Learning Attention Model from Human for Visuomotor Tasks

Luxin Zhang[1]* Ruohan Zhang[2] Zhuode Liu[2] Mary M. Hayhoe[3] Dana H. Ballard[2]
[1] Department of Intelligence Science, Peking University, Beijing, China, 100871
[2] Department of Computer Science, The University of Texas at Austin
[2] Center for Perceptual Systems, The University of Texas at Austin
* zhangluxin@pku.edu.cn; (86)17888837918

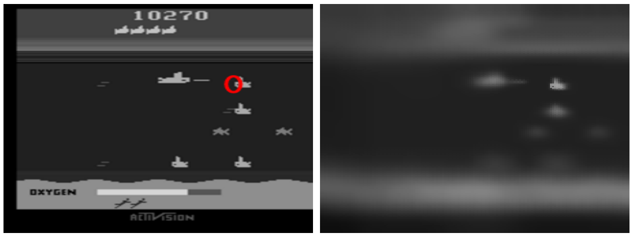
Abstract

A wealth of information regarding intelligent decision making is conveyed by human gaze and visual attention, hence, modeling and exploiting such information might be a promising way to strengthen algorithms like deep reinforcement learning. We collect high-quality human action and gaze data while playing Atari games. Using these data, we train a deep neural network that can predict human gaze positions and visual attention with high accuracy.

Introduction

Humans have high acuity foveal vision in the central 1-2 visual degrees of the visual field (i.e., covering the width of a finger at arms length), with resolution decreasing exponentially in the periphery. While the machine perceives images like in Fig. 1a, a human would see Fig. 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 (Rothkopf, Ballard, and Hayhoe 2007). In this way, a wealth of information is encoded in human gaze behaviors–for example, the priority of one object over another in performing an action.

Previous work in computer vision has formalized visual attention modeling as a saliency prediction problem where saliency is mostly derived from image statistics, such as intensity, color, and orientation in the classic Itti-Koch model (Itti, Koch, and Niebur 1998). This approach has made tremendous progress due to large benchmark datasets and deep neural networks (Bylinskii et al. 2016). In contrast, top-down models emphasizes the effects of task-dependent variables on visual attention (Hayhoe and Ballard 2005). (Rothkopf, Ballard, and Hayhoe 2007) have shown that varying task instructions drastically alters the gaze distributions on different categories of objects (e.g., task-irrelevant objects are ignored even though they are salient). The top-down attention model is hence closely related to reinforcement learning since they both concern visual state features that matter the most for acquiring the reward.



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

Figure 1: (a) An original 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 (the stimulus as perceived by the human) using the algorithm in (Perry and Geisler 2002).

Regardless of their approaches, these works argue that there is much valuable information encoded in gaze behaviors. It should be said that the two approaches are not mutually exclusive, since attention is modulated in both saliency-driven and volition-controlled manners (Itti, Koch, and Niebur 1998). As mentioned before, deep neural networks have been a standard approach to predict bottom-up saliency. In contrast, top-down gaze models often rely on manually defined task variables. Our approach seeks to combine these approaches and use the representation learning power of deep networks to extract task-relevant visual features, given task-driven gaze data.

Experiments and Results

We collected human data using eight Atari games in the Arcade Learning Environment (Bellemare et al. 2012). 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 (see Appendix 1 in supplementary materials) is a three-channel convolution-deconvolution network. The inputs to the top

	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 (Itti, Koch, and Niebur 1998) algorithm is compared to versions of our algorithm. Overall the Image+Motion model achieves the best accuracy.

channel are the images where the preprocessing procedure follows (Mnih et al. 2015) 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 (Farnebäck 2003)) which is included since human gaze is sensitive to movement. The bottom channel includes bottom-up saliency map from the classic Itti-Koch model (Itti, Koch, and Niebur 1998). 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.

For a performance comparison we use the classic bottom-up saliency model (Itti, Koch, and Niebur 1998) 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`. Example prediction results for all games are shown in Appendix 2.

Conclusion and Future Work

The high accuracy achieved in predicting gaze in our work implies that, given a cognitively demanding visuomotor task, human visual attention can be modeled accurately using an end-to-end learning algorithm. This suggests that popular deep saliency models could be used to model top-down visual attention, given task-driven data.

It is worth noting that the results are obtained using a single human subject's data. We are in progress of collecting data from multiple subjects. Given high accuracy of our model, it is possible to test whether the visual attention models generalize across human subjects for the same game.

Such learned model could potentially be helpful for visuomotor behavior learning algorithms. The results of (Mnih et al. 2015) have demonstrated the effectiveness of end-to-end learning of visuomotor tasks, where the Deep Q-Network (DQN) excels at games that involve a single task. However, for multitasking games such as Seaquest and MsPacman the performance is still below human levels. In addition, DQN takes millions of samples to train. The above issues could be potentially alleviated by incorporating our attention model that help extract features to speedup learning and to indicate task priority.

References

[Bellemare et al. 2012] Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2012. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*.

[Bylinskii et al. 2016] Bylinskii, Z.; Recasens, A.; Borji, A.; Oliva, A.; Torralba, A.; and Durand, F. 2016. Where should saliency models look next? In *European Conference on Computer Vision*, 809–824. Springer.

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

[Hayhoe and Ballard 2005] Hayhoe, M., and Ballard, D. 2005. Eye movements in natural behavior. *Trends in cognitive sciences* 9(4):188–194.

[Itti, Koch, and Niebur 1998] Itti, L.; Koch, C.; and 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 et al. 2015] Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

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

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