DOI: 10.62105/2949-6349-2024-1-S1-97-99

EDN IGICHO

АПРОБАЦИЯ ВЫЯВЛЕННОЙ УЯЗВИМОСТИ САМООБУЧАЮЩЕЙСЯ ПРОГРАММЫ ДЛЯ ИГРЫ В ГО «KATAGO»

Филиппов В.В.[1]**,** мастер спорта России, *sensei@go-igo.ru*

Зборовская Т.В.[1]**,** *t.zborovskaya@gmail.com*

[1] Московский государственный университет имени М.В. Ломоносова, Москва, Россия

Аннотация. В данной статье апробируется выявленная К. Пелрайном с соавт. уязвимость нейросети KataGo, играющей в го на уровне 9 дана. Построение групп согласно разработанному четырехуровневому алгоритму обнаруживает неспособность искусственного интеллекта понимать умышленное принесение в жертву крупных групп и расставлять приоритеты при наличии нескольких подобных жертв. Значимость обнаруженной стратегии заключается в том, что с ее применением человек одерживает верх над потенциально непобедимой программой при игре в силу I спортивного разряда.

Ключевые слова: игра го, бадук, вэйци, KataGo, нейронные сети

APPROBATION OF DETECTED WEAKNESS OF SELF-TAUGHT GO GAME AI KATAGO

Filippov V.V.[1], Master of Sports of Russia, *sensei@go-igo.ru*

Zborovskaya T.V.[1], *t.zborovskaya@gmail.com*

[1] Lomonosov Moscow State University, Moscow, Russia

Abstract. The article describes testing a weakness of the go-playing AI KataGo (9d) detected by K. Pelrine et al. Building groups following the developed 4-level algorithm reveals the incapability of AI to understand intentional sacrifice of larger groups and to prioritize when discovering several similar sacrifices on the board. The value of the developed strategy lies in human victory over a potentially unbeatable program even regarding the fact that human is playing on the level of 1q (2000 Elo).

Keywords: go game, baduk, weiqi, KataGo, neural networks

Обоснование. Первой технологией, разработанной для игры в го и способной осуществлять расчеты в диапазоне 171-значных чисел (количество допустимых комбинаций в одной партии в го), стала нейросеть AlphaGo от компании Google DeepMind [1]. Для проверки программы были проведены матчи с профессиональными игроками Фань Хуэем (2015, 5-0 в пользу AlphaGo) и Ли Седолем (2016, 4-1 в пользу AlphaGo). AlphaGo обучалась, просматривая партии настоящих игроков; нейросети нового поколения – такие как KataGo [2], Leela [3], FineArt [4] – обучаются, играя сами с собой и друг с другом. В ноябре 2022 года Келлин Пелрайн и его коллеги при помощи компьютерной программы обнаружили возможность победить KataGo. В феврале 2023 выходит научная публикация о предложенном методе. Цель нашего исследования – апробировать стратегию Пелрайна с соавт. усилиями человека и убедиться в возможности победы человека над программой.

Методы. Обученная Пелрайном с соавт. программа-соперник выигрывает против KataGo последней актуальной версии (база b40c256-s11840935168-d2898845681) в 72%–100% случаев при помощи стратегии, основанной на уязвимости многослойной структуры из мертвых групп. При построении такой четырехуровневой структуры программа теряется в приоритетах, не может правильно рассчитать количество ходов, необходимых для спасения своих групп, стремится ликвидировать 1-й уровень «матрешки» и к моменту, когда кольцо из мертвых групп 2-го уровня смыкается, уже проигрывает по темпу [5]. Изучив алгоритм, мы воспроизвели следующие 4 этапа стратегии:

Этап 1. Построение мертвой группы черных в окружении кольца белой группы.

Этап 2. Окружение кольца белых при помощи 5 мертвых групп черных. Программа окружает черных оставшимися камнями белых. Задействуется все игровое поле.

Этап 3. Третье кольцо мертвых черных групп постепенно смыкается вокруг второго кольца единой белой группы. Из-за построения «матрешки» в программе происходит сбой оценки позиции, белые игнорируют действия соперника, и черные завершают окружение. Все камни белых гибнут, а черные оживают. Белые терпят сокрушительное поражение.

Этап 4. На конечном этапе партии работа нейронной сети нарушается настолько, что уровень игры программы падает с профессионального до уровня знания правил.

Результаты. Предложенную К. Пелрайном с соавт. стратегию удалось успешно апробировать. Человек одерживает верх над программой KataGo, воспроизводя предложенный алгоритм.

Заключение. В отличие от шахмат, в го победа нейросети над человеком, провозглашенная в 2016 году, на данном этапе является неокончательной, так как описанные комбинации легко конструируются человеком и могут производиться игроком I спортивного разряда.

Список литературы

1. Официальный сайт нейросети AlphaGo [Электронный ресурс]. URL: https://www.deepmind.com/research/highlighted-research/alphago (дата обращения: 20.09.2023).
2. Установочные файлы нейросети KataGo на портале разработчиков GitHub. URL: https://github.com/lightvector/KataGo/releases?q=v1.12.4&expanded=true (дата обращения: 20.09.2023).
3. Официальный сайт программы Leela. URL: https://www.sjeng.org/leela.html (дата обращения: 20.09.2023).
4. Lauder E. Tencent's made a Go-playing AI to rival Google's AlphaGo. URL: https://aibusiness.com/companies/tencent-s-made-a-go-playing-ai-to-rival-google-s-alphago (дата обращения: 20.09.2023).
5. Wang T.T., Gleave A., Tseng T., Pelrine K., Belrose N. et al. Adversarial policies beat superhuman Go AIs // arXiv:2211.00241 [cs.LG]. URL: https://goattack.far.ai/adversarial-policy-katago (дата обращения: 20.09.2023). DOI: 10.48550/arXiv.2211.00241
6. Financial Times. Man beats machine at Go in human victory over AI. URL: https://www.ft.com/content/175e5314-a7f7-4741-a786-273219f433a1 (дата обращения: 27.11.2023).

References

1. AlphaGo by Google DeepMind. URL: https://www.deepmind.com/research/highlighted-research/alphago (accessed 20.09.2023).
2. KataGo on GitHub.
URL: https://github.com/lightvector/KataGo/releases?q=v1.12.4&expanded=true (accessed 20.09.2023).
3. Leela official website. URL: https://www.sjeng.org/leela.html (accessed 20.09.2023).
4. Lauder E. Tencent's Made a Go-Playing AI to Rival Google's AlphaGo. URL: https://aibusiness.com/companies/tencent-s-made-a-go-playing-ai-to-rival-google-s-alphago (accessed 20.09.2023).
5. Wang T.T., Gleave A., Tseng T., Pelrine K., Belrose N. et al. Adversarial policies beat superhuman Go AIs // arXiv:2211.00241 [cs.LG]. URL: https://goattack.far.ai/adversarial-policy-katago (accessed 20.09.2023). DOI: 10.48550/arXiv.2211.00241
6. Financial Times. Man beats machine at Go in human victory over AI. URL: https://www.ft.com/content/175e5314-a7f7-4741-a786-273219f433a1 (accessed 27.11.2023).