Knowledge-Based Reasoning in Computer Vision

CSC 2539

Paul Vicol

Outline

- Knowledge Bases
- Motivation
- Knowledge-Based Reasoning in Computer Vision
    - Visual Question Answering
    - Image Classification

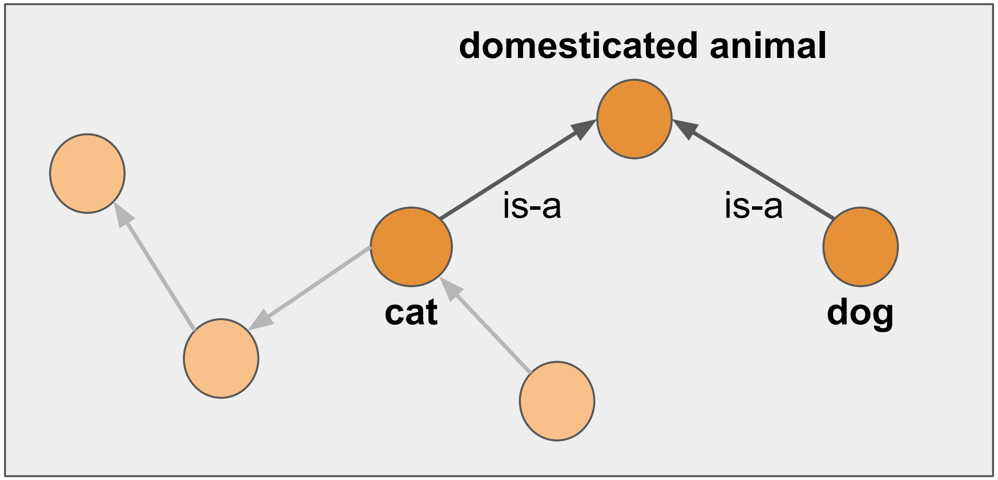
Knowledge Bases

- **KB:** Knowledge in a structured, computer-readable form
- Many KBs exist: different types of information → useful for different tasks
  - Commonsense knowledge: *a person is a physical entity*
  - Encyclopedic knowledge: *a dog is a mammal*
  - Visual knowledge: *a person can wear shorts*
- Advances in structured KBs have driven *natural language question answering*
  - IBM's Watson, Evi, Amazon Echo, Cortana, Siri, etc.

Knowledge Bases

- Knowledge is represented by a *graph* composed of *triples* **(arg1, rel, arg2):**

KB:



Query:

```
?x : (?x, is-a, domesticated animal)
```

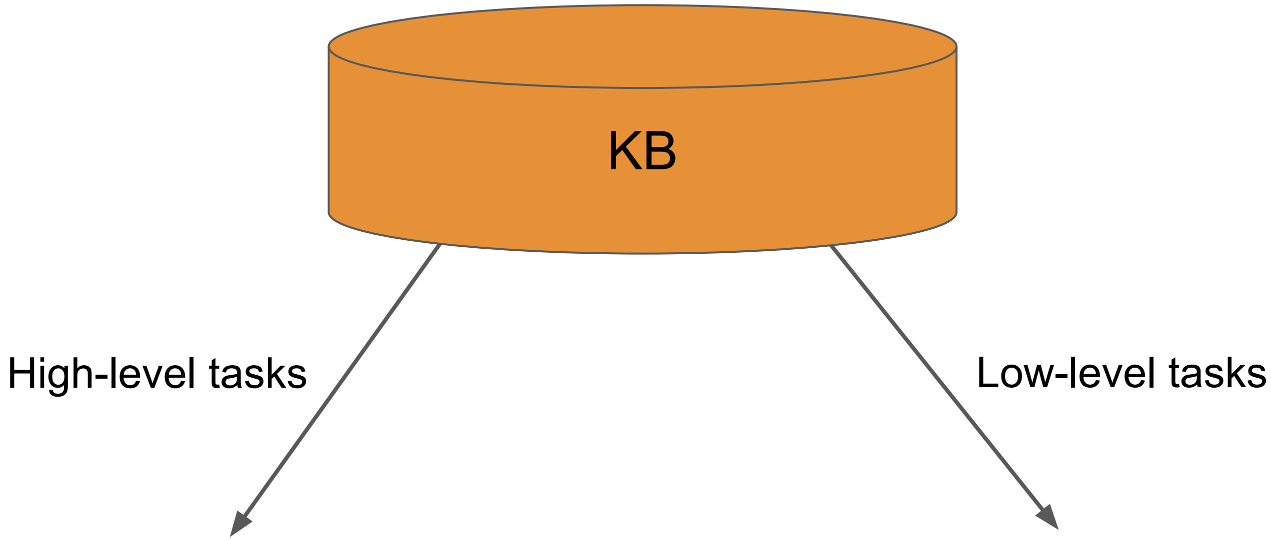
Result:

cat, dog, horse, pig, ...

List of all domesticated animals

Knowledge Bases in Computer Vision

- How can external knowledge be used in computer vision?
- In humans, vision and reasoning are intertwined
  - You use your external knowledge of the world all the time to understand what you see



VQA

Enable reasoning with external knowledge to answer complicated questions that go beyond what is visible in an image

Image Classification, etc.

Enable using knowledge about the world to identify objects based on their properties or relationships with other objects.

Visual Question Answering (VQA)

- The task of VQA involves understanding the content of images, but often requires **prior non-visual information**
- For general questions, VQA requires reasoning with external knowledge
  - Commonsense, topic-specific, or encyclopedic knowledge
  - Right image: need to know that umbrellas provide shade on sunny days

A Purely Visual Question

Q: What color is the train?

A: Yellow

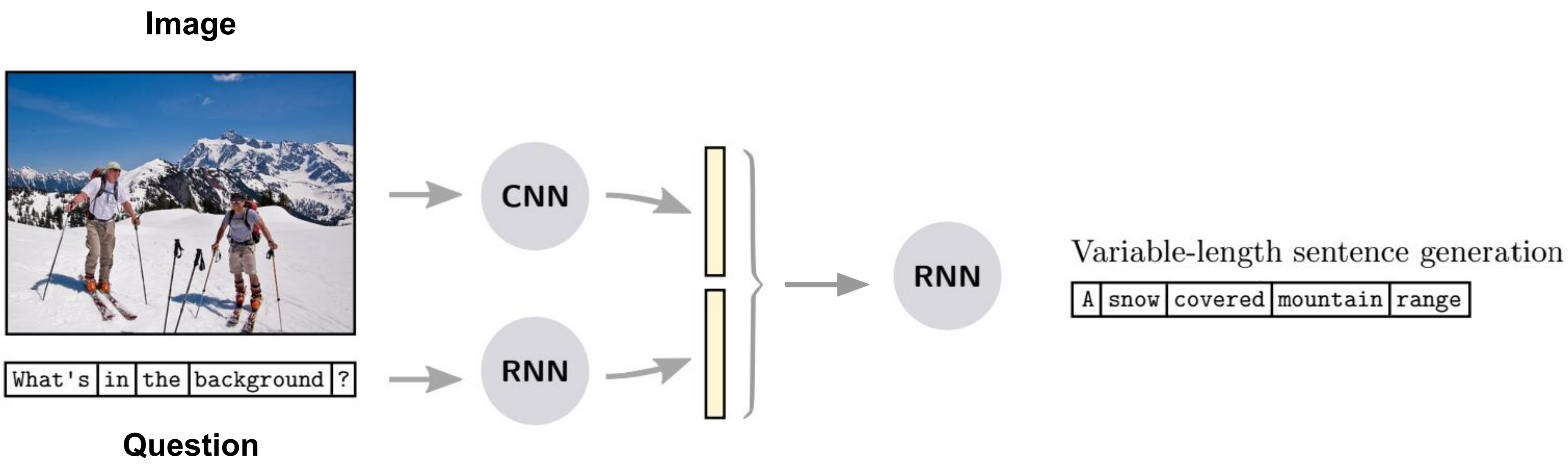
A More Involved Question

Q: Why do they have umbrellas?

A: Shade

The Dominant Approach to VQA

- Most approaches combine CNNs with RNNs to learn a mapping directly from input images and questions to answers:



“Visual Question Answering: A Survey of Methods and Datasets.” Wu et al. https://arxiv.org/pdf/1607.05910.pdf. 2016.

Limitations of the Straightforward Approach

+ Works well in answering simple questions directly related to the image content
    - "What color is the ball?"
    - "How many cats are there?"

- Not capable of *explicit reasoning*
- *No explanation* for how it arrived at the answer
    - Using image info, or using the prevalence of an answer in the training set?
- Can only capture knowledge that is in the training set
    - *Some knowledge is* provided by class labels or captions in MS COCO
    - Only a limited amount of information can be encoded within an LSTM
    - Capturing this would require an implausibly large LSTM

- **Alternative strategy:** Decouple the *reasoning* (e.g. as a neural net) from the storage of knowledge (e.g. in a *structured KB*)

"Explicit Knowledge-Based Reasoning for VQA." Wang et al. https://arxiv.org/pdf/1511.02570.pdf. 2015.

VQA Methods

Method	Knowledge Based	Explicit Reasoning	Structured Knowledge	Number of KBs
CNN-LSTM	×	×	×	0
Ask Me Anything	✓	×	×	1
Ahab	✓	✓	✓	1
FVQA	✓	✓	✓	3

Ask Me Anything: Introduction

- Combines image features with external knowledge

[Method]

1. Construct a *textual* representation of an image
2. Merge this representation with *textual* knowledge from a KB
3. Feed the merged information to an LSTM to produce an answer

Attributes:
umbrella
beach
sunny
day
people
sand
laying
blue
green
mountain

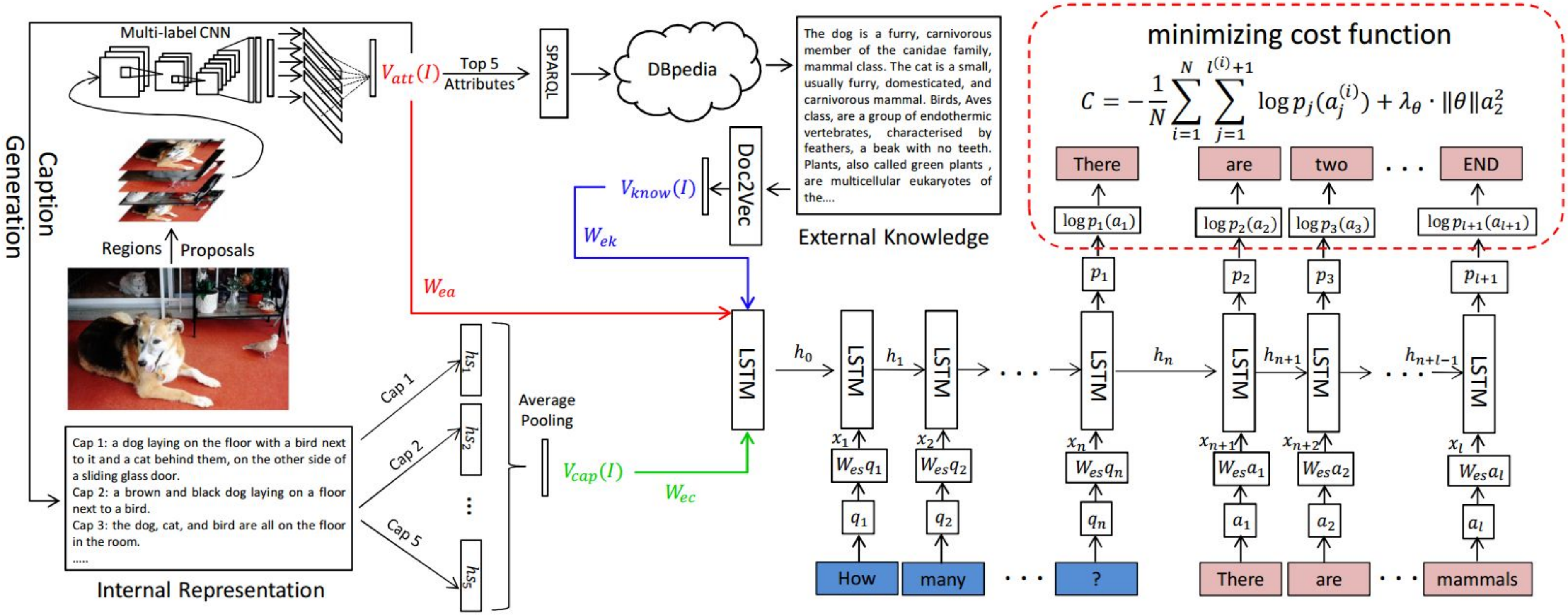
Internal Textual Representation:
A group of people enjoying a sunny day at the beach with umbrellas in the sand.

