

---

# Learning Multiagent Communication with Backpropagation

---

**Sainbayar Sukhbaatar**

Dept. of Computer Science  
Courant Institute, New York University  
sainbar@cs.nyu.edu

**Arthur Szlam**

Facebook AI Research  
New York  
aszlam@fb.com

**Rob Fergus**

Facebook AI Research  
New York  
robfergus@fb.com

## Abstract

Many tasks in AI require the collaboration of multiple agents. Typically, the communication protocol between agents is manually specified and not altered during training. In this paper we explore a simple neural model, called CommNN, that uses continuous communication for fully cooperative tasks. The model consists of multiple agents and the communication between them is learned alongside their policy. We apply this model to a diverse set of tasks, demonstrating the ability of the agents to learn to communicate amongst themselves, yielding improved performance over non-communicative agents and baselines. In some cases, it is possible to interpret the language devised by the agents, revealing simple but effective strategies for solving the task at hand.

## 1 Introduction

Communication is a fundamental aspect of intelligence, enabling agents to behave as a group, rather than a collection of individuals. It is vital for performing complex tasks in real-world environments where each actor has limited capabilities and/or visibility of the world. Practical examples include elevator control [4] and sensor networks [6]; communication is also important for success in robot soccer [26]. In any partially observed environment, the communication between agents is vital to coordinate the behavior of each individual. While the model controlling each agent is typically learned via reinforcement learning [2, 29], the specification and format of the communication is usually pre-determined. For example, in robot soccer, the bots are designed to communicate at each time step their position and proximity to the ball.

In this work, we propose a model where cooperating agents learn to communicate amongst themselves before taking actions solely from reward in the environment. Each agent is controlled by a deep feed-forward network, which additionally has access to a communication channel carrying a continuous vector at each time step. Through this channel, they receive the summed transmissions of other agents. However, what each agent transmits on the channel is not specified a-priori, being learned instead. Because the communication is continuous, the model can be trained via back-propagation, and thus can be combined with standard single agent RL algorithms. The model is simple, versatile, and scalable, allowing dynamic variation at run time in both the number and type of each agent while avoiding exponential blow-ups in state space and number of agents. This allows it to be applied to a wide range of problems involving partial visibility of the environment, where the agents learn a task-specific communication that aids performance.

## 2 Problem Formulation

We consider the setting where we have  $J$  agents, all cooperating to maximize reward  $R$  in some environment. We make the simplifying assumption of full cooperation that each agent receives  $R$ , independent of their contribution. In this setting, there is no difference between each agent having its own controller, or viewing them as pieces of a larger model controlling all agents. Taking the latter perspective, our controller is a large feed-forward neural network that maps inputs for all agents to their actions, each agent occupying a subset of units. A specific connectivity structure between layers

(a) instantiates the broadcast communication channel between agents and (b) propagates the agent state.

Because the agents will receive reward, but not necessarily supervision for each action, reinforcement learning is used to maximize expected future reward. We explore two forms of communication within the controller: (i) discrete and (ii) continuous. In the former case, communication is an action, and can naturally be handled by reinforcement learning. In the continuous case, the signals passed between agents are no different than hidden states in a neural network; thus credit assignment for the communication can be performed using standard backpropagation (within the outer RL loop). We use policy gradient [36] with a state specific baseline for delivering a gradient to the model. Denote the states in an episode by  $s(1), \dots, s(T)$ , and the actions taken at each of those states as  $a(1), \dots, a(T)$ , where  $T$  is the length of the episode. The baseline is a scalar function of the states  $b(s, \theta)$ , computed via an extra head on the model producing the action probabilities. Beside maximizing the expected reward with policy gradient, the models are also trained to minimize the distance between the baseline value and actual reward. Thus after finishing an episode, we update the model parameters  $\theta$  by

$$\Delta\theta = \sum_{t=1}^T \left[ \frac{\partial \log p(a(t)|s(t), \theta)}{\partial \theta} \left( \sum_{i=t}^T r(i) - b(s(t), \theta) \right) - \alpha \frac{\partial}{\partial \theta} \left( \sum_{i=t}^T r(i) - b(s(t), \theta) \right)^2 \right]. \quad (1)$$

Here  $r(t)$  is reward given at time  $t$ , and the hyperparameter  $\alpha$  is for balancing the reward and the baseline objectives, which set to 0.03 in all experiments.

### 3 Communication Model

We now describe the model used to compute  $p(\mathbf{a}(t)|\mathbf{s}(t), \theta)$  at a given time  $t$  (omitting the time index for brevity). Let  $s_j$  be the  $j$ th agent's view of the state of the environment. The input to the controller is the concatenation of all state-views  $\mathbf{s} = \{s_1, \dots, s_J\}$ , and the controller  $\Phi$  is a mapping  $\mathbf{a} = \Phi(\mathbf{s})$ , where the output  $\mathbf{a}$  is a concatenation of discrete actions  $\mathbf{a} = \{a_1, \dots, a_J\}$  for each agent. Note that this single controller  $\Phi$  encompasses the individual controllers for each agents, as well as the communication between agents.

