Machine versus Human Attention in Deep Reinforcement Learning Tasks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep reinforcement learning (RL) algorithms are powerful tools for solving visuomotor decision tasks. However, the trained models are often difficult to interpret, because they are represented as end-to-end deep neural networks. In this paper, we shed light on the inner workings of such trained models by analyzing the pixels that they attend to during task execution, and comparing them with the pixels attended to by humans executing the same tasks. To this end, we investigate the following two questions that, to the best of our knowledge, have not been previously studied. 1) How similar are the visual features learned by RL agents and humans when performing the same task? and, 2) How do similarities and differences in these learned features explain RL agents' performance on these tasks? Specifically, we compare the *saliency maps* of RL agents against visual attention models of human experts when learning to play Atari games. Further, we analyze how hyperparameters of the deep RL algorithm affect the learned features and saliency maps of the trained agents. The insights provided have the potential to inform novel algorithms for closing the performance gap between human experts and RL agents.

1 Introduction

Researchers have devoted much effort to understanding deep neural networks (DNNs). Since DNNs are partially inspired by the biological nervous systems, researchers have compared these neural networks with human brain and sensory systems by asking two typical types of questions. The first one considers representation: How similar are the visual features learned by DNNs and humans when performing the same tasks? The second one concerns explainability: How do similarities and differences in these learned features explain DNNs' performance on their tasks?

The first question motivates a seminal line of research that compares features learned by DNNs with features learned by humans. In computer vision, a linear model that uses the feature map activations generated by a trained CNN can accurately predict the neural activities in early visual cortex, indicating that the two systems have learned similar visual features [17, 74, 75]. In language learning, a similar connection has been found between deep language models and cortical areas [29, 32]. However, this type of comparison has just emerged in decision-making research [41, 11], which motivates us to compare DNNs trained for deep reinforcement learning (RL) tasks with human decision-making. We ask: Do deep RL agents and humans agree on what visual features are important? In other words, do agents pay *attention* to the same visual features as humans do?

The explainability question is motivated by the Explainable AI (XAI) for deep RL agents [27, 1, 52]. Deep RL has achieved many successes, but few of us fully understand these agents. These agents often learn a mapping from raw images to actions end-to-end where it is not clear why a particular decision is made. Our work addresses the explainability of these black-box models: Do RL agents make mistakes because they fail to attend to important visual features that matter for making the

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

correct decision? An expert human's attention data could serve as a useful reference for identifying important visual features, which has been validated in object recognition tasks [47].

To answer these two questions, we situate our research in the domain of Atari games [4]. These games span a variety of dynamics, visual features, reward mechanisms, and difficulty levels for both humans and AIs. Two recent lines of research have made our work possible. On the one hand, a human attention dataset on Atari games has been collected using eye trackers [81]. On the other hand, tools have been introduced to generate visual interpretations of RL agents [21] so one can visualize an RL agent's *saliency map* given an image, which is a topographically arranged map that assigns an importance weight to each image pixel and is often treated as the "attention map" of the RL agent. However, these tools can only provide qualitative results – we still need to manually inspect the saliency maps to interpret the performance. A useful comparison standard like human attention had not been considered hence systematic quantification was impossible.

In this work, we present the first study that compares the RL agent's attention with human attention. We analyze how learning and hyperparameters of the RL algorithm affect the learned features and saliency maps. We show that RL agents' mistakes are related to failing to attend to the correct visual targets. We further study whether RL agents' attention generalizes to unseen states. We conclude by discussing insights gained for RL researchers in both cognitive science and AI.

2 Related Work

Human vs. machine attention in vision and language tasks. Human expert's gaze is very efficient and accurate for solving vision tasks. The peak angular speed of the human eye during a saccade reaches up to 900 degrees per second [56]. This allows humans to move their foveae to the right place at the right time to attend to important features [15]. Therefore, human expert's gaze serves as a good standard in many vision-related tasks for evaluating machine attention, or as a learning target for training machine attention [54, 80]. This approach is widely used in computer vision, see Nguyen et al. [48] for a review. For example, visual saliency researchers train DNNs to predict human visual attention. Saliency research is a well-developed field and we direct interested readers to recent review papers [5, 8, 24, 7]. One such paper has compared visual saliency models with human visual attention [36]. In vision-related language tasks, such as image captioning and visual question answering, it was found that the saliency maps of DNN models are different from human attention [13, 69, 25]. Understanding and quantifying such differences have provided insights on the performance, especially in failure scenarios, of these vision-language models.

Visual explanation for deep RL. Two classes of methods are widely used to generate visual interpretations of DNNs in the form of saliency maps: *gradient-based* and *perturbation-based*. Gradient-based methods compute saliency maps by estimating the input features' influence on the output using the gradient information [64, 65, 40, 78, 63, 68, 61, 10, 82]. These methods are for visualizing general DNNs but have been used to interpret deep RL agents [35, 72, 62, 33, 71]. We did not use gradient-based saliency maps for our analysis because they lack physical meaning and could be difficult to interpret, since changing an input in the direction of the gradient tends to move it off from the manifold of realistic input images [21, 23]. Perturbation-based methods alter parts of the input image and measure how much the output is affected by the change. Hence there have been different methods of alternating the input [77, 20, 12, 83, 55]. These methods have been applied to Atari deep RL agents and can generate qualitatively meaningful saliency maps [21, 31, 22, 53]. However, without human attention as a reference, it is difficult to quantitatively analyze these saliency maps, which motivates our work. There are also methods that change the architecture of the deep RL network by augmenting it with an explicit artificial attention module, so that one can directly access its attention map [46, 76, 45]. Researchers have taken this approach and compared RL agents' attention with human attention [49]. However, these methods do not apply to general deep RL algorithms since they need to modify the original network architectures and retrain the new ones.

