Emergence of Grounded Compositional Language in Multi-Agent Populations

Igor Mordatch, Pieter Abbeel

Environment

- multi-agent reinforcement learning
- cooperative agents
- partially observable state (POMDP), but collectively, state is fully observed
- agents act independently
- takes place in 2d continuous euclidean space
- end-to-end differentiable
- fixed episode length

State

Entities:

- M landmarks
  - position (relative!)
  - color
- N agents
  - all of the above
  - velocity
  - gaze (pointer)

Other state:

- memory bank (private)
- goals (private)
  - action
    - go to, look at, or do nothing (one hot)
  - target agent
  - location
- utterance (public)
  - one hot encoding in vocab of size K

Actions

- accelerate along some vector
- set new look-at
- softmax over symbols to emit
- update memory banks

Dynamics

position and velocity updated as usual
γ is a damping factor (friction)
f is used for collision forces, should be smooth

$$\begin{bmatrix} \mathbf{p} \\ \dot{\mathbf{p}} \\ \mathbf{v} \end{bmatrix}_i^t = \begin{bmatrix} \mathbf{p} + \dot{\mathbf{p}}\Delta t \\ \gamma\dot{\mathbf{p}} + (\mathbf{u}_p + \mathbf{f}(\mathbf{x}_1, ..., \mathbf{x}_N))\Delta t \\ \mathbf{u}_v \end{bmatrix}_i^{t-1}$$

- Symbols emitted show up in the environment in the next time step
- Memory banks are updated as well

Reward

$$r_i^t = -(\begin{bmatrix} \|\mathbf{p}_r^t - \bar{\mathbf{r}}\|^2 \\ \|\mathbf{v}_r^t - \bar{\mathbf{r}}\|^2 \\ 0 \end{bmatrix}^{\mathsf{T}} \mathbf{g}^{\text{type}} + \|\mathbf{u}_i^t\|^2 + \|\mathbf{c}_i^t\|^2)$$

$$R = r_c + r_g + \sum_t \sum_i r_i^t$$

Training Algorithm

Make sure everything is fully differentiable, and then run backpropagation.

1. generate batch-size = 1024 random starting environments
2. run dynamics forward and compute total reward
3. run backprop and apply gradients

Gumbel-Softmax Estimator

ϵ is drawn from the Gumbel distribution
For a categorical distribution with parameters $p_1, \cdots p_k$, we sample $\epsilon_1, \cdots, \epsilon_k$.
Then $\log p_i + \epsilon_i$ will be greater than all other $\log p_j + \epsilon_j$ with probability p_i.

