COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

Data Analytics at Texas A&M (DATA Lab)

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.

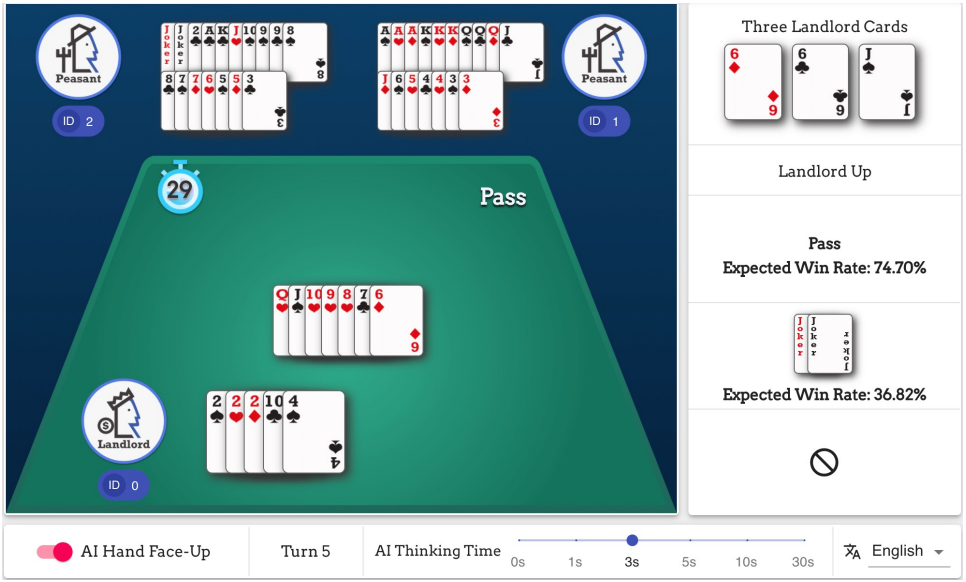
ICML 2021

Paper

Code

DouDizhu: A Popular and Challenging Chinese Poker

- ***Popularity:*** *There are more than **800 million** registered users and **40 million** daily active players on the Tencent mobile platform for DouDizhu.*
- ***Both Competition and Collaboration:*** *Two Peasants fight against Landlord.*
- ***Large State Space:*** *There are up to 10^{83} information sets with very large sizes.*
- ***Large Action Space:*** *There are **27, 472** possible actions due to combinations of cards.*



Challenge: Large and Variable Action Space

Action Type	Number of Actions
Solo	15
Pair	13
Trio	13
Trio with Solo	182
Trio with Pair	156
Chain of Solo	36
Chain of Pair	52
Chain of Trio	45
Plane with Solo	21,822
Plane with Pair	2,939
Quad with Solo	1,326
Quad with Pair	858
Bomb	13
Rocket	1
Pass	1
Total	27,472

- ***Difficulty:*** *Previous work shows that DQN and A3C can only be sightly better than random policies [1][2] in DouDizhu.*
- ***Existing Efforts:*** *CQN [1] uses decomposition but it can not beat simple rules; DeltaDou [3] abstracts the action space with heuristics but it is too slow (two months training).*

[1] You, Yang, et al. "Combinational Q-Learning for Dou Di Zhu." arXiv preprint arXiv:1901.08925 (2019).

[2] Zha, Daochen, et al. "RLCard: A Platform for Reinforcement Learning in Card Games." IJCAI. 2020.

