KVQA: Knowledge-aware Visual Question Answering

Sanket Shah[*†1] **Anand Mishra**[*2] **Naganand Yadati**[2] **Partha Pratim Talukdar**[2]
[1]**IIIT Hyderabad, India** [2]**Indian Institute of Science, Bangalore, India**
sanket.shah@research.iiit.ac.in anandmishra@iisc.ac.in naganand@iisc.ac.in ppt@iisc.ac.in
http://malllabiisc.github.io/resources/kvqa/

吴江恒（jiangh_wu@163.com）

计算机科学与工程学院· KGCODE LAB

Outline

- 背景——为什么需要新的VQA数据集？
- 诞生——KVQA，VQA needs knowledge
- 基线——Facenet+MemNet

传统VQA数据集

Conventional VQA (Antol et al. 2015; Goyal et al. 2017; Trott, Xiong, and Socher 2018)
Q: How many people are there in the image?
A: 3

传统的VQA数据集遵循“答案能在图像中直接找到”，问题类型多集中于图片实体的方位、数量、属性等等。

如果需要常识

Commonsense knowledge-enabled VQA (Wang et al. 2017; 2018; Su et al. 2018; G. Narasimhan and Schwing 2018)
Q: What in the image is used for amplifying sound?
A: Microphone

这类VQA问题需要了解人类世界的常识，如：麦克风能放大说话的音量。

如果考虑命名实体

而两者都只停步于概念层级，不涉及命名实体层级。

而考虑命名实体层面的问题如下：

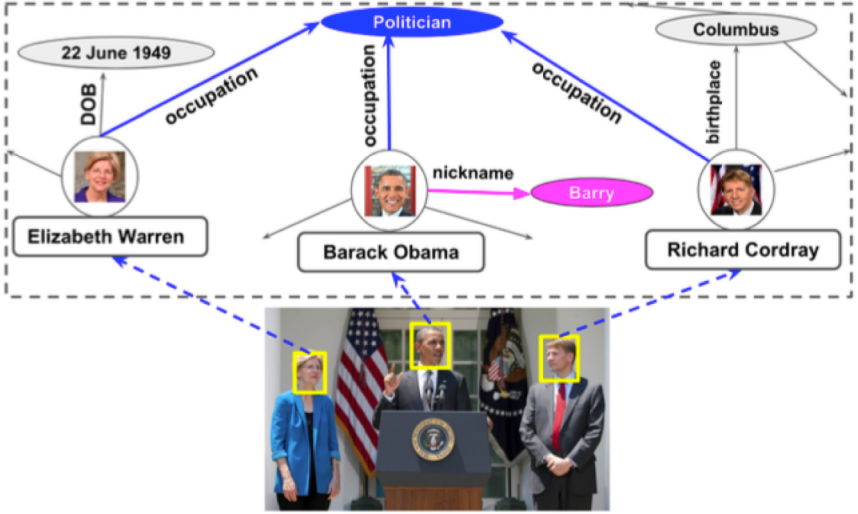
(World) knowledge-aware VQA (KVQA, this paper):
Q: Who is to the left of Barack Obama?
A: Richard Cordray
Q: Do all the people in the image have a common occupation?
A: Yes
Q: Who among the people in the image is called by the nick-name Barry?
A: Person in the center

显然，它们更接近人类感兴趣的问题。

如果考虑命名实体

回答涉及命名实体的问题，就需要知识图谱来辅助了：

(World) knowledge-aware VQA (KVQA, this paper):
Q: Who is to the left of Barack Obama?
A: Richard Cordray
Q: Do all the people in the image have a common occupation?
A: Yes
Q: Who among the people in the image is called by the nickname Barry?
A: Person in the center



