

CNN Mixture-of-Depths

Rinor Cakaj^{1,2}, Jens Mehnert¹, and Bin Yang²

¹ Robert Bosch GmbH, Daimlerstrasse 6, 71229 Leonberg, Germany
{Rinor.Cakaj, JensEricMarkus.Mehnert}@de.bosch.com

² University of Stuttgart, Pfaffenwaldring 47, 70550 Stuttgart, Germany
bin.yang@iss.uni-stuttgart.de

Abstract. We introduce Mixture-of-Depths (MoD) for Convolutional Neural Networks (CNNs), a novel approach that enhances the computational efficiency of CNNs by selectively processing channels based on their relevance to the current prediction. This method optimizes computational resources by dynamically selecting key channels in feature maps for focused processing within the convolutional blocks (Conv-Blocks), while skipping less relevant channels. Unlike conditional computation methods that require dynamic computation graphs, CNN MoD uses a static computation graph with fixed tensor sizes which improve hardware efficiency. It speeds up the training and inference processes without the need for customized CUDA kernels, unique loss functions, or fine-tuning. CNN MoD either matches the performance of traditional CNNs with reduced inference times, GMACs, and parameters, or exceeds their performance while maintaining similar inference times, GMACs, and parameters. For example, on ImageNet, ResNet86-MoD exceeds the performance of the standard ResNet50 by 0.45% with a 6% speedup on CPU and 5% on GPU. Moreover, ResNet75-MoD achieves the same performance as ResNet50 with a 25% speedup on CPU and 15% on GPU.

Keywords: CNN, Mixture-of-Depths, Computational Efficiency, Inference Speed

1 Introduction

Over recent years, convolutional neural networks (CNNs) have demonstrated significant advancements in a variety of computer vision applications, such as image recognition [11, 34], object detection [28, 29], and image segmentation [23, 43]. Despite their remarkable performance, CNNs often require substantial computational power and extensive memory usage, posing considerable challenges when deploying advanced models on devices with limited computational resources [19].

Pruning techniques are widely utilized to reduce the model size and computational demands of CNNs by removing redundant weights or filters according to established criteria [12–14, 20, 24, 40, 44, 45]. However, these methods uniformly process all inputs, failing to adjust for the varying complexities of different inputs, which can lead to performance decreases [19].

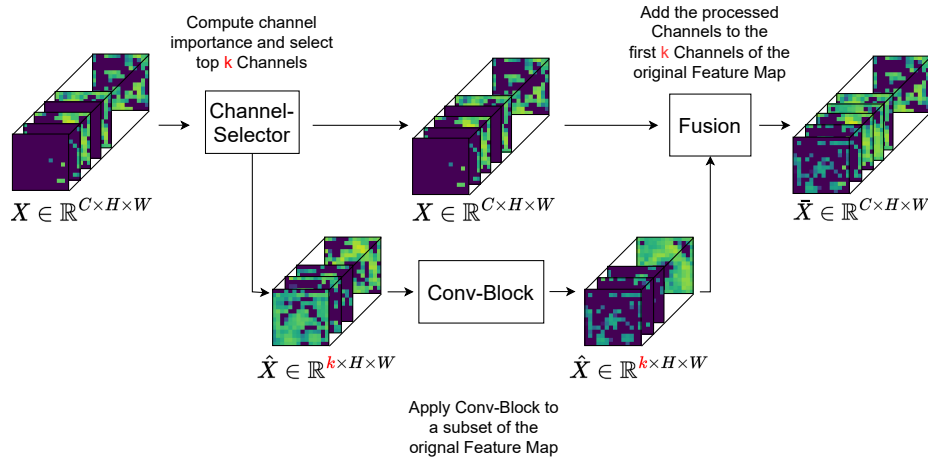


Fig. 1: Illustration of the CNN MoD mechanism, which starts with the Channel Selector module. This module computes the importance scores of each channel in the input feature map, $X \in \mathbb{R}^{C \times H \times W}$, and selects the top k channels for focused processing in the Conv-Block. These selected channels are then processed by a Conv-Block designed to operate on a reduced dimension, $\hat{X} \in \mathbb{R}^{k \times H \times W}$, enhancing computational efficiency. The processed channels are added to the first k channels of the original feature map through a fusion operation, instead of being added back to their original positions. This selective reintegration of refined channels with the unprocessed channels helps to preserve the dimensions of the original feature map ($X \in \mathbb{R}^{C \times H \times W}$).

Dynamic computing, or conditional computation, adapts computational resources to the complexity of inputs to enhance efficiency [19,36,37]. However, the integration of these methods into hardware is challenging due to their reliance on dynamic computation graphs, which are often incompatible with systems optimized for static computation workflows [8,37]. For instance, Wu et al. [41] report increased processing times when layers are conditionally executed via a separate policy network. Despite theoretical reductions in computational demands, practical gains on hardware like GPUs or FPGAs are limited, as non-uniform tasks can interfere with the efficiencies of standard convolution operations [18,33]. Moreover, Ma et al. [25] show that floating point operations (FLOPS) are insufficient for estimating inference speed, as they often exclude element-wise operations like activation functions, summations, and pooling.

