Abstract

In this dissertation, we investigate methods for efficiently learning human-like policies under limited data availability. Here, a policy refers to a probability distribution over actions conditioned on states. We address both group-level and individual-level human-like behavior and also explore potential applications of such policies. One such application is human win rate prediction, which aims to estimate the probability that a human player would win the game if playing from a given position. We examine how human-like policies can be effectively applied in human win rate prediction. The primary target game is Shogi, a representative two-player perfect-information game for which large-scale game records are more difficult to obtain than in Chess. For broader comparison and validation, Go and Chess are also included as additional target domains.

First, we investigate the creation of human-like policies at the group-level perspective. We propose a Blend model that mixes the outputs of a supervised learning model and a reinforcement learning model so that human-like moves can be predicted even when human game records are limited. Experiments conducted on Chess, Go, and Shogi demonstrate that the proposed Maia-S, trained with roughly 1/100 of the 12 million-game dataset used in prior work (Maia), can achieve improved human move matching accuracies by combining the supervised model (Maia-S) with an AlphaZero-like reinforcement learning policy. Across the three games, the Blend model improves move-matching accuracy by 0.3–2.9 percentage points for intermediate players and 2.0–6.0 percentage points for advanced players. Additionally, in Chess, the Blend model improves move-matching accuracies by 0.2–1.1 percentage points even when using the original, large-data Maia model. In Go, the Blend model outperforms KL-regularized search, a method that improves move-matching accuracies by searching under calibrated supervised policies. While KL-regularized search itself improves upon Maia, the Blend model yields even greater gains. In Shogi, an analysis of positions where the Blend model is particularly effective reveals that, in positions where supervised models tend to suggest inferior moves, the Blend model more often aligns with human decisions. This indicates that blending compensates for the weaknesses of both supervised and reinforcement-learning-based policies, as originally expected.

Next, we examine how to realize human-like policies at the individual level. We address the challenge that many players have only a small number of game records available. To address this limitation, we propose the Similarity-Guided Fine-Tuning (SGFT) model, which leverages both the target player's games and games similar to them. Two methods for computing similarity are introduced: a feature-based approach using handcrafted game features, and an embedding-based approach using player-behavior embeddings. A two-stage fine-tuning framework is proposed: first fine-tuning with similar games, and then fine-tuning it with the target player's own games. In addition to move-matching accuracy – the primary metric used in prior studies – this dissertation introduces a new evaluation measure based on game features, designed to quantify how well a model imitates individual playing styles. Experiments on Shogi show that SGFT models outperform one-stage fine-tuning, which corresponds to the standard fine-tuning procedure where the target player's game records and similar game records are trained jointly. Moreover, embedding-based SGFT models outperform feature-based ones. The feature-based evaluation further reveals that even Transfer Maia, a prior individual-level imitation method, does not significantly improve certain style-sensitive indicators. These results suggest that SGFT has strong potential for improving human-recognizable, style-specific characteristics.

Furthermore, we discuss the potential applications of human-like policies in playing, teaching, and commentary. A well-known issue in game AI is that AI-estimated win rates often diverge from the win rates that humans would actually achieve. Focusing on human win rate prediction, we propose an inner-product model. It computes the win rate by combining a policy with the predicted win rates of the successor positions. The proposed method surpasses existing models across multiple evaluation criteria – including result accuracy, cross-entropy, and expected calibration error – demonstrating the effectiveness of human-like policies for commentary applications in game AI.

In summary, we present an integrated framework for efficiently learning human-like policies under the practical constraint of limited data. The proposed methods are applicable across multiple games and provide a solid foundation for future research in human-like gameplay, educational AI, and commentary systems.

Keywords: human-like behavior, game AI, personalized modeling, supervised learning, reinforcement learning, fine-tuning