EMERGENCE OF GROUNDED COMPOSITIONAL LANGUAGE IN MULTI-AGENT POPULATIONS

IGOR MORDATCH, PIETER ABBEEL

(SUBMITTED ON 15 MAR 2017)

PRESENTED BY:

SAMREEN M. HASSAN

12.04.2017

IDEA

- For agents to intelligently interact with humans
- If, and how, grounded compositional language can emerge as a means to achieve goals in multi-agent populations.

INTRODUCTION

- Machine Learning has enabled significant advances in NLP including machine translation, question answering etc.
- FOR AGENTS TO INTELLIGENTLY INTERACT WITH HUMANS
- Use of grounded compositional language in multi agent populations.
- Verbal Communication Language - represented as streams of abstract discrete symbols uttered by agents
- coherent structure, defined vocabulary and syntax
- Non verbal communication – pointing & guiding

WHY?

- Development of agents that are capable of communication and flexible language use.
- The supervised approaches that learn to imitate language from examples of human language are present but…

1. they do not capture language's functional aspects
2. Are ambiguous
3. Require human involvement

- VIEW - an agent possesses an understanding of language, when it can use language to accomplish goals in its environment.

IN THIS STUDY…

- physically-situated multi-agent learning environment and learning methods that bring about emergence of a basic compositional language.
- Language represented as streams of abstract discrete symbols uttered by agents over time with a coherent structure (defined vocabulary & syntax)
- agents utter communication symbols alongside performing actions to get a joint reward
- no pre-designed meanings associated with the uttered symbols
- no explicit language usage goals (correct utterances)
- No explicit roles (speaker/listener)

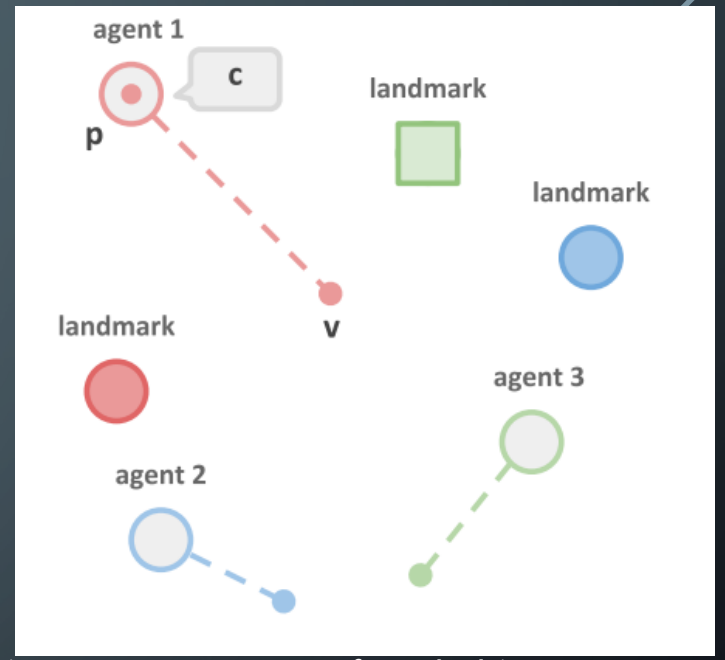
IN THIS STUDY…

- A population of agents is situated as moving particles in a continuous 2-D environment, possessing properties such as color and shape.
- The goals of the population are such as moving to a location, and use of language in order to coordinate on those goals
- A reinforcement learning problem - Agents perform some action 'a' and communication utterances 'c' according to identical policy for all agents
- The language - assigns symbols to separately refer to ENVIRONMENTAL LANDMARKS, ACTION VERBS and AGENTS.
- Non-verbal communication such as pointing and guiding when language communication is unavailable.
- training on a variety of tasks and environment configurations simultaneously.

