Exploring Expert Specialization through Unsupervised Training in Sparse Mixture of Experts

Strahinja Nikolic[1], Ilker Oguz[2], and Demetri Psaltis[3]
[1]School of Engineering (STI), [2]Laboratory of Applied Photonics Devices (LAPD), [3]Optics Laboratory
École Polytechnique Fédérale de Lausanne (EPFL)
{strahinja.nikolic, ilker.oguz, demetri.psaltis}@epfl.ch

Abstract

Understanding the internal organization of neural networks remains a fundamental challenge in deep learning interpretability. We address this challenge by exploring a novel Sparse Mixture of Experts Variational Autoencoder (SMoE-VAE) architecture. We test our model on the QuickDraw dataset, comparing unsupervised expert routing against a supervised baseline guided by ground-truth labels. Surprisingly, we find that unsupervised routing consistently achieves superior reconstruction performance. The experts learn to identify meaningful sub-categorical structures that often transcend human-defined class boundaries. Through t-SNE visualizations and reconstruction analysis, we investigate how MoE models uncover fundamental data structures that are more aligned with the model's objective than predefined labels. Furthermore, our study on the impact of dataset size provides insights into the trade-offs between data quantity and expert specialization, offering guidance for designing efficient MoE architectures.

1 Introduction

Mixture of Experts (MoE) architectures, first introduced by Jacobs et al. [1], decompose computation across specialized sub-networks, or "experts." This paradigm has recently achieved remarkable success in scaling deep learning models to unprecedented sizes, enabling state-of-the-art performance in areas like large language modeling [2, 3, 5]. The promise of MoE models extends beyond computational efficiency: the explicit routing of inputs to different experts can offer a natural window to help us understand how neural networks organize and process data.

However, as MoE models become more powerful and complex, understanding what each expert has learned or how routing decisions relate to meaningful data structure remains a fundamental challenge [5, 6, 8]. This opacity is particularly concerning as these systems are increasingly deployed in critical applications where understanding model behavior is essential. This gap between performance and interpretability points to the need for new methods to analyze the internal mechanisms of MoE models.

Recent theoretical advances have begun to illuminate the fundamental mechanisms underlying expert specialization. The work by [6] demonstrates that MoE architectures possess an inherent capacity to train experts specializing in different data clusters when trained with gradient descent. Their analysis reveals that this specialization emerges automatically from the optimization dynamics, suggesting that expert assignment patterns are likely to reflect fundamental properties of the data distribution.

Yet, such analysis remain sparse in the literature. For instance, a recent comprehensive survey [5] argues that to ensure MoE models are transparent and trustworthy, the field urgently needs new methods to visualize and explain what individual experts learn and how they interact. Improving

Preprint. Under review.

this aspect of interpretability is considered critical for bolstering our understanding and advancing the responsible development of MoE architectures.

To address this challenge, we combine Sparse Mixture of Experts with Variational Autoencoders (VAEs), creating an architecture tailored for the direct analysis of expert specialization. Our approach is partially inspired by MoE-Sim-VAE [7], which demonstrated the potential of MoE-VAEs for clustering. However, we repurpose this architecture with a distinct goal: to analyze the phenomenon of expert specialization itself and the mechanisms that drive it. The architecture's inherent ability to discover underlying data structures makes it an ideal tool for this investigation.

Our investigation reveals a surprising finding: when experts are allowed to specialize according to the natural structure present in data (unsupervised routing), they form clusters that are more effective for the model's objective than when grouped by human-provided labels (supervised routing). As we will demonstrate, these data-driven clusters are closer to being linearly separable and result in superior reconstruction performance. This suggests that expert specialization in MoE architectures can serve as a powerful tool for data structure discovery, uncovering organizational principles that differ from conventional categorizations.

Contributions. We make the following key contributions:

- We introduce SMoE-VAE, a sparse Mixture-of-Experts Variational Autoencoder tailored for interpretability and analysis of expert specialization. The architecture uses a shared high-capacity encoder with lightweight decoder experts, a MLP as a gating unit which receives latent space vectors as input, and is trained with entropy and batch-level load balancing loss.
- To the best of our knowledge, we present the first controlled comparison between supervised (ground-truth-label–guided) and unsupervised routing. Unsupervised routing consistently achieves lower reconstruction loss and discovers clusters in the latent space that are closer to being linearly separable compared to supervised training with a labeled database.
- We provide an analysis framework that explains *why* unsupervised training prevails: (i) latent-space visualizations (t-SNE) contrasting expert assignments and class labels, (ii) linear probes that quantify separability of expert vs. class partitions, and (iii) qualitative visualization of reconstructions revealing cross-class and intra-class specializations.
- We study the joint effect of the number of experts and the number of samples per expert on performance. We show that gains cannot be attributed to data quantity alone; the structure of the data each expert sees and the resulting degree of specialization are critical, with over-fragmentation and under-fragmentation both leading to reduced performance.

2 Related Work

2.1 Mixture of Experts and Interpretability

Mixture of Experts (MoE) architectures have emerged as a powerful paradigm for scaling neural networks while maintaining computational efficiency. Originally introduced by Jacobs et al. [1], MoE models partition the input space among specialized experts, with a gating network determining the routing of inputs to appropriate experts. Recent advances have demonstrated the effectiveness of sparse MoE in large language models [2, 3], where only a subset of experts are activated for each input.

However, as highlighted in recent surveys, the interpretability of MoE models remains a significant challenge. The inherent complexity of MoE models, coupled with their dynamic gating of inputs to specialized experts, poses substantial obstacles to understanding their decision-making processes [5]. This interpretability gap becomes particularly problematic in applications where comprehending the rationale behind model decisions is essential. Current approaches to MoE interpretability focus primarily on analyzing expert utilization patterns and load balancing, but provide limited insight into what each expert has actually learned to specialize in.

