**Inteligência Artificial­­­ -** Marco Moliterno Pena Piacentini

**A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. (Silver et al. 2018).**

Publicado na revista Science em 2018, o artigo (1) apresenta o algoritmo AlphaZero, capaz de alcançar performance superior aos humanos em jogos de xadrez, shogi (xadrez japonês) e go (xadrez chinês). Alpha Zero é uma versão mais ampla do AlphaGo Zero, aplicativo com performance superior à humana no jogo go. Ambos os aplicativos utilizam apenas as regras do jogo como base e são treinados unicamente com reinforcement learning a partir de self-play. Alpha Zero utiliza o mesmo algoritmo e a mesma arquitetura para os três jogos, obtendo resultados superiores aos humanos em todos eles, o que demonstra que algoritmos de reinforcement learning de uso geral podem aprender a partir do zero, sem utilizar conhecimento ou dados humanos.

Outros programas de xadrez computacional avaliam posições com base em jogadas de jogadores humanos talentosos, combinadas com busca alfa-beta sobre uma ampla árvore de busca com um grande número de heurísticas e adaptações. Os primeiros programas a vencer campeões humanos foram desenvolvidos com esta lógica: no xadrez, o Deep Blue em 1997 e no shogi, que é um jogo mais complexo, o Elmo, 2017. Já AlphaGo Zero e Alpha Zero são alimentados apenas com informações da estrutura do tabuleiro e com as regras do jogo, utilizam aprendizado unicamente em self-play, empregam rede neural que utiliza a posição do tabuleiro como input e fornece como output o vetor de probabilidade para cada ação e um valor escalar de cada outcome esperado, combinado com algoritmo árvore de busca Monte Carlo (MCTS) para avaliação das posições, aprimorando constantemente a escolha com resultados anteriores.

Alpha Zero utiliza a mesma arquitetura rede neural de AlphaGo Zero, mas segue outra estratégia para se tornar genérico. Go tem regras simétricas e outcome binário – ganhar ou perder, mas xadrez e shogi têm regras assimétricas (ex. roque é diferente do lado do rei e da rainha) e podem terminar empatados, de forma que, para ser genérico, Alpha Zero não assume simetria como AlphaGo Zero, não aumenta o training data, não transforma a posição do tabuleiro durante o MCTS. Em AlphaGo, a performance da jogada é comparada à melhor jogada de todas as iterações anteriores e a substitui se vencer com margem de 55%, diferente de Alpha Zero, que mantém uma única rede neural que é continuamente atualizada, sem aguardar que a iteração seja completada.

Instâncias diferentes de Alpha Zero foram treinadas para xadrez, shogi e go por 700,000 steps, começando por parâmetros de iniciação randômicos. Foram utilizados 5.000 TPU’s de primeira geração para gerar os self-plays e 16.000 TPU’s de segunda geração para treinar as redes neurais. Os treinamentos duraram aproximadamente 9 horas para xadrez, 12 horas para shogi e 13 dias para go. Alpha Zero teve melhor performance do que programas especializados, em tempo variável - 4 horas para xadrez, contra Stockfish, 2 horas para shogi, contra Elmo e 30 horas para go, contra AlphaGo. Cada programa rodou no hardware em que foi desenhado: Stockfish e Elmo usaram CPUs e Alpha Zero usou com 4 TPUs de primeira geração e 44 CPUs.

Esses resultados de Alpha Zero nos aproximam de atingir uma ambição da inteligência artificial: um sistema de jogo computacional que pode aprender a vencer qualquer jogo.

**Bibliografia:**

1. SILVER, David et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. **Science**, v. 362, n. 6419, p. 1140-1144, 2018.