Disentangling the Role of Categorization in Multi-agent Emergent Communication

Washington Garcia (UF)

Kevin Butler (UF)

Scott Clouse (AFRL/ACT3)

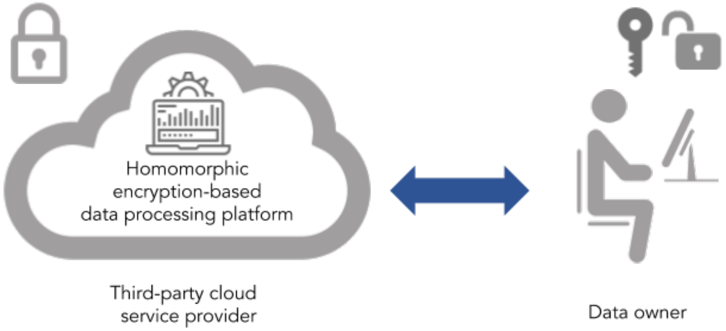
- Work in progress:
  - Plausible deniability and differential privacy in eye tracking (submitted to IEEE VR'22)
  - On-manifold ddversarial examples in hard-label scenarios (submitted to ICLR'22)
  - FHE for ROS (early work in progress)
  - Adversarial learning for counterfactual prediction (submitted to Nature Machine intelligence)
  - Emergent communication (to be submitted to ACL'22, *this talk*)
- Collaborators (UF unless otherwise noted):
  - **Washington Garcia, Caroline Fedele**, Brendan David-John, Eakta Jain, Washington Garcia, Aaditya Prakash, Somesh Jha (UW), Pin-Yu Chen (IBM), ***Scott Clouse (AFRL/ACT3)***

Computation Privacy in ROS

- Assuring **computation privacy** with fully homomorphic encryption (FHE)
  - allows for arbitrary operations on encrypted data
- FHE capability on aerial/limited-resource systems (UAVs)
  - secure outsourcing of crypto/computationally expensive tasks
  - secure robotic operations by integrating FHE into ROS operations

Method:

- Palisade crypto toolkit for FHE, integrated with ROS software
- NVIDIA (ARM) Xavier AGX boards
  - Testing various FHE schemes to determine optimal approach for different data types
  - Testing masked communication, navigation, other ROS applications



/autonomous agent

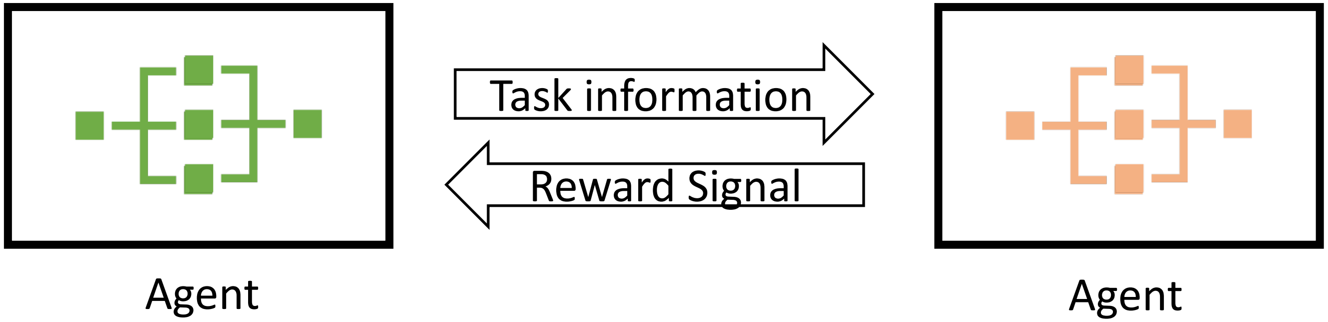
https://miro.medium.com/max/618/1*NfpI6c7Uk93-sRzWyktFjw.png

Motivation

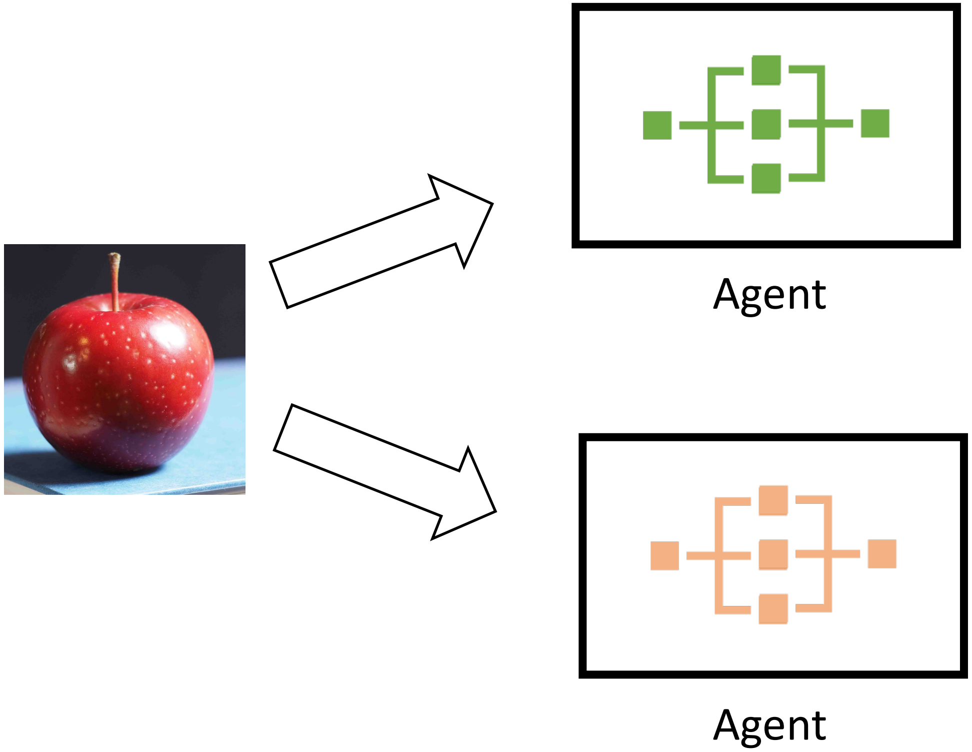
Machine agents are hoped to eventually communicate with each other in what is termed a “machine culture”.

To enable cooperation, machine agents should understand the intention of another agent’s utterances.