From a theoretical perspective, recent work by [6] provides crucial insights into the mechanisms of expert specialization. They address why experts diversify rather than collapse into a single unit and

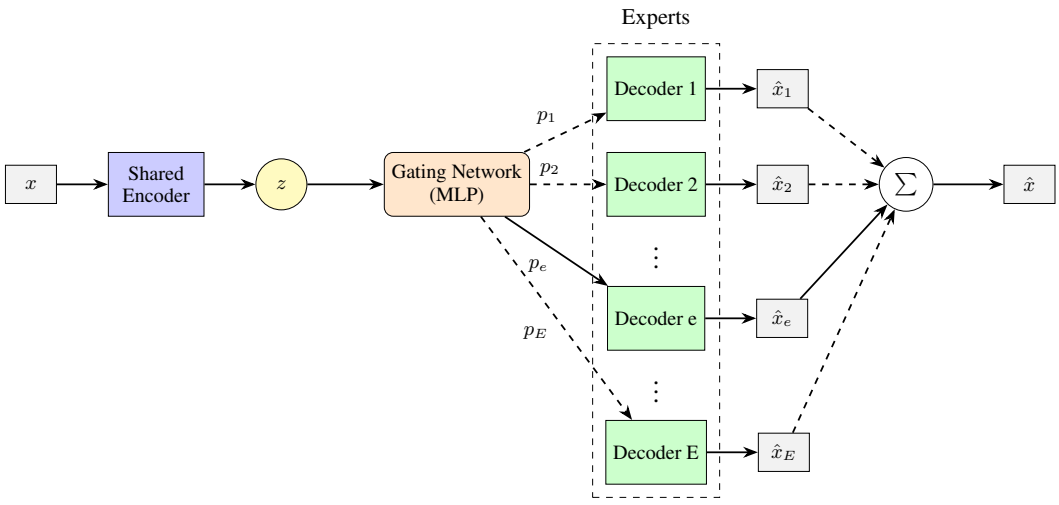


Figure 1: The SMoE-VAE architecture. A shared encoder maps the input x to a latent representation z. The gating network receives z and routes it to specialized decoders. During training all decoders are activated and the final output is the sum of all outputs weighted by gating network probabilities, while during inference only one expert is activated to produce $\hat{x}$.

how the router learns to dispatch data. Their work suggests that experts specialize to exploit underlying cluster structures in the data, and the router learns to identify cluster-specific features. However, their analysis relies on synthetic datasets with linearly separable clusters and uses a linear gating mechanism. This leaves open questions about specialization in more complex, high-dimensional settings with non-linear structures, which our work explores.

One paper that touches on this topic is work from C.Riquelme et al. [16] where they used sparsely gated MoE for Computer Vision. By having access to a huge dataset ($\sim$305M) of labeled images they were able to show that deeper routing decisions correlate better with image classes. This indicates that even in large networks, with multiple MoE layers, the routers are able to partition the input space in a way that correlates well with human labels.

An important question in MoE design concerns the choice of routing strategy. While prior work has shown that learned routing outperforms fixed strategies [8], a deeper understanding of the resulting specializations is needed. The availability of labeled data provides a unique opportunity to analyze these unsupervised, data-driven clusters by comparing them to a supervised baseline where routing is determined by ground-truth categories. Such a comparison, serves as a powerful analytical tool. It allows us to rigorously evaluate the clusters discovered by unsupervised methods against human-defined categories, offering insights into why and how unsupervised expert specialization can be so effective.

2.2 MoE combined with Variational Autoencoders for Data Structure Analysis

Variational Autoencoders [9] provide an ideal testbed for our investigation, as their explicit latent space representation and generative capabilities enable direct visualization of how experts partition and specialize on different aspects of the data distribution. Kopf A. et al. introduced MoE-Sim-VAE [7] which has demonstrated the potential of combining MoE architectures with VAEs for clustering purposes. Their approach places multiple decoder experts after the latent bottleneck and encourages expert specialization through similarity-based constraints. This work establishes that expert specialization in generative models can serve as an effective clustering mechanism. Their approach requires external guidance, such as pre-computed similarity matrices or explicit clustering objectives, to achieve meaningful expert assignment.

One approach related to ours is MIXAE [14], a framework that uses a mixture of autoencoders for unsupervised clustering. In contrast to our architecture, where only the decoders are experts, MIXAE treats entire autoencoders as experts. This design increases the complexity of the gating unit, which must process concatenated latent codes from all autoencoders. Similar to MoE-Sim-

VAE, the primary focus of MIXAE is on clustering performance rather than on the interpretability of the MoE architecture itself, leaving the inner workings of expert specialization largely unexplored.

Another significant contribution, which differs from the previous two mentioned, is the Variational Mixture-of-Experts Autoencoders (MMVAE) [4], which explores multi-modal deep generative models using MoE architectures. The MMVAE approach demonstrates how expert specialization can emerge across different data modalities, providing insights into how MoE architectures can naturally partition complex, multi-modal data distributions. While their focus is on cross-modal generation and representation learning, their work establishes important principles about how expert assignment can reveal underlying data structure.

Our work is motivated by the hypothesis that when properly designed, SMoE-VAE architectures can serve as powerful tools for data structure analysis, discovering clusters and specializations that may be more fundamental to the data distribution than human-imposed labels. Rather than optimizing for generation quality alone, we focus on understanding what experts learn when given the freedom to specialize according to the structure present in the data. The visual nature of image data makes this approach particularly interesting, as expert specializations can be directly assessed through reconstructed images, enabling immediate feedback about what each expert has learned to generate.

