Mixture of Experts for Image Classification: What's the Sweet Spot?

Anonymous authors
Paper under double-blind review

Abstract

Mixture-of-Experts (MoE) models have shown promising potential for parameter-efficient scaling across domains. However, their application to image classification remains limited, often requiring billion-scale datasets to be competitive. In this work, we explore the integration of MoE layers into image classification architectures using open datasets. We conduct a systematic analysis across different MoE configurations and model scales. We find that moderate parameter activation per sample provides the best trade-off between performance and efficiency. However, as the number of activated parameters increases, the benefits of MoE diminish. Our findings offer practical guidance for efficient model design using MoE for image classification tasks.

1 Introduction

Recent advances in machine learning, particularly in the domains of natural language processing (NLP)(Vaswani et al., 2017; Kenton & Toutanova, 2019) and computer vision (Dosovitskiy et al., 2020), have been primarily driven by scaling up model size, computational budgets, and training data. Although these large-scale models demonstrate impressive performance, they are often expensive to train and consume considerable energy resources (Strubell et al., 2019). As a result, the research community has become increasingly interested in exploring more efficient training and serving paradigms. One such promising solution is the use of sparse expert models, with Mixture-of-Experts (MoE) (Shazeer et al., 2017; Lepikhin et al., 2020) emerging as a popular variant.

MoE models introduce sparsity by partitioning the set of parameters into multiple parallel sub-models, called "experts." During training and inference, the gating part of the models routes input examples to specific expert(s), ensuring that each example only interacts with a subset of the network parameters. As the computational cost is partially correlated with the number of parameters activated for a given sample rather than the total number of parameters, this approach facilitates the scaling-up of the model, while keeping computational costs under control, making it an attractive option for a wide range of applications.

Despite their success in various domains (Hihn & Braun, 2021; Costa-jussà et al., 2022; Zoph et al., 2022; Fedus et al., 2022), the application of MoE models in image classification and computer vision, in general, remains limited, and it often requires very large datasets (Riquelme et al., 2021; Mustafa et al., 2022; Komatsuzaki et al., 2022) to be competitive against state-of-the-art approaches. In this work, we focus on leveraging the potential of MoE models for image classification on ImageNet-1k and ImageNet-21k (Russakovsky et al., 2015). We study the efficiency of integrating MoE within two renowned architectures, ConvNext (Liu et al., 2022) and Vision Transformer (ViT) (Touvron et al., 2022). We conduct a series of experiments considering various architecture configurations. Likewise, we investigate the impact of various components, including the number of experts and their sizes, the gate design, and the layer positions, among others. Our experimental findings indicate that optimal design is contingent on the specific network architecture, and that consistently situating the MoE layer within the final two even blocks invariably yields substantial improvements for moderate model size. Nevertheless, when scaling up the approach to large models and datasets close to the state-of-the-art, the benefits of using Mixture-of-Experts for image classification gradually vanish.

Contributions. In this work, we provide a systematic study of applying Mixture-of-Experts (MoE) models to image classification tasks, focusing on the ImageNet-1k and ImageNet-2k benchmarks. Unlike prior efforts that emphasize large-scale deployments with vast compute resources, we explore MoE integration within mid-sized ConvNext and Vision Transformer (ViT) architectures, identifying effective design principles for efficient training and inference. Our contributions are threefold: (1) we evaluate the impact of expert count, expert size, and gating mechanisms on accuracy and computational cost across both architectures; (2) we discover that placing MoE layers in the final two even-numbered blocks yields consistent gains, offering a simple yet robust heuristic for model design; and (3) we reveal a diminishing return of MoE models at larger scales, highlighting the limits of sparsity-based approaches for vision tasks under near state-of-the-art conditions. Our findings offer practical guidelines for future work seeking to balance performance and efficiency in vision model design.

2 Related Work

The Mixture-of-Experts (MoE) model, introduced by (Jacobs et al., 1991), partitions complex tasks into hopefully simpler sub-tasks handled by expert models, whose predictions are combined to produce the final output. This framework has been successfully integrated into various neural network architectures, particularly transformers for NLP tasks (Shazeer et al., 2017). Given this success in NLP, interest has increased in applying MoE to the diverse and complex computer vision tasks.

The transformative potential of MoE architectures is underscored by their successful integration across various domains, with the Vision Transformer (ViT) being a prime example. Riquelme et al. (2021) introduced V-MoE, an MoE-augmented ViT model, for image classification tasks on a massive dataset containing hundreds of millions of examples. They showed that V-MoE not only matches the performance of prior state-of-the-art architectures but also requires half the computational resources during inference. Lou et al. (2021) presented a sparse MoE MLP model based on the MLP-Mixer architecture (Tolstikhin et al., 2021). While the MLP-Mixer architecture does not display performance accuracy, the MoE-enhanced version surpassed its dense counterpart in experiments on ImageNet and CIFAR. Hwang et al. (2023) demonstrated the effectiveness of SwinV2-MoE, a MoE-based model built upon Swin Transformer V2 architecture. They reported superior accuracy in downstream computer vision tasks. Similarly, Puigcerver et al. (2024) introduced a soft MoE mechanism, demonstrating improved performance and training speed compared to classical MoE on billion-scale datasets.

3 Sparsely Activated MoE

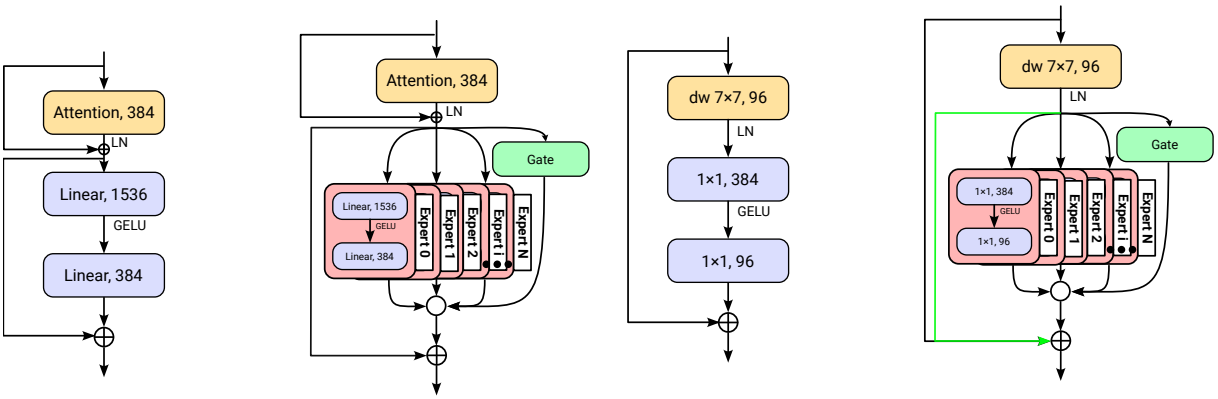
Based on conditional computations, MoE aims at activating specific parts of the model depending on the input. The core idea is to assign experts to different regions of the input space, thereby increasing the model capacity by augmenting the number of parameters without incurring significant computational overhead (Jacobs et al., 1991).