3 Method

We now discuss methods for modeling human attention, training RL agents, and extracting attention information from the trained RL agents. We then discuss how we select data for comparing human versus RL attention, and define the comparison metrics.

Human attention data and model We use human expert gaze data from Atari-HEAD dataset [81]. The original game runs continually at 60Hz [4], a speed that is challenging even for professional gamers. Human eye movements were rushed and inaccurate at this speed hence could not serve as a useful reference. In [81], however, the human data were collected in a semi-frame-by-frame mode, a design that allowed enough time for the players to attend to all relevant objects on the screen and make decisions. The slowed speed allowed players to achieve world expert level performance [81] which is about 35x better than the human performance reported in previous deep RL literature [43, 26].

To get human saliency maps for the image states generated by RL agents, we need an accurate human attention model. RL agents make many more mistakes than expert humans and cannot reach the late stages of the games. This creates a state distribution that does not match human data, which contains mostly good states. Having a human attention model allows us to perform comparisons in states encountered by RL agents only, which is especially important when we later analyze RL agents' failure states.

Predicting human attention is a challenging task in Atari games. The human gaze is rarely on the player's avatar or contingency regions [4] so simple object detectors would not work. Human players often select and focus on a few among multiple visually identical objects, and divide their attention if multiple objects are relevant for decision making. Recently, variants of convolution-deconvolution network models have achieved the best results on predicting human attention from gaze data [39, 79, 50, 14]. Here we follow their approach to train models with human gaze data to produce saliency maps as the ground truth. As a control analysis, we include two attention models in addition to human gaze. The first one captures the motion information, measured by Farneback optical flow between two consecutive images [18]. The second model captures salient low-level image features, including color, orientation, and intensity (weighted equally), computed by the classic Itti-Koch saliency model [30].

Reinforcement learning agent and attention model As a case study, we use a popular deep RL algorithm named Proximal Policy Optimization (PPO) [60] with default hyperparameters [28]. For each experimental condition (discussed later), we train 5 models with different random seeds to capture the variance in training and ensure reproducibility. For the Atari gaming environment, we use the basic version that has no frame skipping and no stochasticity in action execution (NoFrameskip-v4 version). In order to capture variance in training and ensure reproducibility, we select six popular Atari games instead of using all Atari games and train 630 agents in total (300 GPU days on GeForce GTX 1080/1080 TI).

We use a perturbation-based method [21] to extract attention maps from PPO agents, which has been validated by several subsequent studies [23, 62]. 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 region around and produces a perturbed image $\Phi(I, i, j)$. A saliency score for this pixel (i, j) can be defined as how much the blurred image changes the policy π:

$$S_\pi(i, j) = \frac{1}{2}||\pi(I) - \pi(\phi(I, i, j))||^2 \tag{1}$$

Instead of calculating the score for every pixel, [21] found that computing a saliency score for pixel i mod 5 and j mod 5 produced good saliency maps at lower computational cost for Atari games. The final saliency map P is normalized as $P(i, j) = \frac{S_\pi(i,j)}{\sum_{i,j} S_\pi(i,j)}$.

Comparison metrics Next, we compile a set of game images to compute human and RL saliency maps. For each game, we let a trained PPO agent (with default hyperparameters) play the game until terminated, and uniformly sampled 100 images from the recorded trajectory. We will refer to this set of images as the standard image set.

We then define two metrics for comparing saliency maps when using RL data: Pearson's Correlation Coefficient (CC) and Kullback-Leibler Divergence (KL). Let Q denote the human saliency map predicted by the human attention network. CC is between -1 and 1 captures the linear relation between two distributions Q and P:

$$CC(P, Q) = \frac{\sigma(P, Q)}{\sigma_P \times \sigma_Q} \tag{2}$$

where $\sigma(P, Q)$ denotes the covariance, and σ_P and σ_Q are the standard deviations of P and Q respectively. CC penalizes false positives and false negatives equally.

However, we may not want to penalize the RL agent if it attends to regions that the human gaze model is not currently focused on. The human gaze data only reveals the "overt" attention, and humans can still pay "covert" attention to entities in the working memory [51, 59]. In other words, being attended to by the human gaze is a sufficient (but not necessary) condition for the features to be important. Thus we need a second metric that penalizes the 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 is an ideal candidate in this case [7]:

$$KL(P, Q) = \sum_i \sum_j Q(i, j) \log \left(\epsilon + \frac{Q(i, j)}{\epsilon + P(i, j)} \right) \tag{3}$$

where ϵ is a small regularization constant (chosen to be 2.2204e-16 [7]) and determines how much zero-valued predictions are penalized.

4 Results

To make meaningful comparisons, we first ensure that the human attention model is accurate, and that PPO agents' attention maps are consistent over repeated runs. We then compare human attention with PPO attention obtained from different learning stages and from agents that are trained with different discount factors. We then analyze PPO agents' attention in failure and unseen states. At last, we show comparison results for other deep RL algorithms.

