MoA+: Mixture of Autoencoders with Varying Concentrations for Enhanced Image Clustering

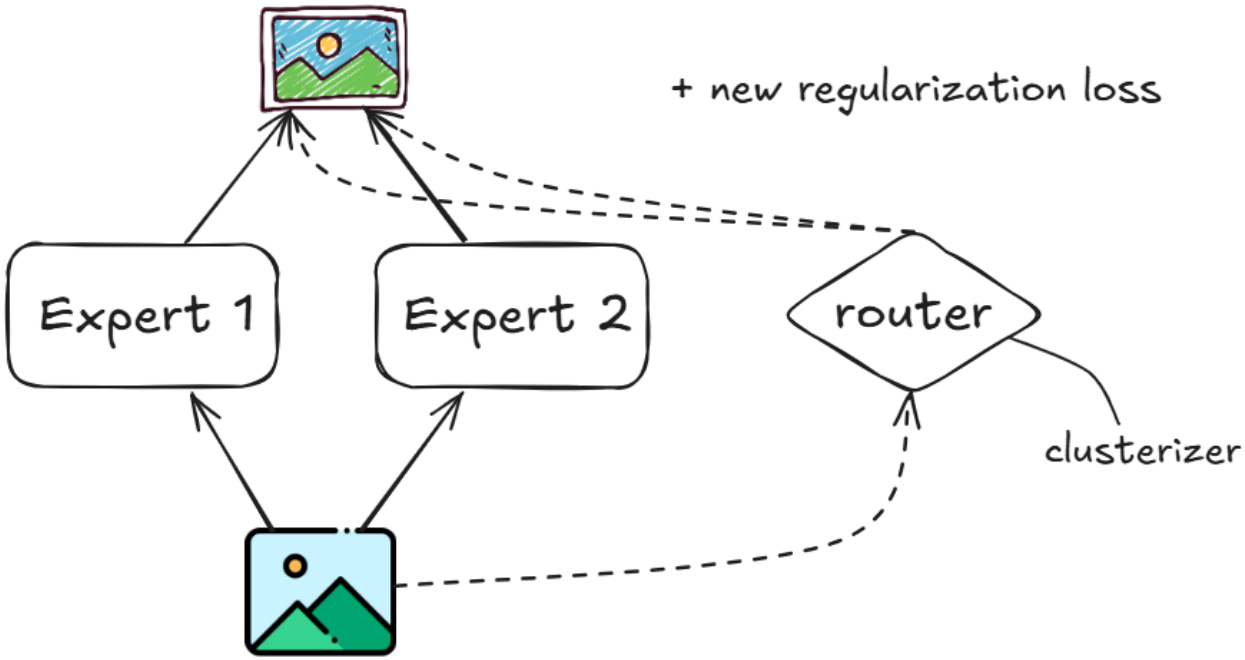
Vadym Tunik

Taras Shevchenko National University of Kyiv

May 23, 2025

MoA+: Core Concepts and Motivation

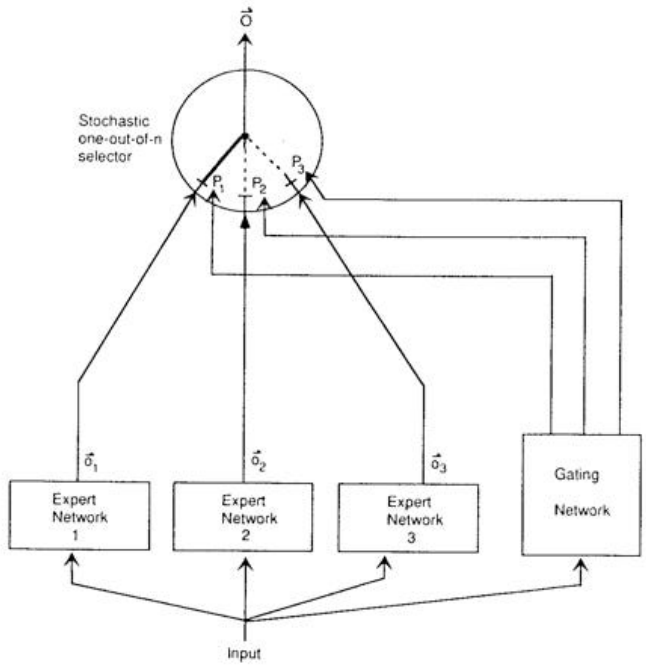
- **Contribution**: MoA+ (Mixture of Autoencoders with Varying Concentrations) for stable clustering by router.
- **Problem**: Standard Mixture of Experts (MoE) models suffer from "expert dominance" - imbalanced training.
- **Solution**: MoA+ is a synthesis of MoE architecture and Mixture models with Varying Concentrations (MVC).