According to Shazeer et al. (2017), an MoE includes a router G and a set of experts E. The router learns a sparse assignment between the input and the experts, while the experts process the inputs like standard neural network modules do. Let x be the input, and let $(E_i)_{i\in[1,N]}$ be the N experts of the MoE layer. Following (Shazeer et al., 2017), the output of this MoE layer is given by:

$$\sum_{i\in \mathrm{Top}_k(x)} G(x)_i \cdot E_i(x)$$

where, G is a conv1x1 gate employing a softmax function, and $G(x)_i$ is its ith output. To promote sparsity in the MoE layer, the number of participating experts is restricted to $k < N$ for each input: $\mathrm{Top}_k(x)$ contains the indices of the k highest $G(x)_i$. Note that involving $k > 1$ experts increases the computational cost compared to the sparse ($k = 1$) counterpart. We tested several gate designs, but none of them clearly outperformed the simple conv1x1 gate (Sec. 4.3.4).

However, enforcing this MoE design is insufficient to yield an effective MoE. To exploit compute capacity optimally, we encourage uniform usage of each expert by using a load balancing auxiliary loss proposed



(a) **Left**: ViT base block. **Right**: Our modified MoE ViT block. LN: Layer normalization. GELU: Gaussian Error Linear Units.

(b) **Left**: ConvNext original base block. **Right**: The modified MoE ConvNext block. The green link represents the added skip connection. DW: Depth-wise. LN: Layer normalization.

Figure 1: Vision Transformer and ConvNext architectures.

in (Shazeer et al., 2017), which promotes balanced expert assignments and equal importance of experts (for more details, see Section 4 of Shazeer et al. (2017)). We also use Batch Prioritized Routing (BPR) introduced by (Riquelme et al., 2021).

3.1 Vision Transformer

The Vision Transformer (ViT) (Dosovitskiy et al., 2020) represents a paradigm shift in computer vision. It moves away from traditional convolution-based architectures in favor of the Transformer's attention mechanism, initially pioneered for natural language processing tasks (Vaswani et al., 2017). In the ViT model, an image is segmented into fixed-size and non-overlapping patches. These patches are linearly embedded into vectors and then passed through a series of Transformer blocks. Each Transformer block is predominantly made up of two main components: a multi-head self-attention mechanism and a multi-layer perceptron (MLP). Within the context of ViT, the MLP ratio (denoted mlp_ratio in the following) refers to the ratio of the hidden dimensions of the MLP to the embedding dimensions. This ratio is pivotal, as it determines the capacity and computational cost of the MLP component. We replace some predefined MLP block by a MoE, as depicted in Fig. 1a. Each MoE operates at the patch level: the experts are spatially distributed.

3.2 ConvNext

ConvNext (Liu et al., 2022) has modernized the classical convolutional neural networks (CNN) design by incorporating insights from Vision Transformers (ViT). The central component of the ConvNext architecture is a block containing a 7×7 depth-wise convolution followed by a 1×1 convolutional layer (Fig. 1b-left). This block is replicated across four stages, each with decreasing resolution. In this work, we replace some predefined ConvNext block, namely the multilayer perceptron, with a MoE block, and introduce a skip connection, as illustrated in Fig. 1b-right. Consequently, each expert becomes a fully convolutional block operating at the feature map level, resulting in spatially distributed experts. Hence, this design not only increases the network capacity but also allows each expert to specialize in specific spatial locations and engages multiple experts per image.

4 Experiments on ImageNet

This section presents the results obtained by the architectures described above (that incorporate MoE into the ConvNext and ViT architectures), trained on ImageNet-1k or pre-trained on ImageNet-21k datasets. It

Table 1: Accuracy for different ImageNet-1K trained models and the MoE strategies "Every 2" and "Last 2" (see text). "Top k" corresponds to the number of experts involved. Throughput is measured on V100 GPUs, following (Touvron et al., 2021). For ImageNet-1k, non-isotropic ConvNext models feature an mlp_ratio of 2, which contributes to their improved flops-per-sample efficiency compared to their dense counterparts.

Architecture	#Params ($\times 10^6$)	Per sample #Params$_{act}$	FLOPs	Throughput (im/s)	IN-1K Accuracy
ConvNeXt-T (Liu et al., 2022)	28.6	28.6	4.5G	814	82.1
ConvNeXt-T-4 Last 2 Top 1	34.5	25.6	4.2G	768	82.1
ConvNeXt-S (Liu et al., 2022)	50	50	8.7G	466	83.1
ConvNeXt-S-4 Last 2 Top 1	56.1	47.3	8.5G	442	83.1
ConvNeXt-B (Liu et al., 2022)	88.6	88.6	15.4G	299	**83.8**
ConvNeXt-B-4 Last 2 Top 1	99.1	83.4	15.0G	289	83.5
ConvNeXt-S *(iso.)*	22.3	22.3	4.3G	1100	79.7
ConvNeXt-S-8 *(iso.)* Last 2 Top 1	38.9	22.3	4.3G	1031	**80.3**
ConvNeXt-B *(iso.)*	82.4	82.4	16.9G	336	**82.0**
ConvNeXt-B-8 *(iso.)* Last 2 Top 1	115.4	82.4	16.9G	303	81.6
ViT-S	22.0	22.0	4.6G	1083	79.8
ViT-S-8 Last 2 Top 2	38.6	25.0	5.3G	892	80.5
ViT-S-8 Every 2 Top 2	71.7	33.1	6.9G	724	**80.7**
ViT-B	86.6	86.6	17.5G	329	**82.8**
ViT-B-8 Every 2 Top 2	284.9	129.9	26.3G	227	82.5

includes the experimental setup, the results on the ImageNet-1k validation set, and the impact of various MoE configurations.

4.1 Experimental Setup

All the results presented in this section rely on MoE models trained on the ImageNet datasets. Although the context is different due to the changes in network structures presented in Secs. 3.1 and 3.2, we set the training hyperparameters similar to (Touvron et al., 2022) for ViT and (Liu et al., 2022) for ConvNext (a grid search looking for better settings did not bring significant improvement).

Furthermore, when working with the ImageNet-1k dataset, we use a strong data-augmentation pipeline, including Mixup (Zhang et al., 2018), Cutmix (Yun et al., 2019), RandAugment (Cubuk et al., 2020), and Random Erasing (Zhong et al., 2020), over 300 epochs. Likewise, we utilize drop path, weight decay, and expert-specific weight decay as regularization strategies. However, for the pretraining phase on ImageNet-21k, we exclude Mixup and Random Erasing from the data augmentation pipeline because they did not improve the results, as also reported in (Touvron et al., 2022; Tu et al., 2022). We pre-train the model for 90 epochs, adhering to the original ConvNext approach (Liu et al., 2022) and, consistently with (Liu et al., 2022) and (Touvron et al., 2022), the final results of ImageNet-1k are obtained through fine-tuning for 30 epochs for ConvNext and 50 for ViT. Comprehensive details of all the hyperparameters are provided in Tab. 10 in App. A.

4.2 Base results on ImageNet

Tab. 1 presents the results obtained on ImageNet-1k validation set by a model that has been entirely trained on ImageNet-1k, for isotropic architecture (e.g., ViT, ConvNext *iso.*) and a hierarchical architecture, namely ConvNext. We see some significant improvements in accuracy compared to the "no-MoE" results for small model sizes, especially for anisotropic models.

Tab. 2 presents the results of models that are pre-trained on ImageNet-21k, and tested on the same ImageNet-1k validation set than above. Here, MoE does bring some improvement for moderate numbers of activations per sample. However, this improvement decreases for large numbers of activations per sample, as depicted in Fig. 2.

