Efficiently Guiding Imitation Learning Agents with Human Gaze

Akanksha Saran  
Department of Computer Science  
University of Texas at Austin  
asaran@cs.utexas.edu

Elaine S. Short  
Department of Computer Science  
Tufts University  
elaine.short@tufts.edu

Ruohan Zhang  
Department of Computer Science  
University of Texas at Austin  
zharu@utexas.edu

Scott Niekum  
Department of Computer Science  
University of Texas at Austin  
sniekum@cs.utexas.edu

ABSTRACT

Human gaze is known to be an intention-revealing signal in human demonstrations of tasks. In this work, we use gaze cues from human demonstrators to enhance the performance of agents trained via three popular imitation learning methods — behavioral cloning (BC), behavioral cloning from observation (BCO), and Trajectory-ranked Reward EXtrapolation (T-REX). Based on similarities between the attention of reinforcement learning agents and human gaze, we propose a novel approach for utilizing gaze data in a computationally efficient manner, as part of an auxiliary loss function, which guides a network to have higher activations in image regions where the human’s gaze fixated. This work is a step towards augmenting any existing convolutional imitation learning agent’s training with auxiliary gaze data. Our auxiliary coverage-based gaze loss (CGL) guides learning toward a better reward function or policy, without adding any additional learnable parameters and without requiring gaze data at test time. We find that our proposed approach improves the performance by 95% for BC, 343% for BCO, and 390% for TREX, averaged over 20 different Atari games. We also find that compared to a prior state-of-the-art imitation learning method assisted by human gaze (AGIL), our method achieves better performance, and is more efficient in terms of learning with fewer demonstrations. We further interpret trained CGL agents with a saliency map visualization method to explain their performance. At last, we show that CGL can help alleviate a well-known causal confusion problem in imitation learning.

KEYWORDS

Human Gaze; Imitation Learning; Learning from Demonstration

1 INTRODUCTION

Learning agents can outperform humans at tasks such as Atari game playing when provided with well-defined goals or rewards using reinforcement learning (RL) [25, 36, 39]. However, designing reward functions by hand can be difficult for complex tasks, even for experts. Imitation learning (IL) [1, 34] is an alternative methodology which overcomes this difficulty by inferring an optimal policy from demonstrations. A challenge in training and utilizing IL agents in the real world is learning from few demonstrations to minimize the burden on end-users, while also sufficiently resolving ambiguity in user intentions and avoiding overfitting. Gaze, an additional informative modality from the demonstrator, can help extract more information out of the same number of demonstrations [50], in addition to information from state-action pairs.

Human attention in the form of eye gaze has been known to encode top-down attention versus bottom-up salience when performing goal-directed tasks [15, 23, 30, 40]. Gaze has been shown to improve the performance of imitation learning algorithms [52], particularly for autonomous driving [8, 46] and Atari game playing [50]. However, most prior approaches utilizing gaze for IL algorithms either use gaze heat maps as input to the agent’s learning model in addition to the world state [24, 50], or predict gaze heatmaps in conjunction with learning the policy [46]. By contrast, we propose using an auxiliary gaze loss during training of imitation learning algorithms to improve the performance of existing methods without increasing model complexity, data requirements, or requiring test-time gaze.

Our methodology utilizes a demonstrator’s gaze fixations on the image as part of a surrogate loss function (coverage-based gaze loss or CGL) during the training phase. Encoding priors in loss functions for label-free supervision of neural networks has been suggested by Stewart et al. [38]. Similarly, we use an auxiliary gaze loss to guide the learning of any agent using image-based state representations and convolutional layers as part of its model architecture. Inspired by our experimental results highlighting the similarity of RL agent attention and human attention (Sec. 3.1), we propose a coverage-based gaze loss (CGL). CGL guides a network to attend to the demonstrator’s gaze locations and helps improve the performance of three IL methods on 20 Atari games. A critical advantage of our approach in contrast to several prior approaches utilizing gaze, is that gaze is not required at test time and instead used as a weak supervisory signal.

