VisKe: Visual Knowledge Extraction and Question Answering by Visual Verification of Relation Phrases

Fereshteh Sadeghi, Santosh K. Divvala, Ali Farhadi

Presented by David Fouhey
Motivation

Radiator!  Beroe!
But what about?
But what about?
But what about?

1. Is it unusual?
2. Is it dangerous?
3. What happens next?
4. Does she know?
5. Where is it?
6. P(crocodile) vs. P(alligator)?
Common Answer – Write Them Down

Table from Chao, Wang, Mahalcea, Deng CVPR ‘15

O(n^3) for triplets

Tons of exceptions
Crocodilians are largely carnivorous, and the diets of different species can vary with snout shape and tooth sharpness. Species with sharp teeth and long slender snouts, like the Indian gharial and Australian freshwater crocodile, are specialised for feeding on fish, insects, and crustaceans, while extremely broad-snouted species with blunt teeth, like the Chinese alligator and broad-snouted caiman, specialise in eating hard-shelled molluscs. Species whose snouts and teeth are intermediate between these two forms, such as the saltwater crocodile and American alligator, have generalised diets and opportunistically feed on invertebrates, fish, amphibians, other reptiles, birds, and mammals.[11][75]
Common Answer – Read a book

<table>
<thead>
<tr>
<th>Google</th>
<th>black sheep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web</td>
<td>Images</td>
</tr>
<tr>
<td>About 41,700,000 results (0.35 seconds)</td>
<td></td>
</tr>
</tbody>
</table>

VS.

<table>
<thead>
<tr>
<th>Google</th>
<th>white sheep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web</td>
<td>Images</td>
</tr>
<tr>
<td>About 30,000,000 results (0.33 seconds)</td>
<td></td>
</tr>
</tbody>
</table>

![Baseball game](image1.png) ![Baseball game](image2.png)
A complementary idea: Use Vision
VisKE

Can we use web data to **verify** relationships like

- Bears fish salmon
- Cats sit in boxes
- Ducks lay eggs
- People sit on sofas

Formalized as scoring a relationship $R(S,O)$ via visual data, e.g., $Cat(Sit,Box)$
Related Web-supervised Approaches

  – Learns phrases (e.g., jumping horse) on web data.

• **NEIL: Extracting Visual Knowledge from Web Data.** Chen, Shrivastava, Gupta ICCV ’13
  – Semi-supervised learning on web data constrained by relationships

  – Original language-based common-sense gatherer
Idea

Hypothesize: Google N-Grams
Verify: Score using graphical model with potentials from detections on web data
Approach Overview

• Given: $R(S,O)$
• Learn visual models for $S,O,SV,VO,SVO$
• Run these models on new data
• Fit model to detection patterns
• Parameters of model tell relationship + fitting likelihood tells how good the model is
An Aside – Visual Phrases

Sadeghi, Farhadi, CVPR ‘11
Approach

\[ P(\mathcal{R}, S, O, SV, VO, SVO) \propto \prod_{x \in \{O, S, SV\}} \Phi(\mathcal{R}, SVO, x) \times \prod_{y \in \{SV, S\}} \Phi(\mathcal{R}, VO, y) \times \prod_{z \in \{S, O, SV, VO, SVO\}} \Psi(z), \quad (1) \]

\[ \Phi^i(\mathcal{R}, x, y) = \begin{cases} \max_\theta \mathcal{L}(x, y, \bar{I}; \theta) & \mathcal{R} \equiv i \\ 1 & \text{otherwise} \end{cases} \quad (2) \]
Pairwise Potential

\[ \Phi^i(\mathcal{R}, x, y) = \begin{cases} \max_{\theta} & \mathcal{L}(x, y, \bar{I}; \theta) \\ 1 & \mathcal{R} \equiv i \\ \text{otherwise} \end{cases} \]  \hspace{1cm} (2)

Each mutual detection turned into:

\[ \{dx, dy, ov, ov_1, ov_2, h_1, w_1, h_2, w_2, a_1, a_2\} \]

Fit a diagonal Gaussian to detections over data

Compute total likelihood of detection relationships
Concerns?

1. What are problems with verification?
2. Web supervised
3. The rare vs the impossible / generalization: do alligators eat raccoons?
4. Intro talks about grounding the objects/verbs. Has it achieved this?
Results
Q: What do alligators eat?
A: Solve \( \text{argmax}_y \ P(\text{Alligator, Eat, } y) \)
Results – Analogies

\[
\{dx, dy, ov, ov_1, ov_2, h_1, w_1, h_2, w_2, a_1, a_2\} \{dx, dy, ov, ov_1, ov_2, h_1, w_1, h_2, w_2, a_1, a_2\}^T
\]
Quantitative Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Base Set</th>
<th>Permute Set</th>
<th>Combined Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual Phrase [32]</td>
<td>49.67</td>
<td>14.12</td>
<td>42.49</td>
</tr>
<tr>
<td>Co-detection Model</td>
<td>49.24</td>
<td>14.65</td>
<td>43.14</td>
</tr>
<tr>
<td>Google Ngram Model [1]</td>
<td>46.17</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Language Model [22]</td>
<td>56.20</td>
<td><strong>22.68</strong></td>
<td>50.23</td>
</tr>
<tr>
<td>VisKE</td>
<td><strong>62.11</strong></td>
<td>20.93</td>
<td><strong>54.67</strong></td>
</tr>
</tbody>
</table>

Table 1. Results (M.A.P.) on the Relation Phrase Dataset. While

<table>
<thead>
<tr>
<th>Model</th>
<th>M.A.P.</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenIE [13]</td>
<td>73.03</td>
</tr>
<tr>
<td>Co-detection Model</td>
<td>76.65</td>
</tr>
<tr>
<td>Visual Phrase [32]</td>
<td>78.45</td>
</tr>
<tr>
<td>Language Model [22]</td>
<td>83.65</td>
</tr>
<tr>
<td>VisKE</td>
<td>85.80</td>
</tr>
</tbody>
</table>

Table 3. Results on the OpenIE dataset.
What’s Missing?