Table 2: Accuracy for ImageNet-21K pre-trained models and different adaptations with ImageNet-1K: *IN-1K* is standard fine-tuning, *Linear prob* is the training of a linear layer for the output probability layer (everything else being frozen), and *0-shot* is the direct application of the pre-trained model to ImageNet-1K. "Every 2" and "Last 2" are the corresponding MoE strategies (see text). "Top k" corresponds to the number of experts involved.

Architecture	#Params ($\times 10^6$)	Per sample #Params$_{act}$	FLOPs	IN-1K Accuracy	Linear prob Accuracy	0-shot Accuracy
ConvNeXt-T (Liu et al., 2022)	28.6	28.6	4.5G	82.9	81.8	44.3
ConvNeXt-T-8 Last 2 Top 1	70.0	28.7	4.5G	**83.5**	**82.3**	**44.6**
ConvNeXt-S (Liu et al., 2022)	50.3	50.3	8.7G	84.6	83.2	45.2
ConvNeXt-S-8 Last 2 Top 1	91.6	50.3	8.7G	**84.9**	**83.6**	**45.3**
ConvNeXt-B (Liu et al., 2022)	88.6	88.6	15.4G	**85.8**	**84.9**	45.8
ConvNeXt-B-8 Last 2 Top 1	162.0	88.6	15.4G	85.7	84.8	**45.9**
ViT-S	22.0	22.0	4.6G	82.6	81.7	44.2
ViT-S-8 Every 2 Top 2	71.7	33.1	6.9G	**83.0**	**81.9**	**44.7**
ViT-B	86.6	86.6	17.5G	85.2	84.0	**45.7**
ViT-B-8 Every 2 Top 2	284.9	129.9	26.3G	85.2	84.0	45.6

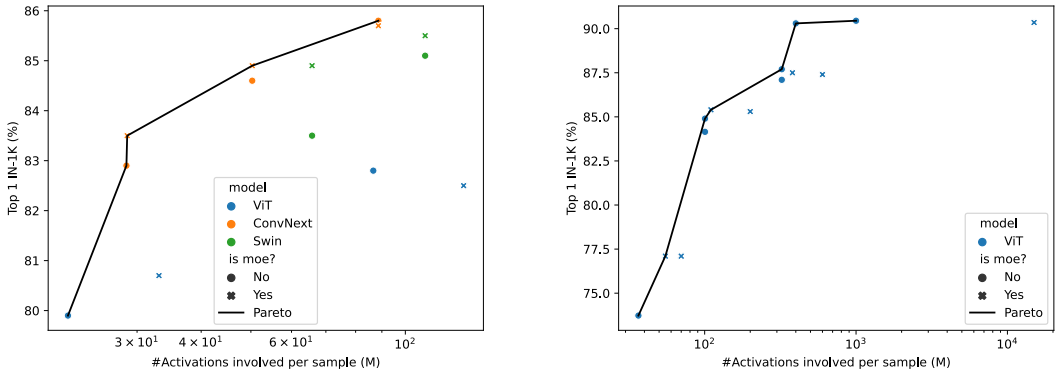


Figure 2: **Left**: Pareto-front for ImageNet21k, x-axis = number of activations per sample. ViT models have been presented in MoE versions only after additional pretraining, and are therefore not presented. MoE seems to be Pareto optimal for a number of activations per sample below 90M. **Right**: Pareto-front for ViT models on JFT-300M. Overall, MoE is never validated for a number of activations per sample above $\approx$ 100M.

4.3 Sensitivity with Respect to Design Choices

This section presents different sensitivity studies that investigate the specifics of the design choices of the MoE layers. We explore the impact of the position of these layers inside the architecture (Sec. 4.3.1), the effect of varying the number of experts (Sec. 4.3.2), and finally, we assess the necessary design changes when transitioning from the ImageNet-1k to ImageNet-21k dataset (Sec. 4.3.3) and the impact of the routing architecture choices (Sec. 4.3.4).

4.3.1 Impact of the Position of MoE Layers

Tab. 3 displays the effects of various placements of MoE layers, comparing the following configurations: (a) **Every 2**: a MoE layer replaces every second layer; (b) **Stage**: a MoE layer replaces the final layer of each stage, resulting in four MoE layers throughout the network for ConvNext; (c) **Last 2**: a MoE layer replaces the final layer of each of the last two stages; (d) **Last 3**: Same as "Last 2", but with an additional MoE layer inserted in the middle of stage 3. As shown in Tab. 3, "Last 2" strategy is the most robust choice: it performs well across all architectures. On the contrary, "Every 2" performs worst for ConvNext architecture, and best for ViT. Consequently, in the following, all experiments use "Last 2" for ConvNext and "Every 2" for ViT.

Table 3: Comparative results for different positions of MoE layers: ImageNet-1k training on ConvNext-T and ViT-S, all employing 8 experts.

Architecture	MoE	# Params		FLOPs	IN-1K
		$\times 10^6$	Sample		
ConvNext-T	**no MoE**	28.6	28.6	4.5G	**82.1**
ConvNext-T-8	every 2	54.3	17.0	3.8G	81.8
	stage	47.4	25.5	4.0G	**82.1**
	last 3	49.9	25.0	4.1G	**82.1**
	last 2	46.3	25.6	4.2G	**82.1**
ConvNeXt-S *(iso.)*	no moe	22.3	22.3	8.7G	79.7
ConvNeXt-S-8 *(iso.)*	every 2	96.7	22.3	8.7G	79.6
	last 2	38.9	22.3	8.7G	**80.3**
ViT-S	no MoE	22.0	22.0	4.2G	79.9
ViT-S-8	**every 2**	71.7	33.1	5.3G	**80.7**
	last 2	38.6	25.0	6.9G	80.5

1

Table 4: Comparative results for different numbers of experts. ImageNet-1k training using the "Last 2" strategy.

Architecture	# experts	#Params ($\times 10^6$)	Per samples #Params$_{act}$	IN-1K Top 1 acc.
ConvNext-T	**no MoE**	28.6	28.6	**82.1**
ConvNext-T	**4**	54.3	25.5	**82.1**
	8	47.4	25.5	**82.1**
	16	46.3	25.5	81.7
ViT-S	no MoE	22.0	22.0	79.9
ViT-S	4	29.1	25.0	79.8
	8	38.6	25.0	**80.5**
	16	57.5	25.0	80.2

Table 5: ImageNet-21k training on ConvNext-T employing 8 experts and an MLP ratio of 4 unless specified explicitly.

MoE strategy	#Params ($\times 10^6$)	Per samples #Params$_{act}$	FLOPs	IN-1K Top 1 acc.
no MoE	28.6	28.6	4.5G	82.9
every 2	97.5	28.6	4.5G	82.9
stage	78.2	28.6	4.5G	**83.5**
last 3	70.0	28.6	4.5G	**83.5**
last 2	70.0	28.6	4.5G	**83.5**
last 2 mlp_ratio 2	46.3	25.6	4.2G	83.4

4.3.2 Influence of the Number of Experts

Tab. 4 presents the performance of ConvNext and ViT architectures on ImageNet-1K training with the "Last 2" strategy MoE layers. It provides a comprehensive view of how the number of experts influences the size (parameter count) of the models and their Top-1 accuracy on this dataset.