We evaluate our auxiliary gaze loss function on 20 Atari games with three different IL approaches. Our experiments show that CGL can improve performance for both inverse reinforcement learning (IRL) and behavioral cloning (BC) frameworks averaged across 20 games: 95% for behavioral cloning (BC) [50], 343% for behavioral cloning from observation (BCO) [42], and 390% for T-REX [6], compared to not using any gaze information at all. Moreover, we show...
2 RELATED WORK

2.1 Imitation Learning for Atari Games

When learning from demonstrations, Atari game playing has been attempted with various imitation learning approaches. Behavioral cloning [3, 9, 29] is a class of imitation learning methods, where an agent learns a policy by using the demonstrated states and actions from the expert as input data for supervised learning. Behavioral cloning from observation (BCO) [42] is a two-phase, iterative imitation learning technique — first allowing the agent to acquire self-supervised experience in a task-independent pre-demonstration phase, which is then used to learn a model for a specific task policy only from state observations of expert demonstrations (without access to actions). The self-supervision produces an inverse dynamics model to infer actions, given state observations. This model is then used to infer expert actions from state-only demonstrations. The inferred actions along with state information are then used to perform imitation learning for the agent’s policy. GAIL [19] is an adversarial imitation learning approach trained by alternating the learning updates between a generator policy network and a discriminator network distinguishing between the demonstrated and generated trajectories. It achieved state-of-the-art performance for low-dimensional domains. BCO shows comparable performance to GAIL on low-dimensional MuJoCo benchmarks [41] with increased learning speed.

Behavioral cloning does not explicitly model the goals or intentions of the demonstrations which a succinct reward function attempts to capture in inverse reinforcement learning (IRL), another class of imitation learning approaches. Typically, such a succinct inferred reward function makes IRL have better generalization properties compared to behavioral cloning [29]. Most deep learning-based IRL methods either require access to demonstrated actions [20] or do not scale to high-dimensional tasks such as video games [12, 13, 28]. Tucker et al. [43] showed that their adversarial IRL method is difficult to train and fails at high-dimensional tasks of Atari game playing, even with extensive parameter tuning. Aytar et al. [2] learn a reward function from observations for three Atari games. They guide the agent to exactly imitate the checkpoints from provided demonstrations, assuming access to high-quality demonstrations. A recent IRL method called T-REX [6] is a reward learning from observation algorithm, that extrapolates beyond a set of ranked and potentially suboptimal demonstrations. T-REX outperforms other imitation learning methods such as BCO and GAIL, on Atari and MuJuCo benchmarks [41] and also demonstrates the ability to extrapolate intentions of a suboptimal demonstrator. However, a performance gap exists between reinforcement and imitation learning methods for Atari game playing. In our work, we propose to reduce this gap by incorporating an additional information modality in the form of human gaze for imitation learning.

2.2 Utilizing Gaze for Learning

Prior studies have shown that human fovea moves to the correct place at the right time to extract task-relevant information, making visual attention a feature selection mechanism for humans [15, 30]. Human gaze information can be used in many ways to help AI agents learn a variety of tasks [51]. Novice human learners can benefit from observing experts’ gaze [44] for learning complex surgical skills. Yamani et al. [47] showed that viewing expert gaze videos can improve the hazard anticipation ability of novice drivers. Saran et al. [31, 32] showed the advantage of incorporating a human demonstrator’s gaze for learning robotics manipulation tasks. Gaze has been exploited in prior imitation learning approaches for autonomous driving [8, 46] and Atari Games [50], but to the best of our knowledge, our work is the first attempt to incorporate gaze in a deep IRL algorithm (TREX).

A common method of incorporating human attention is to simply use the gaze map as an additional image-like input [24] or predict the gaze heatmap and further use high-resolution parts of the image to improve learning [46, 50]. Zhang et al. [50] show improved learning on Atari games for imitation learning (AGIL).
but use predicted heatmaps corresponding to demonstration states as part of the input. In comparison, gaze-modulated dropout (GMD) was proposed by Chen et al. [8] to implicitly incorporate gaze into an IL framework for autonomous driving, instead of using gaze as an additional input. An estimated gaze distribution is used to modulate the dropout probability of units at different spatial locations in the first two convolutional layers. While GMD requires gaze both at train and test time, our auxiliary loss only requires gaze data at train time. We use AGIL and GMD as baseline methods for comparison with CGL in our experiments.

