Prototype Validation in Multi-agent Communication

Washington Garcia (UF)

Kevin Butler (UF)

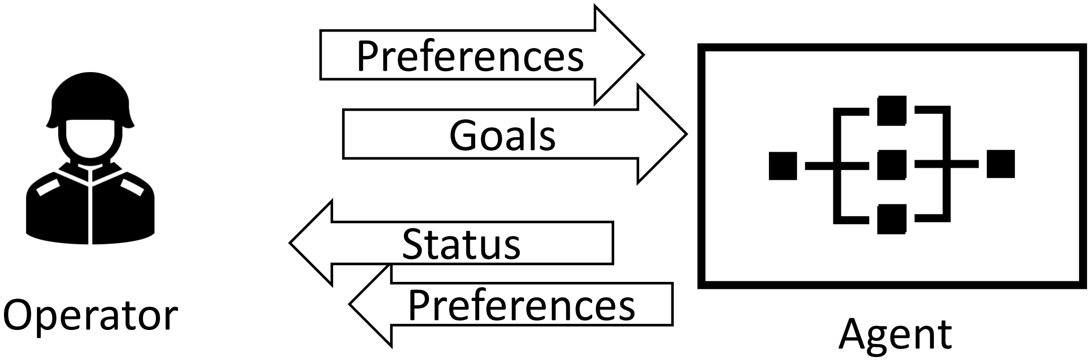
Scott Clouse (AFRL/ACT3)

Background

Machine agents are expected to solve tasks on behalf of human operators (and perhaps other agents).

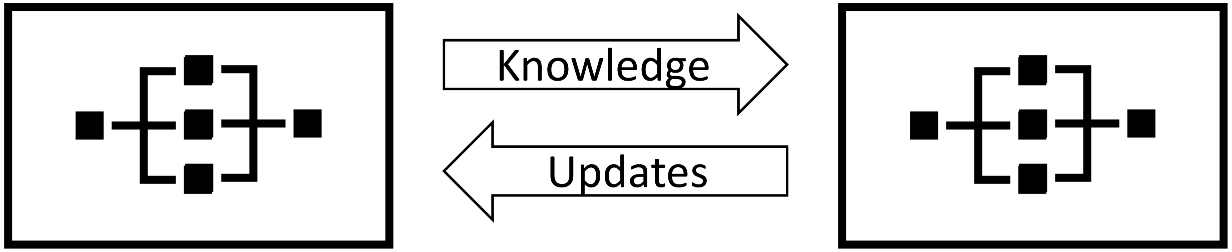
In order to cooperate, operators must be able to communicate effectively with the machine agents.

Beyond that, operator and machine should eventually be able to possess “shared” experiences and beliefs (embodied cognition view).

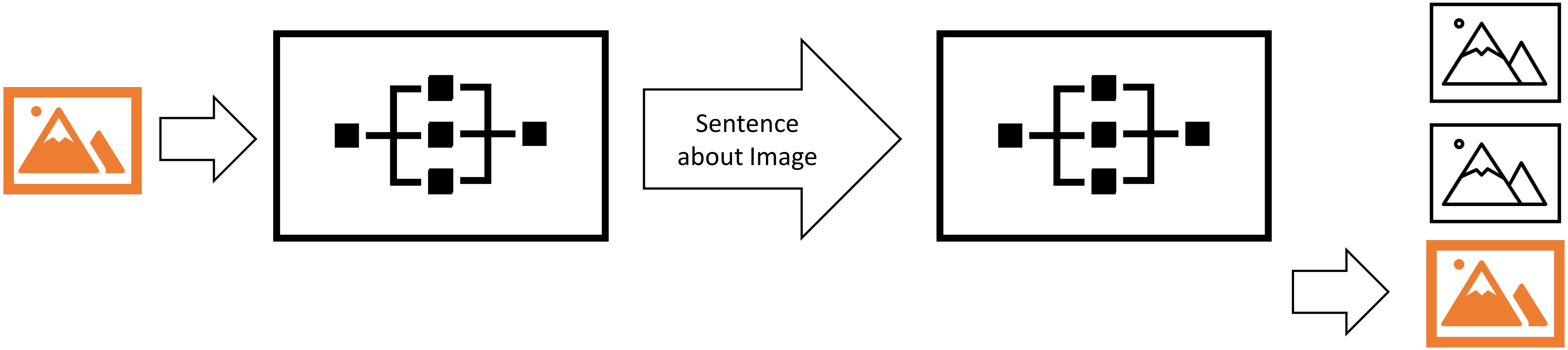


Background

For now let us consider a simpler case of agent-agent communication.

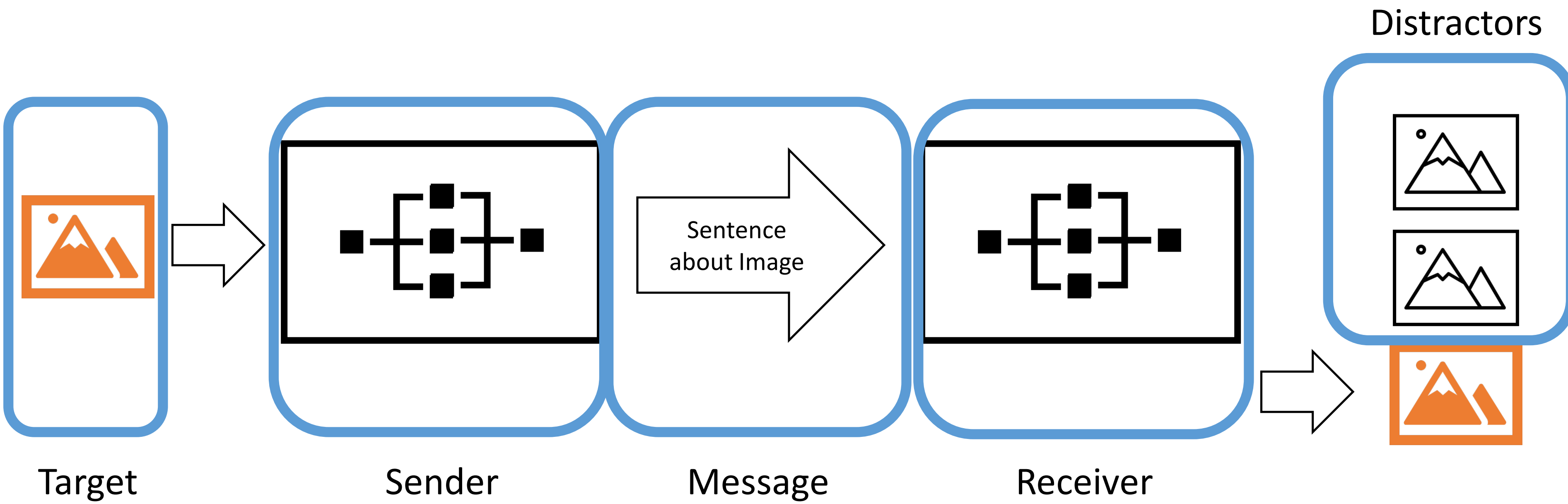


Borrow a scenario from psychology, the Lewis Signaling Game:



