tex}

\togglepaper[42]
% the chapter number will be eventually provided by editors

\begin{document}

%%% provide a shorter version of your title in case it doesn't fit a single line in the
running head
% \shorttitlerunninghead{your short title}

%%% uncomment the following line if you are a single author or all authors have the
same affiliation
% \SetupAffiliations{mark style=none}

\maketitle

% Just comment out the input below when you're ready to go.

\input{example-osl.tex}

\section*{Abbreviations}

\begin{tabularx}{.5\textwidth}{lX}
```

Chapter 42

Open Slavic Linguistics style guidelines

Radek Šimík[a], Berit Gehrke[b], Denisa Lenertová[b], Roland Meyer[b] & Luka Szucsich[b]
[a]Charles University [b]Humboldt-Universität zu Berlin

Abstract goes here and should not have more than 150 words.

Keywords: here, come, your, 4 to 6 keywords or keyphrases

For a start: Do not forget to give your Overleaf project (this paper) a recognizable name. This one could be called, for instance, Šimík et al: OSL template. You can change the name of the project by hovering over the gray title at the top of this page and clicking on the pencil icon.

1 Introduction

Language Science Press is a project run for linguists, but also by linguists. You are part of that and we rely on your collaboration to get at the desired result. Publishing with LangSci Press might mean a bit more work for the author (and for the volume editor), esp. for the less experienced ones, but it also gives you much more control of the process and it is rewarding to see the quality result.

Please follow the instructions below closely, it will save the volume editors, the series editors, and you alike a lot of time.

This stylesheet is a further specification of three more general sources: (i) the Leipzig glossing rules (Bickel et al. 2021), (ii) the generic style rules for linguistics (https://www.eva.mpg.de/fileadmin/content_files/staff/haspelmt/pdf/GenericStyleRules.pdf), and (iii) the Language Science Press guidelines (Nordhoff & Müller 2021).[1] It is advisable to go through these before you start

[1] Notice the way in-text numbered lists should be written – using small Roman numbers enclosed in brackets.

Radek Šimík and Berit Gehrke and Denisa Lenertová and Roland Meyer and Luka Szucsich. Open Slavic Linguistics style guidelines. To appear in: Change Volume Editor & in localcommands.tex Change volume title in localcommands.tex Berlin: Language Science Press. [preliminary page numbering]

与审稿人交流规范

- 让人看了暖心的审稿人意见

Reject – More holes than my grandad' s string vest!

拒发，这货的漏洞比我爷爷网眼背心上的洞还多！

block the author' s email ID so they can' t use the online system in future.

建议锁定该作者的电子邮件*ID*，避免此人日后继续投稿。。。

It is early in the year, but difficult to imagine any paper overtaking this one for lack of imagination, logic, or data –it is beyond redemption.

新的一年才刚开始，但我怎么就感觉很难找到比这篇更扯、更没逻辑的文章呢？谁都拯救不了这篇文章了！

与审稿人交流规范

- 批判接受意见

  有的意见是合理的，但有的意见略有偏颇，因为审稿人的研究方向可能和你的研究方向不同

- 心存感激

  审稿是义务工作，审稿人是不会从中获益的，所以要对审稿人付出的时间心存感激

- 礼貌回复

  如果收到了前面的审稿意见，不要着急怼回去，论文必然无法通过，耐心一一客观、礼貌回复

英文规范小结

- 规范论文的诞生口诀！
  - 精炼的句子
  - 地道的表达
  - 严谨的逻辑
  - 美观的排版
  - 礼貌的交流

学术研究
之日常积累

学术研究之日常积累

日常 5L 积累：

- ☼ Paper List
- ☼ Idea List
- ☼ Math List
- ☼ English List
- ☼ Code List

欲穷千里目，更上一层楼

CogModal GROUP

Paper List

Paper list：

- 如何选论文？
- 如何精读 & 粗读论文？
- 如何进行文献整理 & 总结？

Paper List——如何找论文？

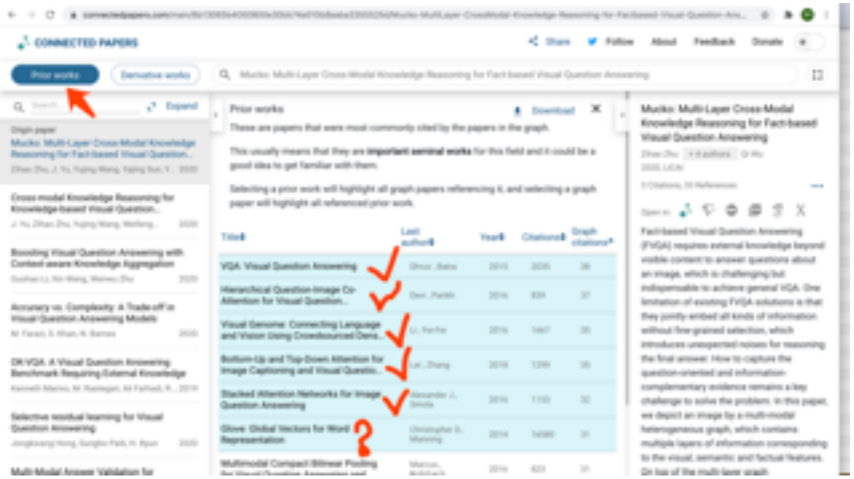
- 看领域内权威的**综述**
- 刷**顶会**期刊
- 刷 **arxiv**
- 从**公众号、知乎**等
- 从 **AMiner、Connect papers** 等

Deep Variational Information Bottleneck

Disentangling by Factorising

ICML, (2018)

AMiner


Prior works

These are papers that were most commonly cited by the papers in the graph.

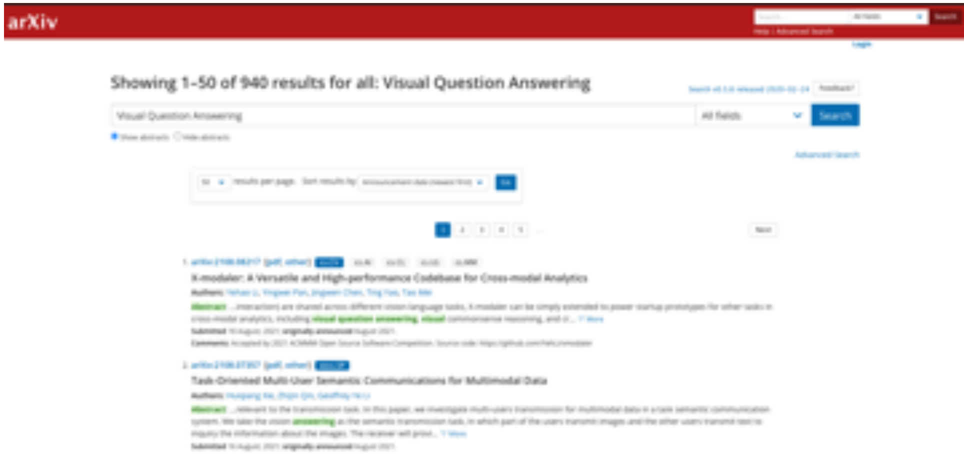
This usually means that they are **important seminal works** for this field and it could be a good idea to get familiar with them.

Selecting a prior work will highlight all graph papers referencing it, and selecting a graph paper will highlight all referenced prior work.

