科研与英文学术论文写作

第五讲 英文学术论文之写作思路

——实验、结论和参考文献

于静 副研究员

中国科学院信息工程研究所

课程主页：https://mmlab-iie.github.io/course/

2022.07 @ Bilibili

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

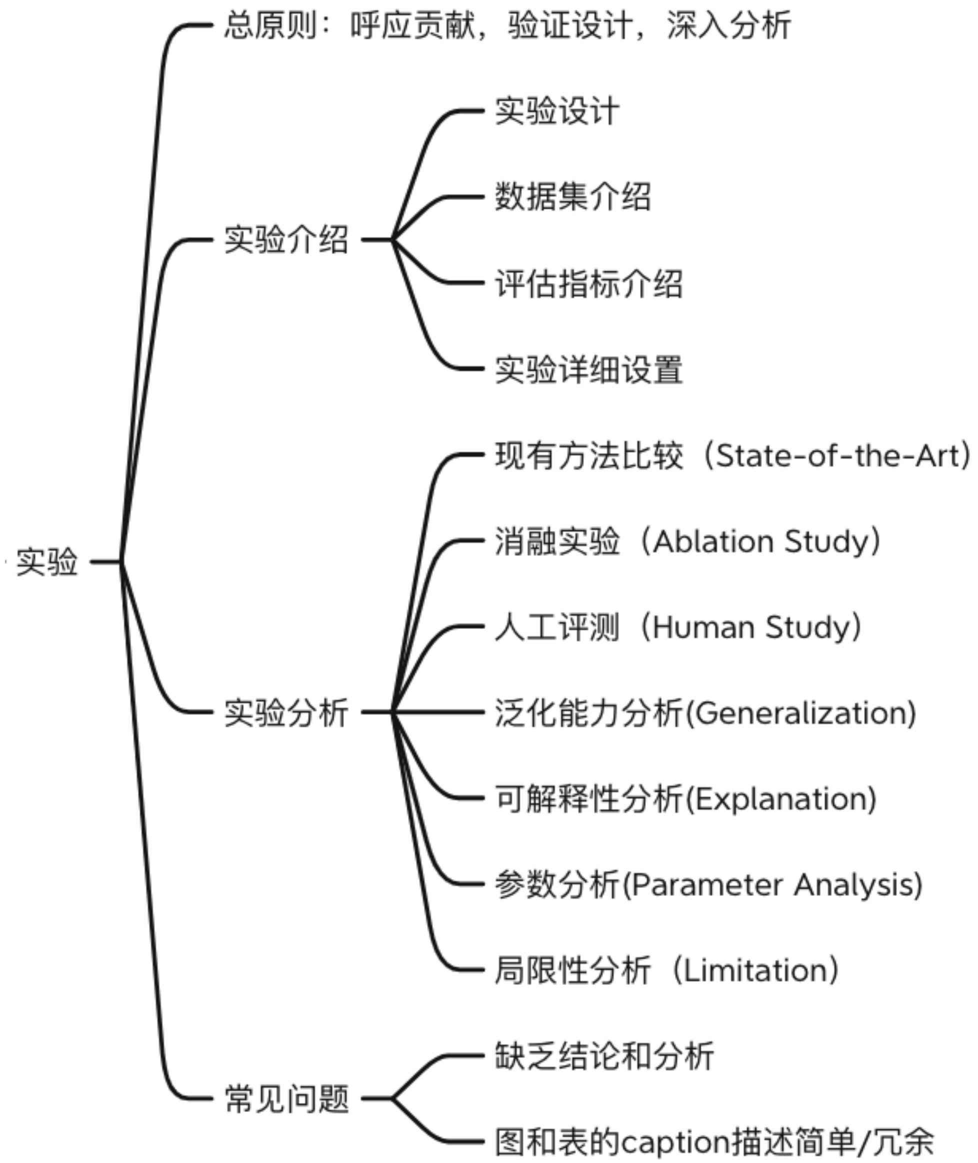
中国科学院大学
University of Chinese Academy of Sciences

一篇论文的组成——实验

基本要求

- **一致：** 支撑理论/方法、动机
- **核心：** 提供重要实验结果
- **诚实：** 不只展示最佳个例
- **分析：** 给出结果的合理解释
- **局限：** 给出方法的能力边界

学习优秀论文的实验写法！



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

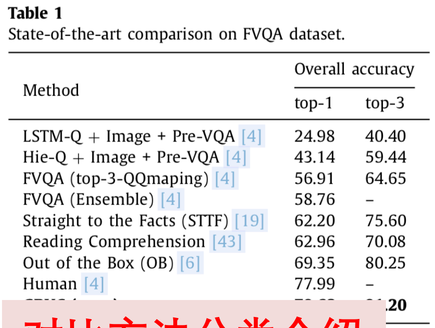


Table 1
State-of-the-art comparison on FVQA dataset.