PROBLEM FORMULATION

- Cooperative partially observable Markov Game – multi agent extension of Markov decision process
- N agents
- set of states S
- Actions A_1 ,…, A_N
- observations O_1, …, O_N
- Initial states are determined by a distribution $\rho : S \rightarrow [0,1]$.
- State transitions are determined by a function $T : S \times A_1 \times ,\ldots, \times A_N \rightarrow S$
- rewards are given by the function $r_i : S \times A_i \rightarrow R$
- observations are given by function $\mathbf{o}_i : S \rightarrow O_i$
- To choose actions, each agent i uses a stochastic policy $\boldsymbol{\pi}_i : O_i \times A_i \rightarrow [0,1]$
- The expected shared return for all agents $R(\boldsymbol{\pi}) = \exp\left[\sum_{t=0}^{T} \sum_{i=0}^{N} r\left(s_i^t , a_i^t\right) \right]$

GROUNDED COMMUNICATION ENVIRONMENT



- physically-simulated 2-D environment in continuous space and discrete time
- 'N' agents, 'M' landmarks
- physical location in space 'p'
- Physical characteristics 'shape and color type'
- location to gaze 'v'
- physical state of an entity 'x'
- agents utter verbal communication symbols 'c' at every timestep from vocabulary 'C' of size 'K' (no correct meaning of symbols)
- Agent's internal goals vector 'g'
- Agent's internal recurrent memory bank 'm'
- full state of the environment is given by $s = [\, x_{1,\ldots,(N+M)}\ c_{1,\ldots,N}\ m_{1,\ldots,N}\ g_{1,\ldots,N}\,] \in S$
- observation for agent i is $o_i(s) = [\, {}_i x_{1,\ldots,(N+M)}\ c_{1,\ldots,N}\ m_i\ g_i\,]$ where ${}_i x_1$ is the observation of entity j's physical state in agent i's reference frame.