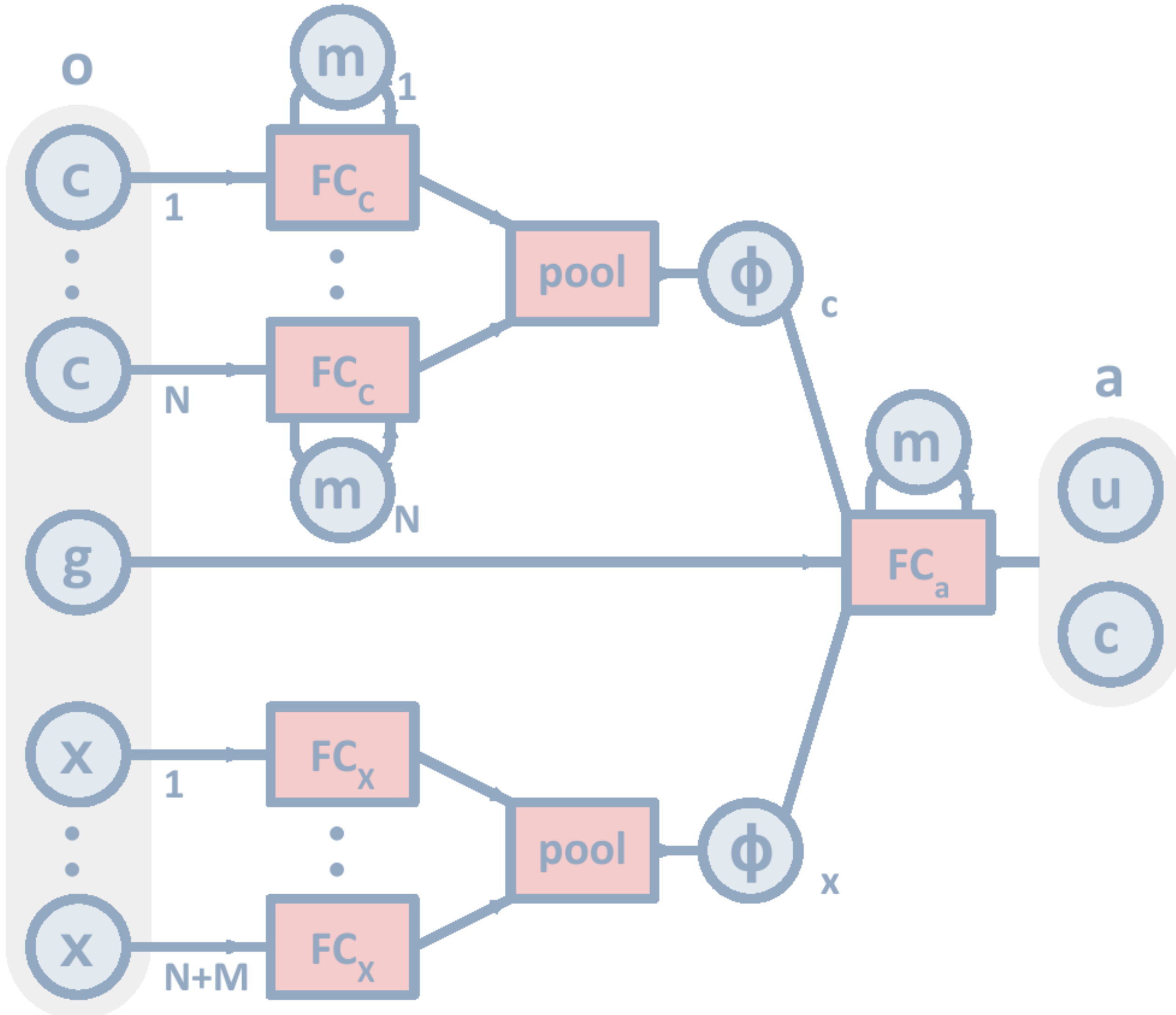
Forward pass:

$$c = \text{sample}(\text{softmax}(\log \vec{p} + \vec{\epsilon}))$$

Backward pass:

$$\frac{dc}{dp} = \frac{d}{dp}\text{softmax}(\log \vec{p} + \vec{\epsilon})$$

Network Architecture



Modules

- allows for training on variable number of landmarks and agents
- each module has independent memory unit
- each module has two layers of 256 units, and a memory size of 32
- input size: ? + 32 (m)
- output size: 256 + 32 (Δm)
- shared weights across all modules of the same type
- unshared memories

Memory update:

$$\mathbf{m}^t = \tanh(\mathbf{m}^{t-1} + \Delta\mathbf{m}^{t-1} + \varepsilon)$$

Softmax pooling is used to deal with multiple agents/landmarks
Gaussian output noise

Prediction Reward

Auxiliary output is prediction of other agent's goals.

$$r_g = -\sum_{\{i,j|i\neq j\}} \|\hat{\mathbf{g}}_{i,j}^T - \mathbf{g}_j^T\|^2$$

MSE even for categorical goal-type and target agent.