External Knowledge:
An umbrella is a canopy designed to protect against rain or sunlight. Larger umbrellas are often used as points of shade on a sunny beach. A beach is a landform along the coast of an ocean. It usually consists of loose particles, such as sand....

Question Answering:
Q: Why do they have umbrellas? **A :** Shade.

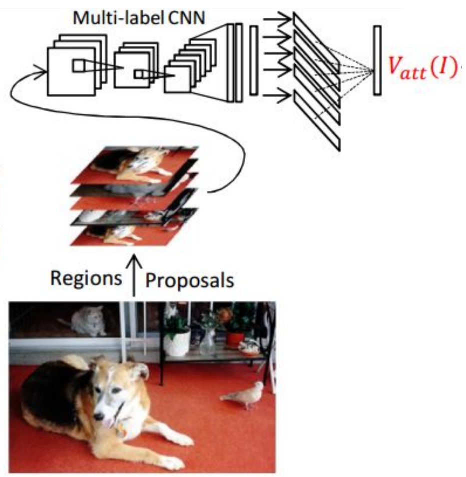
"Ask Me Anything: VQA Based on Knowledge from External Sources" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: Architecture



"**Ask Me Anything: VQA Based on Knowledge from External Sources**" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

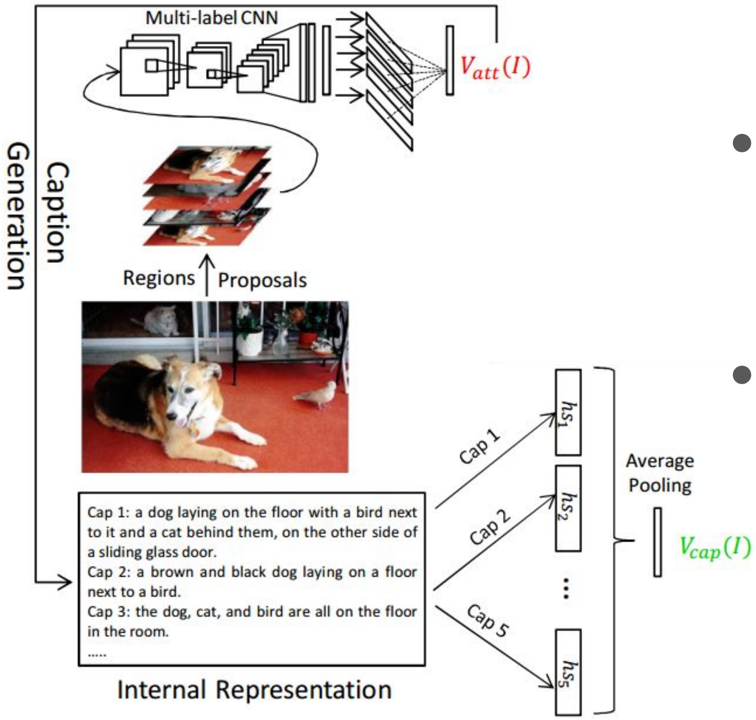
Ask Me Anything: Attribute-Based Representation



- Describe image content in terms of a set of *attributes*
- Attribute vocabulary derived from words in MS COCO captions
  - Attributes can be *objects* (nouns), *actions* (verbs), or *properties* (adjectives)
- Region-based multi-label classification → CNN outputs a probability distribution over 256 attributes *for each region*
- Outputs for each region are max-pooled to produce a single prediction vector $V_{att}(I)$

"Ask Me Anything: VQA Based on Knowledge from External Sources" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: Caption-Based Representation



- Based on the attribute vector $V_{att}(I)$, generate five different captions
  - The captions constitute the *textual representation* of the image
- Average-pooling over the hidden states of the caption-LSTM after producing each sentence yield $V_{cap}(I)$

"Ask Me Anything: VQA Based on Knowledge from External Sources" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: Example Captions from Attributes

Top 5 Attributes:
players, catch, bat, baseball, swing

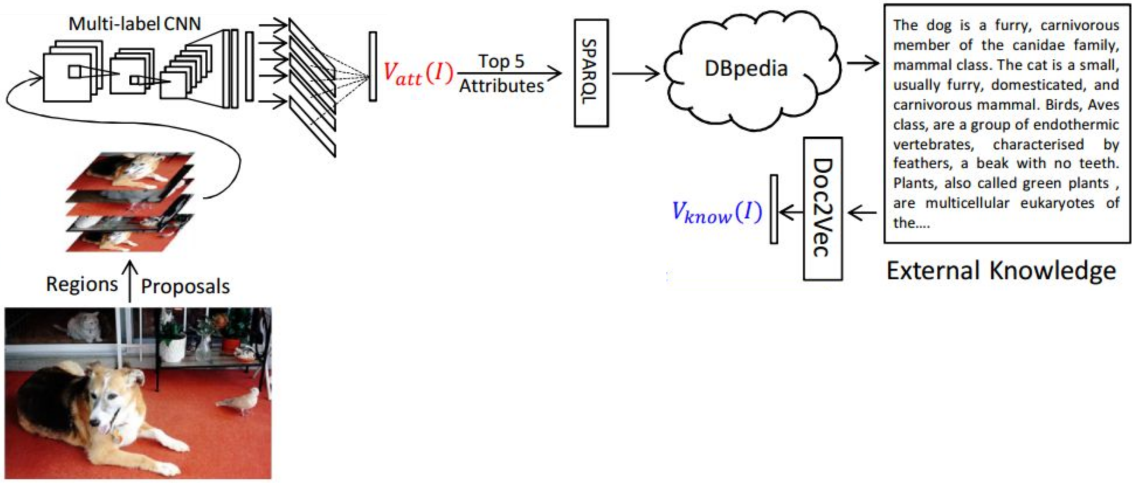
Generated Captions:
A baseball player swing a bat at a ball.
A baseball player holding a bat on a field.
A baseball player swinging a bat on a field.
A baseball player is swinging a bat at a ball.
A batter catcher and umpire during a baseball game.

Top 5 Attributes:
field, two, tree, grass, giraffe

Generated Captions :
Two giraffes are standing in a grassy field.
A couple of giraffe standing next to each other.
Two giraffes standing next to each other in a field.
A couple of giraffe standing next to each other on a lush green field.

"Ask Me Anything: VQA Based on Knowledge from External Sources" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

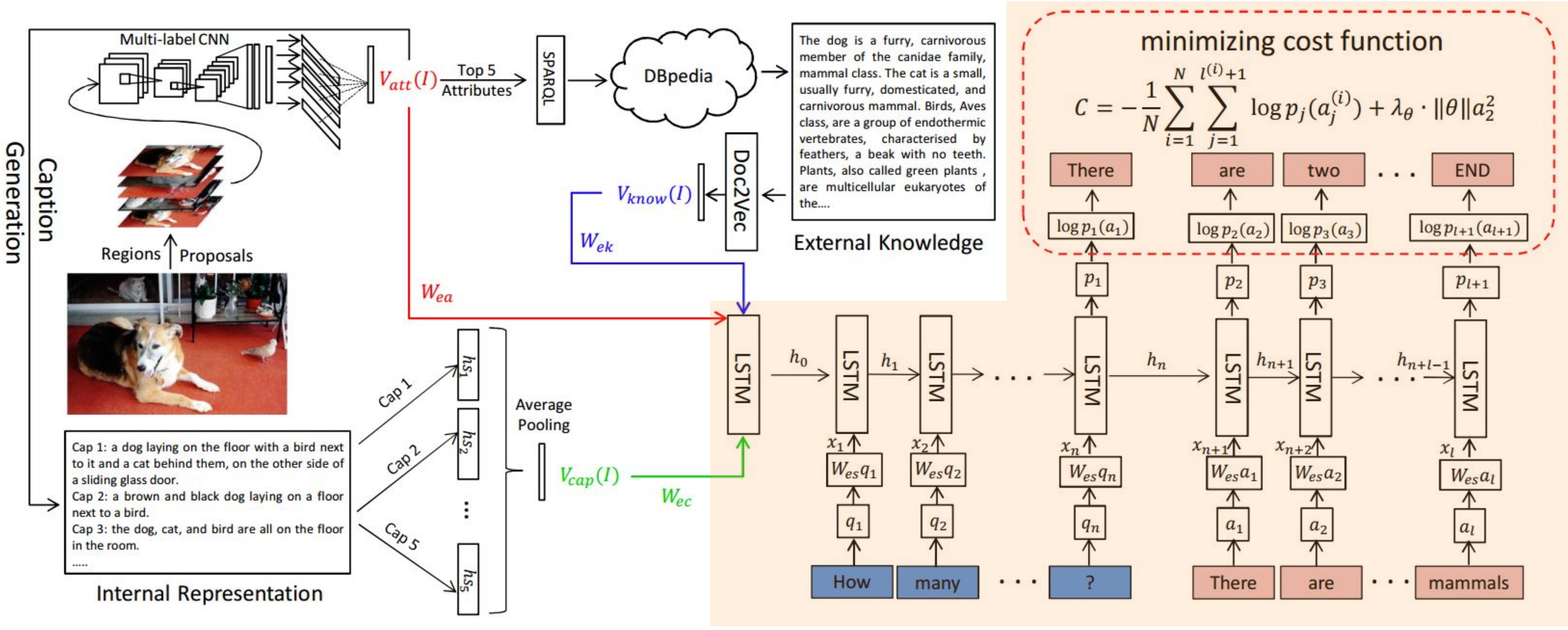
Ask Me Anything: External Knowledge



- *Pre-emptively* fetch information related to the top 5 attributes
  - Issue a SPARQL query to retrieve the textual “comment” field for each attribute
- A comment field contains a paragraph description of an entity
- Concatenate the 5 paragraphs → Doc2Vec → $V_{know}(I)$

