APRENDIZADO POR REFORÇO APLICADO EM UMA SIMULAÇÃO PARA CONTROLE INTELIGENTE DE SEMÁFOROS

Fernando Rodrigues Junior[1], Marcel Campos Inocencio[2]

Resumo: A mobilidade urbana representa um desafio para os centros urbanos, tendo impactos negativos na qualidade de vida e no meio ambiente. Os semáforos, sozinhos, não são eficientes o bastante para controlar o tráfego de maneira satisfatória. Nesse contexto, o uso da tecnologia desempenha um papel fundamental na criação de soluções inteligentes e no desenvolvimento de cidades inteligentes. A aplicação de algoritmos de inteligência artificial pode otimizar a fluidez e a segurança do trânsito. A utilização de simulações desempenha um papel importante na validação das soluções propostas, permitindo redução de custos e avaliação de eficácia. Nesta pesquisa, foi conduzida uma abordagem exploratória e aplicada com base tecnológica visando otimizar o tráfego urbano em interseções semafóricas. A pesquisa investiga a utilização do aprendizado de máquina por reforço para controlar os semáforos, empregando o algoritmo de Otimização de Política Proximal em conjunto com a realização de uma simulação. A simulação foi desenvolvida de forma a permitir o controle de diversos aspectos, como o número de veículos na simulação, o fluxo de tráfego em cada via e a taxa de geração e remoção de veículos. Foram realizados testes em três cenários distintos, nos quais foi possível explorar as habilidades do agente em se adaptar a diferentes fluxos de tráfego. Os resultados obtidos indicam que o algoritmo de Otimização de Política Proximal pode ser usado para treinar agentes inteligentes e ao ser comparado com o método convencional empregado no Brasil, foi observado resultados positivos, apresentando um aumento de 10,03%, 19,56% e 22,15% nos respectivos testes.

Palavras-chave: Aprendizado por reforço. Otimização de política proximal. Controle de semáforos. Simulação de tráfego. Cidades inteligentes. Mobilidade urbana.

[1] Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (UNESC), Criciúma – SC - Brasil. fernandorodriguesdev@unesc.net

[2] Orientador, Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (UNESC), Criciúma - SC - Brasil. marcel.inocencio@unesc.net

ABSTRACT: Urban mobility represents a significant challenge for urban centers, with negative impacts on quality of life and the environment. Traffic lights alone are not efficient enough to control traffic satisfactorily. In this context, the use of technology plays a crucial role in creating intelligent solutions for urban mobility and the development of smart cities. The application of artificial intelligence algorithms can optimize traffic flow and safety. Furthermore, the use of simulations plays an important role in validating proposed solutions, allowing for cost reduction and effectiveness evaluation. In this research, an exploratory and applied approach was conducted based on technology to optimize urban traffic at signalized intersections. The research investigates the use of reinforcement learning to control traffic lights, employing the Proximal Policy Optimization algorithm in conjunction with a simulation. The simulation was developed to allow control over various aspects, such as the number of vehicles in the simulation, traffic flow on each lane, and the rate of vehicle generation and removal. Tests were conducted in three distinct scenarios, which made it possible to explore the agent's abilities to adapt to different traffic flows. The results obtained indicate that the Proximal Policy Optimization algorithm can be used to train intelligent agents. When compared to the conventional method employed in Brazil, positive results were observed, with an increase of 10.03%, 19.56%, and 22.15% in the respective tests.

Keywords: Reinforcement learning. Proximal policy optimization. Traffic light control. Traffic simulation. Smart cities. Urban mobility.

1 INTRODUÇÃO

A mobilidade urbana tem se estabelecido como um dos principais desafios enfrentados pelos centros urbanos. O congestionamento do trânsito provoca estresse para a população e impõe um custo significativo ao meio ambiente, afetando diretamente a qualidade de vida da comunidade. O aumento do tráfego em vias urbanas, aliado à falta de investimentos adequados em infraestrutura, tais como a expansão das rodovias, tem aumentado a complexidade do fluxo de veículos, o que, por sua vez, contribui para o aumento do consumo de combustíveis, além de ocasionar congestionamentos e acidentes (OLIVEIRA; ARAÚJO, 2016).

Na ausência de um fluxo adequado de veículos, é observada uma ampliação dos níveis de poluição sonora e visual, especialmente durante os períodos de maior demanda. O estresse resultante desse cenário acarreta consequências tanto físicas quanto psicológicas, levando os condutores a procurar rotas alternativas e atalhos muitas vezes inexistentes, resultando em acidentes e agravando ainda mais os congestionamentos. Essas adversidades sociais, por sua vez, se manifestam no comprometimento do desempenho profissional, diminuição das interações familiares e ocorrência frequente de atrasos (OLIVEIRA; ARAÚJO, 2016).

Os semáforos foram criados com o intuito de amenizar engarrafamentos e garantir segurança aos motoristas e pedestres. No entanto, por si só não são suficientes para assegurar a eficiência no escoamento do fluxo de veículos, demandando assim a implementação de um gerenciamento adequado, o qual, em muitas instâncias, se mostra inexistente. Considerando a realidade, verifica-se que os métodos utilizados no Brasil não acompanham a evolução tecnológica e não levam em consideração as flutuações imprevisíveis do tráfego, resultando na ineficácia do controle de tráfego (MENENDEZ; SILVA; PITANGA, 2022).

Os centros urbanos crescem e evoluem diariamente e a tecnologia está cada vez mais acessível e presente na vida das pessoas. Valendo-se dos desafios gerados pelo rápido crescimento urbano e seus impactos, a busca por soluções inteligentes parte do princípio de que a tecnologia é fundamental para o desenvolvimento de uma cidade inteligente, modernizando e oferecendo melhor infraestrutura à população (ALVES; DIAS; SEIXAS, 2019). Uma abordagem promissora para lidar com essas questões envolve a aplicação de técnicas de gerenciamento avançadas, como algoritmos de inteligência artificial, visando otimizar a fluidez e a segurança do trânsito (MENENDEZ; SILVA; PITANGA, 2022).

