DouZero: Mastering DouDizhu with Self-Play Deep Reinforcement Learning

Daochen Zha†, Jingru Xie‡, Wenye Ma‡, Sheng Zhang*, Xiangru Lian‡, Xia Hu†, Ji Liu‡
†Department of Computer Science and Engineering, Texas A&M University
‡AI Platform, Kwai Inc.
*Georgia Institute of Technology

DouDizhu: The Most Popular yet Challenging Poker game in China

- **Popularity.** There are more than 800 million registered users and 40 million daily active players on the Tencent mobile platform alone for DouDizhu.
- **Competition and Collaboration.** Two Peasants need to cooperate to fight against Landlord.
- **Large State/Action Space.** There are a large number of information sets and 27,472 possible actions due to combinations of cards, where the legal actions vary significantly from turn to turn.

![Figure 1: A hand and its corresponding legal moves.](image)

Existing Algorithms are not Satisfactory

- **DQN and A3C.** The commonly used algorithms are mainly designed for simple and small action spaces. They are shown to not perform well in DouDizhu’s large and complex action space in previous work.
- **Previous Efforts in DouDizhu.** They reduce the action spaces by using either hierarchical action space or heuristics for abstraction. However, both of them rely on human knowledge and could be sub-optimal. Moreover, they are computationally expensive. DeltaDou, the previous state-of-the-art, takes two months for training.

Traditional Monte-Carlo Methods

To optimize a policy $\pi$, Monte-Carlo Methods estimate Q-value $Q(s, a)$ by iteratively executing the following procedure:

- Generate an episode using $\pi$.
- For each $(s, a)$ appeared in the episode, calculate and update $Q(s, a)$ with the return averaged over all the samples concerning $s, a$.
- For each $s$ in the episode, $\pi(s) \leftarrow Q(s, a)$.

Deep Monte-Carlo (DMC)

We enhance traditional Monte-Carlo methods with deep neural networks, action encoding, and parallel actors.

![Figure 2: Left: Cards for both states and actions are encoded into a 4 x 15 one-hot matrix, where columns correspond to the 13 ranks and the jokers, and each row corresponds to the number of cards of a specific rank or joker. Right: The Q-network of DouZero consists of an LSTM to encode historical moves and six layers of MLP with hidden dimension of 512. The network predicts a value for a given state-action pair based on the concatenated representation of action and state.](image)

The Strongest DouDizhu AI Up-to-Date

Given an algorithm A and an opponent B, we use two metrics to compare the performance of A and B.

- **WP** (Winning Percentage): The number of the games won by A divided by the total number of games.
- **ADP** (Average Difference in Points): The average difference of points scored per game between A and B.

Takeaways

- Simple Monte-Carlo (MC) methods can be made to deliver strong performance in a very hard domain.
- A reasonable experimental pipeline for DouDizhu domain with only days of training on 4 GPUs.
- An online demo at https://douzero.org/.