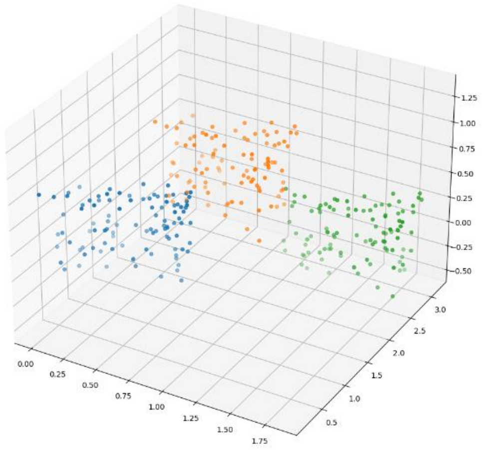
Evaluating the impact of data representation on t-SNE projections

Liviu-Ștefan Neacșu-Miclea

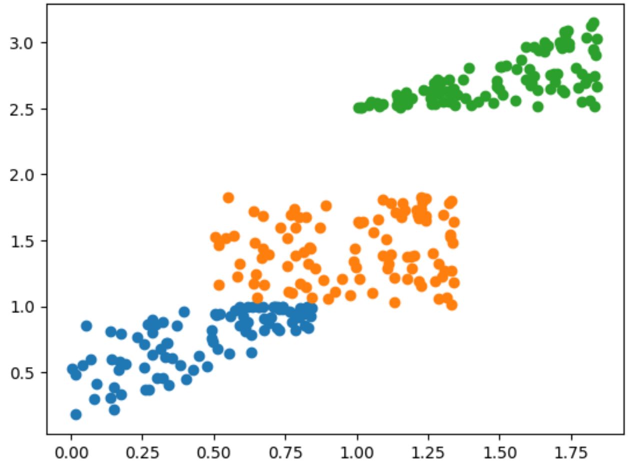
t-SNE

- t-Distributed Stochastic Neighbor Embedding
  - Statistical visualization tool
  - Projects data to lower dimensional spaces
- Tackles the crowding problem of previous SNE methods



$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}$$

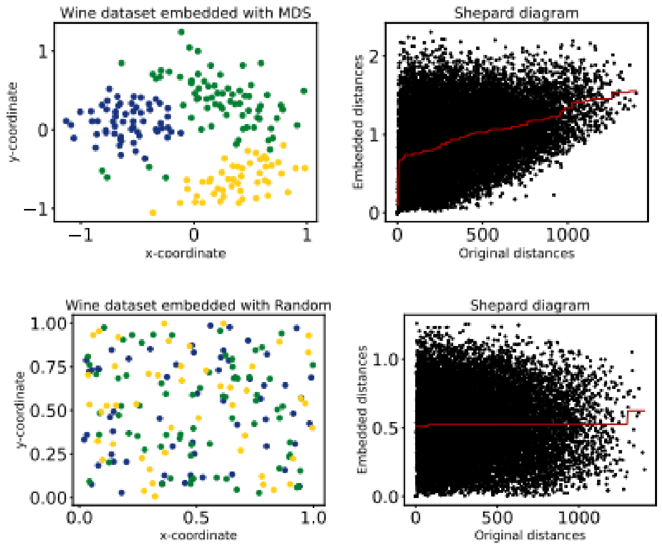
$$q_{ij} = \frac{\left(1 + \left|\left|y_i - y_j\right|\right|^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \left|\left|y_k - y_l\right|\right|^2\right)^{-1}}$$



