Gaze & Motion Prioritization for Atari Games

Krishanu Agrawal
Electrical & Computer Engineering
kagarwal87@gatech.edu

Ezra Ameperosa
Electrical & Computer Engineering
eameperosa3@gatech.edu

Mudit Gupta
College of Computing
mgupta303@gatech.edu

Himanshu Vairagade
Electrical & Computer Engineering
himanshuv@gatech.edu

I. Background and Motivation

Many agents have been designed, created, and trained to play video games or imitate other tasks, but many recent works show that incorporating attention mechanisms, specifically human gaze, into an agent's training can greatly increase performance [1]. Motion within visual input data can also be used to drive the attention of RL/IL agents.

Though both of these mechanisms can improve agent performance, we hope to investigate whether the conditional application of these attention mechanisms alongside an auxiliary loss can be used to further improve performance. This work, if successful, can help differentiate Atari games that simply require focusing on motion versus games that require focusing on specific regions as learned by human gaze data.

II. Relevance and Related Works

A. Atari-HEAD: Demonstration Dataset

Visual attention, a significant challenge in reinforcement learning (RL) and imitation learning (IL), allows humans to reduce the set of prominent features from visual input [2]. Visual attention driven by the human gaze could prove to be a powerful tool the agent could use to learn a decision policy that could help bridge attention and control. The collected data, Atari-HEAD (Atari Human Eye-Tracking and Demonstration) [3], supplies near-optimal demonstrations with millions of human gaze samples across 20 different games.

B. AGIL: Attention Guided Imitation Learning

[4] show that properly using human gaze data can greatly increase an IL agent's performance. They use a 3-channel network to learn a model of attention. These three channels are raw image data, motion data based on the algorithm proposed in [5], and saliency data based on a bottom-up deep network. This three-channel gaze network learns a model of the gaze (relative to a gaussian kernel representation for raw gaze data points). The action prediction network then uses two convolutional channels, one which takes in raw images and a second which takes in images scaled by the learned gaze prediction.

C. Selective Use of Predicted Gaze for Attention

Thammeneni et. al. explain that gaze is not always useful in driving the attention of an IL agent [6]. Eye-movement cascades and object fixation often convey little information about focus). The authors propose Selective Eye-gaze Augmentation (SEA), which uses three separate networks. A gaze prediction auto-encoder network is trained independently to learn the gaze heatmap for each frame with KL-divergence loss. A gating network is then trained alongside the action prediction network which concatenates the latent-space embedding provided by the encoder from the gaze network with the gated values of the gaze prediction. The gating network uses modified GRU units with weights affected by gaze usage and action prediction error such that it learns optimal usage of gaze. In many cases, the gating network doesn't forward the gaze prediction. SEA did worse than AGIL in 4/6 tested games, likely because the gating network slows down training [6].

D. Auxilary Loss to Drive Attention Towards Human Gaze

The authors of [7] propose an auxiliary gaze loss during the training of ~~imitation learning~~ algorithms to improve performance of existing methods without increasing model complexity, data requirements, or requiring test-time gaze. The novelty of their work is that gaze is not required at test time and instead is used as a weak supervisory signal, in contrast to several prior approaches. The loss function, Coverage-based Gaze Loss (CGL), is similar to feature counting where it observes where the human's gaze fixated and compares it to the algorithm's attention to features. CGL is such that the network "attend[s] to the demonstrator's gaze location" which punishes the algorithm for not attending to the human gaze but does not punish the algorithm if it is fixated on other features as well.

E. Human Gaze Assisted Artificial Intelligence

For decision-making tasks, the human gaze can be closely associated with task reward [1], and a good reward function should ~~be able to~~ explain human gaze behavior. Therefore, it is important to model gaze, reward, and action in a single model. Another important question is whether there can be variability introduced in the gaze data since given the same stimulus humans might pay attention to different visual entities [1]. This is an important factor for this project in which gaze data from different humans is used.

III. Proposed Methods / Algorithm Description

[6] was among the first to show that conditional use of gaze predictions can be used to further improve IL agent performance. We plan to create a similar gating mechanism for usage of motion-based heatmaps as well. By providing both of these mechanisms separately, the agent can learn when to focus

on gaze data versus when to focus on motion data. As done by [4] and [6], the gaze-prediction network will be trained separately via KL-divergence loss relative to the ground truth gaze position data. The embedding from within this network will then be sent to two separate gating units responsible for passing through or zeroing out the gaze or motion heatmaps respectively. The gated outputs will then be convolved and again concatenated with the embedding from the gaze network before being sent into final FC layers with softmax for action prediction.

CGL will then be applied to the embedding relative to both the predicted gaze and motion heatmaps, which [7] reported as a potential future improvement to their work.

