UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM MESTRADO E DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

MESTRADO ACADÊMICO EM CIÊNCIA DA COMPUTAÇÃO

RÔMULO FREIRE FÉRRER FILHO

TREINANDO AGENTES DE FORRAGEAMENTO MULTISSENSORIAIS USANDO CURRÍCULO

FORTALEZA

2024

RÔMULO FREIRE FÉRRER FILHO

TREINANDO AGENTES DE FORRAGEAMENTO MULTISSENSORIAIS USANDO CURRÍCULO

Dissertação apresentada ao Curso de Mestrado Acadêmico em Ciência da Computação do Programa de Pós-Graduação em Mestrado e Doutorado em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Computação Gráfica.

Orientador: Prof. Dr. Joaquim Bento Cavalcante Neto.

FORTALEZA

2024

RÔMULO FREIRE FÉRRER FILHO

TREINANDO AGENTES DE FORRAGEAMENTO MULTISSENSORIAIS USANDO CURRÍCULO

Dissertação apresentada ao Curso de Mestrado Acadêmico em Ciência da Computação do Programa de Pós-Graduação em Mestrado e Doutorado em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Computação Gráfica.

Aprovada em: 29 de Novembro de 2024

BANCA EXAMINADORA

Prof. Dr. Joaquim Bento Cavalcante Neto (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Yuri Lenon Barbosa
Universidade Federal do Ceará (UFC)

Prof. Dr. Creto Augusto Vidal
Universidade Federal do Ceará (UFC)

Prof. Dr. José Gilvan Rodrigues Maia
Universidade Federal do Ceará (UFC)

RESUMO

O Aprendizado por Reforço Profundo (ARP), do inglês *Deep Reinforcement Learning* (DRL) tem demonstrado grande sucesso no desenvolvimento de agentes capazes de resolver tarefas complexas em jogos. No entanto, a maioria dos agentes de jogos utiliza apenas sensores visuais para coletar informações sobre o ambiente. Trabalhos mais recentes mostraram que agentes que utilizam sensores auditivos podem ter um desempenho superior em comparação com agentes que dependem apenas da visão. Neste trabalho, avalia-se a capacidade de utilizar múltiplos sensores em uma tarefa de forrageamento. Também é proposta uma estratégia de treinamento baseada em currículo para desenvolver agentes que utilizam o áudio de forma eficaz como fonte de informação em cenários de forrageamento. Primeiramente, demonstra-se que agentes com capacidades visuais e auditivas apresentam um desempenho semelhante ao de agentes com apenas sensor visual, indicando que os primeiros ignoram o áudio. Em seguida, é mostrado que, ao utilizar um currículo de dificuldade gradualmente crescente, o agente consegue usar de forma competente as informações auditivas disponíveis, tornando-se mais robusto para sobreviver em cenários onde as informações visuais não estão disponíveis. Os resultados indicam que agentes podem ser treinados para utilizar o áudio como uma fonte de informação de forma eficaz, utilizando uma estratégia de treinamento baseada em currículo, melhorando assim sua capacidade de lidar com tarefas mais variadas em comparação com agentes que possuem apenas visão.

Palavras-chave: Aprendizado por Reforço Profundo; Agentes Inteligentes; Aprendizado por Currículo; Agentes Multissensoriais; Forrageamento

ABSTRACT

Deep Reinforcement Learning (DRL) has demonstrated significant success in the development of agents capable of solving complex game tasks. However, the majority of game agents rely solely on visual sensors to gather information about the environment. Recent studies have indicated that agents utilizing audio sensors can outperform those that depend exclusively on vision. This work evaluates the ability to utilize multiple sensors in a foraging task. It also proposes a curriculum-based training strategy to develop agents that effectively use audio as a source of information in foraging scenarios. First, it is demonstrated that agents with both visual and auditory capabilities perform similarly to agents with only visual sensors, indicating that the former ignores the audio input. Initially, it is demonstrated that agents equipped with both vision and hearing capabilities perform similarly to agents with only visual sensors, suggesting that the former tend to ignore the audio input. Subsequently, it is shown that by implementing a gradually increasing difficulty curriculum, the agent effectively utilizes the available audio information, enhancing its robustness in scenarios where visual information is absent. The results indicate that agents can be trained to effectively use audio as a source of information through a curriculum-based training strategy, thereby improving their capacity to handle a wider range of tasks compared to agents with only vision.

Keywords: Deep Reinforcemente Learning; Intelligent Agents; Curriculum Learning; Multisensory Agents; Foraging

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ALE	*Atari Learning Environment*
AR	Aprendizado por Reforço
ARP	Aprendizado por Reforço Profundo
DL	*Deep Learning*
DQN	*Deep Q-Learning*
DRL	*Deep Reinforcement Learning*
FFT	*Fast Fourier Transform*
GAE	*Generalized Advantage Estimation*
HG	*Health Gathering*
HGS	*Health Gathering Supreme*
MDP	*Markov Decision Process*
PPO	*Proximal Policy Optimization*
RL	*Reinforcement Learning*
TD	*Temporal Difference*

LISTA DE SÍMBOLOS

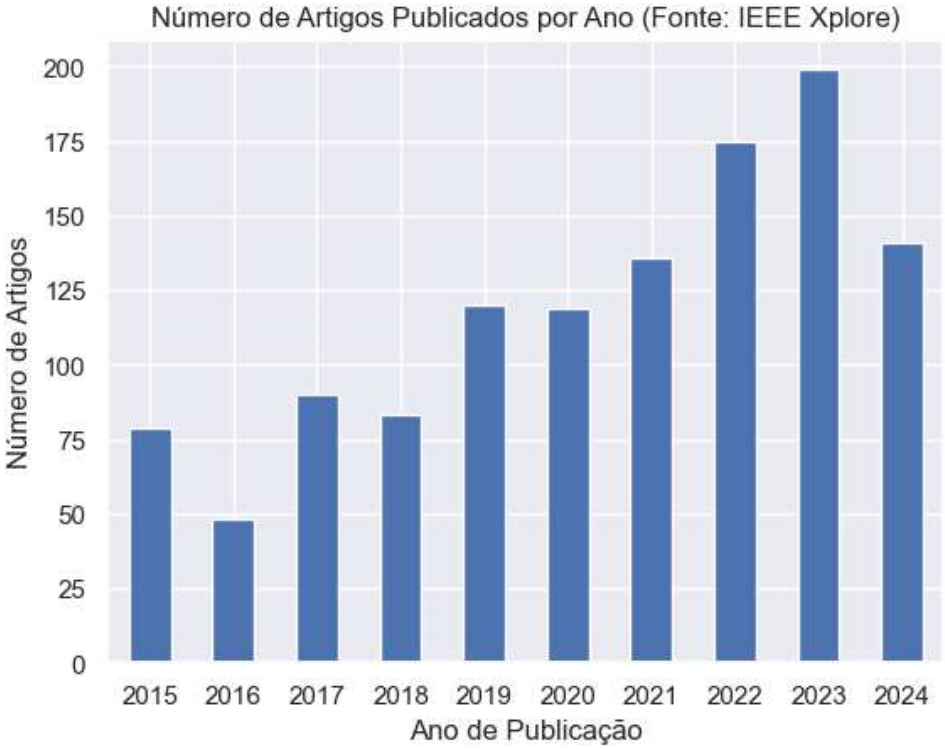
S	Conjunto de Estados
A	Conjunto de Ações
P	Função de Transição
R	Função de Recompensa R
π	Política que mapeia estados em ações
θ	Parâmetros da Rede Neural
ε	Hiperparâmetro que controla o intervalo de clipping

SUMÁRIO

1 INTRODUÇÃO

A pesquisa sobre o desenvolvimento de agentes inteligentes em jogos tem crescido significativamente desde o surgimento do *Deep Reinforcement Learning* (DRL) (MNIH *et al.*, 2015), como visto pelo número de publicações sobre o tema apresentado na Figura 1.

Figura 1 – Artigos publicados por ano.



Fonte: Dados retirados da plataforma IEEE Xplore[1] usando as palavras-chave *intelligent agents* e *games*. Imagem elaborada pelo autor.

Essa abordagem revolucionou a maneira como os agentes aprendem a interagir com ambientes complexos, permitindo a conquista de jogos clássicos como Xadrez e Go (SILVER *et al.*, 2017), além de jogos icônicos da *Atari* (SCHRITTWIESER *et al.*, 2019). O avanço das técnicas de DRL também possibilitou o domínio de jogos digitais complexos, como *Dota 2* (OPENAI *et al.*, 2019) e *Starcraft 2* (VINYALS *et al.*, 2019), evidenciando o potencial dessa metodologia para resolver problemas desafiadores.

Os agentes de DRL são impulsionados por redes neurais convolucionais profundas, que se destacam em tarefas visuais, permitindo que esses agentes aprendam a jogar jogos do zero, sem qualquer conhecimento prévio das regras. Essa capacidade de aprendizado a partir da experiência oferece uma nova perspectiva sobre como a inteligência artificial pode ser aplicada

[1] https://ieeexplore.ieee.org

em ambientes dinâmicos e imprevisíveis.

Por outro lado, historicamente, o processo de aprendizado humano e animal é intrinsecamente baseado na integração dos diferentes sentidos, como visão, audição, olfato, paladar e tato, permitindo que esses seres percebam e interpretem o mundo ao seu redor de maneira abrangente e precisa (KEELEY, 2002). Essa multimodalidade sensorial é fundamental para a sobrevivência, pois permite a detecção de recursos, a identificação de perigos e a interação social, evidenciando como a combinação de estímulos sensoriais enriquece a experiência de aprendizado e aprimora a capacidade de adaptação a ambientes dinâmicos (MUNOZ; BLUMSTEIN, 2012).

Embora o paradigma tradicional de Aprendizado por Reforço tenha sido inspirado no aprendizado animal, a maioria das pesquisas em DRL tem se concentrado em tarefas predominantemente visuais. Apesar dos avanços significativos no desenvolvimento de agentes para jogos, esses sistemas ainda dependem consideravelmente da entrada visual. Em diversas aplicações, os agentes recebem informações numéricas adicionais, como o nível de energia atual ou a disponibilidade de recursos (munição, poções, etc.) e utilizam sensores proprioceptivos para coletar dados sobre seu próprio estado (BAKER *et al.*, 2020).

No entanto, os estudos que exploram a utilização de capacidades auditivas para obter informações adicionais sobre o ambiente ainda são escassos (WOUBIE *et al.*, 2019; HEGDE *et al.*, 2021; GAINA; STEPHENSON, 2019; GAN *et al.*, 2019; GIANNAKOPOULOS *et al.*, 2021; CHEN *et al.*, 2020), tais trabalhos serão apresentados em mais detalhes na Seção 3.2. Essa limitação destaca uma lacuna importante na pesquisa, pois a integração de diferentes modalidades sensoriais poderia enriquecer o desempenho dos agentes, permitindo uma percepção mais robusta e uma tomada de decisão mais informada em cenários complexos. A inclusão da audição como um componente na percepção do ambiente pode aumentar a adaptabilidade dos agentes em contextos dinâmicos, refletindo mais fielmente as capacidades de aprendizado observadas em seres humanos e animais.

Em contraste, a música e os efeitos sonoros desempenham um papel significativo em vários aspectos dos videogames. Desde os efeitos simples presentes no primeiro jogo comercial que já utilizava sons, *Computer Space*, de 1971 (BURCHAM, 1996), até a proposta de jogos exclusivamente auditivos (HUGILL; AMELIDES, 2016), os avanços tecnológicos permitiram que abordagens modernas aproveitassem o desenvolvimento de design sonoro complexo para promover uma variedade de interações multissensoriais, impactando os jogadores de diferentes maneiras (GRIMSHAW *et al.*, 2013). Por exemplo, o uso diversificado de áudio pode aumentar

consideravelmente a imersão do jogador, provocando reações estéticas e emocionais, além de facilitar a acessibilidade e a inclusão (GUILLEN *et al.*, 2021).

Além disso, no mundo natural, os organismos utilizam múltiplas modalidades sensoriais para coletar informações. Em particular, a audição é uma fonte importante de dados que auxilia, por exemplo, na navegação ambiental, na busca por recursos e na detecção de presas ou predadores. Em tarefas de forrageamento, as informações auditivas disponíveis são essenciais para ajudar os animais a reunir alimento enquanto evitam obstáculos. Inspirados por essa capacidade biológica, esforços recentes têm buscado dotar agentes de jogo com percepção auditiva, permitindo que eles compreendam e interajam melhor com seu entorno (GAINA; STEPHENSON, 2019; PARK *et al.*, 2021).

De maneira semelhante, incluir áudio como fonte de informação também é benéfico para agentes de videogame. Trabalhos anteriores demonstraram que agentes que escutam fontes auditivas podem ter um desempenho superior em comparação com agentes que utilizam apenas visão (HEGDE *et al.*, 2021). No entanto, simplesmente treinar um agente que receba informações visuais e auditivas pode não ser suficiente para que ele utilize o áudio de forma eficaz. Um desafio difícil é aprender a usar corretamente as diversas informações fornecidas por diferentes entradas sensoriais. Ao contrário dos ambientes tradicionais de DRL, onde geralmente a entrada visual predomina, os ambientes multissensoriais exigem que os agentes processem e integrem informações de múltiplas fontes simultaneamente.

1.1 Objetivos do trabalho

Dado o contexto apresentado anteriormente, este trabalho tem como objetivo principal propor uma estratégia de currículo para treinar agentes multissensoriais que utilizam de forma eficaz tanto o áudio quanto a visão como fontes de informação em uma tarefa de forrageamento. Com tal fito, determinam-se os seguintes objetivos específicos:

- Demonstrar que agentes são capazes de aprender utilizando apenas o som como fonte de informação;
- Mostrar que agentes com capacidades visuais e auditivas conseguem sobreviver com sucesso em cenários baseados em forrageamento;
- Evidenciar que agentes que possuem apenas visão apresentam um desempenho semelhante ao dos agentes com ambas as capacidades em cenários que podem ser resolvidos apenas com informações visuais, indicando que o áudio é ignorado;

– Apresentar como um currículo pode ser utilizado para treinar gradualmente agentes, passando de percepção monossensorial para multissensorial, permitindo-lhes navegar em ambientes que podem incluir informações visuais ou apenas sonoras;
– Por fim, demonstrar que, embora o desempenho seja semelhante em cenários que podem ser resolvidos apenas com informações visuais, um agente treinado para utilizar o áudio de forma eficaz também terá sucesso em cenários sem informações visuais, enquanto outros agentes ficarão paralisados.

A Figura 2 apresenta uma visão geral do trabalho realizado.

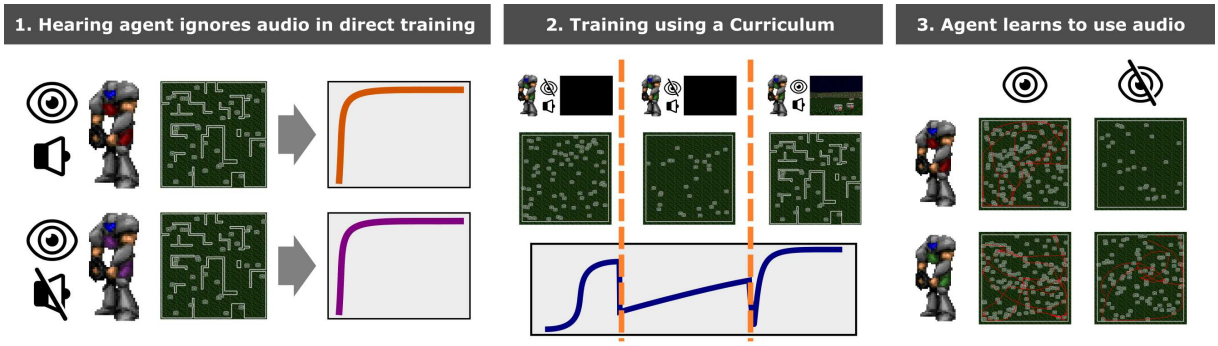


Figura 2 – Visão geral do trabalho desenvolvido. O primeiro passo é o treinamento direto de um agente multissensorial e sua comparação com o agente monossensorial. Em seguida, treina-se o agente utilizando currículo. Por último, comparam-se os agentes, demonstrando a robustez do currículo proposto.

1.2 Organização do trabalho

Esta dissertação está dividida em 6 capítulos da seguinte forma. No Capítulo 2, é apresentada uma visão geral de conceitos importantes dos tópicos relevantes para este trabalho. No Capítulo 3, é realizada uma revisão da literatura existente sobre DRL multissensorial, apresentando agentes e ambientes que tratam do tema. No Capítulo 4, é introduzida a abordagem proposta para treinar agentes de DRL multissensoriais utilizando aprendizado por currículo, juntamente com a descrição da configuração experimental e dos resultados. No Capítulo 5, são discutidos os resultados e as descobertas sobre os mecanismos subjacentes que impulsionam o aprendizado em contextos multissensoriais. Por fim, no Capítulo 6, o trabalho é concluído e são delineadas direções futuras.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão apresentados conceitos-chaves para a compreensão do trabalho desenvolvido. Inicialmente, serão descritos os métodos de aprendizado, como Aprendizado por Reforço, do inglês *Reinforcement Learning* (RL) e Aprendizado por Reforço Profundo, do inglês *Deep Reinforcement Learning* (DRL), bem como o algoritmo utilizado neste trabalho, Otimização de Política Proximal, do inglês *Proximal Policy Optimization* (PPO) , e a técnica de Aprendizagem Curricular. Em seguida apresentar-se-ão os ambientes e tarefas utilizados como base para o treinamento de agentes e a biblioteca de algoritmos de DRL, *Sample Factory*.

2.1 Métodos de Aprendizado

Nesta seção, serão abordados os métodos de aprendizado essenciais para agentes inteligentes, destacando o *Reinforcement Learning* (RL) e sua extensão, o *Deep Reinforcement Learning* (DRL). Também será apresentado o algoritmo *Proximal Policy Optimization* (PPO), utilizado no desenvolvimento do trabalho, e a técnica de *Curriculum Learning*.