One obvious choice for  $\Phi$  is a fully-connected multi-layer neural network, which could extract features  $\mathbf{h}$  from  $\mathbf{s}$  and use them to predict good actions with our RL framework. This model would allow agents to communicate with each other and share views of the environment. We refer to this type of communication as *dense* later in the experiments. However, it is inflexible with respect to the composition and number of agents it controls; cannot deal well with agents joining and leaving the group and even the order of the agents must be fixed. On the other hand, if no communication is used then we can write  $\mathbf{a} = \{\phi(s_1), \dots, \phi(s_J)\}$ , where  $\phi$  is a per-agent controller applied independently. This communication-free model satisfies the flexibility requirements<sup>1</sup>, but is not able to coordinate agents' actions. We refer to this as the *none* communication model.

#### 3.1 Controller Structure

We now detail our architecture for  $\Phi$  that has the modularity of the communication-free model but still allows communication.  $\Phi$  is built from modules  $f^i$ , which take the form of multilayer neural networks. Here  $i \in \{0, \dots, K\}$ , where  $K$  is the number of communication steps in the network.

Each  $f^i$  takes two input vectors for each agent  $j$ : the hidden state  $h_j^i$  and the communication  $c_j^i$ , and outputs a vector  $h_j^{i+1}$ . The main body of the model then takes as input the concatenated vectors  $\mathbf{h}^0 = [h_1^0, h_2^0, \dots, h_J^0]$ , and computes:

$$h_j^{i+1} = f^i(h_j^i, c_j^i) \quad (2)$$

$$c_j^{i+1} = \frac{1}{J-1} \sum_{j' \neq j} h_{j'}^{i+1}. \quad (3)$$

In the case that  $f^i$  is a single linear layer followed by a nonlinearity  $\sigma$ , we have:  $h_j^{i+1} = \sigma(H^i h_j^i + C^i c_j^i)$  and the model can be viewed as a feedforward network with layers  $\mathbf{h}^{i+1} = \sigma(T^i \mathbf{h}^i)$  where  $\mathbf{h}^i$  is the concatenation of all  $h_j^i$  and  $T^i$  takes the block form (where  $C^i = C^i/(J-1)$ ):

<sup>1</sup>Assuming  $s_j$  includes the identity of agent  $j$ .

$$T^i = \begin{pmatrix} H^i & \bar{C}^i & \bar{C}^i & \dots & \bar{C}^i \\ \bar{C}^i & H^i & \bar{C}^i & \dots & \bar{C}^i \\ \bar{C}^i & \bar{C}^i & H^i & \dots & \bar{C}^i \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{C}^i & \bar{C}^i & \bar{C}^i & \dots & H^i \end{pmatrix},$$

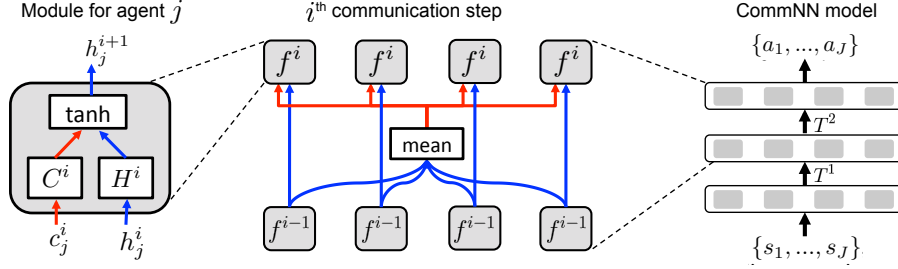


Figure 1: An overview of our communication model. Left: view of module  $f^i$  for a single agent  $j$ . Note that the parameters are shared across all agents. Middle: a single communication step, where each agents modules propagate their internal state  $h$ , as well as broadcasting a communication vector  $c$  on a common channel (shown in red). Right: full model, showing input states  $s$  for each agent, two communication steps and the output actions for each agent.

A key point is that  $T$  is *dynamically sized* since the number of agents may vary. This motivates the the normalizing factor  $J - 1$  in equation (3), which rescales the communication vector by the number of communicating agents. Note also that  $T^i$  is permutation invariant, thus the order of the agents does not matter.

At the first layer of the model an encoder function  $h_j^0 = r(s_j)$  is used. This takes as input state-view  $s_j$  and outputs feature vector  $h_j^0$  (in  $\mathbb{R}^{d_0}$  for some  $d_0$ ). The form of the encoder is problem dependent, but for most of our tasks it is a single layer neural network. Unless otherwise noted,  $c_j^0 = 0$  for all  $j$ . At the output of the model, a decoder function  $q(h_j^K)$  is used to output a distribution over the space of actions.  $q(\cdot)$  takes the form of a single layer network, followed by a softmax. To produce a discrete action, we sample from this distribution:  $a_j \sim q(h_j^K)$ .

Thus the entire model (shown in Fig. 1), which we call a Communication Neural Net (CommNN), (i) takes the state-view of all agents  $s$ , passes it through the encoder  $h^0 = r(s)$ , (ii) iterates  $h$  and  $c$  in equations (2) and (3) to obtain  $h^K$ , (iii) samples actions  $a$  for all agents, according to  $q(h^K)$ . We refer to this type communication as *continuous* type because communication is based on continuous-valued vectors.

