A Joint Model of Language and Perception for Grounded Attribute Learning

Cynthia Matuszek*, Nicholas FitzGerald*, Liefeng Bo, Luke Zettlemoyer, Dieter Fox
Learning Attributes

- Physically grounded systems $\rightarrow$ opportunities for learning

- **Motivation:**
  robots learning from *interaction*
  - With people
  - With environment

- Need a model to learn from real-world input
  - Language and sensory data about novel, physical things

- **Goal:** learn new ideas from interacting with the world.
Some Related Work

- **Vision**
  - Object recognition *(Felzenszwalb et al., TPAMI 2009; et al., CVPR 2009)*
  - Visual attributes of objects *(CVPR Farhadi et al., 2009; Parikh & Grauman, ICCV 2011)*
  - Kernel descriptors *(Bo et al., NIPS 2010; IROS 2011)*

- **Semantic Parsing**
  - Inducing semantic parsers *(Liang et al., ACL 2011; Wong & Mooney, ACL 2007; ...)*
  - CCG parsers *(Zettlemoyer & Collins, ACL 2009; Kwiatkowski et al., EMNLP 2011; ...)*

- **Grounded Language Acquisition and Parsing for Robotics**
  - Parsing NL in known world and action models *(Matuszek et al., ISER 2012; Tellex et al. AAAI 2011; Kollar et al., HRI 2010; ...)*
  - Parsing NL for RoboCup and navigation *(Mooney et al, Chen & Mooney, AAAI 2011)*
  - Language grounding for semantic mapping *(Kruijff & Zender, HRI 2006)*

- And many more!
Outline

- Introduction & Motivation
- Related Work
- Task Description
- Background
- Joint Model & Model Learning
- Experimental Results
- Discussion and Future Work
Task Description

- Learning to select objects described by attribute

- Learn **previously unknown attributes**
  - **Yellow**: new word describing new idea

- NLP-style semantic parsing: mapping NL to formal representation

- Grounded in real-world perception

“Which are the yellow objects?”
Task Description

Language
“yellow”

Formal Expression
\text{color-NEW-1}(x)

Classifier
\text{C}_{\text{color-new-1}}(x) = \{0,1\}

Ground Truth
Outline

- Introduction & Motivation
- Task Description
- Related Work
- **Background**
  - Semantic parsing model; perceptual model
- Joint Model & Model Learning
- Experimental Results
- Discussion and Future Work
Categorial Combinatory Grammars

- Capture syntax and semantics of language
- Parse sentences to expressions in $\lambda$-calculus
- Space of possible parses defined by:
  - lexical entries
  - red $\vdash N : \lambda x \cdot \text{color-red}(x)$
  - block $\vdash N \setminus N : \lambda x \cdot x$
- along with combinatory rules.
- Probabilistic CCGs define a log-linear model over:

$$p(y,z \mid x; \theta, \Lambda) = \frac{e^{\theta \cdot \phi(x,y,z)}}{\sum_{y',z'} e^{\theta \cdot \phi(x,y',z')}}$$
Visual model is a set of binary classifiers, one/attribute

- Each perceptual classifier is applied independently

Kernel descriptors
- Trained using linear SVM

$$P(w_{o,c} = 1 | o; \Theta^P) = \frac{e^{\Theta_P^o \cdot \phi(o)}}{1 + e^{\Theta_P^o \cdot \phi(o)}}$$
Outline

- Introduction & Motivation
- Task Description
- Related Work
- Background
- Joint Model & Model Learning
- Experimental Results
- Discussion and Future Work
Goal: compute the probability of an indicated set

\[ P(G \mid x, O) = \sum_z \sum_w P(G, z, w \mid x, O) \]

World model (product of possible classifier assignments to all objects \( o \) in \( O \)):

\[ P(w \mid O) = \prod_{o \in O} \prod_{c \in C} P(w_{o,c} \mid o) \]

Joint Model:

\[ P(G, z, w \mid x, O) = P(z \mid x)P(w \mid O)P(G \mid z, w) \]
Joint Model

- Goal: compute the probability of an indicated set

\[ P(G \mid x, O) = \sum \sum P(G \mid w) \]

- World model (product of possible assignments to all objects \( o \) in \( O \))

\[ P(w \mid O) = \prod \prod P(w_o, c \mid x, O) \]

- Joint Model:

\[ P(G, z, w \mid x, O) = P(z \mid x)P(w \mid O)P(G \mid z, w) \]
Inference

\[ P(G,z,w \mid x,O) = P(z \mid x) \prod_{o \in O} \prod_{c \in C} P(w_{o,c} \mid o) P(G \mid z,w) \]

Semantic Parsing

\[ \lambda x. \text{color-blue}(x) \]

"These are the ones that are blue"

Joint Probability

Parsing Model

Vision Model

Grounding Query
Language helps determine attribute relations

New language can be ambiguous: “This is \textit{<new word>}.”

- New color attribute?
  “This is red.”
- New shape attribute?
  “This is square.”
- Synonym?
  “This is saffron.”
- No attribute at all
  “This is toy.”

Vision helps decide among these possible classifiers
Supervised Learning

- With labeled sentence meaning, object groups, alignment
- Decomposes into two independent learning problems:

**Semantic Parsing**

- "These are the ones that are blue"
- \(\lambda x. \text{color-blue}(x)\)

- "This is a round toy"
- \(\lambda x. \text{shape-spheroid}(x)\)

**Attribute Classification**

- SVM on kernel descriptors

- \(P(z \mid x)\)
- supervised semantic parser

- \(P(w_{o,c} \mid o)\)
- classification of all attributes
Unsupervised Learning

- Labeling is expensive – can it be avoided?