Intro to Hinton's Mixture of Experts (MoE)

Mixture of Experts (MoE):

- **Definition**: MoE is a machine learning framework that combines multiple specialized submodels, or "experts," with a gating network.
- **Key Idea**: Experts specialize in specific data subsets; the gating network (router) selects the best expert(s) for each input.
- **Relevance**: Introduced for supervised learning by Jacobs, Jordan, Nowlan, and Hinton (1991), its principles are now extended to unsupervised tasks like image clustering, often by using autoencoders as experts.



Standard Mixture of Experts and Shortcomings

The standard MoE for unsupervised clustering involves experts as convolutional autoencoders reconstructing input images and a gating network assigning probabilities p_i^k. The loss function is typically:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - \sum_{k=1}^{M} p_i^k \hat{X}_{i,k})^2$$

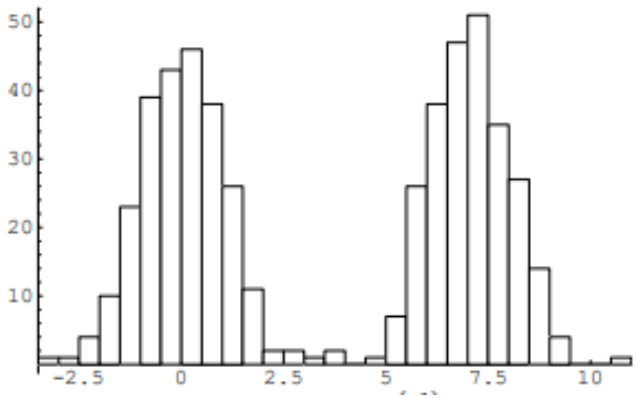
However, a significant shortcoming is expert dominance, where one expert may be assigned high probabilities for most data points, leading to imbalanced training. This can result in that expert overfitting while others are undertrained, causing poor clustering.

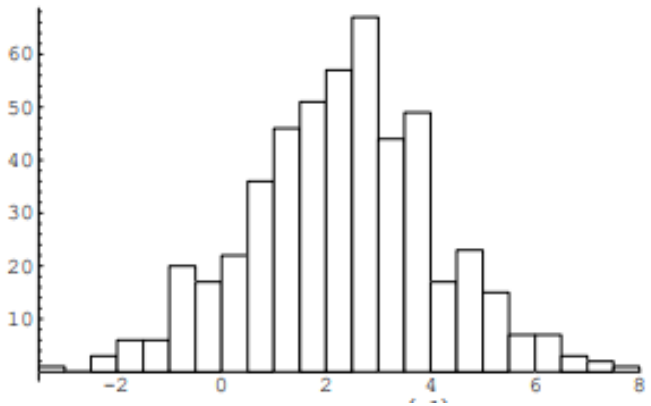
- **Shortcoming**: Expert dominance leads to imbalanced training and can hinder the discovery of meaningful clusters.
- **Result**: Ineffective separation of clusters, impacting performance metrics like NMI.

Intro to Mixture Model with Varying Concentrations

In the mixture model with varying concentrations, we consider the distribution as $P(\xi_i \in A) = \sum_{k=1}^{M} p_{i:n}^k F_k(A)$, where F_k are the component distributions, $p_{i:n}^k$ are the varying mixing probabilities, M is the number of components of the mixture.

- **Varying Mixing Probabilities**: Unlike standard models (e.g., GMMs) with fixed p, here $p_{i:n}^k$ varies per data point.
- **Minimax Weights**: Defined as $A_n = (P_n^T P_n)^{-1} P_n$ for the weighted empirical distribution function $\hat{F}_n^{(k)}(x, a) := \sum_{i=1}^{n} a_{i:n}^k I[\xi_{i:n} < x]$, adjusting p to balance contributions.





Synthesis of MoE and MM with Varying Concentrations

The synthesis in the MoA+ framework integrates the standard MoE architecture with the mixture model with varying concentrations by modifying the loss function. This synthesis aims to improve clustering performance by ensuring all experts are utilized effectively, particularly in unsupervised image clustering tasks.

