Neural Computing and Applications (2023) 35:23401–23410
https://doi.org/10.1007/s00521-021-06367-y

S.I.: HUMAN-IN-THE-LOOP MACHINE LEARNING AND ITS APPLICATIONS

Selective eye-gaze augmentation to enhance imitation learning in Atari games

Chaitanya Thammineni[1] · Hemanth Manjunatha[1] · Ehsan T. Esfahani[1]

Received: 2 February 2021 / Accepted: 26 July 2021 / Published online: 13 August 2021

Abstract
This paper presents the selective use of eye-gaze information in learning human actions in Atari games. Extensive evidence suggests that our eye movements convey a wealth of information about the direction of our attention and mental states and encode the information necessary to complete a task. Based on this evidence, we hypothesize that selective use of eye-gaze, as a clue for attention direction, will enhance the learning from demonstration. For this purpose, we propose a selective eye-gaze augmentation (SEA) network that learns when to use the eye-gaze information. The proposed network architecture consists of three sub-networks: gaze prediction, gating, and action prediction network. Using the prior 4 game frames, a gaze map is predicted by the gaze prediction network, which is used for augmenting the input frame. The gating network will determine whether the predicted gaze map should be used in learning and is fed to the final network to predict the action at the current frame. To validate this approach, we use publicly available Atari Human Eye-Tracking And Demonstration (Atari-HEAD) dataset consists of 20 Atari games with 28 million human demonstrations and 328 million eye-gazes (over game frames) collected from four subjects. We demonstrate the efficacy of selective eye-gaze augmentation compared to the state-of-the-art Attention Guided Imitation Learning (AGIL) and Behavior Cloning (BC). The results indicate that the selective augmentation approach (the SEA network) performs significantly better than the AGIL and BC. Moreover, to demonstrate the significance of selective use of gaze through the gating network, we compare our approach with the random selection of the gaze. Even in this case, the SEA network performs significantly better, validating the advantage of selectively using the gaze in demonstration learning.

Keywords Imitation learning · Human-in-the-loop learning · Learning by demonstration

1 Introduction

The most common form of human augmentation (guidance) in the learning frameworks is to learn *policy* directly from human actions. In comparison with *reinforcement learning*, the *imitation learning (IL)* framework has shown significant advantages as they do not required handcrafted reward functions [1]. By learning directly from human actions, IL can reduce the huge cost of learning from scratch [2, 3]. Moreover, by utilizing *human in the loop learning* in IL, human attention can be used to reduce the state and action space's size to guide the IL. For instance, in visual learning tasks, the gaze position indicates a human's immediate attention to process urgent state information. Several research groups have successfully utilized the eye-gaze maps to guide the learning process [4–7]. In these works, the predicted gaze heat-map is used to select the critical features in a given state resulting in higher accuracy in imitating human actions [7, 8]. Incorporating the human attention model into behavioral cloning has shown to improve the Atari game's performance by 115% [7]. Nonetheless, the effective use of eye-gaze data remains unexplored. In this work, we investigate the selective use of eye-gaze information to enhance the behavior cloning in the Atari platform.

✉ Ehsan T. Esfahani
ehsanesf@buffalo.edu

Chaitanya Thammineni
sthammin@buffalo.edu

Hemanth Manjunatha
hemanthm@buffalo.edu

[1] Human In the Loop System Laboratory, University at Buffalo, Buffalo, NY 14260, USA

The rationale for the use of eye-gaze is based on decades of evidence that strongly suggests that attention facilitates action selection [9–13]. However, not all eye movements provide information on attention and action. For instance, during eye-saccades, no visual information is gathered, and the perceptual attention can occur with eye-saccades. Similarly, fixation on an object does not guarantee that attention is directed [14]. Consequently, over the course of a task, the eye movements should be selectively processed to understand the relevant instances to perform an action. To this end, our main idea is to augment the input frame with the human-gaze only when required and, in any other cases, use an unaugmented input frame.

In nature, an organism usually selects a subset of information by directing their sensory organs toward specific stimuli (over attention) and internally focusing on the particular part of these specific stimuli (covert attention) [13] to act upon or to select an action among the available set of actions [13, 15, 16]. Such behavior integrates deeply with our daily activities, such as reaching tasks [17], sports, and driving [18]. Moreover, many studies have shown that eye movement (providing attention) and motor demonstrations are an inter-weaved phenomenon where the visual system extracts the necessary information to complete a task [17, 19, 20]. Arguing in the same lines, we hypothesize that selective augmentation of eye-gaze (thus selective attention) information should provide vital information about the action itself and should enhance the performance of the action imitation learning. For this purpose, we propose a neural network architecture that learns when to use the eye-gaze information selectively.

The network architecture (Fig. 1) for augmenting the eye-gaze data has three sub-networks: gaze prediction network, gating network, and action prediction network. The *gaze prediction network* takes four frames of the game as an input and predicts the eye-gaze distribution (gaze map) over the last frame in the sequence. This predicted gaze map is used for augmenting the input frame. The *gating network* is used to specify whether the predicted gaze map is used in learning or not. We can achieve this by multiplying the binary output (0 or 1) from the gating network with the predicted gaze map. The *action prediction network* uses information from two channels to predict the current frame's action. The two channels are embedded input frame information and a gated gaze-map. If the gate output is 0, only the input frame information is used for action prediction. There are two loss functions corresponding to gaze prediction and action prediction, and the gaze-prediction loss is *independent* of action prediction loss. Thus, we can decouple the training process and separately train the gaze-prediction network. Therefore, we can use a pre-trained gaze-prediction network while training the action prediction network. The effectiveness of our model is tested on the Arcade Learning Environment (ALE) over 6 different Atari games. Atari games served by ALE has become a widely utilized benchmark [21] for evaluating the development of general, domain-independent AI technology providing an opportunity for fair comparison to the state-of-the-art algorithms.