IV. Data

We plan on using the Atari-HEAD [2] for data and will need minimal resources for data collection. Atari-HEAD is a large-scale data set that consists of human actions while simultaneously recording eye movements for the Atari benchmark. Of the 57 games in the Arcade Learning Environment (ALE) [8], Atari-HEAD consists of data collected from over 117 total hours of gameplay across 20 games and 8 million human demonstrations. The data was gathered by having human subjects play the ALE frame-by-frame to avoid state-action mismatch. At each frame, the game would pause and wait for the human to input an action. During this pause between actions, the human's gaze, and time between actions were tracked along with the state the human was viewing, the action, and the reward received.

The ALE has been well studied and uses a wealth of data cleansing augmentations for training an agent. Ideally, we would use typical methods (e.g. re-scaling images, cropping, conversion to grayscale, etc.) in the evaluation of our performance. We will use open-source software such as PyTorch or TensorFlow to aid in data cleansing and augmentation.

V. Timeline

Since we identified Atari-HEAD as our source of data, we will not need to spend time attaining data or acquiring IRB approval. Instead, our team will be using the time reserved for data collection to research the Atari-HEAD data and clean and augment the data. For instance, some of the data does not have gaze positions which could be a consequence of the human subject blinking while taking an action — such instances will need to be excluded from training data.

The initial steps will be to design and train an autoencoder for predicting the gaze heatmaps. The convolutional layers/embedding can then be forwarded to the action prediction network which will always have the gaze/motion heatmaps concatenated. The CGL [7] can be added at this stage to drive attention toward the gaze/motion heatmaps.

Our midterm goal would be to have a trained agent which always makes use of the given information (predicted gaze/motion) regardless of any gating function.

The next milestone would be to design/implement the gating networks for both gaze and motion heatmaps. Once the network has been finalized and validated, we will perform a meta-analysis on which games favored gaze-based attention versus motion-based attention.

Our final goal will be data showing better/worse performance of our combined attention mechanism approach as well as an analysis of which games seem to favor gaze- or motion-driven attention.

VI. Risk and Risk Mitigation

The team will need to spend time parsing the data and identifying any data that would need to be excluded from training. The time needed to parse through the data for all 20 games is well beyond the time allowed for this project. For our work, we will focus our development/tuning on a few games but will apply our algorithm to multiple games in the Atari-HEAD data to evaluate how our work affects training on other games. The team will identify the baseline game by a multitude of factors such as the amount of data, the human's in-game performance/score, etc.

VII. Expected Results and Impact

We hypothesize that a network that can gate either gaze or motion heatmaps will perform better than networks that do not incorporate either or just one of them. In video games, the gaze is often vital in understanding human intent, so providing an agent with the ability to learn where to focus, its performance should be improved, especially under the assumption that prompting focus on regions with motion is smart. We expect that incorporating CGL will help our agent not only be able to selectively consider gaze/motion information but also prioritize them when it does so.

Having an agent that can prioritize gaze, motion, or both will allow us to analyze different Atari games and determine which games require which type of information more prominently.

References

[1] R. Zhang, A. Saran, B. Liu, Y. Zhu, S. Guo, S. Niekum, D. Ballard, and M. Hayhoe, "Human gaze assisted artificial intelligence: A review," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (C. Bessiere, ed.), pp. 4951–4958, International Joint Conferences on Artificial Intelligence Organization, 7 2020. Survey track.

[2] R. Zhang, C. Walshe, Z. Liu, L. Guan, K. S. Muller, J. A. Whritner, L. Zhang, M. M. Hayhoe, and D. H. Ballard, "Atari-head: Atari human eye-tracking and demonstration dataset," 2019.

[3] R. Zhang, C. Walshe, Z. Liu, L. Guan, K. S. Muller, J. A. Whritner, L. Zhang, M. Hayhoe, and D. Ballard, "Atari-HEAD: Atari Human Eye-Tracking and Demonstration Dataset," Sept. 2019.

[4] R. Zhang, Z. Liu, L. Zhang, J. A. Whritner, K. S. Muller, M. M. Hayhoe, and D. H. Ballard, "Agil: Learning attention from human for visuomotor tasks," 2018.

[5] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proceedings of the 13th Scandinavian Conference on Image Analysis*, SCIA'03, (Berlin, Heidelberg), p. 363–370, Springer-Verlag, 2003.

[6] C. Thammineni, H. Manjunatha, and E. T. Esfahani, "Selective eye-gaze augmentation to enhance imitation learning in atari games," 2020.

[7] A. Saran, R. Zhang, E. S. Short, and S. Niekum, "Efficiently guiding imitation learning agents with human gaze," 2020.

[8] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," *CoRR*, vol. abs/1207.4708, 2012.