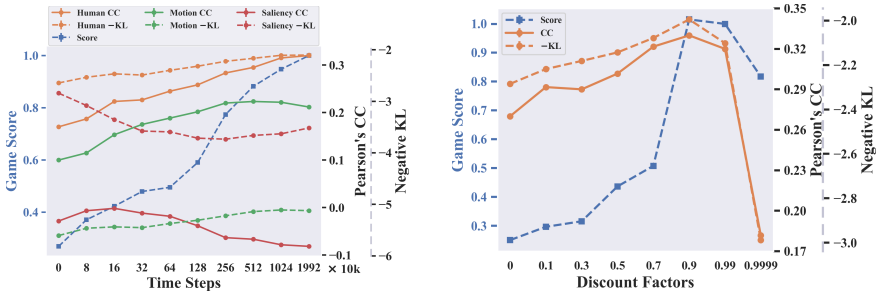
Accuracy of human attention model We implement the convolution-deconvolution gaze prediction model [81] to generate a human saliency map for image I_t at timestep t, given a stack of four consecutive images $I_{t-3}, I_{t-2}, I_{t-1}, I_t$ as input. We use 80% gaze data for training and 20% for testing. The model can accurately predict the human gaze. On testing data, we obtained Area under ROC Curve (AUC) score of 0.968 ± 0.005, CC of 0.562 ± 0.030, and KL of 1.411 ± 0.114 $(n = 6)$, averaged over all games. Prediction accuracy for individual games can be found in Appendix 1 of the supplementary materials. A visualization of the gaze data and prediction results can be found in the attached video file. This accuracy is considered high in saliency research [7].

Consistency of RL attention We then show that saliency maps of the RL agents trained under the same experiment setting, provided by [28], are highly consistent despite the stochasticity in the training process. The stochasticity is controlled by a random seed that is used to initialize both the game environment and the network. For each game, we use 5 random seeds (0-4), train an agent using each seed, and generate 5 saliency maps with these trained agents. For each image in the standard set, we compute pair-wise CCs between the 5 saliency maps (10 CCs in total). The average value for these 10 CCs, across 100 images and 6 games, is $0.924(\pm 0.001, n = 6000)$. Given such high consistency, the saliency maps we use later are the averaged results of these 5 saliency maps.

4.1 RL versus human attention: The effects of learning

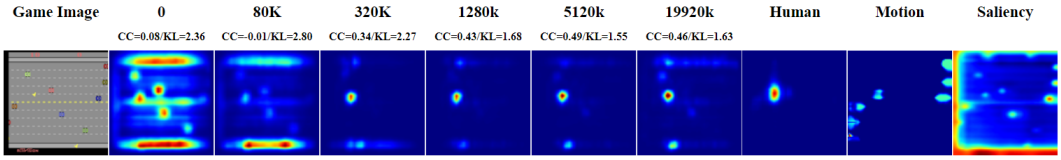
So far we have verified that the gaze model can accurately predict human attention and that RL attention is consistent across repeated runs. We now address the primary research question: How similar are the visual features learned by RL agents and humans when performing the same task? We first study how the attention of PPO agents evolves over time, compared with human attention. For each game, we saved neural network weights at different time steps during training. Then we use these saved models to generate saliency maps on the standard image set.

Fig. 1a shows the aggregated results for all games. CC values and (negative) KL values between human and RL increase during learning, indicating that the RL agents' attention gradually becomes more human-like. We visualize the change of RL attention and human attention in Fig. 2. Networks without any training (time step 0) have saliency maps that are positively correlated with humans (CC$= 0.170$). For example, the second column in Fig. 2 shows that these networks are already sensitive to low-level salient visual features without any learning, which is consistent with previous findings [75, 21]. After training, human and RL saliency maps are more positively correlated (CC$= 0.320$), meanwhile RL and human are more similar to each other than the control models (higher CC/-KL values after training).

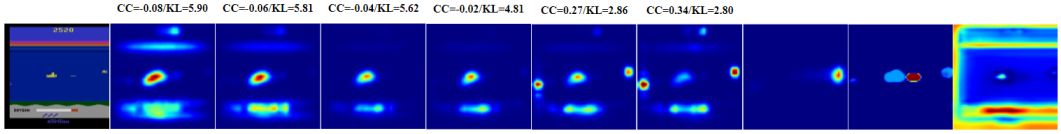


(a) (b)

Figure 1: (a) Changes in human and RL attention similarity across learning time steps. PPO agents gradually learn to pay attention to important visual features and become more human-like. We also show two control attention models: motion (optical flow) and saliency (low-level image features). The x-axis is log scale and KL values are negated for better visualization. The CC, KL, and score results for individual games, as well as more examples, can be found in Appendix 2. (b) Changes in human and RL attention similarity across different discount factors. Choosing different discount factors affects the RL agent's attention and performance. The CC, KL, and score results for individual games, as well as more examples, can be found in Appendix 3.



(a) Freeway: The RL agent gradually learns to focus its attention on the yellow chicken crossing the highway.



(b) Seaquest: The agent learns to attend to the yellow enemy on the right.

Figure 2: Attention of RL agents changes during learning and becomes more human-like. CC/KL values are calculated between the RL agent's attention map and the human attention map.