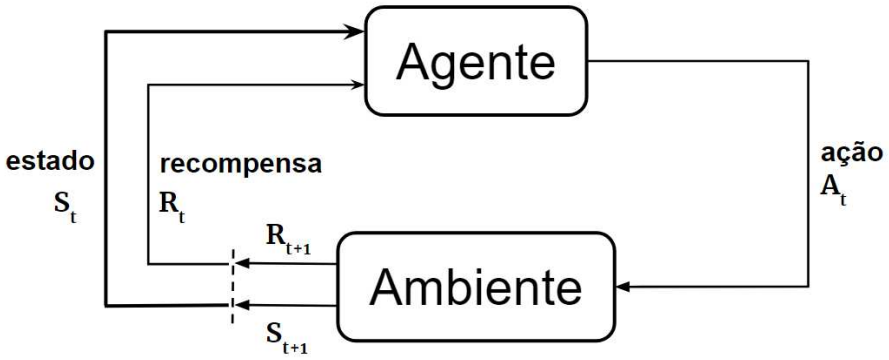
2.1.1 Aprendizado por Reforço

O Aprendizado por Reforço é um paradigma de aprendizado de máquina que se concentra em como agentes devem executar ações em um ambiente de modo a maximizar alguma recompensa cumulativa (SUTTON; BARTO, 2018). Este paradigma é inspirado na psicologia comportamental e é essencialmente diferente de outros métodos de aprendizado de máquina, como o aprendizado supervisionado e não supervisionado. Enquanto no aprendizado supervisionado o modelo aprende a partir de um conjunto de dados etiquetados, no RL o agente aprende a partir das interações com o ambiente, recebendo recompensas ou penalidades como resposta.

Formalmente, um problema de *Reinforcement Learning* pode ser modelado como um Processo de Decisão de Markov (*Markov Decision Process* (MDP)), definido por um conjunto de estados S, um conjunto de ações A, uma função de transição P, que descreve a probabilidade de transição de um estado para outro, e uma função de recompensa R. O objetivo do agente é encontrar uma política π, uma função que mapeia estados em ações, que maximiza a recompensa esperada a longo prazo. A Figura 3 demonstra um fluxo de interação entre o Agente e o Ambiente, entidades que foram descritas por Sutton e Barto (2018). Em um instante t, o agente recebe um

estado S_t do ambiente e em seguida executa uma ação A_t. O ambiente então muda de estado, levando o agente para S_{t+1} e retornando uma recompensa R_{t+1}.

Figura 3 – Diagrama de interação entre Agente e Ambiente.



Fonte: Adaptado de Sutton e Barto (2018).

A função de valor, que quantifica o retorno esperado a partir de um estado específico ou de um par estado-ação, desempenha um papel fundamental no RL. Para aprender a função de valor, várias abordagens podem ser empregadas, como o método de Diferença Temporal (*Temporal Difference* (TD)), que realiza atualizações incrementais baseadas nas transições observadas no ambiente. Algoritmos amplamente reconhecidos no RL, como *Q-Learning* (WATKINS; DAYAN, 1992) e *SARSA* (RUMMERY; NIRANJAN, 1994), fazem uso de técnicas baseadas na função de valor para refinarem suas políticas de ação de maneira eficiente e efetiva.

A eficácia dos algoritmos de RL depende da capacidade do agente de generalizar a partir de experiências passadas para novos estados e situações. Para isso, são frequentemente utilizadas funções de aproximação, como redes neurais, que permitem ao agente lidar com espaços de estado e ação de alta dimensionalidade.

2.1.2 Aprendizado por Reforço Profundo

O Aprendizado por Reforço Profundo combina o RL tradicional com técnicas de Aprendizado Profundo, (*Deep Learning* (DL)), utilizando redes neurais profundas para aproximar a política π ou a função de valor (MNIH *et al.*, 2013; MNIH *et al.*, 2015), possibilitando que os agentes lidem com problemas de alta dimensionalidade e complexidade. Essa técnica tem se mostrado uma alternativa prática para a utilização de RL em aplicações reais, obtendo sucessos notáveis em tarefas complexas, como jogos, robótica e controle de sistemas dinâmicos (ARULKUMARAN *et al.*, 2017). Dentre os principais algoritmos de *Deep Reinforcement*

Learning, destaca-se a Otimização de Política Proximal (*Proximal Policy Optimization* (PPO)), utilizado neste trabalho e apresentado na seção a seguir.

2.1.3 Otimização de Política Proximal

A Otimização de Política Proximal é um algoritmo de DRL que pertence à classe dos métodos baseados em política. Isso significa que ele otimiza diretamente a função $\pi(a|s;\theta)$ isto é, a política que diz a ação a ser executada dado o estado s com parâmetros θ da rede neural que a define. Desenvolvido por Schulman *et al.* (2017), o PPO é projetado para ser uma alternativa eficiente e mais simples aos métodos de gradiente de política de região de confiança, como *Trust Region Policy Optimization (TRPO)* (SCHULMAN, 2015). O PPO tem se destacado pela sua estabilidade e facilidade de implementação, tornando-se um dos algoritmos mais populares no campo de DRL (ARULKUMARAN *et al.*, 2017).

2.1.3.1 Função Objetivo Recortada

Em métodos de gradiente de política tradicionais, grandes atualizações nos parâmetros da política podem levar a uma degradação significativa do desempenho, especialmente em ambientes complexos. Uma das principais inovações do PPO é o uso de uma função objetivo recortada, projetada para limitar a magnitude das atualizações de política entre dois passos consecutivos, evitando grandes mudanças que possam desestabilizar o treinamento. O *clipping* limita a razão de probabilidade $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$, que quantifica a relação entre a probabilidade de uma ação ser escolhida pela nova política em comparação com a política antiga, a um intervalo especificado por $[1-\varepsilon, 1+\varepsilon]$. A função objetivo, L, é dada por

$$L^{CLIP}(\theta) = \mathbb{E}_t\left[\min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t, \text{clip}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right)\right], \tag{2.1}$$

onde $\pi_{\theta_{old}}$ é a política antiga, $\hat{A}_t$ é a vantagem estimada, e ε é um hiperparâmetro que controla o intervalo de clipping. A vantagem estimada $\hat{A}_t$ reduz a variância da estimativa do gradiente e é tradicionalmente calculada por meio do método *Generalized Advantage Estimation* (GAE) (SCHULMAN *et al.*, 2015)

$$\hat{A}_t = \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}, \tag{2.2}$$

onde $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ é o erro temporal (*TD error*), e λ é um hiperparâmetro que controla o *trade-off* entre viés e variância.

A principal diferença quando se compara o PPO a métodos baseados em funções de valor (*value-based methods*), como o DQN (MNIH *et al.*, 2013; MNIH *et al.*, 2015), é que o PPO aprende diretamente a função de política que mapeia estados em ações, balanceando a disputa entre prospecção e exploração por meio da estocasticidade da política aprendida.

2.1.4 Aprendizagem Curricular

A técnica de Aprendizagem Curricular (*Curriculum Learning*) (BENGIO *et al.*, 2009) consiste em dividir um determinado problema a ser atacado por agentes em instâncias menores que apresentem dificuldade incremental entre si, de tal forma a melhorar o aprendizado final dos mesmos. A ideia se inspira em um processo comumente utilizado no aprendizado humano, iniciando por uma tarefa mais simples e gradualmente avançando para situações mais complexas, que exijam um melhor domínio das habilidades. Os principais benefícios dessa abordagem incluem:

- **Aceleração do treinamento**: Ao começar com tarefas mais simples, o agente pode aprender rapidamente os conceitos básicos, o que pode acelerar o processo de treinamento em tarefas mais complexas;
- **Melhoria na generalização**: O aprendizado incremental pode ajudar o agente a generalizar melhor para novas tarefas, pois ele constrói uma base sólida de conhecimento a partir de tarefas mais simples; e
- **Estabilidade do treinamento**: *Curriculum Learning* pode reduzir a variância e melhorar a estabilidade do treinamento, uma vez que o agente é menos propenso a se deparar com situações extremamente difíceis no início do treinamento.

Estruturando o aprendizado de maneira incremental, começando com tarefas mais simples e progredindo para tarefas mais complexas, *Curriculum Learning* permite que os agentes adquiram habilidades de forma mais rápida e robusta, resultando em melhor desempenho e generalização, e tem superado estruturas tradicionais de treinamento em diversos cenários (WANG *et al.*, 2022).

2.2 Ambientes e Tarefas de Aprendizado

A partir de agora, serão apresentadas as ferramentas utilizadas para o desenvolvimento deste trabalho. Primeiro discute-se o conceito de forrageamento, tarefa que os agentes

terão de executar. Em seguida, as bibliotecas *ViZDoom* e *Sample Factory*, completando o arcabouço necessário para compreensão da metodologia desenvolvida.

2.2.1 Forrageamento

O *forrageamento* é a atividade de busca por recursos alimentares disponíveis no ambiente natural. Esse processo envolve uma série de decisões complexas, como quais alimentos escolher, quando mudar para um novo local de alimentação e como otimizar a eficiência para maximizar o ganho energético (WERNER; HALL, 1974). A habilidade de aprender é fundamental nessa busca, uma vez que os animais frequentemente alteram seu comportamento com base em experiências passadas. Esse processo adaptativo, conhecido como *inovação em forrageamento*, envolve experimentar novos tipos de alimentos ou utilizar novas técnicas de obtenção de recursos (WESTNEAT; FOX, 2010).

A **Teoria do Forrageamento Ótimo** propõe que os organismos evoluíram para maximizar a eficiência energética de seu comportamento de forrageamento. Isso envolve a otimização do balanço entre o custo energético da busca e processamento do alimento, e a energia obtida do alimento consumido (WERNER; HALL, 1974).

No contexto do *Aprendizado por Reforço* (Reinforcement Learning), o problema de forrageamento pode ser formalizado como um *Processo de Decisão de Markov* (MDP), em que:

- **Estados**: representam as diferentes condições ou posições no ambiente, incluindo informações sobre a localização de recursos;
- **Ações**: movimentos ou escolhas do agente, como permanecer no *patch* atual ou mover-se para outro;
- **Recompensas**: associadas ao consumo de recursos ou custos energéticos das ações; e
- **Política**: estratégia que o agente segue para decidir as ações com base nos estados.

O agente explora o espaço em busca de alvos ou recursos alimentares, coletando aquilo que procura. Neste cenário, o agente deve aprender políticas que maximizem sua recompensa acumulada, que pode estar relacionada ao ganho energético ou a outro critério de desempenho relevante. Dessa forma, o agente precisa aprender uma política que equilibre a exploração e a prospecção de recursos conhecidos e desconhecidos, similar ao dilema enfrentado por animais em forrageamento real.

2.2.2 ViZDoom

Apresentado por Kempka *et al.* (2016), o *ViZDoom* é uma plataforma de pesquisa em *Reinforcement Learning* baseada no clássico jogo de tiro em primeira pessoa *Doom* (Figura 4), cuja principal inovação foi a possibilidade de treinar agentes em ambientes 3D, permitindo simulações mais realistas. Uma característica que se destaca no *ViZDoom* é a grande facilidade

Figura 4 – Captura de tela do jogo Doom.

Fonte: Página do *Doom* na *Steam*[1].

de customização. Por meio de um sistema de edição de cenários, é possível criar diversos ambientes para realização de experimentos diferentes, com níveis de complexidade variados. O jogo também conta com uma grande comunidade de criadores de *mods*, além de permitir o treinamento de múltiplos agentes em cenários compartilhados. Esse conjunto de fatores contribuíram para uma rápida adoção da plataforma para desenvolvimento de pesquisa e agentes de RL. Outras funções disponíveis na ferramenta são:

- Suporte para múltiplas plataformas;
- Compatibilidade com Python e C++;
- Compatibilidade com ambientes Gym (BROCKMAN *et al.*, 2016);
- Leve e rápido, permitindo a simulação de 7000 quadros por segundo;
- Disponibilização de *Buffer* Visual, *Buffer* de Profundidade e *Buffer* de Áudio, sendo este

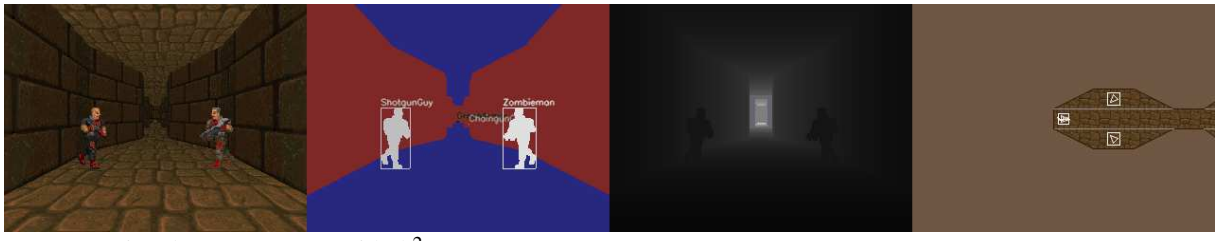
[1] https://store.steampowered.com/app/2280/Ultimate_Doom/

último um dos principais motivos para a escolha dessa ferramenta para este trabalho; e

– Gravação de episódios e renderização sem tela.

A Figura 5 mostra exemplos dos *Buffers* disponíveis.

Figura 5 – Exemplos dos *Buffers* disponíveis no *ViZDoom*. Da esquerda para a direita, *Buffer* Visual, *Buffer* de Rótulos, *Buffer* de Profundidade e Visão de Mapa/Objetos.



Fonte: Página do *ViZDoom* no Github[2].

2.2.3 Sample Factory

Sample Factory é uma biblioteca de RL que implementa algoritmos *Policy Gradients*, como PPO, focada em ser muito eficiente e veloz (PETRENKO *et al.*, 2020). Compatível com os principais ambientes de DRL, incluindo o *ViZDoom* (KEMPKA *et al.*, 2016), Atari (MNIH *et al.*, 2013) e MuJoCo (TODOROV *et al.*, 2012), o *Sample Factory* usa uma abordagem assíncrona para o processamento, dividindo a coleta de amostras entre diferentes *workers* que compartilham um mesmo modelo. Dessa forma, é possível atingir simulações com cerca de $100,000$ quadros por segundo, tanto em ambientes 2D quanto 3D.

No artigo que apresentou o *Sample Factory*, Petrenko *et al.* (2020) destacam um desempenho impressionante no ambiente *ViZDoom*, superando abordagens anteriores de *Deep Reinforcement Learning* em eficiência e velocidade de aprendizado. Seus experimentos revelaram que o sistema pode alcançar um desempenho equivalente ao de jogadores humanos em cenários de *deathmatch* no *Doom*, demonstrando sua capacidade de aprender comportamentos complexos apenas a partir de dados em *pixels*.

O *Sample Factory* apresenta várias características e contribuições significativas que aumentam sua aplicabilidade na pesquisa em *Deep Reinforcement Learning*. Utilizando uma arquitetura de alto desempenho caracterizada por processamento assíncrono e ampla paralelização, a biblioteca facilita processos de aprendizado eficientes. Além disso, integra-se efetivamente a uma variedade de ambientes de simulação e virtuais. O sistema alcançou desempenho de ponta em tarefas complexas de DRL, demonstrando sua capacidade de treinar agentes capazes de

[2] https://github.com/Farama-Foundation/ViZDoom

exibir comportamentos sofisticados diretamente a partir de dados em *pixels* brutos, eliminando, assim, a necessidade de engenharia manual de características. Por meio de seu design inovador e resultados empíricos impressionantes, o *Sample Factory* se destaca como um ativo valioso para o avanço da pesquisa na área de *Reinforcement Learning* e DRL.

3 TRABALHOS RELACIONADOS

Neste capítulo serão apresentados trabalhos relacionados ao controle de agentes multissensoriais por meio de *Deep Reinforcement Learning*, com foco em agentes capazes de enxergar e ouvir. Também serão apresentadas referências ao desenvolvimento de agentes de *Reinforcement Learning* e ambientes para treinamento e avaliação desses agentes.

3.1 *Deep Reinforcement Learning* (DRL)

Nesta seção serão apresentados trabalhos que usam agentes DRL para controlar personagens em jogos digitais ou ambientes similares. Grande parte dos trabalhos iniciais que apresentam agentes inteligentes capazes de vencer jogos utilizam como único sensor a visão do agente. Entretanto, também serão apresentados trabalhos que utilizaram a audição como sensor, seja de forma complementar à visão, ou como única fonte de informação disponível.

O trabalho de Mnih *et al.* (2013) foi pioneiro ao combinar o paradigma de Aprendizado por Reforço com técnicas de *Deep Learning*. O *Deep Q-Learning* (DQN), utiliza uma rede neural convolucional para extrair informações de uma entrada de alta dimensionalidade, os pixels da tela do jogo, e aprender a jogar um conjunto de jogos do console Atari 2600 (BELLEMARE *et al.*, 2013). Desde então, diversos novos algoritmos foram apresentados, dos quais destacam-se *Advantage Actor-Critic* (A2C) (MNIH *et al.*, 2016), *Asynchronous Advantage Actor-Critic* (A3C) (MNIH *et al.*, 2016), *Soft Actor Critic (SAC)* (HAARNOJA *et al.*, 2018) e *Proximal Policy Optimization* (PPO) (SCHULMAN *et al.*, 2017). Tais algoritmos foram utilizados em diferentes ambientes e tarefas, variando entre jogos 2D, jogos 3D, navegação visual, entre outros (ARULKUMARAN *et al.*, 2017; SERAFIM *et al.*, 2020; VINYALS *et al.*, 2019; GORDILLO *et al.*, 2021; FILHO *et al.*, 2021; ZHU *et al.*, 2017; ZENG *et al.*, 2020).

Em todos os trabalhos citados, apenas a visão do agente é utilizada como fonte de informação do ambiente. Contudo, quando um jogador humano experiencia um jogo, também faz uso de outros sentidos que reforçam a imersão no cenário apresentado. Sons são muito utilizados em jogos, seja para alertar o jogador da presença de algum inimigo, a coleta de algum item importante, ou para provocar emoções e criar tensão. Dessa forma, ao utilizar somente os pixels como entrada para o agente, deixa-se de lado um conjunto de outros estímulos que podem auxiliar o jogador. Os trabalhos apresentados a seguir, focam em trazer a experiência sensorial da audição para os agentes.

3.2 Agentes DRL com áudio

Woubie *et al.* (2019) utilizam sons em conjunto com os pixels da tela para treinar um conjunto de agentes em uma tarefa de localização, que consiste em encontrar uma fonte sonora posicionada aleatoriamente dentro de salas em um ambiente 3D. Os autores utilizaram o *ViZDoom* (KEMPKA *et al.*, 2016) para os experimentos e uma representação do cenário criado pode ser vista na Figura 6. Para incorporar o áudio ao agente, foram testadas duas abordagens. Na primeira, o valor do *Pitch* é extraído de um clipe de áudio e enviado diretamente para a CNN. Já a segunda abordagem utiliza uma normalização do sinal de áudio puro. Assim, os autores criaram três agentes, um que utiliza somente a visão e outros dois que utilizam visão e audição, cada um seguindo uma das abordagens de captura de áudio mencionadas.

