



Redes Neurais Profundas para Aprendizado por Reforço, testado em jogos de Atari

Guilherme de Lázari da Costa e Silva, Ramon de Oliveira, Eduardo Valle

Resumo

Redes neurais profundas têm resultados impressionantes em tarefas de classificação e de regressão. Mais recentemente, essas redes estão sendo exploradas para aprendizado por reforço, contexto em que nosso trabalho se insere.

Palavras-chave:

aprendizado de máquina, inteligência artificial, aprendizado por reforço

Introdução

Redes neurais convolucionais trouxeram um pulo de performance para a tarefa de reconhecimento de objetos. Esse ganho se deve tanto a inovações no funcionamento dessas redes, quanto ao aumento exponencial do poder de computação disponível.

Aproveitando-se desses dois aspectos, técnicas de aprendizado por reforço estão sendo exploradas com o uso de redes convolucionais em suas arquiteturas. Mais especificamente tem-se usado essas redes para jogos de Atari, utilizando como entrada apenas a imagem da tela.

Este trabalho tem como objetivo reproduzir o resultado desses estudos e entender quais avanços permitiram o melhor desempenho nessas tarefas de aprendizado por reforço.

Resultados e Discussão

O algoritmo de aprendizado por reforço que implementamos foi o Q-Learning. Esse algoritmo supõe que, em cada estado do jogo, cada ação tem um retorno esperado no longo prazo. A função que mapeia o espaço de estados e ações no retorno esperado é chamada de função Q. O objetivo do algoritmo é aprender essa função Q.

Uma rede neural é um tipo de aproximador universal de funções. No nosso contexto, o papel delas é aproximar a função Q. As redes aproximam a função através de um modelo paramétrico em que os parâmetros (pesos) são aprendidos à medida em que obtemos dados – no nosso caso, em que observamos recompensas e penalidades sofridas após a escolha de certas ações.

Figura 1. Arquiteturas de redes-neurais como Q-function [1]

Existem dois modos de modelar esse problema com uma rede neural: ou o estado e a ação são entradas da rede (como mostra a parte esquerda da Figura 1), ou o estado é a entrada e existe uma saída para cada uma das ações possíveis. (como mostrado na direita da Figura 1). O segundo modo é preferido, pois exige uma única passagem pela rede para saber o Q-value para todas as ações. Além disso o segundo modo permite uma rede específica para o formato de dados do estado.

Nossa implementação para o jogo de Breakout treinou durante 3 dias e conseguiu atingir uma performance acima da esperada para um ser humano. Não conseguimos, porém, alcançar uma performance tão alta quanto a relatada pelos autores do trabalho original. Essa dificuldade, frequente na comunidade, deve-se ao fato de que as redes são notoriamente difíceis de treinar, exigindo muitos ajustes *ad hoc*, vários dos quais não são documentados nos artigos publicados e fazem parte de um certo *folklore* dos praticantes da área.

Conclusão

Juntar aprendizado por reforço com *end-to-end-learning* tem o potencial de causar grande impacto na forma como novos problemas são resolvidos, e até mesmo na automatização de novas tarefas. Carros autônomos, por exemplo, são fortes candidatos para se beneficiarem com esse tipo de abordagem.

Agradecimentos

Agradecimento ao programa PIBIC do CNPq por ter financiado a pesquisa e à Universidade Estadual de Campinas, pela infraestrutura.

Mnih, Kavukcuoglu K., Silver D., et al: *Human-level control through deep reinforcement learning* 2015