Title	Last author	Year	Citations	Graph citations
VQA: Visual Question Answering ✓	Dhruv ..Batra	2015	[illegible]	[illegible]
Hierarchical Question-Image Co-Attention for Visual Question... ✓	Devi ..Parikh	2016	[illegible]	[illegible]
Visual Genome: Connecting Language and Vision Using Crowdsourced Dens... ✓	Li ..Fei-Fei	2016	[illegible]	[illegible]
Bottom-Up and Top-Down Attention for Image Captioning and Visual Questio... ✓	Lei ..Zhang	2018	[illegible]	[illegible]
Stacked Attention Networks for Image Question Answering ✓	Alexander J. Smola	2016	[illegible]	[illegible]
Glove: Global Vectors for Word Representation ?	Christopher D. Manning	2014	[illegible]	[illegible]
Multimodal Compact Bilinear Pooling	Marcus..	2016	[illegible]	[illegible]

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

Connect papers


Showing 1–50 of 940 results for all: Visual Question Answering

Visual Question Answering

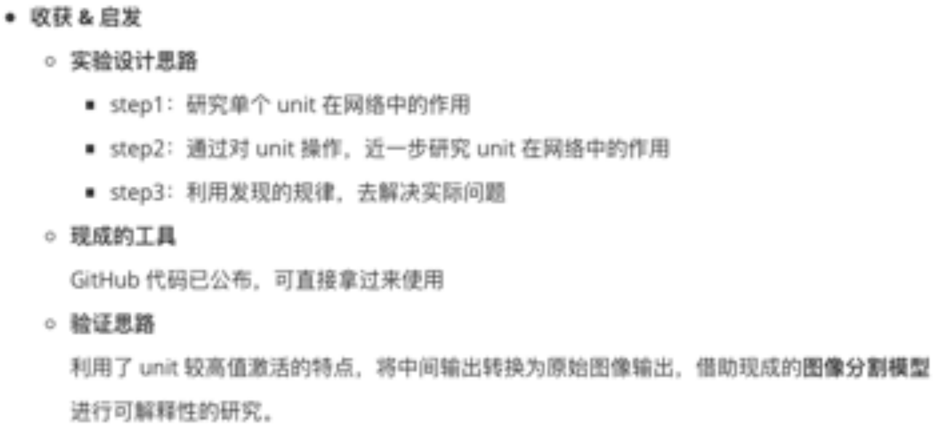
1. X-modaler: A Versatile and High-performance Codebase for Cross-modal Analytics

2. Task-Oriented Multi-User Semantic Communications for Multimodal Data

Arxiv

Paper List——精读范文

- 记录论文的**标题、出处**和**研究团队**等
- 记录读完论文的**收获与启发**
- 记录论文的**动机**
- 记录论文的**模型结构**
- 记录论文的**实验设计**

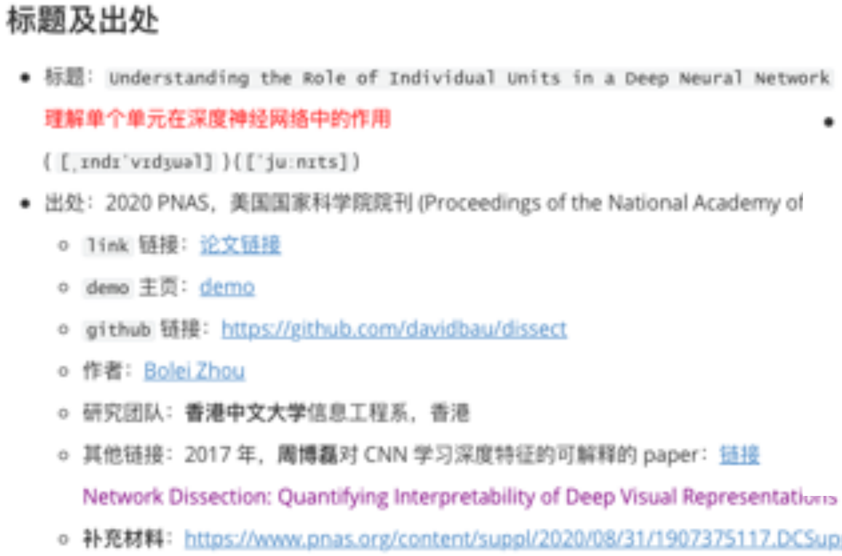


- 收获 & 启发
  - 实验设计思路
    - step1：研究单个 unit 在网络中的作用
    - step2：通过对 unit 操作，近一步研究 unit 在网络中的作用
    - step3：利用发现的规律，去解决实际问题
  - 现成的工具

    GitHub 代码已公布，可直接拿过来使用
  - 验证思路

    利用了 unit 较高值激活的特点，将中间输出转换为原始图像输出，借助现成的**图像分割模型**进行可解释性的研究。

收获 & 启发



标题及出处

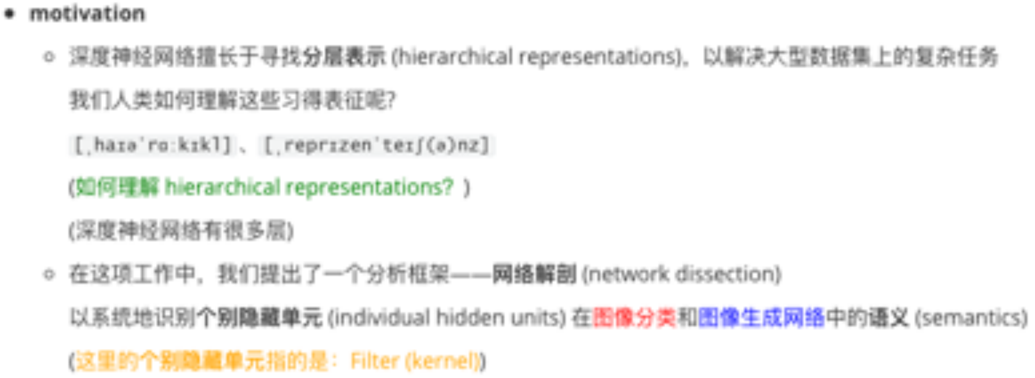
- 标题：`Understanding the Role of Individual Units in a Deep Neural Network`

  理解单个单元在深度神经网络中的作用

  (`[ˌɪndɪˈvɪdʒuəl]`)(`[ˈjuːnɪts]`)
- 出处：2020 PNAS，美国国家科学院院刊 (Proceedings of the National Academy of
  - `link` 链接：论文链接
  - `demo` 主页：demo
  - `github` 链接：https://github.com/davidbau/dissect
  - 作者：Bolei Zhou
  - 研究团队：**香港中文大学**信息工程系，香港
  - 其他链接：2017 年，**周博磊**对 CNN 学习深度特征的可解释的 paper：链接

    Network Dissection: Quantifying Interpretability of Deep Visual Representations
  - 补充材料：https://www.pnas.org/content/suppl/2020/08/31/1907375117.DCSupplemental

标题 & 出处



- **motivation**
  - 深度神经网络擅长于寻找**分层表示** (hierarchical representations)，以解决大型数据集上的复杂任务

    我们人类如何理解这些习得表征呢?

    `[ˌhaɪəˈrɑːkɪkl]`、`[ˌreprɪzenˈteɪʃ(ə)nz]`

    (如何理解 hierarchical representations?)

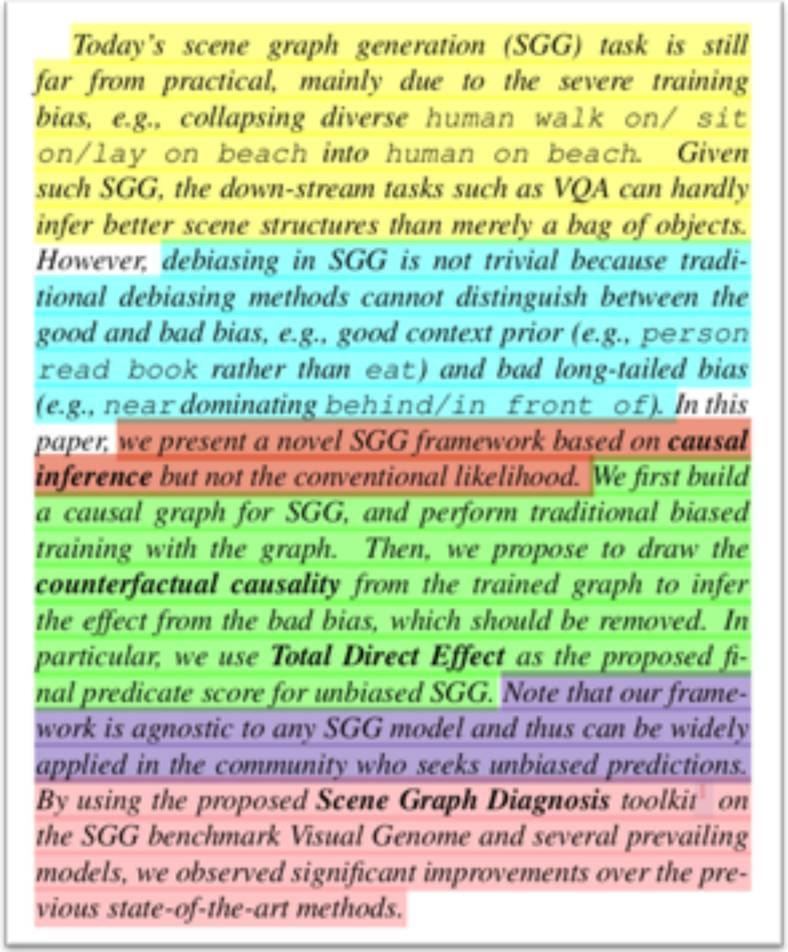
    (深度神经网络有很多层)
  - 在这项工作中，我们提出了一个分析框架——**网络解剖** (network dissection)

    以系统地识别**个别隐藏单元** (individual hidden units) 在图像分类和图像生成网络中的语义 (semantics)

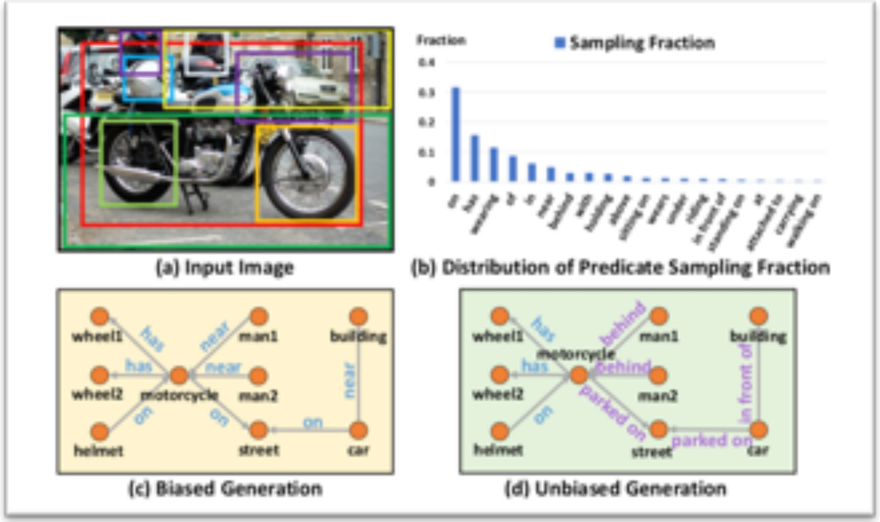
    (这里的个别隐藏单元指的是：Filter (kernel))

动机

Paper List——精读范文



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** toolkit[1] on the SGG benchmark Visual Genome and several prevailing models, we observed significant improvements over the previous state-of-the-art methods.