Os agentes foram treinados utilizando o DQN por 600 mil passos. Os resultados dos experimentos mostram que os agentes multissensoriais foram capazes de aprender mais facilmente a realizar a tarefa, apesar de todos terem sido capazes de completá-la. Nos cenários do *ViZDoom* utilizados, a cada passo de tempo vivo, o agente recebe uma recompensa negativa. Ou seja, quanto maior for a recompensa do agente, mais rápido ele terminou o episódio, indicando um resultado melhor. A Figura 7 mostra a diferença de acumulo de recompensa entre os três agentes.

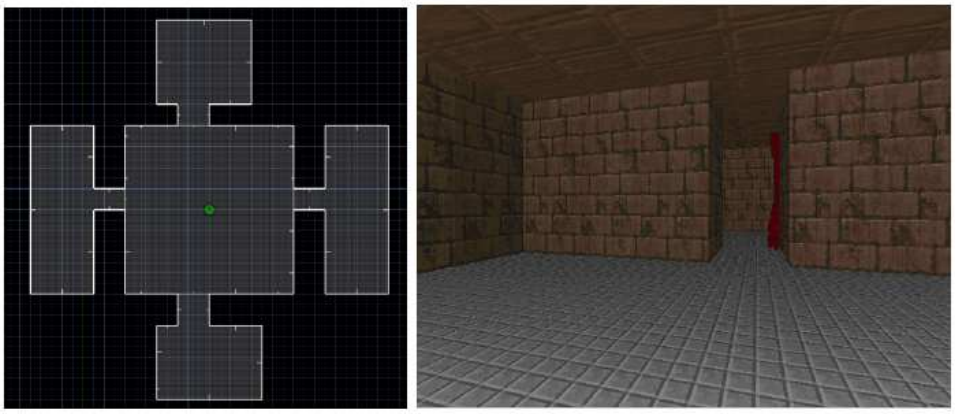


Figura 6 – No lado esquerdo, uma representação 2D vista de cima do cenário criado. Do lado direito, a visão do agente. O pilar vermelho é a fonte sonora que o agente deve localizar. Fonte: Woubie *et al.* (2019).

Hegde *et al.* (2021) também utilizaram o *ViZDoom* como ambiente de treinamento para agentes inteligentes multissensoriais. No trabalho, os autores apresentam uma versão modificada do ambiente, na qual três codificadores de áudio foram implementados e testados

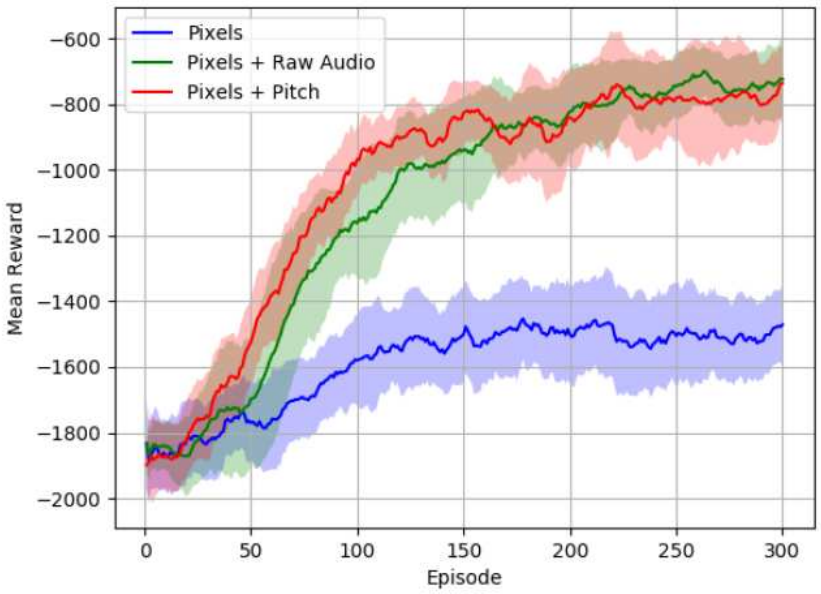


Figura 7 – O gráfico mostra que a recompensa média dos dois agentes que escutam e enxergam foi significativamente maior que a recompensa média do agente surdo, demonstrando uma capacidade mais rápida de atingir o objetivo proposto. Fonte: Woubie *et al.* (2019).

entre si. O primeiro utiliza duas camadas convolucionais de uma dimensão em uma entrada com amostragem reduzida, o segundo aplica uma Transformada de Fourier no *buffer* de áudio, obtendo a amostra no domínio da frequência. Em seguida, é feita uma redução de amostragem utilizando uma camada de *Max-Pooling* e o resultado é submetido a duas camadas totalmente conectadas de 256 neurônios. Por último, também foi testado um codificador de espectrograma, muito comum em trabalhos de processamento de fala (GARCIA-ROMERO *et al.*, 2020). Nesse cenário, as amostras também são transformadas para o domínio da frequência por meio de uma Transformada de Fourier de Curto Termo (*Short-Term Fourier Transform - STFT*) e os espectrogramas gerados são submetidos a duas camadas convolucionais de duas dimensões cada. A Figura 8 apresenta um resumo dos três codificadores desenvolvidos.

Além dos codificadores, o trabalho também apresentou quatro cenários onde os agentes foram treinados e avaliados. O primeiro trata-se de uma tarefa de reconhecimento de música. Há um conjunto de fontes sonoras em uma sala, cada uma tocando uma música diferente e o agente tem como objetivo encontrar a fonte que toca uma música específica pré-determinada. Esse cenário foi utilizado para avaliar os codificadores e escolher qual seria utilizado nos outros experimentos. No cenário seguinte, uma instrução sonora é repetida continuamente indicando o objetivo do agente. Essa instrução indica qual pilar deve ser tocado dentro de um conjunto de salas com diferentes pilares, por exemplo, "vá para o pilar vermelho"ou "vá para o pilar verde".

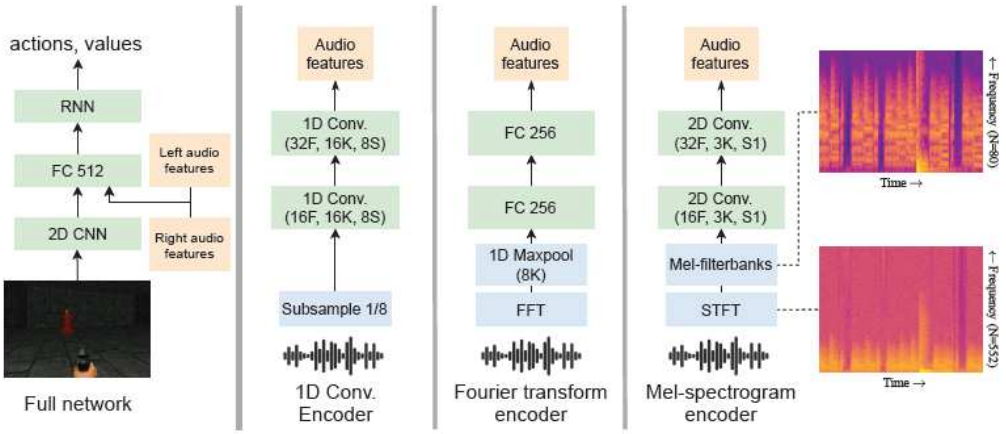


Figura 8 – Ilustração da arquitetura de rede neural utilizada e de cada codificador de áudio testado. K é o tamanho do kernel, F o número de filtros, S o *stride*, FC são as camadas totalmente conectadas e STFT e FFT as Transformadas de Fourier. Fonte: Adaptado de (HEGDE *et al.*, 2021).

O terceiro cenário é similar ao anterior, entretanto a instrução é dita somente uma vez no início do episódio e cabe ao agente memorizá-la e encontrar o pilar correto. O quarto e último cenário consiste em um duelo entre agentes surdos que enxergam e agentes que escutam e enxergam. O objetivo dos agentes é eliminar o adversário.

Os experimentos mostraram que os agentes que escutam obtêm resultado superior a um agente surdo. A Figura 9 mostra as taxas de sucesso dos cenários de seguir instruções. No cenário de duelo, a prevalência do agente capaz de ouvir se repete, obtendo uma taxa de vitória superior ao agente surdo. Os resultados podem ser vistos na tabela da Figura 10.

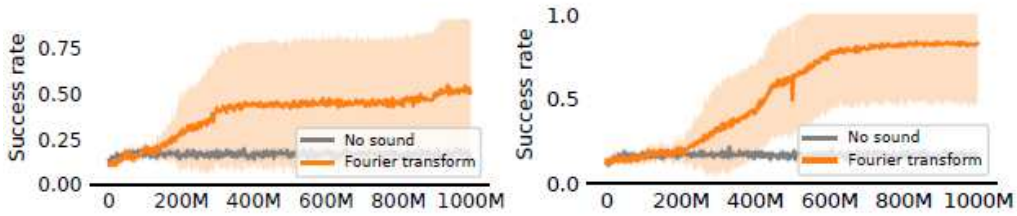


Figura 9 – O gráfico esquerdo apresenta os resultados do cenário 2, onde a instrução é repetida continuamente. O gráfico direito mostra os resultados do cenário 3, no qual a instrução é dita somente uma vez, no início do episódio. Fonte: Adaptado de (HEGDE *et al.*, 2021).

Gaina e Stephenson (2019) apresentaram uma extensão do ambiente *General Video Game AI (GVGAI)* (PÉREZ-LIÉBANA *et al.*, 2019) permitindo utilizar sons como entrada para agentes. Como parte do trabalho, três cenários foram desenvolvidos e disponibilizados para experimentos. São eles: um jogo similar a *Space Invaders* do Atari 2600, chamado *Aliens*, onde uma nave deve atirar em personagens alienígenas e desviar dos projéteis lançados; um jogo de

TABLE I
RESULTS OF 1V1 MATCHES BETWEEN OUR AGENT THAT HAS ACCESS TO THE SOUND AND A VISION-ONLY AGENT. "SOUND (DIS.)" IS THE MAIN AGENT WITH SOUND INPUTS DISABLED DURING THE EVALUATION.

Match	Wins	Losses	Draws
Sound vs No sound	53	31	16
Sound vs Sound (dis.)	74	17	9

Figura 10 – Comparação entre um agente que escuta e um agente surdo. O agente *Sound (dis.)* é o mesmo agente *Sound*, porém com a entrada de áudio desativada. Fonte: Adaptado de (HEGDE *et al.*, 2021).

navegação em labirinto; e um jogo de luta.

Dois agentes foram treinados com o algoritmo Q-Learning e apenas a audição como sensor. O que os diferencia é que em um deles foram incluídas informações de intensidade de observação. Nenhum dos agentes foi capaz de vencer o cenário do jogo de luta ou o labirinto. Já no cenário *Aliens*, ambos obtiveram taxas de sucesso superiores a um agente aleatório utilizado como base, 49% para o agente com informações de intensidade e 37% para o outro.

Gan *et al.* (2019) apresentam uma tarefa de navegação áudio-visual incorporada, na qual um agente deve controlar um robô e conduzi-lo até uma fonte sonora colocada em uma posição aleatória no cenário (Figura 11). Os autores não utilizaram técnicas de DRL para resolver a tarefa apresentada, porém compararam os resultados obtidos a base de DRL. O método proposto utiliza uma combinação de técnicas de *Deep Learning* para solucionar partes individuais do problema, em seguida combinando o resultado em um grafo e encontrando o menor caminho por meio do algoritmo de Djikstra.

Inicialmente, um classificador de imagens é treinado usando uma arquitetura ResNet (HE *et al.*, 2016), sendo este responsável pela visão do agente. Um segundo classificador processa os espectrogramas gerados dos sons no ambiente. Essas informações são combinadas em um planejador de mapas, que consiste em um grafo parcial do ambiente explorado. A Figura 12 ilustra essa abordagem.

Durante os experimentos, os autores compararam a técnica proposta a um conjunto de agentes treinados com o algoritmo A3C (MNIH *et al.*, 2016). Um agente utiliza somente a visão; outro visão e audição; um terceiro utiliza visão, audição e uma memória LSTM; e o quarto é uma combinação dos três anteriores adicionado do planejador de mapa proposto. Também foram avaliados um agente aleatório e um agente guloso. A Figura 13 mostra a tabela

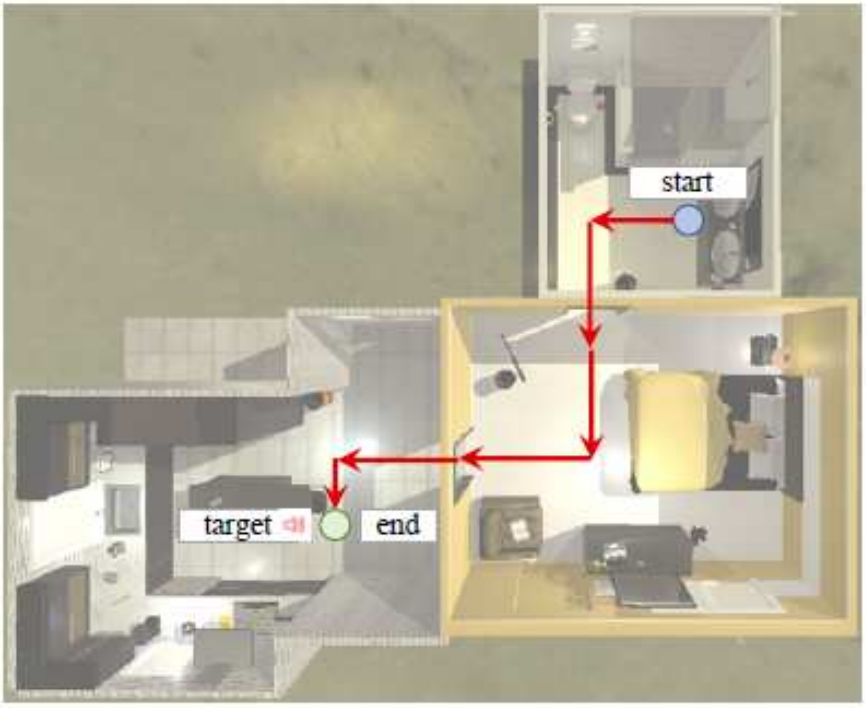


Figura 11 – O agente deve conduzir um robô do ponto azul até o ponto verde. As setas vermelhas mostram o caminho ideal. Fonte: Adaptado de (GAN *et al.*, 2019).

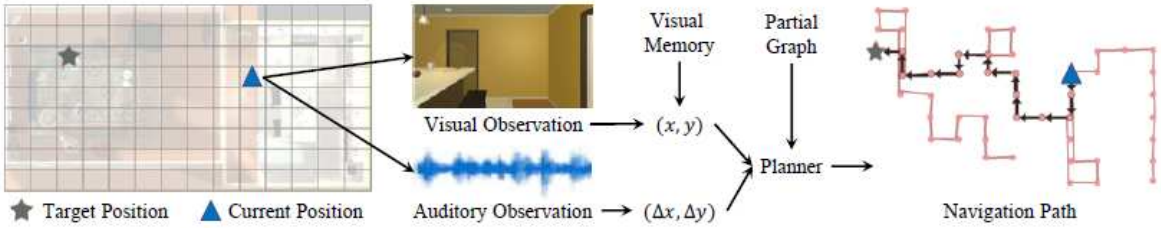


Figura 12 – Combinação de técnicas utilizadas para navegar no ambiente e buscar a fonte sonora alvo. Fonte: Adaptado de (GAN *et al.*, 2019).

de comparação dos agentes. Em todos os cenários do conjunto de dados testado, o agente dos autores obtêm resultado superior aos outros.

Giannakopoulos *et al.* (2021) utilizam agentes somente com áudio em tarefas de navegação baseada em som e localização de fontes sonoras. Também verificam se é possível reaproveitar o aprendizado de uma tarefa em outra. Para tal, criaram dois ambientes usando Unity3D (JULIANI *et al.*, 2018):

- **Navegação baseada em áudio:** um agente é posto em uma sala contendo de 1 a 5 fontes sonoras (além disso a sala está vazia). Uma dessas fontes é determinada como alvo e o agente deve se movimentar em direção a ela sem colidir com as outras (Figura 14).
- **Localização de fontes sonoras:** o cenário anterior é repetido, entretanto o agente passa a ficar fixo no centro da sala e seus movimentos são limitados a rotação e ajuste de altura. O

Approach	Sound	Room 1	Room 2	Room 3	Room 4	Room 5	Average
Random Search	Ring	6/3.5	10/6.8	8/4.6	5/2.6	8/5.1	7.4/4.5
	Alarm	5/3.3	9/5.7	7/4.2	7/4.1	7/3.9	7/4.2
	Clock	8/5.2	9/6.6	6/3.4	4/2.1	7/4.4	6.8/4.3
Greedy Search (A)	Ring	12/10.5	15/13.5	13/11.8	10/8.2	12/11.4	12.4/11.1
	Alarm	9/8.2	13/11.2	10/8.4	8/7.1	11/10.2	10.2/9.0
	Clock	13/12.0	15/13.4	12/11.0	12/9.6	11/10.1	12.6/11.2
A3C (V)	Ring	5/4.7	12/11.5	10/9.1	8/7.8	10/8.9	9/8.4
	Alarm	7/6.0	11/9.3	9/8.9	7/6.2	8/7.5	8.4/7.6
	Clock	7/6.1	8/7.5	9/7.8	9/8.6	11/9.8	8.8/8.0
A3C (V+A)	Ring	18/15.7	25/18.5	15/11.7	14/11.5	20/14.9	18.4/14.5
	Alarm	19/15.9	22/17.1	16/14.5	14/12.5	19/14.6	18/14.9
	Clock	18/13.0	27/23.6	18/13.8	16/13.8	22/16.1	20.2/16.1
A3C (V+A+M)	Ring	27/18.4	36/25.7	33/22.2	29/20.8	31/21.2	31.2/21.7
	Alarm	24/16.5	37/25.0	29/20.4	31/21.4	33/23.2	30.8/21.3
	Clock	29/20.3	34/25.6	27/19.3	34/24.3	28/19.9	30.4/21.9
A3C (V+A+Mapper)	Ring	24/17.9	30/21.3	21/15.6	19/14.3	24/17.4	23.6/17.3
	Alarm	22/16.6	29/20.1	23/16.9	20/15.2	25/17.9	23.8/17.3
	Clock	25/18.5	31/21.7	22/16.2	20/14.9	24/17.2	24.4/17.7
A3C(V+A+M+Mapper)	Ring	30/20.7	38/25.7	35/23.8	33/23.3	33/23.8	33.8/23.5
	Alarm	28/19.1	39/26.4	33/22.5	34/23.5	36/24.4	34.0/23.2
	Clock	31/21.3	37/24.9	31/22.5	36/24.9	32/23.5	33.4/23.4
Ours (no Exp.)	Ring	69/42.2	55/37.4	53/35.4	56/38.6	59/40.3	58.4/38.8
	Alarm	67/41.7	53/35.9	55/36.8	56/37.9	58/39.7	57.8/38.4
	Clock	70/43.9	58/39.8	56/37.7	54/36.4	61/41.6	59.8/40.0
Ours (w/ Exp.)	Ring	**78/58.0**	**66/47.1**	**62/43.4**	**68/55.0**	**74/51.3**	**69.6/51.0**
	Alarm	**76/57.0**	**62/45.5**	**65/45.7**	**66/45.1**	**69/54.4**	**67.6/49.5**
	Clock	**83/60.1**	**70/56.7**	**69/48.7**	**62/51.1**	**70/55.1**	**70.8/54.4**

Figura 13 – Taxas de sucesso (%) / Sucesso ponderado por Path Length (SPL) de diferentes métodos de navegação audiovisual. Fonte: Adaptado de (GAN *et al.*, 2019).

objetivo é localizar todas as fontes sonoras na sala, apontando o microfone do agente na direção de cada uma delas (Figura 15).