3 Method

3.1 SMoE-VAE Architecture

Our approach combines Variational Autoencoders with Sparse Mixture of Experts to enable interpretable analysis of expert specialization patterns. The architecture consists of three components: a shared convolutional encoder, a gating network, and multiple decoder experts that specialize on different aspects of the data distribution.

3.1.1 Encoder-Decoder Design

A key architectural feature is the asymmetric capacity allocation between encoder and decoder components. We employ a single, high-capacity convolutional encoder to process input images, ensuring effective feature extraction and meaningful latent space clustering. This design choice is crucial since all expert routing decisions depend on the quality of latent representations produced by the shared encoder.

In contrast, we use smaller, specialized decoder experts that focus on reconstructing specific data patterns. This asymmetry serves two purposes: (1) it concentrates representational capacity where it is most needed for clustering, and (2) it encourages experts to specialize on distinct reconstruction tasks rather than learning broad representations.

3.1.2 Gating Network

The gating network operates on latent representations $z \in \mathbb{R}^d$ to produce expert selection probabilities. It consists of a three-layer fully connected multilayer perceptron:

$$\begin{aligned} h_1 &= \text{ReLU}(W_1 z + b_1) \\ h_2 &= \text{ReLU}(W_2 h_1 + b_2) \\ \text{logits} &= W_3 h_2 + b_3 \end{aligned}$$

where $W_1 \in \mathbb{R}^{64\times d}$, $W_2 \in \mathbb{R}^{32\times 64}$, and $W_3 \in \mathbb{R}^{E\times 32}$ with E being the number of experts. Note that, since the gating network operates on the latent space the additional computational cost of this network is minimal compared to the rest of the network.

3.1.3 Soft vs Hard Gating Strategy

To bridge the gap between training and inference, we employ a dual gating strategy. During training, we use soft gating where expert outputs are weighted by softmax probabilities:

$$p_e = \frac{\exp(\text{logits}_e)}{\sum_{i=1}^{E} \exp(\text{logits}_i)}$$
$$\hat{x} = \sum_{e=1}^{E} p_e \cdot \text{Decoder}_e(z)$$

During inference, we switch to hard gating for efficiency and interpretability:

$$e^* = \arg\max_e \text{logits}_e$$
$$\hat{x} = \text{Decoder}_{e^*}(z)$$

This approach ensures that the model learns to make confident expert selections while maintaining differentiability during training.

3.2 Loss Function Design

Our loss function combines the standard VAE objective [9] with novel regularization terms that encourage both load balancing and decisive expert selection:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{KL}} + \alpha \mathcal{L}_{\text{gating}}$$

where $\mathcal{L}_{\text{recon}}$ is the Mean Squared Error (MSE) reconstruction loss, $\mathcal{L}_{\text{KL}}$ is the KL divergence regularization [9], and $\mathcal{L}_{\text{gating}}$ combines load balancing and entropy regularization.

3.2.1 Load Balancing Loss

To prevent expert collapse and ensure diverse specialization, we introduce a load balancing term that encourages uniform expert utilization across batches:

$$\bar{p}_e = \frac{1}{N} \sum_{n=1}^{N} p_{n,e}$$
$$\mathcal{L}_{\text{balance}} = E \cdot \text{MSE}(\bar{p}, \mathbf{u})$$

where $\bar{p} = [\bar{p}_1, \ldots, \bar{p}_E]$ represents average expert probabilities across the batch, p_n are softmax probabilities for sample with index n, $\mathbf{u} = [1/E, \ldots, 1/E]$ is the uniform distribution, and N is the batch size. The scaling factor E ensures consistent regularization strength across different numbers of experts.

3.2.2 Entropy Regularization

To encourage sharp expert selections and minimize the train-inference gap, we minimize the entropy of per-sample expert distributions:

$$\mathcal{L}_{\text{entropy}} = \frac{1}{N} \sum_{n=1}^{N} H(p_n)$$
$$H(p_n) = -\sum_{e=1}^{E} p_{n,e} \log(p_{n,e} + \epsilon)$$

where $\epsilon = 10^{-8}$ provides numerical stability. Minimizing entropy encourages the model to produce near-Dirac delta distributions, where one expert receives most of the probability mass. This ensures that soft gating during training closely approximates hard gating during inference.

The combined gating loss is:

$$\mathcal{L}_{\text{gating}} = \lambda_{\text{balance}}\mathcal{L}_{\text{balance}} + \lambda_{\text{entropy}}\mathcal{L}_{\text{entropy}}$$

4 Experimental Setup

4.1 Choice of Dataset and Preprocessing

We conduct our experiments on the QuickDraw dataset [10], a large-scale collection of hand-drawn sketches created by users worldwide. QuickDraw is particularly well-suited for our interpretability analysis for several key reasons. First, it provides abundant data with tens of thousands of samples per category, enabling robust training of expert networks and reliable statistical analysis. Second, the availability of ground truth category labels allows for direct comparison between supervised and unsupervised expert routing approaches. Third, and most importantly for our research, QuickDraw contains natural imperfections and variations that enable meaningful sub-clustering within categories: sketches sometimes exhibit ambiguous category membership (e.g., simplified cat faces that resemble generic faces), and the same object can be drawn in multiple styles (e.g., cats drawn as face-only sketches versus full-body representations). These characteristics allow unsupervised expert routing to discover subclusters that can be more reconstruction-relevant than rigid categorical boundaries.

From the full dataset, we select five diverse categories that exhibit varying levels of visual complexity and inter-category similarity: *face*, *eye*, *cat*, *snowflake*, and *pencil*. This selection provides a balanced mix of organic shapes (faces, eyes, cats), geometric patterns (snowflakes), and linear objects (pencils), enabling comprehensive analysis of expert specialization across different visual structures.