The experimental results suggest that four experts yield the best performance for ConvNext, while eight experts are optimal for ViT. This is showcased by the Top-1 accuracy rates, which are consistent at 82.1% for ConvNext with four experts and reach a peak of 80.5% for ViT with eight experts. However, we note that increasing the number of experts to 16 has a detrimental effect on both architectures. Specifically, for ConvNext, the top-1 accuracy slightly drops to 81.7%, and for ViT, the performance plateaus at 80.2%.

These findings highlight that while MoE layers can enhance model performance, there is a delicate trade-off to be struck in terms of the number of experts. Exceeding the optimal number can lead to suboptimal results, negating the potential benefits of the MoE integration. Furthermore, in our detailed analysis, described in Sec. 5.4, we explore the interplay between the number and the size of each expert. Our investigation revealed that, for isotropic networks such as ViT and ConvNext *iso.*, reducing the size of experts adversely affects performance, while for ConvNext, it has no significant impact. Moreover, during the course of our experiments, we observed that while the Top-1 configuration was superior for ConvNext, the Top-2 configuration for ViT was comparable or even better than the ConvNext Top-1 configuration, leading us to employ Top-1 for ConvNext and Top-2 for ViT.

Table 6: ImageNet-21k training on ConvNext-T: comparison between various numbers of experts, with an MLP ratio of 4.

# experts	#Params ($\times 10^6$)	Per samples #Params$_{act}$	FLOPs	IN-1K Top 1 acc.
no MoE	28.6	28.6	4.5G	82.9
8	54.3	28.6	4.5G	83.5
16	117.1	28.6	4.5G	**83.6**
32	211.6	28.6	4.5G	83.4

4.3.3 Results on ImageNet-21k

Tabs. 5 and 6 present the results of models trained on ImageNet-21k, reporting ImageNet-1k accuracy after fine-tuning. We analyse the effect of MoE layer positioning, expert size (Tab. 5), and expert count (Tab. 6). The results demonstrate that, when large volumes of data are accessible, a greater number of experts can be effectively deployed. Notably, the use of sixteen experts, which was sub-optimal for ImageNet-1k, does not negatively affect performance on ImageNet-21k. This suggests that increasing the number of experts, together with the volume of data, could lead to valuable enhancements.

In terms of MoE layer configuration, our data showed no performance improvements with the addition of more layers, and the "every 2" strategy again yielded the poorest results for ConvNext-based architecture.

4.3.4 Impact of the Routing Architectures

We investigated three distinct routing architectures for our study:

- "Conv": A straightforward 1×1 convolution, similar to a linear gate.
- "Cos": A gate that utilizes cosine similarity for routing, as detailed in (Chi et al., 2022). It is a two-layer architecture, with one linear layer projecting the features to a lower-dimensional space and another performing cosine similarity against some learned latent codes.
- "L2": This is identical to the "Cos" gate, except that it employs the L2-distance for similarity instead of cosine.

Table 7: Performance of MoE-Integrated Models vs Non-MoE Baselines: ablation study on the gating network architecture. MoE layers are placed on the last two even layers.

ConvNext-S					
Gate	IN-1K	IMV2	IMA	IMR	IMSK
no MoE	**82.1**	70.8	24.2	**47.2**	**33.8**
Conv	**82.1**	**71.0**	23.8	46.2	32.7
Cos	81.9	70.6	22.3	45.6	33.0
L2	82.0	71.6	**25.0**	45.9	32.8
ConvNext-S *(iso.)*					
Gate	IN-1K	IMV2	IMA	IMR	IMSK
no MoE	79.7	68.6	13.0	46.4	34.0
Conv	**80.3**	**68.8**	**13.5**	**46.6**	**34.4**
Cos	79.7	68.4	12.2	45.3	33.0
L2	80.1	68.6	13.3	46.1	34.3
ViT-S					
Gate	IN-1K	IMV2	IMA	IMR	IMSK
no MoE	79.8	**69.1**	19.8	**43.4**	29.7
Conv	**80.5**	**69.1**	**20.3**	43.2	**29.9**
Cos	79.7	67.8	15.9	41.6	29.3
L2	80.2	68.8	19.0	41.9	29.2

We conducted training on the ImageNet-1K dataset using various small networks, each employing different routing mechanisms as displayed earlier. The results, as shown in Tab. 7, indicate that the simple convolutional (conv) configuration generally yields slightly better performance in most cases. However, when evaluating these models on out-of-distribution (OOD) datasets, the results are more varied. The inclusion of OOD evaluation is crucial, as an effective routing mechanism should ideally facilitate an expert decomposition that generalizes well. Notably, the MoE models that showed superior performance on ImageNet-1K also displayed enhanced robustness in the OOD evaluations.

5 Discussion

5.1 Hierarchical vs Isotropic Models

First, one can note that the best MoE designs typically apply experts only to the last layers. ViT operates at the same resolution at each layer after the input images have been cut into patches. This isotropic design implies that MoE is applied in layers that are not particularly large. On the other hand, with hierarchical

Table 8: Robustness of MoE w.r.t. to domain change, for two different pretrainings, both fine-tuned on ImageNet-1k: Top-1 accuracies on various datasets.

Model	IN-1K	IMV2	IMA	IMR	IMSK
			Imagenet1k trained		
ConvNext-T	**82.1**	70.8	**24.2**	**47.2**	**33.8**
ConvNext-T-4	**82.1**	**71.0**	23.8	46.2	32.7
ConvNext-S	**83.1**	**72.5**	**31.3**	**49.6**	**37.0**
ConvNext-S-4	**83.1**	72.2	30.2	49.0	37.3
ConvNext-B	**83.8**	**73.4**	**36.7**	**51.3**	**38.2**
ConvNext-B-4	83.5	72.8	33.9	48.6	36.6
ViT-S	79.9	68.8	19.8	43.4	29.7
ViT-S-8	**80.7**	**70.1**	**21.1**	**43.9**	**30.9**
ViT-B	**82.8**	**72.1**	**32.4**	**51.2**	**36.9**
ViT-B-8	82.5	71.5	32.0	46.6	35.2
ConvNext-S *(iso.)*	79.7	68.6	13.0	46.4	34.0
ConvNext-S-8 *(iso.)*	**80.3**	**68.8**	**13.5**	**46.6**	**34.4**
ConvNext-B *(iso.)*	**82.0**	**71.1**	**21.2**	**50.0**	**38.1**
ConvNext-B-8 *(iso.)*	81.6	70.6	19.1	48.5	35.7
			Imagenet21k pretrained		
ConvNext-T	82.9	72.4	**36.2**	51.1	38.5
ConvNext-T-8	**83.5**	**72.8**	32.6	**51.3**	**40.8**
ConvNext-S	84.6	74.7	**44.8**	57.5	43.6
ConvNext-S-8	**84.9**	**75.5**	44.4	**55.7**	**45.5**
ConvNext-B	**85.8**	76.0	**54.6**	**62.0**	**48.8**
ConvNext-B-8	85.7	**76.3**	51.8	59.6	48.4
ViT-S	82.6	72.6	**38.9**	50.8	39.0
ViT-S-8	**83.0**	**72.7**	35.3	**51.0**	**39.4**
ViT-B	85.2	**76.1**	**56.0**	**61.5**	**46.9**
ViT-B-8	85.2	75.4	48.1	59.3	45.9

models such as ConvNext, applying MoE to the last layers considerably increases model size. Indeed, as presented in Tab. 6, employing eight experts doubles the size of the network despite using only two layers of MoE.