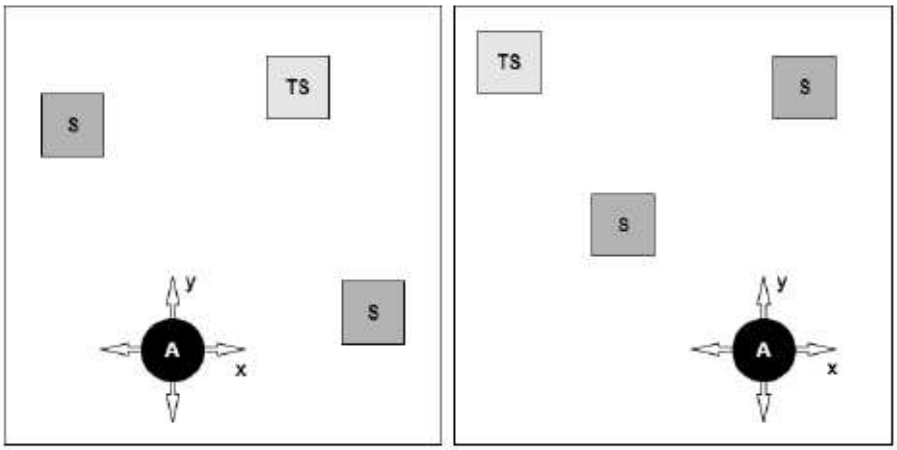


Figura 14 – Cenário de navegação baseada em áudio. A é o agente, S as fontes sonoras e TS a fonte alvo. Fonte: Adaptado de (GIANNAKOPOULOS *et al.*, 2021).

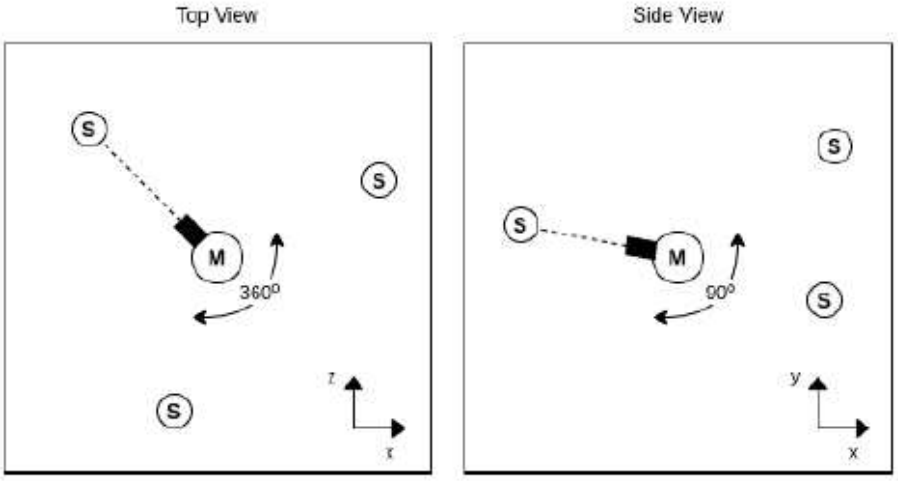


Figura 15 – Cenário de localização de fontes sonoras. M é o agente e S as fontes sonoras. Fonte: Adaptado de (GIANNAKOPOULOS *et al.*, 2021).

O trabalho apresenta uma comparação com um testador humano e um agente aleatório, sendo ambos superados pelo agente autônomo treinado com o algoritmo PPO (SCHULMAN *et al.*, 2017). A arquitetura da rede neural é apresentada na Figura 16. Além disso, é feita uma análise da aplicação de *Transfer Learning* (WEISS *et al.*, 2016) entre as duas tarefas apresentadas. A partir disso, os autores concluem que o desafio principal para um agente que escuta é a localização da fonte sonora, pois é uma habilidade necessária para as duas tarefas.

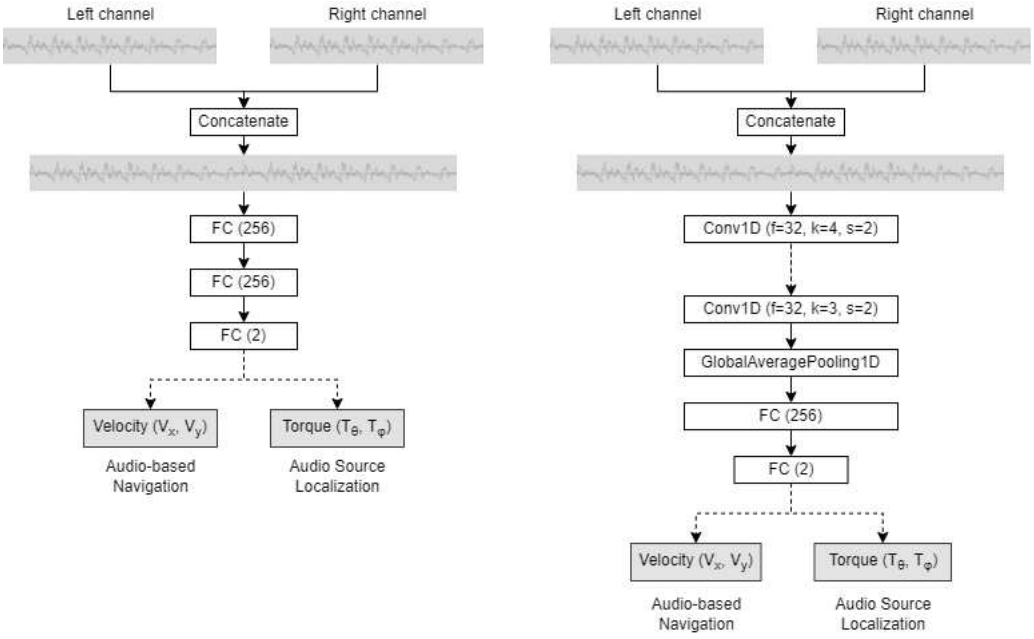


Figura 16 – Arquiteturas de rede neural utilizadas nos experimentos. Fonte: Adaptado de (GIANNAKOPOULOS *et al.*, 2021).

Chen *et al.* (2020) apresentam um *framework* para experimentos em navegação áudio-visual em ambiente 3D complexos. O trabalho apresenta o *SoundSpaces*, um *plugin* para

a ferramenta Habitat (SAVVA *et al.*, 2019) (SZOT *et al.*, 2021) que permite a renderização de áudio em cenários realistas. A plataforma Habitat consiste em um conjunto de escaneamentos 3D de cenários reais, por exemplo, casas, escritórios e apartamentos. A renderização de áudio é feita levando em consideração a geometria dos cenários, os materiais dos objetos, entre outras informações. Dessa forma, é possível inserir sons arbitrários em qualquer objeto do ambiente 3D.

Os autores definem três tarefas para a execução dos experimentos:

- ***PointGoal*:** o agente recebe um vetor de deslocamento indicando a distância e direção do alvo;
- ***AudioGoal*:** o agente recebe apenas o áudio sendo emitido pelo objeto alvo; e
- ***AudioPointGoal*:** o agente recebe uma combinação das tarefas anteriores, ou seja, tanto um vetor de deslocamento quanto um o áudio sendo emitido.

Em todas elas o objetivo é navegar por um cenário complexo em busca de uma fonte de áudio alvo. A visão está presente em todos os cenários, enquanto a audição só está presente nos dois últimos e captura um áudio binaural. Para capturar o som, o agente simula dois microfones posicionados na altura de um humano.

O som ambiente capturado é transformado em espectrogramas utilizando a STFT (GRIFFIN; LIM, 1984) e enviado para uma rede neural convolucional para processamento. A imagem do ambiente é representada por uma matriz RGB de 128 por 128 pixels. A arquitetura geral do agente pode ser vista na Figura 17. O agente foi treinado utilizando PPO (SCHULMAN *et al.*, 2017) e as recompensas são distribuídas de forma a recompensar o agente por terminar a tarefa mais rápido. Por isso, a cada passo de tempo o agente recebe uma recompensa de -0.01. O episódio termina quando o agente executa uma ação de parada e, caso esteja na posição do alvo, recebe uma recompensa de 10.

Os experimentos mostram que os agentes multissensoriais apresentam resultados superiores àqueles que contam apenas com um único sensor (visão). Os autores treinaram modelos com módulos de visão variados (cego, somente RGB e somente profundidade) e, em todos os casos, os que combinavam visão com audição obtiveram resultados superiores. A Figura 18 mostra a comparação entre eles em cada tarefa proposta, bem como uma comparação com sistemas base.

O trabalho também investiga o efeito de diferentes fontes sonoras no treinamento dos agentes. Dividindo o conjunto de áudios em grupos de treino, teste e validação, os autores avaliam

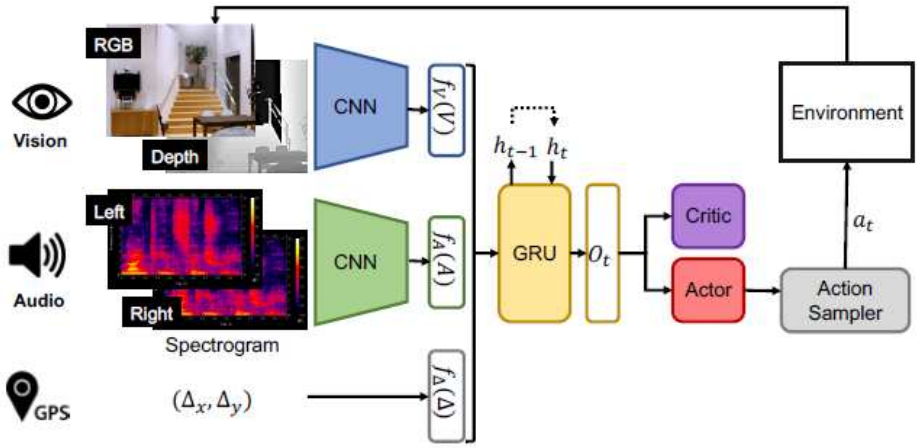


Figura 17 – Arquitetura de rede neural utilizada. Fonte: Adaptado de (CHEN *et al.*, 2020).

		Replica		Matterport3D	
		PointGoal	AudioPointGoal	PointGoal	AudioPointGoal
Baselines	RANDOM	0.044	0.044	0.021	0.021
	FORWARD	0.063	0.063	0.025	0.025
	GOAL FOLLOWER	0.124	0.124	0.197	0.197
Varying visual sensor	Blind	0.480	**0.681**	0.426	**0.473**
	RGB	0.521	**0.632**	0.466	**0.521**
	Depth	0.601	**0.709**	0.541	**0.581**

Figura 18 – Comparação de resultados dos agentes avaliados. Em todos os casos, o agente que escuta obteve resultado superior aos agentes surdos. Fonte: Adaptado de (CHEN *et al.*, 2020).

os agentes treinados em fontes sonoras desconhecidas e em ambientes não vistos previamente. Os resultados mostram que um agente treinado em um conjunto de fontes sonoras, mesmo quando submetido a sons desconhecidos, consegue um resultado similar ou superior a um agente surdo. A Figura 19 mostra as trajetórias dos agentes em alguns cenários, onde pode ser visto que o agente *AudioPointGoal*, que utiliza o conjunto completo de sensores, chega mais próximo da trajetória ideal na maioria dos casos.

3.3 Ambientes de *Deep Reinforcement Learning*

Nesta seção serão apresentados alguns ambientes de treinamento e avaliação de agentes de Aprendizado por Reforço. Esses ambientes possuem características ou modificações e *plugins* que permitem o treinamento de agentes utilizando sons como entrada.

O *Atari Learning Environment* (ALE) é um ambiente amplamente utilizado na área de aprendizado por reforço, introduzido por (BELLEMARE *et al.*, 2013). Ele oferece uma plataforma para treinar e avaliar algoritmos de aprendizado de máquina utilizando jogos clássicos do console *Atari 2600*. O ALE apresenta uma coleção diversificada de jogos que variam em complexidade e dinâmica, permitindo que os agentes aprendam a partir de uma variedade de

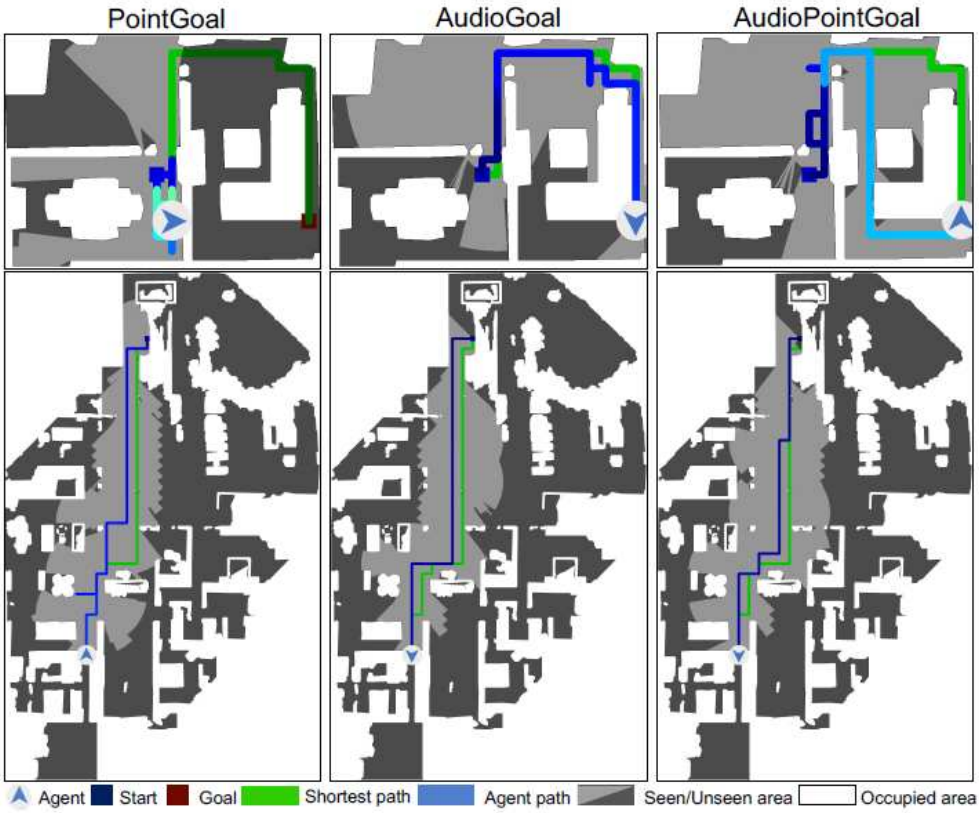


Figura 19 – Trajetórias dos 3 agentes em dois cenários diferentes. O agente *AudioPointGoal* é o que mais se aproxima do trajeto ideal, em verde. Fonte: Adaptado de (CHEN *et al.*, 2020).

cenários, tais como jogos de ação, quebra-cabeças e corrida. Esses jogos servem como tarefas de *benchmark*, devido à simplicidade em termos de entrada (imagens em 2D) e à possibilidade de um agente interagir diretamente com o ambiente através de comandos discretos. Um dos principais benefícios do ALE é a sua capacidade de fornecer um vasto conjunto de jogos, possibilitando uma avaliação comparativa e robusta dos algoritmos de aprendizado por reforço.

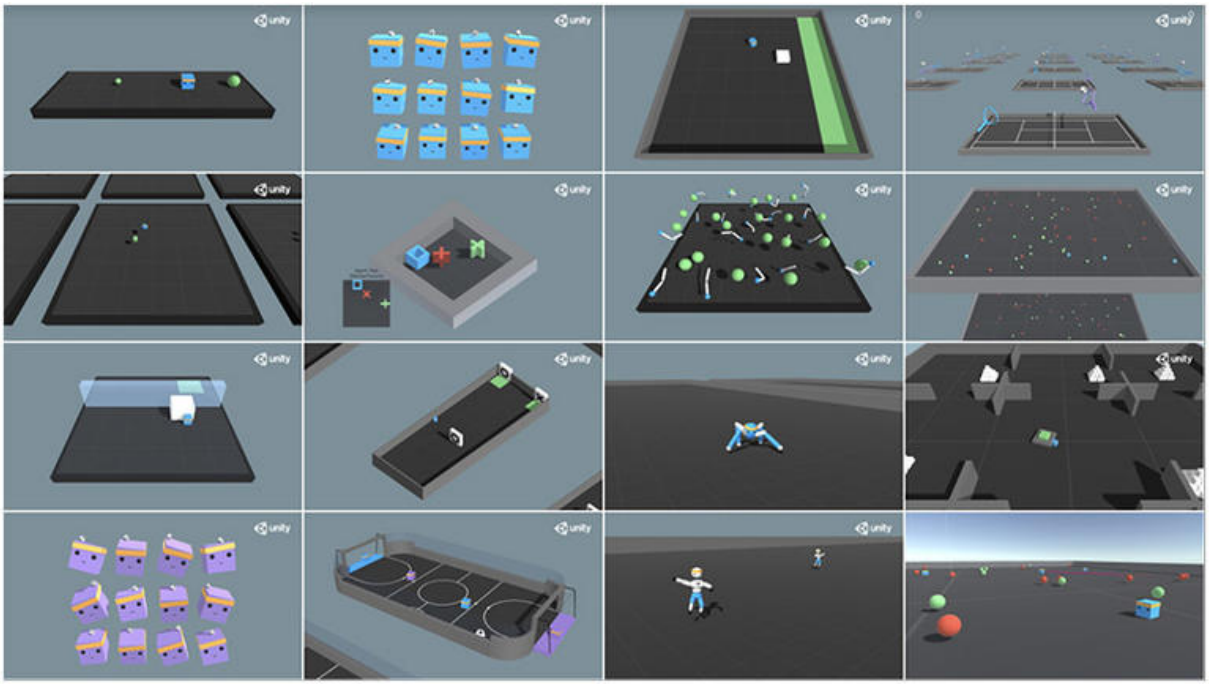
No entanto, o ALE apresenta algumas limitações que o tornam menos adequado para certos tipos de estudos, este incluso. Primeiramente, os jogos do *Atari 2600* são exclusivamente bidimensionais (2D), o que reduz a complexidade espacial dos cenários apresentados aos agentes. Isso pode limitar a generalização de agentes treinados no ALE para ambientes tridimensionais (3D) mais complexos, onde a percepção de profundidade e a navegação espacial são fundamentais.

