

基于深度学习的视觉问答技术



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

于 静

2020-5-16

中科院信息工程研究所

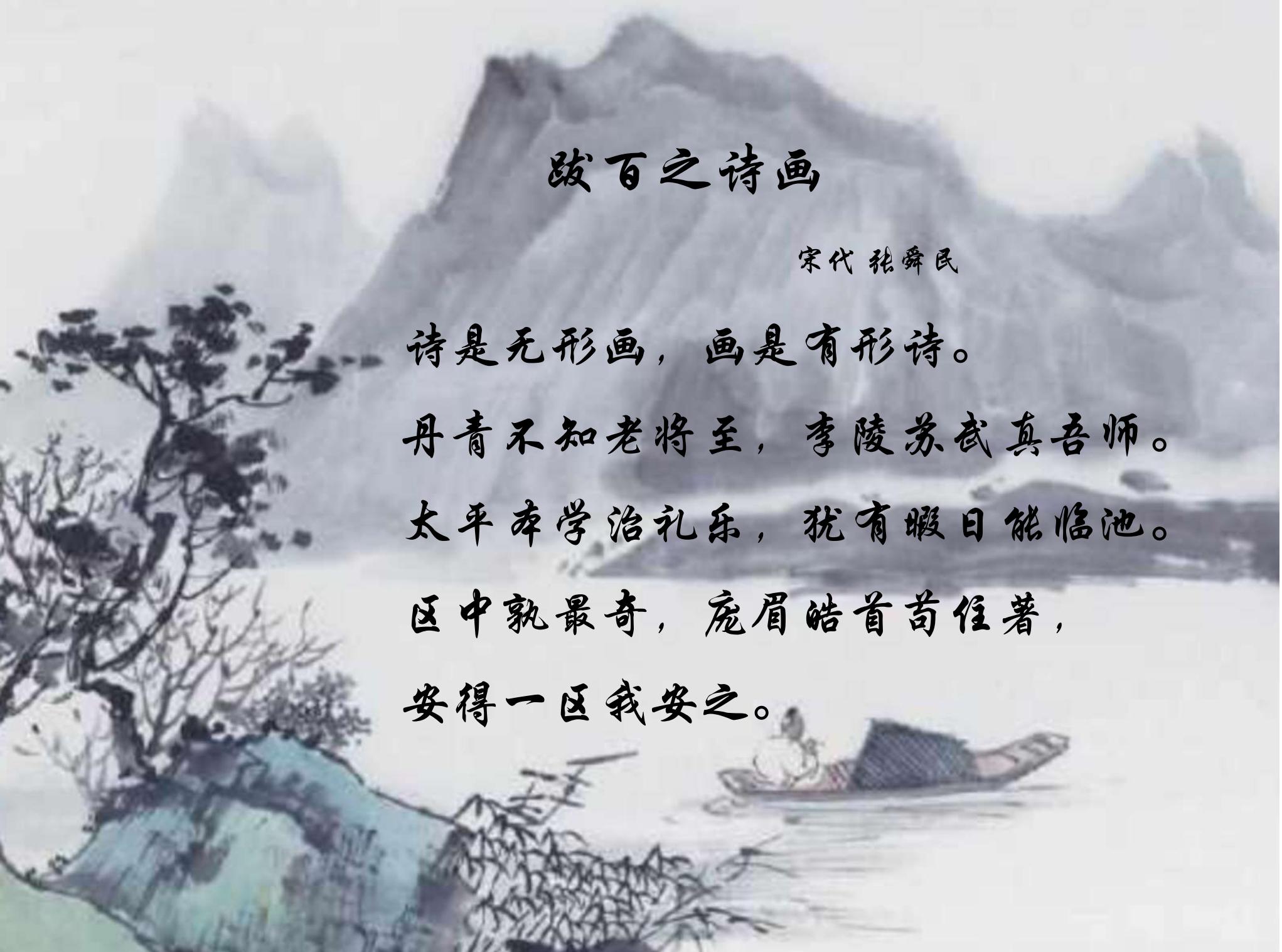
最新PDF: <https://mmlab-iie.github.io/>

报告提纲

- 1 / 多模态机器学习概述
- 2 / 视觉问答技术
- 3 / 视觉对话技术
- 4 / 总结与展望

报告提纲

- 1 / 多模态机器学习概述
- 2 / 视觉问答技术
- 3 / 视觉对话技术
- 4 / 总结与展望



跋百之诗画

宋代 张舜民

诗是无形画，画是有形诗。

丹青不知老将至，李陵苏武真吾师。

太平卒学治礼乐，犹有暇日能临池。

区中孰最奇，庞眉皓首尚住著，

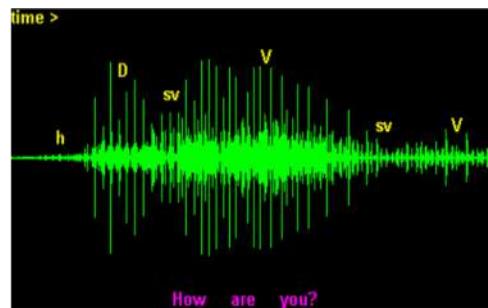
安得一区我安之。

1. 多模态机器学习概述

(1) 什么是多模态 (Multimodal) ?

模态 (Modality) 指信息产生或感知的形式

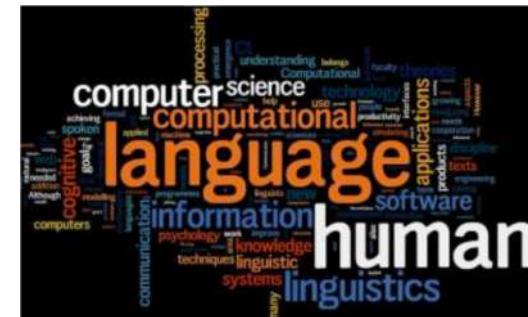
- 信息模态 (Information Modality) 指某种特定类型的信息及其存储形式



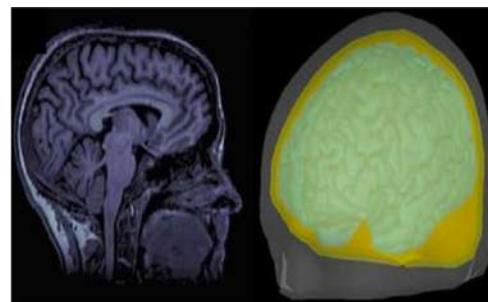
语音



图像



自然语言



医疗影像



多媒体

$$\begin{aligned} \text{OD} &= a, \sigma^2(SI) = \frac{1}{\sigma^2} \int_{a, \sigma^2} (\xi_i) = \frac{1}{\sqrt{2\pi}\sigma} \\ \int_{R_n} T(x) \cdot \frac{\partial}{\partial \theta} f(x, \theta) dx &= M \left(T(\xi) \cdot \frac{\partial}{\partial \theta} \ln L(\xi, \theta) \right) \int_{R_n} \\ \int_{R_n} T(x) \cdot \left(\frac{\partial}{\partial \theta} \ln L(x, \theta) \right) \cdot f(x, \theta) dx &= \int_{R_n} T(x) \left(\frac{\partial}{\partial \theta} \frac{f(x, \theta)}{f(x, \theta)} \right) dx \\ \frac{\partial}{\partial \theta} M T(\xi) &= \frac{\partial}{\partial \theta} \int_{R_n} T(x) f(x, \theta) dx = \int_{R_n} \frac{\partial}{\partial \theta} T(x) f(x, \theta) dx \end{aligned}$$

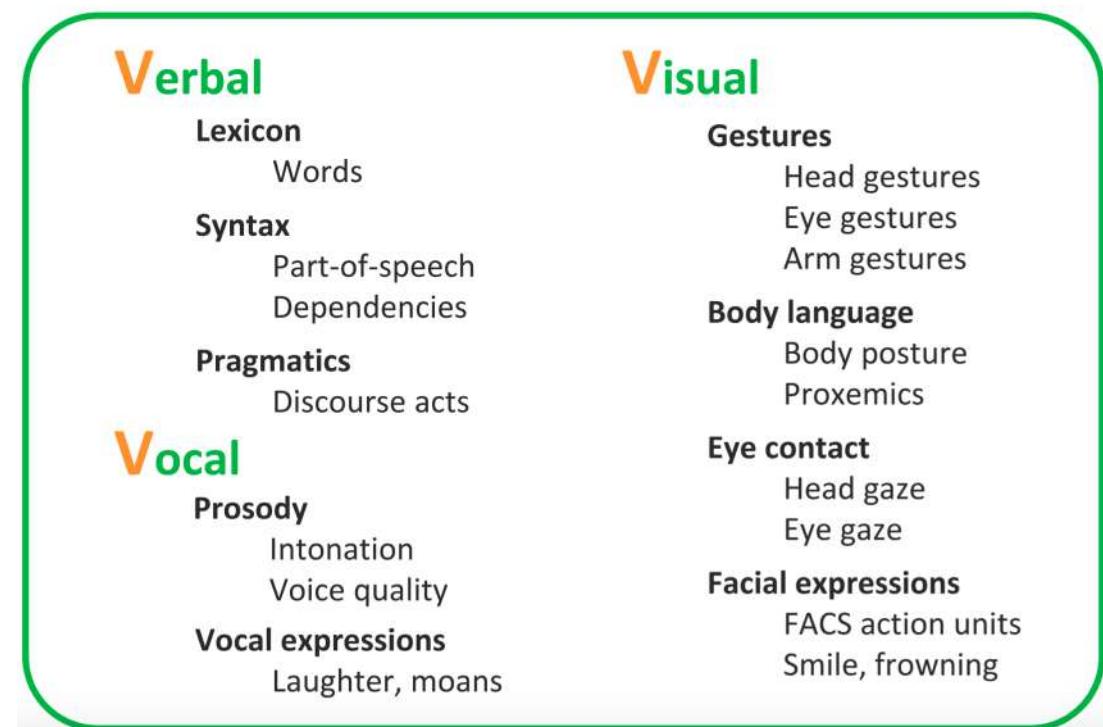
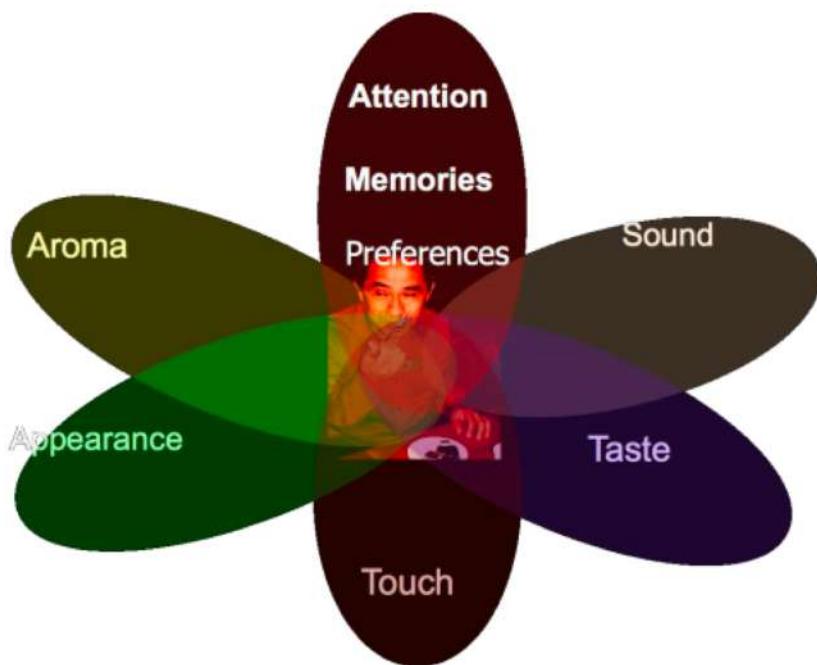
知识

1. 多模态机器学习概述

(1) 什么是多模态 (Multimodal) ?

模态 (Modality) 指信息产生或感知的方式

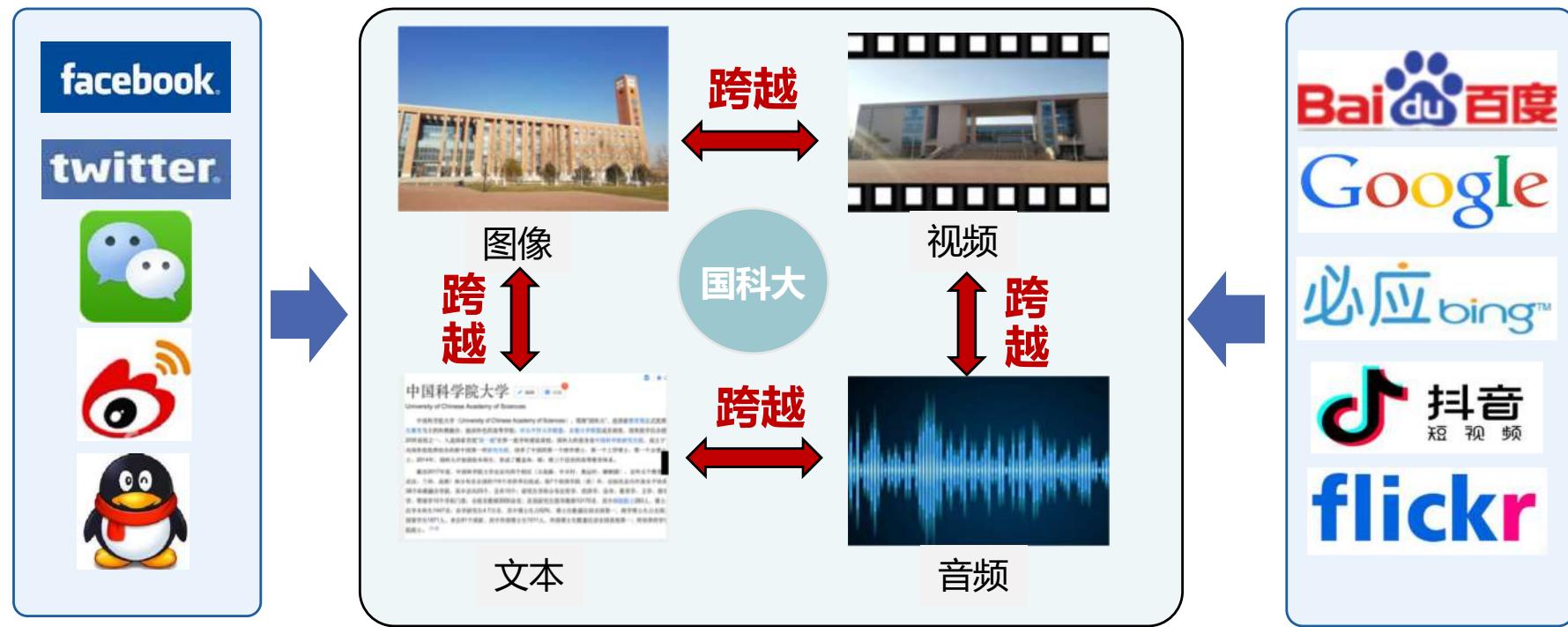
- 信息模态 (Information Modality) 指某种特定类型的信息及其存储形式
- 感知模态 (Sensory Modality) 指感知信息的形式及交互的形式



1. 多模态机器学习概述

(2) 多模态机器学习的研究背景

网络上海量的多媒体数据迅速增长并累积，图像、文本、视频、音频等构成相互融合的跨媒体形态：形式上多源异构，语义上相互关联



跨越语言、视觉、听觉等不同类型的媒体数据，对现实世界中的知识实现更加泛化的分析和推理，对推动人工智能的发展具有重要意义。

1. 多模态机器学习概述

(3) 多模态学习的研究问题

- 表示 (Representation)
- 对齐 (Alignment)
- 融合 (Fusion)
- 转换 (Translation)
- 协同 (Co-Learning)

Multimodal Machine Learning: A survey and Taxonomy

2019 TPAMI

By Tadas Baltrusaiis, Chaitanya Ahuja, and
Louis-Philippe Morency (CMU)

<https://arxiv.org/abs/1705.09406>

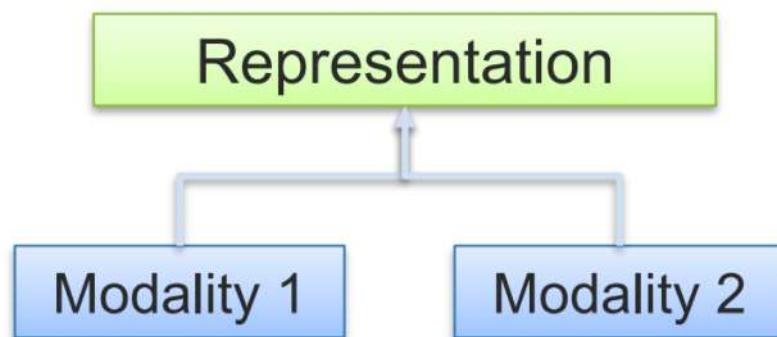
- ✓ 5 项技术挑战
- ✓ 37 种模型分类
- ✓ 253篇参考文献

1. 多模态机器学习概述

(3) 多模态学习的研究问题一：表示

定义：学习如何**表示**和**归纳**多模态的数据，充分挖掘数据中**互补**和**冗余**的信息

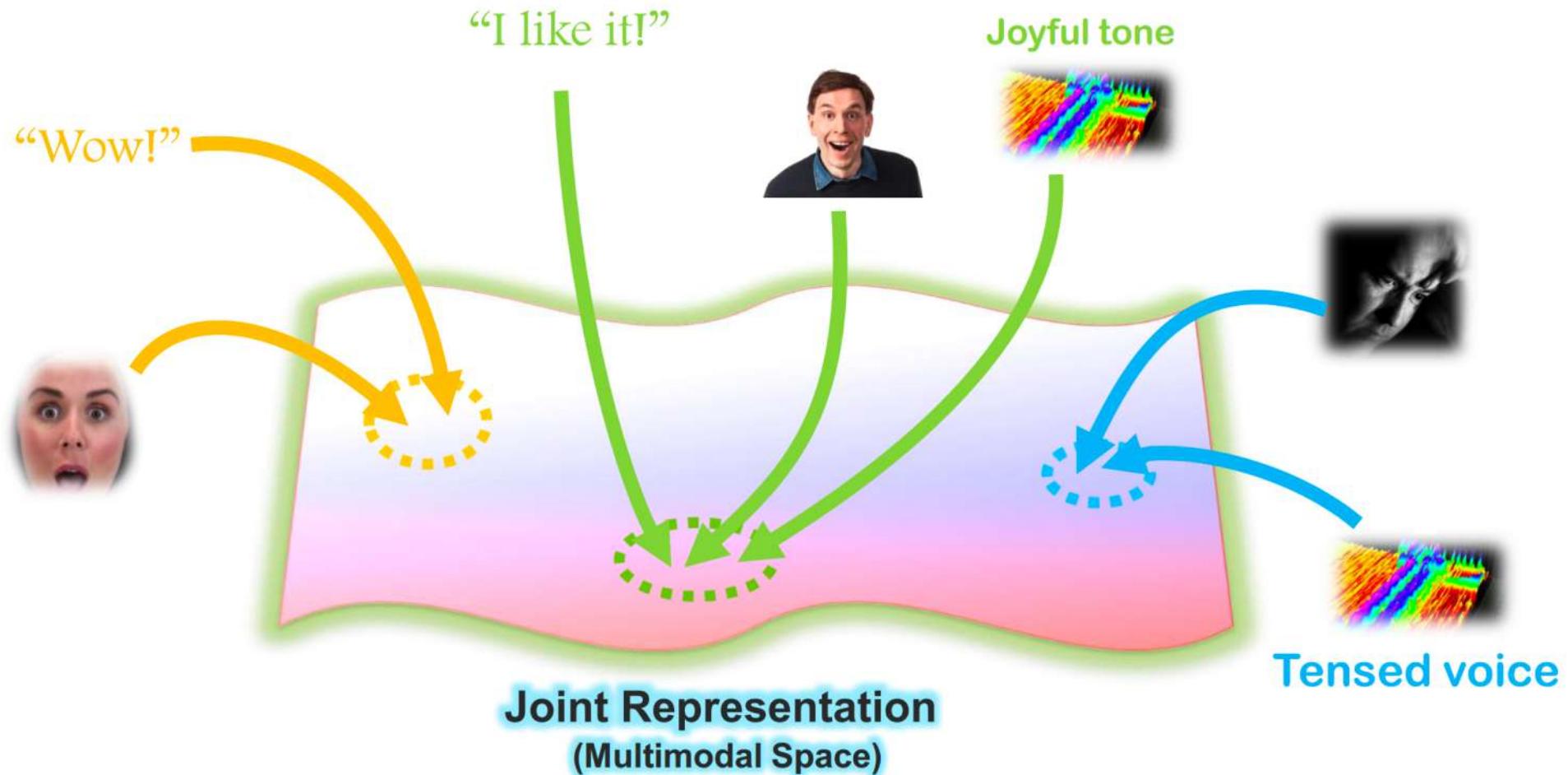
A: 联合表示（Joint Representation）



1. 多模态机器学习概述

A: 联合表示 (Joint Representation)

定义：将表达相同语义的不同模态数据用**统一特征表示**，使其包含不同模态中的互补信息



1. 多模态机器学习概述

A: 联合表示 (Joint Representation)

Audio-visual speech recognition

[Ngiam et al., ICML 2011]

- Bimodal Deep Belief Network

Image captioning

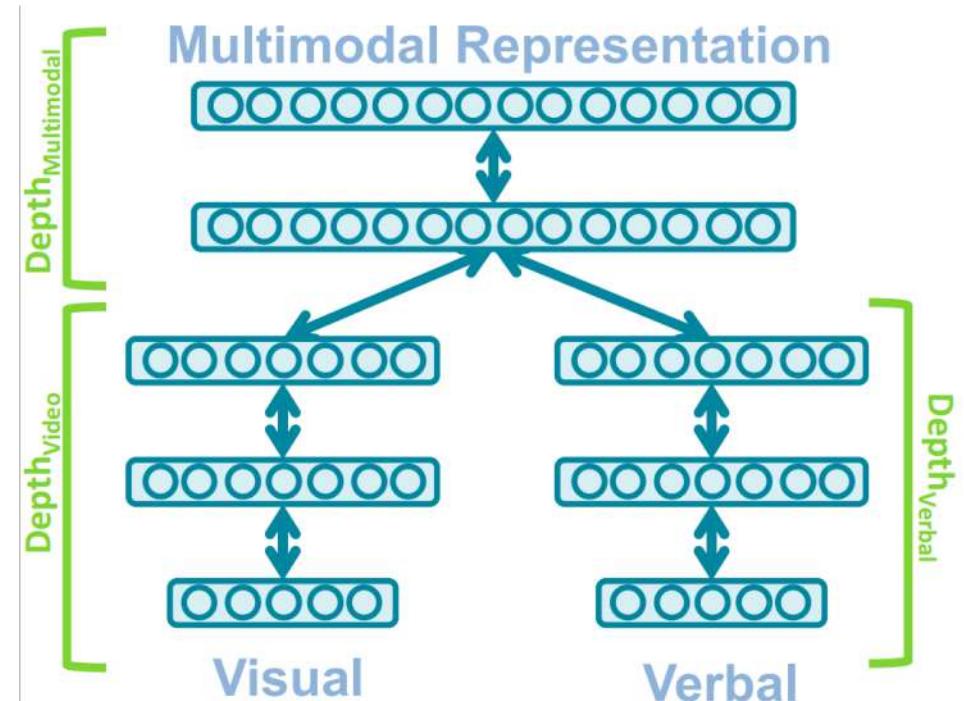
[Srivastava and Salakhutdinov, NIPS 2012]

- Multimodal Deep Boltzmann Machine

Audio-visual emotion recognition

[Kim et al., ICASSP 2013]

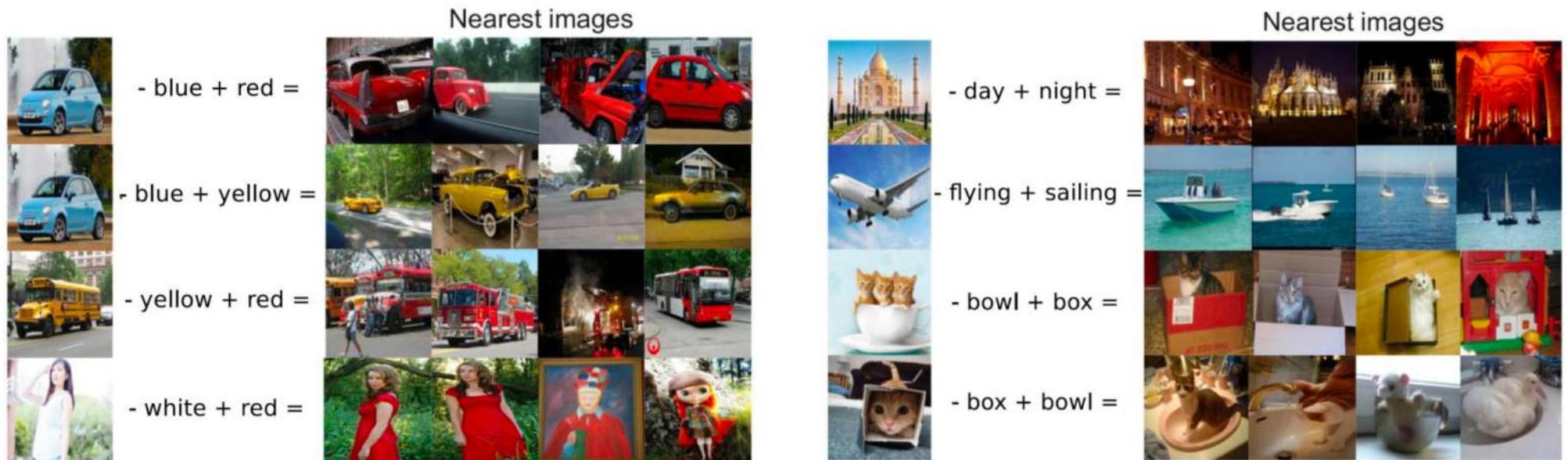
- Deep Boltzmann Machine



1. 多模态机器学习概述

A: 联合表示 (Joint Representation)

多模态联合表示可以实现多模态数据在算数空间内的计算



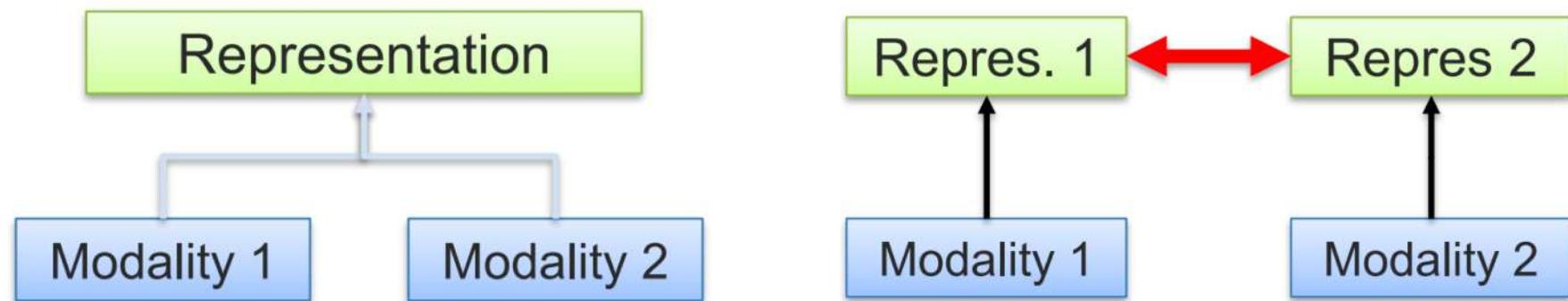
[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

1. 多模态机器学习概述

(3) 多模态学习的研究问题一：表示

定义：学习如何**表示**和**归纳**多模态的数据，充分挖掘数据中**互补**和**冗余**的信息

A: 联合表示（Joint Representation） B: 关联表示（Coordinated Representation）



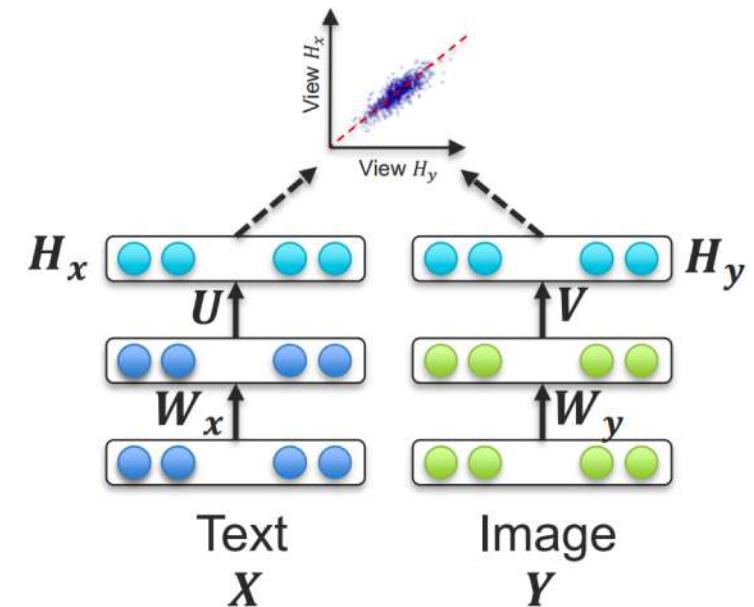
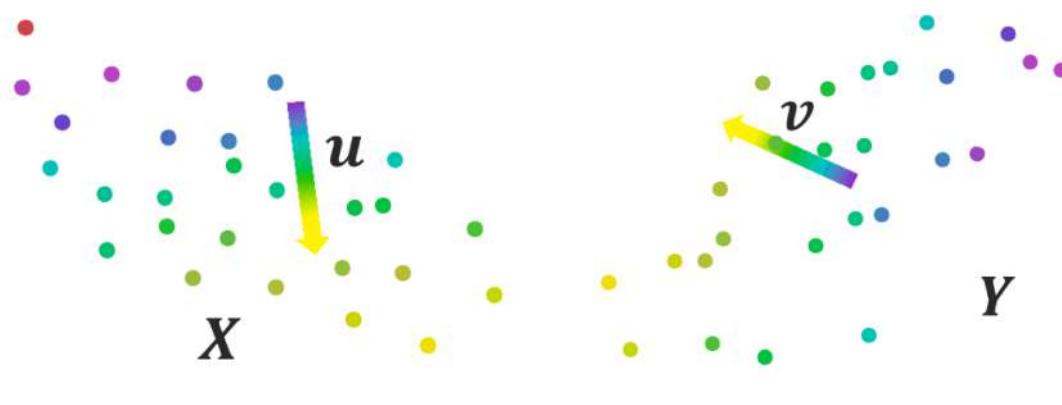
1. 多模态机器学习概述

B: 关联表示 (Coordinated Representation)

定义: 通过线性映射, 使不同模态数据的表示**相关性最大**

模型: Deep canonical correlation Analysis (Deep CCA)

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\boldsymbol{u}, \boldsymbol{v}}{\operatorname{argmax}} \operatorname{corr}(\boldsymbol{u}^T \boldsymbol{X}, \boldsymbol{v}^T \boldsymbol{Y})$$

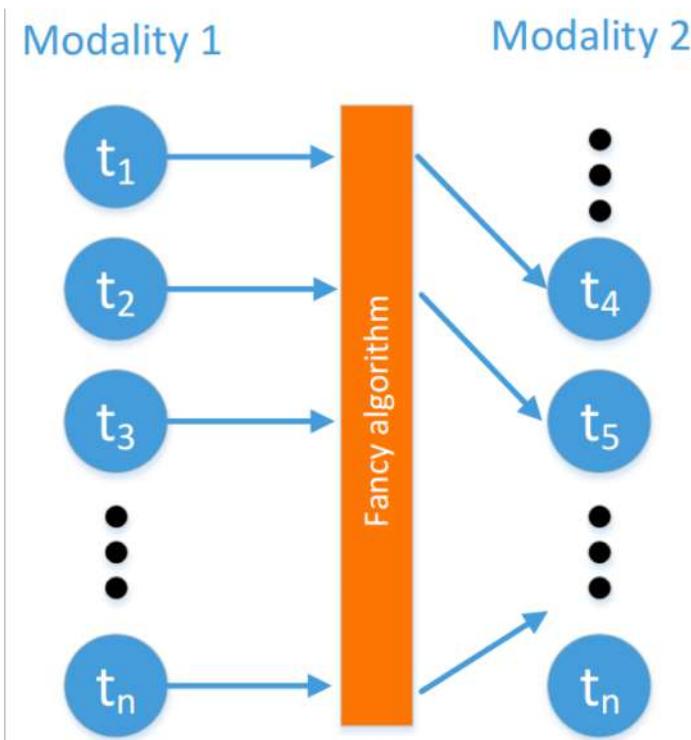


[Andrew et al., ICML 2013]

1. 多模态机器学习概述

(3) 多模态学习的研究问题二：对应

定义：挖掘不同模态数据的组成元素间的**对应关系**



A: 显式对齐 (Explicit Alignment)

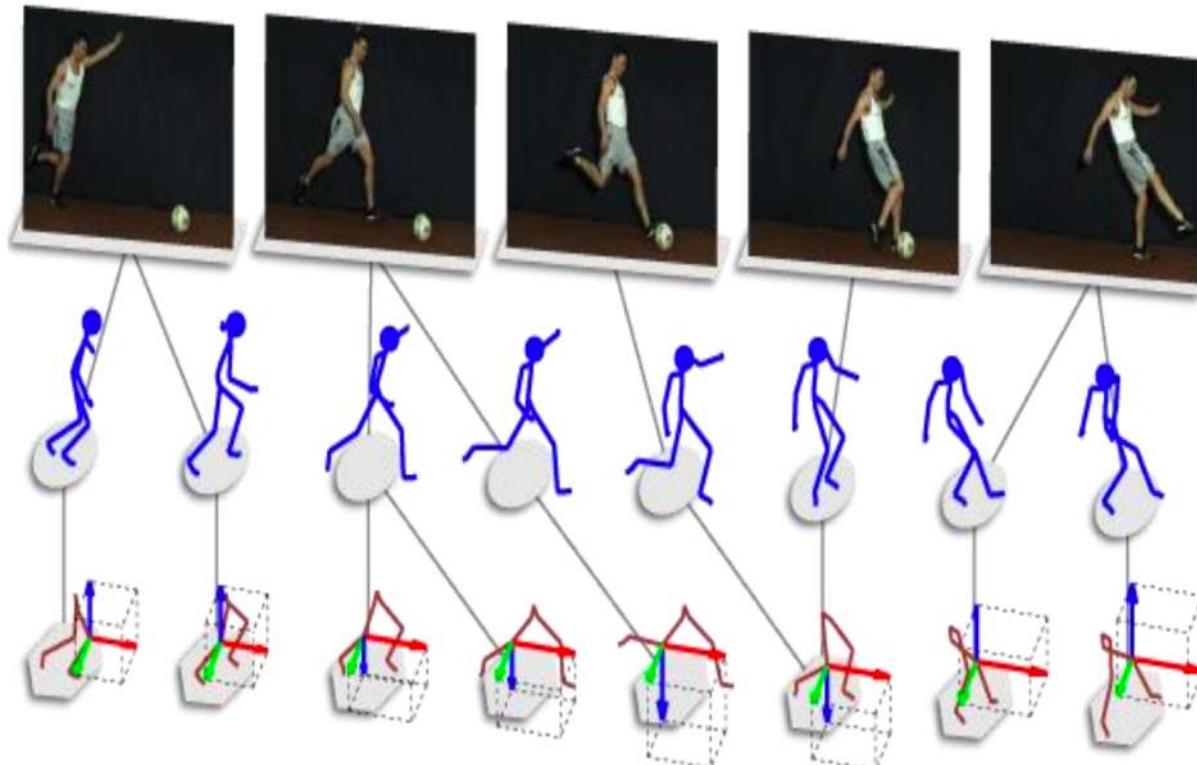
直接找出不同模态数据元素间的对应关系

B: 隐式对齐 (Implicit Alignment)