All sketches are preprocessed by converting vector drawings to 28×28 grayscale images, ensuring computational efficiency while preserving essential visual features. Unless otherwise specified, we use 70,000 samples per category, providing a dataset of 350,000 total images for training and evaluation. In order to evaluate the impact of the database, we systematically vary the number of samples from 5% to 100% of the full dataset while maintaining balanced representation across all categories.

4.2 Model Configuration and Training Protocol

Our implementation uses PyTorch [13] and employs convolutional neural networks for both encoder and decoder components. The shared encoder processes 28×28 input images and maps them to a 32-dimensional latent space, while individual decoder experts reconstruct images from the latent representations.

Training is conducted for 20 epochs across all experiments to ensure fair comparison while avoiding overfitting. We use the Adam optimizer [12] with a learning rate of 10^{-4}. The loss function hyperparameters are set as follows: $\beta = 0.1$ for KL divergence regularization, $\lambda_{\text{balance}} = 200$ for load balancing, and $\lambda_{\text{entropy}} = 400$ for entropy regularization. These values were chosen to balance reconstruction quality with effective expert specialization and load balancing.

For supervised baseline comparisons, we train the gating network to route samples based on ground truth category labels rather than learned latent representations, while maintaining identical model capacity and training procedures.

For dataset size impact studies, we report optimal expert counts and corresponding reconstruction losses across different data regimes. All results are averaged over multiple random seeds to ensure statistical reliability, with error bars representing standard deviation across runs.

5 Results and Analysis

5.1 Unsupervised vs Supervised Expert Routing

Our primary finding demonstrates that unsupervised expert routing consistently outperforms supervised routing based on human-provided labels. Figure 2 presents test reconstruction loss as a function of the number of active experts, comparing our unsupervised SMoE-VAE approach (blue

line with error bars) against a supervised baseline where expert assignment is determined by ground truth class labels.

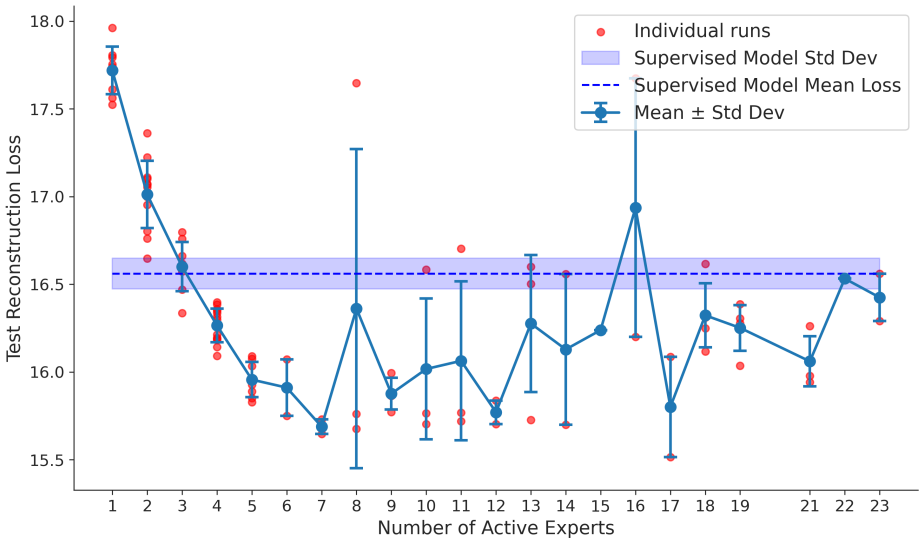


Figure 2: Test reconstruction loss: unsupervised expert routing (blue) vs. supervised routing based on ground truth labels (purple dashed). Unsupervised peaks around 7 experts and outperforms the supervised baseline constrained to 5 experts.

The results reveal several key insights. First, the unsupervised approach achieves significantly lower reconstruction loss across most expert configurations, with the optimal performance around 7 experts reaching a MSE loss on test set near 15.7, better than the supervised approach's 16.6. The supervised model, constrained to exactly 5 experts corresponding to the 5 QuickDraw categories, represents the standard approach of aligning experts with human-defined labels.

Second, the unsupervised approach exhibits a clear performance trend: reconstruction loss decreases substantially from 1 to approximately 7 experts, then gradually degrades beyond this optimal range, suggesting that too many experts lead to over-fragmentation of the data representation. Notably, the optimal number of experts (7) differs from the number of ground truth categories (5), providing evidence that the model discovers a more nuanced data organization than human categorical labels.

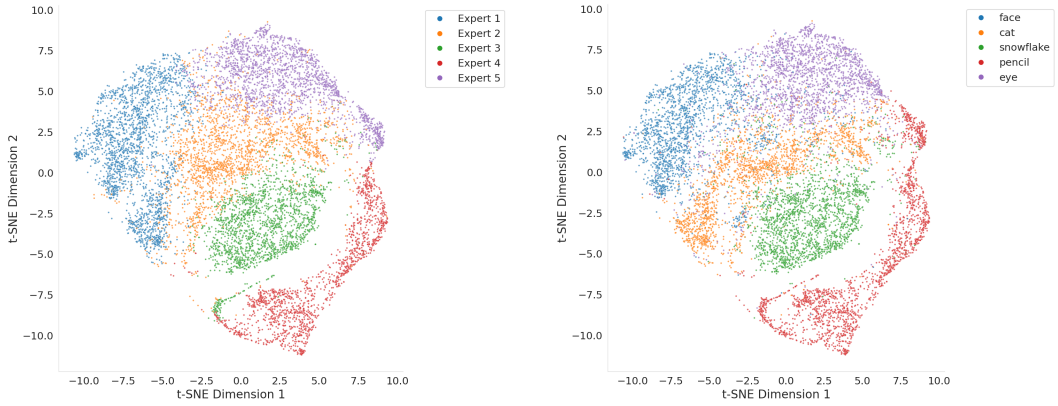
The performance variance (shown by error bars) demonstrates that both approaches achieve consistent results across multiple training runs, but the unsupervised method consistently outperforms the supervised baseline.