Além disso, o *Atari Learning Environment* não inclui observações sonoras dos jogos, o que impede a utilização de pistas auditivas que poderiam enriquecer o processo de aprendizado do agente. Assim, inviabilizando o uso do ALE em estudos de agentes multimodais, dado que ele limita o aprendizado a apenas um fluxo de dados visuais em 2D.

Juliani *et al.* (2018) apresentaram um conjunto de ferramentas para serem utilizadas

com o motor gráfico *Unity 3D*, chamado *Unity 3D - Machine Learning Agents Toolkit (ML-Agents)* (Figura 20). A ferramenta permite utilizar cenários criados na *Unity*[1] para treinar agentes DRL. Como trata-se de um motor gráfico e não de um jogo específico, o usuário pode criar cenários que se adaptem melhor aos objetivos buscados, modelando mapas, mecânicas, física, entre outros fatores. Além dos ambientes, o *ML-Agents* possui um conjunto de algoritmos base já implementados, como o PPO (SCHULMAN *et al.*, 2017), A3C (MNIH *et al.*, 2016) e SAC (HAARNOJA *et al.*, 2018). Diferente do ALE, é possível utilizar o áudio da plataforma

Figura 20 – Cenários base disponibilizados na plataforma *ML-Agents*.



Fonte: Página do *ML-Agents* no Github[2].

ML-Agents como parte da observação do agente, entretanto as iterações são limitadas ao tempo real de execução de um jogo ou cenário, tornando o processo de aprendizado lento e custoso.

O ambiente utilizado neste trabalho, *ViZDoom*, foi apresentado no Capítulo 2, portanto não está incluso nessa seção. Entretanto, ele soluciona ambos problemas apresentados pelos ambientes anteriores, sendo capaz de executar múltiplos ambientes de forma rápida e utilizando sons como informação para o treinamento de agentes DRL.

[1] https://unity.com/pt

3.4 Visão Geral

Neste capítulo, foram apresentados diversos trabalhos que exploram o uso de *Deep Reinforcement Learning* (DRL) para controle de agentes em ambientes digitais, com destaque para a utilização de sensores multimodais, como visão e audição. Apesar de os agentes visuais terem mostrado resultados promissores, a ausência de outras modalidades sensoriais limita a experiência completa dos agentes em comparação com jogadores humanos, que se beneficiam de múltiplos estímulos, incluindo o som.

Para abordar essa lacuna, estudos recentes exploraram a integração do áudio em agentes DRL. Trabalhos como os de Woubie *et al.* (2019) e Hegde *et al.* (2021) mostraram que agentes multissensoriais, que utilizam visão e audição, podem superar aqueles que utilizam apenas a visão em tarefas como localização de fontes sonoras e navegação em ambientes 3D, usando plataformas como o *ViZDoom*.

Embora pesquisas anteriores tenham demonstrado o benefício da adição de áudio para melhorar o desempenho dos agentes em tarefas específicas, a integração eficaz dessas modalidades permanece um desafio. Este trabalho busca explorar como um agente pode aprender a processar e combinar informações visuais e auditivas de forma coordenada, maximizando o aproveitamento das pistas sonoras para melhorar a navegação e a tomada de decisão em ambientes complexos.

4 METODOLOGIA E EXPERIMENTOS

Neste capítulo, serão apresentados os experimentos realizados durante o desenvolvimento deste trabalho. Inicialmente, será feita uma descrição dos cenários desenvolvidos e utilizados para o treinamento e a avaliação dos agentes. Em seguida, serão expostas as estratégias utilizadas para treinamento dos agentes, bem como as características da implementação destes. Por fim, será definido o currículo proposto neste trabalho para possibilitar o treinamento de agentes multissensoriais.

4.1 Cenários de Treinamento

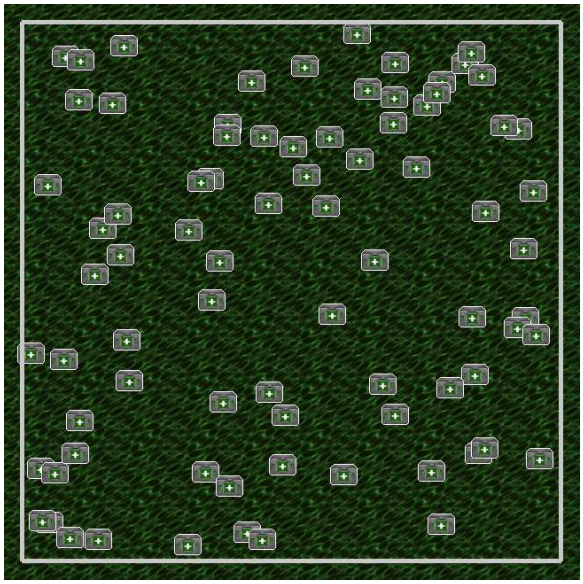
Esta seção apresenta os cenários utilizados para o treinamento e a avaliação dos agentes. Todos os cenários são focados na tarefa de forrageamento, descrita na Seção 2.2.1 do Capítulo 2, e se baseiam em um cenário previamente disponível na plataforma *ViZDoom*.

4.1.1 Health Gathering (HG) e Health Gathering Supreme (HGS)

O cenário *Health Gathering* é a base para todas as variações que serão apresentadas ao longo desta seção. Trata-se de uma sala sem paredes internas, apenas cercada nos seus limites por 4 paredes, formando um quadrado. Não há teto e o chão é feito de ácido, o que periodicamente causa dano aos personagens ao longo do tempo. A Figura 21 mostra uma vista superior do cenário e também a visão de um personagem neste mapa.

Espalhados nesse mapa, encontram-se diversos *Medkits*, itens que recuperam pontos de vida do personagem. O *sprite* de um *Medkit* pode ser visto na Figura 22. Devido ao chão ácido, o objetivo dos agentes neste cenário é aprender a identificar e coletar tais *Medkits* para manterem-se vivos durante o máximo de 2100 passos de jogo. Periodicamente, novos *Medkits* aparecem no mapa, substituindo os que foram previamente coletados ou apenas adicionando ao total.

A variação *Health Gathering Supreme*, vista na Figura 23, adiciona um grau de complexidade extra ao cenário HG ao incluir paredes internas no mapa da sala quadrada. Dessa forma, os *Medkits* podem estar escondidos da visão do jogador, necessitando um esforço maior de busca para atingir o objetivo de manter-se vivo. O jogador deve então ser capaz de procurar em diversas salas internas, muitas vezes sem saída, de forma que remete a labirintos tradicionais, exigindo mais agilidade e objetividade na busca pelos itens de cura. Todos os outros aspectos do



(a) Visão superior do cenário ***Health Gathering***. Os ícones espalhados no mapa representam os *Medkits*.

(b) Ponto de vista do agente nos cenários ***Health Gathering***.

Figura 21 – Todos os cenários baseados no *Health Gathering* terão visualizações similares às apresentadas nesta figura, podendo ter pequenas variações de acordo com o objetivo do cenário.

(a) *Sprite* de *Medkit* visto pelo agente durante a execução.

(b) *Sprite* de *Medkit* utilizado nas visualizações de trajetória.

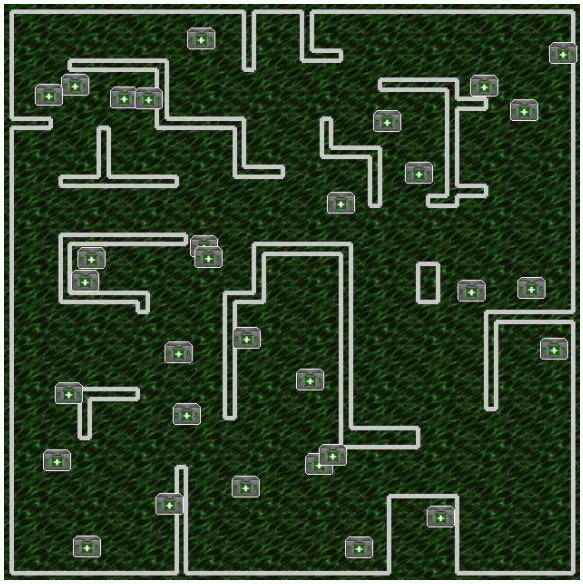
Figura 22 – *Sprites* de *Medkits* no *ViZDoom*.

cenário se mantêm.

4.1.2 Variantes Sound

Os cenários apresentados na Seção 4.1.1 são a base para as variantes utilizadas nos experimentos desenvolvidos neste trabalho. Entretanto, um dos principais pontos proposto para ser explorado é a capacidade dos agentes de interpretarem múltiplos sensores, especialmente o sensor sonoro como auxílio ao visual. Os cenários *Health Gathering* e *Health Gathering Supreme* não possuem em sua implementação original fontes sonoras, tornando-os inutilizáveis para a pesquisa de agentes multissensoriais, demandando a implementação de cenários personalizados.

Para permitir a resolução dos cenários através do áudio, cada *Medkit* foi programado para emitir um som específico, que é reproduzido em *loop* enquanto o item estiver presente no mundo do jogo. O agente precisa aprender a identificar, localizar e coletar esses itens para aumentar sua sobrevivência em cada episódio. Além disso, o motor do jogo possibilita o controle



(a) Visão superior do cenário ***Health Gathering Supreme***. Os ícones espalhados no mapa representam os *Medkits*.

(b) Ponto de vista do agente nos cenários ***Health Gathering Supreme***.

Figura 23 – Todos os cenários baseados no *Health Gathering Supreme* terão visualizações similares às apresentadas nesta figura, podendo ter pequenas variações de acordo com o objetivo do cenário. A principal diferença da versão *supreme* pode ser vista ao comparar com a Figura 21 e observar as paredes internas no mapa.

do nível de atenuação do som. O padrão ***ATTN_STATIC*** (ZDOOM, 2020) foi o escolhido, pois faz com que o som diminua rapidamente à medida que a distância aumenta, exigindo que o agente se aproxime da fonte sonora para identificá-la. Além disso, a duração máxima de cada episódio foi ampliada para 4200 passos de jogo, o que é o dobro dos 2100 originais do *ViZDoom*. Isso permite uma observação melhor do uso dos diferentes sensores em cada cenário. Já o tamanho do mapa permaneceu inalterado.

Para determinar o número de *Medkits* disponíveis em cada mapa, um agente foi treinado em cada cenário e seu desempenho foi analisado enquanto o número de itens era reduzido a cada execução, utilizando os seguintes valores: 30, 20, 15 e 7. Com 7 *Medkits*, nem mesmo o cenário mais fácil era solucionável. O número 30 foi o escolhido, pois tornava todos os cenários solucionáveis, mas não triviais, quando combinado com as mudanças na taxa de aparecimento dos *Medkits*, explicadas abaixo, juntamente com as principais diferenças entre os três cenários.

– **Health Gathering Sound + 30 (HG+30)**: Semelhante ao cenário original, mas com som. Começa com 30 *Medkits* aleatoriamente espalhados pela sala e novas unidades aparecem a cada 25 passos no jogo. Esse cenário possui muitos recursos para o agente

coletar, permitindo que recompensas sejam recebidas com mais frequência, tornando-o mais simples e fácil.

- **Health Gathering Sound = 30 (HG=30)**: Difere do anterior por manter apenas 30 *Medkits* por vez no jogo. Não gera novos ao longo do tempo. Novos *kits* aparecem apenas quando um agente coleta um que já existe, mantendo assim a mesma quantidade o tempo todo. Esse cenário é mais difícil devido à quantidade limitada de recursos para o agente reunir, o que exige que se explore mais para receber uma recompensa.
- **Health Gathering Supreme Sound + 30 (HGS+30)**: Esse cenário é baseado em *Health Gathering Supreme*, que se diferencia dos dois anteriores por ter paredes internas, criando salas separadas. Aqui, a mecânica de aparecimento é a mesma de **HG+30**. Assim, embora abundante em recursos, o agente deve navegar por este cenário labiríntico para encontrá-los, e usar o som para localizar *kits* através de uma parede pode ajudar na identificação.

4.1.3 Cenário de Busca Imediata

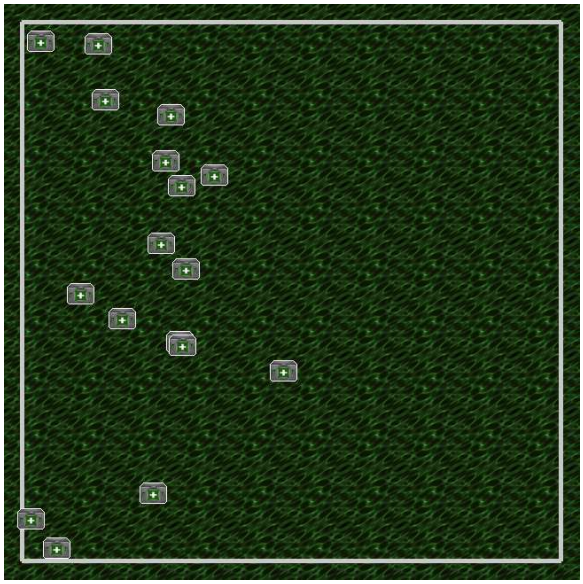
Desenvolvido para verificar a capacidade de um agente de reagir a um estímulo sonoro vindo de uma direção aleatória, esse cenário consiste em uma sala quadrada e vazia, onde o agente sempre aparece posicionado no centro. A cada episódio, um único *Medkit* é colocado em um ponto aleatório na sala, alguns segundos após o surgimento do agente.

O objetivo desse cenário é que o jogador reaja à aparição do *Medkit*, virando-se na direção dele e se locomovendo para coletá-lo. Os *sprites* e sons foram mantidos, porém o dano causado pelo chão ácido foi reduzido, bem como o tamanho da sala.

4.1.4 Cenário Meio a Meio

Esse cenário é uma variação do HG, programado para que os *Medkits* só sejam gerados em um lado da sala. Novamente é uma sala quadrada, o personagem começa sempre no ponto central e periodicamente novos *Medkits* surgem apenas em um dos lados deste quadrado. No início de cada episódio, um dos dois lados é sorteado para o aparecimento dos itens.

Aqui o objetivo é verificar se o agente é capaz de, utilizando apenas a audição, se manter no lado mais benéfico, visto que, caso se mova para o lado oposto, não encontraria itens suficientes para se manter vivo. Os *sprites* e sons foram mantidos. A Figura 24 mostra o cenário e sua vista superior.



(a) Visão superior do cenário **Meio a Meio**. Os ícones espalhados no mapa representam os *Medkits*.

(b) Ponto de vista do agente nos cenários **Meio a Meio**.

Figura 24 – Cenário Meio a Meio, derivado do HG. Destaca-se a distribuição de *Medkits* em apenas um lado do mapa.

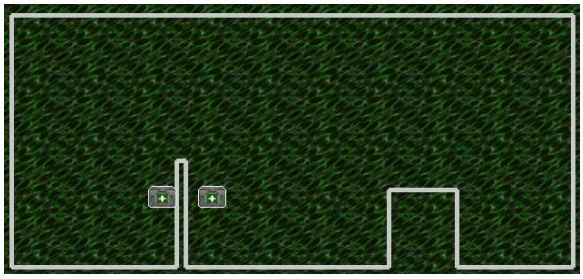
4.1.5 Cenário de Oclusão

O Cenário de Oclusão foi criado como um desafio para verificar se os primeiros agentes treinados com som eram capazes de utilizá-lo efetivamente. Baseado no HGS, consiste em um corredor onde há uma parede, o que oculta o único *Medkit* no mapa. O agente que não utiliza o som dificilmente coletará o *Medkit*, pois ao andar em linha reta e passar pela parede o item não entra no campo de visão do agente, sendo necessário virar totalmente para a esquerda para vê-lo.

O comportamento esperado de um agente capaz de ouvir, é que, ao passar pela parede, escute o som emitido pelo *Medkit* e se vire para coletá-lo. Aqui, a vida do agente foi diminuída, para forçá-lo a sempre buscar o *Medkit* rapidamente, pois caso não o colete assim que passar pela parede, não haverá tempo suficiente para retornar. A Figura 25 mostra o cenário e uma vista superior deste.

4.2 Implementação dos Agentes e Estratégia de Treinamento

Um dos principais objetivos deste trabalho foi desenvolver um agente que utilizasse tanto a imagem quanto o som do ambiente para realizar a tarefa de coleta. Inicialmente, um agente equipado com sensores para ambos os tipos de entrada foi implementado, ou seja, capaz



(a) Visão superior do cenário **Oclusão**. Os ícones espalhados no mapa representam os *Medkits*.

(b) Ponto de vista do agente nos cenários **Oclusão**. Na imagem é possível ver o *Medkit* escondido atrás da parede.

Figura 25 – Cenário Oclusão, derivado do HGS.

de ver e ouvir o ambiente em que está inserido. A entrada visual do agente consiste no conjunto de *pixels* da tela do jogo redimensionado para 320×240. Já o som é representado por uma matriz de tamanho 2520×2, correspondendo aos canais esquerdo e direito para obtenção de som estéreo.

A biblioteca *Sample Factory* fornece um codificador padrão para imagem baseado em arquiteturas de redes neurais convolucionais[1]. Esse codificador foi utilizado para o processamento visual do agente, sem modificações. Para o áudio, utilizou-se um método baseado em Transformada Rápida de Fourier (FFT - do inglês, *Fast Fourier Transform*) (GRIFFIN; LIM, 1984) que foi desenvolvido anteriormente por Hegde *et al.* (2021). Primeiro, aplica-se a FFT aos dois canais de som individualmente, obtendo a representação do áudio no domínio da frequência. Em seguida, a matriz resultante passa por uma camada de *pooling* seguida, originalmente, de duas camadas totalmente conectadas de tamanho 256×256. Entretanto, foram inclusas mais duas camadas iguais à saída da rede, totalizando quatro. Todos os experimentos utilizaram uma taxa de aprendizado de 0.0001, um tamanho de *batch* de 2048, 12 *workers* contando com 16 ambientes cada, e foram executados em uma máquina local equipada com um processador Intel Core i7 10700, 32 GB de memória RAM e uma placa de vídeo RTX 2070 SUPER da Nvidia.