To combine the performance benefits of dynamic computing with the operational efficiency of static computation graphs, we present the CNN Mixture-of-Depths (MoD), inspired by the Mixture-of-Depths approach for Transformers [27]. This innovative approach combines the performance benefits of dynamic computing with the operational efficiency of static computation graphs. The fundamental principle of CNN MoD is that only a selected subset of channels within the feature maps is essential for effective convolutional processing at given lay-

ers in the network. By focusing on these essential channels, CNN MoD boosts computational efficiency without reducing network performance.

The MoD mechanism, shown in Figure 1, enhances CNN feature map processing by dynamically selecting the most important channels for focused computation within Conv-Blocks. MoD begins with the Channel Selector, illustrated in Figure 2, which is designed similarly to the Squeeze-and-Remember block [15]. It evaluates the importance of each channel within the input feature map $X \in \mathbb{R}^{C \times H \times W}$. Adaptive average pooling first reduces each channel to $\tilde{X} \in \mathbb{R}^{C \times 1 \times 1}$, which is then processed by a two-layer fully connected neural network with a bottleneck design. A sigmoid activation function generates scores indicating the importance of each channel.

Following this, the Channel Selector uses a top k selection mechanism to identify the k most crucial channels based on the computed importance scores. These channels are sent to the Conv-Block, designed to operate on the reduced dimension $\hat{X} \in \mathbb{R}^{k \times H \times W}$, thus improving computational efficiency. To ensure that the gradients from the processed output \hat{X} effectively optimize the channel selection process, the processed channels are scaled by their respective importance scores. This allows the gradients to flow back to the Channel Selector, enabling the learning of channel importance throughout training.

The final step involves a fusion operation, in which the processed channels are added to the first k channels of the original feature map. This fusion not only preserves the original dimensions of the feature map but also enhances feature representation by combining processed and unprocessed channels.

The reduction in the number of processed channels within Conv-Blocks is controlled by a hyperparameter $c \geq 1$, where $k = \lfloor \frac{C}{c} \rfloor$ defines the number of channels to process. Here, C is the total number of input channels in the Conv-Block. For example, in a typical ResNet [11] architecture, a Bottleneck block that usually processes 1024 channels will only process 16 channels ($k = 16$) when $c = 64$. This selective processing significantly reduces the computational load by focusing on a smaller subset of channels. Additionally, the sizes of the kernels within the Conv-Blocks are adjusted to match the reduced number of input channels, as detailed in Section 3.2.

Empirical evaluations indicate that the optimal integration of MoD within the CNN architecture involves alternating them with standard Conv-Blocks. In architectures like ResNets [11] or MobileNetV2 [31], Conv-Blocks are organized into modules containing multiple Conv-Blocks of the same type (i.e., with the same number of output channels). Each module begins with a standard block, and is then followed by an MoD Block. This alternating arrangement is based on the principles of the Mixture-of-Depths for Transformers [27]. This does not imply adding additional MoD Blocks to the existing sequence. Instead, every second Conv-Block in the original architecture is replaced by a MoD Block, ensuring the overall depth of the architecture remains unchanged.

CNN MoD achieves performance comparable to traditional CNNs but with reduced inference times, GMACs, and parameters, or it surpasses them while maintaining similar inference times, GMACs, and parameters. On the ImageNet

dataset [30], our ResNet75-MoD matches the accuracy of the standard ResNet50 [11] and provides speed-ups of 15% on GPU and 25% on CPU. Similar results can also be achieved in semantic segmentation and object detection tasks, as detailed in Section 4.

2 Related Work

This section reviews two strategies for enhancing CNN computational efficiency: static pruning and dynamic computing. At the end, we highlight the advantages of our CNN MoD approach, which combines the benefits of both strategies, offering improvements over traditional methods.

2.1 Static Pruning

Static pruning techniques aim to reduce the computational burden of CNNs by eliminating redundant model parameters. Early work in this area primarily targeted weight pruning, which selectively removes individual weights that have little effect on the model’s output [9, 10, 39]. However, these methods tend to create irregular sparsity patterns that do not align well with hardware optimizations [19]. More recently, structured pruning approaches like channel pruning have gained prominence. These methods offer a more hardware-compatible form of sparsity by discarding entire channels based on their assessed importance [13, 21]. For example, FPGM, or Filter Pruning via Geometric Median, identifies and removes filters that are closest to the geometric median within a layer, targeting those considered less crucial [13]. Similarly, HRank evaluates filters based on the rank of their generated feature maps, pruning those that contribute the least to the output’s information content [21].

Despite their benefits, these static pruning methods permanently remove computations, potentially reducing the model’s capacity to represent complex features. This permanent reduction can be particularly limiting for complex images that may require more detailed processing, suggesting that a one-size-fits-all approach to pruning might be suboptimal for diverse real-world applications. Additionally, these methods often require fine-tuning or specialized loss functions, further complicating their implementation.

2.2 Dynamic Computing

Dynamic computing, or conditional computation, dynamically adjusts computational resources according to the complexity of the input, potentially maintaining high model accuracy while reducing computation. Architectures like BranchyNet [35] and MSDNet [16] implement early exits for simpler inputs to decrease average computational load. Furthermore, policy-driven methods such as BlockDrop [41] and GaterNet [3] dynamically decide which network blocks to execute using a policy network, which adapts in real-time to the input.

