

Deep Reinforcement Learning meets Graph Neural Networks: An optical network routing use case

Paul Almasan

Barcelona Neural Networking Center
Universitat Politècnica de Catalunya

Arnau Badia-Sampera

Barcelona Neural Networking Center
Universitat Politècnica de Catalunya

Pere Barlet-Ros

Barcelona Neural Networking Center
Universitat Politècnica de Catalunya

José Suárez-Varela

Barcelona Neural Networking Center
Universitat Politècnica de Catalunya

Krzysztof Rusek

Barcelona Neural Networking Center, Universitat
Politècnica de Catalunya
AGH University of Science and Technology

Albert Cabellos-Aparicio

Barcelona Neural Networking Center
Universitat Politècnica de Catalunya

ABSTRACT

Recent advances in Deep Reinforcement Learning (DRL) have shown a significant improvement in decision-making problems. The networking community has started to investigate how DRL can provide a new breed of solutions to relevant optimization problems, such as routing. However, most of the state-of-the-art DRL-based networking techniques fail to *generalize*, this means that they can only operate over network topologies seen during training, but not over new topologies. The reason behind this important limitation is that existing DRL networking solutions use standard neural networks (e.g., fully connected), which are unable to learn graph-structured information. In this paper we propose to use Graph Neural Networks (GNN) in combination with DRL. GNN have been recently proposed to model graphs, and our novel DRL+GNN architecture is able to learn, operate and generalize over arbitrary network topologies. To showcase its generalization capabilities, we evaluate it on an Optical Transport Network (OTN) scenario, where the agent needs to allocate traffic demands efficiently. Our results show that our DRL+GNN agent is able to achieve outstanding performance in topologies unseen during training.

1 INTRODUCTION

Recent advances in Deep Reinforcement Learning (DRL) showed a significant improvement in decision-making and automated control problems [15, 19]. In this context, the network community is investigating DRL as a key technology to provide a new breed of network optimization problems (e.g. routing) with the goal of enabling self-driving networks [13]. However, existing DRL-based solutions still fail to *generalize* when applied to network-related scenarios. This hampers

the ability of the DRL agent to make good decisions when facing a network state not seen during training.

Indeed, most of existing DRL proposals for networking can only operate over the same network topology seen during training [5, 11] and thus, cannot generalize and efficiently operate over unseen network topologies. The main reason behind this strong limitation is that computer networks are fundamentally represented as graphs. For instance, the network topology and routing policy are typically represented as such. However, state-of-the-art proposals [5, 12, 21, 22] use traditional neural network (NN) architectures (e.g., fully-connected, Convolutional Neural Networks) that are not well suited to model graph-structured information.

Recently, Graph Neural Network (GNN) [18] have been proposed to model and operate on graphs with the aim of achieving relational reasoning and combinatorial generalization. In other words, GNNs facilitate the learning of relations between entities in a graph and the rules for composing them. GNN have been recently proposed in the field of networking modelling and optimization [17].

In this paper, we present a pioneering DRL+GNN architecture for networking. Our novel architecture is able to operate and optimize problems while generalizing to unseen topologies. Specifically, the GNN used by the DRL agent is inspired by Message-passing Neural Networks [8]. With this framework we design a GNN that captures *meaningfully* the relation between paths and the links on a network topology.

To analyse the generalization capabilities, we experimentally evaluate our proposed DRL+GNN architecture in an Optical Transport Network (OTN) scenario. Specifically, the DRL+GNN agent needs to learn an efficient policy to maximize the number of allocated traffic demands over an OTN

topology. Our results show that our agent is able to efficiently operate over a network not seen during training with outstanding performance.

At the best of our knowledge this is the first time a DRL agent has been combined with a GNN to address networking problems. We hope that our novel architecture represents a baseline to be adapted to other scenarios.

2 BACKGROUND

The solution proposed in this paper combines two machine learning mechanisms. First, we use Graph Neural Networks (GNN) to model computer networks. Second, we use Deep Reinforcement Learning (DRL) to build agents that learn how to efficiently operate networks following a particular optimization goal.

