CAP 6412 Advanced Computer Vision


Boqing Gong

Feb 25, 2016
Next week: Vision and language

Tuesday (03/01)
Javier Lores

& Secondary papers

Thursday (03/03)
Aisha Urooji

& Secondary papers
Project 1: Due on 02/28

• If you have discussed option 2 with me
  • Send me the meeting minutes / slides --- grading criteria

• If you take option 1
  • In total, >6,000 validation images
  • Test 3 images per class of the validation set
Travel plan

• At Washington DC on 03/01, Tuesday
• Guest lecture by Dr. Ulas Bagci
Today

• Administrivia
• Recurrent Neural Networks (RNNs) (I)
• VQA, by Nandakishore
Why RNN? (1)

- Feed-forward networks
  - Model static input-output concept
    - No time series
  - Exists a single forward direction

- CNN

- Recurrent neural networks
  - Model dynamic state transition
    - Time & sequence data
  - Exists feedback connections

- LSTM, GRU
Why RNN? (2)

• Markov models
  • Model dynamic state transition
    • Time & sequence data
  • Markov (short-range) dependency
  • Moderately sized states

• Recurrent neural networks
  • Model dynamic state transition
    • Time & sequence data
  • Long-range dependency
  • Exponentially expressive states
Why RNN? (3)

- Image caption generator

Image credits: Karpathy, Andrej, and Li Fei-Fei
Why RNN? (3)

• Image caption generator
• Sentence embedding

Why RNN? (3)

- Image caption generator
- Sentence embedding
- Word embedding

Image credits: Karpathy, Andrej, and Li Fei-Fei
Why RNN? (3)

• Image caption generator
• Sentence embedding
• Word embedding
• Activity recognition

Why RNN? (3)

• Image caption generator
• Sentence embedding
• Word embedding
• Activity recognition
• Video representation

Why RNN? (3)

- Image caption generator
- Sentence embedding
- Word embedding
- Activity recognition
- Video representation
- VQA
- Early action detection
- Human dynamics
- Scene labeling

- Language modeling
- Machine translation
(Discrete-time) RNN

• The processing occurs in discrete steps

(Discrete-time) RNN

- The processing occurs in discrete steps

Image credits: Richard Socher
(Discrete-time) RNN

• At a single time step $t$:
(Discrete-time) RNN

- Three time steps and beyond

Image credits: Richard Socher
(Discrete-time) RNN

- Three time steps and beyond

In math:

\[ h_t = \sigma \left( W^{(hh)} h_{t-1} + W^{(hx)} x_t \right) \]

\[ \hat{y}_t = \text{softmax} \left( W^{(S)} h_t \right) \]

Image credits: Richard Socher
(Discrete-time) RNN

- Three time steps and beyond
- A layered feedforward net
- Tied weights for different time steps
- Conditioning (memorizing?) on all previous input
- Memory being cheap to save in RAM
(Discrete-time) RNN

- Compare with holistically nested network

**Difference:**
- Tied weights for RCNN
- #layers not fixed for RCNN
- Input at every layer for RCNN
- Less RAM for RCNN

Image credits: Richard Socher, Saining Xie, and Zhuowen Tu
Today

• Administrivia
• Recurrent Neural Networks (RNNs) (I)
• VQA, by Nandakishore
Upload slides before or after class

• See “Paper Presentation” on UCF webcourse

• Sharing your slides
  • Refer to the originals sources of images, figures, etc. in your slides
  • Convert them to a PDF file
  • Upload the PDF file to “Paper Presentation” after your presentation

Mateusz Malinowski and Mario Fritz
*Max Planck Institute for Informatics*
*Saarbrucken, Germany*

Presented by: Nandakishore Puttashamachar
Nandakishore@knights.ucf.edu
Presentation Outline

• Motivation
• Goal
• Background
• Approach
• Experiments
• Quantitative results
• Contribution
• Summary
Motivation

- Can machines answer on questions about images?
- Evaluating chain of perception, representation and deduction
- A holistic and open-ended test that resembles the famous Turing Test
Goal

To answer *natural-language* queries about images

**Question** - *what is on the desk and behind the black cup?*

**Answer** – *bottle*
What is the most populous city in California?

Los Angeles
What is the most populous city in California?