Further innovations in dynamic computing utilize gating functions to enable selective processing of Conv-Blocks. SkipNet [38] and ConvNet-AIG [36] dynamically skip the processing of whole Conv-Blocks based on the observation that individual blocks can be removed without interfering with other blocks in the residual networks [11]. Some methods [6, 37, 42] exploit spatial sparsity to reduce the computations, e.g. Verelst et al. [37] learn a pixel-wise execution mask for each block and only calculate on those specified locations. Dynamic Dual Gating [19] introduces another layer of complexity by identifying informative features along spatial and channel dimensions, allowing the model to skip unimportant regions and channels dynamically during inference.

Many works on conditional execution primarily highlight reductions in theoretical complexity. However, in practical settings, the application of these dynamic methods often confronts hardware limitations due to their dependency on dynamic computation graphs. This reliance can lead to inefficiencies and extended execution times on hardware systems that are optimized for static computations [8, 18, 33, 41].

Recently, Raposo et al. [27] introduced the Mixture-of-Depths for Transformers, a method that selectively processes tokens within Transformer blocks to effectively reduce computational load. This innovative approach inspired us for our CNN MoD approach.

2.3 CNN MoD: Combining Static and Dynamic Advantages

The CNN MoD approach combines the benefits of static pruning and dynamic computing within a unified framework. Below are the key advantages of our approach:

- **Static Computation Graph:** CNN MoD retains a static computation graph, which enhances both training and inference time.
- **No Custom Requirements:** Does not require customized CUDA kernels, additional loss functions, or fine-tuning.
- **Dynamic Resource Allocation:** Dynamically allocates computational resources to channels based on their importance, effectively optimizing training and inference speeds.

3 Method: CNN Mixture-of-Depths

The term “Mixture-of-Depths” in our approach refers to the selective processing strategy where not every channel is processed in every convolutional block within the same module, resulting in varied processing depths. This strategy allows for some channels to be processed more frequently than others within the same module, thereby optimizing computational resources and enhancing feature extraction efficiency. This section presents the CNN Mixture-of-Depths approach. It consists of three main components:

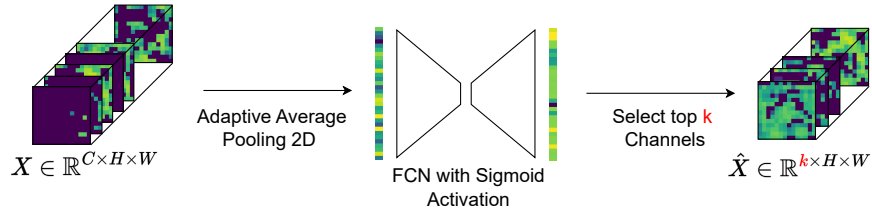


Fig. 2: Illustration of the Channel Selection Process in MoD. The process begins with an input tensor $X \in \mathbb{R}^{C \times H \times W}$, which undergoes adaptive average pooling to reduce spatial dimensions to 1×1 , preserving channel information. The pooled output is processed through a two-layer fully connected network with a reduction factor ($r = 16$), followed by a sigmoid activation to generate channel-wise scores. These scores are used to select the top k channels. This forms a subset of the original tensor with reduced channel dimension but original spatial dimensions.

1. **Channel Selector:** This component selects the top k most important channels from the input feature map based on their relevance to the current prediction. This selection helps to focus computational resources effectively by processing only those channels that are crucial for the current task.
2. **Convolutional Block:** The selected channels are then processed in a Conv-Block. This block can be adapted from existing architectures such as ResNets [11] or ConvNext [22] and is designed to enhance the features of the selected channels. To ensure that the Channel Selector is optimized during training, the processed feature maps are multiplied by their importance scores at the end of this block.
3. **Fusion Operation:** After processing in the Conv-Block, the enhanced channels are integrated back into the original feature map. This is done by adding the processed channels to the first k channels of the feature map, which is then passed on to the subsequent layer without additional adjustments. This step ensures that the processed features are preserved and that the feature map maintains its original dimensions for further processing.

3.1 Channel Selector

Figure 2 shows how the Channel Selector dynamically selects the most significant channels for processing in the Conv-Block. It operates in two main stages:

1. **Adaptive Channel Importance Computation:** Initially, the Channel Selector compresses the input feature map $X \in \mathbb{R}^{C \times H \times W}$ using adaptive average pooling, reducing its dimension to $\tilde{X} \in \mathbb{R}^{C \times 1 \times 1}$. This compressed feature map is processed through a two-layer fully connected network with a bottleneck design, set to a ratio $r = 16$, and concludes with a sigmoid activation to create a score vector $\mathbf{s} \in \mathbb{R}^C$. Each element of \mathbf{s} quantifies the importance of the corresponding channel.

- 2. Top-k Channel Selection and Routing:** Utilizing the importance scores s , the Channel Selector selects the top k channels. These selected channels are routed to the Conv-Block for processing. The original feature map X is routed around the Conv-Block to the fusion process described in Section 3.3.

This selection process enables the Channel Selector to effectively manage computational resources while keeping a static computational graph. This facilitates the dynamic selection of channels.

3.2 Channel Processing Dynamics

After the Channel Selector selects the top k channels, they are processed within the Conv-Block. This block is adaptable from architectures such as ResNets [11] or ConvNext [22], and is designed to process a reduced number of channels.