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

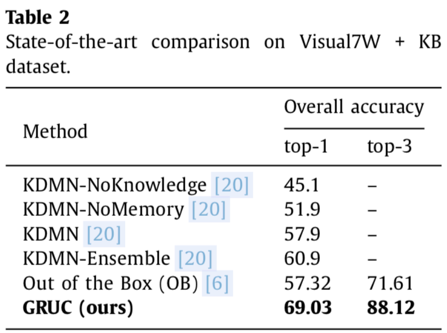


Table 2
State-of-the-art comparison on Visual7W + KB dataset.

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**

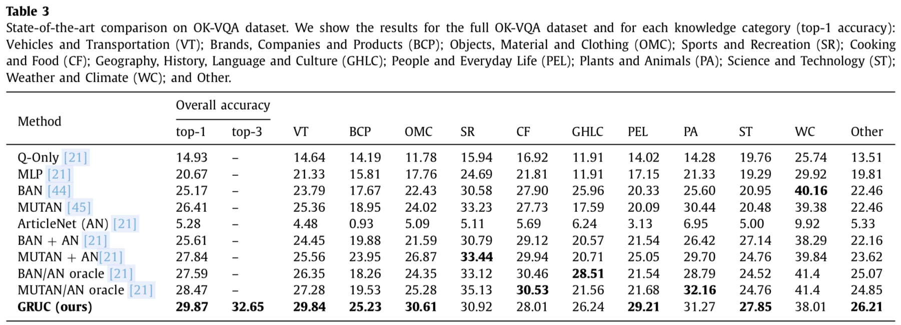
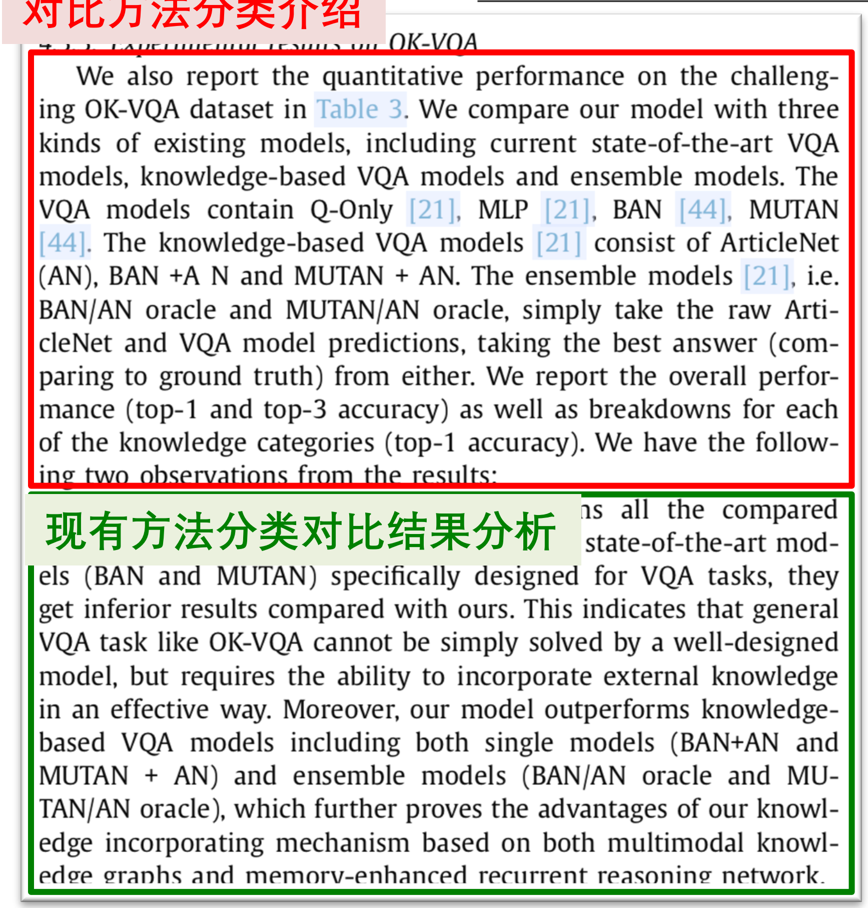


Table 3
State-of-the-art comparison on OK-VQA dataset. We show the results for the full OK-VQA dataset and for each knowledge category (top-1 accuracy): Vehicles and Transportation (VT); Brands, Companies and Products (BCP); Objects, Material and Clothing (OMC); Sports and Recreation (SR); Cooking and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