采用隐式空间的表示对齐不同模态的信息

1. 多模态机器学习概述

A: 显式对齐 (Explicit Alignment) : 时序对齐

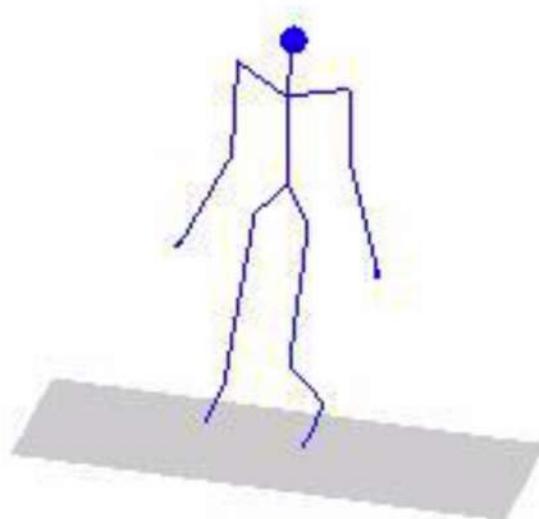


应用

- 对齐异步数据
- 查找不同模态的相似数据
- 多源数据的事件重建

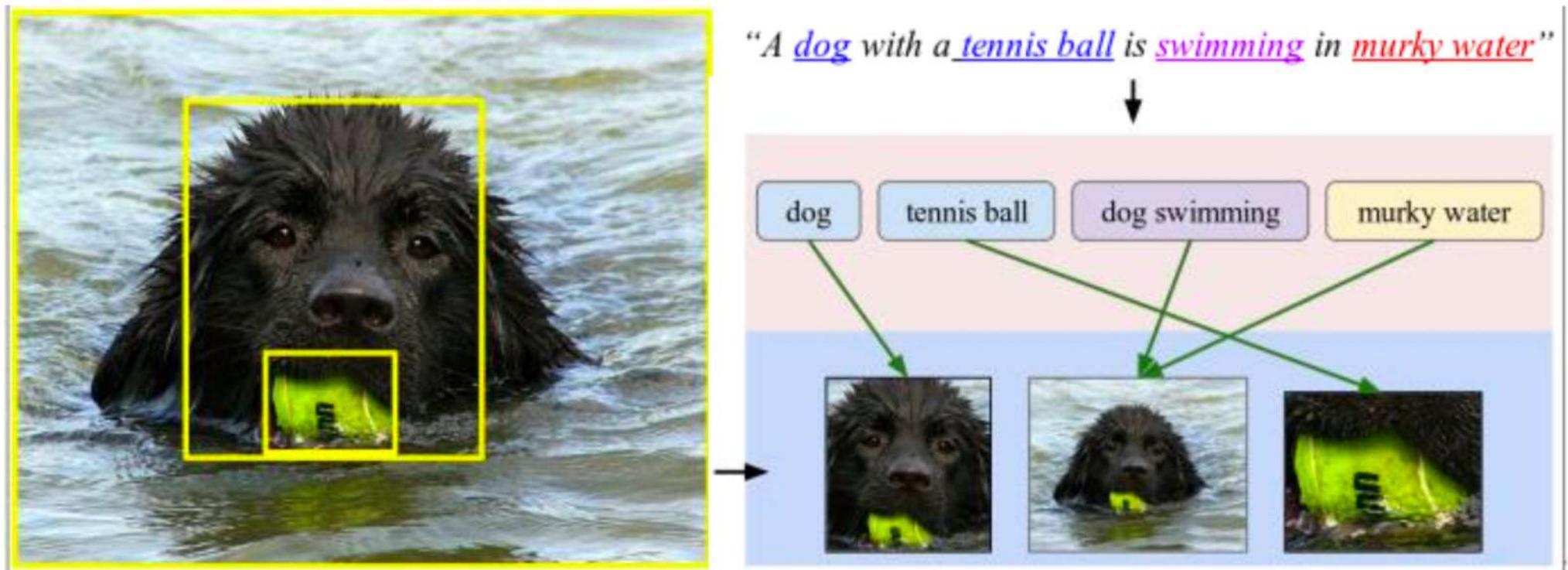
1. 多模态机器学习概述

A: 显式对齐 (Explicit Alignment) : 多模态相似数据



1. 多模态机器学习概述

B: 隐式对齐 (Explicit Alignment) : 解决无法一一对应的问题



[Karpathy et al., Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, <https://arxiv.org/pdf/1406.5679.pdf>]

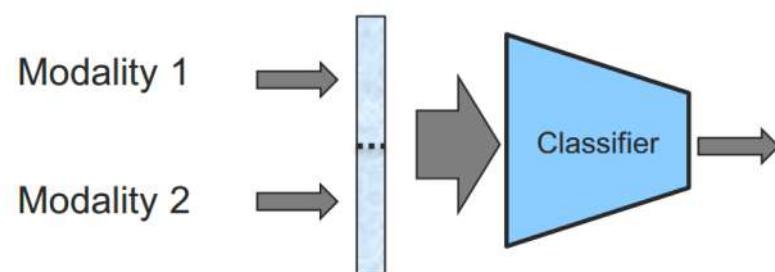
1. 多模态机器学习概述

(3) 多模态学习的研究问题三：融合

定义：融合多种模态的信息完成预测任务

A: 模型无关的方法

1) Early Fusion



2) Late Fusion



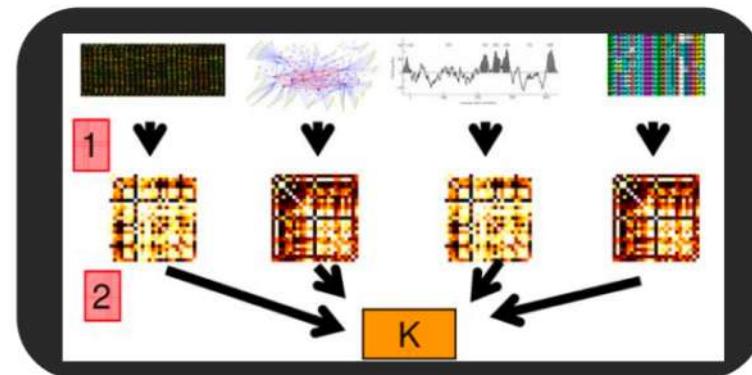
1. 多模态机器学习概述

(3) 多模态学习的研究问题三：融合

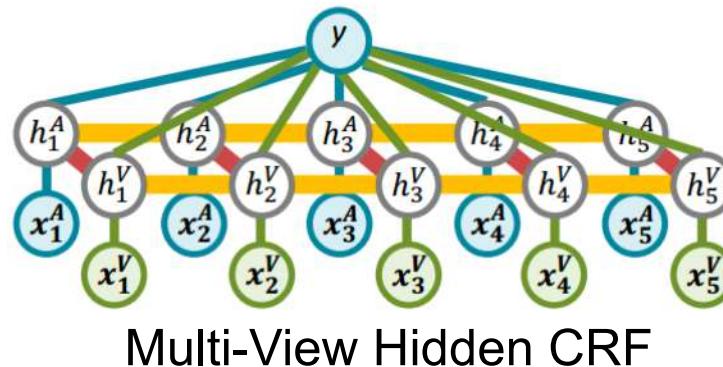
定义：融合多种模态的信息完成预测任务

B: 基于模型的方法

- 1) Deep neural networks
- 2) Kernel based methods
- 3) Graphical models



Multiple kernel learning



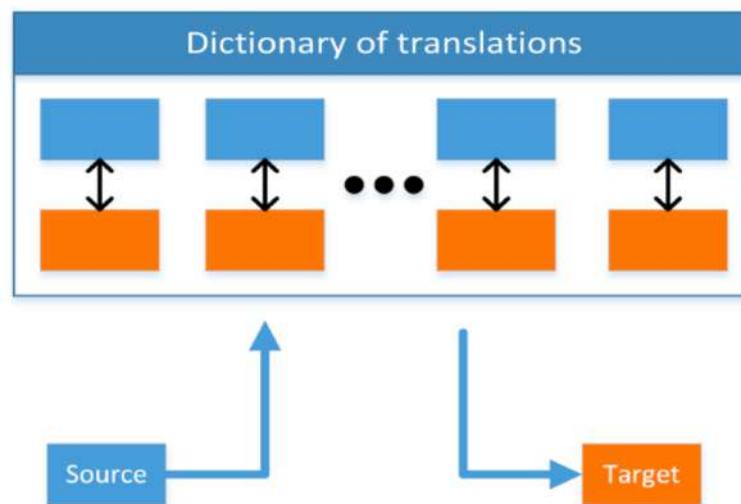
Multi-View Hidden CRF

1. 多模态机器学习概述

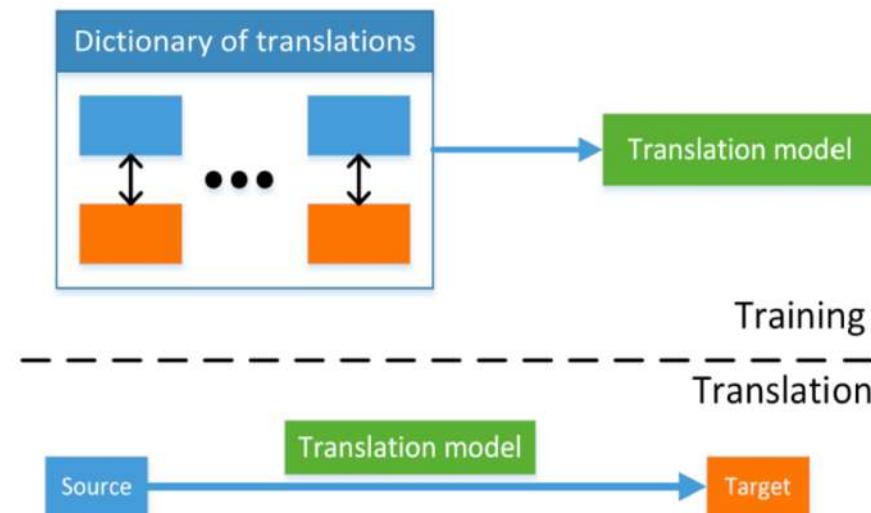
3.多模态学习的研究问题四：转换

定义：开放式或有主题的将一种模态数据转换为另外一种模态数据

A: 基于样本的方法



B: 基于模型的方法



1. 多模态机器学习概述

(3) 多模态学习的研究问题四：转换

定义：开放式或有主题的将一种模态数据转换为另外一种模态数据

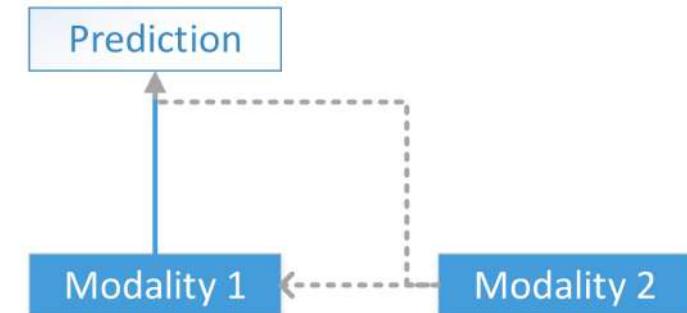


[Marsella et al., Virtual character performance from speech,
SIGGRAPH/Eurographics Symposium on Computer Animation, 2013]

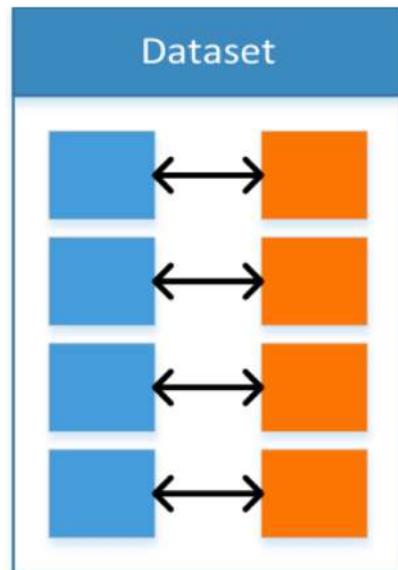
1. 多模态机器学习概述

(3) 多模态学习的研究问题五：协同学习

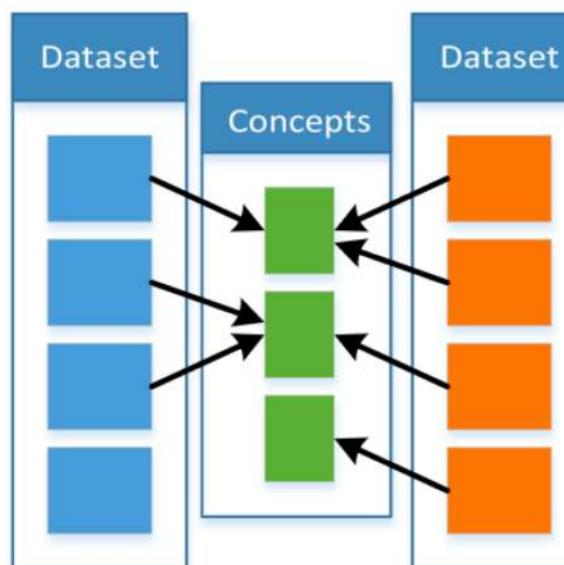
定义：将知识在不同模态数据间进行迁移，包括表示的迁移和预测模型的迁移



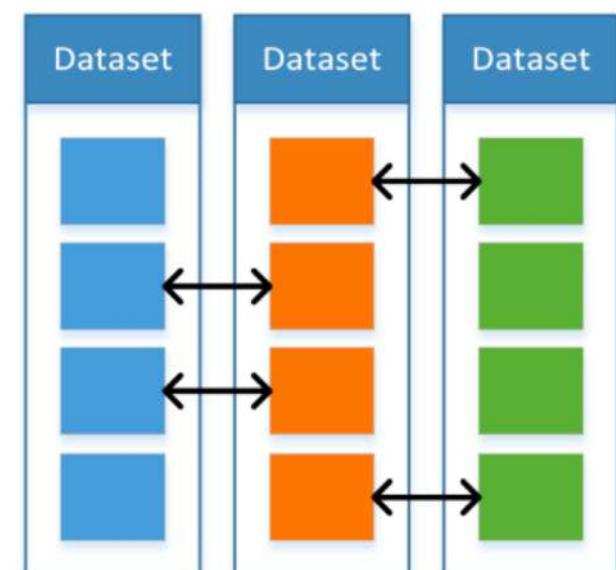
A: 平行数据



B: 非平行数据

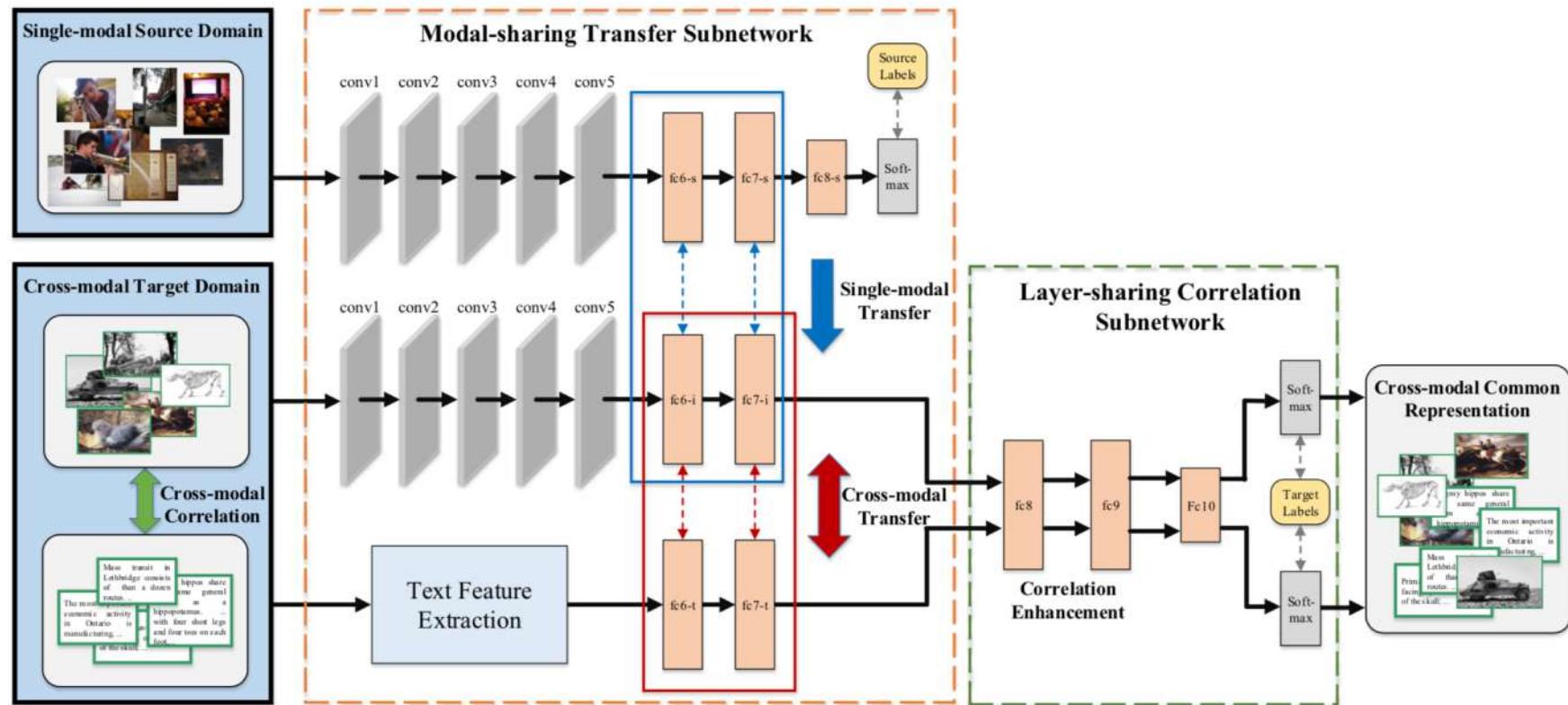


C: 混合数据



1. 多模态机器学习概述

(3) 多模态学习的研究问题五：协同学习



[Huang, et al. Cross-modal Common Representation Learning by Hybrid Transfer Network, IJCAI 2017]

1. 多模态机器学习概述

(3) 多模态学习的研究问题：小结

Representation

- Joint
 - Neural networks
 - Graphical models
 - Sequential
- Coordinated
 - Similarity
 - Structured

Translation

- Example-based
 - Retrieval
 - Combination
- Model-based
 - Grammar-based

- Encoder-decoder
- Online prediction

Alignment

- Explicit
 - Unsupervised
 - Supervised
- Implicit
 - Graphical models
 - Neural networks

Fusion

- Model agnostic
 - Early fusion
 - Late fusion
 - Hybrid fusion

- Model-based
 - Kernel-based
 - Graphical models
 - Neural networks

Co-learning

- Parallel data
 - Co-training
 - Transfer learning
- Non-parallel data
 - Zero-shot learning
 - Concept grounding
 - Transfer learning
- Hybrid data
 - Bridging

1. 多模态机器学习概述

(4) 多模态学习技术的应用

APPLICATIONS	CHALLENGES				
	REPRESENTATION	TRANSLATION	FUSION	ALIGNMENT	CO-LEARNING
Speech Recognition and Synthesis Audio-visual Speech Recognition (Visual) Speech Synthesis	✓ ✓	✓	✓	✓	✓
Event Detection Action Classification Multimedia Event Detection	✓ ✓		✓ ✓		✓ ✓
Emotion and Affect Recognition Synthesis	✓ ✓	✓	✓	✓	✓
Media Description Image Description Video Description	✓ ✓	✓ ✓	✓ ✓	✓ ✓	✓ ✓
Visual Question-Answering	✓		✓	✓	✓
Media Summarization	✓	✓	✓		
Multimedia Retrieval Cross Modal retrieval Cross Modal hashing	✓ ✓	✓		✓	✓

报告提纲

- 1 / 跨模态机器学习概述
- 2 / 视觉问答技术
- 3 / 视觉对话技术
- 4 / 总结与展望

2

视觉问答技术

2.1 视觉问答概述

2.1 视觉问答概述：相关领域

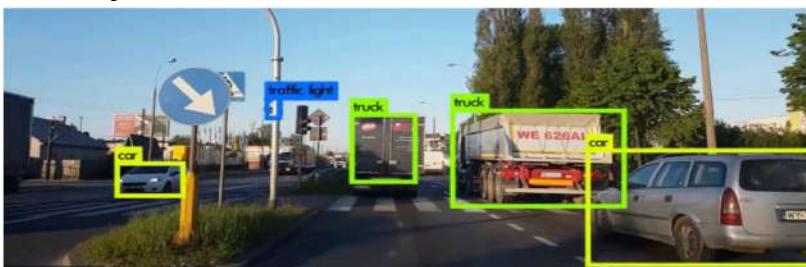
Vision and Language

Computer Vision (CV)

- Image Classification



- Object Detection

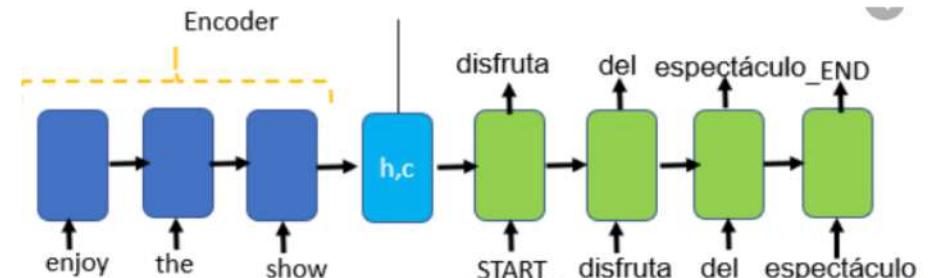


- Segmentation



Natural Language Processing(NLP)

- Language Generation
- Language Understanding
- Language Parsing
- Machine Translation



- Question Answering (QA)

When can I receive the delivery?

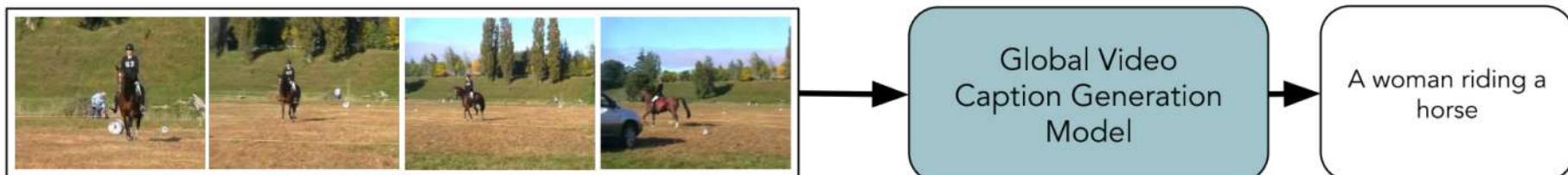
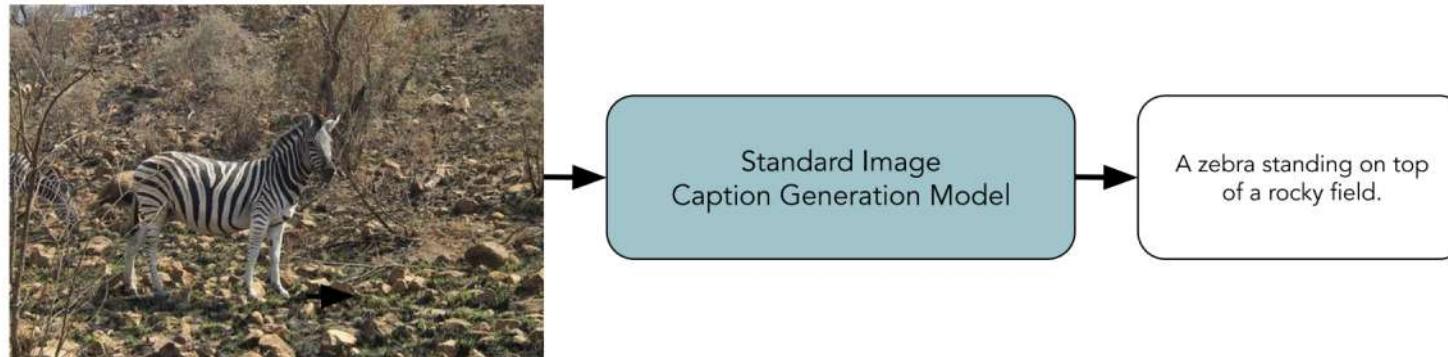


Tuesday or on Saturday

2.1 视觉问答概述：相关领域

Vision and Language

Image Understanding + Language Generation = Image Captioning



2.1 视觉问答概述：任务描述

- 视觉问答（Visual Question Answering, VQA）：根据图像和自然语言形式的问题，自动推理得到准确答案。

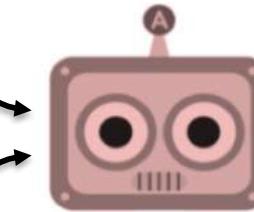


图像

问题

•What are they doing?

- Are there any humans?
- What sport is being played?
- Who has the ball?
- How many players are in the image?
- Who are the teams?
- Is it raining?



视觉问答模型

答案

Playing football.

2.1 视觉问答概述：应用场景

- An aid to visually-impaired
Is it safe to cross the street now?



2.1 视觉问答概述：应用场景

- Surveillance

What kind of car did the man in red shirt leave in?



2.1 视觉问答概述：应用场景

- Interacting with personal assistants
Is my laptop in my bedroom upstairs?



2.1 视觉问答概述：任务类型

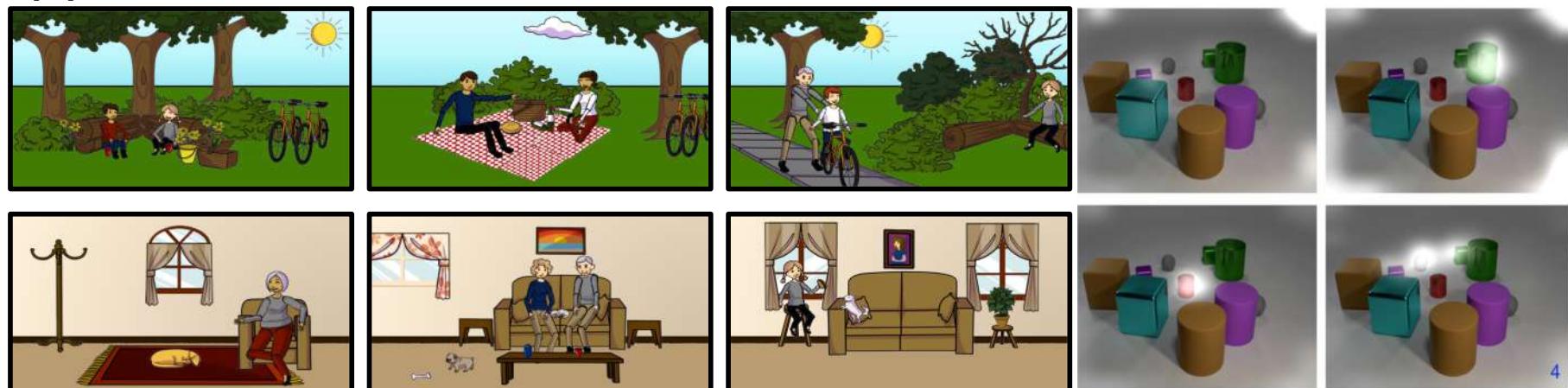
■ 视觉问答的图像类型

(1) 真实图像 (来自MSCOCO)

Tsung-Yi Lin *et al.* "Microsoft COCO: Common Objects in COntext." ECCV 2014.
<http://mscoco.org/>



(2) 生成图像



2.1 视觉问答概述：任务类型

■ 视觉问答的答案类型

(1) Open-ended task

Why is the girl holding an umbrella?



(2) Multiple choice task

What is the bus number?

- | | | | | | |
|--------|-----------------------------------|----------|--------|----------------|----------------|
| a) 3 | b) 1 | c) green | d) 4 | e) window trim | f) blue |
| g) m5 | h) corn, carrots,
onions, rice | i) red | j) 125 | k) san antonio | l) sign
pen |
| m) 478 | n) no | o) 25 | p) 2 | q) yes | r) white |



2.1 视觉问答概述：任务特点

- 视觉问答 (Visual Question Answering, VQA) 的主要特点：
 - 多模态输入：图像和问题
 - 多模态感知：图像目标、视觉关系、问题核心
 - 跨模态认知与推理：视觉 + 语言 + 外部知识
 - 自然语言的准确回答：避免数据偏置

2.1 视觉问答概述：数据集

(1) VQA 1.0 (CVPR 2016 Virginia Tech & Georgia Tech)



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

>0.25 million images
>0.76 million questions
~10 million answers

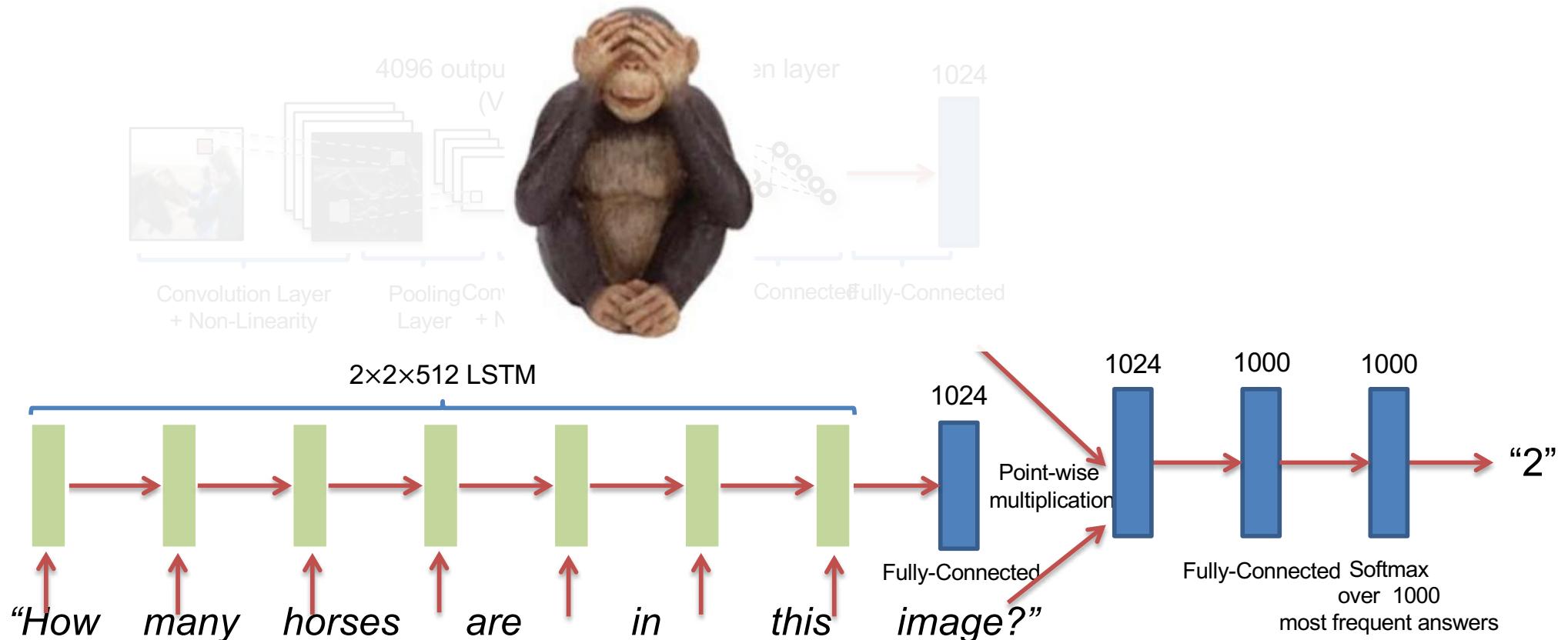
存在局限

- 答案的不一致性、不合理性
- 严重的语言偏置

2.1 视觉问答概述：数据集

(1) VQA 1.0 (CVPR 2016 Virginia Tech & Georgia Tech)

语言偏置问题



2.1 视觉问答概述：数据集

(1) VQA 1.0 (CVPR 2016 Virginia Tech & Georgia Tech)

语言偏置问题

Is there a clock ... ?

'yes' 98%



Is the man wearing glasses
... ?

'yes' 94%



Are the lights on ... ?

'yes' 85%



2.1 视觉问答概述：数据集

(2) VQA 2.0 (CVPR 2017 Virginia Tech & Georgia Tech)

Where is the child sitting?

fridge



arms



减少了VQA 1.0的语言偏置

问题的熵增长到56%

相比VQA 1.0规模更大

~1.7 倍图像-问题对

存在局限

- 严重缺少推理性问题
 - 19.5%包含关系的问题
 - 8%空间推理问题
 - 3%组合推理问题
- 严重的视觉偏置

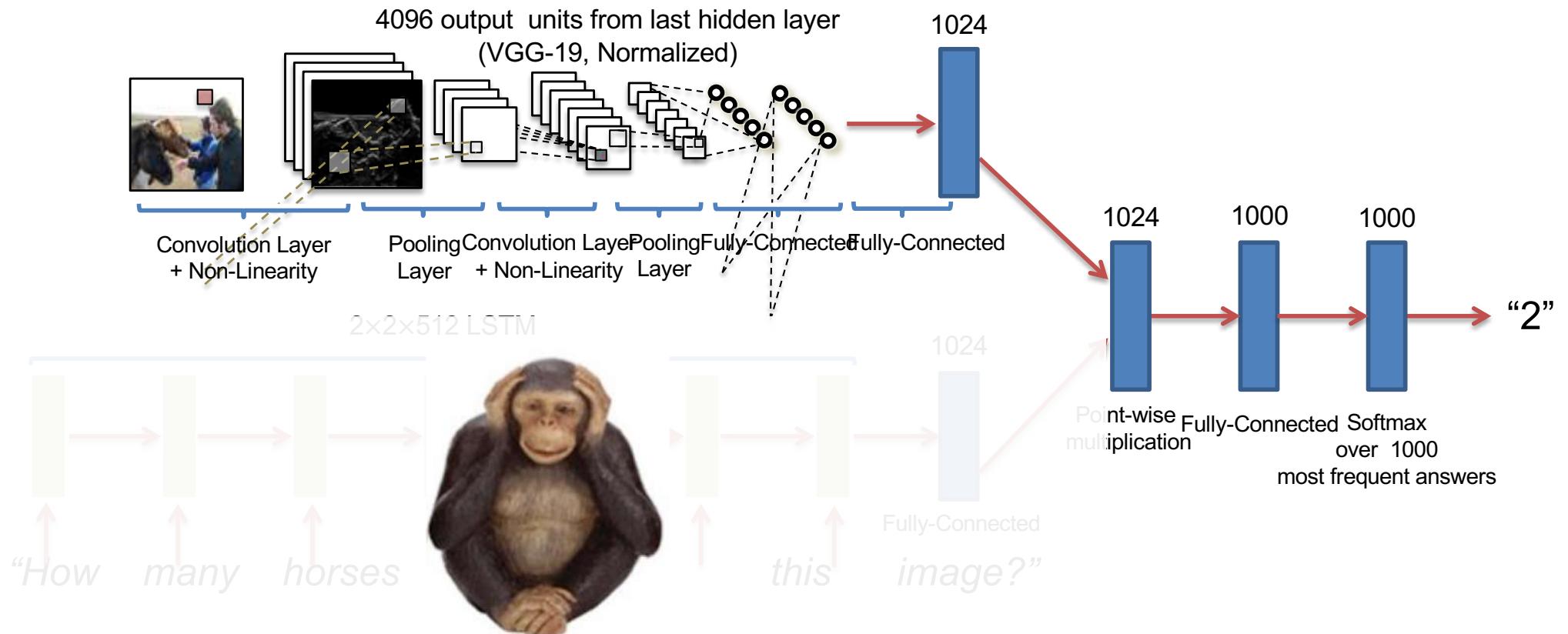
2.1 视觉问答概述：数据集

(2) VQA 2.0 (CVPR 2017 Virginia Tech & Georgia Tech)

视觉偏置问题

“deaf” model能回答对26%的问题！

→ VQA-CP 2.0

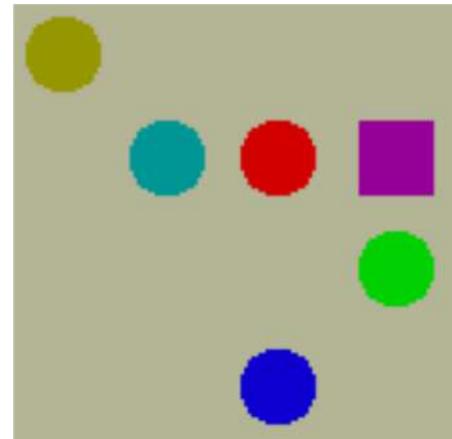


2.1 视觉问答概述：数据集

(3) CLEVR (CVPR 2017 Stanford) Sort-of-CLEVR (NeurIPS 2017 DeepMind)



CLEVR



Sort-of-CLEVR

存在局限

- 人工生成的图像和自然语言问题
- 目标物体的种类、数量、语义关系非常有限
- 复杂模型可能拟合所有组合情况，失去真正推理能力

2.1 视觉问答概述：数据集

(4) 视觉推理：GQA (2018 NeurIPS Stanford)

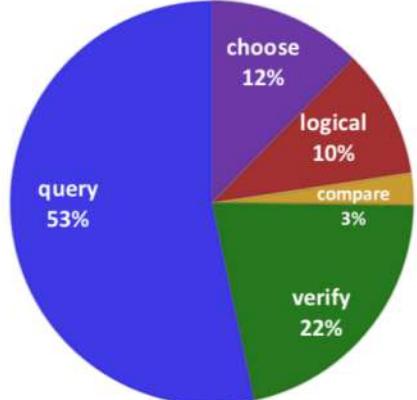


1. Is the **tray** on top of the **table** black or light brown? light brown
2. Are the **napkin** and the **cup** the same color? yes
3. Is the small **table** both oval and wooden? yes
4. Is the **syrup** to the left of the **napkin**? yes
5. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
6. Are there any **cups** to the left of the **tray** that is on top of the **table**? no
7. Could this **room** be a living room? yes

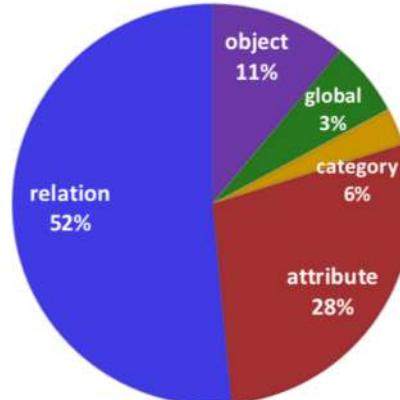
>22 M **compositional questions**
involving a diverse set of reasoning skills

>113K **real-world images**, each comes with a
scene graph to represent its semantics

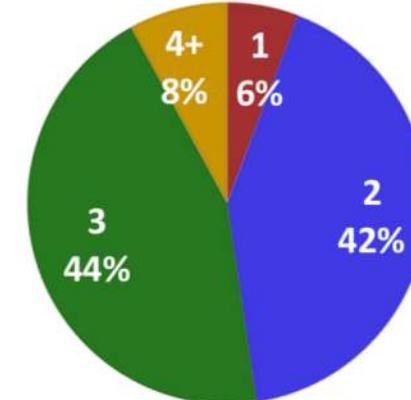
GQA STRUCTURAL TYPE COMPOSITION



GQA SEMANTIC STEPS



GQA SEMANTIC STEPS



2.1 视觉问答概述：数据集

(5) 利用外部知识：FVQA (2018, TPAMI, Univ. of Adelaide)



Question:

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

Factual Knowledge:

<fire hydrant, UsedFor, firefighting>

- ~ 5286 compositional questions
- ~ 2190 real-world images
- ~ 193,449 facts from knowledge base

存在局限

- 固定外部知识库
- 基于规则构造问题
- 数据规模较小

2.1 视觉问答概述：数据集

(5) 利用外部知识：OK-VQA (2019, CVPR, Univ. of Adelaide)



Q: Which American president is associated with the stuffed animal seen here?

