Modeling Conceptual Understanding in Image Reference Games

Prof. Dr. Zeynep Akata

Interpretable ML Tutorial at CVPR 2020

15 June 2020
Outline

Background: Explanation and Learning Are Related

Modeling Conceptual Understanding With Image Reference Games

Conclusion: Explaining Through Communication Is Exciting
Outline

Background: Explanation and Learning Are Related

Modeling Conceptual Understanding With Image Reference Games

Conclusion: Explaining Through Communication Is Exciting
Learning via Explanation

Lombrozo TICS’16
Learning via Explanation

Lombrozo TICS’16
Learning via Explanation

Lombrozo TICS’16
Learning via Explanation

Lombrozo TICS’16
Attributes as Explanations

Lampert et al. CVPR’09

images

attributes

Blue crown
White belly
Black eyes

Red crown
Orange belly
Black eyes

classes

Eastern bluebird
[1 1 1 0 0]

Cardinal
[0 0 1 1 1]

5
Attributes as Explanations

Lampert et al. CVPR'09

Images

- Eastern bluebird
  - [1 1 1 0 0]
  - Blue crown
  - White belly
  - Black eyes

- Cardinal
  - [0 0 1 1 1 1]
  - Red crown
  - Orange belly
  - Black eyes

Classes

- Eastern bluebird
- Cardinal
Attributes as Explanations

Lampert et al. CVPR’09

images

attributes

Blue crown
White belly
Black eyes

Red crown
Orange belly
Black eyes

classes

Eastern bluebird
[1 1 1 0 0]

Cardinal
[0 0 1 1 1] Cardinal
Natural Language as Explanations for Communication
What type of bird is this?
What type of bird is this? It is a Cardinal because it is a red bird with a red beak and a black face. Why not a Vermilion Flycatcher? It is not a Vermilion Flycatcher because it does not have black wings.
Natural Language as Explanations for Communication

What type of bird is this?

It is a **Cardinal** because it is a **red bird** with a **red beak** and a **black face**.
What type of bird is this?

It is a Cardinal because it is a red bird with a red beak and a black face.

Why not a Vermilion Flycatcher?
What type of bird is this? It is a Cardinal because it is a red bird with a red beak and a black face.

Why not a Vermilion Flycatcher? It is not a Vermilion Flycatcher because it does not have black wings.
This red bird has a red beak and a black face.
This red bird has a red beak and a black face.
This red bird has a red beak and a black face.

This red bird has a black beak and a black face.
This red bird has a red beak and a black face.

This red bird has a black beak and a black face.

Explanation Sampler

\[ \mathbf{I}, \mathbf{C} \]
\[ w_0, \ldots, w_{i-1}, w_i \]

Explanation Grounder

\[ \{(A_i, R_i, s_i)\} \]

Phrase-Critic

\( S^p \)

\( S^n \)

\( 2.05 \)

\( 1.02 \)
Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing

Baker et al. Nature'17
Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing

Baker et al. Nature’17
Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing

Baker et al. Nature'17
Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing

Baker et al. Nature’17

Frame 1

Frame 2

Frame 3

Probability

Utility

World

Object

Model

People
Machine Theory of Mind

(a) past traj.

(b) current state
Machine Theory of Mind

Rabinowitz et al. ICML’18

(a) past traj. (b) current state

(c) action

\[
\hat{\pi} = \begin{cases} 
1 & \text{up} \\
0 & \text{down, left, right} 
\end{cases}
\]

consumption

prob.

\[
\begin{cases} 
1 & \text{green} \\
0 & \text{other colors} 
\end{cases}
\]
M$^3$RL: Mind-aware Multi-agent Management Reinforcement Learning

Shu et al. ICLR’19

(a) Nature of the workers
M$^3$RL: Mind-aware Multi-agent Management Reinforcement Learning

(a) Nature of the workers

(b) Incomplete information
M$^3$RL: Mind-aware Multi-agent Management Reinforcement Learning

(a) Nature of the workers
(b) Incomplete information
(c) Contract generation

Shu et al. ICLR’19
Outline

Background: Explanation and Learning Are Related

Modeling Conceptual Understanding With Image Reference Games

Conclusion: Explaining Through Communication Is Exciting
Image Reference Games with Failure in Concept Understanding
Image Reference Games with Failure in Concept Understanding

\[ x_t \quad x_c \]

Red

???
Image Reference Games with Failure in Concept Understanding
**Modeling Conceptual Understanding**

