Human-level but not human-like: Deep Reinforcement Learning in the dark

Rachit Dubey
Princeton University, PRINCETON, New Jersey, United States

Pulkit Agrawal
UC Berkeley, Berkeley, California, United States

Deepak Pathak
UC Berkeley, Berkeley, California, United States

Alyosha Efros
UC-Berkeley, Berkeley, California, United States

Tom Griffiths
University of California, Berkeley, Berkeley, California, United States

Abstract

Deep reinforcement learning (RL) algorithms have recently achieved impressive results on a range of video games, learning to play them at or beyond a human level just from raw pixel inputs. However, do they leverage visual information in the same manner as humans do? Our investigations suggest that they do not: given a static game, we find that a state-of-the-art deep RL algorithm solves that game faster without visual input (only the agent location was provided to the algorithm). We posit that this is because deep RL attacks each problem tabula rasa, i.e. without any prior knowledge, as also suggested by other recent work. We further propose that in certain settings, an agent is better off having no visual input compared to having no visual priors. To demonstrate this, we conduct an experiment with human participants and find that people solve a game that hid all visual input (except agent location) much faster than a game that prevented the use of various visual priors. These results highlight the importance of prior knowledge and provide a compelling demonstration of how the lack of prior knowledge leads to deep RL algorithms approaching a problem very differently from humans.