科研与英文学术论文写作

第九讲 英文学术论文之英文规范
——如何做到简洁与严谨

于静 副研究员

中国科学院信息工程研究所

系列报告主页：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

英语写作规范——八项注意

1. 精简的表达方式
2. 严谨的叙述逻辑

不太容易

3. 专业的学术用语
4. 规范的符号使用
5. 标准的学术术语
6. 客观的图表绘制
7. 正确的文献引用
8. 坚守的学术道德

很容易

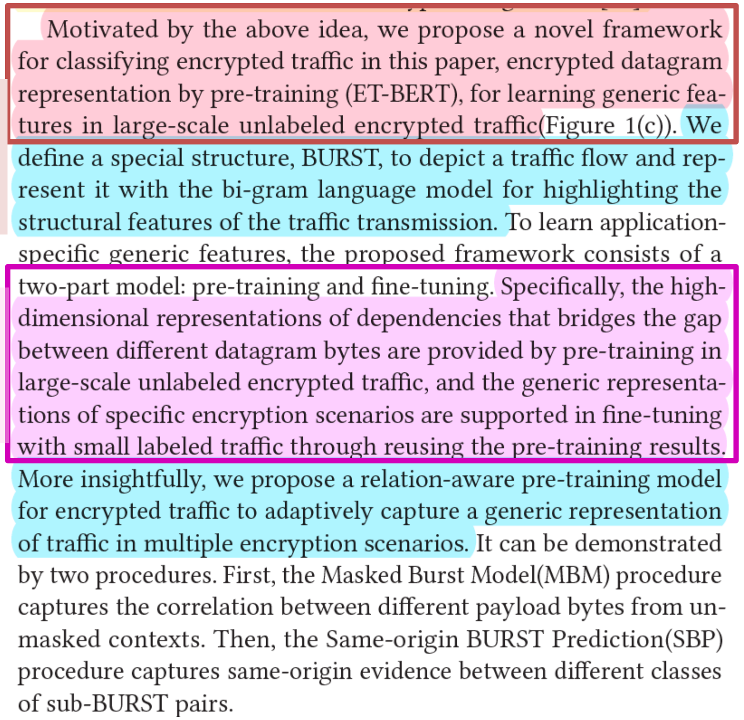
英语写作规范（1）——精简的表达方式

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！

——施一公

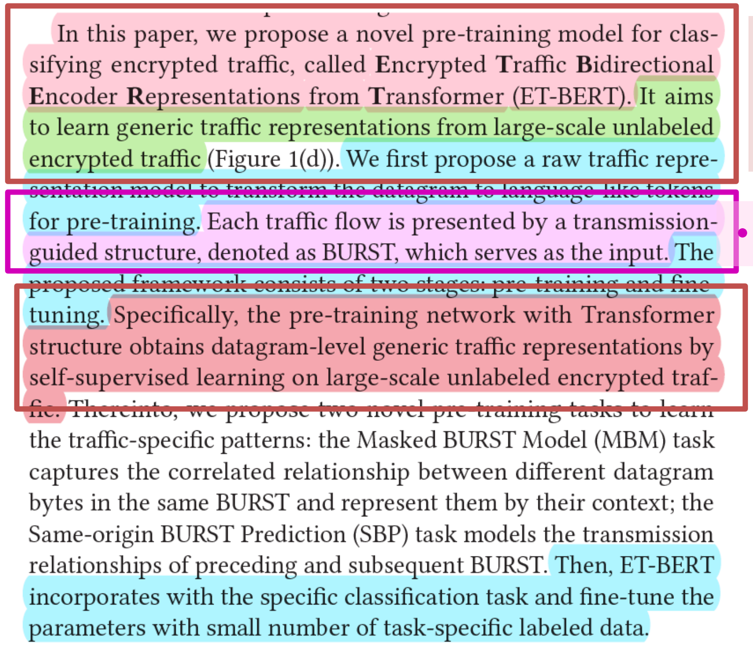
- 表达：一句话3个从句
- 问题：混淆方法和问题
- 无重点！有语病！

- 表达：一句话44个单词
- 问题：加入太多细节
- 无重点！有语病！



Motivated by the above idea, we propose a novel framework for classifying encrypted traffic in this paper, encrypted datagram representation by pre-training (ET-BERT), for learning generic features in large-scale unlabeled encrypted traffic(Figure 1(c)). We define a special structure, BURST, to depict a traffic flow and represent it with the bi-gram language model for highlighting the structural features of the traffic transmission. To learn application-specific generic features, the proposed framework consists of a two-part model: pre-training and fine-tuning. Specifically, the high-dimensional representations of dependencies that bridges the gap between different datagram bytes are provided by pre-training in large-scale unlabeled encrypted traffic, and the generic representations of specific encryption scenarios are supported in fine-tuning with small labeled traffic through reusing the pre-training results. More insightfully, we propose a relation-aware pre-training model for encrypted traffic to adaptively capture a generic representation of traffic in multiple encryption scenarios. It can be demonstrated by two procedures. First, the Masked Burst Model(MBM) procedure captures the correlation between different payload bytes from unmasked contexts. Then, the Same-origin BURST Prediction(SBP) procedure captures same-origin evidence between different classes of sub-BURST pairs.

