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

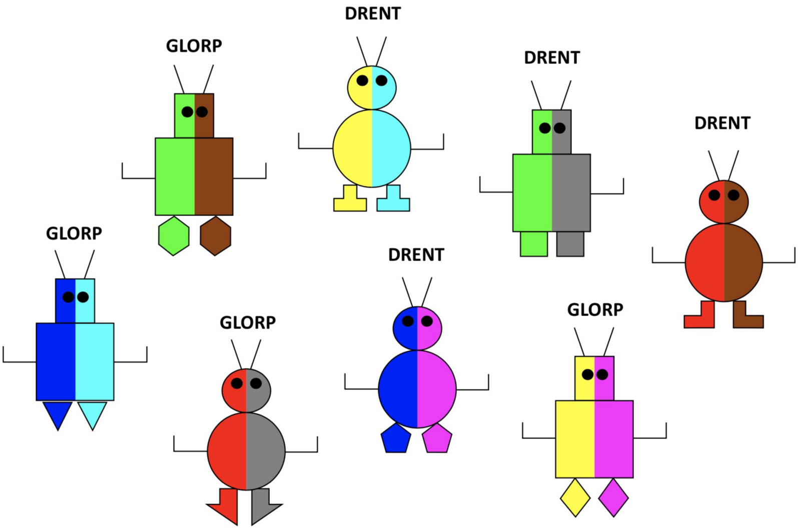
Conclusion: Explaining Through Communication Is Exciting

Outline

Background: Explanation and Learning Are Related

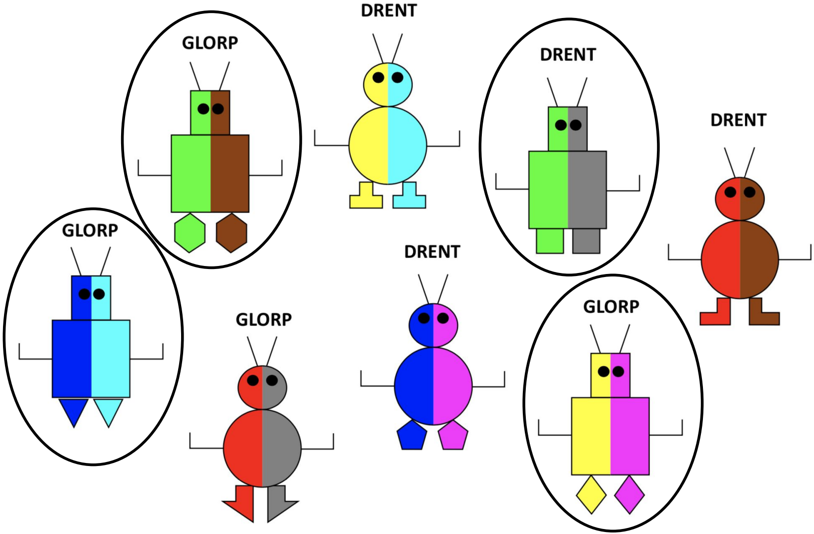
Modeling Conceptual Understanding With Image Reference Games