“Ask Me Anything: VQA Based on Knowledge from External Sources” Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: Full Architecture



- Pass $V_{att}(I)$, $V_{cap}(I)$, and $V_{know}(I)$ as the initial input to an LSTM that reads in the question word sequence and learns to predict the sequence of words in the answer

"Ask Me Anything: VQA Based on Knowledge from External Sources" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: Evaluation

- [Toronto COCO-QA]
  - 4 types of questions (object, number, color, location)
  - Single-word answer
  - Questions derived automatically from human captions on MS-COCO
- [VQA]
  - Larger, more varied dataset
  - Contains "What is," "How many," and "Why" questions
- The model is compared against a CNN-LSTM baseline

"Ask Me Anything: VQA Based on Knowledge from External Sources" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: COCO Evaluation

- **Att+Cap-LSTM** performs better than **Att+Know-LSTM**, so information from captions is more valuable than information from the KB
- **COCO-QA** does not test the use of external information

Toronto COCO-QA	Acc(%)	WUPS@0.9	WUPS@0.0
GUESS[23]	6.65	17.42	73.44
VIS+BOW[23]	55.92	66.78	88.99
VIS+LSTM[23]	53.31	63.91	88.25
2-VIS+BLSTM[23]	55.09	65.34	88.64
Ma et al.[17]	54.94	65.36	88.58
Baseline			
VggNet-LSTM	50.73	60.37	87.48
VggNet+ft-LSTM	58.34	67.32	89.13
Our-Proposal			
Att-LSTM	61.38	71.15	91.58
Att+Cap-LSTM	69.02	76.20	92.38
Att+Know-LSTM	63.07	72.22	90.84
Cap+Know-LSTM	64.31	73.31	90.01
Att+Cap+Know-LSTM	**69.73**	**77.14**	**92.50**

Table 2. Accuracy, WUPS metrics compared to other state-of-the-art methods and our baseline on Toronto COCO-QA dataset.

"Ask Me Anything: VQA Based on Knowledge from External Sources" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: Evaluation

Question Type	Our-Baseline VggNet + LSTM	Our Proposal Att + LSTM	Att+Cap + LSTM	Att+Know + LSTM	Cap+Know + LSTM	A+C+K + LSTM
what is	21.41	34.63	42.21	37.11	35.58	**42.52**
what colour	29.96	39.07	48.65	39.68	40.62	**48.86**
what kind	24.15	41.22	47.93	46.16	44.04	**48.05**
what are	23.05	38.87	47.13	41.13	39.73	**47.21**
what type	26.36	41.71	47.98	44.91	44.95	**48.11**
is the	71.49	73.22	74.63	74.40	73.78	**74.70**
is this	73.00	75.26	76.08	**76.56**	74.18	76.14
how many	34.42	39.14	46.61	39.78	44.20	**47.38**
are	73.51	75.14	76.01	75.75	75.78	**76.14**
does	76.51	76.71	78.07	76.55	77.17	**78.11**
where	10.54	21.42	25.92	24.13	16.09	**26.00**
is there	86.66	**87.10**	86.82	85.87	85.26	87.01
why	3.04	7.77	9.63	11.88	9.99	**13.53**
which	31.28	36.60	**39.55**	37.71	37.86	38.70
do	76.44	75.76	78.18	75.25	74.91	**78.42**
what does	15.45	19.33	21.80	19.50	19.04	**22.16**
what time	13.11	15.34	15.44	**15.47**	15.04	15.34
who	17.07	22.56	25.71	21.23	22.86	**25.74**
what sport	65.65	91.02	93.96	90.86	91.75	**94.20**
what animal	27.77	61.39	70.65	63.91	63.26	**71.70**
what brand	26.73	32.25	33.78	32.44	31.30	**34.60**
others	44.37	50.23	53.29	52.11	51.20	**53.45**
Overall	44.93	51.60	55.04	53.79	52.31	**55.96**

"Ask Me Anything: VQA Based on Knowledge from External Sources" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: VQA Evaluation

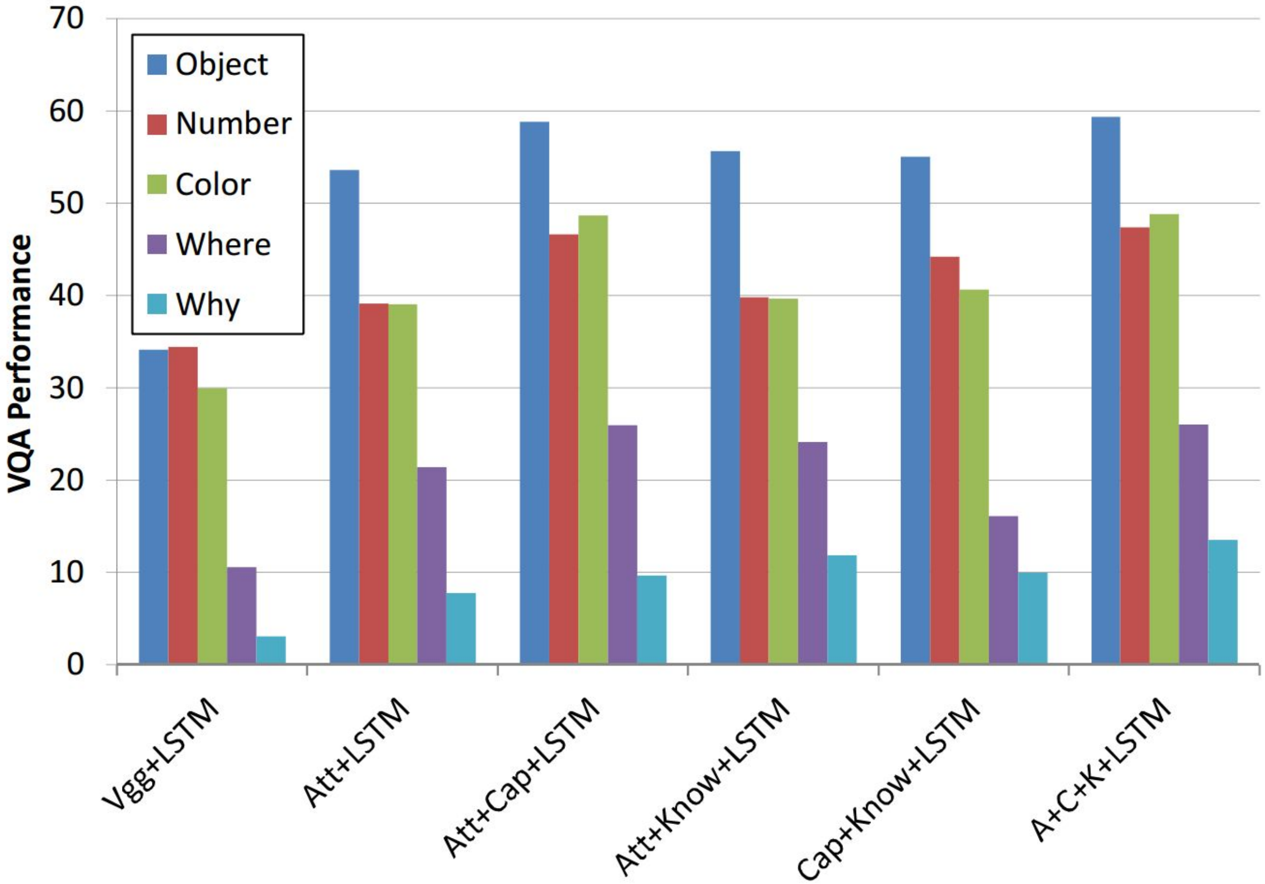
	Test-dev				Test-standard			
	All	Y/N	Num	Others	All	Y/N	Num	Others
Question [1]	40.09	75.66	36.70	27.14	-	-	-	-
Image [1]	28.13	64.01	0.42	3.77	-	-	-	-
Q+I [1]	52.64	75.55	33.67	37.37	-	-	-	-
LSTM Q [1]	48.76	78.20	35.68	26.59	48.89	78.12	34.94	26.99
LSTM Q+I [1]	53.74	78.94	35.24	36.42	54.06	79.01	35.55	36.80
Human [1]	-	-	-	-	83.30	95.77	83.39	72.67
Ours	**59.17**	**81.01**	**38.42**	**45.23**	**59.44**	**81.07**	**37.12**	**45.83**

Table 6. VQA Open-Ended evaluation server results for our method. Accuracies for different answer types and overall performances on test-dev and test-standard datasets are shown.

"Ask Me Anything: VQA Based on Knowledge from External Sources" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: VQA Evaluation

- **“Where”** questions require knowledge of potential locations
- **“Why”** questions require knowledge about people’s motivations
- Adding the KB improves results significantly for these categories



“Ask Me Anything: VQA Based on Knowledge from External Sources” Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: Qualitative Results




	What color is the tablecloth?	How many people in the photo?	What is the red fruit?	What are these people doing?
Ours:	white	2	apple	eating
Vgg+LSTM:	red	1	banana	playing
Ground Truth:	white	2	apple	eating




	Why are his hands outstretched?	Why are the zebras in water?	Is the dog standing or laying down?	Which sport is this?
Ours:	balance	drinking	laying down	baseball
Vgg+LSTM:	play	water	sitting	tennis
Ground Truth:	balance	drinking	laying down	baseball

"Ask Me Anything: VQA Based on Knowledge from External Sources" Wu et al. https://arxiv.org/pdf/1511.06973.pdf. 2016.

Ask Me Anything: Limitations

- Only extracts discrete pieces of text from the KB: **ignores the structured representation**
- No explicit reasoning: cannot provide explanations for how it arrived at an answer

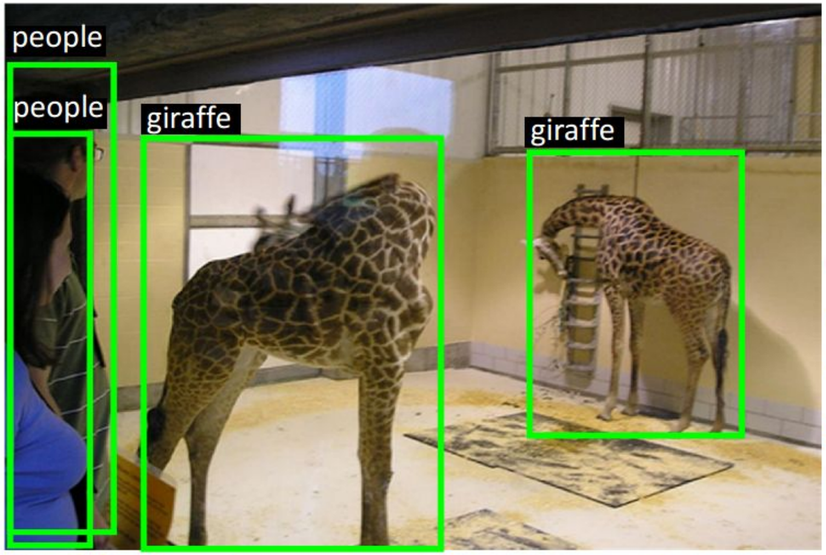
- This approach is evaluated on standard VQA datasets, not special ones that support higher-level reasoning
- **Need a new dataset with more knowledge-based questions**
- **Would also like explicit reasoning**
- Other approaches aim to make use of the *structure* of the KB and perform *explicit reasoning*
  - **Ahab** and **FVQA**
  - They introduce new small-scale datasets that focus on questions requiring external knowledge

"Explicit Knowledge-Based Reasoning for VQA." Wang et al. https://arxiv.org/pdf/1511.02570.pdf. 2015.

