Nome do Aluno: João Gabriel Valentim Rocha

Nome do Orientador: Fábio José Ayres

Título: Estudo comparativo de modelos de otimização de agentes autônomos baseados em aprendizado por reforço.

Palavras chave: *reinforcement learning, genetic algorithms, neural network policies, steering behaviors, policy gradients, deep Q-network.*

1. Descrição do Problema e Revisão de Literatura

Grande parte das técnicas utilizadas para traçar a caminhada de robôs ou carros que dirigem sem a necessidade de motorista são pautadas na otimização de agentes autônomos. A área do conhecimento denominada "Otimização de agentes autônomos" (em inglês: Autonomous Agent optimization) refere-se ao conjunto de técnicas utilizadas para aprimorar uma atividade feita por um agente autônomo, inserido em um ambiente, por meio de ferramentas computacionais, com o objetivo de buscar a solução ótima de uma tarefa determinística ou não (GÉRON, A. 2019). Exemplos do uso de otimização de agentes autônomos incluem a otimização da caminhada de um robô, a trajetória de carros que dirigem sem a necessidade de um motorista, sistemas que podem aprender a jogar jogos eletrônicos (V. MNIH. *et al*. 2013), otimização de um termostato para a economia de energia e negociações financeiras automáticas (GÉRON, A. 2019).

Em muitas aplicações de otimização de agentes autônomos é necessário o emprego de técnicas de aprendizado por reforço. Nos últimos dez anos o campo de aprendizado de máquina passou por uma mudança significativa com o advento das redes neurais profundas (HINTON G., OSINDERO S., TEH Y.W. 2006). Estas são redes neurais com um grande número de camadas escondidas e topologias inovadoras, que tem se mostrado bastante efetivas na construção de modelos preditivos. Particularmente, na área de processamento de imagens, o surgimento de arquiteturas de redes neurais profundas como AlexNet (KRIZHEVSKY A., SUTSKEVER I., HINTON .E. 2012), VGGNet (SIMONYAN K., ZISSERMAN A. 2015) e ResNet (HE K., ZHANG X., REN S., SUN J. 2016) trouxe um nível de desempenho de sistemas computacionais em atividades de entendimento de imagens que é compativel – e por vezes superiores – aos seres humanos (V. MNIH. *et al*. 2015).

Recentemente, modelos de aprendizado por reforço estão sendo cada vez mais aplicados com conceitos de aprendizagem profunda (Aprendizado por reforço profundo), e existe a expectativa de que o uso destes modelos tenha impacto tão significativo em otimização de agentes autônomos quanto os modelos de imagem tiveram sobre o processamento de imagem e visão computacional. Pode-se citarcomo modelos de destaque na literatura cientifica o Deep Q-Networks (DQN) (V. MNIH. *et al*. 2015), o Policy Gradient (PG) (AGARWAL, A. 2020) e o Genetic Algorithm (GA) (MIRJALILI S. 2019). Para exemplificar o avanço desse cenário, com base em 49 jogos do Atari 2600, o modelo DQN foi utilizado como agente e teve desempenho comparável ao de um humano profissional, alcançando mais de 75% da pontuação humana em mais da metade dos jogos (29 jogos), com uma assertividade muito superior aos modelos de aprendizado por reforço convencionais aplicados também ao Atari 2600 (BELLEMARE, M. G., NADDAF, Y., VENESS, J. & BOWLING, M. 2013). Uma outra aplicação das técnicas de aprendizado por reforço, foi o resultado do embate entre o coreano Lee Sedol, 18 vezes campeão mundial do milenar jogo de tabuleiro "Go", e o supercomputador AlphaGo que em uma disputa de 5 jogos, 4 deles foram vencidos pelo AlphaGo (SEDOL,L VS ALPHAGO 2016).

2. Objetivo

Neste projeto serão avaliados os desempenhos dos modelos DQN, PG e GA em tarefas clássicas de otimização de agentes autônomos, bem como será desenvolvida uma metodologia de avaliação e customização de modelos para tarefas específicas, tais como classificação otimização de comportamentos de direção (em inglês: “Steering behaviors”).

Os objetivos específicos deste projeto são:

- Conhecer técnicas de aprendizado por reforço aplicadas a otimização de agentes autônomos;
- Comparar técnicas modernas de otimização de agentes autônomos;
- Desenvolver uma técnica de otimização de comportamentos de direção de agentes autônomos com aprendizado por reforço.