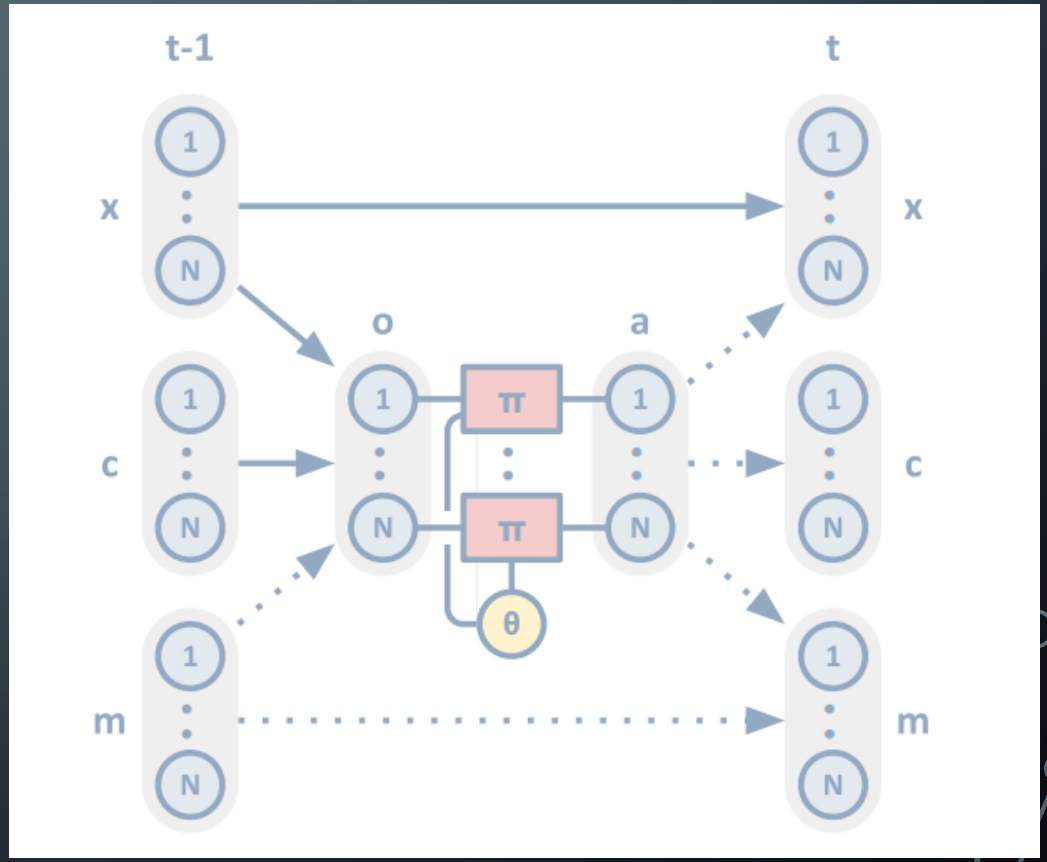
POLICY LEARNING WITH BACKPROPAGATION

- Each agent samples actions from a stochastic policy **π**, which is identical for all agents and defined by parameters ϑ
- Build a end-to-end differentiable model of all agents and environment state over time and calculate $dR/d\vartheta$ *(total return gradient)* with backpropagation.

- Figure:

  transition dynamics of *N* agents from time *t* - 1 to *t*.

  Dashed lines -> one-to-one dependencies between agents
  Solid lines -> all-to-all dependencies.



DISCRETE COMMUNICATION AND GUMBEL-SOFTMAX ESTIMATOR

- To use categorical communication emissions **c** in our setting, it must be possible to differentiate through them.
- The approach, used here, proposes a Gumbel-Softmax distribution, which is a continuous relaxation of a discrete categorical distribution