Conclusion: Explaining Through Communication Is Exciting


DRENT
GLORP
DRENT
DRENT
GLORP
DRENT
GLORP
GLORP

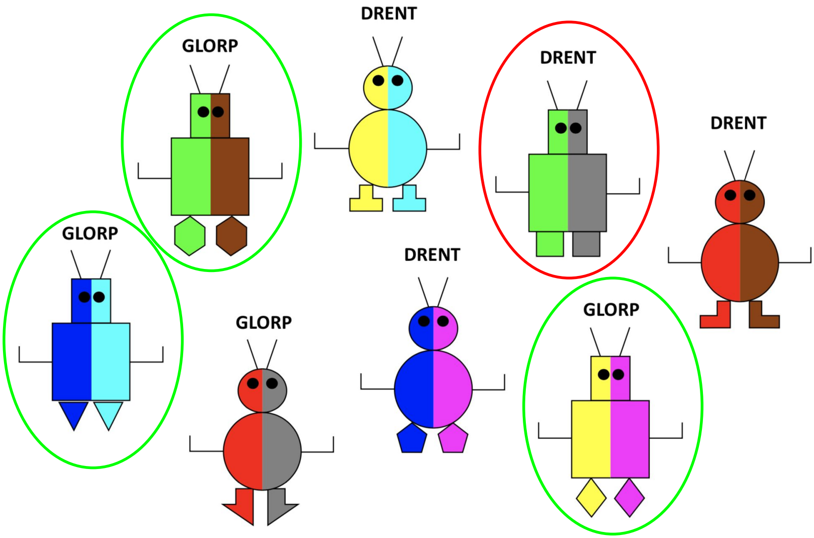
Learning via Explanation

Lombrozo TICS'16

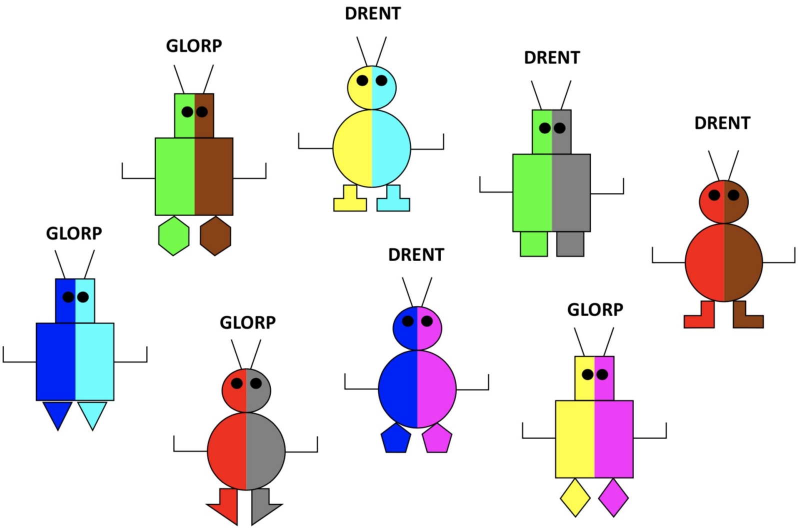


Learning via Explanation

Lombrozo TICS'16

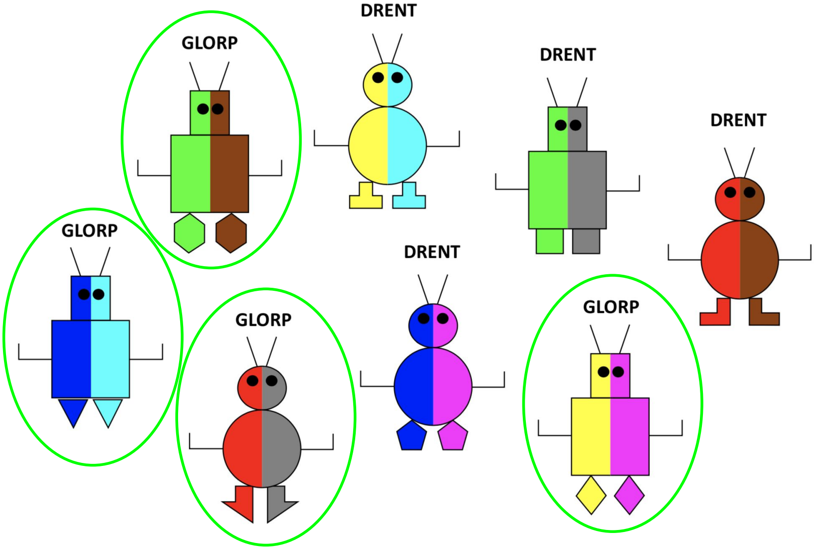


Learning via Explanation



Learning via Explanation

Lombrozo TICS'16



Attributes as Explanations

Lampert et al. CVPR'09

images **attributes** **classes**

Blue crown
White belly
Black eyes

Eastern bluebird

[1 1 1 0 0]

Red crown
Orange belly
Black eyes

[0 0 1 1 1]

Cardinal

Attributes as Explanations

Lampert et al. CVPR'09

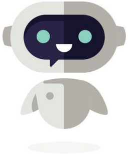
images	attributes	classes
	Blue crown White belly Black eyes	Eastern bluebird [1 1 1 0 0]
	Red crown Orange belly Black eyes	[0 0 1 1 1] Cardinal

Attributes as Explanations

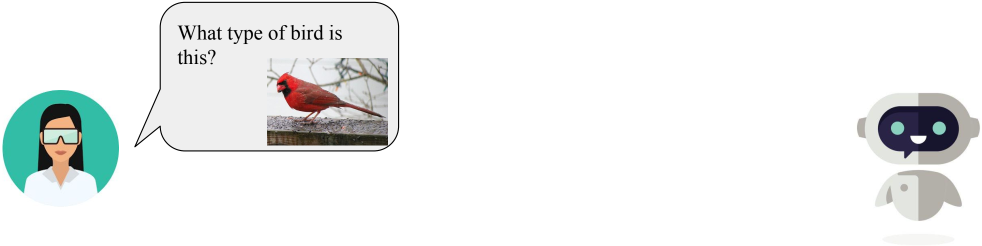
Lampert et al. CVPR'09

images	**attributes**	**classes**
	Blue crown White belly Black eyes	Eastern bluebird [1 1 1 0 0]
	Red crown Orange belly Black eyes	[0 0 1 1 1] Cardinal

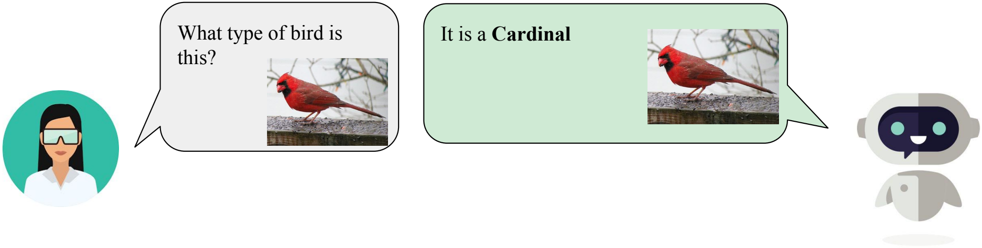
Natural Language as Explanations for Communication



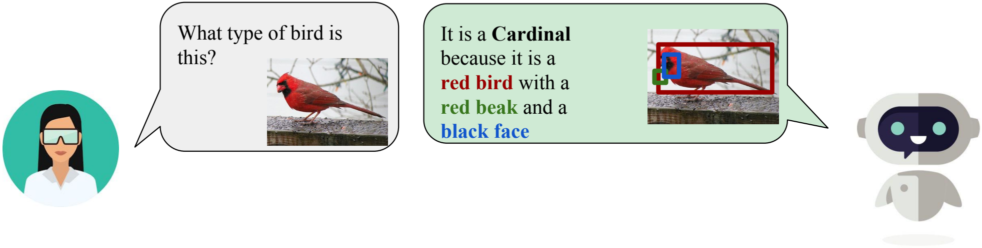
Natural Language as Explanations for Communication



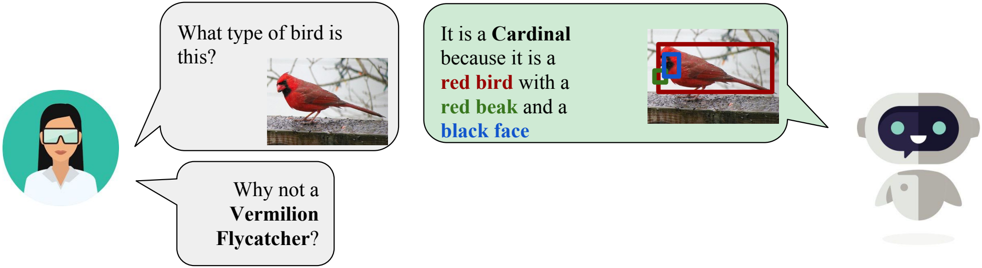
Natural Language as Explanations for Communication