[1] https://github.com/alex-petrenko/sample-factory/blob/abbc4591fcfa3f2cb20b98bb9b0f2a1ee83f47fa/sample_factory/model/encoder.py#L122

4.2.1 Verificando a capacidade de ouvir

No começo do processo de desenvolvimento deste trabalho, foram realizadas algumas provas de conceito para verificar a capacidade dos agentes de aprenderem a ouvir e o quanto a audição seria útil e até mesmo suficiente para a realização de tarefas mais simples de forrageamento. Os cenários Busca Imediata e Meio a Meio, apresentados nas Subseções 4.1.3 e 4.1.4 respectivamente, foram desenvolvidos para realizar estes experimentos iniciais.

4.2.1.1 Agente Meio a Meio

Para validar a capacidade de aprender a ouvir, um agente foi treinado nesse cenário utilizando apenas o áudio como entrada. A cada passo em que o agente está vivo, ele recebe uma recompensa positiva e, caso ele morra, recebe uma recompensa negativa. Não há recompensas por coletar *Medkits*, contudo, o agente deve aprender que coletá-los estende sua vida, aumentando a recompensa acumulada com o tempo.

A Figura 26 mostra a trajetória percorrida pelo agente treinado nesse cenário. O principal ponto que se destaca é a permanência total do agente na zona onde estão localizados os *Medkits*, indicando que o agente, de fato, aprendeu a se manter na área correta utilizando apenas o som para se guiar nesse ambiente. Ou seja, o som é suficiente para o agente aprender uma tarefa de forrageamento simplificada, como a apresentada.

4.2.1.2 Agente de Busca Imediata

Outro ponto investigado foi a capacidade de um agente de reagir a sons direcionais. Para isso, o mesmo agente apresentado na subseção anterior, ou seja, já treinado e capaz de usar sons, foi avaliado no cenário de Busca Imediata. Aqui, apenas um *Medkit* está presente no mapa e o objetivo do agente é coletá-lo o mais rápido possível para se manter vivo. Se o agente for capaz de distinguir a direção de onde o som está vindo, espera-se que ele sempre se vire na direção do item, visto que é o único som presente no ambiente.

A Figura 27 mostra a trajetória do agente em uma execução deste cenário. O *Medkit* visto na imagem surge no mapa de maneira sequencial quando o anterior é coletado. Assim, pode-se ver que o agente sempre segue na direção do próximo item que aparece, mostrando sua capacidade de se guiar por meio de sons tridimensionais.

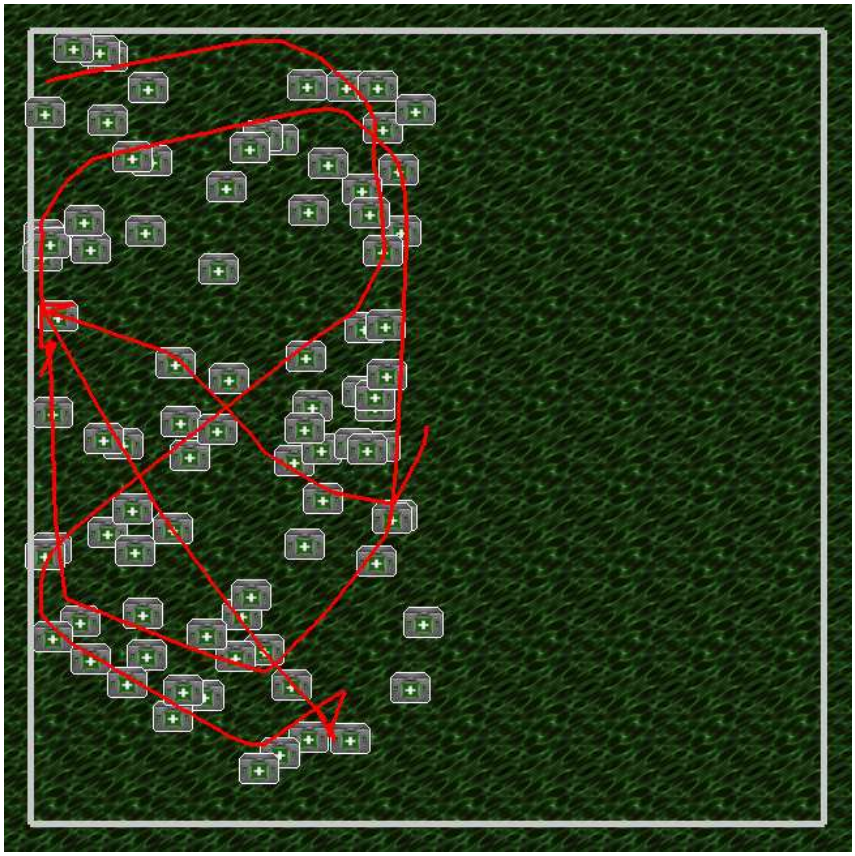


Figura 26 – Trajetória do agente no cenário **Meio a Meio**.

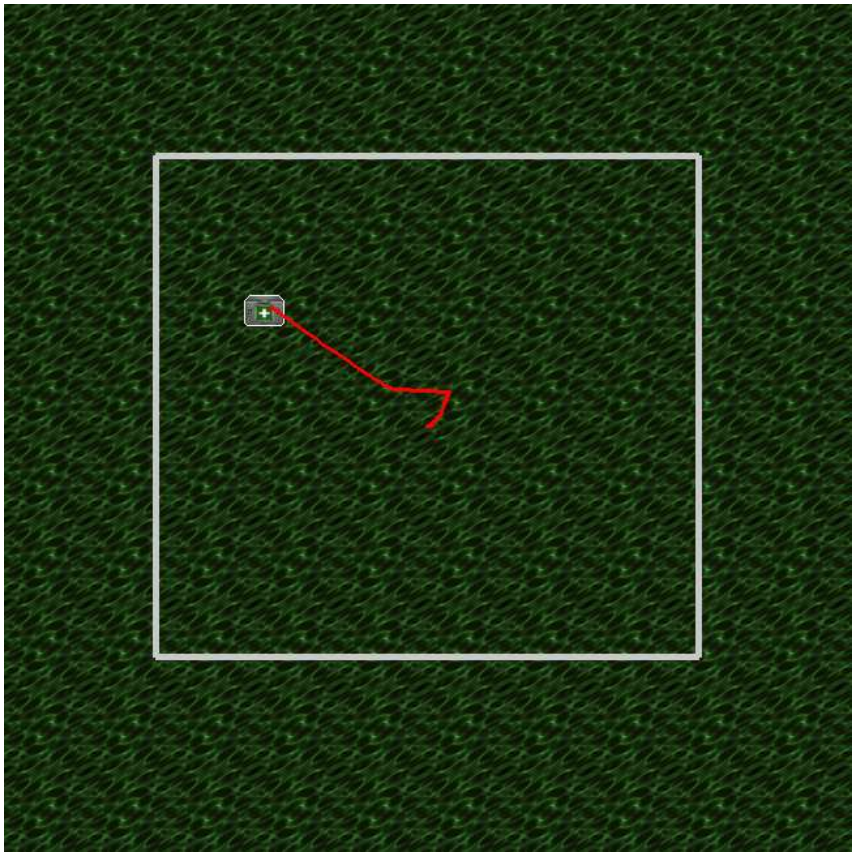


Figura 27 – Trajetória do agente no cenário **Busca Imediata**.

4.2.2 Treinando um agente multissensorial

Confirmada então a habilidade de ouvir dos agentes, iniciaram-se os testes multissensoriais, principal objeto de estudo deste trabalho. Primeiro, para validar os cenários desenvolvidos (**HG+30**, **HG=30**, **HGS+30**), um agente dotado de ambos os sensores, visual e sonoro, denominado *Image and Sound*, foi treinado em todos os três cenários individualmente para garantir que pudesse aprender com cada um deles. Isso foi alcançado, com desempenho máximo e sobrevivência em todos os 4200 passos, após cerca de 400 milhões de passos de treinamento para cada cenário. A trajetória de uma execução exemplo deste agente no cenário **HGS** pode ser observada na Figura 28.

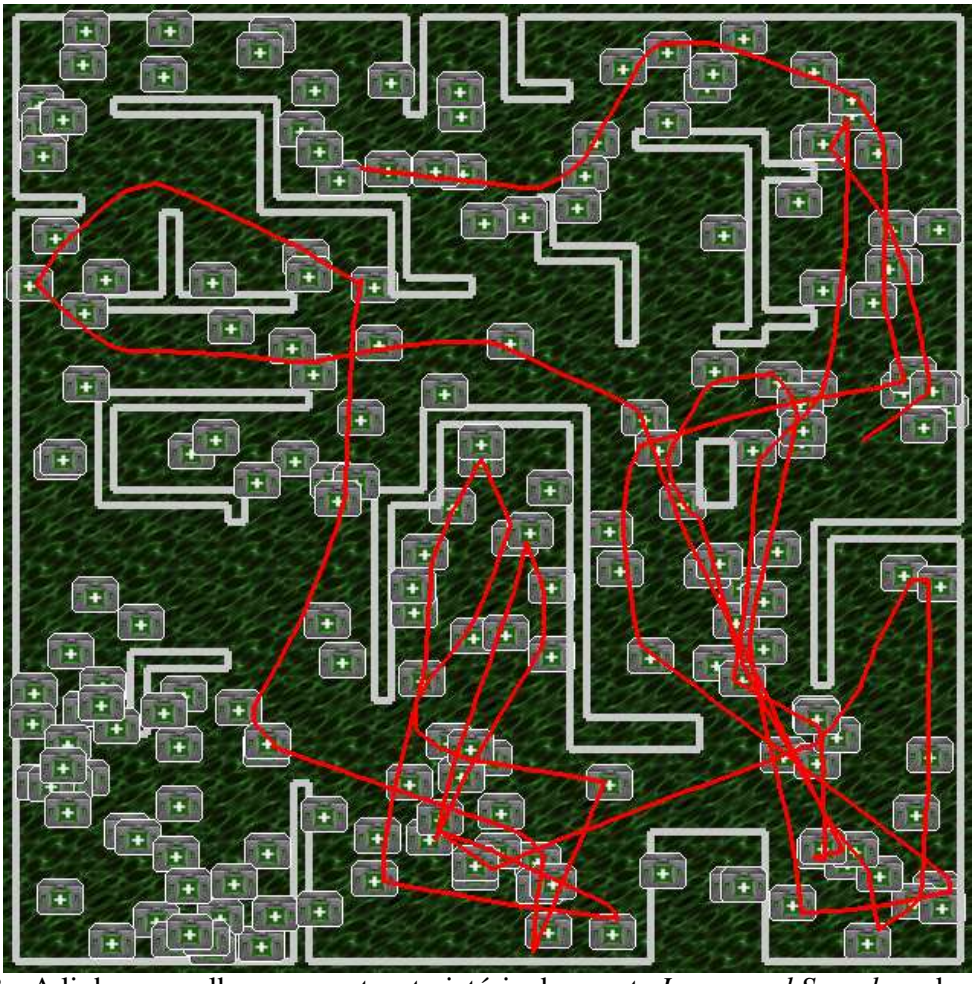


Figura 28 – A linha vermelha representa a trajetória do agente *Image and Sound* sendo executado no cenário **HGS**.

Entretanto, os cenários *Health Gathering* são capazes de serem resolvidos apenas utilizando a visão, o que levou ao questionamento; **O agente *Image and Sound* está usando o áudio ou apenas a visão para solucionar a tarefa proposta?** Para sanar essa dúvida, foram

realizados dois testes:

- **Teste de Oclusão**, em que executa-se o agente treinado utilizando o cenário apresentado na Subseção 4.1.5. Se o agente aprendeu a utilizar o som, ele deve ser capaz de coletar o *Medkit*; e
- **Teste Cego**, no qual remove-se o componente visual do agente previamente treinado, avaliando-o novamente nos mesmos três cenários. O objetivo é verificar se, utilizando apenas o som, o agente é capaz de continuar a resolver a tarefa com eficácia.

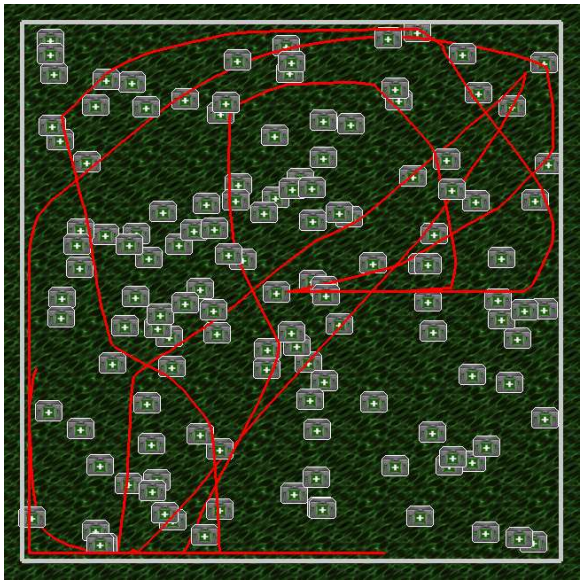
O resultado do Teste de Oclusão mostrou que o agente *Image and Sound* só foi capaz de coletar o *Medkit* em $29,91\%$ das 10.000 tentativas realizadas. Além disso, só pegou o item de imediato após passar pela parede em apenas $19,95\%$ das vezes, tendo nas outras avançado e retornado na direção correta com tempo suficiente para não ser eliminado. Esse foi o primeiro indício de que o som não estava sendo utilizado pelo agente.

Restou então o resultado do Teste Cego para confirmar a hipótese levantada. Ao executar o agente *Image and Sound* nos três cenários, utilizando apenas o sensor sonoro como entrada, observou-se que ele não foi sequer capaz de se movimentar pelo cenário. Durante todos os episódios de teste, o agente permaneceu realizando um comportamento que foi nomeado de **Comportamento de Torre** ou *Turret Behavior*. A conduta do agente consiste em ficar apenas girando em torno do próprio eixo, sem sair do lugar ou coletar qualquer *Medkit*, similar a uma torre sentinela. Uma explicação para esse estado é que o agente está procurando visualmente por algum sinal que o motive a sair do lugar, porém, com o sensor visual desligado, tal sinal nunca é encontrado, levando-o a manter a mesma ação indefinidamente.

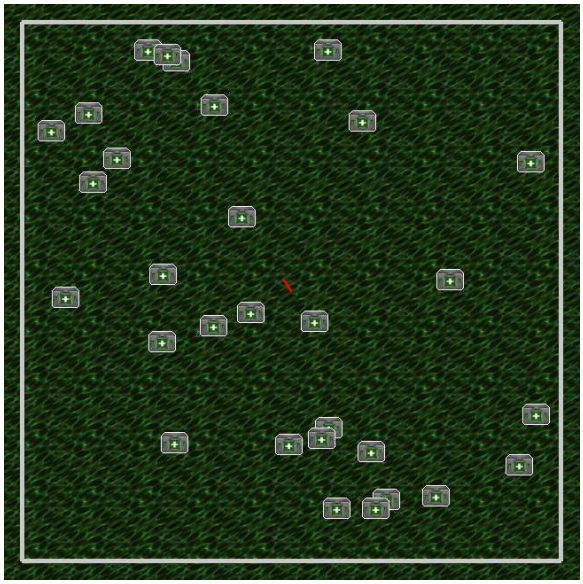
Por último, foi realizado um novo experimento para avaliar o comportamento do primeiro agente. Um novo agente, denominado *Image Only*, foi treinado do zero utilizando apenas a entrada visual e seu desempenho foi comparado ao do agente anterior, que utilizava ambas as entradas. Para executar a avaliação, a entrada de áudio foi zerada. Após 400 milhões de etapas, ele já alcançava desempenho máximo, assim como o agente anterior. A recompensa média recebida por ambos os agentes durante o treinamento foi medida, conforme mostrado na Figura 30. No Capítulo 5 será discutida em mais detalhes a comparação entre os dois agentes.

4.3 Definição do Currículo

Os testes realizados com o agente *Image and Sound* e o experimento do *Image Only* deixaram claro que treinar um sistema multissensorial, fornecendo ambas as informações sem



(a) Agente *Image and Sound* utilizando ambos os sensores disponíveis.



(b) Agente *Image and Sound* com o sensor visual desligado.

Figura 29 – Avaliação do agente *Image and Sound* no cenário **HG+30**, sendo possível observar o comportamento de torre.

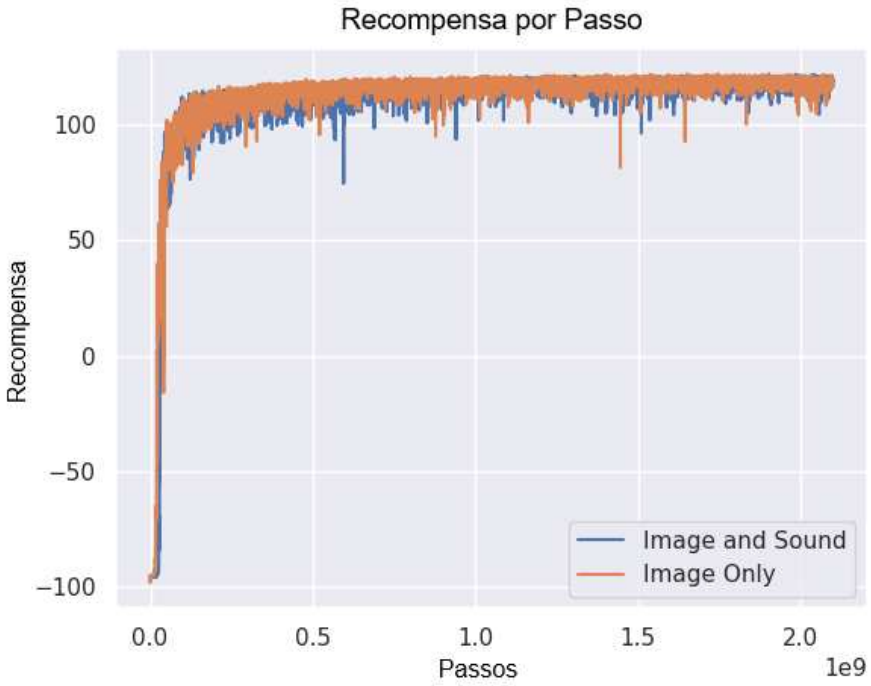


Figura 30 – Recompensa média recebida durante o treinamento pelos agentes *Image and Sound* e *Image Only*.

uma estratégia de treinamento mais robusta, pode não ser suficiente para obter um desempenho eficiente com ambos os sensores. Com isso em mente, buscou-se definir um currículo capaz de equilibrar o aprendizado entre as diferentes fontes de dados, resultando em um agente capaz de aproveitar toda a informação disponível para se adaptar a diferentes cenários sensoriais.