2.1 Graph Neural Networks

Graph Neural Networks (GNNs) is a novel family of neural networks designed to operate over graph-structured information. They were introduced in [18] and numerous variants have been developed since [6, 10, 24]. In its basic form, they consist in associating some initial states to the different elements in a graph and then combine them considering how these elements are related in the graph. An iterative algorithm updates the state elements and uses the final states to produce an output. The particularities of the problem to solve will determine which GNN variant to use, which elements of the graph (edges or nodes) are considered, etc.

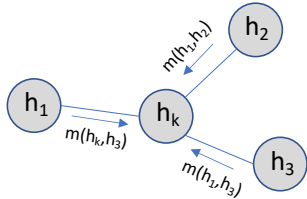


Figure 1: Overview of a single step from the message passing.

Message Passing Neural Networks (MPNN) [8] are a well-known type of GNNs that use an iterative message-passing algorithm to propagate information between the nodes of the graph. In a message-passing step (see Fig. 1), each node k receives messages from all the nodes in its neighbourhood, denoted by $N(k)$. Messages are generated by applying a message function $m(\cdot)$ to the hidden states of node pairs in the graph, and then are combined by an aggregation function, for instance, a sum (Eq. 1). Finally, an update function $u(\cdot)$ is used to compute a new hidden state for every node (Eq. 2).

$$M_k^{t+1} = \sum_{i \in N(k)} m(h_k^t, h_i^t) \quad (1)$$

$$h_k^{t+1} = u(h_k^t, M_k^t) \quad (2)$$

Functions $m(\cdot)$ and $u(\cdot)$ are differentiable functions, and consequently may be learned by neural networks. After a certain number of iterations, the final node states are used by a readout function $r(\cdot)$ to produce an output for the given task. This function can be also implemented by a neural network and is typically tasked to predict properties of individual nodes (e.g., the node's class) or global properties of the graph.

GNNs have been able to achieve relevant performance results in multiple domains where data is typically structured as a graph [2, 8]. Since computer networks are fundamentally represented as graphs, it is inherent in their design that GNNs offer unique advantages for network modelling compared to traditional neural network architectures (e.g., fully connected NN, convolutional NN, etc.). Recent work has shown their potential for network performance prediction [17], even when making predictions on network topologies not seen during training [20].

2.2 Deep Reinforcement Learning

DRL algorithms aims at learning a strategy that leads to maximize the cumulative reward in an optimization problem. DRL agents start from *tabula rasa*. This means that they have no previous expert knowledge about the environment where they operate. They have only a set of possible actions and learn the optimal strategy after an iterative process that explores the action and observation spaces. The learning process consists in a set of actions A and a set of states S . Given a state $s \in S$, the agent will perform an action $a \in A$ that produces a transition to a new state $s' \in S$, and will provide the agent with a reward r . This optimization problem can be modelled as a Markov Decision Process (MDP). However, finding the solution of the MDP requires to evaluate all the possible combinations of state-action pairs.

An alternative to solve the MDP is using Reinforcement Learning (RL). Q-learning [25] is a RL algorithm whose goal is to make an agent learn a policy $\pi : S \rightarrow A$. The algorithm creates a table (a.k.a. q-table) with all the possible combinations of states and actions. At the beginning of the training, the table is initialized (e.g., zeros or random values) and during training the agent updates these values according to the rewards received after selecting an action. These values, called q-values, represent the expected reward assuming the agent is in a state s and performs action a . During training, q-values are updated using the Bellman equation (see Eq. 3) where $r(s, a)$ is the reward obtained from selecting action a in state s , $Q(s', a)$ is the q-value function and $\gamma \in [0, 1]$ is

the discount factor, which represents the importance of the rewards obtained in the future.

$$Q(s, a) = r(s, a) + \gamma \max_a Q(s', a) \quad (3)$$

Deep Q-Network (DQN) [14] is a more advanced algorithm based on Q-learning that uses a deep NN to approximate the q-value function. As the q-table size becomes larger, Q-learning has difficulties to learn a policy from high dimensional state and action spaces. To overcome this problem, they proposed to use a deep NN as q-value function estimator. This method uses an experience replay buffer to store past sequential experiences (i.e. stores tuples of $\{s, a, r, s'\}$). While training the neural network, the temporal correlation is broken by sampling randomly from the experience buffer.