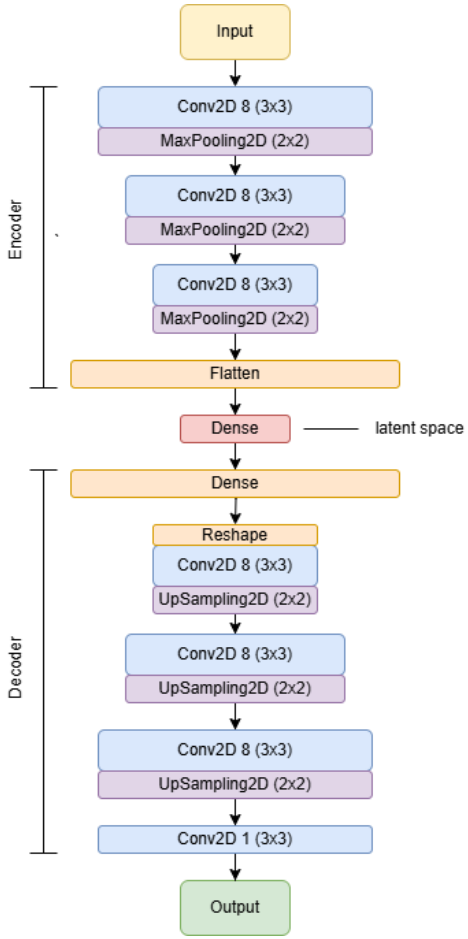
- **Approach**: Replaces direct gating probabilities $p^k_{i:n}$ in the loss calculation with minimax weights $a^k_{i:n}$ derived from these probabilities to weight the individual reconstruction error of each expert.
- **Goal**: Regularizes the gate, balances experts by addressing dominance, and improves clustering.

Modified Loss Definition

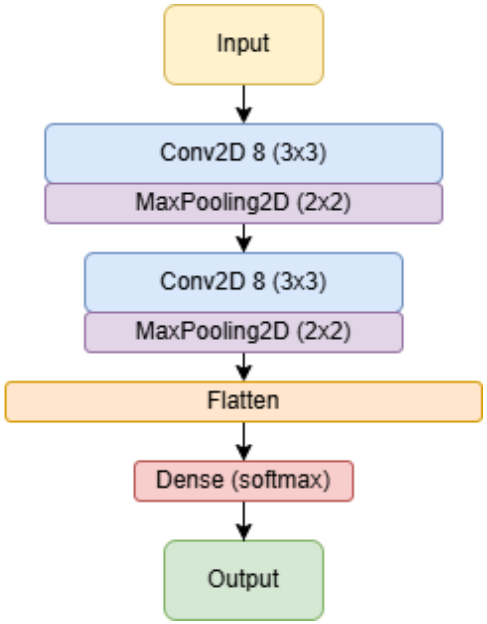
$$\text{loss}_{MoA+} = \left| \sum_{k=1}^{M} \sum_{i=1}^{n} a^k_{i:n} (X_i - \hat{X}_{i,k})^2 \right|$$

Experts and Router in MoA+

Experts: Convolutional autoencoders (CAEs), like U-Net/SegNet, reconstruct X_i to $\hat{X}_{i,k}$ with MSE error.



Router: Convolutional gate network assigns probabilities $P_n = (p^k_{i:n})$, clustering images by selecting the best CAE.



Modified Loss Function

- **Standard Loss**: $MSE_{standard} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \sum_{k=1}^{M} p_i^k \hat{X}_{i,k})^2$.
- **MoA+ Innovation**: Uses minimax weights $a_{i:n}^k = [(\Gamma_n)^{-1} P_n]_{i,k}$, with $\Gamma_n = P_n^T P_n$ to weight the individual reconstruction error of expert.
- **Loss Function**:

$$\text{loss}_{MoA+} = \left| \sum_{k=1}^{M} \sum_{i=1}^{n} a_{i:n}^k (X_i - \hat{X}_{i,k})^2 \right|$$

- **Upper Bound (used for optimization when $M = 2$)**:

$$\leq \sqrt{\sum_{k=1}^{M} \frac{1}{\lambda_k^2}} \sqrt{\sum_{k=1}^{M} (\sum_{i=1}^{n} p_{i:n}^k (X_i - \hat{X}_{i,k})^2)^2}$$

  where λ_k are eigenvalues of Γ_n. Remark: In general, this bounding is not true for $M > 2$, but it can still be used.
- **Bounding purpose**: Direct optimization of the $loss_{MoA+}$ is complex due to the matrix inversion $(P_n^T P_n)^{-1}$ involving router outputs, making gradient computation difficult for $M > 4$.