Our main contribution is a selective eye-gaze augmentation (SEA) network that automatically learns when to use the eye-gaze information for better action prediction. We demonstrate the efficacy of SEA on the publicly available Atari-HEAD dataset, which consists of eye movements of the subjects during the gameplay. The proposed framework is shown to outperform the state-of-the-art Attention Guided Imitation Learning (AGIL) on the same dataset, which

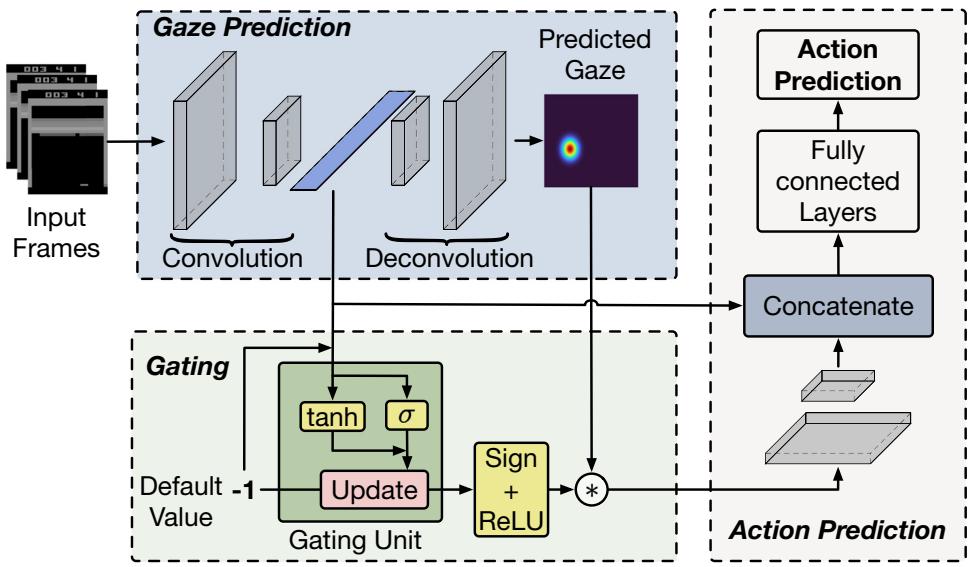

Fig. 1 Architecture of selective eye-gaze augmentation (SEA) network. The network has three modules: (a) Gaze prediction network, which predicts eye gaze given four game frames. (b) Gating network that learns when to use the eye-gaze. (c) Action prediction network which learns the action mapping from the game frame embeddings and eye-gaze information

uses eye-gaze information to learn the human actions in Atari.

2 Related work

Recently, the use of eye-gaze in guiding imitation learning is gaining traction. For example, [22] showed that by utilizing the human attention from eye-tracking as an inductive bias in recurrent neural network (RNN), the performance could be dramatically increased. The study showed that RNN regularised by human attention improved sentiment analysis, grammatical error detection, and abusive language detection. Penkov et al. [23] used eye-tracking to learn a mapping between abstract plan symbols and their physical instances. The study showed that the eye-gaze guided system successfully learns the grounding of abstract plan symbols. Eye-gazes have also been successfully used in driving [24]. For instance, Yuying Chen et al. [25] presented a gaze modulated drop-out method in deep driving networks for application in driving. The study showed that the gaze modulated drop method reduced the steering prediction error by 23.5%. On similar lines, in navigation, Yuying Chen et al. [26] used the graph convolutional networks with attention learned from the human gaze to navigate a robot through a crowd successfully. The study showed that the eye-gaze guided model performed significantly better than the state-of-the-art methods.

The eye-gaze augmentation has also found success in-game platforms like Atari. Zhang et al. [4] introduced a large-scale dataset of human actions in Atari video games with simultaneously recorded eye movements. The study showed that using a learned human gaze model to inform imitation learning resulted in a 115% increase in in-game performance. The above research works provide a wide range of applications demonstrating the efficacy of augmenting human gaze information into imitation learning. On similar lines, Akanksha Saran et al. [6] used gaze cues from human demonstrators to enhance the performance of state-of-the-art inverse reinforcement learning and behavior cloning algorithms without adding any additional learnable parameters to those models. They showed that augmenting existing convolutional architecture with gaze information guided the learning agent toward better reward function and policy. Ruohan Zhang et al. [4, 7] proposed the Attention Guided Imitation Learning (AGIL) framework, in which a learning agent first learns a visual attention model from human gaze data, then learns how to perform the visuomotor task from human decisions. The framework demonstrated the effectiveness of end-to-end learning of visuomotor tasks guided by attention.

3 Selective eye-gaze augmentation network

In this section, we briefly present a description of the dataset used for the study and, subsequently, provide details on the architecture of the three sub-networks of the selective eye-gaze augmentation (SEA) network: gaze prediction network, gating network, and action prediction network.

3.1 Dataset description

To study the efficacy of selective eye-gaze augmentation, we have used a large-scale Atari-HEAD dataset [4], which is collected from four subjects playing 20 different Atari games with varying difficulty levels and game dynamics. During the gameplay subject's, eye movements are recorded using EyeLink 1000 eye tracker at 1000 Hz. The game screen was 64.6 × 40.0 cm (1280 × 840 in pixels), and the average distance between the subject and the screen was 78.7 cm. The subjects were novices who were familiar with the game environment. The dataset contains 117 hours of gameplay, around 28 million human actions, and 328 million eye-gazes. More information on the game statistics and gaze information can be found in [4].

