**CE888 – Assignment 2**

Osama Rahman

School of Computer Science and Electronic Engineering

University of Essex

England

orahmab@essex.ac.uk

[[1]](#footnote-1)

Reinforcement Learning and Interpretability

*Abstract*— As artificial intelligence becomes more prominent in day to day activities from self-driving cars to medical diagnosis, understanding why an AI agent came to a decision is becoming increasingly important. With artificial intelligence, and machine learning models, many systems operate within what is referred to as a *black box.* On the outside we can see what is given as input to a system, and what the system gives as output. However, the stages between input and output are often difficult to comprehend for both experts and non-experts alike. It is important to understand why a conclusion has been made or an action is taken when there are real world consequences, be they financial or to do with human health and safety. Due to this ambiguity and the necessity to understand actions of an AI agent, it is critical for research to be conducted in the area of AI interpretability. This report hopes to enlighten the reader on this topic by providing a background, and an explanation of studies carried out to interpret the actions of an Atari game playing reinforcement learning agent.

# Introduction

This report hopes to provide an understanding of why it is important to be able to interpret the actions of artificially intelligent agents, and machine learning models. The report will also showcase studies carried out to aid in the interpretation of the Atari game playing reinforcement learning agent.

With the rise in popularity of artificial intelligence and machine learning models across many industries, it has become ever important to be able to understand and provide reasonings for why and how an agent or model has come to a conclusion. This is especially important with regards to financial or health matters. With regard to more recent matters it is important to understand why a specific advertisement may have been shown to a user on a social media website. For example, if a machine learning model was able to identify people with alcoholic tendencies, it may begin to show them advertisements for alcohol. While this will drive sales for the advertiser, it is, if anything, immoral. As such, being able to peak into the blackbox is a necessary step to take as AI and

machine learning becomes more powerful, and prominent in our lives. Studies have shown that using such techniques is an efficient way to enhance sales [1]. As such, being able to interpret actions and conclusions will allow the finetuning of such models, and also aid in preventing their misuse.

Reinforcement learning agents are able to learn their environment and actions to take on a rewards-based learning procedure. For example, if an agent takes a good action they will be given some reward. If they take a better action, they will be given a better reward. Should the agent take an action it is not supposed to, or an action the programmer does not want it to take, the agent will be given a negative reward. Interpreting the actions of a game playing agent may seem like an easy task on the outside, however the reasons for an action may be more complex than expected. For example, when playing a game an agent may seem to be attacking enemies to increase its score, however, it may be doing so to increase its playtime, which may result in larger rewards. This thought process holds true for other implementations of artificial intelligence and machine learning models in other industries too. Being able to understand actions and conclusions of agents and models increases our trust in such systems and will enable their improvements.

Research was conducted using 3000 images from 6 games played by the reinforcement learning agent, and actions corresponding to each image. Two main methods were used in conducting research, optical flow analysis, and LIME, a tool which enables insight into a model. Further discussion of these tools and how they were used can be found in the methodology and experiments sections.

# Background

This section will discuss core background knowledge relevant for this research project, and area of interest. There are four main components readers should have a high level understanding of; reinforcement learning; interpretability; optical flow; and LIME.

## Reinforcement Learning

Reinforcement learning in an unsupervised learning technique employed within the realm of artificial intelligence. In essence, the goal of this learning technique is to teach an AI agent to accomplish its goals through the use of positive and negative rewards. Through the acquisition of positive and negative rewards, an agent is able to see which actions it can take to maximise potential rewards. As such, it generates policies which it will follow to gain a reward. The policy which returns the highest reward is referred to as the optimal policy [2].

In most situations an agent will be placed in an environment with little to no knowledge of what to do, and of the environment itself. It will essentially learn through the process of trial and error, learning through positive and negative reinforcement. For example, an agent playing the Atari game Breakout will receive positive reinforcement for hitting a brick with the ball, and negative reinforcement for dropping the ball. To enhance the ability of the agent a larger positive reward could be given if the agent hits multiple bricks, without the ball returning to the paddle.

Reinforcement learning can be split into two subcategories of learning, passive learning and active learning. Within passive learning, the agent has a fixed policy. In a given state the agent will always execute the same action, with the goal of learning how good a policy is [2]. The agent does not know the reward for a policy until it is executed, nor the probability of reaching a future state from its current state. As such this method encourages the exploration of the agents environment, for it to learn the utility associated with each state encountered. The utility () can be calculated by summing the reward at each state () multiplied by some discount factor ( (1).

Active reinforcement learning requires the agent to first learn an entire model, with outcome probabilities for each action. Once this has been completed, the agent has a choice of actions it is able to take. As such, it must learn the optimal utility function to decide on which action it should take at a given step. It is also able to see one step ahead, to increase the utility [2]. For example, say an agent has two choices. Choice A, with a reward of 10, and choice B with a reward of 8. It may very well choose choice B if the reward in the state after is greater than that of choice A. This technique encourages exploration by choosing paths which enable a larger end reward, rather than a large immediate reward.

### A3C Algorithm

This project employs the use of an Atari game playing agent for data collection, and for interpretation analysis. The specific agent used is based on the A3C (

1. [↑](#footnote-ref-1)