These results support our hypothesis that unsupervised expert specialization can discover data structure that is fundamental to the underlying distribution that is not present in the human-imposed categorizations, validating the theoretical predictions from [6] about MoE's natural cluster discovery capabilities in real-world image data.

5.2 Expert Specialization Analysis

To understand why unsupervised expert routing outperforms supervised approaches, we analyze how the learned expert assignments relate to the underlying structure of the latent space. Figure 3 presents t-SNE [11] visualizations of the same latent space representations, colored by expert assignments (left) versus ground truth class labels (right).

The comparison reveals a strong correspondence between the clusters formed by expert assignments and those defined by ground-truth class labels. As seen in Figure 3, the spatial organization of expert-colored clusters (left) closely mirrors the class-colored clusters (right). This visual alignment suggests that the unsupervised expert specialization discovers a latent structure that is highly correlated with the semantic categories in the data. To quantify this relationship, we calculated the correlation between the expert assignments through unsupervised training and the class labels, find-



(a) Colored by Expert Assignment

(b) Colored by Class Labels

Figure 3: t-SNE of latent space. Expert assignments (left) form more coherent clusters than ground truth class labels (right), explaining the superior performance of unsupervised routing.

ing a strong positive correlation of 0.802. This indicates that while the model is not explicitly guided by labels, it learns to partition the data in a way that is semantically meaningful.

This correlation between labels and expert assignments is promising, but it does not explain why unsupervised routing performs better than supervised. One visual clue from Figure 3 is the better quality of the clustering, which means fewer overlaps between the clusters in case of expert assignments. To quantify this difference in clustering quality, we trained linear classifiers to predict both expert assignments and the original class labels from latent representations. The results strongly support the visual observations: classification accuracy reaches 93.4% when predicting expert assignments, compared to only 85.1% when predicting ground truth class labels. This 8.3% improvement demonstrates that expert assignments are more linearly separable in the latent space.

The experts naturally organize according to the intrinsic geometry of the learned latent space, creating more coherent and well-separated regions that are easier to model with individual decoder networks. In contrast, human-defined categories may not respect the natural boundaries that emerge from the data's underlying manifold structure, leading to expert assignments that are less optimal for reconstruction tasks.

5.3 Visual Expert Specialization Patterns

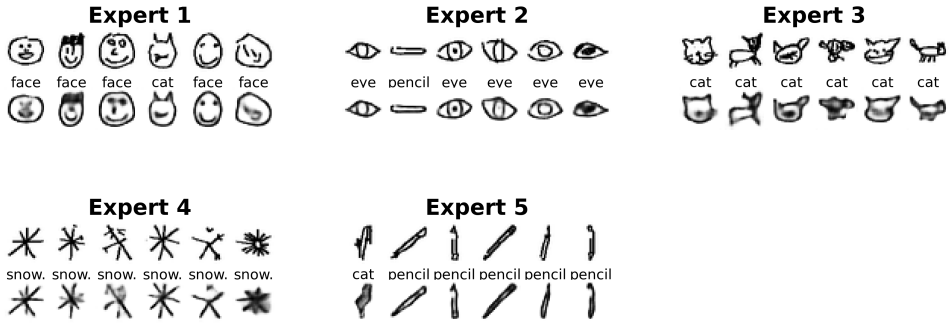


Figure 4: Expert specialization for 5 experts. Each expert shows 5 random input images (top) and their reconstructions (bottom), demonstrating specialization based on visual features rather than semantic class labels.

To directly observe what each expert has learned to specialize on, we visualize the actual images that activate each expert along with their reconstructions. Figure 4 shows this analysis for a model with 5 active experts, while Figure 5 demonstrates the same for 23 active experts. For each expert, we

display 5 randomly selected images that were routed to that expert (top row) and their corresponding reconstructions (bottom row), along with the ground truth class labels and utilization percentages.

The results reveal specialization patterns that transcend class boundaries. In the 5-expert configuration, Expert 1 specializes primarily in faces and cats, Expert 2 focuses on eyes and flat oval structures, Expert 3 handles cats and similar curved shapes, Expert 4 processes snowflakes and Expert 5 is dedicated to pencil-like linear objects. Importantly, the expert assignments capture visual similarity rather than semantic categories, for instance, certain cat drawings that resemble faces are routed to the face expert rather than being constrained by their ground truth labels.

The 23-expert configuration in Figure 5 reveals even more granular specialization patterns. Most striking is the fine-grained organization of pencil-like objects: the model learns separate experts for horizontal pencils, vertical pencils, and pencils at various angles. Similarly, cat experts differentiate between different ways of drawing a cat. This level of specialization demonstrates that when given sufficient capacity, the MoE architecture discovers sub-categorical structures that are highly relevant for reconstruction quality but invisible to human categorical labels.

