The tree autoencoder model,

with application to hierarchical data visualization

Miguel Á. Carreira-Perpiñán Kuat Gazizov

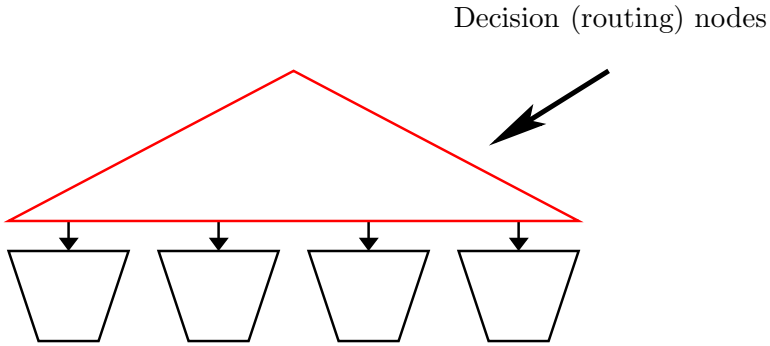
Department of Computer Science and Engineering
University of California, Merced

38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024)

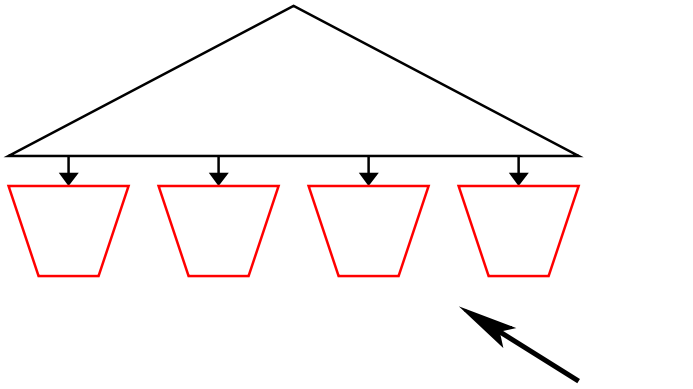
NEURAL INFORMATION PROCESSING SYSTEMS

UCMERCED

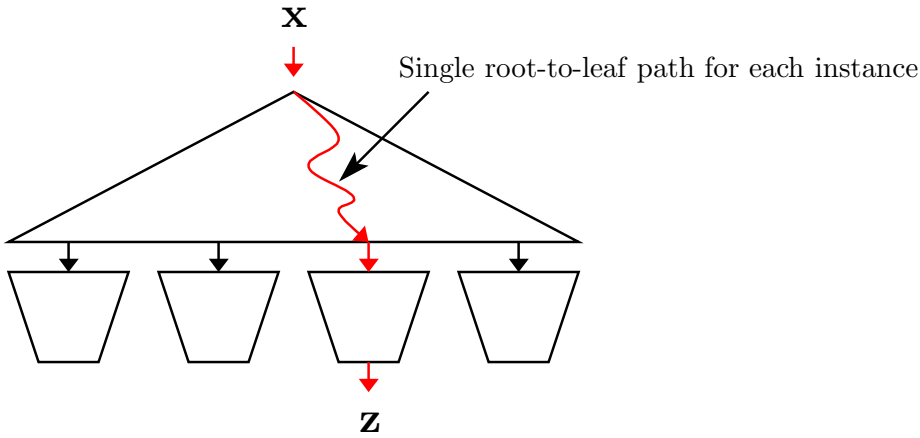
Encoder



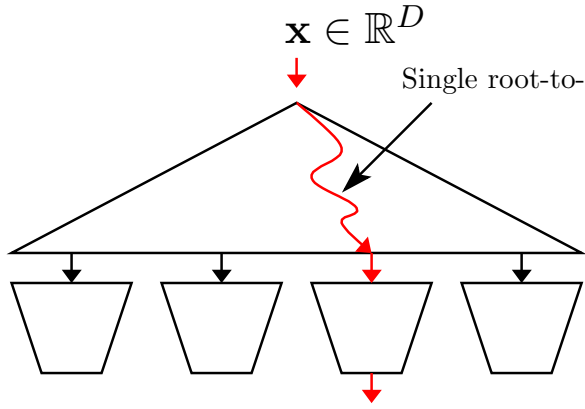
Encoder



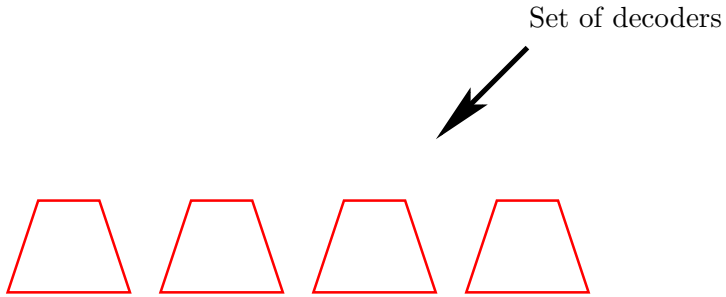
Encoder