A: Teddy Roosevelt

Outside Knowledge

Another lasting, popular legacy of Roosevelt is the stuffed toy bears—teddy bears—named after him following an incident on a hunting trip in Mississippi in 1902.

Developed apparently simultaneously by toymakers ... and named after President Theodore "Teddy" Roosevelt, the teddy bear became an iconic children's toy, celebrated in story, song, and film.

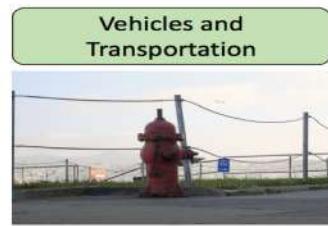
At the same time in the USA, Morris Michtom created the first teddy bear, after being inspired by a drawing of Theodore "Teddy" Roosevelt with a bear cub.

~ 14,055 compositional questions,

covering a variety of knowledge categories such as science technology, history, and sports

~ 14,031 real-world images

~ open-domain knowledge



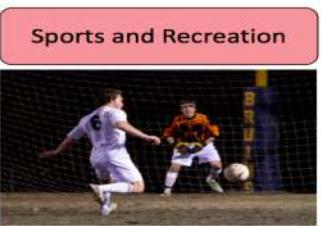
Q: What sort of vehicle uses this item?
A: firetruck



Q: When was the soft drink company shown first created?
A: 1898



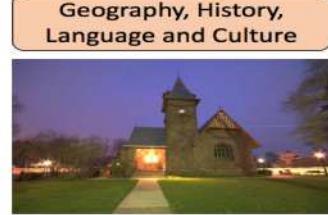
Q: What is the material used to make the vessels in this picture?
A: copper



Q: What is the sports position of the man in the orange shirt?
A: goalie



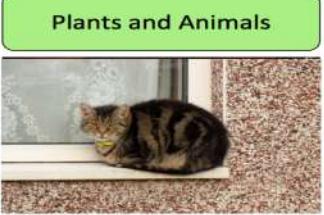
Q: What is the name of the object used to eat this food?
A: chopsticks



Q: What days might I most commonly go to this building?
A: Sunday



Q: Is this photo from the 50's or the 90's?
A: 50's



Q: What phylum does this animal belong to?
A: chordate, chordata



Q: How many chromosomes do these creatures have?
A: 23



Q: What is the warmest outdoor temperature at which this kind of weather can happen?
A: 32 degrees

2.1 视觉问答概述：数据集

其他VQA数据集

- Visual Turing Test [[Geman et al., PNAS 2014](#)]
- DAQUAR [[Malinowski & Fritz, NIPS 2014](#)]
- COCO-QA [[Ren et al., NIPS 2015](#)]
- FM-IQA [[Gao et al., NIPS 2015](#)]
- Visual7W [[Zhu et al., CVPR 2016](#)]
- Visual Genome [[Krishna et al., IJCV 2016](#)]
- VQA-HAT [[Das et al., EMNLP 2016](#)]
- CLEVR [[Johnson et al., CVPR 2017](#)]
- VQA v2.0 [[Goyal et al., CVPR 2017](#)]
- FVQA [[Wang et al., TPAMI 2018](#)]
- GQA [[Hudson et al., CVPR 2019](#)]
- KVQA [[Shah et al., AAAI 2019](#)]
- OK-VQA [[Marino et al., CVPR 2019](#)]
- VQA-360 [[Chou et al., WACV, 2020](#)]

2.1 视觉问答概述：评测指标

■ 视觉问答的评测指标

$$\text{Acc}(\textit{ans}) = \min \left\{ \frac{\#\text{humans that said } \textit{ans}}{3}, 1 \right\}$$



Open-Ended/Multiple-Choice/Ground-Truth

Q: WHAT OBJECT IS THIS

Ground Truth Answers:

(1) television	(6) television
(2) tv	(7) television
(3) tv	(8) tv
(4) tv	(9) tv
(5) television	(10) television

Q: How old is this TV?

Ground Truth Answers:

(1) 20 years	(6) old
(2) 35	(7) 80 s
(3) old	(8) 30 years
(4) more than thirty years	(9) 15 years
old	(10) very old
(5) old	

Q: Is this TV upside-down?

Ground Truth Answers:

(1) yes	(6) yes
(2) yes	(7) yes
(3) yes	(8) yes
(4) yes	(9) yes
(5) yes	(10) yes

2.1 视觉问答概述：学术成果

■ 视觉问答相关论文

Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick
School of Computer Science, The University of Adelaide

{qi.wu01, p.wang, chunhua.shen, anton.vandenhengel, anthony.dick}@adelaide.edu.au

Simple Baseline for Visual Question Answering

Bolei Zhou¹, Yuandong Tian², Sainbayar Sukhbaatar², Arthur Szlam², and Rob Fergus²

¹Massachusetts Institute of Technology
²Facebook AI Research

Compositional Memory for Visual Question Answering

Aiwen Jiang^{1,2}

¹Jiangxi Normal University
²aiwen.jiang@nicta.com.au

Fang Wang²

²NICTA and ANU
²{fang.wang, fatih.porikli}@nicta.com.au

Fatih Porikli²

³Toyota Research Institute North America

Yi Li*^{2,3}

²Toyota Research Institute North America

Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering

Huijuan Xu
UMass Lowell
@cs.uml.edu

Kate Saenko
UMass Lowell
saenko@cs.uml.edu

... and many more

Deep Compositional Question Answering with Neural Module Networks

Jacob Andreas Marcus Rohrbach Trevor Darrell Dan Klein

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

{jda, rohrbach, trevor, klein}@{cs, eecs, eecs, cs}.berkeley.edu

Kevin J. Shih, Saurabh Singh, and Derek Hoiem

University of Illinois at Urbana-Champaign

{kjshih2, ss1, dhoiem}@illinois.edu

Where To Look: Focus Regions for Visual Question Answering

ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering

Kan Chen
University of Southern California
kanchen@usc.edu

Haoyuan Gao
Baidu Research - IDL
gaohaoyuan@baidu.com

Jiang Wang
Baidu Research - IDL
wangjiang03@baidu.com

Wei Xu
Baidu Research - IDL
wei.xu@baidu.com

Liang-Chieh Chen
UCLA
lcchen@cs.ucla.edu

Ram Nevatia
University of Southern California
nevatia@usc.edu

Stacked Attention Networks for Image Question Answering

Zichao Yang¹, Xiaodong He², Jianfeng Gao², Li Deng², Alex Smola¹

¹Carnegie Mellon University, ²Microsoft Research, Redmond, WA 98052, USA
oy@cs.cmu.edu, {xiaohe, jfgao, deng}@microsoft.com, alex@smola.org

2.1 视觉问答概述：国际比赛

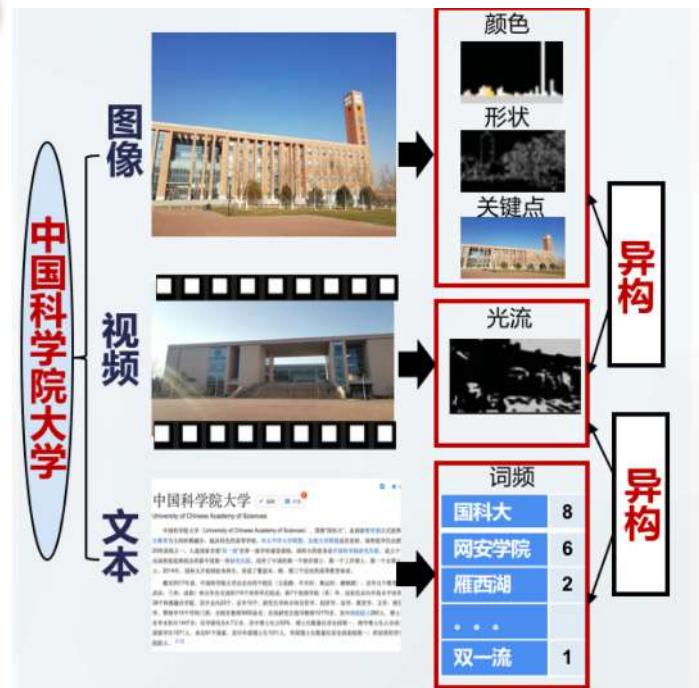
- 视觉问答国际比赛：VQA Challenge, GQA Challenge

The image displays two screenshots of visual question answering challenge websites. The top screenshot is for the VQA Challenge 2020, featuring a red header with the VQA logo and "Visual Question Answering". It includes navigation links for Home, People, Code, Demo, Download, Evaluation, Challenge (selected), Browse, Visualize, Workshop, Sponsors, Terms, and External. The Georgia Tech logo is also present. The main content area says "Welcome to the VQA Challenge 2020!" and shows a deadline of Friday, May 15, 2020, 23:59:59 GMT, with a 00 days 16h 56m 20s countdown. Below this are "Overview" and "Challenge Guidelines" links. A diagram illustrates the AI process: an image of a person with a yellow mustache is input into an "AI System", which outputs the answer "bananas". A question box asks "What is the mustache made of?". The bottom screenshot is for the GQA Challenge 2020, with a dark blue header featuring the GQA logo and navigation links for Home, About, Download, Evaluation, Challenge (selected), Visualize, Paper, Slides, and Contact. The main content area says "Welcome to the GQA Challenge 2020!" and features a large digital clock showing the time as 09:16:57:19.

2.1 视觉问答概述：技术挑战

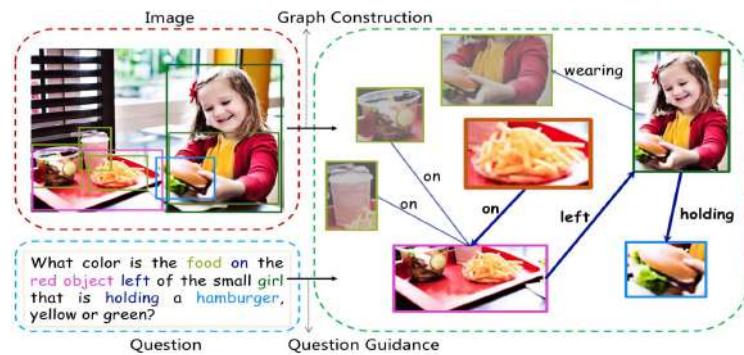
- 视觉信息、语言的多模态信息融合，解决“**异构鸿沟**”

1



- 跨媒体语义关联与推理技术

2



- 多源外部先验知识的利用

问题：这是葡萄牙当地什么时间？

3



2

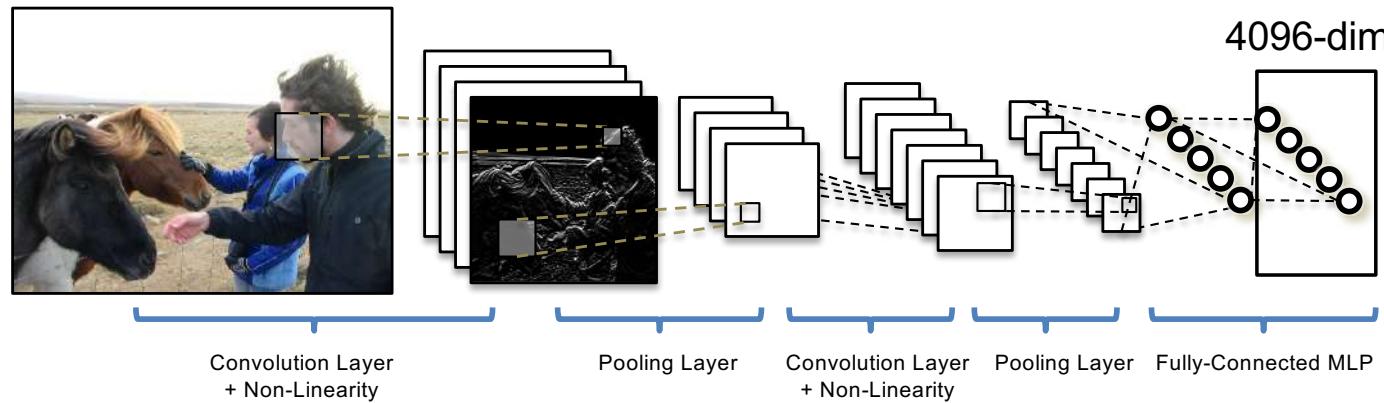
视觉问答技术

2.2 视觉问答模型

2.2 视觉问答模型

基准模型：双通道VQA模型

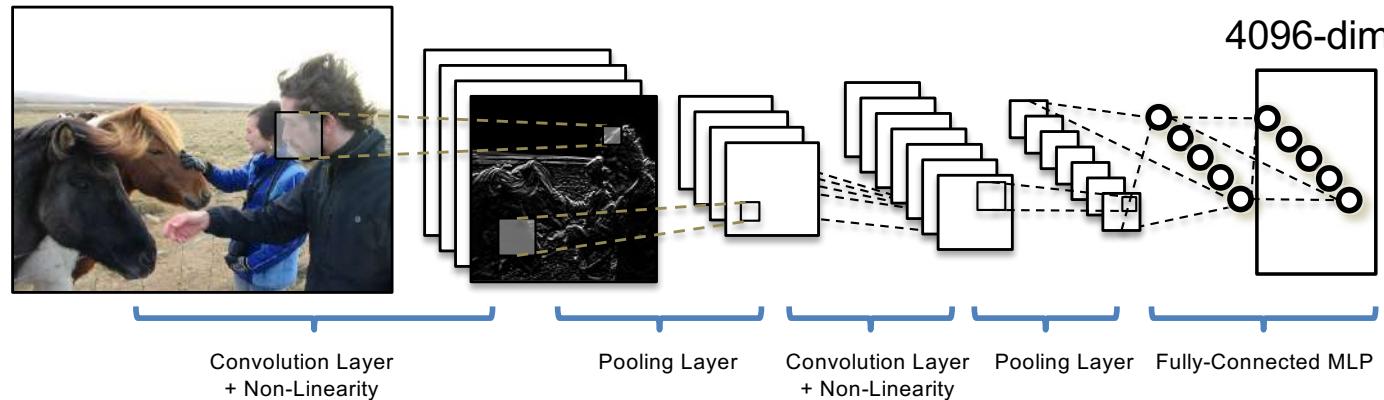
Image Embedding



2.2 视觉问答模型

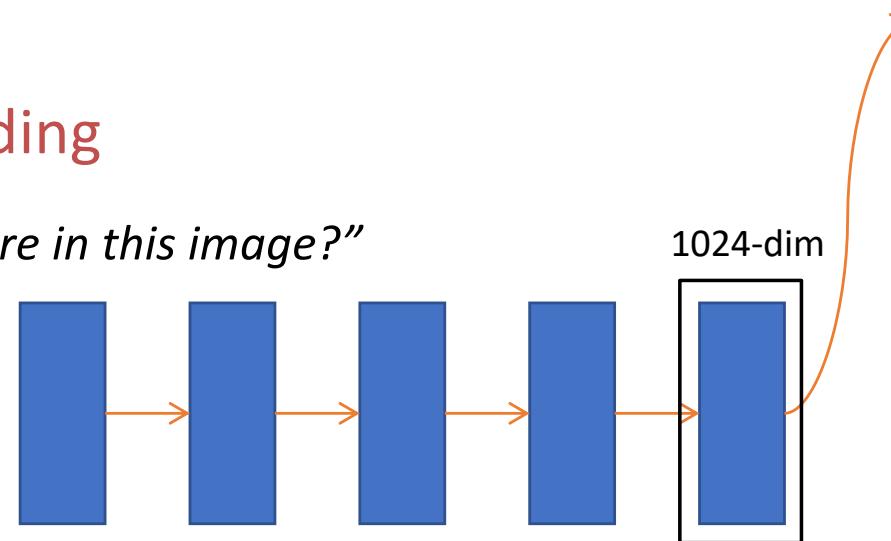
基准模型：双通道VQA模型

Image Embedding



Question Embedding

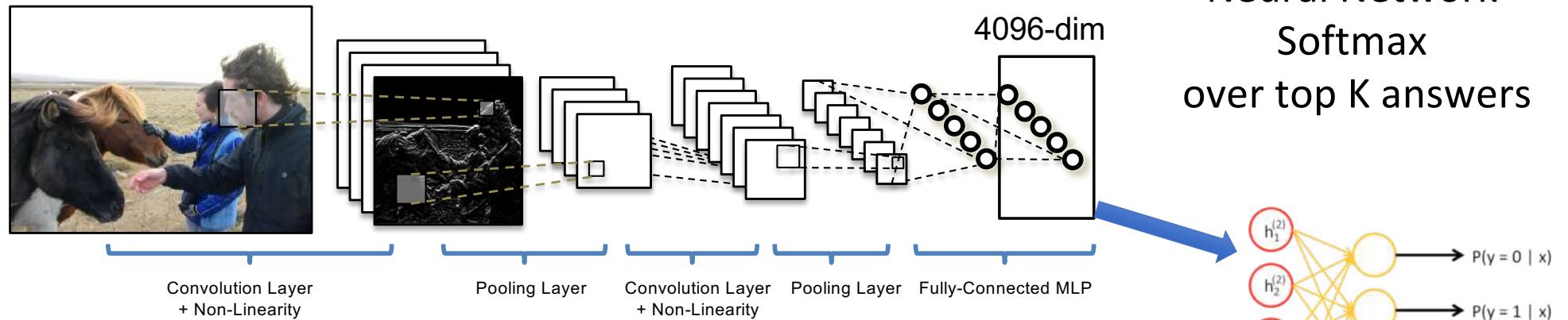
“How many horses are in this image?”



2.2 视觉问答模型

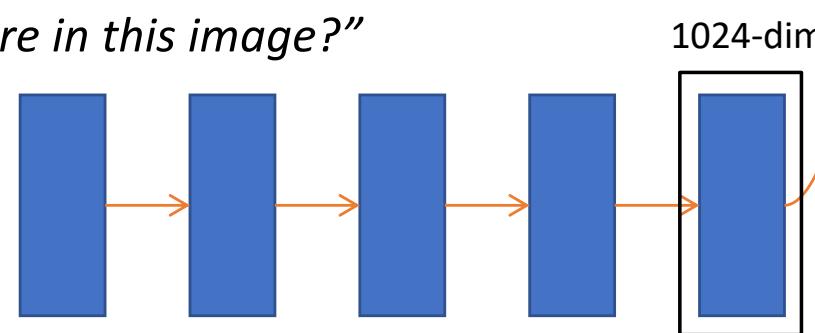
基准模型：双通道VQA模型

Image Embedding

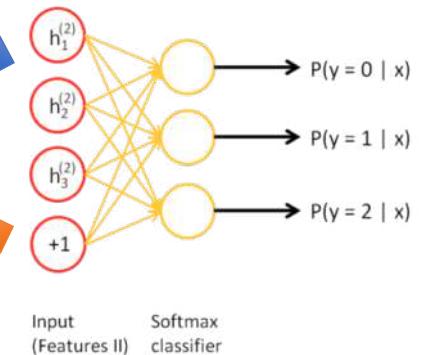


Question Embedding

"How many horses are in this image?"



Neural Network
Softmax
over top K answers

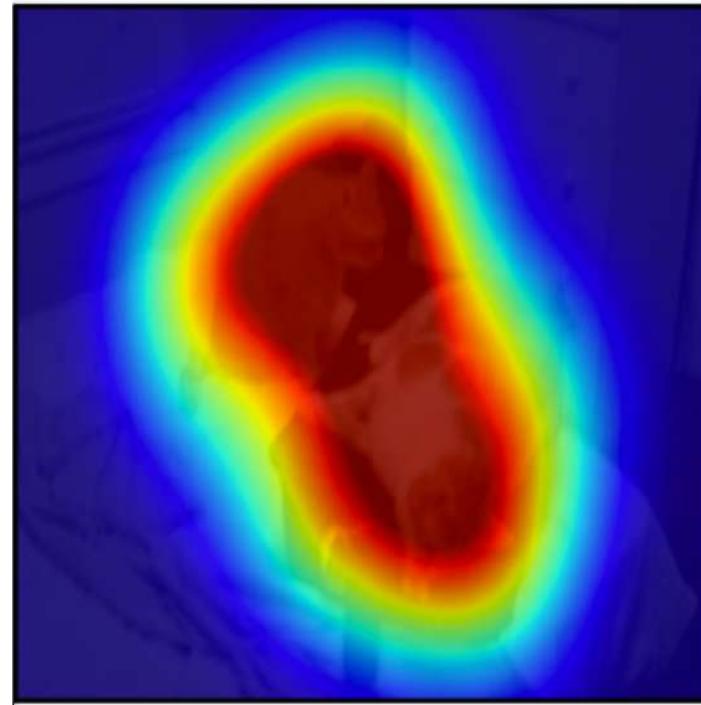


2.2 视觉问答模型

基于注意力机制的VQA模型 (2016 EMNLP)



What is the name of the
cafe? - bagdad

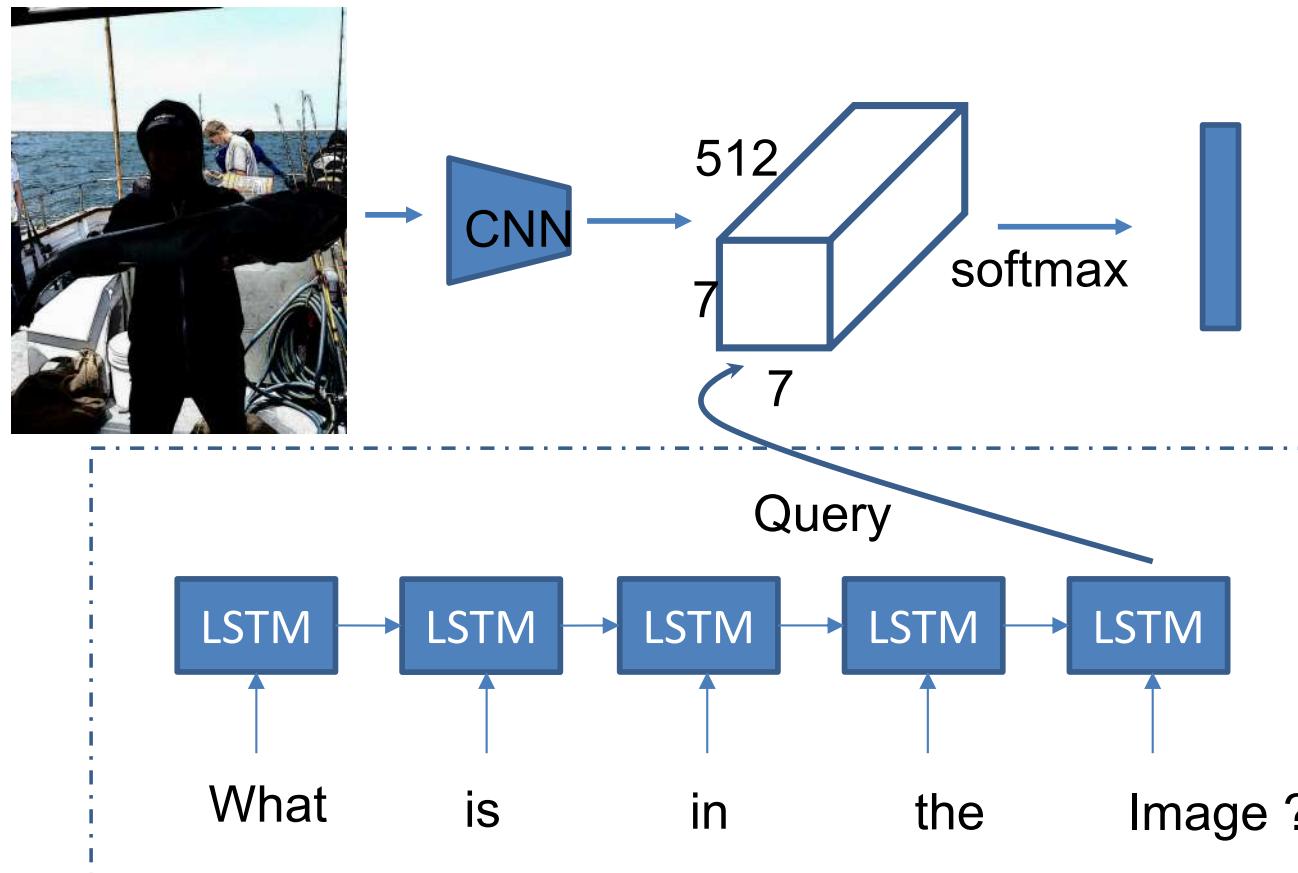


What number of cat is
laying on bed? - 2

2.2 视觉问答模型

回顾：注意力机制

Attention Mechanism (Soft)



2.2 视觉问答模型

基于注意力机制的VQA模型(2016 CVPR)

- 如果一轮注意力不能准确关注到目标区域?



What are sitting in the basket on a bicycle?

2.2 视觉问答模型

基于注意力机制的VQA模型(2016 CVPR)

- 如果一轮注意力不能准确关注到目标区域？**采用多轮注意力**



What are sitting in the basket on a bicycle?

2.2 视觉问答模型

基于注意力机制的VQA模型(2016 NIPS)

- 如何对问题分解、关注问题中最相关的信息？

What
are
sitting
in
the
basket
on
a
Bicycle
?



2.2 视觉问答模型

基于注意力机制的VQA模型(2016 NIPS)

- 如何对问题分解、关注问题中最相关的信息？

What
are
sitting

in
the
basket

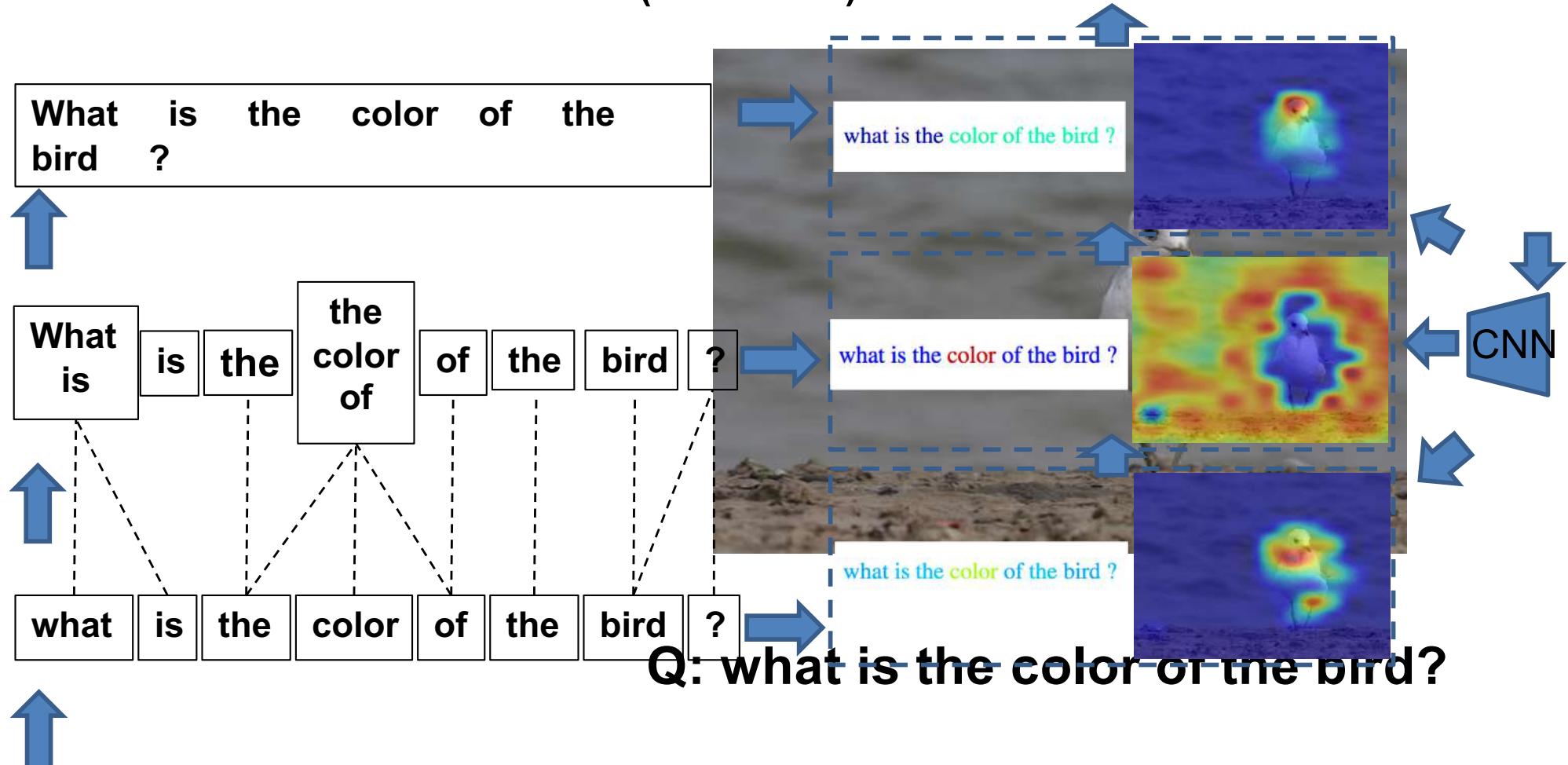
on
a
Bicycle
?



2.2 视觉问答模型

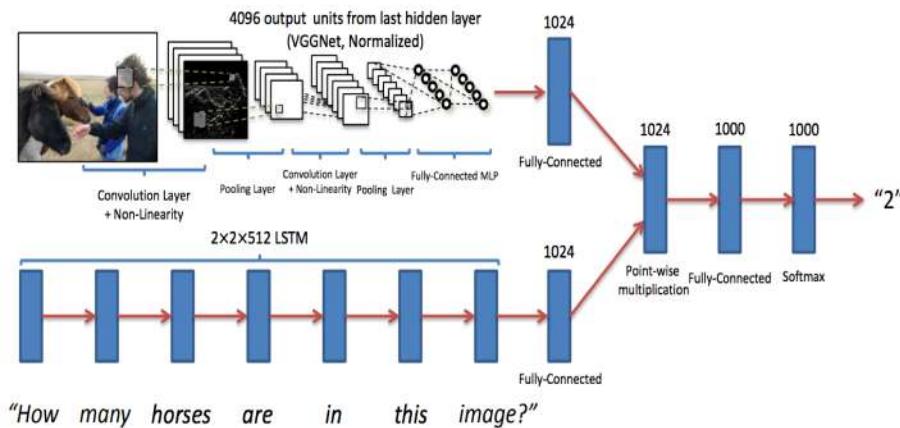
基于注意力机制的VQA模型(2016 NIPS)

- Hierarchical Co-Attention (NIPS 2016)

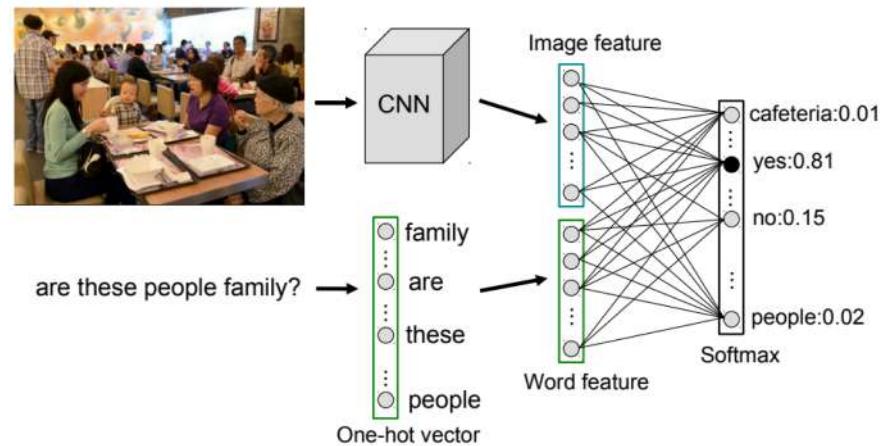


2.2 视觉问答模型

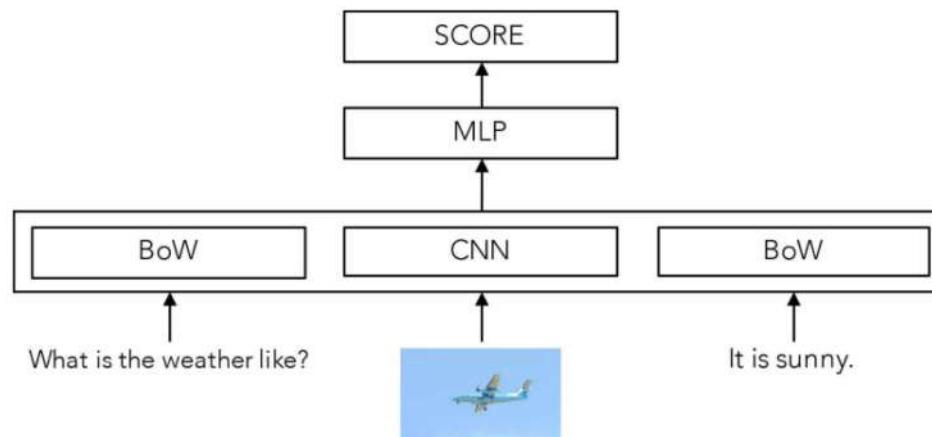
其他基础、简单的VQA模型



Aishwarya, Lu, Antol et.al 2015



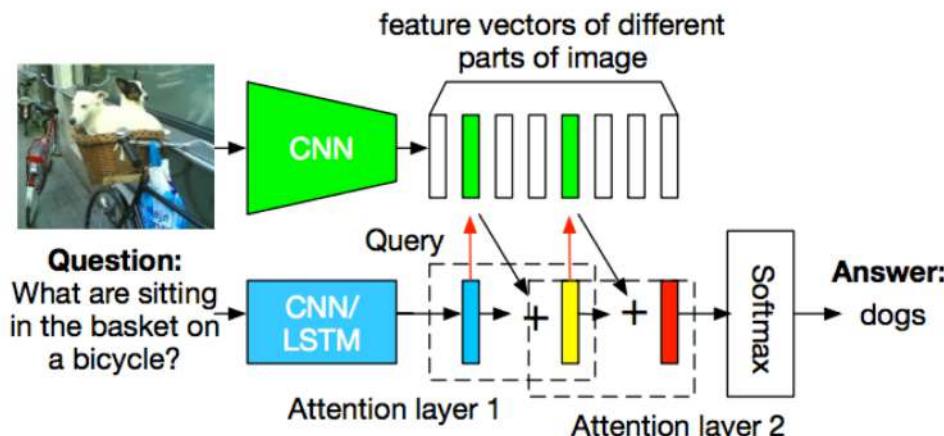
Zhou et.al 2015



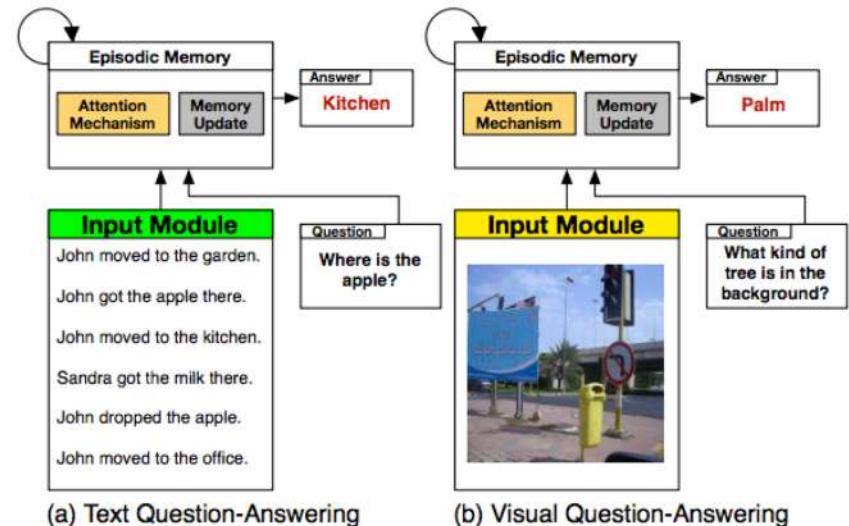
Jabri et.al 2016

2.2 视觉问答模型

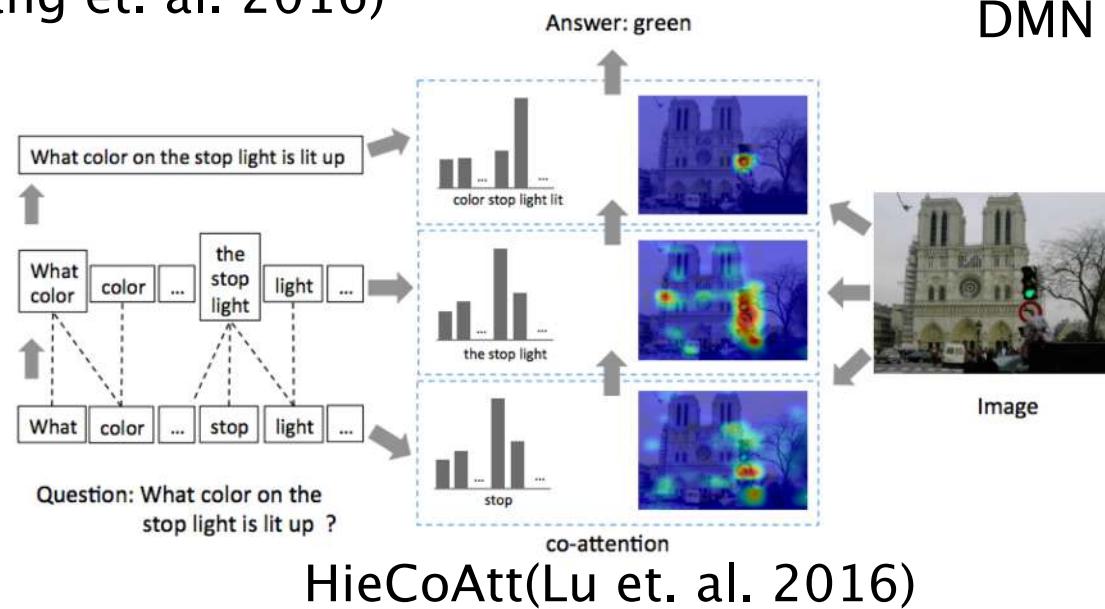
基于注意力机制的VQA模型



SAN (Yang et. al. 2016)



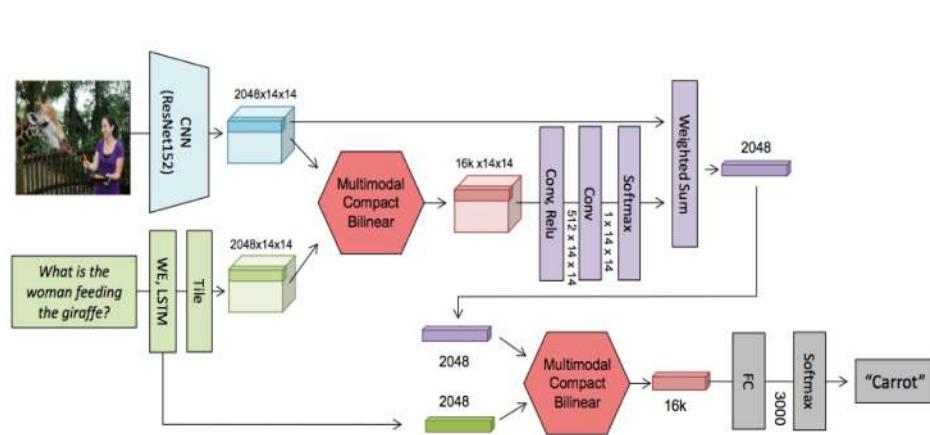
DMN (Xiong et. al. 2016)



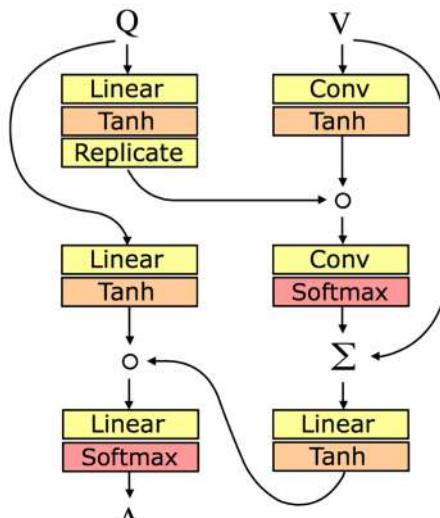
HieCoAtt(Lu et. al. 2016)

2.2 视觉问答模型

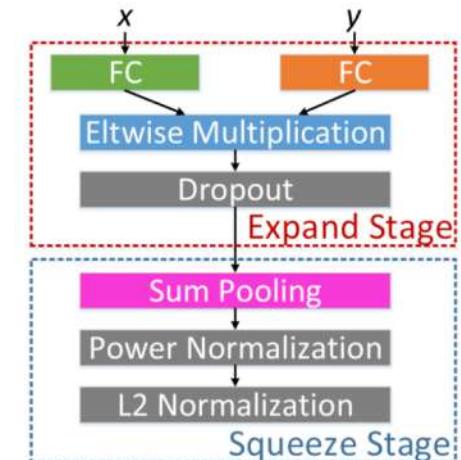
多模态特征学习的VQA模型 (如何融合两种特征、解决跨模态关联问题?)