[3] Jiang, Qiqi, et al. "DeltaDou: Expert-level Doudizhu AI through Self-play." IJCAI. 2019

Monte-Carlo (MC) Methods

- *To optimize a policy π, every-visit MC [1] can be used to estimate Q-table Q(s, a) by iteratively executing the following procedure:*

    1. Generate an episode using π.

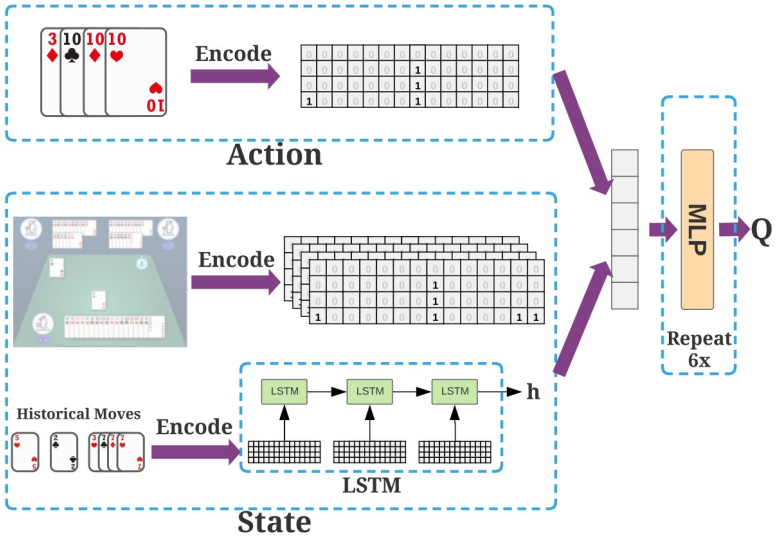
    2. For each s, a appeared in the episode, calculate and update Q(s, a) with the return averaged over all the samples concerning s, a.

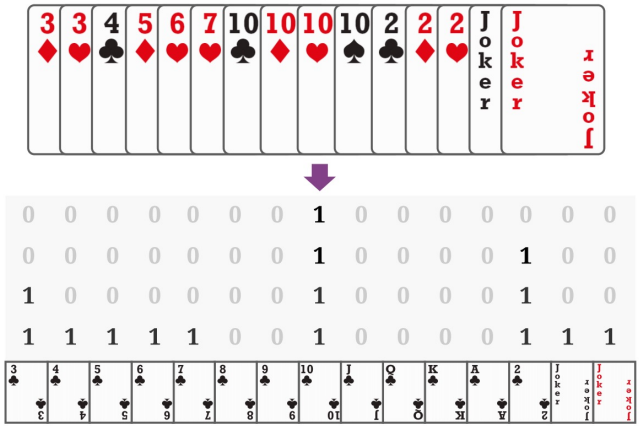
    3. For each s in the episode, π(s) ← arg maxa Q(s, a).

[1] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

Deep Monte-Carlo (DMC)

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





[1] You, Yang, et al. "Combinational Q-Learning for Dou Di Zhu." arXiv preprint arXiv:1901.08925 (2019).

[2] Zha, Daochen, et al. "RLCard: A Platform for Reinforcement Learning in Card Games." IJCAI. 2020.

