Towards Emerging Nonverbal Communication Protocols for Multi-Robot Populations

Kalesha Bullard*, Jakob Foerster*, Douwe Kiela*, Joelle Pineau*†and Franziska Meier*
*Facebook Artificial Intelligence Research
†MILA, McGill University, Montreal, Canada
Email: ksbullard@fb.com

I. Introduction

The ability to communicate effectively with other agents is part of a necessary skill repertoire of intelligent agents and occurs only in multi-agent contexts. Within the field on multi-agent learning (MAL), emergent communication literature primarily focuses on *language acquisition*[1] to communicate agent intent [9, 3, 4, 1, 7]. While language is a powerful and expressive modality for communication, not all agents will have access to this modality. For example, zoomorphic agents, robotic manipulators, and prelingual infants are generally not expected to use language to communicate. Even when language is a possible modality, it is sometimes not available to agents (*e.g.* when agents are too far in proximity for verbal communication or when encountering agents that don't have the ability to hear). Just as important as the inability to use language as a modality for communication is the ability for multi-modal communication. Non-verbal communication is an important co-expressive form of communication, alongside language [8]. Embodied agents inherently have access to another expressive modality for communication, through manual (articulation motion) communication, which we explore in this work. It is, to the best of our knowledge, a first investigation into the emergence of high-dimensional motion generation for nonverbal communication in cooperative agent environments.

We focus on the task of learning to generate goal-conditioned motion by having agents play a gesture-based referential game. Referential games have been extensively explored in the emergent communication literature [3, 2, 4, 7]. The basic idea is that through repeated agent-agent interactions the community of agents converges on a communication protocol, such that each agent in the community can both generate and interpret the language developed. In our problem setting, the messages consist of motion trajectories, with different trajectories corresponding to different intents.

In this work, no explicit supervision is provided, to indicate *how* communicative intents *should be* mapped to the trajectory space. This makes learning a non-verbal communication protocol with desirable properties (*e.g.* generality, interpretability) challenging since the space of possible trajectories is $\mathbb{R}^{J\times m\times T}$, where J is the number of joints of the agent, m the number of degrees of freedom per joint, and T a fixed time horizon. Without data to guide the policy search, converging on a meaningful protocol is a highly underconstrained optimization problem. Yet, we know from observing human-human communication, it is possible to discover and exploit a significantly smaller part of the trajectory space for meaningful exchange of information. Inspired by this, the central hypothesis of this work is: with such a large space of motion trajectories *valid* for communicating intent, to emerge *general* communication skills, it is essential to both (1) constrain the space of likely (or desirable) trajectories and (2) impose structure for how to assign meaning (labels) to generated trajectories.

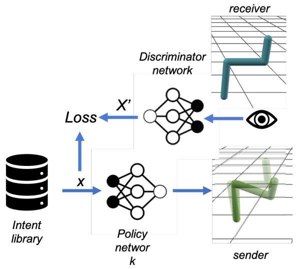


Fig. 1: Overview of Learning System

The overarching goal of this work then is to learn nonverbal communication protocols that can generalize to novel partners. We aim to enable zero-shot communication, whereby at inference time, agents are able to effectively communicate with novel, independently trained partners. Our approach involves two steps: (1) bias (constrain) the policy search by imposing an energy penalty, such that agents incur cost as a function of the amount of torque exerted in generating trajectories and (2) induce a non-uniform distribution over the set of communicative intents, such that all agents will assign labels to generated trajectories consistent with likelihood of intent being communicated. This approach is intended to impose sufficient structure in the protocol learning process, such that even if agents have never had the opportunity to directly interact nor indirectly influence one another during training (*e.g.* agent i never interacts with a communication partner of agent j), agents can still independently learn protocols with similar structure, thus improving communication success at inference time between completely novel partners. The first step incentivizes trajectory generation in a similar way (with minimal energy); the second step encourages assignment of meaning to trajectories in a similar way (pairing most frequently occurring intents with lowest energy trajectories).

[1] Throughout this paper we use the term *language* to refer to cheap-talk protocols using discrete dictionaries.

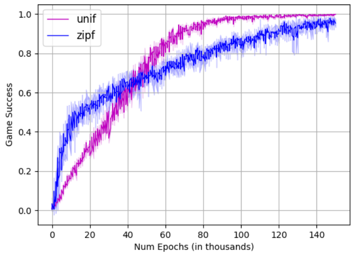


Fig. 2: Game success during protocol training. *No* energy regularization. Comparing uniform and zipf intent distributions. Self-play, N = 500. Curves show mean and standard error, over 5 random seeds.

II. Approach

The protocol learning problem is formalized as a cooperative Markov game with N agents [6], defined by tuple (S, A, T, R, γ). The objective is to infer a policy that maximizes expected shared return R for all agents. The policy state space S consists of the joint configuration of the agent and the intent to be communicated, where the joint configuration of the agent is defined by three-dimensional position (p_x, p_y, p_z) and three-dimensional rotation (r_x, r_y, r_z) of each agent joint j in the agent's set of joints J. The action space A consists of the angular velocities for each joint $j \in J$: $(\Delta r_x, \Delta r_y, \Delta r_z)$. The transitions $T : (s_t, a_t) \rightarrow s_{t+1}$ are deterministically computed through a differentiable forward kinematics (FK) module, ensuring kinematically valid trajectories are generated.

