Modelling "model of others" using Deep Q-Network

Tuan Anh Nguyen
n.a.tuan@apple.ee.uec.jp
The University of Electro-communications

Chie Hieida
hchie@apple.ee.uec.ac.jp
The University of Electro-communications

Takayuki Nagai
tnagai@ee.uec.ac.jp
The University of Electro-communications

ABSTRACT

The ability to anticipate others intention is an essential part of human interaction. Humans do this by observing other people's behaviour and from that clue trying to acquire the "model of others" which is one mechanism of predicting what others may do in a particular situation. In this study, we consider the building block task in which two agents have to build a particular shape together. In this task, two agents have different goal as first and in order to complete the task, the agents have to be able to predict each other's intention and adapt their strategies so that they have the same goal. Using Deep Q-Network (DQN) as the decision-making model of the agents, we first update each agent's model by training them to do the task individually so they can learn how the task can be completed. After that, two agents will be set to do the task together. By observing each other behaviour and learning, we expect that their models will be modified according to these new experiences to the point that they can learn to anticipate other's intention and change their goals to a single one.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; See http://acm.org/about/class/1998/ for the full list of ACM classifiers. This section is required.

Author Keywords

model of others; reinforcement learning; intention inference.

INTRODUCTION

In order to effectively cooperate with other people in certain tasks we have to understand other's intention correctly. As human, however we cannot know exactly other's intention but still cooperate well enough because we can somehow predict that intention by just observing people's behaviour and adjust our own behaviour in an appropriate way. The existence of mirror neurons show that we may use our own internal decision-making model both when we decide our own action as well as when we "read" other's intention. However, sometime our prediction may not match our partner's intention and may lead to misunderstanding. Needless to say that it is less confused when cooperating with people we had worked together before than whom we barely knew about. Learning how to behave in a certain situation may provide the ability to predict other's behaviour in the very same situation, but the more we interact with and observe particular people behaviour the more accurate our estimation can become. We can hypothesise this as first of all we learn our own internal model and then when we interact with other people we use our model to starting predict what they will probably do. We may do this by inputting information about "others" instead of ourself into the decision-making model. As we interact, we slowly learn how to configure the model according to our new observations of their behaviour to get better one. The same thing can be applied to human - robot interaction; if robot can learn and use their own internal models to predict the intention of human then they are more likely to communicate better with us.

However, what exactly is that kind of model and how it works is still not fully understood. In recent years, DQN, which combines the brain-inspired artificial neural network and the process of learning from trial-and-error through reward signal of Reinforcement Learning, has successfully learn how to play a variety of Atari 2600 video games from only screen pixels and game score by itself. This is similar to how human learns to do real life task. Therefore, we think DQN can be used as a simple "model of others" that human uses in predicting other's intention.

We think that the process in which human learns from interacting with environment can be described as in the left side of Figure 1. At first, we have a goal to guide our actions. Our action then changes the state of environment. By perceiving this change, we can compare the new state of environment with our goal and give feedback to the decision-making model about the contribution of our action in achieving the goal. From that feedback (refer as reward), our decision-making model will be trained to select its actions better.

As we interact with others, the learning process slightly changes. We propose the framework of the interaction between two agents A and B as illustrated in the right side of Figure 1 in which agent A is observing and learning to cooperate from the behaviour of agent B. The learning process is the same for agent B. We suppose that they have two different goals at first. They will try to change the state of environment so that it get close to their goals as much as possible. Obviously both agents will not achieve their goal when they act only on their own interest. They need to learn to cooperate. It means two agents need to be encouraged to change its goal to match with that of the other. However, as same as not everyone is easily set aside their own interest, how much each agent