KVQA提供了一个这类问题的数据集，下面介绍它的构成。

VQA领域数据集比较

Dataset name	# images	# QA pairs	Image source	Knowledge Graph type	Named entities
KVQA (this paper)	24,602	183,007	Wikipedia	World knowledge	✓
FVQA (Wang et al. 2018)	1,906	4,608	COCO	Commonsense	✗
KB-VQA (Wang et al. 2017)	700	2,402	COCO + ImgNet	Commonsense	✗
TallyQA (Acharya, Kafle, and Kanan 2019)	165,443	287,907	Visual genome + COCO	✗	✗
CLEVR (Johnson et al. 2017)	100,000	999,968	Synthetic images	✗	✗
VQA-2 (Goyal et al. 2017)	204,721	1,105,904	COCO	✗	✗
Visual Genome (Krishna et al. 2017)	108,000	1,445,322	COCO	✗	✗
TQA (Kembhavi et al. 2017)	-	26,260	Textbook	✗	✗
Visual 7w (Zhu et al. 2016)	47,300	327,939	COCO	✗	✗
Movie-QA (Tapaswi et al. 2016)	-	14,944	Movie videos	✗	✗
VQA (Antol et al. 2015)	204,721	614,163	COCO	✗	✗
VQA-abstract (Antol et al. 2015)	50,000	150,000	Clipart	✗	✗
COCO-QA (Ren, Kiros, and Zemel 2015)	69,172	117,684	COCO	✗	✗
DAQUAR (Malinowski and Fritz 2014)	1,449	12,468	NYU-Depth	✗	✗

KVQA是目前唯一一个蕴含命名实体知识的数据集

KVQA的构成

图片：收集自wikipedia页面，内容为公众人物。图片中人物会被按顺序标注出来并链接到wikidata数据库中

问题：由模板生成，利用SPARQL在wikidata中查找得到答案

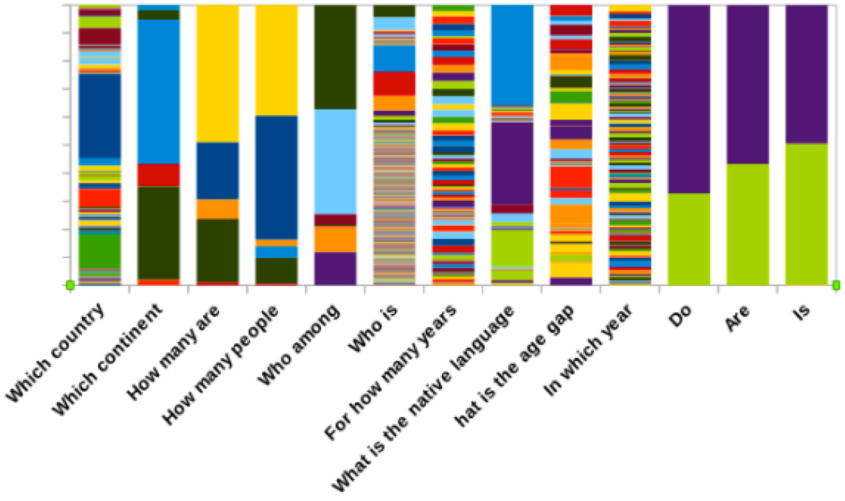
知识库：wikidata，以三元组的形式构成，每个实体和关系都有独立的id

KVQA dataset statistics:	
Number of images	24,602
Number of QA pairs	183,007
Number of unique entities	18,880
Number of unique answers	19,571
Average question length (words)	10.14
Average answer length (words)	1.64
Average number of questions per image	7.44

质量&难度

KVQA在QA对时，进行了如下设置：

1. 问题的种类丰富，包括方位、一跳和多跳、判断、交集、差集、比较、计数、多实体多关系问题
2. 对问题进行改写，保证问法的多样性
3. 答案分布尽量均匀，避免偏置的干扰
4. 答案可以超出图片包含的实体
5. 推理不超过三跳



问法及其对应的答案分布

数据实例

(a) ***Wikipedia caption*:** Khan with United States Secretary of State Hillary Clinton in 2009.

Q: Who is to the left of Hillary Clinton? (*spatial*)
A: **Aamir Khan**

Q: Do all the people in the image have a common occupation? (*multi-entity, intersection, 1-hop, Boolean*)
A: **No**

(b) ***Wikipedia caption*:** Cheryl alongside Simon Cowell on The X Factor, London, June 2010.

Q: What is the age gap between the two people in the image? (*multi-entity, subtraction, 1-hop*)
A: **24 years**

Q: How many people in this image were born in United Kingdom? (*1-hop, multi-entity, counting*)
A: **2**

(c) ***Wikipedia caption*:** BRICS leaders at the G-20 summit in Brisbane, Australia, 15 November 2014

Q: Were all the people in the image born in the same country? (*Boolean, multi-entity, intersection*)
A: **No**

Q: Who is the founder of the political party to which person second from left belongs to? (*spatial, multi-hop*)
A: **Syama Prasad Mookerjee**

(d) ***Wikipedia caption*:** Serena Williams and Venus Williams, Australian Open 2009.

