Optimal transport for deep neural networks pruning

A PostDoc call

Responsible: Enzo Tartaglione

enzo.tartaglione@telecom-paris.fr

Optimal transport, also known as the transportation theory or the Wasserstein metric, is a mathematical framework that addresses the problem of finding the most efficient way to transport mass or resources from one distribution to another, while minimizing a certain cost function [1,2,3]. Initially developed in the 18th century as a logistics and economics tool, optimal transport has gained significant attention in modern mathematics and various scientific disciplines, including computer science and machine learning. At its core, optimal transport seeks to quantify the similarity between two probability distributions by finding the optimal way to redistribute the mass of one distribution to match the other, taking into account the cost of moving mass from one location to another. This elegant and versatile concept has found applications in diverse fields, from image processing and data analysis to economics [11] and neuroscience, making it a powerful and unifying mathematical tool with wide-ranging implications [12].

Optimal transport theory plays indeed a pivotal role in the field of deep learning by providing a powerful mathematical framework for solving complex data alignment and transportation problems [4, 5]. At its core, deep learning aims to extract meaningful representations from data, and optimal transport theory offers a principled way to measure and manipulate the relationships between data points, enabling the alignment of distributions in high-dimensional spaces. This alignment is crucial for tasks such as domain adaptation, style transfer, and generative modeling, where preserving the underlying data structure is essential. Moreover, optimal transport has found applications in training generative adversarial networks (GANs), where it guides the generator to produce realistic samples by minimizing the transportation cost between the generated and real data distributions [5]. Thus, optimal transport theory not only enriches the mathematical foundations of deep learning but also empowers the development of more effective and interpretable deep learning models [6, 10].

Optimal transport theory has found a novel application in the domain of deep neural network pruning, offering a principled approach to efficiently reduce the size of neural networks while preserving their functionality [7,8,9]. When applied to pruning, optimal transport identifies the most informative neurons/channels/connections within a neural network while minimizing the

impact on its performance. By considering the transportation cost between the full network and a pruned version, optimal transport helps strike a balance between model compression and information retention. This approach leads to more structured and interpretable pruning strategies, ensuring that the critical features and connections are retained, ultimately yielding smaller, more efficient neural networks without significant loss in predictive power. Thus, optimal transport theory contributes to the development of streamlined and resource-efficient deep learning models, addressing the ever-growing demand for computationally lightweight neural networks in various real-world applications.

Objectives. The application of optimal transport theory to deep learning pruning faces certain limitations. Firstly, computational complexity remais a significant challenge, particularly for large-scale neural networks. The optimization problems involved in finding the optimal transportation plan can be computationally demanding, making it challenging to apply to extensive models in practice. Additionally, the scalability of optimal transport methods to handle high-dimensional data, such as images or large-scale text datasets, poses an obstacle. However, the potential application to reduce the depth of the deployed deep neural models [13] can constitute a relevant trade-off between optimization complexity VS gain in terms of inference complexity, which is the main objective to be explored in this project.

References

[1] G. Monge. *Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pages 666–704, 1781.
[2] Ablordeppey, Veronica. *The Transportation Problem of a Beverage Industry: A Case Study of Accra Brewery Limited.(ABL)*. Diss. 2012.
[3] Grattan-Guinness, Ivor. *Companion encyclopedia of the history and philosophy of the mathematical sciences*. Routledge, 2002.
[4] Haker, Steven, et al. "Optimal mass transport for registration and warping." *International Journal of computer vision* 60 (2004): 225-240.
[5] Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks." *International conference on machine learning*. PMLR, 2017.
[6] Montesuma, Eduardo Fernandes, Fred Ngole Mboula, and Antoine Souloumiac. "Recent advances in optimal transport for machine learning." *arXiv preprint arXiv:2306.16156* (2023).
[7] Shen, Yucong, et al. "CPOT: CHANNEL PRUNING VIA OPTIMAL TRANSPORT." (2020).
[8] Li, Yunqiang, et al. "Differentiable Transportation Pruning." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
[9] Backurs, Arturs, et al. "Scalable nearest neighbor search for optimal transport." *International Conference on machine learning*. PMLR, 2020.
[10] Flamary, R., et al. "Optimal transport for domain adaptation." *IEEE Trans. Pattern Anal. Mach. Intell* 1 (2016): 1-40.
[11] Peyré, Gabriel, and Marco Cuturi. "Computational optimal transport." *Center for Research in Economics and Statistics Working Papers* 2017-86 (2017).
[12] Kolouri, Soheil, Yang Zou, and Gustavo K. Rohde. "Sliced Wasserstein kernels for probability distributions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
[13] Zhu, Liao, et al. "Can Unstructured Pruning Reduce the Depth in Deep Neural Networks?." ICCV Workshop (2023)