3.2 Gaze prediction network

The gaze prediction network is adopted from Zhang et al. [4]. The input for the network is a game frame of channel c, width w and height h. Since we are using monochromatic images, the channel size is $c = 1$. For the prediction of the gaze over the frame i at a time instance t_i, we use a history of four frames, i.e., $i...i-3$.

The frames are stacked along the channel dimension to form the input tensor $X^i \in \mathbb{R}^{c\times w\times h}$. A 2-layer convolution block is used to generate the embedding $X^i_{em} = ReLU(BN(\boldsymbol{W_c} * X^i))$ where BN denotes batch normalization. This embedding is used in the *gating network* as well as in the *action prediction network*. The embeddings are further deconvoluted ($\boldsymbol{W_d}$, using a 2-layers deconvolution) to generate the gaze prediction map $\boldsymbol{E_i} = ReLU(BN(\boldsymbol{W_d} * X^i_{em}))$ over the ith game frame. We have used softmax with Kullback–Leibler (KL) divergence loss function between the $\boldsymbol{E_i}$ and true human gaze to train the gaze prediction network. We choose KL divergence because we treated the human gaze over images as a probability distribution (a single Gaussian model); thus, KL is an appropriate measure.

Note that the gaze prediction network parameters are not affected by the action error, and consequently, one can decouple the training of the gaze and action prediction networks. In our implementation, the gaze prediction

network is trained first, and then the trained network is used during action prediction training. Another advantage of decoupling gaze and action network is that any pretrained gaze prediction network can be used with/without fine-tuning. Also, this reduces the number of parameters to be trained.

3.3 Gating network

The gating network will identify the instances at which the gaze information should augment the game frames to predict the human actions. It takes the embeddings X^i_{em} as the input and outputs a binary value of 1 or 0, indicating the use or discarding the gaze information, respectively. The gating function can be modeled as $c^i = g(\boldsymbol{W_g} * X^i_{em})$ with $\boldsymbol{W_g}$ as the learnable gating parameter. The gating function g can be implemented by modifying the popular GRU or LSTM [27] units to be non-recurrent. To implement such a behavior, we modify the *GRU* unit, as shown in Eq. 1. This removes the temporal dependency aspects of the GRU and only preserve the gating functionality.

$$g(\boldsymbol{W_g} * X^i_{em}) := \begin{cases} h^i = GRU(W_g, X^i_{em}, -1) \\ c^i = ReLU(sgn(h^i)) \end{cases} \quad (1)$$

In this formulation, the *GRU* unit is implemented with a default hidden state of -1 ($h = -1$). The output of the gating network c^i depends on the sign of *GRU* unit according to $c^i = ReLU(sgn(h))$. The output is then element-wise multiplied with the predicted eye-gaze map E_i, which is subsequently used to augment the input game-frame. Note that the default value of the *GRU* is –1 ($h = -1$), hence the gaze information by default is not utilized for augmenting the game frame unless the *GRU* values change. As shown in Fig. 2, the *GRU* unit weights are influenced by the error in gaze usage and error in action prediction. Consequently, the efficient use of the predicted gaze depends on how well the gaze itself is estimated. The dependence of gating network performance on predicted gaze is the desired behavior because a well-estimated gaze can filter irrelevant game features and enhance necessary features for action prediction.

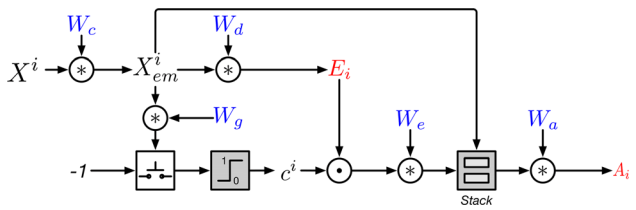

Fig. 2 Learnable weights of SEA network. $\boldsymbol{W_c}$ and $\boldsymbol{W_d}$ are the convolution and deconvolution weight of the gaze prediction network. The convolution operation produces an embedding (X^i_{em}) of the input frames. Thus created embedding (X^i_{em}) is used in gaze prediction, gating network, and action prediction network. $\boldsymbol{W_g}$ is the weights of the gating network. Lastly, $\boldsymbol{W_e}$ and $\boldsymbol{W_a}$ are weights of the action (A) prediction network. Note: ⊛ denotes convolution operation and ⊙ denotes element-wise multiplication. A fully connected layer is a special case of convolution where the dimension of the kernel size is equal to the input tensor

3.4 Action prediction network

The action network uses two types of embedding to predict the actions. The first embedding is from the gating network and calculated as a convolution operation given by $ReLU(BN(\boldsymbol{W_e} * \boldsymbol{E_i} \odot c^i))$ where the c^i is the gating network output, and $\boldsymbol{E_i}$ is the predicted eye-gaze over the input X^i. The second embedding is from gaze prediction network X^i_{em}. The two embeddings are concatenated to form a feature vector to learn the action mapping (Fig. 2). The concatenated feature vector is subsequently forwarded through a sequence of fully connected layers with learnable weights $\boldsymbol{W_a}$. The output of the fully connected layers is one of the 18 feasible actions defined in the Atari game. We use softmax with cross-entropy as the training criterion.