3 NETWORK OPTIMIZATION SCENARIO

Finding the optimal routing configuration for a set of traffic demands is a NP-hard problem [7]. This makes it necessary to explore alternative solutions (e.g., heuristics) with the aim of maximizing the performance at an affordable cost. In this paper, we explore the potential of a GNN-based DRL agent to operate and generalize over routing scenarios involving diverse network topologies. As a first approach, we consider a routing scenario in Optical Transport Networks (OTN). In this scenario, the DRL agent needs to make routing decisions on every traffic demand as it comes.

The DRL agent operates at the level of the electrical domain, over a logical topology where the nodes represent Reconfigurable Optical Add-Drop Multiplexer (ROADM) nodes and edges some predefined lightpaths connecting them (see Fig. 2). Thus, the DRL agent receives a traffic demand defined by the tuple $\{src, dst, bandwidth\}$, and is tasked to route the traffic demand through a particular sequence of lightpaths (i.e., an end-to-end path) that connect the source and the destination of such demand. Traffic demands are considered requests of Optical Data Units (ODU) whose bandwidth demands are defined in the ITU-T Recommendation G.709 [1].

In this scenario, the routing problem is defined as finding the optimal strategy to route incoming source-destination traffic demands with the goal of saving network resources in the long-term. We consider that a demand is properly allocated if there is enough available capacity in all the lightpaths forming the end-to-end path selected. The demands do not expire, occupying the lightpaths until the end of an episode. This implies a challenging task for the DRL agent, since it has not only to identify critical resources on networks (e.g., potential bottlenecks) but also has to deal with the uncertainty in the generation of future traffic demands.

The main reason behind choosing this evaluation scenario is that it is a classical problem for OTN, for which a close-to-optimal heuristic is well known. This will serve as a baseline

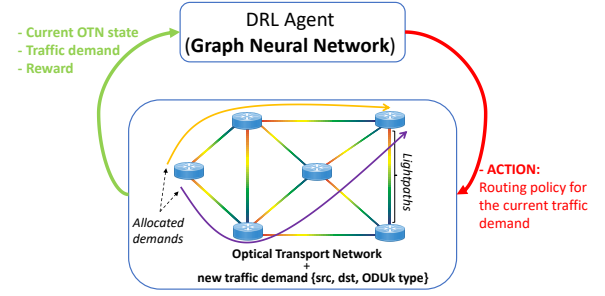


Figure 2: Schematic representation of the DRL agent in the OTN routing scenario.

benchmark to validate the generalization performance of our model.

4 GNN-BASED DRL AGENT DESIGN

In this section we describe the DRL+GNN agent proposed in this paper. Our agent implements the DQN algorithm [14], where the q-value function is modelled by a GNN. Algorithm 1 details the pseudo-code including the training process.

The agent has an external loop that indicates the total number of iterations, and two nested loops that represent evaluation and training periods respectively. The variables $\{s, d, r\}$ represent the network state, the current traffic demand to allocate and the obtained reward respectively. The training loop (lines 2-12) is executed *Training_eps* episodes. For each new demand, the agent selects a discrete action that corresponds to an end-to-end path. After each decision, it receives a reward that is stored into an experience replay buffer along with the state and the action. Then, function *agt.replay()* takes randomly some samples from this buffer to train the GNN model by exploiting the input graph-structured information. The evaluation loop (lines 13-20) is executed during *Evaluation_eps* episodes, where one episode starts with an empty network and finishes when a demand is allocated into a path that do not have enough available capacity for the current traffic demand. Particularly, this occurs when at least one lightpath in the end-to-end path selected does not have enough available capacity. After the evaluation loop is completed, we store the mean reward accumulated over all the evaluation episodes. This is used to represent the evolution of the performance achieved by the DRL agent during the training.

The GNN model is designed following a Message Passing Neural Network [8] fashion. In our case, we consider the link entity and we perform the message passing step between all links. Specifically, for each link l the message function from Eq.1 will take as input the hidden state of the link h_l and the hidden state of a neighbour link h_i where $i \in N(l)$. The same process is repeated for all neighbours of link l and

for all links from the graph. After iterating over all links, the outputs are aggregated using an element-wise sum, resulting on a new feature vector for each link. The resulting vectors are combined with the respective previous link hidden state in a Recurrent NN (RNN). Finally, the outputs of the RNN are aggregated using a sum, passing the result to a fully connected neural network. This will output a single value, indicating the q-value of the allocated demand.

