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* \((\text{arg1}, \text{rel}, \text{arg2})\):

```
KB:

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

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

```
Result: cat, dog, horse, pig, ...
```

List of all domesticated animals
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.

**Knowledge Bases in Computer Vision**

- **High-level tasks**
  - VQA
  - Enable reasoning with external knowledge to answer complicated questions that go beyond what is visible in an image.

- **Low-level tasks**
  - Image Classification, etc.
  - Enable using knowledge about the world to identify objects based on their properties or relationships with other objects.
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:

```
Question: What's in the background?

Image

Variable-length sentence generation
A snow covered mountain range
```
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*)

<table>
<thead>
<tr>
<th>Method</th>
<th>Knowledge Based</th>
<th>Explicit Reasoning</th>
<th>Structured Knowledge</th>
<th>Number of KBs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-LSTM</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>0</td>
</tr>
<tr>
<td>Ask Me Anything</td>
<td>✓</td>
<td>❌</td>
<td>❌</td>
<td>1</td>
</tr>
<tr>
<td>Ahab</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
</tr>
<tr>
<td>FVQA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3</td>
</tr>
</tbody>
</table>
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

**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.
- 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 $V_{\text{att}}(I)$
Based on the attribute vector $V_{att}(I)$, generate five different captions

- The captions constitute the \textit{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: 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.

● **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 → \( V_{\text{know}}(I) \)
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: 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: 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

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

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

### Ask Me Anything: Evaluation

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Our-Baseline VggNet LSTM</th>
<th>Our Proposal Att LSTM</th>
<th>Our Proposal + Cap LSTM</th>
<th>Our Proposal + Att+Know LSTM</th>
<th>Our Proposal + Cap+Know LSTM</th>
<th>Our Proposal + A+C+K LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>what is</td>
<td>21.41</td>
<td>34.63</td>
<td>42.21</td>
<td>37.11</td>
<td>35.58</td>
<td>42.52</td>
</tr>
<tr>
<td>what colour</td>
<td>29.96</td>
<td>39.07</td>
<td>48.65</td>
<td>39.68</td>
<td>40.62</td>
<td>48.86</td>
</tr>
<tr>
<td>what kind</td>
<td>24.15</td>
<td>41.22</td>
<td>47.93</td>
<td>46.16</td>
<td>44.04</td>
<td>48.05</td>
</tr>
<tr>
<td>what are</td>
<td>23.05</td>
<td>38.87</td>
<td>47.13</td>
<td>41.13</td>
<td>39.73</td>
<td>47.21</td>
</tr>
<tr>
<td>what type</td>
<td>26.36</td>
<td>41.71</td>
<td>47.98</td>
<td>44.91</td>
<td>44.95</td>
<td>48.11</td>
</tr>
<tr>
<td>is the</td>
<td>71.49</td>
<td>73.22</td>
<td>74.63</td>
<td>74.40</td>
<td>73.78</td>
<td>74.70</td>
</tr>
<tr>
<td>is this</td>
<td>73.00</td>
<td>75.26</td>
<td>76.08</td>
<td>76.56</td>
<td>74.18</td>
<td>76.14</td>
</tr>
<tr>
<td>how many</td>
<td>34.42</td>
<td>39.14</td>
<td>46.61</td>
<td>39.78</td>
<td>44.20</td>
<td>47.38</td>
</tr>
<tr>
<td>are</td>
<td>73.51</td>
<td>75.14</td>
<td>76.01</td>
<td>75.75</td>
<td>75.78</td>
<td>76.14</td>
</tr>
<tr>
<td>does</td>
<td>76.51</td>
<td>76.71</td>
<td>78.07</td>
<td>76.55</td>
<td>77.17</td>
<td>78.11</td>
</tr>
<tr>
<td>where</td>
<td>10.54</td>
<td>21.42</td>
<td>25.92</td>
<td>24.13</td>
<td>16.09</td>
<td>26.00</td>
</tr>
<tr>
<td>is there</td>
<td>86.66</td>
<td>87.10</td>
<td>86.82</td>
<td>85.87</td>
<td>85.26</td>
<td>87.01</td>
</tr>
