MIT-IBM Watson AI Lab

# Log-Euclidean Signatures for Intrinsic Distances Between Unaligned Datasets

*Tal Shnitzer, **Mikhail Yurochkin, **Kristjan Greenewald, *Justin Solomon

*GDP, CSAIL, MIT **MIT-IBM Watson AI Lab


## Introduction

- Comparing high-dimensional point clouds with no adequate models is challenging.
- The manifold hypothesis motivates comparing such datasets through the geometric properties of their underlying manifold.
- The Laplace-Beltrami operator and heat kernel provide geometrically-motivated data representations in various manifold learning techniques. However, comparing these representations typically requires costly pointwise alignment.
- One approach that avoids these costs is comparing spectral properties of these operators instead, e.g. the heat trace.
- We propose a new spectral method for comparing ***unaligned datasets***, derived by taking into account the symmetric positive-definite (SPD) structure of heat-kernels.

## Representing High-dimensional point clouds

### Representing data using diffusion maps operators[1]

- A manifold learning technique, approximating the heat kernel of the underlying manifold, $e^{-t\Delta}$, for some dataset $X$:

$$K_\epsilon(i,j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\epsilon}\right), \qquad \widehat{K}_\epsilon(i,j) = \frac{K_\epsilon(i,j)}{\sum_\ell K_\epsilon(i,\ell)\cdot\sum_\ell K_\epsilon(\ell,j)}$$

$$W(i,j) = \frac{\widehat{K}_\epsilon(i,j)}{\sum_s \widehat{K}_\epsilon(s,j)\sum_\ell \widehat{K}_\epsilon(i,\ell)}, \qquad \lim_{N\to\infty}\lim_{\epsilon\to 0} W^{t/\epsilon} = e^{-t\Delta}$$

### Riemannian manifold of SPD matrices

- $W$ is SPD. The space of SPD matrices forms a Riemannian manifold, when endowed with a proper metric.
- The log-Euclidean (LE) metric is one suitable choice, providing several computational and algorithmic advantages in our setting.

$$d_{LE}(W_1, W_2) = \|\log(W_1) - \log(W_2)\|_F$$

## Distance for Unaligned Datasets

- Most metrics comparing $W_\ell$ of different datasets, including the LE metric, require full pointwise alignment.
- Our distance is defined by lower-bounding, regularizing and truncating the LE metric, overcoming the alignment need:

$$\|\log W_1 - \log W_2\|_F^2 \geq \sum_{i=1}^{N}\left(\log\lambda_i^{(1)} - \log\lambda_i^{(2)}\right)^2$$

$$\boldsymbol{d_{LES}^2(W_1, W_2) = \sum_{i=1}^{K}\left(\log\left(\lambda_i^{(1)}+\gamma\right) - \log\left(\lambda_i^{(2)}+\gamma\right)\right)^2}$$


## Eigenvalue Approximation

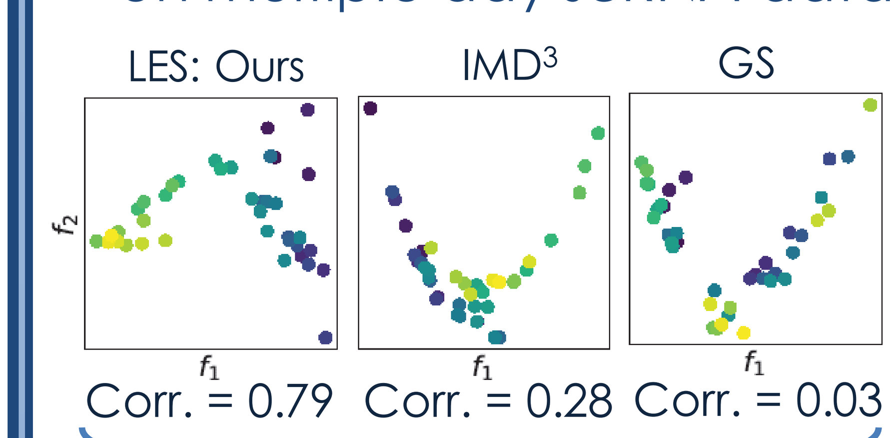
- We estimate the leading eigenvalues with a modified Nyström method[2], resulting in clear error bounds:

$$\mathbb{E}\left[\sum_{i=1}^{K}\left|\log(\lambda_i+\gamma) - \log(\hat{\lambda}_i+\gamma)\right|\right] < \frac{1.5K}{(M-K-1)(\lambda_K+\gamma)}\sum_{i=K+1}^{N}\lambda_i$$

- The regularization term, $\gamma$, facilitates better bounds.

