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

Vadym Tunik

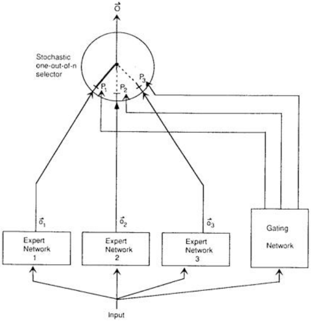
Taras Shevchenko National University of Kyiv

April 10, 2025

Intro to Hinton's Mixture of Experts (MoE)

Mixture of Experts (MoE):

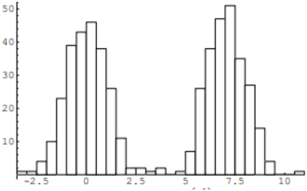
- **Definition**: MoE is a machine learning framework that combines multiple specialized submodels, or "experts," with a gating network.
- **Key Idea**: Experts specialize in specific data subsets; the gating network selects the best expert(s) for each input.
- **Relevance**: Introduced for supervised learning, it's now used for unsupervised tasks like image clustering.

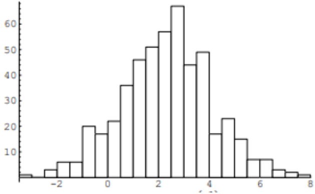


Intro to Mixture Model with Varying Concentrations

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 $\hat{F}_n^{(k)}(x, a) := \sum_{i=1}^{n} a_{i:n}^k I\left[\xi_{i:n} < x\right]$, 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 gating probabilities $p^k_{i:n}$ with minimax weights $a^k_{i:n}$ in the loss.
- **Goal**: Regularizes the gate, balances experts, and improves clustering by tackling dominance.

Modified Loss Definition

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

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:

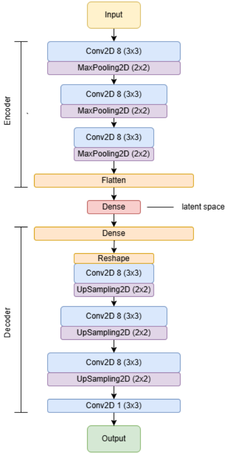
$$MSE = \frac{1}{n} \sum_{i=1}^{n} (X_i - \sum_{k=1}^{M} p_i^k \hat{X}_{i,k})^2$$

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.

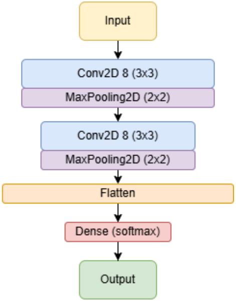
- **Shortcoming**: Expert dominance (ST-MoE, 2022) leads to imbalanced training.
- **Result**: Ineffective separation of clusters, impacting performance metrics like NMI.

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**: Mean Squared Error (MSE).
- **MoA+ Innovation**: Uses minimax weights $a^k_{i:n} = [(\Gamma_n)^{-1} P_n]_{i,k}$, with $\Gamma_n = P_n^T P_n$.
- **Loss Function**:

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

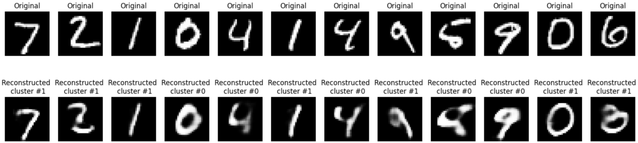
- **Upper Bound (only for M=2)**:

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

  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**: The initial loss is difficult to differentiate because of the parameters inside a and the reconstructed X.

Experimental Results: Setup and Metrics

Tested on MNIST (n grayscale digit images, $M = 2$).



Reconstructed MNIST images by Standard MoE.

- **Metric**: Normalized Mutual Information (NMI): Measures clustering alignment with true labels, ranging from 0 (no agreement) to 1 (perfect agreement).

Model	**NMI**	**Loss**
Standard MoA	0.08	0.018
MoA+	0.8	0.008

MoA+ significantly outperforms, showing balanced clustering.

References

- G. E. Hinton et al., ”Adaptive Mixtures of Local Experts,” *Neural Computation*, vol. 3, no. 1, pp. 79-87, 1991.
- V. Miroshnichenko, R. Maiboroda, ”Asymptotic normality of modified LS estimator,” *Modern Stochastics*, vol. 7, iss. 4, pp. 435-448, 2020.
- B. Zoph et al., ”ST-MoE: Designing Stable and Transferable Sparse Expert Models,” arXiv:2202.08906, 2022.