Topological analysis of neural network

Michele Alessi

Università degli studi di Trieste: Advanced Topics in Machine Learning A.A. 2022-23

Goal of the project

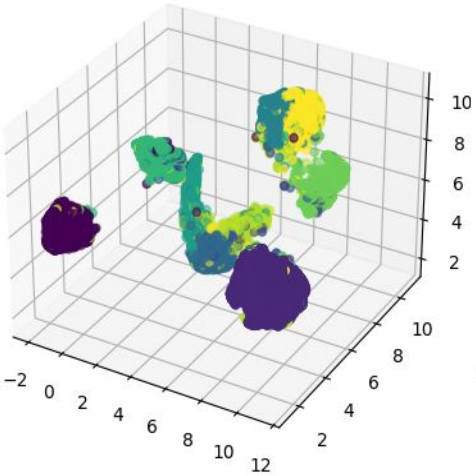
- Study the manifolds of weights and activations of a neural network using different techniques
- Adopt a new methodology to cluster data and infer some properties of the network using a new methodology based on mapper algorithm
- Recover neurons importance to classify a particular label

Step by step procedure:

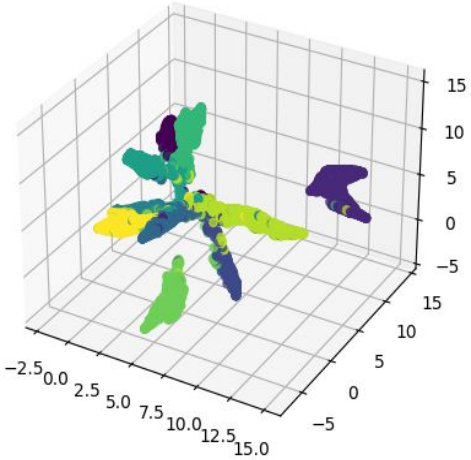
- Train a FCNN with **one hidden layer** to classify MNIST dataset, using different configurations of the network.
  - 100 hidden neurons
  - 1024 hidden neurons
  - Regularized regime
  - Not regularized regime

- After training, extract **weights**
- Use the model on the testset (10 000 images) and store **neurons activations** for each test image.

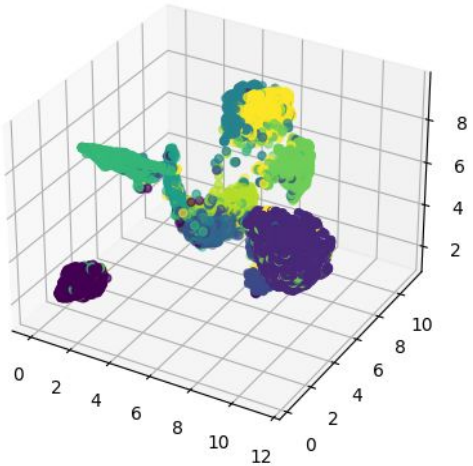
Hidden activ not reg.



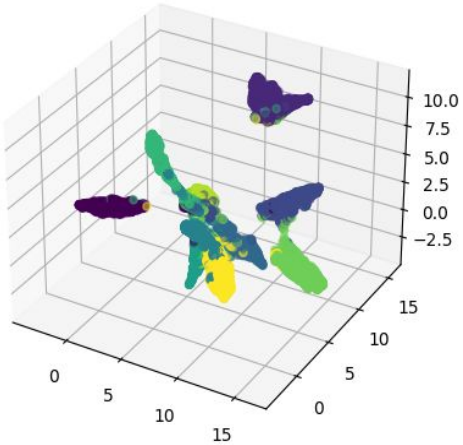
Output activ not reg.



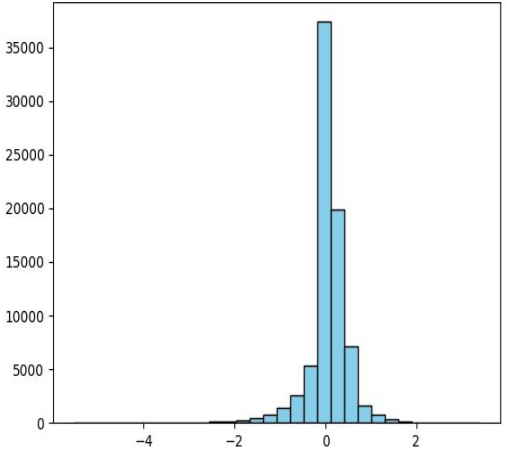
Hidden activ reg.



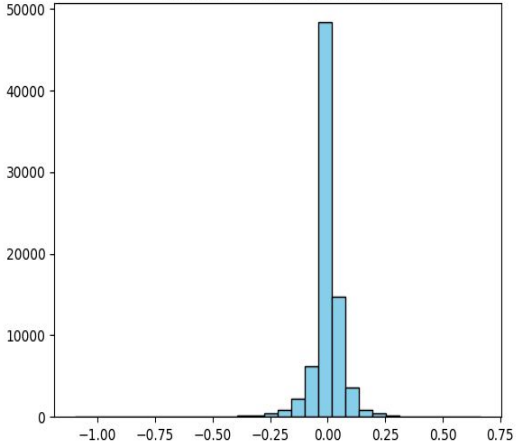
Output activ reg.



Weights distribution not reg.

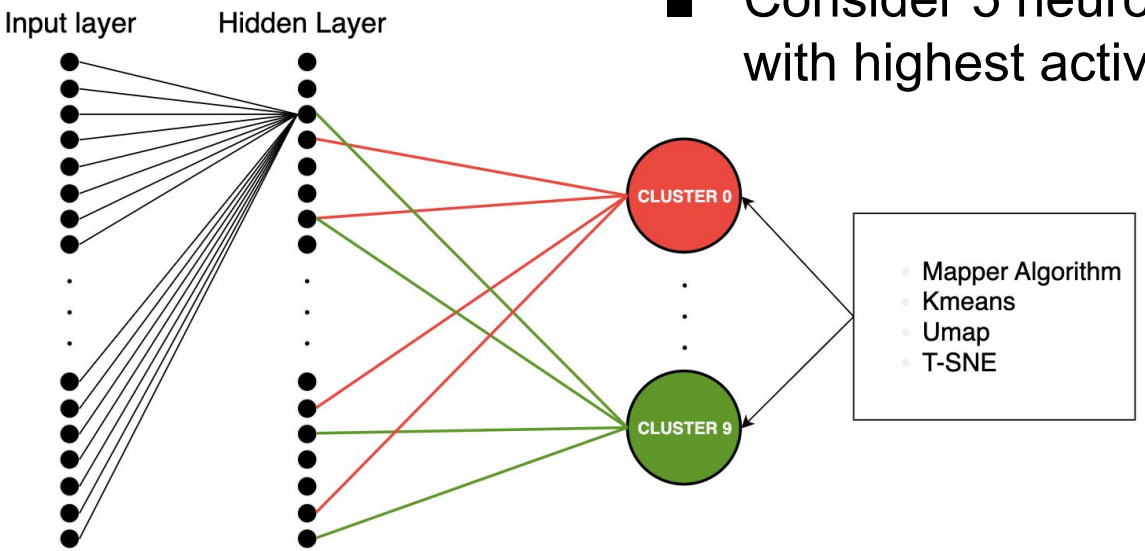


Weights distribution reg.