4.1 Environment

To simulate the environment, we implemented the *env.step()* and *env.reset_env()* functions in the Gym framework [4]. The *env.step()* simulates the transition of the network state after a traffic demand is allocated and generates a new traffic demand for the following step. In case there is no enough free capacity in the path selected, the agent is notified with a flag (*done*). The *env.reset_env()* method is executed after every episode end and is responsible for resetting the environment and generate the first traffic request for the next episode. Also, the agent uses the *agt.rmb()* function to store transitions in the experience replay buffer. For this buffer, we implemented a cyclic replay memory (i.e. once the memory is full, oldest experiences are removed) to avoid that the agent is trained on very old samples as the training evolves. The *agt.act()* method selects one over a set of possible actions given an input state following an ϵ -greedy strategy [14] during training.

In order to limit the dimensionality of the action space, we consider that the agent has only flexibility to allocate the current traffic demand over k candidate paths.

5 EXPERIMENTAL RESULTS

In this section we train and evaluate our GNN-based DRL agent to efficiently allocate traffic demands in an OTN routing scenario.

5.1 Evaluation setup

We implemented the DRL environment using the OpenAI Gym framework [4]. For the sake of simplicity, we consider 3 types of traffic demands (ODU2, ODU3, ODU4) whose bandwidth requirements are expressed in terms of multiples of ODU0 signals (8, 32 and 64 ODU0 signals respectively). When the DRL agent allocates a demand, it receives an immediate reward being the bandwidth (in ODU bandwidth units) of the current traffic demand if it was properly allocated, otherwise the reward is 0. Traffic demands are generated on every step by randomly selecting a source-destination pair in the network and an ODU k type demand that represents the bandwidth.

Preliminary experiments were carried to choose an appropriate optimizer and hyperparameter values for our DRL

Algorithm 1 DRL Agent Training algorithm

```

1: for it in Iterations do
2:   for episode in Training_eps do
3:     s, d, src, dst  $\leftarrow$  env.reset_env()
4:     reward  $\leftarrow$  0
5:     while TRUE do
6:       a, s'  $\leftarrow$  agt.act(s, d, src, dst)
7:       r, done, d', src', dst'  $\leftarrow$  env.step(s')
8:       agt.rmb(s, d, src, dst, a, r, s', d', src', dst')
9:       reward  $\leftarrow$  reward + r
10:      If done == TRUE : break
11:      If len(agt.mem) > batch_size : agt.replay()
12:      d  $\leftarrow$  d', s  $\leftarrow$  s', dst  $\leftarrow$  dst'
13:    for episode in Evaluation_eps do
14:      s, d, src, dst  $\leftarrow$  env.reset_env()
15:      reward  $\leftarrow$  0
16:      while TRUE do
17:        a, s'  $\leftarrow$  agt.act(s, d, src, dst)
18:        r, done, d', src', dst'  $\leftarrow$  env.step(s')
19:        reward  $\leftarrow$  reward + r
20:        If done then break

```

agent. To have a stable learning, we decided to change the weights of the GNN model every 2 episodes and using 5 batches to train. The optimizer used is a Stochastic Gradient Descent [3] method with Nesterov Momentum [23]. Regarding the hyperparameters, we use a learning rate of 10^{-4} , and a momentum of 0.9. For the ϵ -greedy exploration strategy, we start with $\epsilon=1.0$ that is maintained during 10 episodes. Then, ϵ decays exponentially every 5 training episodes. For the experience replay buffer, we select a size of 5,000 samples.

5.2 Training and evaluation

We train the DRL agent on an OTN routing scenario with the 14-node NSFNet topology [9], where we consider that the edges represent lightpaths with capacity for 200 ODU0 signals. During training, the agent receives traffic demands and allocates them on one of the $k=4$ shortest paths available in the action set.

We run 400 episodes, store the output in the experience replay buffer and train the GNN by selecting randomly one sample from the buffer. For the evaluation, we run 50 episodes and compute the average cumulative reward obtained over all of them.