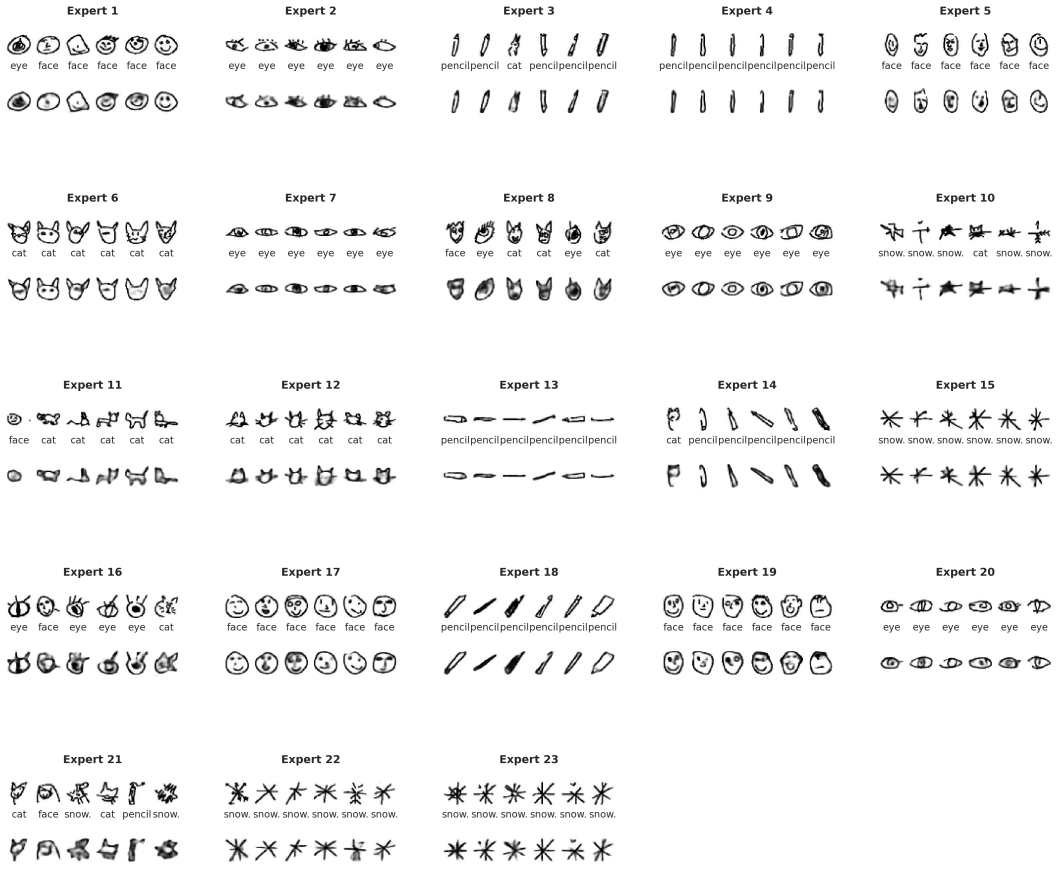


Figure 5: Expert specialization for 23 experts showing fine-grained sub-categorical specializations (e.g., pencils at different orientations and various face/eye sub-types).

These visualizations provide evidence for why unsupervised expert assignment outperforms supervised approaches. The gating network successfully captures visual similarity patterns that span across traditional class boundaries while discovering meaningful sub-categories within classes. A cat drawing that visually resembles a face is appropriately routed to the face expert, enabling better reconstruction than forcing it into a "cat" expert that may not handle face-like features well. This adaptive specialization based on visual structure rather than assigned labels in part explains the superior reconstruction performance observed in our quantitative results.

The question that arises now is why doesn't the 23 expert configuration outperform the 7 expert one as show in Figure 2? We conducted further analysis on dataset size impact to answer this question.

5.4 Dataset Size Impact on Optimal Expert Count

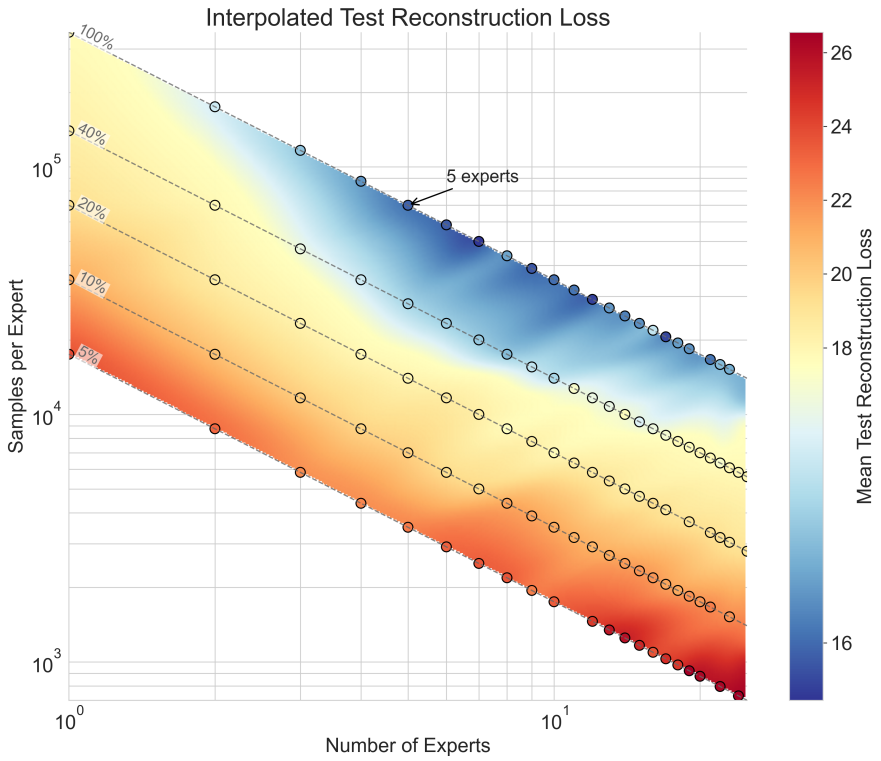


Figure 6: Log-log influence of number of experts and samples per expert on performance (test reconstruction loss). Markers show experimental points; surface is linearly interpolated. A clear minimum appears around 5 experts, matching the number of classes.