Evaluating a projection

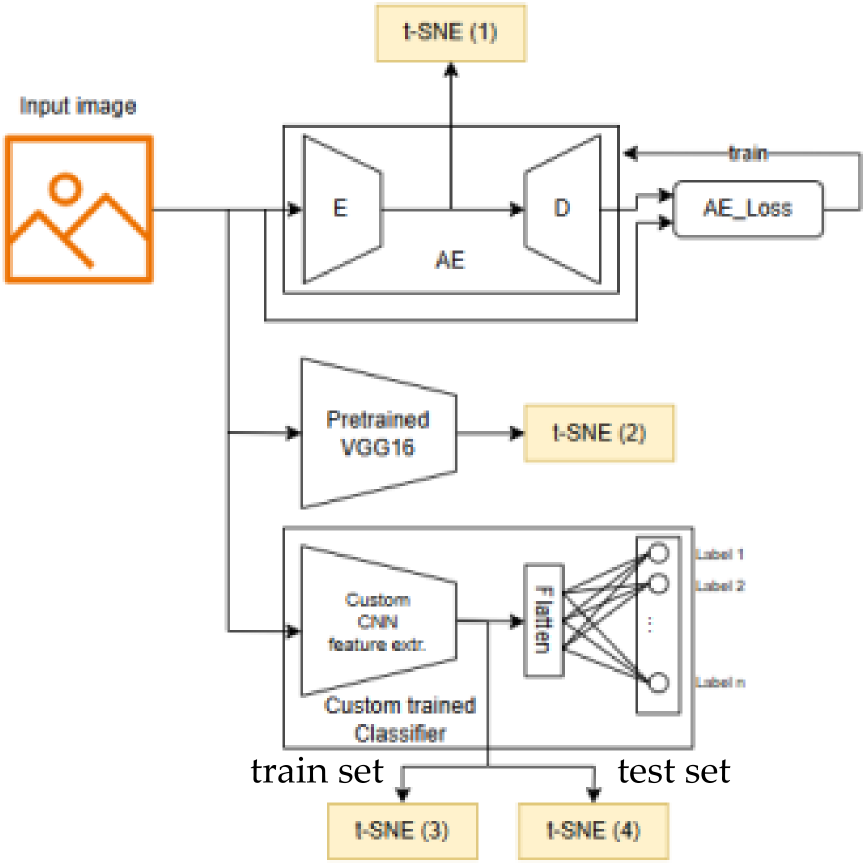
- Metrics
  - Raw Stress (RS) $RS(X,P) = \sum_{i,j} \left(\Delta^X(x_i, x_j) - \Delta^P(p_i, p_j) \right)^2$
    - MSE between pairwise differences in high and low dimensional spaces
  - Normalized Stress (NS) $NS(X,P) = \sqrt{\frac{\sum_{i,j} \left(\Delta^X(x_i, x_j) - \Delta^P(p_i, p_j) \right)^2}{\sum_{i,j} \Delta^X(x_i, x_j)^2}}$
    - Reduce the amplitude of RS
  - Scale-Normalized Stress (SNS) $SNS(X,P) = \min_{\alpha > 0} NS(X, \alpha P)$
  - Shepard Goodness Score (SGS)
    - Sperman rank correlation of the Shepard diagram
  - Non-Metric (Kruskal) Stress (NMS)
    - Measure of distances order preservation
    - Involves isotonic regression on the Shepard diagram

$$NMS(X,P) = \frac{\sum_{i,j} \left(\Delta^{\hat{X}}(\hat{x}_i, \hat{x}_j) - \Delta^P(p_i, p_j) \right)^2}{\sum_{i,j} \Delta^P(p_i, p_j)^2},$$



Shepard diagram of a good and bad clustering (Smelser et al.)

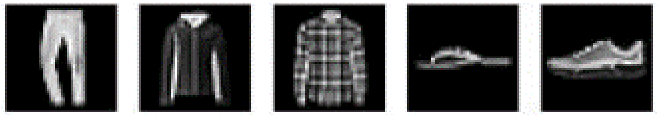
Experiment setup



- t-SNE plot on multiple representations of each datasets:
  - An autoencoder (AE) latent space
  - Pretrained VGG-16
  - Trained CNN classifier (train & test subsets)
- Purpose: exploring the way rearranging the same information affects dimensionality reduction

Datasets

Fashion FMNIST



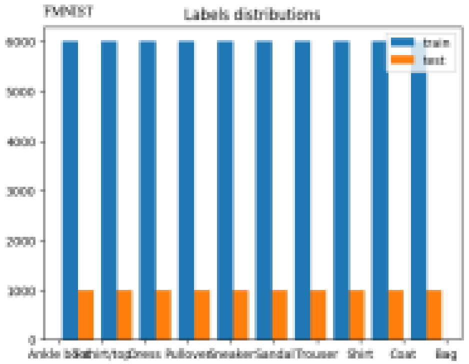
- 10 classes
- Benchmarking dataset
- Curated and balanced
- Many samples

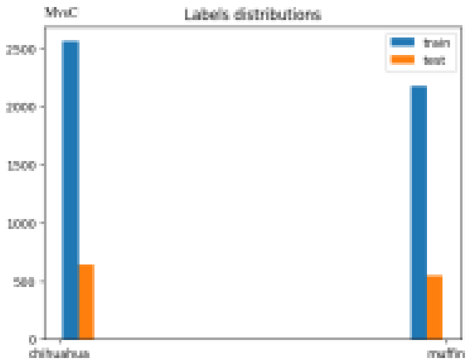
Muffin vs Chihuahua

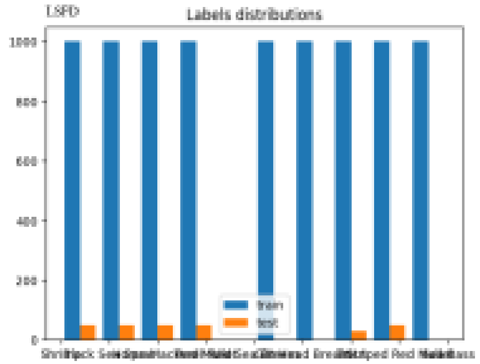
- 2 classes
- Contextual diversity
- Real world images
- Less normalized data