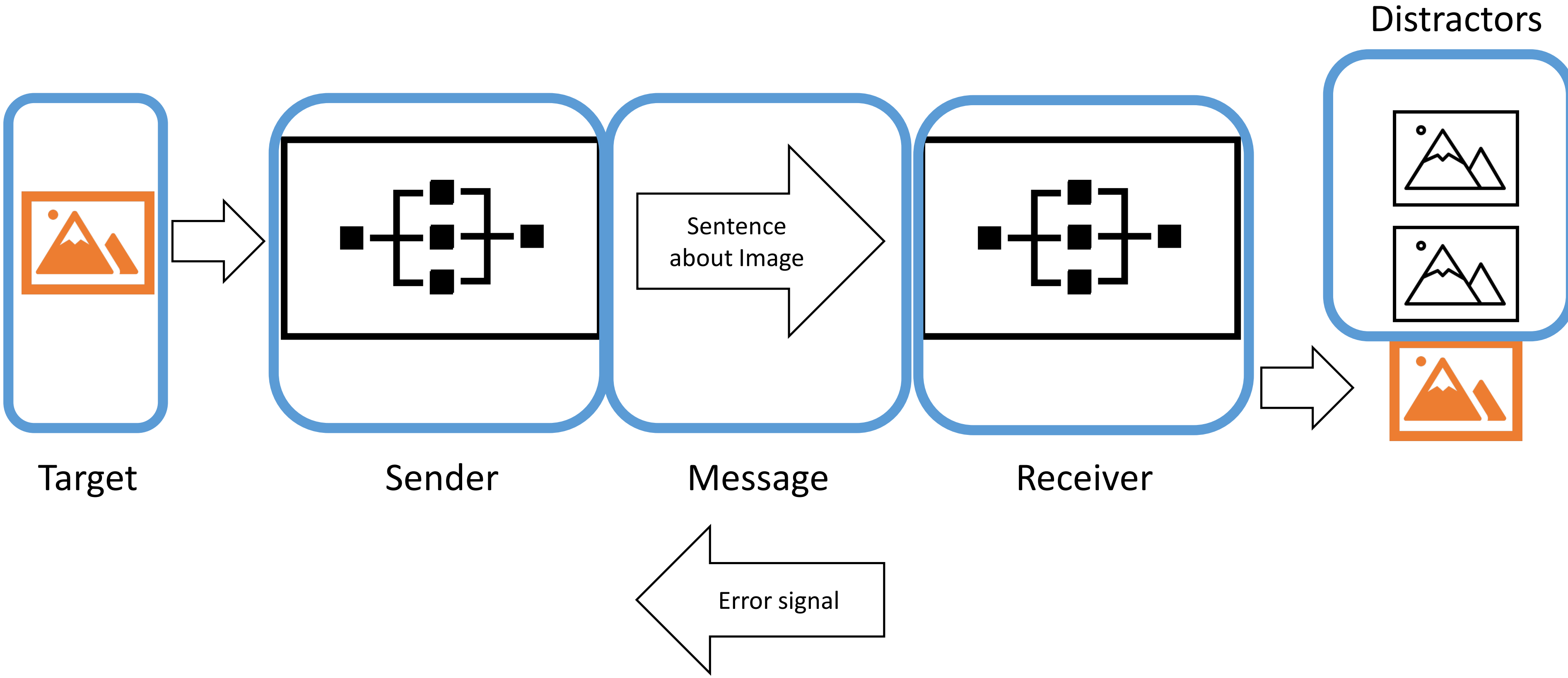
Background

Recent literature have examined "emergent language" between agents who cooperate in this game.



Background

Recent literature have examined “emergent language” between agents who cooperate in this game.



Emergent Language

So far emergent communication literature have been concerned with the message channel:

- Fix the vocabulary the Sender can use, what protocol/language forms between the agents?
- Does the protocol mimic human language? Can it?
  - Compositionality
  - Capacity
  - Bandwidth

What about the vocabulary?

Issues

Until now the vocabulary is fixed, which makes some troubling assumptions:

- Objects can be described by explicit sentence-like representations, consisting of tokens
  - Drawback: Reminiscent of Symbolic AI (i.e., physical symbol system hypothesis)
- The tokens are fixed apriori by the system designer
  - Issue: Designer's cognition is outside of the system, system is still open loop
- Since tokens are fixed, shall we assume they are the best?
  - Issue: How do we judge what tokens are best?

Issues

Until now the vocabulary is fixed, which makes some troubling assumptions:

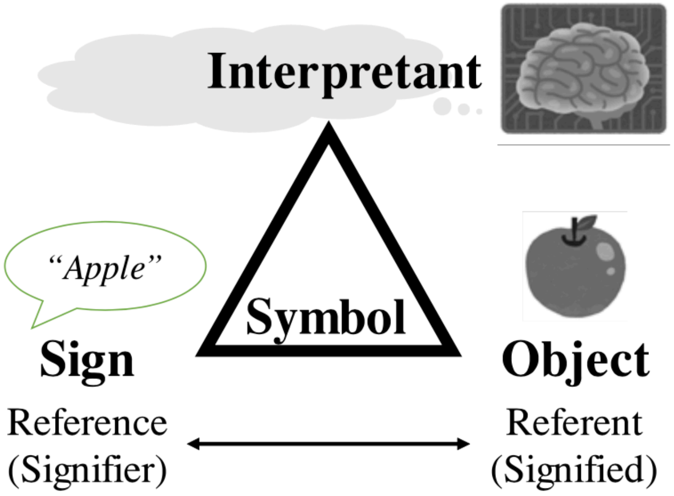
- Objects can be described by explicit sentence-like representations, consisting of tokens
  - Drawback: Reminiscent of Symbolic AI (i.e., physical symbol system hypothesis)
- The tokens are fixed apriori by the system designer
  - Issue: Designer's cognition is outside of the system, system is still open loop
- Since tokens are fixed, shall we assume they are the best?
  - Issue: How do we judge what tokens are best?

Takeaway: If objects can be described as an emergent proto-language, we need emergent vocabulary for the language too.

Approach

Human language developed tokens (words) in an emergent way, this is the study of Semiotics (Chandler 2007, Routledge Basics).

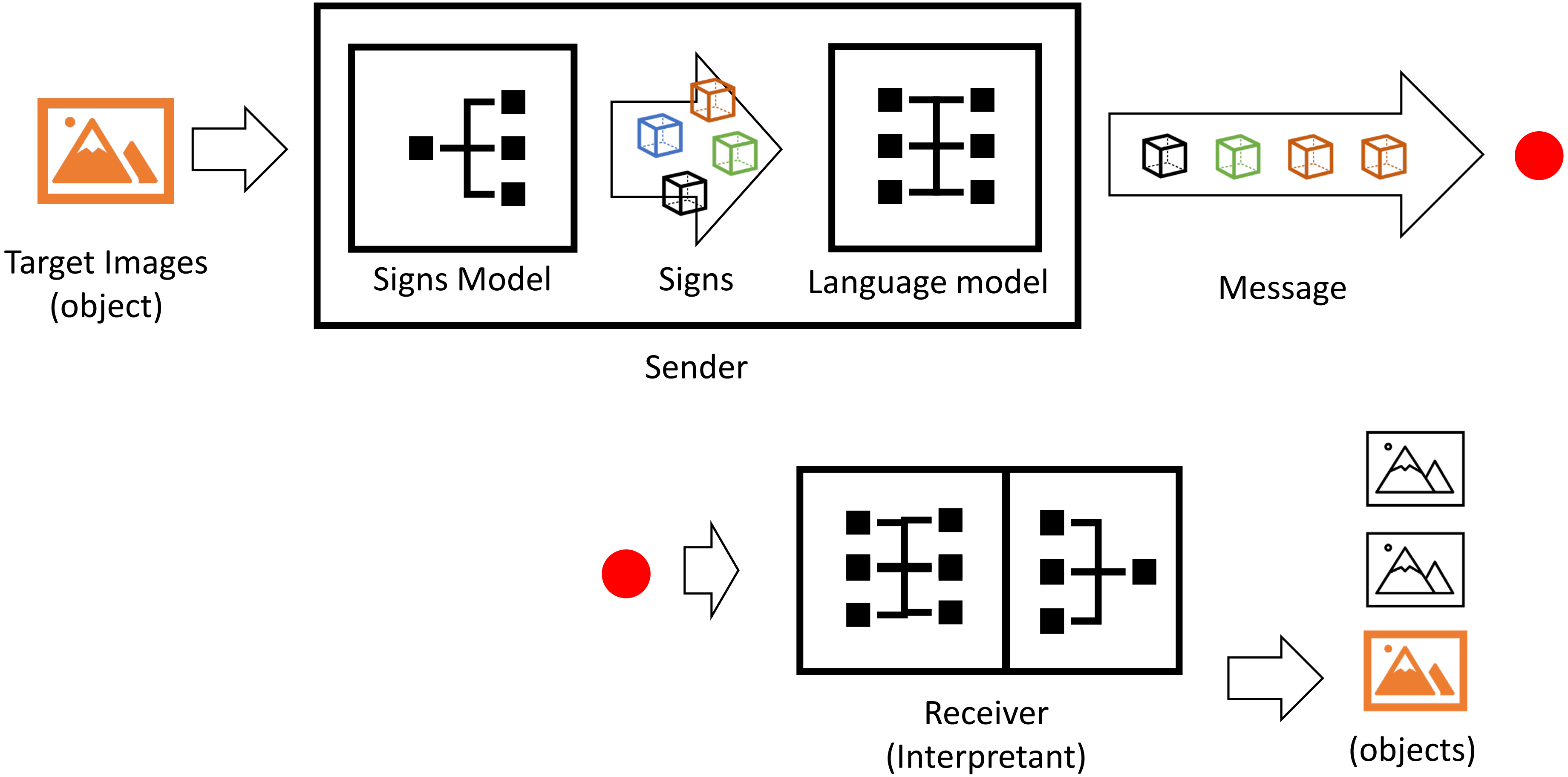
Semiotics define language as symbol systems, symbols are formed by Semiotic triad:



Taniguchi et al. 2019 "Symbol Emergence in Cognitive Developmental Systems"

Approach

Develop the idea of Semiotics within a computational Lewis Signaling game:



Target Images (object)

Signs Model

Signs

Language model

Sender

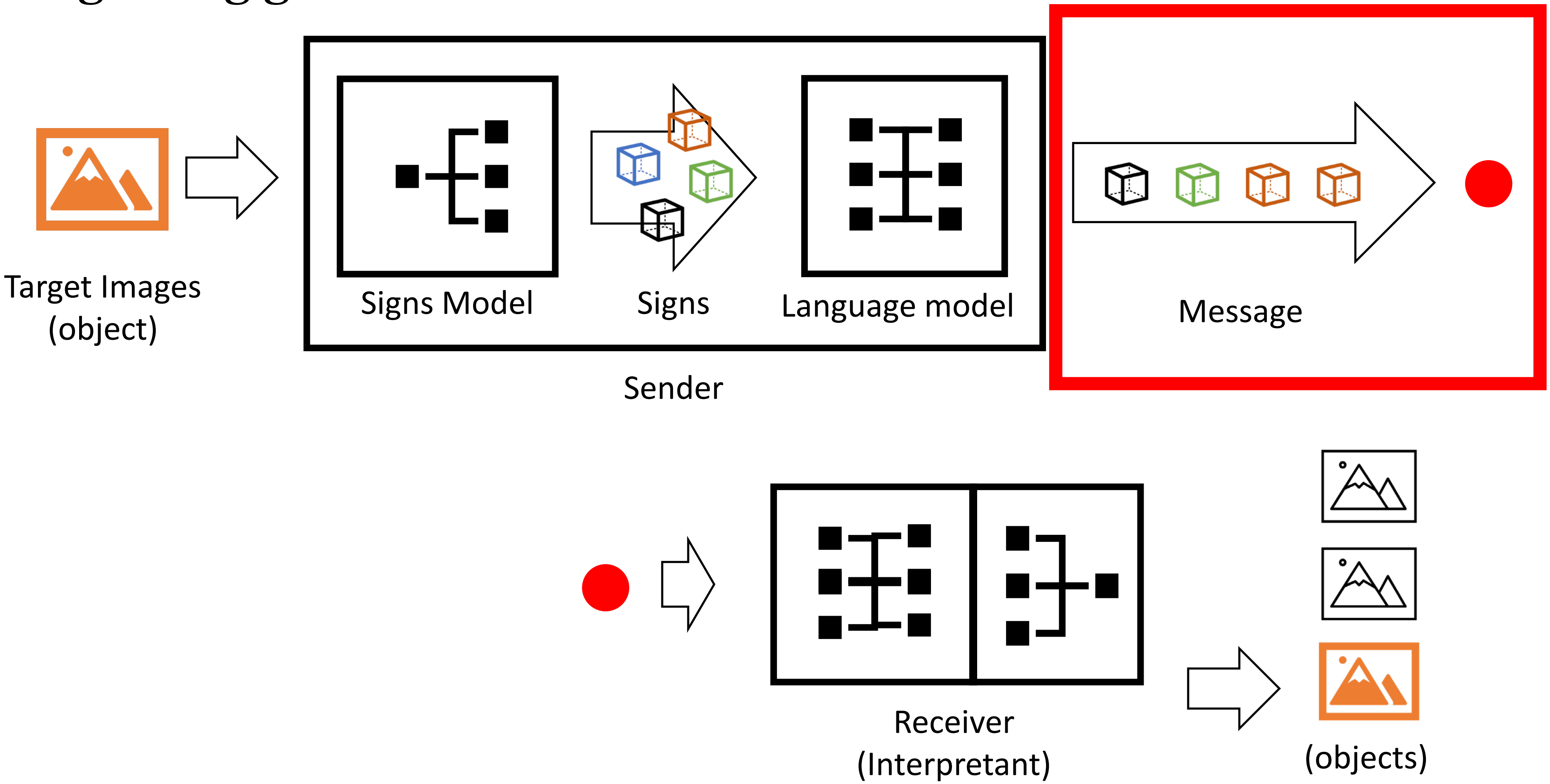
Message

Receiver (Interpretant)

(objects)