最初版本



In this paper, we propose a novel pre-training model for classifying encrypted traffic, called **E**ncrypted **T**raffic **B**idirectional **E**ncoder **R**epresentations from **T**ransformer (ET-BERT). It aims to learn generic traffic representations from large-scale unlabeled encrypted traffic (Figure 1(d)). We first propose a raw traffic representation model to transform the datagram to language-like tokens for pre-training. Each traffic flow is presented by a transmission-guided structure, denoted as BURST, which serves as the input. The proposed framework consists of two stages: pre-training and fine-tuning. Specifically, the pre-training network with Transformer structure obtains datagram-level generic traffic representations by self-supervised learning on large-scale unlabeled encrypted traffic. Thereinto, we propose two novel pre-training tasks to learn the traffic-specific patterns: the Masked BURST Model (MBM) task captures the correlated relationship between different datagram bytes in the same BURST and represent them by their context; the Same-origin BURST Prediction (SBP) task models the transmission relationships of preceding and subsequent BURST. Then, ET-BERT incorporates with the specific classification task and fine-tune the parameters with small number of task-specific labeled data.

最终版本

- 一句话命名方法及问题
- 一句话突出方法目标
- 一句话介绍方法创新

- 一句话表达核心方法思路

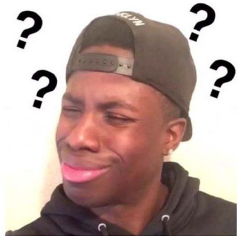
Tips: （1）一句话只表达一个意思！ （2）减少中文翻译英文！ （3）避免重复表达！