- Given *K*-categorical distribution parameters *p*, a differentiable *K*-dimensional one-hot encoding sample *G* from the Gumbel-Softmax distribution can be calculated as:

$$G(\log p)_k = \exp\left((\log p_k + epsilon)/\tau\right) \;/\; \sum_{i=0}^{k} \exp\left((\log p_i + epsilon)/\tau\right)$$

Where -

- ***epsilon*** are samples from Gumbel(0, 1) distribution,

$$epsilon = -\log(-\log(u)),\quad u \sim U[0; 1]$$

- τ *softmax temperature parameter (set to 1)*

To emit a communication symbol, our policy is trained to directly output log*p* , which is transformed to a

symbol emission sample **c** ~ *G*(log*p*).

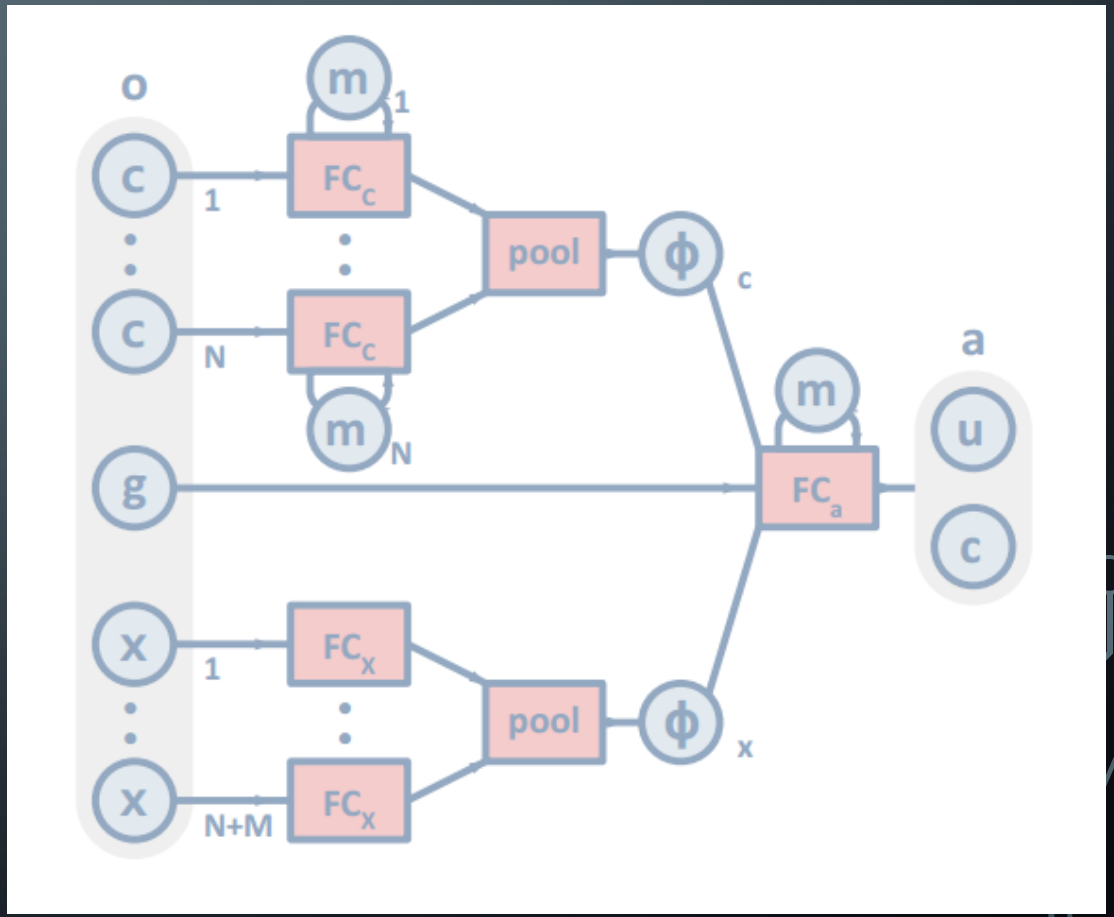
The resulting gradient can be estimated as $dc/d\vartheta = dG/dp \cdot dp/d\vartheta$.