O aprendizado de máquina é um subcampo da Inteligência Artificial com o objetivo de desenvolver técnicas capazes de emular a inteligência humana. Seu propósito é permitir que os sistemas aprendam a partir de dados de treinamento específicos, a fim de automatizar processos de construção de modelos analíticos e solucionar tarefas, baseando-se em experiências prévias (MONARD; BARANAUSKAS, 2003). Aprendizado por reforço é uma técnica que tem como fundamento permitir que a rede treine sem dados iniciais, aprimorando-se por meio de um sistema de recompensas no qual acertos ou erros resultam em consequências

que incrementam a probabilidade de alcançar um determinado objetivo (SILVA et al., 2020).

A adoção dessas abordagens tem o potencial de reduzir congestionamentos, acidentes e emissões de poluentes, proporcionando benefícios significativos para a mobilidade urbana (MENENDEZ; SILVA; PITANGA, 2022).

Para alcançar tais resultados, é necessário implantar tecnologias que permitam a identificação em tempo real da situação do tráfego, juntamente com o desenvolvimento de métodos que possibilitem um controle e gerenciamento eficiente. No entanto, considerando os custos associados a essas ações, a utilização de simulações desempenha um papel importante na validação das soluções propostas. As simulações possibilitam testar e avaliar de forma simplificada a eficácia da otimização dos semáforos em cenários diversos, sem a necessidade de investir recursos públicos em implementações reais. Essa abordagem permite a realização de testes em um ambiente controlado que se assemelha à realidade, resultando em uma redução de tempo e custos associados ao desenvolvimento do projeto (VIEIRA, 2006).

Portanto, o presente estudo tem como objetivo investigar o uso de aprendizado de máquina por reforço, por meio de uma simulação, para a otimização do tráfego urbano em interseções semafóricas.

Os objetivos específicos desta pesquisa consistem em utilizar o algoritmo de Otimização de Política Proximal (PPO) de aprendizado de máquina por reforço; comparar o desempenho do método de gerenciamento de semáforos baseado em aprendizado por reforço com o método convencional de semáforos a tempos fixos atualmente utilizado no Brasil; analisar os resultados das comparações empregadas.

2 TRABALHOS CORRELATOS

Esse tópico apresenta os trabalhos nacionais e internacionais relacionados aos temas abordados no trabalho, os quais foram analisados, com a finalidade de adquirir embasamento sobre os assuntos pontuados.

2.1 CONTRIBUIÇÕES DE APRENDIZADO POR REFORÇO EM ESCOLHA DE ROTA E CONTROLE SEMAFÓRICO

No estudo de Bazzan (2021), é realizada uma análise focada em duas tarefas específicas, controle de semáforos e escolha de rotas, destacando a

relevância das contribuições da inteligência artificial. Dentre os diversos aspectos abordados, há uma ênfase nos benefícios proporcionados pelo uso de novos métodos e tecnologias relacionadas à inteligência artificial, especialmente ao aprendizado de máquina e aprendizado por reforço. Foram apresentados métodos baseados em aprendizado de máquina como: coordenação de semáforos via teoria dos jogos abordagem de aprendizado por reforço baseado em modelo e abordagens baseadas em controle hierárquico, que evidenciam os ganhos em termos de tempo de viagem e eficiência do sistema como um todo.

2.2 DESENVOLVIMENTO DE UM SISTEMA DE CONTROLE DE TRÁFEGO INTELIGENTE BASEADO EM VISÃO COMPUTACIONAL

O trabalho de Cortez (2022) descreve que os semáforos que operam com tempo fixo não são eficientes em diversas situações do trânsito brasileiro e para aproveitar a rede semafórica já existente, o mesmo propôs uma solução de baixo custo. Para isso, utilizou visão computacional para contar os veículos que cruzam o semáforo, permitindo alertar sobre congestionamentos e tomar decisões com base nos dados coletados, por meio da captura de imagens das câmeras de segurança já presentes nas vias. Os semáforos foram controlados utilizando um Raspberry Pi 3 e um computador foi responsável por capturar as imagens do trânsito no semáforo, contar os veículos e calcular o tempo necessário para que eles realizassem a travessia. Obteve como resultado um ganho de até 33% na fluidez do trânsito relativo ao modelo atual de tempo fixo.

2.3 INTELLILIGHT: UMA ABORDAGEM DE APRENDIZADO POR REFORÇO PARA O CONTROLE INTELIGENTE DE SEMÁFOROS

O trabalho de Wei et al (2018) contém a proposta de um modelo de *Deep Reinforcement Learning* que visa testar os modelos com um conjunto de dados de tráfego real em grande escala obtido por meio de câmeras de vigilância. Segundo os autores, os estudos existentes ainda não testaram os métodos nos dados de tráfego reais e se concentram apenas no estudo das recompensas sem interpretar as políticas. Obteve como resultado, comparando com métodos de linha de base em dados do mundo real, o alcance da melhor recompensa, comprimento da fila, atraso

e duração em todos os métodos comparados (controle de tempo fixo, controle de semáforo auto-organizado e aprendizado profundo por reforço para o controle de semáforos), com uma melhoria relativa de 32%, 38%, 19% e 22% correspondentemente em relação ao melhor método de linha de base.

