Learning Communication Protocols

with Deep Reinforcement Learning

Samuel R. Kopp

Emergence of Communication and Language

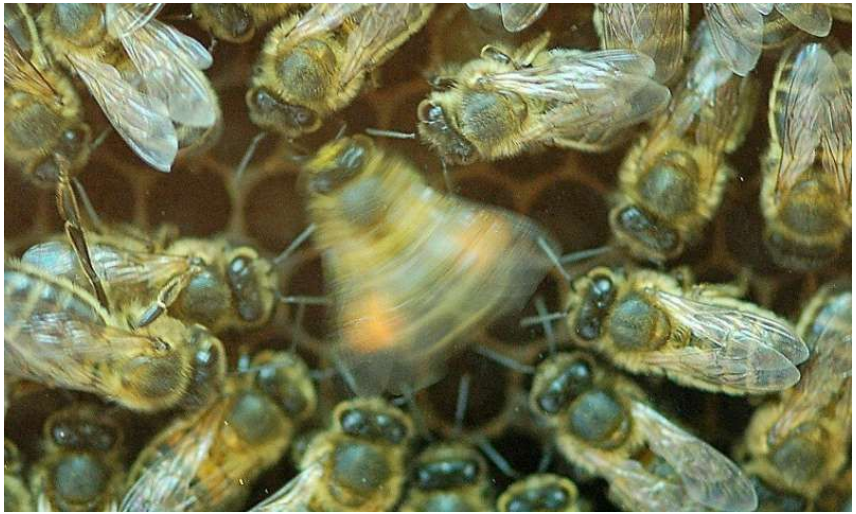
Does language learning need grounding in the world?

What is the origin of language?

How did intelligent agents in the real world learn a common communication protocol?

Can a population of agents learn to communicate to fulfill an objective?

Simple forms of communication in neural networks found in nature



Multi-Agent Environment

XXX			
	XX		
	X		
			XXX

- Grid with patches of “food”
- Objective: gather food on the grid in collaboration
- Agents can broadcast messages to other agents & select cells to explore

Deep Reinforcement Learning

- Task as partially observable Markov decision process (POMDP)
- Agents receive positive/negative reward by choosing actions
- Parameterized policy (neural network) maps states to action
- Actions consist of location in the grid and a message with a given length and vocabulary

Agent architecture

- Single policy:
  - state of agent *i* consists of previous messages and last observation *i*
  - action consists both of message to send and location to explore

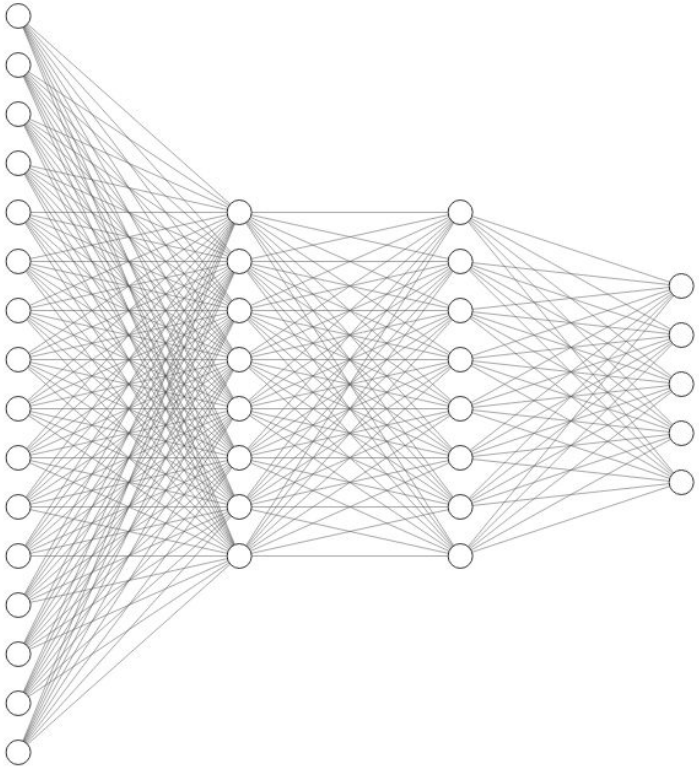
- Dual policy:
  - separate policies for sending messages and selecting location
  - message policy: observation *i* → message *i*
  - exploring policy: messages → location *i*

Proximal Policy Optimization

- State-of-the-art deep RL algorithm, used f.e. for OpenAI Five Dota
- Objective: $J(\theta) = E(R)$
- Standard Actor-Critic gradient: $\nabla_\theta J(\theta) = E[\nabla_\theta log(\pi_\theta(a|s))A^\pi(a|s)]$
- Advantage instead of reward: $A^\pi(a|s) = r - V^\pi$
- PPO gradient:

$$L(s, a, \theta_k, \theta) = min[\frac{\pi_\theta(a|s)}{\pi_k\theta(a|s)}A^\pi(s, a), clip(\frac{\pi_\theta(a|s)}{\pi_k\theta(a|s)}, 1 - \epsilon, 1 + \epsilon)A^\pi(s, a)]$$

messages +
observation



message +
location +
value estimate

N input units → 8 hidden units → 8 hidden units → M output units

Single Policy

sparsity:	popul.:	msg-len:	avg rew:
High	Small	3	~0.31
High	Big	3	~2.52
Low	Small	3	~10.11
Low	Big	3	~55.12
High	Small	0	~0.32
High	Big	0	~2.55
Low	Small	0	~10.09
Low	Big	0	~55.01

Dual Policy

sparsity:	popul.:	msg-len:	avg rew:
High	Small	3	~0.28
High	Big	3	~2.52
Low	Small	3	~10.01
Low	Big	3	~56.01
High	Small	0	~0.31
High	Big	0	~2.51
Low	Small	0	~10.22
Low	Big	0	~55.02

Result is that messages do not improve objective

→ *clearly no meaningful communication*

Conclusion & Further Research

- are RL algorithms fitted to learn policies that happen extremely rarely through random actions?
- would it help to use explicit gradient information of message impact on other agents?
- which exploration strategy does nature use to come up with communication?
- can new individuals learn from already trained agents?