Database → System

Los Angeles
• How does the system interpret the question?
How does the system interpret the question?
Background

most populous city in California

city

populous  in

most  California

Linguistics
syntactic locality

Trees

Computation
efficient interpretation
Approach
• What is a world?

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

Relations - \(\text{above}(A, B), \text{inFront}(A, B)\)
Background

• Single-world
  • Generate a single perceived world $\mathcal{W}$ based on segmentation
  • Given a question $Q$ and a world $\mathcal{W}$, predict an answer $A$
  • Compute the posterior which marginalizes over the latent logical forms $\mathcal{T}$

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

$P(A|\mathcal{T}, \mathcal{W})$ - Denotation of a logical form $\mathcal{T}$ on the world $\mathcal{W}$

\[
P(A|\mathcal{T}, \mathcal{W}) = 1[A \in \sigma_{\mathcal{W}}(\mathcal{T})]
\]

$\sigma_{\mathcal{W}}$ - Evaluate a logical form on the world $\mathcal{W}$
Background

• Single-world

\[ \sigma_W(\mathcal{T}) := \bigcap_j^d \{v : v \in \sigma_W(p), t \in \sigma_W(\mathcal{T}_j), R_j(v, t) \} \]

**where**

\[ \mathcal{T} := \langle p, (\mathcal{T}_1, R_1), (\mathcal{T}_2, R_2), ..., (\mathcal{T}_d, R_d) \rangle \]

• Log-linear distribution over the logical forms

\[ P(\mathcal{T}|Q) \propto \exp(\theta^T \phi(Q, \mathcal{T})) \]

• Feature vector \( \phi \) measures compatibility between \( Q \) and \( \mathcal{T} \)

• Model parameters are learnt by searching over a restricted space of valid trees and with gradient descent updates of the parameter.

• Datalog inference produces answers from latent logical forms.
Approach

• Perceived-world
  
  • \textit{World} \( W \) now consists of “Facts” derived from automatic – semantic image Segmentation \( S \).
  
  • Semantic segmentation is used to collect information about objects such as
    
    • Object class
    • 3D position
    • Color

  • Object hypothesis is represented as n-tuple –

\[
predicate(instance\_id, image\_id, color, spatial\_loc)
\]

where, \( predicate \in \{ \text{table, bag, books, window,} \ldots \} \)

\( instance\_id \) – object’s id
\( image\_id \) – id of the image containing the object
\( spatial\_loc \) – \( min, max \) and mean location of the object along \( X, Y, Z \) axis.

\[
(X_{\text{min}}, X_{\text{max}}, X_{\text{mean}}, Y_{\text{min}}, Y_{\text{max}}, Y_{\text{mean}}, Z_{\text{min}}, Z_{\text{max}}, Z_{\text{mean}})
\]
## Approach

- **Perceived-world**
  - Predicates defining spatial relations between two objects -

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

• Multi-world
  • Previous approach did not consider the “uncertainty” in class labeling
  • Multiple interpretation of the scene/different perceived worlds

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

• Computationally efficient
• Probability of an answer can be estimated on different world independent of each other
Approach
Approach – Scalability and Implementation

- For worlds containing many facts and spatial relations implementation is computationally expensive.
- Batch-based approximation
- Every image induces a set of facts named batch of facts.
- For every test image find k nearest neighbors in the space of training batches.
- Use TFIDF to measure similarity.
- In short, build a training world from k images with similar content to the perceived world of test image.
Experiments

• DAtaset for QUestion Answering on Real-world images (DAQUAR)
  • Images and Semantic Segmentation –
    • NYU-Depth V2 dataset – 1449 RGBD images with annotated semantic segmentations.
    • Every pixel mapped into 40 classes out of which 37 informative object classes are used.