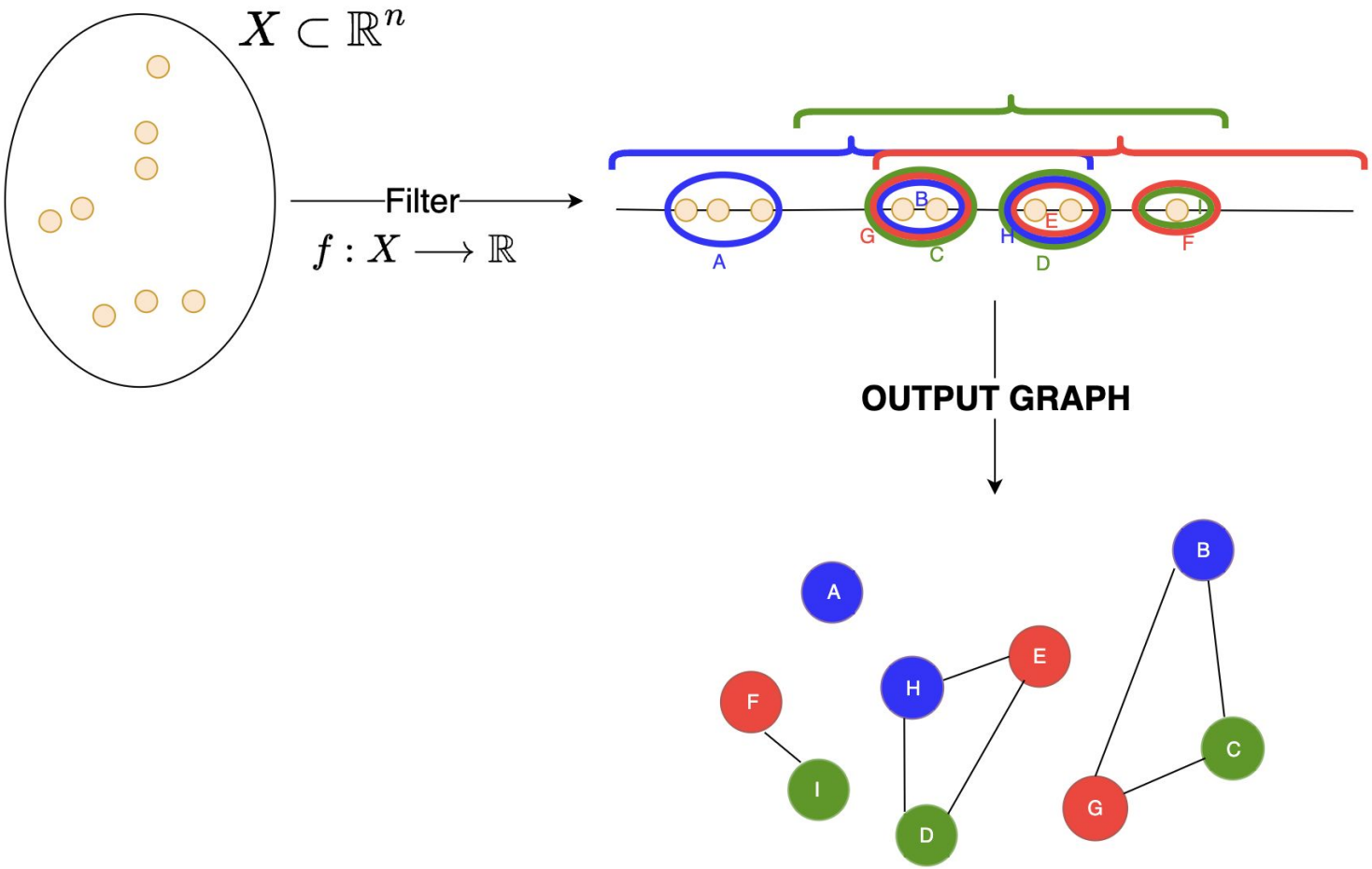
Step by step procedure:

- Cluster outputs activations manifold, using different methods
- For each method, for each cluster:
  - For each point in the cluster:
    - Consider 5 neurons in the hidden layer with highest activations