Large-Scale Fish Dataset

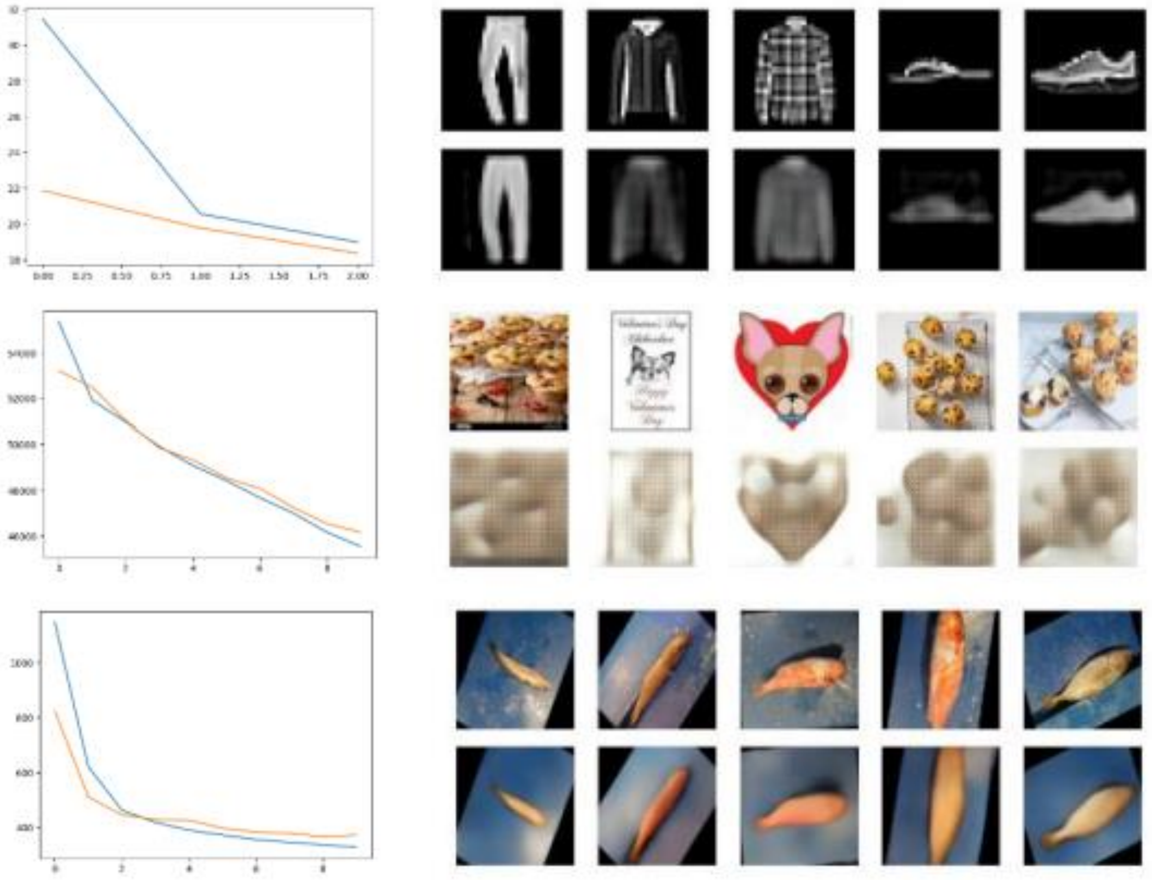
- 9 classes
- Geometrically predictable
- Easier to extract features
- Pre-augmented (just train)