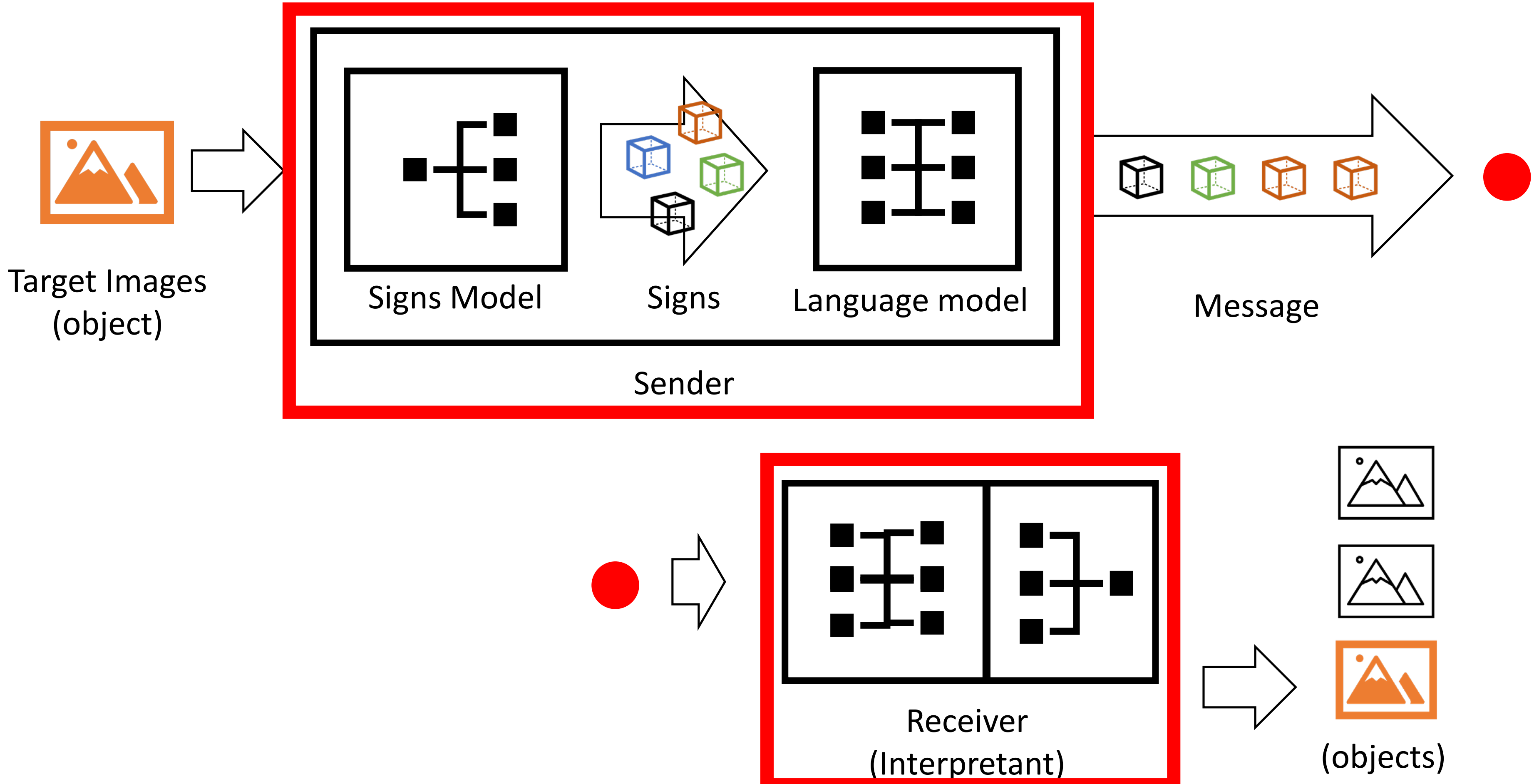
Approach

Develop the idea of Semiotics within a computational Lewis Signaling game:



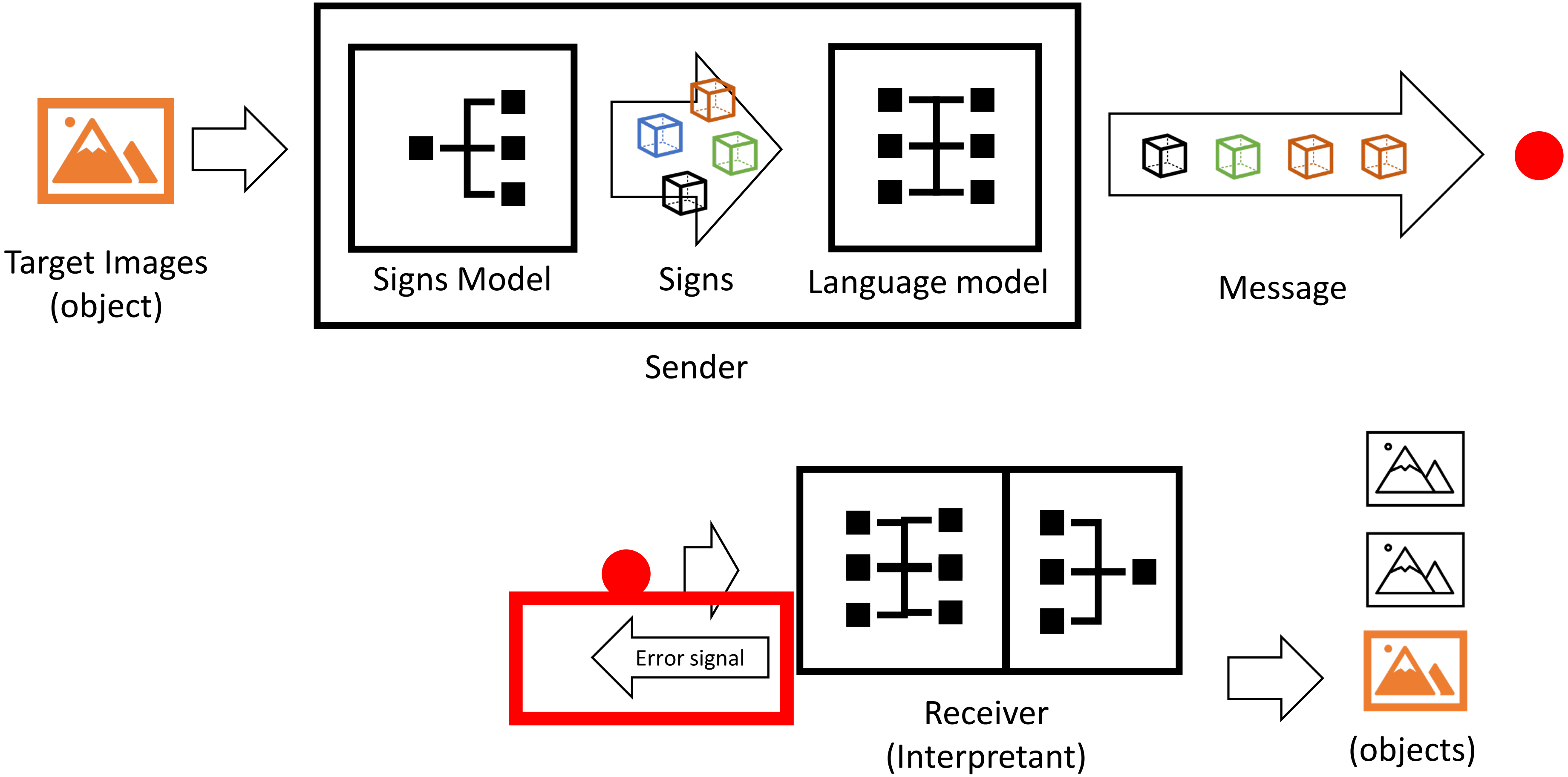
Approach

Develop the idea of Semiotics within a computational Lewis Signaling game:



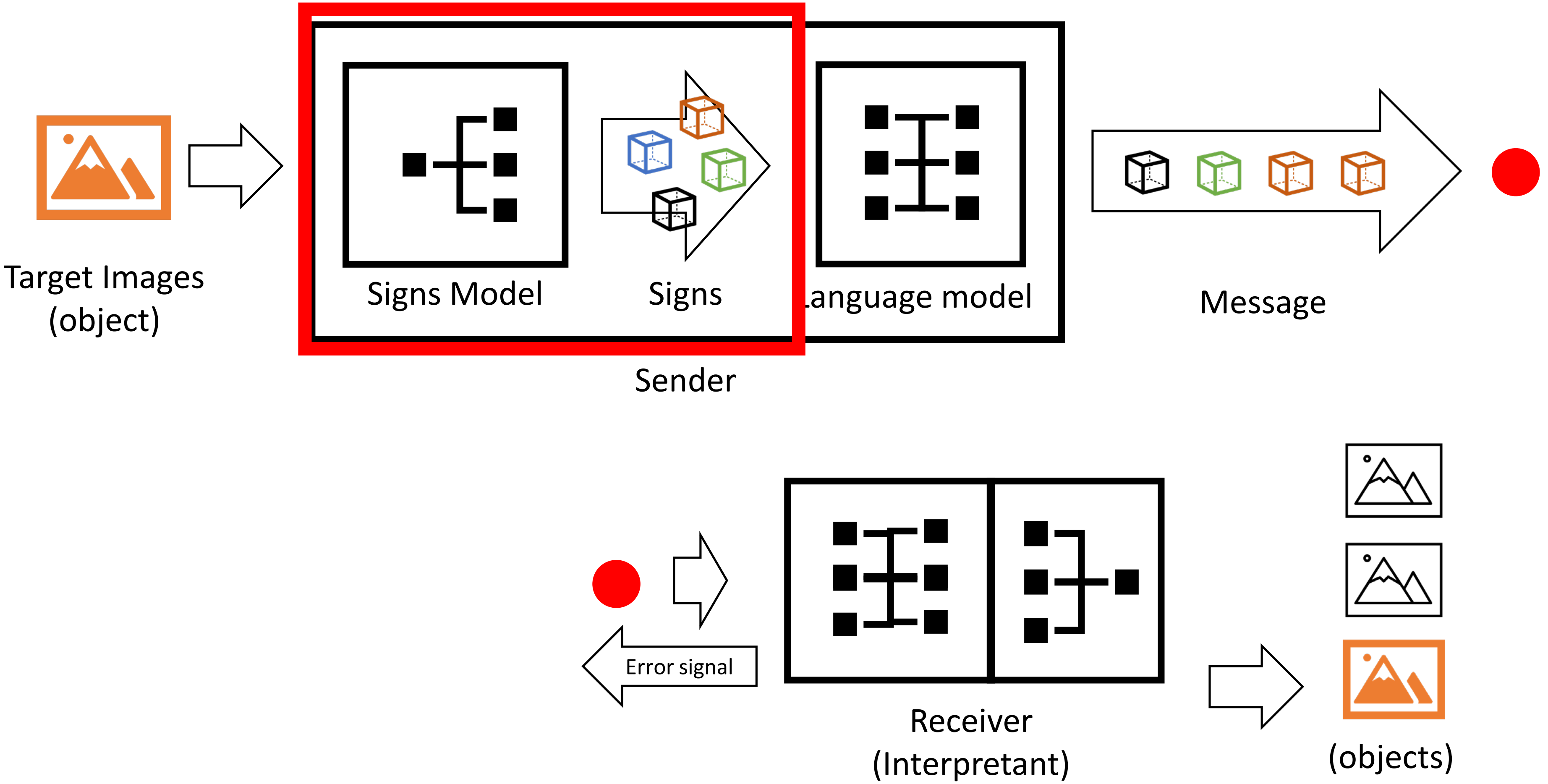
Approach

Develop the idea of Semiotics within a computational Lewis Signaling game:



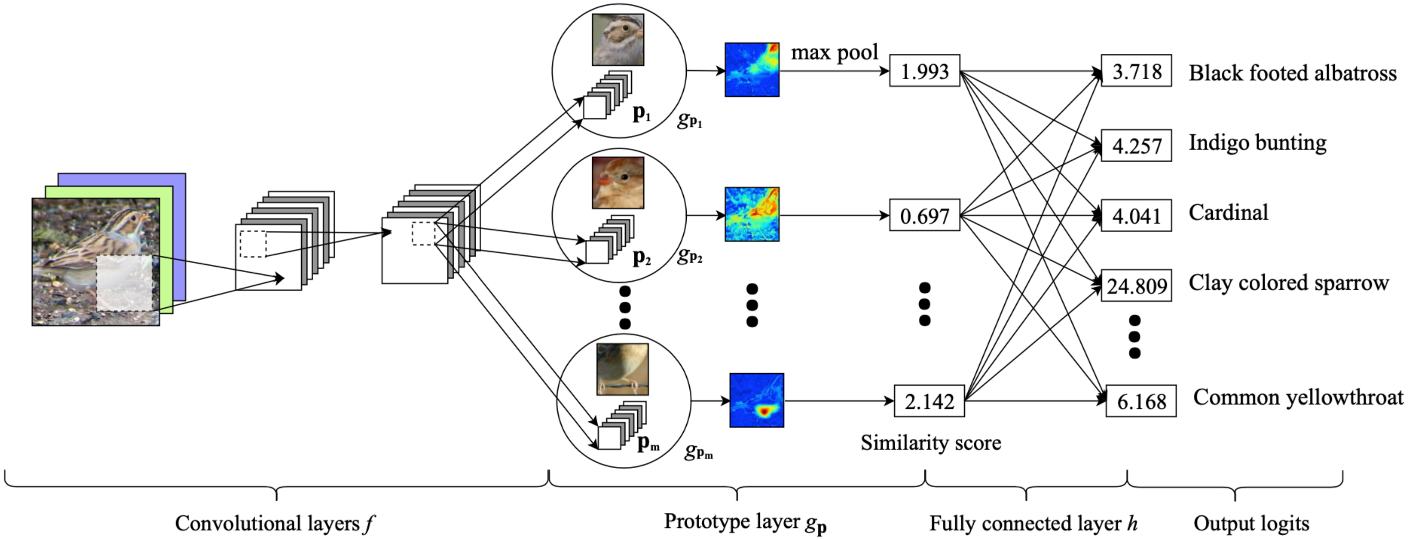
Approach

Develop the idea of Semiotics within a computational Lewis Signaling game:



Signs Modeling

Chen et al. proposed a way to learn “prototypes” from image datasets:

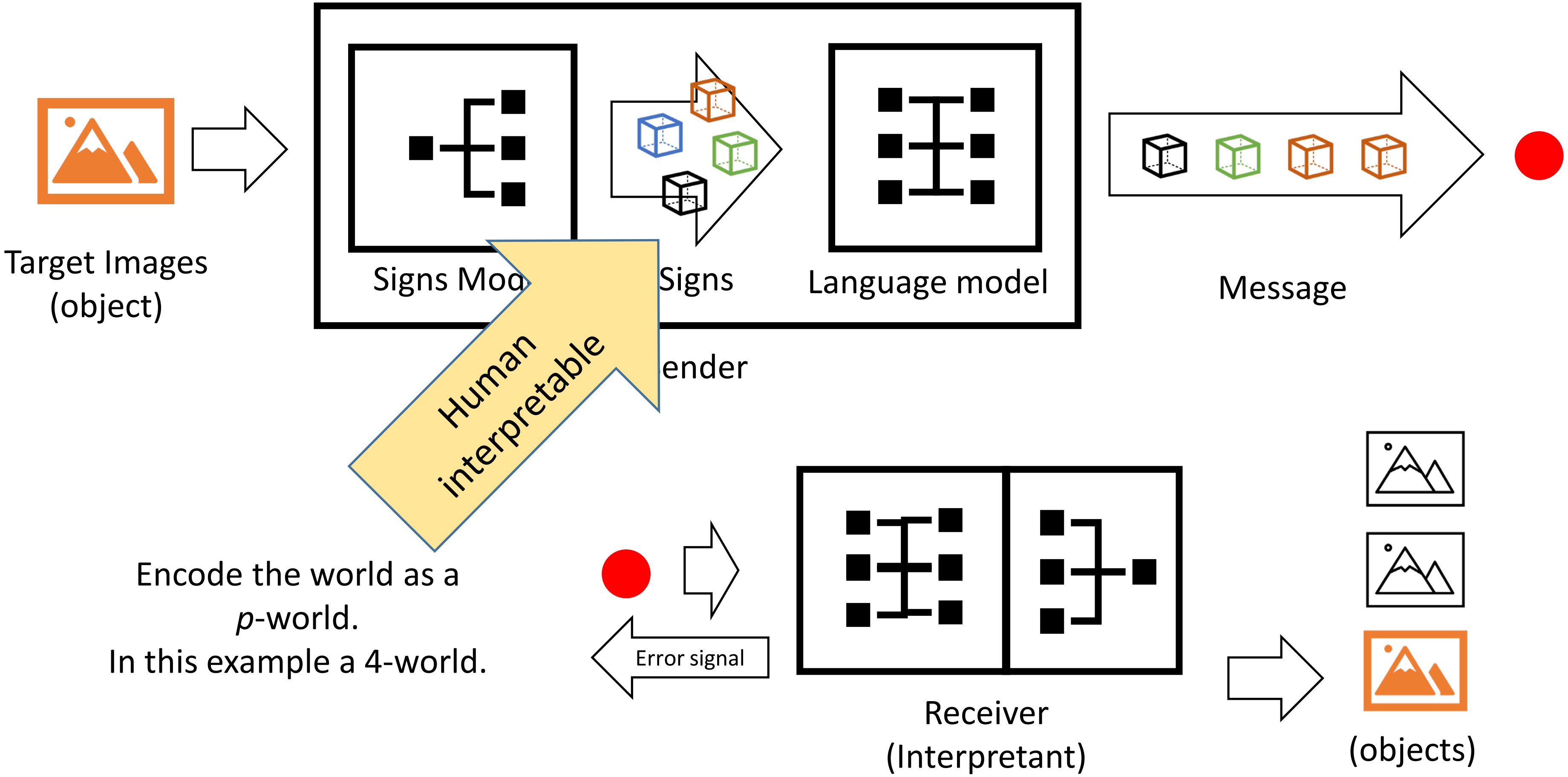


Chen et al. 2018 “This Looks Like That: Deep Learning for Interpretable Image Recognition”

We borrow this method of prototype extraction for signs model.