The number of channels k processed in each Conv-Block is determined by the formula $k = \lfloor \frac{C}{c} \rfloor$, where C represents the total input channels of the block, and c is a hyperparameter determining the extent of channel reduction. For instance, in a standard ResNet [11] Bottleneck block that typically processes 1024 channels, setting $c = 64$ reduces the processing to only 16 channels ($k = 16$). Unlike the standard bottleneck flow where channels transition from 1024 to 256 and back to 1024, in a MoD Bottleneck block, the channel dimensions are significantly reduced from 16 to 4 and then back to 16, focusing computational efforts on the most essential features and enhancing efficiency. In our empirical investigations, we found that the hyperparameter c should be set to the maximal number of input channels in the first Conv-Block and remain the same in every MoD Block throughout the CNN. For example, $c = 64$ for ResNet [11] and $c = 16$ for MobileNetV2 [31].

The final step in the Conv-Block involves multiplying the processed channels with the importance scores obtained from the Adaptive Channel Importance Computation. This step ensures that gradients are effectively propagated back to the Channel Selector during training, which is needed for optimizing the selection mechanism.

3.3 Fusion Mechanism

The Fusion Mechanism combines processed and unprocessed channels to maintain the dimensionality required for subsequent convolutional operations. This approach ensures that the remaining non-selected channels, which contain useful features for later stages, are preserved, thereby maintaining a comprehensive feature set necessary for the model’s performance.

Consider the original input feature map X with dimensions $\mathbb{R}^{C \times H \times W}$, where C represents the total number of channels, and H and W denote the height and width of the feature map, respectively. The processed feature map \hat{X} , resulting from the adapted Conv-Block applied to the selected top k channels, has dimensions $\mathbb{R}^{k \times H \times W}$. To integrate \hat{X} with X , the fusion operation is performed as

follows:

$$\bar{X}[:, : k, :] = \hat{X} + X[:, : k, :], \quad (1)$$

$$\bar{X}[:, k :, :] = X[:, k :, :]. \quad (2)$$

where (1) adds processed features to the first k channels of X , and (2) maintains the remaining unprocessed channels. This formula confirms that \bar{X} , the feature map after fusion, has the same number of channels C as the original input X , preserving the dimensions needed for subsequent layers.

In our experiments, we tested various strategies for integrating the processed channels back into the feature map X . These strategies included adding the processed channels back to their original positions and completely replacing the original channels with the processed ones. However, empirical results did not show improvements with either method. It seems beneficial for the network’s performance to consistently use the same positions within the feature maps for processed information. This observation is confirmed by experiments where the processed channels were added to the last k channels of the feature map X , yielding results comparable to those achieved when added to the first k channels (see Appendix A.5).

3.4 Integration in CNN Architecture

MoD can be integrated into various CNN architectures, such as ResNets [11], ConvNext [22], VGG [32], and MobileNetV2 [31]. These architectures are organized into modules containing multiple Conv-Blocks of the same type (i.e., with the same number of output channels). Through our experiments, we found out that alternating MoD Blocks with standard Conv-Blocks in each module is the most effective integration method. We also explored using MoD Blocks in every block and selectively within specific modules, however, the alternating strategy proved to be the most effective approach. It is important to note that MoD Blocks replace every second Conv-Block, maintaining the original architecture’s depth (e.g., 50 layers in a ResNet50 [11]). Each module starts with a standard block, such as a BasicBlock [11], followed by an MoD Block. This alternating pattern indicates that the network can handle substantial capacity reductions, provided that there are regular intervals of full-capacity convolutions. Furthermore, this method ensures that MoD Blocks do not interfere with the spatial dimension-reducing convolutions that typically occur in the first block of each module.

4 Experiments

This section evaluates the CNN Mixture-of-Depths on various tasks, including supervised image recognition on CIFAR-10/100 [17] and ImageNet [30], semantic segmentation on Cityscapes [5], and object detection on Pascal VOC [7]. Results for CIFAR-10/100 are detailed in Appendix D due to space constraints.

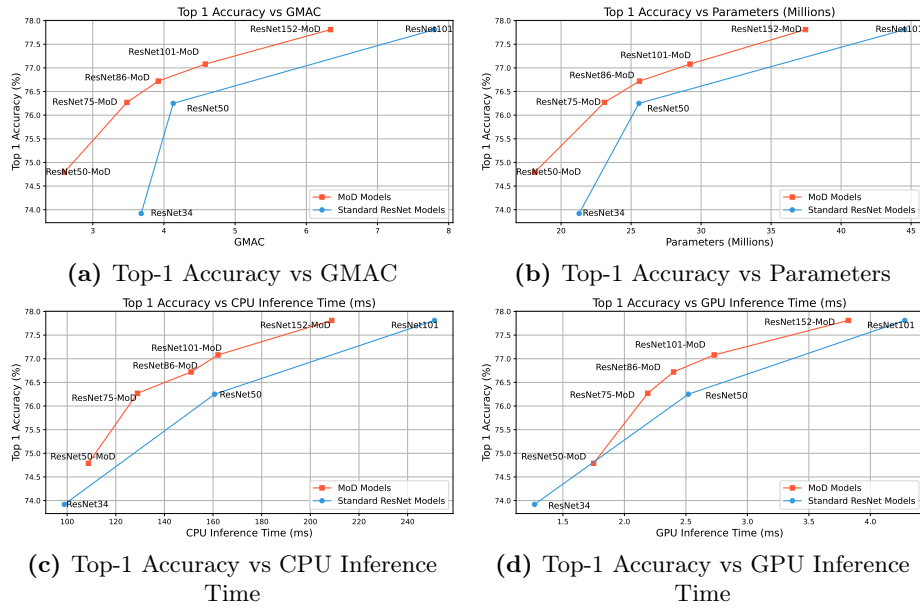


Fig. 3: ResNet MoD models outperform standard ResNets under similar computational constraints, as shown across four panels. Panel (a) shows the higher accuracy per GMAC, highlighting better computational efficiency, while panel (b) illustrates the improved parameter efficiency. Panels (c) and (d) demonstrate ResNet MoD’s superior Top-1 accuracy with comparable or faster inference times on CPU and GPU.

