Attention Guided Deep Imitation Learning

Ruohan Zhang*[1] (zharu@utexas.edu)

Zhuode Liu*[1] (zhuode93@cs.utexas.edu)

Mary M. Hayhoe[2] (hayhoe@utexas.edu)
[2]Center for Perceptual Systems, The University of Texas at Austin, Austin, TX, 78712, USA

Dana H. Ballard[1] (dana@cs.utexas.edu)
[1]Department of Computer Science, The University of Texas at Austin, Austin, TX, 78712, USA

Abstract

When a learning agent attempts to imitate human visuomotor behaviors, it may benefit from knowing the human demonstrator's visual attention. Such information could clarify the goal of the demonstrator, i.e., the object being attended is the most likely target of the current action. Hence it could help the agent better infer and learn the demonstrator's underlying state representation for decision making. We collect human control actions and eye-tracking data for playing Atari games. We train a deep neural network to predict human actions, and show that including gaze information significantly improves the prediction accuracy. In addition, more biologically correct representation enhances prediction accuracy.

Keywords: Imitation Learning; Representation Learning; Eye movements; Visual Attention

Introduction

A learning agent can benefit from the demonstration of human experts. However, directly imitating a humans is challenging for a machine, forestalled by the different underlying human task representation. Although the environment presents the same visual stimulus to humans and the machine, the underlying decision states may not be the same, partially due to the differences in perceptual systems. While most machines use full-resolution cameras, humans have foveal vision with high acuity for only 1-2 visual degrees covering the width of a finger at arms length. This leads to discrepancy in perceived states of human and machine, where the machine perceives images like in Figs. 1a while a human may see Figs. 1c-1d.

How do we help learning agents infer the demonstrator's decision states? A traditional method would ask human experts to provide the representation such as hand-engineered task-relevant features, an approach challenged by deep neural networks due to their power in representation learning which we prefer in this work. The question becomes whether we can provide a large amount of useful information in addition to human actions to help the deep network learn the representation more efficiently.

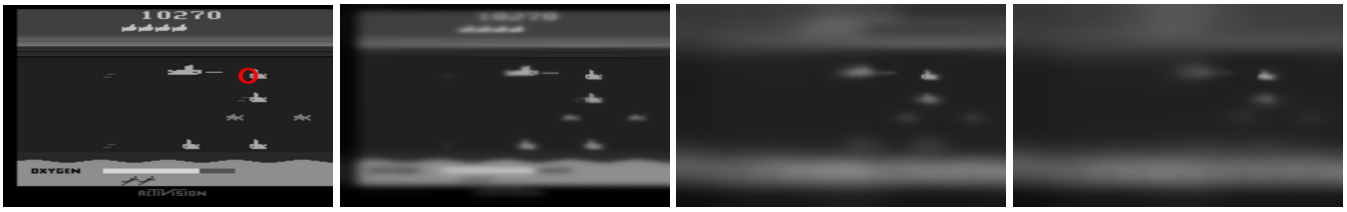
A foveal visual system may seem inferior compared to a full resolution camera, but it leads to an outstanding property of human intelligence: the visual attention mechanism. Humans manage to move their foveae to the correct place at the right time to perceive important task-relevant features (Diaz, Cooper, Rothkopf, & Hayhoe, 2013; Rothkopf, Ballard, & Hayhoe, 2007). In this work, we propose to approach the deep imitation learning problem with human gaze data, which is a good indicator of the demonstrator's attention and can be collected efficiently using modern high-speed eye trackers. We hypothesize that such information can help a learning agent better imitate a human demonstrator's behaviors.

Methods and Results

Data We collected human game-playing actions playing Atari games using the Arcade Learning Environment (Bellemare, Naddaf, Veness, & Bowling, 2012). The games only proceed when the subject takes an action to allow enough response time and obtain good policies. At each time step t, the raw image frame I_t, human keystroke action a_t, and gaze position g_t were recorded. The gaze data was recorded using an EyeLink 1000 eye tracker at 1000Hz. The game screen is 64.6×40.0cm and the distance to the subjects' eyes is 78.7cm. For this work we use 30-minute data each from Breakout, Seaquest, and Ms.Pacman. We are in the progress of collecting more data from different games and more subjects. The dataset will be made publicly available.