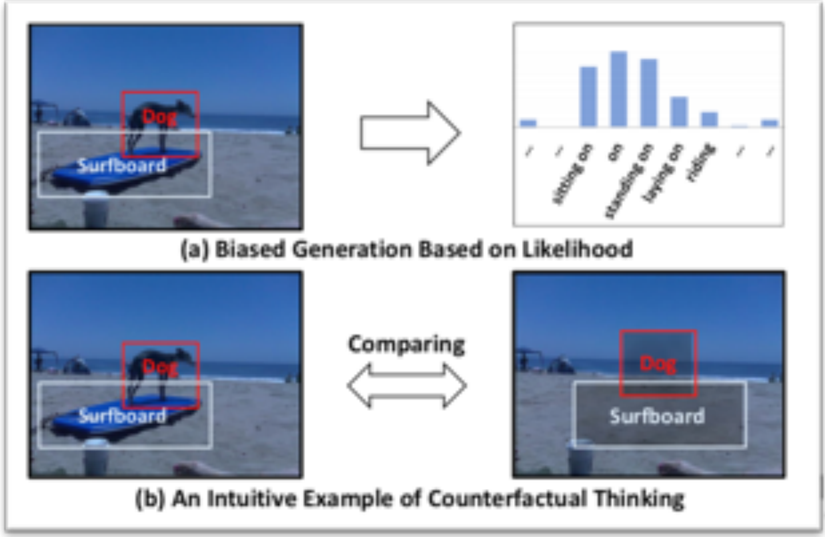


(a) Input Image

(b) Distribution of Predicate Sampling Fraction

(c) Biased Generation

(d) Unbiased Generation

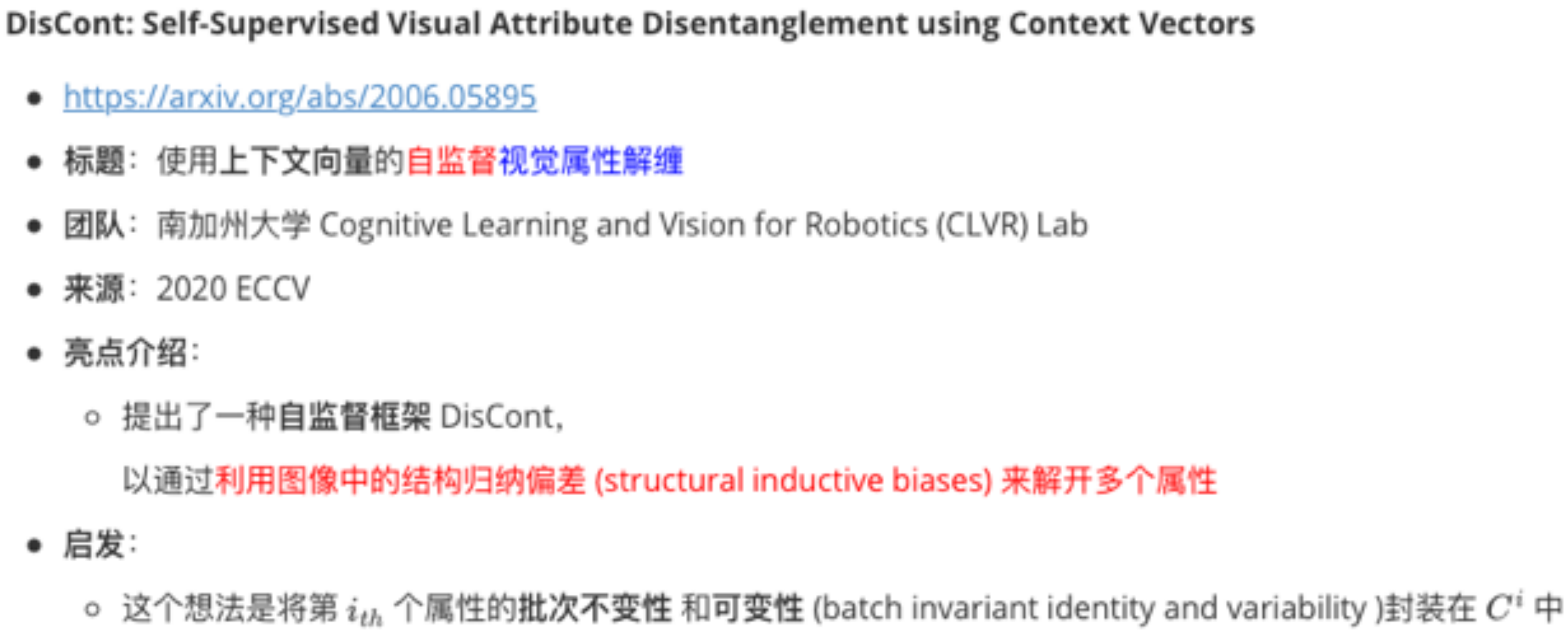


(a) Biased Generation Based on Likelihood

(b) An Intuitive Example of Counterfactual Thinking

Paper List——粗读论文

- 记录论文的链接、团队、来源、亮点(50字左右)、启发等



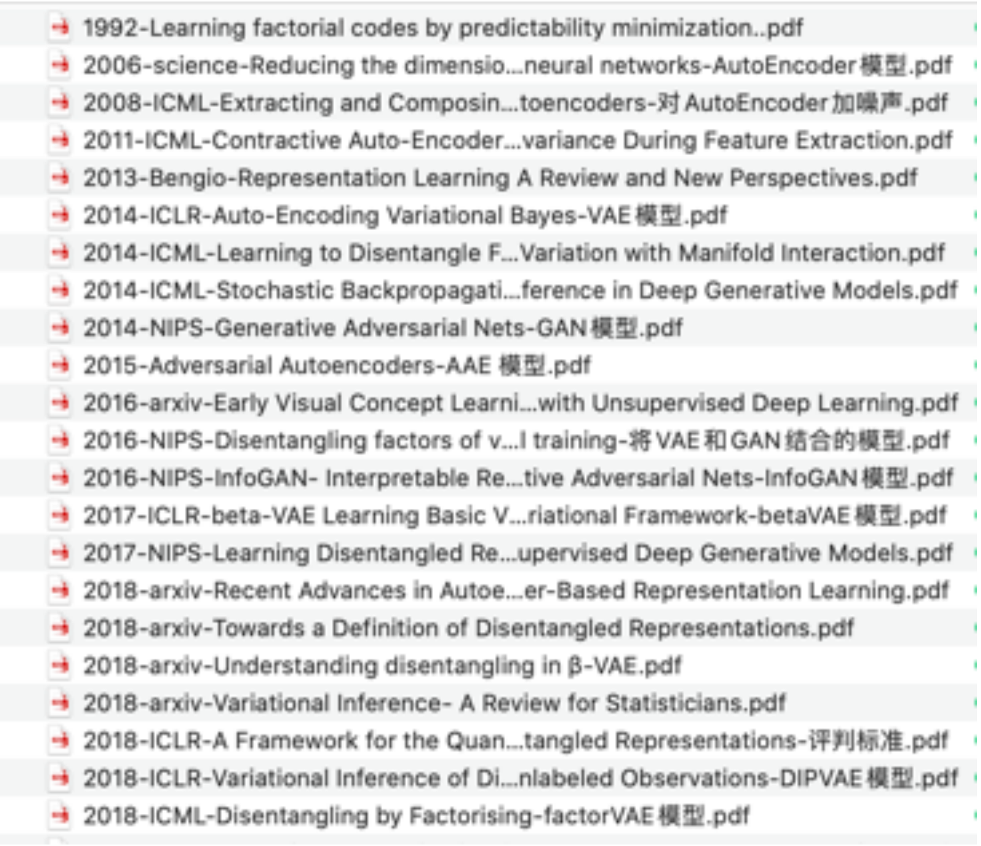
DisCont: Self-Supervised Visual Attribute Disentanglement using Context Vectors

- https://arxiv.org/abs/2006.05895
- 标题：使用上下文向量的自监督视觉属性解缠
- 团队：南加州大学 Cognitive Learning and Vision for Robotics (CLVR) Lab
- 来源：2020 ECCV
- 亮点介绍：
  - 提出了一种自监督框架 DisCont，
    以通过利用图像中的结构归纳偏差 (structural inductive biases) 来解开多个属性
- 启发：
  - 这个想法是将第 i_{th} 个属性的批次不变性 和可变性 (batch invariant identity and variability)封装在 C^i 中

粗读论文的笔记参考

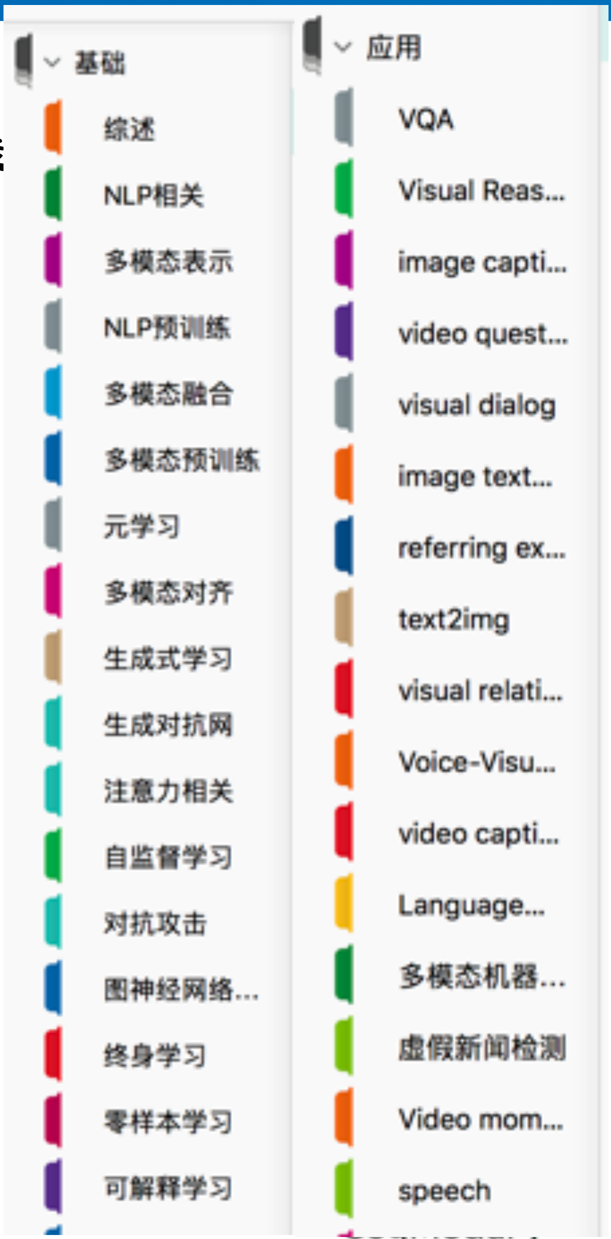
Paper List——文献整理

- 按**内容**整理：将论文按照不同的主题进行分类
- 按**时间**整理：年份 + 来源 + 题目 + 内容简介



1992-Learning factorial codes by predictability minimization..pdf
2006-science-Reducing the dimensio...neural networks-AutoEncoder模型.pdf
2008-ICML-Extracting and Composin...toencoders-对AutoEncoder加噪声.pdf
2011-ICML-Contractive Auto-Encoder...variance During Feature Extraction.pdf
2013-Bengio-Representation Learning A Review and New Perspectives.pdf