DouZero Outperforms the Existing AIs

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

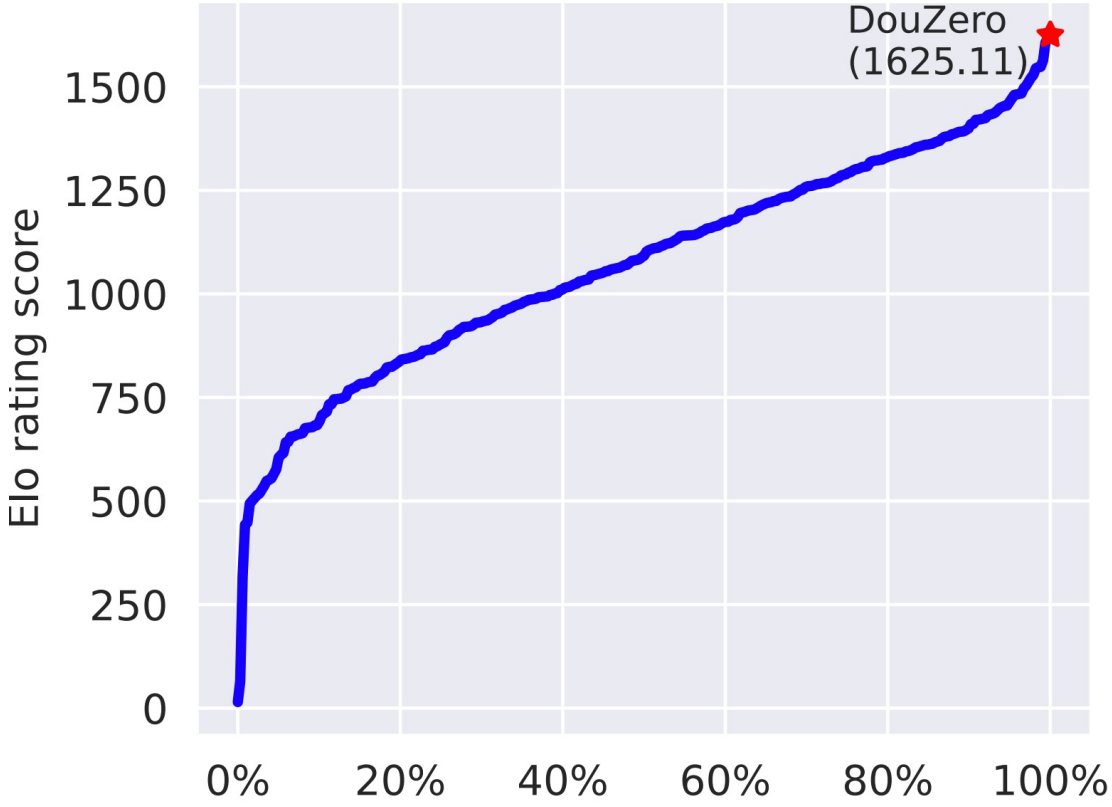
  1. ***WP (Winning Percentage):*** *The number of the game won by A divided by the total number of games.*

  2. ***ADP (Average Difference in Points):*** *The average difference of points scored per game between A and B. The base point is 1. Each bomb will double the score.*

Rank	A \ B	DouZero		DeltaDou		SL		RHCP-v2		RHCP		RLCard		CQN		Random	
		WP	ADP	WP	ADP	WP	ADP	WP	ADP	WP	ADP	WP	ADP	WP	ADP	WP	ADP
1	DouZero	-	-	**0.586**	**0.258**	**0.659**	**0.700**	**0.757**	**1.662**	**0.764**	**1.671**	**0.889**	**2.288**	**0.810**	**1.685**	**0.989**	**3.036**
2	DeltaDou	0.414	-0.258	-	-	**0.617**	**0.653**	**0.745**	**1.500**	**0.747**	**1.514**	**0.876**	**2.459**	**0.784**	**1.534**	**0.992**	**3.099**
3	SL	0.341	-0.700	0.396	-0.653	-	-	**0.611**	**0.853**	**0.632**	**0.886**	**0.813**	**1.821**	**0.694**	**1.037**	**0.976**	**2.721**
4	RHCP-v2	0.243	-1.662	0.257	-1.500	0.389	-0.853	-	-	**0.515**	**0.052**	**0.692**	**1.121**	**0.621**	**0.714**	**0.967**	**2.631**
5	RHCP	0.236	-1.671	0.253	-1.514	0.369	-0.886	0.485	-0.052	-	-	**0.682**	**1.259**	**0.603**	**0.248**	**0.941**	**2.720**
6	RLCard	0.111	-2.288	0.124	-2.459	0.187	-1.821	0.309	-1.121	0.318	-1.259	-	-	**0.522**	**0.168**	**0.943**	**2.471**
7	CQN	0.190	-1.685	0.216	-1.534	0.306	-1.037	0.379	-0.714	0.397	-0.248	0.478	-0.168	-	-	**0.889**	**1.912**
8	Random	0.011	-3.036	0.008	-3.099	0.024	-2.721	0.033	-2.631	0.059	-2.720	0.057	-2.471	0.111	-1.912	-	-