Small Vocabulary Reward

Dirichlet process

Suppose n^t words have already been used at time t and the kth word has been used n_k^t times.

$$p_{t+1}(k) = \frac{n_k^t+\alpha K^{-1}}{\alpha+n^t-1}$$

In order to encourage this, maximize the log-likelihood:

$$r_c = \sum_t \sum_i \log p_t(c_i)$$

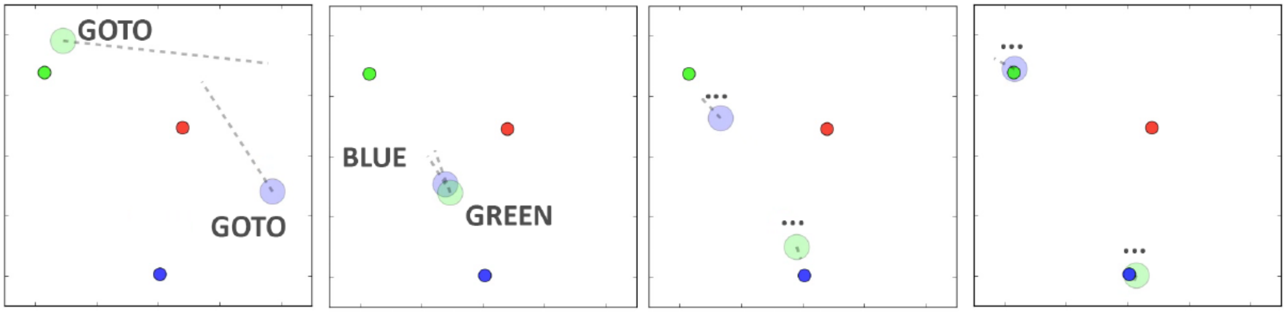
Accumulate counts over:

- time steps
- agents

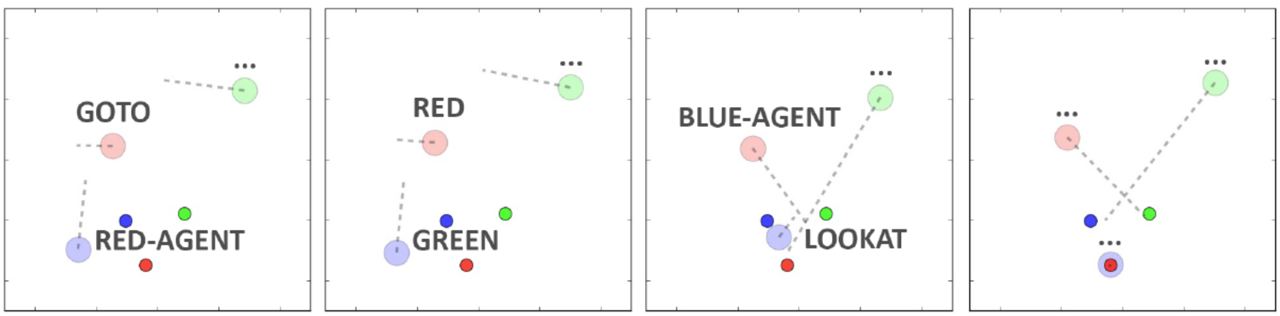
- batches

Results

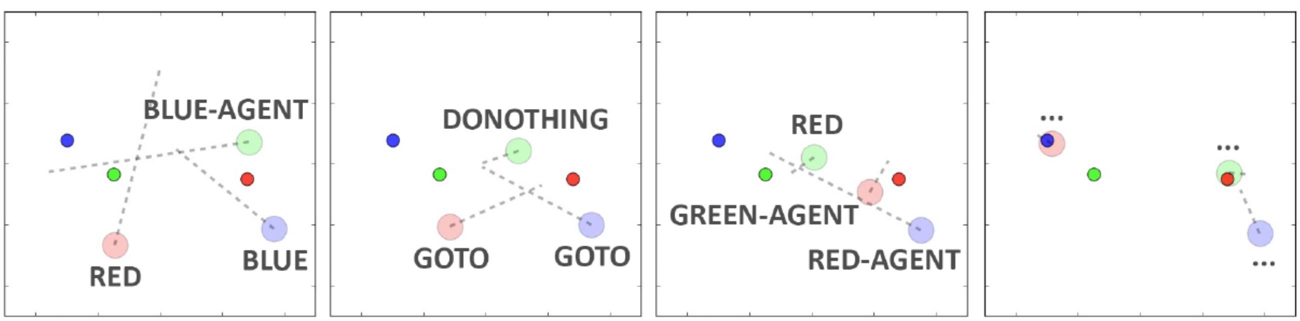
1



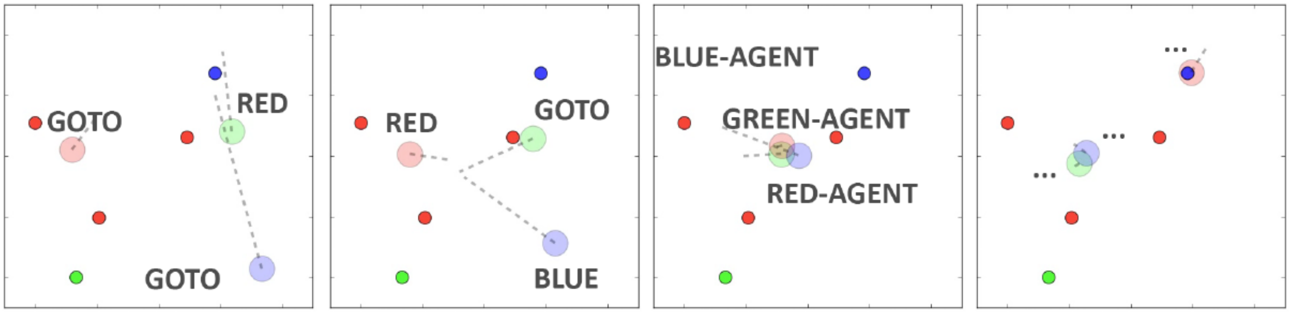
2



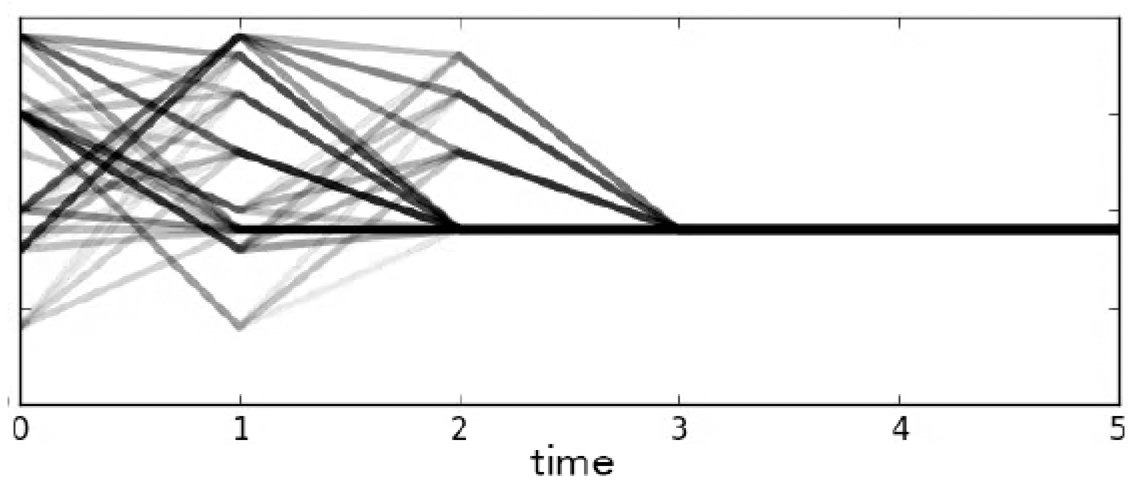
3



4

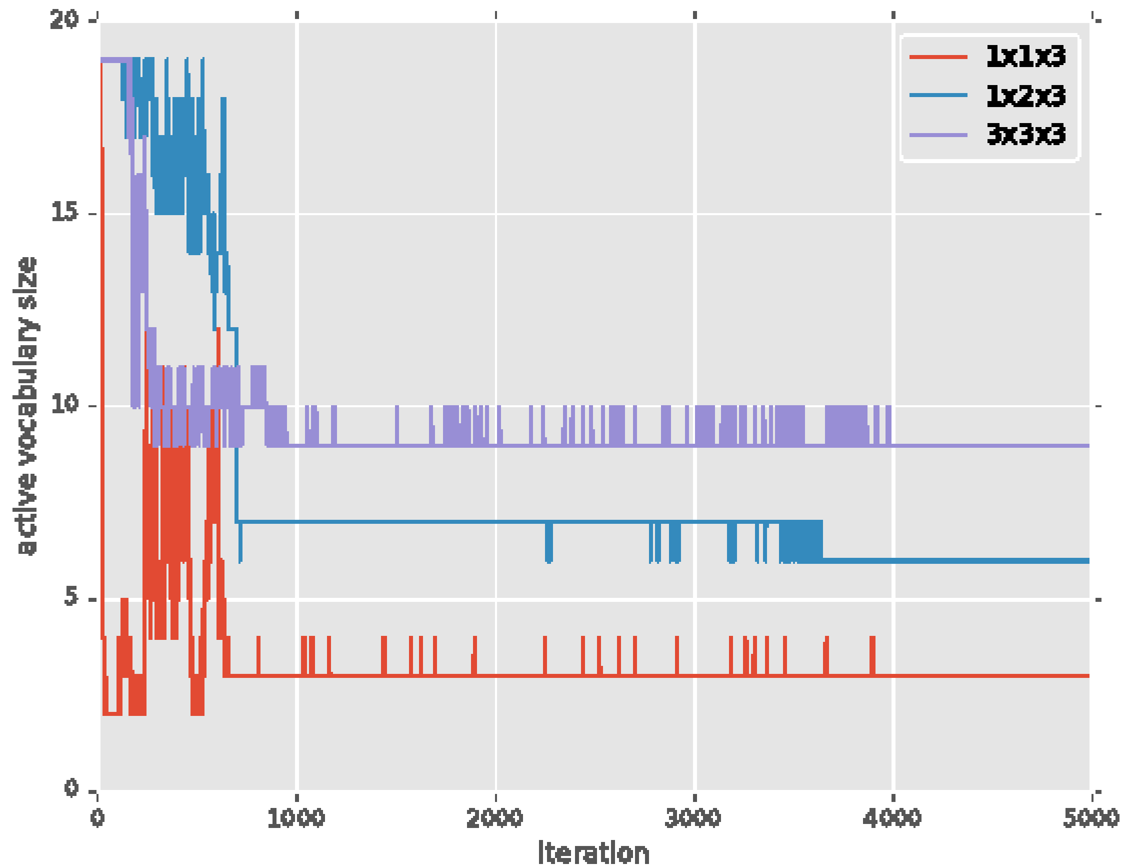


Communication:



More complex requirements require larger language

- 1x1x3: 2 agents, 1 action, 3 landmarks
- 1x2x3: 2 agents, 2 actions, 3 landmarks
- 3x3x3, 4 agents, 3 actions, 3 landmarks



Agents cannot see each other:

Condition	Train Reward	Test Reward
No Communication	-0.919	-0.920
Communication	-0.332	-0.392

Generalization

- Agents will go to the center of two landmarks with the same color if one is referenced.
- If multiple agents have the same color, all will follow instructions from a referring agent.

Nonverbal Communication

- Use of gaze to point to landmarks
- Use of pushing and collision physics to push agents towards landmarks