科研与英文学术论文写作

第三讲 英文学术论文之写作思路
——摘要和引言

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——摘要

基本要求

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

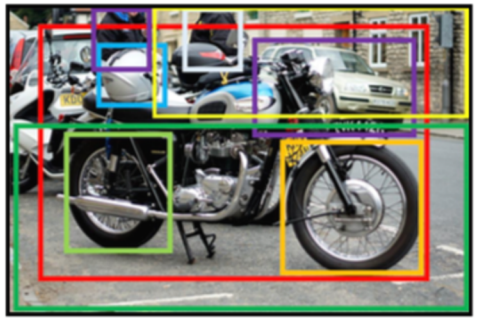
一篇论文的组成——摘要

CVPR 2020

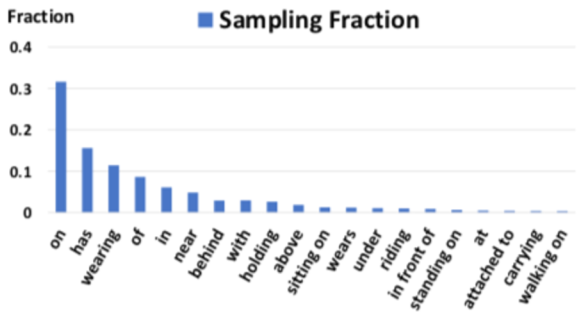
Unbiased Scene Graph Generation from Biased Training

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

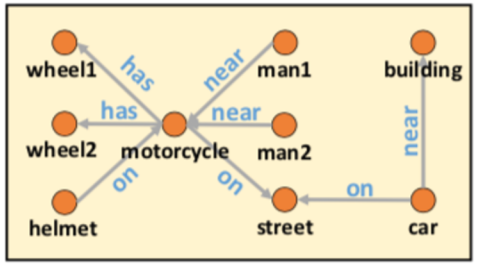
[1]Nanyang Technological University, [2]Damo Academy, Alibaba Group, [3]Renmin University of China, [4]Tsinghua University



(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
CVPR 2020

Today's scene graph generation (SGG) task is still far from practical, mainly due to the severe training bias, e.g., collapsing diverse `human walk on/ sit on/lay on beach` *into* `human on beach`*. Given such SGG, the down-stream tasks such as VQA can hardly infer better scene structures than merely a bag of objects. However, debiasing in SGG is not trivial because traditional debiasing methods cannot distinguish between the good and bad bias, e.g., good context prior (e.g.,* `person read book` *rather than* `eat`*) and bad long-tailed bias (e.g.,* `near` *dominating* `behind/in front of`*). In this paper, we present a novel SGG framework based on* ***causal inference*** *but not the conventional likelihood. We first build a causal graph for SGG, and perform traditional biased training with the graph. Then, we propose to draw the* ***counterfactual causality*** *from the trained graph to infer the effect from the bad bias, which should be removed. In particular, we use* ***Total Direct Effect*** *as the proposed final predicate score for unbiased SGG. Note that our framework is agnostic to any SGG model and thus can be widely applied in the community who seeks unbiased predictions. By using the proposed* ***Scene Graph Diagnosis*** *toolkit*[1] *on the SGG benchmark Visual Genome and several prevailing models, we observed significant improvements over the previous state-of-the-art methods.*

SGG存在的挑战：数据偏置

SGG方法的问题：未区分偏置的好坏

本文方法的思路：因果推断

本文具体的亮点：

构建因果图-->反事实推断-->优化目标

本文方法的优势：

- 模型无关，广泛适用
- 模型效果达到新SOTA

一篇论文的组成——引言

基本要求（摘要的扩展，最难的部分）

- 研究背景与挑战
- 提出问题与原因
- 相关工作与不足
- 本文研究思路
- 本文主要贡献

一篇论文的组成——引言

CCF—A

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

CCF—C

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

一篇论文的组成——引言

IJCAI 2020

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

Zihao Zhu[1,2*], **Jing Yu**[1,2*†], **Yujing Wang**[3], **Yajing Sun**[1,2], **Yue Hu**[1,2] and **Qi Wu**[4]