3 MATERIAIS E MÉTODOS

Nesta pesquisa, foi realizada uma abordagem exploratória e aplicada com base tecnológica para otimizar o tráfego urbano em interseções semafóricas. Para isso, foram aplicados conceitos de aprendizado por reforço, por meio de simulação. O objetivo principal foi utilizar o algoritmo de Otimização de Política Proximal, do inglês *Proximal Policy Optimization* (PPO), para treinar um agente inteligente responsável pelo controle de uma interseção com quatro semáforos.

Otimização de Política Proximal é um algoritmo de aprendizado por reforço, utilizado para otimizar políticas em agentes de tomada de decisão. O PPO equilibra exploração e aproveitamento por meio de uma função de objetivo substituta e otimização com restrições. O algoritmo coleta dados de interação com o ambiente, calcula vantagens para estimar a qualidade das ações e realiza múltiplas épocas de otimização para atualizar a política do agente. O PPO também emprega técnicas como estimativa de função de valor[3] e regularização de entropia[4] para melhorar o desempenho. Essas características tornam o PPO uma abordagem eficiente para problemas complexos, como o controle de semáforos em ambientes urbanos (OPENAI, 2023).

O desenvolvimento da aplicação foi realizado utilizando o motor de jogo Unity, na versão 2021.3.19f1 LTS. Para incorporar técnicas de aprendizado por reforço à aplicação no Unity, foi utilizado o pacote ML-Agents, versão 0.30.0, e a linguagem de programação Python, na versão 3.7. As configurações do computador utilizado para desenvolver o projeto se constituem em: Processador AMD Ryzen 3600, Placa de Video Nvidia GTX 1660 super e 16 gigas de memória RAM.

[3] A estimativa de função de valor envolve a predição do valor esperado de uma política em um estado.
[4] A regularização de entropia incentiva a exploração e diversidade de ações através do aumento da entropia da política.

3.1 AMBIENTE DE APRENDIZADO

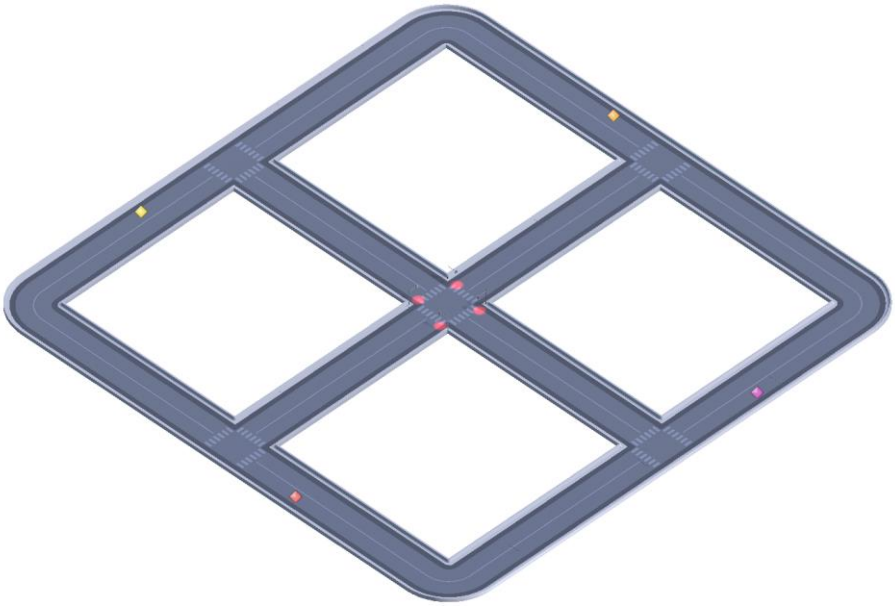
A Unity é um motor de jogo multiplataforma criado pela Unity Technologies, tem como foco o desenvolvimento de mídias interativas em tempo real, sendo utilizada com maior frequência no desenvolvimento de jogos eletrônicos, entretanto também é utilizada em: arquitetura, simulação, cinema, entre outros. Conta ainda com um ambiente de desenvolvimento integrado, fornecendo uma interface por meio da qual é possível utilizar as ferramentas de desenvolvimento em um único lugar. Além disso, conta com uma solução de aprendizado de máquina que visa ajudar desenvolvedores a implementar e testar modelos de forma facilitada (UNITY, 2022).

As cenas desempenham um papel fundamental no desenvolvimento de jogos e aplicativos. Uma cena é um ambiente em que é possível criar, organizar e manipular objetos e elementos específicos do projeto. Ela serve como um recipiente que contém todos os componentes necessários para criar uma determinada parte da aplicação (UNITY, 2022).

Dentro de uma cena na Unity, é possível criar e manipular *GameObjects*, um *GameObject* é uma entidade fundamental que representa um objeto no ambiente virtual. Esses objetos podem variar desde personagens, inimigos, cenários, itens, até luzes, câmeras e efeitos especiais. Cada *GameObject* possui uma série de componentes que definem seu comportamento, aparência e interação com o ambiente virtual. Esses componentes podem incluir *scripts*, colisores, renderizadores, controladores de animação, entre outros. Eles permitem a criação de interatividade e a implementação da lógica da aplicação (UNITY, 2022).

A simulação foi desenvolvida considerando a interconexão de quatro vias de tráfego (figura 1). Todas as vias são de duplo sentido e convergem para uma interseção central. Essa interseção é dividida em dois grupos distintos, denominados grupo A e grupo B. É importante destacar que as mudanças de sinalização luminosa ocorrem simultaneamente nos dois grupos, de forma coordenada. Os automóveis presentes em cada grupo se deslocam em conformidade com essa sincronização, são dotados de autonomia e possuem a capacidade de navegar independentemente, evitando colisões e seguindo as regras de trânsito, como parar nos semáforos e aguardar a autorização para prosseguir, sendo relevante observar que o tempo de segurança é adicionado no arranque dos veículos.