Performance benchmarks were carried out on both GPU and CPU platforms to evaluate the running times. These measurements were taken using the Timer and Compare functions from the `torch.utils.benchmark` package, with tests performed on an Intel(R) Core(TM) i7-10850H CPU @ 2.70GHz and a NVIDIA Quadro RTX 3000 GPU.

4.1 Image Recognition on ImageNet

We evaluated the CNN MoD on the ImageNet dataset [30], comprising 1.2 million training images, 50,000 validation images, and 150,000 test images across 1,000 categories. The experiments were conducted using three different random seeds to ensure the robustness and reproducibility of the results. Implementation details and further results are provided in Appendix A.

Table 1 presents the comparative results for various network configurations, highlighting our approach against advanced pruning and dynamic computing strategies. This includes comparisons with established methods such as DGNet [19], Batch-Shaping [1], ConvNet-AIG [36], HRANK [21], FPGM [13], and DynConv [37]. We report for all models the top-1 classification accuracy on the validation set, along with attributes such as no requirement for fine-tuning (No

FT), simple training mechanisms (Simple Train.), the availability of a GPU implementation (GPU Impl.), a static computational graph (Static Graph), and enhanced training speed (Faster Train.). Here, "Simple Training" indicates that no customized loss functions or complex hyperparameter tuning is required, and "Faster Training" is determined by measuring the relative training time compared to standard models.

The MoD Blocks reduce the computational load in every second block, allowing the construction of deeper architectures with the computational costs comparable to less deep standard models. Specifically, in the ResNet86-MoD, the number of Conv-Blocks in the third layer is 18, compared to just 6 blocks in the standard ResNet50, while maintaining nearly the same GMACs, parameters, and inference times. This expansion is enabled by the MoD Blocks that process fewer channels. Similarly, the ResNet75-MoD increases the number of Conv-Blocks in the third layer to 14. For further details, see Appendix A.2.

As described in Section 3.2, the value of c in all MoD Blocks is set to the maximum number of input channels in the first Conv-Block. For instance, $c = 64$ for ResNet [11] and $c = 16$ for MobileNetV2 [31].

The MoD models achieve performances comparable to traditional CNNs but with reduced inference times or they surpass them while maintaining similar inference times. For instance, our ResNet75-MoD matches the performance of the standard ResNet50 and achieves an 15% speed-up on GPU and 25% on CPU. Similarly, the ResNet50-MoD model significantly enhances processing speeds, achieving a 27% improvement on CPU and 48% on GPU, with a slight impact on accuracy. Furthermore, the ResNet86-MoD model improves accuracy by 0.45% with a speedup of 6% on CPU and 5% on GPU. A test variant, MoD50 rand, employing randomized channel selection, exhibited a performance decline of about 0.8%. This highlights the significance of strategic channel selection in our MoD approach. Figure 3 presents a comparative performance analysis of ResNet MoD models against standard ResNets for different computational metrics.

Dynamic approaches like DGNNet [19] and ConvNet-AIG [36], which lack data on real-world inference speedup, show simulated GMAC reductions only, indicating potential discrepancies in practical applications.

The results presented in Table 2 demonstrate the effectiveness of applying the MoD approach to MobileNetV2 architectures. The MobileNetV2-MoD-L model, with a deeper configuration (see Appendix A.3), maintains nearly the same Top-1 Accuracy as the standard MobileNetV2, while achieving a 11% speed-up on CPU and 10% on GPU. The standard MobileNetV2-MoD, despite a slight decrease in accuracy, offers a significant speed-up of 43% on CPU and 39% on GPU, illustrating the benefits of the MoD approach in balancing performance with computational efficiency.

4.2 Semantic Segmentation on Cityscapes

For semantic segmentation, we utilized the Cityscapes dataset [5], which comprises high-quality, finely annotated images from 50 European cities, divided into 19 semantic classes (2,975 training, 500 validation, and 1,525 test images). The

Table 1: This table presents the MoD models against other dynamic pruning and computational efficiency strategies. The table evaluates each method based on Top-1 Accuracy (Acc %), computational complexity (GMAC), model size (Params, in millions - M) and Speed-up for CPU and GPU (SpeedUp CPU, SpeedUp GPU). Models abbreviated as “R50” refer to ResNet50 architectures. Notably, only MoD models incorporate all the attributes such as No Fine-Tuning (No FT), Simple Training (Simple Train.), GPU Implementation (GPU Impl.), a Static Computational Graph (Static Graph), and Enhanced Training Speed (Faster Train.).