### 3.2 Model Extensions

**Local Connectivity:** An alternative to the broadcast framework described above is to allow agents to communicate to others within a certain range. Let  $N(j)$  be the set of agents present within communication range of agent  $j$ . Then (3) becomes:

$$c_j^{i+1} = \frac{1}{|N(j)|} \sum_{j' \in N(j)} h_{j'}^{i+1}. \quad (4)$$

As the agents move, enter and exit the environment,  $N(j)$  will change over time. In this setting, our model has a natural interpretation as a dynamic graph, with  $N(j)$  being the set of vertices connected to vertex  $j$  at the current time. The edges within the graph represent the communication channel between agents, with (4) being equivalent to belief propagation [22]. Furthermore, the use of multi-layer nets at each vertex makes our model similar to an instantiation of the GGSNN work of Li *et al.* [14].

**Temporal Recurrence:** We also explore having the network be a recurrent neural network (RNN). This is achieved by simply replacing the communication step  $i$  in Eqn. (2) and (3) by a time step  $t$ ,

and using the same module  $f^t$  for all  $t$ . At every time step, actions will be sampled from  $q(h_j^t)$ . Note that agents can leave or join the swarm at any time step. If  $f^t$  is a single layer network, we obtain plain RNNs that communicate with each other. In later experiments, we also use an LSTM as an  $f^t$  module.

**Skip Connections:** For some tasks, it is useful to have the input encoding  $h_j^0$  present as an input for communication steps beyond the first layer. Thus for agent  $j$  at step  $i$ , we have:

$$h_j^{i+1} = f^i(h_j^i, c_j^i, h_j^0). \quad (5)$$

### 3.3 Discrete Communication

An alternate way for agents to communicate is via discrete symbols, with the meaning of these symbols being learned during training. Since discrete outputs are not differentiable, reinforcement learning is used to train in this setting. However, unlike actions in the environment, an agent has to output a discrete symbol at every communication step. But if these are viewed as *internal* time steps of the agent, then the communication output can be treated as an action of the agent at a given (internal) time step and we can directly employ the policy gradient [36].

At communication step  $i$ , agent  $j$  will output the index  $w_j^i$  corresponding to a particular symbol, sampled according to:

$$w_j^i \sim \text{Softmax}(Dh_j^i) \quad (6)$$

where matrix  $D$  is the model parameter. Let  $\hat{w}$  be a 1-hot binary vector representation of  $w$ . In our broadcast framework, at the next step the agent receives a bag of vectors from all the other agents (where  $\wedge$  is the element-wise OR operation):

$$c_j^{i+1} = \bigwedge_{j' \neq j} \hat{w}_{j'}^i \quad (7)$$

## 4 Related Work

Our model combines a deep network with reinforcement learning [9, 20, 13]. Several recent works have applied these methods to multi-agent domains, such as Go [16, 25] and Atari games [30], but they assume full visibility of the environment and lack communication. There is a rich literature on multi-agent reinforcement learning (MARL) [2], particularly in the robotics domain [18, 26, 6, 21, 3]. Amongst fully cooperative algorithms, many approaches [12, 15, 34] avoid the need for communication by making strong assumptions about visibility of other agents and the environment. Others use communication, but with a pre-determined protocol [31, 19, 38, 17].

A few notable approaches involve learning to communicate between agents under partial visibility: Kasai *et al.* [10] and Varshavskaya *et al.* [33], both use distributed tabular-RL approaches for simulated tasks. Giles & Jim [7] use an evolutionary algorithm, rather than reinforcement learning. Guestrin *et al.* [8] use a single large MDP to control a collection of agents, via a factored message passing framework where the messages are learned. In common with our approach, this avoids an exponential blowup in state and action-space. In contrast to these approaches, our model uses a deep network for both agent control and communication.

From a MARL perspective, the closest approach to ours is the concurrent work of Foerster *et al.* [5]. This also uses a deep reinforcement learning in multi-agent partially observable tasks, specifically two riddle problems (similar in spirit to our *levers* task) which necessitate multi-agent communication. Like our approach, the communication is learned rather than being pre-determined. However, the agents communicate in a discrete manner through their actions. This contrasts with our model where multiple continuous communication cycles are used at each time step to decide the actions of all agents. Furthermore, our approach is amenable to dynamic variation in the number of agents.

The Neural GPU [39] has similarities to our model but differs in that a 1-D ordering on the input is assumed and it employs convolution, as opposed to the global pooling in our approach (thus permitting unstructured inputs). Our model can be regarded as an instantiation of the GNN construction of Scarselli *et al.* [24], as expanded on by Li *et al.* [14]. In particular, in [24], the output of the model is the fixed point of iterating equations (4) and (2) to convergence, using recurrent models. In [14], these recurrence equations are unrolled a fixed number of steps and the model trained via backprop through time. In this work, we do not require the model to be recurrent, neither do we aim to reach steady state. Additionally, we regard Eqn. (4) as a pooling operation, conceptually making our model a single feed-forward network with local connections.