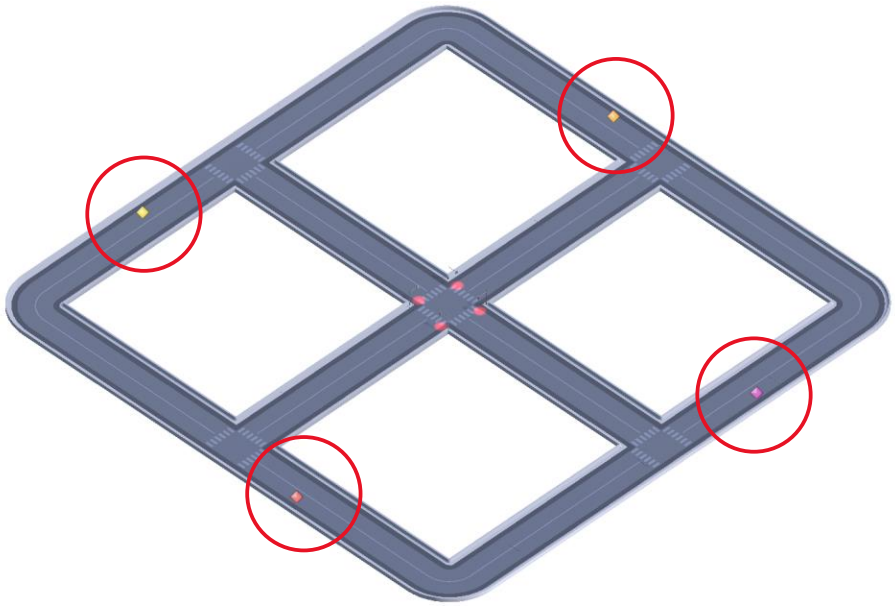
Figura 1 – Ambiente de aprendizado



Fonte: Do autor.

A geração de veículos na simulação foi realizada por meio de quatro geradores posicionados em locais específicos (figura 2). Esses geradores têm a capacidade de criar carros com base em dois parâmetros: tempo de geração e número máximo de carros gerados. O tempo de geração define o intervalo de tempo, em segundos, em que um novo carro deve ser gerado. Já o número máximo de carros gerados é utilizado para manter o número de veículos em cena constante, controlando a quantidade total de carros presentes.

Figura 2 – Geradores de veículos

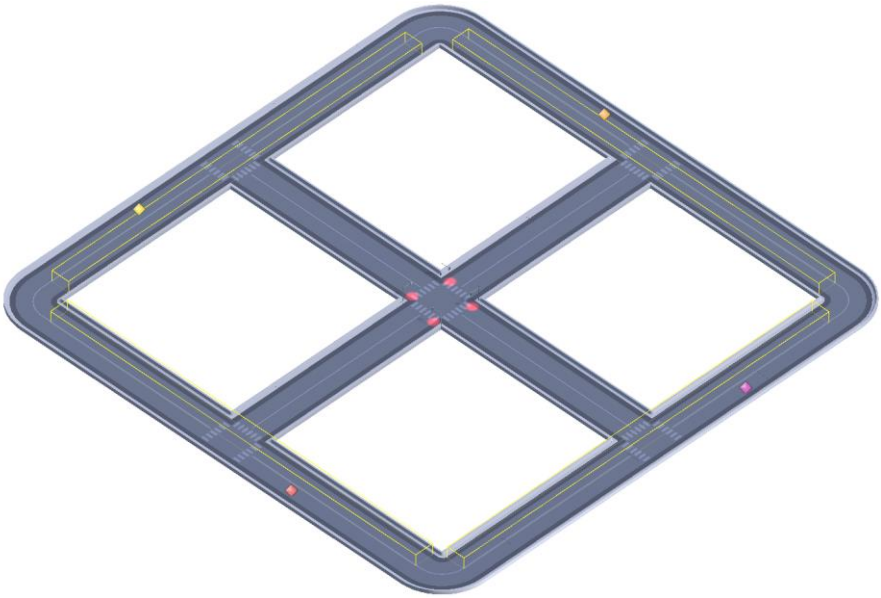


Fonte: Do autor.

Para simular o movimento dos veículos em direção a seus destinos, foi implementado um mecanismo de remoção de carros que tem como objetivo remover os que estão em cena, mas não estão nas vias principais com os semáforos (figura 3). Essa abordagem envolve a utilização de dois parâmetros, tempo de início, que determina quando a exclusão irá iniciar e intervalo de exclusão, ambos em segundos, foram posicionados quatro na cena. Essa estratégia visa garantir um fluxo constante de veículos nas vias controladas pelos semáforos. Dessa forma, ao alterar os valores dos parâmetros do gerador, é possível variar o volume de tráfego sem afetar o funcionamento dos semáforos.

Os geradores são responsáveis por manter o fluxo controlado de veículos na simulação, enquanto o mecanismo de destruição de carros garante que apenas os veículos relevantes para a interação com os semáforos estejam presentes, proporcionando uma representação mais realista do tráfego e permitindo a análise de desempenho dos semáforos em condições variadas.