Method	No	Simple	GPU	Static	Faster	Top1	GMAC	Params	Inference (ms)		SpeedUp	
	FT	Train.	Impl.	Graph	Train.	Acc (%)	(M)	(M)	CPU	GPU	CPU	GPU
ResNet152-MoD	✓	✓	✓	✓	✓	77.81	6.34	37.46	208.80	3.82	1.20	1.12
ResNet101						77.81	7.80	44.55	251.05	4.28	1.00	1.00
ResNet101-MoD	✓	✓	✓	✓	✓	77.08	4.58	29.21	162.03	2.73	1.55	1.57
ResNet86-MoD	✓	✓	✓	✓	✓	76.72	3.92	25.60	150.96	2.40	1.06	1.05
R50-DGNet [19]	✓	✗	✗	✗	✗	76.41	1.65*	29.34	— ¹	— ¹	—	—
ResNet75-MoD	✓	✓	✓	✓	✓	76.27	3.48	23.10	128.90	2.19	1.25	1.15
ResNet50						76.25	4.13	25.56	160.66	2.52	1.00	1.00
R50-Batch-Shaping [1]	✓	✗	✗	✗	✗	75.70	2.07*	15.31	—	—	1.31 ²	—
R50-ConvNet-AIG [36]	✓	✗	✗	✗	✗	75.45	2.59*	26.56	—	—	1.29 ²	—
R50-HRANK [21]	✗	✗	✓	✓	✗	74.98	2.30	16.15	— ³	— ³	—	—
ResNet50-MoD	✓	✓	✓	✓	✓	74.79	2.60	18.11	108.74	1.75	1.48	1.44
R50-FPGM [13]	✗	✗	✓	✓	✗	74.83	1.91	—	—	—	—	1.61 ⁴
R50-DynConv [37]	✓	✗	✓	✗	✗	74.40	2.25	25.56	—	—	—	— ⁵
R50-FPGM [13]	✓	✗	✓	✓	✗	74.13	1.91	—	—	—	—	1.61 ⁴
MoD50 rand.	✓	✓	✓	✓	✓	74.02	2.59	17.11	100.06	1.43	1.77	1.43

¹ No CPU/GPU Speed-Up given for ResNet50.

² CPU speed-ups for Batch-Shaping, ConvNet-AIG are from Batch-Shaping paper.

³ HRANK lacks implementation to compute CPU/GPU inference.

⁴ Only GPU speed-up reported for FPGM.

⁵ GPU speed-up for DynConv depends on a custom CUDA kernel.

* GMAC values are simulated estimates.

experiments were conducted using three different random seeds. Implementation details are provided in Appendix B.

Table 3 presents the performance of Fully Convolutional Networks (FCNs) with ResNet-based MoD backbones for semantic segmentation on the Cityscapes validation dataset. The FCN with a ResNet86-MoD backbone matches the inference times of the standard FCN-R50 model while enhancing the mean Intersection over Union (mIoU) by 0.95%, demonstrating MoD’s ability to boost segmentation accuracy without additional computational costs. The FCN-R75-MoD provides similar accuracy as standard models but with computational efficiency improvements, achieving a 9% speedup on CPU and 7% on GPU.

4.3 Object Detection on Pascal VOC

For our object detection experiments, we utilized the PASCAL VOC dataset, employing the mmdetection library [2] to configure and train our models.

Table 2: Comparative evaluation of standard and modified MobileNetV2 models on the ImageNet dataset. This table details Top-1 Accuracy, computational complexity (MMAC), model size (Params, in millions), and inference performance on CPU and GPU.

Method	Top1	MMAC	Params	Inference (ms)		SpeedUp	
	Acc (%)			(M)	CPU	GPU	CPU
MobileNetV2	71.63	320.36	3.5	47.76	0.78	—	—
MobileNetV2-MoD-L	71.55	344.76	3.34	42.94	0.71	1.11	1.10
MobileNetV2-MoD	69.34	220.56	2.94	33.43	0.56	1.43	1.39

Table 3: Comparison of FCN models on the Cityscapes validation dataset, contrasting standard ResNet50 backbones against MoD-enhanced versions. The table demonstrates that MoD models deliver performance comparable to or better than the standard FCN, with either reduced or similar inference times. Metrics include FLOPS (T), parameters (M), and inference durations (ms).

Model	Perf. Metrics (%)			FLOPS	Params	Inference (ms)		SpeedUp	
	aAcc	mIoU	mAcc			(T)	(M)	CPU	GPU
FCN-R86-MoD	95.75	73.75	81.75	0.377	47.15	8272.40	202.65	1.01	1.02
FCN-R50	95.68	72.80	80.25	0.393	47.11	8389.35	206.30	—	—
FCN-R75-MoD	95.72	72.72	80.56	0.359	44.66	7687.20	192.50	1.09	1.07
FCN-R50-MoD	95.37	71.58	79.54	0.322	39.66	6757.20	166.50	1.24	1.24

The experiments were conducted using three different random seeds. Detailed implementation notes are provided in Appendix C.

The outcomes of our experiments with Faster R-CNN models on the PASCAL VOC dataset are presented in Table 4. The Faster-RCNN with a ResNet86-MoD backbone demonstrates the MoD approach’s effectiveness by achieving a 0.37% improvement in mean Average Precision (mAP) and a 0.4% increase in Average Precision at 50% IoU threshold (AP50) compared to the standard Faster-RCNN-R50 model. Although GPU inference speed slightly decreases to 0.96x of the baseline, CPU processing speed is enhanced, showing a 1.10x improvement over the standard model. Similarly, the Faster-RCNN-R75-MoD model maintains accuracy close to the baseline while boosting inference efficiency, achieving an 11% speedup on CPU and 5% on GPU.