Approach

Develop the idea of Semiotics within a computational Lewis Signaling game:



Encode the world as a *p*-world.
In this example a 4-world.

Experimental Setup

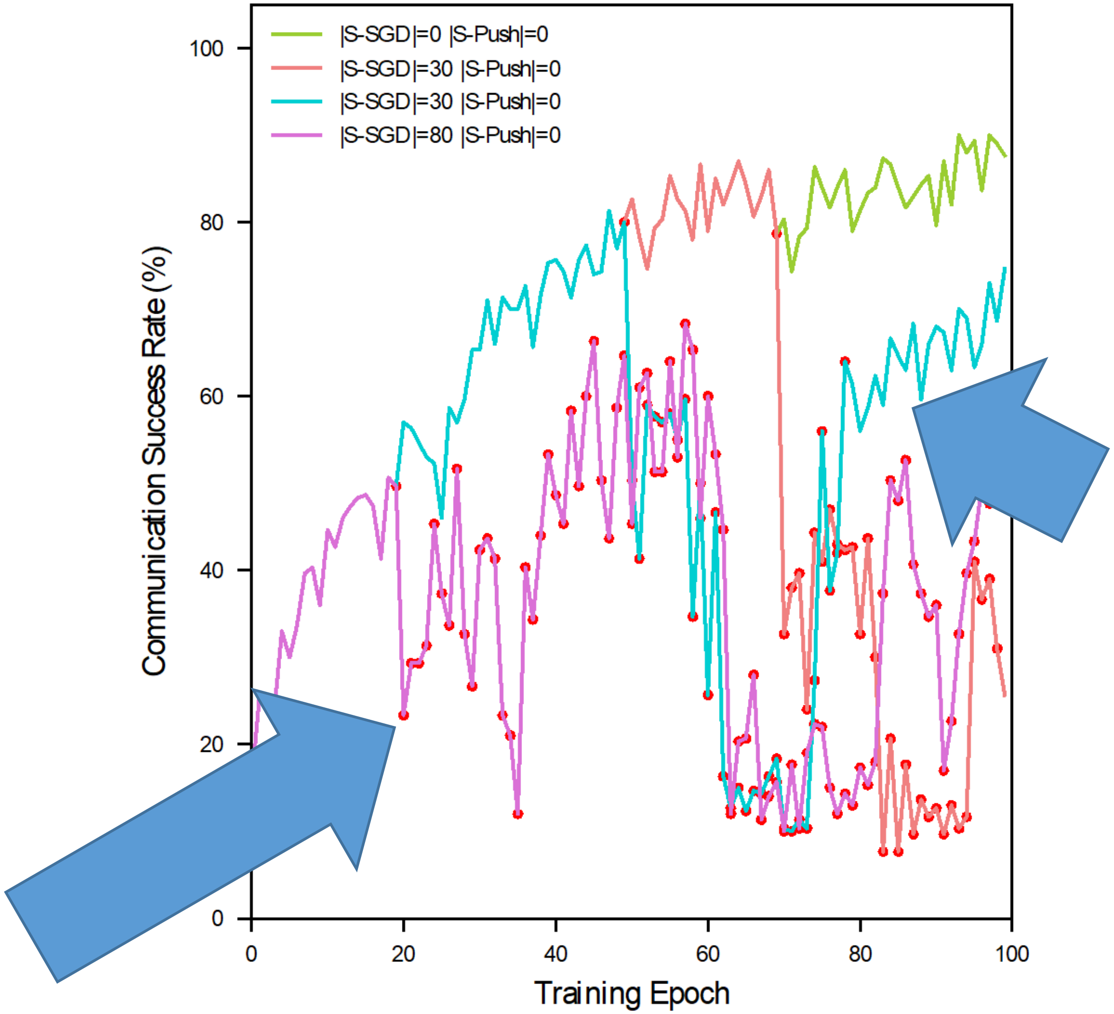
Our experiments approach some open questions about this computational model:

- What is the tradeoff between traditional learning (agents only) and semiotic learning (full end-to-end tuning)
- Does the signs model retain usefulness in the original classifier task?
- How does the social feedback on signs affect the protocol?
  - Topographical Similarity (TS) [Compositionality]
  - *p* positional disentanglement (p-pos) [Compositionality]
  - *p* bag of symbols disentanglement (p-bos) [Compositionality]
  - Entropy

Results

Signaling success:

How early is too early?



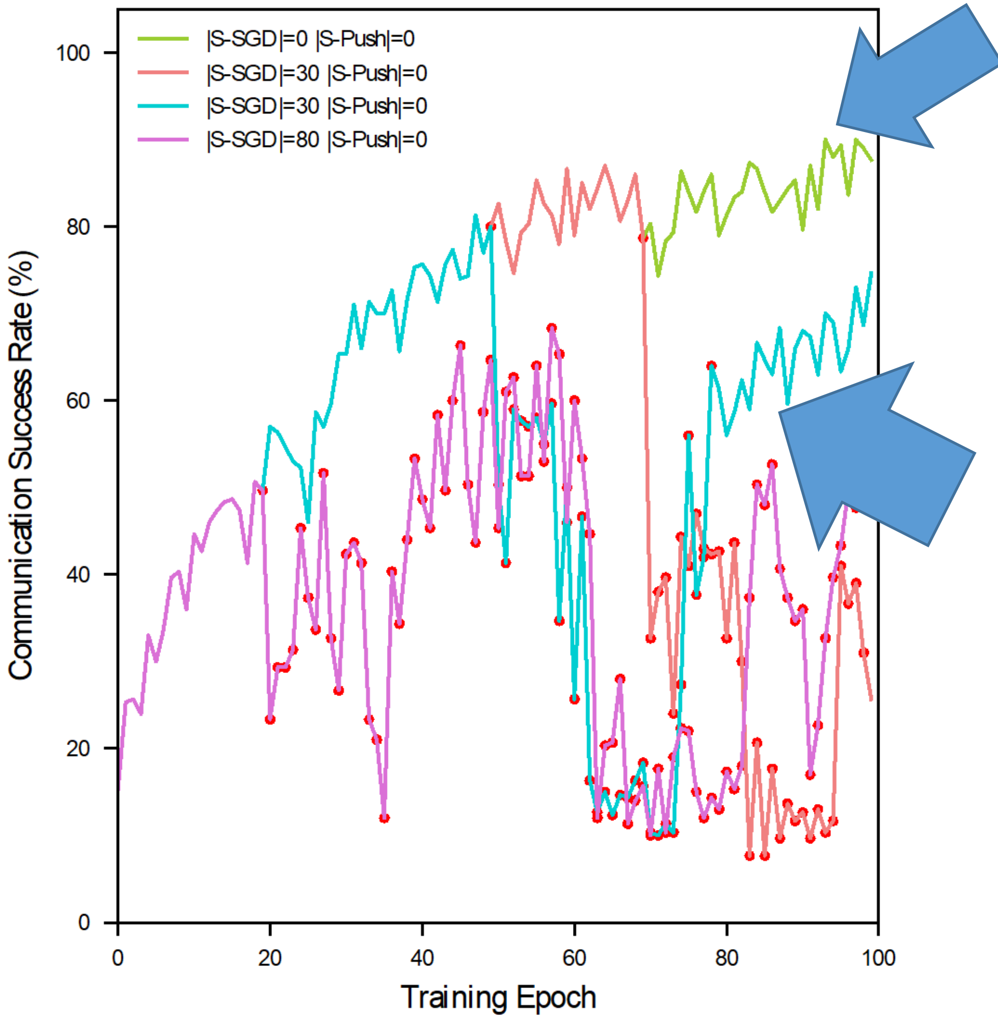
Results

Signaling success:

How early is too early?

