Visual Attention Guided Deep Imitation Learning

Ruohan Zhang[*] **Zhuode Liu**[*]
The University of Texas at Austin

Luxin Zhang
Peking University

Karl S. Muller **Mary M. Hayhoe** **Dana H. Ballard**
The University of Texas at Austin

Abstract

When an intelligent agent learns to imitate human visuomotor behaviors, it may benefit from knowing where the human is allocating visual attention, which can be inferred from their gaze. A wealth of information regarding intelligent decision making is conveyed by human gaze allocation; hence, exploiting such information has the potential to improve the agent's performance. With this motivation, we collect high-quality human action and gaze data while playing Atari games. We first train a deep neural network that can predict human gaze positions and visual attention with high accuracy (the gaze network) and then train another deep neural network to predict human actions (the policy network). Including the gaze predictions from the gaze network in the policy network significantly improves the action prediction accuracy. We conclude that it is feasible to learn human attention in the given visuomotor tasks, and that combining the learned attention model with imitation learning yields promising results.

1 Introduction

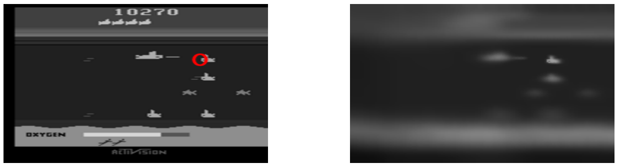
A learning agent can benefit from imitating human experts [1]. For deep imitation learning in visuomotor tasks–where the input could be in raw pixel space–an algorithm needs to consider the following factors: 1). Humans have a unique sensory system that is different from machines' and this leads to different perceived decision states. 2). The traits of this sensory system lead to gaze behaviors and visual attention–intelligent mechanisms that are not yet available to the learning agent. Without this mechanisms, it is difficult for the agent to infer which visual features are being attended and are relevant for the decision at the moment.

Humans have high acuity foveal vision in the central 1-2 visual degrees of the visual field, with resolution decreasing in the periphery. This leads to a discrepancy in the perceived states of a human and a machine, where the machine perceives images like in Fig. 1a while a human may see Figs. 1b. 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 in order to emphasize important task-relevant features [8]. In this way, a wealth of information is encoded in human gaze behaviors–for example, the importance of one object over another in performing an action.

We hypothesize that a promising approach to strengthen a deep imitation learning algorithm is to model human visual attention through gaze, and subsequently include such a model in the imitation learning process. This would allow the learning agent to use gaze information to help decipher the internal state representation used by the human. By extracting features that are most important for tasks, the learning agent may better imitate a human demonstrator's behaviors.

[*] Both authors contributed equally. Contact: zharu@utexas.edu

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.



(a) Original image with gaze (b) Human perceived image

Figure 1: (a) A game frame for Atari Seaquest with a red circle indicating the human gaze position. The gaze position is used to generate a foveated image (b) that is biologically plausible retinal representations of the visual stimulus (perceived by the human) using the algorithm in [7].

2 Gaze Prediction

We collected human data using eight Atari games in the Arcade Learning Environment [2]. The raw image frame, the human keystroke action, and the gaze position were recorded. The gaze data was recorded using an EyeLink 1000 eye tracker at 1000Hz.

We formalize visual attention modeling as an end-to-end saliency prediction problem, whereby a deep network can be used to predict a probability distribution of the gaze (saliency map). The full network architecture (Fig. 2) is a three-channel convolution-deconvolution network. The inputs to the top channel are the images where the preprocessing procedure follows [6] and hence consist of a sequence of 4 frames stacked together where each frame is 84 × 84 in grayscale. The mid channel models motion information (i.e., optical flow calculated using [3]) which is included since human gaze is sensitive to movement. The bottom channel includes bottom-up saliency map from the classic Itti-Koch model [4]. We then average the output of the all channels. The output of the network is a gaze saliency map trained with Kullback-Leibler divergence as the loss function. We use the same architecture and hyperparameters for all eight games.

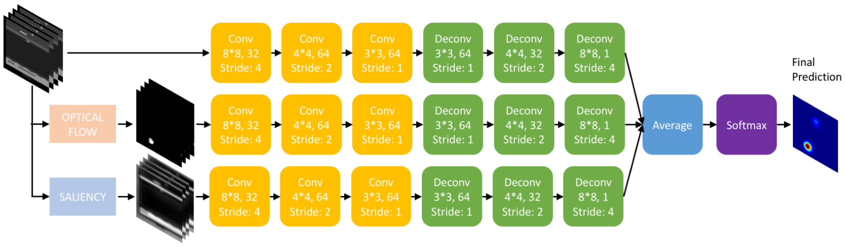


Figure 2: The three-channel gaze network. The top channel takes in images, the mid channel takes in the corresponding optical flow , and the bottom channel takes in the bottom-up image saliency. The final output is a gaze saliency map that indicates the predicted probability distribution of the gaze.