POLICY ARCHITECTURE

- policy class – stochastic neural network
- policy outputs –

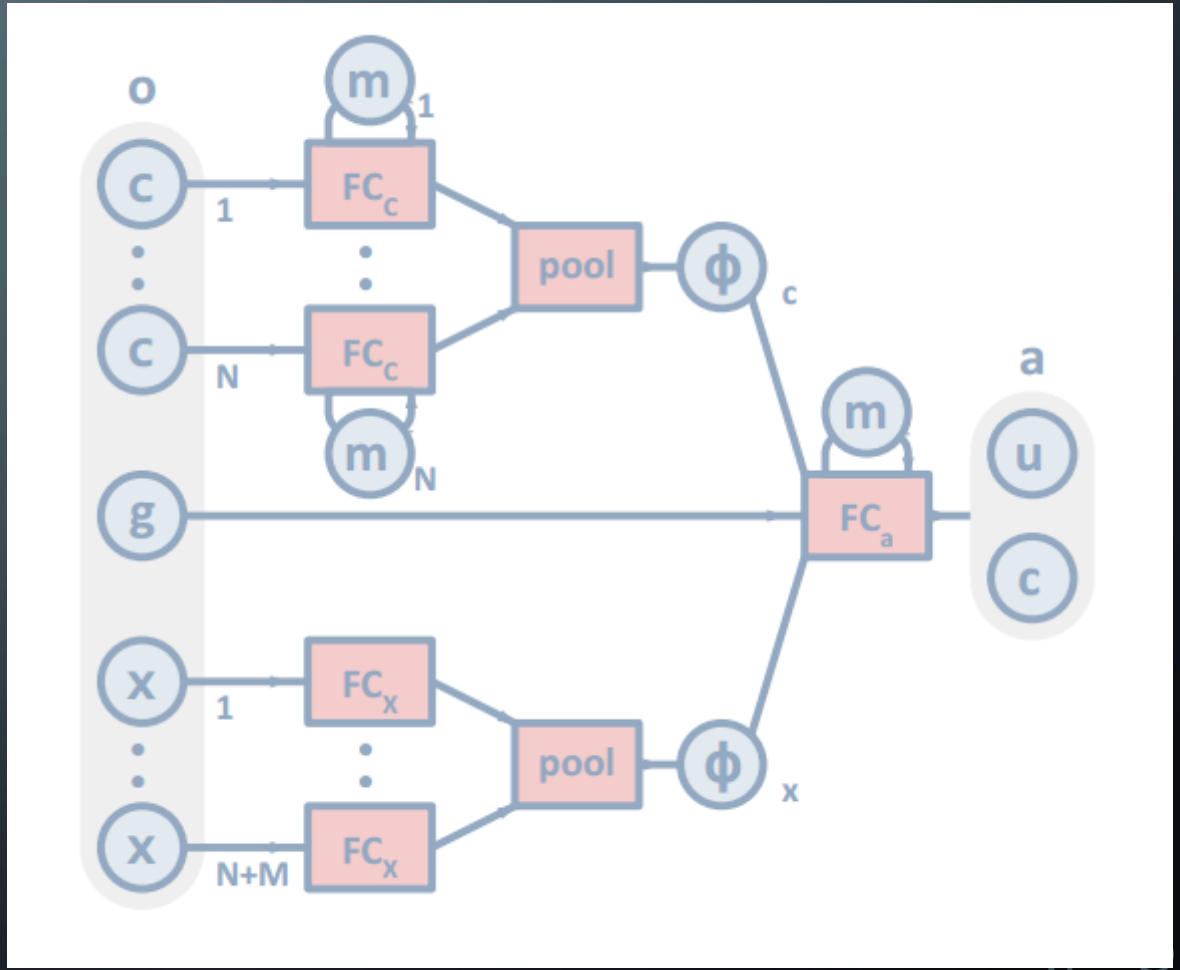
1. samples of an agent's physical actions u
2. communication symbol utterance **c**
3. internal memory updates $\Delta\mathbf{m}$