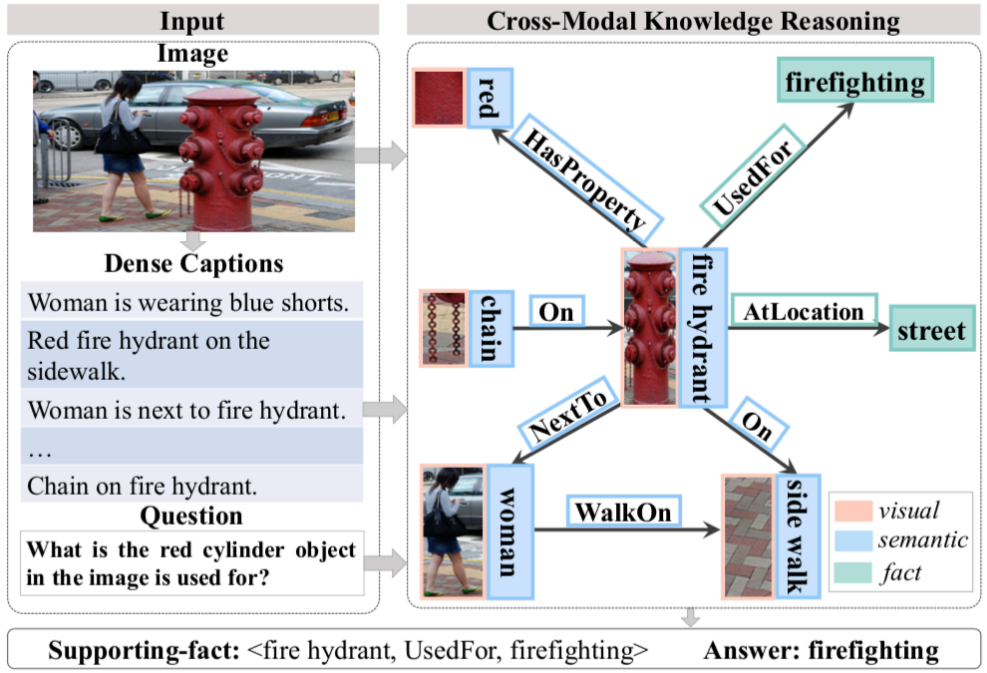
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[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



一篇论文的组成——引言

Abstract

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

1 Introduction

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

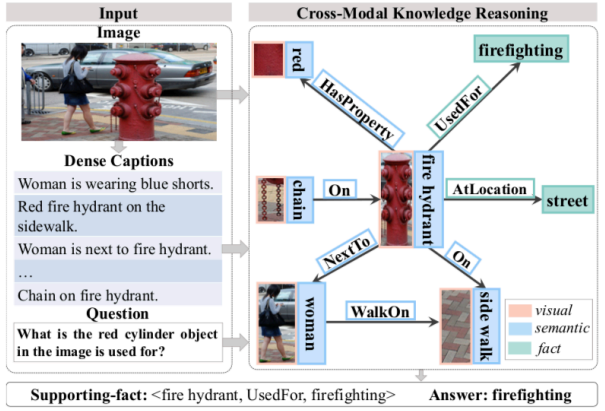


Figure 1: An illustration of our motivation. We represent an image by multi-layer graphs and cross-modal knowledge reasoning is conducted on the graphs to infer the optimal answer.

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.

有主要实验结果后可以动笔！

1 Introduction

大背景

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

背景：VQA需要知识
挑战：引入互补知识

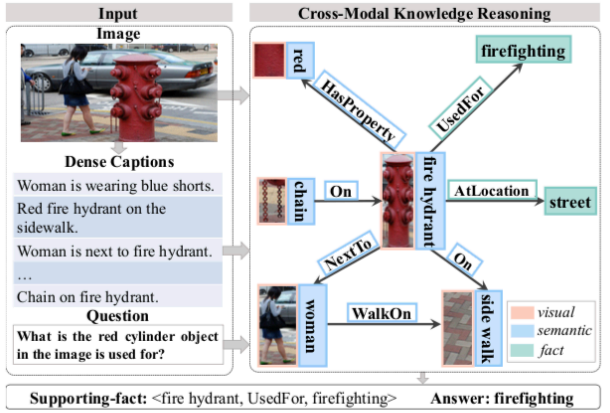


Figure 1: An illustration of our motivation. We represent an image by multi-layer graphs and cross-modal knowledge reasoning is conducted on the graphs to infer the optimal answer.

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

现有方法问题：分类归纳

ttribute '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
ion. However, th
equally to each gr
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.

从具体方法归纳问题

可解决的小问题

In
gene
corresponding to different modalities. The proposed model is focused on ***Mu**lti-Layer **C**ross-Modal **K**nowledge 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,
he Cros
compler

方法介绍：呼应挑战和问题

如何解决问题

贡献：凝练方法的普适性

The r
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.

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

准确、专业评价

欢迎大家在B站、知乎专栏、邮件留言交流！

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