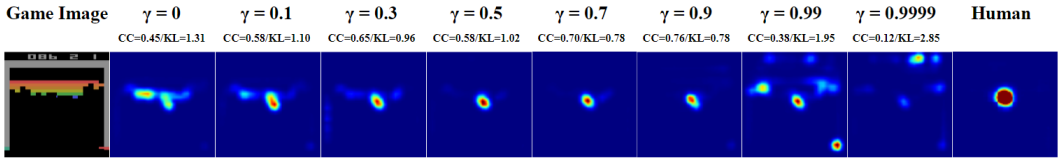
We also show aggregated game performance in normalized game scores in Fig. 1a. For each game, we normalize the game scores obtained during learning (averaged over 50 episodes) by dividing them by the final scores. We find a strong positive correlation between the human CC/KL values and the game score (average Pearson's correlation coefficient of $0.813/0.790$). The correlations for all games are statistically significant ($p < 0.05$, Appendix 2), indicating that small changes in similarity with human attention are reliable predictors of performance change. For comparison, the correlation values between CC/KL and game score for motion baseline are $0.404/0.251$, for saliency baseline are $-0.258/-0.688$.

This result sheds light on an important research topic: bottom-up versus top-down attention. The two sides debate how much human or machine attention is driven by bottom-up image features captured by the saliency model [30], and how much it is driven by top-down task signals such as reward [57, 6]. Our result suggests that learning is a key factor in this debate. In the early stages, attention is more driven by image features, indicated by the higher similarity between RL attention and saliency baseline. Then top-down reward signals shape the attention during learning by making reward-associated objects more salient and irrelevant objects less salient, as shown in Fig. 2.

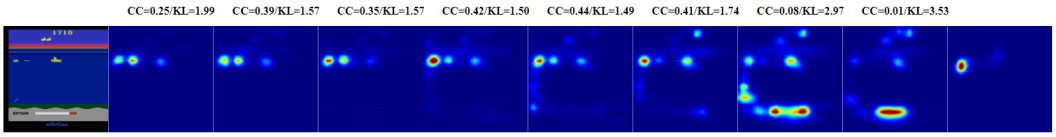
4.2 RL versus human attention: Discount factors

We then analyze how hyperparameters of the PPO algorithm affect the attention of the trained agents. We have seen that reward shapes attention during training, therefore a reasonable hypothesis is that varying reward-related parameters will likely affect attention. The parameter we can vary in these

games is the discount factor $\gamma \in [0, 1)$, which determines how much the RL agent weighs future reward compared to immediate reward. The default γ is 0.99 for all games [28]. We train PPO agents with $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9, 0.9999\}$ and generate saliency maps on the standard image set. The results are shown in Fig. 1b. Overall, the RL attention is most similar to human's when $\gamma = 0.9$. From the visualizations in Fig. 3 we can see that with small γs, the agents tend to be myopic and only focus on very few objects. With high γs, the agents divide their attention into several objects. Since the human gaze model only captures the overt attention and involves typically 1-3 objects, intermediate γ values produce saliency maps that are similar to humans.



(a) Breakout: The human pays attention to the ball.



(b) Seaquest: The human pays attention to an approaching enemy from the left and oxygen level at the bottom.

Figure 3: Effect of different discount factors on Ms.Pac-Man and Seaquest agents' attention. Agents trained with intermediate discount factor values have saliency maps that are more human-like.

For each γ value, we normalize the game score by dividing it by the score obtained by the $\gamma = 0.99$ agent. For most games, this default value has the best performance. However, the performance increases with CCs and negative KL values only up to $\gamma = 0.7$ in Fig. 1b. Beyond that, the RL agents need to learn to attend to objects that matter in the long-term, which are likely to be captured by human covert attention and do not show up in our human attention model.

Nonetheless, deviating from the default $\gamma = 0.99$ can lead to better performance. In Seaquest, the agent with $\gamma = 0.9$ achieves 15% higher score than the default agent. Fig. 3b shows that with a lower $\gamma = 0.9$, the agent can focus, like humans, on an immediate threat from the left. Setting $\gamma = 0.99$ or 0.9999 distracts the agent to attend to the oxygen bar at the bottom that is important in the long run but less urgent now. A similar result was found for Ms.Pac-Man with an 18% higher score from the $\gamma = 0.9$ agent than the $\gamma = 0.99$ agent.

This result has an important implication. Atari games are episodic tasks with true $\gamma = 1$ but lower γ values often lead to better performance in practice [60, 43][1]. Fig. 1b shows that $\gamma = 0.9999$ agents perform poorly. We have verified that all $\gamma = 0.9999$ agents (5 random seeds) converged after 200M samples, although to sub-optimal policies (Appendix 3). Again, the reason is that being a little myopic helps the agent focus on the most urgent targets. This result provides another reason, from a perception perspective, for why RL agents need to adjust their planning horizon by reducing the discount factor–confirming the theoretical results provided by [34].

4.3 RL versus human attention: Failure states

We now turn to the second research question that concerns explainability in deep RL: Why do deep RL agents make mistakes? Did RL agents fail to attend to the right objects, or did they attend to the right objects but make wrong decisions? We compile a second image set, named failure image set, by recording the game frames right before the RL agent loses a "life" which incurs a large penalty. We locate 100 such instances for each game. Freeway is excluded since the PPO agent learned a policy that is nearly optimal.