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

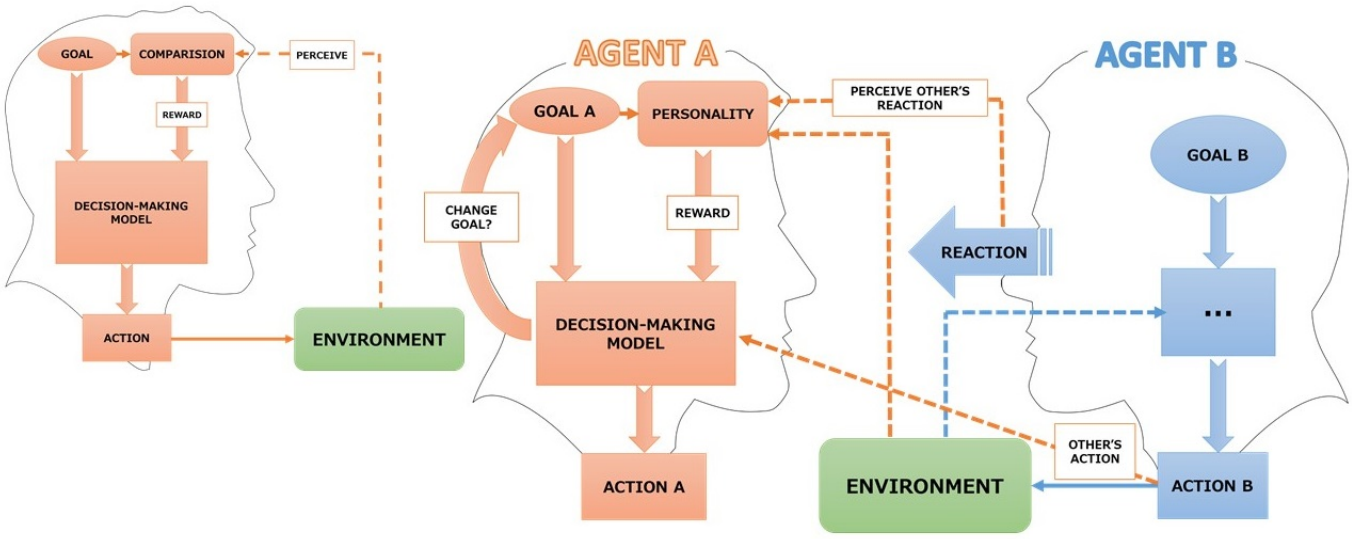


Figure 1: Framework of interacting and learning

is willing to change its goal and cooperate is not the same. In this framework, we refer to that as the personality of the agent. When agent B takes an action, environment changes to a new state. Agent A can perceive this change and compare that with its goal. Besides, agent A can also perceive agent B's reaction to the change. The reaction can be in the form of verbal communication, facial expression etc. From these two feedbacks, depending on its personality, a reward will be generated in the combination of how much agent A values agent B's reaction over its own interest. This reward then will be used as feedback to the decision-making model to reinforce the good action. The same evaluate process happens when agent A takes the action itself. The fact that human can understand an action is performed by ourself or other can be expressed in this framework as an input signal of self or other to the model with respect to who performed the action.
To examine the above assumption, we simplify the process of interaction and cooperation into a simple simulation task of block building. Each agent has its own goal of building a particular shape. When these two shapes are the same, the task is easy to complete. However, when these two shapes are different, one agent has to be able to predict the other's goal and adjust its goal accordingly. If we can use DQN as the decision-making model that helps the agents success in this situation then probably we can learn something about the model that human uses in predicting the other's intention. The detail of the building block task will be described in next section. After that, the learning process using DQN and the simulation of that process will be discussed.

BUILDING BLOCK TASK

The task aims at simulating a simple situation required interaction and cooperation between two agents. The goal of this task is to use white piece of blocks to build a pre-assigned shape (refer as goal-shape) on a 4x4 grid board as showed in Figure 2(a). Two agent with two different goal-shapes will take turn to stack blocks until either one of two goal-shapes is completed. Therefore, the success of the task then depends on whether two agents can successfully predict the other's goal and adapt their behaviour accordingly.
The agents in this task can perform 20 actions: 16 actions corresponding to 16 positions on the grid board that it can stack blocks into and 4 actions to select which goal-shape in four pre-assigned goal-shapes to build. However, as real world has many constrain due to the laws of physics, the environment in this task also has following restrictions:

1. A block can only be stacked above another block or at the bottom of the board.
2. If the agent choose a position that already had another block, the action will be perceived as taking out that block. However, only a block that has no block above it can be taken out.
3. If it take more than a certain steps to complete the task then the task is terminated and reset.