- Initialization
  - Train an initial supervised model from labeled scenes (sentence/logic and object/attributes)

- Add $N$ new, unknown attribute classifiers
  - Initialize to a small, near-uniform distribution
  - Pair with every unknown word/phrase

- Objective: $LL(D; \Theta^L, \Theta^P) = \sum_{i=1 \ldots n} \ln P(G_i|x_i, O_i; \Theta^L, \Theta^P)$
Using an EM-style algorithm:

1. Compute latent $P(z \mid x)$ and $P(w_{c,o} \mid o)$

$$P(z, w \mid x_i, O_i, G_i; \Theta^L, \Theta^P) = \frac{P(z \mid x_i; \Theta^L)P(w \mid O_i; \Theta^P)P(G_i \mid z, w)}{\sum_{z'} \sum_{w'} P(z' \mid x_i; \Theta^L)P(w' \mid O_i; \Theta^P)P(G_i \mid z', w')}$$

2. Re-estimate parameters of parsing and vision models

$$\Delta^L = \sum_{z'} \sum_{w'} P(z', w' \mid x_i, O_i, G_i; \Theta^L, \Theta^P) \left( E_{P(y \mid x_i, z'; \Theta^L)} \left[ \phi^L_j(x_i, y, z') \right] - E_{P(y, z \mid x_i; \Theta^L)} \left[ \phi^L_j(x_i, y, z) \right] \right)$$

$$\Delta^P_c = \sum_{z'} \sum_{w'} P(z', w' \mid x_i, O_i, G_i; \Theta^L, \Theta^P) \left( \sum_{o \in O_i} \left[ w'_{o,c} - P(w'_{o,c} = 1 \mid \phi(o); \Theta^P) \right] \phi(o) \right)$$
Outline

- Introduction & Motivation
- Task Description
- Related Work
- Background
- Joint Model & Model Learning
- Experimental Results
- Discussion and Future Work
1: Initialization

“This is an orange ball.”

\[ \lambda \text{x}. \text{color-orange}(x) \land \text{shape-spheroid}(x) \]

2: Training

“All of these toys are yellow.”

3: Testing

“It’s the yellow block.”

“\color{blue}{\text{color-blue}(x)}? \quad \text{shape-rect}(x)? \quad \text{shape-cyl.}(x)? \ldots”

“color-NEW(x)”
What is the Parent Saying?

Watch the video, then **describe what the parent is saying to the child**, in complete sentences.

- Pretend you are a parent teaching a child about something.
- The question is:

  **How does the parent describe this group of objects?**

Your answer should be the sentence(s) the parent said while pointing to these things.

"This one’s an orange ball."

\[ \lambda x . \text{orange}(x) \land \text{spheroid}(x) \]
Experimental Evaluation

- 142 scenes, 6 color and 6 shape attributes
- ~1,000 NL sentences from Mechanical Turk
- Ground truth formulas and classifier assignments

- 20 splits into
  - 30% training items for initialization (3 colors, 3 shapes)
  - 55% training items for training (3 new colors, 3 new shapes)
  - 10% test cases with new colors+shapes

Precision = 82%, Recall = 71%, F1 = 76%
## Object Classifier Results

**x-axis:** Novel NL words

**y-axis:** Selected classifiers (including null)

<table>
<thead>
<tr>
<th>NL Token</th>
<th>NEW0</th>
<th>NEW1</th>
<th>NEW2</th>
<th>NEW3</th>
<th>NEW4</th>
<th>NEW5</th>
<th>null</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>3.27</td>
<td>-0.34</td>
<td>-0.37</td>
<td>-0.16</td>
<td>-0.16</td>
<td>-0.17</td>
<td>0.00</td>
</tr>
<tr>
<td>green</td>
<td>-0.39</td>
<td>-0.30</td>
<td>3.47</td>
<td>-0.19</td>
<td>-0.19</td>
<td>-0.19</td>
<td>0.00</td>
</tr>
<tr>
<td>blue</td>
<td>-0.34</td>
<td>2.97</td>
<td>-0.31</td>
<td>-0.16</td>
<td>-0.16</td>
<td>-0.16</td>
<td>0.00</td>
</tr>
<tr>
<td>thing</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.29</td>
</tr>
<tr>
<td>cube</td>
<td>-0.43</td>
<td>0.31</td>
<td>-0.37</td>
<td>-0.23</td>
<td>0.00</td>
<td>2.78</td>
<td>0.00</td>
</tr>
<tr>
<td>that</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.42</td>
</tr>
<tr>
<td>arch</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.09</td>
<td>-0.14</td>
<td>0.60</td>
<td>-0.15</td>
<td>0.00</td>
</tr>
<tr>
<td>triangle</td>
<td>0.34</td>
<td>-0.30</td>
<td>0.04</td>
<td>1.92</td>
<td>-0.18</td>
<td>-0.19</td>
<td>0.00</td>
</tr>
<tr>
<td>toys</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Initialization Data

- Primarily initializing language model
- Training focused on learning new attributes/words
Failure cases

- **Bad parses:**
  
  “This is a red, toy rectangle.”

- **Bad classification:**
  
  Cylinders (lengthwise) look like rectangular solids
  
  Humans made the same errors – data is noisy!
Failure cases

◆ Incorrect human input
  ◆ Typos
  ◆ Visual errors

“This is a blue toy shaped like a half-pipe.”

◆ Unexpected human input

It’s brown, with blue specular highlights. (This confused the classifiers, too)

“This object is a fake piece of green lettuce. Do not try to eat!”
Accurate language and attribute models can be learned from data:
- Language, raw percepts, and target objects

Language and vision combine to learn previously-unrepresented ideas
- Extending the world model by interaction, not programming.

Future Work
- Scale complexity of language, sensory streams
- Extend to learning tasks, goals, and more attributes
- Implement in interactive setting
Thanks!

Questions?