A Multi-World Approach to Question Answering about Real-World Scenes

Mateusz Malinowski, Mario Fritz
Outline

1. Goal
2. Dataset
3. Performance Measure
4. Technical Approach
Motivation

• “full scene understanding”
  • semantic segmentation
  • image captioning

• Q & A is the most complete
Full Scene Understanding?

Semantic Segmentation

Image Captioning

a car parked outside of a grassy field
Goal

To answer natural-language queries about images

Question: what is on the desk and behind the black cup?
Answer: bottle
Dataset
Dataset: Images

- 1449 RGB-D Images and pixel labels from NYU-Depth v2
- 894 object categories (!)
  - restricted to 37 for most evaluations

*Figure from Silberman et al, 2012*
Dataset: Q & A

Human Dataset:
• 12,000 Q&A pairs (~9 per image)
• questions unconstrained
• Each answer must be one of
  • a color
  • a number
  • a set of object categories - e.g. \{bed, couch\}

Synthetic Dataset:
• 420 Q&A pairs
• generated from templates
• answers can also be
  • scene types - e.g. bedroom
  • sets of images
Question: how many plastic toy containers are below the table in front of the wall?

Answer: 6
Question: what is on the desk?

Answer: {desk_mat, paper, book, napkin_dispenser}
**Question:** what color are the paper trays in the bookshelf on the left side of the wall divider not on the desk in front of the computer chair?

**Answer:** black
Performance Measure
Performance Measure

Difficulties:
• near-synonyms, e.g. couch vs futon
• hypernyms, e.g. person vs woman vs skateboarder
• comparing sets, e.g. \{pillow, book\} vs \{pillow\}
Performance Measure

Difficulties:
• near-synonyms
• hypernyms
• comparing sets

WUP score → WUPS score

Fuzzy ⊆
What’s wrong with WUP?

\[
WUP(a, b) = \frac{2 \ast \text{depth}(lca(a, b))}{\text{depth}(a) + \text{depth}(b)} \in [0, 1]
\]

\[
WUP(\text{lamp}, \text{table}) = \frac{2 \ast \text{depth}(\text{furniture})}{\text{depth}(\text{lamp}) + \text{depth}(\text{table})} = 0.88
\]

\[
WUP(\text{couch}, \text{futon}) = 0.52
\]
What about distributed representations?

- generalize to multi-word answers like “red jacket” and “female tennis player”
- usually trained on huge text corpora, with no visual information
Asymmetry

**Question:** Who is holding the racquet?
**GT Answer:** female tennis player
**Answer:** person

\[d(person, \text{female tennis player}) \sim 0\]

**Question:** Who is speaking?
**GT Answer:** person
**Answer:** female tennis player

\[d(\text{female tennis player}, person) \sim \infty\]
Performance Measure

Needs to:

• include visual similarity

• be asymmetric
Technical Approach
Two sources of uncertainty

1. Vision: what is in the image?

2. Language: what does the question ask?

Figure from “Perceptual Organization and Recognition of Indoor Scenes from RGB-D images”, Gupta et al, 2013

Figure from “Learning Dependency-Based Compositional Semantics”, Liang et al, 2013
Graphical Model

\[ P(A|Q, S) = \sum_{W} \sum_{T} P(A|T, W) P(T|Q) P(W|S) \]
Representation

What is a world?

- Facts, e.g. \(\text{chair}(\text{segment, color, } X_{\{\min, \text{mean, max}\}}, Y_{\{\min, \text{mean, max}\}}, Z_{\{\min, \text{mean, max}\}})\)

- Relations, e.g. \(\text{above}(A, B)\), \(\text{inFront}(A, B)\)