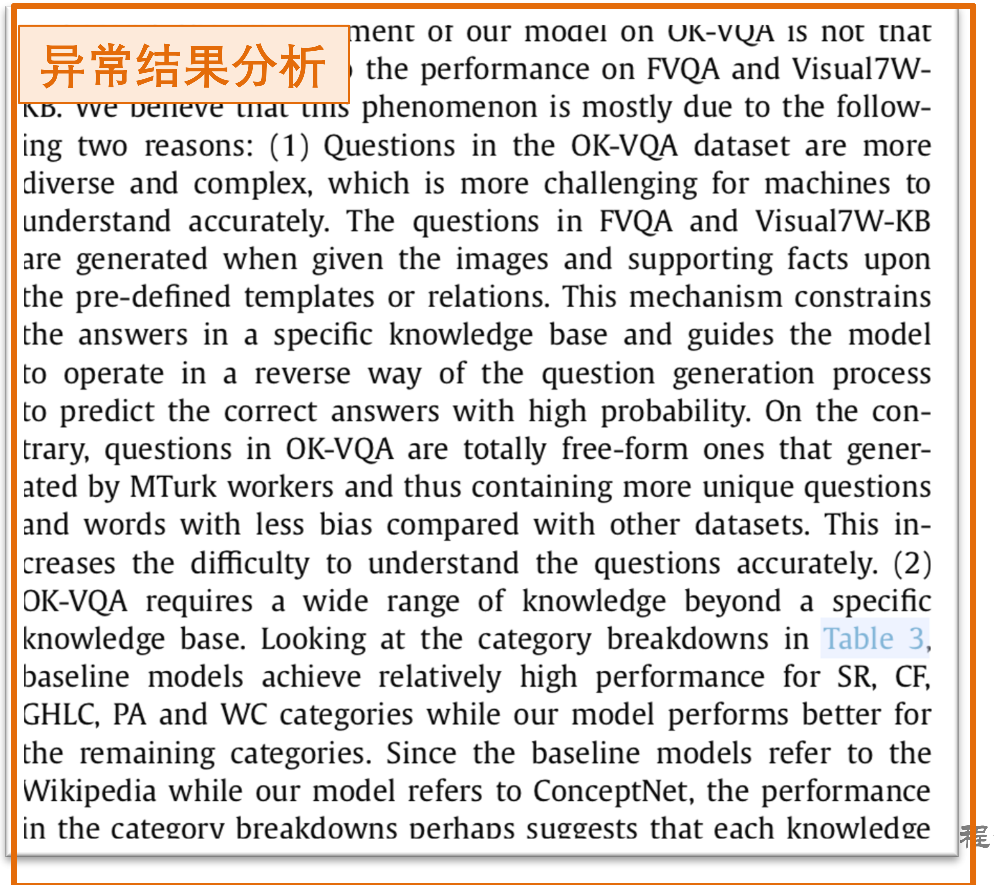
对比方法分类介绍

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

[illegible] ns all the compared [illegible] 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

[illegible] ment of our model on OK-VQA is not that [illegible] 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

Ablation

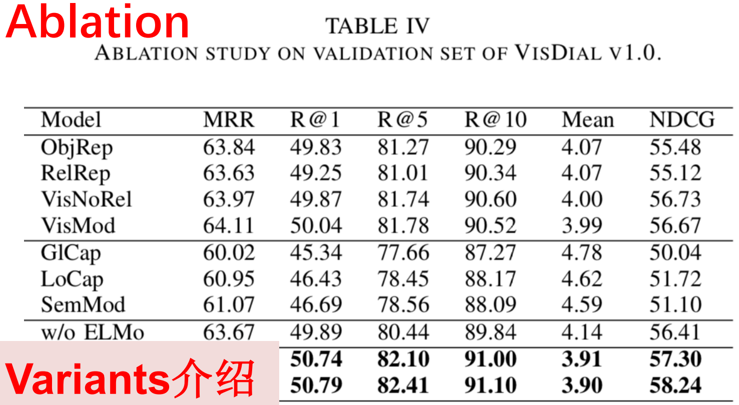
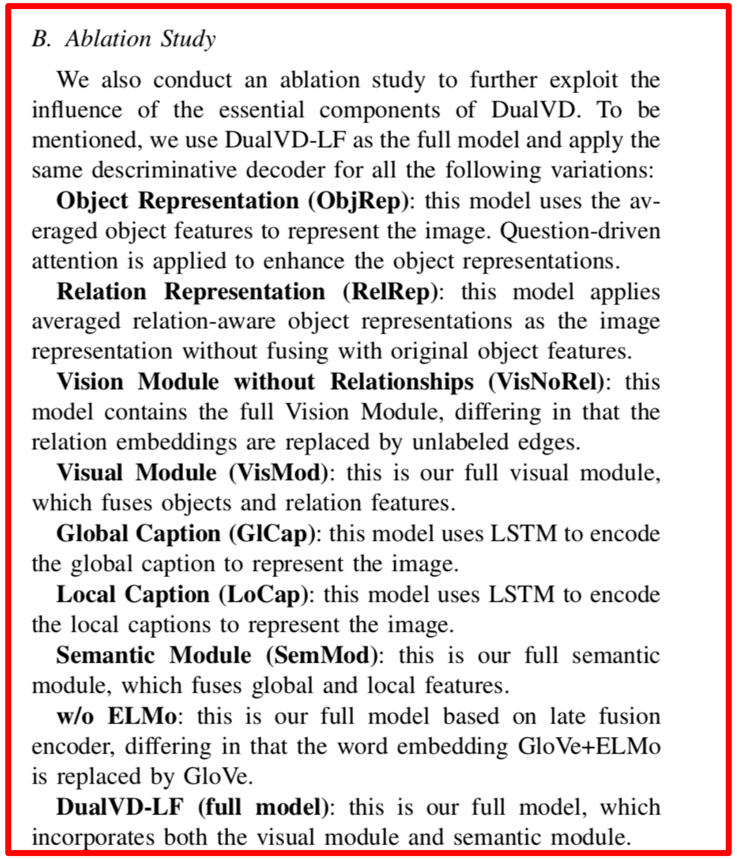