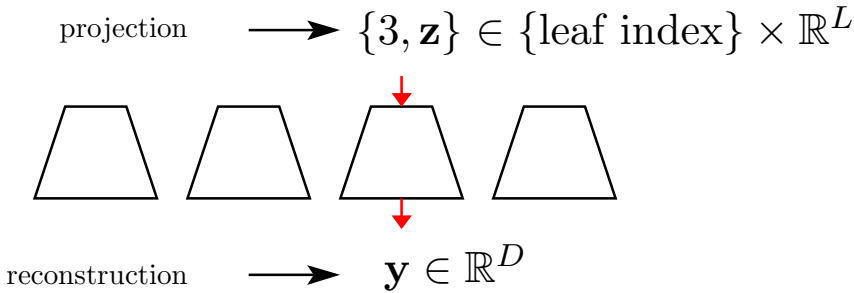
Encoder


$\mathbf{x} \in \mathbb{R}^D$
Single root-to-leaf path for each instance
projection
$\{3, \mathbf{z}\} \in \{\text{leaf index}\} \times \mathbb{R}^L$

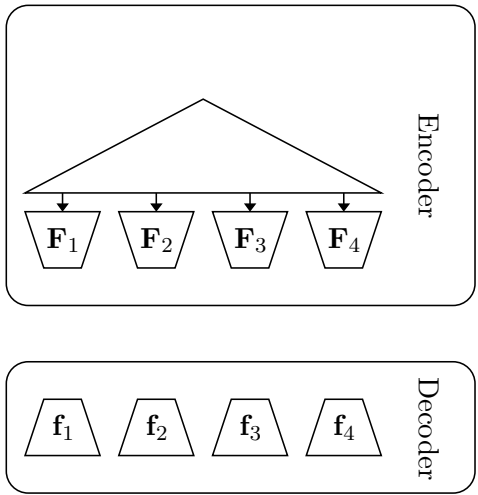
Decoder



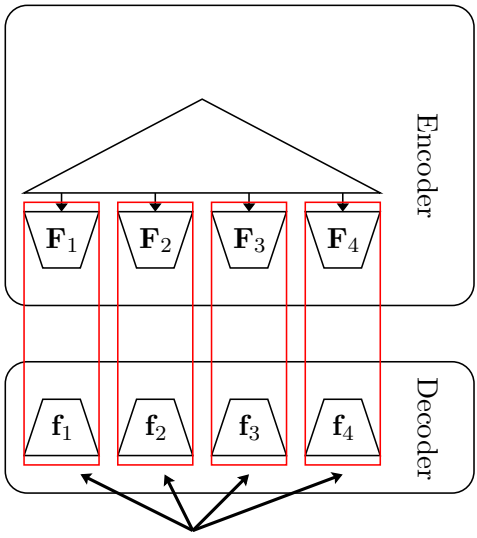
Decoder


projection
$\{3, \mathbf{z}\} \in \{\text{leaf index}\} \times \mathbb{R}^L$
reconstruction
$\mathbf{y} \in \mathbb{R}^D$

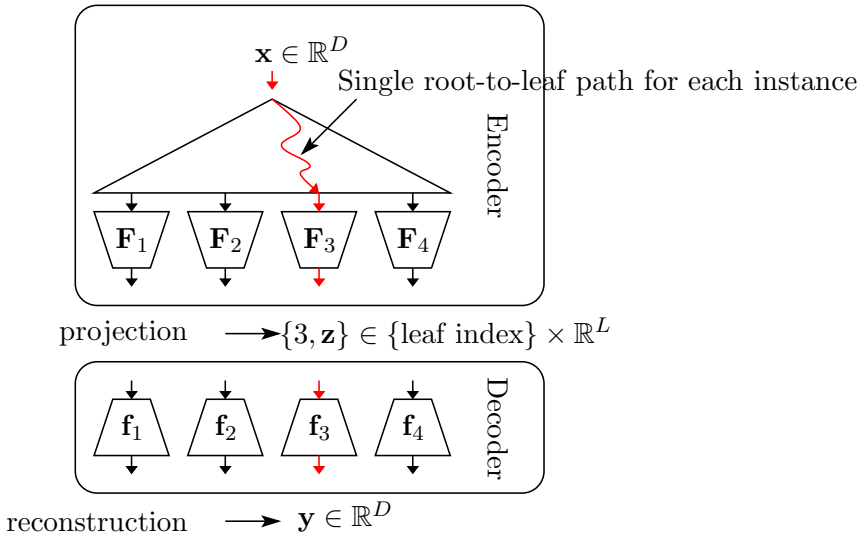
Tree autoencoder



Tree autoencoder



Tree autoencoder



Training

- Objective function:

$$E(\boldsymbol{\Theta}) = \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{T}(\mathbf{x}_n; \boldsymbol{\Theta})\|_2^2 + \lambda \sum_{i \in \text{decision nodes}} \|\mathbf{w}_i\|_1$$

- Since the problem is non-differentiable, gradient descent methods are not applicable
- We use Tree Alternating Optimization

Advantages

- It optimizes a clear objective function – reconstruction error
- It does not require a neighborhood graph
  - Unlike t-SNE or UMAP, which is very costly
- It does not increase distances
  - Distances are preserved or reduced, not artificially increased
- It is highly interpretable
- It has out-of-sample mapping
- It is efficient during inference
  - Logarithmic in the number of leaves
- It is linear in the number of instances.
  - The approximate cost is $\Theta(ND^2)$ for shallow trees and $\mathcal{O}(ND)$ for deep trees

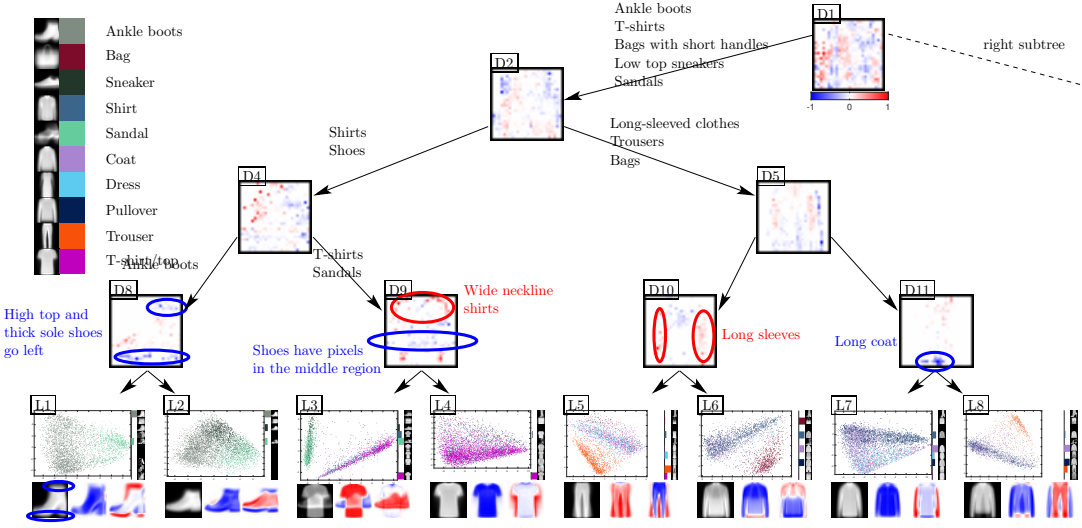
Disadvantages

- It needs access to the explicit feature vectors, unlike graph-based methods, which need instead pairwise affinities.
- The optimization converges to a local optimum, which depends on the initialization
- As with any DR method, PCA trees reveal only some structure, so practitioners should try multiple, complementary methods