3 APPROACH

To enable existing IL algorithms to take advantage of human gaze signals accompanying demonstrations, we first try to understand the correlation between human gaze and well-performing RL agents for six Atari games (Sec. 3.1). Based on our findings, which show a similarity between human gaze and RL attention, we formalize our auxiliary coverage-based gaze loss term to mimic the attention mechanism of human demonstrators (Sec. 3.2). This loss term can guide any convolutional network to attend towards features that human demonstrators attend to. Our approach does not increase the model complexity of existing algorithms in terms of the number of learnable parameters, and can be easily applied to the training of any neural network with convolutional layers. We then describe other baseline approaches incorporating human gaze information for IL, which we evaluate and compare against our proposed approach (Sec. 3.3).

3.1 Human Gaze coverage of RL Agent Attention

Prior studies in the cognitive science literature have established the concept of selective attention for decision making in primates [26]. While choosing an action among the available set of actions, animals select a subset of information by directing their sensory organs towards specific stimuli (overt attention) and focus on specific parts of the stimulus internally (covert attention) to act upon. Human gaze data only reveals overt attention which is directly connected to a sensory organ. However, humans can still pay covert attention to entities in the working memory [27]. In other words, being attended by the human gaze model is a sufficient (but not necessary) condition for the features to be important. An example of this is shown in Fig. 2 where a high-performing RL agent on the game Breakout attends to more features of the state space in addition to what the human gaze attends to (human gaze predicted by a reliable generative human gaze model [50], RL attention calculated by [14]). Hence for our analysis comparing the attention of RL agents and humans, we define a coverage metric (Equation 1) that penalizes the RL agent only if it fails to pay attention to human attended regions, or equivalently, a metric that is sensitive to false negatives if we treat human attention as the ground truth. KL divergence is an ideal candidate in this case [7]. Let $P$ denote the human attention map and $Q$ denote the RL attention map. The coverage metric is computed as follows:

$$KL(P||Q) = \sum_i \sum_j P(i, j) \log \left( \frac{P(i, j) + \epsilon}{Q(i, j) + \epsilon} \right)$$

Figure 2: A trained PPO agent’s attention map (b) and corresponding human gaze map (c) for the same input image (a) for the Breakout Atari game. The RL attention is directed towards more regions than the human gaze.

where $\epsilon$ is a small regularization constant chosen to be $2.2204E^{-16}$ following convention [7]. A lower value of the KL-divergence based coverage metric would signify a stronger correlation between human gaze and RL attention.

We compute this coverage metric between the RL attention map and the corresponding human gaze map for 100 images per game. The overall metric reported for a game is averaged over all 100 images. The images are uniformly sampled from a policy rollout of the learned RL agent trained via proximal policy optimization (PPO). To examine if there is a correlation between the attention of the agent and human, we compare human attention maps from one game with RL attention maps of the other five games. We hypothesize that the coverage metric would have a lower value for the gaze maps and RL attention maps from the same game versus other games.

We find that the lowest average coverage metric scores are obtained between RL attention heatmaps and gaze heatmaps for the same game (diagonal of Table 1). Comparison of human gaze heatmaps for one game and RL attention heatmaps of all other game agents is equivalent to comparison of gaze heatmaps with randomly generated attention maps (without knowledge of the game state with which gaze heatmap is computed). This one-to-one comparison for each of the 100 time steps along a trajectory implies a correlation between the human gaze and RL attention for the same task. The analysis here informs our auxiliary loss function to guide the attention of imitation learning algorithms to have coverage over human gaze.

3.2 Coverage-based Gaze Loss

Based on the coverage metric (Equation 1), we propose adding an auxiliary loss term in addition to the existing loss function for a network, modifying the training procedure of any IL algorithm. Our loss term will have a higher penalty if the network does not attend to parts of the image that the demonstrator focused on, but will have no penalty for activations where the demonstrator did not pay attention. We refer to the proposed loss function as a coverage-based gaze loss (CGL).