• New dataset of question and answers –
  • Two types of annotation – Synthetic and Human
  • Synthetic QA pairs – generated automatically based on these templates –

<table>
<thead>
<tr>
<th>Description</th>
<th>Template</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>counting</td>
<td>How many {object} are in {image_id}?</td>
<td>How many cabinets are in image1?</td>
</tr>
<tr>
<td>counting and colors</td>
<td>How many {color} {object} are in {image_id}?</td>
<td>How many gray cabinets are in image1?</td>
</tr>
<tr>
<td>room type</td>
<td>Which type of the room is depicted in {image_id}?</td>
<td>Which type of the room is depicted in image1?</td>
</tr>
<tr>
<td>superlatives</td>
<td>What is the largest {object} in {image_id}?</td>
<td>What is the largest object in image1?</td>
</tr>
<tr>
<td>set</td>
<td></td>
<td></td>
</tr>
<tr>
<td>counting and colors</td>
<td>How many {color} {object}?</td>
<td>How many black bags?</td>
</tr>
<tr>
<td>negations type 1</td>
<td>Which images do not have {object}?</td>
<td>Which images do not have sofa?</td>
</tr>
<tr>
<td>negations type 2</td>
<td>Which images are not {room_type}?</td>
<td>Which images are not bedroom?</td>
</tr>
<tr>
<td>negations type 3</td>
<td>Which images have {object} but do not have a {object}?</td>
<td>Which images have desk but do not have a lamp?</td>
</tr>
</tbody>
</table>
Experiments

NYU-Depth V2 dataset: image, Z axis, ground truth and predicted semantic segmentations.
Experiments

• DAtaset for QUestion Answering on Real-world images (DAQUAR)
  • Performance measure –
    • Standard Accuracy –
      \[ \frac{1}{N} \sum_{i=1}^{N} 1\{A^i = T^i\} \]

  \[ A^i, T^i \] are the i-th answer and ground truth respectively

  Set of objects - \[ T = \{t_1, t_2, \ldots\} \] and \[ i \in \{1, 2, \ldots, N\} \]

  \[ \text{WUPS}(A, T) = \frac{1}{N} \sum_{i=1}^{N} \min \left\{ \prod_{a \in A^i} \max_{t \in T^i} \text{WUP}(a, t), \prod_{t \in T^i} \max_{a \in A^i} \text{WUP}(a, t) \right\} \cdot 100 \]

  • WUPS is a soft measure to find the similarity between generated answer and the ground truth label
Quantitative Results

• How the Multi world approach will perform under uncertain segmentation and unknown logical forms.

<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>13.75%</td>
</tr>
</tbody>
</table>

• Severe drop in performance switching from human to automatic segmentation.
Quantitative Results

- Performance of the system with Human QA pairs (HumanQA)

<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>Single</td>
<td>37</td>
<td>60.27%</td>
<td>61.04%</td>
<td>78.96%</td>
</tr>
</tbody>
</table>

- Human annotations exhibit more variations in contrast to synthetic approach.
- Longer questions and includes more spatially related objects.
Quantitative Results

• WUPS score for different thresholds –
QA: (What is behind the table?, window)
Spatial relation like ‘behind’ are dependent on the reference frame. Here the annotator uses observer-centric view.

QA: (What is beneath the candle holder, decorative plate)
Some annotators use variations on spatial relations that are similar, e.g., ‘beneath’ is closely related to ‘below’.

QA: (What is in front of the wall divider?, cabinet)
Annotators use additional properties to clarify object references (i.e., wall divider). Moreover, the perspective plays an important role in these spatial relations interpretations.

Q: (What is behind the table?, sofa)
Spatial relations exhibit different reference frames. Some annotations use observer-centric, others object-centric view.
QA: (How many lights are on?, 6)
Moreover, some questions require detection of states ‘light on or off’.

QA: (What is in front of the curtain behind the armchair?, guitar)

QA2: (What is in front of the curtain?, guitar)
Spatial relations matter more in complex environments where reference resolution becomes more relevant. In cluttered scenes, pragmatism starts playing a more important role.

Q: What is at the back side of the sofas? Annontators use wide range spatial relations, such as ‘backside’ which is object-centric.

QA: (How many doors are in the image?, 5)
QA2: (How many doors are in the image?, 5)
Different interpretation of ‘door’ results in different counts: 1 door at the end of the hall vs. 5 doors including lockers.

QA: (How many drawers are there?, 8)
The annotators use their common-sense knowledge for amodal completion. Here the annotator infers the 8th drawer from the context.

Q: (How many doors are open?, 1)
Notion of states of object (like open) is not well captured by current vision techniques. Annotators use such attributes frequently for disambiguation.
Contributions

• An approach and a dataset of question answer pairs
• Combine language with perception in a multi-world Bayesian framework
The Big Picture
Thank you