Para isso, um agente foi treinado utilizando uma estratégia de currículo desenvolvida para forçá-lo a aprender a se guiar usando tanto informações visuais quanto sonoras. O currículo consistiu em uma combinação dos três cenários mencionados na Seção 4.1.2, do mais fácil ao mais difícil: **HG+30** $\rightarrow$ **HG=30** $\rightarrow$ **HGS+30**. No entanto, nos dois primeiros cenários de treinamento, apenas o sensor de som está ativo, forçando o agente a aprender a sobreviver utilizando apenas o som emitido pelos *Medkits*. Ao chegar ao terceiro cenário, o sensor de visão também é incluído, visto que é necessário para aprender a evitar as paredes internas do labirinto.

O agente, nomeado *Curriculum*, foi treinado por 500 milhões de etapas no cenário **HG+30**, 1 bilhão de etapas no **HG=30** e, finalmente, 600 milhões de etapas no **HGS+30**. Ele alcançou resultados satisfatórios em cada cenário individualmente e, ao final de todo o treinamento, foi capaz de obter o mesmo desempenho que os dois agentes anteriores, aprendendo a sobreviver durante toda a duração do episódio. A Figura 31 mostra a recompensa recebida durante toda a execução do currículo, e a Figura 32 mostra a trajetória do agente operando no cenário **HGS**.

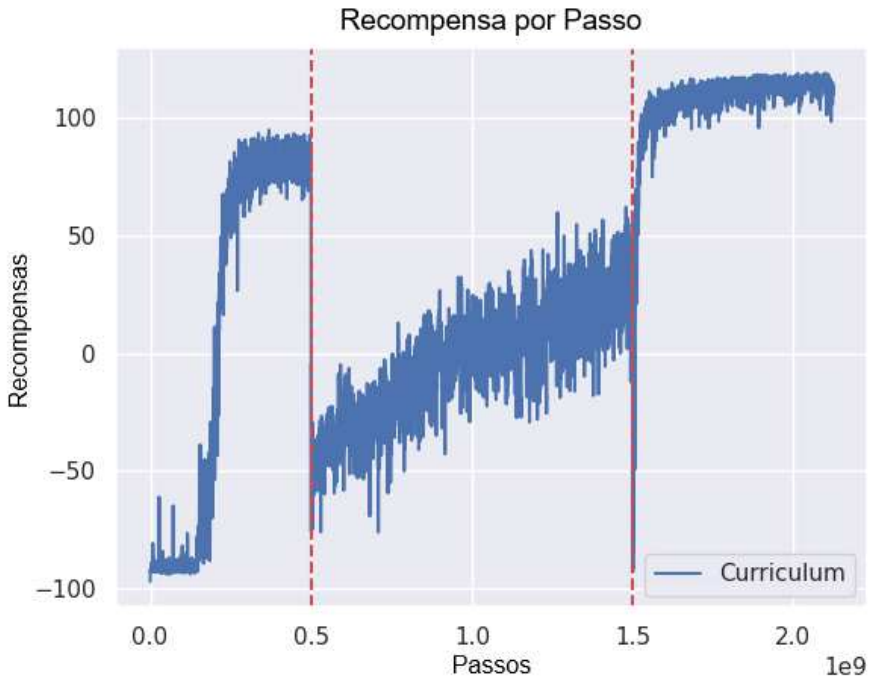


Figura 31 – Recompensa recebida durante o treinamento pelo agente *Curriculum*. As linhas verticais vermelhas indicam quando houve mudança na fase do currículo.

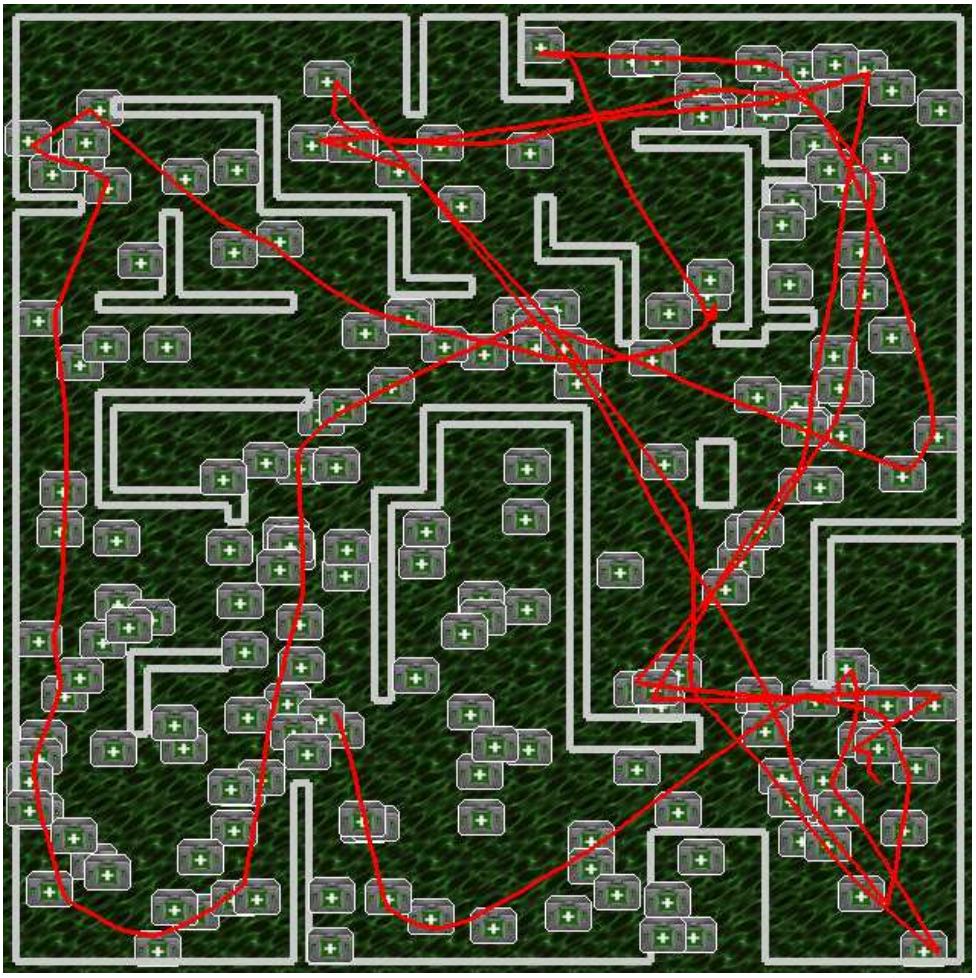


Figura 32 – Trajetória do agente *Curriculum* enquanto operava no cenário **HGS**.

5 DISCUSSÃO DOS RESULTADOS

O Capítulo 4 apresentou a metodologia e os experimentos realizados ao longo do desenvolvimento deste trabalho, objetivando o treinamento de um agente capaz de realizar forrageamento multissensorial. Neste capítulo, os principais resultados obtidos nos experimentos serão discutidos e algumas análises baseadas nesses resultados serão apresentadas.

5.1 Avaliando os agentes treinados

A Seção 4.2.2 apresentou o agente *Image and Sound*, treinado na tarefa de forrageamento utilizando som e imagem como entrada. Como visto na Figura 28, tal agente foi capaz de aprender a completar a tarefa de forrageamento, porém não era capaz de se guiar utilizando o áudio quando a imagem não estava presente, como evidenciado pelo Teste Cego.

A hipótese é que, no cenário de *Health Gathering*, o uso da visão é suficiente para completar a tarefa, como demonstrado pelo segundo agente, denominado *Image Only*. Dessa forma, ao iniciar o treinamento com ambos os sensores (visual e sonoro) ativados, espera-se que o agente passe a depender exclusivamente da entrada visual. Os experimentos indicam que o sensor visual fornece informações mais relevantes para o agente nessa tarefa, levando-o a ignorar o sensor sonoro e a concentrar-se apenas no que vê.

O resultado do Teste Cego, visto na Figura 29b, reforça a hipótese acima, pois, mesmo possuindo o sensor de som, o agente é incapaz de completar a tarefa. Comparando esse resultado com o comportamento apresentado pelo *Image Only* no Teste Cego, visto na Figura 33, vê-se que ambos consideram somente a imagem para decidir que ação executar, resultando em agentes igualmente monossensoriais.

Por outro lado, um agente treinado com o *Curriculum* proposto ainda alcança a mesma performance final de sobrevivência em todos os 4200 passos, mas uma avaliação cega realizada da mesma maneira mostra que ele realmente aprendeu a usar o som para localizar os *Medkits*. Quando a visão do agente é desativada, ele perde um pouco de desempenho, porém ainda é capaz de sobreviver, não replicando o *Comportamento de Torre* dos agentes anteriores, conforme demonstrado na Figura 34. Isso proporciona uma maior adaptabilidade ao agente, pois em um cenário onde existe oscilação na quantidade de informação disponível, o resultado final não é tão negativamente afetado.

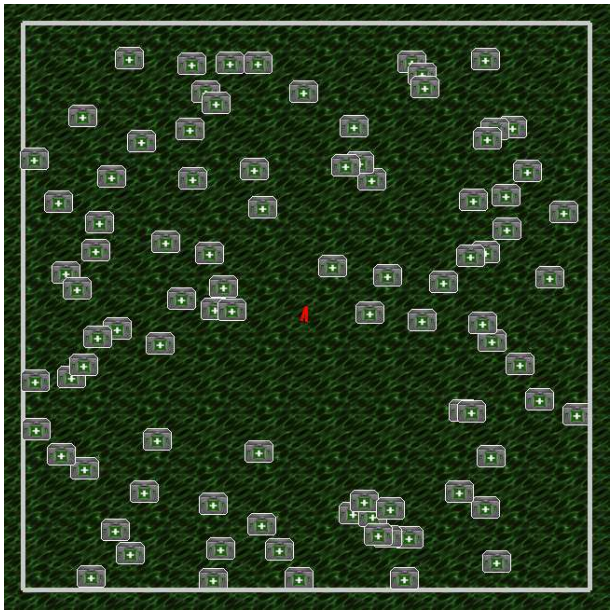
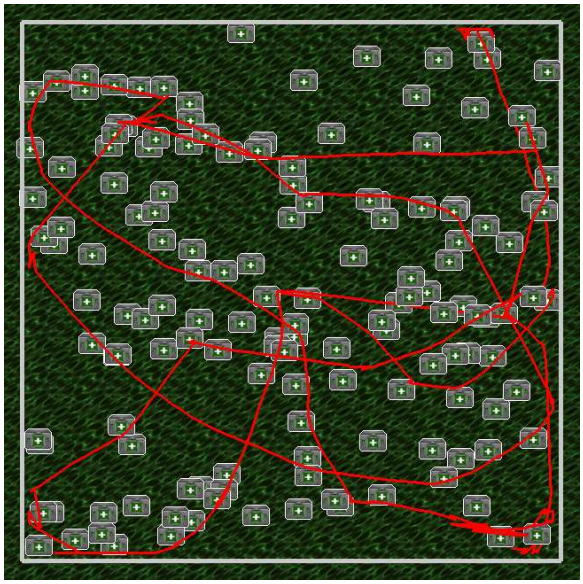
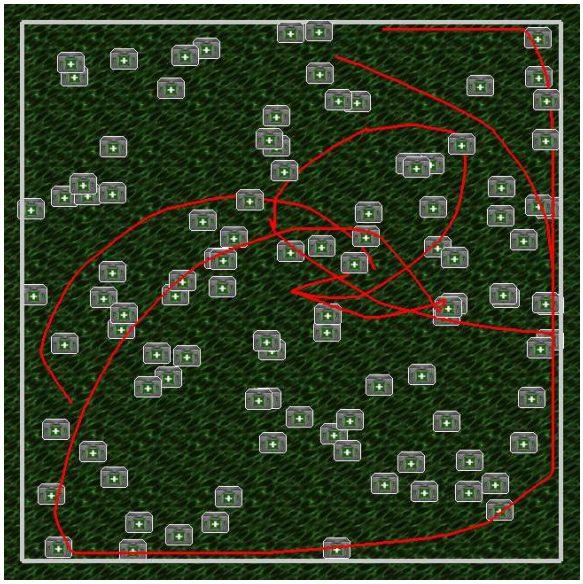


Figura 33 – Trajetória do agente *Image Only* durante avaliação do **Teste Cego**.



(a) Execução do agente *Curriculum* com ambos os sensores disponíveis.



(b) Execução do agente *Curriculum* no cenário **HG+30** com a entrada visual desativada.

Figura 34 – Avaliação do agente *Curriculum* em duas situações distintas. A Figura 34b mostra que mesmo na ausência da visão o agente conseguiu explorar o ambiente em busca de recursos para continuar vivo.

5.2 Distribuição de Ações

Outra forma de avaliar a capacidade dos agentes treinados de completar a tarefa de forrageamento e de fazer uso de ambos sensores é por meio da observação das ações executadas durante suas rodadas de avaliação. A Tabela 1 mostra a distribuição de ações executadas pelos três agentes no cenário mais complexo, **HGS+30**, após 1000 rodadas de execução com seus respectivos sensores ativos (visual e sonoro para os agentes multissensoriais e somente visual

para o agente *Image Only*).

	Image and Sound	*Image Only*	*Curriculum*
Nenhuma Ação	2.28%	1.57%	3.06%
Esquerda	18.17%	14.57%	16.05%
Direita	20.84%	24.81%	22.15%
Frente	57.66%	58.20%	57.66%
Trás	1.05%	0.84%	1.08%

Tabela 1 – Distribuição de ações no cenário **HGS+30**.

O equilíbrio visto reflete o aprendizado dos três agentes neste cenário, no qual todos foram capazes de obter o resultado máximo. Para tal, tiveram que aprender a utilizar as diferentes ações possíveis, a partir das variações de entradas que recebiam. No entanto, observa-se que os agentes com capacidades multissensoriais realizam a ação de retroceder com mais frequência, o que indica que podem ser capazes de se mover em direção aos *Medkits* sem precisar virar e vê-los. Além disso, o agente *Image Only* executa a ação de avançar com mais frequência e permanece parado com menos regularidade, o que corrobora o fato de que ele precisa se mover para ter os *Medkits* em seu campo de visão.

Ao comparar a distribuição das ações *Esquerda* e *Direita* apenas, nota-se que o agente com o *Curriculum* apresenta um comportamento mais equilibrado, seguido pelo agente *Image and Sound*, enquanto o agente *Image Only* prefere virar para *Direita* na maior parte do tempo. Essa distribuição pode ser observada na Tabela 2. Os resultados dessa distribuição de ações sugerem que os agentes com sensores de áudio apresentam um comportamento mais fluido, movendo-se e girando em direção aos kits médicos de maneira mais uniforme. Isso contrasta com os agentes que utilizam apenas visão, os quais se movem frequentemente para ver os kits médicos e apresentam uma direção de rotação preferencial.

	Image and Sound	*Image Only*	*Curriculum*
Esquerda	42.02%	37.00%	**46.58%**
Direita	57.98%	63.00%	**53.42%**

Tabela 2 – Distribuição das ações Esquerda e Direita no cenário **HGS+30**.

A Tabela 3 mostra a diferença na distribuição de ações entre os três agentes em execução cega por 1000 episódios, mais uma vez demonstrando que o agente com *Curriculum* é capaz de utilizar o som para decidir qual ação realizar. O *Comportamento de Torre* também é exemplificado na Tabela 3, onde cerca de 98% de todas as ações realizadas pelos agentes *Image and Sound* e *Image Only* foram a ação de *Virar à Direita*, como se estivesse em busca

de uma entrada visual para começar a se mover para frente. Contrastando diretamente com o comportamento observado no cenário onde a visão estava ativada. Por outro lado, o agente *Curriculum* apresenta uma distribuição de ações muito mais ampla e similar ao que é visto na Tabela 1, enfatizando sua capacidade de se adaptar em cenários onde há alteração nos sensores de entrada.

	Image and Sound	*Image Only*	***Curriculum***
Nenhuma Ação	0.41%	0.48%	**5.83%**
Esquerda	0.48%	0.27%	**1.93%**
Direita	98.29%	98.15%	**30.72%**
Frente	0.23%	0.09%	**57.20%**
Trás	0.59%	1.01%	**4.32%**

Tabela 3 – Distribuição de ações no cenário **HG=30** com sensores visuais desativados.

5.2.1 Resumo dos Experimentos

Os experimentos realizados produziram diversas descobertas significativas, que estão resumidas abaixo

- **Capacidade de ouvir:** Os experimentos *Meio a Meio* e *Busca Imediata* mostraram a capacidade que agentes de Aprendizado por Reforço possuem de aprender utilizando informações sonoras apenas. Em todos os cenários, o agente treinado conseguiu executar a tarefa proposta sem dificuldade, utilizando apenas o áudio do jogo como entrada, validando esse sensor para os experimentos mais robustos;
- **Desempenho do agente *Image and Sound***: O agente inicial, equipado com sensores visuais e auditivos, aprendeu com sucesso a completar as tarefas de forrageamento em todos os cenários, alcançando um desempenho máximo e sobrevivendo a todos os 4200 passos após cerca de 400 milhões de passos de treinamento por cenário. No entanto, na ausência de entrada visual, ele não conseguiu navegar de forma eficaz, destacando uma dependência excessiva de indicadores visuais;
- **Desempenho do agente *Image Only***: Um segundo agente, treinado exclusivamente com entradas visuais, alcançou desempenho semelhante ao do agente multissensorial. Isso indica que a informação visual por si só é suficiente para as tarefas de forrageamento nos cenários projetados, explicando a falha do agente multissensorial quando privado de entrada visual;
- **Estratégia de treinamento com *Curriculum***: Para abordar o uso limitado do sensor

de audio, foi introduzida uma estratégia de treinamento com currículo que obrigou o agente a depender inicialmente de informações auditivas antes de integrar a entrada visual. O agente foi treinado em três fases, começando com entrada apenas de som em cenários mais simples e progredindo para entrada combinada no cenário mais complexo. O agente treinado com currículo demonstrou um desempenho robusto em todos os cenários. Notavelmente, manteve o uso de informações auditivas em testes sem visão, evitando assim o *Turret Behavior* observado nos outros agentes;

– **Distribuição de ações e análise comportamental**: A análise da distribuição de ações mostrou que o agente treinado com currículo apresentou uma distribuição de ações mais equilibrada em comparação aos agentes multissensoriais e de apenas visão. Isso indica um processo de tomada de decisão mais refinado, utilizando de forma eficaz tanto as entradas visuais quanto as sonoras.

