Deep Clustering based on a Mixture of Autoencoders

Shlomo E. Chazan, *Sharon Gannot* and *Jacob Goldberger*

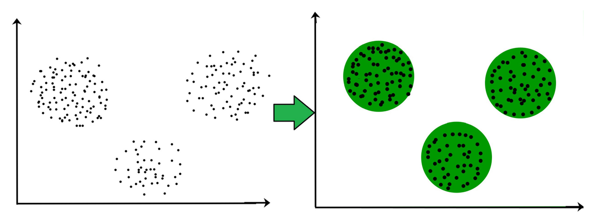
Faculty of Engineering, Bar-Ilan University Ramat-Gan, Israel

{Shlomi.Chazan, Sharon.Gannot , Jacob.Goldberger}@biu.ac.il

Main Contribution

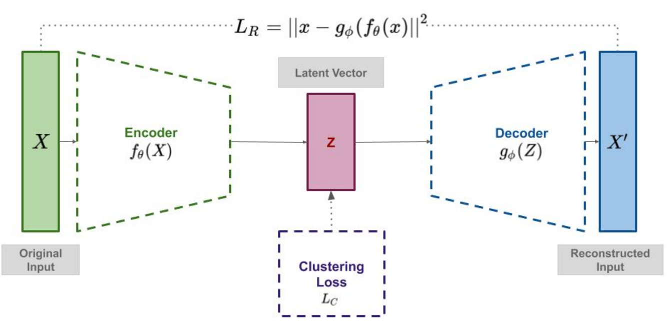
- We introduce Mixture of Deep Auto Encoders for the clustering task
- End-to-end deep learning based approach for clustering
- Each autoencoder is an expert in one cluster
- The gate network carries out the clustering itself

Clustering problem



Deep-clustering approach

- Clustering high-dimensional datasets is hard since the inter-point distances become less informative in **high-dimensional** spaces
- Dimensionality reduction with **DNN**
- **K-means** is applied in the embedded space
- ✗ **Collapsing problem**
- ✗ **Regularization** is required



Deep-clustering drawbacks

- ✗ The DNN is only used to **find low-dimensional feature space**
- ✗ Requires regularization - the embedded space information can be entirely **irrelevant** to the clustering process
- ✗ Requires **fine-tuning** for each dataset

Why not using fully DNN-based clustering ?