Speaker

Listener (color-blind)

\[ x^k_t \]

\[ x^k_c \]

**Reward**

"It's image"

\[ +1 \]

\[ -1 \]

Agent Embedding

Cone beak

Red beak

Yellow feet

Cone beak

Red beak

Yellow feet

Cone beak

Yellow feet

Yellow feet

Cone beak

or
Modeling Conceptual Understanding

Speaker

$\phi_S(x_t^k)$

Yellow feet
Red beak
Cone beak

Listener (color-blind)

$\phi_L(x_t^k)$

Yellow feet
Red beak
Cone beak

$\phi_S(x_c^k)$

Yellow feet
Red beak
Cone beak

$\phi_L(x_c^k)$

Yellow feet
Red beak
Cone beak

Reward

“It’s image”

+1

-
Modeling Conceptual Understanding

Corona, Alaniz, Akata NeurIPS’19

Speaker

Listener (color-blind)

Red beak

Reward

“It’s image

+1

−1

Agent Embedding

\[ \phi_S(x^k_t) \]

\[ \phi_L(x^k_c) \]

\[ x^k_c \leftrightarrow x^k_t \]

\[ \pi([z_{s,k}, h_k]) \]

\[ g_k \]

\[ x^k_t \leftrightarrow x^k_c \]

\[ z_{l,k} \]

\[ \phi_S(x^k_c) \]

\[ \phi_L(x^k_t) \]

yellow feet

red beak

cone beak

Yellow feet

Red beak

cone beak
Perceptual Modules (PM)

1. Extract image-level features using a CNN
2. Predict attribute-level features

\[ \phi(x) = f(\text{CNN}(x)) \]
Perceptual Modules (PM)

1. Extract image-level features using a CNN
2. Predict attribute-level features

\[ \phi(x) = f(CNN(x)) \]

- each element in \( \phi(x) \in [0, 1]^{|A|} \)
  represents a separate attribute
- \(|A|\): \# of visual attribute labels
- Speaker is one: \( \phi_S \),
  Multiple listeners: \( \phi_L \)

<table>
<thead>
<tr>
<th>images</th>
<th>attributes</th>
<th>classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1 1 1 0 0]</td>
<td>Blue crown</td>
<td>Eastern bluebird</td>
</tr>
<tr>
<td>[0 0 1 1 1]</td>
<td>Red crown Orange belly Black eyes</td>
<td>Cardinal</td>
</tr>
</tbody>
</table>
Agent Embedding (AE)

Speaker: Select attribute $a_k$ from

$$z_s^a = \phi^a_S(x^k_t) - \phi^a_S(x^k_c).$$

Listener: Select attribute $a_k$ from

$$z_l^a = \phi^a_L(x^k_t) - \phi^a_L(x^k_c).$$

receives reward $r_k \in \{-1, 1\}$
Agent Embedding (AE)

Speaker: Select attribute $a_k$ from

$$z_s^a = \phi^a_S(x_t^k) - \phi^a_S(x_c^k).$$

Listener: Select attribute $a_k$ from

$$z_l^a = \phi^a_L(x_t^k) - \phi^a_L(x_c^k).$$

receives reward $r_k \in \{-1, 1\}$

AE module: LSTM, AE $h_k$: LSTM hidden state

$$h_k = \text{LSTM}(h_{k-1}, o_k)$$

$o_k$: One-hot vector, the index of the non-zero entry is $a_k$ and its value is $r_k$. 
Policy Learning

Concatenate image-pair difference and AE

\[ s_k = \left[ \phi(x_t^k) - \phi(x_c^k); h_k \right] \]

Predict \( V(s_k, a_k) \) of using \( a_k \) to describe \( x_t^k \):

\[ \mathcal{L}_V = \frac{1}{N + M} \sum_{N+M} \text{MSE}(V(s_k, a_k), r_k) \]
Policy Learning: Different Policies Implemented Here

1. Epsilon Greedy Policy: Randomly sample $a_k$ with prob. $\epsilon$ or greedily choose $a_k$

   $$a_k = \arg \max_{a \in A} V(s_k, a)$$

2. Active Policy: Train using policy gradient

   $$\mathcal{L}_a = \frac{1}{N} \sum_{N} -R \log \pi_S(s_t, a_t) \text{ with } R = -\frac{1}{M} \sum_{M} \text{MSE}(V(s_k, a_k), r_k) \quad (1)$$