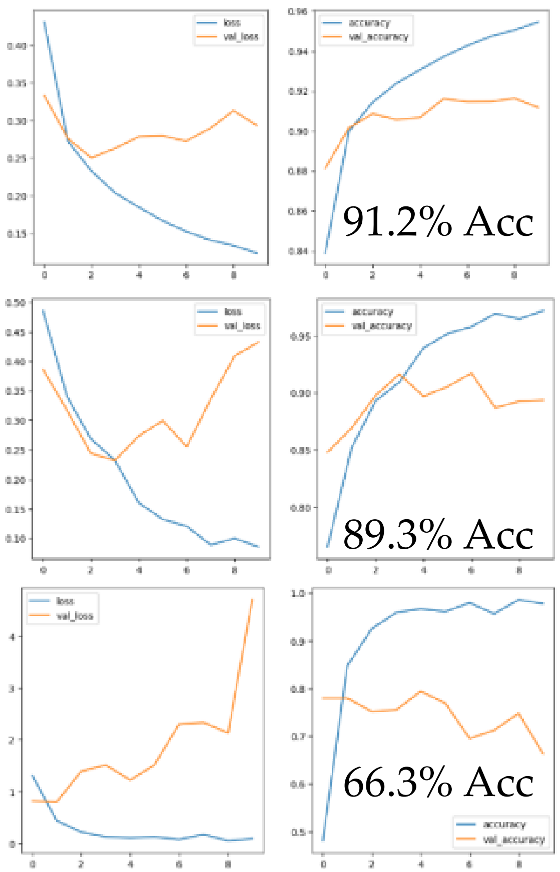


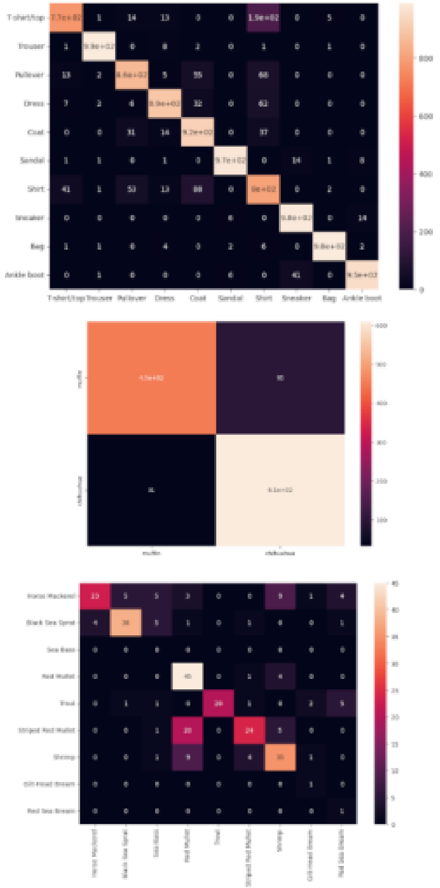


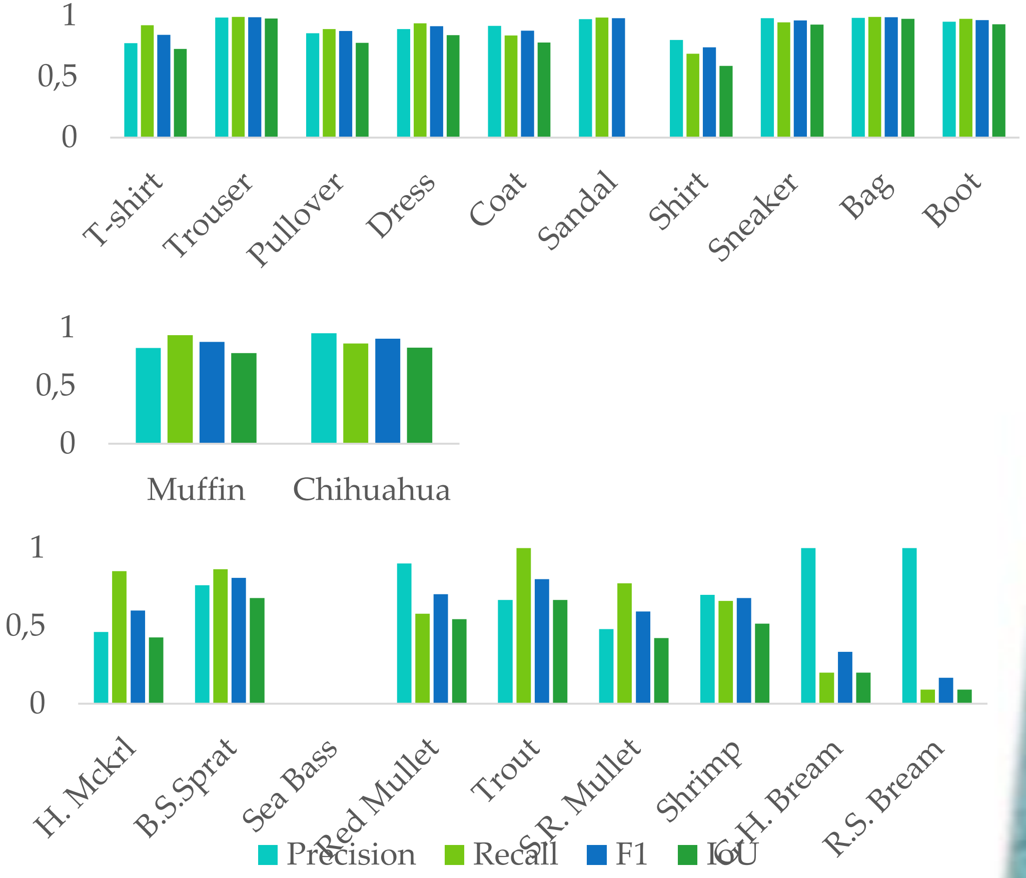
Models Training Results - Autoencoder



Models Training Results – CNN classifier



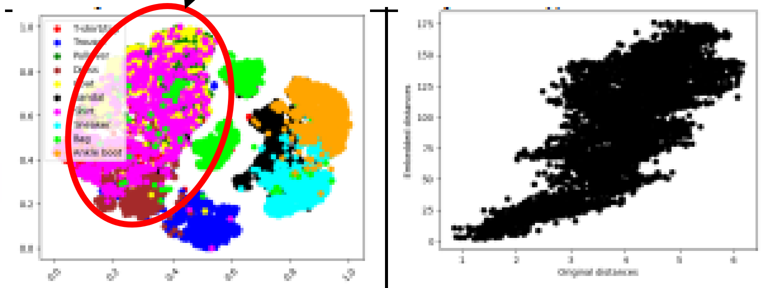




Projection Results – Fashion MNIST

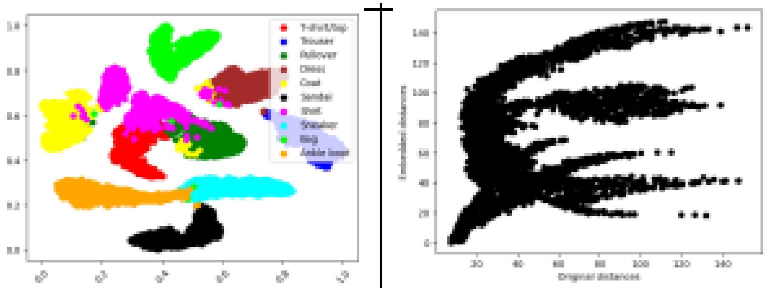
"Mega-cluster" –
Shirt, T-Shirt, Coat, Pullover

Stripe-like point formations
in Shepard diagram