2014-ICLR-Auto-Encoding Variational Bayes-VAE模型.pdf
2014-ICML-Learning to Disentangle F...Variation with Manifold Interaction.pdf
2014-ICML-Stochastic Backpropagati...ference in Deep Generative Models.pdf
2014-NIPS-Generative Adversarial Nets-GAN模型.pdf
2015-Adversarial Autoencoders-AAE 模型.pdf
2016-arxiv-Early Visual Concept Learni...with Unsupervised Deep Learning.pdf
2016-NIPS-Disentangling factors of v...l training-将VAE和GAN结合的模型.pdf
2016-NIPS-InfoGAN- Interpretable Re...tive Adversarial Nets-InfoGAN模型.pdf
2017-ICLR-beta-VAE Learning Basic V...riational Framework-betaVAE模型.pdf
2017-NIPS-Learning Disentangled Re...upervised Deep Generative Models.pdf
2018-arxiv-Recent Advances in Autoe...er-Based Representation Learning.pdf
2018-arxiv-Towards a Definition of Disentangled Representations.pdf
2018-arxiv-Understanding disentangling in β-VAE.pdf
2018-arxiv-Variational Inference- A Review for Statisticians.pdf
2018-ICLR-A Framework for the Quan...tangled Representations-评判标准.pdf
2018-ICLR-Variational Inference of Di...nlabeled Observations-DIPVAE模型.pdf
2018-ICML-Disentangling by Factorising-factorVAE模型.pdf

按时间整理



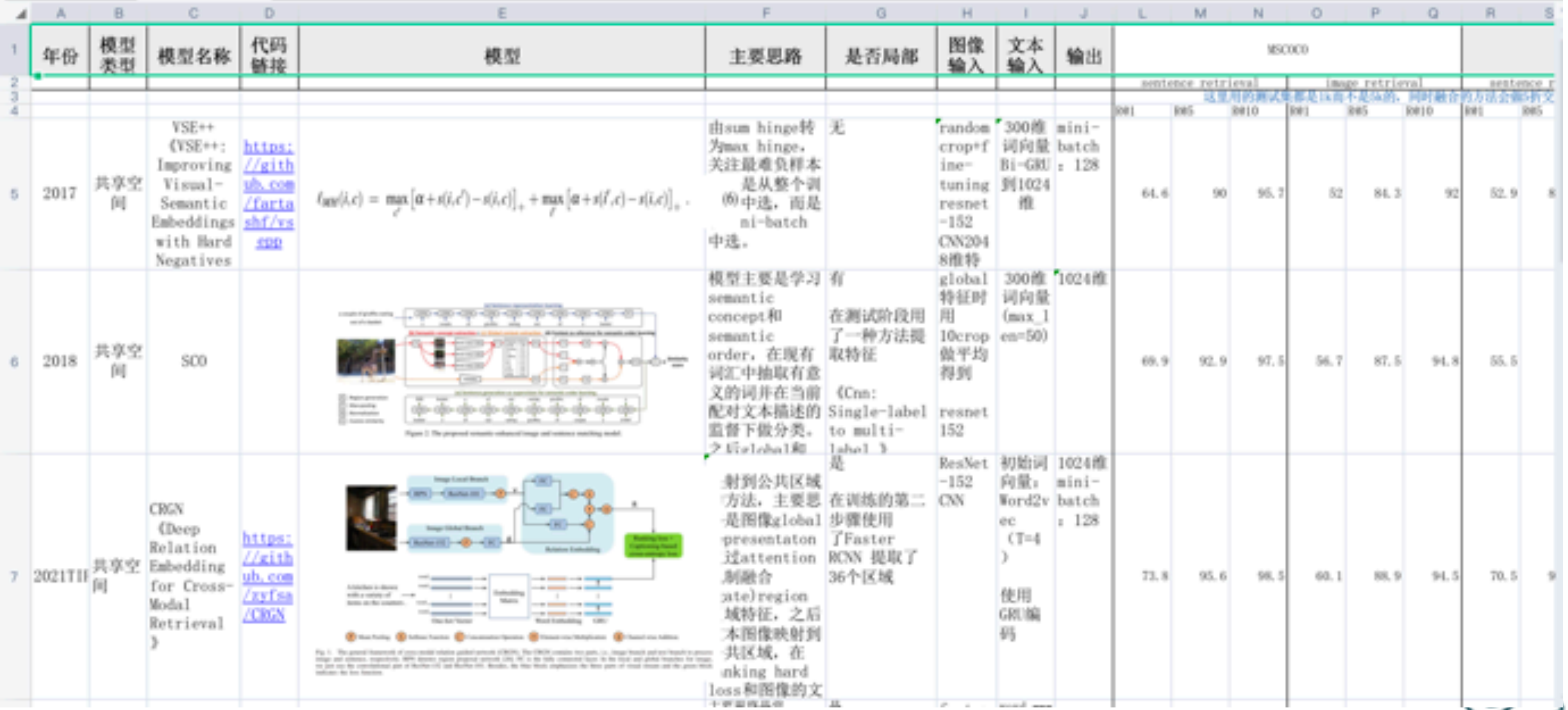
基础
- 综述
- NLP相关
- 多模态表示
- NLP预训练
- 多模态融合
- 多模态预训练
- 元学习
- 多模态对齐
- 生成式学习
- 生成对抗网
- 注意力相关
- 自监督学习
- 对抗攻击
- 图神经网络...