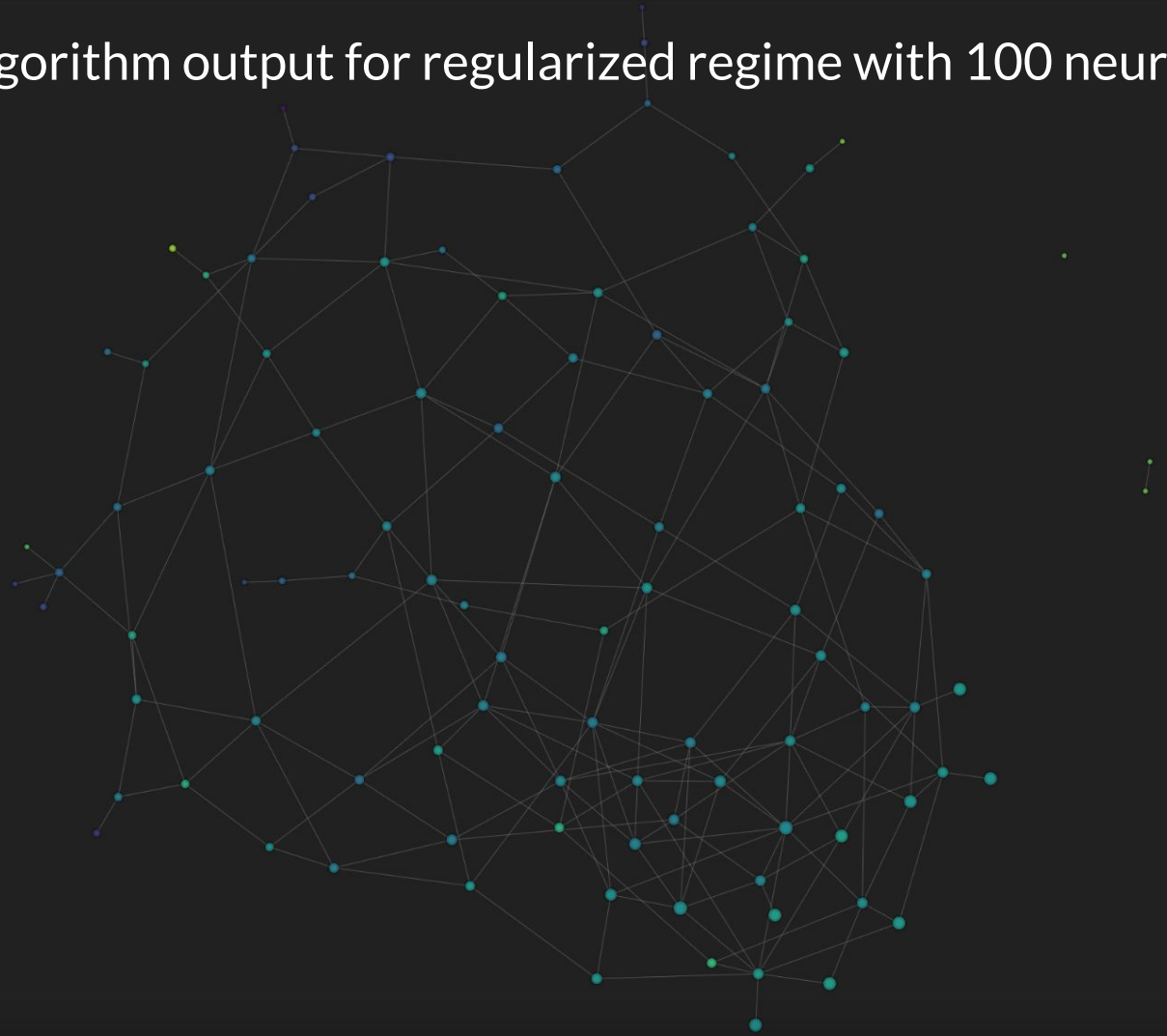


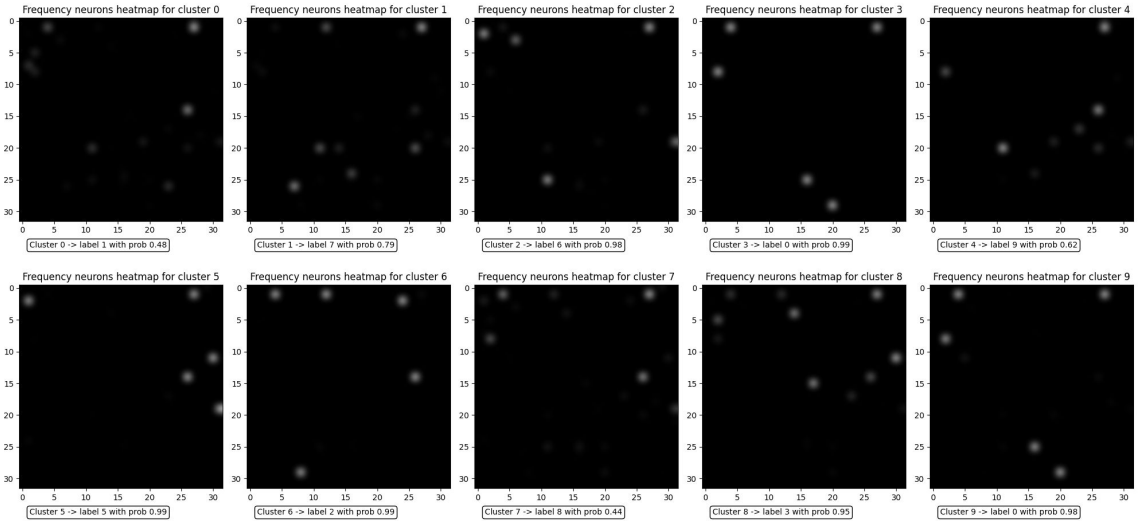
We end up with information about which neurons in the hidden layer tend to be activated the most for each cluster.

Mapper Algorithm



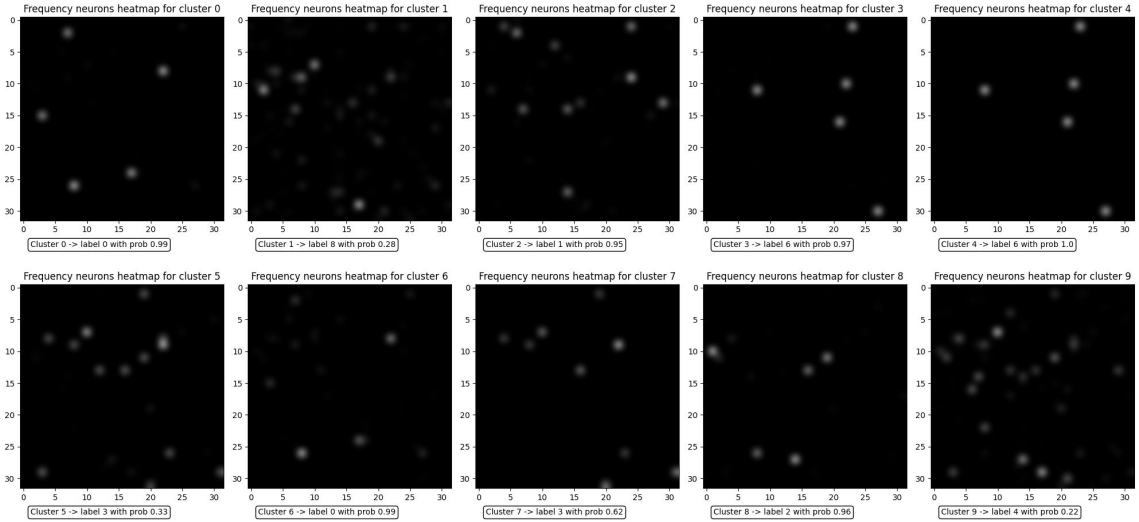
Mapper Algorithm output for regularized regime with 100 neurons



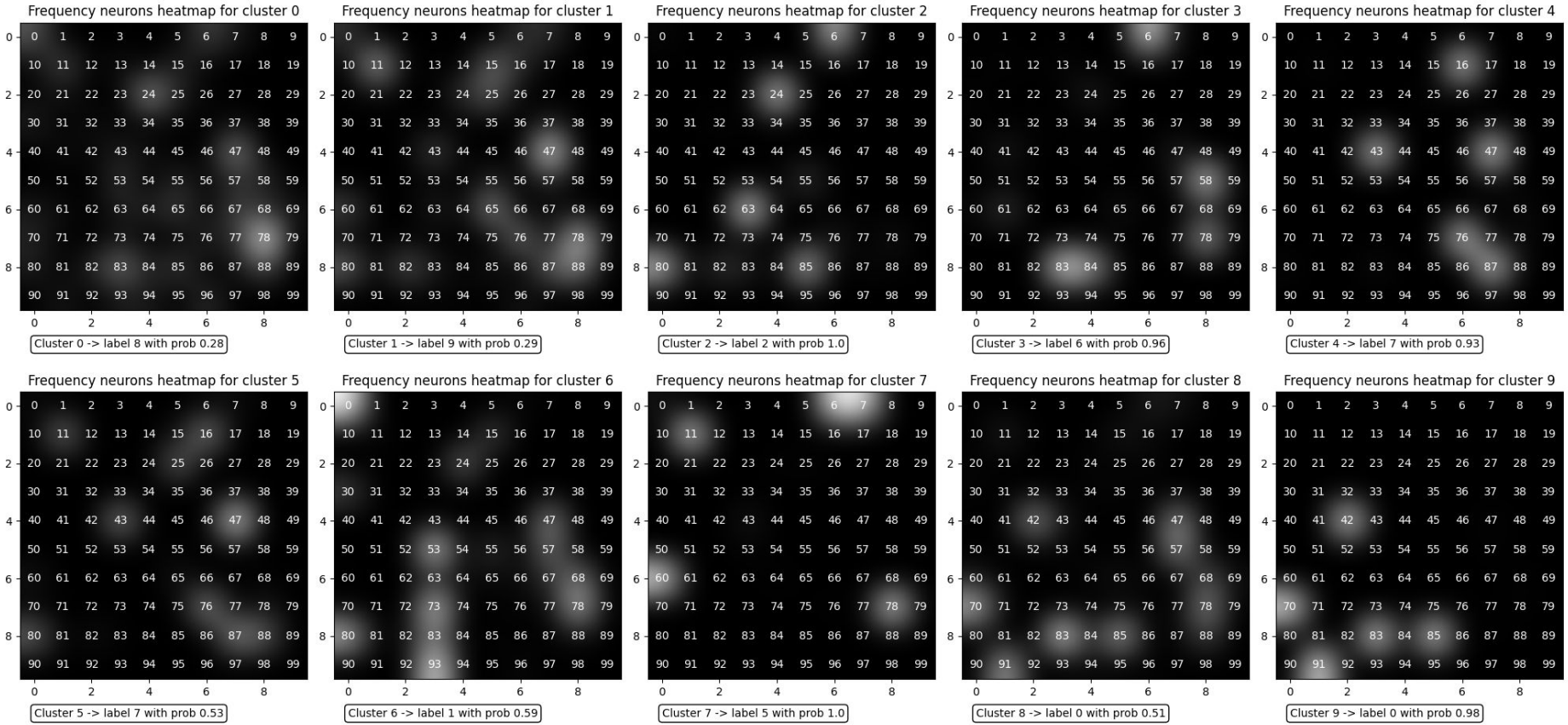


Mapper Heatmap for hidden-layer neurons activations - Regularized regime with 1024 neurons