Previous results raise a question: what factors, beyond the number of semantic categories, influence optimal number of experts? A primary candidate is the size of the dataset, or more specifically, the number of samples available for each expert to learn from.

To investigate this relationship, we conducted a systematic analysis across various data regimes, with the results visualized in Figure 6. This figure presents a log-log plot where the x-axis represents the number of experts and the y-axis represents the number of samples per expert. The color intensity corresponds to the mean test reconstruction loss, with blue indicating lower loss and red indicating higher loss. The surface is generated by linearly interpolating between experimental data points (shown as circles), while the dashed diagonal lines represent fixed percentages of the total dataset size. A complementary view of this data, including error bars, is provided in Appendix A.

The plot reveals several key trends. A distinct valley of low reconstruction loss is centered near 5 experts, which aligns with the number of classes in our dataset. This suggests that, as a baseline, the model's performance is optimized when the number of experts roughly matches the number of primary data categories.

If we observe the plot vertically (fixing the number of experts and increasing samples per expert) we confirm a well-known principle: more data leads to better performance, as evidenced by the color shifting from red to blue. A more interesting pattern emerges when we analyze the plot horizontally, increasing the number of experts while keeping the number of samples per expert constant. If the number of samples per expert was the sole factor that influenced expert performance (lower loss) we would expect to observe somewhat constant performance along the horizontal axis. Instead, we observe a significant improvement in performance as we move to the right. This phenomenon can be attributed to two factors. First, although each decoder expert is trained on a fixed number of samples,

the shared encoder is exposed to a larger and more diverse dataset, allowing it to learn more robust and effective latent representations for the entire network.

Second, and more fundamentally, increasing the number of experts allows for greater specialization. With fewer experts, each is forced to model a more complex, multi-modal data distribution (e.g., a single expert may have to learn to reconstruct both "cat" and "pencil" sketches). In contrast, with more experts, each can specialize on a simpler, more uni-modal subset of the data. To validate this hypothesis, we conducted an additional experiment: a single expert trained on the five-class dataset achieved a reconstruction loss of 24.0. When the same network was trained on single-class datasets of the same total size, average loss was significantly lower, at 18.4. For comparison, the same network trained on 20 times more data achieves loss of 17.7, which is only marginally better than the previous. This confirms that it is more effective for experts to learn from homogeneous data distributions, explaining why the performance improves as the model is given more experts to partition the data, even though the number of samples per expert remains constant.

6 Conclusion

We have presented a novel SMoE-VAE architecture that enables interpretable analysis of expert specialization through visual reconstruction patterns. Our key finding demonstrates that for the QuickDraw database, unsupervised expert routing consistently outperforms supervised routing based on human-provided labels, revealing that experts naturally discover data structures which are more informative, when allowed to specialize according to intrinsic patterns rather than categorical boundaries.

Through comprehensive evaluation on the QuickDraw dataset, we showed that experts develop coherent specializations that span across traditional class boundaries while identifying reconstruction-relevant subclusters within classes. The t-SNE analysis aand visual reconstruction patterns provide compelling evidence that expert assignments based on learned data structure achieve superior linear separability (93.4% vs 85.1%) and reconstruction quality compared to human-defined categorizations.

Our analysis of dataset size effects reveals a nuanced relationship between performance, data quantity, and the degree of expert specialization. We found that expert performance is more sensitive to the homogeneity of the data it models than to the absolute number of samples it is trained on. Increasing the number of experts allows for greater specialization on simpler, more uni-modal data subsets, which improves reconstruction quality even when the number of samples per expert is held constant. However, this benefit is balanced by the risk of data starvation; for a fixed dataset size, increasing the number of experts indefinitely degrades performance. These findings highlight a critical trade-off in MoE design and provide guidance for selecting an appropriate number of experts based on the underlying structure and size of the dataset.

At this point, it is worth mentioning that our experiments had a persistent challenge with expert collapse, where roughly half of the available experts remain inactive during training. This suggests potential issues with network initialization or the early training procedure that warrant further investigation.

Another key point is that our load balancing and entropy regularization terms differ from standard approaches in the literature, with [14] using almost the same loss formulation. However, we note that the theoretical work of [6] achieved low entropy probability distributions with equal load balance, providing strong indications that our conclusions would remain valid with alternative loss formulations.

Looking toward future research, several important directions emerge from this work. First, synthetic datasets could provide controlled environments for a deeper analysis of expert specialization mechanisms. Second, the behavior of SMoE-VAE architectures under dataset imbalance conditions remains unexplored and could reveal important robustness characteristics. Third, extending our approach to multiple layers of MoE could unlock more sophisticated hierarchical specialization patterns. Lastly, one obvious extension of our work is to validate the findings across different image databases.

Our work sugests that expert specialization can serve as a lens for understanding fundamental data organization principles. The finding that unsupervised routing discovers a more informative struc-

ture than human categorizations has broader implications for how we conceptualize optimal computational organization in neural networks. This methodology opens new avenues for interpretable analysis of complex architectures and provides a foundation for future investigations into the relationship between learned representations and labeled data structure.

A Different look on Figure 6

Figure 7 provides a complementary view to Figure 6, making it easier to see differences in test reconstruction loss in absolute numbers, with explicit error bars indicating the standard deviation across multiple runs. Each curve corresponds to a different percentage of the total dataset, from 5% to 100%. The x-axis represents the number of samples available to each expert, which means that the number of active experts increases from right to left along each curve. Note that the lowest curve (with 100% of data used) uses the same data as the curve in Figure 2.