<tr>
<td>which</td>
<td>31.28</td>
<td>36.60</td>
<td>39.55</td>
<td>37.71</td>
<td>37.86</td>
<td>38.70</td>
</tr>
<tr>
<td>do</td>
<td>76.44</td>
<td>75.76</td>
<td>78.18</td>
<td>75.25</td>
<td>74.91</td>
<td>78.42</td>
</tr>
<tr>
<td>what does</td>
<td>15.45</td>
<td>19.33</td>
<td>21.80</td>
<td>19.50</td>
<td>19.04</td>
<td>22.16</td>
</tr>
<tr>
<td>what time</td>
<td>13.11</td>
<td>15.34</td>
<td>15.44</td>
<td>15.47</td>
<td>15.04</td>
<td>15.34</td>
</tr>
<tr>
<td>who</td>
<td>17.07</td>
<td>22.56</td>
<td>25.71</td>
<td>21.23</td>
<td>22.86</td>
<td>25.74</td>
</tr>
<tr>
<td>what sport</td>
<td>65.65</td>
<td>91.02</td>
<td>93.96</td>
<td>90.86</td>
<td>91.75</td>
<td>94.20</td>
</tr>
<tr>
<td>what animal</td>
<td>27.77</td>
<td>61.39</td>
<td>70.65</td>
<td>63.91</td>
<td>63.26</td>
<td>71.70</td>
</tr>
<tr>
<td>what brand</td>
<td>26.73</td>
<td>32.25</td>
<td>33.78</td>
<td>32.44</td>
<td>31.30</td>
<td>34.60</td>
</tr>
<tr>
<td>others</td>
<td>44.37</td>
<td>50.23</td>
<td>53.29</td>
<td>52.11</td>
<td>51.20</td>
<td>53.45</td>
</tr>
<tr>
<td>Overall</td>
<td>44.93</td>
<td>51.60</td>
<td>55.04</td>
<td>53.79</td>
<td>52.31</td>
<td>55.96</td>
</tr>
<tr>
<td></td>
<td>Test-dev</td>
<td></td>
<td>Test-standard</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>----------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>Y/N</td>
<td>Num</td>
<td>Others</td>
<td>All</td>
<td>Y/N</td>
</tr>
<tr>
<td>Question [1]</td>
<td>40.09</td>
<td>75.66</td>
<td>36.70</td>
<td>27.14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Image [1]</td>
<td>28.13</td>
<td>64.01</td>
<td>0.42</td>
<td>3.77</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q+I [1]</td>
<td>52.64</td>
<td>75.55</td>
<td>33.67</td>
<td>37.37</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LSTM Q [1]</td>
<td>48.76</td>
<td>78.20</td>
<td>35.68</td>
<td>26.59</td>
<td>48.89</td>
<td>78.12</td>
</tr>
<tr>
<td>LSTM Q+I [1]</td>
<td>53.74</td>
<td>78.94</td>
<td>35.24</td>
<td>36.42</td>
<td>54.06</td>
<td>79.01</td>
</tr>
<tr>
<td>Human [1]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>83.30</td>
<td>95.77</td>
</tr>
<tr>
<td>Ours</td>
<td>59.17</td>
<td>81.01</td>
<td>38.42</td>
<td>45.23</td>
<td>59.44</td>
<td>81.07</td>
</tr>
</tbody>
</table>

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.

- “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: Qualitative Results

<table>
<thead>
<tr>
<th>Question</th>
<th>Ours</th>
<th>Vgg+LSTM</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>What color is the tablecloth?</td>
<td>white</td>
<td>red</td>
<td>white</td>
</tr>
<tr>
<td>How many people in the photo?</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>What is the red fruit?</td>
<td>apple</td>
<td>banana</td>
<td>apple</td>
</tr>
<tr>
<td>What are these people doing?</td>
<td>eating</td>
<td>playing</td>
<td>eating</td>
</tr>
<tr>
<td>Why are his hands outstretched?</td>
<td>balance</td>
<td>play</td>
<td>balance</td>
</tr>
<tr>
<td>Why are the zebras in water?</td>
<td>drinking</td>
<td>water</td>
<td>drinking</td>
</tr>
<tr>
<td>Is the dog standing or laying down?</td>
<td>laying down</td>
<td>sitting</td>
<td>laying down</td>
</tr>
<tr>
<td>Which sport is this?</td>
<td>baseball</td>
<td>tennis</td>
<td>baseball</td>
</tr>
</tbody>
</table>

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

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

**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.

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?”
- Questions constructed by humans filling in 23 templates:

  - AnimalClass: What is the \( \text{taxonomy} \) of the \( \text{animal} \)?
  - LocIntro: Where was the \( \text{obj} \) invented?
  - YearIntro: When was the \( \text{obj} \) introduced?
  - FoodIngredient: List the ingredient of the \( \text{food} \).
  - LargestObj: What is the largest/smallest \( \text{concept} \)?

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