TABLE IV
ABLATION STUDY ON VALIDATION SET OF VISDIAL V1.0.

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.

不同维度对比分析

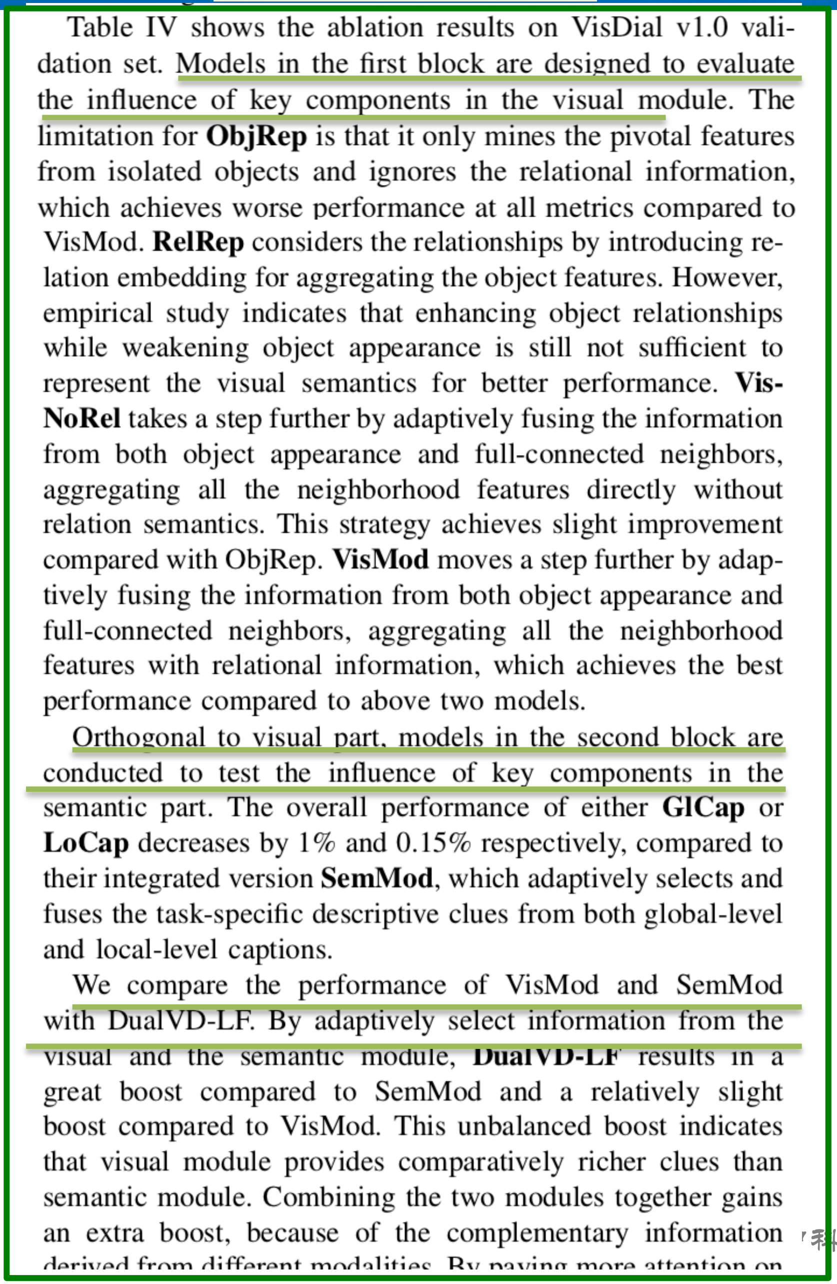


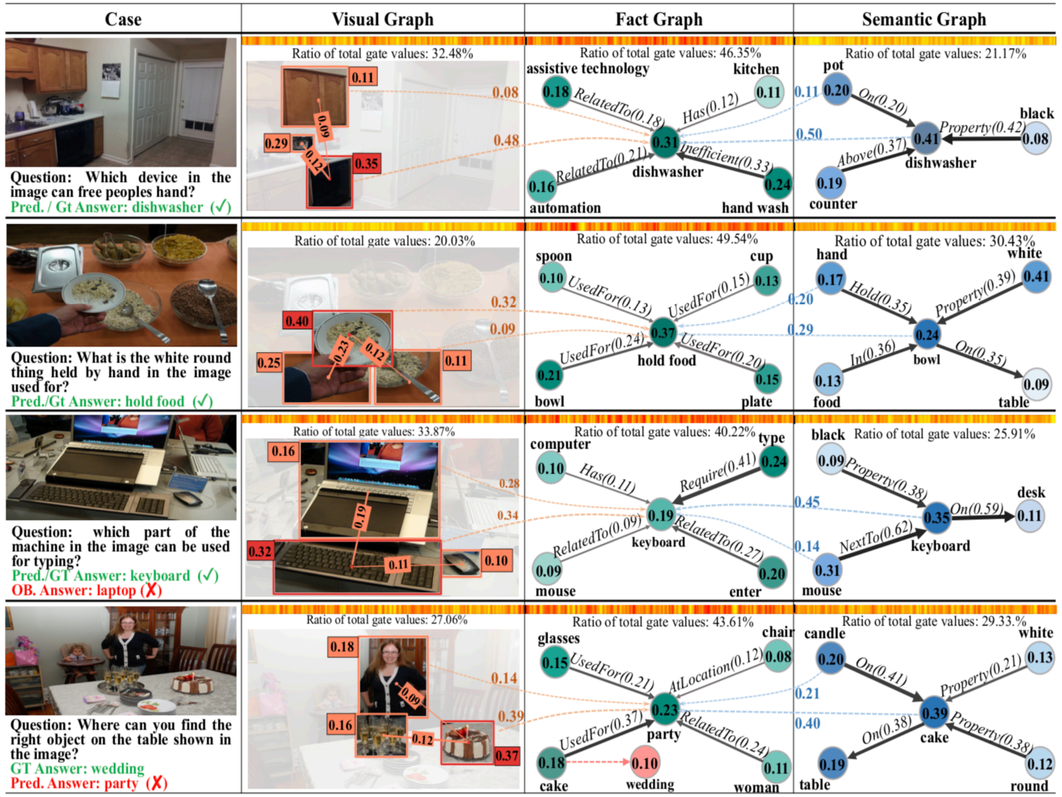
Table IV shows the ablation results on VisDial v1.0 validation set. Models in the first block are designed to evaluate the influence of key components in the visual module. The limitation for **ObjRep** is that it only mines the pivotal features from isolated objects and