CGL operates on the human gaze heatmap and the output of the last convolutional layer. For consistency in comparison with baselines, gaze heatmaps are generated using convolution-deconvolution networks trained on real human gaze data [50]. Activation feature maps from the last convolutional layer [35] of image classification CNNs are shown to have the best compromise between high-level
<table>
<thead>
<tr>
<th>RL Attention → Human Attention ↓</th>
<th>asterix</th>
<th>breakout</th>
<th>centipede</th>
<th>ms_pacman</th>
<th>phoenix</th>
<th>seaquest</th>
</tr>
</thead>
<tbody>
<tr>
<td>asterix</td>
<td>1.72</td>
<td>6.97</td>
<td>4.64</td>
<td>3.13</td>
<td>14.61</td>
<td>6.34</td>
</tr>
<tr>
<td>breakout</td>
<td>5.26</td>
<td>2.09</td>
<td>4.94</td>
<td>3.80</td>
<td>11.13</td>
<td>5.57</td>
</tr>
<tr>
<td>centipede</td>
<td>4.43</td>
<td>6.40</td>
<td>1.86</td>
<td>3.34</td>
<td>9.86</td>
<td>5.71</td>
</tr>
<tr>
<td>ms_pacman</td>
<td>4.53</td>
<td>6.46</td>
<td>5.49</td>
<td>1.78</td>
<td>13.34</td>
<td>5.92</td>
</tr>
<tr>
<td>phoenix</td>
<td>4.29</td>
<td>10.75</td>
<td>5.19</td>
<td>3.59</td>
<td>3.55</td>
<td>6.49</td>
</tr>
<tr>
<td>seaquest</td>
<td>5.07</td>
<td>7.70</td>
<td>5.97</td>
<td>3.51</td>
<td>14.10</td>
<td>3.03</td>
</tr>
</tbody>
</table>

Table 1: Comparison between PPO network attention and human gaze attention using KL divergence. The values represent the average KL divergence between gaze heat maps and RL attention heat maps for 100 uniformly sampled images from a policy rollout of the PPO agent.
\( \hat{r}(s) \) represents the gaze map corresponding to the state \( s \) from the trajectory snippet \( r_t \) and \( c_t(s) \) represents the collapsed and normalized version of the last convolutional layer’s output for the same state \( s \). The loss function accumulates gaze over the entire trajectory snippet for both trajectories used as input to the network.

We use the default implementation of T-REX from Brown et al. [6]. The reward network has four convolutional layers. The gaze loss is computed over the last convolutional layer output – a spatial map of size \( 16 \times 7 \times 7 \) (normalized, collapsed and upsampled to the size of the gaze heatmaps \( 84 \times 84 \)). At the end, a fully connected layer with 64 hidden units is used to determine the ranking between a pair of demonstrations.

Similar to the implementation of Brown et al. [6], the trajectories are first subsampled by maximizing over every 3rd and 4th frame, from which a stack of 4 consecutive frames with pixel values normalized between 0 and 1 is passed as input to the reward network. The snippets are ranked based on ground truth rewards or cumulative game scores of the trajectories they are sampled from. A PPO agent is then trained using the learned reward to obtain a policy for gameplay.

### 3.3 Other Techniques to Incorporate Gaze

Here we describe two alternative methods incorporating human gaze for imitation learning, which we compare against.

#### 3.3.1 Gaze-modulated Dropout (GMD)

As a baseline for learning from human gaze, we implement GMD [8] for the first two convolutional layers of the BCO policy network Torabi et al. [42]. The BCO policy network does not originally use dropout layers. Gaze maps are generated using a convolution-deconvolution network [50], trained separately for each game on the Atari-HEAD dataset [53]. The gaze prediction network uses as input a stack of 4 consecutive game frames, each of size \( 84 \times 84 \). Details of the network architecture are similar to Zhang et al. [50]. We employ this network for gaze prediction, as it has been shown to work well for the Atari domain, instead of the Pix2Pix network [21] used by Chen et al. [8] for the autonomous driving domain. The generated gaze map is then used as a mask for the additional dropout layer added after first two the convolutional layers as described by Chen et al. [8]. Units of the convolutional layer near the estimated gaze location are assigned a lower dropout probability than units far from the estimated gaze location. This is similar to conventional dropout [37], but with non-uniform dropout probability for spatial units corresponding to different parts of the image space.

