Dataset Transferability Analysis via Optimal Transport

Davi Sales Barreira

June 9, 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).

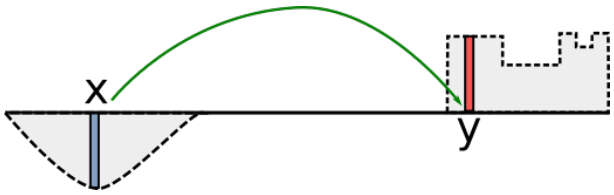

Figure 1: Illustration of the original Monge Problem, where all the mass is excavated from location x is transported to a deterministic location y. The transport assignment map is represented by the arrow in green.

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.

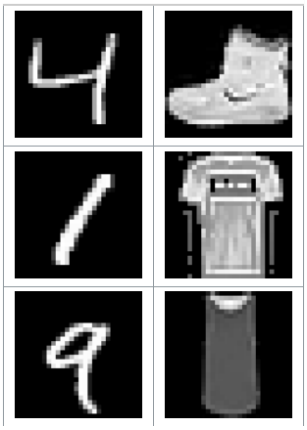

Figure 2: In the left, samples from MNIST. In the right, samples from FMNIST.

3 Technology Used

This project used the Julia programming language to perform the Optimal Transport calculations and all the datasets transformations. Also, instead of Observable, we've utilized a tool called "Pluto", which is a notebook in Julia very similar to Observable. The visualization tool was built inside the Pluto notebook which enabled the creations of buttons and interactive calculations. Similar to Observable, the Pluto notebook does not access directly the data in the computer, so in order to plot the images from the datasets, we also shipped a small code in Julia to create a server to serve the data.

Inside Pluto notebook, we wrote javascript code with instruction to create the visualization graphs with D3 and VegaLite.

4 Visualization Tool

The visualization tool developed in this project has three main components. The first one, shown in Figure 3, is composed of two views. In the view in the left, we use the UMAP dimensionality reduction in order to visualize the datasets in two dimensions.The edges in blue represent the optimal map for transporting the empirical distribution of one dataset into another. Note that the OTDD is related to such edges, since it measures the Optimal Transport distance between the datasets. In this view, the user can use a brush to do a first selection of the datasets which he wants to perform the data augmentation by pressing the "Select Initial Samples" button. Also, note that there is a small selector in the top left, where the user may change the marker type from "Image" to "Circle", which will more clearly show which sample belongs to which dataset.

The view on the right contains a heatmap with the dataset labels in each axis. This heatmap contains the number of samples that are transported in terms of their labels. Qualitatively, when two datasets have good transferability, it is expected that such heatmap matrix will be sparse. The reason for this is that , in the ideal scenario for transfer learning, all samples with the same label would all be "transported" to samples in the second dataset also all with the same label (e.g. all 1's in MNIST go to "dresses" in the FMNIST). Therefore, the main reason for this view on the right is to help the user understand which labels have more space for improvement.

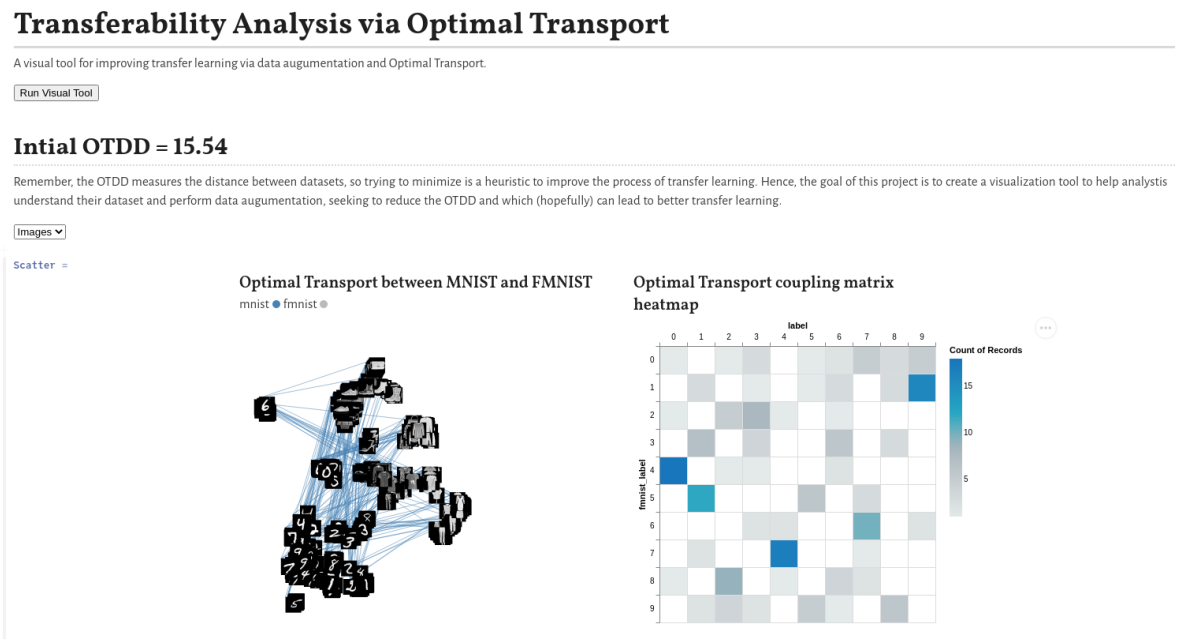

Figure 3: In the left, samples from MNIST. In the right, samples from FMNIST.

