# Proposal - Davi Sales Barreira

May 11, 2021

## 1 Optimal Transport for Dataset Distances to Aid Transfer Learning

### 1.1 Description and Background Motivation

This project is mainly based on the paper "Geometric Dataset Distances via Optimal Transport" by Alvarez and Fusi (2020). In this article, the authors proposed a way of measuring the distance between datasets and showed that this distance was correlated to performance in terms of transfer learning. This distance, which used Optimal Transport, could be used as a parameter to evaluate how well a model trained on a dataset could be used in another dataset. The farther the dataset were, the worst in terms of transfer learning the model would be.

Besides the work of Alvarez and Fusi (2020), many other works have been developed using Optimal Transport (OT) for Transfer Learning. Hence, the idea is to create a set of visualizations exploring how the datasets compare in terms of OT metrics, thus, aiding modelers when deciding on how to perform transfer learning between datasets. These visualizations should, for example, help modelers identify possible types of data augmentation that could perhaps improve the transfer of knowledge (e.g. should one normalize the dataset, should one add cropping, etc).

The authors of this papers also wrote a blog post [https://www.microsoft.com/en-us/research/blog/measuring-dataset-similarity-using-optimal-transport/] which explored this idea. The goal is to sistemitize some of the visualizations produced in this blog post, and increment with some other OT metrics and new visualizations. Therefore, by the end of this project, a set of standard visualizations shall be proposed in order to evaluate datasets in terms of their transfer capabilities.

### 1.2 Tasks

The initial task is to calculate the distance between the datasets using the different OT metrics. While Alverez and Fusi (2020) focused only on their new Geomtric Dataset Distance, we would use other metrics which are also proposed in the literature to perform Transfer Learning. This would give more robustness in terms of transfer capabilities.

Once this is done, we want to somehow visualize the data in a lower dimension, using PCA or other method. This would allow use to see how the dataset is been transported into the other. Next, we'd like to visualize a heatmap in terms of the OT plans, in order to understand how categorical data is been transported. This helps us understand which categories might be ambiguous between datasets, and could be perhaps removed all together.

Next, we would propose some standard data augmentations and visualize it's impact in terms of the OT plan.

Finally, we would implement some interactive components, such as showing the actual image when the user hover over the datapoint.

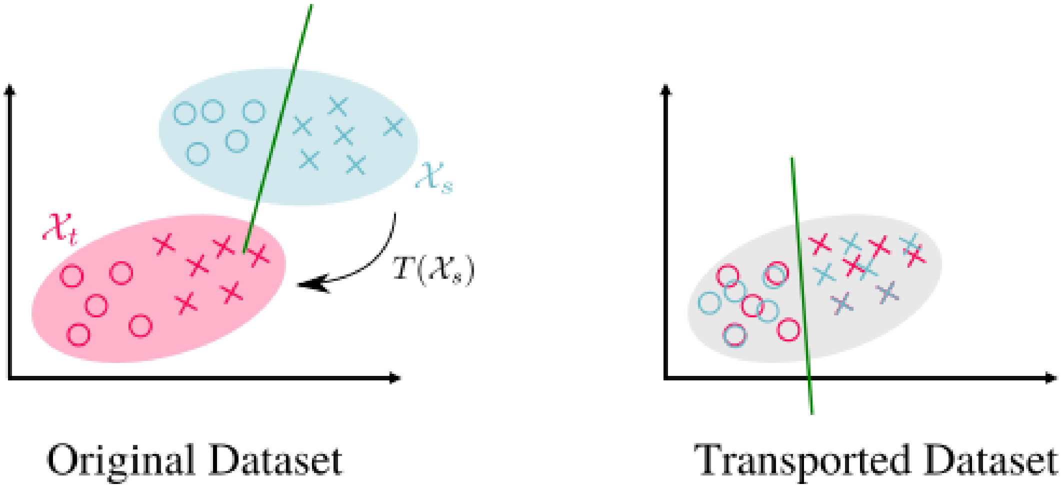
### 1.3 Data

In terms of dataset, we'll be using standard Machine Learning datasets focusing on image classification, such as MNIST, FashionMNIST, USPS, etc. Hence, the collection and cleaning is not necessary. The only transformation necessary will be in terms of augmenting the datasets, performing image manipulation such as centering, cropping, etc.

### 1.4 Models

The visualizations are proposed to aid Transfer Learning tasks, but it actually focuses on preprocessing. Optimal Transport metrics help evaluate the tranfer capability of the dataset, before actually performing the Machine Learning tranining and tranfering. The models that will be implemeted are actually the Optimal Transport algorithms for calculating the distances between the datasets and the OT plans.

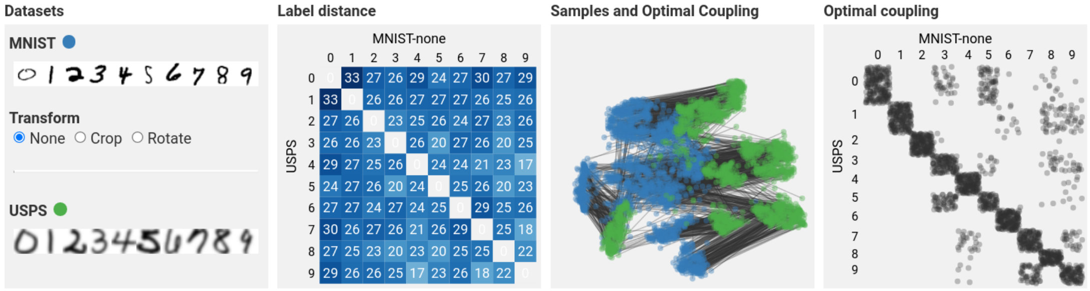
Below we show an schematic image of how OT is used for Transfer Learning.



## 1.5 Design

The visualization that will be produced is similar to the one below. The datasets will be visualized in 2D using t-SNE (or PCA). Once this is done, the Optimal Transport plan will be obtained and the distance between the datasets will be produced. The lines in black represent the Optimal Transport plan. This Optimal Transport plan will also be shown in the matrix format (the utmost right graph in the figure below).

The visualization on the write shows the different ways which transforming the data affects the Optimal Transport distance between the datasets and thus, affects the transfer capability. I intend to provide more data transformation options, and other metrics besides the one introduced by Alvarez and Fusi (2020).



Datasets

MNIST ●

Transform

● None ○ Crop ○ Rotate

USPS ●

Label distance

MNIST-none

| USPS | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 33 | 27 | 26 | 29 | 24 | 27 | 30 | 27 | 29 |
| 1 | 33 | 0 | 26 | 26 | 27 | 27 | 27 | 26 | 25 | 26 |
| 2 | 27 | 26 | 0 | 23 | 25 | 26 | 24 | 27 | 23 | 26 |
| 3 | 26 | 26 | 23 | 0 | 26 | 20 | 27 | 26 | 20 | 25 |
| 4 | 29 | 27 | 25 | 26 | 0 | 24 | 24 | 21 | 23 | 17 |
| 5 | 24 | 27 | 26 | 20 | 24 | 0 | 25 | 26 | 20 | 23 |
| 6 | 27 | 27 | 24 | 27 | 24 | 25 | 0 | 29 | 25 | 26 |
| 7 | 30 | 26 | 27 | 26 | 21 | 26 | 29 | 0 | 25 | 18 |
| 8 | 27 | 25 | 23 | 20 | 23 | 20 | 25 | 25 | 0 | 22 |
| 9 | 29 | 26 | 26 | 25 | 17 | 23 | 26 | 18 | 22 | 0 |

Samples and Optimal Coupling

Optimal coupling