Q: Who among the people in the image is the eldest? (multi-entity, comparison)
A: **Person in the left**

Q: Who among the people in the image were born after the end of World War II? (multi-entity, multi-relation, comparison)
A: **Both**

训练&评价

1. 划分：721的比例随机划分训练、测试与验证集，进行5次

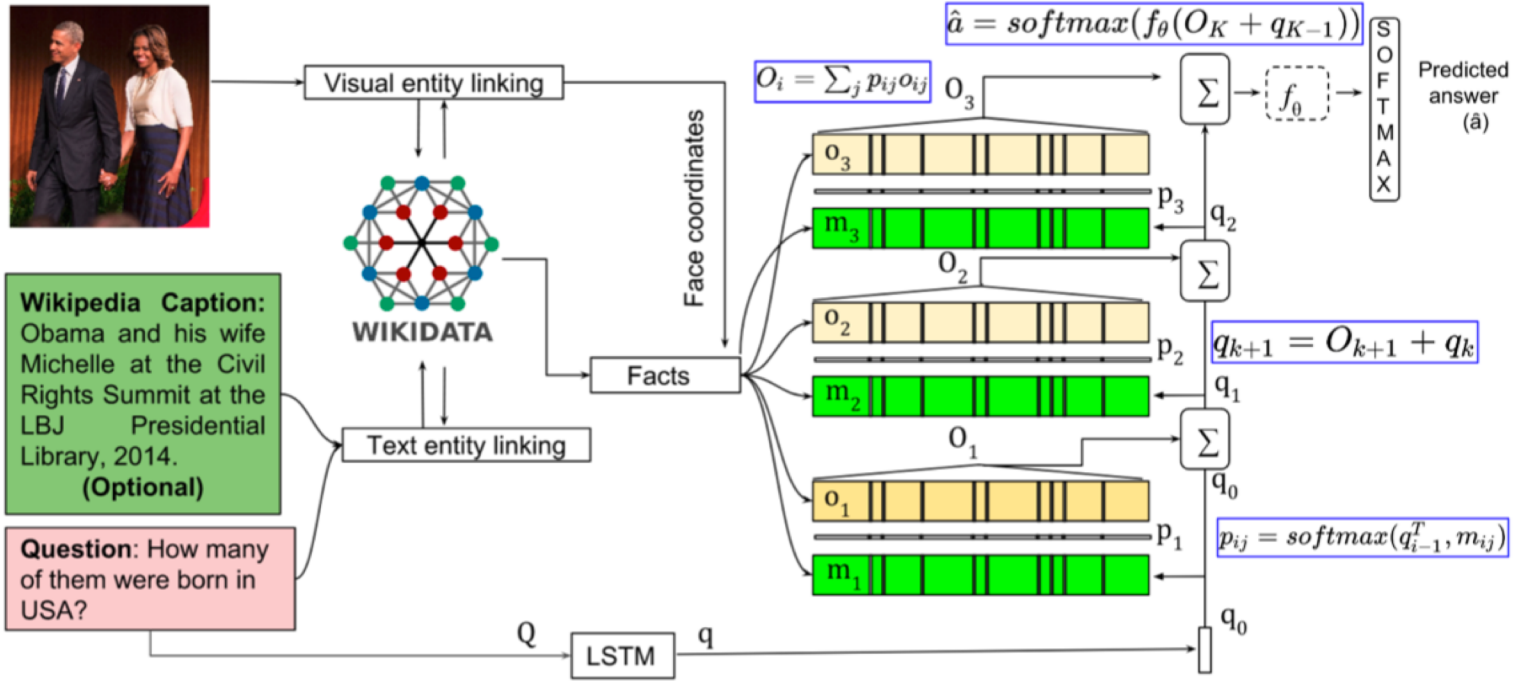
2. 评价标准：所有划分的mean accuracy

3. 封闭世界&开放世界

   - 封闭世界：涉及18K个实体，事实只包含18种预先指定的关系
   - 开放世界：涉及69K个实体，事实包含200种常见关系

KVQA解决方案

Visual Entity Linking + VQA over KG



Visual Entity Linking

主要任务是大规模的人脸识别

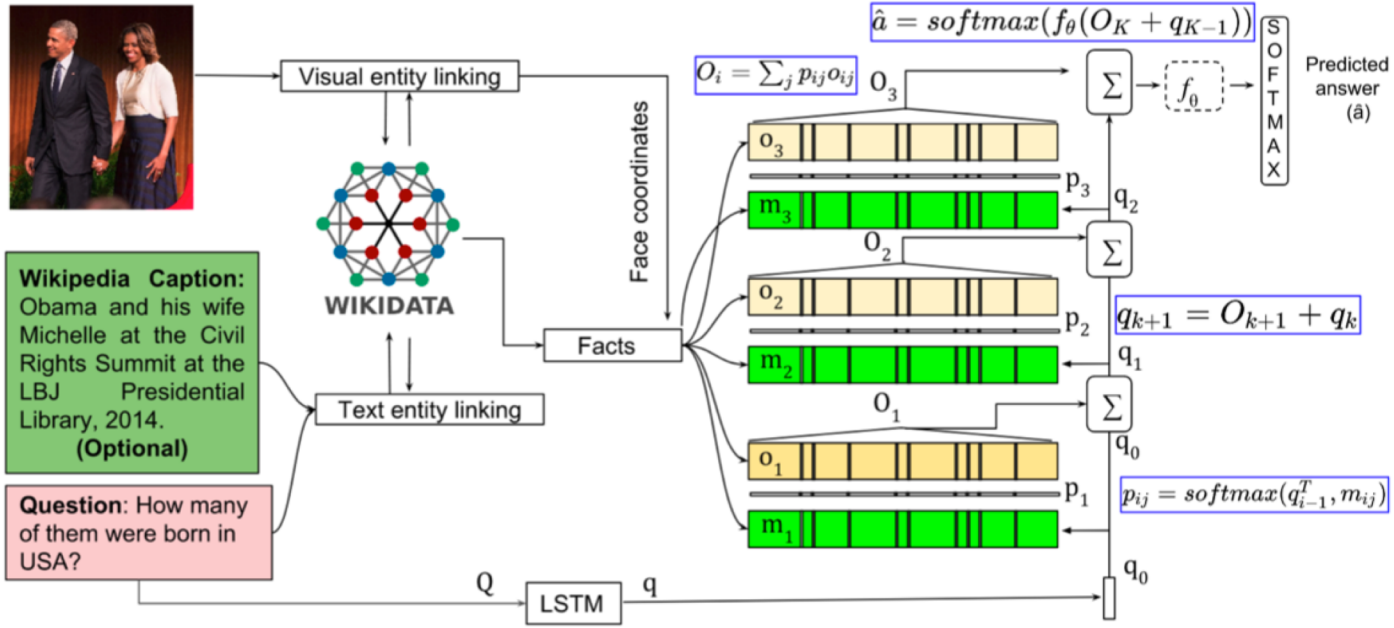
利用表现较好的Facenet (2017- 05-12)

Method	Precision		Recall	
	Top-1	Top-5	Top-1	Top-5
MS-Captionbot	95.2	NA	16.6	NA
Facenet [Closed world]	81.0	-	82.2	87.2
Facenet [Open world]	73.5	-	71.4	76.5

VQA over KG

包括实体链接、从KG中获取相关知识、知识与问题的表示、得到答案。

使用了MenNet作为baseline



VQA over KG

