Explainable AI in Atari: Visualizing the decision-making process of Deep Reinforcement Learning using Grad-CAM

Introduction

Deep Reinforcement Learning (DRL) models have demonstrated decent performance in training agents to perform complex tasks, one notable success being playing Atari games. Using Deep Q-Networks (DQN), the agents in [1] outperformed expert human players. However, the decision-making process of such DRL models remains opaque, due to the black-box nature and complexity of the architecture. The lack of transparency makes it difficult to fully understand how the models arrive at each decision, which limits the potential for further optimization and development in real-world applications.

In recent years, the advancements in Explainable AI (XAI) have offered solutions to address the lack of interpretability in deep models. Providing visual explanations enables the identification of features or regions that influence the model's decisions. Grad-CAM is one such effective technique specifically for explaining neural network models through gradient-based localization [2]. While XAI techniques have been applied to explain many Computer Vision and Natural Language Processing models, there's not extensive research in explaining DRL models for gaming environments such as Atari. Existing research mostly focuses on static and individual frame explanations, which might not fully capture the convoluted and dynamic decision-making process [3]. This project seeks to bridge the gap by providing video-based explanations that will help with understanding the decision-making process of the agents, therefore enhancing trust and improving performance of the models, which would ultimately pave the way for developing real-world applications.

Gap in Existing Research

The complex nature of the decision-making strategies of agents in DRL models offers hurdles for users wanting to understand the predicted action for solving tasks in Atari games, raising concerns about reliability and real-world application risks. Consequently, these models require interpretability and transparency. Despite XAI emerging as a transformative approach for providing insights to the decision-making process of AI models, due to the complexity of the DRL models the existing research still lacks effective explanations methods [4]. Very few researchers have tried to utilize the existing XAI techniques to explain DRL models for Atari domain. Paper [3] and [5] utilized Grad-CAM to visualize the behaviors of the AI players in Atari games. However, paper [3] redesigned the DeepMind's A3C model and trained the agents for various Atari Games to make the model compatible with Grad-CAM which limits the use of Grad-CAM to the already existing pre-trained DRL model for Atari games. Paper [6] proposed to transform the raw input pixels into symbolic, interpretable features and then train a surrogate model to explain the behavior of the DRL agents in the Atari domain. Although their proposed surrogate model learned to accurately approximate the behavior of the Atari games, this approach underperforms in complex game scenarios. Paper [7] proposed an attention-based mask generation process to policy and state values in Atari 2600 games but their method requires to set optimal value of threshold for the pseudo-maskatt in Mask-attention loss. Overall, existing research on explaining Atari games has notable limitations, predominantly focusing on the A3C method, which restricts generalizability across different DRL techniques.

Proposed Solution

In this project, we propose to use Grad-CAM to explain the behavior of RL agents in the domain of Atari games. The following steps outlines the process:

1. First, we will run a simulation of 100 seconds from selected Atari games and capture 100 frames per second using OpenCV Library for Python [8].
2. Each frame is passed as an input to Grad-CAM, showing the state of each agent.
3. During the forward pass of the input images, feature maps are created through activations of the convolutional layers.
4. Then using a task specific network, a score is generated for each of the actions.
5. The gradient is computed by taking the derivative of the action score with respect to the feature map. It is then backpropagated to the convolutional layer for identifying the contributing features.
6. This is followed by calculating global average pooling of the gradients and neuron importance weights are obtained.
7. The result is then multiplied by the feature map and an activation function ELU is applied to produce a gradient-based class localization map.
8. This localization map is then overlayed on top of the input state and a high quality heatmap is generated. This heatmap indicates the regions that influence the agents to take a particular action in a certain state.
9. We repeat steps 2-8 for each of the frames and finally, all the frames with heatmap overlay are compiled into a video explanation of the decision-making process.

Success Metrics

For evaluation, we will visually assess the effectiveness and faithfulness of the explanation, identifying the success or failure of the explainer in pinpointing the features that contribute to the agent’s decision towards an action in a specific state.

To evaluate the generalizability of the proposed method, we will conduct experiments on 3 Atari games. We will compare the generated heatmaps for each of the game environments and visually assess the effectiveness. We will also use other XAI techniques, including SmoothGrad, PatternNet, Layer-wise Relevance Propagation (LRP), LIME for the purpose of comparison.

Since Grad-CAM has the ability to locally focus on the most salient part of the input images, we anticipate that it will show the best performance compared to the other methods mentioned. With the outcome of this project, we seek to demonstrate the effectiveness of video-based explanations for gaming environment in providing insights on the behavior of a complex DRL model, which can hopefully inspire the potential of developing video-based explanations in real-world situations.

References

1. Mnih, V. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.
2. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. International journal of computer vision, 128, 336-359.
3. Joo, H. T., & Kim, K. J. (2019, August). Visualization of deep reinforcement learning using grad-CAM: how AI plays atari games?. In 2019 IEEE conference on games (CoG) (pp. 1-2). IEEE.
4. Andrulis, J., Meyer, O., Schott, G., Weinbach, S., & Gruhn, V. (2020). Domain-Level Explainability--A Challenge for Creating Trust in Superhuman AI Strategies. arXiv preprint arXiv:2011.06665.
5. Weitkamp, L., van der Pol, E., & Akata, Z. (2019). Visual rationalizations in deep reinforcement learning for atari games. In Artificial Intelligence: 30th Benelux Conference, BNAIC 2018,‘s-Hertogenbosch, The Netherlands, November 8–9, 2018, Revised Selected Papers 30 (pp. 151-165). Springer International Publishing.
6. Sieusahai, A., & Guzdial, M. (2021, October). Explaining deep reinforcement learning agents in the atari domain through a surrogate model. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (Vol. 17, No. 1, pp. 82-90).
7. Itaya, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H., & Sugiura, K. (2024). Mask-Attention A3C: Visual Explanation of Action-State Value in Deep Reinforcement Learning. IEEE Access.
8. Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.