## 5 Experiments

### 5.1 Lever Pulling Task

We start with a very simple game that requires the agents to communicate in order to win. This consists of  $m$  levers and a pool of  $N$  agents. At each round,  $m$  agents are drawn at random and they must each choose a lever to pull, simultaneously with the other  $m - 1$  agents, after which the round ends. The goal is for each of them to pull a *different* lever. Correspondingly, all agents receive reward proportional to the number of distinct levers pulled. Each agent can see its own identity, and nothing else, thus  $s_j = j$ .

We implement the game with  $m = 5$  and  $N = 500$ . We use a CommNN with two communication steps ( $K = 2$ ) and skip connections from (5). The encoder  $r$  is a lookup-table with  $N$  entries of 128D. Each  $f^i$  is a two layer neural net with ReLU nonlinearities that takes in the concatenation of  $(h^i, c^i, h^0)$ , and outputs a 128D vector. The decoder is a linear layer plus softmax, producing a distribution over the  $m$  levers, from which we sample to determine the lever to be pulled. We also use a communication free version of the model that has  $c$  zeroed during training. The results are shown in Table 1. The metric is the number of distinct levers pulled divided by  $m = 5$ , averaged over 500 trials, after seeing 50000 batches of size 64 during training. We explore both reinforcement (see (1)) and direct supervision (where the sorted ordering of agent IDs are used as targets). In both cases, the CommNN with communication performs significantly better than the version without it.

Communication	Training method	
	Supervised	Reinforcement
None	0.59	0.59
Continuous	0.99	0.94

Table 1: Results of lever game (#distinct levers pulled)/(#levers) for our CommNN and no communication models, using two different training approaches. Allowing the agents to communicate enables them to succeed at the task.

### 5.2 Cooperative Games

In this section, we consider two multi-agent tasks in the MazeBase environment [27] that use reward as their training signal. The first task is to control cars passing through a traffic junction to maximize the flow while minimizing collisions. The second task is to control multiple agents in combat against enemy bots.

We experimented with several module types. With a feedforward MLP, the module  $f^i$  is a single layer network and  $K = 2$  communication steps are used. For an RNN module, we also used a single layer network for  $f^t$ , but shared parameters across time steps. Finally, we used an LSTM for  $f^t$ . In all modules, the hidden layer size is set to 50. All the models use skip-connections. Both tasks are trained for 300 epochs, each epoch being 100 weight updates with RMSProp [32] on mini-batch of 288 game episodes (distributed over multiple CPU cores). In total, the models experience  $\sim 8.6$ M episodes during training. We repeat all experiments 5 times with different random initializations, and report mean value along with standard deviation. The training time varies from a few hours to a few days depending on task and module type.

#### 5.2.1 Traffic Junction

This consists of a 4-way junction on a  $14 \times 14$  grid as shown in Fig. 2(left). At each time step, new cars enter the grid with probability  $p_{\text{arrive}}$  from each of the four directions. However, the total number of cars at any given time is limited to  $N_{\text{max}} = 10$ . Each car occupies a single cell at any given time and is randomly assigned to one of three possible routes (keeping to the right-hand side of the road). At every time step, a car has two possible actions: *gas* which advances it by one cell on its route or *brake* to stay at its current location. A car will be removed once it reaches its destination at the edge of the grid.

Two cars *collide* if their locations overlap. A collision incurs a reward  $r_{\text{coll}} = -10$ , but does not affect the simulation in any other way. To discourage a traffic jam, each car gets reward of  $\tau r_{\text{time}} = -0.01\tau$  at every time step, where  $\tau$  is the number time steps passed since the car arrived. Therefore, the reward at time  $t$  is:

$$r(t) = C^t r_{\text{coll}} + \sum_{i=1}^{N^t} \tau_i r_{\text{time}},$$

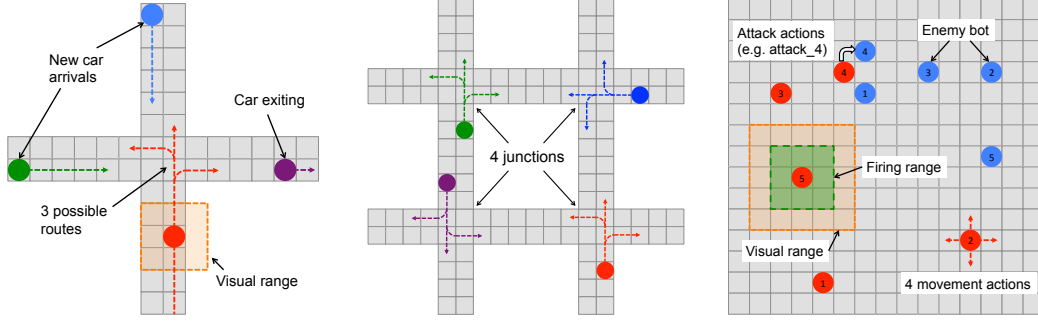


Figure 2: Left: Traffic junction task where agent-controlled cars (colored circles) have to pass the through the junction without colliding. Middle: A harder version with four connected junctions. Right: The combat task, where model controlled agents (red circles) fight against enemy bots (blue circles). In both tasks each agent has limited visibility (orange region), thus is not able to see the location of all other agents.