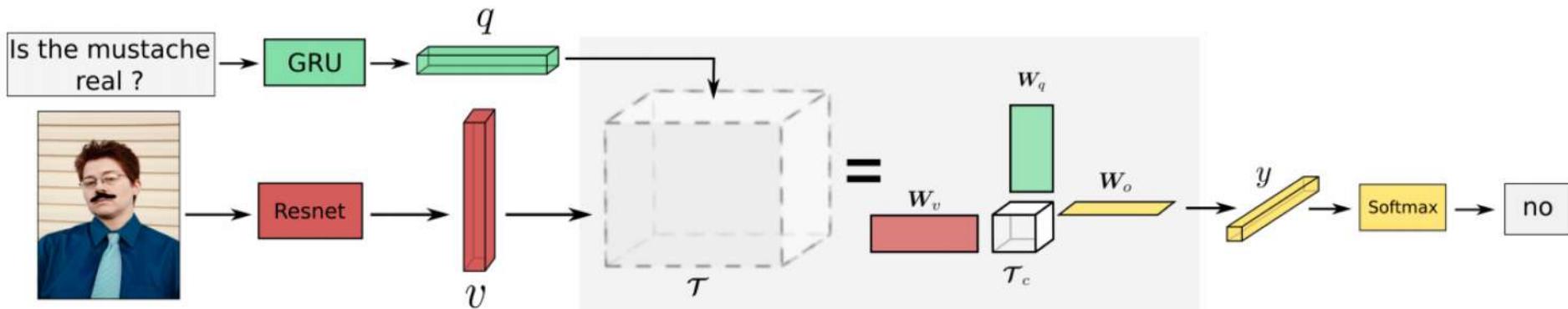
[MCB\(Fukui et al 2016\)](#)



[MLB \(Kim et al 2016\)](#)



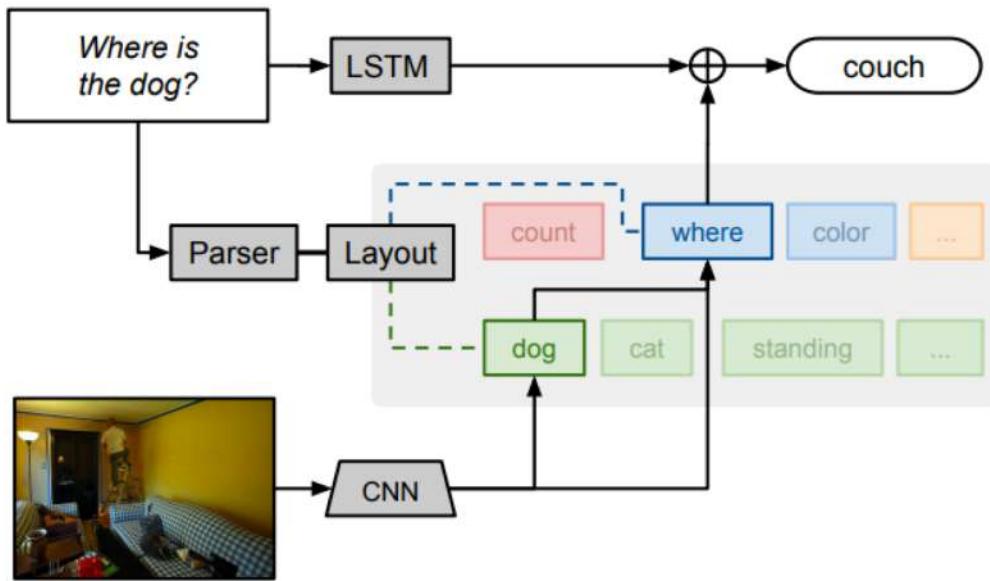
[MFB \(Yu et al 2018\)](#)



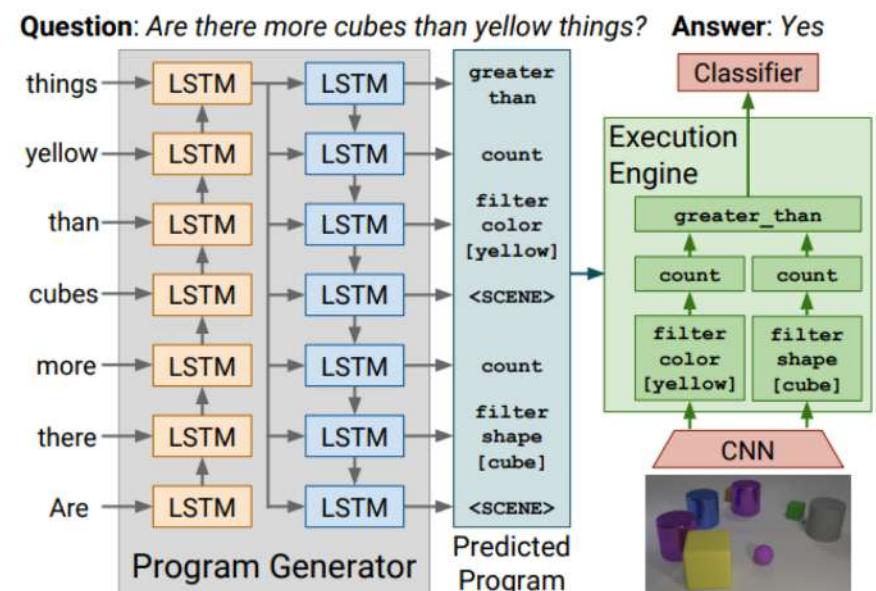
[MUTAN \(Younes et al 2017\)](#)

2.2 视觉问答模型

Modular Network / Programmer (如何解决组合泛化问题?)



[Modular Network \(Andreas et. al. 2015\)](#)



[Jonhson et. al. 2017](#)

2.2 视觉问答模型

- 挑战：如何解决多模态知识推理问题？

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

IJCAI 2020

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

1



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

2



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

3

Microsoft
Research
微软亚洲研究院

4


THE UNIVERSITY
ofADELAIDE

Content

1. Introduction to VQA
2. FVQA Dataset
3. Related Work
4. Motivation
5. Methodology
6. Experiments
7. Future Work

Introduction to VQA

Example:



Question:

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

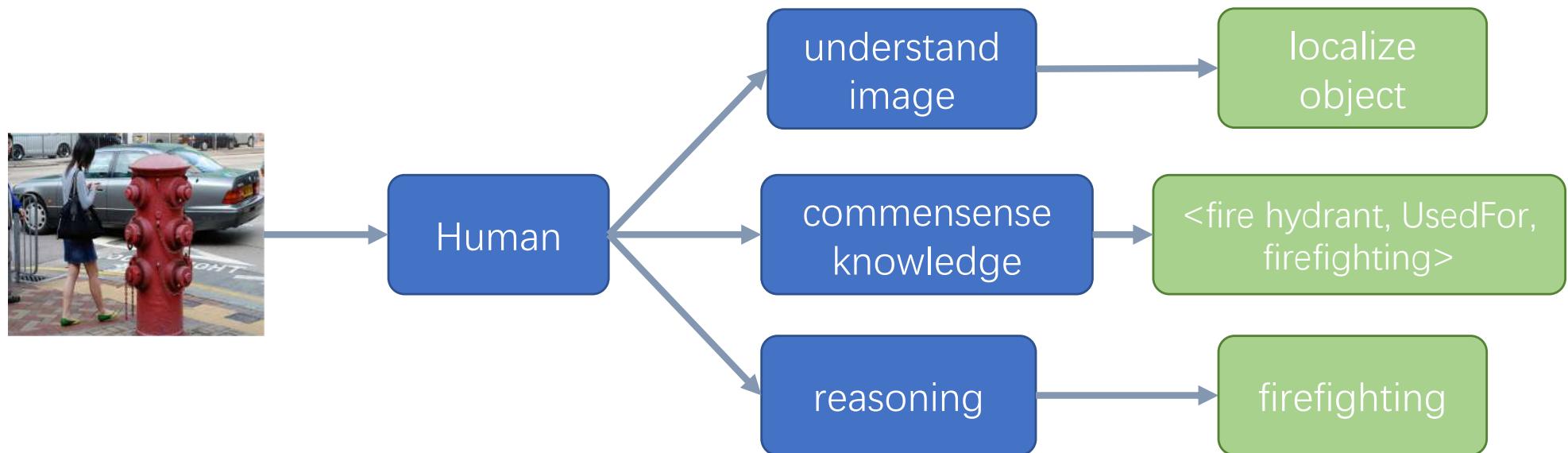
Visual info is not enough!

Factual Knowledge:

<fire hydrant, UsedFor, firefighting>

- Most of visual questions require external knowledge to answer.
- Existing VQA datasets rarely involve questions that require external knowledge to answer.

FVQA Dataset



FVQA Dataset

Dataset	Number of images	Number of questions	Num. question categories	Average quest. length	Average ans. length	Knowledge Bases	Supporting-Facts
DAQUAR [12]	1,449	12,468	4	11.5	1.2	-	-
COCO-QA [9]	117,684	117,684	4	8.6	1.0	-	-
VQA-real [5]	204,721	614,163	20+	6.2	1.1	-	-
Visual Genome [11]	108,000	1,445,322	7	5.7	1.8	-	-
Visual7W [10]	47,300	327,939	7	6.9	1.1	-	-
Visual Madlibs [8]	10,738	360,001	12	6.9	2.0	-	-
VQA-abstract [5]	50,000	150,000	20+	6.2	1.1	-	-
VQA-balanced [67]	15,623	33,379	1	6.2	1.0	-	-
KB-VQA [61]	700	2,402	23	6.8	2.0	1	-
Ours (FVQA)	2,190	5,826	32	9.5	1.2	3	✓

- collects a knowledge base of 190,000 facts ($\langle e_1, r, e_2 \rangle$) from DBpedia, ConceptNet, WebChild.
- each \langle question-answer \rangle pair is associated with a supporting-fact.
- about 97.6% of collected questions require commonsense knowledge.

[Peng Wang et al.] TPAMI 2018 FVQA: Fact-based Visual Question Answering

Related Work

- query-mapping based methods
- learning based methods

Related Work

query-mapping based methods

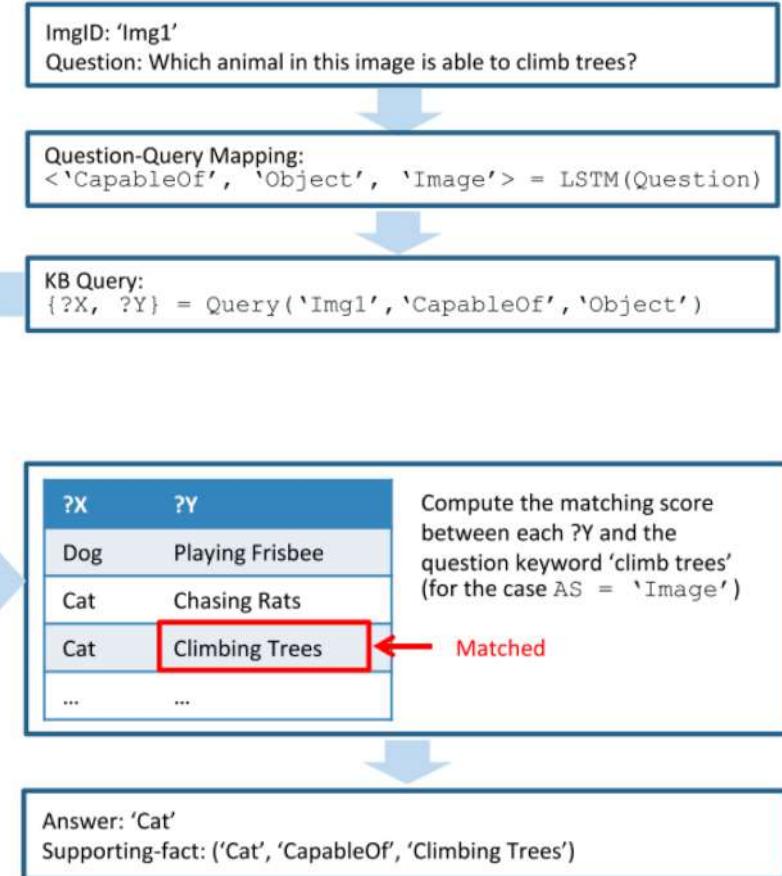
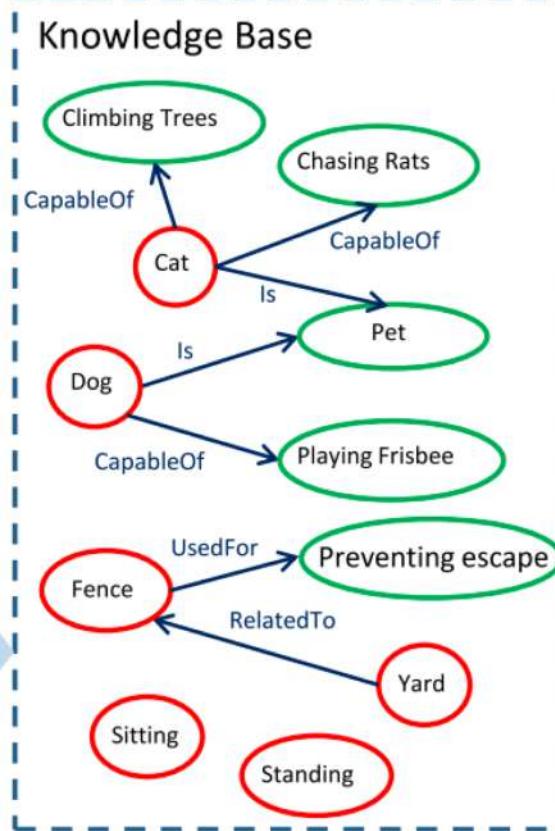
Img1



Object Detector
Scene Classifier
Attribute Classifier

Extracted Visual Concepts:

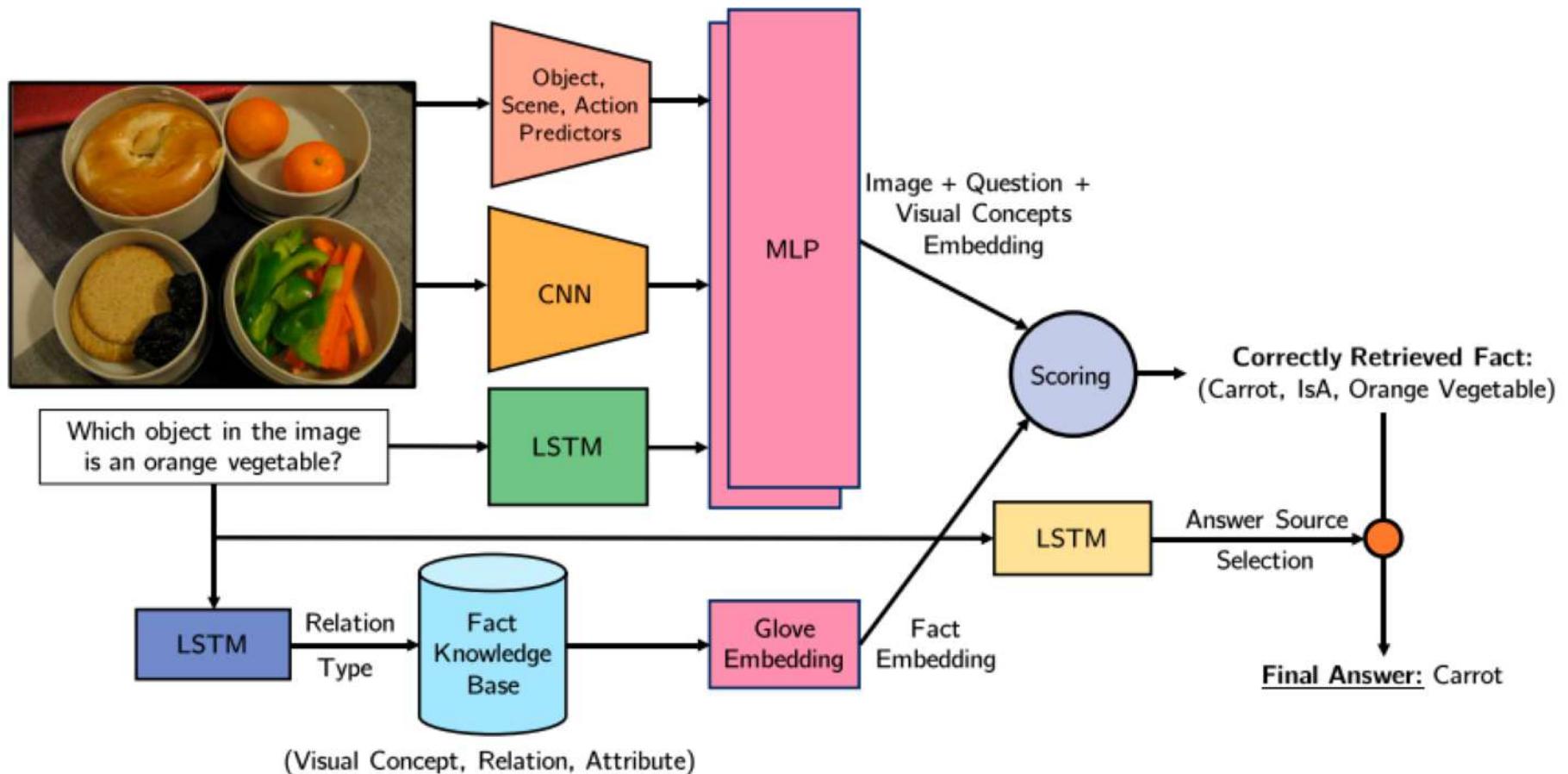
Object	Cat
	Dog
	Fence
Scene	Yard
Action	Sitting
	Standing
	Standing



[Peng Wang et al. TPAMI 2018] FVQA: Fact-based Visual Question Answering

Related Work

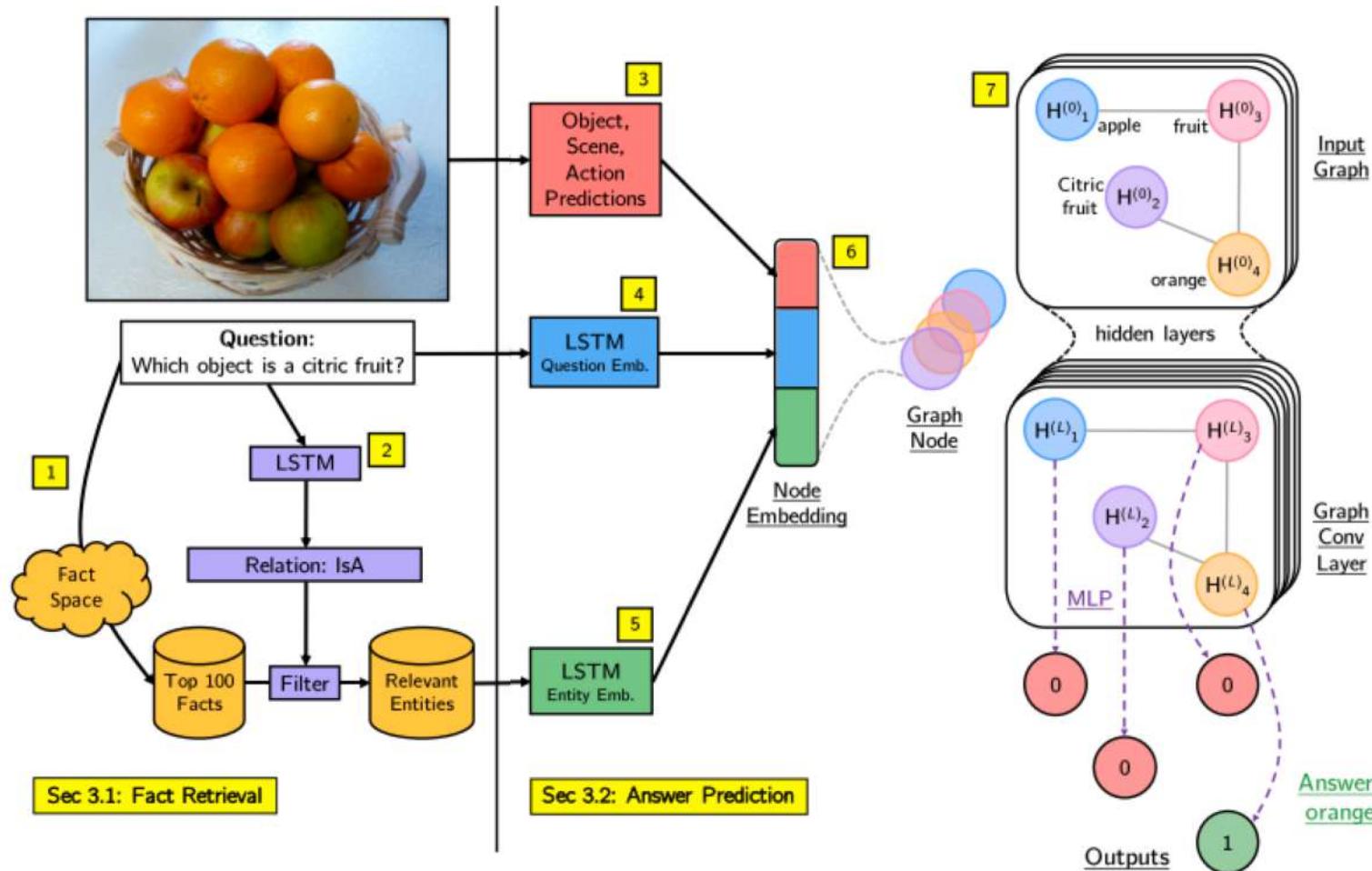
learning based methods



[Narasimhan et al. ECCV 2018 Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering]

Related Work

learning based methods

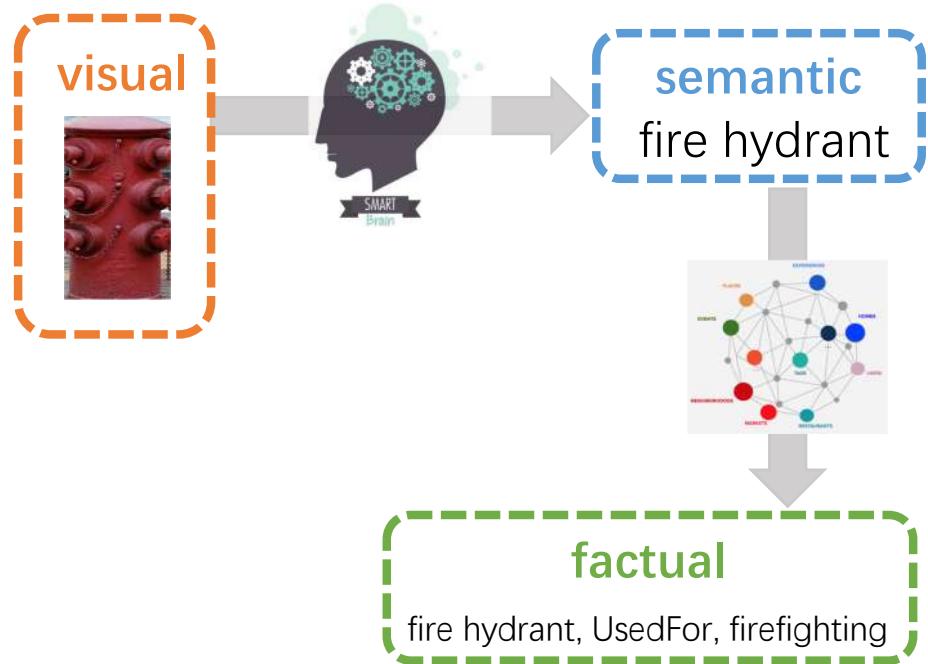


[Narasimhan et al. NIPS 2018 *Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering*]

Limitations

- Vulnerable to misconceptions
- Reasoning without visual information
- Ignoring relational and semantic information
- Incorporating visual information without selection

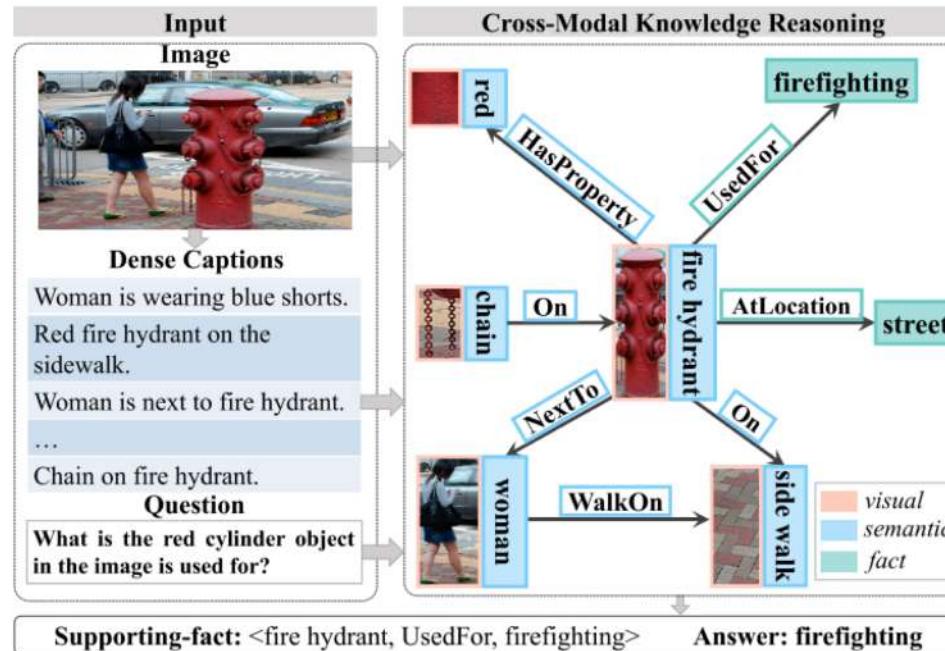
Motivation



Human:

1. visually localize ‘the red cylinder’
2. semantically recognize it as ‘fire hydrant’
3. connects the knowledge that ‘fire hydrant is used for firefighting’

Motivation



Goal

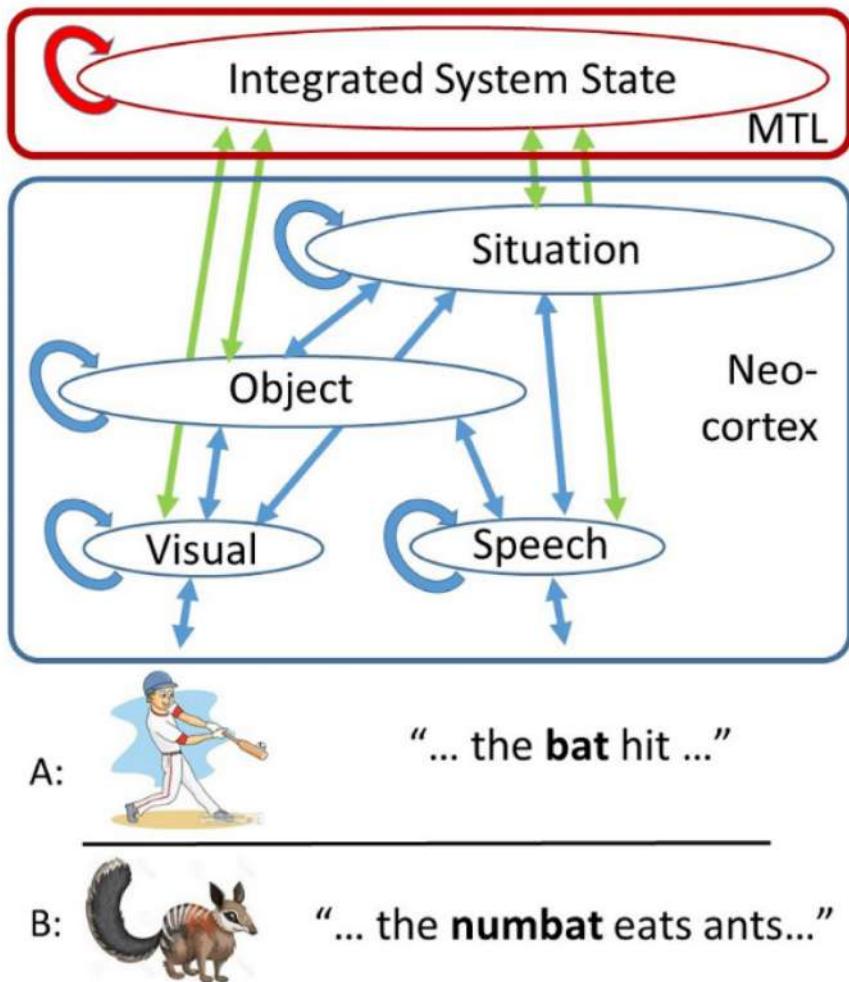
Find the optimal node from the factual as the final answer.

Challenge

How to collect the **question-oriented** and **information-complementary** evidence from visual, semantic and knowledge perspectives ?

Motivation

Human Cognition Theory



Medial temporal lobe system:

- provide a network that stores an integrated embedding of the neocortical system state

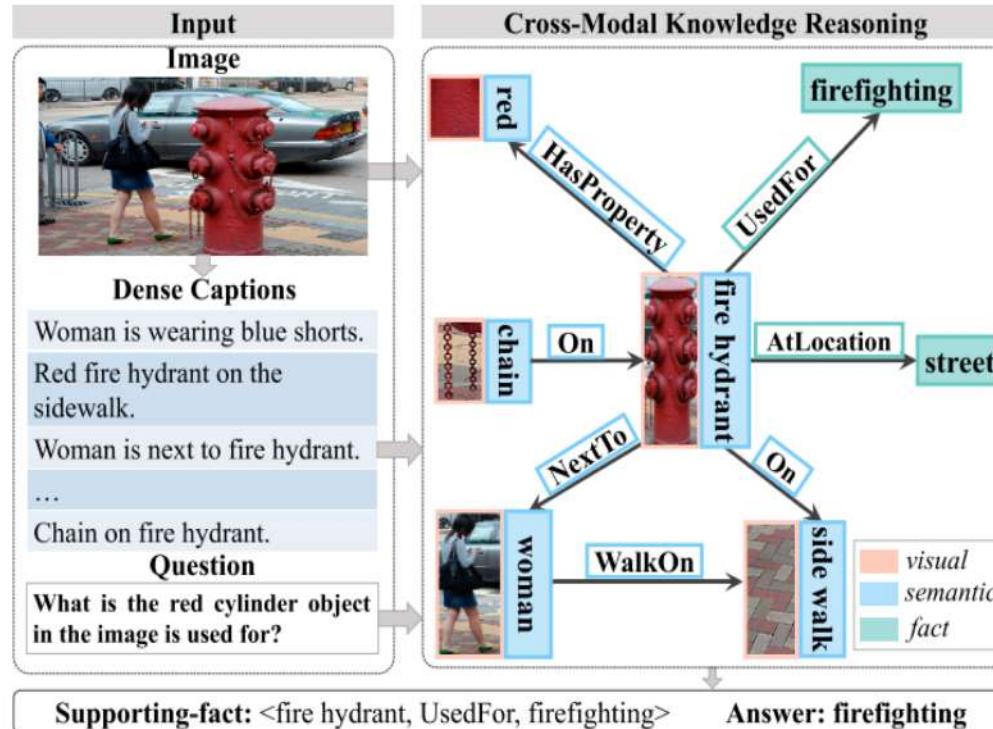
Neocortical system:

- Each oval forming an embedding (representation) of a specific kind of information.
- Blue arrows represent learned connections that allow the embeddings to constrain each other.

[McClelland et al., Extending machine language models towards human-level language understanding, arxiv, 2019]

Motivation

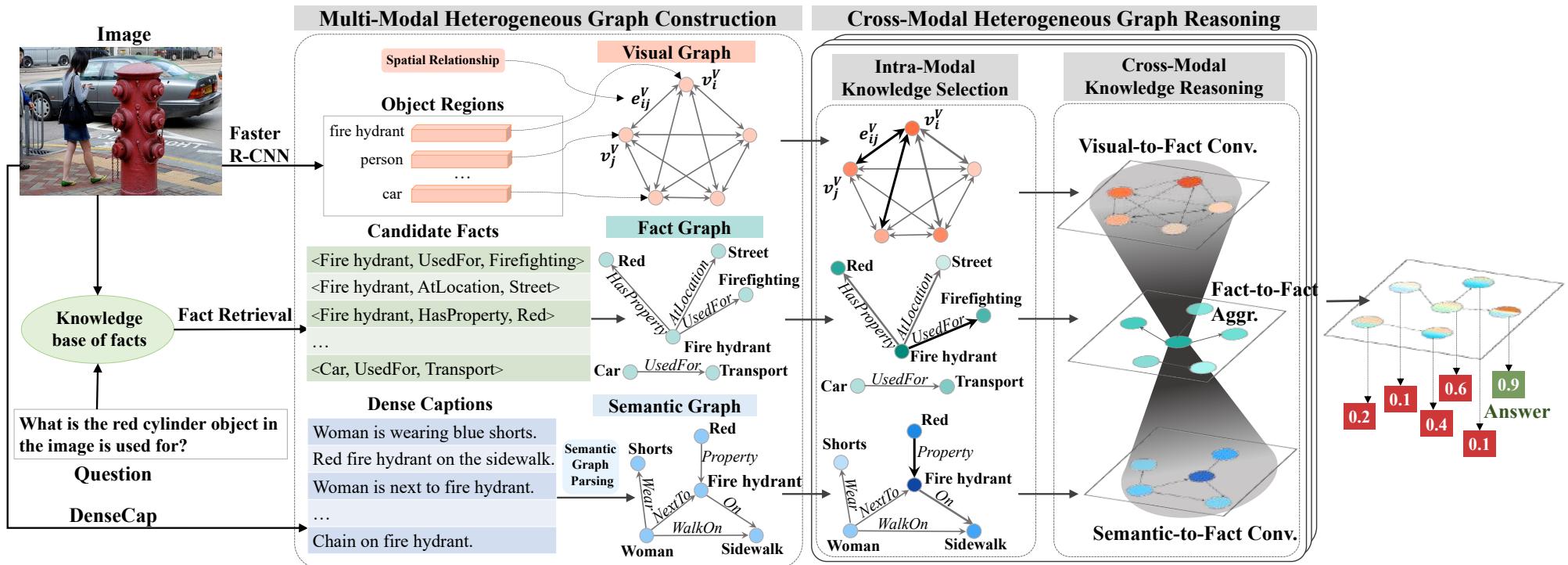
Multi-Layer Cross-Modal Knowledge Reasoning (Mucko)



- **Multi-layer graph representation**
 - **visual layer:** object appearance and visual relationships
 - **semantic layer:** high-level abstraction
 - **fact layer:** knowledge of facts
- **Heterogeneous graph convolutional network**
adaptively collect complementary evidence in the multi-layer graphs.