- policy must consolidate multiple incoming communication symbol streams emitted by other agents (c) and incoming observations of physical entities. (x)



POLICY ARCHITECTURE

- the policy instantiates a collection of identical processing modules
- Each processing module is a fully-connected multi-layer perceptron
- The outputs of individual processing modules are pooled with a softmax operation into feature vectors ***φ***c *(communication)* and ***φ***x *(physical observation)*
- The pooled features and agent's private goal vector are passed to the final processing module that outputs distribution parameters [u,c] from which action samples are generated as **u** = *u* + ***epsilon*** and **c** ~ *G*(c), where ***epsilon*** is a zero-mean Gaussian noise
- our agents continually emit a stream of symbols over time, so recurrent memory is advantageous
- all fully-connected modules with 256 hidden units and 2 layers each are used in all the experiments
- Size is feature vectors ***φ*** is 256 and
- size of each memory module is 32.



AUXILLARY PREDICTION REWARD

- In agent *i*'s policy, each communication processing module *j* additionally outputs a prediction $\mathbf{g}\hat{}_{i,j}$ of agent *j*'s goals.
- At the end of the episode, a reward is added for predicting other agent's goals, which in turn encourages communication utterances.

- reward has the form
  $r_g = - \sum_{\{i,j \mid i!=j\}} || \mathbf{g}\hat{}_{i,j}^{T} - \mathbf{g}_j^{T} ||^2$

COMPOSITIONALITY AND VOCABULARY SIZE

- How compositional syntax form?
- Hypothesis 1:

a. process of language transmission and acquisition from one generation of agents to the next iteratively.

b. symbol utterances from the previous generation + infer meaning of unseen symbols.

c. implemented with predesigned rules lead to formation of a compositional vocabulary.

- Alternatively:

a. emergence of compositionality requires the number of concepts describable by a language to be above a factor of vocabulary size

b. maximum vocabulary size $K = 20$ in all experiments

c. For small maximum size, the policy optimization became stuck in a local minima

d. use a large vocabulary size limit but use a soft penalty function to prevent the formation of unnecessarily large vocabularies.

EXPERIMENTS

HOW VARIATION IN GOALS, ENVIRONMENT CONFIGURATION, AGENTS PHYSICAL CAPABILITIES AFFECT COMMUNICATION STRATEGIES

EXPERIMENT

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

Table 1. Training and test physical reward for setting with and without communication.