For a performance comparison we use the classic bottom-up saliency model [4] as the baseline. Then an ablation study is performed where the model consists only one or two channels of the original network. The performance are evaluated using Area Under the Curve (AUC) and Kullback-Leibler divergence (KL) as in Table 1. Overall, the prediction results of all our models are highly accurate across all games and largely outperform the bottom-up saliency baseline, indicated by the high AUC (above 0.90 for all games) and low KL values obtained. Using image or motion alone each gives reasonable performance, while combining these two channels produces the best results for most games. Including bottom-up saliency into the model does not improve the performance in general. This indicates that in the given tasks, the top-down visual attention is different than and hard to be inferred from the traditional bottom-up image saliency. We encourage readers to view the video demo of the prediction results at `https://youtu.be/PR_1wVOj3BU`.

3 Policy Learning

Given a gaze network that can accurately predict visual attention, we can incorporate it into a policy network that learns a human's control policy. A deep network is trained with supervised learning to

	Saliency(S)		Image(I)		Motion(M)		I+M		I+S		I+M+S	
	AUC↑	KL↓	AUC↑	KL↓	AUC↑	KL↓	AUC↑	KL↓	AUC↑	KL↓	AUC↑	KL↓
Breakout	0.494	4.375	0.970	1.304	0.968	1.357	0.970	**1.294**	0.969	1.301	0.969	1.299
Freeway	0.560	4.289	0.973	1.261	0.972	1.287	0.972	**1.257**	0.973	1.260	0.973	1.260
Enduro	0.447	4.517	0.988	0.834	0.987	0.844	0.988	**0.832**	0.988	0.834	0.988	0.835
Riverraid	0.494	4.235	0.962	1.609	0.951	1.829	0.962	1.593	0.961	1.613	0.962	**1.592**
Seaquest	0.352	4.744	0.963	1.464	0.959	1.540	0.964	1.438	0.963	1.470	0.964	**1.437**
MsPacman	0.426	4.680	0.932	1.985	0.919	2.217	0.935	**1.959**	0.933	1.995	0.936	1.961
Centipede	0.691	3.774	0.956	1.711	0.958	1.686	0.961	**1.622**	0.957	1.709	0.960	1.645
Vecnture	0.607	3.868	0.957	1.749	0.956	1.648	0.964	1.512	0.956	1.727	0.964	**1.510**

Table 1: Quantitative results of predicting human gaze across eight games. Random prediction baseline: AUC = 0.500, KL = 6.159. As a benchmark, the classic [4] algorithm is compared to versions of our algorithm. Overall the Image+Motion model achieves the best accuracy.

classify human actions given the current frame. The baseline network architecture follows the Deep Q-Network [6] and will be referred as the NoGaze network.

We then treat the predicted gaze heatmap from gaze network as a saliency mask and multiply the mask with image frame element-wise. The mask has the effect of emphasizing the stimulus being attended. The resulting output is then fed into the same network structure. We will refer to this model as the Masking model. A masked image highlights the visual information being perceived at the moment, but it may lose other task-relevant information that human stored in memory. To compensate for this effect, we feed both the original image and the masked image into a two-channel deep network. The model is referred as the Masking+Image model. The final architecture is shown in Fig. 3. The prediction results are shown in Table 2. It is clear that the Masking+Image model has an advantage over the NoGaze baseline. In particular, results for games that often require multitasking show large improvements in performance: 15.1% on Seaquest, 16% on MsPacman, 5.1% on Centipede, and 6.7% on Venture. We conclude that including gaze information significantly improves the performance in terms of policy matching accuracy.

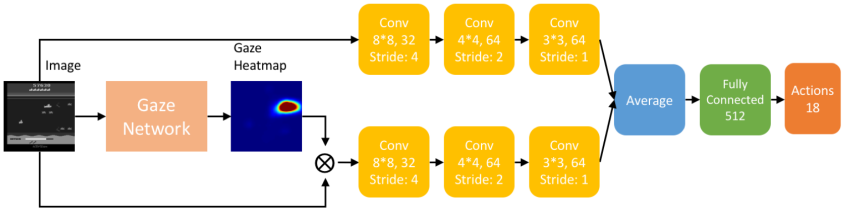


Figure 3: The final policy network architecture for imitating human actions. The top channel takes in the current image frame and the bottom channel takes in the masked image which is an element-wise product of the original image and predicted gaze saliency map by the gaze network.