Communication type	MLP	Modules RNN	LSTM
None	20.6 ± 14.1	19.5 ± 4.5	9.4 ± 5.6
Continuous	2.2 ± 0.6	7.6 ± 1.4	1.6 ± 1.0
Dense	12.5 ± 4.4	-	-
Discrete	20.2 ± 11.2	-	-

Communication type	Other game versions Easy (MLP)	Hard (RNN)
None	15.8 ± 12.5	26.9 ± 6.0
Continuous	0.3 ± 0.1	22.5 ± 6.1
Cont. local	-	21.1 ± 3.4
Discrete	1.1 ± 2.4	-

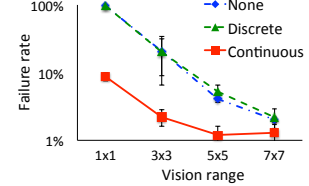


Table 2: Traffic junction task. Left: failure rates (%) for different types of communication and module function  $f(\cdot)$ . Continuous consistently improves performance, over the dense baseline and no communication. Middle: Game variants. In the easy case, discrete communication does help, but still less than continuous. On the hard version, local communication (see Section 3.2) does at least as well as broadcasting to all agents. Right: As visibility in the environment decreases, the importance of communication grows.

where  $C^t$  is the number of collisions occurring at time  $t$ , and  $N^t$  is number of cars present. The simulation is terminated after 40 steps and is classified as a failure if one or more collisions have occurred. Details of the input representation, training and other game variations can be found in Appendix A.

In Table 2, we show the probability of failure of a variety of different module/communication method pairs. Continuous communication between cars significantly reduces the failure rate for all module types. Discrete communication did not give any benefit, except for the easy game. We also tried a dense communication baseline by allowing the matrix  $T$  to be arbitrary, resulting in a single large fully-connected network controlling all agents. However, this did not work as well as continuous communication (a video showing this model before and after training can be found at <https://youtu.be/onK98y-UNHQ>). We also explore how partial visibility within the environment effects the advantage given by communication. As the vision range of each agent decreases, the advantage of communication increases. Impressively, with zero visibility (the cars are driving blind) the continuous communication model is still able to succeed 90% of the time.

### 5.2.2 Analysis of Communication

We now attempt to understand what the agents communicate when performing the junction task. We start by recording the hidden state  $h_j^i$  of each agent and the corresponding *communication vectors*  $\tilde{c}_j^{i+1} = C^{i+1}h_j^i$  (the contribution agent  $j$  at step  $i + 1$  makes to the hidden state of other agents). Fig. 3(left) and Fig. 3(right) show the 2D PCA projections of the communication and hidden state vectors respectively. These plots show a diverse range of hidden states but far more clustered communication vectors, many of which are close to zero. This suggests that while the hidden state carries information, the agent often prefers not to communicate it to the others unless necessary. This is a possible consequence of the broadcast channel: if everyone talks at the same time, no-one can understand. See Appendix B for norm of communication vectors and brake locations.

To better understand the meaning behind the communication vectors, we ran the simulation with only two cars and recorded their communication vectors and locations whenever one of them braked.

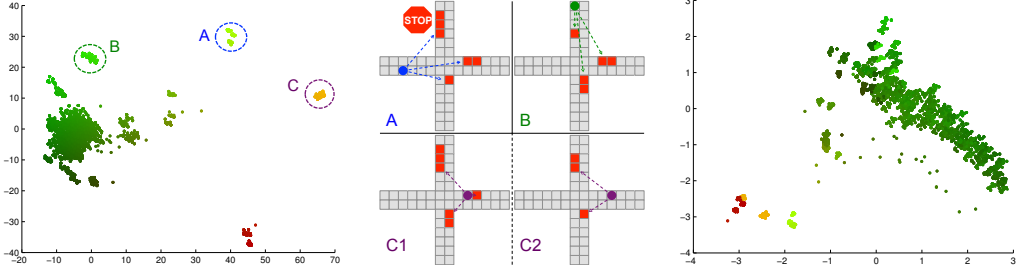


Figure 3: Left: First two principal components of communication vectors  $\tilde{c}$  from multiple runs on the traffic junction task Fig. 2(left). While the majority are “silent” (i.e. have a small norm), distinct clusters are also present. Middle: for three of these clusters, we probe the model to understand their meaning (see text for details). Right: First two principal components of hidden state vectors  $h$  from the same runs as on the left, with corresponding color coding. Note how many of the “silent” communication vectors accompany non-zero hidden state vectors. This shows that the two pathways carry different information.

Vectors belonging to the clusters A, B & C in Fig. 3(left) were consistently emitted when one of the cars was in a specific location, shown by the colored circles in Fig. 3(middle) (or pair of locations for cluster C). They also strongly correlated with the other car braking at the locations indicated in red, which happen to be relevant to avoiding collision.

