COS 484: Natural Language Processing

Language Grounding

Fall 2019
Recap: Language representations

Contextualized Word Representations

- **ELMo** = Embeddings from **Language Models**

  Deep contextualized word representations
  https://arxiv.org/cs
  by ME Peters - 2018 - Cited by 1683 - Related articles
  Deep contextualized word representations. ... Our word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on a large text corpus.

- **BERT** = Bidirectional **Encoder** Representations from **Transformers**

  BERT: Pre-training of Deep Bidirectional Transformers for ...
  https://arxiv.org/cs
  by J Devlin - 2018 - Cited by 2259 - Related articles
  Oct 11, 2018 - Unlike recent language representation models, BERT is designed to pre-train deep ... As a result, the pre-trained BERT model can be fine-tuned with just one additional output ... Which authors of this paper are endorsers?
Symbol grounding problem

- Miller and Johnson-Laird (1976) — Language and Perception
- Harnad (1990) — Symbol grounding problem
  - How do we connect “symbols” to the world in the right way?

In a pure symbolic model the crucial connection between the symbols and their referents is missing; an autonomous symbol system, though amenable to a systematic semantic interpretation, is ungrounded. In a pure connectionist model, names are connected to objects through invariant patterns in their sensory projections, learned through exposure and feedback, but the crucial compositional property is missing; a network of names, though grounded, is not yet amenable to a full systematic semantic interpretation. In the hybrid system proposed here, there is no longer any autonomous symbolic level at all; instead, there is an intrinsically dedicated symbol system, its elementary symbols (names) connected to nonsymbolic representations that can pick out the objects to which they refer, via connectionist networks that extract the invariant features of their analog sensory projections.

- Neural networks (connectionism) help us connect symbolic reasoning to sensory inputs

(slide adapted from Greg Durrett)
Color test

- What color is this?
Grounding color

- Bayesian model for grounded color semantics
- 829 color descriptions

(McMahan and Stone, 2014)
Gricean maxims

- Cooperative, effective communication

- **Maxim of quantity:** Give as much information as need, and no more

- **Maxim of quality:** Provide truthful information, supported by evidence

- **Maxim of relation:** Be relevant, say things pertinent to discussion

- **Maxim of manner:** Be clear, brief and orderly, avoid obscurity and ambiguity
Types of grounding

- Perception
  - Visual: $green = [0,1,0]$ in RGB
  - Auditory: $loud = >120 \text{ dB}$
  - Taste: sweet = $>\text{some threshold level of sensation on taste buds}$
  - High-level concepts:

\[\text{cat} \quad \text{dog}\]
Types of grounding

- **Temporal concepts**
  - *late evening* = after 6pm
  - *fast, slow* = describing rates of change

- **Actions**
  - running
  - eating
Types of grounding

- Relations
  - Spatial: 
    - *left, on top of, in front of*
  - Functional: 
    - *Jacket:* keeps people warm
    - *Mug:* holds water
- Size: 
  - Whales are *larger* than lions
A chair
A chair

- green
- armless
- medium size
- used to sit on

light
frangible
plush

Context is very important!
Semantics does not exist in isolation

Language

- Logical forms
- Parse trees
- Vector representations

Perception

Interaction
Some grounding tasks

‣ Vision
  ▶ Captioning
  ▶ VQA
  ▶ Spatial reasoning

‣ Interaction
  ▶ Instruction following
  ▶ Text-based games
the girl is licking the spoon of batter

- Describe an image in a sentence
Image captioning

- Describe an image in a sentence
- Requires recognizing objects, attributes, relations in image
- Caption must be fluent

Applications?