Mapper Heatmap for hidden-layer neurons activations - **Non** regularized regime with 1024 neurons

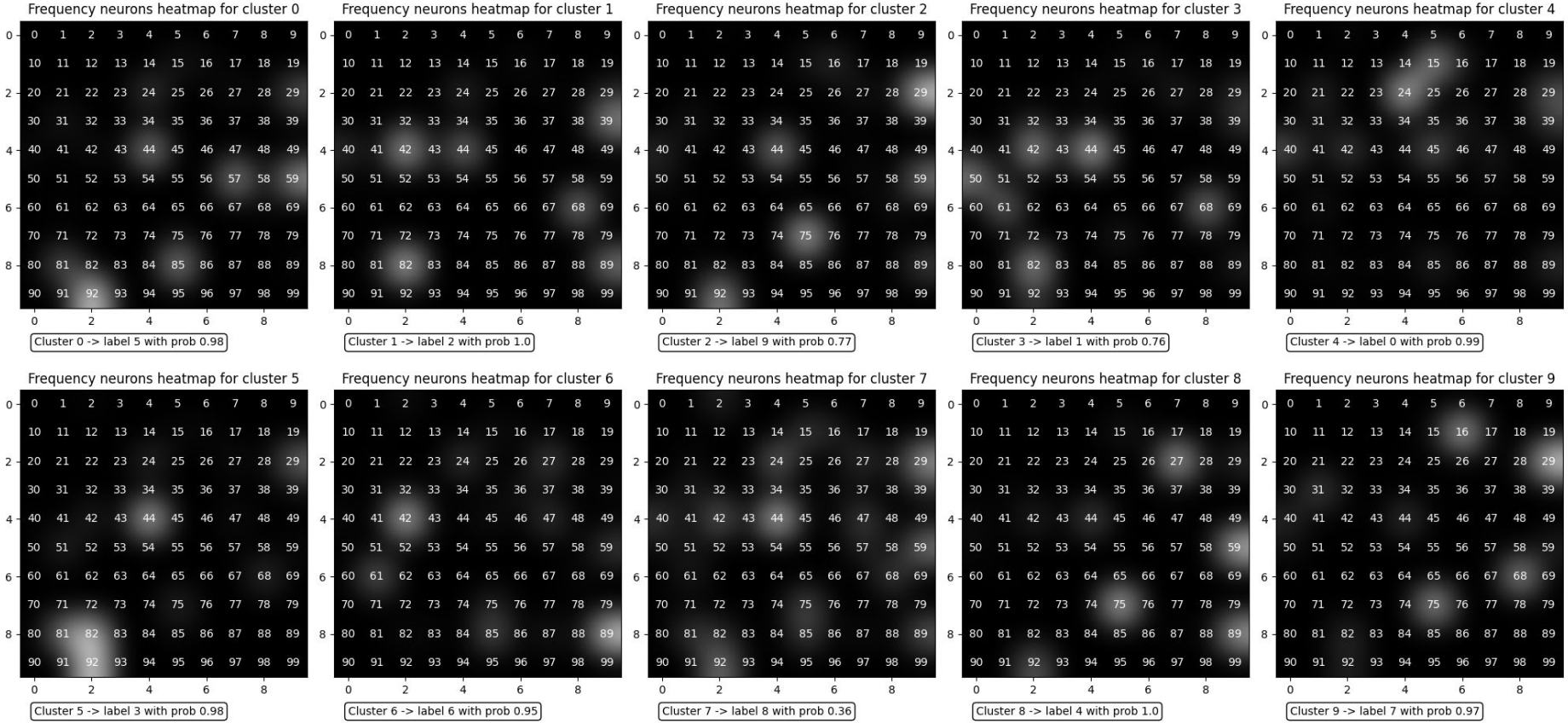


Mapper Heatmap for hidden-layer neurons activations - **Non** regularized regime with 100 neurons



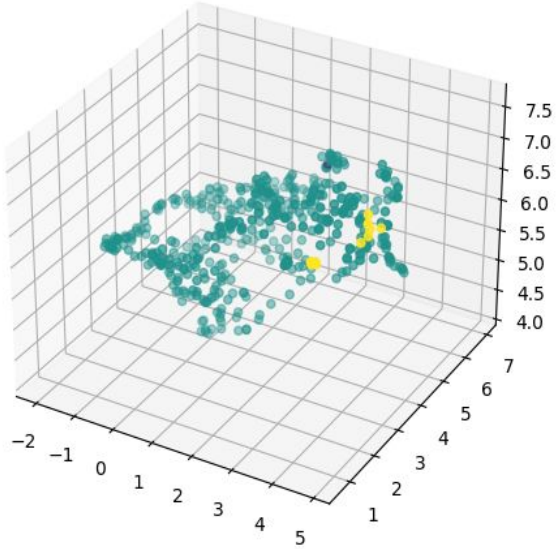
Mapper Heatmap for hidden-layer neurons activations - Regularized regime with 100 neurons