### 5.2.3 Combat Task

We simulate a simple battle involving two opposing teams in a  $15 \times 15$  grid as shown in Fig. 2(right). Each team consists of  $m = 5$  agents and their initial positions are sampled uniformly in a  $5 \times 5$  square around the team center, which is picked uniformly in the grid. At each time step, an agent can perform one of the following actions: move one cell in one of four directions; attack another agent by specifying its ID  $j$  (there are  $m$  attack actions, each corresponding to one enemy agent); or do nothing. If agent A attacks agent B, then B’s health point will be reduced by 1, but only if B is inside the firing range of A (its surrounding  $3 \times 3$  area). Agents need one time step of cooling down after an attack, during which they cannot attack. All agents start with 3 health points, and die when their health reaches 0. A team will win if all agents in the other team die. The simulation ends when one team wins, or neither of teams win within 40 time steps (a draw).

The model controls one team during training, and the other team consist of bots that follow a hard-coded policy. The bot policy is to attack the nearest enemy agent if it is within its firing range. If not, it approaches the nearest visible enemy agent within visual range. An agent is visible to all bots, if it is inside the visual range of any individual bot. This shared vision gives an advantage to the bot team. When input to a model, each agent is represented by a set of one-hot binary vectors  $\{i, t, l, h, c\}$  encoding its unique ID, team ID, location, health points and cooldown. A model controlling an agent also sees other agents in its visual range ( $3 \times 3$  surrounding area). The model gets reward of -1 if the team loses or draws at the end of the game. In addition, it also get reward of  $-0.1$  times the total health points of the enemy team, which encourages it to attack enemy bots.

Table 3 shows the win rate of different module choices with various types of communication. Among different modules, the LSTM achieved the best performance. Continuous communication improved all module types. With the MLP module, we tried dense and discrete communication types but they degraded performance relative to no communication. We also explored several variations of the task: varying the number of agents in each team by setting  $m = 3, 10$ , and increasing visual range of agents to  $5 \times 5$  area. The result on those tasks are shown on the right side of Table 3. Using continuous communication (CommNN model) consistently improves the win rate, even with the greater environment observability of the  $5 \times 5$  vision case.

Communication type	Modules		
	MLP	RNN	LSTM
None	34.2 $\pm$ 1.3	37.3 $\pm$ 4.6	44.3 $\pm$ 0.4
Continuous	44.5 $\pm$ 13.4	44.4 $\pm$ 11.9	49.5 $\pm$ 12.6
Dense	17.7 $\pm$ 7.1	-	-
Discrete	26.9 $\pm$ 9.4	-	-

Communication type	Other game variations (MLP)		
	$m = 3$	$m = 10$	$5 \times 5$ vision
None	29.2 $\pm$ 5.9	30.5 $\pm$ 8.7	60.5 $\pm$ 2.1
Continuous	51.0 $\pm$ 14.1	45.4 $\pm$ 12.4	73.0 $\pm$ 0.7

Table 3: Win rates (%) on the combat task for different communication approaches and module choices. Continuous consistently outperforms the other approaches. The dense baseline does worse than using no communication at all. On the right we explore the effect of varying the number of agents  $m$  and agent visibility. Even with 10 agents on each team, communication clearly helps.

### 5.3 bAbI tasks

We apply our model to the bAbI [35] toy Q & A dataset, which consists of 20 tasks each requiring different kind of reasoning. The goal is to answer a question after reading a short story. We can formulate this as a multi-agent task by giving each sentence of the story its own agent. Communication among agents allows them to exchange useful information necessary to answer the question.

The input is  $\{s_1, s_2, \dots, s_J, q\}$ , where  $s_j$  is  $j$ 'th sentence of the story, and  $q$  is the question sentence. We use the same encoder representation as [28] to convert them to vectors. The  $f(\cdot)$  module consists of a two-layer MLP with ReLU non-linearities. After  $K = 3$  communication steps, we sample an output word  $y$  from a softmax decoder layer. The model is trained in a supervised fashion using a cross-entropy loss between  $y$  and the correct answer  $y^*$ . The hidden layer size is set 100 and weights are initialized from  $N(0, 0.2)$ . We train the model 100 epochs with learning rate 0.003 and mini-batch size 32 with Adam optimizer [11] ( $\beta_1 = 0.9, \beta_2 = 0.99, \epsilon = 10^{-6}$ ). We used 10% of training data as validation set to find optimal hyper-parameters for the model.

Results on the 10K version of the bAbI task are shown in Table 4, along with other baselines (see Appendix C for a detailed breakdown). Our model is doing better than LSTM baseline, but worse than the MemN2N model [28], which is specifically designed to solve reasoning over long stories. However, it successfully solves most of the tasks, including ones that require information sharing between two or more agents through communication.

	Mean error (%)	Failed tasks (err. > 5%)
LSTM [28]	36.4	16
MemN2N [28]	4.2	3
DMN+ [37]	2.8	1
CommNN (MLP, no communication)	15.2	9
CommNN (MLP module)	7.1	3

Table 4: Experimental results on bAbI tasks

## 6 Discussion and Future Work

We have introduced CommNN, a simple controller for MARL that is able to learn continuous communication between a dynamically changing set of agents. Evaluations on four diverse tasks clearly show the model outperforms models without communication, “dense” baselines, and models using discrete communication. Despite the simplicity of the broadcast channel, examination of the traffic task reveals the model to have learned a sparse communication protocol that conveys meaningful information between agents.