(MS COCO, Chen et al., 2015)
Captioning as multi-modal translation

French → Text encoder (e.g. RNN) → Text decoder (e.g. RNN) → English

Image encoder → Text decoder (e.g. RNN) → English

(Donahue et al., 2015, Vinyals et al., 2015)
Convolutional Neural Networks

Left: A regular 3-layer Neural Network. Right: A ConvNet arranges its neurons in three dimensions (width, height, depth), as visualized in one of the layers. Every layer of a ConvNet transforms the 3D input volume to a 3D output volume of neuron activations. In this example, the red input layer holds the image, so its width and height would be the dimensions of the image, and the depth would be 3 (Red, Green, Blue channels).

CNN for image classification
Convolutional Neural Networks

CNN for image classification

(source: towardsdatascience.com)
Captioning with attention

Two men playing frisbee in a dark field.

(Anderson et al., 2018)
Captioning with attention

![Diagram of captioning with attention](Anderson et al., 2018)

<table>
<thead>
<tr>
<th></th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>c5</td>
<td>c40</td>
<td>c5</td>
<td>c40</td>
<td>c5</td>
<td>c40</td>
<td>c5</td>
<td>c40</td>
</tr>
<tr>
<td>Review Net [48]</td>
<td>72.0</td>
<td>90.0</td>
<td>55.0</td>
<td>81.2</td>
<td>41.4</td>
<td>70.5</td>
<td>31.3</td>
<td>59.7</td>
</tr>
<tr>
<td>Adaptive [27]</td>
<td>74.8</td>
<td>92.0</td>
<td>58.4</td>
<td>84.5</td>
<td>44.4</td>
<td>74.4</td>
<td>33.6</td>
<td>63.7</td>
</tr>
<tr>
<td>PG-BCM [24]</td>
<td>75.4</td>
<td>-</td>
<td>59.1</td>
<td>-</td>
<td>44.5</td>
<td>-</td>
<td>33.2</td>
<td>-</td>
</tr>
<tr>
<td>SCST:Att2all [34]</td>
<td>78.1</td>
<td>93.7</td>
<td>61.9</td>
<td>86.0</td>
<td>47.0</td>
<td>75.9</td>
<td>35.2</td>
<td>64.5</td>
</tr>
<tr>
<td>LSTM-A3 [49]</td>
<td>78.7</td>
<td>93.7</td>
<td>62.7</td>
<td>86.7</td>
<td>47.6</td>
<td>76.5</td>
<td>35.6</td>
<td>65.2</td>
</tr>
<tr>
<td>Ours: Up-Down</td>
<td><strong>80.2</strong></td>
<td><strong>95.2</strong></td>
<td><strong>64.1</strong></td>
<td><strong>88.8</strong></td>
<td><strong>49.1</strong></td>
<td><strong>79.4</strong></td>
<td><strong>36.9</strong></td>
<td><strong>68.5</strong></td>
</tr>
</tbody>
</table>

*(Anderson et al., 2018)*
Video captioning

An overview of the S2VT video to text architecture.

(Venugopalan et al., 2015)
Visual Question Answering

- Require *multi-modal* knowledge and reasoning

- Well-defined *evaluation metric* (accuracy)

(Agrawal et al., 2015)
Visual Question Answering

“How many horses are in this image?”

Any issues?

(Agrawal et al., 2015)
Better multimodal reasoning

FiLM
(Perez et al., 2017)

Neural module networks
(Andreas et al., 2016)
On deeper examination:

- Just using language is a pretty good prior!
- “Do you see a ..” = yes (87% of the time)
- “How many…” = 2 (39%)
- “What sport …” = tennis (41%)

Balanced VQA  
(Goyal et al., 2017)
Spatial Relations

Golland et al. (2010)

- How would you indicate O1 to someone with relation to the other two objects? (not calling it a vase, or describing its inherent properties)
- What about O2?
- Requires modeling listener — “right of O2” is insufficient though true
Spatial Reasoning

Put the Toyota block in the same row as the SRI block, in the first open space to the right of the SRI block
Move Toyota to the immediate right of SRI, evenly aligned and slightly separated
Move the Toyota block around the pile and place it just to the right of the SRI block
Place Toyota block just to the right of The SRI Block
Toyota, right side of SRI