Each community is composed of a set of articulated agents. The agent module contains a predefined kinematic structure and two learnable models: a policy (for generating motion) and a discriminator (for observing and interpreting motion). All networks for all agents are co-trained in parallel through the referential game. Policy networks are given a communicative intent (*e.g.* "yes", "no"), in the form of an embedding vector, and produce a trajectory rollout as output. The policy network is a multi-layer perceptron. At each time step t, it inputs the joint configuration of the agent, $(p_x, p_y, p_z, r_x, r_y, r_z)\ \forall j \in J$, concatenated with an intent embedding. It outputs an action a_t, $(\Delta r_x, \Delta r_y, \Delta r_z)\ \forall j \in J$. After T steps, the episode terminates, and the joint state sequence is concatenated to compose a trajectory. This sequence is given as input to the discriminator, also a multi-layer perceptron. It outputs a softmax over all possible intents in the library and is updated using cross-entropy loss between the predicted and ground truth intents. Figure 1 illustrates an overview of the system just described.

For experimental conditions, we employ an energy loss to regularize protocol learning. It is the total torque exerted by the agent when generating a trajectory (Equation 1).

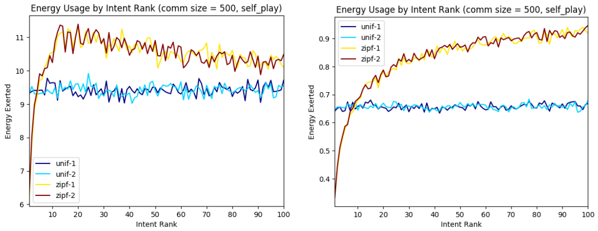
$$L_{engy} = \| I * (\hat{\omega}_{2:T} - \hat{\omega}_{1:T-1}) \|_2^2 \tag{1}$$

Here, I is the moment of inertia and w the angular velocity, for all joints. The total loss is a linear combination of prediction (cross-entropy) and energy (torque) losses.

III. Preliminary Experiments

A. Task Description

We design an *Intent Recognition* referential task for our experiments and assume a library of 100 communicative intents given a priori. Receiver agents are able to directly observe positions and orientations of sender joints. However, Gaussian noise is added to each action to simulate imperfect actuation. A robot arm topology is used (illustrated in Figure 1), motivated by the goal of multi-robot interaction. Agent joints are not constrained, allowing maximal range of motion.



(a) N=500 (Baseline)

(b) N=500 (Energy Regularization)

Fig. 3: Energy Exertion as a function of Intent Rank. All protocols trained through Self-Play, where population size N = 500.

B. Energy Regularization for Paired-Play

Preliminary experiments were aimed at examining successful learning of nonverbal, manual protocols, and understanding whether utilization of (a) energy regularization and (b) a nonuniform distribution over communicative intents, are able to induce a clear learning bias that we can exploit for zero-shot communication. In particular, we induce a Zipf distribution over intents, as it has been used in linguistics for modeling distributions over words and natural language utterances [5, 10]. Figure 2 shows average sample efficiency of learning baseline protocols, which use *no* energy regularization. Figure 3 plots energy utilization for four protocol learning experiments based upon intent rank. All protocols trained using self-play.

The learning curve results imply we can successfully train nonverbal communication protocols. The energy utilization plots show consistent trends, comparing baseline protocols to energy regularized protocols. First, employing energy regularization decreases average torque used in generating communication by 1-2 orders of magnitude; this has important implications for robots which must generate motion on physical hardware. Second, using a uniform intent distribution, whether or not with energy regularization, does not facilitate differentiation *between* intents, for an agent not part of the protocol learning process. This is shown by the *unif* curves which exhibit a constant trend across intents, meaning no differentiation along the energy spectrum. Lastly, by inducing a Zipf distribution over intents, we introduce a well-founded inductive bias into the protocol learning process that can be exploited for generalizing communication. This bias appears stronger when the nonuniform intent distribution is paired with an energy regularizer derived from first principles of motion (physical forces exerted by agent). That is, the difference between zipf and uniform curves in regularized protocols (right) is more significant than in baseline protocols (left).

Overall, these preliminary experimental findings suggest we can successfully train a protocol through the manual modality and then exploit patterns in energy exertion, when employing a nonuniform intent distribution with energy regularization, to enable better zero-shot communication of the learned protocol.

References

[1] Laura Graesser, Kyunghyun Cho, and Douwe Kiela. Emergent linguistic phenomena in multi-agent communication games. *arXiv preprint arXiv:1901.08706*, 2019.

[2] Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in neural information processing systems*, pages 2149–2159, 2017.

[3] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*, 2016.

[4] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv preprint arXiv:1804.03984*, 2018.

[5] Wentian Li. Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*, 38(6):1842–1845, 1992.

[6] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[7] Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 693–701. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[8] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[9] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.