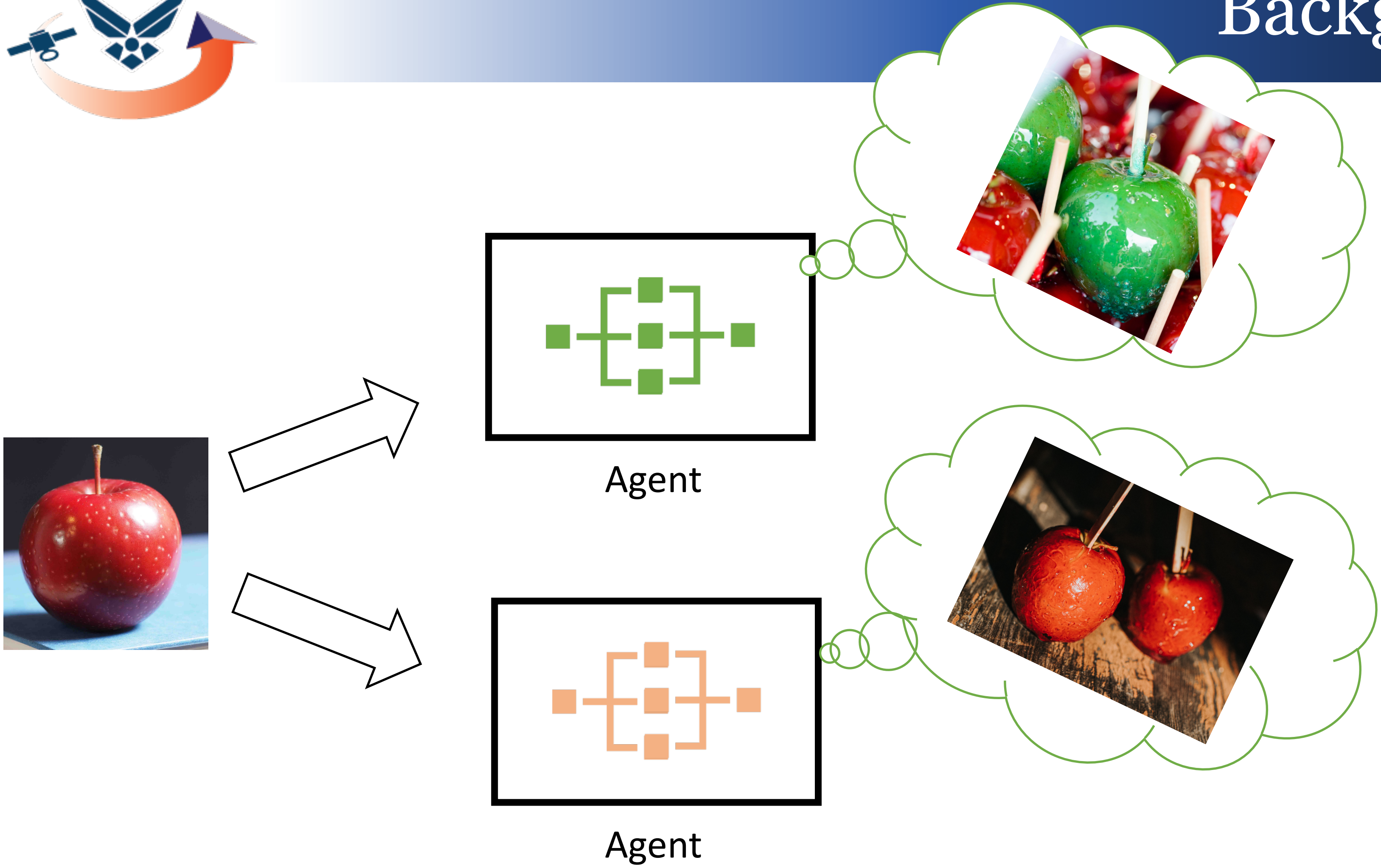
However, in a heterogenous multi-agent system, what level of understanding do we really need?