Natural Language as Explanations for Communication



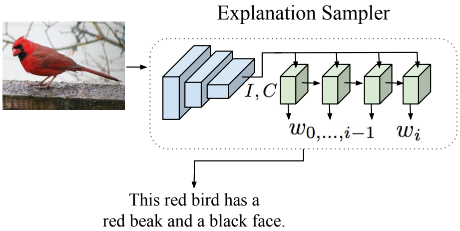
Natural Language as Explanations for Communication



Natural Language as Explanations for Communication

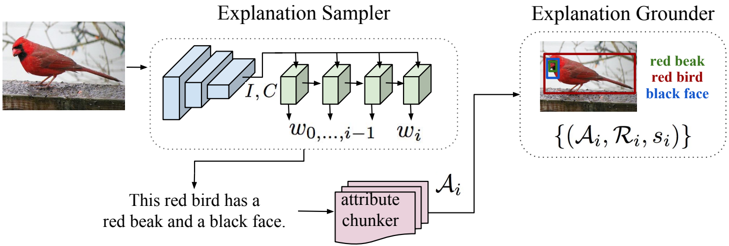
Grounding Visual Explanations

Hendricks et al. ECCV'16 & ECCV'18



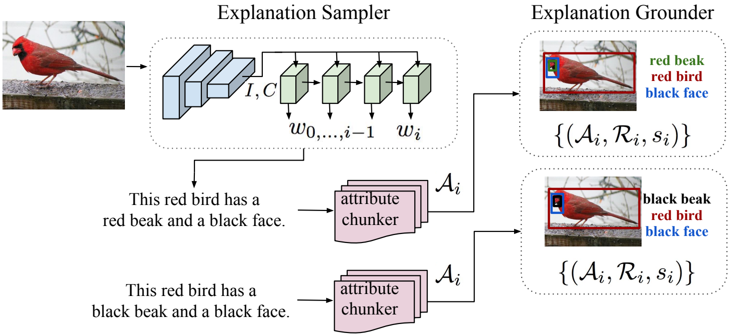
Grounding Visual Explanations

Hendricks et al. ECCV'16 & ECCV'18



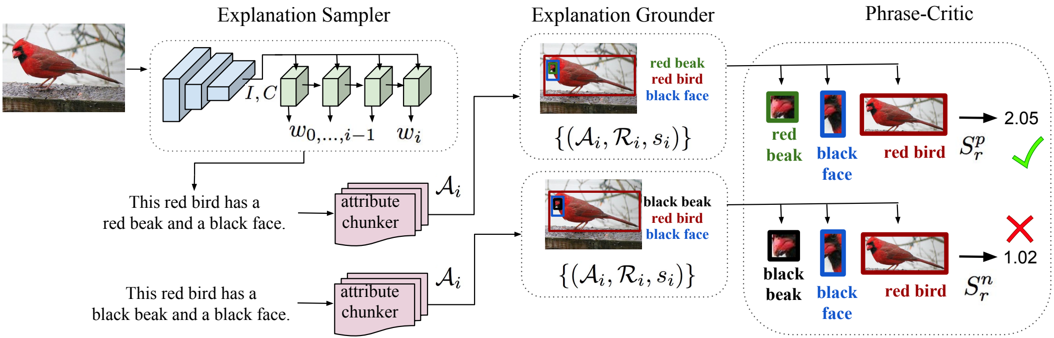
Grounding Visual Explanations

Hendricks et al. ECCV'16 & ECCV'18



Grounding Visual Explanations

Hendricks et al. ECCV'16 & ECCV'18



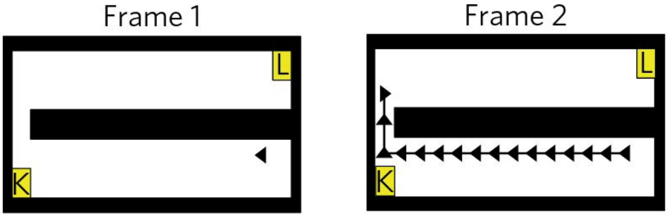
Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing

Baker et al. Nature'17

Frame 1

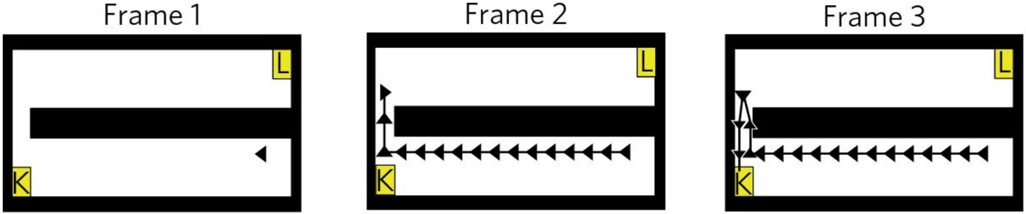
Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing

Baker et al. Nature'17



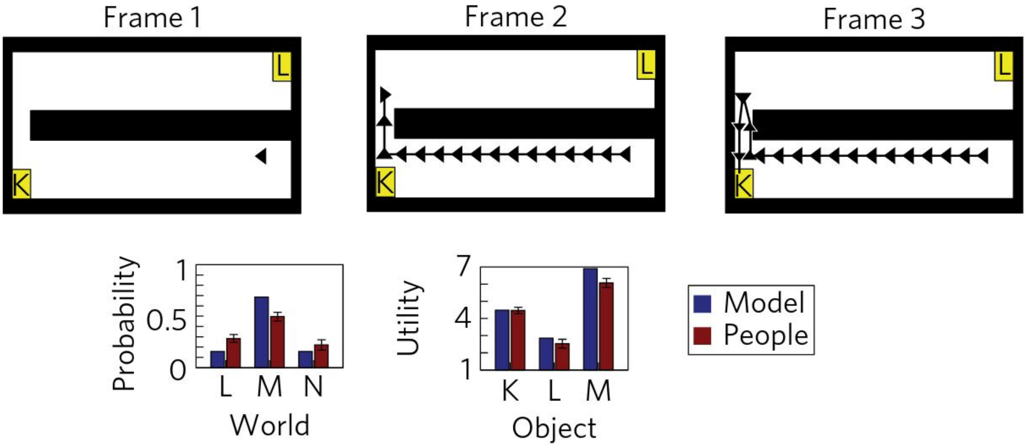
Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing

Baker et al. Nature'17

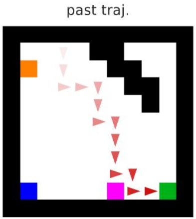


Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing

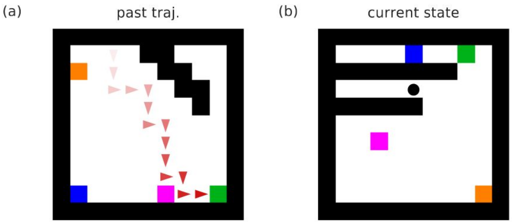
Baker et al. Nature'17

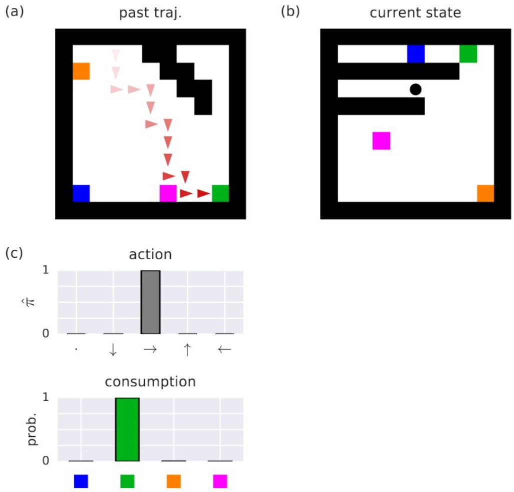


(a)

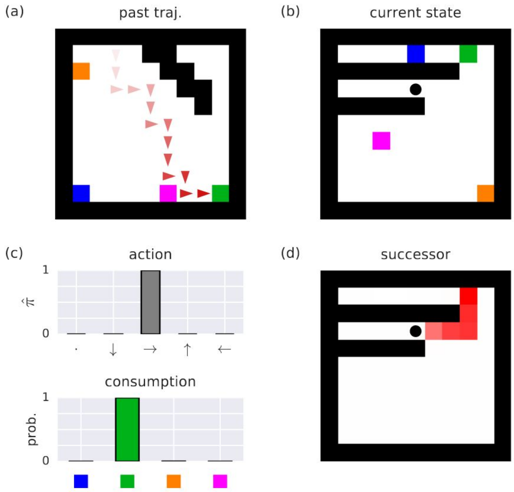


Machine Theory of Mind



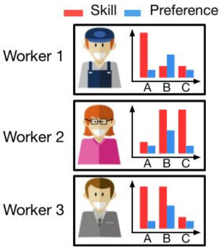

(a)
past traj.
(b)
current state
(c)
action
1
$\hat{\pi}$
0
· ↓ → ↑ ←
consumption
1
prob.
0

Machine Theory of Mind



M^{3}RL: Mind-aware Multi-agent Management Reinforcement Learning

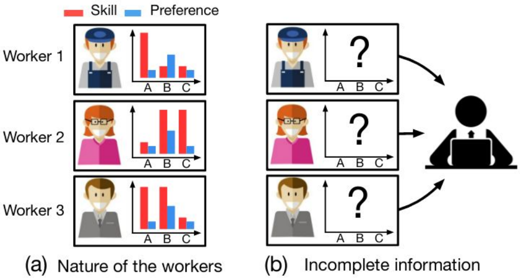
Shu et al. ICLR'19



(a) Nature of the workers

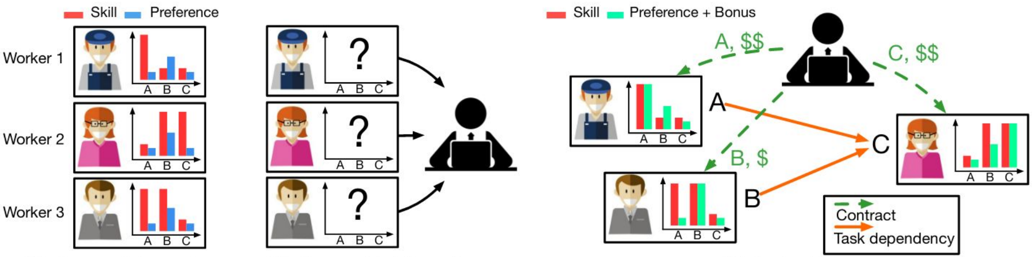
M^{3}RL: Mind-aware Multi-agent Management Reinforcement Learning

