Towards a Competitive 3-Player Mahjong AI using Deep Reinforcement Learning

Xiangyu Zhao xz398@cam.ac.uk

Sean B. Holden sbh11@cl.cam.ac.uk

Cambridge, United Kingdom

Published in 2022 IEEE Conference on Games (CoG)

Outline

- Introduction
- What is Mahjang?
- Difference between Normal Play and Sanma
- Data
- About AI
- AI evaluation
- Result
- Conclusion
- Future Work

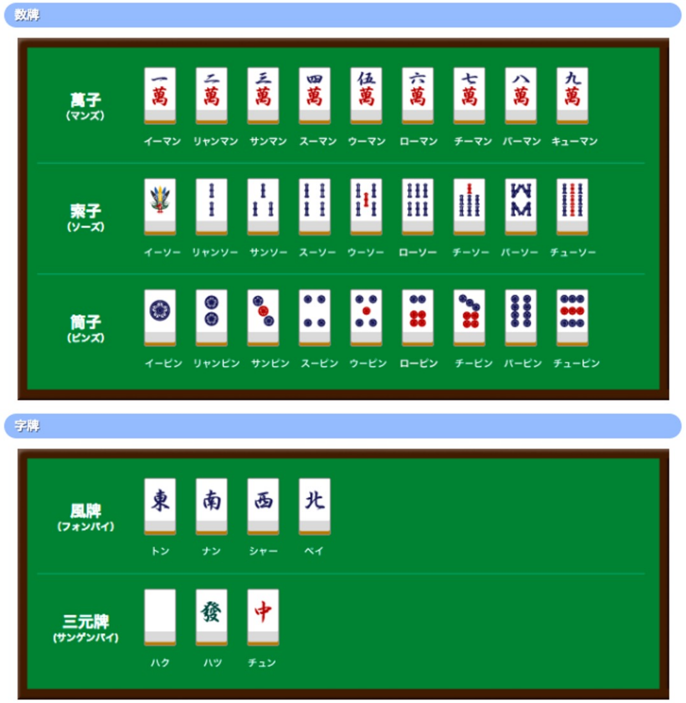
Introduction

- Mahjong is an incomplete information game that originated in China at the end of the 19th century.
    - It is played by more than 2 players and has significant hidden information;
    - It has complex playing and scoring rules;
    - It has a huge number of winning hands in various patterns, allowing flexible in-game strategy adaptation.

- Therefore, the goal of this research is to create an AI with reduced hidden information in a three-player rule called Sanma.

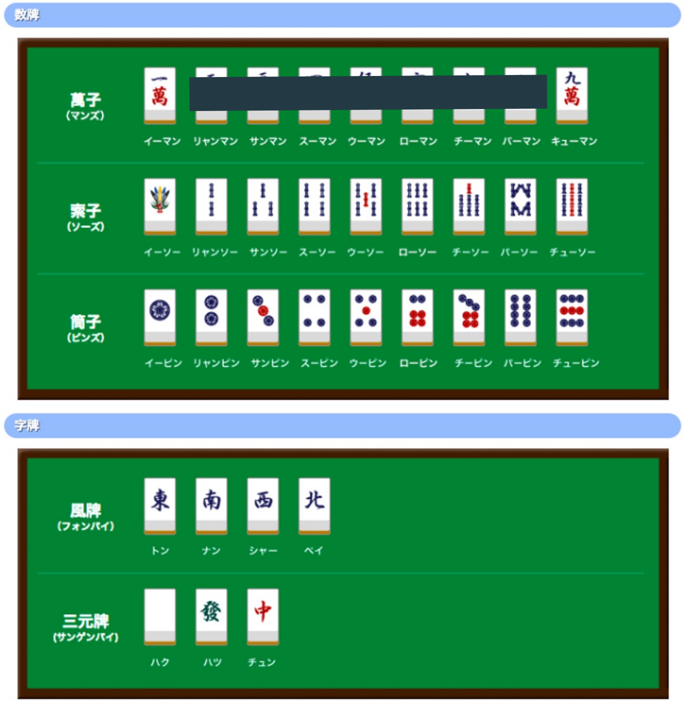
What is Mahjong?