4 Experiments and results

This section provides the experimental setup and results of 3 sub-networks of the SEA across six games from the Atari-HEAD dataset. Each game in the dataset consists of 20 trials (5 trials per subject), out of which 15 trials were used for training, and five trials were used for inference purposes. The training period was 30 hours over different games. The hyperparameters and training details for three sub-networks of SEA are presented in their respective subsections.

4.1 Gaze prediction network results

For gaze prediction, we use a stack consisting of the current frame and the previous three frames. The frames are converted to gray-scale and downsampled to 84 × 84 pixel size before stacking. We closely follow the gaze model architecture in [4], where the gaze is a probability distribution over the 2D image. The probability distribution is calculated by Gaussian estimate with mean and covariance. In terms of human gaze samples, we use the last point of the true eye-gazes on the current frame as the ground truth. The true point gaze is converted to a continuous Gaussian probability distribution with a mean-centered at the gaze point, and standard deviation σ of one visual degree [28]. We use KL divergence as the error criterion. Such training results in a point estimate with learned mean and variance. Figure 3 shows the predicted and actual human gaze in five

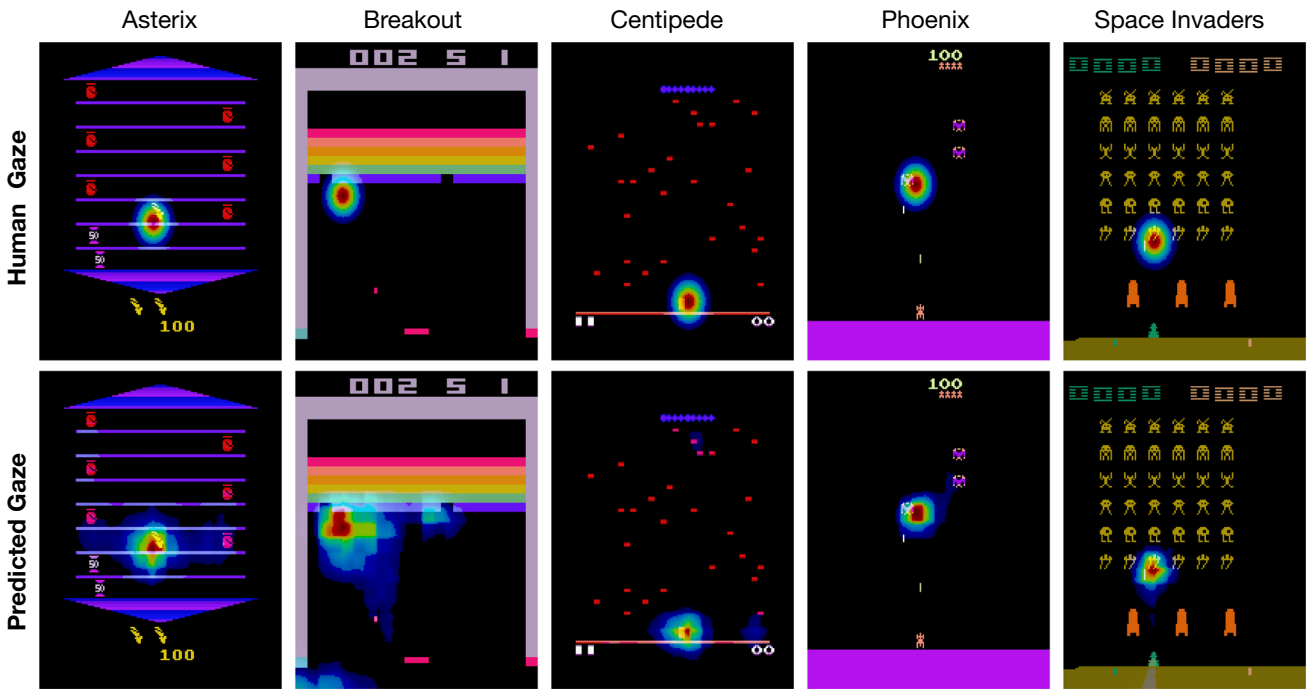

Fig. 3 Comparison of human gaze and predicted gaze from gaze prediction network

different games. The spread of the predicted gaze is more pronounced in some games like Breakout.

Further, we analyze whether the gaze prediction network is capable of understanding the game dynamics. In this regard, Fig. 4 shows the predicted gaze and the ball movement in a Breakout game. The predicted gaze follows the ball closely before hitting the paddle (Fig. 4, frame 1 to frame 6). As the ball leaves the paddle, the gaze prediction shifts toward the bricks even before the ball reaches the brick. We believe that this behavior is not due to the gaze prediction network architecture but due to the human gaze data used for training. As the human gaze encodes causal