Motivation

Framework

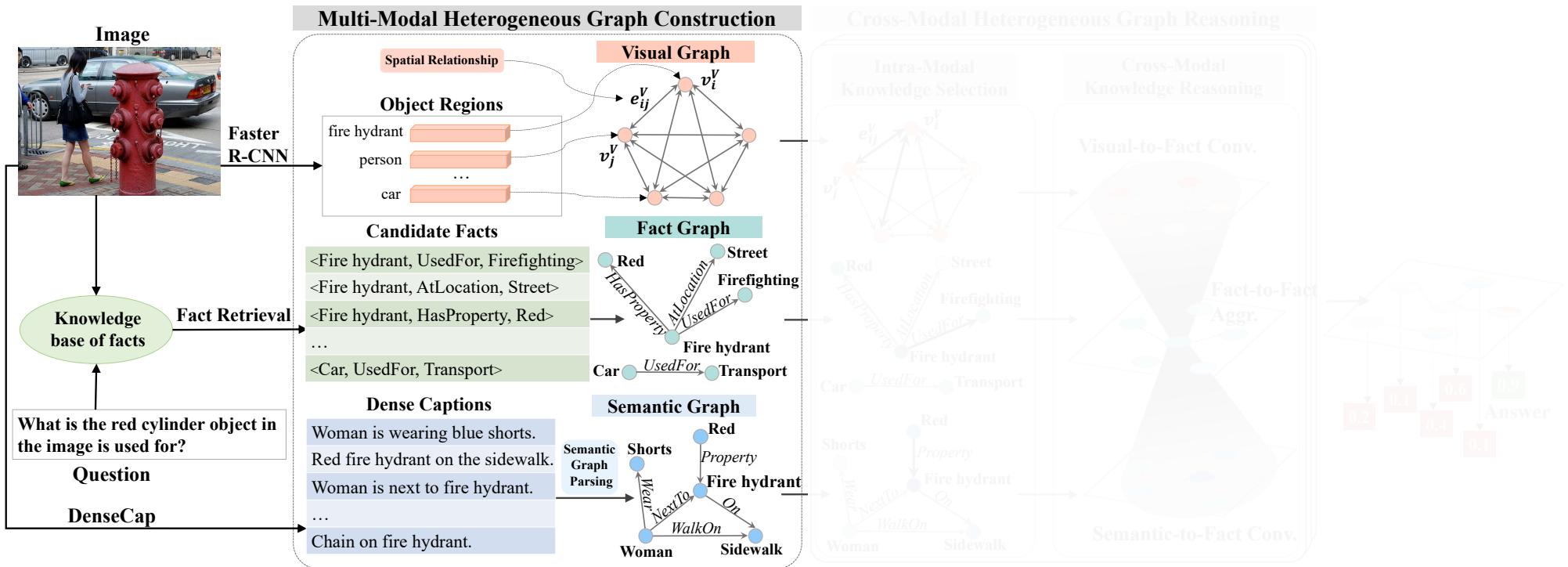


Two stage :

- Multi-Modal Heterogenous Graph Construction
- Cross-Modal Heterogeneous Graph Reasoning

Motivation

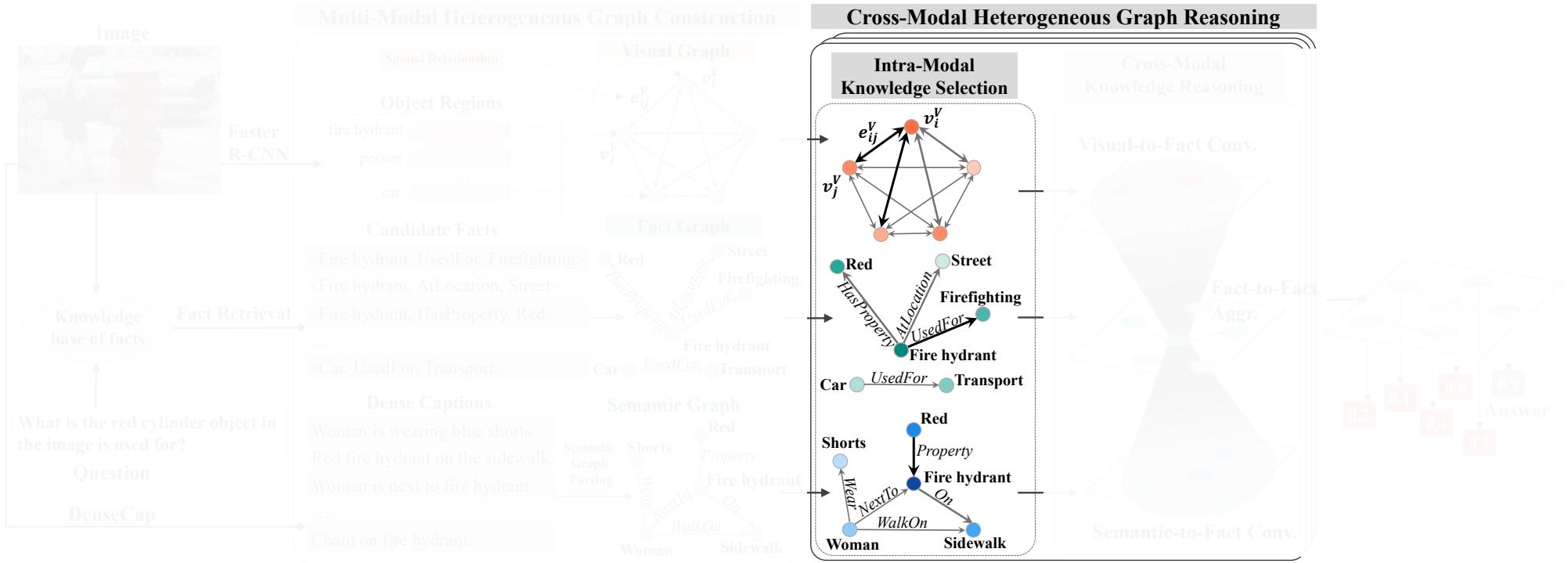
Framework



➤ Multi-Modal Heterogenous Graph Construction

Motivation

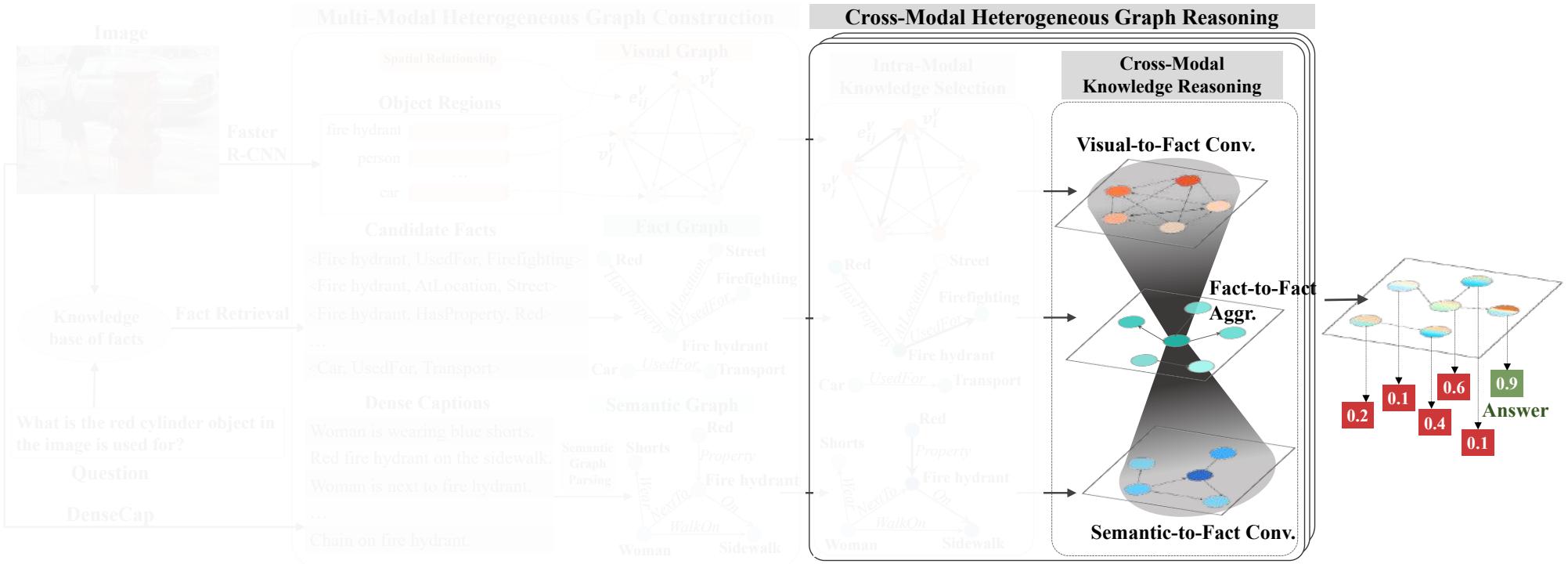
Framework



- Cross-Modal Heterogeneous Graph Reasoning
 - ◆ Intra-Modal Knowledge Selection

Motivation

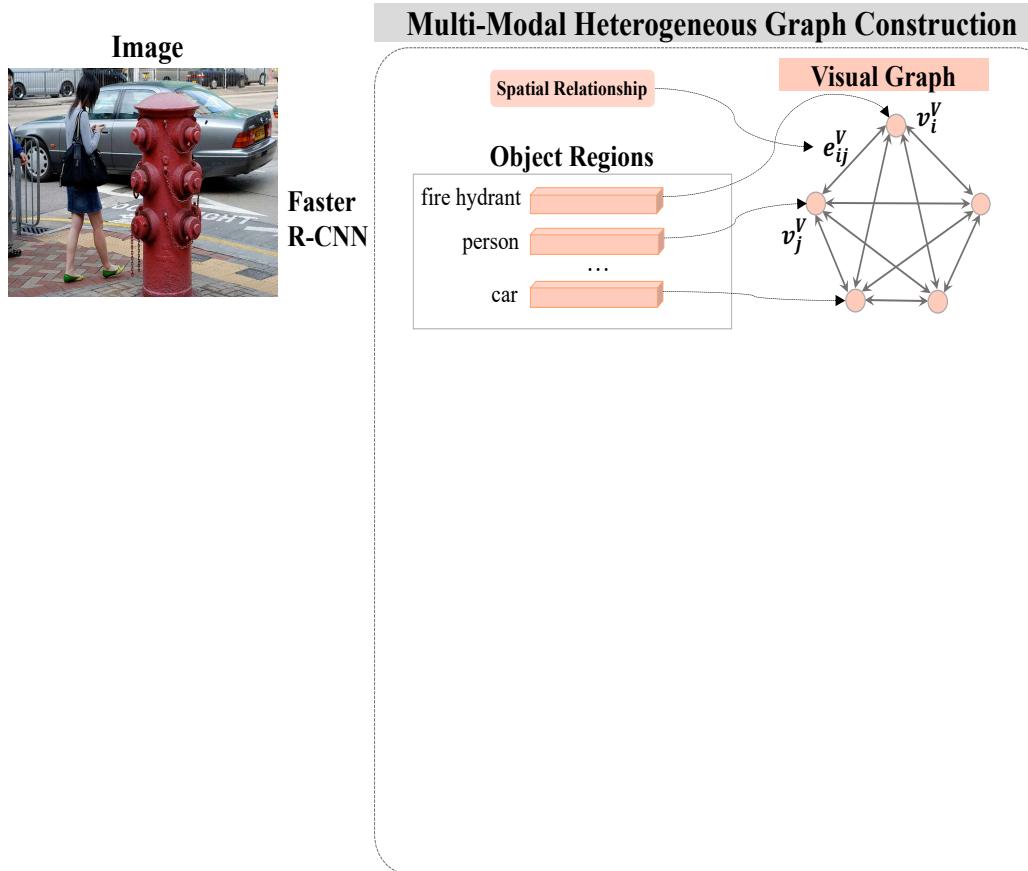
Framework



- Cross-Modal Heterogeneous Graph Reasoning
 - ◆ Cross-Modal Knowledge Reasoning

Motivation

Multi-Modal Heterogeneous Graph Construction

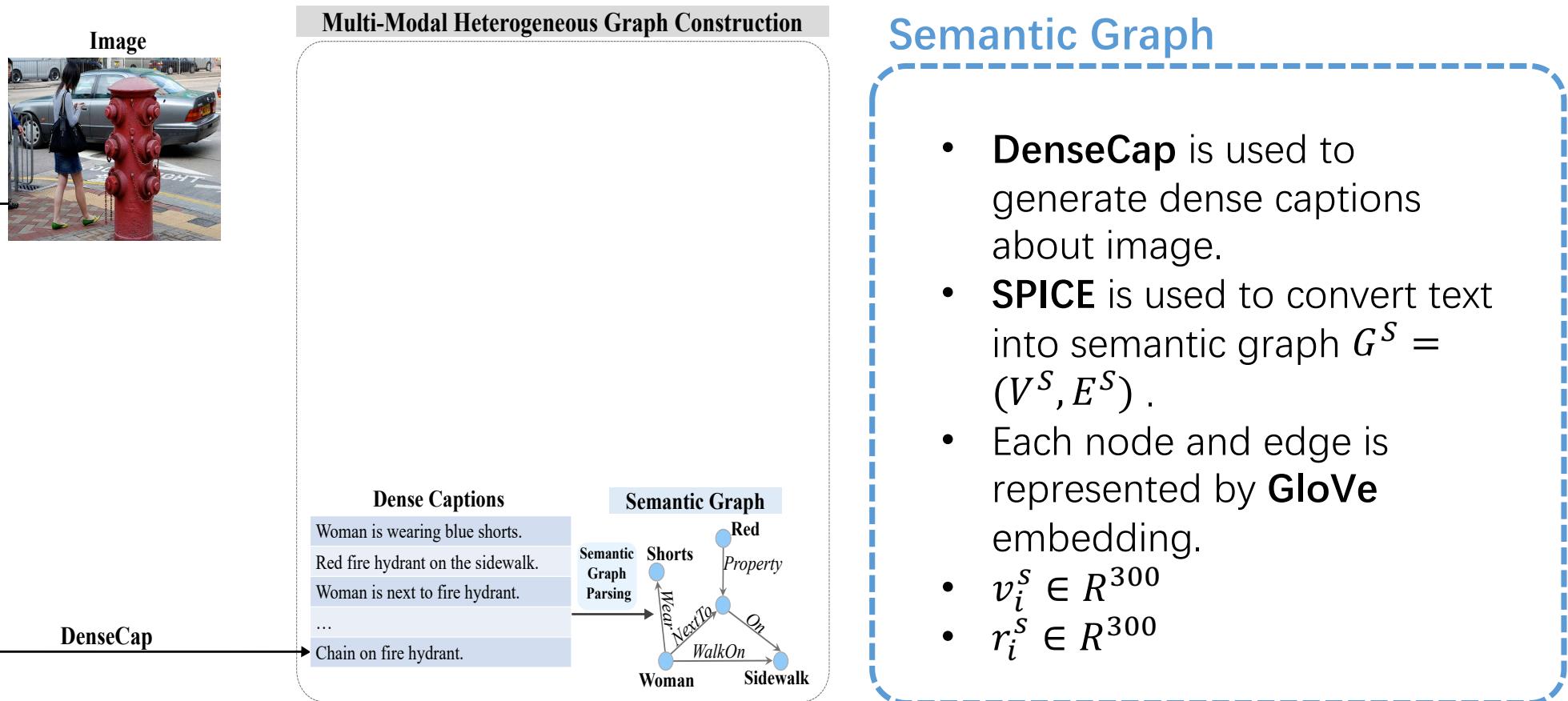


Visual Graph

- Faster-RCNN is used to extract a set of objects, $O = \{o_i\}_{i=1}^K (K = 36)$.
- Each object has a 2048-d feature.
- Construct a visual graph $G^V = (V^V, E^V)$ over O
- $v_i^V \in R^{2048}$
- spatial relationship $r_i^V = \left[\frac{x_j - x_i}{w_i}, \frac{y_j - y_i}{h_i}, \frac{w_j}{w_i}, \frac{h_j}{h_i}, \frac{w_j h_j}{w_i h_i} \right]$

Motivation

Multi-Modal Heterogeneous Graph Construction



Motivation

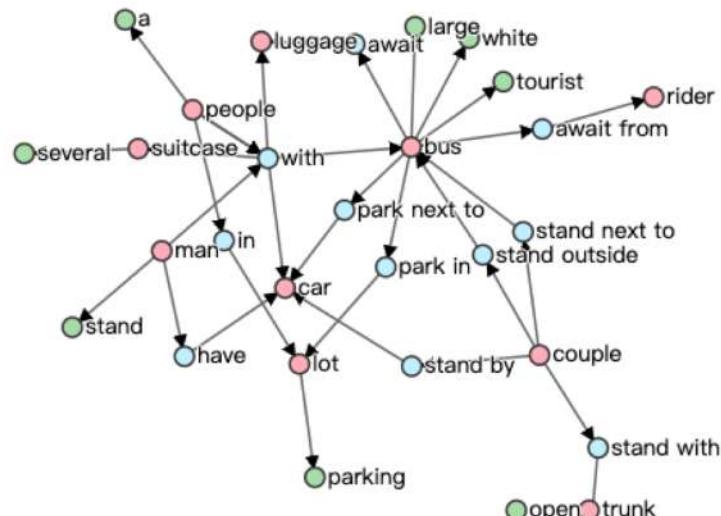
SPICE

Reference captions

- "A white bus parked in a parking lot next to a car."
- "A couple of people standing outside a bus."
- "People in a parking lot with luggage, a bus, and some cars."
- "A man with several suitcases stands next to a bus while another couple stands by their car with an open trunk."
- "A large tourist bus is awaiting luggage from riders."

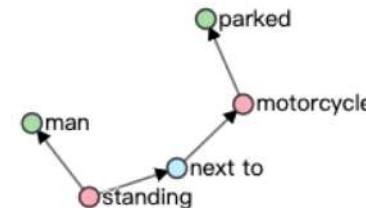


Reference scene graph



Candidate caption & scene graph

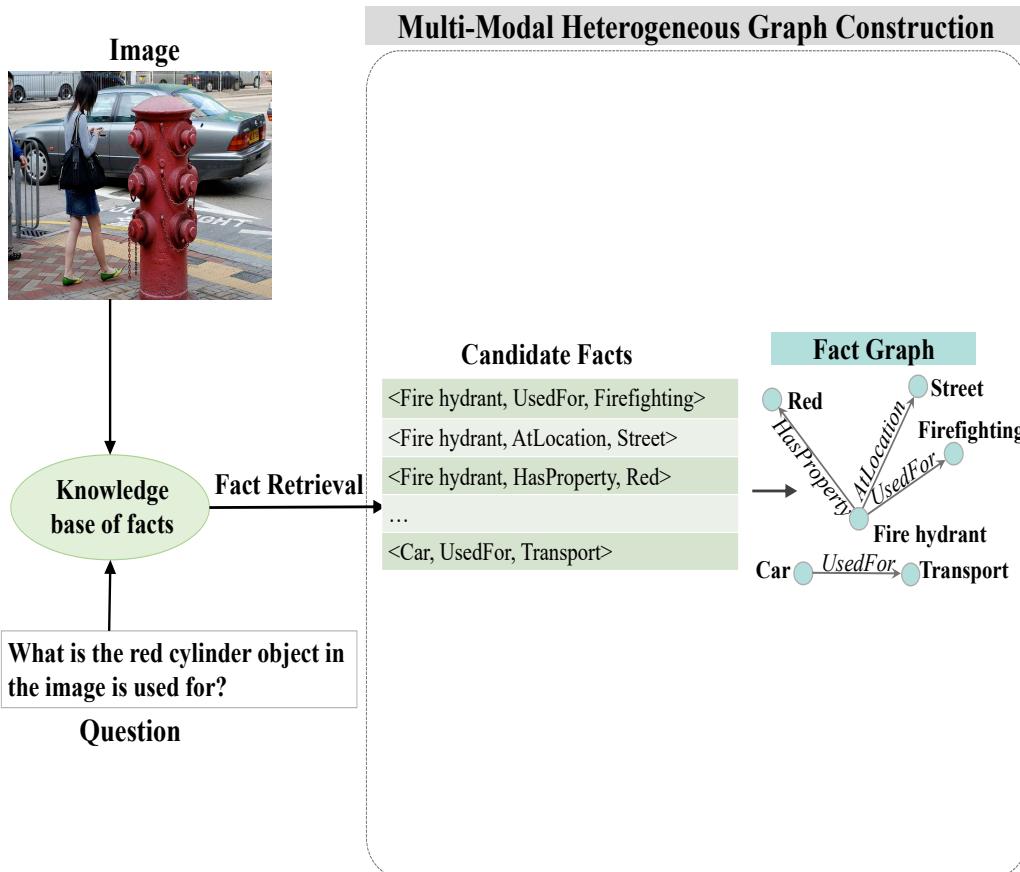
- "a man standing next to a parked motorcycle"



SPICE F-Score: 0, Pr: 0, Re: 0

Motivation

Multi-Modal Heterogeneous Graph Construction

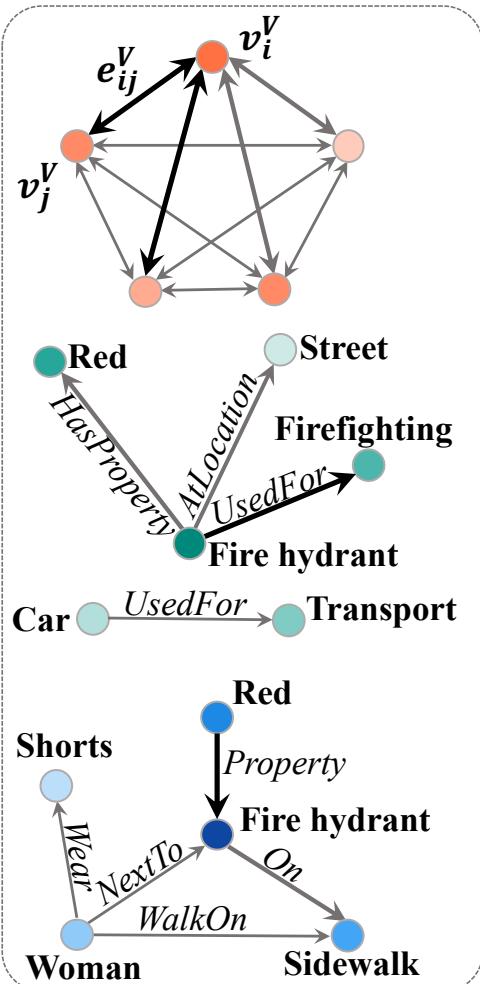


Fact Graph

- for each *fact* $\langle e1, r, e2 \rangle$ of KB, compute the cosine similarities of $(e1, e2)$ and $(o1, o2, \dots, o36)$
- average these similarities to assign a score to the *fact*
- sort and select top-k facts according to scores.
- train a relation classifier to predict relation type based on the question
- filter the facts according to relation type.

Motivation

Intra-Modal Knowledge Selection

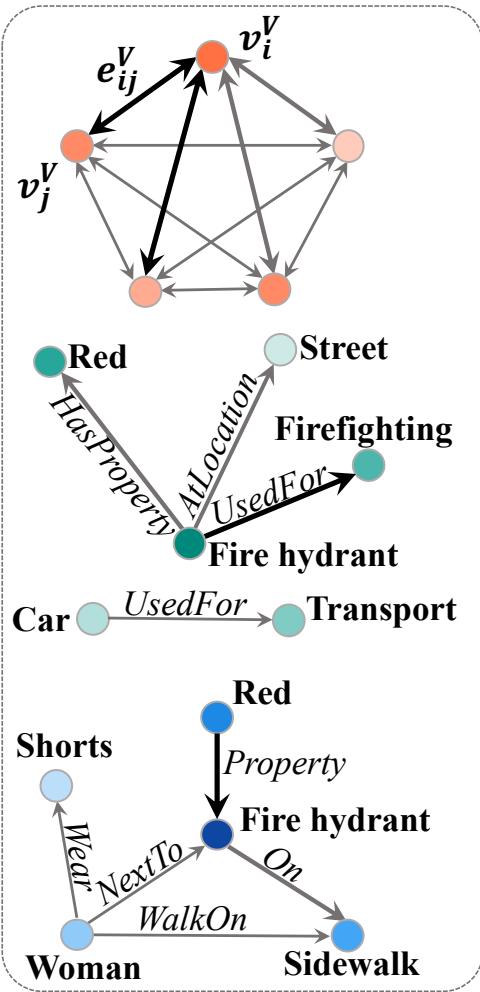


Question-guided Node Attention: evaluate the relevance of each node corresponding to the question by attention mechanism.

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

Motivation

Intra-Modal Knowledge Selection



Question-guided Node Attention: evaluate the relevance of each node corresponding to the question by attention mechanism.

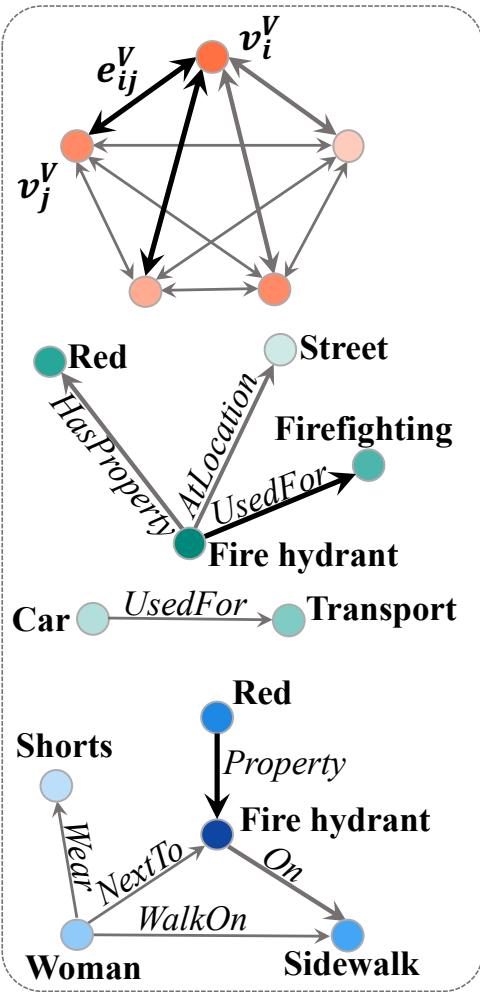
$$\alpha_i = \text{softmax}(\mathbf{w}_a^T \tanh(\mathbf{W}_1 \mathbf{v}_i + \mathbf{W}_2 \mathbf{q}))$$

Question-guided Edge Attention: evaluate the importance of edge constrained by the neighbor node v_j regarding to v_i as:

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

Motivation

Intra-Modal Knowledge Selection



Question-guided Node Attention: evaluate the relevance of each node corresponding to the question by attention mechanism.

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

Question-guided Edge Attention: evaluate the importance of edge constrained by the neighbor node v_j regarding to v_i as:

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

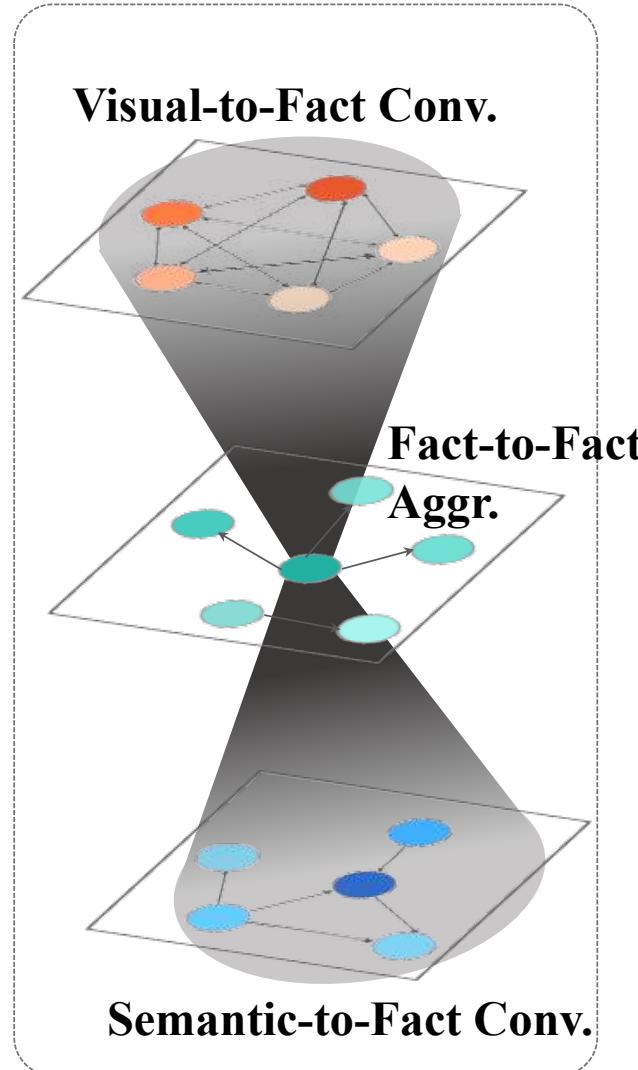
Intra-Modal Graph Convolution: gather and update the neighborhood information and the representation of v_i as:

$$\mathbf{m}_i = \sum_{j \in \mathcal{N}_i} \beta_{ji} \mathbf{v}'_j$$

$$\hat{\mathbf{v}}_i = \text{ReLU}(\mathbf{W}_7[\mathbf{m}_i, \alpha_i \mathbf{v}_i])$$

Motivation

Cross-Modal Knowledge Reasoning

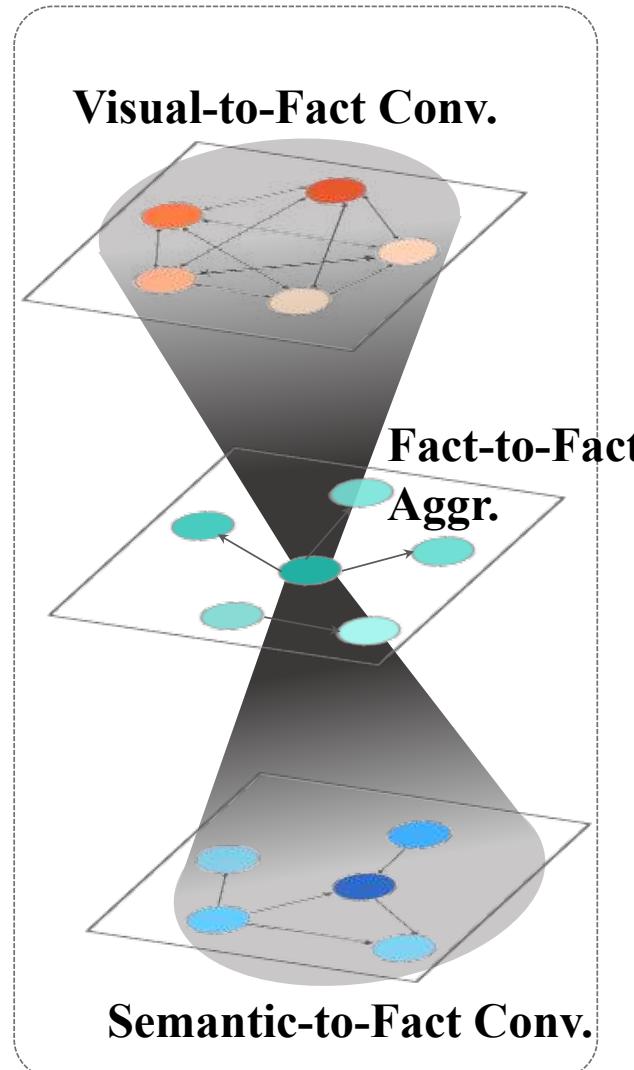


Visual-to-Fact Convolution: gather complementary information from visual graph by cross-modal convolutions.

$$\gamma_{ji}^{V-F} = \text{softmax}(\mathbf{w}_c \tanh(\mathbf{W}_8 \hat{\mathbf{v}}_j^V + \mathbf{W}_9 [\hat{\mathbf{v}}_i^F, \mathbf{q}]))$$

Motivation

Cross-Modal Knowledge Reasoning



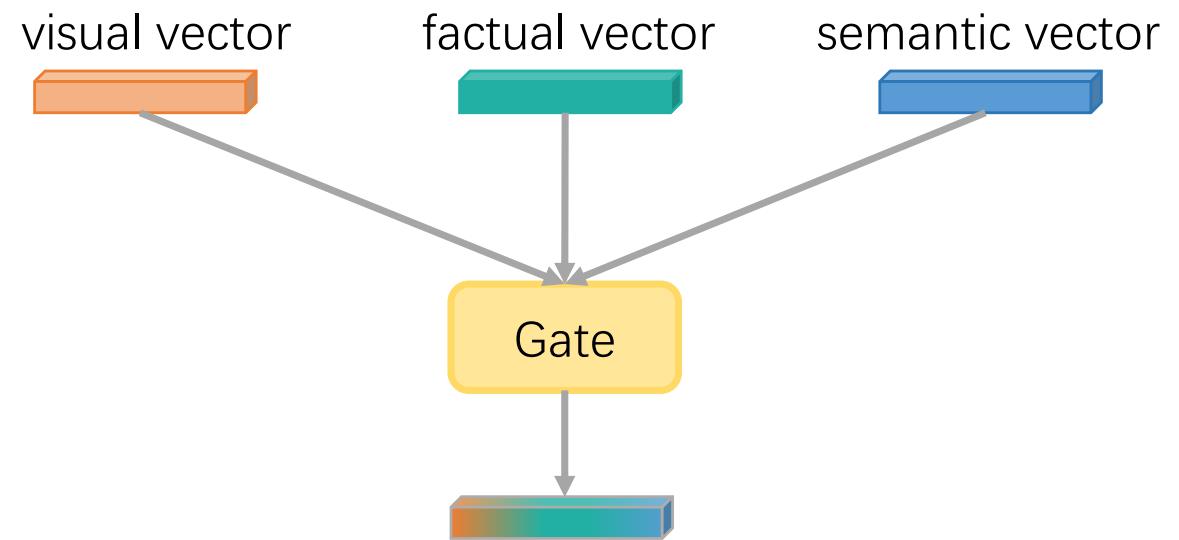
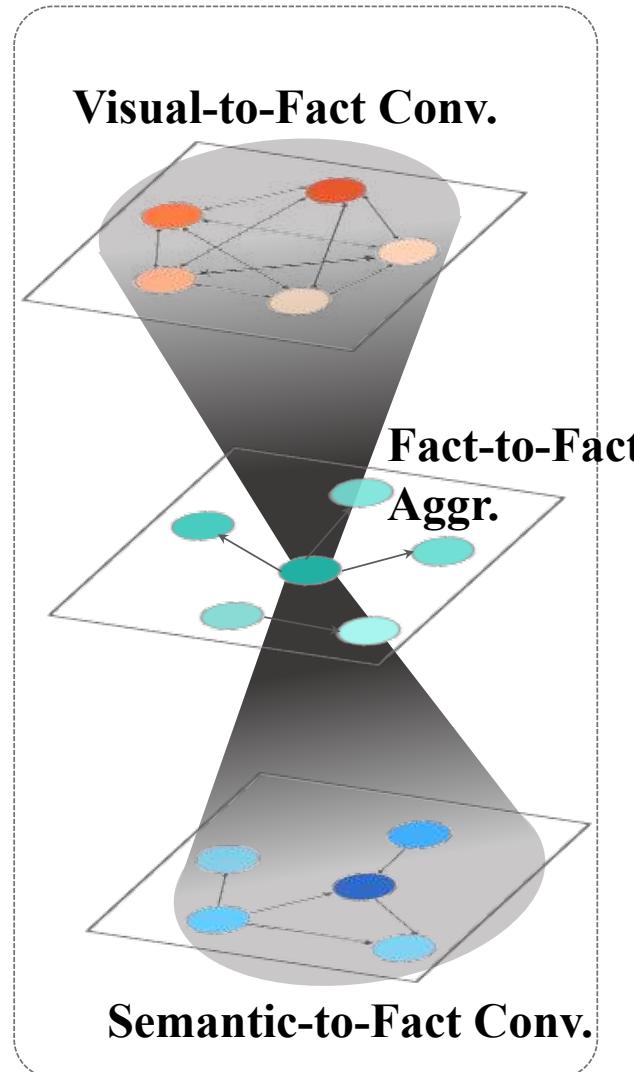
Visual-to-Fact Convolution: gather complementary information from visual graph by cross-modal convolutions.

$$\gamma_{ji}^{V-F} = \text{softmax}(\mathbf{w}_c \tanh(\mathbf{W}_8 \hat{\mathbf{v}}_j^V + \mathbf{W}_9 [\hat{\mathbf{v}}_i^F, \mathbf{q}]))$$

Semantic-to-Fact Convolution: gather complementary information from semantic graph by cross-modal convolutions.