5.2 Positions and Numbers of MoE Layers

There are two prevalent strategies in the literature concerning the position of MoE layer: "Every 2" and "Last 2". In the context of hierarchical architectures such as ConvNext, the "Every 2" strategy dramatically increases the number of weights while simultaneously yielding inferior results. However, for ViT "Last 2" and "Every 2" are two viable strategies. This is in line with the results of V-MoE (Riquelme et al., 2021) on large-scale datasets. Overall, our results confirm that "Last 2" is a solid starting point, as it consistently yields positive outcomes across all tested architectures. However, it may not always be the optimal choice for every architecture. For instance, in the case of ViT, "Every 2" demonstrates better results. Regarding the number of experts, we noted that model performance quickly saturates with the number of experts, as reproduced in Tab. 6, and detailed in Fig. 3.

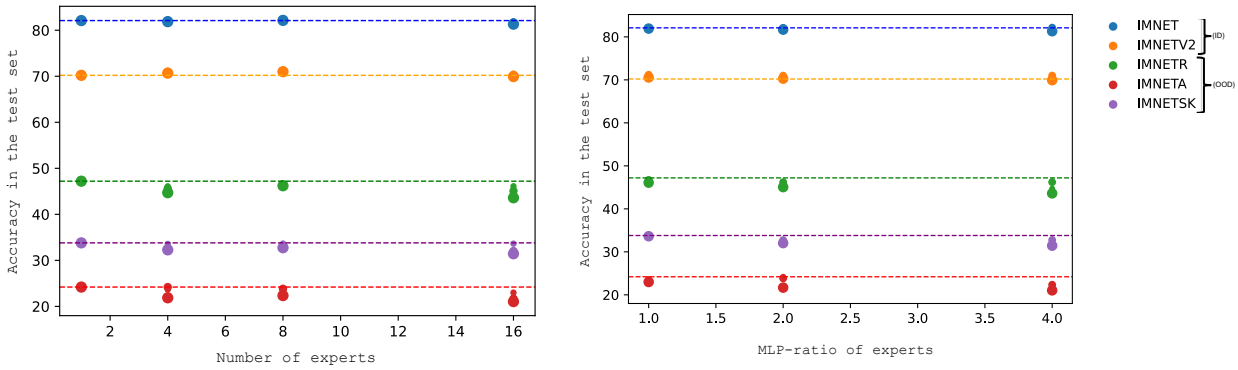
5.3 Robustness Evaluation

We evaluated MoE models, trained or fine-tuned on ImageNet-1k, against a range of robustness benchmark datasets, such as ImageNet-A (IMA) (Hendrycks et al., 2021b), ImageNet-R (IMR) (Hendrycks et al., 2021a), ImageNet-Sketch (IMSK) (Wang et al., 2019), and ImageNet V2 (IMV2) (Recht et al., 2019), the latter being used as a measure of overfitting.

Tab. 8 reports Top-1 accuracy for all datasets. It shows that MoE models tend to underperform their dense counterparts in ImageNet-1k. However, we note that MoE models are better for the robustness metric; they were actually already better on ImageNet-1k. So, these results actually emphasize the success of MoE when

the model is sufficiently small to be improved by a MoE, as discussed in the previous section: the domain generalization holds only in the regime in which MoE improves the base results.

5.4 Impact of Expert Size and Number



(a) # experts ($\in \{1, 4, 8, 16\}$) vs. performance. **Points' size**: MLP-ratio ($\in \{1, 2, 4\}$).

(b) Experts' MLP-ratio ($\in \{2, 4\}$) vs. performance. **Points' size**: # experts ($\in \{1, 4, 8, 16\}$).

Figure 3: Exploring the interplay between the size and count of experts in a MoE layer for ConvNext-T on ImageNet-1K. Baseline results (without MoE) are denoted by dotted lines. For this small dataset, MoE does not bring much improvement (see Fig. 2 for bigger datasets).

Two critical tunable parameters in the MoE layers are the number of experts per layer and the size of each expert. The latter is defined by the mlp_ratio, which scales the hidden dimension of the feed-forward network (FFN) (Sec. 3.1). Fig. 3 illustrates the sensitivity of ConvNeXt's performance to variations in these parameters, with the number of experts ranging from 4 to 16 and the mlp_ratio from 1 to 4. By default (original network), this ratio is set to 4, meaning that values below 4 correspond to smaller per-sample FFN capacities than in the original network.

To assess generalization, we report top-1 accuracy across three dataset categories: (i) in-distribution (ID) datasets—ImageNet and ImageNetV2 (Recht et al., 2019); (ii) out-of-distribution (OOD) datasets—ImageNet-R (Hendrycks et al., 2021a) and ImageNet-Sketch (Wang et al., 2019); and (iii) adversarial datasets—ImageNet-A (Hendrycks et al., 2021b). The objective is to determine whether performance benefits more from using many smaller experts or fewer larger ones.

Our experiments on ConvNeXt reveal that while MoE layers can slightly improve or preserve performance on ID datasets, they tend to degrade performance on OOD datasets. In particular, increasing the number of experts to 16 consistently reduces accuracy across all dataset types. Regarding expert size, smaller experts appear marginally more robust to increases in expert count. Overall, the best tradeoff for ID is achieved with 4 to 8 experts and an mlp_ratio of 2 or 4. Based on these findings, we adopt a configuration of 4 experts with an mlp_ratio of 2 for experiments on ImageNet-1K. For ConvNeXt isotropic and ViT architectures, however, reducing the mlp_ratio consistently harms performance, and in the case of ViT, it also introduces training instabilities. Also, OOD performance is typically degraded in Fig. 3.

5.5 On Data-Augmentation

As shown in the previous section, MoE is sensitive to the amount of data at hand. This sensitivity stems from the inherent structure of MoE, where increasing the number of experts results in each expert operating on a smaller data fraction. Consequently, each expert tends to learn specific data clusters, as discussed by (Chen et al., 2022).

Modern training methodologies often employ robust data augmentation techniques to enhance model generalization. However, the interplay between these augmentations and the data clusters learned by the MoE model remains relatively uncharted. For instance, techniques like Mixup (Zhang et al., 2018), which blend

Table 9: ViT-B trained on ImageNet-1k with different hyperparameters. In particular, for the top row (gray) we follow the training recipe from Deit (Touvron et al., 2021) and for the bottom one we follow Deit III (Touvron et al., 2022).

Architecture	# Params ($\times 10^6$)	Per sample $\#\text{Params}_{act}$	IN-1K Acc.
ViT-B-8 (Touvron et al., 2021)	86.6	86.6	81.8
ViT-B-8 Every 2 Top 2	284.9	129.9	**82.2**
ViT-B (Touvron et al., 2022)	86.6	86.6	**82.8**
ViT-B-8 Every 2 Top 2	284.9	129.9	82.5

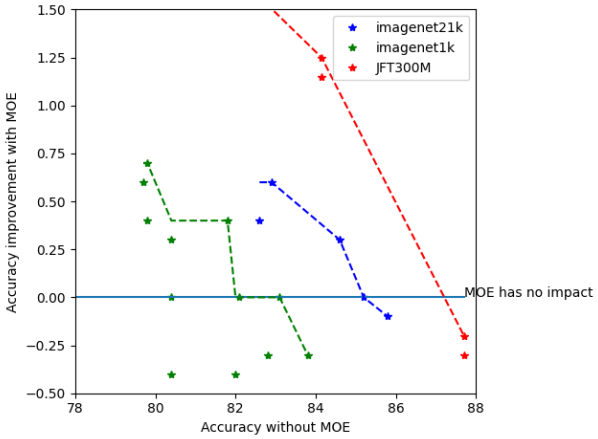


Figure 4: Improvement from MoE (y-axis) vs. baseline accuracy without MoE (x-axis) for image classification, across pretraining sizes. Larger pretraining yields greater gains; higher baseline accuracy reduces impact. Dashed lines show maximal improvement per accuracy level. Results are drawn from Tabs. 1, 2 and 11 (in App. B).

images, could blur the distinctions between these clusters. Similarly, the Random Resize method, with its aggressive magnification capability, may distort the original data clusters. Understanding the precise influence of these augmentations on the MoE model's learning process is essential, as it holds implications for the model's ultimate performance. Some evidence of this can be seen in Tab. 9. While one training recipe might show MoE outperforming its dense counterparts (as indicated in the top rows of the table), a different, more effective recipe could lead to the opposite outcome (Bottom rows of the table).