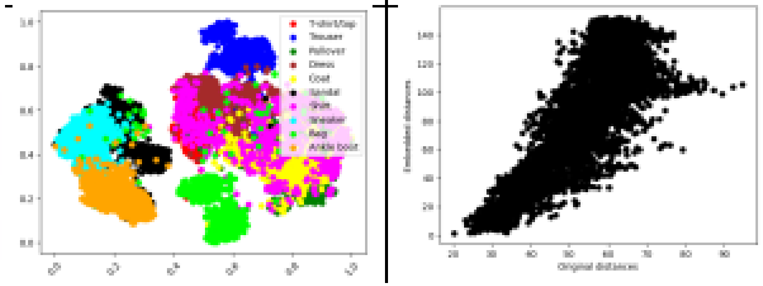
NS = 22.7105
SGS = 0.6678
SNS = 0.3114
NMS = 0.3077

t-SNE (1) -AE



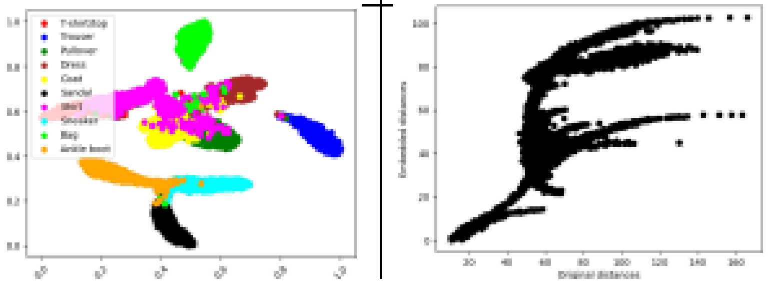
NS = 0.5829
SGS = 0.5244
SNS = 0.4246
NMS = 0.3563

t-SNE (3) –Cls. train.



NS = 0.5895
SGS = 0.44816
SNS = 0.3838
NMS = 0.3692

t-SNE (2) -VGG



NS = 0.3884
SGS = 0.6331
SNS = 0.3879
NMS = 0.3456

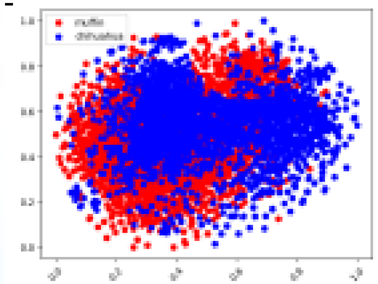
t-SNE (3) –Cls. val.