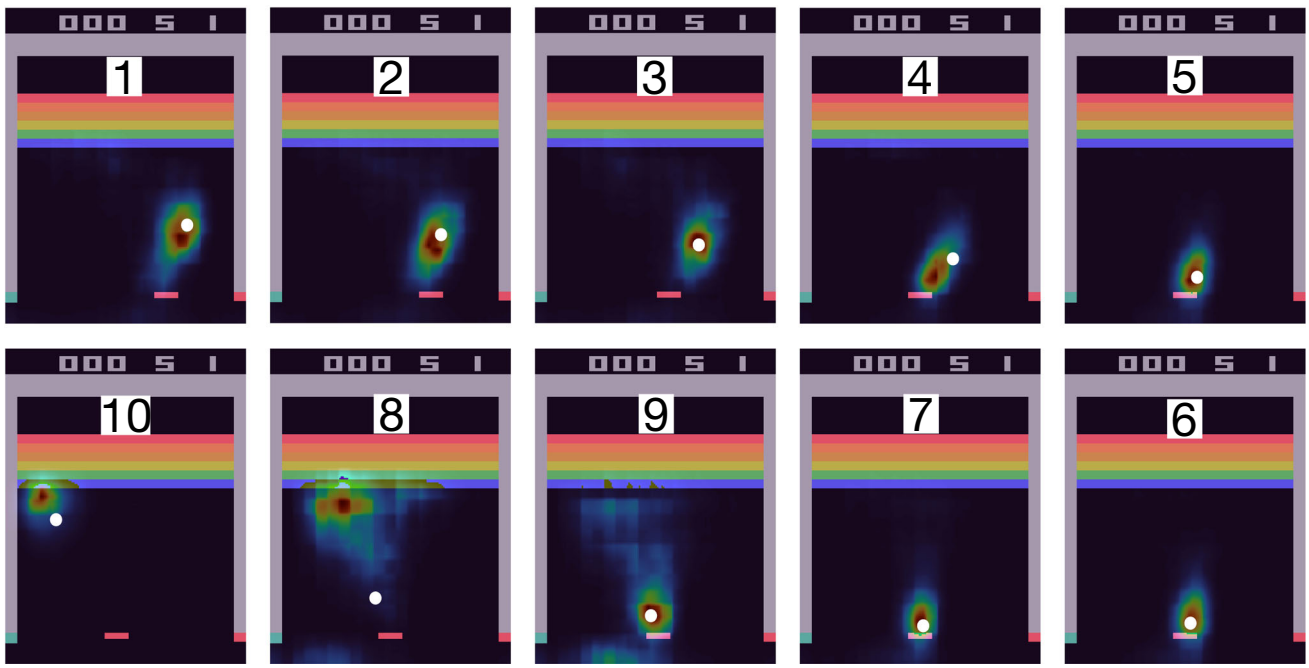

Fig. 4 Predicted gaze over a series of frames in the game of Breakout. The predicted gaze shifts toward the bricks as soon as the ball leaves the paddle (Ball is highlighted for more clarity)

[29, 30] relationship, the gaze prediction network, trained in the human gaze, can learn the causal link to an extent. Such a causal relationship might be hard to learn if we do not use any human data.

4.2 Gating network results

One of the core tenets underlying the selective gaze utilization is that by augmenting the game frames with gaze only when required, we can achieve higher performance than using gaze all the time. To achieve this, we employ the gating network that learns to use the gaze selectively when desired.

Figure 5 shows the use of gaze over a small gameplay duration for two games. The gate output of 1 results in the gaze being used, and the gaze output of 0 results in not being used. We can see that throughout the game, the gating network selects gaze as required and not all the time. For instance, the gating network output was sparse, and switching happened less frequently for SeaQuest. However, for Phoenix, gate output switching is much more frequent.

To further highlight the gating network dynamics, let's consider the Breakout game results (Fig. 6). As it can be seen, the gate output is 0 (thus, no eye gaze is used) when the ball is moving away from the paddle (frame 1 in Fig. 6), at this stage, no action is needed. As the ball starts moving toward the paddle (frames 2 and 3) after hitting the brick, the paddle position should be adjusted by moving the paddle right or left. At this moment, the gate output turns on, thus using the gaze data. The gate remains open until the ball hits the paddle and leaves. These results highlight the efficacy of the gating network. However, the gating network behavior is dependent on the dynamics of the game, which is evident in Table 1. For these games, we can see that the maximum utilization of gaze is under 40% of the total number of frames seen during gameplay.

4.3 Action prediction network results

For action prediction, we train a network using game-frames and predicted eye-gaze information described in Sect. 3.4. We have used an independently trained gaze prediction (see Sect. 3.2) to augment the input game-frames. The predicted gaze is modulated by the gating network (Sect. 3.3) before sending it to the action prediction network. The hyperparameters for action prediction network training are the same as the gaze prediction network. For performance comparison, we use three different baselines: Behavior Cloning (BC), Attention Guided Imitation Learning (AGIL), and Random gated SEA. The random gated version is exactly the SEA network, except the gating function is not learned. Instead, the output is randomly chosen as 1 or 0 from an uniform distribution. The same random behavior is used during the learning and inference phase. In the AGIL approach, the input-game frames are masked with predicted eye-gaze, and this masked game-frame is then used to predict the action. In BC, no eye-gaze information is used; the action is predicted using only the game-frame stack.

Regarding action classification accuracy, Table 2 (note, the highest accuracies are highlighted) provides SEA network performance compared to the baseline approaches. The SEA network performs well only in two games (Asterix and MsPacman), while the AGIL outperforms SEA in the other four games. It should be noted that action classification accuracy is not correlated with the game score. There are two main fundamental differences in the nature of the gameplay and action classification that can potentially cause this performance mismatch. First, the gameplay is dynamical, i.e., actions depend on the previous state, while such a dependency is absent in the classification problem. Hence, the data used in classification are different from the data coming from gameplay. Second, there is often more than one viable action for a given state in the gameplay. This is even seen in the actions recorded by different human subjects. In the gameplay, any action from a valid alternative action set will result in a score, while in