英语写作规范（2）——严谨的叙述逻辑

假设审稿人还有一个小时审稿deadline，刚打开你的论文，读到…

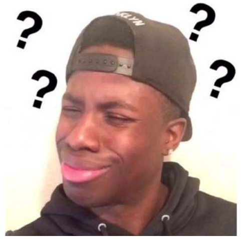
这么一句话

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

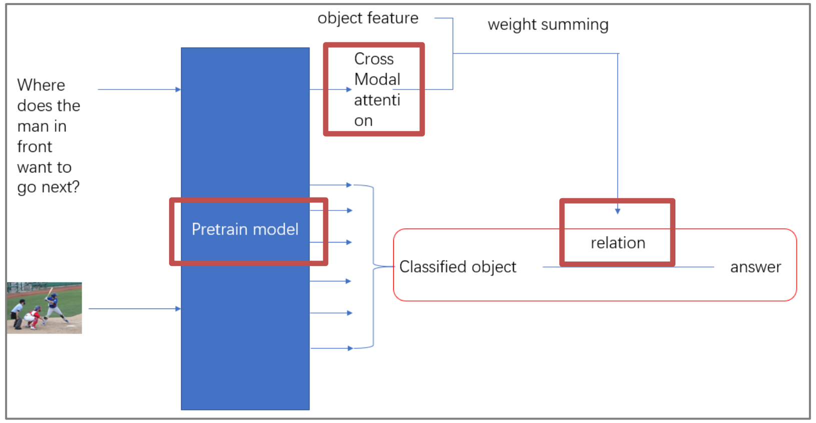


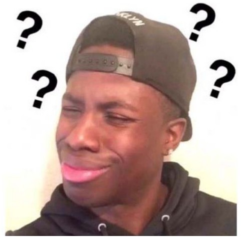
这么一个公式

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



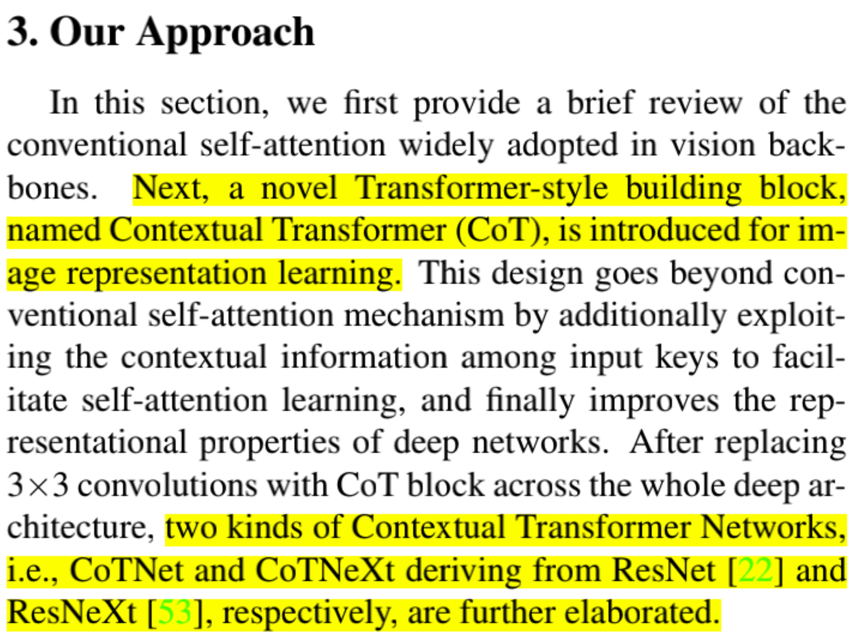
这么一个图





英语写作规范（2）——严谨的叙述逻辑

Tips: （1）在术语使用前定义解释！

