Abstract

Artificial Intelligence (AI) is often suggested to replace repetitive tasks, automating and managing administrative load or even assisting healthcare and scientific studies. To have reliable machine learning (ML) models, we need to fully understand their internal operations. In this thesis, we examine the scientific understanding of AI on three levels: the input structure through pre-processing, training by optimisation symmetry and post-training interventions.

First we examine the image classification task, albeit from a different perspective. The study we introduce focuses on understanding how preprocessing interactions with a model can reveal information about the model itself. Using a lens of structure on the data presented to the model we analyse the impact on its training. We demonstrate that proper data pre-processing impacts the training of one-dimensional convolutional neural networks and provide insight into the sequencing the input for vision transformer models.

Second, we investigate the optimisation of Reinforcement Learning (RL) models. Using the perspective of a context-dependent dynamic, we test the underlying assumption related to the expected values in the formulation of the Bellman equation which governs the optimisation process. Our research places RL models in ergodically broken and path-dependent contexts, where the optimal policy is not determined by ensemble averages, but by growth rates. We demonstrate that traditional models fail to find optimal policies, and offer an alternative method to accurately adjust the training of agents.

Finally, we turn to post-training interventions of Large Language Models (LLM). By extracting internal feature representations of the model we can align their behaviour. We study these operations in light of the transformation they pose on the output, considering the impact of linearity and affinity properties. We show that, to some extent, these interventions are linear transformations, but are heavily dependent on the model itself. We further demonstrate the impact on the semantic quality of models under such transformations.