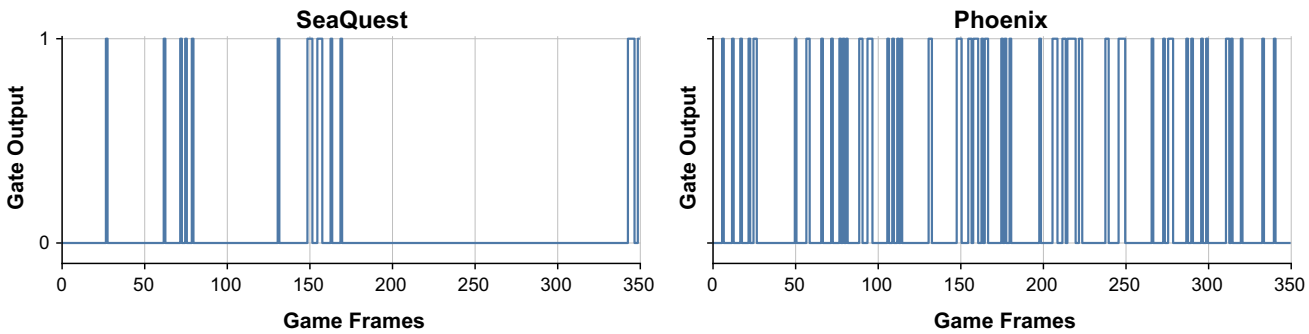

Fig. 5 Gate output for SeaQuest and Phoenix at different games frames

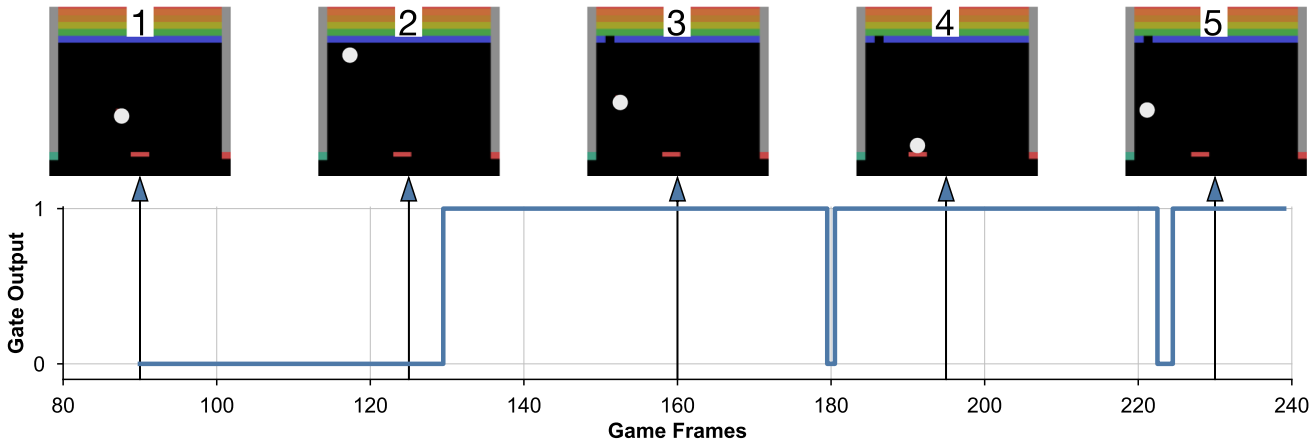

Fig. 6 Gate output from gating network with frames at different time instances. The gate output is 1 when the ball moves toward the paddle. When the ball starts moving toward the paddle, the SEA network starts using the eye-gaze. The gate output is 1 till the ball leaves the paddle. Note: The results are shown from the part of the game (frame 80 to frame 240)

Table 1 Percentage of gaze usage over the entire gameplay in different games

Game	Asterix	Breakout	Centipede	MsPacman	Phoenix	SeaQuest
Gaze usage (%)	0.368	0.365	0.369	0.266	0.227	0.068

Table 2 Action classification accuracy of SEA in comparison with majority action, BC, and AGIL

Game	Majority action	BC	AtariHEAD-AGIL	SEA
Asterix	0.365	0.68	0.532	**0.621**
Breakout	0.8	0.79	**0.816**	0.595
Centipede	0.581	0.37	**0.628**	0.57.4
MsPacman	0.266	0.555	0.678	**0.681**
Phoenix	0.291	0.33	**0.658**	0.545
SeaQuest	0.208	0.47	**0.505**	0.37

the classification problem, it will be counted as a miss and hence reducing the classification accuracy.

To shed more light on this matter, let's consider the game of Breakout. The player must use the paddle below to guide (by moving right or left) the ball knocking down as many bricks as possible. When the ball is moving upward, the action taken will have no effect on the score, but it contributes to the classification accuracy. For this specific example, we have observed about 10% higher classification accuracy (50% vs. 40.5%) at the moments that actions directly affect the game score (ball moving downward vs. going upward). For some of the games, this can be further generalized into two categories: taking any actions vs. taking no action at all. In other words, the classification performance can be calculated by grouping the human-demonstrations as action and no-action, which removes the discrepancy caused by different players taking different actions for the same game state.

To further clarify, Table 3 provides F1-scores when classification is done between no action and action; here, the exact subject's action is irrelevant (hence high F1 score). However, when we consider what exact action the subject did, the F1 score decreases. This further validated that even though the subjects took different actions, the game score was not affected, and hence a good classification accuracy need not reflect a good game score or vice-versa.