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{closeAbove}(A, B))</td>
<td>(\text{above}(A, B)) and ((Y_{\min}(B) &lt; Y_{\max}(A) + \epsilon))</td>
</tr>
<tr>
<td>(\text{closeLeftOf}(A, B))</td>
<td>(\text{leftOf}(A, B)) and ((X_{\min}(B) &lt; X_{\max}(A) + \epsilon))</td>
</tr>
<tr>
<td>(\text{closeInFrontOf}(A, B))</td>
<td>(\text{inFrontOf}(A, B)) and ((Z_{\min}(B) &lt; Z_{\max}(A) + \epsilon))</td>
</tr>
<tr>
<td>(X_{\text{aux}}(A, B))</td>
<td>(X_{\text{mean}}(A) &lt; X_{\max}(B)) and (X_{\min}(B) &lt; X_{\text{mean}}(A))</td>
</tr>
<tr>
<td>(Z_{\text{aux}}(A, B))</td>
<td>(Z_{\text{mean}}(A) &lt; Z_{\max}(B)) and (Z_{\min}(B) &lt; Z_{\text{mean}}(A))</td>
</tr>
<tr>
<td>(h_{\text{aux}}(A, B))</td>
<td>(\text{closeAbove}(A, B)) or (\text{closeBelow}(A, B))</td>
</tr>
<tr>
<td>(v_{\text{aux}}(A, B))</td>
<td>(\text{closeLeftOf}(A, B)) or (\text{closeRightOf}(A, B))</td>
</tr>
<tr>
<td>(d_{\text{aux}}(A, B))</td>
<td>(\text{closeInFrontOf}(A, B)) or (\text{closeBehind}(A, B))</td>
</tr>
<tr>
<td>(\text{leftOf}(A, B))</td>
<td>(X_{\text{mean}}(A) &lt; X_{\text{mean}}(B))</td>
</tr>
<tr>
<td>(\text{above}(A, B))</td>
<td>(Y_{\text{mean}}(A) &lt; Y_{\text{mean}}(B))</td>
</tr>
<tr>
<td>(\text{inFrontOf}(A, B))</td>
<td>(Z_{\text{mean}}(A) &lt; Z_{\text{mean}}(B))</td>
</tr>
<tr>
<td>(\text{on}(A, B))</td>
<td>(\text{closeAbove}(A, B)) and (Z_{\text{aux}}(A, B)) and (X_{\text{aux}}(A, B))</td>
</tr>
<tr>
<td>(\text{close}(A, B))</td>
<td>(h_{\text{aux}}(A, B)) or (v_{\text{aux}}(A, B)) or (d_{\text{aux}}(A, B))</td>
</tr>
</tbody>
</table>

Figure from Malinowski and Fritz, 2014
Simplifying Assumptions

\[ P(A|Q, S) = \sum_{W} \sum_{T} P(A|T, W) P(T|Q) P(W|S) \]

1.  \( P(T|Q) \propto \exp(w^T \phi(T, Q)) \)

2.  \( P(W|S) = P(s_1 = c_f(1), \ldots, s_n = c_f(n)) = \prod_i P(s_i = c_f(i)) \)

3.  Sample 25 possible worlds

4.  \( P(A|T, W) \sim 3\text{-}\text{nearest neighbour “batch” approximation} \)
Results
Results

<table>
<thead>
<tr>
<th>Segmentation</th>
<th>World(s)</th>
<th>#classes</th>
<th>Accuracy</th>
<th>WUPS at 0.9</th>
<th>WUPS at 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanSeg</td>
<td>Single</td>
<td>894</td>
<td>7.86%</td>
<td>11.86%</td>
<td>38.79%</td>
</tr>
<tr>
<td>HumanSeg</td>
<td>Single</td>
<td>37</td>
<td>12.47%</td>
<td>16.49%</td>
<td>50.28%</td>
</tr>
<tr>
<td>AutoSeg</td>
<td>Single</td>
<td>37</td>
<td>9.69%</td>
<td>14.73%</td>
<td>48.57%</td>
</tr>
<tr>
<td>AutoSeg</td>
<td>Multi</td>
<td>37</td>
<td>12.73%</td>
<td>18.10%</td>
<td>51.47%</td>
</tr>
<tr>
<td>Human Baseline</td>
<td>Single</td>
<td>894</td>
<td>50.20%</td>
<td>50.82%</td>
<td>67.27%</td>
</tr>
<tr>
<td>Human Baseline</td>
<td>Multi</td>
<td>37</td>
<td>60.27%</td>
<td>61.04%</td>
<td>78.96%</td>
</tr>
</tbody>
</table>

My Baseline

“how many” —> 2

“what color” —> white

else ————> {table}

17.85%
Error Analysis - Language

Q: How many red chairs are there?
H: 0
M: 6
C: blinds

Q: How many chairs are at the table?
H: wall
M: 4
C: chair
## Error Analysis - Vision

### synthetic question-answer pairs (SynthQA)

<table>
<thead>
<tr>
<th>Segmentation</th>
<th>World(s)</th>
<th># classes</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanSeg</td>
<td>Single with Neg. 3</td>
<td>37</td>
<td>56.0%</td>
</tr>
<tr>
<td>HumanSeg</td>
<td>Single</td>
<td>37</td>
<td>59.5%</td>
</tr>
<tr>
<td>AutoSeg</td>
<td>Single</td>
<td>37</td>
<td>11.25%</td>
</tr>
<tr>
<td>AutoSeg</td>
<td>Multi</td>
<td>37</td>
<td><strong>13.75%</strong></td>
</tr>
</tbody>
</table>
Summary

- Very interesting, high-level vision problem
- Very difficult, large dataset
- Unclear performance measure
- Authors’ approach doesn’t work