ignores the relational information,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

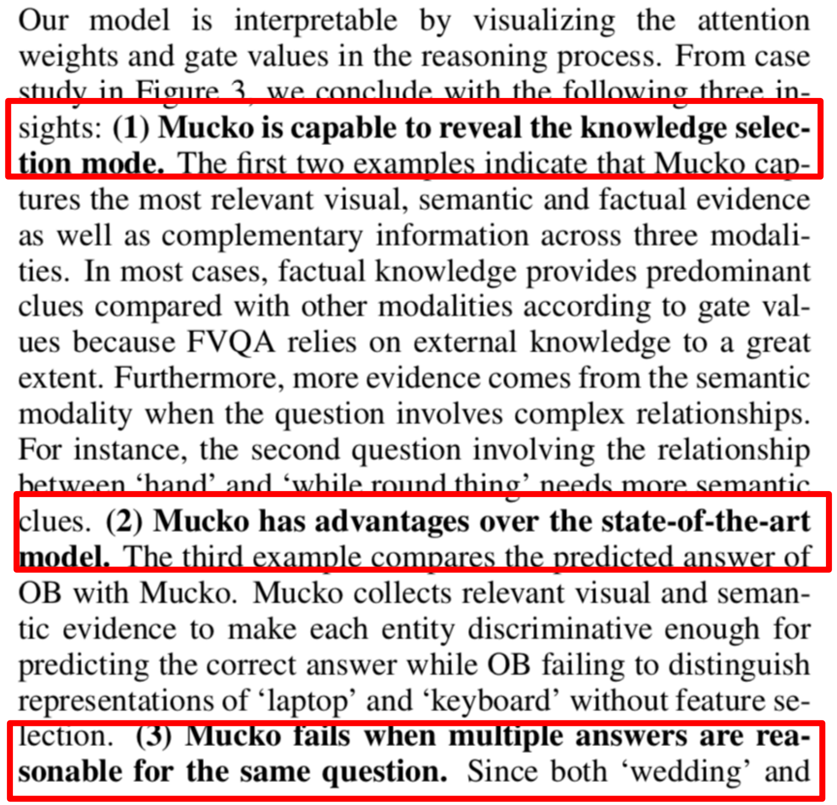
Orthogonal to visual part, models in the second block are conducted to test the influence of key components in the semantic part. The overall performance of either **GlCap** or **LoCap** decreases by 1% and 0.15% respectively, compared to their integrated version **SemMod**, which adaptively selects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