- 终身学习
- 零样本学习
- 可解释学习

应用
- VQA
- Visual Reas...
- image capti...
- video quest...
- visual dialog
- image text...
- referring ex...
- text2img
- visual relati...
- Voice-Visu...
- video capti...
- Language...
- 多模态机器...
- 虚假新闻检测
- Video mom...
- speech

按内容整理

Paper List——文献整理

- 可以将领域内的**经典模型**方法记录在 Excel 表格中，便于查找。



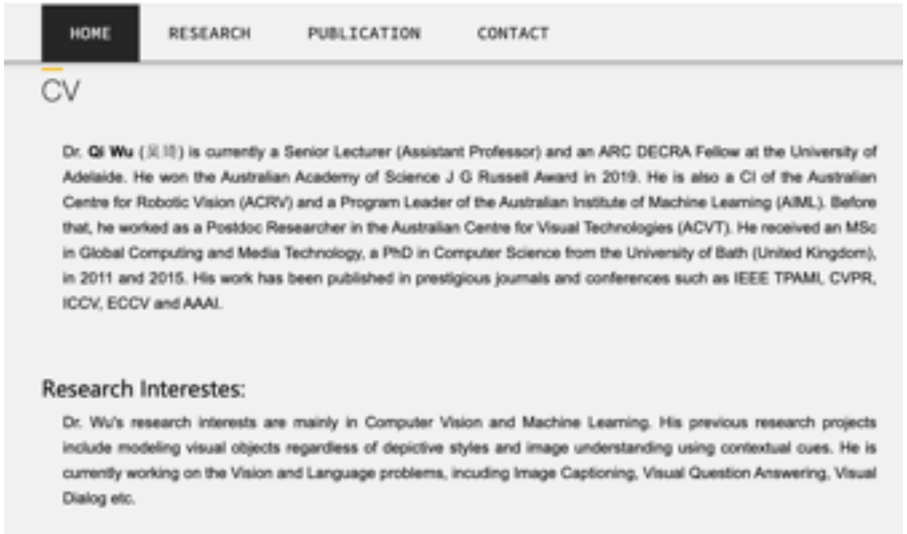
年份	模型类型	模型名称	代码链接	模型	主要思路	是否局部	图像输入	文本输入	输出	MSCOCO							
										sentence retrieval			image retrieval			sentence r	
										这里用的测试集都是1k而不是5k的，同时融合的方法会做5折交							
										R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5
2017	共享空间	VSE++ 《VSE++: Improving Visual-Semantic Embeddings with Hard Negatives	https://github.com/fartashf/vsepp	$\ell_{MH}(i,c) = \max_{c'}\left[\alpha+s(i,c')-s(i,c)\right]_+ + \max_{i'}\left[\alpha+s(i',c)-s(i,c)\right]_+ .$ (6)	由sum hinge转为max hinge，关注最难负样本是从整个训中选，而是ni-batch中选。	无	random crop+fine-tuning resnet-152 CNN2048维特	300维词向量Bi-GRU到1024维	mini-batch: 128	64.6	90	95.7	52	84.3	92	52.9	8
2018	共享空间	SCO			模型主要是学习semantic concept和semantic order，在现有词汇中抽取有意文的词并在当前配对文本描述的监督下做分类。之后global和	有 在测试阶段用了一种方法提取特征 《Cnn: Single-label to multi-label》	global特征时用10crop做平均得到 resnet152	300维词向量(max_len=50)	1024维	69.9	92.9	97.5	56.7	87.5	94.8	55.5	
2021TIP	共享空间	CRGN《Deep Relation Embedding for Cross-Modal Retrieval》	https://github.com/zyfsa/CRGN		射到公共区域方法，主要思是图像global presentaton过attention制融合ate)region域特征，之后本图像映射到共区域，在nking hard loss和图像的文	是 在训练的第二步骤使用了Faster RCNN 提取了36个区域	ResNet-152 CNN	初始词向量：Word2vec（T=4） 使用GRU编码	1024维 mini-batch: 128	73.8	95.6	98.5	60.1	88.9	94.5	70.5	9

经典模型整理

Idea List

- 有想法 or 问题，及时找老师或师兄师姐交流，及时记录
- 多看领域内的论坛、tutorial和workshop等 。可从B站、知乎、CCF 数字图书馆等也有很多资源可看
- 多与同领域内的同龄人多交流。
- 跟上大牛的步伐。可从大牛的个人主页获取信息。

CCF 数字图书馆



CV

Dr. Qi Wu (吴琦) is currently a Senior Lecturer (Assistant Professor) and an ARC DECRA Fellow at the University of Adelaide. He won the Australian Academy of Science J G Russell Award in 2019. He is also a CI of the Australian Centre for Robotic Vision (ACRV) and a Program Leader of the Australian Institute of Machine Learning (AIML). Before that, he worked as a Postdoc Researcher in the Australian Centre for Visual Technologies (ACVT). He received an MSc in Global Computing and Media Technology, a PhD in Computer Science from the University of Bath (United Kingdom), in 2011 and 2015. His work has been published in prestigious journals and conferences such as IEEE TPAMI, CVPR, ICCV, ECCV and AAAI.

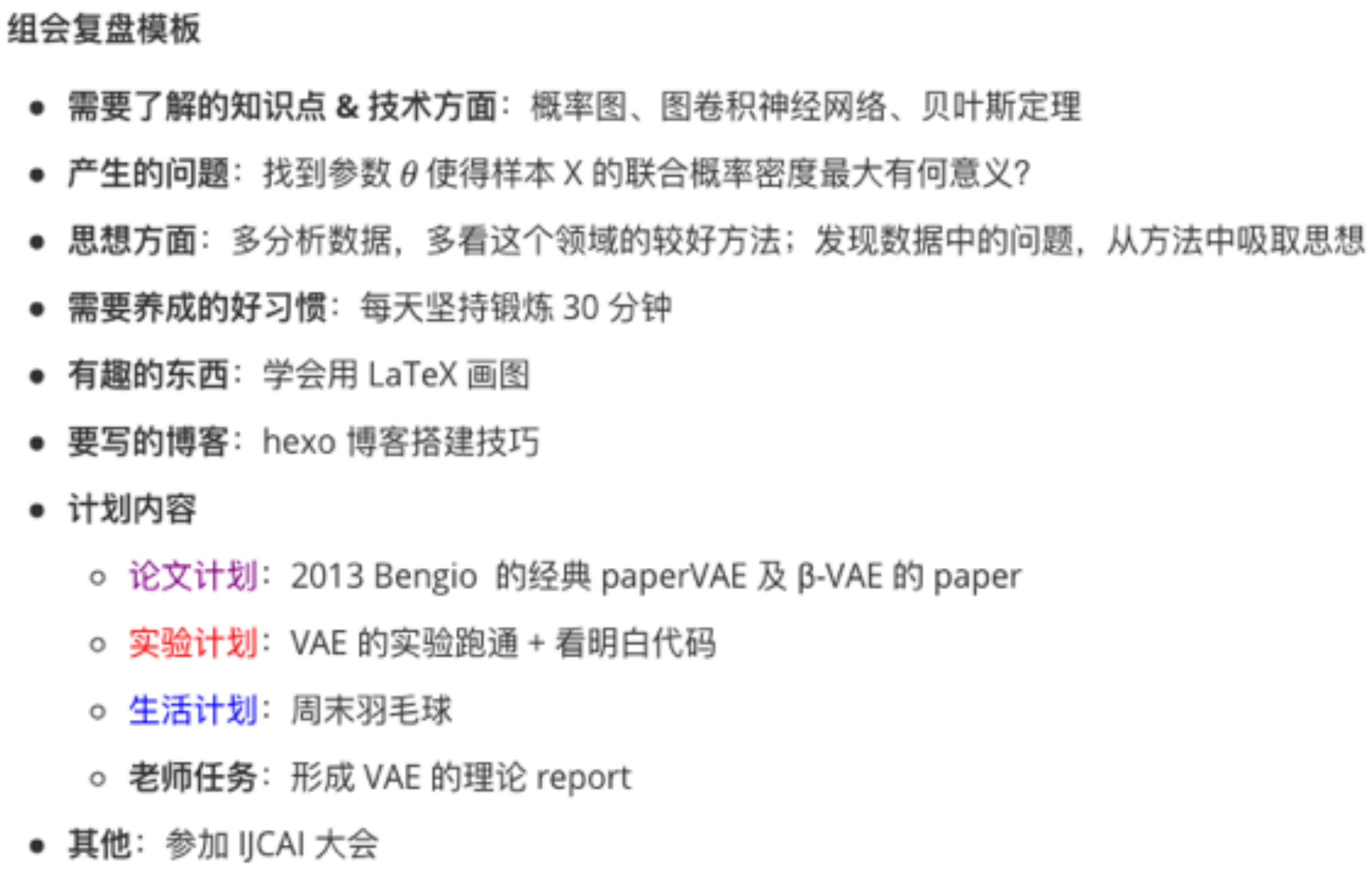
Research Interestes:

Dr. Wu's research interests are mainly in Computer Vision and Machine Learning. His previous research projects include modeling visual objects regardless of depictive styles and image understanding using contextual cues. He is currently working on the Vision and Language problems, incuding Image Captioning, Visual Question Answering, Visual Dialog etc.

