Avaliação de Agentes de Aprendizagem Profunda em Problemas de Roteamento e Alocação Espectral em Redes Ópticas Elásticas Multibandas

Paulo César Oliveira Brito, Thiago César Ferreira Bezerra de Carvalho, Noêmia Cíntia Sales Santos da Silva, Ana Júlia Fernandes de Brito Ameno, Helder Alves Pereira

***Resumo*— Este trabalho apresenta uma avaliação de agentes de aprendizagem profunda, utilizados em problemas de roteamento e alocação espectral em redes ópticas elásticas multibandas. A metodologia implementada avalia o estado da arte, por meio da variação de hiperparâmetros, dos agentes utilizados e da realização de novos treinamentos. Os resultados obtidos mostram que podem ser obtidos ganhos de performance nos agentes utilizados, obtendo-se melhores resultados em termos de probabilidade de bloqueio de chamadas, aproximadamente 9%, e média final de recompensa, aproximadamente $44{,}58$, ambos valores superiores aos encontrados em treinamentos realizados na literatura.**

***Palavras-Chave*— Agente, Aprendizagem Profunda por Reforço, Otimização, Rede Óptica Multibanda.**

***Abstract*— This paper presents an evaluation of deep learning agents, used in routing and spectral assignment problems in multiband elastic optical networks. The implemented methodology evaluates the state of the art, through the variation of hyperparameters, the used agents and the accomplishment of new trainings. The results obtained show that performance gains can be obtained in the agents used, obtaining better results in terms of calls blocking probability, approximately 9%, and final average reward, approximately 44.58, both higher values to those found in training conducted in the literature.**

***Keywords*— Agent, Multiband Optical Network, Optimization, Reinforcement Deep Learning.**

I. Introdução

Com o aumento do tráfego de dados e na demanda por serviços de grande largura de banda e alta velocidade, redes ópticas tradicionais enfrentam desafios significativos em termos de escalabilidade e eficiência [1]. As redes ópticas elásticas (EON – *Elastic Optical Network*) têm desempenhado um papel fundamental na evolução e aprimoramento dessas redes, oferecendo maior flexibilidade e capacidade de adaptação às necessidades do tráfego de dados [2].

Entretanto, apesar das vantagens oferecidas pelas EONs, o projeto e a operação eficiente dessas redes representam desafios de alta complexidade [1]. A solução do problema de roteamento, seleção de banda de transmissão, formato de modulação e espectro (RBMLSA – *Routing, Band, Modulation Level and Spectrum Assignment*) torna-se crucial para alcançar uma maior eficiência nessas redes [3]. Para enfrentar esse desafio, o uso de modelos de inteligência artificial, em particular do paradigma de aprendizagem por reforço (RL – *Reinforcement Learning*), tem atraído considerável atenção da comunidade científica [4]–[11].

Dessa forma, agentes de RL têm seus comportamentos controlados por hiperparâmetros, ou seja, variáveis que especificam como o agente deverá interpretar os dados do ambiente em que está inserido para efetuar a melhor ação [2]. A otimização adequada desses hiperparâmetros torna-se essencial para alcançar um desempenho ótimo dos modelos de aprendizagem por reforço, levando a uma alocação eficiente dos recursos nas redes EONs [12].

Este artigo apresenta uma avaliação de hiperparâmetros para o agente *Trust Region Policy Optimization* (TRPO), disponível na biblioteca *Stable-Baselines*, quando operando no ambiente de simulação de redes ópticas multibandas da biblioteca Optical-RL-Gym [2], utilizando a técnica de aprendizagem por reforço (vide Seção IV). Está organizado da seguinte forma: na Seção II, apresenta-se o estado da arte sobre técnicas de aprendizagem profunda utilizadas no cenário de redes ópticas multibandas (MB). Na Seção III, descreve-se os agentes e, na Seção IV, suas respectivas configurações nos treinamentos que foram realizados. Na Seção V, apresentam-se os resultados e, por fim, na Seção VI, são feitas as considerações finais.