Figura 3 – Representação dos colisores do destruidor de veículos em amarelo



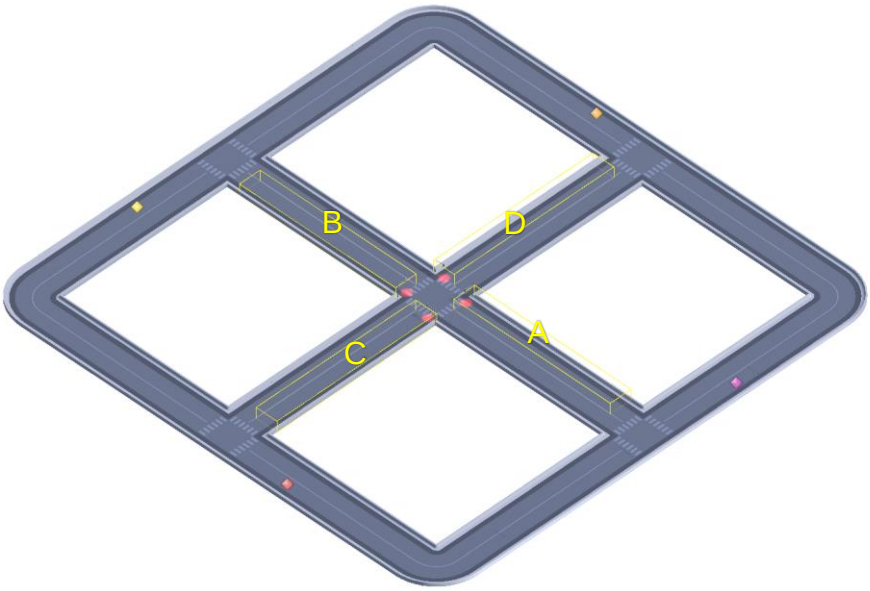
Fonte: Do autor.

3.2 O AGENTE INTELIGENTE

Os agentes no ambiente de aprendizagem operam por meio de etapas, a cada uma delas, coletam observações e as direcionam para sua política decisória, recebendo um vetor de ação. Na Unity, o agente é um *GameObject*, que estende a classe *Agent* do ML-Agents (RUBAK, 2021).

O agente é configurado com uma política de aprendizado baseada no algoritmo PPO, o qual permite que o agente atualize sua estratégia de ação com base nas observações coletadas e nas recompensas recebidas (JULIANI et al., 2018). Com o objetivo de coletar observações relevantes, foram definidos quatro colisores, um para cada semáforo (figura 4). Estes colisores permitem ao agente obter informações sobre a quantidade de veículos presentes em cada via, fornecendo dados para a tomada de decisões, estipulou-se a separação dos colisores em: A e B para o grupo A e C e D para o grupo B, determinando assim suas respectivas vias.

Figura 4 – Representação dos colisores do agente inteligente em amarelo e suas respectivas denominações.



Fonte: Do autor.

O agente foi configurado para a tomada de decisões com um conjunto de quatro observações e três possíveis ações. As observações são baseadas na contagem de veículos presentes na via em que o semáforo está localizado. As ações disponíveis correspondem à abertura do semáforo para cada grupo de veículos, bem como uma ação vazia que tem por objetivo incentivar o agente a realizar uma quantidade menor de ações.

O agente recebe uma recompensa de 0.32 pontos quando o número de veículos em cada grupo é menor que seis, que é dois terços dos possíveis veículos na via, o que visa encorajá-lo a manter o fluxo de tráfego constante. Por outro lado, ele recebe uma penalidade de -0.04 pontos quando um grupo possui menos de seis

veículos e o outro grupo tem mais de dez veículos, evitando assim que o agente beneficie apenas uma das vias.

Além disso, o agente é recompensado com 0,32 pontos ao executar a ação de trocar o semáforo em um intervalo de sete segundos. Isso garante que o tempo de transição entre as configurações do semáforo não seja muito curto, a fim de evitar acidentes. Por fim, o agente recebe uma recompensa positiva de 0.04 pontos quando realiza a ação vazia.

Os pesos das recompensas foram determinados com base em testes conduzidos durante o processo de treinamento, resultando na conclusão de que a maximização do aprendizado do agente ocorre ao manter a recompensa em 1 ponto quando todas as condições estão corretas. Essa configuração têm o objetivo de orientar o agente a tomar decisões que levem a um fluxo de tráfego equilibrado e seguro, considerando o número de veículos em cada grupo, o tempo de troca dos semáforos e a realização de ações vazias de forma adequada.

Para o treinamento do agente, foi estabelecida uma taxa fixa de geração de veículos, com um novo veículo sendo gerado a cada intervalo de cinco segundos. Cada gerador de carros tem uma capacidade máxima de doze veículos, totalizando um máximo de quarenta e oito carros presentes na cena. Além disso, o destruidor de veículos foi programado para começar sua ação após cinquenta segundos e repetir a cada intervalo de cinco segundos.

Essas configurações permitem simular um ambiente em que os carros são gerados em intervalos regulares e limitados, enquanto o destruidor de veículos atua a cada cinco segundos para remover um carro da cena. Essa dinâmica, de geração e remoção de veículos, cria um fluxo de tráfego em constante mudança, proporcionando desafios para que o agente aprenda a tomar decisões adequadas em relação à abertura dos semáforos e ao controle do tráfego.

Os tempos de geração e exclusão de veículos foram definidos por meio dos testes realizados no desenvolvimento da simulação, uma quantidade excessiva de veículos em cena, prejudica o desempenho computacional da simulação e pode ser difícil para o agente aprender e tomar decisões adequadas, devido à alta densidade e complexidade do tráfego. Portanto, limitar a quantidade de veículos em cena é uma estratégia para equilibrar o desempenho da simulação e garantir que o agente possa aprender de forma eficaz, enfrentando um fluxo de tráfego desafiador, mas gerenciável.

Foram realizadas diversas iterações de treinamento do agente, e por meio das falhas ocorridas, foi possível refinar o sistema de recompensas do agente. O que é corroborado por Rubak (2021), quando afirma que: Ao ajustar os pesos das recompensas, é importante encontrar um equilíbrio que leve o agente a aprender de forma eficaz, evitando comportamentos indesejados. Se as recompensas positivas forem muito altas em relação às negativas, o agente pode optar por ações que levem a ganhos rápidos, mas não sejam eficientes em relação ao objetivo geral. Por outro lado, se as recompensas negativas forem excessivamente altas, o agente pode ficar desencorajado e ter dificuldade em explorar diferentes estratégias.