- Actions – 3 (*go to* location, *look at* location, *do nothing*)
- Goal description vector – Goal(agent 'i') –> (action, location, agent 'r' who performs)

But an agent can't share the goal vector directly with another agent

- *Different reference frames*
- *Can communicate only in discrete symbols*
- *No shared global positioning reference*

- Place goal locations on landmarks which are observed by all agents
- Agent 'i' communicates landmark reference to agent 'r'
- No communication strategies – all agents go towards centroid of all landmarks
- Reward performance very similar during test

EXPERIMENT - Syntactic Structure

- Environment 1:

I. 2 agents, multiple landmarks and actions

II. symbols forming for each of the landmark colors and each of the action types.

III. A typical conversation and physical agent configuration

Green Agent: GOTO, GREEN, ...
Blue Agent: GOTO, BLUE, ...

The action type verb *GOTO* is uttered first because actions take time to accomplish

When the agent receives *GOTO* symbol, it starts moving toward the centroid of all the landmarks

then moves towards the specific landmark when it receives its color identity

EXPERIMENT - Syntactic Structure

- Environment 2:

I. >3 agents, multiple landmarks and actions

II. symbols forming for each of the landmark colors and each of the action types.

III. agents need to form symbols for referring to each other - agent colors

IV. A typical conversation and physical agent configuration

Red Agent: GOTO, RED, BLUE-AGENT, ...
Green Agent: ..., ..., ..., ...
Blue Agent: RED-AGENT, GREEN, LOOKAT, ...

The Agents may not omit any utterances when they are the subject of their private goal

The agents largely settle on using a consistent set of symbols for each meaning, due to vocabulary size penalties

- SIMPLIFIED Environment:

one landmark or one type of action to take, no symbols are formed

Figure:

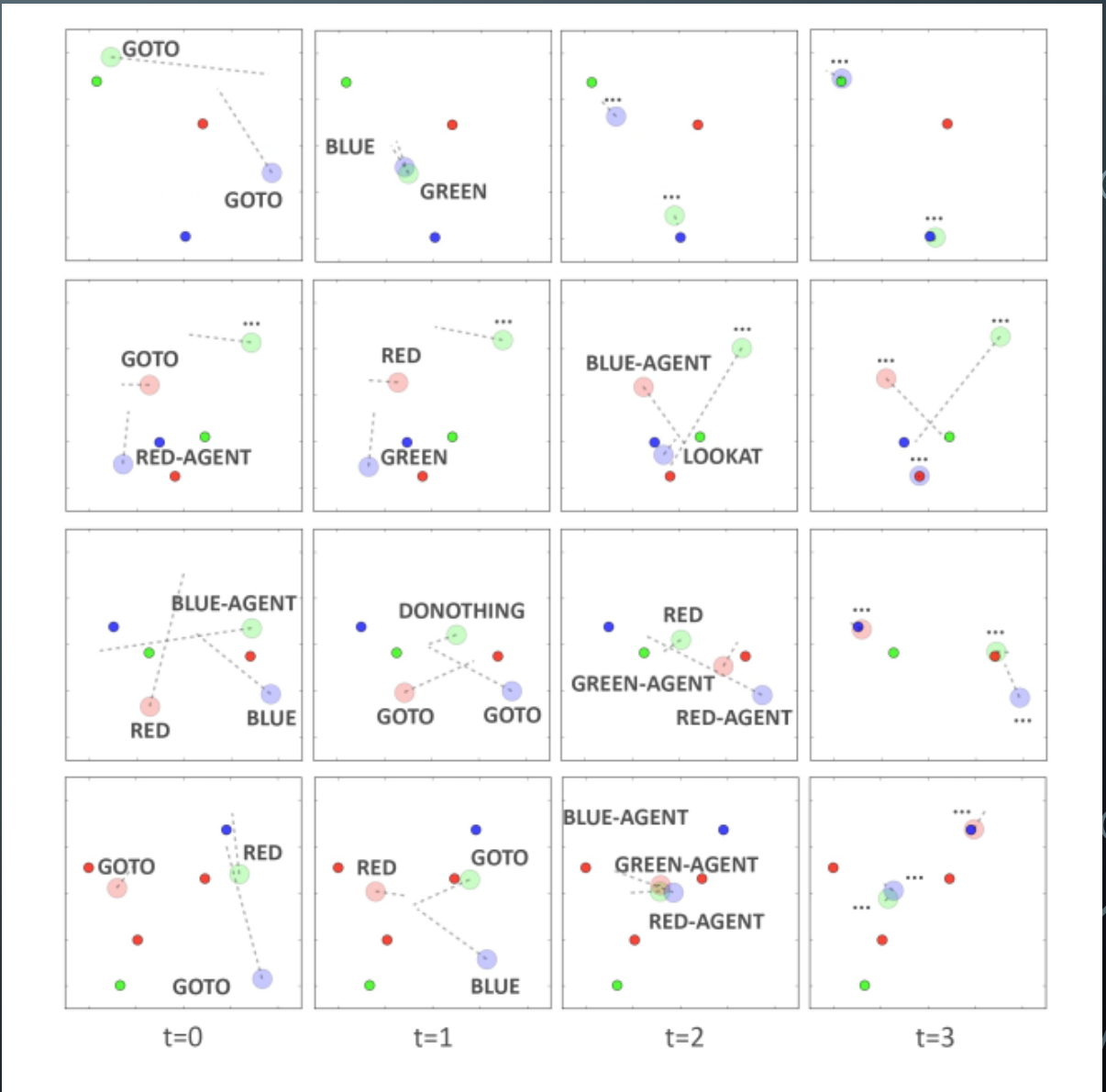
A collection of typical sequences of events in our environments shown over time.