II. Aprendizagem por Reforço em Redes Ópticas Multibandas

Esta seção apresenta o estado da arte com relação à aplicação da técnica de aprendizagem por reforço em redes ópticas multibandas. A área de aprendizagem por reforço busca propor abordagens computacionais para solucionar problemas formulados como processos de tomadas de decisões sequenciais [4]–[8]. Em aprendizagem por reforço, um agente inteligente aprende através da interação com o ambiente. Em cada instante de tempo, o agente observa o estado atual, escolhe e executa uma ação, observando em seguida uma recompensa e o estado seguinte [9]–[11].

Em [13], Natalino *et al.* apresentaram o Optical RL-Gym, um kit de ferramentas de código aberto já equipado com um conjunto básico de casos de uso de rede óptica empacotados como ambientes de RL. O objetivo principal do Optical RL-Gym é de reduzir o tempo e a complexidade para começar

Paulo César Oliveira Brito faz parte do Programa de Pós-Graduação em Engenharia Elétrica, Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande – PB, e-mail: paulo.brito@ee.ufcg.edu.br.

Thiago César Ferreira Bezerra de Carvalho, Noêmia Cíntia Sales Santos da Silva, Ana Júlia Fernandes de Brito Ameno e Helder Alves Pereira fazem parte da Unidade Acadêmica de Engenharia Elétrica, Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Campina Grande-PB, e-mails: {thiago.carvalho,noemia.silva,ana.julia.ameno}@ee.ufcg.edu.br e helder.pereira@dee.ufcg.edu.br.

a aplicar modelos de RL na solução de problemas de redes ópticas. O artigo evidencia técnicas de uso do ambiente em que seriam implementados os agentes a serem estudados, demonstrando o seu funcionamento e avaliando o desempenho frente às problemáticas de redes ópticas. Os autores também apresentaram a integração entre o ambiente e os agentes, demonstrada pela avaliação de quatro modelos de agentes de última geração, verificando que os agentes DQN, TRPO e PPO2 apresentaram melhores resultados de performance no cenário analisado.

Por sua vez, Morales *et al.* [1] implementaram e adicionaram dois novos ambientes ao Optical RL-Gym para resolver o problema de RBMLSA em EONs dinâmicas. Os resultados demonstraram a viabilidade do ambiente, pois o número de bandas disponíveis representadas pelo cenário analisado reduziram significativamente o bloqueio de chamadas. Além disso, os resultados mostraram que o agente TRPO obteve melhores resultados, com desempenho mais estável nas diferentes topologias de rede testadas. Além disso, a análise dos resultados mostrou que as técnicas de RL podem alcançar bons resultados em termos de probabilidade de bloqueio de chamadas e tornam-se uma solução atrativa para resolver o problema de RBMLSA.

El Sheikh *et al.* [2] compararam uma abordagem heurística com uma abordagem baseada em inteligência artificial, especificamente utilizando o agente DQN, implementado no ambiente do Optical RL-Gym. Ao realizar as simulações, os autores observaram que o agente era consistentemente mais rápido do que a abordagem heurística, porém a heurística apresentava um número menor de bloqueios. Os autores concluíram que, ao desenvolver trabalhos com inteligência artificial, não era necessário apenas focar na velocidade do agente, mas em sua eficiência em termos de bloqueios na rede.

Gonzalez *et al.* [12] projetaram funções de recompensa para melhorar o desempenho e aprendizagem dos agentes, utilizando RL no cenário MB-EON. Foram apresentadas quatro novas funções de recompensa no Optical RL-Gym, em que cada função de recompensa induzia a comportamentos dos agentes de forma distinta ao processar as solicitações de chamadas, levando em conta a disponibilidade de largura de banda, uso, compactação e informações de *feedback*, fornecidas pelo agente. Dentre as recompensas apresentadas, a que mostrou melhor desempenho, levou o agente a escolher a faixa de frequências menos ocupada.