5.6 Alternate Views

With the ImageNet benchmark increasingly reaching saturation, the focus in contemporary computer vision models has shifted towards computational efficiency and scaling prowess. An essential question arises: *Does the MoE model enhance this aspect in image classification?* Our experiments with ImageNet reveal an intriguing insight: integrating MoE into models like ConvNext and ViT enhances performance, but this improvement tends to plateau in models with over 100M parameters. At first glance, this might seem to contradict findings from some papers (Riquelme et al., 2021; Mustafa et al., 2022). However, with additional context, this observation aligns with the existing literature, as evidenced by results in Figs. 2 and 4, and Tab. 11 (App. B). Specifically, as the base dense network grows larger, the performance gap between MoE and its dense counterpart narrows.

Exploring this further in large-scale datasets, such as JFT 3 billion (Zhai et al., 2022), reveals similar findings. For instance, the vision transformer model SoViT, with 400M parameters (Alabdulmohsin et al., 2023), exhibits comparable performance to the 15B V-MoE (Riquelme et al., 2021) when both use ViT-based architectures. Similarly, when considering the JFT-3B dataset (Zhai et al., 2022), the 5.6B LiMoE-H model (Mustafa et al., 2022) and Coca-Large (Yu et al., 2022) lead to analogous results when evaluated on a per-sample parameter basis. Interestingly, in the vision-language modeling domain, Lin et al. (2024) observe that increasing the number of experts for the image modality does not lead to diminishing returns, aligning with our findings regarding the efficacy of experts in scaling performance.

To summarize, while the MoE model does not appear to redefine the state-of-the-art when matched against dense models on both ImageNet and billion-scale datasets, its real strength emerges in enhancing smaller models. By integrating MoE, these compact models can achieve better performance, pushing their capabilities closer to the forefront of current benchmarks. Yet, it is essential to acknowledge that, in the broader context,

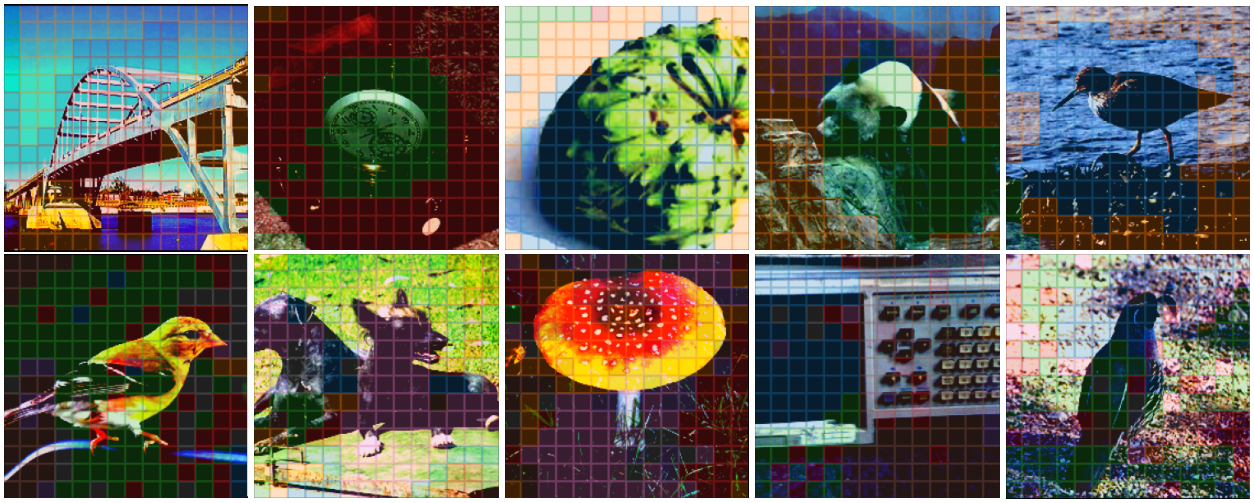


Figure 5: The top row features the ConvNext-T-4 model, while the bottom row showcases the ViT-S-8 model, both of which have been trained on the ImageNet1k dataset. The images displayed have been generated by upsampling the gating output, retaining only the top expert for clarity. Each distinct color represents a unique expert. These specific images are samples from the ImageNet validation set, extracted from the last third stage of the ConvNext-T-4 and the final layer of the ViT-S-8 models, respectively.

MoE might not drive significant advancements in the overall state-of-the-art. This nuanced perspective is further enforced in the appendix, with the data showcased in Tab. 11 in App. B.

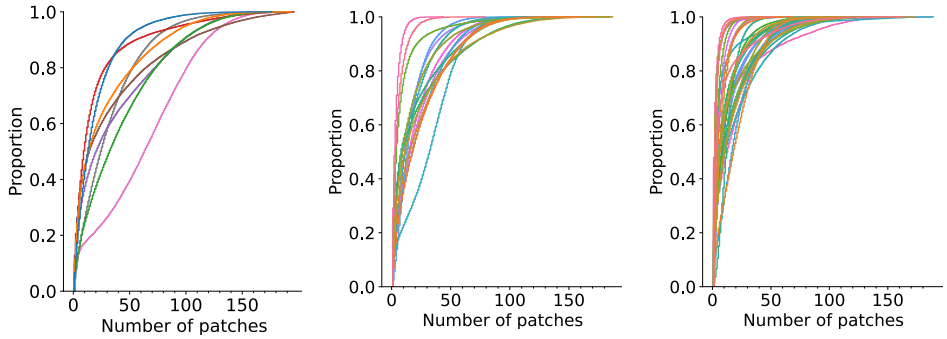
5.7 Model Inspection

The objective of this section is to examine the routing component of the model. We will present the top expert for each routing layer. Note that when considering the top-2 configuration, we only report, for this visualization, the top-1 of these two experts.

Visual inspection: Fig. 5 provides a visualization of the portions of images that are assigned to the different experts. This reveals that some experts specialize in specific elements of an image, such as animals, buildings, or objects. This trend is especially apparent when considering the top-1 MoE, but the distinction becomes less clear with the top-2 MoE. In both instances, the clusters formed by each expert are challenging to interpret, and the spatial locations of each expert tend to lack clarity and can be quite ambiguous.