Portanto, encontrar o balanço adequado dos pesos das recompensas é crucial para obter um treinamento eficiente e levar o agente a aprender comportamentos desejados em ambiente específico. Esse processo envolve um ajuste iterativo e experimentação, buscando alcançar um equilíbrio que resulte em um agente que tome ações consistentes com os objetivos estabelecidos.

Durante as fases iniciais do treinamento, foi identificado um problema em que o agente, em determinado ponto, deixava de executar ações. Inicialmente, os pesos atribuídos quando o número de veículos em cada grupo era inferior a seis eram de 1 ponto. Isso levava o agente a parar de tomar ações, interrompendo o fluxo de tráfego, mas garantindo um ganho constante. Para solucionar esse problema, foi estabelecido um peso menor, de 0,32 por grupo, que, somado, totalizava 0,64, assegurando que ambas as ações fossem igualmente importantes.

No entanto, tais mudanças mitigaram o erro até certo ponto, indicando que apenas essa modificação não seria suficiente para evitar que o agente beneficiasse um semáforo. Introduziu-se um peso negativo, de -0,04 quando um grupo possuía menos de seis veículos e o outro grupo possuía mais de dez veículos, que representava a situação em que o semáforo não estava alternando corretamente.

Além disso, encontrou-se um problema de alternância excessiva dos semáforos, visando maximizar os ganhos. Para lidar com essa questão, foi implementada uma recompensa positiva quando o agente executava ações com uma duração mínima de sete segundos que somado aos quatro segundos de amarelo, totaliza onze segundos, que de acordo com o DENATRAN (2014) está entre os valores típicos observados na prática (variando entre 10 e 20 segundos), em relação aos tempos de segurança, valores mínimos aceitáveis para a duração dos períodos de sinal verde. Além disso, uma terceira ação foi introduzida, concedendo 0,04 pontos ao

agente, com o objetivo de incentivar a redução do número de ações executadas e maximizar seus ganhos.

O treinamento final do agente, utilizado para a realização dos testes desenvolvidos nesta pesquisa, totalizou quinhentas mil etapas, com uma duração total de aproximadamente cinco horas. Cada etapa do treinamento envolveu interações do agente com o ambiente. Durante o treinamento, foram coletadas informações sobre o desempenho do agente, incluindo a média e o desvio padrão das recompensas obtidas. Durante esse período, o agente apresentou um progresso significativo, melhorando seu desempenho ao longo do tempo.

No início do treinamento, a média das recompensas obtidas pelo agente foi de aproximadamente 150,971, com um desvio padrão de 16,185. À medida que o treinamento prosseguia, a média das recompensas diminuiu para cerca de 105,700 na etapa cem mil, entretanto em seguida começou a aumentar. Ao final do treinamento, o agente alcançou uma média de recompensa de aproximadamente 151,788, com um desvio padrão de 12,565 indicando que aprendeu a realizar suas ações de forma consistente, com resultados previsíveis e satisfatórios.

4 RESULTADOS E DISCUSSÃO

Durante a realização dos testes, foram coletados os seguintes dados: quantidade de carros que passaram no cruzamento, fluxo de carros por minuto e indicador de congestionamento, obtido por meio da média de tempo na fila após vinte segundos em espera, determinado com base nas observações realizadas nos testes, nos quais foi constatado que a formação de filas de congestionamento mais significativas ocorre quando o tempo médio de espera excede vinte segundos.

O indicador de congestionamento supervisiona o tráfego veicular em quatro vias que são equipadas com semáforos, realizando cálculos do tempo médio de espera na fila e registrando a quantidade de ocasiões em que esse valor excede vinte segundos, o que indica a ocorrência de congestionamento. Quando há um fluxo intenso de veículos e a capacidade das vias não é suficiente para acomodar de maneira eficiente todos os veículos, a fila de veículos tende a se alongar e, consequentemente, o tempo de espera na fila aumenta, caracterizando um congestionamento de grande magnitude. Essa métrica tem o propósito de identificar a frequência com que o congestionamento se manifesta. Caso esse valor seja

elevado, tal fato indica a recorrência do congestionamento, sugerindo a existência de problemas relacionados à capacidade das vias ou à gestão do tráfego.

A coleta de dados foi executada por meio de três testes de avaliação, onde cada iteração apresentou uma duração de dez minutos em tempo real. Para efeitos de análise, a velocidade do processo foi acelerada em dez vezes, resultando em um período total de coleta de dados de uma hora e quarenta minutos a cada iteração. O tempo de segurança entre a troca de sinalização foi de quatro segundos para ambos os modelos e o tempo de duração do semáforo em tempo fixo foi definido como trinta segundos.

No primeiro cenário de teste, o fluxo de carros foi mantido constante em todas as vias. A taxa de geração de carros foi configurada para um carro a cada cinco segundos, com um limite máximo de doze carros por gerador, totalizando um máximo de quarenta e oito carros em cena. Os destruidores de carros foram programados para iniciar sua atuação após cinquenta segundos do início da simulação, removendo um carro a cada intervalo de cinco segundos. Essa estratégia foi adotada com o objetivo de garantir que os carros sejam inseridos na cena inicialmente e, em seguida, permaneçam constantes ao longo do tempo.

A seguir são apresentados os resultados do primeiro cenário (tabela 1 e tabela 2).

Tabela 1 – Resultado do desempenho do agente inteligente no primeiro cenário de teste.