Semiotic training vs.

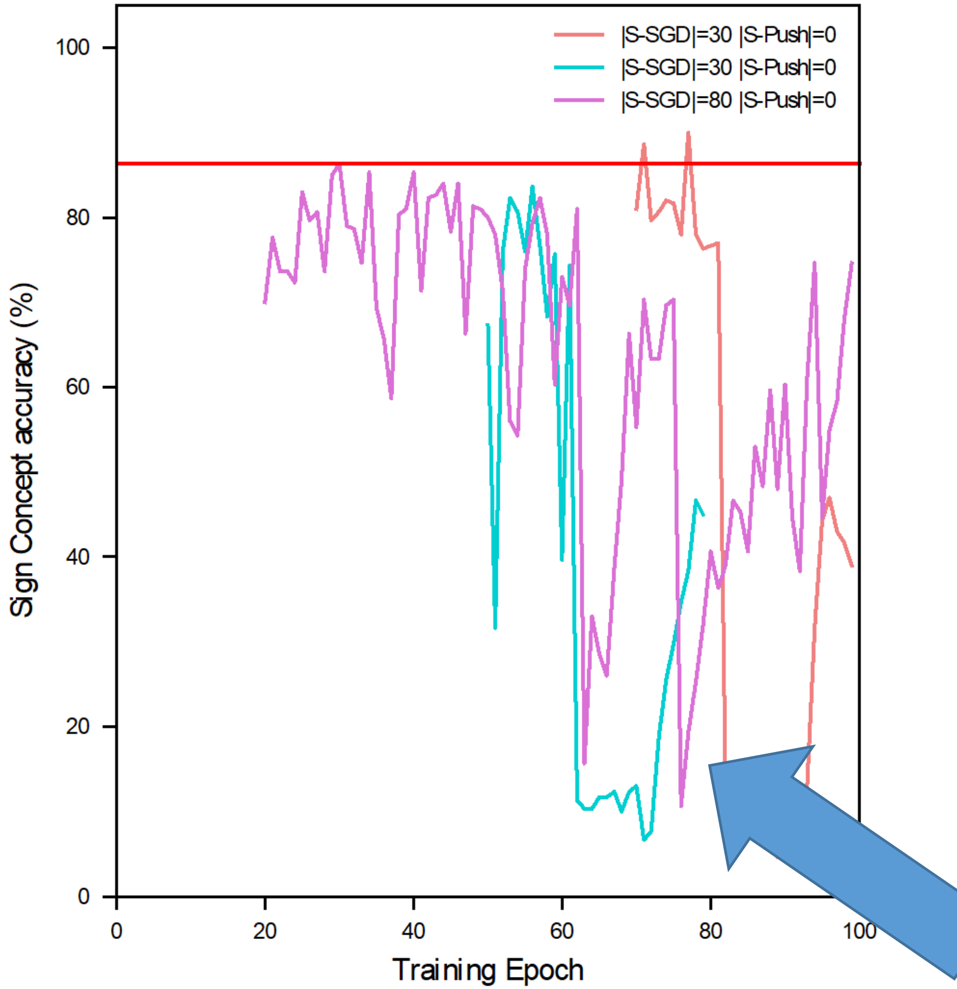
No semiotic training



Results

Original Task Success:

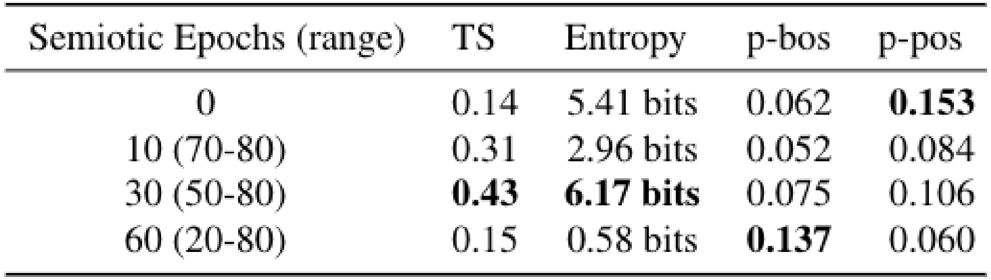
Rapid fall with eventual recovery



Results

Compositionality:

Select models at 80 epochs and run each metric.



Semiotic Epochs (range)	TS	Entropy	p-bos	p-pos
0	0.14	5.41 bits	0.062	**0.153**
10 (70-80)	0.31	2.96 bits	0.052	0.084
30 (50-80)	**0.43**	**6.17 bits**	0.075	0.106
60 (20-80)	0.15	0.58 bits	**0.137**	0.060

TS - Clear advantage to using semiotic training

Entropy – Semiotic in midrange epochs = higher entropy

p-bos – More semiotic epochs = higher p-bos

p-pos – No semiotic epochs …. ?

Lingering Questions

Semiotic training helps us validate prototypes towards signs that are more useful in a proto-language.

The proto-language is arguably more “human-like” due to using interpretable patches of data as signs, but also increasing metrics of compositionality.

Future work:

- Determine why performance is situational
- Ablation studies to narrow down good training strategy
- Extension to RL setting (policy/action grounding): AFRL/ACT3 Summer 2021

Thank you

w.garcia@ufl.edu

UF UNIVERSITY of FLORIDA

Duke UNIVERSITY

TEXAS The University of Texas at Austin

UC SANTA CRUZ

A Causal View on Manifold-Gradient Mutual Information

Washington Garcia

Shantanu Ghosh

Kevin Butler

(UF)

Our Previous Work

Approximation of Gradient-Manifold Mutual Information for a robust model

$$I(\mathcal{G},\mathcal{M})_\epsilon = \frac{2}{\sqrt{2\pi\sigma^2}}\sum_{i=1}^{||\mathcal{M}^+||}\exp(-\frac{(x_i^*-\theta)^2}{2\sigma^2})\cdot\beta_i^+\Delta_i + \frac{2}{\sqrt{2\pi\sigma^2}}\sum_{i=1}^{||\mathcal{M}^+||}\exp(-\frac{(x_i^*+\theta)^2}{2\sigma^2})\cdot\beta_i^-\Delta_i.$$

where

$$\beta_i^+ = -\frac{(x_i^*-\theta)^2}{2\sigma^2} - \log(\lambda_+) - \log\left(\exp(-\frac{(x_i^*-\theta)^2}{\sigma^2}) + \exp(-\frac{(x_i^*+\theta)^2}{\sigma^2})\right),$$

$$\beta_i^- = -\frac{(x_i^*+\theta)^2}{2\sigma^2} - \log(\lambda_-) - \log\left(\exp(-\frac{(x_i^*-\theta)^2}{\sigma^2}) + \exp(-\frac{(x_i^*+\theta)^2}{\sigma^2})\right),$$