Reach the cell above the westernmost rock

Robotic Manipulation

(Bisk et al., 2016, Misra et al., 2017)
Instruction Following

- Want to be able to follow instructions in a virtual environment
- “Go along the blue hall, then turn left away from the fish painting and walk to the end of the hallway”

(MacMahon et al., 2006)

(slide adapted from Greg Durrett)
Instruction Following

Instruction: “Go away from the lamp to the intersection of the red brick and wood”

Basic: Turn ( ),
       Travel ( steps: 1 )

Landmarks: Turn ( ),
           Verify ( left: WALL , back: LAMP , back: HATRACK , front: BRICK HALL ) ,
           Travel ( steps: 1 ) ,
           Verify ( side: WOOD HALL )

- Train semantic parser on (utterance, action) pairs
- Language is grounded in actions in the world

(Chen and Mooney, 2011)
Grounding semantics in control applications

1. Use feedback from control application to understand language

Walk across the bridge

Reward +1

Alleviate dependence on large scale annotation

2. Use language to improve performance in control applications

Score: 7

Score: 107

1. Ghosts chase and try to kill you
2. Collect all the pellets
3. ...
Reinforcement Learning

• Delayed feedback

⇒ How to perform credit assignment for individual actions

• Large number of possible action sequences

⇒ Need for effective exploration

Improved language understanding translates to improved task performance
Playing Civilization by reading game manuals

Neural network for policy

(Branavan et al., 2012)
Text-based games

You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here.

- open mailbox
- go east
- search field

Underlying game state (h1)
You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here.

- open mailbox
Text-based games

Opening the mailbox reveals a leaflet.

Underlying game state (h2)
You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here.
You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here.

Underspecified descriptions
Opportunity

Grounded language learning

You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here.

open mailbox
Opportunity

Grounded language learning

In-game rewards provide unstructured feedback

+10 gold
+5 health
Traditional RL framework

Markov decision process

**State** $s = \text{Observed Environment}$

**Action** $a = \text{Command to execute}$

**State 1**
- **Location**: Field
- **Wind level**: 3
- **Time**: 12pm
- **Mailbox**: closed

**Action**
- **open mailbox**

**State 2**
- **Location**: Field
- **Wind level**: 3
- **Time**: 12pm
- **Mailbox**: open

**Policy**

$$\pi : s \rightarrow a$$

**Reward**

$$+1$$

**Action value function**

$$Q(s, a)$$
Text-based games

**Partially observed Markov decision process**

**State** \( s \) = Observed Environment

**Action** \( a \) = Command to execute

**State 1**
You are standing in an open field west of a white house. There is a small mailbox here.

**Action**
open mailbox

**State 2**
Opening the mailbox reveals a leaflet.

**Location**: Field

**Wind level**: 3

**Policy**
\[ \pi : s \rightarrow a \]

**Reward**
+1
Model: Representation Generator

Recurrent Neural Network to map text to vector representation

$T \rightarrow Q(s, a)$
Model: Action Scorer

Neural Network for control policy

Recurrent Neural Network to map text to vector representation

(Narasimhan et al., 2015)
Results

<table>
<thead>
<tr>
<th>Quest completion (%)</th>
<th>LSTM-DQN</th>
<th>BI-DQN</th>
<th>BOW-DQN</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>98.0</td>
<td>55.0</td>
<td>40.0</td>
<td>15.0</td>
</tr>
</tbody>
</table>

Graph showing quest completion over epochs for different methods: LSTM-DQN, BI-DQN, BOW-DQN, and Random.
Visualizing Learnt Representations

t-SNE visualization of vectors learnt by agent
Semantics does not exist in isolation

Language → Logical forms → Parse trees → Vector representations → Perception

Language → Interaction
Final Projects

• Poster Presentations:
  • **Date:** Monday, January 13, 2020
  • **Time:** 10 am - 12 pm
  • **Place:** Friend Center Convocation Room
  • Final reports due on Dean’s date, **January 14**