大牛个人主页

Idea List

- 及时的进行复盘总结。可以记录不会的知识点、遇到的问题、好的想法、论文阅读计划、实验计划等。



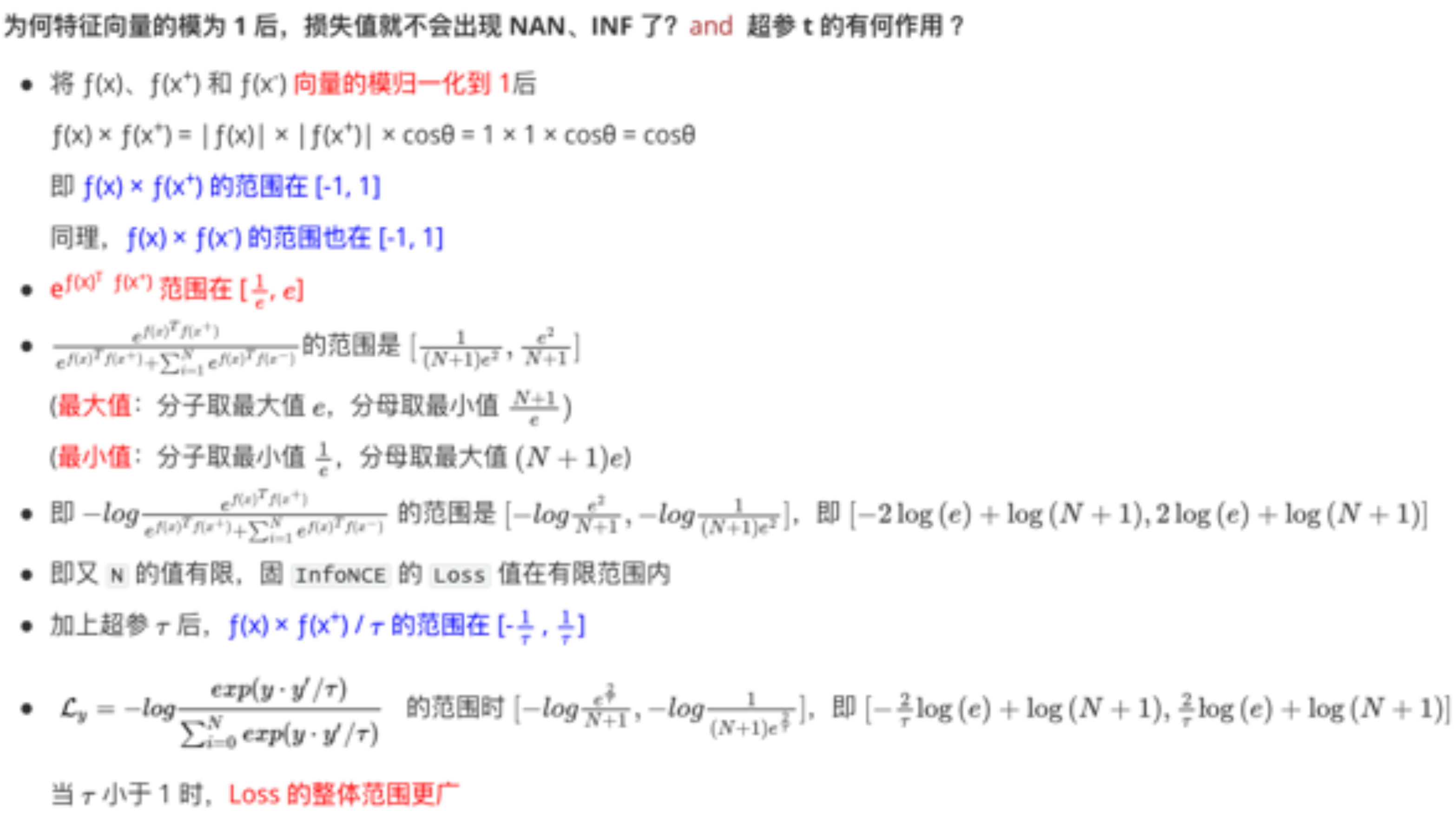
组会复盘模板

- 需要了解的知识点 & 技术方面：概率图、图卷积神经网络、贝叶斯定理
- 产生的问题：找到参数 θ 使得样本 X 的联合概率密度最大有何意义?
- 思想方面：多分析数据，多看这个领域的较好方法；发现数据中的问题，从方法中吸取思想
- 需要养成的好习惯：每天坚持锻炼 30 分钟
- 有趣的东西：学会用 LaTeX 画图
- 要写的博客：hexo 博客搭建技巧
- 计划内容
  - 论文计划：2013 Bengio 的经典 paperVAE 及 β-VAE 的 paper
  - 实验计划：VAE 的实验跑通 + 看明白代码
  - 生活计划：周末羽毛球
  - 老师任务：形成 VAE 的理论 report
- 其他：参加 IJCAI 大会

复盘总结示例

Math List

- 积累**基础的**数学知识
- 对模型中重要的**数学公式**进行**推导**



为何特征向量的模为 1 后，损失值就不会出现 NAN、INF 了? and 超参 t 的有何作用？

- 将 $f(x)$、$f(x^+)$ 和 $f(x^-)$ 向量的模归一化到 1后

  $f(x) \times f(x^+) = |f(x)| \times |f(x^+)| \times \cos\theta = 1 \times 1 \times \cos\theta = \cos\theta$

  即 $f(x) \times f(x^+)$ 的范围在 [-1, 1]

  同理，$f(x) \times f(x^-)$ 的范围也在 [-1, 1]

- $e^{f(x)^T f(x^+)}$ 范围在 $[\frac{1}{e}, e]$
- $\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+\sum_{i=1}^{N} e^{f(x)^T f(x^-)}}$ 的范围是 $[\frac{1}{(N+1)e^2}, \frac{e^2}{N+1}]$

  (最大值：分子取最大值 e，分母取最小值 $\frac{N+1}{e}$)

  (最小值：分子取最小值 $\frac{1}{e}$，分母取最大值 $(N+1)e$)

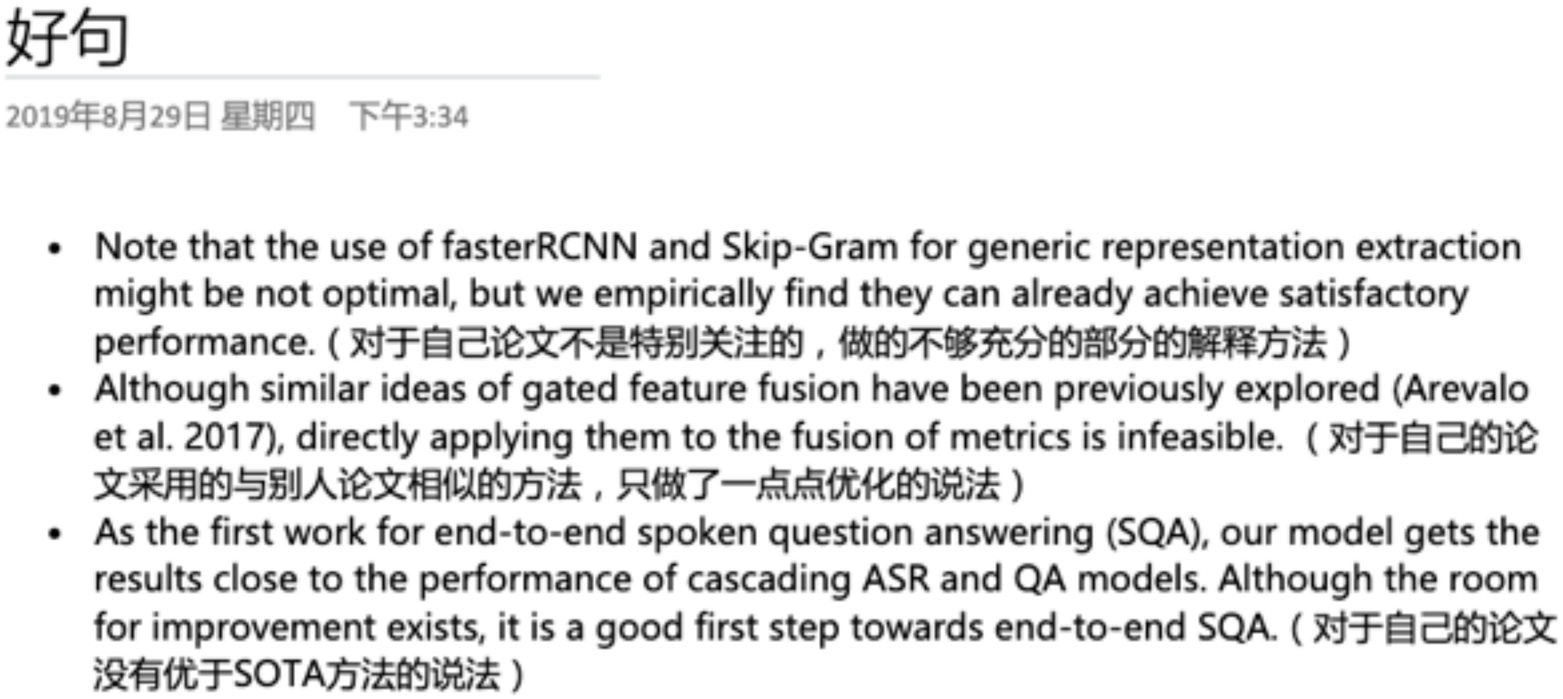
- 即 $-log\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+\sum_{i=1}^{N} e^{f(x)^T f(x^-)}}$ 的范围是 $[-log\frac{e^2}{N+1}, -log\frac{1}{(N+1)e^2}]$，即 $[-2\log(e)+\log(N+1), 2\log(e)+\log(N+1)]$
- 即又 `N` 的值有限，固 `InfoNCE` 的 `Loss` 值在有限范围内
- 加上超参 τ 后，$f(x) \times f(x^+) / \tau$ 的范围在 $[-\frac{1}{\tau}, \frac{1}{\tau}]$
- $\mathcal{L}_y = -log\frac{exp(y \cdot y'/\tau)}{\sum_{i=0}^{N} exp(y \cdot y'/\tau)}$ 的范围时 $[-log\frac{e^{\frac{2}{\tau}}}{N+1}, -log\frac{1}{(N+1)e^{\frac{2}{\tau}}}]$，即 $[-\frac{2}{\tau}\log(e)+\log(N+1), \frac{2}{\tau}\log(e)+\log(N+1)]$

  当 τ 小于 1 时，Loss 的整体范围更广

数学知识总结示例

English List

- 积累专业知识相关的英语词汇
- 积累论文中的好句、好词



好句

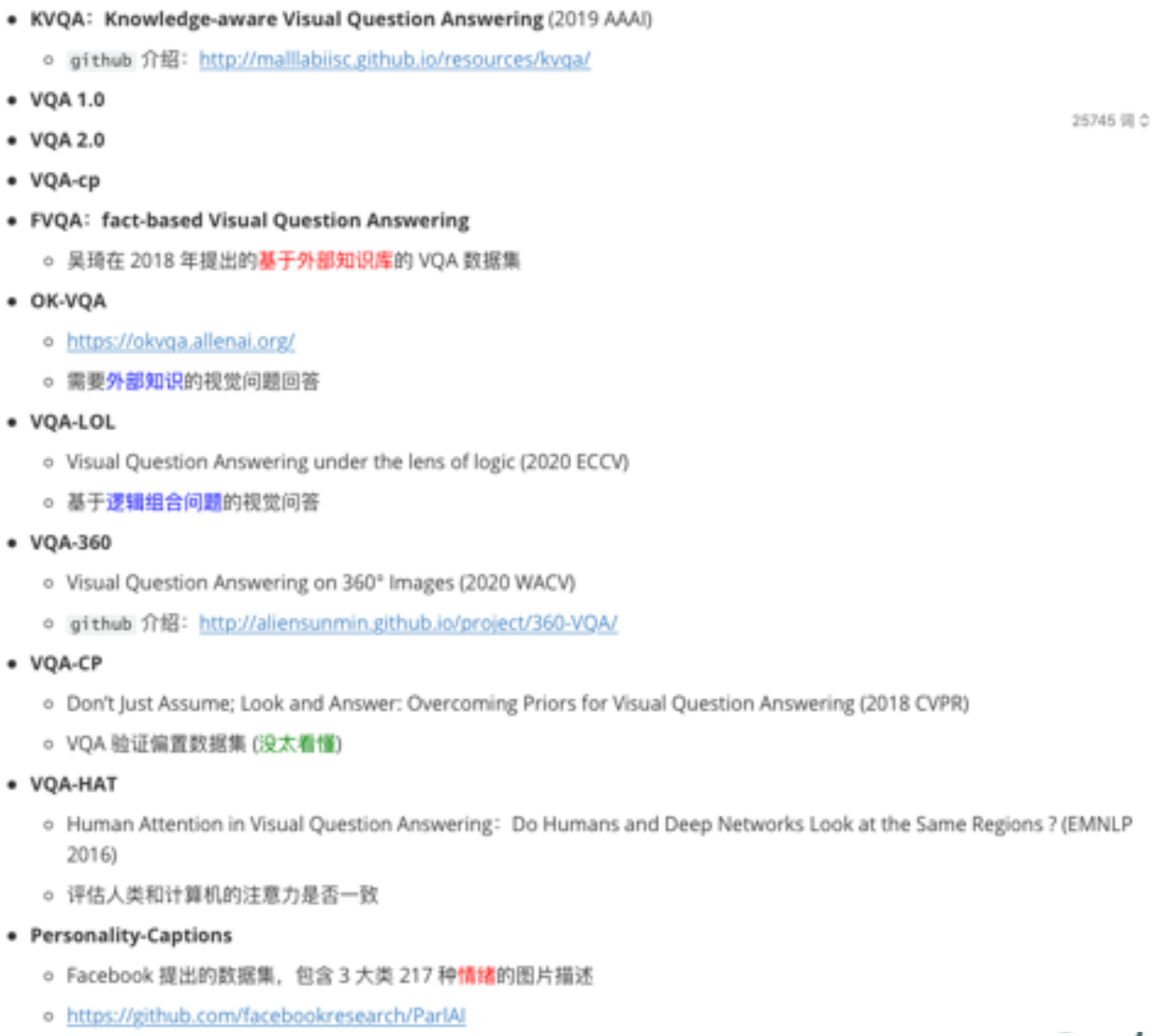