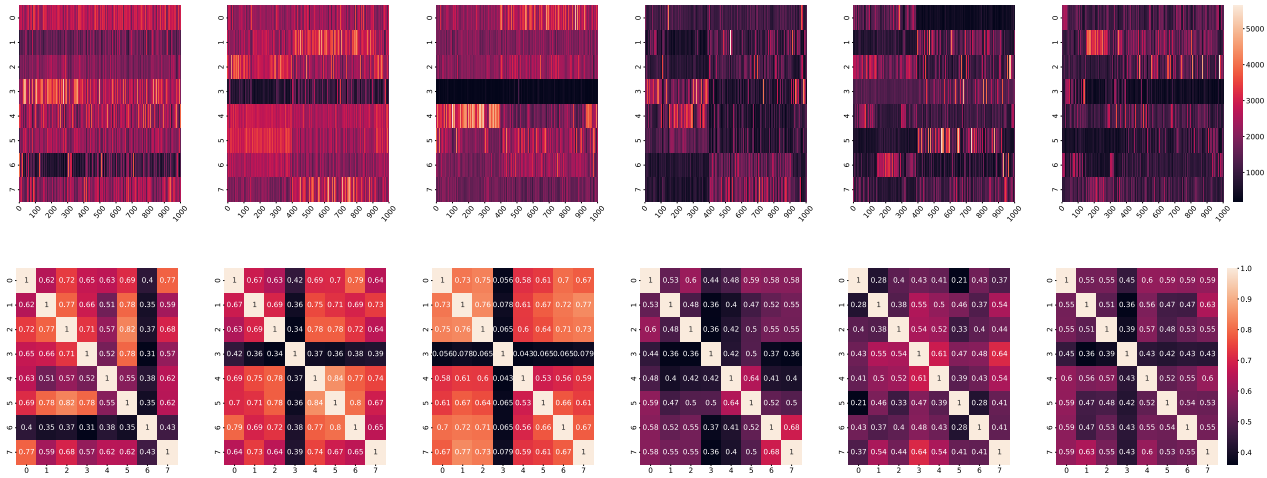
Number of experts involved per image: We quantified the average number of experts contributing to processing an image. In the initial layers, the majority of experts are typically engaged for each image. With fewer than eight experts, most are utilized, even in deeper layers. For instance, in the ViT model, with eight experts, an average of seven experts contributed per image, while in the ConvNext model, there are six. As we increase the number of available experts per MoE layer, the number of contributing experts per image also rises. For example, in ConvNext, with 16 experts, an average of 10.5 experts are involved, and with 32 experts, this number increases to 15.8. This suggests that each expert is likely focused on a limited number of image patches. This observation is confirmed by the cumulative distribution function (CDF) of each expert per image, as depicted in Fig. 6

Correlation between experts and labels: Fig. 7 (top) presents the correlation between the number of expert occurrences and individual classes in ImageNet-1k. The class IDs in ImageNet-1k are organized in such a way that adjacent classes often share similar attributes. For instance, class IDs ranging from 151 to 268 are all dedicated to different dog breeds. From Fig. 7, we note that the occurrences in the first three MoE layers do not match the ImageNet-1k classes, with most experts appearing uniformly across different classes. However, in the last three MoE layers, there is a noticeable trend of more experts aligning with specific ImageNet-1k classes. This distinction is especially pronounced between experts focusing on animals (classes up to ID 397, such as experts 0 and 6 in Fig. 7-(e)) and those centered on objects (classes beyond



(a) 8 experts (b) 16 experts (c) 32 experts

Figure 6: Cumulated Distribution Function of the number of patches used per image for ConvNext-S trained on ImageNet-21k, with 8, 16, and 32 experts. Each colour corresponds to one expert. Sample detailed interpretation: for 8 experts, the "red" expert sees between 1 and 30 patches (out of 196) for 80% of the images on which it is active. In particular (and even more so when there are many experts), an expert frequently covers a limited part of the image. For instance, if the y-axis is close to 1 at x-axis=100 patches, covering more than 100 patches over 196 is rare: experts don't specialize much. As the number of experts increases, this tendency of experts being used only on a few patches increases too. For 32 experts, most of them focus on less than 30 patches, 80% of the time they are active.



(a) (b) (c) (d) (e) (f)

Figure 7: Routing analysis of ViT-S-8, trained on ImageNet-1k. Top row: per-class occurrence of the experts, with the abscissa indicating the class ID and the ordinate representing the expert index (8 experts). Bottom row: similarity score between the experts. The graphs are arranged from left to right from the MoE layer closest to the input (a) to the layer closest to the output (f).

ID 397, like experts 1 and 5 in Fig. 7-(e)). As the number of experts increases, this class-specific alignment still holds, though it remains challenging to clearly define the roles of these experts since they often span several scattered classes rather than forming tight, contiguous groups that would fit each class, as illustrated in Fig. 8.

Expert Similarity: In order to quantify the similarity between experts, we use a method derived from the Jaccard index. For each image, we count the occurrences c_i of each expert E_i. The similarity between two experts, E_i and E_j, is defined as $S_{ij} = \frac{|E_i \cap E_j|}{|E_i \cup E_j|} = \frac{\sum^{E_i \cap E_j} c_i + c_j - |c_i - c_j|}{\sum^{E_i \cup E_j} c_i + c_j}$, where $E_i \cap E_j$ represents the images where both E_i and E_j appear together, while $E_i \cup E_j$ refers to all images where either E_i or E_j is present