Motivation

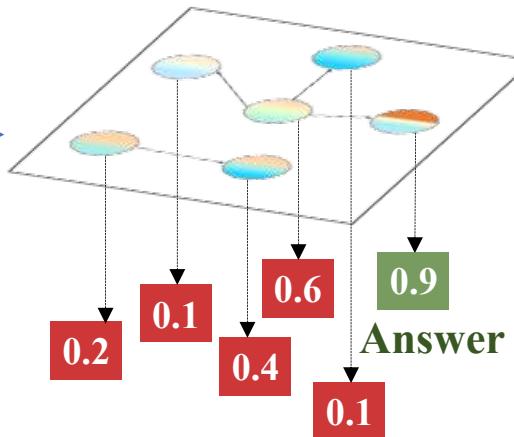
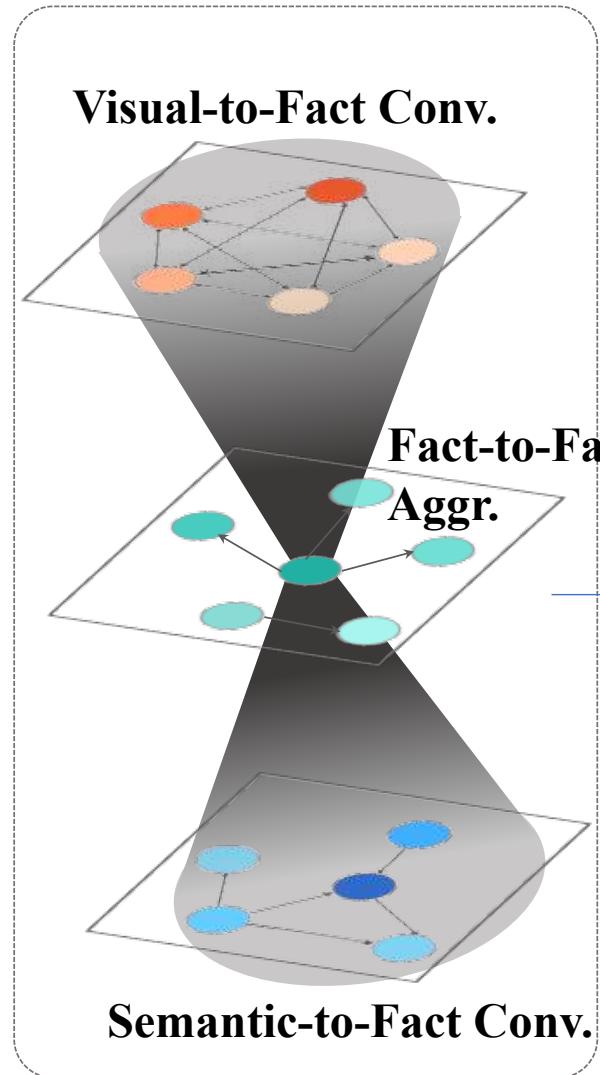
Cross-Modal Knowledge Reasoning



$$\begin{aligned}gate_i &= \sigma(\mathbf{W}_{10}[\mathbf{m}_i^{V-F}, \mathbf{m}_i^{S-F}, \hat{\mathbf{v}}_i^F]) \\ \tilde{\mathbf{v}}_i^F &= \mathbf{W}_{11}(gate_i \circ [\mathbf{m}_i^{V-F}, \mathbf{m}_i^{S-F}, \hat{\mathbf{v}}_i^F])\end{aligned}$$

Motivation

Cross-Modal Knowledge Reasoning



GNN+MLP

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

Experiments

Datasets

FVQA: The FVQA dataset consists of 2,190 images, 5,286 questions and a knowledge base of 193,449 facts. The knowledge base is constructed by extracting the top visual concepts from all the images in the dataset and querying those concepts from three knowledge bases, including DBpedia ConceptNet and WebChild.

Visual7W+KB: generated based on the test images in Visual7W by filling a set of question-answer templates that need to reason on both content and external knowledge. Visual7W+KB consists of 16,850 open-domain question-answer pairs based on 8,425 images in Visual7W test split.

OK-VQA: The questions in OK-VQA are manually generated by MTurk workers, which are not derived from specific knowledge bases. It requires the model to retrieve supporting knowledge from open-domain resources, which is much closer to the general VQA but more challenging for existing models. It contains 14,031 images and 14,055 questions.

Experiments

Evaluation Metrics	Top-1 accuracy Top-3 accuracy
---------------------------	----------------------------------

Experiment Content

- Comparison with State-of-the-Art Methods
- Ablation Study
- Visualization

Experiments

Comparison with State-of-the-Art Methods

State-of-the-art comparison on FVQA dataset

Method	Overall Accuracy	
	top-1	top-3
LSTM-Question+Image+Pre-VQA [10]	24.98	40.40
Hie-Question+Image+Pre-VQA [10]	43.14	59.44
FVQA (top-3-QQmapping) [10]	56.91	64.65
FVQA (Ensemble) [10]	58.76	-
Straight to the Facts (STTF) [9]	62.20	75.60
Reading Comprehension [6]	62.96	70.08
Out of the Box (OB) [8]	69.35	80.25
Human [10]	77.99	-
Mucko	73.06 ± 0.39	85.94 ± 0.46

top-1: ↑ 3.7%

top-3: ↑ 5.7%

Experiments

Comparison with State-of-the-Art Methods

State-of-the-art comparison on Visual7w+KB dataset

Method	Overall Accuracy	
	top-1	top-3
KDMN-NoKnowledge [5]	45.1	-
KDMN-NoMemory [5]	51.9	-
KDMN [5]	57.9	-
KDMN-Ensemble [5]	60.9	-
Out of the Box (OB) ¹ [8]	57.32	71.61
Mucko (ours)	68.88 ± 0.52	85.13 ± 0.67

top-1: ↑ **11.5%**
top-3: ↑ **13.5%**

Experiments

Comparison with State-of-the-Art Methods

State-of-the-art comparison on OKVQAdataset

Method	Overall Accuracy	
	top-1	top-3
Q-Only [7]	14.93	-
MLP [7]	20.67	-
BAN [3]	25.17	-
MUTAN [1]	26.41	-
ArticleNet (AN) [7]	5.28	-
BAN + AN [7]	25.61	-
MUTAN + AN [7]	27.84	-
BAN/AN oracle [7]	27.59	-
MUTAN/AN oracle [7]	28.47	-
Mucko (ours)	29.20 ± 0.31	30.66 ± 0.55

top-1: ↑ 0.7%

Experiments

Ablation Study

Method	Overall Accuracy	
	top-1	top-3
Mucko (full model)	73.06	85.94
1 w/o Intra-Modal Knowledge Selection	70.50	81.77
2 w/o Semantic Graph	71.28	82.76
3 w/o Visual Graph	69.12	78.05
4 w/o Semantic Graph & Visual Graph	20.43	29.10
5 S-to-F Concat.	67.82	76.65
6 V-to-F Concat.	69.93	80.12
7 V-to-F Concat. & S-to-F Concat.	70.68	82.04
8 w/o relationships	72.10	83.75

Model 1: evaluate the influence of Intra-Modal Knowledge Selection.

Experiments

Ablation Study

Method		Overall Accuracy	
		top-1	top-3
Mucko (full model)		73.06	85.94
1	w/o Intra-Modal Knowledge Selection	70.50	81.77
2	w/o Semantic Graph	71.28	82.76
3	w/o Visual Graph	69.12	78.05
4	w/o Semantic Graph & Visual Graph	20.43	29.10
5	S-to-F Concat.	67.82	76.65
6	V-to-F Concat.	69.93	80.12
7	V-to-F Concat. & S-to-F Concat.	70.68	82.04
8	w/o relationships	72.10	83.75

Model 2-4: evaluate the influence of each layer of graphs.

1. Both semantic and visual graphs are beneficial to provide valuable evidence for answer inference.
2. The visual information has greater impact than the semantic part.

Experiments

Ablation Study

Method		Overall Accuracy	
		top-1	top-3
Mucko (full model)		73.06	85.94
1	w/o Intra-Modal Knowledge Selection	70.50	81.77
2	w/o Semantic Graph	71.28	82.76
3	w/o Visual Graph	69.12	78.05
4	w/o Semantic Graph & Visual Graph	20.43	29.10
5	S-to-F Concat.	67.82	76.65
6	V-to-F Concat.	69.93	80.12
7	V-to-F Concat. & S-to-F Concat.	70.68	82.04
8	w/o relationships	72.10	83.75

Model 5-7: evaluate the effectiveness of the proposed cross-modal graph convolutions.

1. Concatenating the mean pooling of all the semantic/visual node features with each entity feature
2. Proves the benefits of cross-modal convolution in gathering complementary evidence from different modalities.

Experiments

Ablation Study

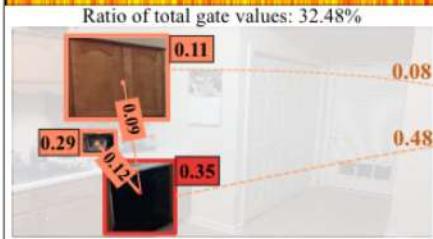
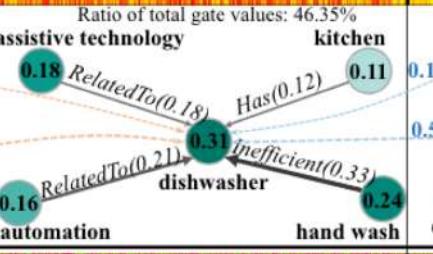
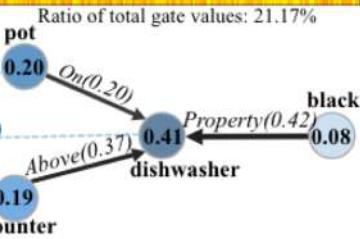
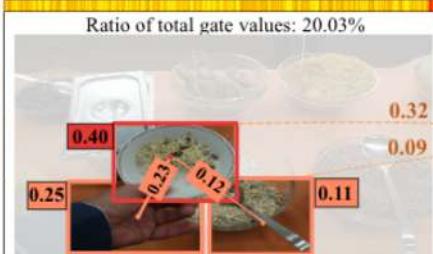
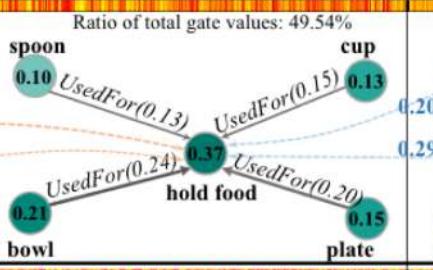
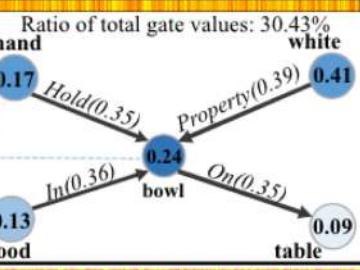
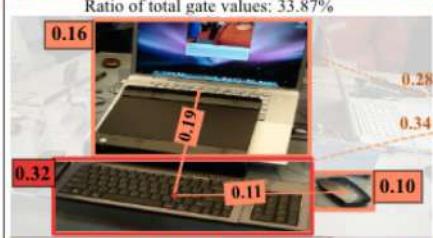
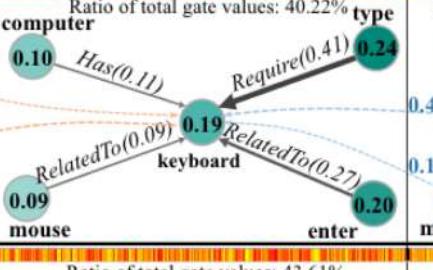
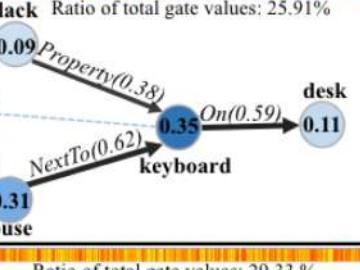
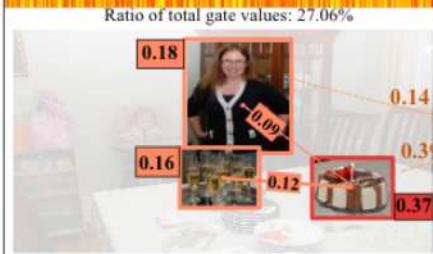
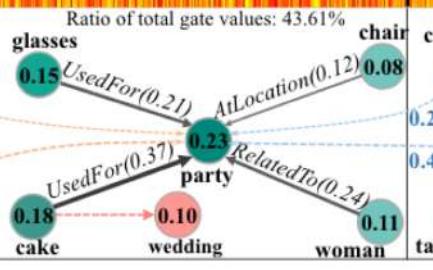
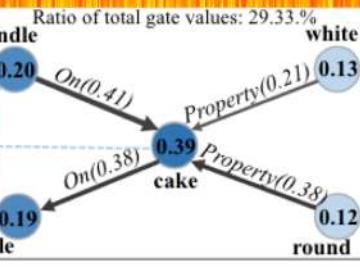
Method		Overall Accuracy	
		top-1	top-3
Mucko (full model)		73.06	85.94
1	w/o Intra-Modal Knowledge Selection	70.50	81.77
2	w/o Semantic Graph	71.28	82.76
3	w/o Visual Graph	69.12	78.05
4	w/o Semantic Graph & Visual Graph	20.43	29.10
5	S-to-F Concat.	67.82	76.65
6	V-to-F Concat.	69.93	80.12
7	V-to-F Concat. & S-to-F Concat.	70.68	82.04
8	w/o relationships	72.10	83.75

Model 8: evaluate the influence of the relationships in the heterogeneous graph.

1. It proves the benefits of relational information, though it is less influential than the modality information.

Experiments

Visualization

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

Experiments

Visualization

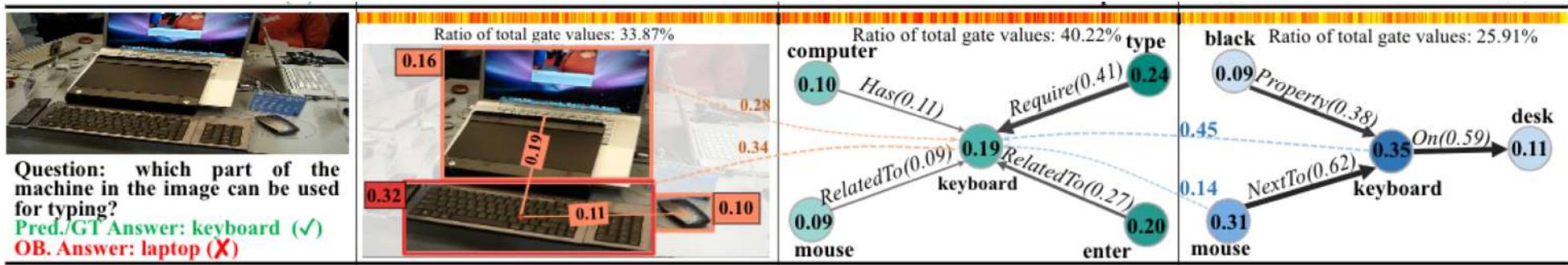
Case	Visual Graph	Fact Graph	Semantic Graph
 <p>Question: Which device in the image can free people's hand? Pred. / Gt Answer: dishwasher (✓)</p>	 Ratio of total gate values: 32.48%	<pre> graph LR AT[assistive technology] -- RelatedTo(0.18) --> K[kitchen] K -- Has(0.12) --> HW[hands wash] A[automation] -- RelatedTo(0.21) --> D[dishwasher] D -- Inefficient(0.33) --> HW </pre>	<pre> graph LR P[pot] -- On(0.20) --> D[dishwasher] D -- Property(0.42) --> B[black] C[counter] -- Above(0.37) --> D </pre>
 <p>Question: What is the white round thing held by hand in the image used for? Pred./Gt Answer: hold food (✓)</p>	 Ratio of total gate values: 20.03%	<pre> graph LR S[spoon] -- UsedFor(0.13) --> H[hand] C[cup] -- UsedFor(0.15) --> H B[bowl] -- UsedFor(0.24) --> HF[hold food] P[plate] -- UsedFor(0.20) --> HF H -- Hold(0.35) --> HF </pre>	<pre> graph LR H[hand] -- Hold(0.35) --> HF[hold food] W[white] -- Property(0.39) --> B[bowl] F[food] -- In(0.36) --> B B -- On(0.35) --> T[table] </pre>

Mucko is capable to reveal the knowledge selection mode.

- In most cases, factual knowledge provides predominant clues compared with other modalities because FVQA relies on external knowledge to answer.
- More evidence comes from the semantic modality when the question involves complex relationships.

Experiments

Visualization



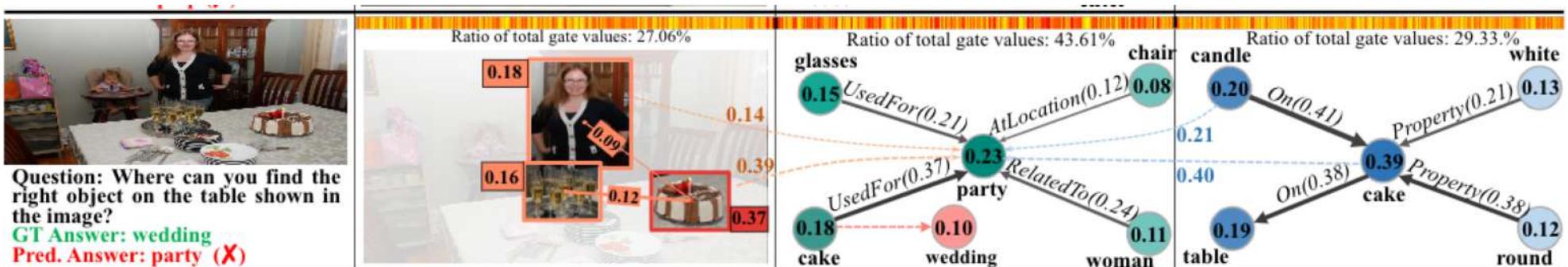
Mucko has advantages over the state-of-the-art model.

- 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.

Experiments

Visualization

Error example



Mucko fails when multiple answers are reasonable for the same question.

- Since both “wedding” and “party” may have cakes, the predicted answer “party” in the last example is reasonable from human judgement.

Conclusion

- 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.

报告提纲

- 1 / 多模态机器学习概述
- 2 / 视觉问答技术
- 3 / 视觉对话技术
- 4 / 总结与展望

3.1 什么是视觉对话?



Is there smoke in any room around you?



Yes, in one room

Go there and look for people



...



Did anyone enter the room last week?



Yes, 127 instances logged on camera



Were any of them carrying a black bag?



...

▲ 采用自然语言对机器人指挥控制（现场救援指挥）

▲ 时空不对等用户辅助智能分析（安防领域）



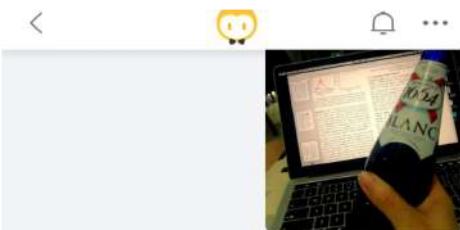
Peter just uploaded a picture from his vacation in Hawaii



Great! Is he at the beach?



No, on a mountain

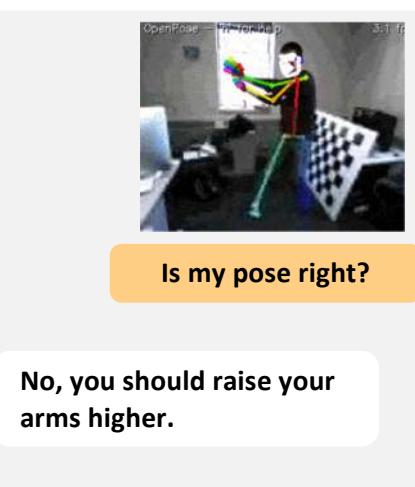


亲，小蜜还在学习图片的意思，刚学会用图片淘宝的技能，你是要我帮你找图片上类似的宝贝吗？

图片上的酒有卖吗？



Is my pose right?



No, you should raise your arms higher.

▲ 视觉信息不对等的智能聊天（小爱同学，天猫精灵）

▲ 基于多模态信息的智能客服
(阿里小蜜，度蜜，智能健身教练)

3.1 什么是视觉对话？

➤ 视觉对话任务

输入

- 图像 I
- 图像描述 C 和 $t-1$ 轮的对话历史
$$H = \{C, (Q_1, A_1), \dots, (Q_{t-1}, A_{t-1})\}$$
- 当前轮问题 Q_t

输出

- 视觉对话任务要求从100个候选答案集合 $A = \{A_1, A_2, \dots, A_{100}\}$ 中选择最佳答案



VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Captioning

Two people are in a wheelchair and one is holding a racket.

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman

3.2 视觉对话的挑战

Challenge 1: Multiple Answers



Q: Do you see any birds?

Valid Answers

No No birds

I do not see any birds

Nope Not at all

No, Not that I can see

Challenge 2: Visual Reference



Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman

3.2 视觉对话的挑战

Challenge 3: Visual Content Understanding



C: A man doing a grind on a skateboard.

Q1: Is the ~~man~~ on the skateboard?

A1: Yes, he is.

3

Q4: Is he younger or older?

A4: He is in the middle-aged.

Q5: Is there **sky** in the picture?

A5: Yes, the ~~sky~~ is deep blue.

Challenge 4: Generative Method



C: A large bus is tipping over on the street near buildings.

Q1: Is this a yellow school bus?

A1: No, it is a city bus.

Q2: See any stop signs?

A2: No, there are no signs at all.

Q3: Any people?

A3: Yes, there are in the people are in the people are in the people are in the people are.

3.3 视觉对话的模型

- 针对挑战三：如何使模型关注对话中不断变化的图像内容？

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

AAAI 2020

Xiaoze Jiang^{1,2}, Jing Yu^{1*}, Zengchang Qin², Yingying Zhuang², Xingxing Zhang³, Yue Hu¹ and Qi Wu⁴

1



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

2



北京航空航天大学
BEIHANG UNIVERSITY

3

Microsoft
Research
微软亚洲研究院

4



THE UNIVERSITY
of ADELAIDE

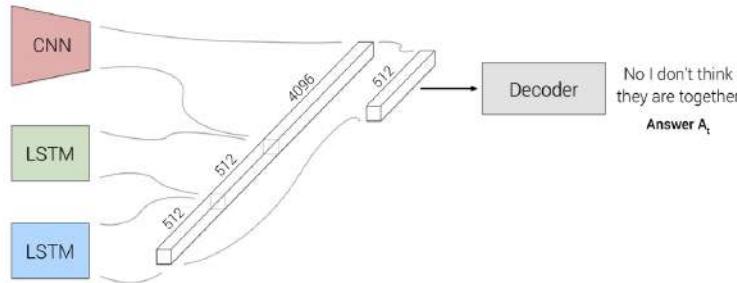
Motivation



Do you think the woman is with him?
Question Q_i

The man is riding his bicycle on the sidewalk. Is the man wearing a helmet? No he does not have a helmet on... Are there any people nearby? Yes there's a woman walking behind him.

t rounds of history
(concatenated)



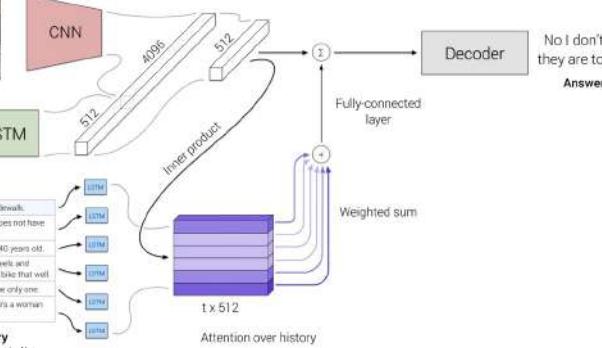
No I don't think they are together
Answer A_i



Do you think the woman is with him?
Question Q_i

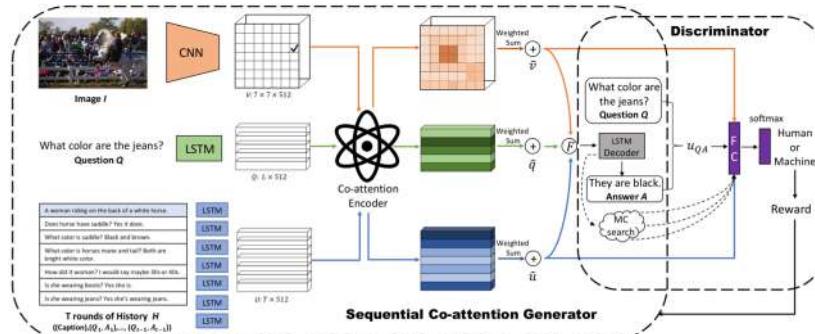
The man is riding his bicycle on the sidewalk. Is the man wearing a helmet? No he does not have a helmet on... How old is the man? He looks around 40 years old... What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well... Is anyone else riding a bike? No he's the only one... Are there any people nearby? Yes there's a woman walking behind him.

((Caption), (Q₁, A₁), ..., (Q_t, A_t))



No I don't think they are together
Answer A_i

▲ Late Fusion [1] (LF)



The role of visual information has been less studied !

▲ Co-Attention[2] (CoAtt)

[1] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, pages 1080–1089, 2017.

[2] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. Are you talking tome? reasoned visual dialog generation through adversarial learning. In *CVPR*, pages 6106–6115, 2018.

Motivation



Image

C: A man doing a grind on a skateboard.

Q1: Is the man on the skateboard?

A1: Yes, he is.

...

Q4: Is he younger or older?

A4: He is in the middle-aged.

Q5: Is there sky in the picture?

A5: Yes, the sky is deep blue with some clouds.

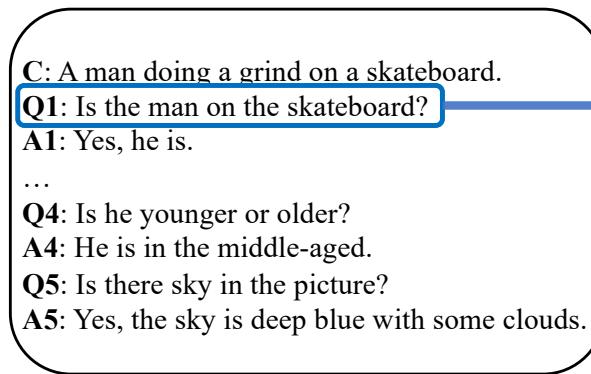
History

- Visual Dialogue task demands the agent to adaptively focus on diverse visual content with respect to the current question.
- The **key challenge** in Visual Dialogue task is thus to learn a more comprehensive and semantic-rich image representation, which may have adaptive attentions on the image for variant questions.

Motivation



Image



History



the man



skateboard

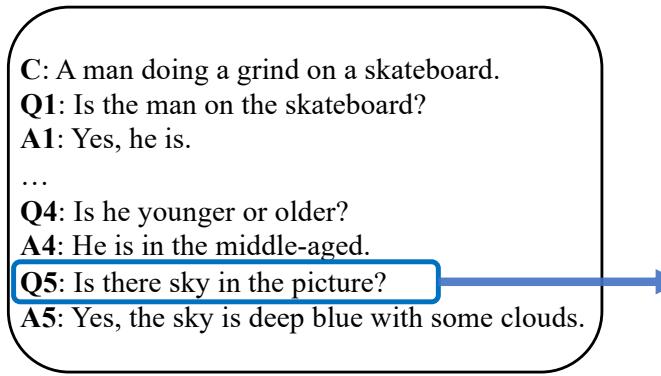
Prospect

- Visual Dialogue task demands the agent to adaptively focus on diverse visual content with respect to the current question.
- The **key challenge** in Visual Dialogue task is thus to learn a more comprehensive and semantic-rich image representation, which may have adaptive attentions on the image for variant questions.

Motivation



Image



History



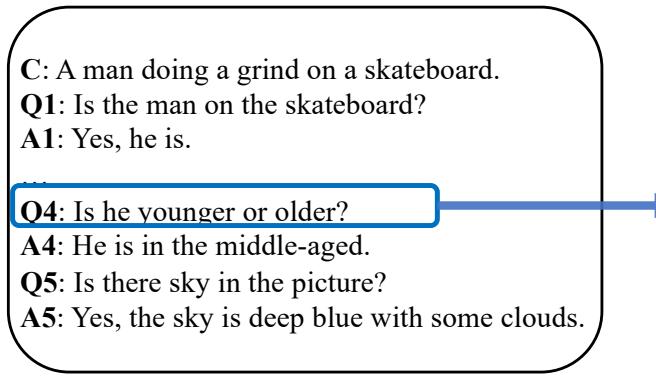
Background
sky

- Visual Dialogue task demands the agent to adaptively focus on diverse visual content with respect to the current question.
- The **key challenge** in Visual Dialogue task is thus to learn a more comprehensive and semantic-rich image representation, which may have adaptive attentions on the image for variant questions.

Motivation



Image



History

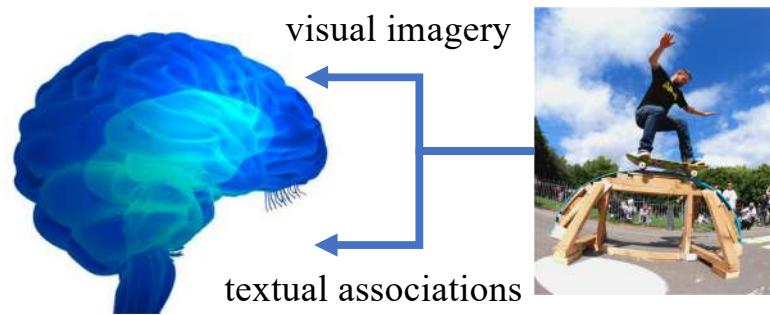


the middle-aged man

Higher-level semantics

- Visual Dialogue task demands the agent to adaptively focus on diverse visual content with respect to the current question.
- The **key challenge** in Visual Dialogue task is thus to learn a more comprehensive and semantic-rich image representation, which may have adaptive attentions on the image for variant questions.

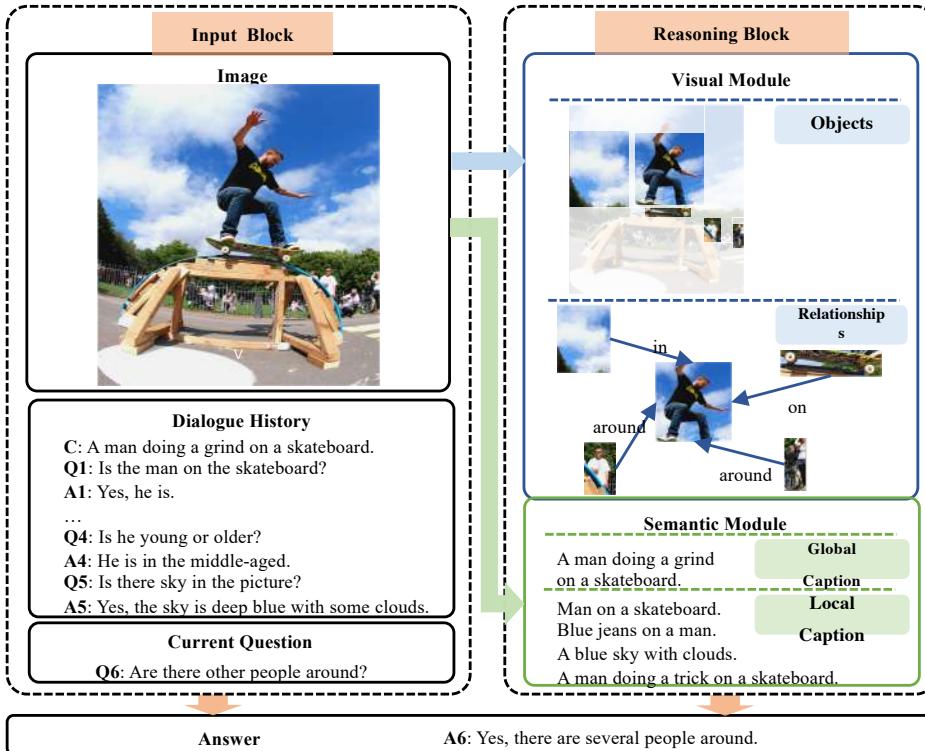
Motivation



- Dual-coding theory [1]:
Our brain encodes information in two ways: visual imagery and textual associations.
- When asked to act upon a concept, our brain re-trives either images or words, or both simultaneously.
- The ability to encode a concept by two different ways strengthens the capacity of memory and understanding.

[1] A. Paivio, “*Imagery and Verbal Process.*” New York: Holt, Rinehart and Winston., 1971.

Motivation



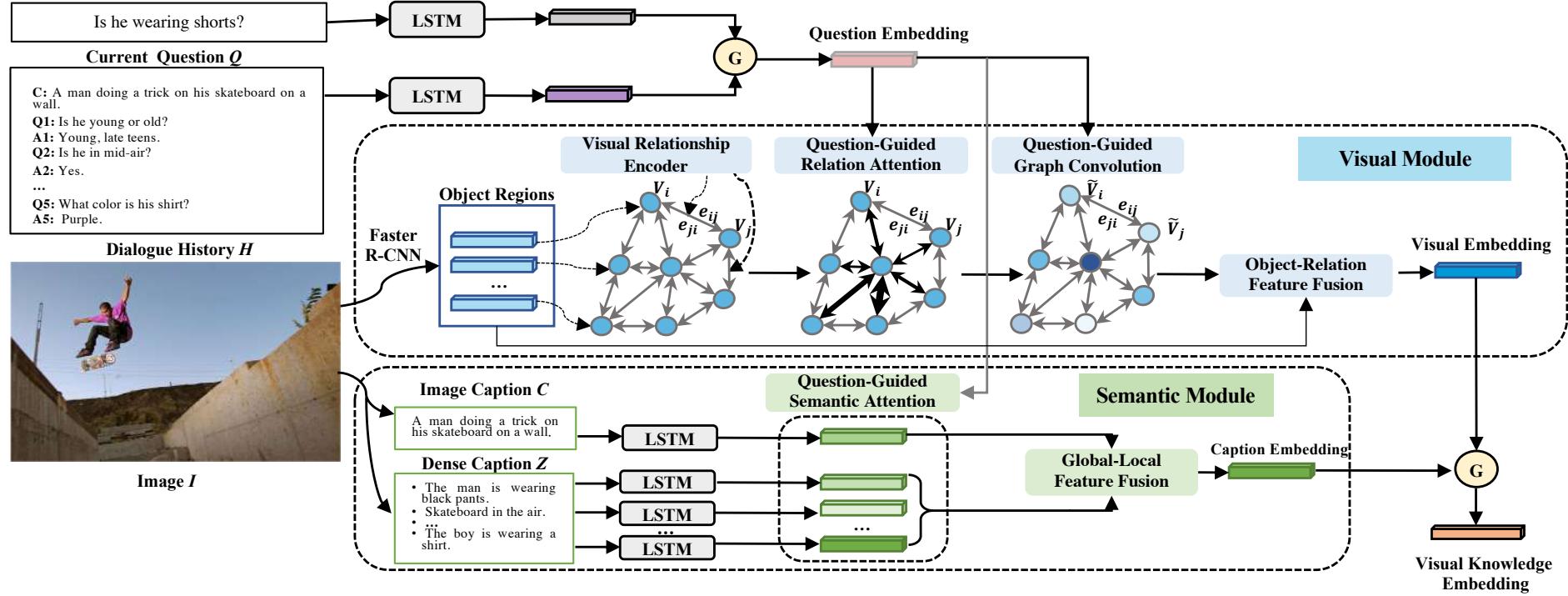
➤ Inspired by the cognitive process, we first propose a novel framework to comprehensively depict an image from both visual and semantic perspectives.

➤ Based on the dual encoding framework, we propose a new method to adaptively select question-relevant information from the image in a hierarchical mode:

(1) **intra-model selection**: captures the visual and semantic information individually from the object-relational visual features and global-local semantic features

(2) **inter-modal selection**: obtains the joint visual-semantic knowledge by correlating vision and semantics.