Fig. 4d shows two failure cases for Breakout. In the first case, the attention map of the RL agent is extremely different from the human's, and in the other, they are highly similar. In the first case, the

[1] Strictly speaking, γ is a property of the MDP, not of the agent. Performance across MDPs with different γs are not directly comparable in this sense. Varying discount factor and clipping reward are common reward engineering designs that alter the true MDP return to achieve better performance in practice.

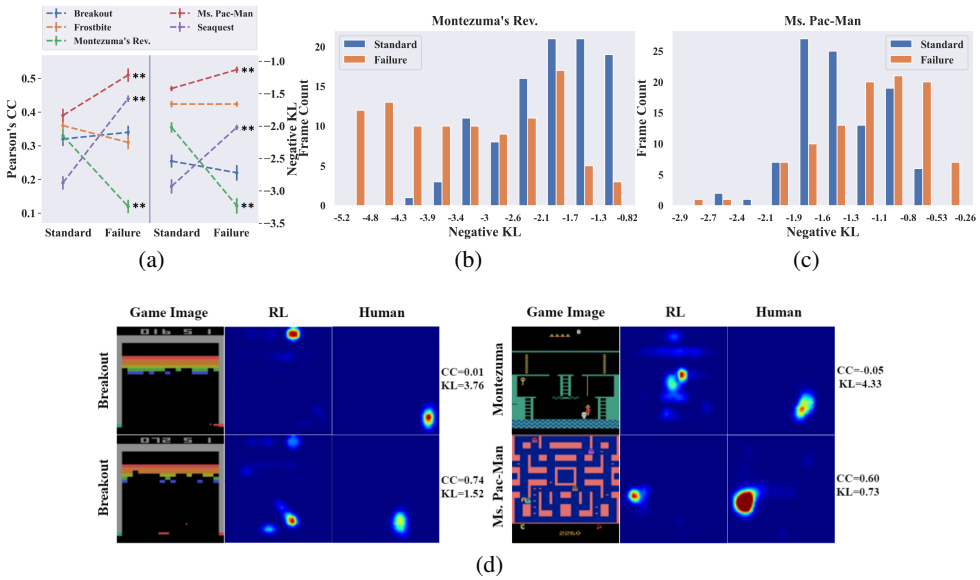


Figure 4: RL vs. human attention in states where RL agents made mistakes.(a) How attention similarities change in failure states compared to normal states. *indicates a significance test result of $p \leq 0.05$; ** indicates $p \leq 0.01$. Error bars are the standard errors of the mean ($n = 100$). (b,c) The histograms of negative KL values for 100 standard vs. 100 failure states indicate that the RL agents are more likely to neglect (Montezuma's Revenge) or attend to (Ms.Pac-Man) human attended regions. (d) We show examples of RL and human saliency maps for Breakout (top: wrong attention; bottom: right attention but wrong decision), Montezuma's Revenge (wrong attention), and Ms.Pac-Man (right attention but wrong decision).

RL agent failed to attend to the ball as the human model did, this may be the reason why it failed to move the paddle to catch the ball in the next frame (it is still possible to catch that ball). In the second case, both the agent and the human attend to the ball. In this way, we can interpret failure cases using human attention as a reference. If RL attention and human attention are more different in failure than normal states, then the model has not learned to pay attention to the right object.

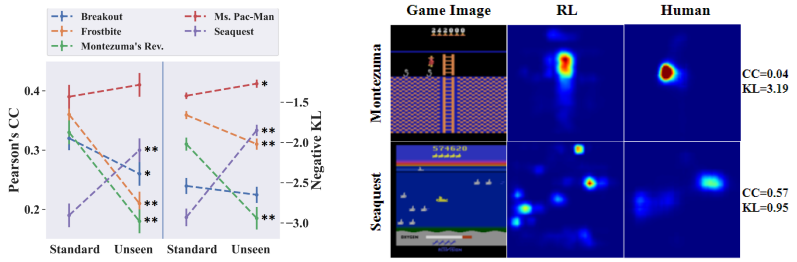
Fig. 4a shows the quantitative results for all games. For Montezuma's Revenge, the RL attention in failure states becomes less similar to human attention compared to the standard image set, indicated by significantly decreased CC/KL values. Fig. 4b shows the (negative) KL histogram of 100 failure states vs. normal states, from which we can see that in the failure states the RL agents are more likely to neglect human attended regions. Fig. 4d shows that the agent fails to attend to the enemy as the human does.

On the contrary, for Ms. Pac-Man and Seaquest (Fig. 4a), in failure states, attention maps of RL agents and humans are more similar than in the normal states. Fig. 4c shows the (negative) KL histogram of 100 failure states vs. normal states in which the human and RL agents are more likely to agree on the objects to be attended to. These are often situations with objects that are more dangerous (see Appendix 4). Hence it is easy for RL agents to attend to the right object but the decision is hard. For example, in Fig. 4d Ms.Pac-Man, the agent attended to the Pac-Man and the enemy ghost similarly to humans. However, the RL agents made the wrong decision and ran into the ghost.

We conclude that the mistakes made by the RL agents are more likely related to perceptual errors in certain games. Similarity measurements with human attention can help identify and interpret the failures made by RL agents.

4.4 RL versus human attention: Unseen data

Next, we study whether the RL agent's attention generalizes to unseen states. The unseen states are late-game states obtained from human experts' data [81] which RL agents have not encountered. We refer to this set as the unseen image set (100 images per game). Again, Freeway is excluded since the agent is nearly optimal.