Note: clusters 0, 2 and 5 have same neurons activated: 92, 44, 29

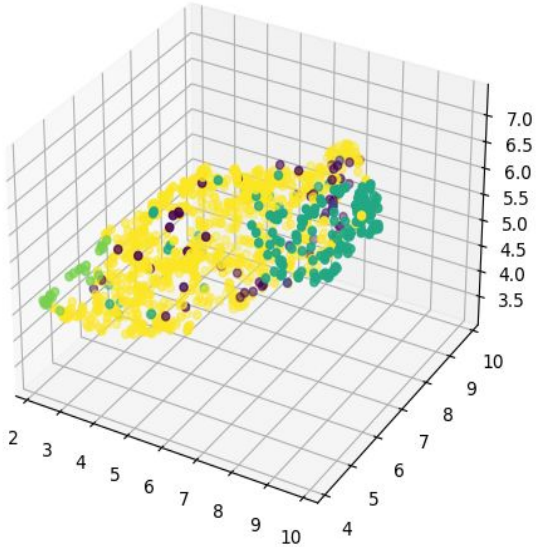


Corresponding manifold for activations of points in clusters 0, 2, 5

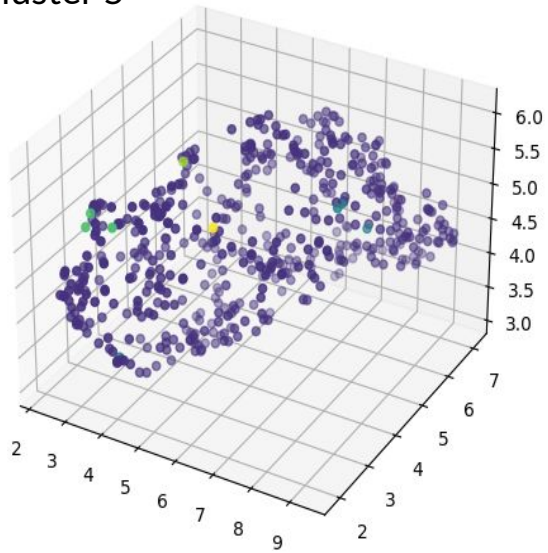
Cluster 0



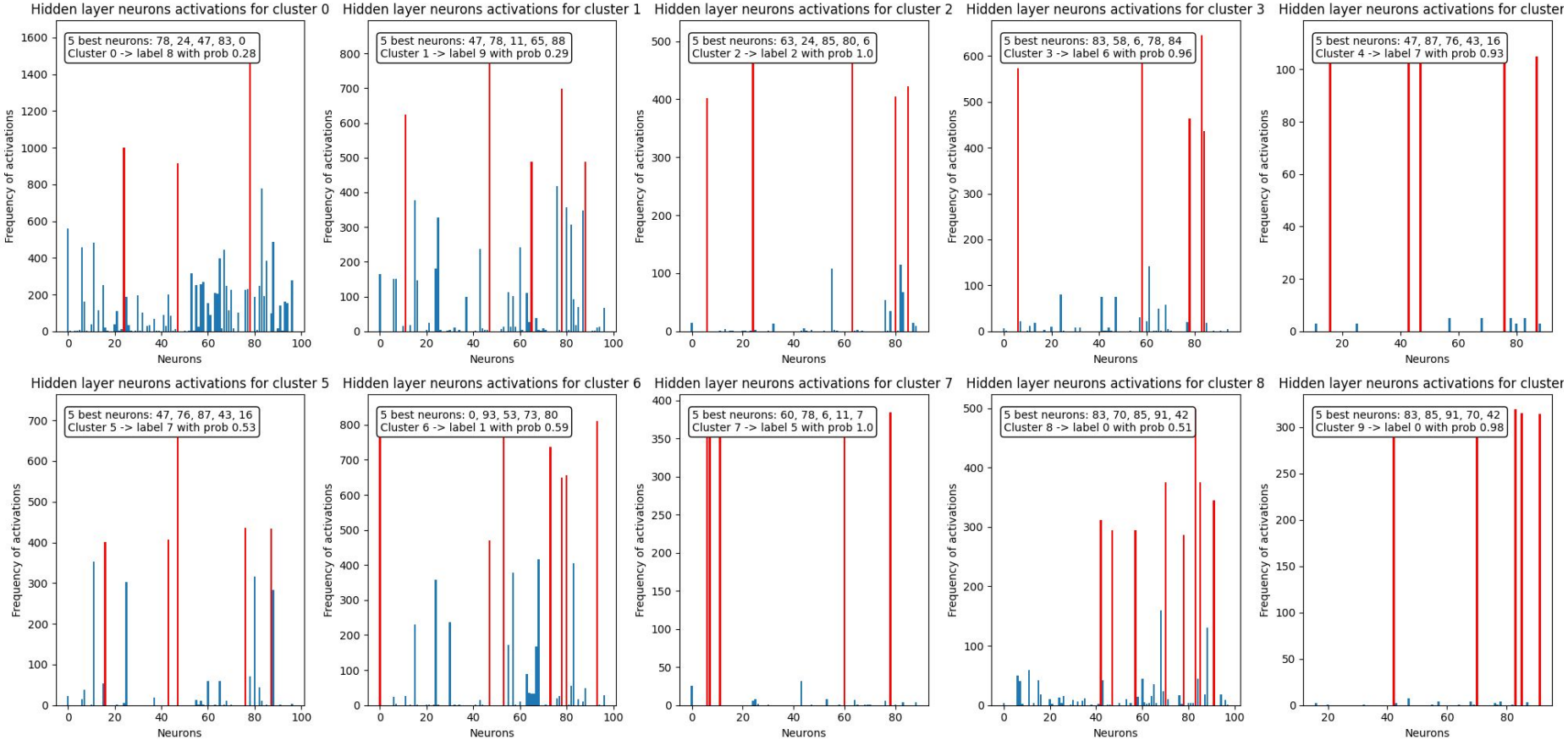
Cluster 2



Cluster 5

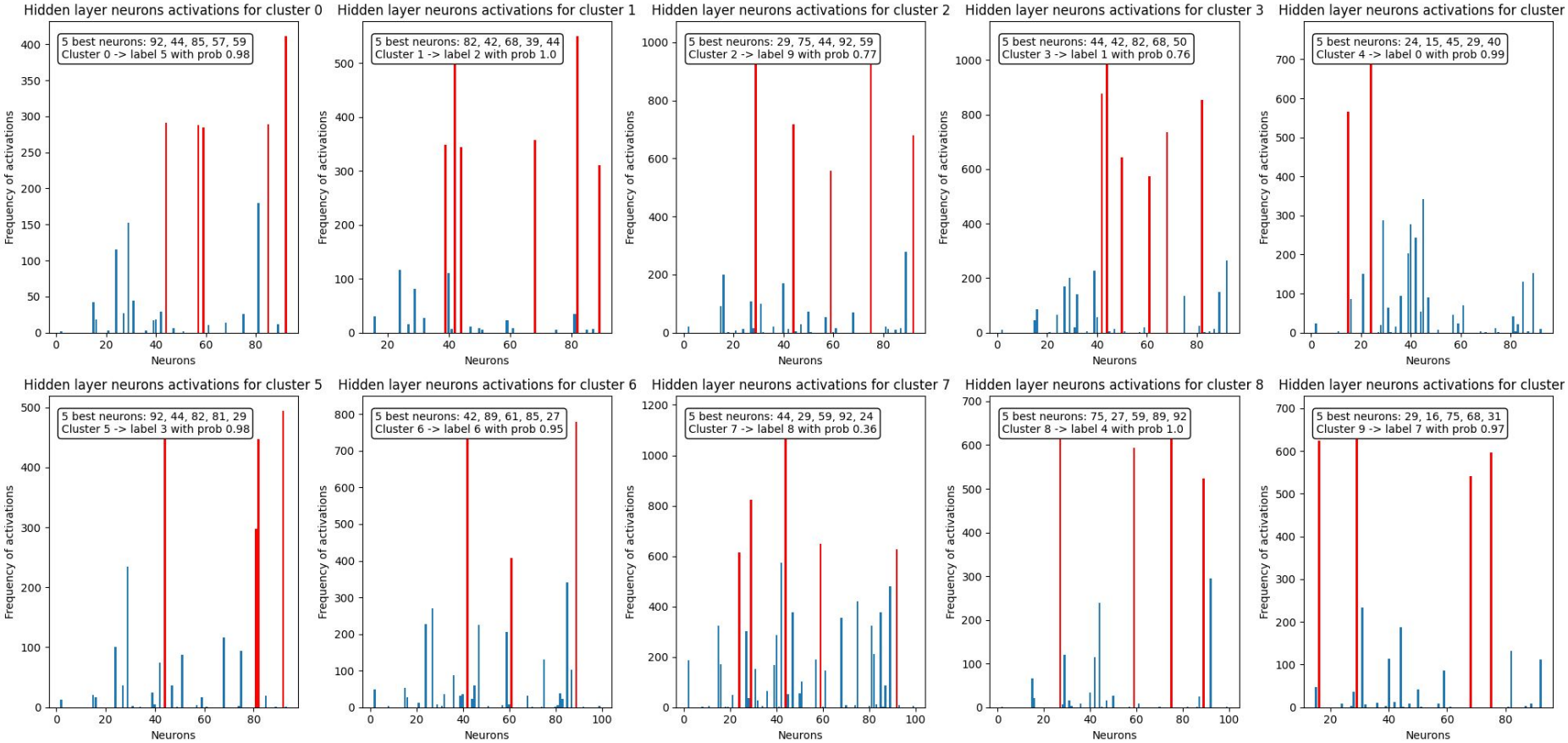


Barplot for mapper - **Non** regularized regime with 100 neurons



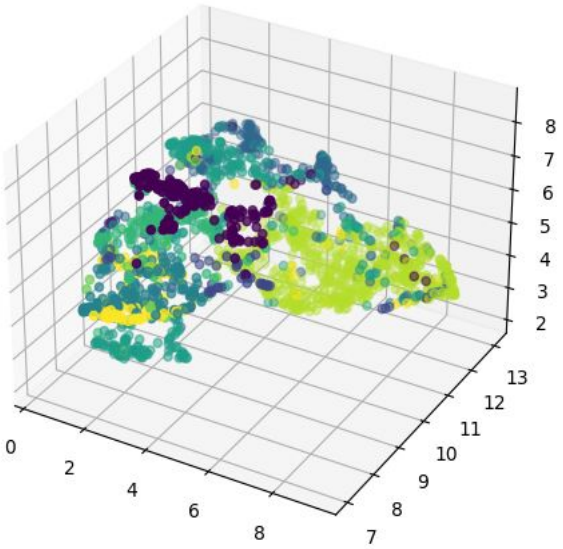
Barplot for mapper - Regularized regime with 100 neurons