#### 3.3.2 Attention Guided Imitation Learning (AGIL)

AGIL adds more parameters to a BC network to utilize gaze. The output of the gaze prediction network is used as input to an additional convolutional pathway in a modified version of standard behavioral cloning. AGIL consists of two channels of 3 convolutional layers. One channel takes as input a single image frame (game state) and another uses a masked image which is an element-wise product of the original image and predicted gaze saliency map. Finally, the outputs of the two channels are averaged to predict one of the 18 actions within ALE [4]. We use the same hyperparameters provided by Zhang et al. [50] for the implementation of AGIL.

### 4 EXPERIMENTS AND RESULTS

We use demonstrations from 20 games in ALE [4] with varying dynamics and features. Demonstrations and corresponding human gaze data are from the publicly available Atari-HEAD dataset [53]. We augmented three imitation learning algorithms with CGL – BC, BCO, and T-REX. These algorithms were implemented in the OpenAI Gym platform [5], which contains Atari 2600 video games with high-dimensional observation space (raw pixels). All reported results were game scores averaged over 30 different rollouts (episodes) of the learned policy, similar to the procedure followed by Zhang et al. [50]. We used the default settings from OpenAI baselines [11] for parameters of ALE [4]. All experiments are conducted on server clusters with NVIDIA 1080, 1080Ti, Titan V, or DGX GPUs.

For evaluation, we intend to show improvement in terms of game scores using CGL. We calculate the improvement factor over baseline in the following way: improvement = (new score - baseline score) / baseline score. If both the baseline score and the new score are zero, improvement is zero. However, for some games baseline game scores are zeroes but new scores are non-zero. In such cases, the improvement will not be calculated. We report average improvement (including games in which improvements are negative) across 20 games. Details on the experiments and individual games scores can be found in the supplementary material [33]. Note that the improvement factors are underestimated, due to the way we handle zero score games.

#### 4.1 CGL Improves BC

BC+CGL outperforms BC on 19 out of 20 games with 15 minutes of demonstration data (Fig. 3). On average, the improvement is 95.1%. With all 300 minutes of human gameplay data, BC+CGL outperforms BC on all 20 games with an average improvement of 86.2%. The hyperparameter \( \alpha \) is tuned using a grid search from a set of 7 values — 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9. Batch size \( N = 50 \) and Adadelta optimizer [49] is used for all models based on BC.

#### 4.1.1 Efficiency of CGL in terms of Learnable Parameters

Previous methods (such as AGIL) incorporate human attention by introducing extra parameters to the model due to additional neural network modules added. To tease apart whether improvement in these approaches comes from increased parameters to standard behavioral cloning or from the gaze information itself, we perform the following experiment. We re-train the AGIL network, but instead of using gaze heatmaps, we pass the original image as input to the gaze pathway, referred to as the BC-2ch model. This helps us disambiguate if more parameters in the model help more versus the gaze information itself. As shown in Fig. 3, we find that the BC-2ch model does result in improved performance over BC, indicating that part of AGIL’s improvement over BC is due to additional parameters. This hints at the fact that increasing model complexity alone without using any additional information as input proves beneficial.

#### 4.1.2 CGL Provides Stronger Guidance than AGIL

In Fig. 3, we also show that on average, CGL outperforms the previously best method to incorporate gaze (AGIL [50]) by a large margin. This study suggests that CGL performs significantly better than AGIL with fewer model parameters, and the advantage is even more evident with a limited amount of human demonstration data (95.1% improvement...
over BC versus 10.1%). The sample efficiency of CGL is critical as it can be beneficial for applying this method to challenging imitation learning problems, where collecting demonstrations is cumbersome, expensive and time-critical.

4.2 CGL Improves BCO

BCO+CGL outperforms basic BCO on 14 out of 20 games with 15 minutes of human demonstration data (Fig. 4). On average, the improvement is 160.9%. The hyperparameter $\alpha$ is tuned using a grid search from a set of 9 values — 0.001, 0.005, 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9. The Adam optimizer\cite{Kingma2014AdamAM} is used to solve for the network parameters with a batch size $N = 32$. There are six games for which BCO achieves scores of zero but BCO+CGL can achieve non-zero scores. With all 300 minutes of human demonstration data, BCO+CGL outperforms BCO on 12 out of 20 games with an average improvement of 343.6%. The largest improvement comes from the game Centipede, which increases the average improvement by a large margin. However, we found that BCO is unable to learn an accurate inverse dynamics model for up to seven of the 20 games. For these games, the baseline BCO policy model scores zero and the utilization of gaze is also unable to overcome the shortcomings of the inverse dynamics model.