Shu et al. ICLR'19



M^{3}RL: Mind-aware Multi-agent Management Reinforcement Learning

Shu et al. ICLR'19



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

Outline

Image Reference Games with Failure in Concept Understanding

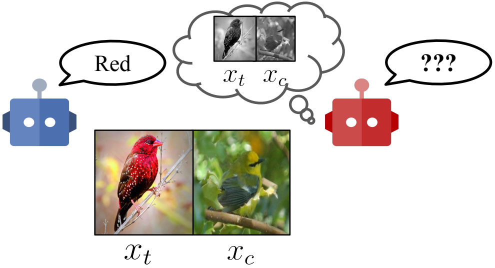


Image Reference Games with Failure in Concept Understanding

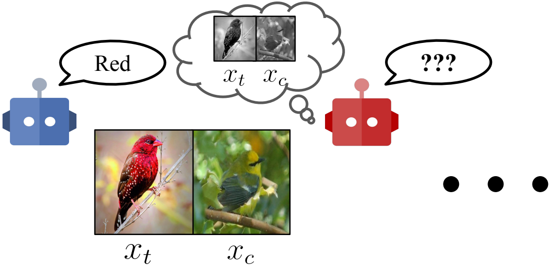
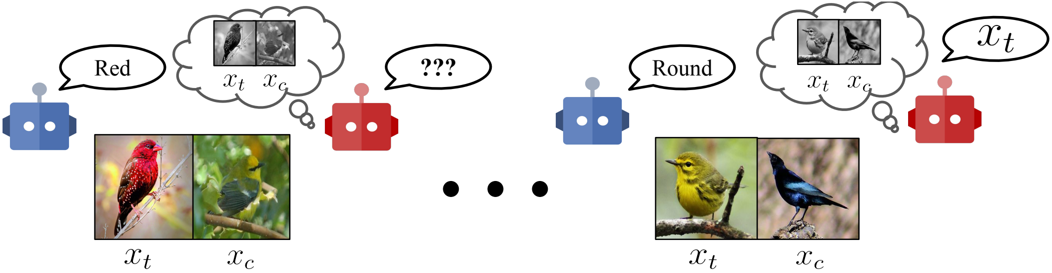
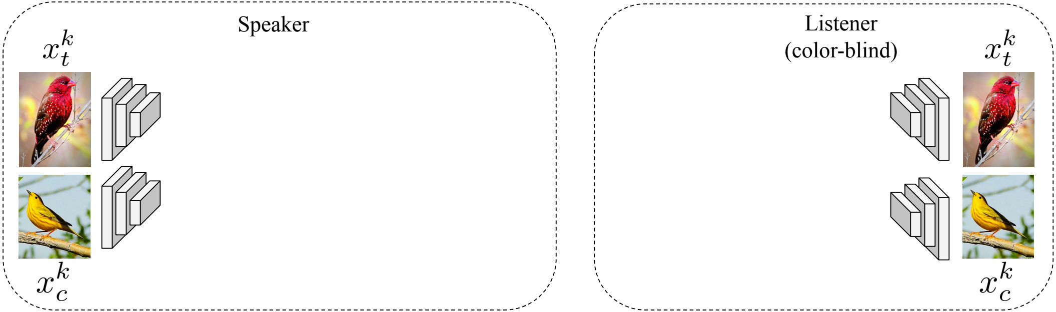


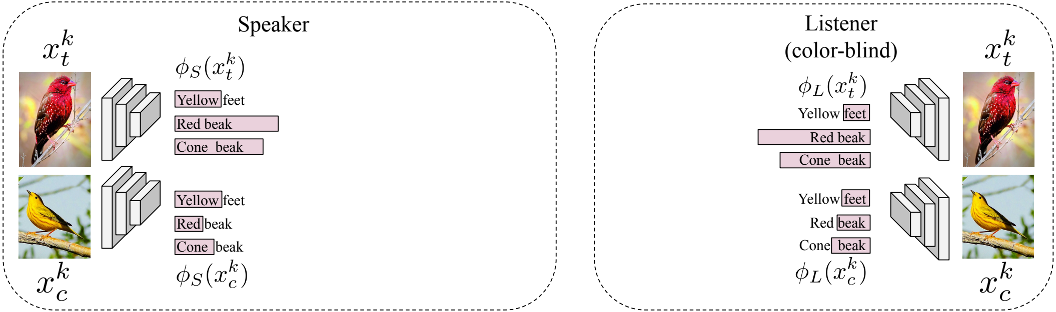
Image Reference Games with Failure in Concept Understanding