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

一篇论文的组成——实验

Interpretation

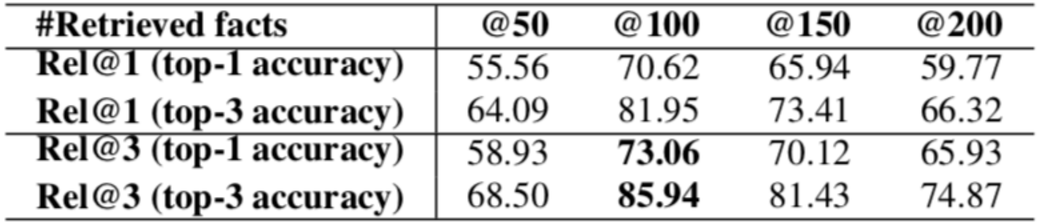

Case
Visual Graph
Fact Graph
Semantic Graph
Question: Which device in the image can free peoples hand?
Pred. / Gt Answer: dishwasher (✓)
Ratio of total gate values: 32.48%
Ratio of total gate values: 46.35%
assistive technology
kitchen
RelatedTo(0.18)
Has(0.12)
RelatedTo(0.21)
Inefficient(0.33)
dishwasher
automation
hand wash
Ratio of total gate values: 21.17%
pot
On(0.20)
black
Property(0.42)
Above(0.37)
dishwasher
counter
Question: What is the white round thing held by hand in the image used for?
Pred./Gt Answer: hold food (✓)
Ratio of total gate values: 20.03%
Ratio of total gate values: 49.54%
spoon
cup
UsedFor(0.13)
UsedFor(0.15)
UsedFor(0.24)
UsedFor(0.20)
hold food
bowl
plate
Ratio of total gate values: 30.43%
hand
white
Hold(0.35)
Property(0.39)
In(0.36)
On(0.35)
bowl
food
table
Question: which part of the machine in the image can be used for typing?
Pred./GT Answer: keyboard (✓)
OB. Answer: laptop (✗)
Ratio of total gate values: 33.87%
Ratio of total gate values: 40.22%
computer
type
Has(0.11)
Require(0.41)
RelatedTo(0.09)
RelatedTo(0.27)
keyboard
mouse
enter
black
Ratio of total gate values: 25.91%
Property(0.38)
desk
On(0.59)
NextTo(0.62)
keyboard
mouse
Question: Where can you find the right object on the table shown in the image?
GT Answer: wedding
Pred. Answer: party (✗)
Ratio of total gate values: 27.06%
Ratio of total gate values: 43.61%
glasses
chair
UsedFor(0.21)
AtLocation(0.12)
UsedFor(0.37)
RelatedTo(0.24)
party
cake
wedding
woman
Ratio of total gate values: 29.33 %
candle
white
On(0.41)
Property(0.21)
On(0.38)
Property(0.38)
cake
table
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