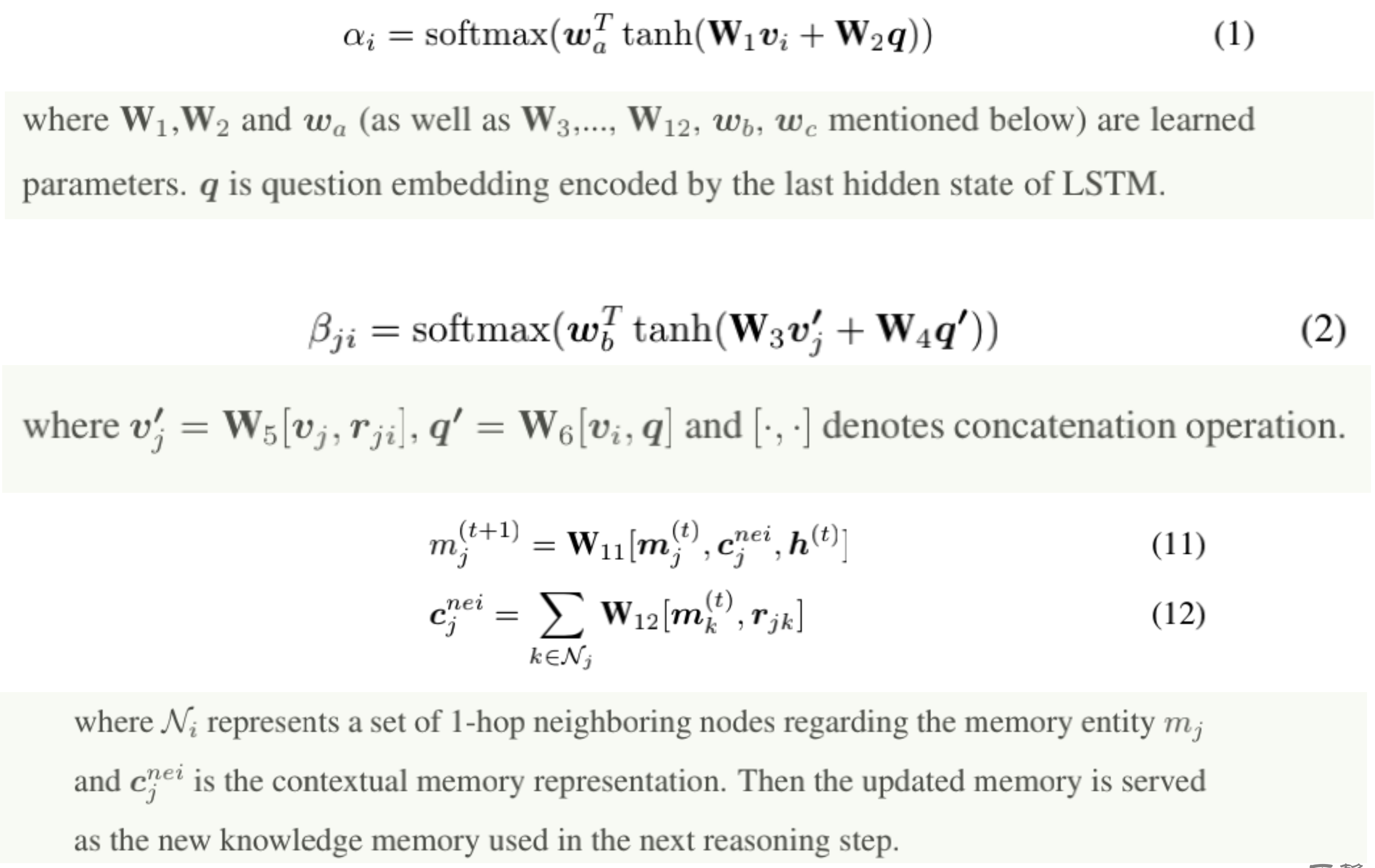


3. Our Approach

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.

英语写作规范（2）——严谨的叙述逻辑

Tips: （2）给出公式后集中对定义及公式中的符号进行解释



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

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.

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

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.