Pandya [3] buscou melhorias na eficiência espectral e na alocação de recursos por meio da implementação de duas etapas: (1) treinar um agente e (2) utilizar esse agente para tornar a alocação de recursos mais eficiente. A primeira etapa consistia em treinar um agente por meio de aprendizagem por reforço, utilizando diferentes solicitações de serviço com requisitos variados de largura de banda, duração e qualidade de serviço. Durante o treinamento, os parâmetros do agente eram ajustados para maximizar a eficiência espectral e a utilização dos recursos da rede. Na segunda etapa, o agente treinado era utilizado para alocar recursos em tempo real para novas solicitações de serviço. O agente realizava a predição da quantidade de recursos necessários para cada solicitação de serviço e alocava recursos de forma dinâmica e eficiente, maximizando a eficiência espectral e a utilização de recursos da rede.

Diferente do que foi realizado por Pandya [3], neste artigo, apresenta-se uma avaliação de hiperparâmetros para o agente *Trust Region Policy Optimization* (TRPO), disponível na biblioteca *Stable-Baselines*, quando operando no ambiente de simulação de redes ópticas multibandas da biblioteca Optical-RL-Gym [2], utilizando a técnica de aprendizagem por reforço (vide Seção IV). A Tabela I apresenta um resumo com as contribuições da literatura que utilizaram a técnica de aprendizagem por reforço no cenário de redes ópticas multibandas.

III. Agentes e Hiperparâmetros

Diante dos artigos apresentados na Seção II, pode-se verificar a utilização de agentes da biblioteca *Stable-baselines* e a realização de treinamentos para criação de modelos a partir da configuração de hiperparâmetros, que são imprescindíveis para determinar a melhor performance de tais agentes. A seguir, encontram-se as definições de cada agente considerado neste artigo e os hiperparâmetros que são configurados para a realização dos respectivos treinamentos.

A. DQN

Baseia-se em um modelo que armazena todas as experiências anteriores, em termos de transições estado-ação, na memória e então reutiliza toda vez que a função Q neural é atualizada, estilo um *buffer* de repetição, uma rede de destino e recorte de gradiente, e faz uso desses diferentes dados para estabilizar o aprendizado com redes neurais. O agente DQN utiliza uma rede neural profunda para aproximação de valores. Como etapa básica no treinamento, o estado inicial do agente é alimentado na rede neural e retorna o valor Q de todas as ações possíveis, como na saída.

B. TRPO

É uma abordagem iterativa para otimizar políticas com melhoria monotônica garantida. Esse agente é comumente utilizado para multiprocessamento. Essa é uma estratégia mais robusta quando comparada, por exemplo, com o DQN, visto que, durante o treinamento, o agente TRPO interage com o ambiente em várias etapas usando a política atual antes de usar minilotes para atualização e as propriedades críticas em várias épocas.

C. PPO2

É implementado a partir de uma combinação de dois outros agentes, o A2C e o TRPO. O PPO2 é uma implementação do OpenAI feita para uma unidade de processamento gráfico, e usa ambientes vetorizados comparado com o PPO1. Diferentemente do agente TRPO, o agente PP02 utiliza várias épocas de atualizações de minilote, podendo aumentar o custo de processamento a depender do experimento a ser realizado.

Para configuração desses agentes, hiperparâmetros são configurados e esse processo pode ser feito por meio de testagem

TABELA I
CONTRIBUIÇÕES DA LITERATURA QUE UTILIZARAM A TÉCNICA DE APRENDIZAGEM POR REFORÇO NO CENÁRIO DE REDES ÓPTICAS MULTIBANDAS.

Referência	Contribuições
[13]	Agentes DQN, TRPO e PPO2 com os melhores resultados.
[1]	Agente TRPO apresenta o melhor resultado.
[2]	Heurística apresenta melhor desempenho com menor número de bloqueios.
[12]	Funções de recompensa são projetadas e avaliadas.
[3]	Ajustes de hiperparâmetros dos agentes garantem melhorias na eficiência espectral e na alocação de recursos.

dos resultados ao longo dos treinamentos ou podem ser utilizados valores padrão (*default*). Porém, ao utilizar os valores *default*, não se pode concluir que os agentes apresentarão os melhores resultados de performance. A Tabela II apresenta os principais hiperparâmetros e seus significados.