## Gene Expression Analysis

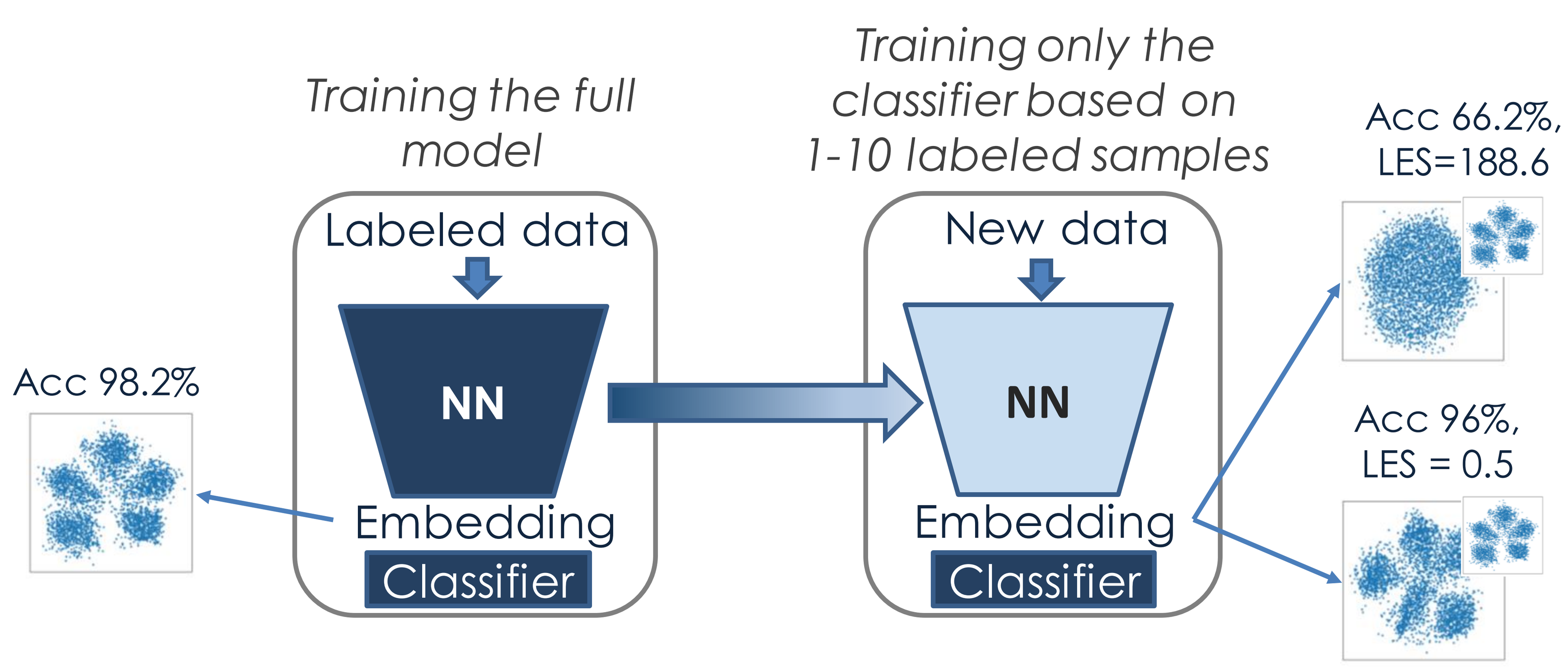
- Recovering time trajectory of cell differentiation based on multiple day scRNA data




Data visualization:


## Analysis of Neural Network Embeddings

### Predicting Success of NN Embedding in Few-Shot Learning



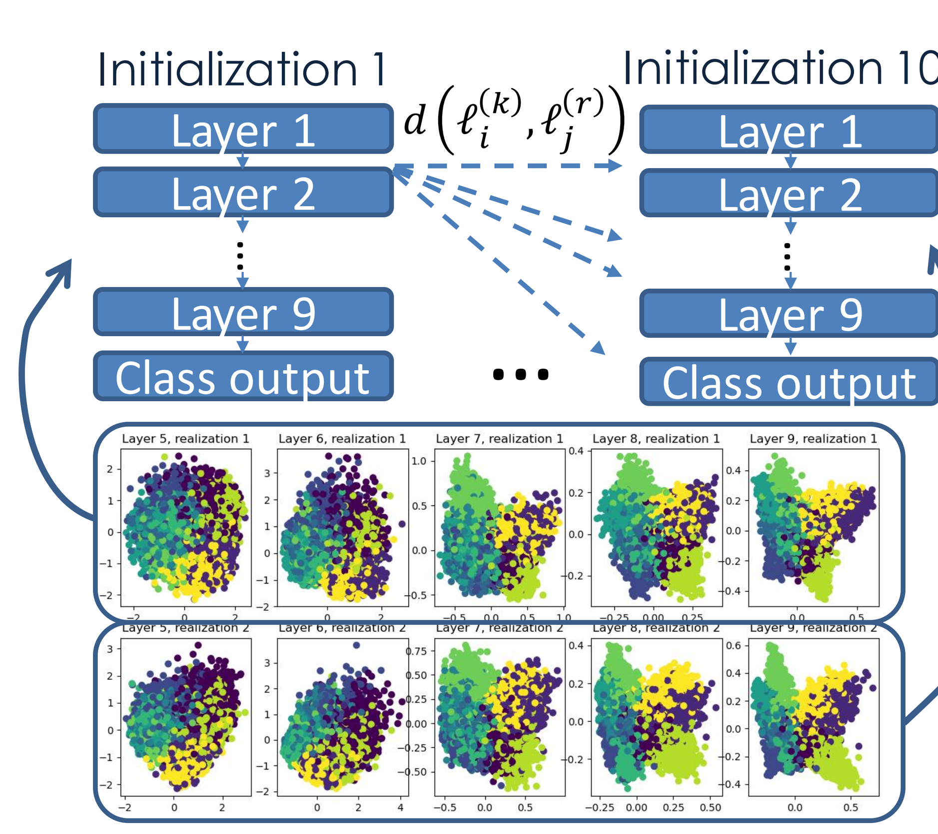

Correlations between classification accuracy and LES distance of embeddings:

| | FC100 1-shot | FC100 5-shot | FC100 10-shot | CIFAR-FS 1-shot | CIFAR-FS 5-shot | CIFAR-FS 10-shot |
|---|---|---|---|---|---|---|
| Acc. | 38.19±0.48% | 54.45±0.49% | 60.52±0.48% | 70.79±0.69% | 83.98±0.44% | 87.11±0.40% |
| LES | **-0.934**±0.034 | **-0.945**±0.032 | **-0.935**±0.032 | -0.671±0.128 | **-0.698**±0.155 | **-0.657**±0.175 |
| IMD | -0.184±0.250 | -0.303±0.225 | -0.210±0.235 | -0.151±0.225 | -0.015±0.220 | -0.032±0.281 |
| OT | 0.739±0.102 | 0.605±0.126 | 0.579±0.133 | 0.453±0.170 | 0.269±0.214 | 0.180±0.231 |
| GW | -0.919±0.034 | -0.921±0.046 | -0.914±0.045 | **-0.677**±0.126 | -0.672±0.167 | -0.582±0.171 |

### Comparing NN Layer Embeddings

Layer position classification accuracy:

| | LES | IMD[3] | CKA[4] |
|---|---|---|---|
| Same input | 96.5% | 85.2% | 97.3% |
| Different input | 95.8% | 81% | - |




1 R.R. Coifman and S. Lafon, "*Diffusion Maps*", ACHA, 2006.
2 J.A. Tropp et al. "Fixed rank approximation of a positive-semidefinite matrix from streaming data", NeurIPS, 2017
3 A. Tsitsulin et al. "The shape of data: Intrinsic distance for data distributions", ICLR, 2020
4 S. Kornblith et al. "Similarity of neural network representations revisited", ICML, PMLR, 2019

**Accepted to ICML 2022**
This work is part of the Model Fusion with Optimal Transport MIT-IBM project