5 Game performance analysis

The trained SEA network and the baselines are evaluated thirty times for each game. During the evaluation, the same random seed is kept across SEA and other baselines. The averaged games scores and the standard deviations are listed in Table 4 (highest scores are highlighted). To

Table 3 F1 scores of classification with all the actions and dropping no-action/invalid actions

Game	Asterix	Breakout	Centipede	MsPacman	Phoenix	SeaQuest
Action versus no action	0.95	0.76	0.93	0.98	0.90	0.97
All the actions	0.62	0.65	0.57	0.70	0.48	0.37

Table 4 Game scores of SEA in comparison with BC, AGIL, and random gated SEA

	SEA		BC		AGIL		Random SEA	
Game	Raw score	z-score	Raw score	z-score	Raw score	z-score	Raw score	z-score
Asterix	**608.3 ± 148.3**	0.96	246.7 ± 166.8	−0.87	410 ± 153.0	−0.042	408.3 ± 122.5	−0.051
Breakout	7.13 ± 2.68	0.733	1.76 ±1.54	−1.03	**7.26 ± 1.61**	0.787	3.4 ± 1.36	−0.492
Centipede	**13023 ± 4333**	1.26	528 ± 331	−1.16	12099 ± 4512	−0.214	7086 ± 2376	0.112
MsPacman	**1258 ± 385.7**	1.07	265 ± 85.5	−1.099	1008 ± 255.0	0.523	542 ± 152.1	−0.494
Phoenix	**4905 ± 1106**	0.558	4461 ± 1361	0.196	3503 ± 1098	−0.588	4019 ± 751.8	−0.166
SeaQuest	**304.6 ±43.7**	1.069	104.00± 24.44	−1.407	232.67 ± 37.77	0.181	230.67 ± 34.92	0.156

statistically compare the performance of the SEA with the other methods, we treat each game as a separate domain on which all the learning algorithm are tested. However, the range of score in each game is not the same. Hence, we calculate a z-score of each game (across all the algorithms) to remove the game specific effects. A Friedman's Chi-square test on the average z-score (Listed in Table 4) is conducted which resulted in a significant difference between the performance of different algorithms (p value 0.0081). Consequently, we conduct post-hoc pair-wise Welch's t-tests with Bonferroni correction for multiple comparisons (adjusted $\alpha : 0.05/3 = 0.016$) to protect against type-I error inflation. The results of post-hoc analysis are shown in Table 5. It can be seen that the SEA approach outperforms all other methods. It should be noted that AGIL provides a slightly better average score in Breakout comparing to SEA, but there is no statistical difference between the scores. The detailed comparison of the SEA outcome with the baseline methods is discussed next.

BC versus SEA In behavior cloning (BC), the policy is a simple action imitation through a straightforward classification of actions. On the other hand, SEA selectively augments the game frame with gaze for action classification. We can emulate the BC approach in SEA by keeping the gate-out zero. The performance gain in SEA, when compared to BC, is because the supplemented gaze information helps to guide the system toward important aspects of the frame similar to [4]. As shown in Table 5, SEA significantly outperforms BC (p value: 2.6E-5). Interestingly, even randomly using the gaze information (Random gated SEA) outperforms the BC approach (Table 4) in 5 out of 6 games, indicating the advantage of using the human-gaze to enhance imitation learning.

AGIL versus SEA Attention guided imitation learning [4] (AGIL) explicitly calculates the gaze and masks the game frame with calculated gaze to predict the action. From Table 4, it is evident that AGIL outperforms the simple behavior cloning (BC). However, SEA is developed on the main hypothesis that learning to augment a game frame with the gaze selectively should perform better than augmentation at all times, especially if the gaze is directly integrated without considering its dynamic and type (e.g., fixation vs. saccade). Consequently, we can see in Table 5 that SEA outperforms AGIL (p value: 0.0074). If the SEA model chooses to ignore all the gaze information, the model performance should fall back to behavior cloning (BC), and using all the gaze data should result in at least the performance of AGIL. It is not surprising that the AGIL performance is slightly higher than the random gated SEA network (even lower in the Phoenix game case). This further questions the overall benefits of using gaze augmentation blindly.

Table 5 p values between different algorithms corrected for multiple comparison

	SEA versus BC	SEA versus AGIL	SEA versus Random SEA
p value	2.6E−5* (< 0.016)	0.0045 * (< 0.016)	3.6E−5 * (< 0.016)

*Significant difference

SEA versus Random gated SEA By comparing the SEA with AGIL, we showed the benefits of selective gaze usage for demonstration learning. However, it is also critical to show that the proposed gating network is indeed learning the instances at which the gaze information should be used. This is done through the comparison of SEA with a random gated SEA. As seen in Table 5, the SEA approach outperforms the random gaze augmentation method p value: 3.856E−5, which indicates that the learned gating function output has a pattern (i.e., is not random), and the gating behavior indeed depends on the game dynamics under consideration.

6 Conclusions

In this work, we propose a selective eye-gaze augmentation (SEA) approach in which the network learns when to use gaze information to enhance the demonstration learning. We demonstrate the efficacy of selective eye-gaze augmentation on 6 Atari games. The game data consist of human demonstrations and eye movement over the game frames. The SEA network uses a gating mechanism whose output is either 1 or 0. When the output is 1, the input game-frame is augmented with predicted eye-gaze, and on the contrary, if the output is 0, only the input game-frame is used to predict the appropriate action. It can be thought of as a more generalized version of simple behavior cloning (BC, no eye-data) and attention guided imitation learning (AGIL, mask all the game frames with eye-data). We can emulate both BC and AGIL networks by modulating the gating behavior.

The SEA network outperforms both BC and AGIL approaches in several Atari-games. To demonstrate selective eye-gaze augmentation's effectiveness, we considered a case where the eye-gaze is randomly augmented with a game-frame. The random eye-gaze augmentation performed significantly better than BC in several games. However, the performance was not better when compared with the AGIL approach. Thus, the results indicate the benefits of gaze in enhancing learning and highlight the importance of selective gaze usage.