Ao visualizar a Tabela II, pode ser verificada a existência de hiperparâmetros distintos entre os 3 agentes (DQN, PPO2 e TRPO), dificultando o processo de configuração e de avaliação comparativa entre eles. Artigos apresentados na Seção II já consideraram esses agentes e concluíram que o agente TRPO possui melhor performance quando comparado com os outros agentes, DQN e PPO2 [1], [3]. Porém, esses mesmos artigos relataram que melhores resultados com o agente TRPO podem ser alcançados com a configuração de seus hiperparâmetros na realização de novos treinamentos.

IV. CONFIGURAÇÃO DOS TREINAMENTOS

Artigos apresentados na Seção II relataram o agente TRPO como o de melhor desempenho, com resultados consistentes em todos os cenários e topologias de redes consideradas [1], [3]. Assim, foi definido esse agente para realização de novos treinamentos por meio do processo de avaliação dos hiperparâmetros com o objetivo de trazer contribuições positivas nos resultados de performance que estão associados à recompensa retornada para o agente ao executar ações no ambiente.

O processo de treinamento envolve o teste de valores dos hiperparâmetros, podendo obter resultados positivos ou negativos a depender de como foram configurados. Em redes EONs, a probabilidade de bloqueio de chamadas é uma métrica utilizada para avaliar o desempenho da rede. No contexto do presente estudo, ao receber uma requisição de chamada entre dois nós, o agente que opera a rede óptica deve ser capaz de direcionar: (1) a rota física por onde a conexão será estabelecida – o caminho na rede por onde os dados irão trafegar e (2) o espectro de frequência por onde os dados irão ser transmitidos nessa rota. Noutras palavras, o agente deve resolver os problemas de (1) roteamento e (2) alocação espectral. Uma chamada pode ser bloqueada se nenhuma rota estiver disponível para estabelecer a conexão ou se, uma vez estabelecida a rota, estejam indisponíveis os espectros de frequência.

De acordo com essa metodologia, o ambiente de simulação envia ao agente que opera a rede o sinal de recompensa. Ou seja, se o agente conseguiu encontrar uma rota disponível e se conseguiu alocar corretamente o espectro de frequência da chamada, o sinal é positivo – representando uma ação correta do agente. Caso contrário, o sinal é negativo – representado uma escolha errada que não deve ser repetida.

É importante observar que, para realização desses treinamentos, foi tomada como base a configuração inicial do artigo [2]. Não foram realizadas modificações no ambiente, considerou-se uma carga de 1000 Erlangs, a seguinte configuração inicial do TRPO:

- MlpPolicy;
- env;
- gamma=$0{,}95$;
- timesteps_per_batch=100000;
- max_kl=$0{,}01$;
- cg_iters=10;
- lam=$0{,}98$;
- entcoeff=$0{,}0$;
- cg_damping=$0{,}01$;
- vf_stepsize=$0{,}0003$;
- vf_iters=3;
- verbose=0;
- tensorboard_log=None;
- _init_setup_model=True;
- policy_kwargs=None;
- full_tensorboard_log=False;
- seed=None;
- n_cpu_tf_sess=1

e a topologia NSFNet com as distâncias em quilômetros, ilustrada na Figura 1 [14].

V. RESULTADOS

Para o treinamento 1, foram obtidos os valores de $39{,}21$ para a média final de recompensa (MFR) e de 17% para a probabilidade de bloqueio de chamadas (PBC), quando os hiperparâmetros gamma, lam e vf_iters receberam os seguintes valores: $0{,}99$; $0{,}97$ e 3, respectivamente. Para o treinamento 2, foram obtidos os valores de $44{,}23$ para a média final de recompensa e de 10% para a probabilidade de bloqueio de chamadas, quando os hiperparâmetros gamma, lam e vf_iters receberam os seguintes valores: $0{,}95$; $0{,}97$ e 3, respectivamente. Para o treinamento 3, foram obtidos os valores de $44{,}58$ para a média final de recompensa e de 9% para a probabilidade de bloqueio de chamadas, quando os hiperparâmetros gamma, lam e vf_iters receberam os seguintes valores: $0{,}94$; $0{,}97$ e 3, respectivamente. Para o treinamento 4, foram obtidos os valores de $43{,}80$ para a média final de recompensa e de 14% para a probabilidade de bloqueio de chamadas, quando os

TABELA II
HIPERPARÂMETROS UTILIZADOS NAS SIMULAÇÕES.