Speaker
x_t^k
x_c^k
Listener
(color-blind)
x_t^k
x_c^k

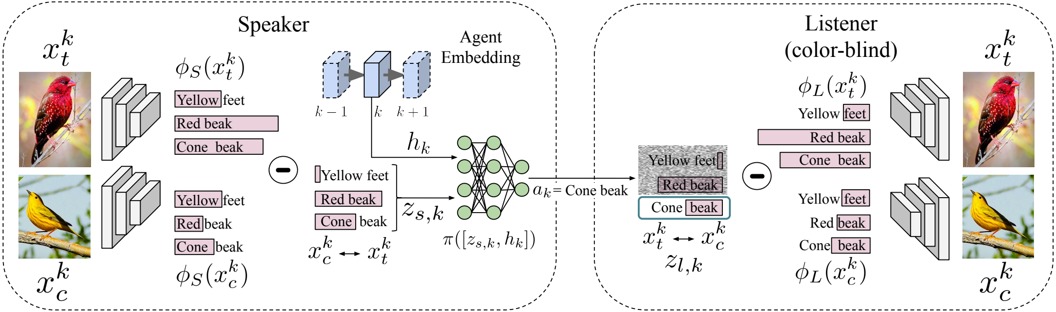
Modeling Conceptual Understanding

Corona, Alaniz, Akata NeurIPS'19



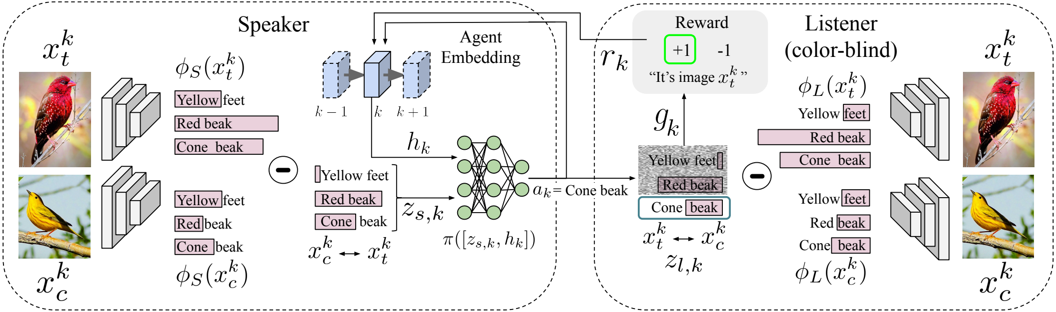
Modeling Conceptual Understanding

Corona, Alaniz, Akata NeurIPS'19

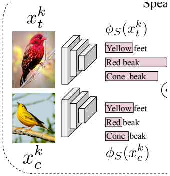


Modeling Conceptual Understanding

Corona, Alaniz, Akata NeurIPS'19



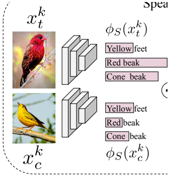
Perceptual Modules (PM)



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

$$\phi(x) = f(\mathsf{CNN}(x))$$

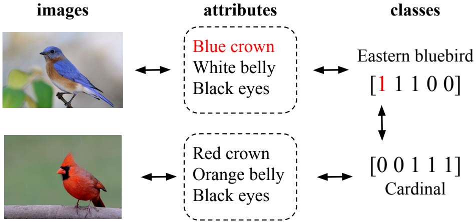
Perceptual Modules (PM)



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

$$\phi(x) = f(\text{CNN}(x))$$

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



Agent Embedding (AE)

Speaker: Select attribute a_k from

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

Listener: Select attribute a_k from

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

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

Agent Embedding (AE)

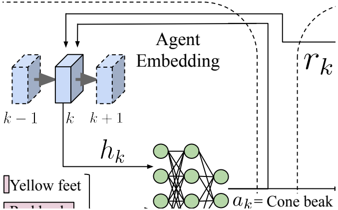
Speaker: Select attribute a_k from

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

Listener: Select attribute a_k from

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

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

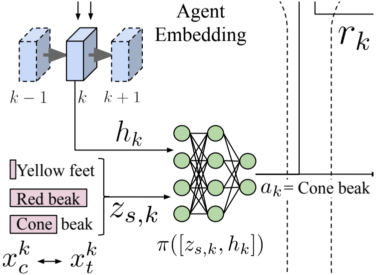


AE module: LSTM, AE h_k: LSTM hidden state

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

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. ϵ or greedily choose a_k

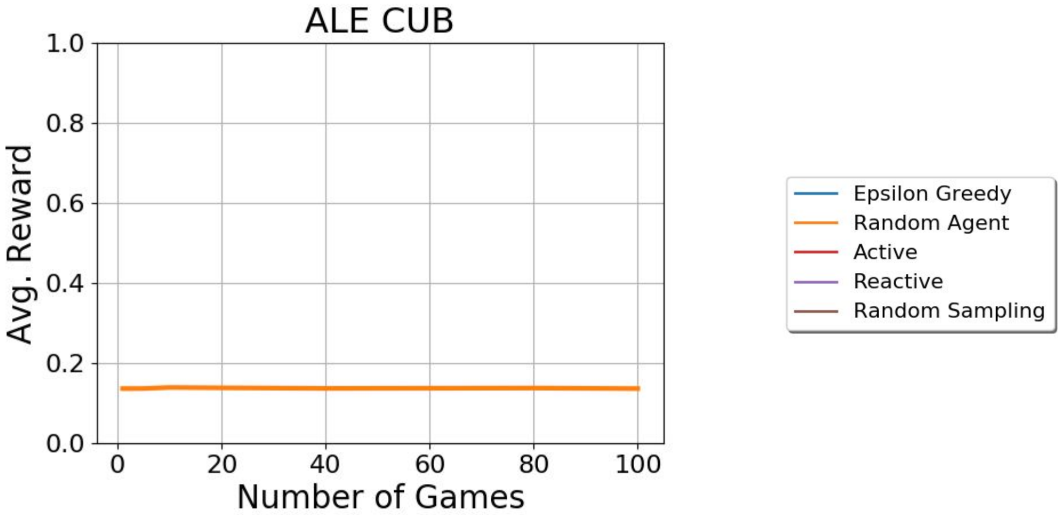
$$a_k = arg\max_{a \in A} V(s_k, a)$$

2. Active Policy: Train using policy gradient

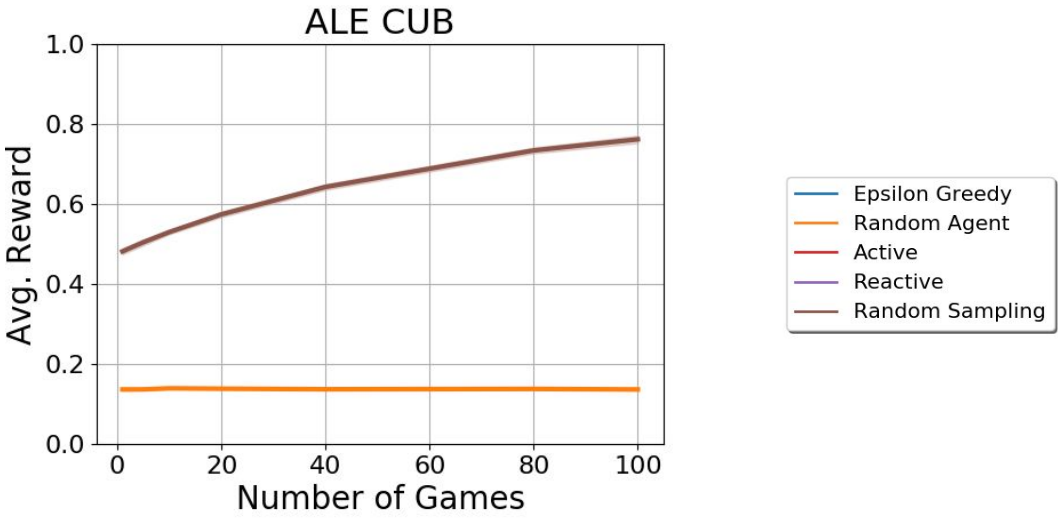
$$\mathcal{L}_a = \frac{1}{N} \sum_{N} -R \log \pi_S(s_t, a_t) \text{ with } R = -\frac{1}{M} \sum_{M} \text{MSE}(V(s_k, a_k), r_k) \tag{1}$$