(a) (b)

Figure 5: Human versus RL attention in states that RL agents have not seen. (a) How attention similarities change in unseen states compared to seen states in different games. (b) Unseen states for Montezuma's Revenge (top), the RL agent's attention is very different from human attention due to a new object on the left. For Seaquest (bottom), they are similar.

Fig. 5a shows the results. For Frostbite and Montezuma's Revenge, the similarities drop significantly, mostly due to new objects that the agents have never encountered. Fig. 5b shows an example for Montezuma's Revenge. The agents attended to the ladder, a previously seen object, but failed to attend to a new enemy object on the left like the human did. For Seaquest and Ms. Pac-Man, the RL attention and human attention are more similar on unseen data. This is because there are no new objects in these unseen states – only objects move faster and appear in larger numbers. The player often encounters dangerous states that are close to failure, similar to the ones in Section 4.3. Therefore we observe a similar increase in similarity for Ms.Pac-Man and Seaquest. Breakout is an interesting case. The CC value drops significantly due to unseen spatial layouts of the objects. The KL does not change much because there are no new objects so the agent maintains its attention on the ball and the paddle. More examples are in Appendix 5.

The results suggest that the obstacle to achieving expert human-level performance for certain games is first a perception challenge – the agents need to learn to recognize, attend to, and then learn to act upon new objects. This is easy for humans due to their prior knowledge but challenging for RL agents [37, 70, 16]. For the other games, it is a decision-learning problem – the agent's attention is fairly generalizable and it needs to learn a good policy for challenging states.

4.5 Other Atari agents

The above analyses were done for a particular RL algorithm–PPO. Next we apply our method to other RL algorithms, including C51 [3], Rainbow [26], DQN [43], and A2C [42], as well as two evolutionary algorithms, GA [66] and ES [58]. We use trained models from Dopamine [9] and Atari Model Zoo [67] in which each algorithm has 3-5 trained models. There is a strong positive correlation between model performance (in terms of the game score, averaged over 50 episodes each) and similarity measurement (in terms of CC with human attention on the standard image set). The average correlation coefficient for five games is 0.631 (excluding Montezuma since most algorithms have zero scores). The results for individual games can be found in Appendix 6. This result suggests that the positive correlation between model performance and similarity with human attention generalizes to other deep RL algorithms, hence we expect the rest of the analyses we performed with PPO, such as failure diagnosis, can be done for these algorithms as well.

5 Discussion

We provide visual explanations for deep RL agents using human experts' attention as a reference. We have discussed how RL attention develops and becomes more human-like during training, and how varying the discount factor affects learned attention. We show that human attention can be useful in diagnosing an RL agent's failures. We have also identified further challenges in closing the performance gap between human experts and RL agents. Our analysis is restricted to saliency map comparisons, but other approaches are possible for measuring the similarity of representations learned by different RL agents [73]. Our human attention models, all compiled datasets, and tools for comparing RL attention with human attention are made available for future research in this direction.

For researchers who are interested in RL algorithms, we have gained at least three important insights. First, since the task performance and similarity to human attention are highly correlated, one could use human attention as prior knowledge to guide the learning process of RL agents, e.g., by encouraging them to attend to the correct objects early in the learning process. This could be especially helpful for games like Seaquest, in which the agent has not learned to focus on the right object after 5120k time steps (Fig. 2). Two human studies using Atari games suggested that prior knowledge, such as perceptual prior, is why humans learn faster and better in these games [70, 16]. Then the next question is how to incorporate human attention into DNNs which has been studied in several computer vision tasks [54]. Our results indicate that RL agents first learn to attend to a few important objects like humans, then learn to attend to other objects. Therefore a desirable loss function would encourage the agents to focus on the regions that the human attended to, but would not penalize the agents for attending to more regions eventually [59].

Secondly, our results provide visual explanations for the agents' performance when varying the discount factors and highlight the importance of choosing proper planning horizons with appropriate discount factors. Recent works confirm this by showing that it is beneficial to have an adaptive discount factor [2] or multiple discount factors [19].

At last, failure analysis could identify states where agent's attention drastically differs from expert humans. These are the states that may need human intervention/correction in a human-in-the-loop RL paradigm. By publishing our human gaze prediction model researchers can diagnose their algorithms with the states of their interest.

For researchers who are interested in using RL as models for cognition, it is perhaps both surprising and encouraging to see that RL agents trained from scratch with only images and reward signals can develop attention maps that are similar to humans, especially when considering that they have very little prior knowledge. This result is similar to previous research that shows CNNs trained from image classification tasks can learn features that are similar to the ones in the visual cortex [17, 74, 75]. Consistent with previous findings, we show that model task performance and feature similarity are highly correlated [75]. Our results are complementary to the recent findings using human brain imaging data when playing Atari games [11], suggesting that deep RLs can learn biologically plausible representations and can be used as models for human gaze, decision, and brain activities.

In this work, for experiments that we can use either RL or human data, we focused on the explainable RL issue and performed the analyses using RL data. If the focus is on using RL to model human decisions, one can take our approach with human data [11]. We have performed the analyses on the effects of learning and discount factors using human states and eye fixation data and confirmed that using human data does not affect our main conclusions (Appendix 7).