Ahab: Explicit Knowledge-Based Reasoning for VQA

- Performs *explicit reasoning* about the content of images
- Provides *explanations* of the reasoning behind its answers

1. Detect relevant image content
2. Relate that content to information in a KB
3. Process a natural language question into a KB query
4. Run the query over the combined image and KB info
5. Process the response to form the final answer



Attributes:
glass
house
room
standing
walking
wall
zoo

Scenes:
museum
indoor

Visual Question: How many giraffes in the image?
Answer: Two. Reason: Two giraffes are detected.

Common-Sense Question: Is this image related to zoology?
Answer: Yes. Reason: Object/Giraffe --> Herbivorous animals --> Animal --> Zoology; Attribute/Zoo --> Zoology.

KB-Knowledge Question: What are the common properties between the animal in this image and the zebra?
Answer: Herbivorous animals; Animals; Megafauna of Africa.

"Explicit Knowledge-Based Reasoning for VQA." Wang et al. https://arxiv.org/pdf/1511.02570.pdf. 2015.

KB-VQA Dataset

- Contains *knowledge-level questions* that require explicit reasoning about image content
- Three categories of questions:
    1. **Visual:** Can be answered directly from the image: “Is there a dog in this image?”
    2. **Common sense:** Should not require an adult to consult an external source: “How many road vehicles are in this image?”
    3. **KB knowledge:** Requires an adult to use Wikipedia: “When was the home appliance in this image invented?”

~50%

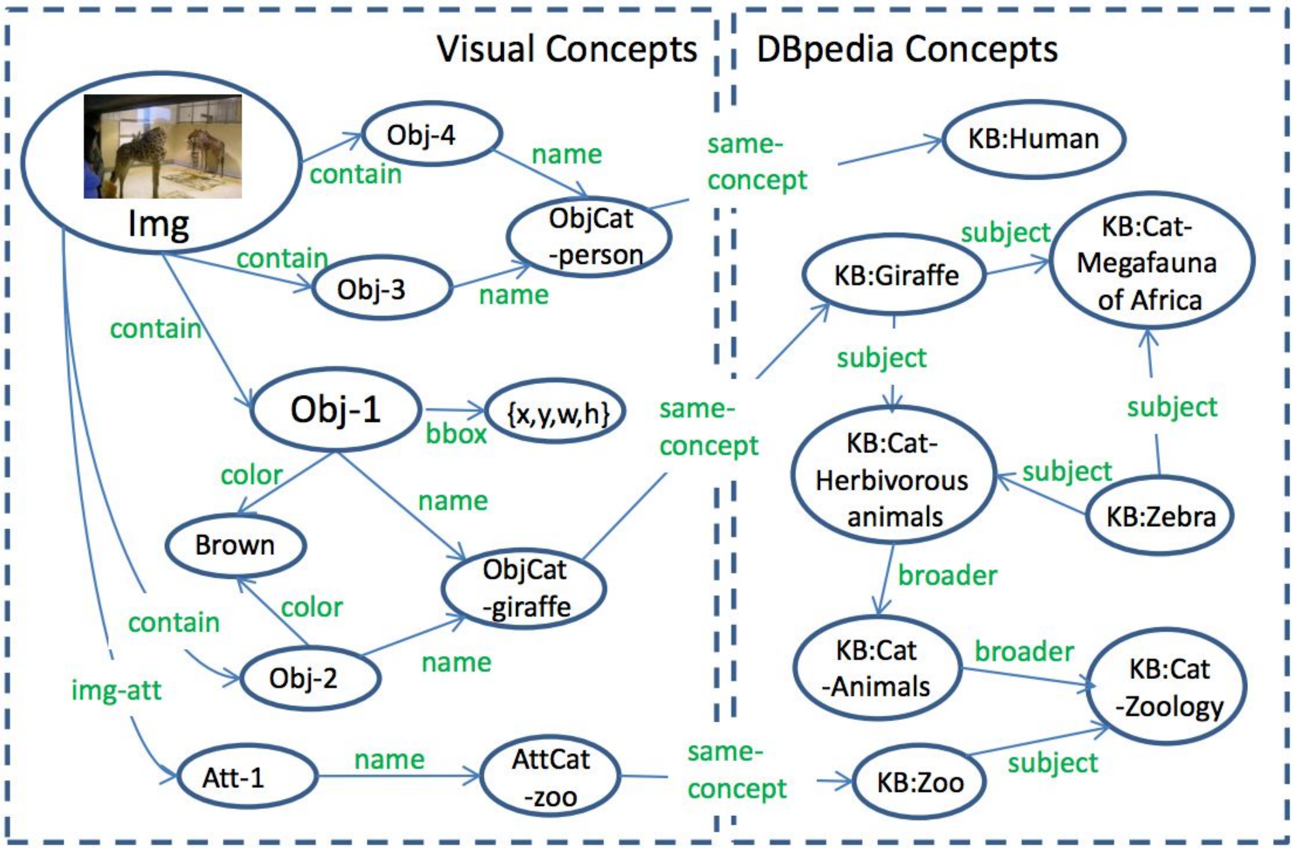
- Questions constructed by humans filling in *23 templates:*

AnimalClass	What is the ⟨*taxonomy*⟩ of the ⟨*animal*⟩?	46
LocIntro	Where was the ⟨*obj*⟩ invented?	40
YearIntro	When was the ⟨*obj*⟩ introduced?	32
FoodIngredient	List the ingredient of the ⟨food⟩.	31
LargestObj	What is the largest/smallest ⟨*concept*⟩?	27

“Explicit Knowledge-Based Reasoning for VQA.” Wang et al. https://arxiv.org/pdf/1511.02570.pdf. 2015.

Ahab Method: RDF Graph Construction

- Detect concepts in the image and *link them* to the KB
- Resulting RDF graph includes image contents + info from DBpedia



"Explicit Knowledge-Based Reasoning for VQA." Wang et al. https://arxiv.org/pdf/1511.02570.pdf. 2015.

Ahab Method: Parsing Questions

Input NLQ	What is the common properties between the animal on the right side of this image and the zebra?
Parsing	• Template: *CommProp.* • *<obj> = "right animal"; <concept> = "zebra".*
Mapping	• "right animal" → `Obj1-1`→`Obj-girraffe`→`KB:Giraffe` • "zebra" → `KB:Zebra`
Query	`?x:((KB:Giraffe, subject/?broader, ?x) AND` `(KB:Zebra, subject/?broader, ?x))`
Answer &Reason	Answer: Herbivorous animals, Animals, Magafauna of Africa Reason: Giraffe, Zebra → Herbivorous animals → animals; Giraffe, Zebra → Magafauna of Africa

"Explicit Knowledge-Based Reasoning for VQA." Wang et al. https://arxiv.org/pdf/1511.02570.pdf. 2015.

Ahab: Results for Question Types

- Ahab outperforms the baseline on all question types

Gap to reach human performance

Visual

Visual

Visual

Outperforms humans

KB

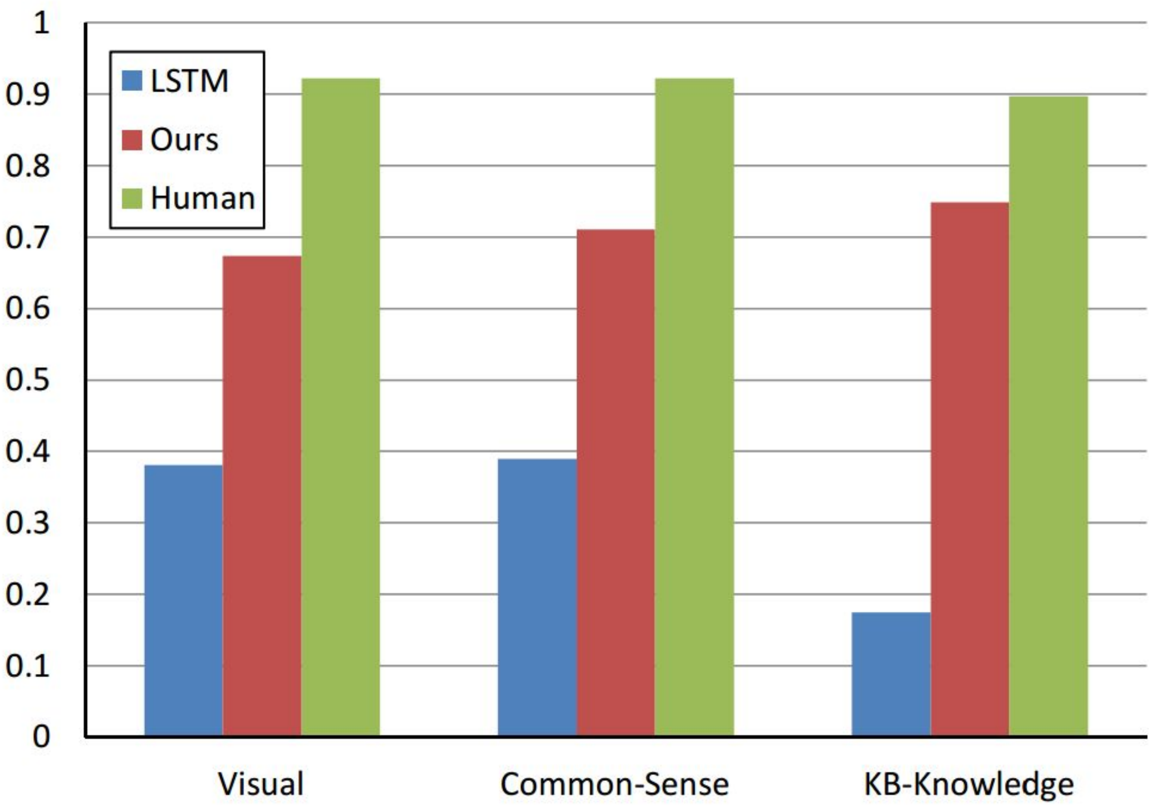
KB

Question Type	Accuracy(%) LSTM	Accuracy(%) Ours	Accuracy(%) Human	Correctness (Avg.) LSTM	Correctness (Avg.) Ours	Correctness (Avg.) Human
IsThereAny	64.9	86.9	93.6	3.6	4.5	4.7
IsImgRelate	57.0	82.2	97.1	3.3	4.2	4.9
WhatIs	26.9	66.9	94.5	2.1	3.7	4.8
ImgScene	30.4	69.6	85.9	2.3	3.8	4.5
ColorOf	14.6	29.8	93.2	1.7	2.5	4.7
HowMany	32.5	56.1	90.4	2.3	3.3	4.6
ObjAction	19.7	57.1	90.5	1.8	3.5	4.7
IsSameThing	54.9	77.5	91.5	3.2	4.2	4.6
MostRelObj	32.1	80.4	92.9	2.3	4.2	4.6
ListObj	1.9	63.0	100	1.1	3.6	4.8
IsTheA	74.5	80.4	92.2	3.9	4.2	4.7
SportEquip	2.1	70.8	79.2	1.2	3.9	4.2
AnimalClass	0.0	87.0	95.7	1.0	4.5	4.8
LocIntro	2.5	67.5	95.0	1.1	3.6	4.8
YearIntro	0.0	46.9	93.8	1.0	2.9	4.8
FoodIngredient	0.0	58.1	74.2	1.0	3.4	4.3
LargestObj	0.0	66.7	96.3	1.0	3.8	4.8
AreAllThe	29.6	63.0	81.5	2.3	3.7	4.3
CommProp	0.0	76.9	76.9	1.0	4.1	4.2
AnimalRelative	0.0	88.2	76.5	1.1	4.4	4.1
AnimalSame	41.2	70.6	94.1	2.6	3.8	4.8
FirstIntro	25.0	25.0	75.0	2.0	1.5	4.1
ListSameYear	25.0	75.0	50.0	1.8	4.2	3.0
Overall	36.2	69.6	92.0	2.5	3.8	4.7

"Explicit Knowledge-Based Reasoning for VQA." Wang et al. https://arxiv.org/pdf/1511.02570.pdf. 2015.