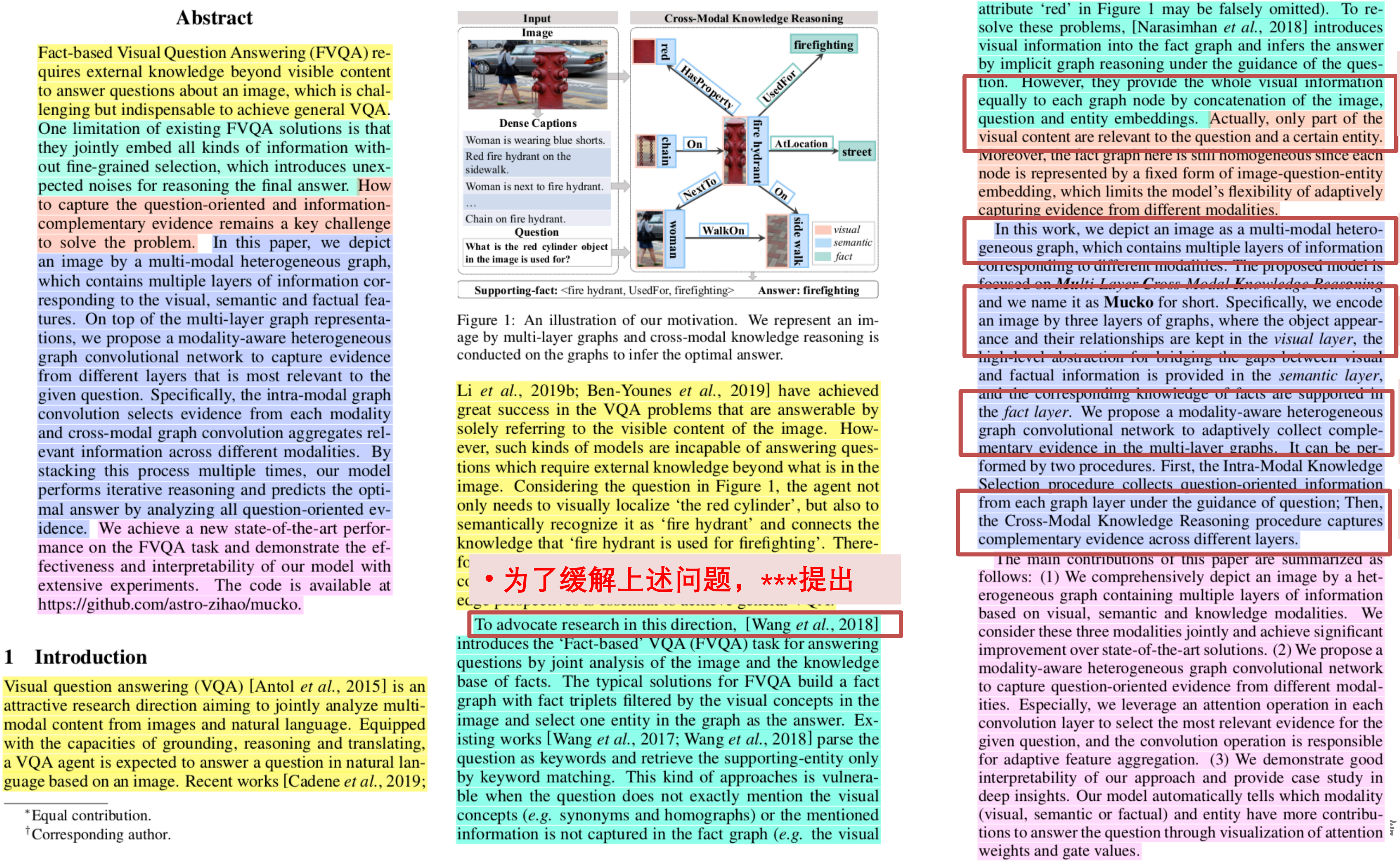
$$m_j^{(t+1)} = \mathbf{W}_{11}[\boldsymbol{m}_j^{(t)}, \boldsymbol{c}_j^{nei}, \boldsymbol{h}^{(t)}] \qquad (11)$$

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

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.

英语写作规范（2）——严谨的叙述逻辑

Tips: （3）有清晰的段落结构， 段落/章节之间有过度！



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

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

[*] Equal contribution.

[†] Corresponding author.