Each row is an independent trial.

Large circles represent agents and small circles represent landmarks.

Communication symbols are shown next to the agent making the utterance.

The labels for abstract communication symbols are chosen purely for Visualization.



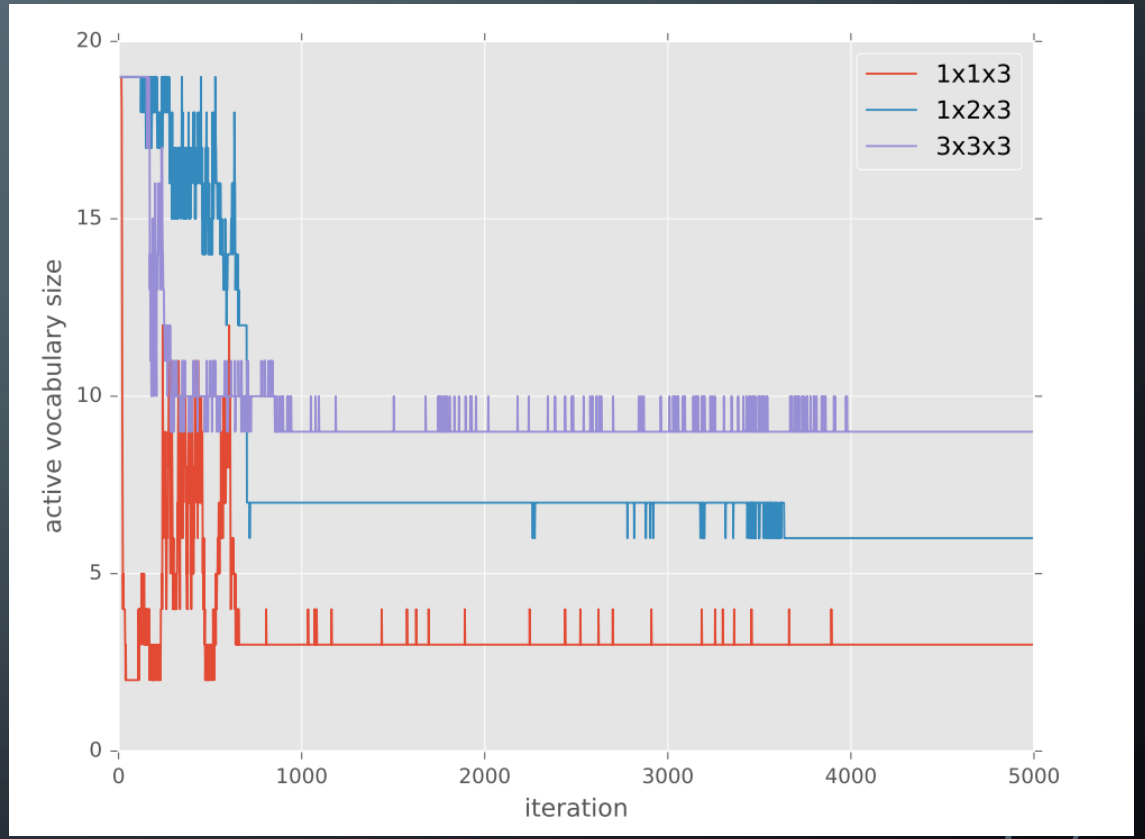
EXPERIMENT – Symbol Vocabulary Usage

- Used word activation counts to settle on the appropriate compositional word counts.
- During training large vocabulary sizes are taken
- Figure: Word activation counts for different environments over training iterations

i. *1x1x3* case refers to environment with 2 agents and one action, which requires only communicating 1 out of 3 landmark identities.

ii. *1x2x3* contains 2 types of actions

iii. *3x3x3* case contains 3 agents that require explicit referencing



EXPERIMENT – Generalization To Unseen Configurations

- Trained agents can be placed into arbitrarily-sized groups and still function reasonably.
- Multiple agents in the environment with the same color identity, all agents of the same color will perform the same task if they are being referred to.
- When agents of a particular color are asked to perform two conflicting tasks, they will perform the average of the conflicting goals assigned to them.
- When there are multiple landmarks of the same color, the agents receiving the landmark color utterance go towards the centroid of all landmark of the same color.
- When there are distractor landmarks of novel colors, the agents never go towards them.

- So they behave sensibly in unseen configurations too.

NON VERBAL COMMUNICATION

- Alternative strategy apart from language use
- Agents observe other agent's position and gaze location.

➢ Pointing – *gazing in its direction*

➢ Guiding – *guide the goal recipient*

➢ Pushing – *pushing to the target location*

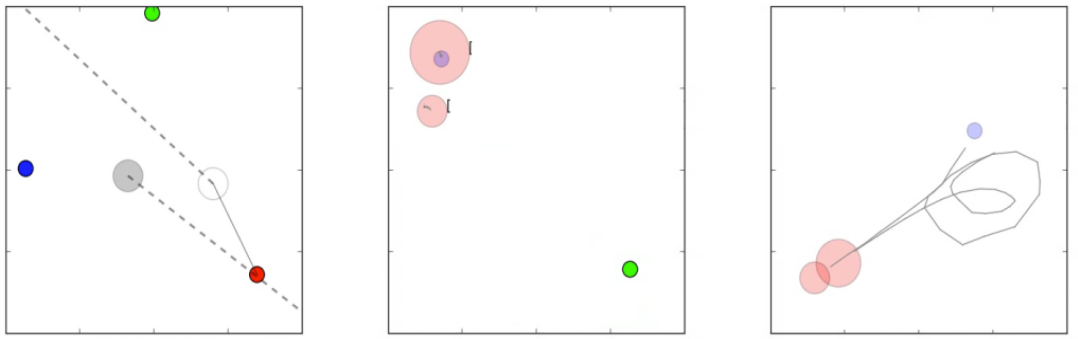


Figure 7. Examples of non-verbal communication strategies, such as pointing, guiding, and pushing.

EXPERIMENT VIDEOS

- https://sites.google.com/site/multiagentlanguage/

CONCLUSION & FUTURE STEPS…

- emergence of an abstract compositional language formed without any exposure to human language use in a multi-agent environment and using different learning methods
- how variation in environment configuration and physical capabilities of agents affect the communication strategies that arise.
- *experiment with larger number of actions that necessitate more complex syntax and larger vocabularies.*
- *integrate exposure to human language to form communication strategies that are compatible with human use.*

- Thank you for your attention.