Ahab: Evaluation

- Human-based evaluation (because questions are open-ended, especially KB-knowledge questions)
- Compared to a baseline CNN-LSTM model



- The performance gap between Ahab and the LSTM *increases for questions requiring external knowledge*

"Explicit Knowledge-Based Reasoning for VQA." Wang et al. https://arxiv.org/pdf/1511.02570.pdf. 2015.

Ahab: Qualitative Results

Q1: Which object in this image is most related to entertainment?
A1: TV.
R1: Television → Performing Arts → Entertainment.

Q4: How many road vehicles in this image?
A4: Three.
R4: There are two trucks and one car.

Q3: Is there any tropical fruit?
A3: Yes.
R3: Banana → Tropical fruit.

Q6:List close relatives of the animal. **A6**: Donkey, horse, mule, asinus, hinny, wild ass, kiang .etc

Q2: Is the image related to sleep?
A2: Yes.
R2: Attribute-bedroom → sleep; Object-bed → sleep.

Q5: Tell me the ingredient of the food in the image.
A5: Meat, bread, vegetable, sauce, cheese, spread.

Q: Is there any root vegetable?
A: True . **R**: Carrot--> Category: Root vegetables.

Q: Are all the vehicles in this image wheeled vehicles?
A: True
R: Found objects: car, motorcycle.

"Explicit Knowledge-Based Reasoning for VQA." Wang et al. https://arxiv.org/pdf/1511.02570.pdf. 2015.

Ahab: Discussion

- \+ Capable of reasoning about the content of images and interactively answering a wide range of questions about them
- \+ Uses a structured representation of the image content, and relevant info about the world from a large external KB
- \+ Capable of explaining its reasoning in terms of entities in the KB, and the connections between them
- \+ Ahab can be used with any KB for which a SPARQL interface is available

- \- Relies heavily on pre-defined question templates
  - \- Special query handling for each predicate, e.g. `CommProp`, `IsImgRelate`, etc.
- \- Uses only one KB (DBpedia)

"Explicit Knowledge-Based Reasoning for VQA." Wang et al. https://arxiv.org/pdf/1511.02570.pdf. 2015.

Fact-based Visual Question Answering (FVQA)

Question: What can the red object on the ground be used for ?
Answer: Firefighting
Support Fact: Fire hydrant can be used for fighting fires.

- *Recognition:* red object = fire hydrant
- *Required Knowledge:* a fire hydrant can be used for fighting fires
- FVQA dataset contains *supporting facts* for each example:

<FireHydrant, CapableOf, FightingFire>

What you need to know to answer the question

“FVQA: Fact-based Visual Question Answering.” Wang et al. https://arxiv.org/pdf/1606.05433.pdf. 2016.

Fact-based Visual Question Answering (FVQA)

- Current datasets have focused on questions which are answerable by direct analysis of the question and image alone.
- The FVQA dataset which requires much deeper reasoning.
- FVQA contains questions that require external information to answer.
- Extend conventional VQA dataset with supporting facts, represented as triplets, such as **<Cat, CapableOf, ClimbingTrees>**.

Differences from Ahab

- The FVQA model ***learns*** a mapping from questions to KB queries
  - By *classifying* questions into categories, and extracting parts
- **Uses 3 KBs:** DBpedia, ConceptNet, WebChild

"FVQA: Fact-based Visual Question Answering." Wang et al. https://arxiv.org/pdf/1606.05433.pdf. 2016.

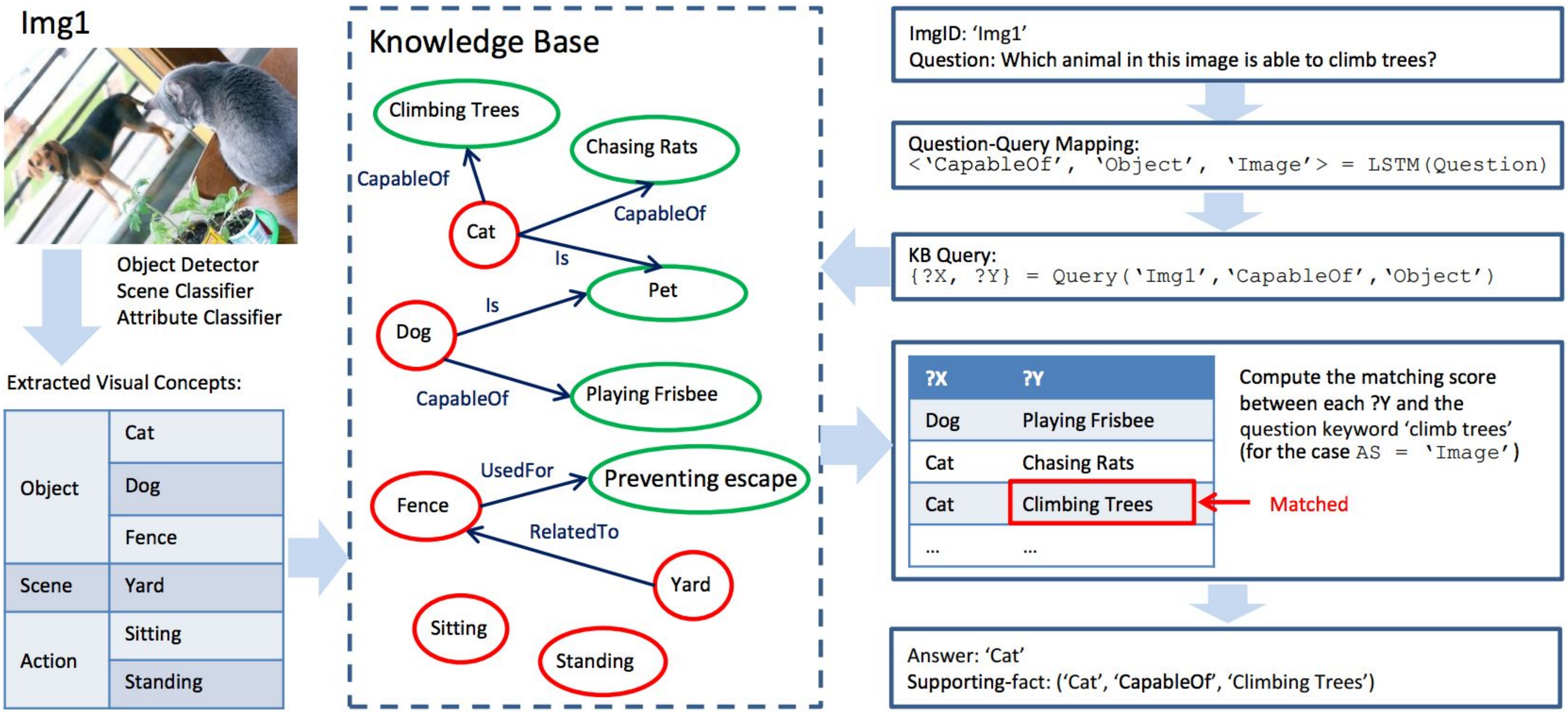
FVQA Ground-Truth Examples

	Which object in this image is able to stop cars	Can you name the beer that we usually enjoy with the fruit in the image?	Which instrument in this image is usually used in polka music	Why do they need a bow tie?
GT Fact:	Traffic light can stop cars	Lemon is related to corona	Accordions are used in polka music	Bow ties are worn at formal events
Ground Truth:	Traffic light	Corona	Accordion	Formal events

	Whether this animal runs slower or faster than horse?	What drink is made with this fruit?	Whether the game is a summer or winter Olympic?	How many times you should use this stuff per day?
GT Fact:	Camel are slower than horse	Grenadine is related to pomegranates	Balance beam belongs to the category of Summer Olympic disciplines	A toothbrush should be used twice a day
Ground Truth:	Slower	Grenadine	Summer Olympic disciplines	Used twice a day

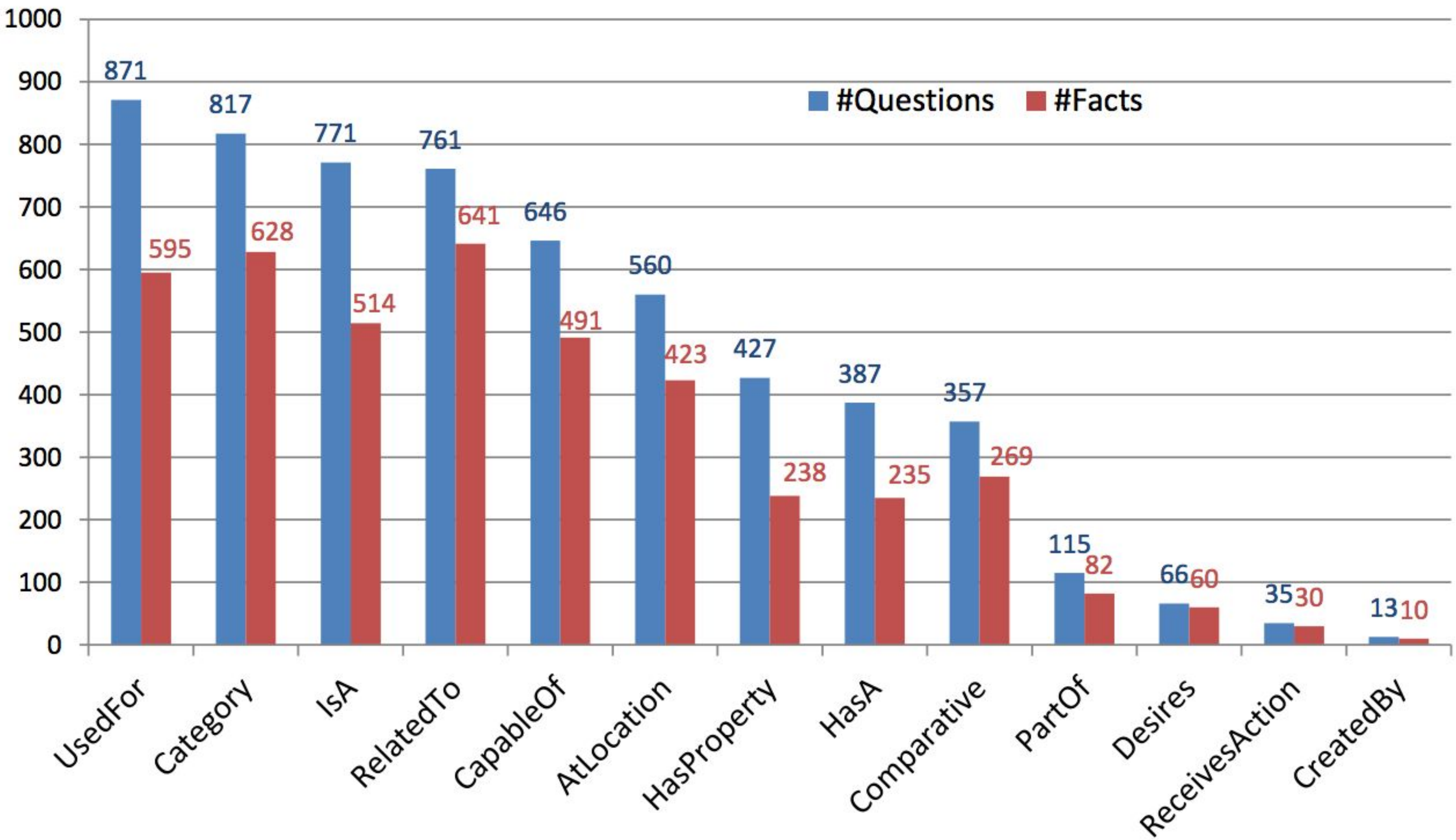
"FVQA: Fact-based Visual Question Answering." Wang et al. https://arxiv.org/pdf/1606.05433.pdf. 2016.