Experiments: Setups and Metrics

We compare two models: the classical Hinton's MoE and our MoA+. These models are similar by architecture, with the primary difference being the loss function used for training.

- **Datasets**:
  - MNIST: A standard benchmark of 70,000 grayscale handwritten digit images (28×28).
  - MAD (Military Activity Dataset): A specialized dataset of melspectrograms from audio recordings of military activities (e.g., firefights, communications), including from the Russo-Ukrainian War.
- **Task**: Binary clustering ($M = 2$) for both datasets.
- **Metrics**:
  - Normalized Mutual Information (NMI): Measures clustering alignment with true labels, ranging from 0 (no agreement) to 1 (perfect agreement). This is the primary metric for clustering quality.
  - Mean Squared Error (MSE): Measures reconstruction quality.
- **Hyperparameters**: Same configuration for both models in the experiments.

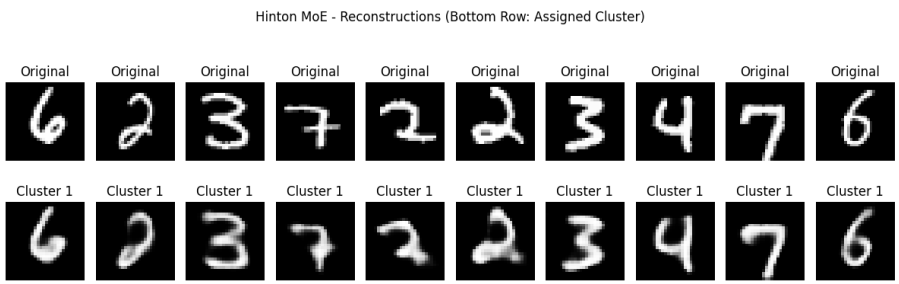
Experimental Results [MNIST] - Metrics & Reconstructions

Performance Metrics:

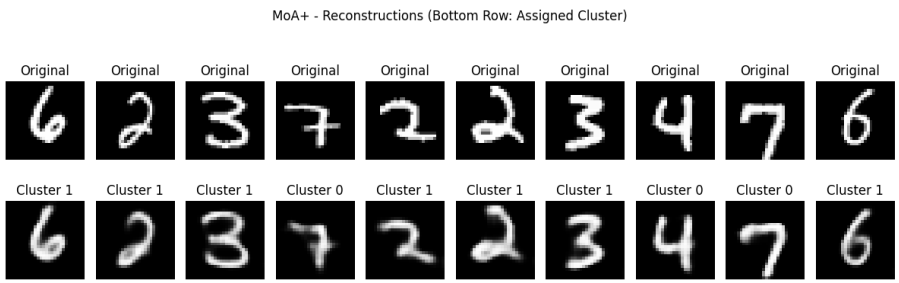
Model	MSE	NMI
Hinton MoE	0.013	0.000
MoA+	0.015	**0.276**

MoA+ significantly outperforms Hinton MoE in NMI, indicating better clustering, with comparable MSE. The slight increase in MSE for MoA+ is acceptable given the substantial improvement in NMI. This suggests MoA+ focuses more on balanced expert training for clustering rather than solely minimizing reconstruction error via expert dominance.

Reconstruction Examples (MNIST Digits):



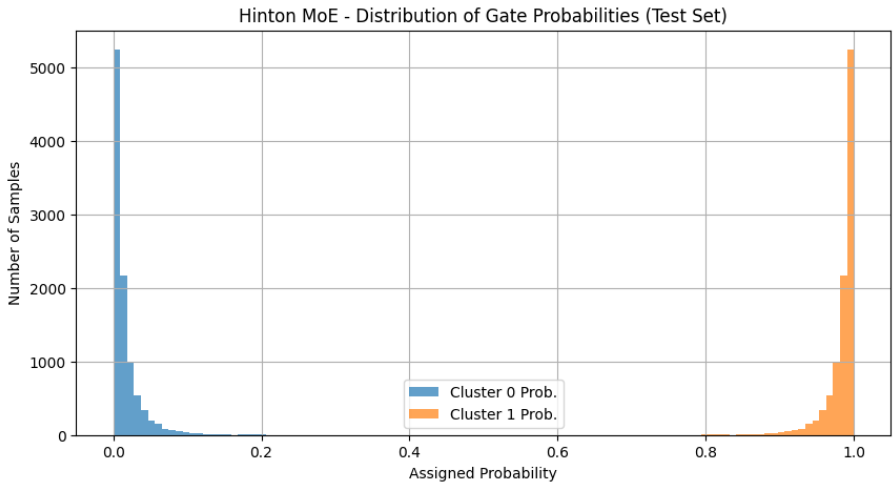
Hinton MoE Reconstructions.