Tempo	10 Minutos	10 Minutos	10 Minutos	Média
Fluxo	63,08	62,36	63	62,8
Quantidade de carros	6243	6168	6234	6215,0
Congestionamento A	0	0	2	0,7
Congestionamento B	0	0	3	1,0
Congestionamento C	0	170	141	103,7
Congestionamento D	0	115	0	38,3

Fonte: Do autor.

Tabela 2 – Resultado do desempenho do semáforo a tempo fixo no primeiro cenário de teste.

Tempo	10 Minutos	10 Minutos	10 Minutos	Média
Fluxo	56,76	57,14	57,34	57,1
Quantidade de carros	6020	5976	5982	5992,7
Congestionamento A	1185	988	946	1039,7

Congestionamento B	1198	921	503	874,0
Congestionamento C	895	1063	1063	1007,0
Congestionamento D	1550	585	431	855,3

Fonte: Do autor.

Ao comparar-se o desempenho entre o funcionamento do semáforo convencional e o semáforo controlado por um agente inteligente, constatou-se que o agente inteligente exibiu uma melhora significativa no fluxo de veículos. A diferença entre os dois fluxos foi estimada em aproximadamente 10,03% em relação ao semáforo de tempo fixo. No que diz respeito ao indicador de congestionamento, o semáforo de tempo fixo demonstrou uma frequência consideravelmente maior de ocorrências de congestionamento, o que indica um desempenho inferior em comparação ao agente inteligente.

No segundo cenário de testes, manteve-se o fluxo constante em três das vias, com a taxa de geração de um carro a cada cinco segundos e um número máximo de dez carros por gerador. Para a quarta via (via A pertencente ao grupo A), utilizou-se um gerador com taxa de um carro a cada dois segundos e um número máximo de dezoito carros, mantendo um total máximo de quarento e oito carros em cena.

Os dados deste cenário são apresentados nas tabelas 3 e 4.

Tabela 3 – Resultado do desempenho do agente inteligente no segundo cenário de teste.

Tempo	10 Minutos	10 Minutos	10 Minutos	Média
Fluxo	70,84	71,18	70,63	70,9
Quantidade de carros	7041	7042	7037	7040,0
Congestionamento A	0	0	0	0,0
Congestionamento B	0	0	0	0,0
Congestionamento C	1535	615	871	1007,0
Congestionamento D	1107	336	516	653,0

Fonte: Do autor.

Tabela 4 – Resultado do desempenho do semáforo a tempo fixo no segundo cenário de teste.

Tempo	10 Minutos	10 Minutos	10 Minutos	Média
Fluxo	59,44	59,51	58,93	59,3
Quantidade de carros	5927	5931	5842	5900,0
Congestionamento A	8278	8259	7591	8042,7
Congestionamento B	1589	1242	1549	1460,0

Congestionamento C	1101	854	1691	1215,3
Congestionamento D	363	769	765	632,3

Fonte: Do autor.

Ao aumentar o fluxo de veículos em apenas uma via, notou-se, no método convencional, um considerável incremento nos casos de congestionamento na via (A), com maior fluxo. Por meio do uso de um agente inteligente, registrou-se um acréscimo de 19,56% no fluxo de veículos. Observou-se ainda, que na via com maior fluxo, a ocorrência de congestionamento foi eliminada, o que indica que o agente inteligente priorizou à via com maior fluxo (e, por extensão, à via (B) oposta a mesma). Esse comportamento resultou em uma maior quantidade de congestionamentos nas vias C e D, pertencentes ao grupo oposto, onde o fluxo foi aumentado.

Para o terceiro cenário de testes, manteve-se o fluxo constante em duas das vias, com a taxa de geração de um carro a cada cinco segundos e um número máximo de dez carros por gerador. Para a terceira via (A pertencente ao grupo A) e quarta via (C pertencente ao grupo B), utilizou-se um gerador com taxa de um carro a cada dois segundos e um número máximo de catorze carros, mantendo um total máximo de quarento e oito carros em cena.

Os resultados do terceiro cenário estão apresentados nas tabelas 5 e 6.

Tabela 5 – Resultado do desempenho do agente inteligente no terceiro cenário de teste.

Tempo	10 Minutos	10 Minutos	10 Minutos	Média
Fluxo	72,65	73,25	72,48	72,8
Quantidade de carros	7188	7373	7172	7244,3
Congestionamento A	0	21	54	25,0
Congestionamento B	0	25	35	20,0
Congestionamento C	164	19	66	83,0
Congestionamento D	108	244	108	153,3

Fonte: Do autor.

Tabela 6 – Resultado do desempenho do semáforo a tempo fixo no terceiro cenário de teste.

Tempo	10 Minutos	10 Minutos	10 Minutos	Média
Fluxo	60,26	59,27	59,38	59,6
Quantidade de carros	6077	5904	5917	5966,0
Congestionamento A	5963	6480	5929	6124,0
Congestionamento B	1250	957	1532	1246,3

Congestionamento C	4371	4213	4975	4519,7
Congestionamento D	1628	1371	979	1326,0

Fonte: Do autor.

O desempenho do agente foi semelhante ao primeiro cenário, apresentando um aumento de 22,15% em relação ao método convencional, com ocorrências de congestionamento reduzidas. Esses resultados eram esperados devido à coordenação dos semáforos, permitindo que o agente atuasse de maneira semelhante ao primeiro cenário, porém com um leve aumento no fluxo de veículos. Por outro lado, o uso de semáforos de tempo fixo mostrou-se prejudicado pelo aumento do fluxo de veículos em ambas as vias, resultando em um considerável aumento nos congestionamentos destas vias.

Embora o agente tenha recebido treinamento em um ambiente de fluxo constante, seu desempenho no gerenciamento do tráfego revelou-se positivo em todas as situações, obtendo resultados semelhantes as pesquisas de Cortez (2022) e Wei et al (2018). Vale destacar que o agente também se encaixa nos aspectos mencionados por Bazzan (2021), por utilizar aprendizado por reforço em seu treinamento e aplicar no controle semafórico, contribuindo para melhora no tráfego.