FVQA Architecture



"FVQA: Fact-based Visual Question Answering." Wang et al. https://arxiv.org/pdf/1606.05433.pdf. 2016.

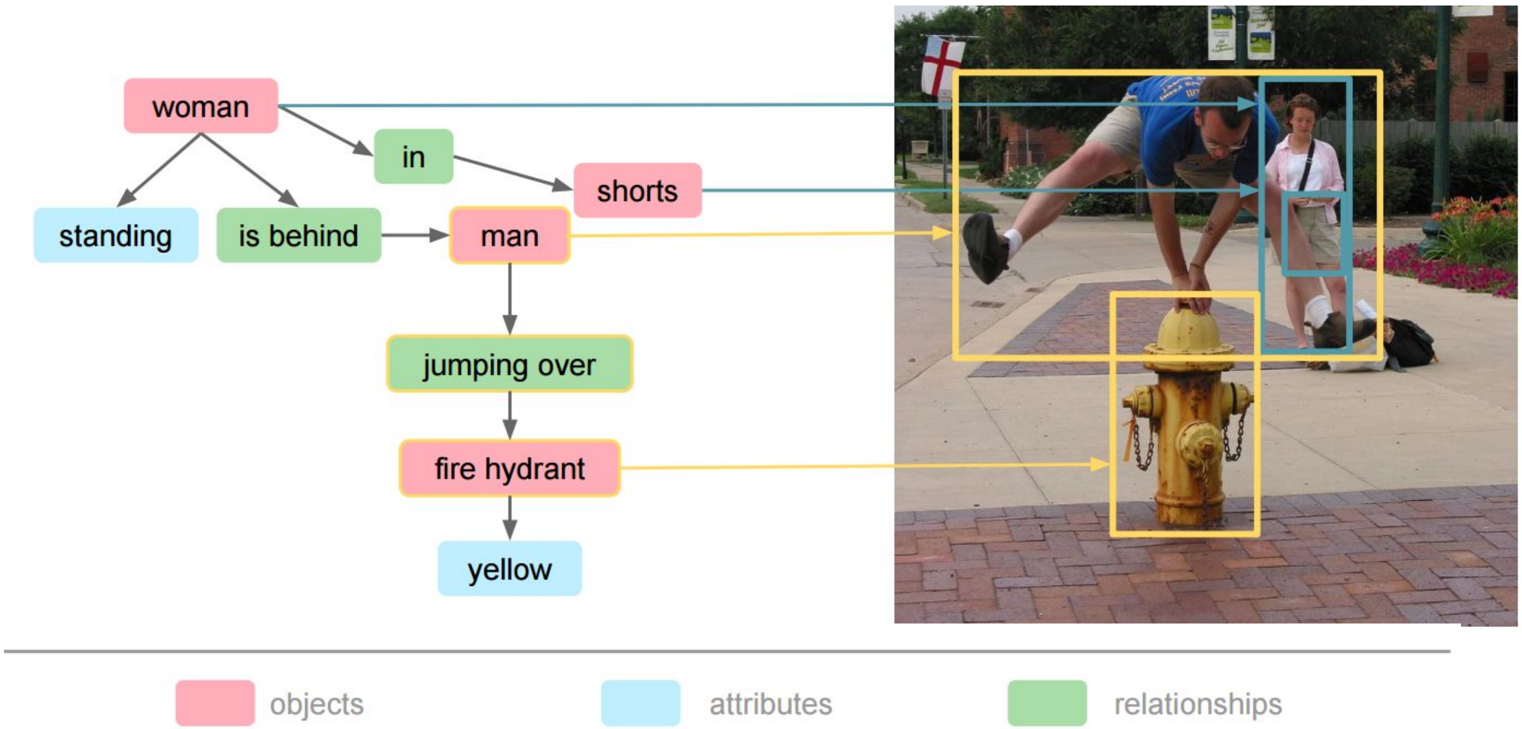
FVQA Predicates



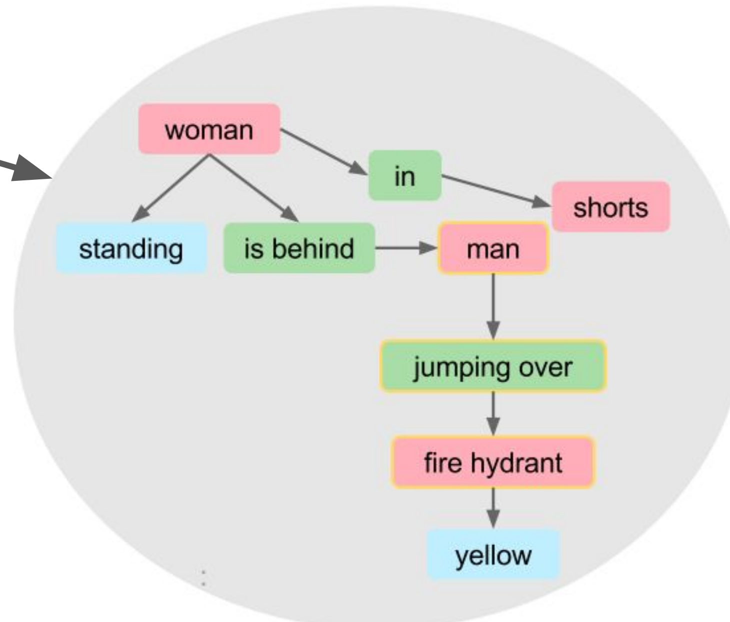
"FVQA: Fact-based Visual Question Answering." Wang et al. https://arxiv.org/pdf/1606.05433.pdf. 2016.

Graphs as Knowledge Bases and Structured Image Representations

- **Graphs** are a powerful way to express relationships between concepts
- **Knowledge graphs** capture *general world knowledge*
- **Scene graphs** capture the semantic content of *an image*
- Can combine these to integrate image-level details with general knowledge



Scene Graphs + KBs



Q: What color is the man's shirt?

A: Blue

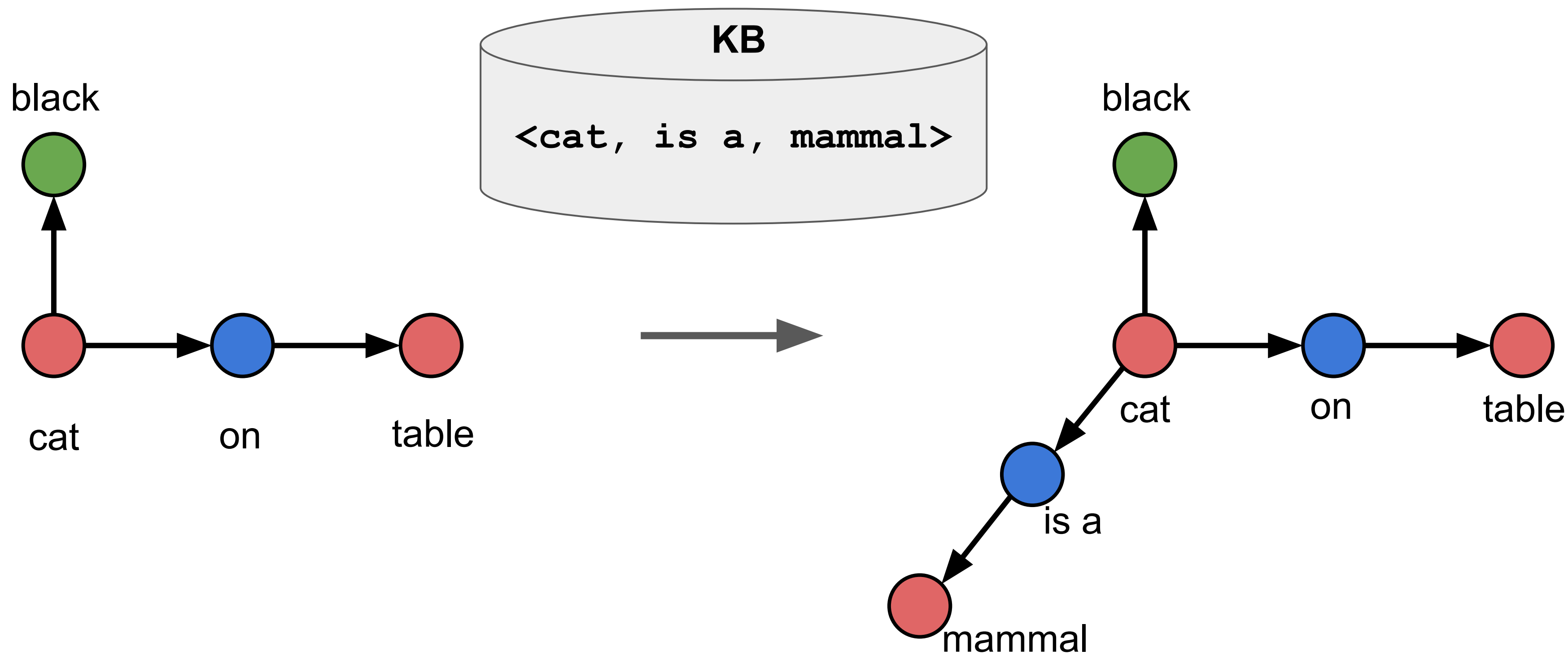
Blue $\in$?

woman
man
shorts
fire
hydrant
stand
yellow
in
is
behind
jump
over

"Visual Question Answering: A Survey of Methods and Datasets." Wu et al. https://arxiv.org/pdf/1607.05910.pdf. 2016.

