Dataset Transferability Analysis via Optimal Transport

Davi Sales Barreira

June 8, 2021

1 Introduction

Optimal Transport (OT) theory is a field of mathematics that studies the problem of optimally transporting quantities from one configuration to another given a cost function. The origin of the field is commonly attributed to the french mathematicians Gaspard Monge (1746-1818) whose original motivating problem was "what is the optimal way to transport soil extracted from one location and move to another where it will be used, for example, on a construction?" (see Figure 1).

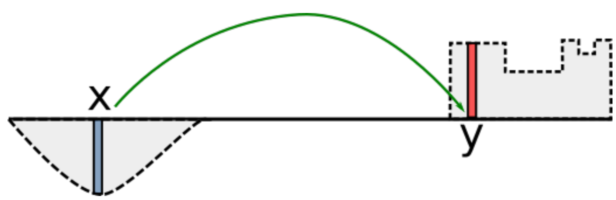

Figure 1: Illustration of the original Monge Problem, where all the mass is excavated from location x is transported to a deterministic location y. The transport assignment map is represented by the arrow in green.

This seemingly narrow subject has actually many applications beyond what one may see at first. In the field of Machine Learning, Optimal Transport theory has been gaining attention, specially in subareas such as transfer learning ([Flamary et al., 2014], [Courty et al., 2014], [Damodaran et al., 2018], [Solomon et al., 2014], [Shen et al., 2018]). One of the main ways in which OT is used in Machine Learning is in order to define a distance metrics between probability distributions. Note that two datasets may be interpreted as empirical distributions in high-dimensions, and one can use Optimal Transport in order to obtain the minimal cost of transporting one dataset distribution into the other. This cost can be thought of as a distance measure between datasets.

This project is based on the work of Alvarez-Melis and Fusi [2020], where the authors proposed one way of measuring the distance between datasets using Optimal Transport. Their metric, which was called Optimal Transport Dataset Distance (OTDD), was shown to be correlated to performance in terms of transfer learning, i.e. the lower the OTDD, the better was the transfer learning between two datasets. Hence, the OTDD could be used as a parameter to evaluate how well the transfer learning would be between two datasets.

Suppose that you have many datasets which you can train to then use a transfer learning method in order to make prediction in another dataset. Hence, Alvarez-Melis and Fusi [2020] proposed the use of OTDD as a metric in order to evaluate which dataset would be best suited.

In this project, we make use of the OTDD metric to evaluate the transferability between two datasets, but instead of comparing many datasets, we develop a tool that allows users to explore the differences between the datasets, and perform data augmentations in order to improve the transferability between the datasets, i.e. reduce the OTDD distance.

2 Datasets

We utilize the MNIST and FashionMNIST (FMNIST) datasets, which are benchmark datasets for Machine Learning. The MNIST dataset is composed of handwritten digits from zero to nine, each picture is in gray-scale and consists in 28 by 28 pixels. The FMNIST consists of small photos of clothes divided in 10 categories (e.g. shoes, dress, shirt, etc) Each image is also in gray-scale and 28 by 28 pixels. Figure 2 presents some samples from the datasets.

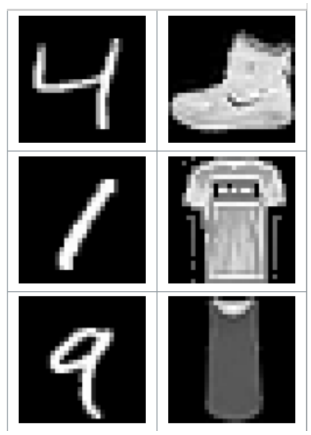

Figure 2: In the left, samples from MNIST. In the right, samples from FMNIST.

References

David Alvarez-Melis and Nicolò Fusi. Geometric dataset distances via optimal transport. *arXiv preprint arXiv:2002.02923*, 2020.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

Rémi Flamary, Nicolas Courty, Devis Tuia, and Alain Rakotomamonjy. Optimal transport with laplacian regularization: Applications to domain adaptation and shape matching. In *NIPS Workshop on Optimal Transport and Machine Learning OTML*, 2014.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Wasserstein propagation for semi-supervised learning. In *International Conference on Machine Learning*, pages 306–314. PMLR, 2014.