2019年8月29日 星期四 下午3:34

- Note that the use of fasterRCNN and Skip-Gram for generic representation extraction might be not optimal, but we empirically find they can already achieve satisfactory performance. (对于自己论文不是特别关注的，做的不够充分的部分的解释方法)
- Although similar ideas of gated feature fusion have been previously explored (Arevalo et al. 2017), directly applying them to the fusion of metrics is infeasible. （对于自己的论文采用的与别人论文相似的方法，只做了一点点优化的说法）
- As the first work for end-to-end spoken question answering (SQA), our model gets the results close to the performance of cascading ASR and QA models. Although the room for improvement exists, it is a good first step towards end-to-end SQA. (对于自己的论文没有优于SOTA方法的说法)

英语总结示例

Code List

- 数据集积累
- 基准模型积累
- 代码 trick 积累
- 实验结果积累

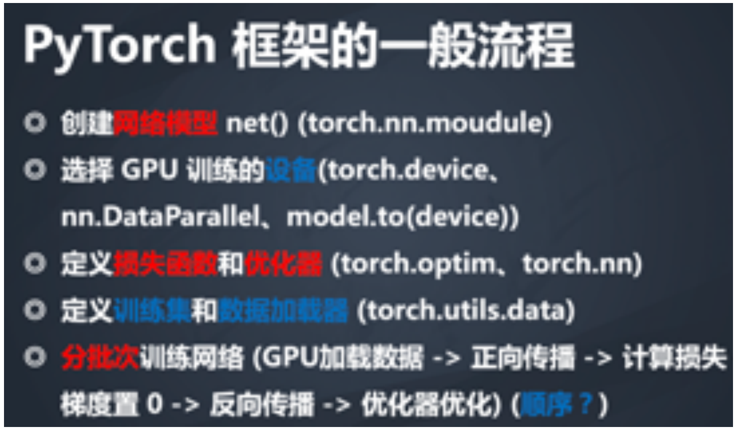


- **KVQA：Knowledge-aware Visual Question Answering** (2019 AAAI)
  - `github` 介绍：http://malllabiisc.github.io/resources/kvqa/
- **VQA 1.0**
- **VQA 2.0**
- **VQA-cp**
- **FVQA：fact-based Visual Question Answering**
  - 吴琦在 2018 年提出的基于外部知识库的 VQA 数据集
- **OK-VQA**
  - https://okvqa.allenai.org/
  - 需要外部知识的视觉问题回答
- **VQA-LOL**
  - Visual Question Answering under the lens of logic (2020 ECCV)
  - 基于逻辑组合问题的视觉问答
- **VQA-360**
  - Visual Question Answering on 360° Images (2020 WACV)
  - `github` 介绍：http://aliensunmin.github.io/project/360-VQA/
- **VQA-CP**
  - Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering (2018 CVPR)
  - VQA 验证偏置数据集 (没太看懂)
- **VQA-HAT**
  - Human Attention in Visual Question Answering：Do Humans and Deep Networks Look at the Same Regions ? (EMNLP 2016)
  - 评估人类和计算机的注意力是否一致
- **Personality-Captions**
  - Facebook 提出的数据集，包含 3 大类 217 种情绪的图片描述
  - https://github.com/facebookresearch/ParlAI

数据集整理

Code List

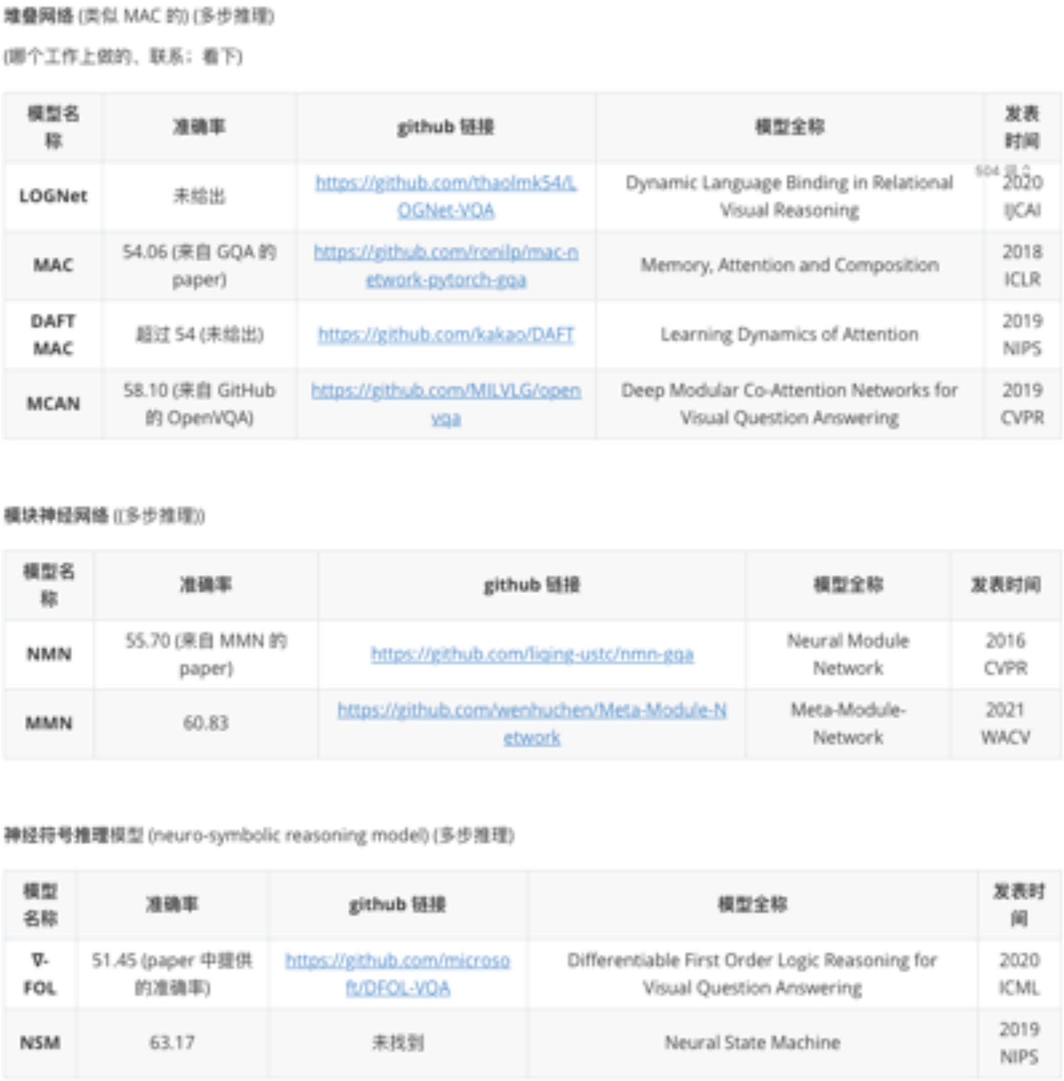
- 数据集积累
- 基准模型积累