Hiperparâmetro	**Significado**	**Agente**
gamma	Valor de desconto	DQN, TRPO, PPO2
learning_rate	Taxa de aprendizado para o otimizador de adam	DQN, PPO2
buffer_size	Tamanho do *buffer* de *replay*	DQN
exploit_fraction	Fração de todo o período de treinamento durante o qual a taxa de exploração é recozida	DQN
buffer_size	Amostra em lote do *buffer* de repetição para treinamento	DQN
timesteps_per_batch	Número de passos a serem executados por lote	TRPO
max_kl	Limite de perda de Kullback-Leibler	TRPO
cg_iters	Número de iterações para o cálculo do gradiente conjugado	TRPO
lam	fator GAE	TRPO, PPO2
entcoeff	Peso para a perda de entropia	TRPO, PPO2
verbose	Nível de verbosidade	PPO2
nminibatches	Número de *minibatches* de treinamento por atualização	PPO2

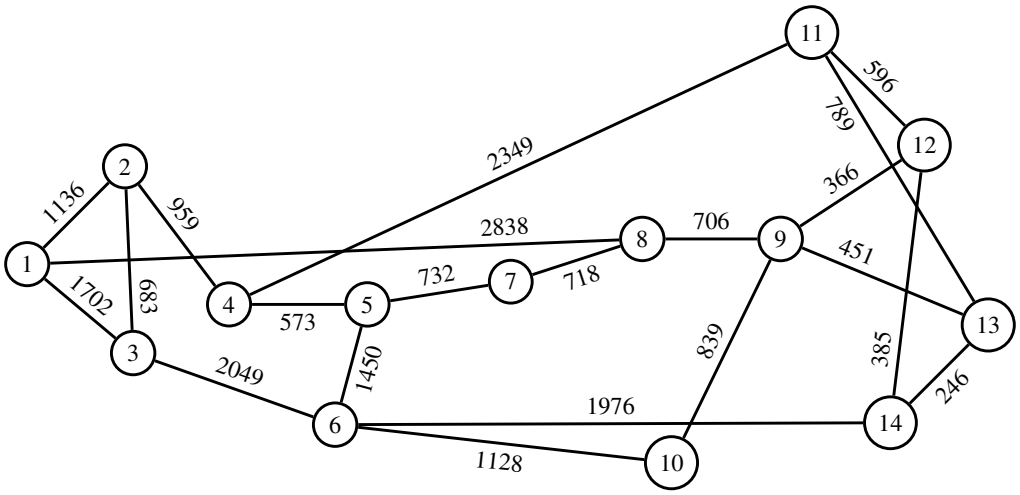

Fig. 1. Topologia NSFNet com as distâncias em quilômetros.

hiperparâmetros gamma, lam e vf_iters receberam os seguintes valores: 0,94; 0,97 e 5, respectivamente.

Ao término da realização dos treinamentos, é possível verificar os ganhos reais em melhorias na média final de recompensa e na probabilidade de bloqueio de chamadas do agente. Comparando-se o treinamento 1 e o treinamento 3, é possível visualizar a diminuição da probabilidade de bloqueio de chamadas de 17% para 9%, como também, aumento na média final de recompensa de 40,1 para 43,8. A Tabela III apresenta as médias finais de recompensa e probabilidade de bloqueio de chamadas obtidas para cada um dos quatro treinamentos.

Fig. 2 mostra a distribuição da recompensa ao longo dos treinamentos. É possível verificar as melhorias na média final de recompensas ao longo dos treinamentos. Por exemplo, para o treinamento 1, verifica-se uma média igual a 37, enquanto que para o treinamento 4 de 40,5.

Fig. 3 ilustra a distribuição da probabilidade de bloqueio de chamadas ao longo dos treinamentos. Percebe-se que, ao longo de cada treinamento, com a configuração dos hiperparâmetros, os resultados se mostram melhores, por exemplo, a média da probabilidade de bloqueio de chamadas obtida no treinamento 1 é de aproximadamente 0,125. Já no treinamento 4, essa média diminui para 0,10.