In Figure 3 we show the evaluation score of the GNN-based model during training. We also show the evolution of ϵ during the training. As we can observe, when *epsilon* starts to decay (i.e., around episode 10) there is a stable increase in the score achieved by the agent. This suggests that, at this point, the GNN is already in a positive phase of training and

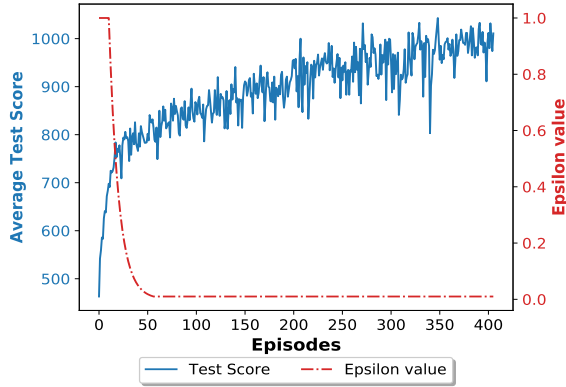


Figure 3: Average of the *Evaluation_eps=50* episodes test score during training.

it is possible to use its q-value estimates to make a smarter exploration of the action space.

5.3 Generalization over other network scenarios

To evaluate the generalization capability of our agent, we select the version of the agent with highest score during the training and evaluate it on a scenario of the 17-node GBN topology [16]. Note that the agent has not seen any sample from this topology during training. In order to benchmark its performance, we compare it against the "Shortest Available Path" (SAP) policy. This routing policy typically represents a performance close to the optimal MDP solution in our OTN routing scenario.

Figure 4 shows the performance of our GNN-based DRL agent on GBN against the SAP heuristic and the RAND policy, which represents a random path selection over all shortest available paths. This policy can be seen as a load balancing policy as the path is chosen uniformly. The *y-axis* represents the score achieved over 50 evaluation episodes (*x-axis*). The horizontal lines indicate the average score obtained over all the episodes by each strategy. This plot reveals the ability of our DRL agent to maintain a good performance even when it operates in a routing scenario with a different topology not seen during training.

6 CONCLUSION

In this paper, we presented a DRL architecture based on GNNs that is able to generalize to unseen network topologies. The use of GNNs to model the network environment allow the DRL agent to operate in different networks than those used for training. We believe that the lack of generalization was the main obstacle preventing the deployment of existing DRL-based solutions in production networks. Thus,

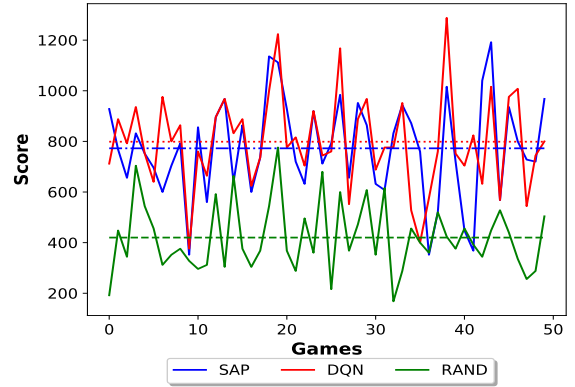


Figure 4: Comparison of the GNN-based DRL agent performance with the SAP and RAND strategies.

the proposed architecture is the first step towards the development of a new generation of DRL-based products for networking.

In order to show the generalization capabilities of our DRL+GNN architecture, we selected a classical problem in the field of OTN, for which a close-to-optimal heuristic is well known. This served as a baseline benchmark to validate the generalization performance of our model. Our results show that our model is able to sustain a similar accuracy in a network never seen during training. Previous DRL solutions based on traditional neural network architectures are not able to generalize to other topologies.

Our ongoing work is focused on applying our DRL+GNN architecture to more complex routing and networking problems. Given the generalization performance shown by GNNs for modelling more complex scenarios, we are confident similar results will also be obtained when combined with DRL.

ACKNOWLEDGMENTS

This work was supported by the Spanish MINECO under contract TEC2017-90034-C2-1-R (ALLIANCE), the Catalan Institution for Research and Advanced Studies (ICREA), by FI-AGAUR grant by the Catalan Government and by the Polish Ministry of Science and Higher Education with the subvention funds of the Faculty of Computer Science, Electronics and Telecommunications of AGH University. The research was also supported in part by PL-Grid Infrastructure.

REFERENCES

- [1] 2019. ITU-T Recommendation G.709/Y.1331: Interface for the optical transport network. <https://www.itu.int/rec/T-REC-G.709/>.
- [2] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. 2016. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*. 4502–4510.