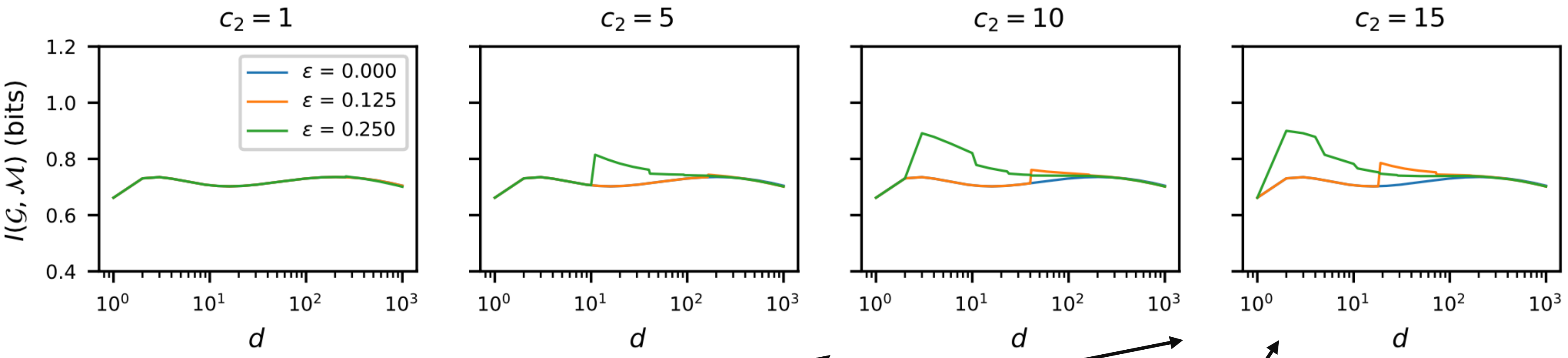
$$\lambda_+ = \frac{1}{\sqrt{2\pi\sigma}}\sum_{j=1}^{n}\exp\left(-\frac{1}{2}\cdot\frac{(x_j^*-\theta)^2}{\sigma^2}\right)\Delta_i, \qquad \text{(marginal for g = 1)}$$

$$\lambda_- = \frac{1}{\sqrt{2\pi\sigma}}\sum_{j=1}^{n}\exp\left(-\frac{1}{2}\cdot\frac{(x_j^*+\theta)^2}{\sigma^2}\right)\Delta_i, \qquad \text{(marginal for g = −1)}$$

for $\Delta_i = x_i - x_{i-1}$ and positive $x_i^* \in [x_{i-1}, x_i]$, $x_j^* \in [x_{j-1}, x_j]$

Our Previous Work

Consider three ϵ-robust models: $\epsilon \in \{0, 0.125, 0.250\}$

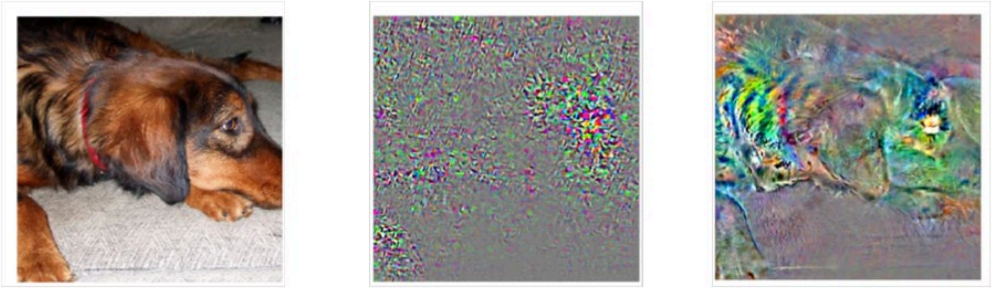


Lower dim, Higher c_2 = Higher “clean” MI

Greatest MI with ”most robust” ϵ setting (green vs. orange)

Discussion

Our experimental results potentially explain the phenomena empirically observed by Engstrom et al. (2018) and Tsipras et al. (2018):

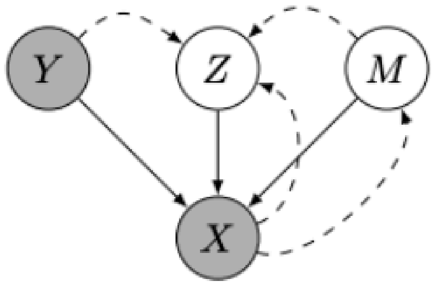


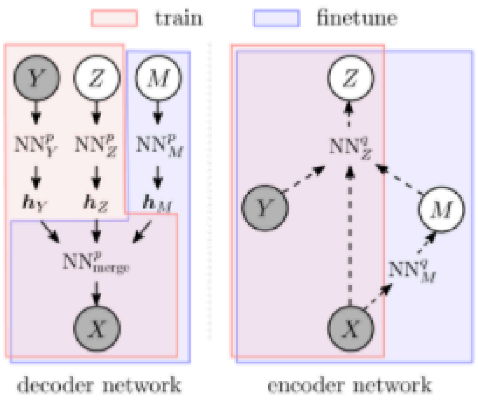
“Adversarial Robustness: Theory and Practice. Part 4”, Madry and Kolter

Higher “clean” MI = Leakage of semantic priors

Background

DEEP CAMMA MODEL





ELBO for clean data(do(m=0))

$$\text{ELBO}(x, y) := \mathbb{E}_{q_\phi(z,m|x,y)}\left[\log \frac{p_\theta(x, y, z, m)}{q_\phi(z, m|x, y)}\right]$$

ELBO for clean data(do(m))

$$\text{ELBO}(x, y, do(m = 0)) := \mathbb{E}_{q_{\phi_1}(z|x,y,m=0)}\left[\log \frac{p_\theta(x|y, z, m = 0)p(y)p(z)}{q_{\phi_1}(z|x, y, m = 0)}\right]$$

Complete loss objective

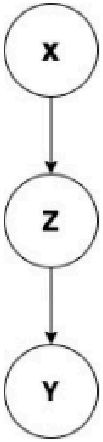
$$\mathcal{L}_{\text{aug}}(\theta, \phi) = \lambda \mathbb{E}_{\mathcal{D}}[\text{ELBO}(x, y, do(m = 0))] + (1 - \lambda)\mathbb{E}_{\mathcal{D}'}[\text{ELBO}(x, y)]$$

Prediction

$$p(y^*|x^*) = \frac{p(x^*|y^*)p(y^*)}{p(x^*)} \approx \text{softmax}_{c=1}^{C}\left[\log \sum_{k=1}^{K} \frac{p_\theta(x|y, z_c^k, m^u)p(y_c)p(z)}{q_{\phi_1}(z_c^k|x^*, y_c, m^u)}\right]$$

Deep VIB & Mutual Information

DEEP VARIATIONAL INFORMATION BOTTLENECK



- Predict Y from X
- Find useful representations Z from X to predict Y
- Learn to maximize mutual information between Z and Y such that the dependency between X and Z should be less by minimizing the mutual information between X and Z

Loss function to optimize

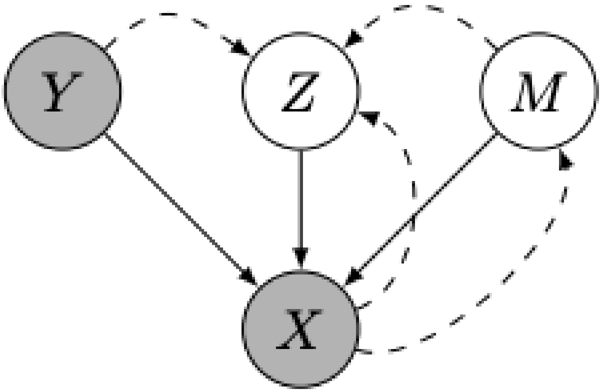
$$\max_{\boldsymbol{\theta}} I(Z,Y;\boldsymbol{\theta}) \text{ s.t. } I(X,Z;\boldsymbol{\theta}) \leq I_c$$

$$R_{IB}(\boldsymbol{\theta}) = I(Z,Y;\boldsymbol{\theta}) - \beta I(Z,X;\boldsymbol{\theta})$$

Using variational inference

$$J_{IB} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{\epsilon \sim p(\epsilon)} \left[-\log q(y_n|f(x_n,\epsilon))\right] + \beta \, \mathrm{KL}\left[p(Z|x_n), r(Z)\right]$$

Current Approach



Represent M from X so that mutual information between M and X should be low
(M being the features that can be manipulated)

Represent M from Z so that mutual information between M and X should be high
(Causal adversarial training with MI)

Thank you

w.garcia@ufl.edu
shantanughosh@ufl.edu