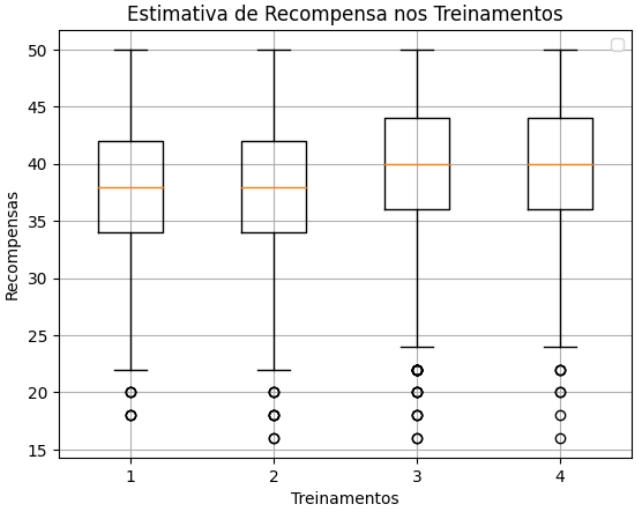

Fig. 2. Distribuição da recompensa ao longo dos treinamentos.

TABELA III
MÉDIAS FINAIS DE RECOMPENSA E PROBABILIDADE DE BLOQUEIO DE CHAMADAS OBTIDAS PARA CADA UM DOS QUATRO TREINAMENTOS.

	Treinamento 1	**Treinamento** 2	**Treinamento** 3	**Treinamento** 4
gamma	0,99	0,95	0,94	0,94
lam	0,97	0,97	0,97	0,97
vf_iters	3	3	3	5
MFR	39,21	44,23	44,58	43,80
PBC	17%	10%	9%	14%

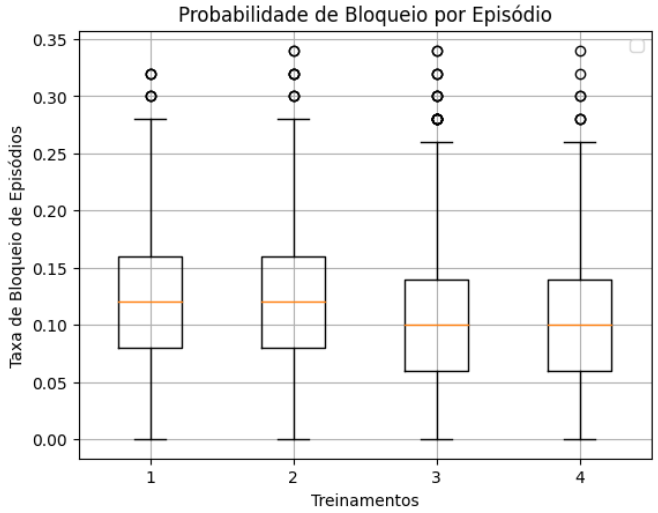

Fig. 3. Distribuição da probabilidade de bloqueio de chamadas ao longo dos treinamentos.

VI. CONCLUSÃO

Este artigo apresentou uma avaliação de hiperparâmetros de um agente de aprendizagem profunda, todos utilizados em problemas de roteamento e alocação espectral em redes ópticas elásticas multibandas.

Nesse contexto, modificações nos hiperparâmetros e novos treinamentos foram realizados de modo que, para o treinamento 3, foi possível verificar que a média final de recompensa foi igual a 44,58, valor superior ao obtido por meio do treinamento inicialmente realizado, com valor de 40,1.

Ainda foi possível observar que a probabilidade de bloqueio de chamadas no treinamento 3 obteve menor valor (9%) quando comparada com a probabilidade de bloqueio de chamadas obtida no treinamento inicial (37%).

Desse modo, conclui-se que a modificação nos hiperparâmetros e realização de novos treinamentos tornou possível obter modelos de agentes TRPO com melhores resultados, em termos de média final de recompensas e de probabilidade de bloqueio de chamadas.

AGRADECIMENTOS

Os autores deste trabalho agradecem à Fundação de Apoio à Pesquisa do Estado da Paraíba (FAPESQ/PB – Termo de Outorga nº 3067/2021) o apoio financeiro e à UFCG o apoio institucional.