- Mahjong is a tile game and uses card-like objects called tiles: there are three suits with 1-9 written on them, four tiles each with a direction and three special tiles.
  - The hand has 14 tiles.
  - Making an winning Hand
  - The player with the most points wins.


数牌
萬子（マンズ）
イーマン リャンマン サンマン スーマン ウーマン ローマン チーマン パーマン キューマン
索子（ソーズ）
イーソー リャンソー サンソー スーソー ウーソー ローソー チーソー パーソー チューソー
筒子（ピンズ）
イーピン リャンピン サンピン スーピン ウーピン ローピン チーピン パーピン チューピン
字牌
風牌（フォンパイ）
トン ナン シャー ペイ
三元牌（サンゲンパイ）
ハク ハツ チュン

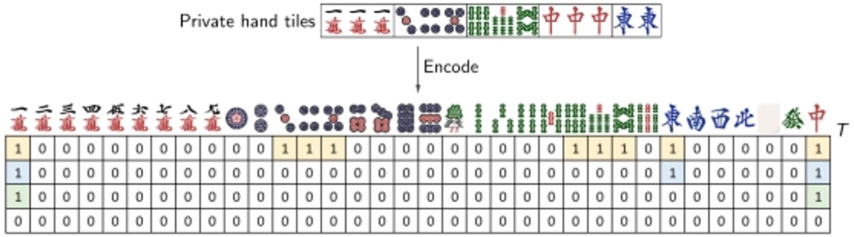
Difference between Normal Play and Sanma

- Tiles 2 to 8 of "manzu" are not used.
- The action "Chi" is removed.
- The action " Kita" is added as a simple action
- In addition, players tend to play more aggressively because sanma reduces the amount of hidden information.



Data

- The layout of the mahjong board is not standardized, so we need to encode information that can be observed by the CNN.
- The figure on the right shows the correspondence between the sequences and the tiles



🀇	🀈	🀉	🀊	🀋	🀌	🀍	🀎	🀏	🀙	🀚	🀛	🀜	🀝	🀞	🀟	🀠	🀡	🀐	🀑	🀒	🀓	🀔	🀕	🀖	🀗	🀘	東	南	西	北	🀆	發	中
1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

An example of private hand encoding using four channels

Data

・This time, a model is created and trained for each of the sanma's actions.

・A model called the Action Model is used to parameterize Meowjong's policy for each action (discard, pon, kan, kita, reach).

・The model structure uses a CNN consisting of four convolutional layers followed by a fully connected layer

About AI

- The AI created in this study is called Mewojong

- Supervised Learning
  - Players in the top 0.1% of the ranking of the mahjong app Tenho
  - Using 5,000 rounds of logs from 2020 data.

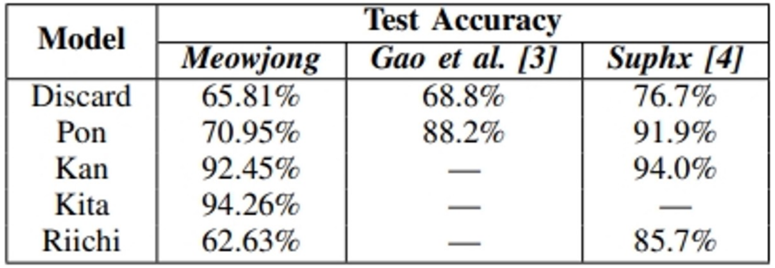
- Reinforcement Learning
  - Inherit supervised learning and let the discard model learn through 400 episodes of self-play
  - The Monte Carlo policy gradient method was used for RL learning

AI evaluation

- The table shows the testing accuracy of the model and the testing accuracy achieved in previous studies