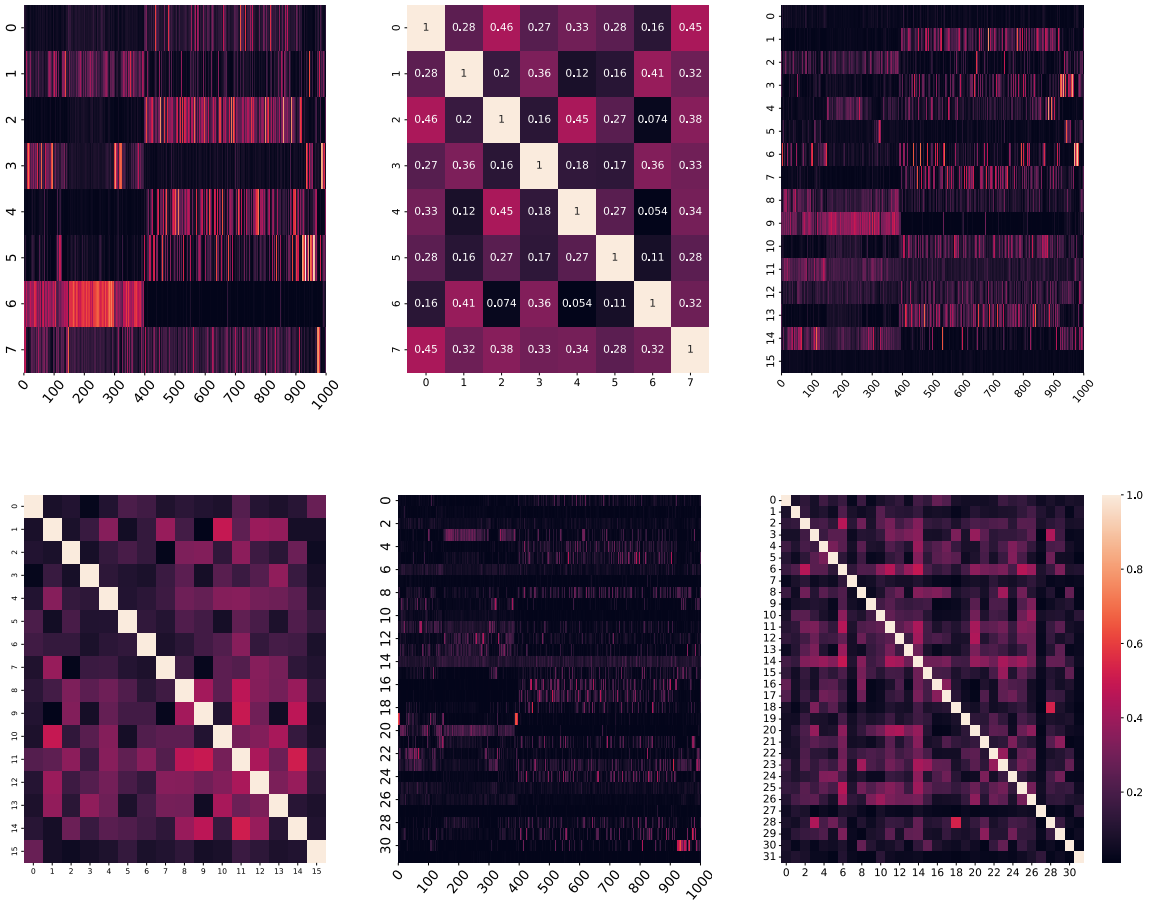


Figure 8: ConvNext-S trained on ImageNet-21k with 8, 16, and 32 experts trained on ImageNet-21k routing analysis. The right columns display the frequency of each expert across different classes, where the x-axis represents class IDs, and the y-axis denotes the expert number. The left column provides a similarity score between experts.

at least once. The absolute value term $|c_i - c_j|$ acts as a corrective factor to account for scenarios where one expert may be deployed for a minimal number of patches while the other is utilized for the majority, indicating a lack of synergy between the two experts. In the bottom panel of Fig. 7, we observe a distinct pattern in the utilization of experts across the stages of the MoE. In the initial layers, there is a prevalent co-occurrence of multiple experts, which highlights a concentration on local image features across all the images. Conversely, as we analyze the network deeper, the frequency of expert co-occurrence diminishes. This shift indicates a transition in focus from local image elements to more semantically rich aspects, with individual experts concentrating on specific semantic attributes unique to certain images.

Are experts well partitioned and clustered? For a general task such as ImageNet-1k, there is no clear, natural decomposition into experts, making it difficult to precisely assess if one expert decomposition is better than another, aside from comparing final accuracy. This comparison, however, provides limited insight into the quality of the experts. The literature (Chen et al., 2023; Mittal et al., 2022; Mohammed et al., 2022) suggests that, without explicit enforcement, MoE models struggle to separate tasks or find natural clusters beneath them. For instance, when a task separation was explicitly incorporated into the loss, there was a marked improvement in both accuracy and the ease with which experts could be pruned for specific tasks (Chen et al., 2023). Similarly, Mittal et al. (2022) found that the enforced decomposition resulted in

substantial gains, while the absence of such enforcement left the model unable to discover optimal partitions. Consequently, we make three observations on ImageNet. (i) As the number of experts increases, there is a corresponding rise in the number of experts actively involved. Even with 32 experts, almost half of the experts are used per image in the last MoE layers. This seems counterintuitive since ImageNet is centered around objects; at least for deep layers, it should plateau at some point. Instead, experts are not well aligned with classes and spread widely, as depicted in Fig. 7. (ii) An analysis of the cumulative distribution function (CDF) of experts' engagement (Fig. 6) reveals a prevalent trend in which most experts are assigned to less than 20 patches of an image, equating to less than 10% of the total image area, 50% of the time. Moreover, these patches are not necessarily contiguous. (iii) When we look at the visualization of each expert, like in Fig. 5, we observe a concentration of experts in the image for MoE layers near the output. The interpretation of each expert's role becomes challenging: An expert does not seem to adhere closely to an object, or to a part of the location of an object.

6 Conclusion

Implementing effective Mixture-of-Experts (MoE) models for image classification tasks remains a challenging and open problem. Our experiments show that, particularly in large-scale models, MoE for image classification offers limited gains over state-of-the-art methods in accuracy, robustness, generalization, or per-sample efficiency across both convolutional networks and vision transformers. This contrasts with NLP tasks, where MoE improves performance by enhancing capacity for knowledge storage and retrieval. In image classification, which relies more on feature extraction and extrapolation, MoE models show only modest benefits, mainly in smaller models with fewer activated parameters. These findings suggest that the effectiveness of MoE in image classification tasks is highly context- and scale-dependent, with advantages diminishing as model capacity grows.

References

Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting ViT in shape: Scaling laws for compute-optimal model design. *arXiv:2305.13035*, 2023. 10, 18

Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11828–11837, 2023. 13

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. *arXiv:2208.02813*, 2022. 9

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, et al. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 35:34600–34613, 2022. 7

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv:2207.04672*, 2022. 1

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020. 4

A. Dosovitskiy, L. Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, M. Dehghani, Matthias Minderer, G. Heigold, S. Gelly, Jakob Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2020. 1, 3

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022. 1

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a. 8, 9

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b. 8, 9

Heinke Hihn and Daniel A Braun. Mixture-of-variational-experts for continual learning. *arXiv:2110.12667*, 2021. 1

Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, HoYuen Chau, Peng Cheng, Fan Yang, Mao Yang, and Yongqiang Xiong. Tutel: Adaptive mixture-of-experts at scale. In D. Song, M. Carbin, and T. Chen (eds.), *Machine Learning and Systems*, volume 5, pp. 269–287, 2023. 2, 18

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 2

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, volume 1, pp. 4171–4186, 2019. 1

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *arXiv:2212.05055*, 2022. 1

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv:2006.16668*, 2020. 1

Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv:2407.21770*, 2024. 10

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. 1, 3, 4, 5, 17, 18

Yuxuan Lou, Fuzhao Xue, Zangwei Zheng, and Yang You. Cross-token modeling with conditional computation. *arXiv:2109.02008*, 2021. 2

Sarthak Mittal, Yoshua Bengio, and Guillaume Lajoie. Is a modular architecture enough? *Advances in Neural Information Processing Systems*, 35:28747–28760, 2022. 13

Muqeeth Mohammed, Haokun Liu, and Colin Raffel. Models with conditional computation learn suboptimal solutions. In *I Can't Believe It's Not Better Workshop: Understanding Deep Learning Through Empirical Falsification*, 2022. 13

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with LIMoE: the language-image mixture of experts. *arXiv:2206.02770*, 2022. 1, 10

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *12th International Conference on Learning Representations*, 2024. 2

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet ? In *International Conference on Machine Learning*, pp. 5389–5400, 2019. 8, 9

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 1, 2, 3, 8, 10, 18

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv:1701.06538*, 2017. 1, 2, 3

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv:1906.02243*, 2019. 1

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. MLP-Mixer: An all-MLP architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 2

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021. 4, 10

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit III: Revenge of the ViT. In *17th European Conference Computer Vision*, pp. 516–533, 2022. 1, 4, 10, 17, 18

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European Conference on Computer Vision*, pp. 459–479. Springer, 2022. 4

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, 2017. 1, 3

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019. 8, 9

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv:2205.01917*, 2022. 10

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision*, 2019. 4

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022. 10, 18

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 4, 9

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *AAAI Conference on Artificial Intelligence*, 34(07):13001–13008, 2020. 4

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. Designing effective sparse expert models. *arXiv:2202.08906*, 2022. 1