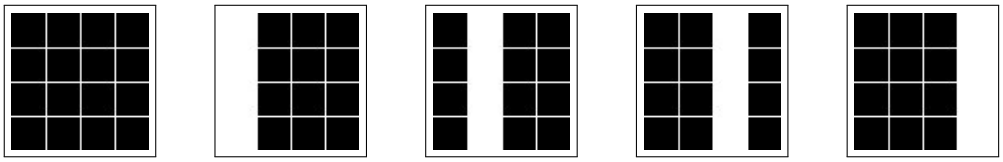
At first, each agent will be trained individually to learn how to complete the task in regard of the above restrictions and generates their own decision-making model. We call this the training phase. Then, two agent are set to do the task together and learn from the behaviour of the other in the cooperating phase. We expect that by observing the other's behaviour and using that new experiences to update their own models, the agents can automatically learn how to use these models to predict the other's goal as well as adjust their strategies to cooperate.

MODEL AND TRAINING PROCESS

As mentioned before, we will use DQN as the decision-making model for each agent and train that model to be able to predict the other's goal.
DQN is a convolutional neural network (CNN) trained to approximate the function of estimating the value of an action given a particular input state which is also referred as

Layer	Input	Filter size	Stride	Activation	Output
conv1	84x84x2	8x8	4	ReLu	20x20x32
conv2	20x20x32	4x4	2	ReLu	9x9x64
conv3	9x9x64	3x3	1	ReLu	7x7x64
fc4	7x7x64			ReLu	512
fc5	512			Linear	20

Table 1: Network architecture of DQN used in the task



(a) Grid board (b) Goal No.1 (c) Goal No.2 (d) Goal No.3 (e) Goal No.4

Figure 2: Grid board and training phase goal-shapes

Q-function. DQN algorithm stores all of the agent's experiences as tuples of transition <state, action, reward, next state> and randomly samples these experiences to trained the CNN. For our learning process, each element of the experiences is described below.

- The state s will be a set of two 84x84 pixels images: the current state of the grid board, the goal-shape.
- The action a will be one of 20 actions that the agent can perform as mentioned in the task description.
- The reward r is the feedback of how much the action contributes to the purpose of completing its goal-shape and is decided as follow:
  - If the agent successfully stacked a block, for the position belong to the goal-shape, it receives $r = 2$; for the position not belong to the goal-shape, it receives $r = -1$.
  - If the agent successfully took out a block, for the position that belong to the goal-shape, it receives $r = -4$; for the position not belong to the goal-shape, it receives $r = 1$.
  - If the agent failed to stack or take out block, it receives $r = -1$.
  - If the agent select one of the four goal-shape changing actions, it receives $r = -5$.
  - If the agent successfully complete the task, it receives $r = 40$. Over the step limit has a penalty of $r = -10$.

Except for the input and output, the DQN's network architecture is the same as in [1] and is specified in Table 1. The two agent's DQNs are also trained with double Q-learning method [3] which helps reducing the overestimation of Q-value of normal DQN.
We use the RL-Glue framework [2] to simulate the reinforcement learning experiment.

SIMULATION EXPERIMENT

In order to cooperate in the task, each agent need to "understand" how the task can be completed. Therefore, they will be trained to do the task by themselves first. We refer to this as training phase. Then they will be set to do the task together in cooperative phase.

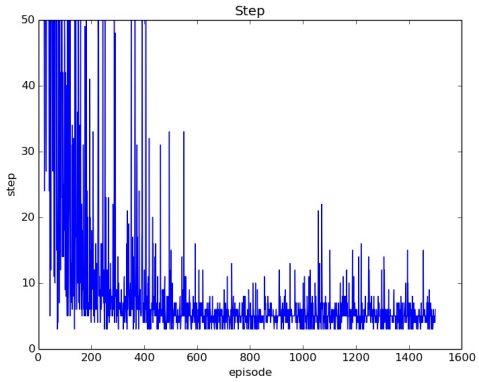
Training Phase

In this phase two agents are trained separately. The goal-shapes used in this training phase are the four shapes showed in Figure 2 and were randomly assigned to the agent at the beginning of each episode. To diversify the experiences of the agent, we randomly place one block at one of four positions at the bottom at the beginning of each episode before agent can take the first action.
We trained each agent for over 150000 episodes. The steps limit of each episode was set to 50 steps. We record the total steps the agent took and total reward that it received in each episode. The result is showed in Figure 3 with each point in both graph is the average value of every 100 episodes. The result shows that the agent successfully learned how to complete the task.