REFERÊNCIAS

[1] P. Morales, P. Franco, A. Lozada, N. Jara, F. Calderón, J. Pinto-Ríos, and A. Leiva, "Multi-band environments for optical reinforcement learning gym for resource allocation in elastic optical networks," in *International Conference on Optical Network Design and Modeling (ONDM)*. IEEE, 2021, pp. 1–6.

[2] N. E. D. El Sheikh, E. Paz, J. Pinto, and A. Beghelli, "Multi-band provisioning in dynamic elastic optical networks: a comparative study of a heuristic and a deep reinforcement learning approach," in *International Conference on Optical Network Design and Modeling (ONDM)*. IEEE, 2021, pp. 1–3.

[3] R. Jashvantbhai Pandya, "Machine learning-oriented resource allocation in c+ l+ s bands extended sdm–eons," *IET Communications*, vol. 14, no. 12, pp. 1957–1967, 2020.

[4] B. Tang, J. Chen, Y.-C. Huang, Y. Xue, and W. Zhou, "Optical network routing by deep reinforcement learning and knowledge distillation." Optica Publishing Group, 2021, p. T4A.82. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=ACPC-2021-T4A.82

[5] B. Tang, Y.-C. Huang, Y. Xue, and W. Zhou, "Deep reinforcement learning-based rmsa policy distillation for elastic optical networks," *Mathematics*, vol. 10, no. 18, 2022. [Online]. Available: https://www.mdpi.com/2227-7390/10/18/3293

[6] L. Xu, Y.-C. Huang, Y. Xue, and X. Hu, "Deep reinforcement learning-based routing and spectrum assignment of eons by exploiting gcn and rnn for feature extraction," *Journal of Lightwave Technology*, vol. 40, no. 15, pp. 4945–4955, 2022.

[7] B. Tang, Y.-C. Huang, Y. Xue, and W. Zhou, "Heuristic reward design for deep reinforcement learning-based routing, modulation and spectrum assignment of elastic optical networks," *IEEE Communications Letters*, vol. 26, no. 11, pp. 2675–2679, 2022.

[8] C. Hernández-Chulde, R. Casellas, R. Martínez, R. Vilalta, and R. M. noz, "Evaluation of deep reinforcement learning for restoration in optical networks." Optica Publishing Group, 2022, p. Th2A.19. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=OFC-2022-Th2A.19

[9] L. Xu, Y.-C. Huang, Y. Xue, and X. Hu, "Hierarchical reinforcement learning in multi-domain elastic optical networks to realize joint rmsa," *Journal of Lightwave Technology*, vol. 41, no. 8, pp. 2276–2288, 2023.

[10] X. Guo, F. Yan, X. Xue, B. Pan, G. Exarchakos, and N. Calabretta, "Qos-aware data center network reconfiguration method based on deep reinforcement learning," *Journal of Optical Communications and Networking*, vol. 13, no. 5, pp. 94–107, Maio 2021. [Online]. Available: https://opg.optica.org/jocn/abstract.cfm?URI=jocn-13-5-94

[11] S. Nallaperuma, Z. Gan, J. W. Nevin, M. Shevchenko, and S. J. Savory, "Interpreting multi-objective reinforcement learning for routing and wavelength assignment in optical networks," *Journal of Optical Communications and Networking*, vol. 15, pp. 497–506, 2023.

[12] M. Gonzalez, F. Condon, P. Morales, and N. Jara, "Improving multi-band elastic optical networks performance using behavior induction on deep reinforcement learning," in *IEEE Latin-American Conference on Communications (LATINCOM)*, 2022, pp. 1–6.

[13] C. Natalino and P. Monti, "The optical rl-gym: An open-source toolkit for applying reinforcement learning in optical networks," in *International Conference on Transparent Optical Networks (ICTON)*. IEEE, 2020, pp. 1–5.

[14] A. A. Santos-Júnior, H. A. Pereira, and R. C. A. Almeida-Júnior, "Análise do impacto de penalidades físicas na banda c para uma rede óptica elástica," in *XL Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, vol. 1, 2022, pp. 1–5.