Note: cluster 7 has noise and bad accuracy, clusters 8 and 9 are the more *localized*

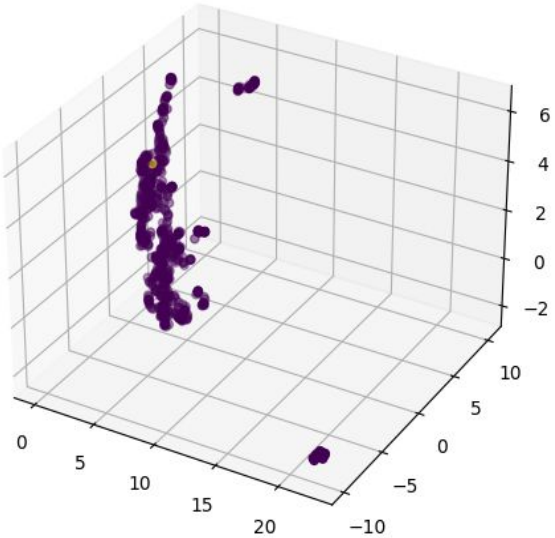


Corresponding manifold for activations of points in clusters 7, 8, 9

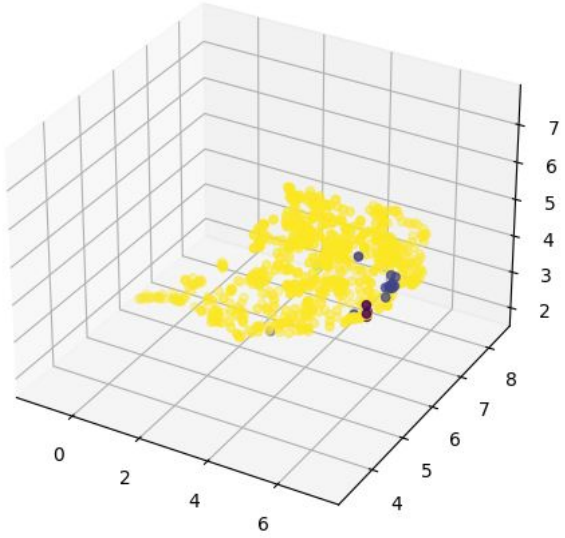
Cluster 7



Cluster 8

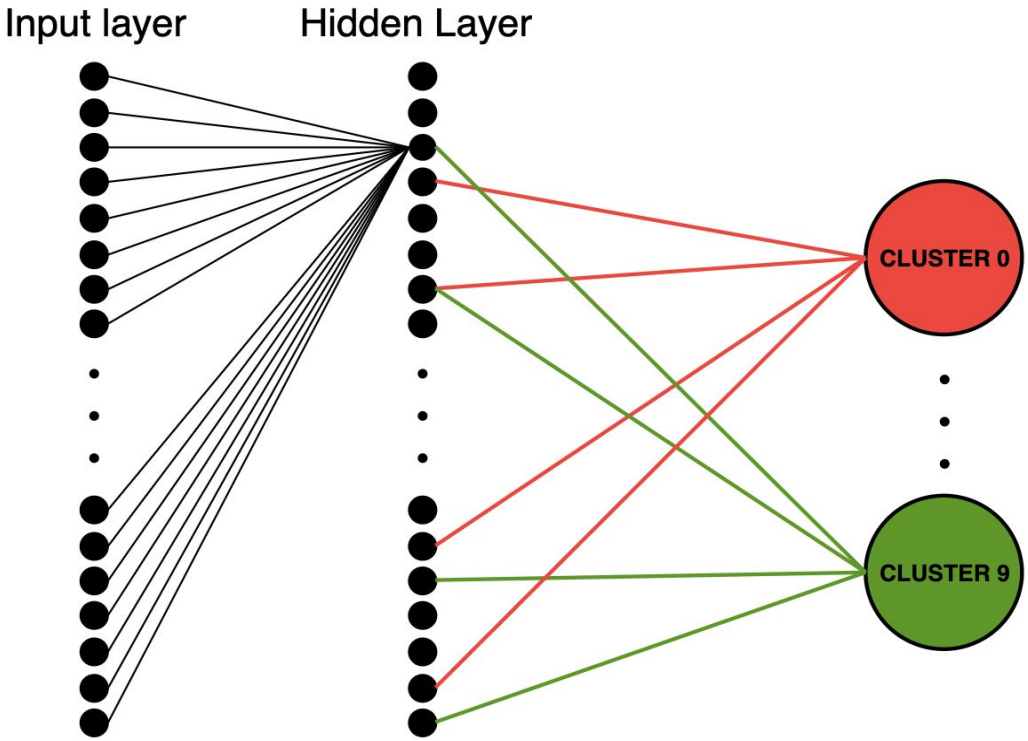


Cluster 9



Step by step procedure:

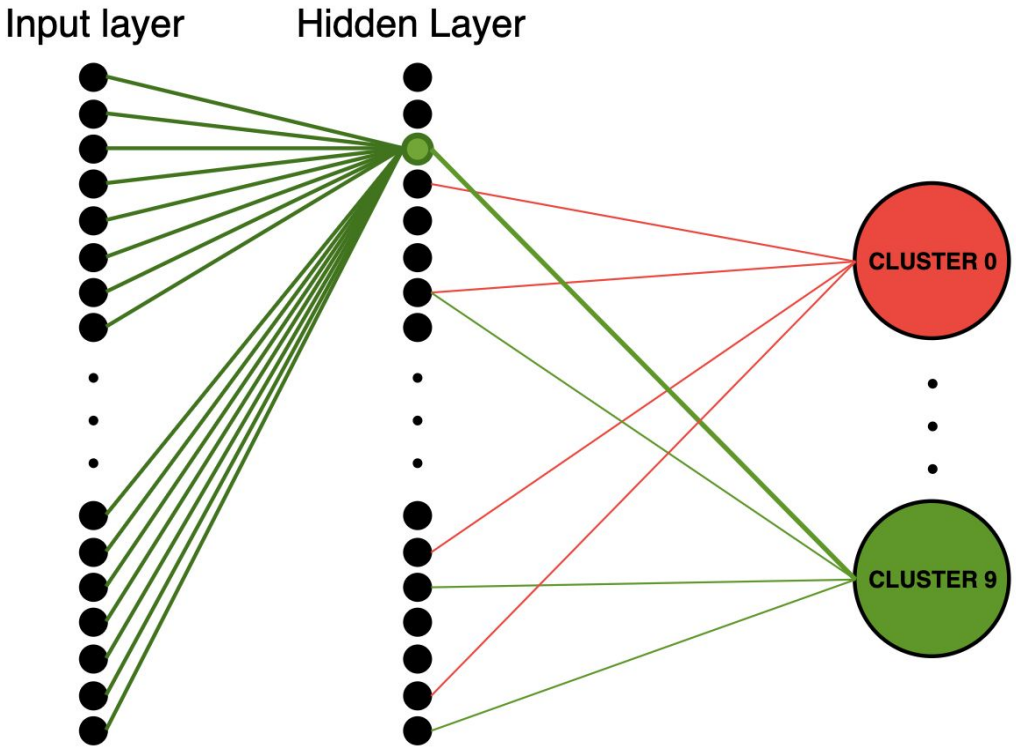
Knowing most activated neurons for each cluster, consider weights associated with those neurons.