4.2.1 CGL Provides Stronger Guidance than GMD. We test GMD and CGL with BCO and find that on average across 20 games, CGL outperforms GMD both with 15 minutes and 300 minutes of demonstration data (Fig. 4). One may expect that convolutional dropout helps generalization by reducing over-fitting. However, it is far less advantageous, since the shared-filter and local-connectivity architecture in convolutional layers is a drastic reduction in the number of parameters and this already reduces the possibility to overfit\cite{He2015DeepRL}. Empirical results by Wu et al.\cite{Wu2017ConvolutionalDR} confirm that the improvement in generalization to test data from convolutional dropout is often inferior to max-pooling or fully-connected dropout.

4.2.2 CGL Provides Stronger Guidance than Implicit Motion Information. Prior work has established that human gaze encodes attention which is different from salient regions in a scene (such as motion)\cite{Krause2006VisualML,Yao2013VisualML,Seitz2012VisualML}. We test whether our proposed approach extracts the additional information from human gaze data, in comparison to what might already be encoded in the visual game state, such as motion. We replace the gaze heat maps used by CGL with heatmaps representing the normalized motion in an input image frame stack (the difference between the last and first frame in a stack shown in Fig. 5). We test this motion-based loss for BCO (Fig. 4). Overall, the gains from CGL are higher than those from motion, more so with 300 minutes of demonstration data. Our results are in line with prior work\cite{Pastalkovski2019ImitationVL} which shows that using optical flow between two frames as the attention map provides only moderate performance improvements in Atari Games. Incorporating both gaze and motion information simultaneously in an auxiliary loss can be investigated as part of future work.

4.3 CGL Improves T-REX

T-REX is an inverse reinforcement learning algorithm which compares pairs of trajectory snippet to learn the reward. Along with full 300 minutes of demonstration data, we evaluate with 30 minutes of demonstration data to compare trajectories from two different demonstrations (each demonstration in Atari-HEAD\cite{Perez2018AtariHEAD} was at least 15 minutes long). The hyperparameter $\alpha$ is tuned using a grid search from a set of 9 values — 0.001, 0.005, 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9 and the Adam optimizer\cite{Kingma2014AdamAM} is used. T-REX+CGL outperforms basic T-REX on 15 out of 20 games both with 30 minutes and 300 minutes worth of training data\cite{Lai2017T-REX}. On average, the improvement due to CGL is 390.4% with 30-minutes of data, and 373.6% with 300-minutes of data (Table 2). There are four games that T-REX achieves scores of zero but T-REX+CGL can achieve non-zero scores. To the best of our knowledge, CGL is the first method to augment the learning of an IRL algorithm with human gaze.

4.4 Best Performing Models for each Game

We then summarize the best game scores obtained from various algorithms presented above. The results are shown in Table 3. We notice that CGL augmented methods achieve the best results in 15 out of 20 games. For comparison, we also show DQN scores\cite{Mnih2015Human-levelCO, Mnih2019Human-levelCO}
Figure 5: Motion in the visual game state, i.e. the difference between the last and first frame in an input image stack, cannot alone explain human attention. This is further highlighted with minimal performance gains when CGL utilizes the motion heatmap instead of the human gaze heatmap.

Table 2: Average (across 20 games) improvement over the T-REX baseline. Result is presented as mean±standard error of the mean (N=20).

<table>
<thead>
<tr>
<th>Improvement over T-REX (%)</th>
<th>T-REX+CGL</th>
</tr>
</thead>
<tbody>
<tr>
<td>30min data</td>
<td>390.4 ± 203.5</td>
</tr>
<tr>
<td>300min data</td>
<td>373.6 ± 206.5</td>
</tr>
</tbody>
</table>

as a reference (the evaluation methods are slightly different). With human gaze information (especially with CGL), imitation learning algorithms start to match and even outperform DQN. Note that DQN is trained with 200M samples per game, while IL methods are at most trained with 360K samples (300-minutes of human data).