One aspect of our model that we did not fully exploit is its ability to handle heterogenous agent types and we hope to explore this in future work. Furthermore, we believe the model will scale gracefully to large numbers of agents, perhaps requiring more sophisticated connectivity structures; we also leave this to future work.

### Acknowledgements

The authors wish to thank Daniel Lee and Y-Lan Boureau for their advice and guidance.



## References

- [1] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.
- [2] L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(2):156–172, 2008.
- [3] Y. Cao, W. Yu, W. Ren, and G. Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics*, 1(9):427–438, 2013.
- [4] R. H. Crites and A. G. Barto. Elevator group control using multiple reinforcement learning agents. *Machine Learning*, 33(2):235–262, 1998.
- [5] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson. Learning to communicate to solve riddles with deep distributed recurrent q-networks. *arXiv*, abs/1602.02672, 2016.
- [6] D. Fox, W. Burgard, H. Kruppa, and S. Thrun. Probabilistic approach to collaborative multi-robot localization. *Autonomous Robots*, 8(3):325–344, 2000.
- [7] C. L. Giles and K. C. Jim. Learning communication for multi-agent systems. In *Innovative Concepts for Agent Based Systems*, pages 377–390. Springer, 2002.
- [8] C. Guestrin, D. Koller, and R. Parr. Multiagent planning with factored mdps. In *NIPS*, 2001.
- [9] X. Guo, S. Singh, H. Lee, R. L. Lewis, and X. Wang. Deep learning for real-time atari game play using offline monte-carlo tree search planning. In *NIPS*, 2014.
- [10] T. Kasai, H. Tenmoto, and A. Kamiya. Learning of communication codes in multi-agent reinforcement learning problem. *IEEE Conference on Soft Computing in Industrial Applications*, pages 1–6, 2008.
- [11] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations*, 2015.
- [12] M. Lauer and M. A. Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings of the Seventeenth International Conference on Machine Learning, ICML ’00*, 2000.
- [13] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *arXiv:1504.00702*, 2015.
- [14] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel. Gated graph sequence neural networks, 2015.
- [15] M. L. Littman. Value-function reinforcement learning in markov games. *Cognitive Systems Research*, 2(1):55–66, 2001.
- [16] C. J. Maddison, A. Huang, I. Sutskever, and D. Silver. Move evaluation in go using deep convolutional neural networks. In *ICLR*, 2015.
- [17] D. Maravall, J. De Lope, and R. Dominguez. Coordination of communication in robot teams by reinforcement learning. *Robotics and Autonomous Systems*, 61(7):661–666, 2013.
- [18] M. Matari. Reinforcement learning in the multi-robot domain. *Autonomous Robots*, 4(1):73–83, 1997.
- [19] F. S. Melo, M. Spaan, and S. J. Witwicki. Querypomdp: Pomdp-based communication in multiagent systems. In *Multi-Agent Systems*, pages 189–204, 2011.
- [20] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- [21] R. Olfati-Saber, J. Fax, and R. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.
- [22] J. Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *AAAI*, pages 133–136, 1982.
- [23] B. Peng, Z. Lu, H. Li, and K. Wong. Towards Neural Network-based Reasoning. *ArXiv preprint: 1508.05508*, 2015.
- [24] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2009.
- [25] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- [26] P. Stone and M. Veloso. Towards collaborative and adversarial learning: A case study in robotic soccer. *International Journal of Human Computer Studies*, (48), 1998.
- [27] S. Sukhbaatar, A. Szlam, G. Synnaeve, S. Chintala, and R. Fergus. Mazebase: A sandbox for learning from games. *CoRR*, abs/1511.07401, 2015.
- [28] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. *NIPS*, 2015.
- [29] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1998.
- [30] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, and R. Vicente. Multiagent cooperation and competition with deep reinforcement learning. *arXiv:1511.08779*, 2015.
- [31] M. Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *ICML*, 1993.

- [32] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.
- [33] P. Varshavskaya, L. P. Kaelbling, and D. Rus. *Distributed Autonomous Robotic Systems 8*, chapter Efficient Distributed Reinforcement Learning through Agreement, pages 367–378. Springer Berlin Heidelberg, 2009.
- [34] X. Wang and T. Sandholm. Reinforcement learning to play an optimal nash equilibrium in team markov games. In *Advances in neural information processing systems*, pages 1571–1578, 2002.
- [35] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. In *arXiv preprint: 1502.05698*, 2015.
- [36] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine Learning*, pages 229–256, 1992.
- [37] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. *CoRR*, abs/1603.01417, 2016.
- [38] C. Zhang and V. Lesser. Coordinating multi-agent reinforcement learning with limited communication. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS ’13, pages 1101–1108, 2013.
- [39] Łukasz Kaiser and I. Sutskever. Neural gpu learn algorithms, 2015.