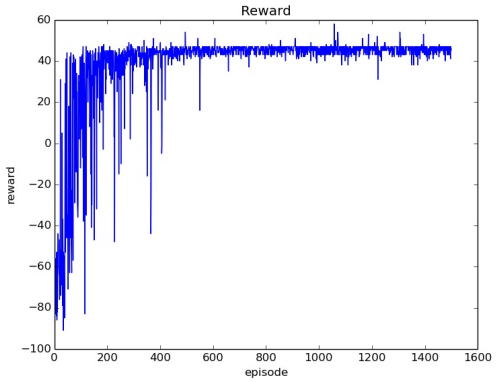
Cooperating Phase

After two agent successfully learned the task separately, we will set them to do that on the same grid board and they will take turn to take action. Two agent will have two different goal-shapes at the beginning of each episode. During the cooperating phase, the agents will not only learn from their own experiences but also from the other agent behaviour. An episode ends when either one of them achieve their goal-shapes at that time or they take more than a certain steps to do so. One episode of the task in this phase is conducted as follow:

- First, agent A takes action a_t in state s_t and transform the environment to state s_{t+1}. The reward agent A receives is the sum reward between its own reward and reward from agent B's feedback. In detail, by taking action a_t agent A



(a) Tracking of agent's avarage steps



(b) Tracking of agent's average reward

Figure 3: Result of training phase

receives a reward r^A. At the same time, agent B observes the new state and also virtually receives a reward r^B. Both r_t^A and r_t^B are calculated in the same ways as described in the training phase except for the penalty of changing goal may be set different for each agent depending on their personality. The sum reward for agent A is then $r_t^A = \alpha_1 r_t^A + \alpha_2 r_t^B$ in which $\alpha_1 + \alpha_2 = 1$. The more we want agent A to cooperate the smaller we weight its reward by decreasing α_1. The sum reward for agent B is computed the same but with different α_1 and α_2.

- Next, agent B observes the state s_{t+1} as the consequence of agent A's action and decide to take action a_{t+1} to transform the environment to state s_{t+2}. Agent B receives the reward r_{t+1}^B for taking action a_{t+1} and agent A receives reward r_{t+1}^A because of that action. The sum reward for agent A is again $r_{t+1}^A = \alpha_1 r_{t+1}^A + \alpha_2 r_{t+1}^B$. The reward for agent B is also computed similarly.

- Then we set $t \leftarrow t+2$ and repeat from step 1 until the episode ends.

The DQN of both agents will then be trained using the experiences of both itself and the observation of the other agent's behaviour. During the interaction, each agent's model will receive an input signal to tell it that which experiences are the consequence of which agent's actions.

DISCUSSION AND CONCLUSION

In this article, for the purpose of studying the mechanism of predicting other people's intention, we proposed a plan to conduct a cooperative task between two agents with different goal at first. The plan is divided into two phases: training phase in which each agent learns their own decision-making model by completing the task individually and cooperating phase in which they update their model according to their new experiences of doing the task together. DQN is used in the training phase to be the decision-making model. The result shows that with DQN the agent had learned the task well. We then will continue to the next phase of setting up two agents doing the task together in the future simulation.

Although the building block task is a simple task, if the agents can successfully adapt their goal to the other's one we will have a model that can be used both to decide one own action as well as estimate the other's action. Then we can further study on the ability to apply the model to other tasks in real life to help robot cooperate well with human. Moreover, by adjusting the values of α_1, α_2 as well as the penalty that each agent receives when changing goal, we can investigate on which kinds of personalities are suitable to cooperate together. From that result, we may further consider how agent can build a good relationship with human through cooperating.

ACKNOWLEDGEMENT

This work was supported by JSPS Grant-in-Aid for Scientific Research on Innovative Areas (Cognitive Interaction Design) and JSPS KAKENHI Grant Number JP16J04930.

REFERENCES

1. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei a Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533. DOI: http://dx.doi.org/10.1038/nature14236

2. Brian Tanner and Adam White. 2009. RL-Glue : Language-Independent Software for Reinforcement-Learning Experiments. *Journal of Machine Learning Research* 10 (2009), 2133–2136.

3. Hado van Hasselt, Arthur Guez, and David Silver. 2015. Deep Reinforcement Learning with Double Q-learning. *arXiv:1509.06461 [cs]* (2015). http://arxiv.org/abs/1509.06461$\delimiter"026E30F$nhttp://www.arxiv.org/pdf/1509.06461.pdf