3. Metodologia (Proposta)

O projeto deverá seguir as seguintes etapas:

a) Estudo acerca de aprendizado por reforço, aprendizado por reforço profundo e novas arquiteturas para otimização de agentes autônomos:

Embasamento computacional, matemático e estatistico com o intuito de saber quais as melhores aplicações de cada modelo para agentes autônomos.

b) Nesta etapa do projeto, é imprescindível que os problemas a serem abordados sejam definidos, para então aplicar os conceitos associados aos modelos GA, PG e DQN:
   a. **Definição do(s) problema(s):** Escolher o problema ou o conjunto de problemas no qual serão aplicadas as técnicas de aprendizado por reforço.
   b. **Aplicação dos modelos**: Estudo e implementação dos mecanismos a serem utilizados para aplicar os modelos GA, PG e DQN dentro do contexto do(s) problema(s) inserido(s) na etapa anterior.

c) Avaliação de desempenho dos modelos GA, PG e DQN:

A priori, os modelos que serão usados para realizar tarefas de otimização de agente autônomos são GA, PG e DQN, contudo, com base na etapa anterior poderá ser selecionado algum outro modelo que também venha acrescentar na solução de uma dada tarefa.

Para cada modelo serão definidos:

   a. **Tarefas e aplicações:** Escolher um ou mais tipos de agentes autônomos e quais comportamentos de direção. A escolha dependerá da etapa de estudo das técnicas e de suas potencialidades, além da estapa de definição do problema a ser analisado.

   b. **Avaliação de desempenho:** Ao definir o problema específico, é possível estabelecer a métrica específica a ser utilizada, portanto, o modelo é então avaliado e sua performance na tarefa é documentada. Existe uma miríade de agentes autônomos e ambientes para avaliação de desempenho disponíveis na literatura, a depender da tarefa a ser avaliada. Nesta fase do projeto será escolhido o conjunto apropriado.

d) Desenvolvimento de metodologia de customização dos modelos para usos específicos

e) Criação de uma base de treinamento, validação e teste agentes autônomos com comportamento de direção

f) Demonstração do uso de aprendizado por reforço e aprendizado por reforço profundo para o conjunto de problemas determinado nas etapas anteriores.

4. Resultados Esperados

Como resultados desta pesquisa teremos uma análise do estado-da-arte em otimização de agentes autônomos, e o desenvolvimento de um modelo de aprendizado profundo para a análise de agentes autônomos com comportamento de direção. Espera-se que este modelo seja útil a produção de insights a respeito de aplicações na área de otimização de agentes autônomos com comportamento de direção.

5. Referências Bibliográficas

V. MNIH. Playing Atari with Deep Reinforcement Learning, **Deep Mind, ArXiv,** 2013. https://arxiv.org/pdf/1312.5602v1.pdf.

V. MNIH. Human-level control through deep reinforcement learning, **Nature** 2015. https://storage.googleapis.com/deepmind-data/assets/papers/DeepMindNature14236Paper.pdf.

SHIFFMAN, D. The Nature of Code. **Simulating Natural Systems With Processing**, 2012. https://natureofcode.com/book/

GÉRON, A. Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow. **Concepts, Tools and Techniques to Build Intelligent Systems**, 2nd Ed, Páginas 437-470, 2019.

HINTON, G.; OSINDERO, S.; TEH, Y. W. A fast learning algorithm for deep belief nets. **Neural Computation**, Volume 18 Issue 7, Páginas 1527-1554, 2006.

KRIZHEVSKY A., SUTSKEVER I., HINTON .E. Imagenet classification with deep convolutional neural networks. **Advances in Neural Information Processing Systems,** Páginas 1097-1105, 2012.

SIMONYAN K., ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. **International Conference on Learning Representations,** 2015.

HE K., ZHANG X., REN S., SUN J. Deep residual learning for image recognition. **Computer Vision and Pattern Recognition**, Páginas 770-778, 2016.

AGARWAL A., SHAM M KAKADE, JASON D LEE, MAHAJAN G. Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes. **Proceedings of Thirty Third Conference on Learning Theory**. PMLR 125:64-66, 2020.

MIRJALILI S. (2019) Genetic Algorithm. In: **Evolutionary Algorithms and Neural Networks. Studies in Computational Intelligence**, vol 780. Springer, Cham. https://doi.org/10.1007/978-3-319-93025-1_4.

BELLEMARE, M. G., NADDAF, Y., VENESS, J. & BOWLING, M. The arcade learning environment: An evaluation platform for general agents. **J. Artif. Intell. Res. 47, 253–279** (2013).

BELLEMARE, M. G., VENESS, J. & BOWLING, M. Investigating contingency awareness using Atari 2600 games. **Proc. Conf. AAAI. Artif. Intell. 864–871** (2012).

CRAIG W. REYNOLDS. 1987. Flocks, herds and schools: **A distributed behavioral model**. SIGGRAPH Comput. Graph. 21, 4 (July 1987), 25–34. DOI:https://doi.org/10.1145/37402.37406

CRAIG W. REYNOLDS. Steering Behaviors For Autonomous Characters. **Game developers conference**. 1999. http://www.red3d.com/cwr/steer/gdc99/

SEDOL,L VS ALPHAGO. Google DeepMind Challenge Match: **Lee Sedol vs AlphaGo**. 2016. https://www.youtube.com/watch?v=yCALyQRN3hw&t=22412s

Cronograma de atividades

	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Estudo de técnicas											
Avaliação de desempenho											
Metodologia customizada											
Criação base de dados											
Aplicação a Agentes autônomos											
Atividades de escrita de relatório											