Background



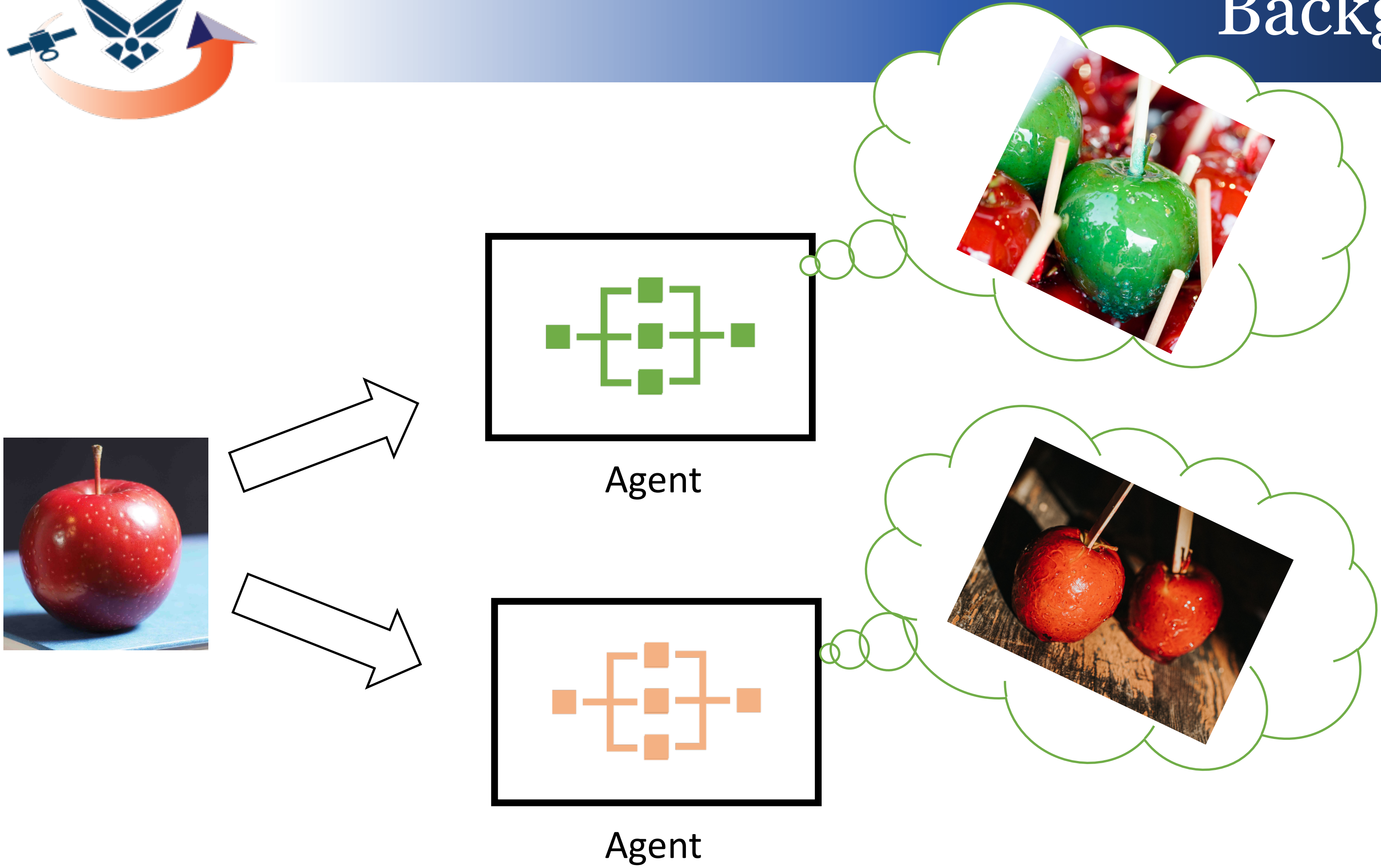
Background



Agent

Agent

Background

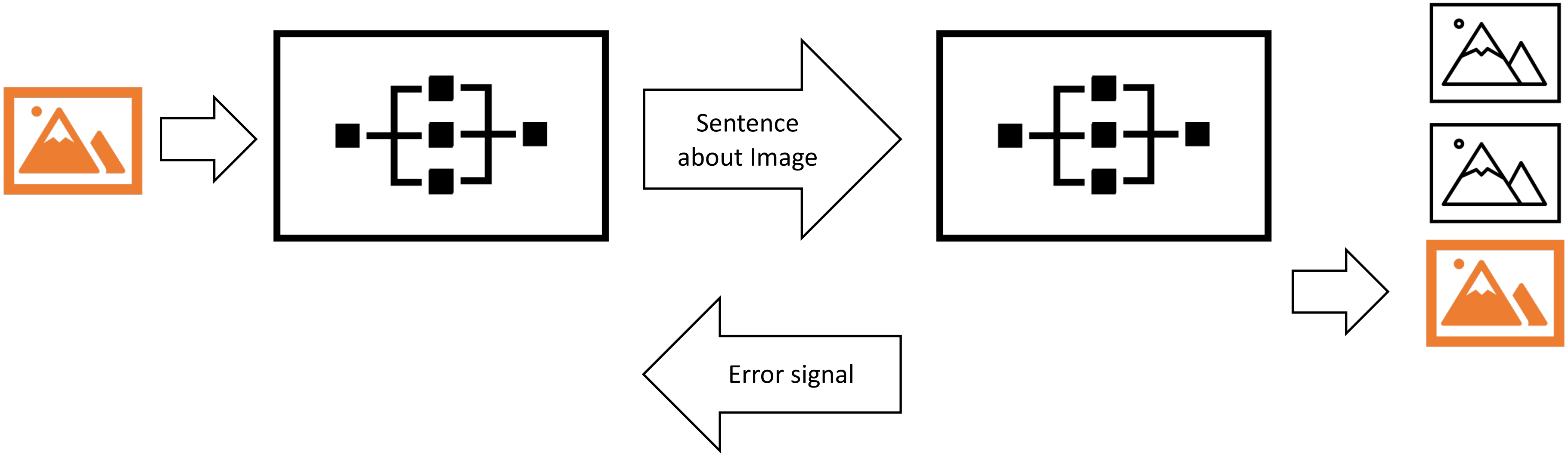

Agent

Agent

Takeaway: If agents can have different reasoning processes, can they still communicate?

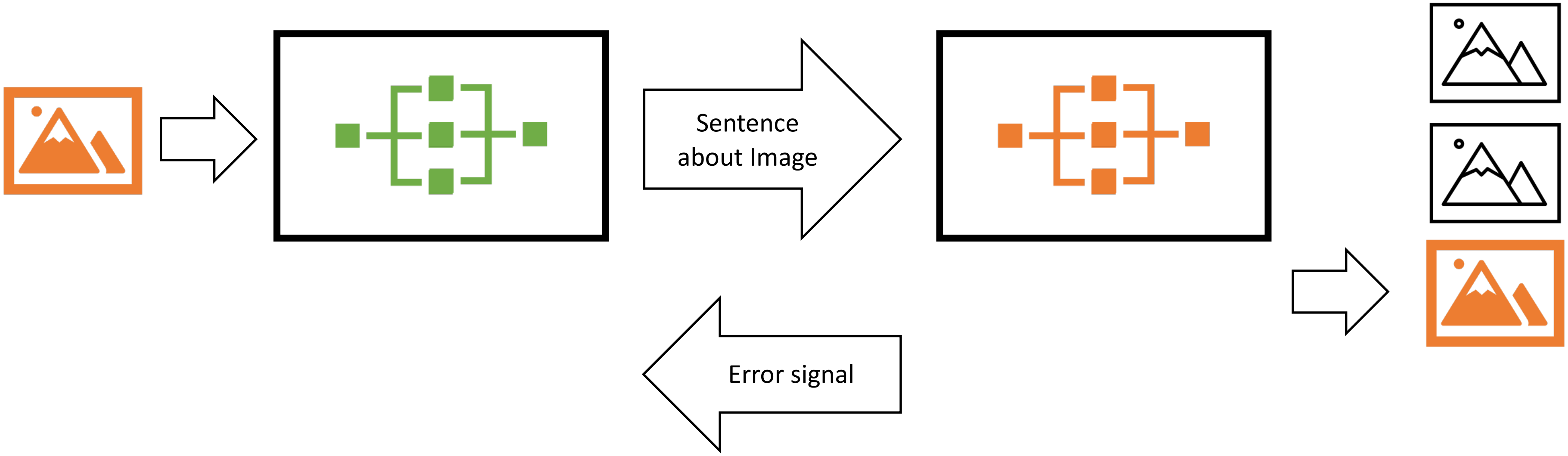
Approach

We consider two agents playing a Lewis Signaling Game:



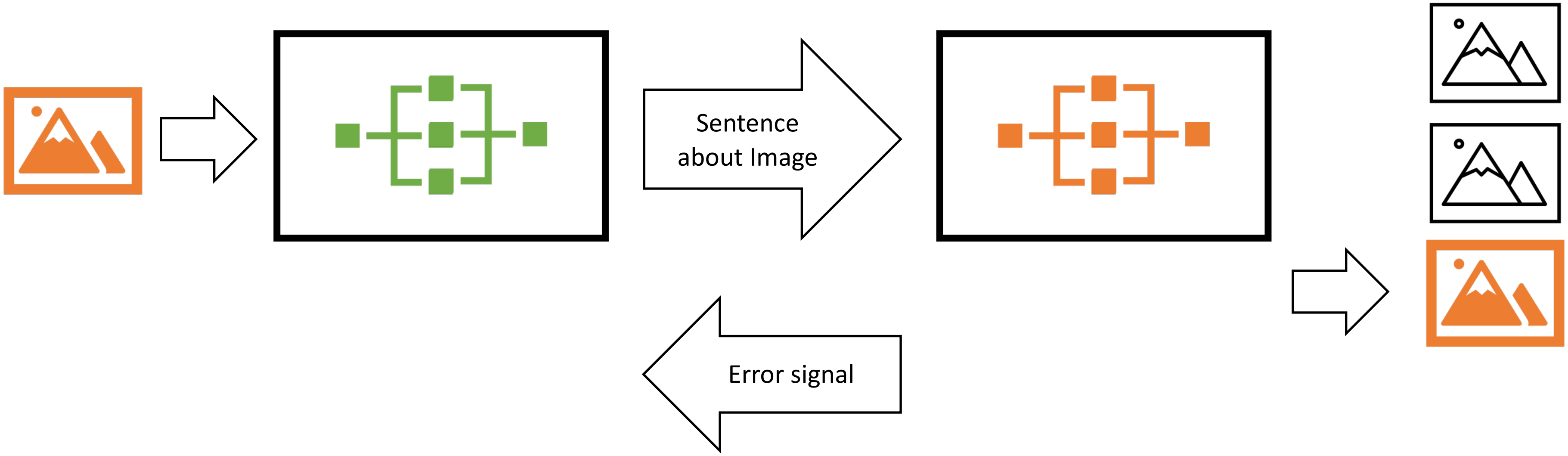
Approach

We consider two agents playing a Lewis Signaling Game:



Approach

We consider two agents playing a Lewis Signaling Game:



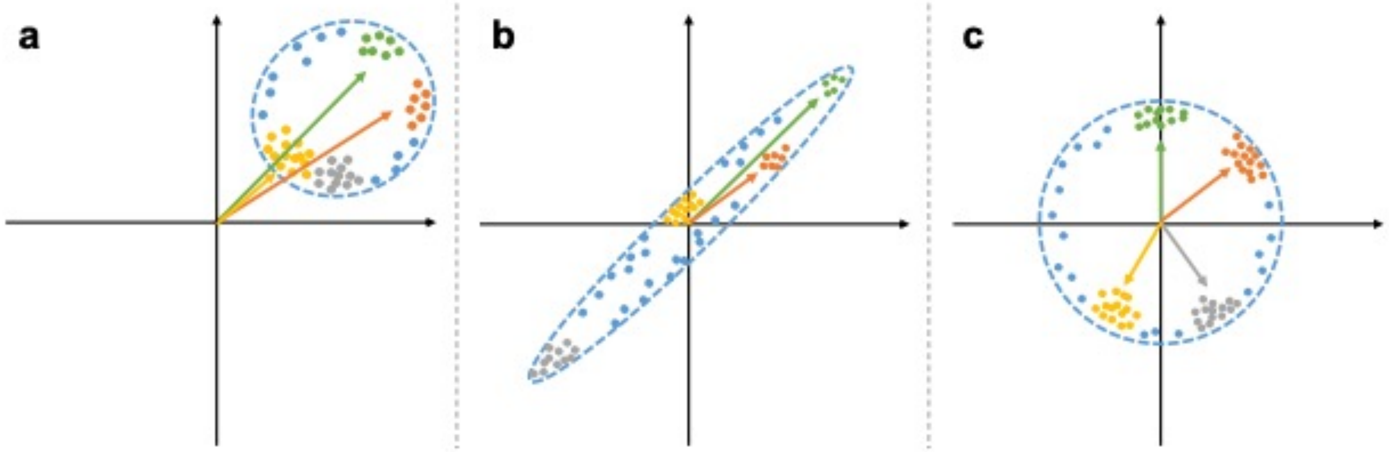
Beyond model weights or architectures, how do we know the agent is reasoning in a certain way?

Approach

Disentangled representations (DR) enable tuning the reasoning process.

DR generally split the learning domain into *k* concept classes (which can be different from dataset classes).

-> learn latent representation with concept separation



Concept Whitening for Interpretable Image Recognition (Chen et al. 2020)

Approach

Disentangled representations (DR) enable tuning the reasoning process.

DR generally split the learning domain into *k* concept classes (which can be different from dataset classes).

-> learn latent representation with concept separation

ProtoPNet (Chen et al. 2018)

- **Unsupervised** disentanglement - using prototypical image patches from the data to represent concepts.

Concept Whitening (CW) (Chen et al. 2020)

- **Supervised** disentanglement - using pre-defined concept examples.

Experimental Setup

We consider three reasoning processes:

- Traditional Conv. Nets (e.g., VGG, **ResNet**, DenseNet)
- ProtoPNet (Chen et al. 2018)
- Concept Whitening (CW) (Chen et al. 2020)

Study the effect of tuning the concept realization of each agent with respect to the communication success rate on two datasets:

1. 10-class subset of UCSD Birds dataset (CUB10)
2. 64-class mini-ImageNet for few-shot Learning (MI64)

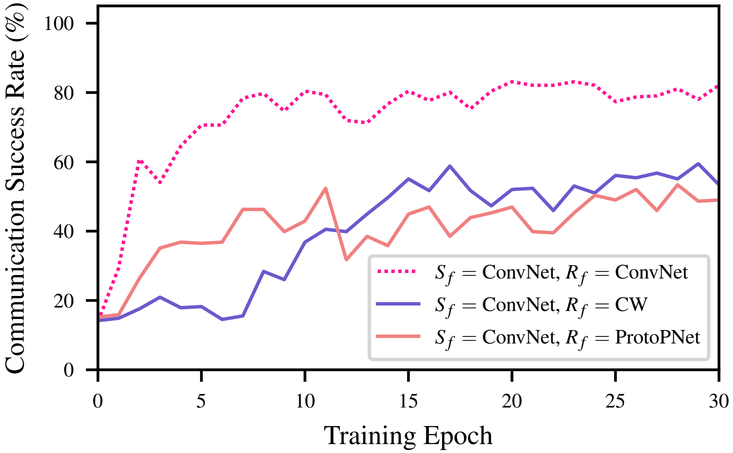
   -> MI64 allows studying interaction on unseen objects.