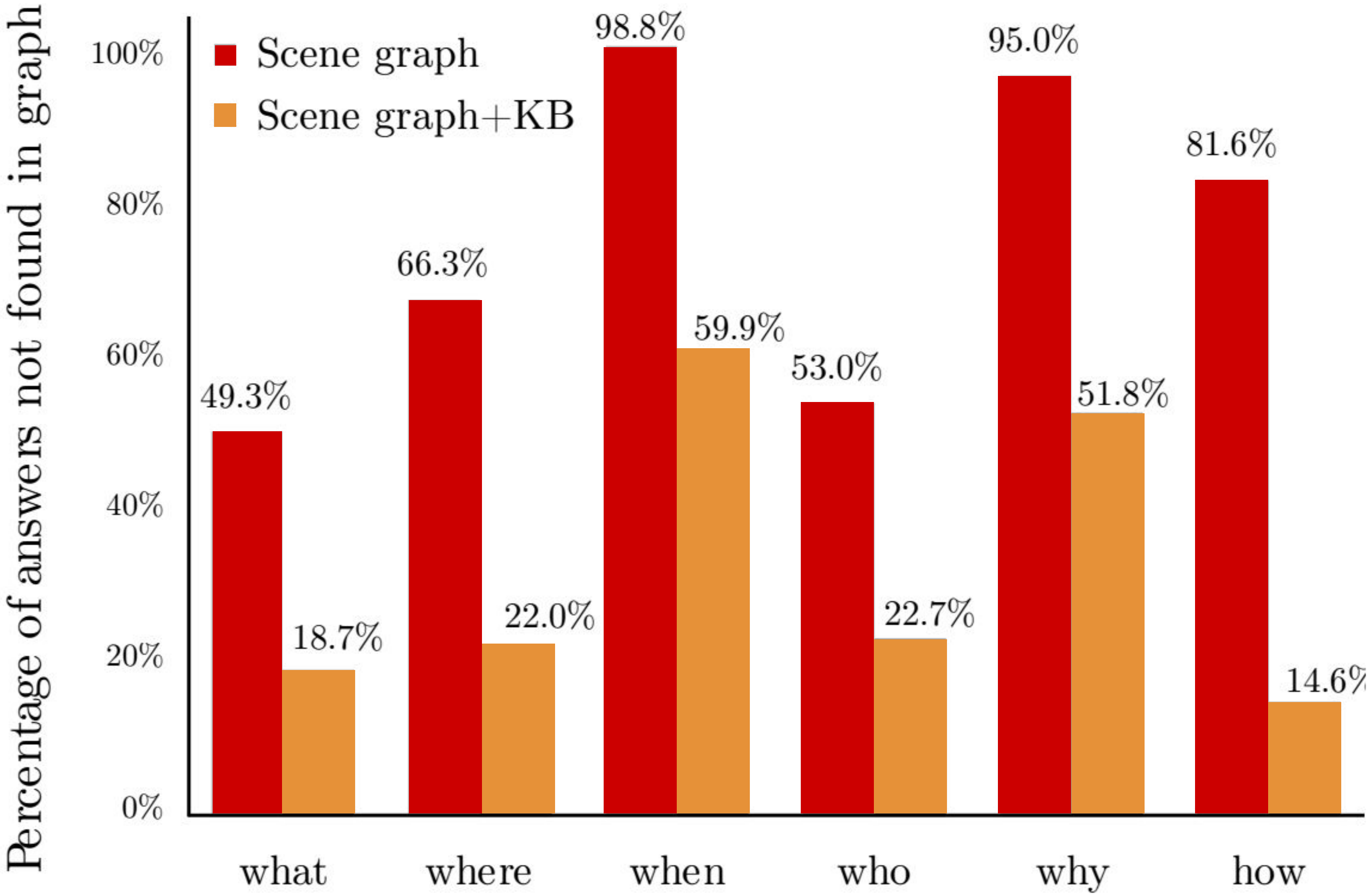
Scene Graphs + KBs

- The information in KBs is complementary to *annotations* in scene graphs
- Can *combine* visual datasets (Visual Genome) with large-scale KBs that provide *commonsense* information about visual and non-visual concepts



Scene Graphs + KBs: Results

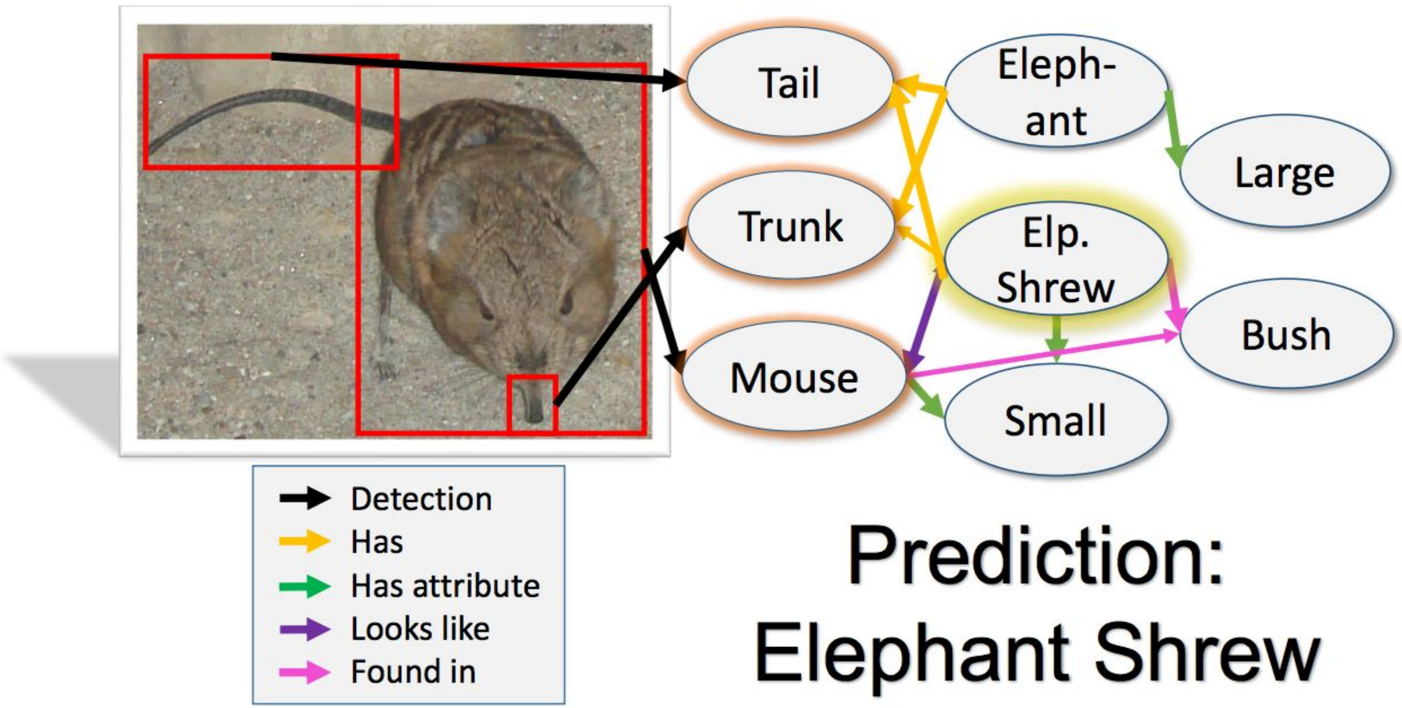
- Effect of graph completion on Visual Genome questions:



"Visual Question Answering: A Survey of Methods and Datasets." Wu et al. https://arxiv.org/pdf/1607.05910.pdf. 2016.

The More You Know: Using Knowledge Graphs for Image Classification

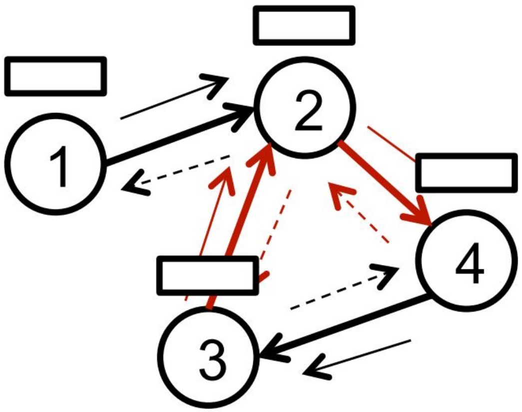
- Can exploit structured visual knowledge to improve image classification
- Humans use *knowledge of the world* to recognize unfamiliar objects
  - Gained from experience and language



"The More You Know: Using Knowledge Graphs for Image Classification." Marino et al. https://arxiv.org/pdf/1612.04844.pdf. 2016.

The More You Know: Graph Search Neural Network

- A *Graph Neural Network (GNN)* takes an arbitrary graph as input
- It propagates information between nodes to predict node-level outputs



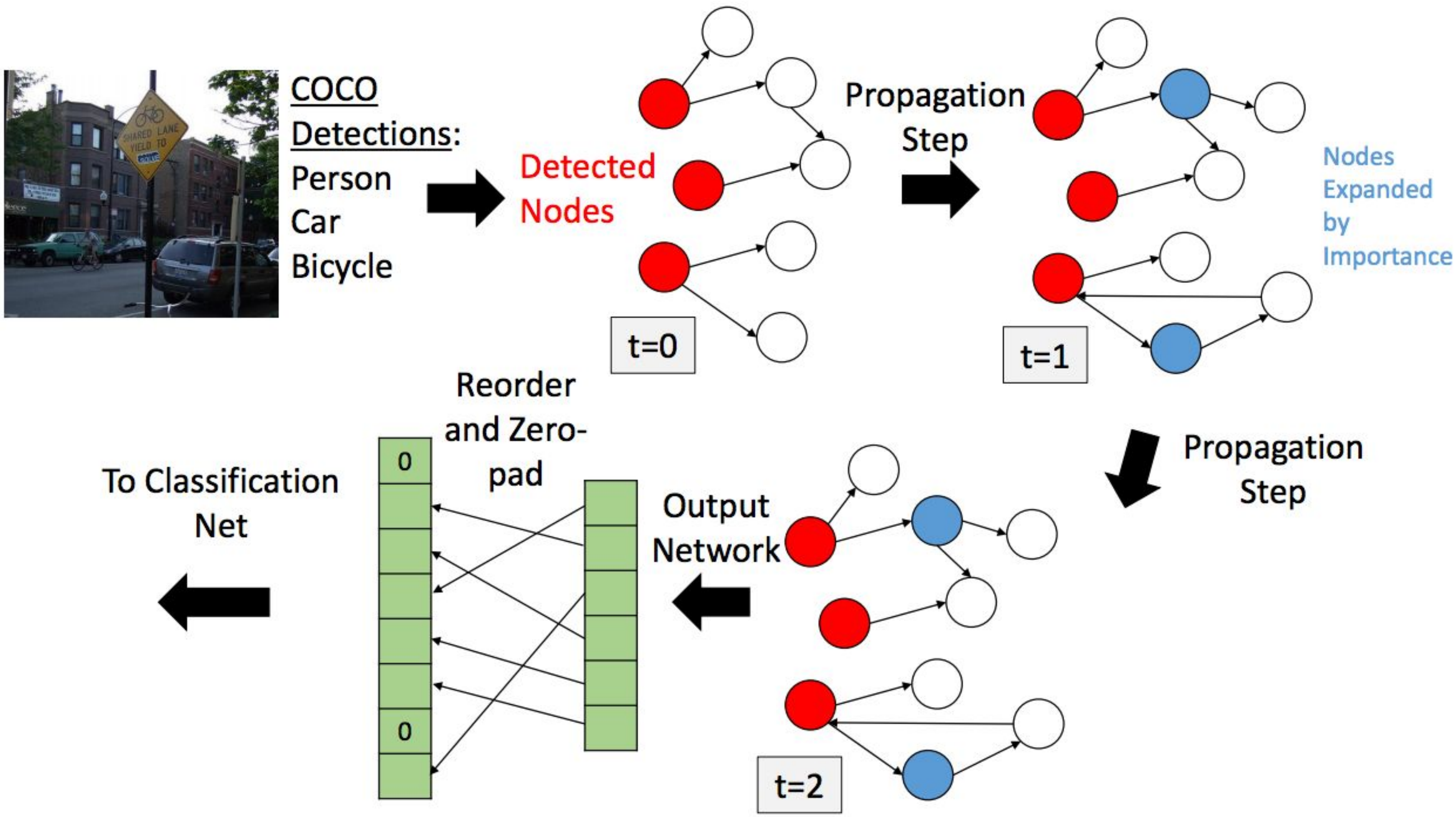
- Updates the representations for each node in a graph at each time step
- Not *scalable* to graphs with thousands of nodes → knowledge graphs