Multiple factors are important for interpreting our results and could explain the remaining differences between human attention and RL attention. The first one is the overt vs. covert attention issue discussed earlier. Humans store information in memory and do not need to constantly move their eyes to attend to all task-relevant objects. To complete the human attention map, in addition to gaze data, one will need to retrieve human covert attention from brain activity data, a technique that became possible recently [38, 11]. Another factor is human intrinsic reward. Humans are likely to have internal reward functions that are different from the ones provided by the game environment, and reward is known to affect attention [57, 38]. Hypothetically specific algorithmic or network architecture designs that capture important cognitive aspects of human decision-making could lead to more similar saliency maps.

A closely related research direction compares a human player's *policy* with an RL agent's learned policy [44] which could further allow us to better understand the similarities and differences between humans and RL agents. However, as we have shown here, the difference in decisions could be due to perception, which needs to be considered while comparing policies. Our approach lays the groundwork for future research in this direction.

Ethics statement Our work studies how attentional mechanisms of humans and decision-making reinforcement learning agents are similar or different. The hope is that our work will help us better understand the differences and similarities between humans and AI. Comparing humans with AIs allows us to better understand their strengths and limitations. The study involves collecting and modeling human eye-tracking data, we removed all personally identifiable information such that the private information of human subjects is protected.

References

[1] Alnour Alharin, Thanh-Nam Doan, and Mina Sartipi. Reinforcement learning interpretation methods: A survey. *IEEE Access*, 8:171058–171077, 2020.

[2] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *International Conference on Machine Learning*, pages 507–517. PMLR, 2020.

[3] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[4] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[5] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

[6] Ali Borji and Laurent Itti. Defending yarbus: Eye movements reveal observers' task. *Journal of Vision*, 14(3):29–29, 2014.

[7] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019.

[8] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where should saliency models look next? In *European Conference on Computer Vision (ECCV)*, pages 809–824. Springer, 2016.

[9] Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G. Bellemare. Dopamine: A Research Framework for Deep Reinforcement Learning. 2018.

[10] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[11] Logan Cross, Jeff Cockburn, Yisong Yue, and John P O'Doherty. Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. *Neuron*, 2020.

[12] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6967–6976, 2017.

[13] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

[14] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[15] Gabriel Diaz, Joseph Cooper, Constantin Rothkopf, and Mary Hayhoe. Saccades to future ball location reveal memory-based prediction in a virtual-reality interception task. *Journal of Vision*, 13(1):20–20, 2013.

[16] Rachit Dubey, Pulkit Agrawal, Deepak Pathak, Thomas L Griffiths, and Alexei A Efros. Investigating human priors for playing video games. *arXiv preprint arXiv:1802.10217*, 2018.

[17] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

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

[19] William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.

[20] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.

[21] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. In *International Conference on Machine Learning*, pages 1792–1801, 2018.

[22] Christopher Grimm, Dilip Arumugam, Siddharth Karamcheti, David Abel, Lawson LS Wong, and Michael L Littman. Modeling latent attention within neural networks. *arXiv preprint arXiv:1706.00536*, 2017.

[23] Piyush Gupta, Nikaash Puri, Sukriti Verma, Sameer Singh, Dhruv Kayastha, Shripad Deshmukh, and Balaji Krishnamurthy. Explain your move: Understanding agent actions using focused feature saliency. *arXiv preprint arXiv:1912.12191*, 2019.

[24] Sen He, Hamed R Tavakoli, Ali Borji, Yang Mi, and Nicolas Pugeault. Understanding and visualizing deep visual saliency models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10206–10215, 2019.

[25] Sen He, Hamed R Tavakoli, Ali Borji, and Nicolas Pugeault. Human attention in image captioning: Dataset and analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8529–8538, 2019.

[26] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[27] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, page 106685, 2020.

[28] Ashley Hill, Antonin Raffin, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, Rene Traore, Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, and Yuhuai Wu. Stable baselines. `https://github.com/hill-a/stable-baselines`, 2018.

[29] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[30] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998.

[31] Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, and Katia Sycara. Transparency and explanation in deep reinforcement learning neural networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 144–150, 2018.

[32] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In *Advances in Neural Information Processing Systems*, pages 6628–6637, 2018.

[33] Theo Jaunet, Romain Vuillemot, and Christian Wolf. Drlviz: Understanding decisions and memory in deep reinforcement learning. *arXiv preprint arXiv:1909.02982*, 2019.

[34] Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1181–1189, 2015.

[35] Ho-Taek Joo and Kyung-Joong Kim. Visualization of deep reinforcement learning using grad-cam: How ai plays atari games? In *2019 IEEE Conference on Games (CoG)*, pages 1–2. IEEE, 2019.

[36] Qiuxia Lai, Wenguan Wang, Salman Khan, Jianbing Shen, Hanqiu Sun, and Ling Shao. Human\textit {vs} machine attention in neural networks: A comparative study. *arXiv preprint arXiv:1906.08764*, 2019.

[37] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

[38] Yuan Chang Leong, Angela Radulescu, Reka Daniel, Vivian DeWoskin, and Yael Niv. Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, 93(2):451–463, 2017.

[39] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference on Computer Vision (ECCV)*, pages 619–635, 2018.

[40] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

[41] Josh Merel, Diego Aldarondo, Jesse Marshall, Yuval Tassa, Greg Wayne, and Bence Ölveczky. Deep neuroethology of a virtual rodent. *arXiv preprint arXiv:1911.09451*, 2019.