Table 4: Performance comparison of F-RCNN models on the Pascal VOC validation dataset, featuring standard and MoD-enhanced ResNet backbones. The table shows that MoD variants match or exceed the standard F-RCNN’s efficiency, achieving similar or better object detection performance with faster or equivalent inference times. Metrics include FLOPS (T), parameters (M), and inference durations (ms).

Model	Perf. Metrics (%)		FLOPS (T)	Params (M)	Inference (ms)		SpeedUp	
	mAP	AP50			CPU	GPU	CPU	GPU
Faster-RCNN-R86-MoD	77.67	77.67	0.108	41.486	2813.9	83.5	1.10	0.96
Faster-RCNN-R50	77.30	77.27	0.111	41.446	3101.5	80.5	—	—
Faster-RCNN-R75-MoD	77.04	77.03	0.104	38.989	2784.5	76.4	1.11	1.05
Faster-RCNN-R50-MoD	75.10	75.10	0.949	33.994	2453.3	69.3	1.26	1.16

5 Channel Selection and Regularization Analysis

5.1 Channel Selector Mechanism

In this section, we investigate the practical operation of the Channel Selector within our CNN MoD approach. This analysis focuses on the Channel Selector within a MoD Block in the third module of a ResNet75-MoD. To provide a clearer understanding of the Channel Selector’s behavior, we analyze how frequently channels are chosen in the top k routing mechanism. We use five different classes from the ImageNet validation set: plane, truck, church, cliff, and pug. For each class, fifty samples are drawn from the validation set, and we count the number of times each channel is selected. The percentage of occurrences for each channel is plotted in Figure 4. For comparison, we also plot the percentages across all 1000 classes in Figure 4a.

Our analysis shows that the Channel Selector selects different channels for different classes, indicating that it adapts to enhance the detection of class-specific features. Additionally, some channels are selected more frequently across various classes, suggesting a potential for further computational load reduction through channel pruning techniques.

5.2 Regularization Effect of Selective Processing

An interesting byproduct of the CNN MoD approach is its regularization effect, caused by the selective processing of channels in the modified Conv-Blocks. Traditional CNN architectures typically process all channels within each Conv-Block, which can lead to learning overly specific features that may not generalize well to new data. In contrast, the MoD framework selectively reduces the number of channels processed, focusing on those most important ones for the current task.

This selective processing naturally encourages the network to prioritize and refine features that are more likely to be generalizable across various images. As a result, MoD acts as a form of natural regularization, pushing the network

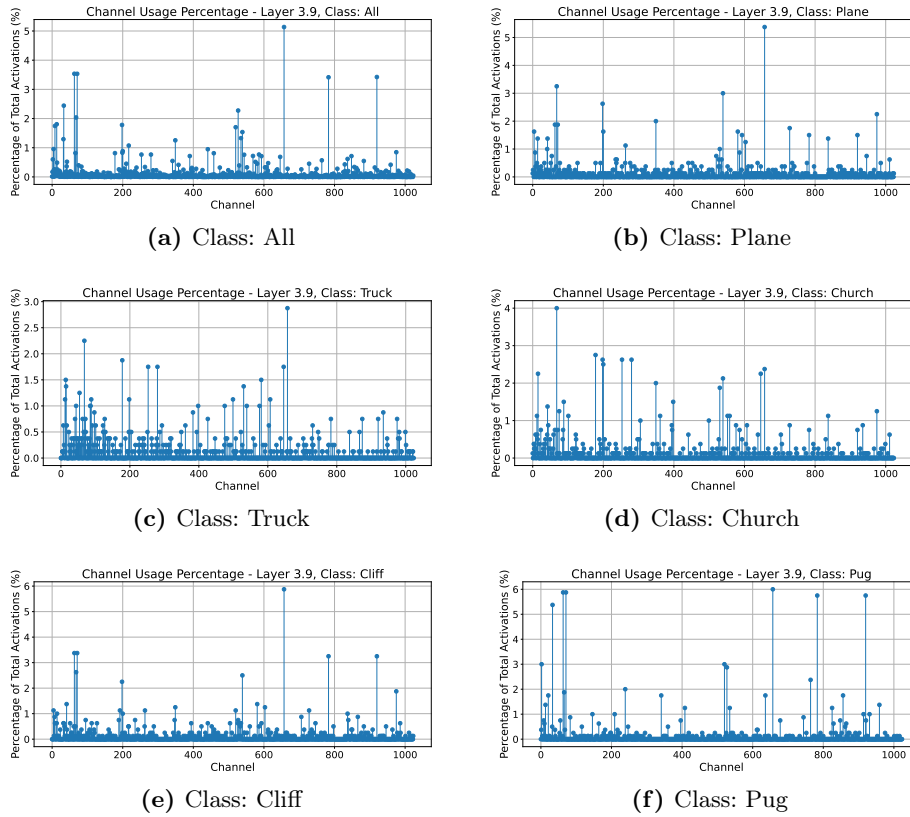


Fig. 4: Channel selection frequencies within the third module of a ResNet75-MoD for five diverse ImageNet classes: plane, truck, church, cliff, and pug. The analysis shows the percentage of times channels are selected by the Channel Selector out of the total selections in the layer, based on fifty samples per class. These percentages are compared to a baseline derived from all 1000 classes (4a), indicating that the Channel Selector selects different channels for different classes.

to extract more useful information from a limited set of inputs channels and thereby promoting the learning of robust, broadly applicable features. This is particularly advantageous for preventing overfitting, a common challenge in deep learning models, without additional explicit regularization strategies.