- **Solution**: *Graph Search Neural Network (GSNN)* selects a subset of the graph in each time step, and updates representations only for those nodes
  - A way to *efficiently* use knowledge graphs to improve image classification

"The More You Know: Using Knowledge Graphs for Image Classification." Marino et al. https://arxiv.org/pdf/1612.04844.pdf. 2016.

The More You Know: Graph Search Neural Network

1. Initial nodes are selected based on objects detected in the image
2. In each step, neighbours of the nodes in the current graph are added to the graph.
3. After a few expansion steps the node outputs are used in a classification pipeline



"The More You Know: Using Knowledge Graphs for Image Classification." Marino et al. https://arxiv.org/pdf/1612.04844.pdf. 2016.

The More You Know: Visual Genome Knowledge Graph

- Visual Genome contains over 100,000 images with many categories that fall in the long tail, each labelled with a scene graph

- Visual Genome knowledge graph → commonsense visual knowledge mined from common relationships in scene
  - Examples: grass is green; people can wear hats
- VG does not contain *semantic relationships* between concepts
  - e.g., a *dog* is an *animal*
- To incorporate hypernym information, VG graphs are *fused* with a WordNet graph representing a hierarchy of concepts

"The More You Know: Using Knowledge Graphs for Image Classification." Marino et al. https://arxiv.org/pdf/1612.04844.pdf. 2016.

The More You Know: Evaluation

- Mean Average Precision for multi-label classification on the VGFS dataset

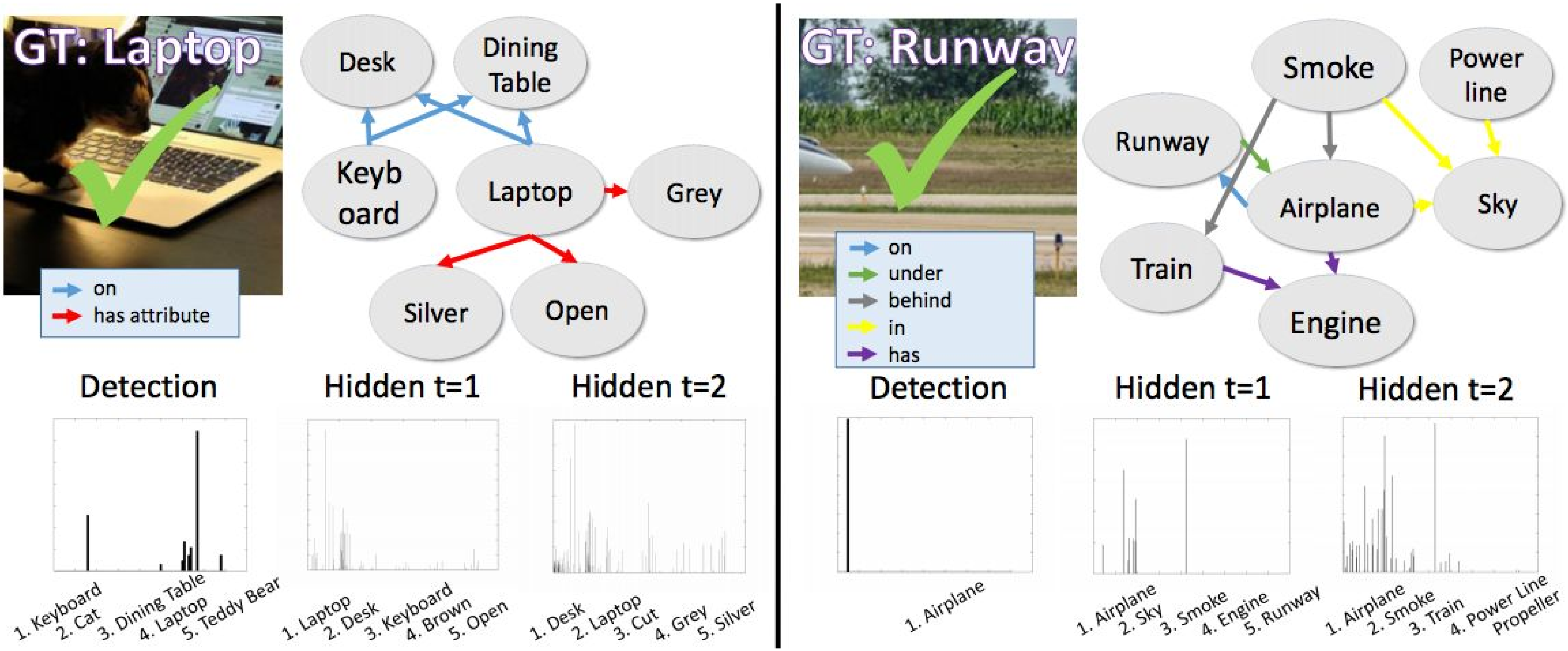
Method	1-shot	5-shot
VGG	5.96	8.07
VGG+Det	4.77	9.09
GSNN-VG	6.60	9.85
GSNN-VG+WN	**7.30**	**11.11**

- Mean Average Precision for multi-label classification on VGML using 500, 1,000, 5,000 training examples and the full training dataset

Method	500	1k	5k	full
VGG	10.28	12.09	18.58	19.15
VGG+Det	10.54	13.05	22.42	23.52
GSNN-VG	11.63	15.19	21.51	23.49
GSNN-VG+WN	**11.74**	**16.14**	**24.75**	**26.70**

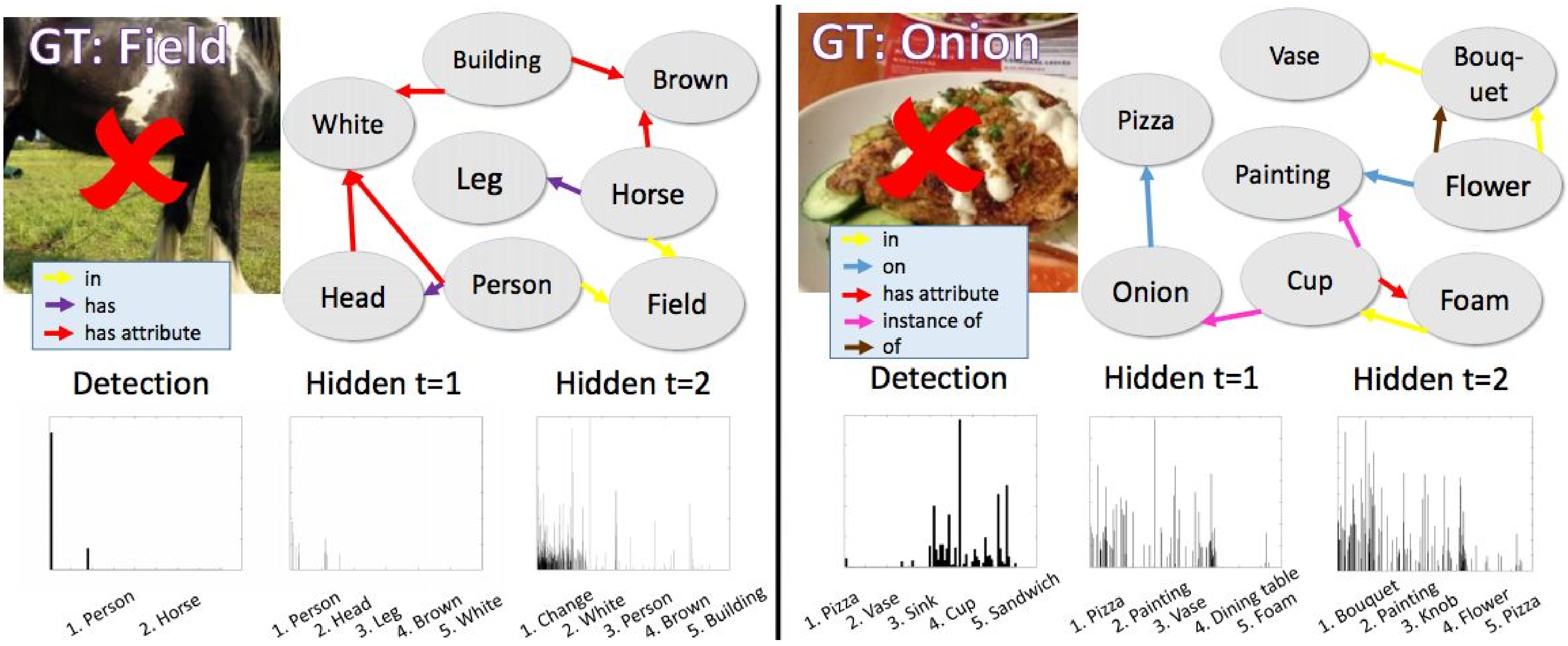
"The More You Know: Using Knowledge Graphs for Image Classification." Marino et al. https://arxiv.org/pdf/1612.04844.pdf. 2016.

The More You Know: Success Cases



"The More You Know: Using Knowledge Graphs for Image Classification." Marino et al. https://arxiv.org/pdf/1612.04844.pdf. 2016.

The More You Know: Failure Cases



"The More You Know: Using Knowledge Graphs for Image Classification." Marino et al. https://arxiv.org/pdf/1612.04844.pdf. 2016.

VQA Methods

Method	Knowledge Based	Explicit Reasoning	Structured Knowledge	Number of KBs
CNN-LSTM	×	×	×	0
Ask Me Anything	✓	×	×	1
Ahab	✓	✓	✓	1
FVQA	✓	✓	✓	3

Thank you!