It should be noted that the SEA gating network is independent of the task and its associated visual complexity. As a future direction, more evaluation studies may be conducted for games with a higher level of visual complexity to examine the proposed network's performance. Further, the present SEA architecture does not consider any temporal dependence. Hence, we plan to implement the SEA network (sub-networks) with temporal dependence using recurrent neural networks in our future works. Finally, one can also extend the SEA to a reinforcement learning setting where the selective use of game frames can be learned online as a potential future direction of this work.

Acknowledgements We gratefully acknowledge NVIDIA Corporation's support with the donation of the Titan Xp GPU used for this research.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Judah K, Fern A, Tadepalli P, Goetschalckx R (2014) Imitation learning with demonstrations and shaping rewards. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 1890–1896
2. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M et al (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484
3. Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, Choi DH, Powell R, Ewalds T, Georgiev P et al (2019) Grandmaster level in Starcraft ii using multi-agent reinforcement learning. Nature 575(7782):350–354
4. Zhang R, Walshe C, Liu Z, Guan L, Muller KS, Whritner JA, Zhang L, Hayhoe MM, Ballard DH (2020) Atari-head: Atari human eye-tracking and demonstration dataset. AAAI Conference on Artificial Intelligence (AAAI)
5. Nikulin D, Ianina A, Aliev V, Nikolenko S (2019) Free-lunch saliency via attention in atari agents. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, pp 4240–4249
6. Saran A, Zhang R, Short ES, Niekum S (2020) Efficiently guiding imitation learning algorithms with human gaze. arXiv preprint arXiv:2002.12500
7. Zhang R, Liu Z, Zhang L, Whritner JA, Muller KS, Hayhoe MM, Ballard DH (2018) Agil: Learning attention from human for visuomotor tasks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 663–679
8. Li Y, Liu M, Rehg JM (2018) In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 619–635
9. Neumann O (2016) Beyond capacity: a functional view of attention. Perspectives on perception and action. Routledge, pp 375–408

10. Houghton G, Tipper SP (2013) A model of selective attention as a mechanism of cognitive control. Localist connectionist approaches to human cognition. Psychology Press, pp 49–84
11. Castiello U (2005) The neuroscience of grasping. Nat Rev Neurosci 6(9):726–736
12. Cisek P (2007) Cortical mechanisms of action selection: the affordance competition hypothesis. Philos Trans R Soc B Biol Sci 362(1485):1585–1599
13. Petrosino G, Parisi D, Nolfi S (2013) Selective attention enables action selection: evidence from evolutionary robotics experiments. Adapt Behav 21(5):356–370
14. Zhao M, Gersch TM, Schnitzer BS, Dosher BA, Kowler E (2012) Eye movements and attention: the role of pre-saccadic shifts of attention in perception, memory and the control of saccades. Vis Res 74:40–60
15. Gibson JJ (2014) The ecological approach to visual perception, classic. Psychology Press
16. Miller J, Hackley SA (1992) Electrophysiological evidence for temporal overlap among contingent mental processes. J Exp Psychol Gen 121(2):195
17. Land M, Mennie N, Rusted J (1999) The roles of vision and eye movements in the control of activities of daily living. Perception 28(11):1311–1328
18. Ahlstrom C, Victor T, Wege C, Steinmetz E (2011) Processing of eye/head-tracking data in large-scale naturalistic driving data sets. IEEE Trans Intell Transp Syst 13(2):553–564
19. Gredebäck G, Falck-Ytter T (2015) Eye movements during action observation. Perspect Psychol Sci 10(5):591–598
20. Flanagan JR, Johansson RS (2003) Action plans used in action observation. Nature 424(6950):769–771
21. Bellemare MG, Naddaf Y, Veness J, Bowling M (2013) The arcade learning environment: An evaluation platform for general agents. J Artif Intell Res 47:253–279. https://doi.org/10.1613/jair.3912
22. Barrett M, Bingel J, Hollenstein N, Rei M, Søgaard A (2018) Sequence classification with human attention. In: Proceedings of the 22nd Conference on Computational Natural Language Learning, pp 302–312
23. Penkov S, Bordallo A, Ramamoorthy S (2017) Physical symbol grounding and instance learning through demonstration and eye tracking. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp 5921–5928
24. Palazzi A, Abati D, Solera F, Cucchiara R et al (2018) Predicting the drivers focus of attention: the dr (eye) ve project. IEEE Trans Pattern Anal Mach Intell 41(7):1720–1733
25. Chen Y, Liu C, Tai L, Liu M, Shi BE (2019) Gaze training by modulated dropout improves imitation learning. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp 7756–7761
26. Chen Y, Liu C, Shi BE, Liu M (2020) Robot navigation in crowds by graph convolutional networks with attention learned from human gaze. IEEE Robot Autom Lett 5(2):2754–2761
27. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555
28. Meur OL, Baccino T (2012) Methods for comparing scan paths and saliency maps: strengths and weaknesses. Behav Res Methods 45(1):251–266. https://doi.org/10.3758/s13428-012-0226-9
29. Adams RA, Bauer M, Pinotsis D, Friston KJ (2016) Dynamic causal modelling of eye movements during pursuit: confirming precision-encoding in v1 using meg. Neuroimage 132:175–189
30. Gerstenberg T, Peterson MF, Goodman ND, Lagnado DA, Tenenbaum JB (2017) Eye-tracking causality. Psychol Sci 28(12):1731–1744