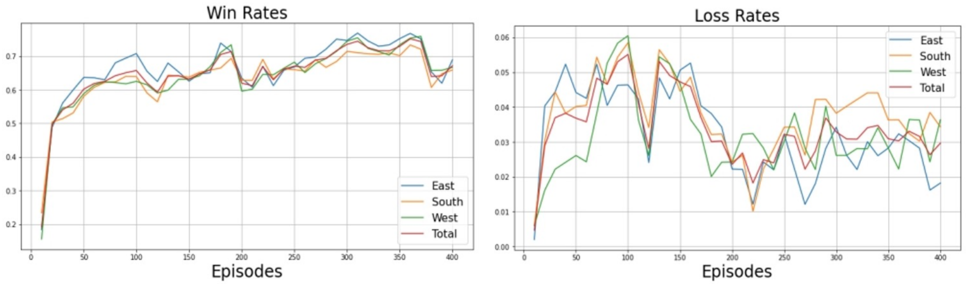
- The figure shows the win and loss rates after 500 rounds of play after every 10 episodes

Supervised AI



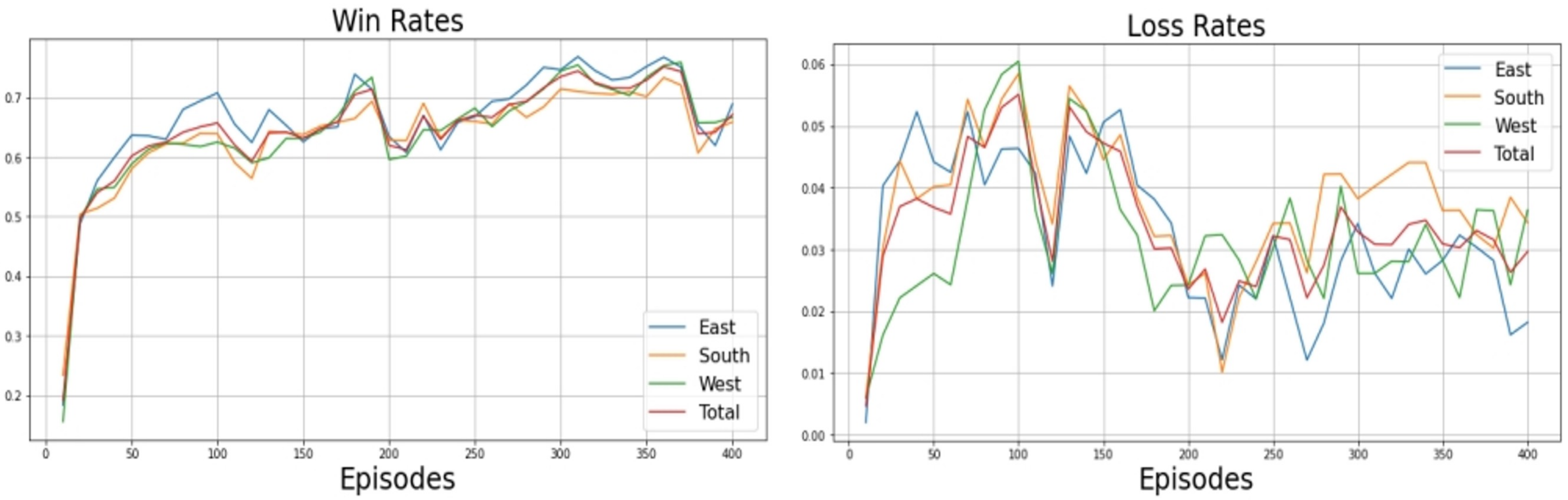
Model	Test Accuracy		
	Meowjong	*Gao et al. [3]*	*Suphx [4]*
Discard	65.81%	68.8%	76.7%
Pon	70.95%	88.2%	91.9%
Kan	92.45%	—	94.0%
Kita	94.26%	—	—
Riichi	62.63%	—	85.7%

Reinforcement AI



AI evaluation

Reinforcement AI



Win Rates

Loss Rates

Experiments

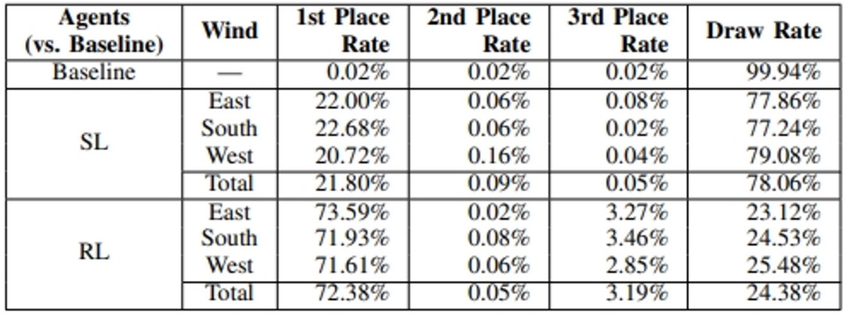
- Each played 5000 rounds per direction
  - Baseline AI only
  - Supervised AI + Baseline AI(2)
  - Reinforced AI + Baseline AI