3. Random Agent policy: Always select $a_k$ at random
4. Reactive policy: Select $a_k$ at random, if $r_k = -1$ sample a different $a_k$
5. Random Sampling: Select $a_k$ at random during $N +$ greedy during $M$
Comparing Learned Policies vs Baselines

![Graph showing comparison between learned policies and baselines. The x-axis represents the number of games, and the y-axis represents the average reward. The graph shows that the learned policies perform consistently better than the baselines.]
Comparing Learned Policies vs Baselines

![Graph showing comparison between learned policies and baselines](image-url)
Comparing Learned Policies vs Baselines

![Graph showing ALE CUB performance across different policies]

- Epsilon Greedy
- Random Agent
- Active
- Reactive
- Random Sampling

The graph illustrates the performance of various policies over the number of games, with the x-axis representing the number of games and the y-axis showing the average reward.
Comparing Learned Policies vs Baselines

![ALE CUB Graph](image)

- **Epsilon Greedy**
- **Random Agent**
- **Active**
- **Reactive**
- **Random Sampling**
Showing Necessity of Agent Embeddings

![Graph showing CUB with Agent Embeddings and Baseline]
Evaluating Cluster Quality

1. Generate AE in 50K episodes
2. Perform K-Means on AE: $C'$ (GT = $C$)
3. Evaluate: variation of information (VI)

$$VI(C, C') = H(C) + H(C') - 2I(C, C')$$

where $H$: entropy, $I$: mutual information

- $VI$ measures amount of information needed to switch from $C$ to $C'$
Modeling Conceptual Understanding Qualitative Results

Game 1

Discrim. Chosen
Brown back
Brown back
Blue underparts
Blue underparts
Rufous belly
Rufous belly
Yellow wing
Yellow wing

[Images showing incorrect choices with an 'X' mark]
## Modeling Conceptual Understanding Qualitative Results

<table>
<thead>
<tr>
<th>Discrim. Chosen</th>
<th>Brown back</th>
<th>Blue underparts</th>
<th>Rufous belly</th>
<th>Yellow wing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown back</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blue underparts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rufous belly</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yellow wing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Game 1

- **Discrim. Chosen**: Orange leg
- **Brown back**: Spotted belly pattern
- **Blue underparts**: Spotted back pattern
- **Rufous belly**: Rufous crown
- **Yellow wing**: Solid belly pattern

### Game 10

- **Discrim. Chosen**: Rufous crown
- **Brown back**: Rufous belly
- **Blue underparts**: Rufous belly
- **Rufous belly**: Rufous crown
- **Yellow wing**: Yellow belly

- **Correct**: Rufous crown
- **Incorrect**: Rufous belly, Rufous belly, Yellow belly

21
## Modeling Conceptual Understanding Qualitative Results

<table>
<thead>
<tr>
<th>Game 1</th>
<th>Game 10</th>
<th>Game 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown back</td>
<td>Orange leg</td>
<td>Orange beak</td>
</tr>
<tr>
<td>Brown back</td>
<td>Spotted belly pattern</td>
<td>Duck-like shape</td>
</tr>
<tr>
<td>Blue underparts</td>
<td>Yellow belly</td>
<td>Yellow belly</td>
</tr>
<tr>
<td>Blue underparts</td>
<td>Spotted back pattern</td>
<td>Has eyebrow</td>
</tr>
<tr>
<td>Rufous belly</td>
<td>Rufous crown</td>
<td>Yellow wing</td>
</tr>
<tr>
<td>Rufous belly</td>
<td>Rufous crown</td>
<td>Solid belly pattern</td>
</tr>
<tr>
<td>Yellow wing</td>
<td>Yellow belly</td>
<td>White belly</td>
</tr>
<tr>
<td>Yellow wing</td>
<td></td>
<td>Forked tail shape</td>
</tr>
</tbody>
</table>

- Discriminative Chosen
- 

$x_t$ $x_c$ $x_t$ $x_c$ $x_t$ $x_c$ $x_t$ $x_c$
Outline

Background: Explanation and Learning Are Related

Modeling Conceptual Understanding With Image Reference Games

Conclusion: Explaining Through Communication Is Exciting
Conclusions

Modeling conceptual understanding is necessary to succeed in some tasks

1. Formulation for modeling other agents’ understanding
2. Allows XAI systems to tailor their explanations to the specific users
3. Learned AEs recovers a clustering over other agents’ conceptual understanding

Modeling Conceptual Understanding in Image Reference Games
Rodolfo Corona, Stephan Alaniz and Zeynep Akata
published at NeurIPS 2019


Thank you!