Clustering representation

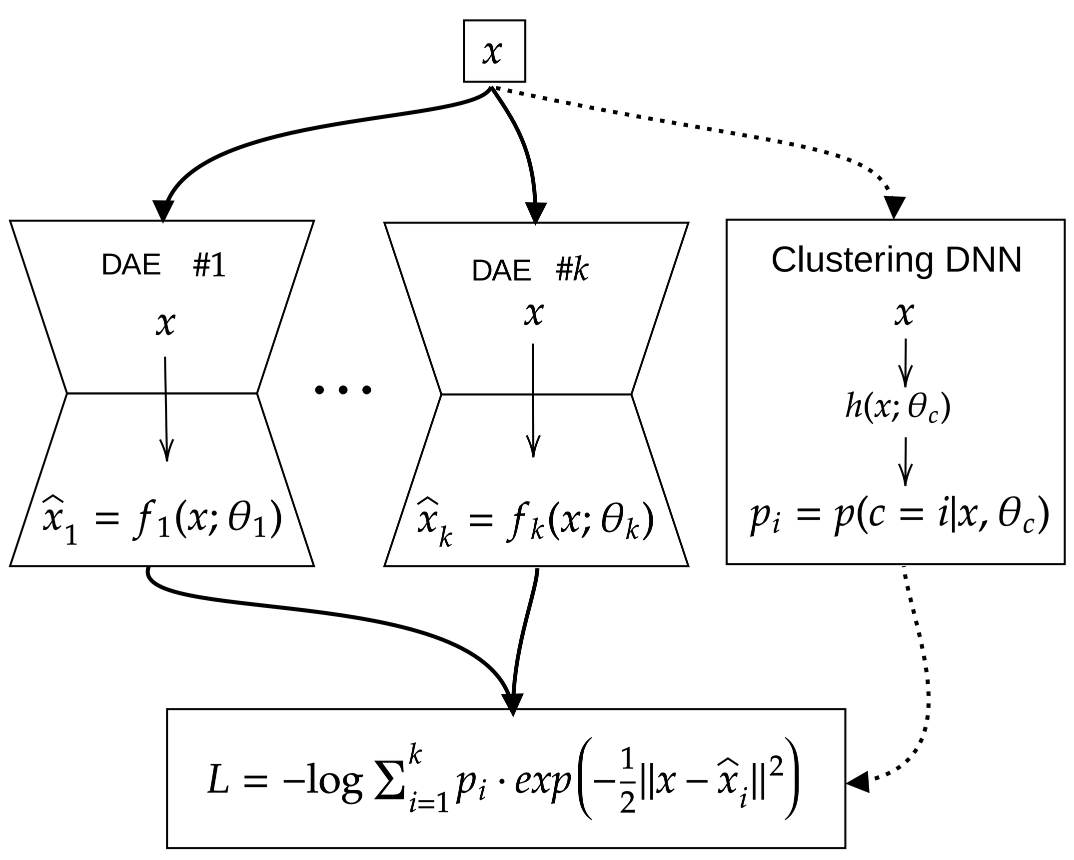
K-means approach

- The K-means algorithm represents each cluster by a **centroid**
- The clustering is carried out by finding the centroid with the **minimum distance** from the data point

Our approach

- ✓ Each cluster is represented by an **autoencoder** that specializes in reconstructing objects belonging to that cluster
- ✓ The clustering is carried out by directing the input object to the **most suitable autoencoder**

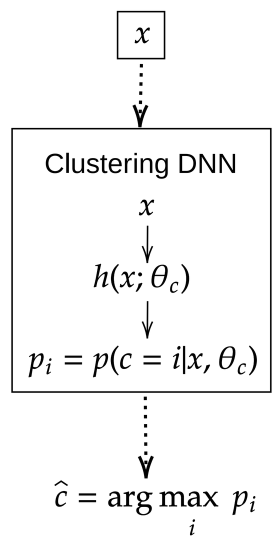
Deep Autoencoder Mixture Clustering (DAMIC)



The clustering procedure

Soft decision

- Only the gate DNN is used for the clustering
- Softmax - for soft decision



Hard decision

$$\hat{c} = \arg\max_{i=1}^{k} p(c = i|x;\theta_c) = \arg\max_{i=1}^{k} (w_i^T h(x) + b_i).$$

DAMIC algorithm

Goal: clustering $x_1, \ldots, x_n \in R^d$ into k clusters.

Network components:

- A network that computes a soft clustering of the data point:

$$p(c = i|x;\theta_c) = \frac{\exp(w_i^T h(x) + b_i)}{\sum_{j=1}^{k} \exp(w_j^T h(x) + b_j)}$$

- A set of autoencoders (one for each cluster):

$$x \to \hat{x}_i = f_i(x;\theta_i), \quad i = 1, \ldots, k$$

Pre-training:

- Train a single autoencoder for the entire dataset
- Apply a k-means algorithm in the embedded space
- Use the k-means labels to train separate DAE for each cluster separately
- Use the k-means labels to train the clustering DNN

Joint Training:

- clustering is obtained by minimizing the reconstruction error:

$$L(\theta_1, \ldots, \theta_k, \theta_c) = -\sum_{t=1}^{n} \log \Big(\sum_{i=1}^{k} p(c_t = i|x_t;\theta_c) \cdot \exp(-\frac{1}{2} \cdot ||x_t - f_i(x_t;\theta_i)||^2) \Big)$$

The final (hard) clustering is:

$$\hat{c}_t = \arg\max_{i=1}^{k} p(c_t = i|x_t;\theta_c), \quad t = 1, \ldots, n.$$

Training procedure

$$\frac{\partial L}{\partial \theta_c} = -\sum_{t=1}^{n} \sum_{i=1}^{k} w_{ti} \cdot \frac{\partial}{\partial \theta_c} \log p(c_t = i|x_t;\theta_c)$$

with

$$w_{ti} = \frac{p(c_t = i|x_t;\theta_c) \exp(-\frac{1}{2} \cdot ||x_t - f_i(x_t;\theta_i)||^2)}{\sum_{j=1}^{k} p(c_t = j|x_t;\theta_c) \exp(-\frac{1}{2} \cdot ||x_t - f_j(x_t;\theta_j)||^2)}$$

Clustering evaluation

NMI Normalized mutual information
ARI Adjusted rand index
ACC Clustering accuracy