## A Traffic Junction

Each car is represented by one-hot binary vector set  $\{n, l, r\}$ , that encode its unique ID, current location and assigned route number respectively. Each agent controlling a car can only observe other cars in the surrounding  $3 \times 3$  neighborhood (however it can communicate to all cars). The state vector for each agent is thus a concatenation of all these vectors, having dimension  $3^2 \times |n| \times |l| \times |r|$ . We use curriculum learning [1] to make the training easier. In first 100 epochs of training, we set  $p_{\text{arrive}} = 0.05$ , but linearly increased it to 0.2 during next 100 epochs. Finally, training continues for another 100 epochs. The learning rate is fixed at 0.003 throughout. We also implemented additional easy and hard versions of the game, the latter being shown in Fig.2(middle).

The easy version is a junction of two one-way roads on a  $7 \times 7$  grid. There are two arrival points, each with two possible routes. During curriculum, we increase  $N_{\text{total}}$  from 3 to 5, and  $p_{\text{arrive}}$  from 0.1 to 0.3.

The harder version consists from four connected junctions of two-way roads in  $18 \times 18$  as shown in Fig.2(center). There are 8 arrival points and 7 different routes for each arrival point. We set  $N_{\text{total}} = 20$ , and increased  $p_{\text{arrive}}$  from 0.02 to 0.05 during curriculum.

## B Traffic Junction Analysis

Here we visualize the average norm of the communication vectors and brake locations over the  $14 \times 14$  spatial grid.

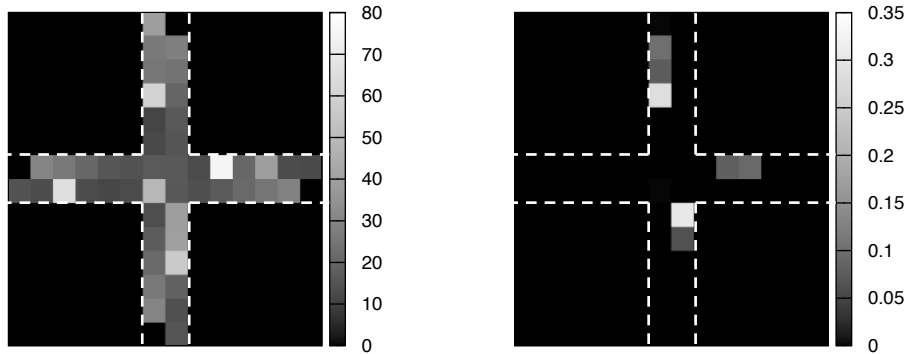


Figure 4: (left) Average norm of communication vectors (right) Brake locations

## C bAbI tasks

Here we give further details of the model setup and training, as well as a breakdown of results in Table 4.

Let the task be  $\{s_1, s_2, \dots, s_J, q, y^*\}$ , where  $s_j$  is  $j$ 'th sentence of story,  $q$  is the question sentence and  $y^*$  is the correct answer word (when answer is multiple words, we simply concatenate them into single word). Then the input to the model is

$$h_j^0 = r(s_j, \theta_0), \quad c_j^0 = r(q, \theta_q).$$

Here, we use simple position encoding [28] as  $r$  to convert sentences into fixed size vectors. Also, the initial communication is used to broadcast the question to all agents. Since the temporal ordering of sentences is relevant in some tasks, we add special temporal word " $t = j$ " to  $s_j$  for all  $j$ .

For  $f$  module, we use a 2 layer network with skip connection, that is

$$h_j^{i+1} = \sigma(W_i \sigma(A^i h_j^i + B^i c_j^i + h_j^0)),$$

where  $\sigma$  is ReLU non-linearity (bias terms are omitted for clarity). After  $K = 3$  communication steps, the model outputs an answer word by

$$y = \text{Softmax}(D \sum_{j=1}^J h_j^K)$$

Since we have the correct answer during training, we will do supervised learning by using cross entropy cost on  $\{y^*, y\}$ . The hidden layer size is set 100 and weights are initialized from  $N(0, 0.2)$ . We train the model 100 epochs with learning rate 0.003 and mini-batch size 32 with Adam optimizer [11] ( $\beta_1 = 0.9, \beta_2 = 0.99, \epsilon = 10^{-6}$ ). We used 10% of training data as validation set to find optimal hyper-parameters for the model.

	Error on tasks (%)							Mean error (%)	Failed tasks (err. > 5%)
	2	3	15	16	17	18	19		
LSTM [28]	81.9	83.1	78.7	51.9	50.1	6.8	90.3	36.4	16
MemN2N [28]	0.3	2.1	0.0	51.8	18.6	5.3	2.3	4.2	3
DMN+ [37]	0.3	1.1	0.0	45.3	4.2	2.1	0.0	2.8	1
Neural Reasoner+ [23]	-	-	-	-	0.9	-	1.6	-	-
CommNN (MLP, no communication)	69.0	69.5	29.4	47.4	4.0	0.6	45.8	15.2	9
CommNN (MLP module)	3.2	68.3	0.0	51.3	15.1	1.4	0.0	7.1	3

Table 5: Experimental results on bAbI tasks. Only showing some of the task with high errors.