[42] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

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

[44] Steven Moore and John C Stamper. Exploring expertise through visualizing agent policies and human strategies in open-ended games. In *EDM (Workshops)*, pages 30–37, 2019.

[45] Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. Towards interpretable reinforcement learning using attention augmented agents. In *Advances in Neural Information Processing Systems*, pages 12329–12338, 2019.

[46] Sajad Mousavi, Michael Schukat, Enda Howley, Ali Borji, and Nasser Mozayani. Learning to predict where to look in interactive environments using deep recurrent q-learning. *arXiv preprint arXiv:1612.05753*, 2016.

[47] Satya M Muddamsetty, Mohammad NS Jahromi, Andreea E Ciontos, Laura M Fenoy, and Thomas B Moeslund. Introducing and assessing the explainable ai (xai) method: Sidu. *arXiv preprint arXiv:2101.10710*, 2021.

[48] Tam V Nguyen, Qi Zhao, and Shuicheng Yan. Attentive systems: A survey. *International Journal of Computer Vision*, 126(1):86–110, 2018.

[49] Dmitry Nikulin, Anastasia Ianina, Vladimir Aliev, and Sergey Nikolenko. Free-lunch saliency via attention in atari agents. *arXiv preprint arXiv:1908.02511*, 2019.

[50] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[51] Michael I Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1):3–25, 1980.

[52] Erika Puiutta and Eric MSP Veith. Explainable reinforcement learning: A survey. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 77–95. Springer, 2020.

[53] Nikaash Puri, Sukriti Verma, Piyush Gupta, Dhruv Kayastha, Shripad Deshmukh, Balaji Krishnamurthy, and Sameer Singh. Explain your move: Understanding agent actions using specific and relevant feature attribution. In *International Conference on Learning Representations*, 2019.

[54] LAI Qiuxia, Salman Khan, Yongwei Nie, Sun Hanqiu, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 2020.

[55] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[56] DA Robinson. The mechanics of human saccadic eye movement. *The Journal of physiology*, 174(2):245–264, 1964.

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

[58] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[59] Akanksha Saran, Ruohan Zhang, Elaine Schaertl Short, and Scott Niekum. Efficiently guiding imitation learning algorithms with human gaze. *arXiv preprint arXiv:2002.12500*, 2020.

[60] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[61] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[62] Wenjie Shi, Zhuoyuan Wang, Shiji Song, and Gao Huang. Self-supervised discovering of causal features: Towards interpretable reinforcement learning. *arXiv preprint arXiv:2003.07069*, 2020.

[63] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

[64] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[65] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[66] Felipe Petroski Such, Vashisht Madhavan, Edoardo Conti, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. *arXiv preprint arXiv:1712.06567*, 2017.

[67] Felipe Petroski Such, Vashisht Madhavan, Rosanne Liu, Rui Wang, Pablo Samuel Castro, Yulun Li, Jiale Zhi, Ludwig Schubert, Marc G Bellemare, Jeff Clune, et al. An atari model zoo for analyzing, visualizing, and comparing deep reinforcement learning agents. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3260–3267. AAAI Press, 2019.

[68] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017.

[69] Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2487–2496, 2017.

[70] Pedro A Tsividis, Thomas Pouncy, Jacqueline L Xu, Joshua B Tenenbaum, and Samuel J Gershman. Human learning in atari. 2017.

[71] Junpeng Wang, Liang Gou, Han-Wei Shen, and Hao Yang. Dqnviz: A visual analytics approach to understand deep q-networks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):288–298, 2018.

[72] Laurens Weitkamp, Elise van der Pol, and Zeynep Akata. Visual rationalizations in deep reinforcement learning for atari games. In *Benelux Conference on Artificial Intelligence*, pages 151–165. Springer, 2018.

[73] Erik Wijmans, Julian Straub, Dhruv Batra, Irfan Essa, Judy Hoffman, and Ari Morcos. Analyzing visual representations in embodied navigation tasks. *arXiv preprint arXiv:2003.05993*, 2020.

[74] Daniel L Yamins, Ha Hong, Charles Cadieu, and James J DiCarlo. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. In *Advances in Neural Information Processing Systems*, pages 3093–3101, 2013.

[75] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[76] Zhao Yang, Song Bai, Li Zhang, and Philip HS Torr. Learn to interpret atari agents. *arXiv preprint arXiv:1812.11276*, 2018.

[77] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.

[78] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[79] Ruohan Zhang, Zhuode Liu, Luxin Zhang, Jake A Whritner, Karl S Muller, Mary M Hayhoe, and Dana H Ballard. Agil: Learning attention from human for visuomotor tasks. In *European Conference on Computer Vision (ECCV)*, pages 692–707. Springer, 2018.

[80] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJCAI: proceedings of the conference*, volume 2020, page 4951. NIH Public Access, 2020.

[81] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl S Muller, Jake A Whritner, Luxin Zhang, Mary M Hayhoe, and Dana H Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, 2020.

[82] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[83] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see Sec. 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see Ethics Statement.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] The paper does not include theoretical results
   (b) Did you include complete proofs of all theoretical results? [N/A] The paper does not include theoretical results
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see Sec. 3 and Appendix 1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please see Sec. 3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Please see ethics statement.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] The paper does not involve human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] The paper does not involve human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] The paper does not involve human subjects.