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
\( \gamma \) is a damping factor (friction)
\( f \) is used for collision forces, should be smooth

\[
\begin{bmatrix}
  \mathbf{p} \\
  \dot{\mathbf{p}} \\
  \mathbf{v}
\end{bmatrix}_{i}^{t} = \begin{bmatrix}
  \mathbf{p} + \dot{\mathbf{p}} \Delta t \\
  \gamma \dot{\mathbf{p}} + (\mathbf{u}_{p} + f(x_1, \ldots, x_N)) \Delta t \\
  \mathbf{u}_v
\end{bmatrix}_{i}^{t-1}
\]

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

Reward

\[
r_{i}^{t} = -\begin{bmatrix}
  \| \mathbf{p}_{r}^{t} - \mathbf{r} \|^2 \\
  \| \mathbf{v}_{r}^{t} - \mathbf{r} \|^2 \\
  0
\end{bmatrix}^{T} g_{\text{type}} + \| \mathbf{u}_{i}^{t} \|^2 + \| \mathbf{c}_{i}^{t} \|^2
\]
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

$\epsilon$ 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 \( (\Delta m)\)
- shared weights across all modules of the same type
- unshared memories

Memory update:
\[ m^t = \tanh(m^{t-1} + \Delta m^{t-1} + \varepsilon) \]

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

**Prediction Reward**

Auxiliary output is prediction of other agent’s goals.

\[ r_g = - \sum_{\{i,j| i \neq j\}} \| \hat{g}_{i,j}^T - g_j^T \|^2 \]

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 \( k \)th word has been used \( n^t_k \) times.

\[ p_{t+1}(k) = \frac{n^t_k + \alpha K^{-1}}{\alpha + n^t - 1} \]

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:

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

More complex requirements require larger language
Agents cannot see each other:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Train Reward</th>
<th>Test Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Communication</td>
<td>-0.919</td>
<td>-0.920</td>
</tr>
<tr>
<td>Communication</td>
<td>-0.332</td>
<td>-0.392</td>
</tr>
</tbody>
</table>

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