- [3] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.
- [4] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. [arXiv:arXiv:1606.01540](https://arxiv.org/abs/1606.01540)
- [5] Xiaoliang Chen, Jiannan Guo, Zuqing Zhu, Roberto Proietti, Alberto Castro, and SJB Yoo. 2018. Deep-RMSA: A deep-reinforcement-learning routing, modulation and spectrum assignment agent for elastic optical networks. In *2018 Optical Fiber Communications Conference and Exposition (OFC)*. IEEE, 1–3.
- [6] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*. 3844–3852.
- [7] Miriam Di Ianni. 1998. Efficient delay routing. *Theoretical Computer Science* 196, 1-2 (1998), 131–151.
- [8] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1263–1272.
- [9] Xiaojun Hei, Jun Zhang, Brahim Bensaou, and Chi-Chung Cheung. 2004. Wavelength converter placement in least-load-routing-based optical networks using genetic algorithms. *Journal of Optical Networking* 3, 5 (2004), 363–378.
- [10] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. [arXiv preprint arXiv:1511.05493](https://arxiv.org/abs/1511.05493) (2015).
- [11] Shih-Chun Lin, Ian F Akyildiz, Pu Wang, and Min Luo. 2016. QoS-aware adaptive routing in multi-layer hierarchical software defined networks: A reinforcement learning approach. In *2016 IEEE International Conference on Services Computing (SCC)*. IEEE, 25–33.
- [12] Albert Mestres, Eduard Alarcón, Yusheng Ji, and Albert Cabellos-Aparicio. 2018. Understanding the modeling of computer network delays using neural networks. In *Proceedings of the 2018 Workshop on Big Data Analytics and Machine Learning for Data Communication Networks*. ACM, 46–52.
- [13] Albert Mestres, Alberto Rodríguez-Natal, Josep Carner, Pere Barlet-Ros, Eduard Alarcón, Marc Solé, Victor Muntés-Mulero, David Meyer, Sharon Barkai, Mike J Hibbett, et al. 2017. Knowledge-defined networking. *ACM SIGCOMM Computer Communication Review* 47, 3 (2017), 2–10.
- [14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. [arXiv preprint arXiv:1312.5602](https://arxiv.org/abs/1312.5602) (2013).
- [15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.
- [16] J. Pedro, J. Santos, and J. Pires. 2011. Performance evaluation of integrated OTN/DWDM networks with single-stage multiplexing of optical channel data units. In *Proceedings of ICTON*. 1–4. <https://doi.org/10.1109/ICTON.2011.5970940>
- [17] Krzysztof Rusek, José Suárez-Varela, Albert Mestres, Pere Barlet-Ros, and Albert Cabellos-Aparicio. 2019. Unveiling the potential of Graph Neural Networks for network modeling and optimization in SDN. In *Proceedings of the 2019 ACM Symposium on SDN Research*. ACM, 140–151.
- [18] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20, 1 (2008), 61–80.
- [19] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharmarajan Kumaran, Thore Graepel, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. [arXiv preprint arXiv:1712.01815](https://arxiv.org/abs/1712.01815) (2017).
- [20] José Suárez-Varela, Sergi Carol-Bosch, Krzysztof Rusek, Paul Almasan, Marta Arias, Pere Barlet-Ros, and Albert Cabellos-Aparicio. 2019. Challenging the generalization capabilities of Graph Neural Networks for network modeling. In *Proceedings of the ACM SIGCOMM 2019 Conference Posters and Demos*. ACM, 114–115.
- [21] José Suárez-Varela, Albert Mestres, Junlin Yu, Li Kuang, Haoyu Feng, Pere Barlet-Ros, and Albert Cabellos-Aparicio. 2019. Routing based on deep reinforcement learning in optical transport networks. In *Optical Fiber Communication Conference*. Optical Society of America, M2A–6.
- [22] José Suárez-Varela, Albert Mestres, Junlin Yu, Li Kuang, Haoyu Feng, Albert Cabellos-Aparicio, and Pere Barlet-Ros. 2019. Routing in optical transport networks with deep reinforcement learning. *Journal of Optical Communications and Networking* 11, 11 (2019), 547–558.
- [23] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*. 1139–1147.
- [24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. [arXiv preprint arXiv:1710.10903](https://arxiv.org/abs/1710.10903) (2017).
- [25] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3-4 (1992), 279–292.