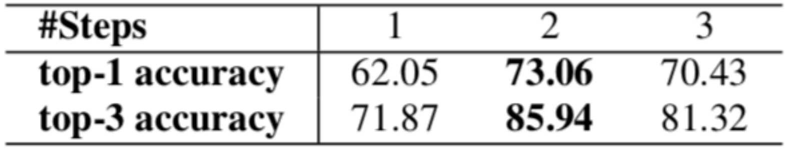
一篇论文的组成——实验

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

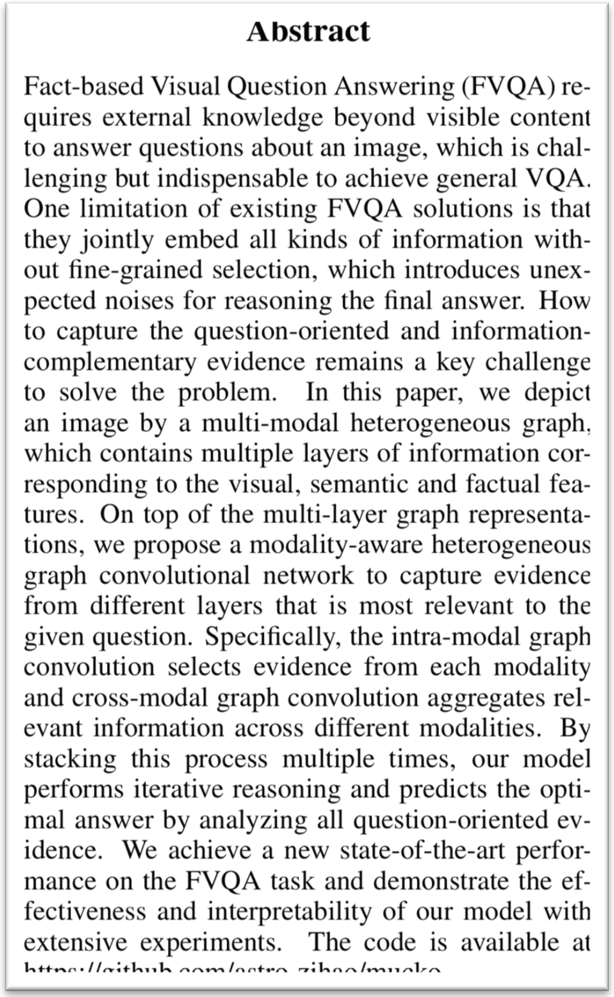
- 泛化能力（更多不同任务）
- 局部模块效果验证
- 存在的局限性
- …

一篇论文的组成——结论

基本要求

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

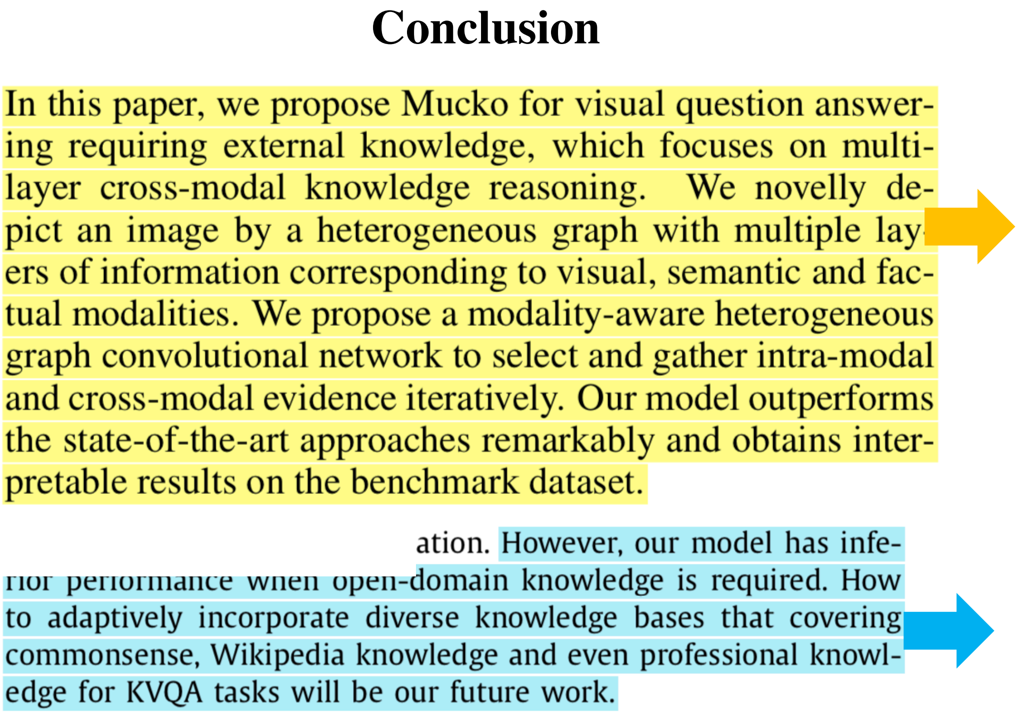
一篇论文的组成——结论



Abstract

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

基本要求

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

一篇论文的组成——参考文献

基本要求

- 不遗漏，查全
- 按照会议/期刊既定格式
- 常见错误：大小写、全称/缩写、漏写、名字错拼

论文写作及修改过程

Project Slides

基于***的视觉问答

姓名：丁阳

项目开始时间：2020.8.12
最后修改时间：2021.7.9

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

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

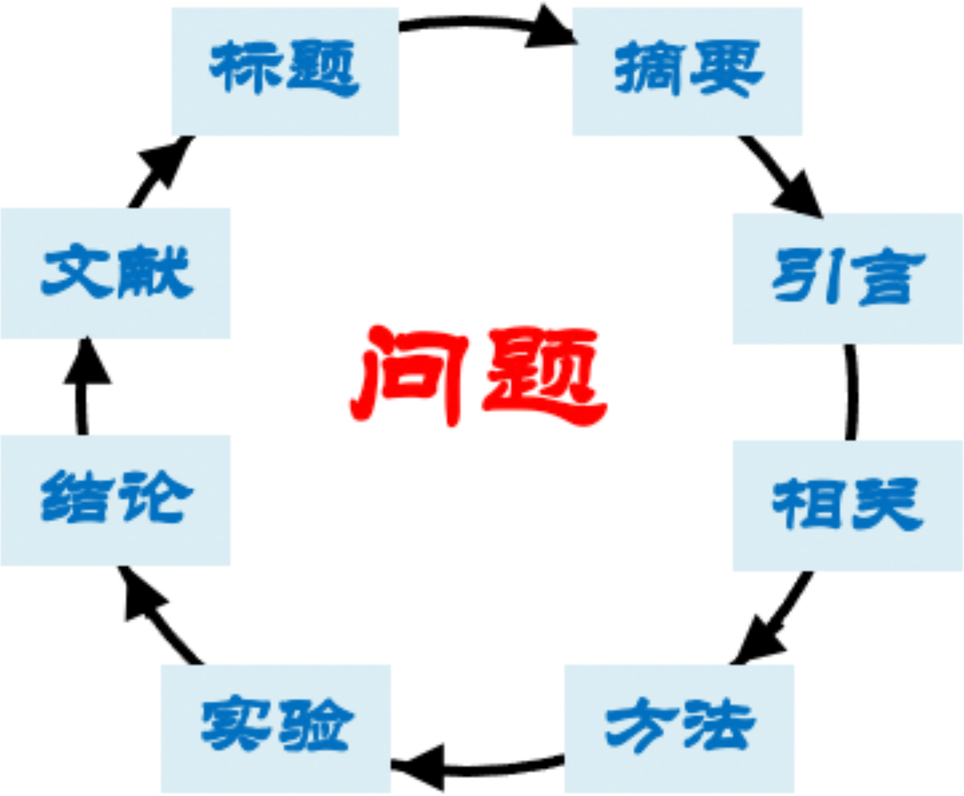
- ✓ CVPR
- ✓ 视觉-语言
- ✓ 视觉问答

