A Search Heuristic Guided Reinforcement Learning Approach to the Traveling Salesman Problem

Benjamin Hogstad, Jonas Falkner, and Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim
hogstad@uni-hildesheim.de, falkner@ismll.uni-hildesheim.de,
schmidt-thieme@ismll.uni-hildesheim.de

Since 1932, the Traveling Salesman Problem (TSP) has entranced mathematicians. The $\mathcal{NP}$-Hard combinatorial optimization problem has often had heuristics carefully formulated to find good, yet not exact solutions. Approximate solutions to the TSP have become an increasingly important focus of research as scalability of exact solutions limits its real life applications. Recently, Reinforcement Learning (RL) approaches to finding approximate solutions to combinatorial optimization problems have become increasingly promising. For the TSP, encoder-decoder neural network architectures are common within RL approaches. [4] implemented a Q-Learning approach using graph embedding. [1], [5], and [2] each implemented policy gradient methods using LSTM ([1], and Transformer network architectures ([5] and [2]). [2] enhanced their approach with a 2-opt procedure [6] to further improve their network's suggested tour. With the exception of [2], the majority of research has not combined existing expert knowledge, with respect to heuristics, to that of RL. [3] applied a joint supervised and reinforcement learning approach to Q learning; though this application was only evaluated on the Atari environments. By utilizing human experts in certain situations, the agent was able to learn to mimic the expert's behavior and eventually surpass it through exploration.

In this work, we evaluate three main approaches: **Supervised**, **Guided**, and **Combined**. In the **Supervised** approach, we apply the principles outlined by [3] to the TSP. The human experts have been replaced with simple heuristics (Nearest Neighbor, Sweep, and 2-opt[6]). Note that this approach introduces three new hyperparameters: λ_{nn}, λ_{sw}, and λ_{2opt} to the traditional Q-learning loss function and apply an importance factor to the respective heuristic experts and supervised loss.

The **Guided** approach utilizes the traditional Q-learning loss function and an enhanced ϵ-*greedy* exploration strategy. When an exploration action would be selected, it is instead selected from one of the following: *Random*, *Nearest Neighbor*, *K Nearest Neighbor*, *Sweep*, or *K Sweep*. Random actions behave as in traditional ϵ-*greedy*. *Nearest Neighbor* (Figure 1) and *Sweep* (Figure 2) actions are those selected by the respective heuristics. *K Nearest Neighbor* and *K Sweep* take K top available nodes according to each heuristic and then from the subset of nodes, a greedy action is taken. Note that K is a new hyperparameter.

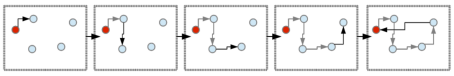

Fig. 1. An example of the Nearest Neighbor construction heuristic for five nodes.

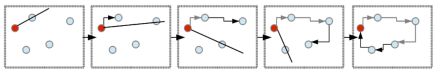

Fig. 2. An example of the Sweep construction heuristic for five nodes.

The **Combined** approach utilizes a hybrid of guided supervision. The agent's exploration is guided and penalized for non heuristic approved actions.

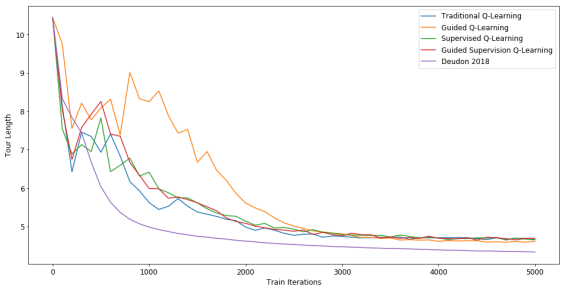

Fig. 3. 20 Node Euclidean TSP preliminary results.

The Neural Network structure is similar to that of [2] with a few minor changes. The context query is a combination of the depot node (the first node in the tour), the current location, and all visited nodes. Furthermore, an additional attention mechanism is implemented in the decoder.

The three approaches were evaluated on a two-dimensional Euclidean TSP set in the unit square and compared against [2] and a traditional Q-Learning model. Figure 3 shows that although the **Guided** approach does not perform well initially it slightly leads to better results than the other Q-Learning models. However, none of the Q-Learning models surpassed [2].

References

1. Bello, I., Pham, H., Le, Q.V., Norouzi, M., Benjgio, S.: Neural combinatorial optimization with reinforcement learning. In: arXiv preprint arXiv: 1611.09940 (2016)
2. Deudon, M., Cournut, P., Lacoste, A., Adulyasak, Y., Rousseau, L.M.: Learning heuristics for the tsp by policy gradient. In: International conference on the integration of constraint programming, artificial intelligence, and operations research, pp. 170?181. Springer, Cham (2018).
3. Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I. Dulac-Arnold, G.: Deep q-learning from demonstrations. In: Thirty-Second AAAI Conference on Artificial Intelligence. (2018)
4. Khalil E., Dai, H., Zhang, Y., Dilkina, B., Song, L.: Learning combinatorial optimization algorithms over graphs. In: Advances in Neural Information Processing Systems pp. 6348?6358. (2017)
5. Kool, W., Hoof, H.V., Welling, M.: Attention solves your TSP, approximately. In: Statistics 1050, p. 22. (2018)
6. Toth, P., Vigo, D.: The vehicle routing problem. Society for Industrial and Applied Mathematics (2002)