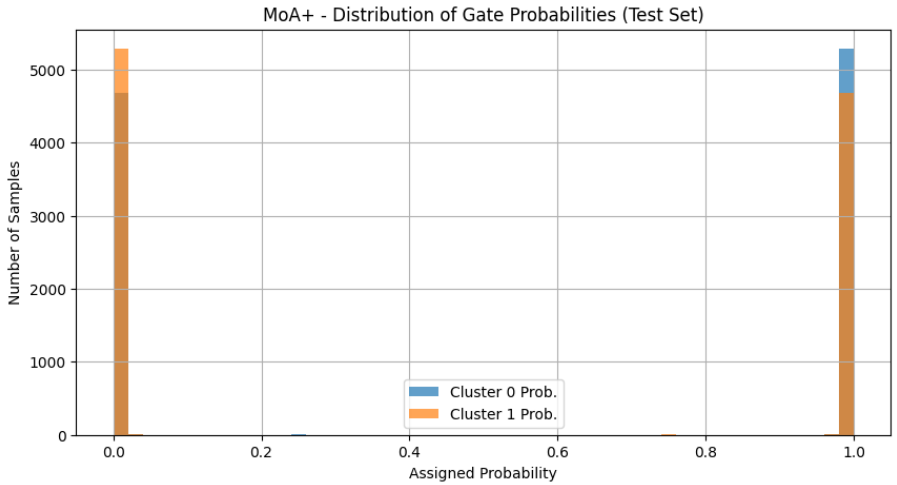
MoA+ Reconstructions.

Experimental Results [MNIST] - Probability Distributions

Probability Distribution Comparison:



Hinton MoE Gate Probabilities. Shows strong imbalance; one expert dominates.



MoA+ Gate Probabilities. Shows two clear peaks, indicating more balanced and confident assignment.

The Hinton MoE router heavily favors one expert, while MoA+ demonstrates more balanced and confident assignments to both experts, aligning with the improved NMI score.

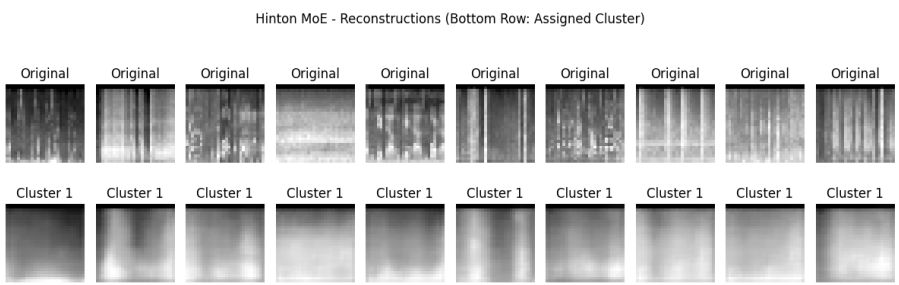
Experimental Results [MAD]

The Military Activity Dataset (MAD) consists of melspectrograms from audio recordings of military activities. A melspectrogram is a visual representation of the spectrum of frequencies in a sound signal, transformed to the mel scale, which mimics human auditory perception.

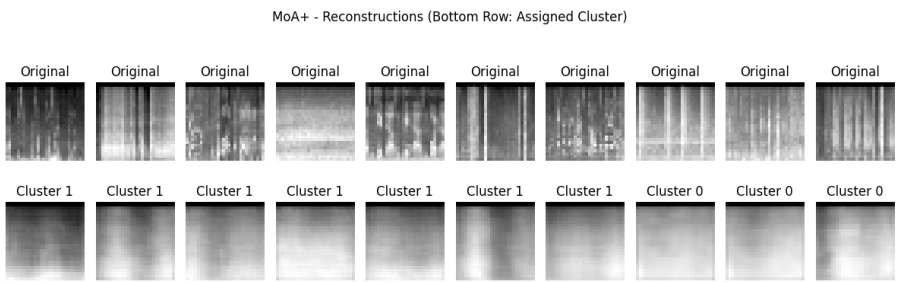
Performance Metrics:

Model	MSE	NMI
Hinton MoE	0.008	0.023
MoA+	0.010	**0.127**

Similar to MNIST, MoA+ shows significantly higher NMI on MAD, with comparable MSE. This highlights MoA+'s advantage on complex, noisy data.