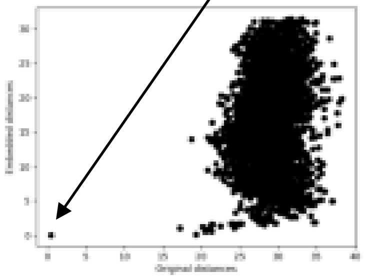
Projection Results – Muffin vs Chihuahua

Unable to distinguish between classes

Duplicate samples

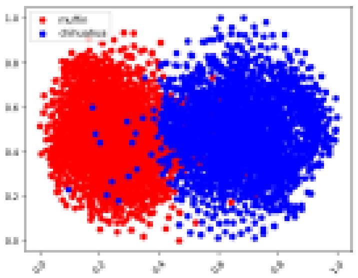
Better separation

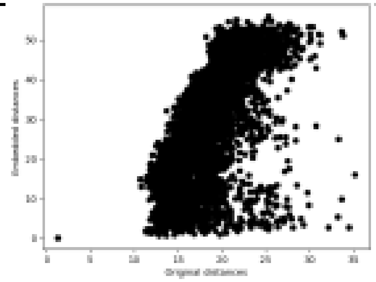




NS = 0.5822
SGS = 0.5081
SNS = 0.4063
NMS = 0.3927

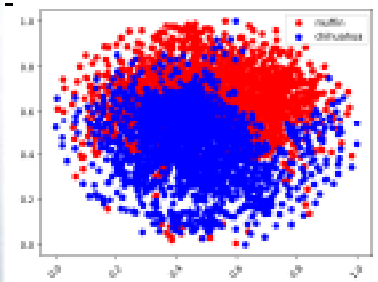
t-SNE (1) -AE

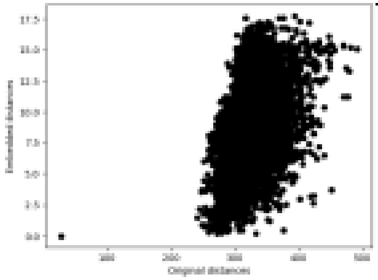




NS = 0.5582
SGS = 0.5613
SNS = 0.4190
NMS = 0.4106

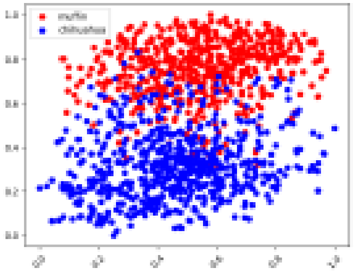
t-SNE (3) –Cls. train.

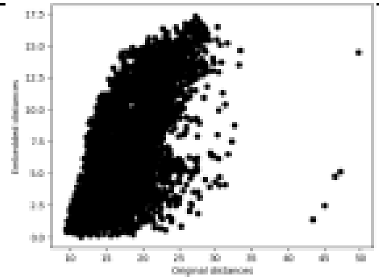




NS = 0.9778
SGS = 0.3268
SNS = 0.42915
NMS – 0.4255

t-SNE (2) -VGG

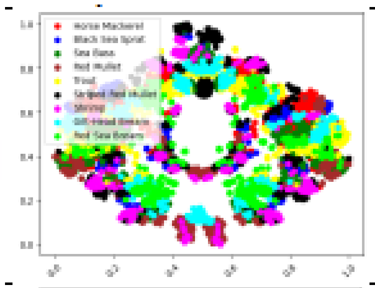


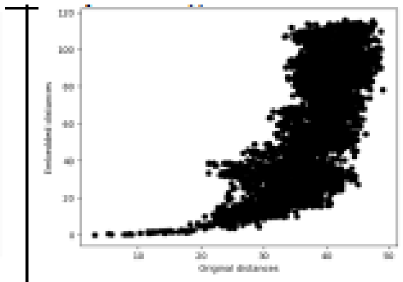


NS = 0.6172
SGS = 0.6077
SNS = 0.4031
NMS = 0.3875

t-SNE (3) –Cls. val.