MemNet和BiLSTM的效果比较，

Oracle为假设VEL已解决的情况

	Method	Oracle	-wikiCap		+wikiCap	
			ORG	PRP	ORG	PRP
Closed World	BLSTM	51.0	47.2	25.0	48.0	27.2
	MemNet	59.2	49.5	32.0	50.2	34.2
Open World	BLSTM	–	16.8	13.1	20.6	14.0
	MemNet	–	36.0	26.2	36.8	30.5

MemNet在各类问题上的表现，

PRP为问题的改写形式

Category	ORG	PRP	Category	ORG	PRP
Spatial	48.1	47.2	Multi-rel.	45.2	44.2
1-hope	61.0	60.2	Subtraction	40.5	38.0
Multi-hop	53.2	52.1	Comparison	50.5	49.0
Boolean	75.1	74.0	Counting	49.5	48.9
Intersect.	72.5	71.8	Multi-entity	43.5	43.0

总结

- KVQA需要额外的实例知识来解决
- 提出了一个解决KVQA的baseline框架

是否可以用图神经网络来改进效果？


धन्यवाद
Hindi
Hindi
多謝
Traditional Chinese
ขอบคุณ
Thai
Спасибо
Russian
Thank You
English
Gracias
Spanish
شكرا
Arabic
Grazie
Italian
Danke
German
Obrigado
Brazilian Portuguese
多谢
Simplified Chinese
Merci
French
நன்றி
Tamil
Tamil
ありがとうございました
Japanese
감사합니다
Korean