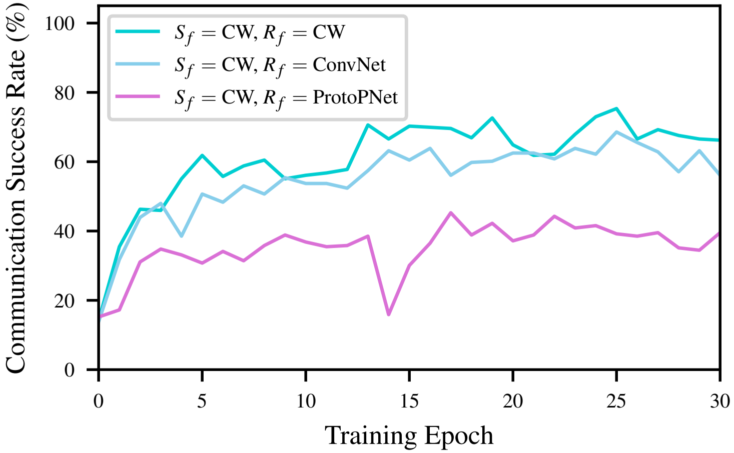
Results

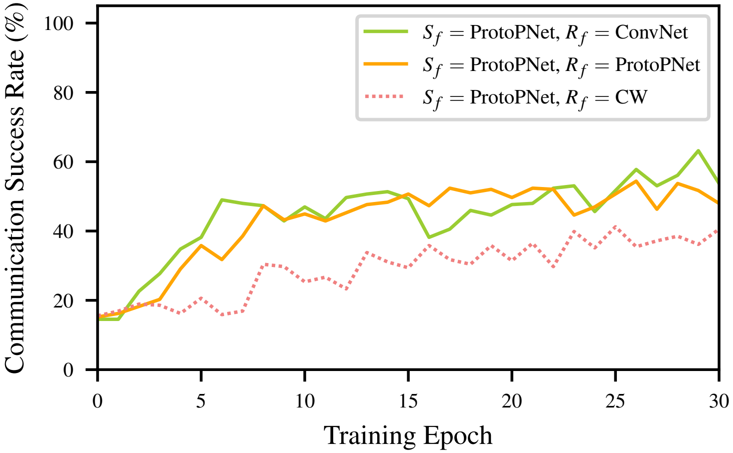
Signaling success (CUB10):

S_f = Sender model

R_f = Recv. model





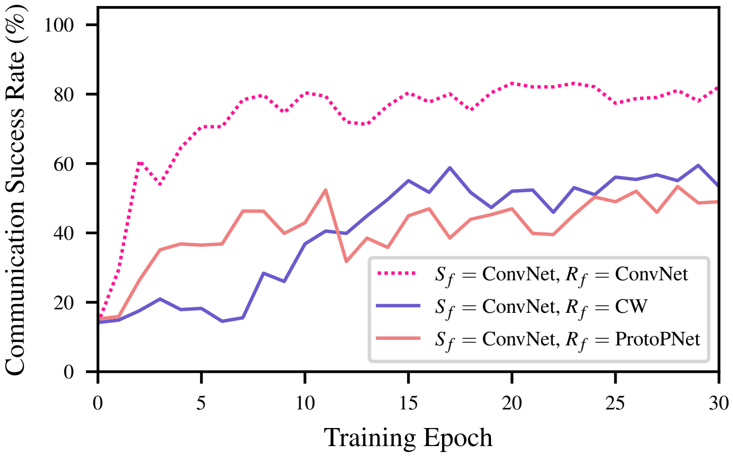


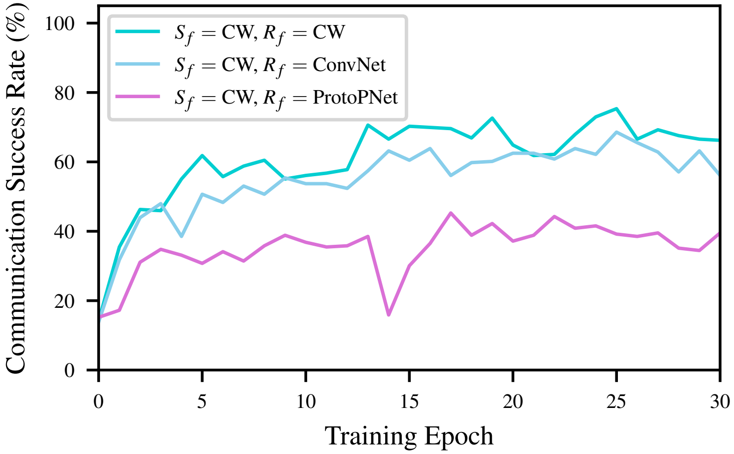
Results

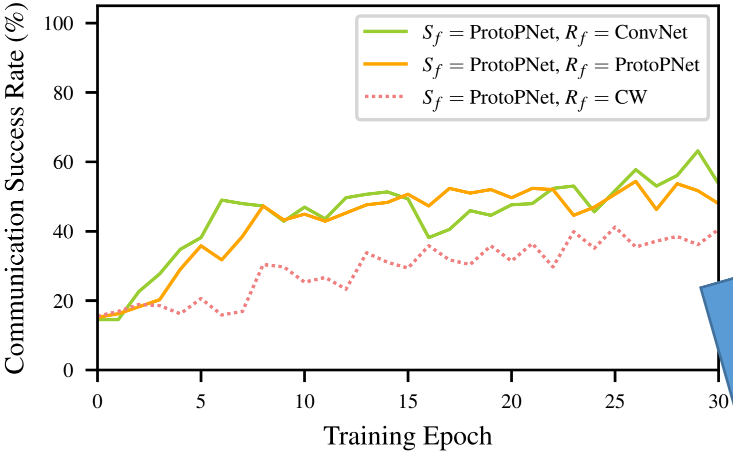
Signaling success (CUB10):