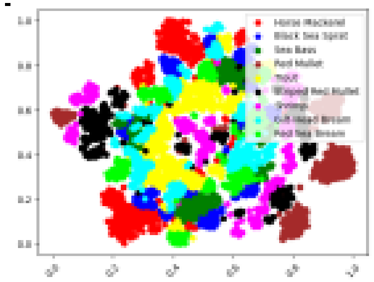
Projection Results – Fish Dataset

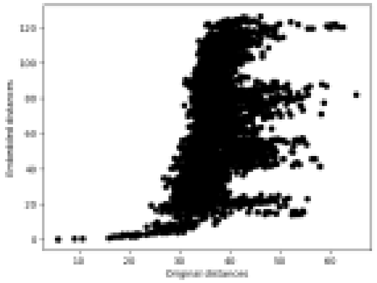




NS = 0.8448
SGS = 0.5431
SNS = 0.3684
NMS = 0.3338

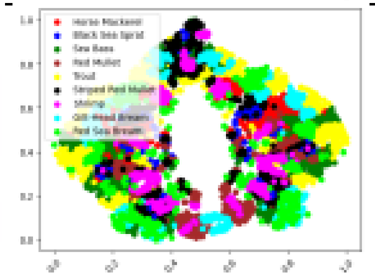
t-SNE (1) -AE

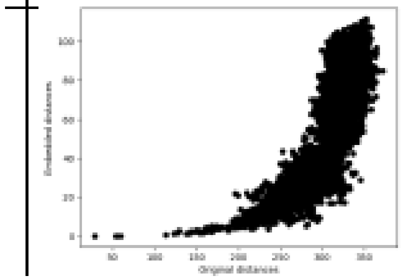




NS = 1.0561
SGS = 0.4176
SNS = 0.3873
NMS = 0.3841

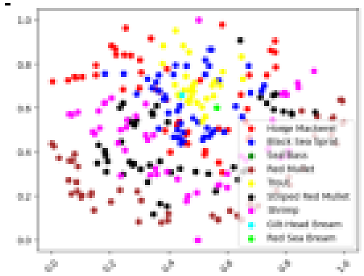
t-SNE (3) –Cls. train.

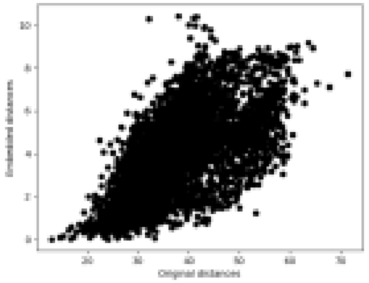




NS = 0.7918
SGS = 0.6626
SNS = 0.3717
NMS = 0.3011

t-SNE (2) -VGG





NS = 0.8909
SGS = 0.6682
SNS = 0.3518
NMS = 0.3287

t-SNE (3) –Cls. val.

Radial structure due to rotation during augmentation

Train-test discrepance when projecting classifier embeddings – caused by differences in the processing methods of the samples subsets of the dataset