Model

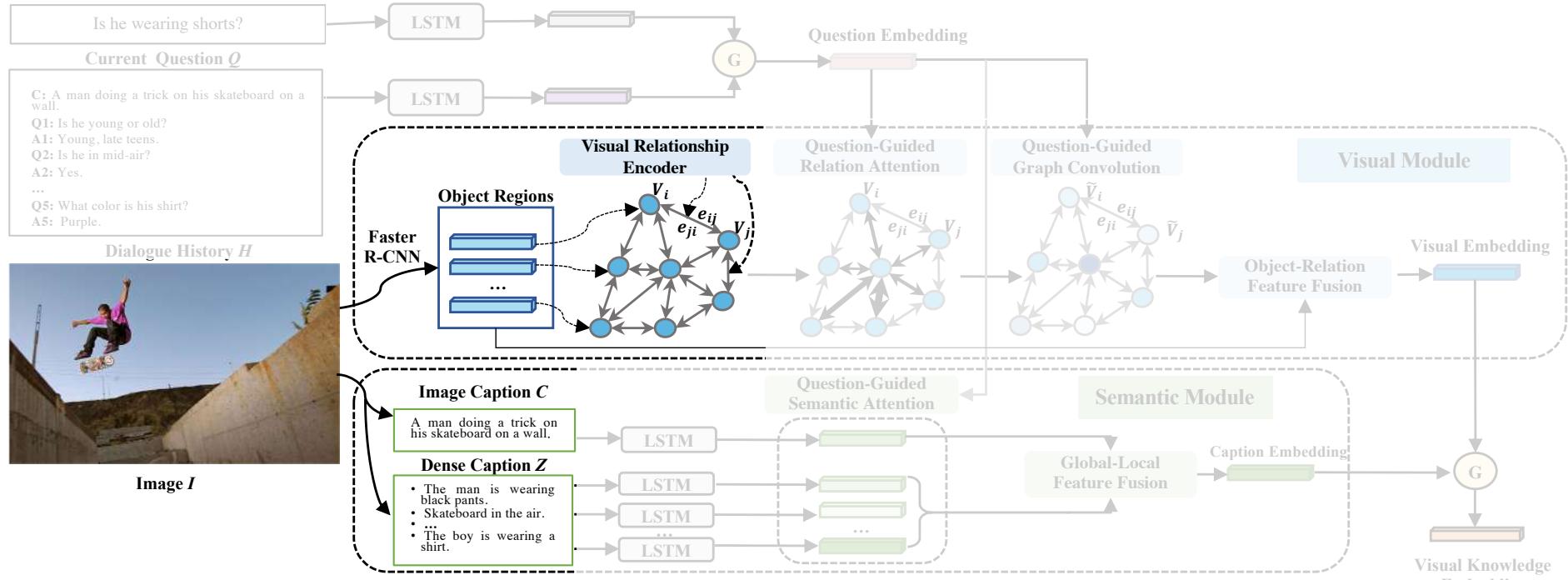


➤ The core structure of the model is divided into two parts:

Visual-Semantic Dual Encoding

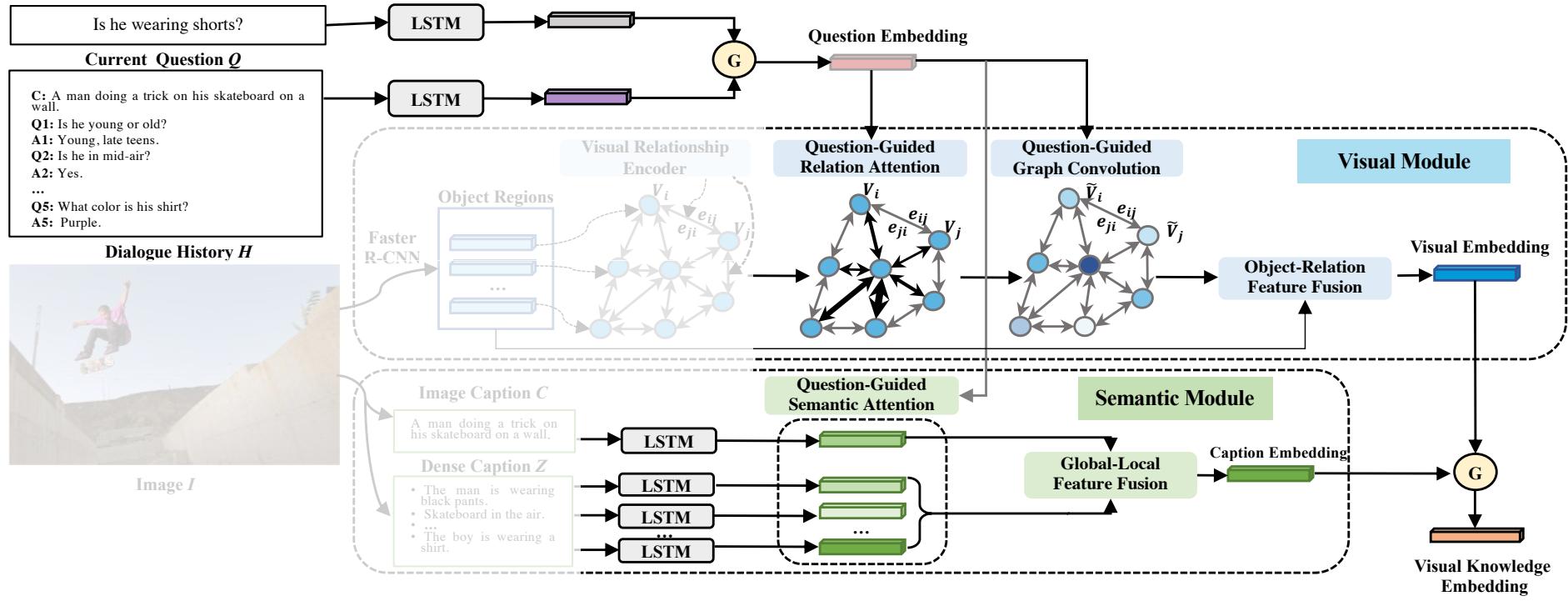
Adaptive Visual-Semantic Knowledge Selection

Model



➤ **Visual-Semantic Dual Encoding**

Model

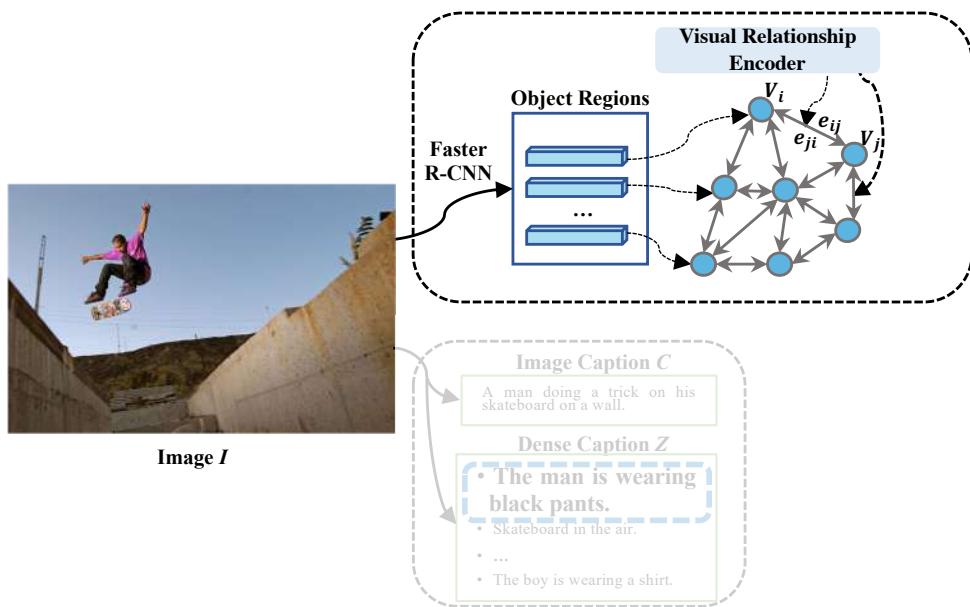


➤ Adaptive Visual-Semantic Knowledge Selection

Model

● Visual-Semantic Dual Encoding **Visual Encoding** and **Semantic Encoding**

- Propose a novel framework to comprehensively depict an image from both visual and semantic perspectives.



• Scene Graph Construction

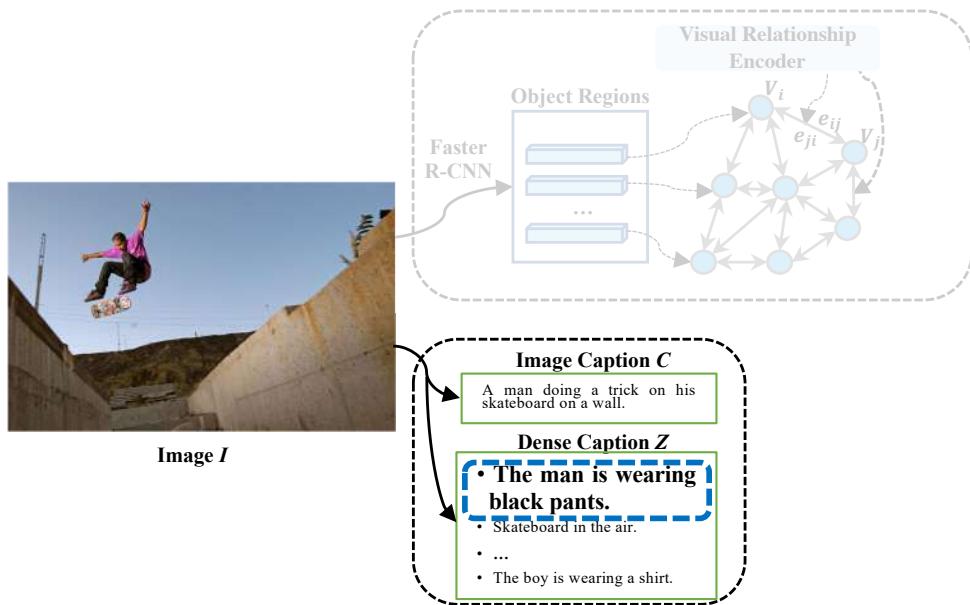
- ✓ **Nodes:** Use a pre-trained Faster-RCNN to detect N objects in an image and then describe each node as a 2048-dimentional vector.
- ✓ **Edges:** Use a pre-trained visual relationship encoder^[1] to encode relationships between the subject and object.

[1] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny “Large-scale visual relationship understanding,” in *AAAI*, 2019.

Model

● Visual-Semantic Dual Encoding **Visual Encoding** and **Semantic Encoding**

- Propose a novel framework to comprehensively depict an image from both visual and semantic perspectives.



• Multi-level Image Captions

Each image is represented as a hierarchical semantic description:

- ✓ **Global image caption:** the captions provided by the dataset.
- ✓ **Local image caption:** the top k dense captions extracted by the **DenseCap**^[1].

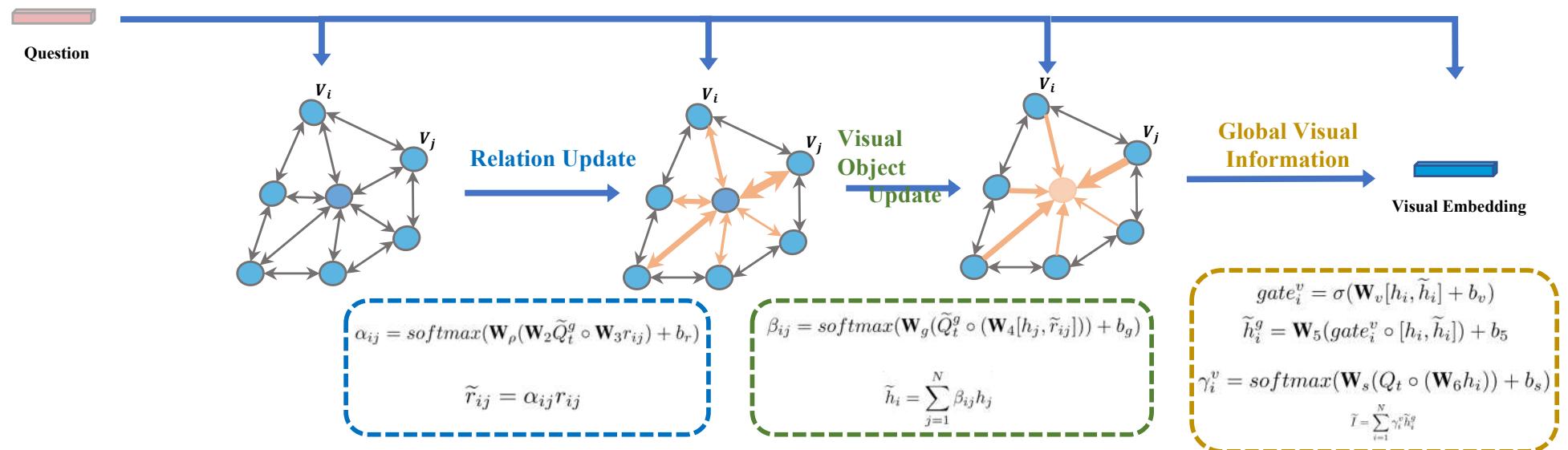
[1] Johnson, J.; Karpathy, A.; and Fei-Fei, L. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 4565–4574.

Model

- Visual-Semantic Knowledge Selection

- (1) Intra-modal selection: Visual selection and Semantic selection
- (2) Inter-modal selection

Visual Module intra-selection

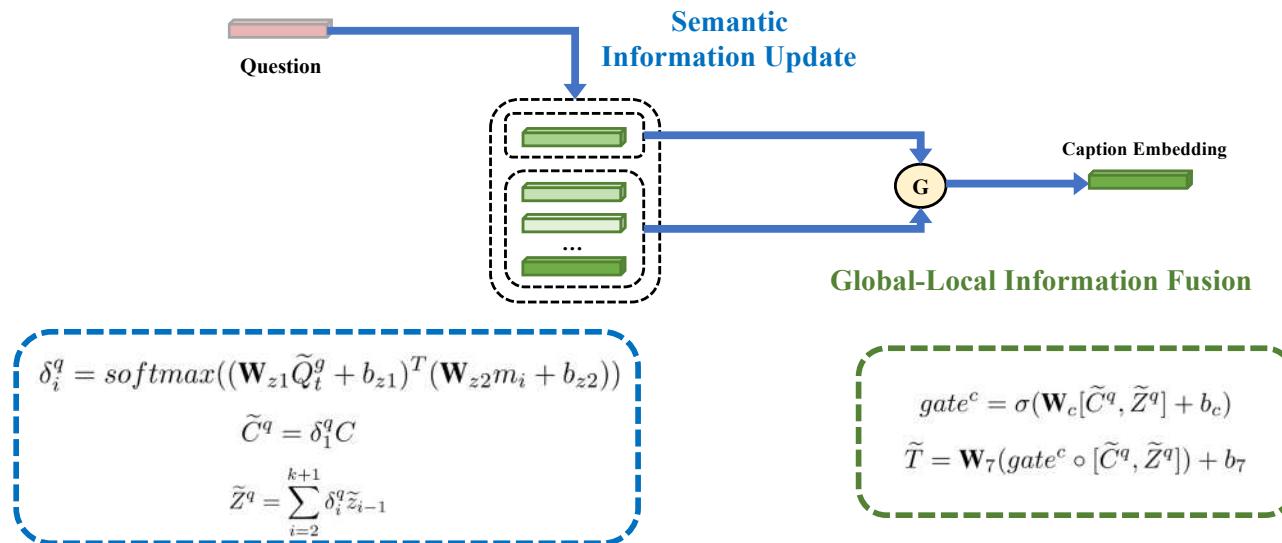


Model

- Visual-Semantic Knowledge Selection

- (1) Intra-modal selection: Visual selection and Semantic selection
- (2) Inter-modal selection

Visual Module intra-selection



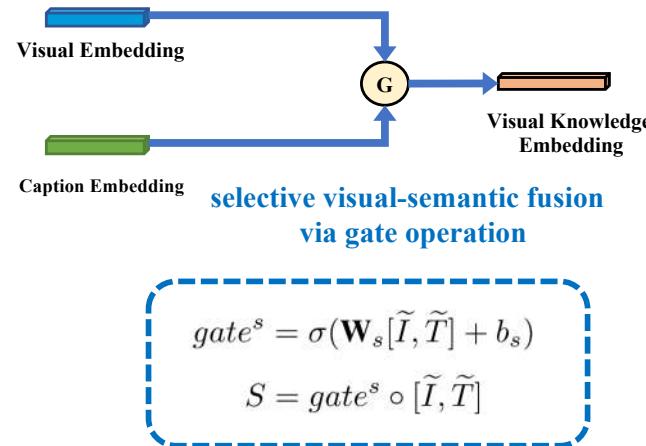
Model

- Visual-Semantic Knowledge Selection

(1) Intra-modal selection: Visual selection and Semantic selection

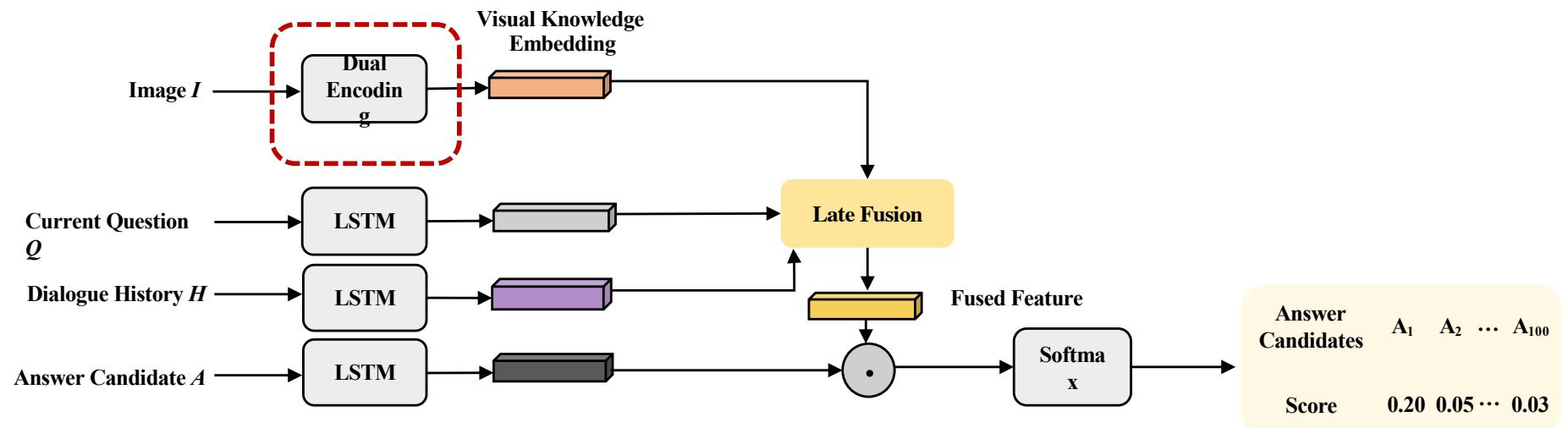
(2) Inter-modal selection

Inter-modal selection



Model

● Late Fusion Encoder



- Our model has complementary advantages with existing research work on dialogue history, it can also be applied to more complex decoders and fusion strategies.
- We utilize the simple late fusion and discriminative decoder to highlight the advantages of our visual encoder.

Results

➤ Dataset

VisDial v0.9 : built on MSCOCO images,
divided into *train*, *test* and *val* set.

VisDial v1.0 : all the splits of VisDial v0.9 serve as the *train* set (120k) ,
test set (8k) and *val* set (2k) consist of dialogues on extra 10k COCO-like
images from Flickr.

➤ Evaluation Metrics

VisDial v0.9 : utilize retrieval metrics: MRR, R@ k ($k = 1, 5, 10$) and Mean

VisDial v1.0 : NDCG is added as an extra metric for more comprehensive analysis.

Lower value for Mean and higher value for other metrics are desired.

➤ Experiments

- (1) Overall Results on VisDial v0.9 and VisDial v1.0
- (2) Ablation study
- (3) Visualization

Results

➤ Compare with State-of-the-art

Table 1: Comparison on validation split of VisDial v0.9.

Model	MRR	R@1	R@5	R@10	Mean
LF (Das et al. 2017)	58.07	43.82	74.68	84.07	5.78
HRE (Das et al. 2017)	58.46	44.67	74.50	84.22	5.72
MN (Das et al. 2017)	59.65	45.55	76.22	85.37	5.46
SAN-QI (Yang et al. 2016)	57.64	43.44	74.26	83.72	5.88
HieCoAtt-QI (Lu et al. 2016)	57.88	43.51	74.49	83.96	5.84
AMEM (Seo et al. 2017)	61.60	47.74	78.04	86.84	4.99
HCIAE (Lu et al. 2017)	62.22	48.48	78.75	87.59	4.81
SF (Jain, Lazebnik, and Schwing 2018)	62.42	48.55	78.96	87.75	4.70
CoAtt (Qi et al. 2018)	63.98	50.29	80.71	88.81	4.47
CorefMN (Kottur et al. 2018)	64.10	50.92	80.18	88.81	4.45
VGNN (Zheng et al. 2019)	62.85	48.95	79.65	88.36	4.57
DualVD	62.94	48.64	80.89	89.94	4.17

Table 2: Comparison on test-standard split of VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF (Das et al. 2017)	55.42	40.95	72.45	82.83	5.95	45.31
HRE (Das et al. 2017)	54.16	39.93	70.47	81.50	6.41	45.46
MN (Das et al. 2017)	55.49	40.98	72.30	83.30	5.92	47.50
LF-Att (Das et al. 2017)	57.07	42.08	74.82	85.05	5.41	40.76
MN-Att (Das et al. 2017)	56.90	42.43	74.00	84.35	5.59	49.58
CorefMN (Kottur et al. 2018)	61.50	47.55	78.10	88.80	4.40	54.70
VGNN (Zheng et al. 2019)	61.37	47.33	77.98	87.83	4.57	52.82
RvA (Niu et al. 2019)	63.03	49.03	80.40	89.83	4.18	55.59
DL-61 (Guo, Xu, and Tao 2019)	62.20	47.90	80.43	89.95	4.17	57.32
DualVD	63.23	49.25	80.23	89.70	4.11	56.32

- ✓ Our model consistently outperforms all the approaches on most metrics and slightly underperforms than the model using multi-step reasoning and complex attention mechanism.

Results

➤ Ablation Study

Table 3: Ablation study of DualVD on 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
DualVD	64.64	50.74	82.10	91.00	3.91	57.30

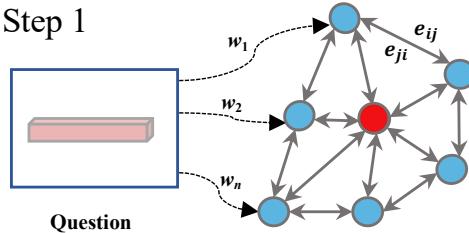
✓ Each component is effective.

Results

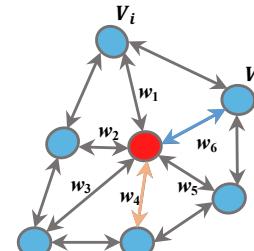
- A critical advantage of DualVD lies in its interpretability: DualVD is capable to predict the attention weights in the visual module, semantic module and the gate values in visual-semantic fusion.

- Visualization of visual objects

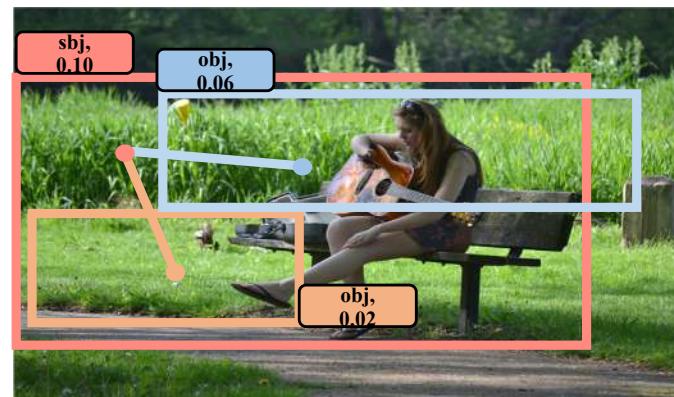
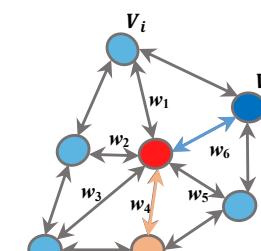
Step 1



Step 2



Step 3

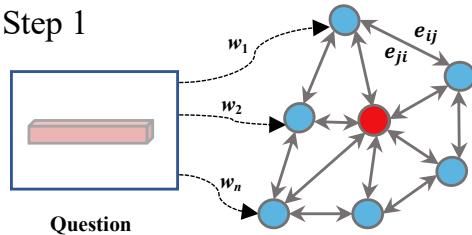


Results

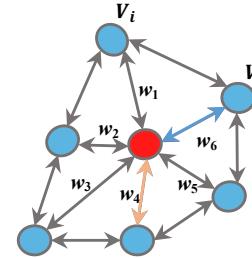
- A critical advantage of DualVD lies in its interpretability: DualVD is capable to predict the attention weights in the visual module, semantic module and the gate values in visual-semantic fusion.

- Visualization of visual objects

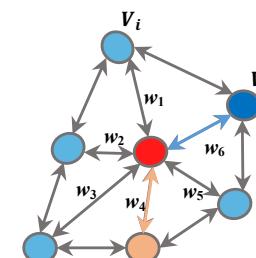
Step 1



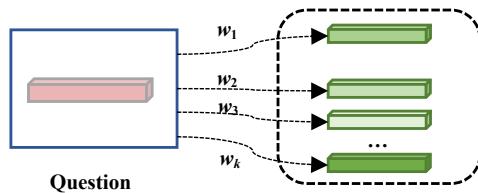
Step 2



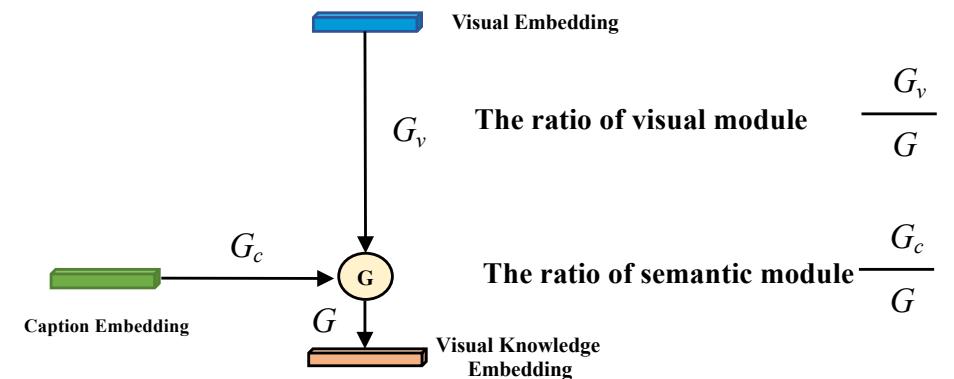
Step 3



- Visualization of caption



- Visualization of Gate Value



Results

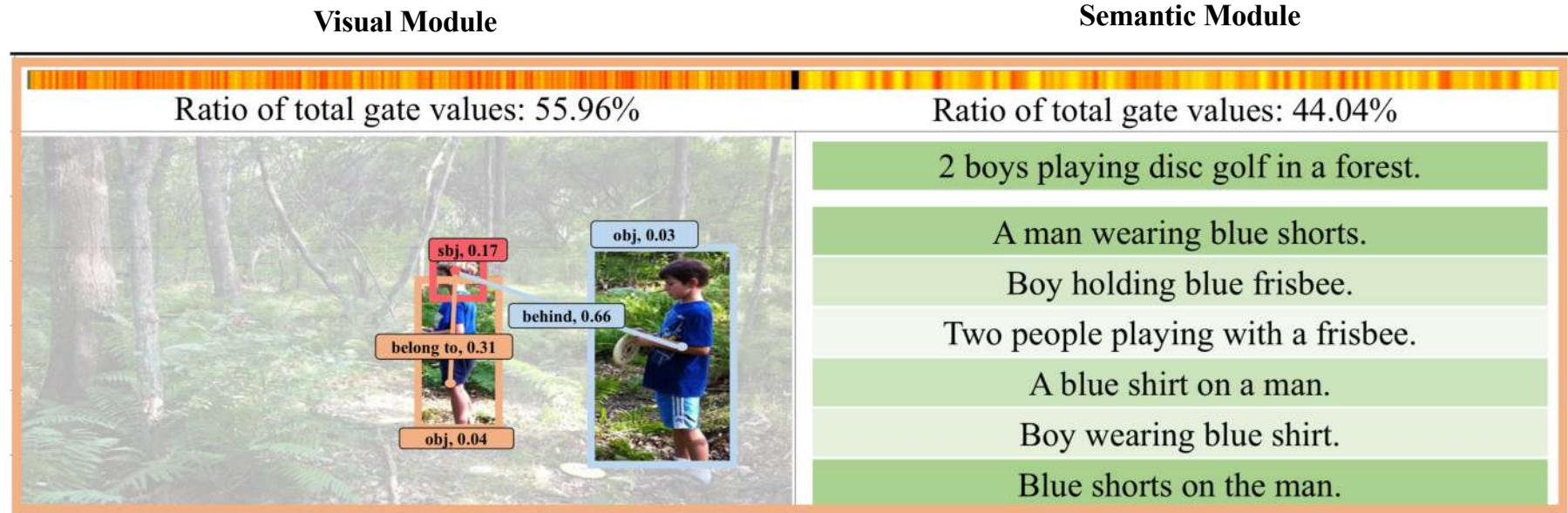
- Case

Image	Dialogue History
 A photograph showing two young boys in a dense forest. They are both wearing blue shirts and shorts, and they are holding yellow frisbees. One boy is facing the other, while the other is looking down at something on the ground.	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 Dose 1 of the boys holding the disc? Answer3 They are both holding discs.

Results

- Case

Question1	Are the boys teenagers?
Answer1	They are young boys.

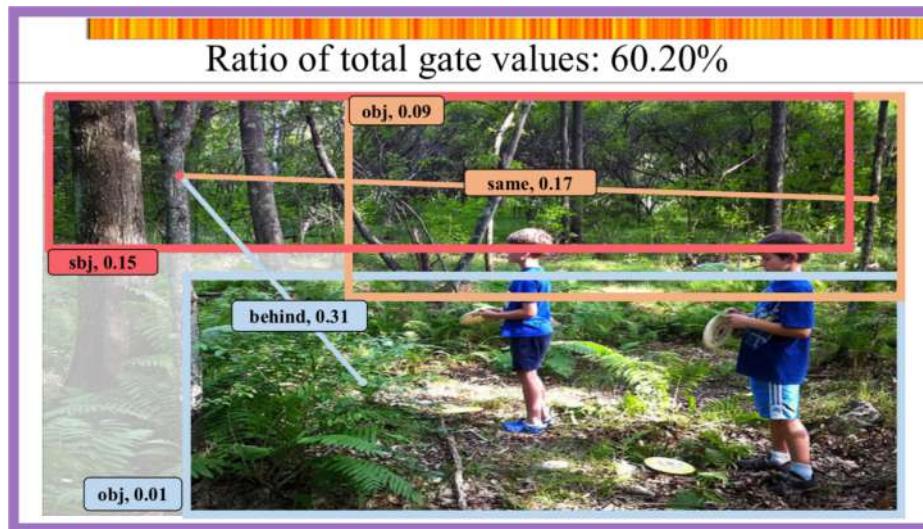


Results

- Case

Question2	Do you see a lot of trees?
Answer2	Yes, a ton of trees.

Visual Module



Semantic Module

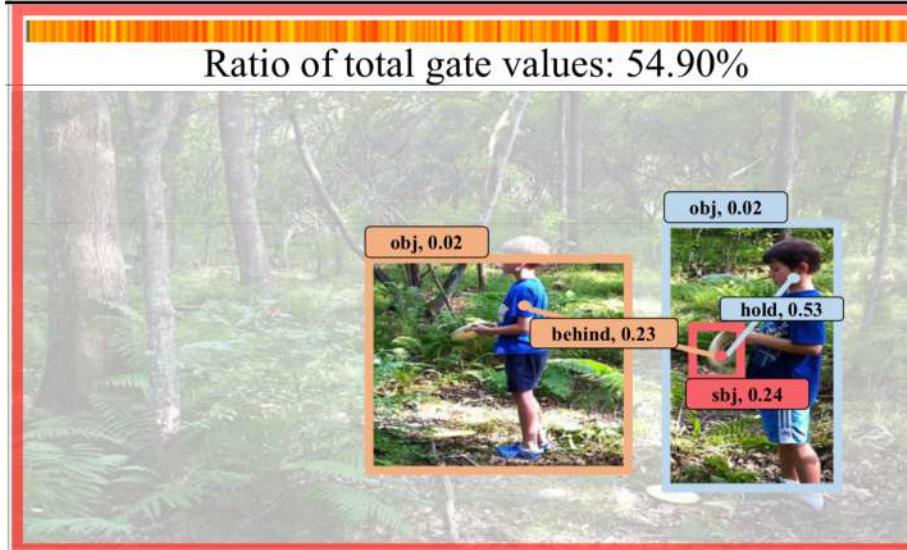


Results

- Case

Question3	Dose 1 of the boys holding the disc?
Answer3	They are both holding discs.

Visual Module



Semantic Module

Ratio of total gate values: 45.10%
2 boys playing disc golf in a forest.
A man wearing blue shorts.
Boy holding blue frisbee.
Two people playing with a frisbee.
A blue shirt on a man.
Boy wearing blue shirt.
Blue shorts on the man.

Results

- Observation
 - The amount of information derived from each module highly depends on the complexity of the question and the relevance of the content.
 - The visual information is more important than semantic information to image understanding in visual dialogue.
 - DualVD is capable to capture the most relevant visual and semantic information regarding the current question.

Challenges

Challenge 3: Visual Content Understanding



C: A man doing a grind on a skateboard.

Q1: Is the man on the skateboard?

A1: Yes, he is.

• • •

Q4: Is he younger or older?

A4: He is in the middle-aged.

Q5: Is there ~~sky~~ in the picture?

A5: Yes, the sky is deep blue.

Challenge 4: Generative Method



C: A large bus is tipping over on the street near buildings.

Q1: Is this a yellow school bus?

A1: No, it is a city bus.

Q2: See any stop signs?

A2: No, there are no signs at all.

Q3: Any people?

A3: Yes, there are in the people are in the people are
in the people are in the people are.

- 针对挑战四：如何生成语义丰富且重复度低的回复？

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

IJCAI 2020

Xiaoze Jiang^{1,2}, Jing Yu^{1,3*}, Yajing Sun^{1,3}, Zengchang Qin², Zihao Zhu^{1,3}, Yue Hu^{1,3} and Qi Wu⁴

1



2



3



4



5



Motivation



Discriminative Method

Answer candidate set

- A1: I cannot see.
- A2: It is blue.
- A3: No, it is a city bus.
- A4: It is a sunny day!
- ...

Select an answer

A3

limited ability in real world



Generative Method

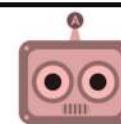
Vocabulary

I, you, hit, blue, sunny, red, better, mouse, house, phone, job, bus, an, the, old, see, is, no, any, at, people, signs, it, all, a, trees, attitude, ...

Generate an answer

No, it is a city bus on the street.

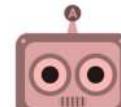
crucial to achieve human-like conversation



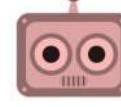
I am your visual dialogue chat bot: vv, nice to meet you! Please enter an image.



Ok ! A large bus is tipping over on the street near buildings.



Is this a yellow school bus?



No, it is a city bus.

See any stop signs?



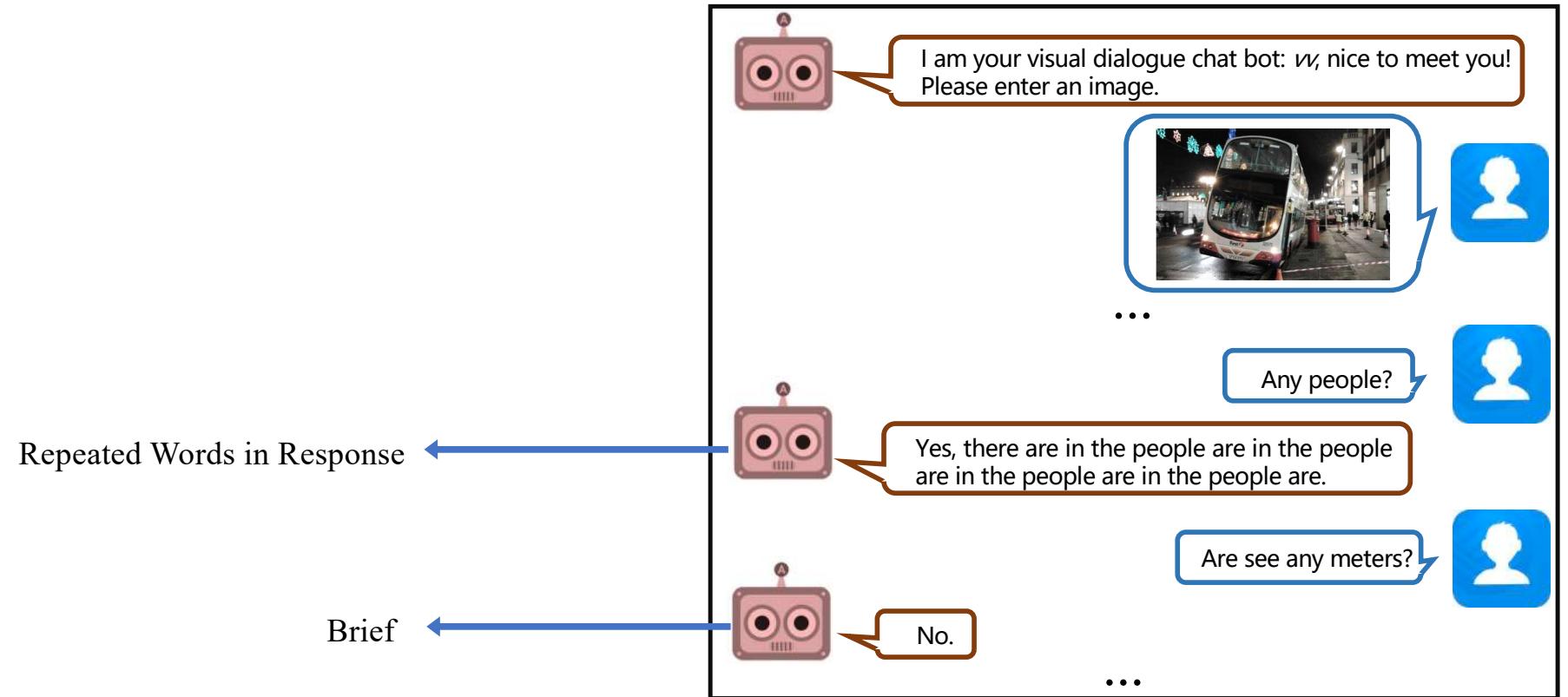
No, there are no signs at all.