5 CONCLUSÃO

Nessa pesquisa aplicou-se o uso de aprendizado por reforço, utilizando o algoritmo de Otimização de Política Proximal, por meio de uma simulação, para a otimização do tráfego urbano em interseções semafóricas. Na avaliação de desempenho do agente realizou-se a comparação de desempenho do método de gerenciamento de semáforos baseado em aprendizado por reforço com o método convencional de semáforos a tempos fixos atualmente utilizado no Brasil.

Os resultados da abordagem proposta, em suma, demonstraram um desempenho superior em relação ao método convencional, sugerindo a viabilidade do uso do algoritmo de Otimização de Política Proximal para o treinamento de agentes inteligentes. Foi constatada eficácia no gerenciamento de trânsito, uma vez que foram observadas adaptações bem-sucedidas aos cenários de teste, resultando em melhorias significativas. Especificamente, foram observados aumentos de 10,03%,

19,56% e 22,15% em comparação ao método de semáforos com tempos fixos, nas respectivas análises.

As dificuldades encontradas no processo de treinamento do agente foram superadas de maneira efetiva através da aplicação de um balanceamento cuidadoso das recompensas, levando em consideração observações obtidas durante testes empíricos, bem como por meio de uma análise da literatura especializada. Essa abordagem permitiu ajustar de forma adequada as recompensas fornecidas ao agente durante o treinamento.

Como trabalhos futuros na área, sugere-se a validação do agente em novos cenários, baseados em dados reais. Ademais, sugere-se que os testes sejam feitos em um cenário com maior número de parâmetros como pedestres, caminhões e condições adversas no trânsito como acidentes, a fim de garantir maior precisão nos resultados obtidos.

REFERÊNCIAS

ALVES, Maria Abadia; DIAS, Ricardo Cunha; SEIXAS, Paulo Castro. *Smart Cities* no Brasil e em Portugal: o estado da arte. **urbe. Revista Brasileira de Gestão Urbana**, v. 11, p. e20190061, 2019. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2175-33692019000100411&tlng=pt>. Acesso em: 22 out. 2022.

BAZZAN, Ana L. C. Contribuições de aprendizado por reforço em escolha de rota e controle semafórico. **Estudos Avançados**, v. 35, n. 101, p. 95–110, 2021. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142021000100095&tlng=pt>. Acesso em: 6 jun. 2023.

CORTEZ, Diogo Eugênio da Silva. **Desenvolvimento de um sistema de controle de tráfego inteligente baseado em visão computacional**. 2022. Dissertação (Mestrado em Tecnologia da Informação) - Universidade Federal do Rio Grande do Norte, [S. l.], 2022. Disponível em: <https://repositorio.ufrn.br/bitstream/123456789/47158/1/Desenvolvimentosistemacontrole_Cortez_2022.pdf>. Acesso em: 2 jun. 2022.

DEPARTAMENTO NACIONAL DE TRÂNSITO (DENATRAN). **Manual de Sinalização de Trânsito Volume V – Sinalização Semafórica**. Brasília: CONTRAN, 2014.

JULIANI, Arthur; BERGES, Vincent-Pierre; TENG, Ervin; et al. Unity: A General Platform for Intelligent Agents. 2018. Disponível em: <https://arxiv.org/abs/1809.02627>. Acesso em: 3 nov. 2022.

MENENDEZ, Oisy Hernandez; SILVA, Natalia Assunção Brasil; PITANGA, Heraldo Nunes. Análise estatística aplicada à gestão do tráfego em interseção semafórica. **Research, Society and Development**, v. 11, n. 3, 7 fev. 2022. DOI http://dx.doi.org/10.33448/rsd-v11i3.26178. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/26178>. Acesso em: 3 abr. 2022.

OPENAI. **Proximal Policy Optimization**. Disponível em: <https://openai.com/research/openai-baselines-ppo>. Acesso em: 25 out. 2022.

OLIVEIRA, Michel Bruno Wanderley; ARAÚJO , Heygon Henrique Fernandes. APLICAÇÃO DE SIMULAÇÃO EM UM CRUZAMENTO RODOVIÁRIO. **XLVIII Simpósio Brasileiro de Pesquisa Operacional**, Vitória, ES., p. 2851-2862, 30 nov. 2016. Disponível em: <http://www.din.uem.br/sbpo/sbpo2016/pdf/156698.pdf>. Acesso em: 22 abr. 2022.

RUBAK, Mario. **Imitation Learning with the Unity Machine Learning Agents Toolkit / vorgelegt von: Mario Rubak**. Wien, 2021. Disponível em: <http://pub.fh-campuswien.ac.at/obvfcwhsacc/6294692>. Acesso em: 3 nov. 2022.

UNITY. **Unity Real-Time Development Platform | 3D, 2D VR & AR Engine**. Disponível em: <https://unity.com/>. Acesso em: 25 out. 2022.

VIEIRA, Guilherme Ernani. Uma revisão sobre a aplicação de simulação computacional em processos industriais. **XIII SIMPEP**, São Paulo, n. 13, 6 nov. 2006. Disponível em: <https://simpep.feb.unesp.br/anais/anais_13/artigos/676.pdf>. Acesso em: 6 jun. 2022.

WEI, H. et al. **IntelliLight: A Reinforcement Learning Approach for Intelligent Traffic Light Control**. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. **Anais**... In: KDD ’18: THE 24TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. London United Kingdom: ACM, 19 jul. 2018. Disponível em: <https://dl.acm.org/doi/10.1145/3219819.3220096>. Acesso em: 28 jun. 2022.