- 为了缓解上述问题，***提出

- 然而，他们方法只…
- 但事实上，需要…

- 本工作中，我们提出…

- 具体来说，我们首先…

- 在上面构建的表征基础上，我们进行…

- 然后，实现…

英语写作规范（2）——严谨的叙述逻辑

Tips: （4）图表文字清晰直接表达内容，与图注、表注、正文一致！

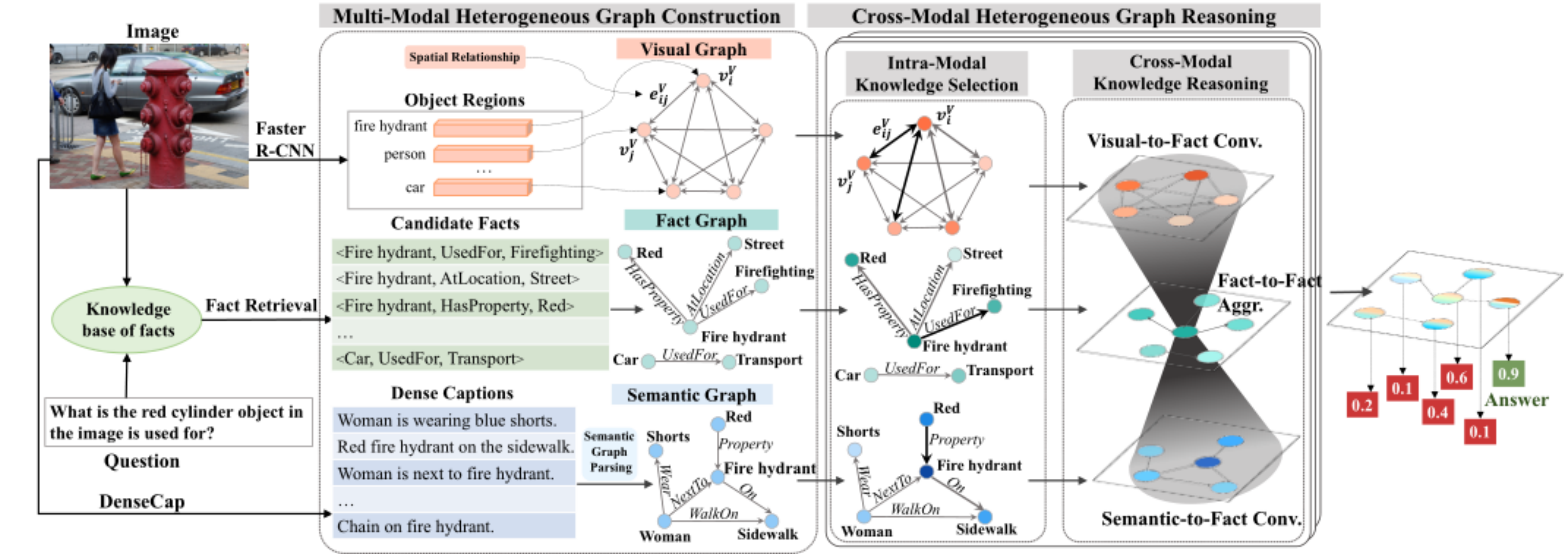
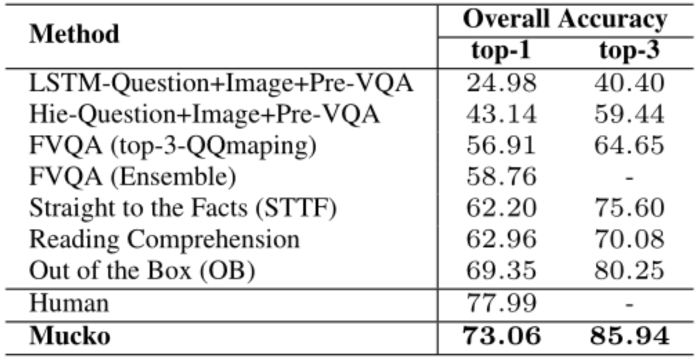
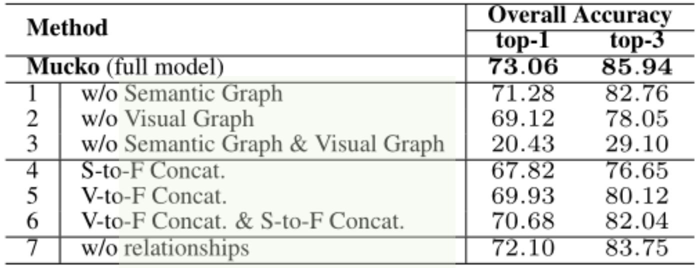


Figure 2: An overview of our model. The model contains two modules: Multi-modal Heterogeneous Graph Construction aims to depict an image by multiple layers of graphs and Cross-modal Hetegeneous Graph Reasoning supports intra-modal and cross-modal evidence selection.



Method	Overall Accuracy	
	top-1	top-3
LSTM-Question+Image+Pre-VQA	24.98	40.40
Hie-Question+Image+Pre-VQA	43.14	59.44
FVQA (top-3-QQmaping)	56.91	64.65
FVQA (Ensemble)	58.76	-
Straight to the Facts (STTF)	62.20	75.60
Reading Comprehension	62.96	70.08
Out of the Box (OB)	69.35	80.25
Human	77.99	-
Mucko	**73.06**	**85.94**

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



Method		Overall Accuracy	
		top-1	top-3
Mucko (full model)		**73.06**	**85.94**
1	w/o Semantic Graph	71.28	82.76
2	w/o Visual Graph	69.12	78.05
3	w/o Semantic Graph & Visual Graph	20.43	29.10
4	S-to-F Concat.	67.82	76.65
5	V-to-F Concat.	69.93	80.12
6	V-to-F Concat. & S-to-F Concat.	70.68	82.04
7	w/o relationships	72.10	83.75

Table 2: Ablation study of key components of Mucko.

欢迎大家在B站留言交流！

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