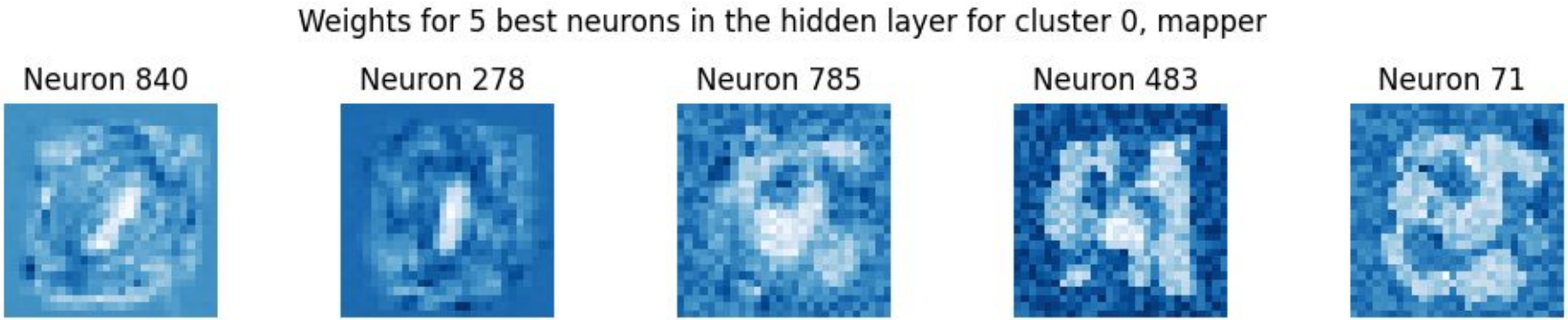
Step by step procedure:

Knowing most activated neurons for each cluster, consider weights associated with those neurons.

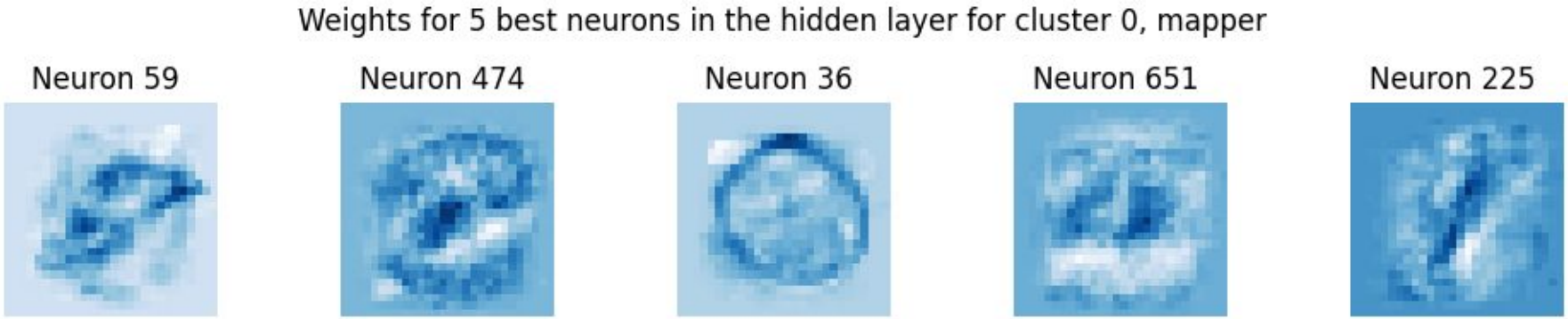
- Plot the weights associated with a particular neurons
- Combine together weights associated with most important neurons with respect to a cluster