	Action Prediction Accuracy			Game Score	
	NoGaze	Masking	Masking+Image	NoGaze	Policy+Gaze
Breakout	81.5 ± 0.3	82.5 ± 0.3	**86.2 ± 0.2**	0.9 ± 1.1	**8.1 ± 5.2**
Freeway	**96.7 ± 0.0**	96.3 ± 0.1	96.4 ± 0.2	**30.2 ± 0.9**	29.4 ± 1.2
Enduro	60.6 ± 0.4	**62.1 ± 0.6**	61.9 ± 0.3	**155.0 ± 39.5**	106.9 ± 26.7
Riverraid	**72.5 ± 0.3**	59.2 ± 0.3	**72.5 ± 0.4**	2448.4 ± 701.3	**3182.4 ± 1050.6**
Seaquest	46.0 ± 1.8	60.5 ± 0.2	**61.6 ± 0.2**	244.0 ± 89.3	**441.9 ± 369.9**
MsPacman	54.6 ± 1.0	70.2 ± 0.2	**70.6 ± 0.3**	509.8 ± 409.8	**937.2 ± 605.8**
Centipede	61.9 ± 0.2	64.1 ± 0.4	**67.0 ± 0.3**	5001.9 ± 3083.1	**7047.5 ± 3659.5**
Venture	46.7 ± 0.2	**54.0 ± 0.4**	53.3 ± 0.3	**307.0 ± 156.4**	179.0 ± 198.6

Table 2: Left: Percentage accuracy in predicting human actions across eight games using different models. Random prediction baseline: 5.56%. The Masking+Img model yields the best prediction accuracy. Right: Game scores for combining the gaze network and the policy network to play the game versus the plain deep imitation learning model. Results are presented in mean ± standard deviation.

Another goal of imitation learning is to learn a good policy to actually perform the task. Hence, we use the trained policy network in Fig. 3 and the gaze network in Fig. 2 to play the games and record the scores. The agent chooses actions probabilistically according to the prediction.

The average games scores over 100 episodes per game are reported in Table 2. The Policy+Gaze model outperforms the baseline on five out of eight games. In general, these are the games that policy network predicts human actions more accurately than the NoGaze model, such as Breakout, Seaquest, MsPacman, and Centipede. An exception is Venture where the prediction accuracy is low for both models. For Seaquest and MsPacman which have the largest improvement on prediction accuracy, the game scores also improve the most. For Riverraid, the prediction results are similar but the Policy+Gaze network learns a better policy indicated by the higher score. The NoGaze model performs better on Enduro and Freeway which are both racing games. By investigating the game playing behaviors, we find that Policy+Gaze model learns to avoid obstacles and is more cautious than the NoGaze model; however in these games a bold agent can achieve better performance.

4 Conclusion and Future Work

A question for imitation learning research is: what should be learned from the human demonstrator? The agent could learn the state representation, or the policy, or the reward function, or some high-level cognitive functionality such as attention. This work represents the first use of attention for imitation learning. Directly imitating human actions does not work well even with a large dataset [5] and people have been developing insights about why this is the case. Our work tackles this problem from a new angle and suggests that failing to take human visual attention into account is one factor that contributes to the poor prediction accuracy and performance hence cannot not be overlooked.

Why does visual attention help action prediction and task performance? First, the gaze information could help to identify and disambiguate the goal of current action when multiple task-relevant objects are present. For example, in Fig. 1b the gaze indicates that the goal of current action involves the enemy on the right side of the screen. While there are many other task-relevant objects nearby, the gaze information helps to indicate which of these is most relevant for the current action. Second, gaze could sometimes be more precise and informative than hand-engineered features. Consider [8] in which they show that humans look at the edges of obstacles to avoid them, and the centers of targets to approach them. For both of these reasons, knowing the gaze information can help the agent to infer the correct decision state of the human.

The results of [6] have demonstrated the effectiveness of end-to-end learning of visuomotor tasks, where the DQN excels at games that involve a single task. However, for games such as Seaquest and MsPacman–which typically involve multiple tasks–the performance is still below human levels. In addition, DQN takes millions of samples to train. The above issues could be potentially alleviated by having an attention model that extracts features to speedup learning and to indicate task priority.

Acknowledgments

This research was supported by NSF grant CNS-1624378 and NIH grant EY05729 .

References

[1] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[2] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 2012.

[3] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. *Image analysis*, pages 363–370, 2003.

[4] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[5] Vitaly Kurin, Sebastian Nowozin, Katja Hofmann, Lucas Beyer, and Bastian Leibe. The atari grand challenge dataset. *arXiv preprint arXiv:1705.10998*, 2017.

[6] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[7] Jeffrey S Perry and Wilson S Geisler. Gaze-contingent real-time simulation of arbitrary visual fields. In *Electronic Imaging 2002*, pages 57–69. International Society for Optics and Photonics, 2002.

[8] Constantin A Rothkopf, Dana H Ballard, and Mary M Hayhoe. Task and context determine where you look. *Journal of vision*, 7(14):16–16, 2007.