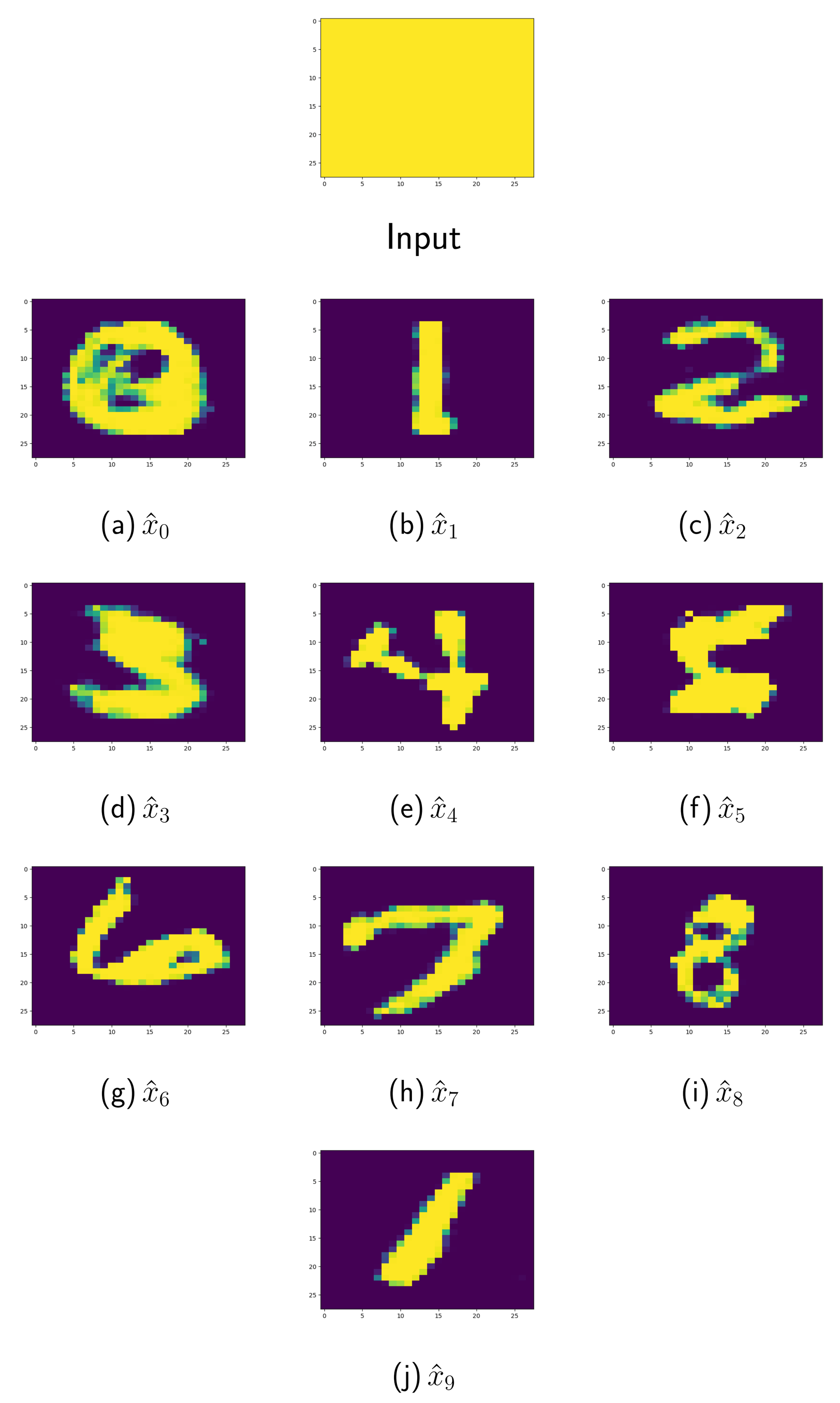
MNIST database

Method	DAMIC	DCN	DAE+KM	DEC	KM
NMI	**0.87**	0.81	0.74	0.80	0.50
ARI	**0.81**	0.75	0.67	0.75	0.37
ACC	**0.89**	0.83	0.80	0.84	0.53

Fashion-MNIST database

Method	DAMIC	DCN	DAE+KM	DEC	KM
NMI	**0.65**	0.55	0.60	0.54	0.51
ARI	**0.49**	0.42	0.45	0.40	0.37
ACC	**0.60**	0.50	0.57	0.51	0.47