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

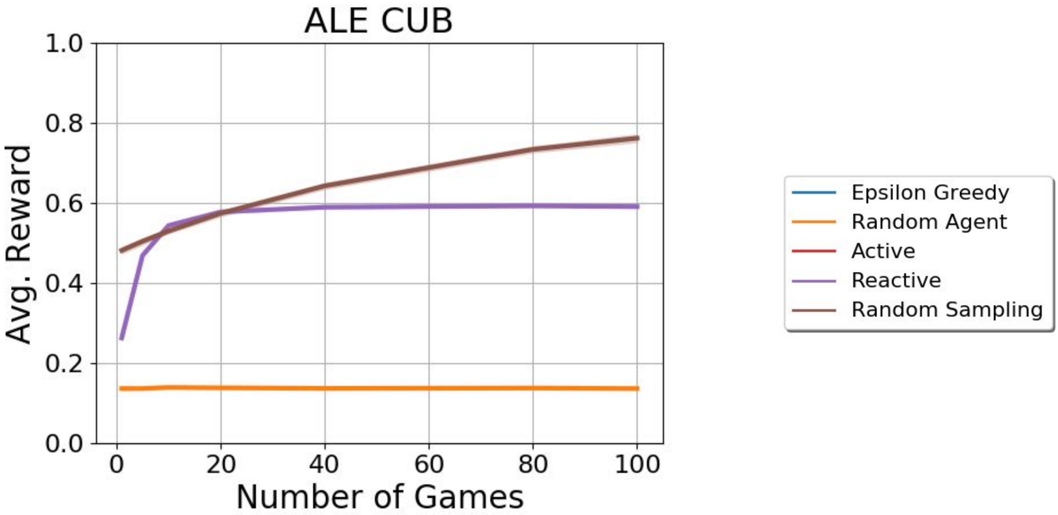
Comparing Learned Policies vs Baselines



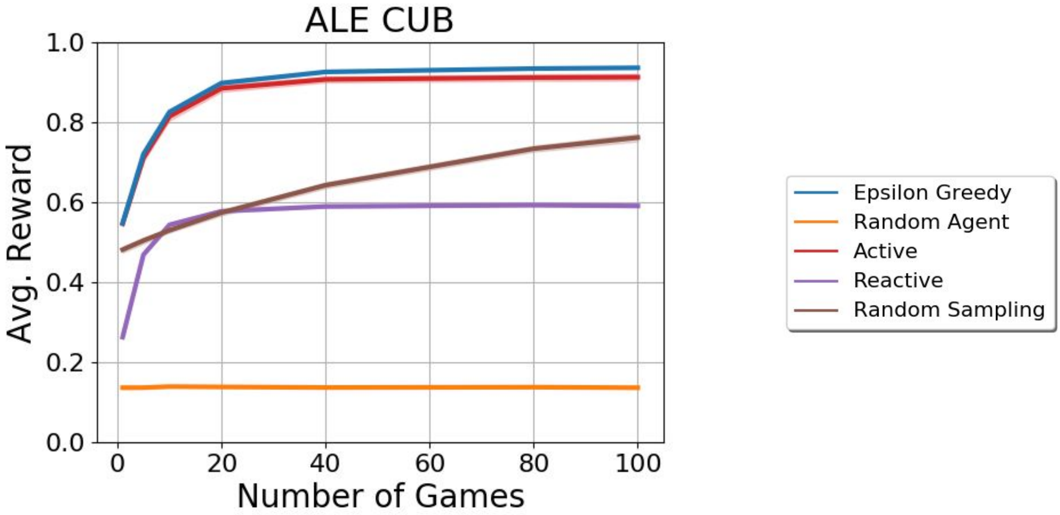
Comparing Learned Policies vs Baselines



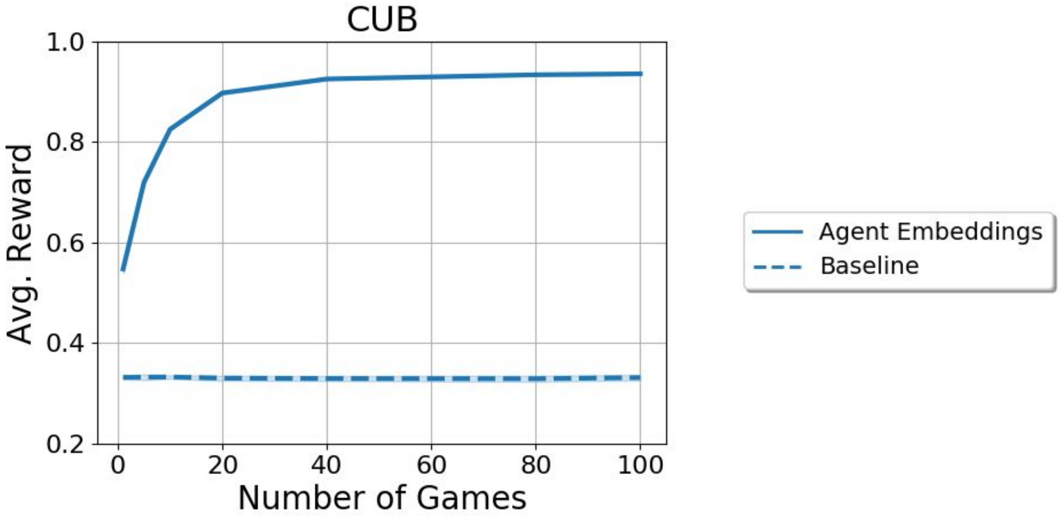
Comparing Learned Policies vs Baselines



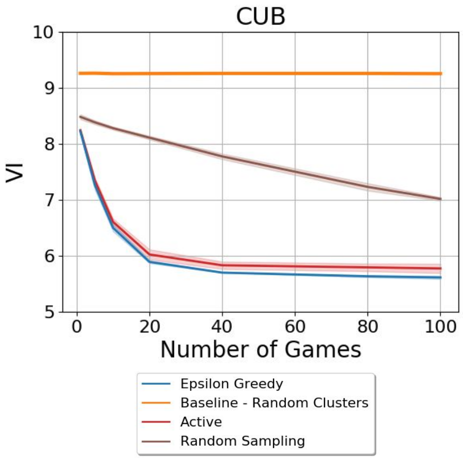
Comparing Learned Policies vs Baselines



Showing Necessity of Agent Embeddings



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

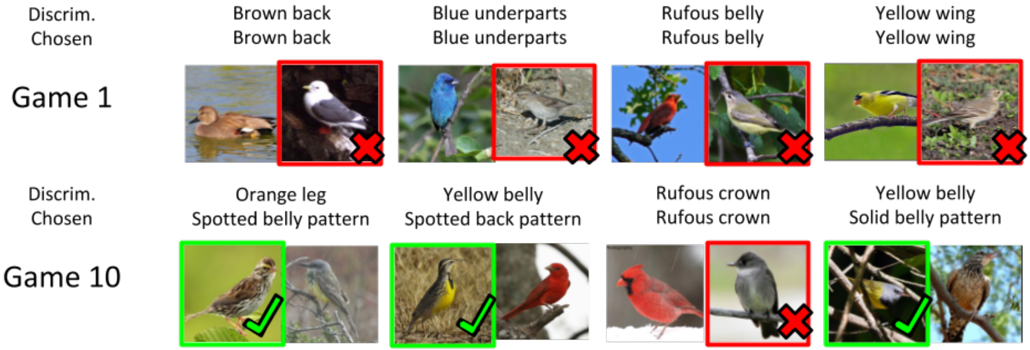
- VI measures amount of information needed to switch from C to C'

Modeling Conceptual Understanding Qualitative Results

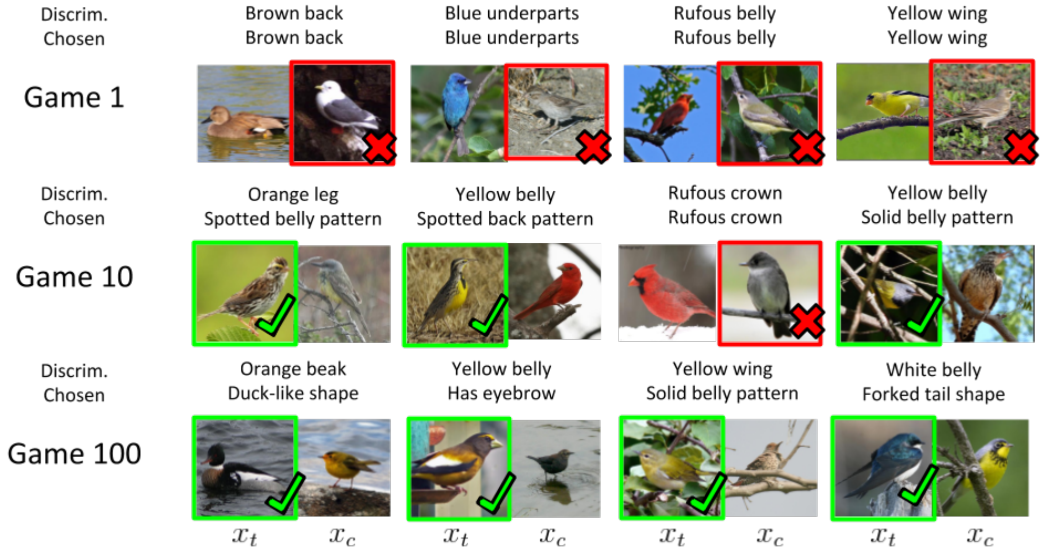
Discrim.	Brown back	Blue underparts	Rufous belly	Yellow wing
Chosen	Brown back	Blue underparts	Rufous belly	Yellow wing

Game 1

Modeling Conceptual Understanding Qualitative Results



Modeling Conceptual Understanding Qualitative Results



Outline

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

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. (2017).
Rational quantitative attribution of beliefs, desires and percepts in human mentalizing.
Nature Human Behaviour, 1.

Christoph H. Lampert, H. N. and Harmeling, S. (2014).
Attribute-based classification for zero-shot visual object categorization.
IEEE TPAMI.

Corona, R., Alaniz, S., and Akata, Z. (2019).
Modeling conceptual understanding in image reference games.
In *Neural Information Processing Systems (NeurIPS)*.

Hendricks, L.-A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016).
Generating visual explanations.
In *European Conference of Computer Vision (ECCV)*.

Hendricks, L. A., Hu, R., Darrell, T., and Akata, Z. (2018).
Grounding visual explanations.
In *European Conference of Computer Vision (ECCV)*.

Lombrozo, T. (2016).
Explanatory preferences shape learning and inference.
In *Trends in Cognitive Science*.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., and Botvinick, M. (2018).
Machine theory of mind.
In *ICML*.

Reed, S., Akata, Z., Lee, H., and Schiele, B. (2016).
Learning deep representations of fine-grained visual descriptions.
In *IEEE Computer Vision and Pattern Recognition (CVPR)*.

Shu, T. and Tian, Y. (2019).
M^3RL: Mind-aware multi-agent management reinforcement learning.
In *ICLR*.

Thank you!