S_f = Sender model

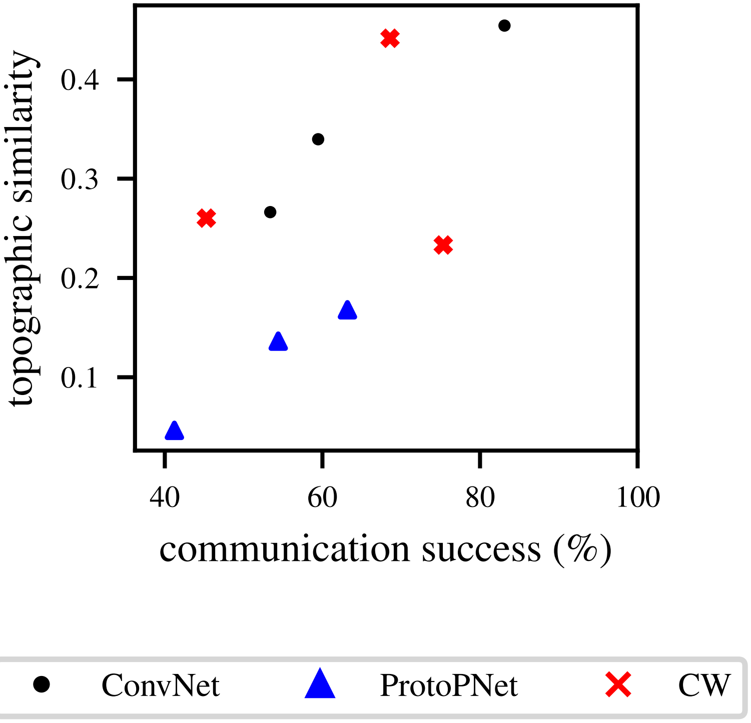
R_f = Recv. model





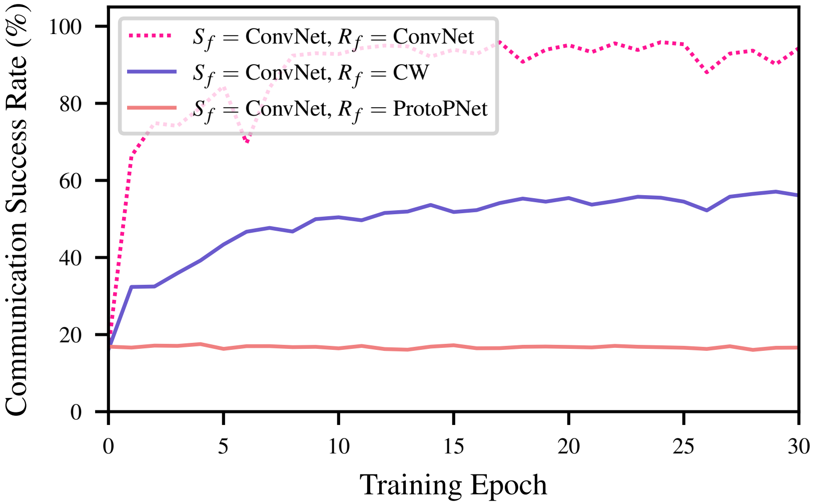


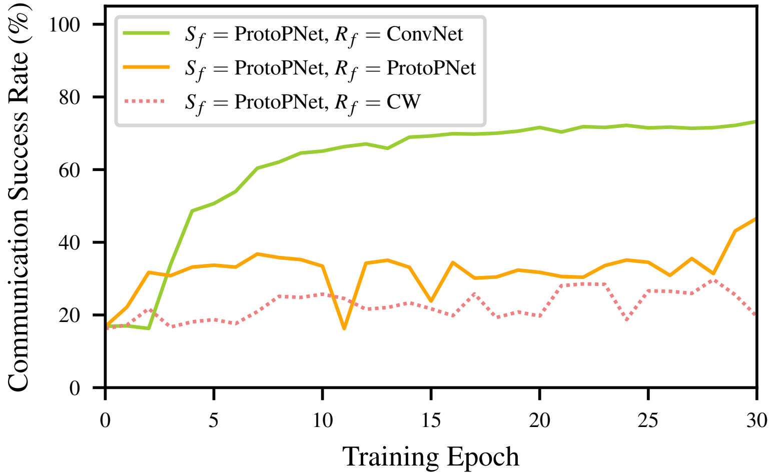
Protocol topographic similarity (score of re-using tokens in the messages to describe similar objects) – CUB10:

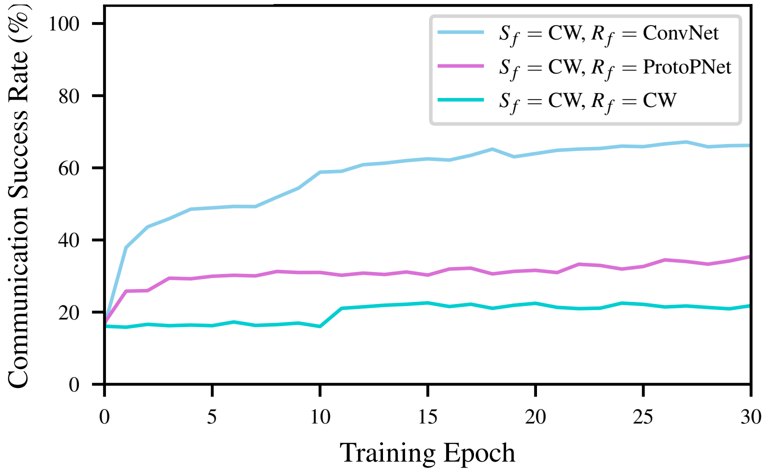


Results

What if the agents *have never seen the objects before*? MI64 dataset:

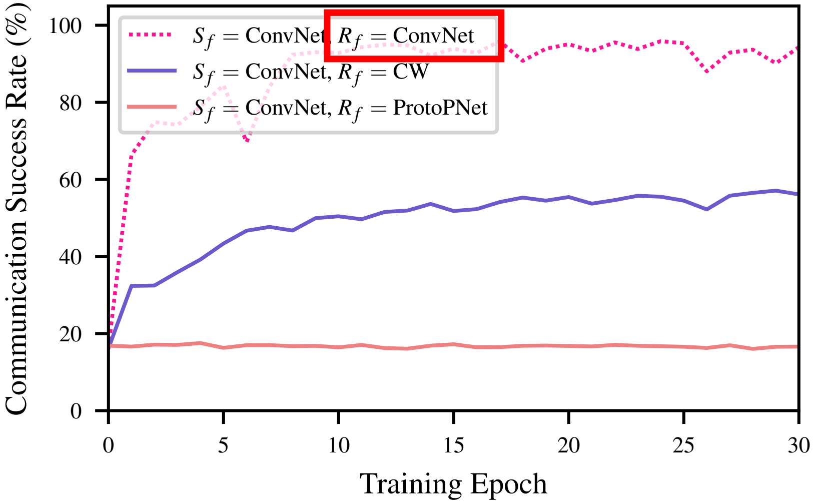


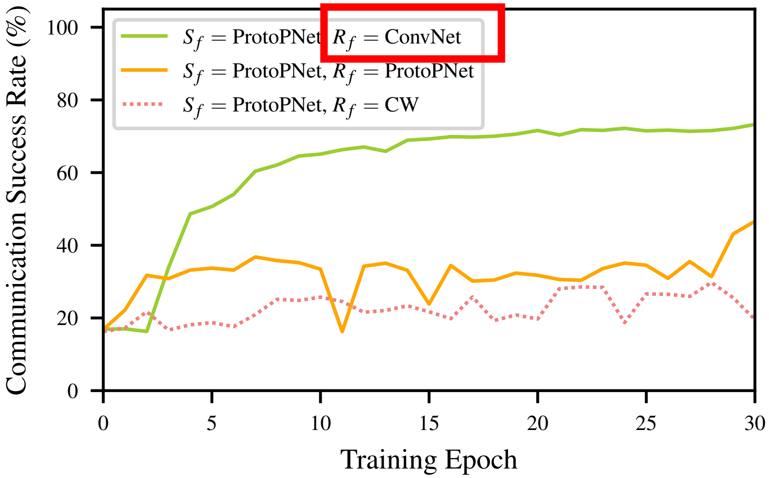


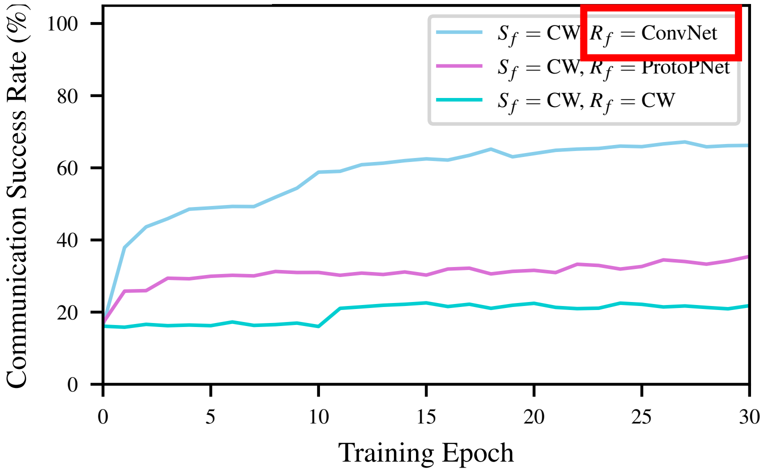


Results

What if the agents *have never seen the objects before*? MI64 dataset:

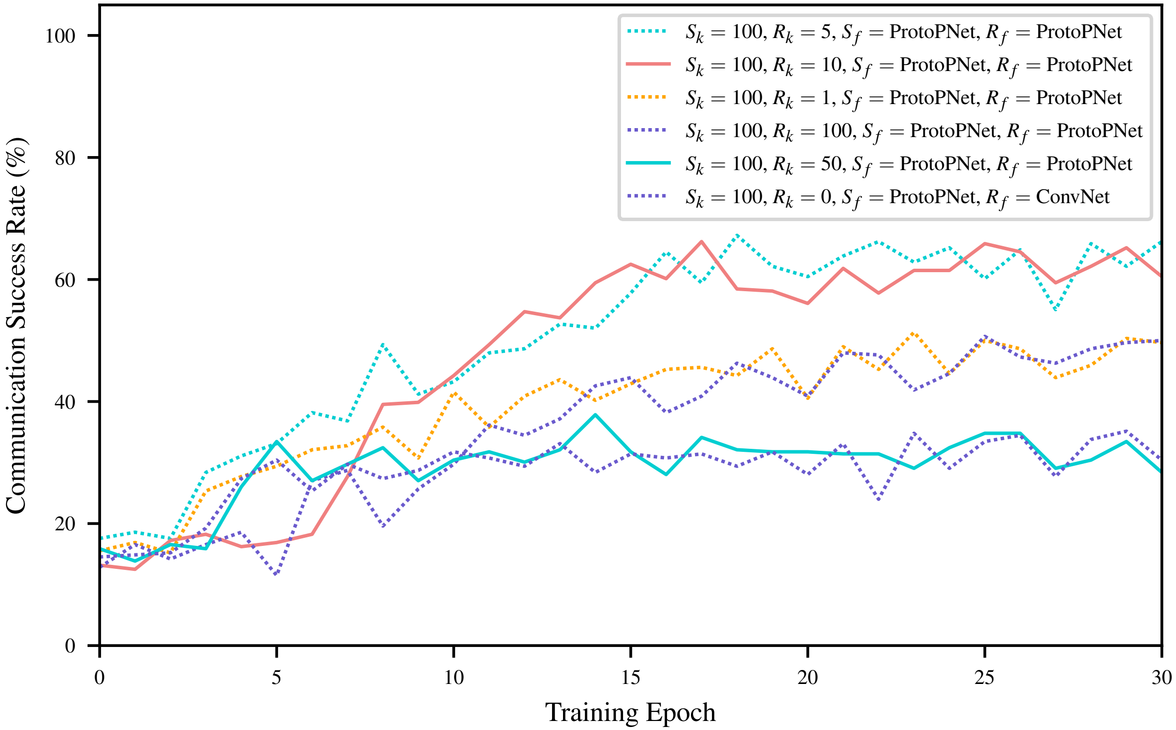






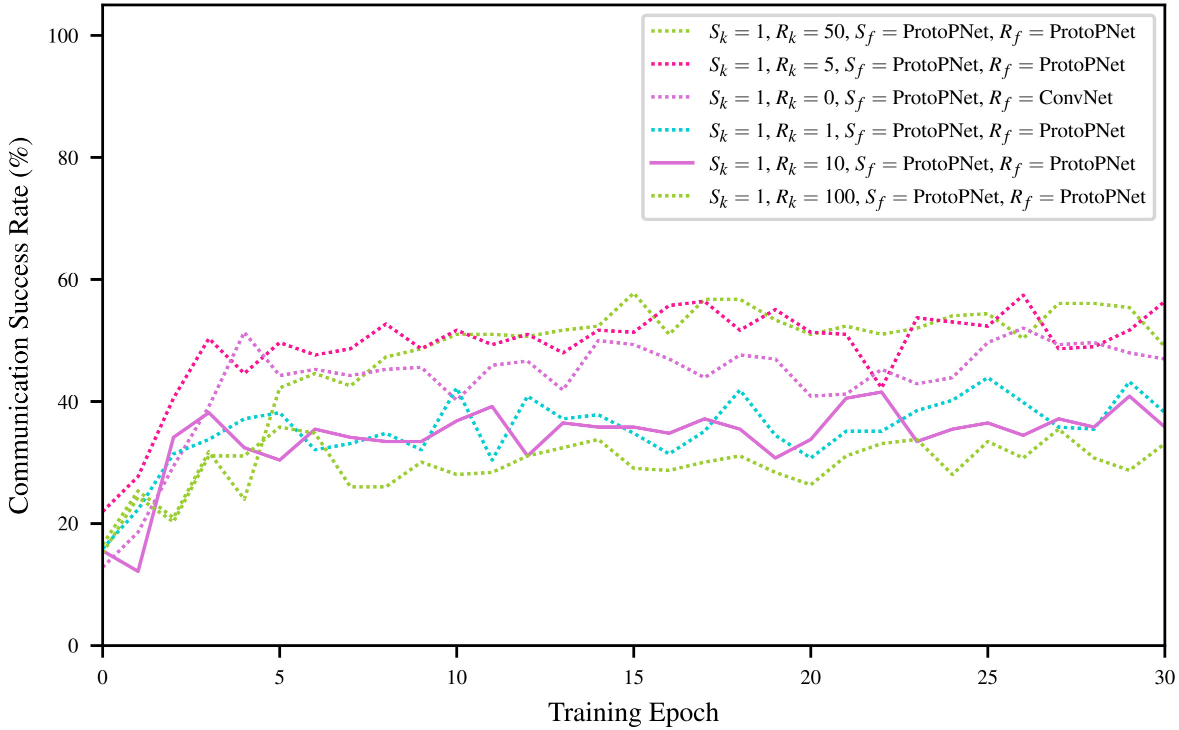
Results

What if agents are the same model (e.g., ProtoPNet), but each realize different level of categorization (number of learned concepts)? CUB10:



Results

What if agents are the same model (e.g., ProtoPNet), but each realize different level of categorization (number of learned concepts)? CUB10:



Spinoff Experiment

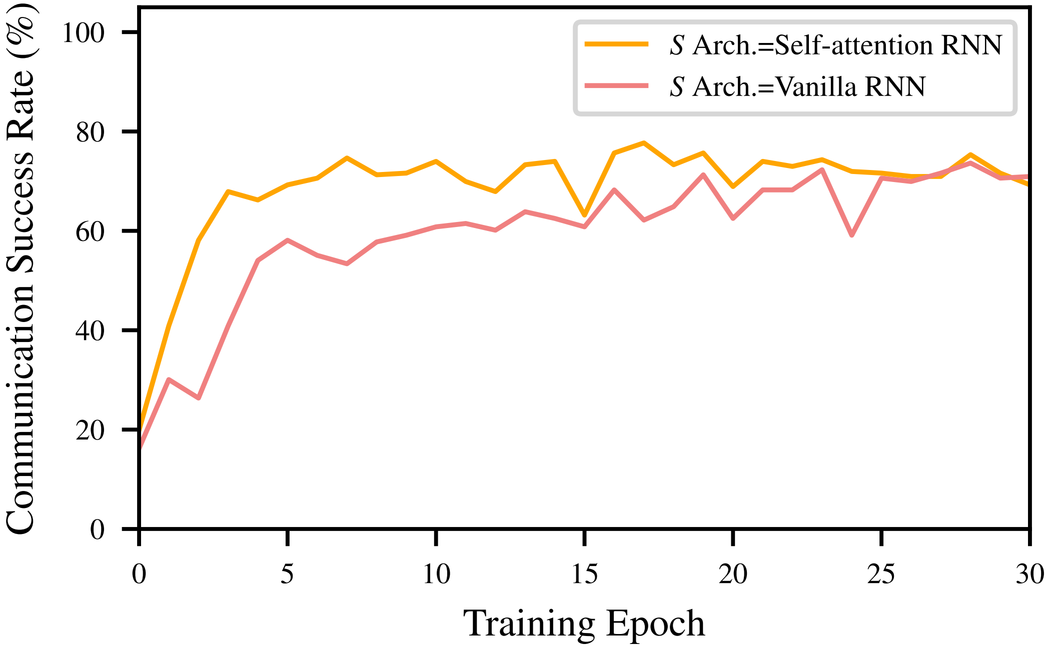
Access to the model's internal disentangled representation is powerful, since we can use it to influence the sender's utterances.

Let's say the vocabulary is exactly the same as the sender's learned concept "vocabulary".

We can give the sender a basic self-attention module that weights utterances using the activated concepts.

Results

Signaling success - CUB10:



Lingering Questions

Combinations of sending and receiving agents produce unintuitive interactions:

- The “smartest” receiving agent is not necessarily the best.
- ConvNets not necessarily compatible with ProtoPNet & CW agents
- Self-attention with learned concepts offers quick ramp-up. Why?

Future work:

- Expanding our ablation studies to 1-length and *k*-length message baselines.
- Self-attention with CW concepts
- Submission to ACL RR
- Leveraging different agent logic (e.g., Dan Guralnik’s UMA models)

Thank you

w.garcia@ufl.edu