The second component is a rescaled view of the data samples selected in the first component. These samples appear as soon as the user presses the "Select Initial Sample" button. In here, the user can do a more thorough visual inspection of the samples. Here there is also a brush for the user to perform another selection. Then, by pressing the "Final Picks" button, the visual tool selects the samples inside the brush and sends them to the third component.

Figure 4: Example of user selecting some samples in the first view, and visualizing them after pressing the "Sample Initial Samples" buttons.

The third component is where the user will apply the modifications to the samples, trying to reduce the OTDD and thus, improving the transferability between the datasets. This component has a "Default" button, two sliders and two checkboxes. The first slider performs an equalization of the data (i.e. it makes images thinner or thicker depending on the value of gamma applied). The second slider performs rotations. The checkers perform either vertical or horizontal mirroring. Finally, the "Default" button restores the original images.

As the user applies this transformations to the data, there is a view in the bottom showing how the final samples will look. The OTDD recalculated as the values in the handles are modified, so the user can visualize in real time how the modifications are affecting the transferability.

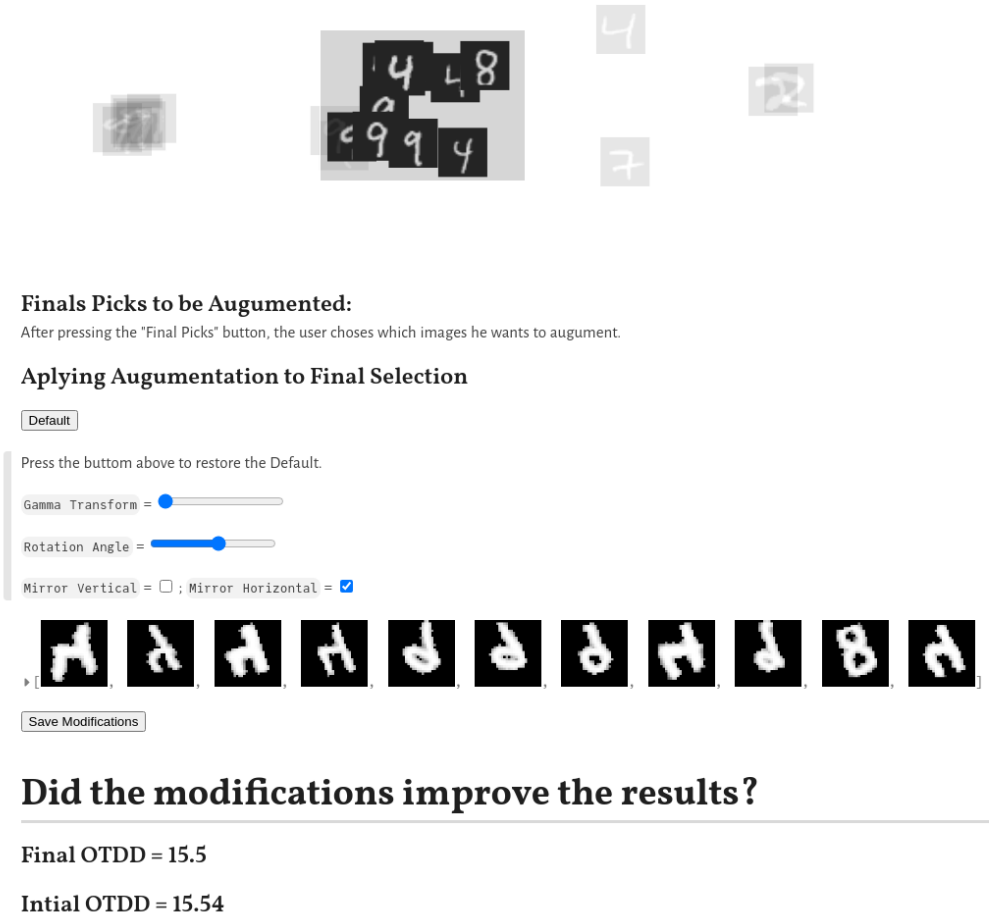

Figure 5: Third component of the visualization tool.

5 Installation and Troubleshooting

In this section, we give the instructions on how to run the visualization tool developed in this project, and how to solve possible issues when trying to run it for the first time.

As we've stated, this project was developed in the Julia programming language, hence, the user must have Julia installed.

The visualizations use a "cdn" to import the javascript packages (i.e. d3 and VegaLite), therefore, it requires internet access to properly work.

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.