Os resultados deste trabalho reforçam a eficácia das estratégias de treinamento com currículo como forma de promover a integração de múltiplos sensores em agentes de *Deep Reinforcement Learning* (DRL).

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho teve como objetivo estudar as capacidades multissensoriais de agentes de *Deep Reinforcement Learning* em tarefas de forrageamento. Nesse sentido, foram desenvolvidos cenários de DRL utilizando a plataforma *ViZDoom*, com diferentes níveis de dificuldade e objetivos. Todos os ambientes continham fontes de informações visuais e sonoras que os agentes podiam utilizar.

Inicialmente, foram executados dois experimentos como forma de validar a capacidade de aprender utilizando apenas o áudio. Ambos resultaram em agentes capazes de realizar as respectivas tarefas apresentadas, permitindo, assim, a realização de experimentos mais robustos. Em seguida, uma abordagem de treinamento multissensorial direto produziu um agente que aprende a completar uma tarefa de forrageamento no qual ambos os sensores estão disponíveis, mas não consegue ter sucesso em um cenário com apenas a entrada de áudio. Isso evidencia uma subutilização do sensor auditivo, de forma a tornar o agente totalmente dependente do estímulo visual.

Para contornar esse problema, foi proposta uma estratégia de treinamento com currículo, visando capacitar um agente a sobreviver de maneira eficaz em um cenário com variabilidade de informação. Os experimentos subsequentes mostraram que, enfatizando inicialmente entradas sensoriais menos dominantes, é possível desenvolver agentes capazes de comportamentos adaptativos e robustos em ambientes sensoriais variados, sem perdas de desempenho.

Trabalhos futuros podem explorar modalidades sensoriais adicionais e estratégias de currículo mais complexas para aprimorar ainda mais a adaptabilidade e o desempenho dos agentes. Novos cenários poderiam ser implementados para aproveitar a capacidade dos agentes de aprender a partir de diferentes graus de informação auditiva. Além disso, no desenvolvimento deste trabalho, foi utilizado apenas um arquivo de áudio durante os experimentos. Uma questão interessante a ser investigada é se um agente consegue aprender diferentes padrões de áudio simultaneamente, assim como variações do mesmo áudio, como volume, velocidade e tonalidade. Essa investigação poderia revelar informações valiosas sobre a integração multissensorial e o aprendizado em ambientes dinâmicos.

O trabalho apresentado nesta dissertação foca em demonstrar que um currículo bem elaborado pode melhorar significativamente as capacidades de aprendizado e adaptação de agentes de *Deep Reinforcement Learning*, garantindo que eles possam utilizar todas as informações sensoriais disponíveis para tomar decisões informadas em ambientes complexos e dinâmicos.

Esta pesquisa contribui para os campos de agentes autônomos em jogos e forrageamento, além de abrir novas possibilidades para trabalhos futuros na integração multissensorial em outras tarefas complexas.

REFERÊNCIAS

ARULKUMARAN, K.; DEISENROTH, M.; BRUNDAGE, M.; BHARATH, A. A brief survey of deep reinforcement learning. **IEEE Signal Processing Magazine**, v. 34, 08 2017.

BAKER, B.; KANITSCHEIDER, I.; MARKOV, T.; WU, Y.; POWELL, G.; MCGREW, B.; MORDATCH, I. Emergent tool use from multi-agent autocurricula. In: **International Conference on Learning Representations (ICLR)**. [*S. n.*], 2020. p. 1–28. Disponível em: https://arxiv.org/abs/1909.07528.

BELLEMARE, M. G.; NADDAF, Y.; VENESS, J.; BOWLING, M. The arcade learning environment: An evaluation platform for general agents. **Journal of Artificial Intelligence Research**, v. 47, p. 253–279, 2013.

BENGIO, Y.; LOURADOUR, J.; COLLOBERT, R.; WESTON, J. Curriculum learning. In: **Proceedings of the 26th Annual International Conference on Machine Learning**. New York, NY, USA: Association for Computing Machinery, 2009. (ICML '09), p. 41–48. ISBN 9781605585161. Disponível em: https://doi.org/10.1145/1553374.1553380.

BROCKMAN, G.; CHEUNG, V.; PETTERSSON, L.; SCHNEIDER, J.; SCHULMAN, J.; TANG, J.; ZAREMBA, W. OpenAI Gym. **ArXiv**, abs/1606.01540, 2016. Disponível em: http://arxiv.org/abs/1606.01540.

BURCHAM, S. E. The video game [r] evolution. **XRDS: Crossroads, The ACM Magazine for Students**, ACM New York, NY, USA, v. 3, n. 2, p. 7–8, 1996.

CHEN, C.; JAIN, U.; SCHISSLER, C.; GARI, S. V. A.; AL-HALAH, Z.; ITHAPU, V. K.; ROBINSON, P.; GRAUMAN, K. Soundspaces: Audio-visual navigaton in 3d environments. In: **ECCV**. [*S. l.*: *s. n.*], 2020.

FILHO, R. F. F.; NOGUEIRA, Y. L. B.; VIDAL, C. A.; CAVALCANTE-NETO, J. B.; SERAFIM, P. B. de S. Gym hero: A research environment for reinforcement learning agents in rhythm games. In: **2021 20th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)**. [*S. l.*: *s. n.*], 2021. p. 87–96.

GAINA, R. D.; STEPHENSON, M. "did you hear that?" learning to play video games from audio cues. In: **2019 IEEE Conference on Games (CoG)**. [*S. l.*: *s. n.*], 2019. p. 1–4.

GAN, C.; ZHANG, Y.; WU, J.; GONG, B.; TENENBAUM, J. B. Look, listen, and act: Towards audio-visual embodied navigation. **CoRR**, abs/1912.11684, 2019. Disponível em: http://arxiv.org/abs/1912.11684.

GARCIA-ROMERO, D.; SELL, G.; MCCREE, A. Magneto: X-vector magnitude estimation network plus offset for improved speaker recognition. In: **Proc. Odyssey 2020 the speaker and language recognition workshop**. [*S. l.*: *s. n.*], 2020. p. 1–8.

GIANNAKOPOULOS, P.; PIKRAKIS, A.; COTRONIS, Y. A deep reinforcement learning approach for audio-based navigation and audio source localization in multi-speaker environments. **CoRR**, abs/2110.12778, 2021. Disponível em: https://arxiv.org/abs/2110.12778.

GORDILLO, C.; BERGDAHL, J.; TOLLMAR, K.; GISSL'EN, L. Improving playtesting coverage via curiosity driven reinforcement learning agents. **2021 IEEE Conference on Games (CoG)**, p. 1–8, 2021.

GRIFFIN, D.; LIM, J. Signal estimation from modified short-time fourier transform. **IEEE Transactions on acoustics, speech, and signal processing**, IEEE, v. 32, n. 2, p. 236–243, 1984.

GRIMSHAW, M.; TAN, S.-L.; LIPSCOMB, S. D. Playing with sound: The role of music and sound effects in gaming. In: **The Psychology of Music in Multimedia**. [*S. l.*]: Oxford University Press, 2013. ISBN 9780199608157.

GUILLEN, G.; JYLHä, H.; HASSAN, L. The role sound plays in games: A thematic literature study on immersion, inclusivity and accessibility in game sound research. In: **Proceedings of the 24th International Academic Mindtrek Conference**. New York, NY, USA: Association for Computing Machinery, 2021. (Academic Mindtrek '21), p. 12–20. ISBN 9781450385145. Disponível em: https://doi.org/10.1145/3464327.3464365.

HAARNOJA, T.; ZHOU, A.; ABBEEL, P.; LEVINE, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: **ICML**. [*S. l.*: *s. n.*], 2018.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [*S. l.*: *s. n.*], 2016. p. 770–778.

HEGDE, S.; KANERVISTO, A.; PETRENKO, A. Agents that listen: High-throughput reinforcement learning with multiple sensory systems. In: **2021 IEEE Conference on Games (CoG)**. [*S. l.*: *s. n.*], 2021. p. 1–5.

HUGILL, A.; AMELIDES, P. Audio-only computer games: Papa sangre. In: ____. **Expanding the Horizon of Electroacoustic Music Analysis**. [*S. l.*]: Cambridge University Press, 2016. p. 355–375.

JULIANI, A.; BERGES, V.-P.; TENG, E.; COHEN, A.; HARPER, J.; ELION, C.; GOY, C.; GAO, Y.; HENRY, H.; MATTAR, M. *et al.* Unity: A general platform for intelligent agents. **ArXiv**, abs/1809.02627, 2018. Disponível em: http://arxiv.org/abs/1809.02627.

KEELEY, B. L. Making sense of the senses: Individuating modalities in humans and other animals. **The Journal of Philosophy**, Journal of Philosophy, Inc., v. 99, n. 1, p. 5–28, 2002. ISSN 0022362X. Disponível em: http://www.jstor.org/stable/3655759.

KEMPKA, M.; WYDMUCH, M.; RUNC, G.; TOCZEK, J.; JAŚKOWSKI, W. Vizdoom: A Doom-based AI research platform for visual reinforcement learning. In: IEEE. **2016 IEEE Conference on Computational Intelligence and Games (CIG)**. [*S. l.*], 2016. p. 1–8.

MNIH, V.; BADIA, A. P.; MIRZA, M.; GRAVES, A.; LILLICRAP, T.; HARLEY, T.; SILVER, D.; KAVUKCUOGLU, K. Asynchronous methods for deep reinforcement learning. In: BALCAN, M. F.; WEINBERGER, K. Q. (Ed.). **Proceedings of The 33rd International Conference on Machine Learning**. New York, New York, USA: PMLR, 2016. (Proceedings of Machine Learning Research, v. 48), p. 1928–1937. Disponível em: https://proceedings.mlr.press/v48/mniha16.html.

MNIH, V.; KAVUKCUOGLU, K.; SILVER, D.; GRAVES, A.; ANTONOGLOU, I.; WIERSTRA, D.; RIEDMILLER, M. A. Playing atari with deep reinforcement learning. **ArXiv**, abs/1312.5602, 2013. Disponível em: http://arxiv.org/abs/1312.5602.

MNIH, V.; KAVUKCUOGLU, K.; SILVER, D.; RUSU, A. a.; VENESS, J.; BELLEMARE, M. G.; GRAVES, A.; RIEDMILLER, M.; FIDJELAND, A. K.; OSTROVSKI, G.; PETERSEN, S.; BEATTIE, C.; SADIK, A.; ANTONOGLOU, I.; KING, H.; KUMARAN, D.; WIERSTRA, D.; LEGG, S.; HASSABIS, D. Human-level control through deep reinforcement learning. **Nature**, v. 518, n. 7540, p. 529–533, 2015. ISSN 0028-0836.

MUNOZ, N. E.; BLUMSTEIN, D. T. Multisensory perception in uncertain environments. **Behavioral Ecology**, v. 23, n. 3, p. 457–462, 01 2012. ISSN 1045-2249. Disponível em: https://doi.org/10.1093/beheco/arr220.

OPENAI; BERNER, C.; BROCKMAN, G.; CHAN, B.; CHEUNG, V.; DEBIAK, P.; DENNISON, C.; FARHI, D.; FISCHER, Q.; HASHME, S.; HESSE, C.; JóZEFOWICZ, R.; GRAY, S.; OLSSON, C.; PACHOCKI, J.; PETROV, M.; PINTO, H. P. de O.; RAIMAN, J.; SALIMANS, T.; SCHLATTER, J.; SCHNEIDER, J.; SIDOR, S.; SUTSKEVER, I.; TANG, J.; WOLSKI, F.; ZHANG, S. Dota 2 with large scale deep reinforcement learning. **ArXiv**, abs/1912.06680, p. 1–66, 2019. Disponível em: https://arxiv.org/abs/1912.06680.

PARK, K.; OH, H.; LEE, Y. Veca: A toolkit for building virtual environments to train and test human-like agents. **arXiv preprint arXiv:2105.00762**, 2021.

PÉREZ-LIÉBANA, D.; LIU, J.; KHALIFA, A.; GAINA, R. D.; TOGELIUS, J.; LUCAS, S. M. M. General video game ai: A multitrack framework for evaluating agents, games, and content generation algorithms. **IEEE Transactions on Games**, v. 11, p. 195–214, 2019.

PETRENKO, A.; HUANG, Z.; KUMAR, T.; SUKHATME, G.; KOLTUN, V. Sample factory: Egocentric 3d control from pixels at 100000 fps with asynchronous reinforcement learning. In: **ICML**. [*S. l.*: *s. n.*], 2020.

RUMMERY, G. A.; NIRANJAN, M. **On-line Q-learning using connectionist systems**. [*S. l.*]: University of Cambridge, Department of Engineering Cambridge, UK, 1994. v. 37.

SAVVA, M.; KADIAN, A.; MAKSYMETS, O.; ZHAO, Y.; WIJMANS, E.; JAIN, B.; STRAUB, J.; LIU, J.; KOLTUN, V.; MALIK, J. *et al.* Habitat: A platform for embodied ai research. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [*S. l.*: *s. n.*], 2019. p. 9339–9347.

SCHRITTWIESER, J.; ANTONOGLOU, I.; HUBERT, T.; SIMONYAN, K.; SIFRE, L.; SCHMITT, S.; GUEZ, A.; LOCKHART, E.; HASSABIS, D.; GRAEPEL, T.; LILLICRAP, T.; SILVER, D. Mastering atari, go, chess and shogi by planning with a learned model. **ArXiv e-prints**, p. 1–21, 2019. Disponível em: https://arxiv.org/abs/1911.08265.

SCHULMAN, J. Trust region policy optimization. **arXiv preprint arXiv:1502.05477**, 2015.

SCHULMAN, J.; MORITZ, P.; LEVINE, S.; JORDAN, M.; ABBEEL, P. High-dimensional continuous control using generalized advantage estimation. **arXiv preprint arXiv:1506.02438**, 2015.

SCHULMAN, J.; WOLSKI, F.; DHARIWAL, P.; RADFORD, A.; KLIMOV, O. Proximal policy optimization algorithms. **arXiv preprint arXiv:1707.06347**, 2017.

SERAFIM, P. B. de S.; NOGUEIRA, Y. L. B.; VIDAL, C. A.; CAVALCANTE-NETO, J. B.; FILHO, R. F. Investigating deep q-network agent sensibility to texture changes on fps games.

In: **2020 19th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)**. [*S. l.*: *s. n.*], 2020. p. 117–125.

SILVER, D.; HUBERT, T.; SCHRITTWIESER, J.; ANTONOGLOU, I.; LAI, M.; GUEZ, A.; LANCTOT, M.; SIFRE, L.; KUMARAN, D.; GRAEPEL, T.; LILLICRAP, T.; SIMONYAN, K.; HASSABIS, D. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. **ArXiv e-prints**, p. 1–19, 2017. Disponível em: http://arxiv.org/abs/1712.01815.

SUTTON, R. S.; BARTO, A. G. **Reinforcement Learning: An Introduction**. Cambridge, MA, USA: A Bradford Book, 2018. ISBN 0262039249.

SZOT, A.; CLEGG, A.; UNDERSANDER, E.; WIJMANS, E.; ZHAO, Y.; TURNER, J.; MAESTRE, N.; MUKADAM, M.; CHAPLOT, D.; MAKSYMETS, O.; GOKASLAN, A.; VONDRUS, V.; DHARUR, S.; MEIER, F.; GALUBA, W.; CHANG, A.; KIRA, Z.; KOLTUN, V.; MALIK, J.; SAVVA, M.; BATRA, D. Habitat 2.0: Training home assistants to rearrange their habitat. **arXiv preprint arXiv:2106.14405**, 2021.

TODOROV, E.; EREZ, T.; TASSA, Y. Mujoco: A physics engine for model-based control. In: **2012 IEEE/RSJ International Conference on Intelligent Robots and Systems**. [*S. l.*: *s. n.*], 2012. p. 5026–5033.

VINYALS, O.; BABUSCHKIN, I.; CZARNECKI, W. M.; MATHIEU, M.; DUDZIK, A.; CHUNG, J.; CHOI, D. H.; POWELL, R.; EWALDS, T.; GEORGIEV, P.; OH, J.; HORGAN, D.; KROISS, M.; DANIHELKA, I.; HUANG, A.; SIFRE, L.; CAI, T.; AGAPIOU, J. P.; JADERBERG, M.; VEZHNEVETS, A. S.; LEBLOND, R.; POHLEN, T.; DALIBARD, V.; BUDDEN, D.; SULSKY, Y.; MOLLOY, J.; PAINE, T. L.; GULCEHRE, C.; WANG, Z.; PFAFF, T.; WU, Y.; RING, R.; YOGATAMA, D.; WÜNSCH, D.; MCKINNEY, K.; SMITH, O.; SCHAUL, T.; LILLICRAP, T. P.; KAVUKCUOGLU, K.; HASSABIS, D.; APPS, C.; SILVER, D. Grandmaster level in starcraft ii using multi-agent reinforcement learning. **Nature**, p. 1–5, 2019.

WANG, X.; CHEN, Y.; ZHU, W. A survey on curriculum learning. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 44, n. 9, p. 4555–4576, 2022.

WATKINS, C. J.; DAYAN, P. Q-learning. **Machine learning**, Springer, v. 8, p. 279–292, 1992.

WEISS, K.; KHOSHGOFTAAR, T. M.; WANG, D. A survey of transfer learning. **Journal of Big data**, SpringerOpen, v. 3, n. 1, p. 1–40, 2016.

WERNER, E. E.; HALL, D. J. Optimal foraging and the size selection of prey by the bluegill sunfish (lepomis macrochirus). **Ecology**, Wiley Online Library, v. 55, n. 5, p. 1042–1052, 1974.

WESTNEAT, D.; FOX, C. W. **Evolutionary behavioral ecology**. [*S. l.*]: Oxford University Press, 2010.

WOUBIE, A.; KANERVISTO, A.; KARTTUNEN, J.; HAUTAMÄKI, V. Do autonomous agents benefit from hearing? **ArXiv**, abs/1905.04192, 2019.

ZDOOM. **ACS PlaySound Documentation**. 2020. https://zdoom.org/wiki/PlaySound [Accessed: (28/05/2024)].

ZENG, F.; WANG, C.; GE, S. S. A survey on visual navigation for artificial agents with deep reinforcement learning. **IEEE Access**, v. 8, p. 135426–135442, 2020.

ZHU, Y.; MOTTAGHI, R.; KOLVE, E.; LIM, J. J.; GUPTA, A.; FEI-FEI, L.; FARHADI, A. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: **2017 IEEE International Conference on Robotics and Automation (ICRA)**. [*S. l.*: *s. n.*], 2017. p. 3357–3364.