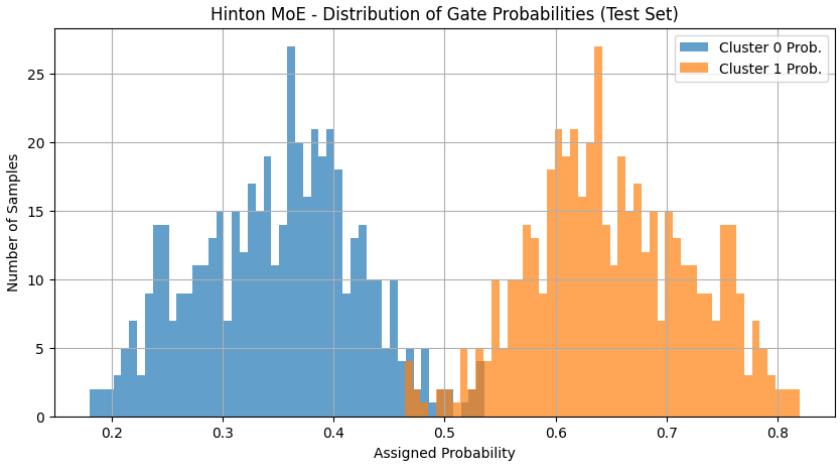
Hinton MoE Reconstructions.



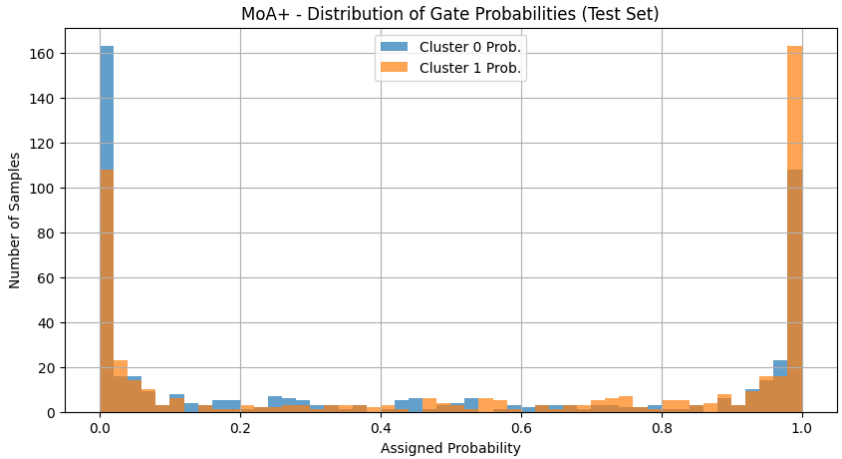
MoA+ Reconstructions.

Experimental Results [MAD] - Distributions

Probability Distributions Comparison:



Hinton MoE Gate Probabilities



MoA+ Gate Probabilities. Shows a tendency towards bimodal distribution, though less distinct than on MNIST.

Percentage Label Distributions Comparison:

TrueLabel	Cluster 0	Cluster 1
0	88.89	41.63
1	11.11	58.37

TrueLabel	Cluster 0	Cluster 1
0	19.23	59.86
1	80.77	40.14

True Labels: 0-Communication, 1-Firefights.

References

- R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton, "Adaptive Mixtures of Local Experts," *Neural Computation*, vol. 3, no. 1, pp. 79-87, 1991.
- R. Maiboroda, V. Miroshnychenko, & O. Sugakova, "Estimation of Concentrations Parameters in the Model of Mixture with Varying Concentrations," *Austrian Journal of Statistics*, 54(1), 1-16, 2025.
- B. Zoph et al., "ST-MoE: Designing Stable and Transferable Sparse Expert Models," arXiv:2202.08906, 2022.
- O. Mikushova, & V. Miroshnychenko, "Differentiable Bounds on Loss Functions in the Mixture Model with Varying Concentrations," *In Proceedings of the XXIII International Scientific Practical Conference «Shevchenkivska Vesna 2025»*, p. 33, Kyiv, Ukraine: Taras Shevchenko National University of Kyiv, 2025.