4.5 Visualizing CGL Attention

We can analyze whether the CGL agents have successfully learned to pay attention to the critical regions highlighted by human saliency maps in two ways. First, we directly visualize the activation map of the networks trained with and without CGL, which has already been shown in Fig. 1 for a trained BCO agent. However, this only shows that the convolutional layer we applied CGL to behaves as expected.

We use a second method to show that the whole trained network has learned to attend to the desired region with CGL. We visualize the attention maps of trained CGL agents with a method commonly used to provide visual interpretations of deep RL agents [14]. The algorithm takes an input image \( I \) and applies a Gaussian filter to a pixel location \((i, j)\) to blur the image. This manipulation adds spatial uncertainty to the surrounding region and produces a perturbed image \( \Phi(i, j) \). A saliency score for this pixel \((i, j)\) can be defined as how much the blurred image changes the network output [14]. Doing this for every pixel results in a saliency map that approximates the “attention” of a network. The results for a case of Breakout where CGL outperforms the baseline T-REX method can be found in Fig. 6, where the T-REX+CGL agent successfully learned to focus on the ball like the human did, while the T-REX agent did not. We also highlight a failure case with MsPacman, where the CGL agent does not outperform baseline T-REX. We find that the network fails to attend to both modes of attention in the human’s gaze map.

Table 3: A summary of the best game scores obtained. DQN scores are from no-op starts evaluation regime table of [16], except for game Riverraid [25]. With human gaze information (especially with CGL), imitation learning algorithms start to match and even outperform DQN.

<table>
<thead>
<tr>
<th>Game</th>
<th>Algorithm (#demo)</th>
<th>Score</th>
<th>DQN Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>alien</td>
<td>AGIL (300min)</td>
<td>2104.7</td>
<td>1620.0</td>
</tr>
<tr>
<td>asterix</td>
<td>T-REX+CGL (30min)</td>
<td>66445.0</td>
<td>4359.0</td>
</tr>
<tr>
<td>bank_heist</td>
<td>BC-2ch (300min)</td>
<td>174.3</td>
<td>455.0</td>
</tr>
<tr>
<td>berzerk</td>
<td>BCO+CGL (15min)</td>
<td>687.67</td>
<td>585.6</td>
</tr>
<tr>
<td>breakout</td>
<td>T-REX+CGL (30min)</td>
<td>438.4</td>
<td>385.5</td>
</tr>
<tr>
<td>centipede</td>
<td>T-REX+CGL (30min)</td>
<td>20762.5</td>
<td>4657.7</td>
</tr>
<tr>
<td>demon_attack</td>
<td>T-REX+CGL (30min)</td>
<td>17589.0</td>
<td>12149.4</td>
</tr>
<tr>
<td>enduro</td>
<td>BC+CGL (30min)</td>
<td>445.1</td>
<td>729.0</td>
</tr>
<tr>
<td>freeway</td>
<td>BC-2ch (30min)</td>
<td>31.4</td>
<td>30.8</td>
</tr>
<tr>
<td>frostbite</td>
<td>BC+CGL (30min)</td>
<td>5897.7</td>
<td>797.4</td>
</tr>
<tr>
<td>hero</td>
<td>BC+CGL (15min)</td>
<td>19023.2</td>
<td>20437.8</td>
</tr>
<tr>
<td>montezuma</td>
<td>BC+CGL (30min)</td>
<td>1720.0</td>
<td>0.0</td>
</tr>
<tr>
<td>ms_pacman</td>
<td>BC+CGL (30min)</td>
<td>2739.7</td>
<td>3085.6</td>
</tr>
<tr>
<td>name_this_game</td>
<td>AGIL (300min)</td>
<td>5817.0</td>
<td>8207.8</td>
</tr>
<tr>
<td>phoenix</td>
<td>AGIL (300min)</td>
<td>5140.0</td>
<td>8485.2</td>
</tr>
<tr>
<td>riverraid</td>
<td>T-REX+CGL (30min)</td>
<td>7370.0</td>
<td>8316.0</td>
</tr>
<tr>
<td>road_runner</td>
<td>BC+CGL (30min)</td>
<td>33510.0</td>
<td>39544.0</td>
</tr>
<tr>
<td>sequeast</td>
<td>T-REX+CGL (30min)</td>
<td>759.3</td>
<td>5860.6</td>
</tr>
<tr>
<td>space_invaders</td>
<td>T-REX+CGL (30min)</td>
<td>1563.7</td>
<td>1692.3</td>
</tr>
<tr>
<td>venture</td>
<td>BC+CGL (15min)</td>
<td>376.7</td>
<td>165.0</td>
</tr>
</tbody>
</table>