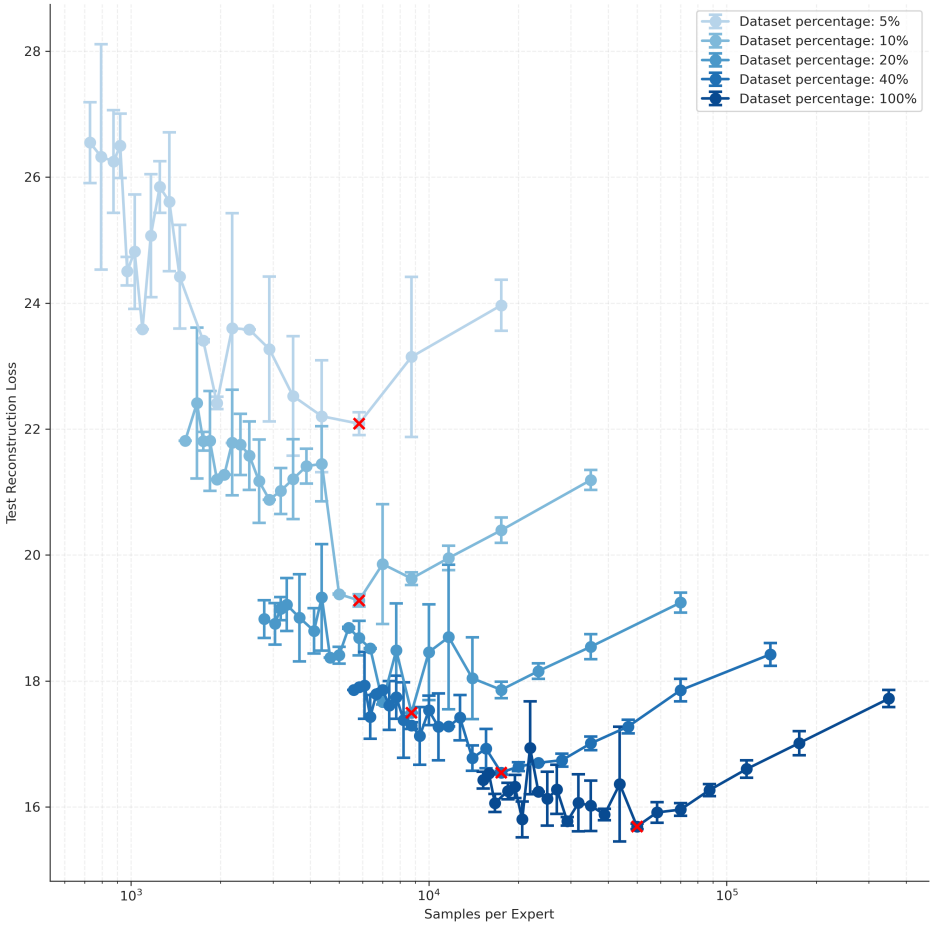


Figure 7: Test reconstruction loss as a function of samples per expert for different dataset sizes. Each curve represents a fixed dataset percentage, with error bars showing standard deviation across runs.

This visualization offers another perspective on the relative importance of data homogeneity versus the sheer volume of samples per expert. For instance, a model trained on 5% of the dataset (the light blue curve) with 17,500 samples allocated to a single expert achieves a reconstruction loss of around 24.0. In contrast, a model trained on 40% of the data (the dark blue curve) but with a same number of samples per expert distributed among 4 experts achieves a significantly lower loss of approximately 18.0. This comparison reinforces the conclusion from the main text: expert performance is more

sensitive to the homogeneity of the data it models than to the absolute number of samples it is trained on.

Furthermore, this plot helps explain why increasing the number of experts does not always lead to better performance. As the number of experts grows (moving from right to left on a given curve), the number of samples per expert decreases, eventually leading to data starvation and degraded performance.

Code and data availability

Our code is available at `https://github.com/strajdzsha/smoe-vae`.

References

[1] 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.

[2] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity," Journal of Machine Learning Research, vol. 23, no. 120, pp. 1–39, 2022.

[3] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," in International Conference on Machine Learning, 2020, pp. 5737–5746.

[4] Y. Shi, N. Siddharth, B. Paige, and P. H. S. Torr, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," in Advances in Neural Information Processing Systems, vol. 32, 2019.

[5] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim and J. Huang, "A Survey on Mixture of Experts in Large Language Models," in IEEE Transactions on Knowledge and Data Engineering, vol. 37, no. 7, pp. 3896-3915, July 2025, doi: 10.1109/TKDE.2025.3554028.

[6] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li, "Towards Understanding the Mixture-of-Experts Layer in Deep Learning," in Advances in Neural Information Processing Systems, 2022, vol. 35, pp. 23049–23062.

[7] A. Kopf, V. Fortuin, V. R. Somnath, and M. Claassen, "Mixture-of-Experts Variational Autoencoder for clustering and generating from similarity-based representations on single cell data," PLOS Comput. Biol., vol. 17, no. 6, pp. 1–17, Jun. 2021. doi: 10.1371/journal.pcbi.1009086.

[8] N. Dikkala, N. Ghosh, R. Meka, R. Panigrahy, N. Vyas, and X. Wang, "On the Benefits of Learning to Route in Mixture-of-Experts Models," in The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in International Conference on Learning Representations, 2014.

[10] D. Ha and D. Eck, "A neural representation of sketch drawings," in International Conference on Learning Representations, 2018.

[11] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 9, no. 11, pp. 2579–2605, 2008.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in International Conference on Learning Representations, 2015.

[13] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," 2019

[14] D. Zhang, Y. Sun, B. Eriksson, and L. Balzano, "Deep unsupervised clustering using mixture of autoencoders," *arXiv preprint arXiv:1712.07788*, 2017.

[15] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *ICLR (Poster)*, 2017, arXiv:1701.06538.

[16] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keysers, and N. Houlsby, "Scaling Vision with Sparse Mixture of Experts," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.