- Each played 5000 rounds per direction
  - Supervised AI only
  - Supervised AI + Reinforced AI(2)
  - Reinforced AI + Supervised AI(2)

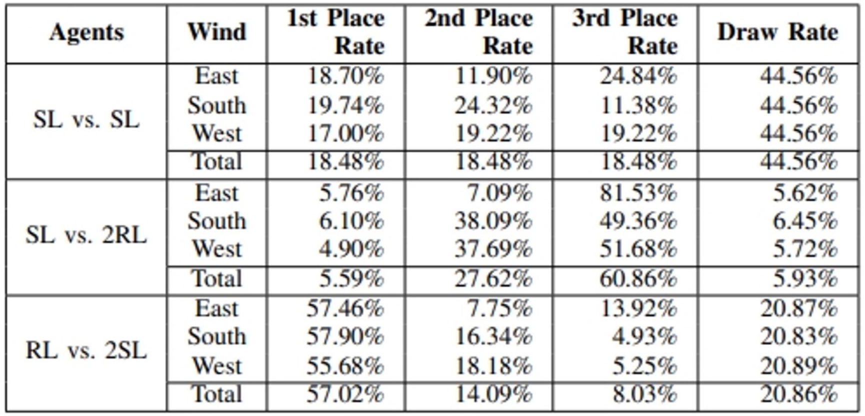
Result

- The table above shows the results of the first experiment.
  - Baseline AIs often play each other to a draw.
  - Both AIs had a higher win rate than the baseline AI

- The table below shows the results of the second experiment
  - The results show that the reinforced AI is stronger than the supervised AI.



Agents (vs. Baseline)	Wind	1st Place Rate	2nd Place Rate	3rd Place Rate	Draw Rate
Baseline	—	0.02%	0.02%	0.02%	99.94%
SL	East	22.00%	0.06%	0.08%	77.86%
	South	22.68%	0.06%	0.02%	77.24%
	West	20.72%	0.16%	0.04%	79.08%
	Total	21.80%	0.09%	0.05%	78.06%
RL	East	73.59%	0.02%	3.27%	23.12%
	South	71.93%	0.08%	3.46%	24.53%
	West	71.61%	0.06%	2.85%	25.48%
	Total	72.38%	0.05%	3.19%	24.38%



Agents	Wind	1st Place Rate	2nd Place Rate	3rd Place Rate	Draw Rate
SL vs. SL	East	18.70%	11.90%	24.84%	44.56%
	South	19.74%	24.32%	11.38%	44.56%
	West	17.00%	19.22%	19.22%	44.56%
	Total	18.48%	18.48%	18.48%	44.56%
SL vs. 2RL	East	5.76%	7.09%	81.53%	5.62%
	South	6.10%	38.09%	49.36%	6.45%
	West	4.90%	37.69%	51.68%	5.72%
	Total	5.59%	27.62%	60.86%	5.93%
RL vs. 2SL	East	57.46%	7.75%	13.92%	20.87%
	South	57.90%	16.34%	4.93%	20.83%
	West	55.68%	18.18%	5.25%	20.89%
	Total	57.02%	14.09%	8.03%	20.86%

Conclusion

- In this paper, we design a data structure that encodes observable states in Sanma and We designed a data structure that encodes observable states and constructed a CNN structure that solves the Sanma decision-making problem.

- Behavior model achieved test accuracy comparable to that of a 4-player Mahjong AI through supervised learning.

Future Work

- In the future we will take into account information from multiple rounds to maximize the final score

- This requires flexibility, for example, if the AI is in first place in the final round, the AI can intentionally lose to the lowest ranked player to keep the first place

Thank you for attention