```
Visual Concepts

Obj-1
- Brown
  - contain
  - color
- ObjCat-giraffe
  - bbox
  - name
  - same-concept

Obj-2
- Att-1
  - name

Obj-3
- ObjCat-person
  - name

Obj-4
- contain

DBpedia Concepts

KB:Human
- KB:Cat-Megafauna of Africa
- subject

KB:Giraffe
- subject

KB:Cat-Herbivorous animals
- subject

KB:Cat-Animals
- subject

KB:Zebra
- KB:Cat-Zoology
- subject

KB:Zoo
- subject

{\{x,y,w,h\}}
- name

same-concept

```

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-giraffe→KB:Giraffe
- “zebra” → KB:Zebra

Query:
```
?x:((KB:Giraffe, subject/?broader, ?x) AND
   (KB:Zebra, subject/?broader, ?x))
```

Answer & Reason:
- Answer: Herbivorous animals, Animals, Maga fauna of Africa
- Reason: Giraffe $\rightarrow$ Herbivorous animals $\rightarrow$ animals
  Zebra $\rightarrow$ Maga fauna of Africa

Ahab: Results for Question Types

- Ahab outperforms the baseline on all question types

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Accuracy (%)</th>
<th>Correctness (Avg.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LSTM</td>
<td>Ours</td>
</tr>
<tr>
<td>IsThereAny</td>
<td>64.9</td>
<td>86.9</td>
</tr>
<tr>
<td>IsImgRelate</td>
<td>57.0</td>
<td>82.2</td>
</tr>
<tr>
<td>WhatIs</td>
<td>26.9</td>
<td>66.9</td>
</tr>
<tr>
<td>ImgScene</td>
<td>30.4</td>
<td>69.6</td>
</tr>
<tr>
<td>ColorOf</td>
<td>14.6</td>
<td>29.8</td>
</tr>
<tr>
<td>HowMany</td>
<td>32.5</td>
<td>56.1</td>
</tr>
<tr>
<td>ObjAction</td>
<td>19.7</td>
<td>57.1</td>
</tr>
<tr>
<td>IsSameThing</td>
<td>54.9</td>
<td>77.5</td>
</tr>
<tr>
<td>MostRelObj</td>
<td>32.1</td>
<td>80.4</td>
</tr>
<tr>
<td>ListObj</td>
<td>1.9</td>
<td>63.0</td>
</tr>
<tr>
<td>IsTheA</td>
<td>74.5</td>
<td>80.4</td>
</tr>
<tr>
<td>SportEquip</td>
<td>2.1</td>
<td>70.8</td>
</tr>
<tr>
<td>AnimalClass</td>
<td>0.0</td>
<td>87.0</td>
</tr>
<tr>
<td>LocIntro</td>
<td>2.5</td>
<td>67.5</td>
</tr>
<tr>
<td>YearIntro</td>
<td>0.0</td>
<td>46.9</td>
</tr>
<tr>
<td>FoodIngredient</td>
<td>0.0</td>
<td>58.1</td>
</tr>
<tr>
<td>LargestObj</td>
<td>0.0</td>
<td>66.7</td>
</tr>
<tr>
<td>AreAllThe</td>
<td>29.6</td>
<td>63.0</td>
</tr>
<tr>
<td>CommProp</td>
<td>0.0</td>
<td>76.9</td>
</tr>
<tr>
<td>AnimalRelative</td>
<td>0.0</td>
<td>88.2</td>
</tr>
<tr>
<td>AnimalSame</td>
<td>41.2</td>
<td>70.6</td>
</tr>
<tr>
<td>FirstIntro</td>
<td>25.0</td>
<td>25.0</td>
</tr>
<tr>
<td>ListSameYear</td>
<td>25.0</td>
<td>75.0</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td><strong>36.2</strong></td>
<td><strong>69.6</strong></td>
</tr>
</tbody>
</table>

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

Ahab: Qualitative Results

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

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

Q3: Is there any tropical fruit?  
A3: Yes.  
R3: Banana $\rightarrow$ Tropical fruit.

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

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

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

Q: Is there any root vegetable?  
A: True.  
R: Carrot $\rightarrow$ Category: Root vegetables.

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

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)

Fact-based Visual Question Answering (FVQA)

- **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

**Question**: What can the red object on the ground be used for?

**Answer**: Firefighting

**Support Fact**: Fire hydrant can be used for fighting fires.
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 \(<\text{Cat, CapableOf, ClimbingTrees}>\).

**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 Ground-Truth Examples

**Which object in this image is able to stop cars?**
- **GT Fact:** Traffic light can stop cars
- **Ground Truth:** Traffic light

**Can you name the beer that we usually enjoy with the fruit in the image?**
- **GT Fact:** Lemon is related to corona
- **Ground Truth:** Corona

**Which instrument in this image is usually used in polka music?**
- **GT Fact:** Accordion are used in polka music
- **Ground Truth:** Accordion

**Why do they need a bow tie?**
- **GT Fact:** Bow ties are worn at formal events
- **Ground Truth:** Formal events

**Whether this animal runs slower or faster than horse?**
- **GT Fact:** Camel are slower than horse
- **Ground Truth:** Slower

**What drink is made with this fruit?**
- **GT Fact:** Grenadine is related to pomegranates
- **Ground Truth:** Grenadine

**Whether the game is a summer or winter Olympic?**
- **GT Fact:** Balance beam belongs to the category of Summer Olympic disciplines
- **Ground Truth:** Summer Olympic disciplines

**How many times you should use this stuff per day?**
- **GT Fact:** A toothbrush should be used twice a day
- **Ground Truth:** Used twice a day

FVQA Architecture

FVQA Predicates

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.

![Diagram of relationships](image_url)
Scene Graphs + KBs

Q: What color is the man’s shirt?
A: Blue

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:

```
<table>
<thead>
<tr>
<th>Question</th>
<th>Scene Graph</th>
<th>Scene Graph + KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>what</td>
<td>49.3%</td>
<td>18.7%</td>
</tr>
<tr>
<td>where</td>
<td>66.3%</td>
<td>22.0%</td>
</tr>
<tr>
<td>when</td>
<td>98.8%</td>
<td>59.9%</td>
</tr>
<tr>
<td>who</td>
<td>53.0%</td>
<td>22.7%</td>
</tr>
<tr>
<td>why</td>
<td>95.0%</td>
<td>51.8%</td>
</tr>
<tr>
<td>how</td>
<td>81.6%</td>
<td>14.6%</td>
</tr>
</tbody>
</table>
```

Can exploit **structured visual knowledge** to improve image classification

Humans use *knowledge of the world* to recognize unfamiliar objects

- Gained from experience and language

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.

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: 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 hyponym information, VG graphs are fused with a WordNet graph representing a hierarchy of concepts

The More You Know: Evaluation

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

<table>
<thead>
<tr>
<th>Method</th>
<th>1-shot</th>
<th>5-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG</td>
<td>5.96</td>
<td>8.07</td>
</tr>
<tr>
<td>VGG+Det</td>
<td>4.77</td>
<td>9.09</td>
</tr>
<tr>
<td>GSNN-VG</td>
<td>6.60</td>
<td>9.85</td>
</tr>
<tr>
<td>GSNN-VG+WN</td>
<td>7.30</td>
<td>11.11</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>500</th>
<th>1k</th>
<th>5k</th>
<th>full</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG</td>
<td>10.28</td>
<td>12.09</td>
<td>18.58</td>
<td>19.15</td>
</tr>
<tr>
<td>VGG+Det</td>
<td>10.54</td>
<td>13.05</td>
<td>22.42</td>
<td>23.52</td>
</tr>
<tr>
<td>GSNN-VG</td>
<td>11.63</td>
<td>15.19</td>
<td>21.51</td>
<td>23.49</td>
</tr>
<tr>
<td>GSNN-VG+WN</td>
<td>11.74</td>
<td>16.14</td>
<td>24.75</td>
<td>26.70</td>
</tr>
</tbody>
</table>


Marino et al. 2016.

The More You Know: Failure Cases

<table>
<thead>
<tr>
<th>Method</th>
<th>Knowledge Based</th>
<th>Explicit Reasoning</th>
<th>Structured Knowledge</th>
<th>Number of KBs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-LSTM</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>0</td>
</tr>
<tr>
<td>Ask Me Anything</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>1</td>
</tr>
<tr>
<td>Ahab</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
</tr>
<tr>
<td>FVQA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3</td>
</tr>
</tbody>
</table>
Thank you!