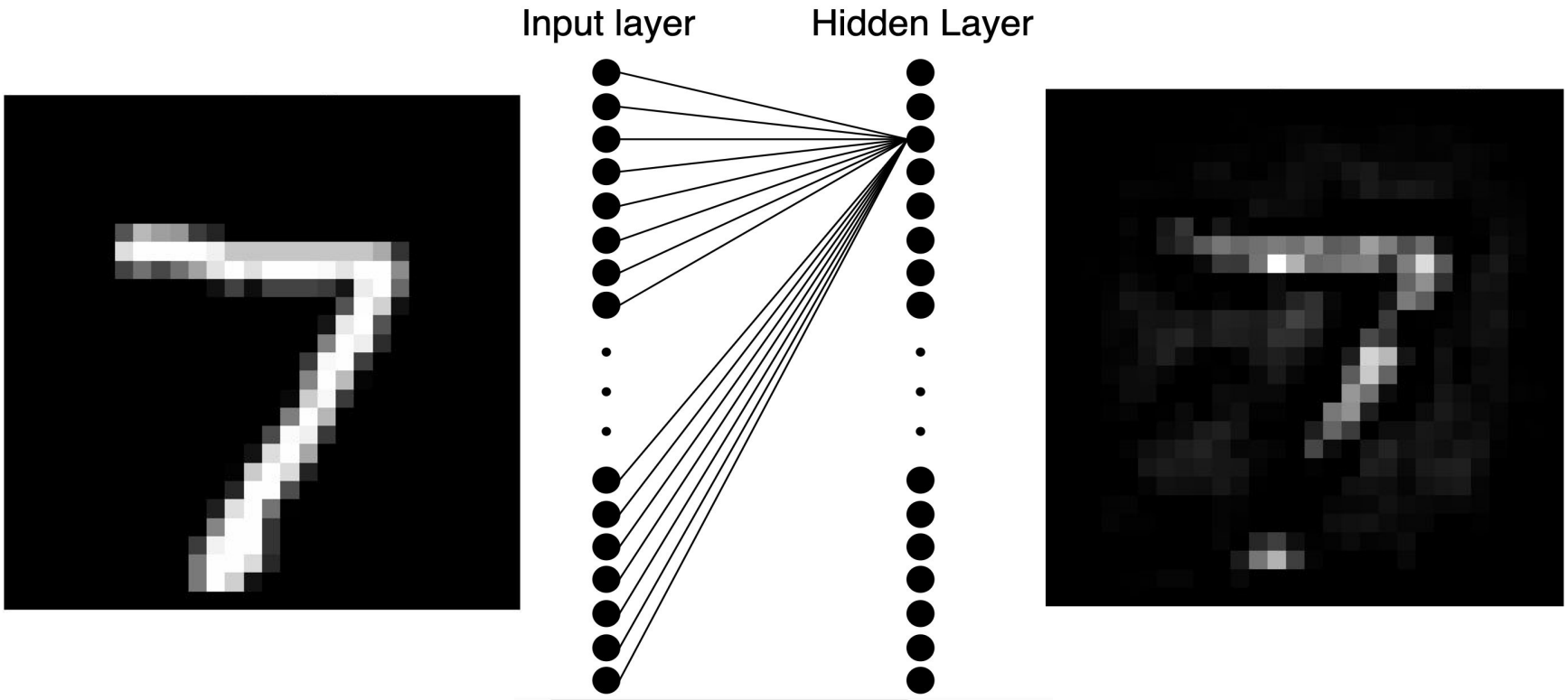
Not reg. regime with 1024 neurons

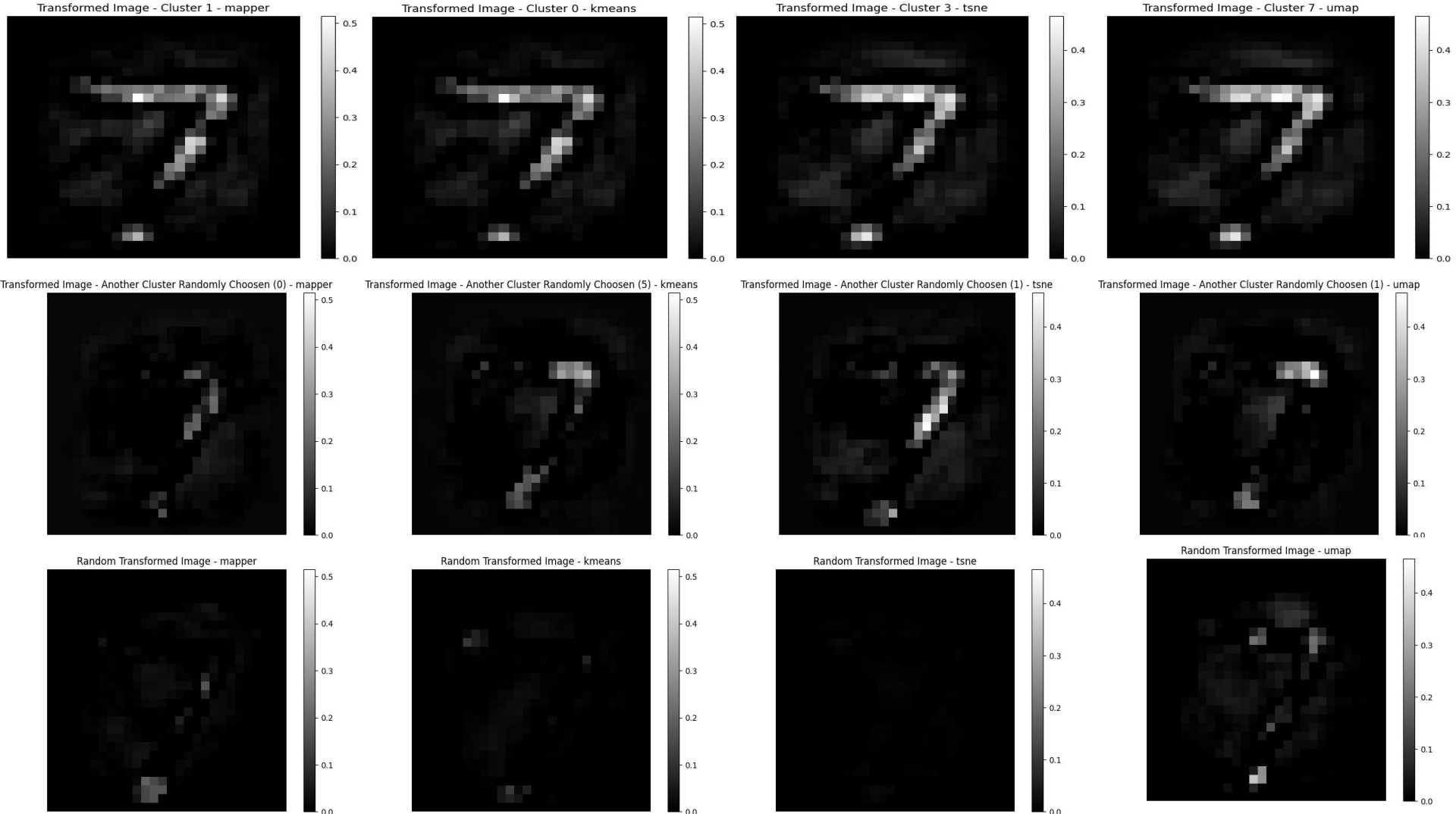


Reg. regime with 1024 neurons

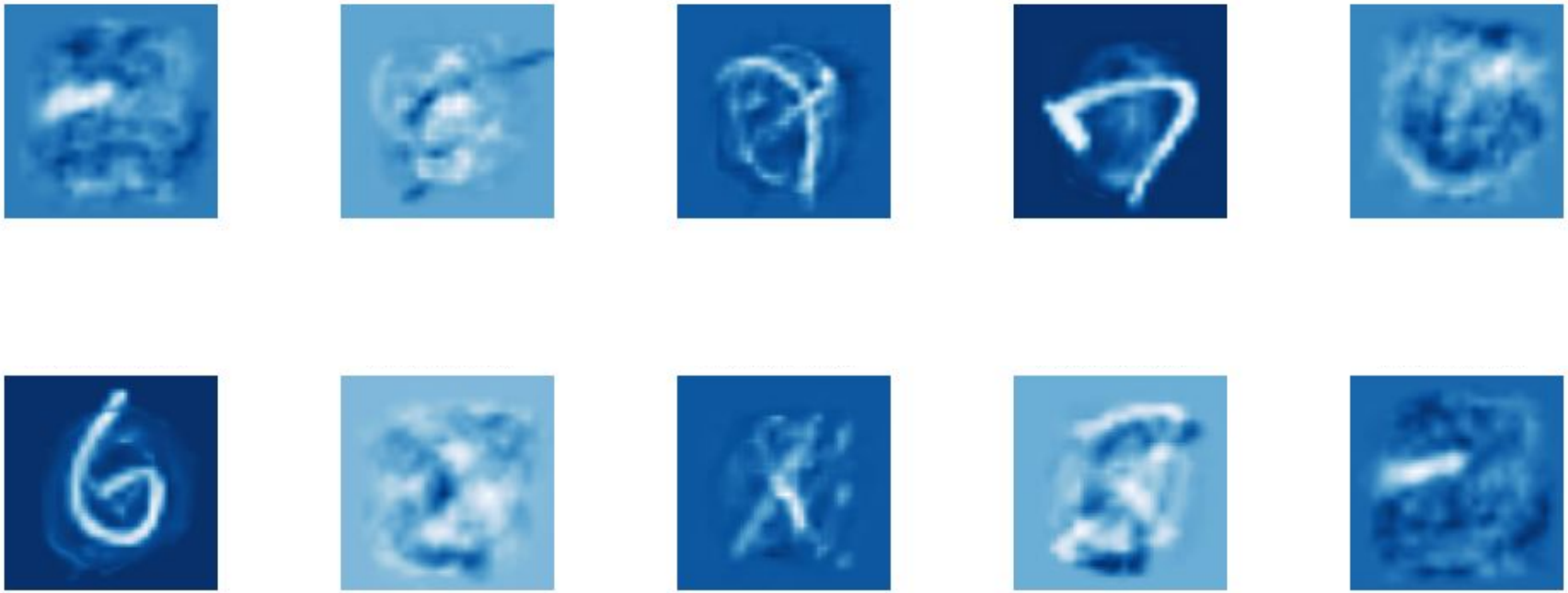


- How neurons associated with a certain cluster *see* an image?
- What happen with neurons belonging to other clusters or with randomly selected neurons?




Transformed Image - Cluster 1 - mapper
Transformed Image - Cluster 0 - kmeans
Transformed Image - Cluster 3 - tsne
Transformed Image - Cluster 7 - umap
Transformed Image - Another Cluster Randomly Choosen (0) - mapper
Transformed Image - Another Cluster Randomly Choosen (5) - kmeans
Transformed Image - Another Cluster Randomly Choosen (1) - tsne
Transformed Image - Another Cluster Randomly Choosen (1) - umap
Random Transformed Image - mapper
Random Transformed Image - kmeans
Random Transformed Image - tsne
Random Transformed Image - umap

Reshape of input images (64x64) - this is how new weights plots look like: *(just to give an insight)*



Further work

- Add hidden layers, study how informations *flow* through the layers using topological techniques
- Try other types of regularization (for ex. L1 regularization to manage the sparsity of neurons activations)
- Study correlations and statistics coming from our data
- Try other dataset (for ex. FashionMNIST)
- Try other architectures (for ex. CNN)
- Apply this analysis to other frameworks (for ex. PINA, from SISSA-MatLab)