Metrics statistics

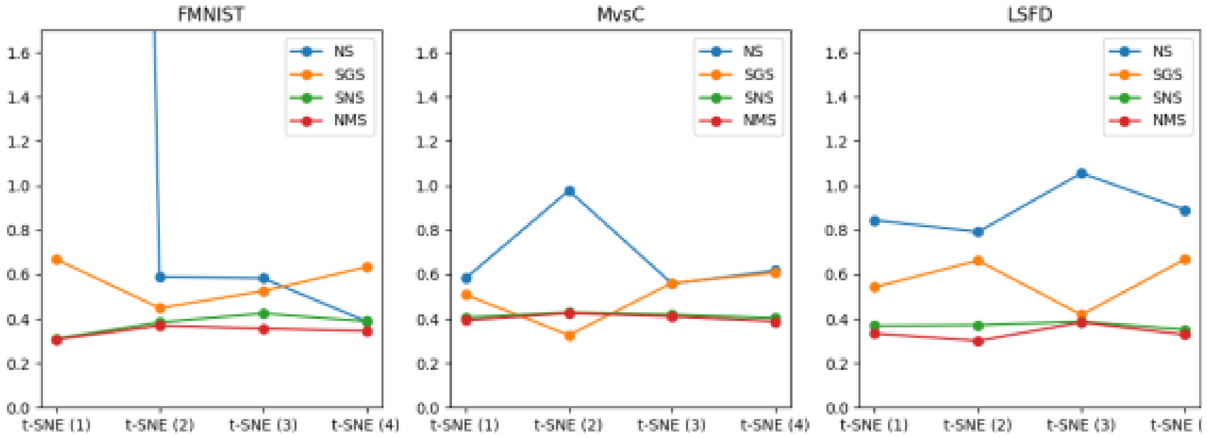


Figure 6. Projection metrics evolution over the four phases

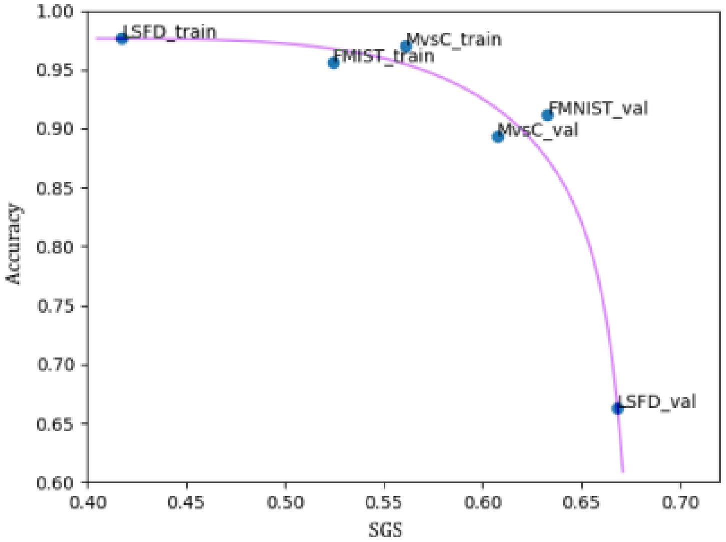
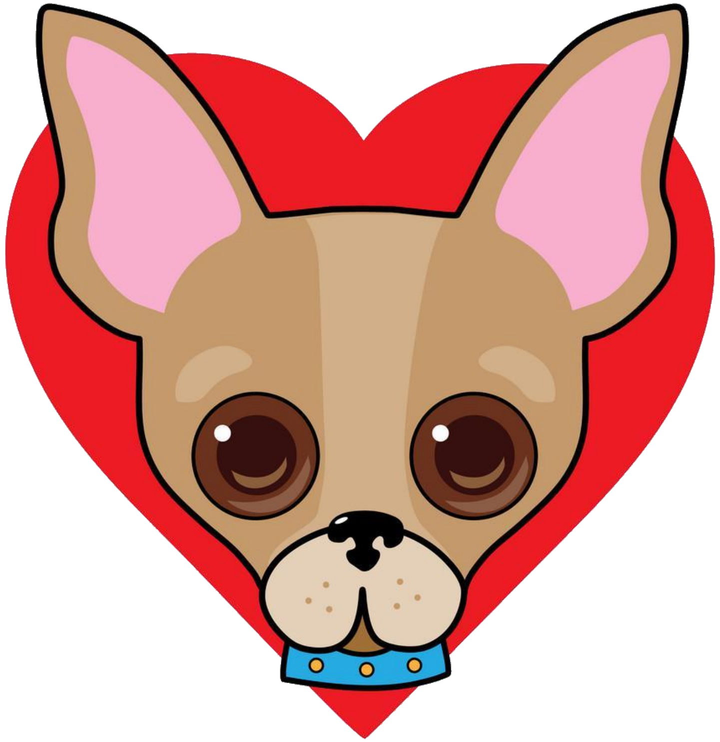


Figure 7. Relation between SGS and Accuracy

Conclusions and improvement opportunities

- t-SNE can reveal cluster structures, but further investigation is needed to reveal their meaning and validity,
- can detect structural patterns in the dataset (e.g. geometric similarities)
- … but it struggles to handle large variations of contexts.
- Combining projection methods with supervised learning may provide an idea why overfitting happens
- Further work
  - Refine models training
  - Try other projection methods (PCA, MDS, UMAP) and metrics (local, per-cluster)
  - Evaluate more datasets

Thank you for your attention!



References

- S. Cortinhas. Muffin vs Chihuahua image classification dataset. https://www.kaggle.com/datasets/samuelcortinhas/muffin-vs-chihuahua-imageclassification. Accessed: 2024-11-17
- M. Espadoto et al. "Toward a Quantitative Survey of Dimension Reduction Techniques". In: IEEE Transactions on Visualization and Computer Graphics 27.3 (2021), pp. 2153–2173. doi: 10.1109/TVCG.2019.2944182.
- B. Ghojogh et al. "Stochastic neighbor embedding with Gaussian and student-t distributions: Tutorial and survey". In: arXiv preprint arXiv:2009.10301 (2020).
- K. J. Smelser, J. Miller, and S. G. Kobourov. ""Normalized Stress" is Not Normalized: How to Interpret Stress Correctly". In: ArXiv abs/2408.07724 (2024). url: https://api.semanticscholar.org/CorpusID:271874691.
- O. Ulucan, D. Ulucan, and M. Turkan. "A Large-Scale Dataset for Fish Segmentation and Classification". In: Oct. 2020. doi: 10.1109/ASYU50717.2020.9259867.
- H. Xiao, K. Rasul, and R. Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms". In: arXiv preprint arXiv:1708.07747 (2017).