Deadline

完整研究

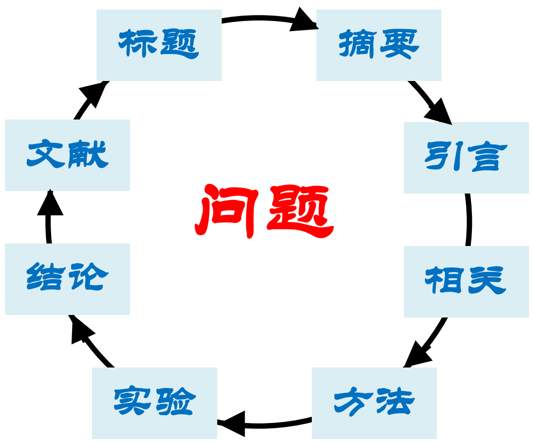
Take Home Message

写作思路



Take Home Message

写作思路


问题

标题

摘要

引言

相关

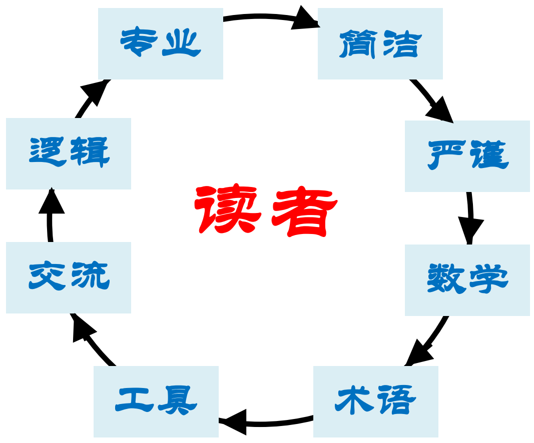
方法

实验

结论

文献

英文规范


读者

专业

简洁

严谨

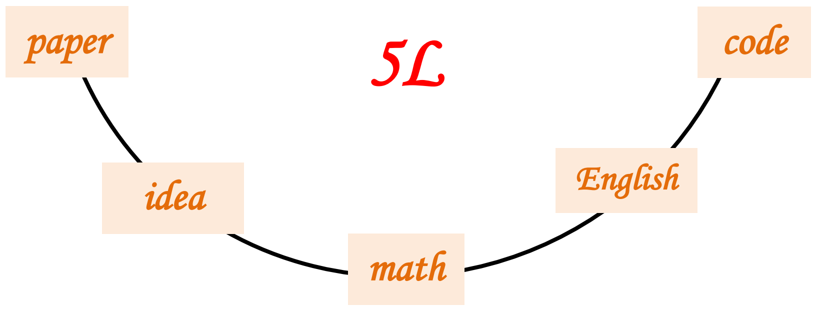
数学

术语

工具

交流

逻辑


5L

paper

idea

math

English

code

日常积累

参考资料

参考书籍

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

参考课程

- 董彬，北京大学，北京国际数学研究中心，《研究型学习》

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

As long as I am dreaming, believing and doing,

I can go anywhere and achieve anything!

于静

邮箱：yujing02@iie.ac.cn

课程主页： https://mmlab-iie.github.io/course/

研究组主页：https://mmlab-iie.github.io/

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

课程主页 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