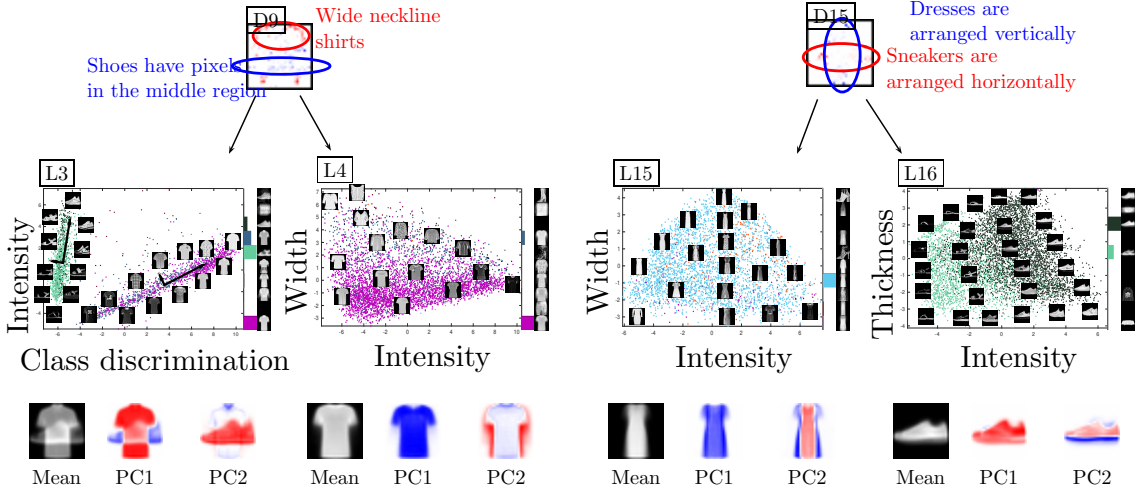
Fashion MNIST

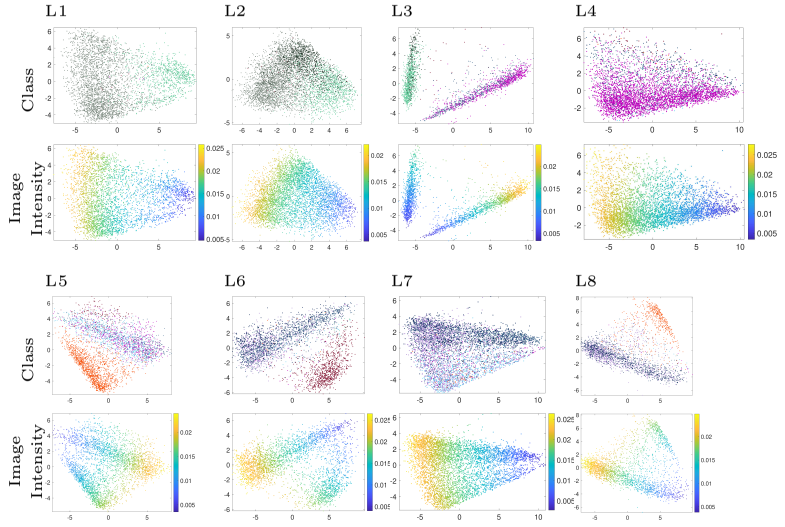
The dataset consists of $60,000$ grayscale images of size 28×28 pixels, organized into 10 classes. Below is an example, where each column corresponds to a different class.

PCA tree in action

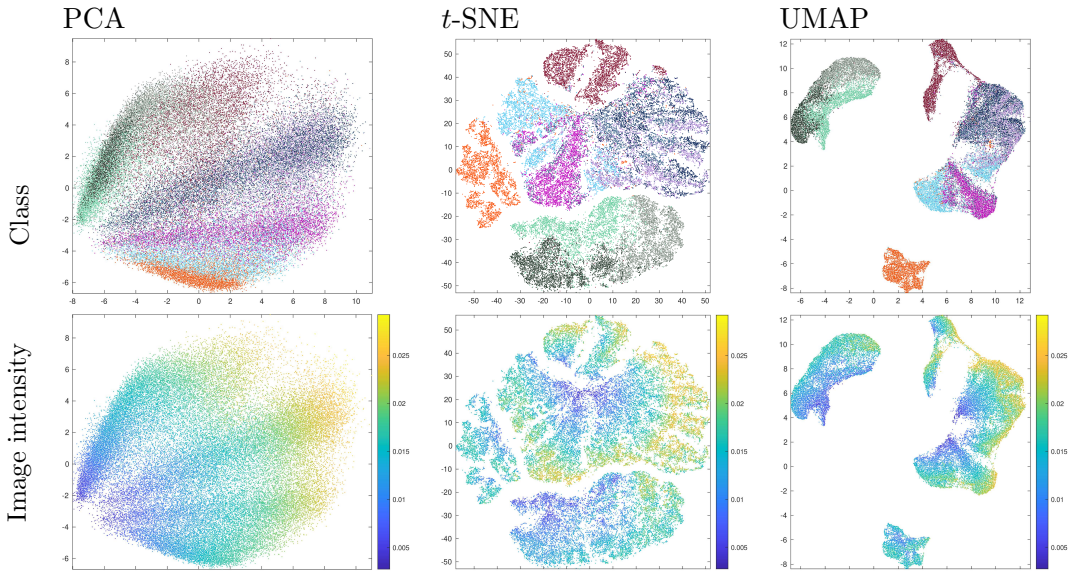


PCA tree in action




L1
L2
L3
L4
L5
L6
L7
L8
Class
Image Intensity

Well established algorithms



Can we trust t-SNE or UMAP?