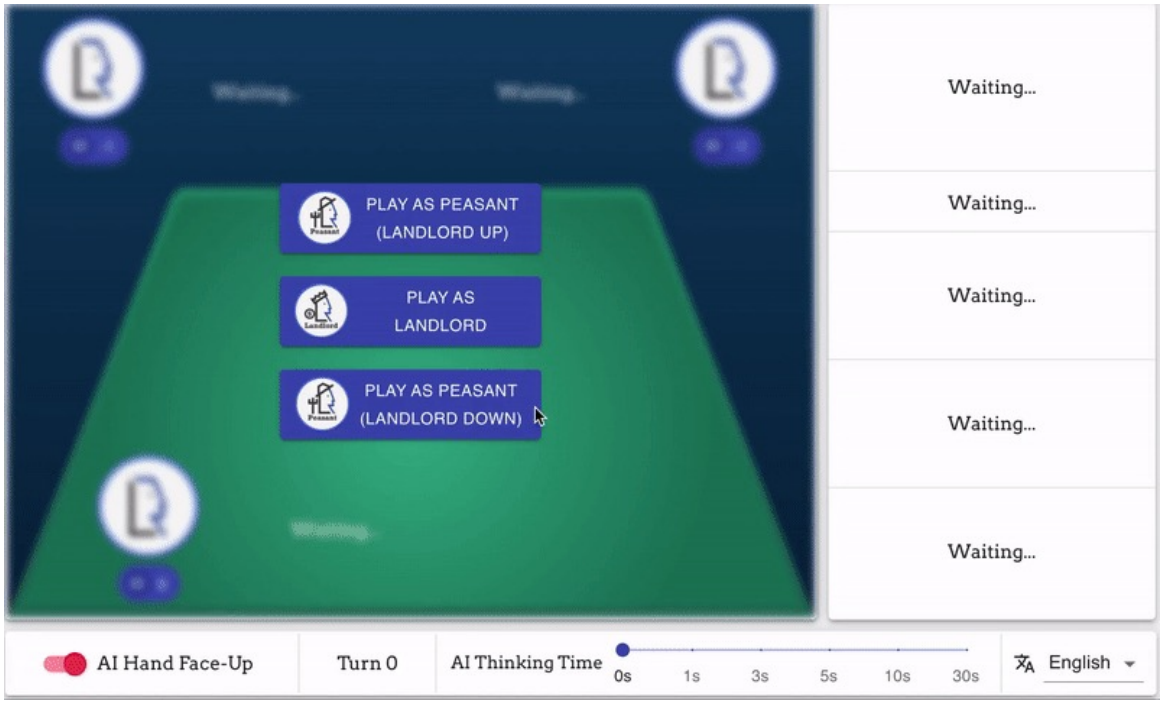
DouZero Outperforms the Existing AIs

- *DouZero ranked the first in the Botzone leaderboard among 344 AI agents.*



Open-Source Projects and Demo

- *Given the simplicity of the our method, we believe there are lots of future opportunities.*
- *We have open-sourced our code to facilitate future research.*
- *We have developed an online demo to play against the AI.*



Takeaways

- *Simple Monte-Carlo (MC) methods can be made to deliver strong performance in a hard domain, enhanced with neural networks and action encoding.*
- *A reasonable experimental pipeline for DouDizhu domain with only days of training on 4 GPUs.*
- *Open-Sourced everything (environment, code, pre-trained model, GUI demo).*
- *We hope our efforts can motivate future research of RL.*

Online Demo: https://douzero.org/

Paper: https://arxiv.org/abs/2106.06135

Code: https://github.com/kwai/DouZero

RLCard (integrated DouZero): https://rlcard.org/

Demo