1115
4.6 Reducing Causal Confusion with CGL

Discriminative models for IL such as BC are non-causal, i.e. the training procedure is unaware of the causal structure of the interaction between the demonstrator and the environment. Causal misidentification is the phenomenon when cloned policies fail by misidentifying the causes of the demonstrator’s actions. A very problematic effect of distributional shift in BC can lead to causal misidentification. This is exacerbated by the causal structure of sequential action: the very fact that current actions cause future observations often introduces complex new nuisance correlates.

Prior work on understanding causal confusion in IL [10] uses past action information (often correlated with current action) to identify if the IL algorithm is in a causal confusion trap. To understand the performance gains of CGL, we investigate if it disambiguates the intent of the user in the demonstrated actions by eliminating causal confusion. We overlay the four frame image stack (state) with actions from the last frame in the previous stack (Fig. 7). This lays a causal confusion trap for the IL agent. If the agent can ignore the new correlated action information that is part of the state space, it hints towards the fact that the agent learns to ignore those features and perform better empirically. We find that on average (across 20 games), CGL agents suffer less when trained with confounded data compared to the BC baseline (-34.0% versus -47.8%), and still perform better than BC agents. Detailed results are available in [33]. This hints to the fact that in addition to directing the attention of the network to learn a better mapping between states and actions, part of the gains from using human gaze data with CGL can come from reducing causal confusion.

5 CONCLUSIONS

In this work, we introduced an auxiliary coverage-based gaze loss (CGL) term which guides the training of any imitation learning network with convolutional layers. Our experiments showed improved performance on several Atari games over standard imitation learning algorithms. Our approach provides these gains without requiring gaze prediction at test time or increasing the model complexity of existing algorithms. We outperform a baseline method (GMD), which also does not increase model complexity. CGL is more efficient in terms of both learnable parameters and data efficiency when compared to a state-of-the-art gaze-augmented IL method (AGIL) which utilizes gaze in the form of additional input to a BC algorithm. AGIL requires gaze prediction at test time and is shown to gain performance by increasing model complexity alone. Our approach improved performance by utilizing gaze without these shortcomings. We also highlighted that utilizing human gaze provides additional information to what is encoded implicitly in the game state (such as motion). Our work confirms prior research showing gaze can help extract more information from a demonstrator than traditional state-action pairs, bridging the gap in performance between IL and RL agents.

6 DISCUSSION AND FUTURE WORK

Human attention can be seen as a form of spatial prior on the visual input. In deep learning research, this prior is often used as a mask to filter out unimportant information (e.g., AGIL). This approach has two main drawbacks. First, it requires the mask at testing time. Secondly, some unattended visual features handled by human memory systems could still be useful for decisions. Therefore, completely filtering out all unattended information seems inappropriate. In this work, we present a novel method to incorporate gaze effectively that only requires access to human gaze data at training time. Moreover, by utilizing a coverage-based loss, this method highlights the attended features while keeping the unattended features available for the learning agent. This novel method in training deep neural networks can be applied to other learning tasks that utilize other forms of spatial priors.

Additionally, human gaze and actions from demonstrations may be correlated in time. Our approach only utilizes gaze per game state, and so do all other approaches we compare against. Utilizing temporal connections in the gaze signal is a direction for future work. We also hope our work encourages the research community to innovate on other novel ideas for efficiently incorporating human attention.

ACKNOWLEDGMENTS

This work has taken place in the Personal Autonomous Robotics Lab (PeARL) at The University of Texas at Austin. PeARL research is supported in part by the NSF (IIS-1724157, IIS-1638107, IIS-1749204, IIS-1925082) and ONR (N00014-18-2243). This work is also partially supported by NIH (EY05729).