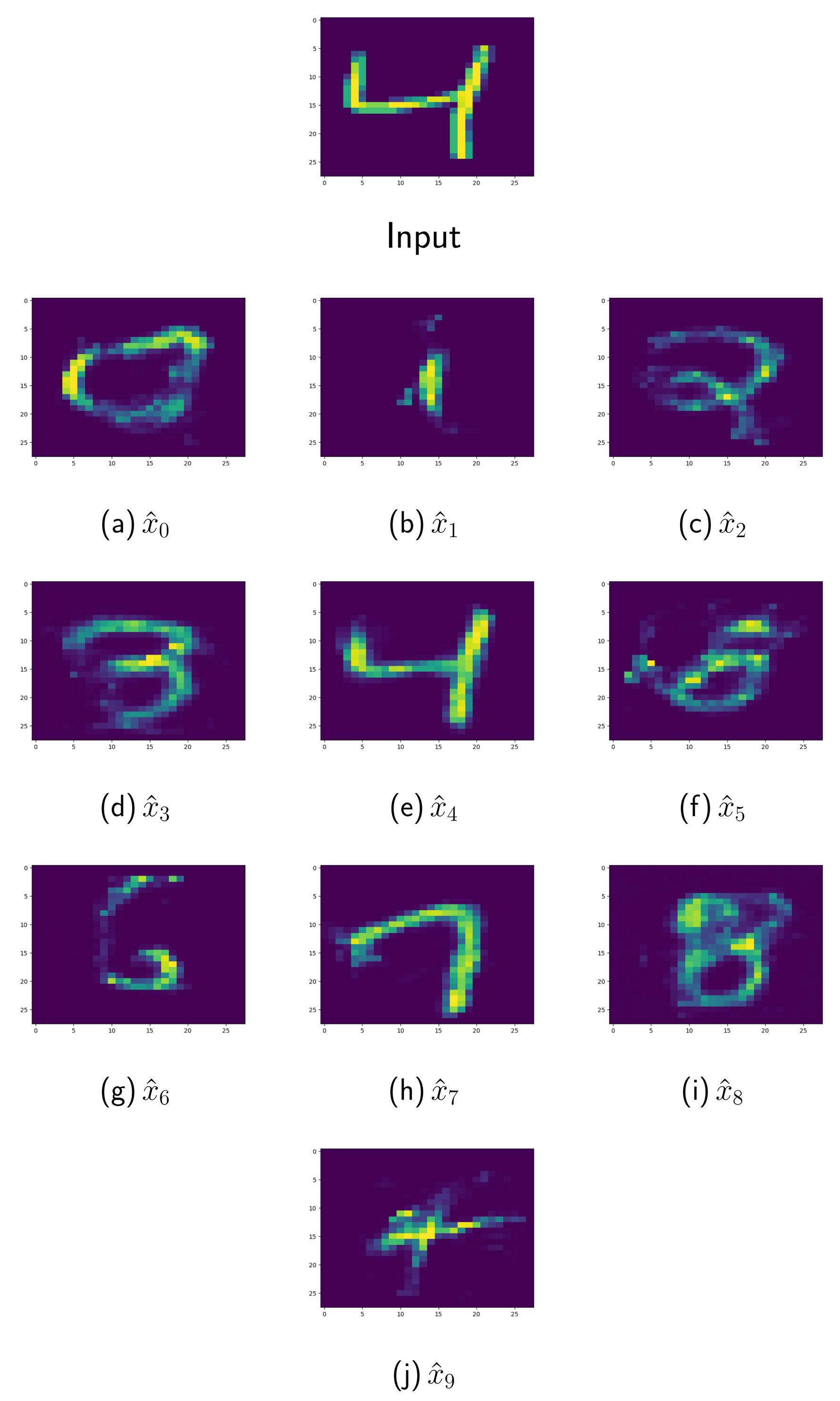
DAE expertise



Input

(a) $\hat{x}_0$ (b) $\hat{x}_1$ (c) $\hat{x}_2$ (d) $\hat{x}_3$ (e) $\hat{x}_4$ (f) $\hat{x}_5$ (g) $\hat{x}_6$ (h) $\hat{x}_7$ (i) $\hat{x}_8$ (j) $\hat{x}_9$

Best reconstruction wins



Input

(a) $\hat{x}_0$ (b) $\hat{x}_1$ (c) $\hat{x}_2$ (d) $\hat{x}_3$ (e) $\hat{x}_4$ (f) $\hat{x}_5$ (g) $\hat{x}_6$ (h) $\hat{x}_7$ (i) $\hat{x}_8$ (j) $\hat{x}_9$

$$p(c = 4|x;\theta_c) = 0.92 \gg p(c \neq 4|x;\theta_c) \approx 0$$

Ablation study on the MNIST database

Method	DAMIC	Pre-training	Joint-training	KM
NMI	**0.87**	0.74	0.71	0.50
ARI	**0.81**	0.67	0.53	0.37
ACC	**0.89**	0.80	0.60	0.53

Conclusions

- End-to-end DNN-based approach for clustering
- The clusters are represented by autoencoder networks
- Loss function does not suffer from the collapsing problem
- There is no need for regularization
- High performance (state-of-the-art in the Fashion-MNIST database)