...

Motivation



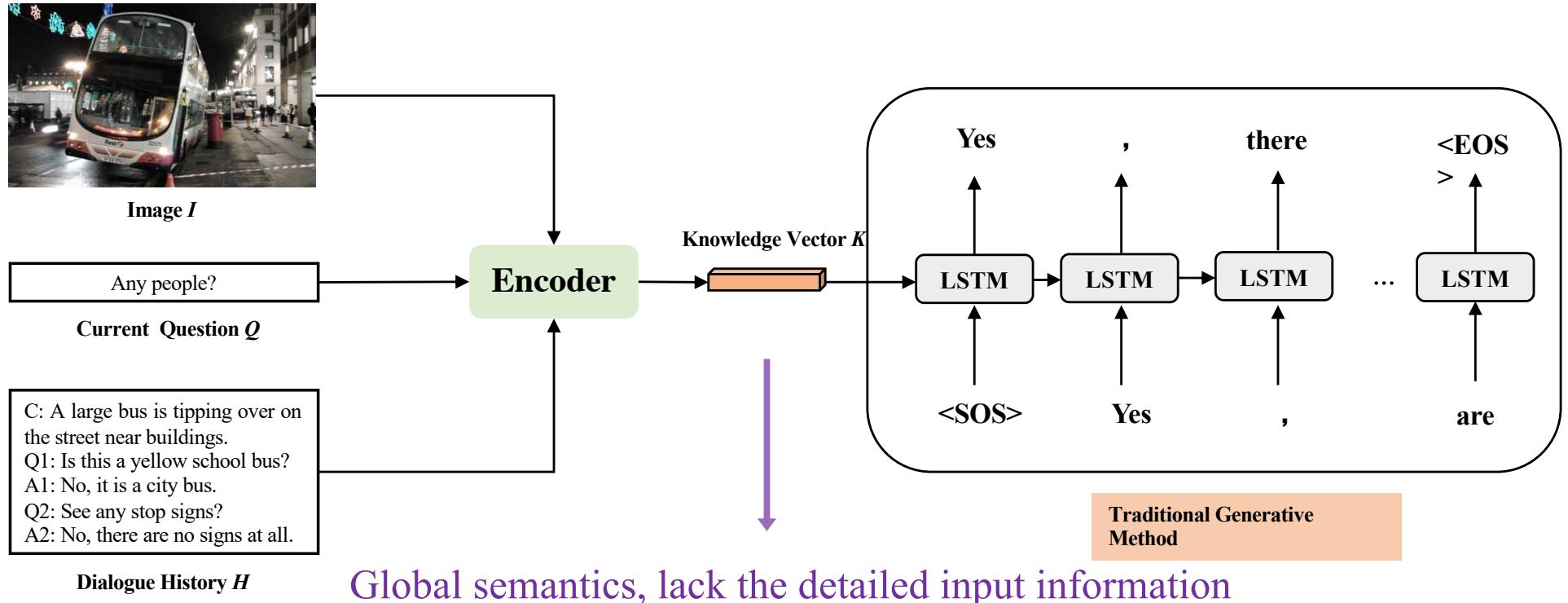
Traditional Generative Method



Motivation



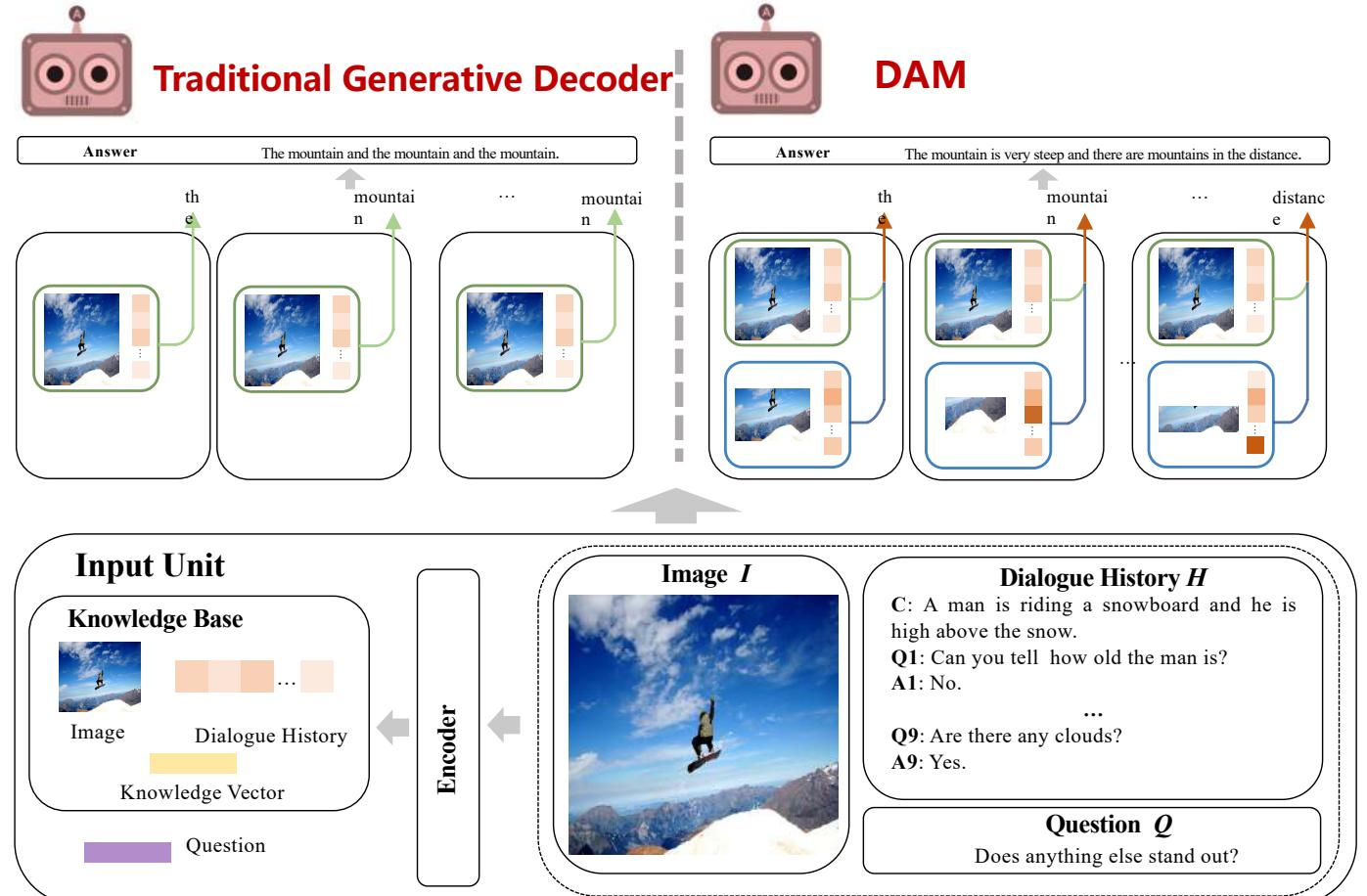
Traditional Generative Method



Motivation

- Each DAM module performs an adaptive combination of the response-level semantics captured from the encoder and the word-level semantics specifically selected for generating each word.

Green color: response-level semantics
Blue color: word-level semantics

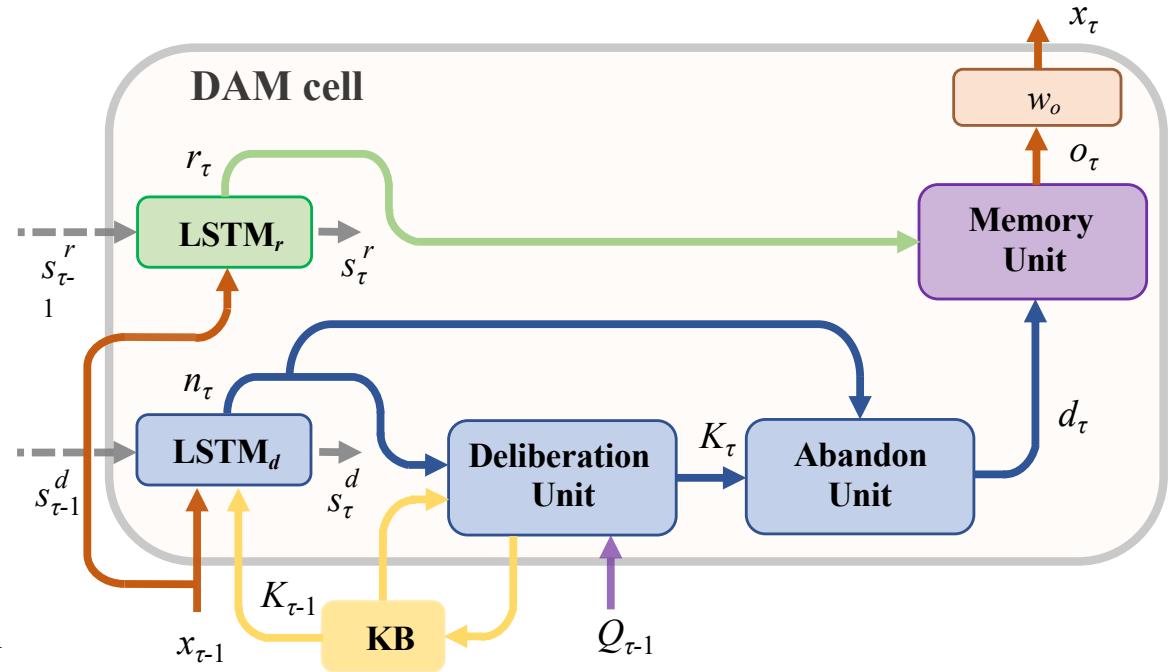


Model



DAM

- Two Level Semantic Decode Layer
 - ✓ Response-level semantic decode layer
 - ✓ Word-level detail decode layer
- Modular Architecture
 - ✓ Deliberation Unit
 - ✓ Abandon Unit
 - ✓ Memory Unit
- Universal Architecture
 - ✓ DAM can be combined with existing visual dialogue models by adapting the Deliberation Unit to the corresponding encoder.



Model

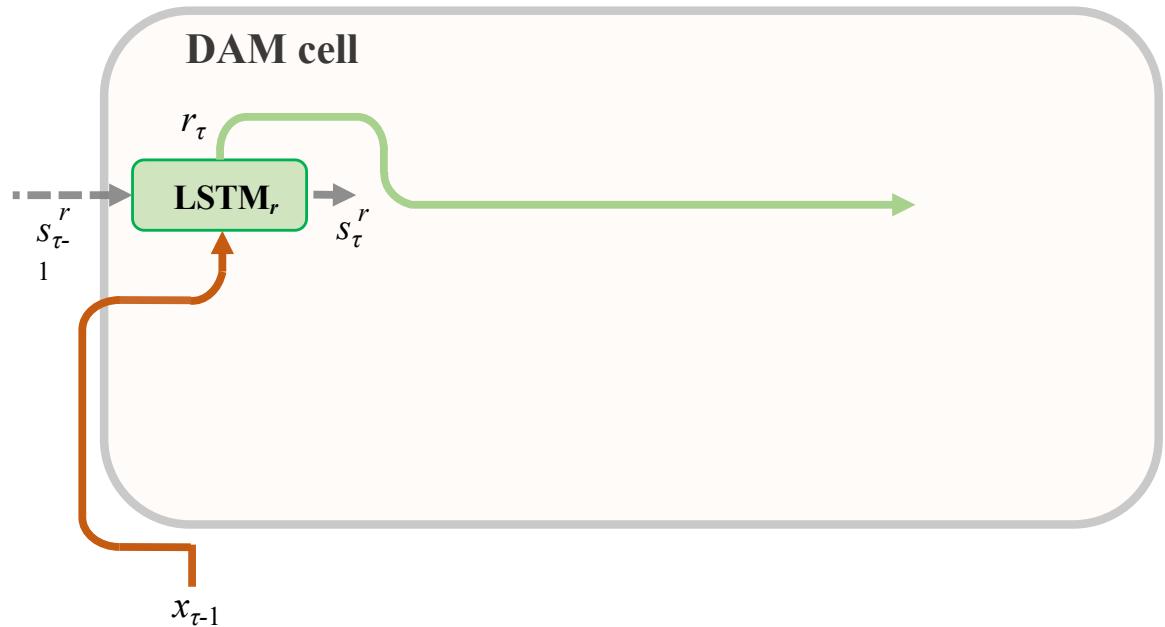


Response-level Semantic Decode Layer (RSL)

- ✓ RSL is responsible for capturing the global information to guarantee the response's fluency and correctness.

$$r_\tau = LSTM_r(x_{\tau-1}, s_{\tau-1}^r)$$

where $x_{\tau-1}$ is the previous generated word,
 s_{τ}^r is the memory state of $LSTM_r$.



Model

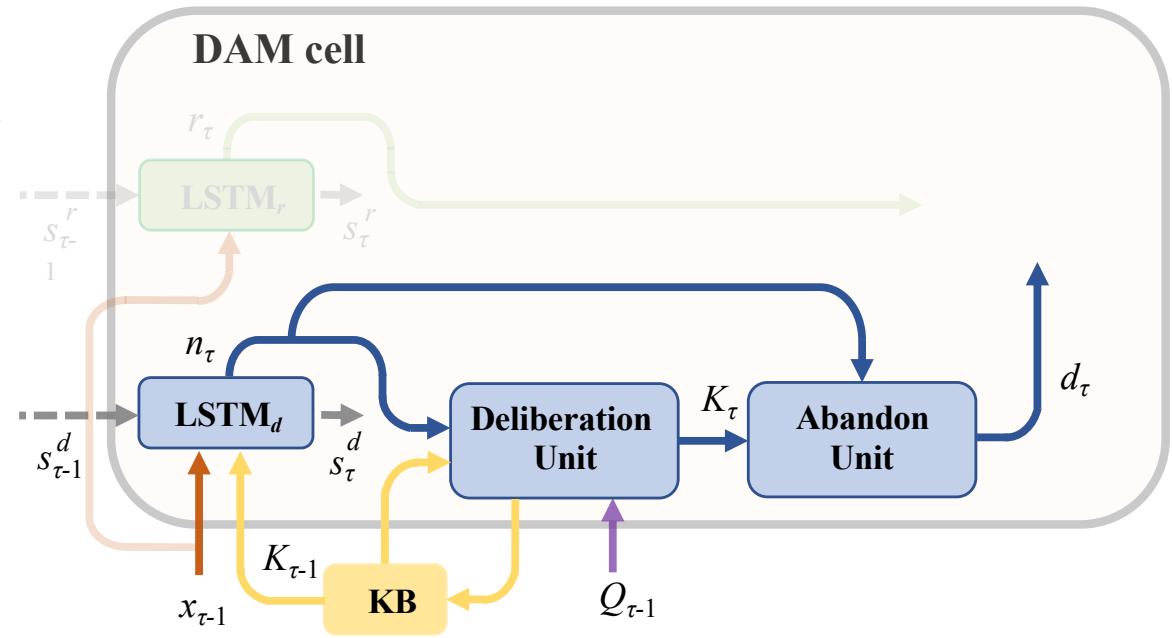


Word-level Detail Decode Layer (WDL)

- ✓ WDL incorporates the essential and unique visual dialogue contents (i.e. question, dialogue history and image) into the generation of current word to enrich the word-level details.

$$n_\tau = LSTM_d([x_{\tau-1}, K_{\tau-1}], s_{\tau-1}^d)$$

where $K_{\tau-1}$ is the updated knowledge vector in the $\tau - 1$ step,
 s_τ^d is the memory state of $LSTM_d$,
[;] denotes concatenation.

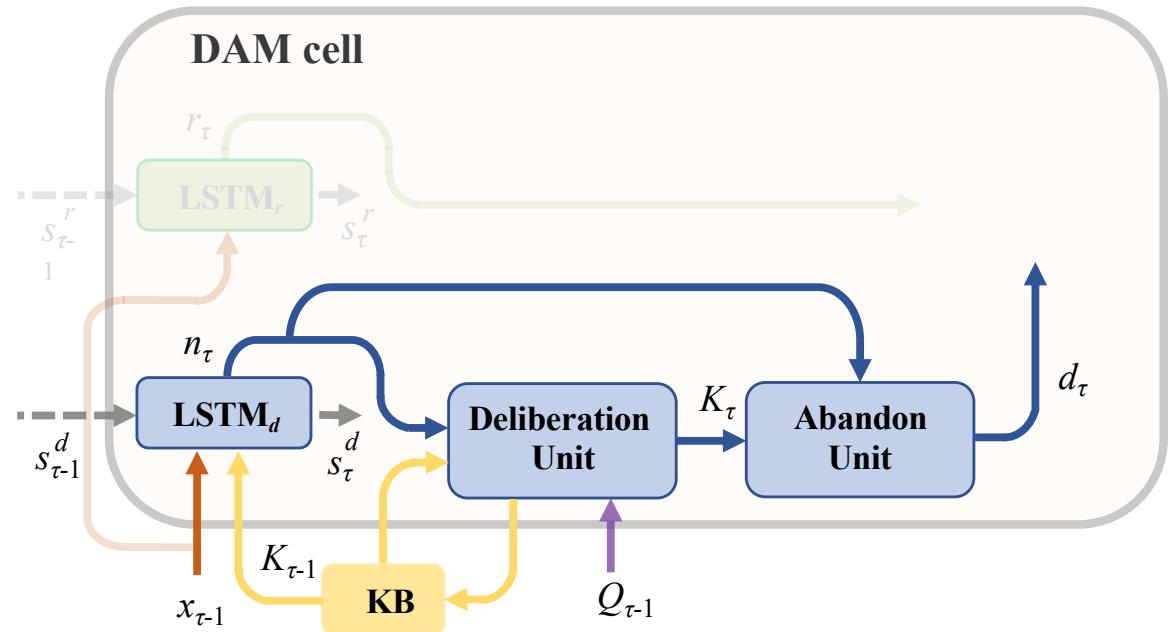


Model



Word-level Detail Decode Layer (WDL)

- Deliberation Unit
- ✓ Deliberation Unit aims to adaptively leverage the encoder structure to extract the most related and detailed information for current word generation.
- ✓ To prove the effectiveness of DAM, we combine it with three typical encoders: LF, MN, DualVD.



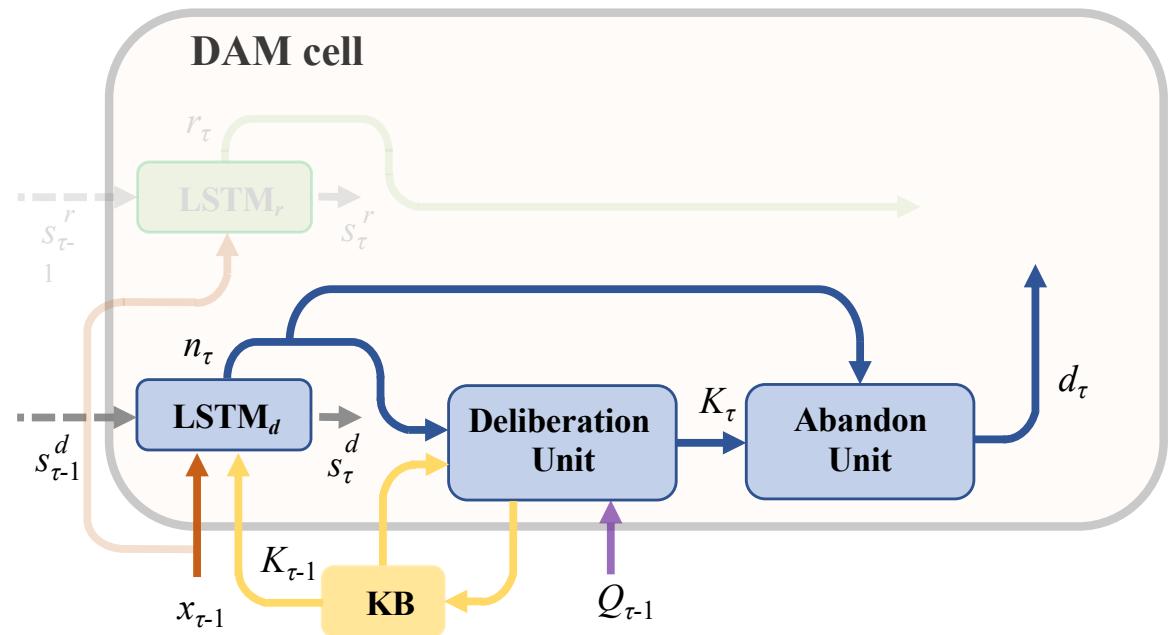
Model



Word-level Detail Decode Layer (WDL)

- Abandon Unit
- ✓ Abandon Unit further filters out the redundant information while enhancing the word-specific information from both the global and local encoded clues.

$$\begin{aligned} gate_{\tau}^a &= \sigma(\mathbf{W}_a[n_{\tau}, K_{\tau}] + b_a) \\ d_{\tau} &= gate_{\tau}^a \circ [n_{\tau}, K_{\tau}] \end{aligned}$$



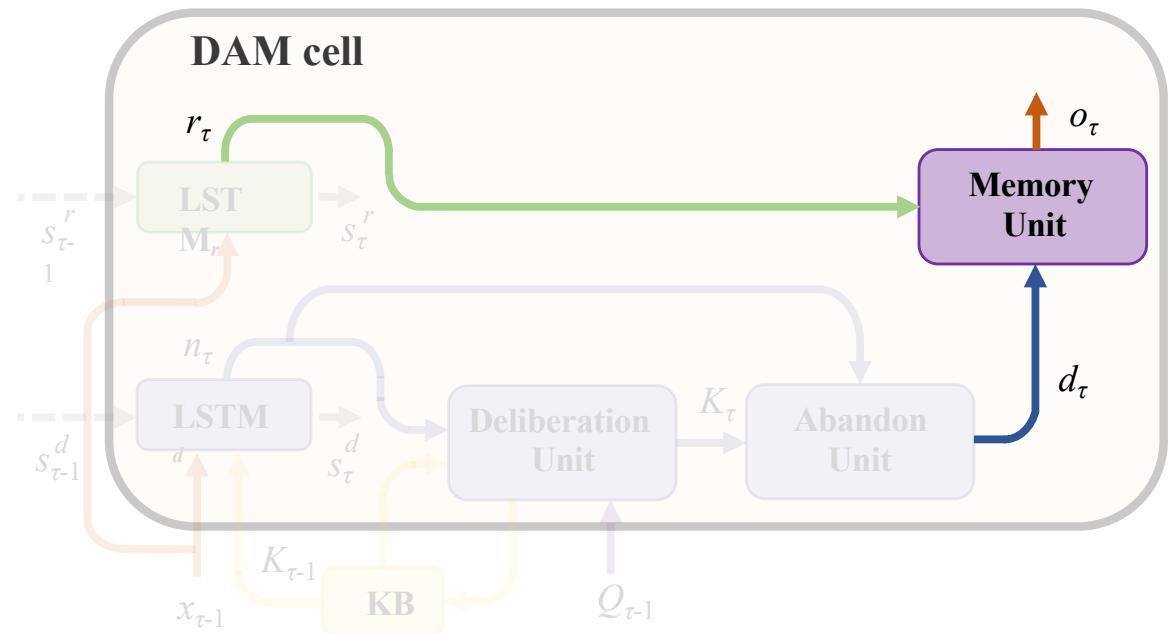
Model



Two Level Information Fusion

- Memory Unit
- ✓ Memory Unit selects response-level information to control the global semantics in response and tracks the word-level information for generating more detailed and less repeated response via a gate operation.

$$\begin{aligned} \text{gate}_\tau^m &= \sigma(\mathbf{W}_m[r_\tau, d_\tau] + b_m) \\ o_\tau &= \text{gate}_\tau^m \circ [r_\tau, d_\tau] \end{aligned}$$



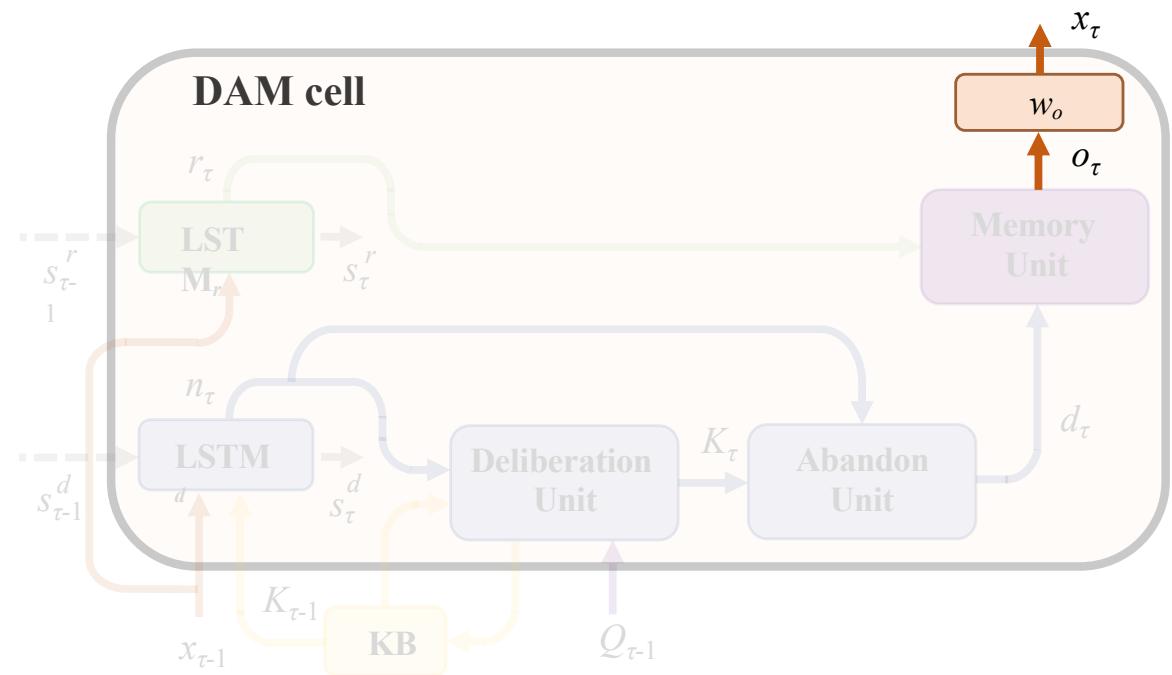
Model



Word Generation

- ✓ The generated word x_τ with the maximum value in the probability distribution is selected as the predicted word. Probability distribution is computed as:

$$P_\tau^o = \text{softmax}(\mathbf{w}_o^T o_\tau + b_o)$$



Model



Variants of Deliberation Unit

- Guided by the question and current generated word state n_τ , Deliberation Unit captures more detailed information from encoder-specific structures.
- To prove the effectiveness of DAM, we combine it with three typical encoders:
 - ✓ LF encoder^[1] for the general feature fusion
 - ✓ MN encoder^[1] for dialogue history reasoning
 - ✓ DualVD encoder^[2] for visual-semantic image understanding

[1] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, pages 1080–1089, 2017.

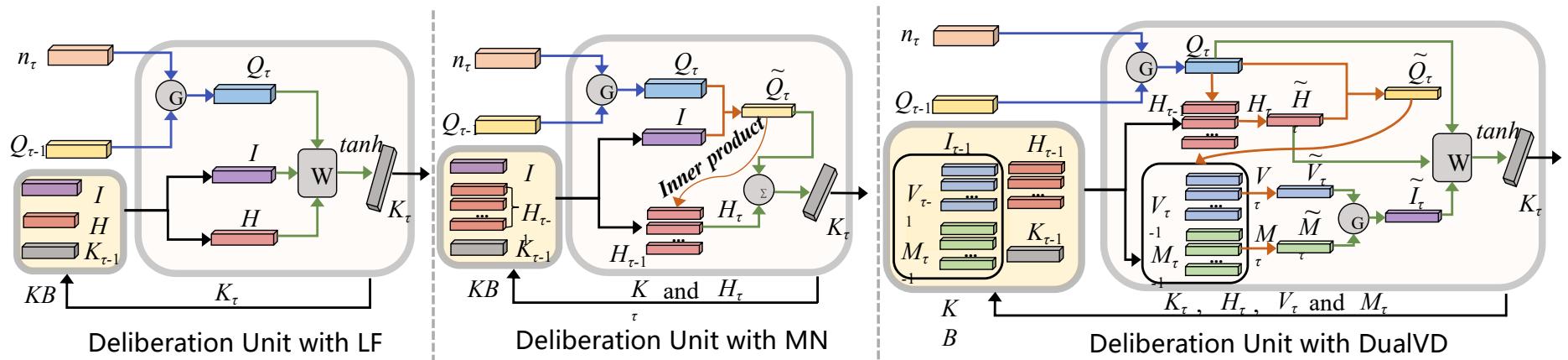
[2] Xiaoze Jiang, Jing Yu, Zengchang Qin, Yingying Zhuang, Xingxing Zhang, Yue Hu, and Qi Wu. Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue. In *AAAI*, 2020.

Model



Variants of Deliberation Unit

- It mainly contains three steps:
 - ✓ Step 1: word-guided question information update (blue lines in the following figure)
 - ✓ Step 2: question-guided information update (orange lines in the following figure)
 - ✓ Step 3: general feature fusion (green lines in the following figure)



Results

➤ Dataset

VisDial v1.0 : *train* set (120k) built on MSCOCO images,
test set (8k) and *val* set (2k) consist of dialogues on extra 10k COCO-like
images from Flickr.

➤ Evaluation Metrics

VisDial v1.0 : utilize retrieval metrics: MRR, R@ k ($k = 1, 5, 10$) and Mean,
NDCG for more comprehensive analysis.

Lower value for Mean and higher value for other metrics are desired.

➤ Experiments

- (1) Overall Results on VisDial v1.0
- (2) Ablation study (including Human Study)
- (3) Qualitative Analysis

Results

➤ Compare with State-of-the-art

Table 1: Result comparison on validation set of VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
HCIAE-G [Lu <i>et al.</i> , 2017]	49.07	39.72	58.23	64.73	18.43	59.70
CoAtt-G [Wu <i>et al.</i> , 2018]	49.64	40.09	59.37	65.92	17.86	59.24
Primary-G [Guo <i>et al.</i> , 2019]	49.01	38.54	59.82	66.94	16.69	-
ReDAN-G [Gan <i>et al.</i> , 2019]	49.60	39.95	59.32	65.97	17.79	59.41
DMRM [Chen <i>et al.</i> , 2020]	50.16	40.15	60.02	67.21	15.19	-
LF-G [Das <i>et al.</i> , 2017]	44.67	34.84	53.64	59.69	21.11	52.23
MN-G [Das <i>et al.</i> , 2017]	45.51	35.40	54.91	61.20	20.24	51.86
DualVD-G [Jiang <i>et al.</i> , 2020]	49.78	39.96	59.96	66.62	17.49	60.08
LF-DAM (ours)	45.08	35.01	54.48	60.57	20.83	52.68
MN-DAM (ours)	46.16	35.87	55.99	62.45	19.57	52.82
DualVD-DAM (ours)	50.51	40.53	60.84	67.94	16.65	60.93

Our models

Baseline models

- ✓ Compared with the baseline models, our models outperform them on all the metrics, which indicates the complementary advantages between DAM and existing encoders in visual dialogue.
- ✓ DualVD-DAM outperforms DMRM on all the other metrics without multi-step reasoning, which is the advantages in DMRM over our models.

Results

➤ Ablation Study

We conduct extensive ablation study to verify the following key points:

- ✓ The Effectiveness of Each Unit
- ✓ The Effectiveness of Two-Level Decode Structure
- ✓ The Effectiveness of Each Operation in Deliberation Unit

Results

✓ The Effectiveness of Each Unit

Table 2: Ablation study of each unit on VisDial v1.0 validation set.

Base Model	Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF-DAM	2LSTM	44.43	34.53	53.55	59.48	21.38	51.99
	2L-M	44.77	34.85	54.06	60.03	21.13	52.04
	2L-DM	45.06	34.90	54.24	60.39	20.87	52.58
	2L-DAM	45.08	35.01	54.48	60.57	20.83	52.68
MN-DAM	2LSTM	45.58	35.27	55.38	61.54	19.96	52.38
	2L-M	45.67	35.29	55.57	61.97	19.91	52.11
	2L-DM	45.77	35.53	55.40	62.05	19.95	52.51
	2L-DAM	46.16	35.87	55.99	62.45	19.57	52.82
DualVD-DAM	2LSTM	49.72	40.04	59.52	66.41	17.62	59.79
	2L-M	50.09	40.38	59.94	66.77	17.31	59.85
	2L-DM	50.20	40.33	60.22	67.48	17.15	59.72
	2L-DAM	50.51	40.53	60.84	67.94	16.65	60.93

- 1) 2L-DAM: this is our full model that adaptively selects related information for decoding.
- 2) 2L-DM: full model w/o Abandon Unit.
- 3) 2L-M: 2L-DM w/o Deliberation Unit.
- 4) 2-LSTM: 2L-M w/o Memory Unit.

Results

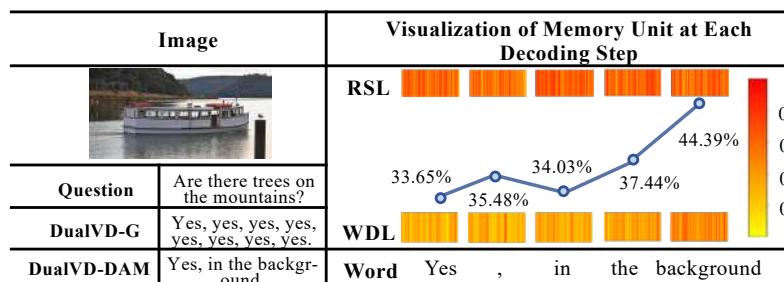
✓ The Effectiveness of Two-Level Decode Structure

- Complementary Advantages

Table 3: Human evaluation of 100 sample responses on VisDial v1.0 validation set.

Model	M1 ↑	M2 ↑	Repetition ↓	Richness ↑
RSL(DualVD-G): RSL only	0.60	0.47	0.20	0.03
WDL: WDL only	0.69	0.54	0.07	0.15
DualVD-DAM	0.75	0.61	0.01	0.13

- Information Compositive Mode



Results

✓ The Effectiveness of Each Operation in Deliberation Unit

Table 4: Ablation study of Deliberation Unit on VisDial v1.0 validation set.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
I-S	50.01	40.25	59.78	66.76	17.67	59.09
I-V	50.03	40.30	59.34	66.90	17.34	58.93
I-SV	50.13	40.34	60.09	67.06	17.34	59.51
H	50.19	40.36	60.09	66.96	17.27	59.92
DualVD-DAM	50.51	40.53	60.84	67.94	16.65	60.93

- 1) **I-S** only uses semantic-level image information for information selection.
- 2) **I-V** only utilizes visual-level image information for information selection.
- 3) **I-SV** jointly exploits semantic and visual information for information selection.
- 4) **H** only leverages dialogue history for information selection.

Results

➤ Qualitative Analysis

- ✓ Information selection quality

Image	Dialogue History	Visualization		
	C: A tarmac with a lot of large blue and white planes parked. Q1: Are there people? A1: I see 2 people. Q2: Is it sunny? A2: It looks like a clear day, yes. Q3: Are there clouds? A3: A couple of clouds, yes. Q4: Are the planes big? A4: They look like large passenger planes. Q5: Are there people boarding? A5: No. Q6: Are there any bags? A6: No. Q7: Are there signs? A7: No.			
	Question Is there a building?	C there	QA1 QA2 QA3 QA4 QA5 QA6 QA7 buildings	QA1 QA2 QA3 QA4 QA5 QA6 QA7 backgrounds
	DualVD-G No.	C there	QA1 QA2 QA3 QA4 QA5 QA6 QA7 buildings	QA1 QA2 QA3 QA4 QA5 QA6 QA7 backgrounds
	DualVD-DAM There are some buildings in the background.	C there	QA1 QA2 QA3 QA4 QA5 QA6 QA7 buildings	QA1 QA2 QA3 QA4 QA5 QA6 QA7 backgrounds

Figure: Visualization of the evidence when generating the response by DualVD-DAM. The essential visual regions and dialogue history for answering the question are highlighted in the last three columns. The attention weights of visual regions and dialogue history are visualized, where clearer region and darker orange color indicates higher attention weight.

报告提纲

- 1 / 多模态机器学习概述
- 2 / 视觉问答技术
- 3 / 视觉对话技术
- 4 / 总结与展望

总结与展望

- 现有两种保存知识的方法，即知识图谱和预训练模型，各自的特点？是否可以构建一个包罗万象的知识库？
- 如何解决跨媒体分析模型的可解释性问题？
- 如何结合符号方法和连接方法，获得更符合人类认知的推理能力？
- 我们如何更客观准确地度量机器的推理能力？

感谢聆听! Q & A



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

于静，助理研究员
中科院信息工程研究所
邮箱: yujing02@iie.ac.cn
地址: 北京市海淀区闵庄路甲89号, 100093