Baseline model We first trained a deep network with standard supervised learning to predict human actions. The network architecture follows the Deep Q-Network (Mnih et al., 2015) which has two convolutional layers followed by two fully connected layers. The image preprocessing procedure follows the same work, hence the input to the network is the Y-channel of a raw image frame (Fig.1). **Premasking** Our second model treats gaze as saliency information in the pixel space. We use gaze position g_t to create a gaze heatmap similar to a saliency map (Itti, Koch, & Niebur, 1998), in which a Gaussian filter ($\sigma = 25$) centered at gaze position is applied to create a mask. Then we directly multiply the mask with I_t element-wise. The mask has the effect of emphasizing the stimulus centered at the gaze. The resulted output is then fed into the same network structure as the baseline.

Foveated rendering We hypothesize that training the network with realistic retinal images may improve prediction, since these images are closer to the true human representation. We fed the visual angle of the game screen (45 degrees), gaze positions, and images into the Space Variant Imaging



(a) Original image with gaze (b) Foveated image, 10 degrees (c) 45 degrees (the true value) (d) 80 degrees

Figure 1: The original game frames for Atari Seaquest with red circles indicating the gaze position. The gaze position is used to generate the foveated images. Visual degree indicates the size of the game screen in the visual field.

system (Perry & Geisler, 2002)[1]. The software provides biological plausible simulation of the foveated retinal images as shown in Figs. 1b. The foveated images are fed into the network.

Results The prediction accuracy is shown in Table 1. Including gaze information, whether by premasking or foveated rendering, significantly improves the performance where the latter one has clear advantage except for Ms.Pacman.

	Seaquest	Ms.Pacman	Breakout
Random	16.67	20.00	25.00
Baseline	41.44±0.34	34.44±0.18	55.37±0.53
Premasking	49.18±0.23	43.22±0.29	59.39±0.27
Foveated	54.49±0.10	41.73±0.25	67.95±0.28

Table 1: Percentage accuracy in predicting human actions across three games (mean ± standard error).

The importance of correct visual angle We vary the visual angle to change clarity of the images, as if the subjects were viewing the game screen from various distances, as shown in Figs. 1c-1d. The actual visual angle is 45 degrees in our experiments. We observe that the prediction accuracy decreases when deviated from the true value, as shown in Fig. 2, but they all outperforms the baseline (41.44%) significantly.

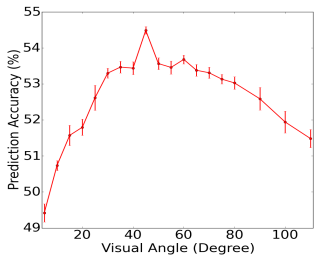


Figure 2: Action prediction accuracy for models trained using foveated images with different visual angles. 45 degrees is the correct visual angle that yields the best performance.

Conclusion and Future Work

Foveal visual attention information helps a deep learning agent perform an imitation learning task. Such information can be used to provide the agent with the stimulus that human actually perceives, or feed into the model as a saliency map. Although a large enough network with enough data may potentially learn the retinal foveated representation without gaze, gaze information simplifies the learning problem by weighting important visual features more.

Although our model has shown improvement over the baseline, there is much room for future work. Due to human visuomotor response time, action a_t may not be conditioned on the image and gaze at time t, but on images and gazes several hundreds milliseconds ago. More importantly, the human memory system allows states of previous fixated objects to be preserved, and an internal model may perform model-based prediction to update the environmental states in memory. These cognitive functionalities could be readily implemented by deep networks models such as a recurrent neural network to allow a better prediction of human actions.

[1] http://www.cps.utexas.edu/svi/

Acknowledgments

The research is supported by NIH Grant EY05729 and NSF Grant CNS 1624378.

References

Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. (2012). The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*.

Diaz, G., Cooper, J., Rothkopf, C., & Hayhoe, M. (2013). Saccades to future ball location reveal memory-based prediction in a virtual-reality interception task. *Journal of vision*, *13*(1), 20–20.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Perry, J. S., & Geisler, W. S. (2002). Gaze-contingent real-time simulation of arbitrary visual fields. In *Electronic imaging 2002* (pp. 57–69).

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of vision*, *7*(14), 16–16.