- 代码 trick 积累
- 实验结果积累



PyTorch 框架的一般流程

- 创建网络模型 net() (torch.nn.moudule)
- 选择 GPU 训练的设备(torch.device、nn.DataParallel、model.to(device))
- 定义损失函数和优化器 (torch.optim、torch.nn)
- 定义训练集和数据加载器 (torch.utils.data)
- 分批次训练网络 (GPU加载数据 -> 正向传播 -> 计算损失梯度置 0 -> 反向传播 -> 优化器优化) (顺序？)

Trick 积累



堆叠网络 (类似 MAC 的) (多步推理)

(哪个工作上做的、联系：看下)

模型名称	准确率	github 链接	模型全称	发表时间
LOGNet	未给出	https://github.com/thaolmk54/LOGNet-VQA	Dynamic Language Binding in Relational Visual Reasoning	2020 IJCAI
MAC	54.06 (来自 GQA 的 paper)	https://github.com/ronilp/mac-network-pytorch-gqa	Memory, Attention and Composition	2018 ICLR
DAFT MAC	超过 54 (未给出)	https://github.com/kakao/DAFT	Learning Dynamics of Attention	2019 NIPS
MCAN	58.10 (来自 GitHub 的 OpenVQA)	https://github.com/MILVLG/openvqa	Deep Modular Co-Attention Networks for Visual Question Answering	2019 CVPR

模块神经网络 ([多步推理])

模型名称	准确率	github 链接	模型全称	发表时间
NMN	55.70 (来自 MMN 的 paper)	https://github.com/liqing-ustc/nmn-gqa	Neural Module Network	2016 CVPR
MMN	60.83	https://github.com/wenhuchen/Meta-Module-Network	Meta-Module-Network	2021 WACV

神经符号推理模型 (neuro-symbolic reasoning model) (多步推理)

模型名称	准确率	github 链接	模型全称	发表时间
∇-FOL	51.45 (paper 中提供的准确率)	https://github.com/microsoft/DFOL-VQA	Differentiable First Order Logic Reasoning for Visual Question Answering	2020 ICML
NSM	63.17	未找到	Neural State Machine	2019 NIPS

模型积累

CogModal GROUP

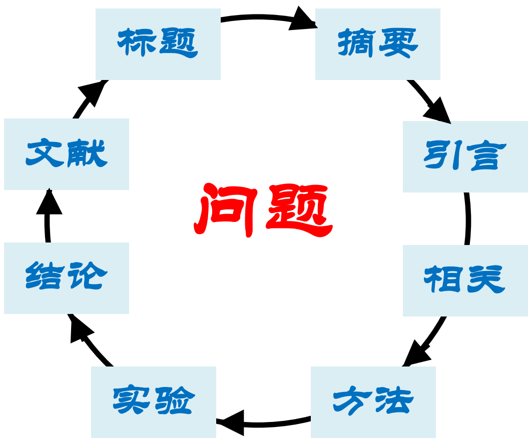
本讲总结
与心得分享

One more thing...

Personal experience about PhD study,
which is definitely not the best strategy......

Take Home Message

写作思路


问题

标题

摘要

引言

相关

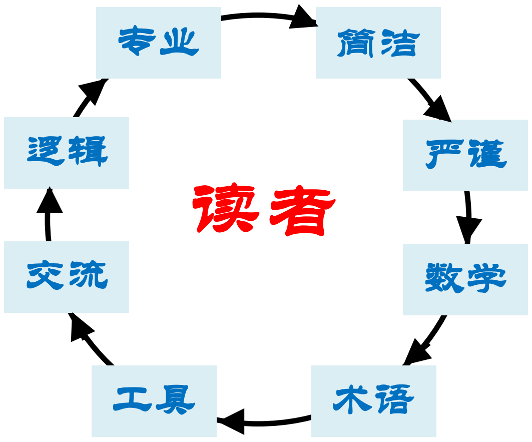
方法

实验

结论

文献

英文规范


读者

专业

简洁

严谨

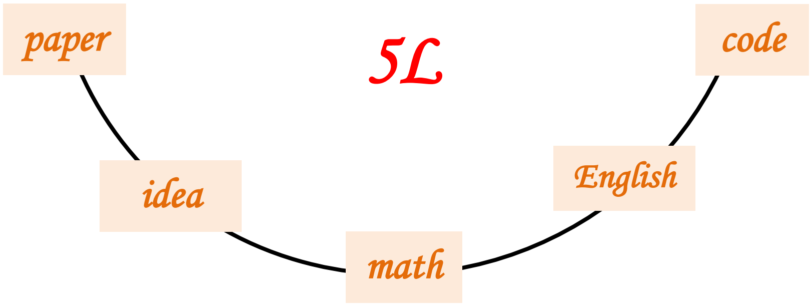
数学

术语

工具

交流

逻辑


5L

paper

idea

math

English

code

日常积累

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