6 Conclusion

In this work, we presented CNN MoD, an approach inspired by the Mixture-of-Depths method developed for Transformers [27]. This technique combines the advantages of static pruning and dynamic computing within a single framework. It optimizes computational resources by dynamically selecting key channels in

feature maps for focused processing within the Conv-Blocks, while skipping less relevant channels.

It maintains a static computation graph, which optimizes the training and inference speed without the need for customized CUDA kernels, additional loss functions, or fine-tuning. These attributes separate MoD from other dynamic computing methods, offering a significant simplification of both training and inference processes. CNN MoD achieves performance comparable to traditional CNNs but with reduced inference times, GMACs, and parameters, or it surpasses them while maintaining similar inference times, GMACs, and parameters in image recognition, semantic segmentation, and object detection. For example, on ImageNet, ResNet86-MoD exceeds the performance of the standard ResNet50 by 0.45% with a 6% speedup on CPU and 5% on GPU. Moreover, ResNet75-MoD achieves the same performance as ResNet50 with a 25% speedup on CPU and 15% on GPU.

The fusion operation remains computationally demanding, primarily due to the implementation of the slicing operation, as revealed by torch.profiler. Future work will aim to refine this component, potentially through the development of a customized CUDA kernel to enhance operational efficiency. Further research into the optimal number of processed channels within layers is also promising for optimizing performance.

References

1. Bejnordi, B.E., Blankevoort, T., Welling, M.: Batch-shaping for learning conditional channel gated networks. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), <https://openreview.net/forum?id=Bke89JBtvB> 9, 11, 25
2. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open MMLab Detection Toolbox and Benchmark. arXiv preprint arXiv:1906.07155 (2019) 11
3. Chen, Z., Li, Y., Bengio, S., Si, S.: You Look Twice: GaterNet for Dynamic Filter Selection in CNNs. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 9172–9180. Computer Vision Foundation / IEEE (2019). <https://doi.org/10.1109/CVPR.2019.00939>, http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_You_Look_Twice_GaterNet_for_Dynamic_Filter_Selection_in_CNNs_CVPR_2019_paper.html 4
4. Contributors, M.: MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and benchmark. <https://github.com/open-mmlab/msegmentation> (2020) 22
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 8, 10, 22
6. Dong, X., Huang, J., Yang, Y., Yan, S.: More is Less: A More Complicated Network with Less Inference Complexity. In: 2017 IEEE Conference on Computer Vision and

- Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 1895–1903. IEEE Computer Society (2017). <https://doi.org/10.1109/CVPR.2017.205>, <https://doi.org/10.1109/CVPR.2017.205> 5
7. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* **88**(2), 303–338 (Jun 2010) 8
 8. Graves, A.: Adaptive Computation Time for Recurrent Neural Networks. *CoRR* **abs/1603.08983** (2016), <http://arxiv.org/abs/1603.08983> 2, 5
 9. Guo, Y., Yao, A., Chen, Y.: Dynamic Network Surgery for Efficient DNNs. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, December 5-10, 2016, Barcelona, Spain. pp. 1379–1387 (2016), <https://proceedings.neurips.cc/paper/2016/hash/2823f4797102ce1a1aec05359cc16dd9-Abstract.html> 4
 10. Han, S., Pool, J., Tran, J., Dally, W.J.: Learning both Weights and Connections for Efficient Neural Network. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, December 7-12, 2015, Montreal, Quebec, Canada. pp. 1135–1143 (2015), <https://proceedings.neurips.cc/paper/2015/hash/ae0eb3eed39d2bcef4622b2499a05fe6-Abstract.html> 4
 11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016. pp. 770–778. IEEE Computer Society (2016). <https://doi.org/10.1109/CVPR.2016.90>, <https://doi.org/10.1109/CVPR.2016.90> 1, 3, 4, 5, 6, 7, 8, 10, 24
 12. He, Y., Kang, G., Dong, X., Fu, Y., Yang, Y.: Soft Filter Pruning for Accelerating Deep Convolutional Neural Networks. In: Lang, J. (ed.) *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, July 13-19, 2018, Stockholm, Sweden. pp. 2234–2240. *ijcai.org* (2018). <https://doi.org/10.24963/IJCAI.2018/309>, <https://doi.org/10.24963/ijcai.2018/309> 1
 13. He, Y., Liu, P., Wang, Z., Hu, Z., Yang, Y.: Filter Pruning via Geometric Median for Deep Convolutional Neural Networks Acceleration. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, Long Beach, CA, USA, June 16-20, 2019. pp. 4340–4349. Computer Vision Foundation / IEEE (2019). <https://doi.org/10.1109/CVPR.2019.00447>, http://openaccess.thecvf.com/content_CVPR_2019/html/He_Filter_Pruning_via_Geometric_Median_for_Deep_Convolutional_Neural_Networks_CVPR_2019_paper.html 1, 4, 9, 11, 25
 14. He, Y., Zhang, X., Sun, J.: Channel Pruning for Accelerating Very Deep Neural Networks. In: *IEEE International Conference on Computer Vision, ICCV 2017*, Venice, Italy, October 22-29, 2017. pp. 1398–1406. IEEE Computer Society (2017). <https://doi.org/10.1109/ICCV.2017.155>, <https://doi.org/10.1109/ICCV.2017.155> 1
 15. Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18-22, 2018. pp. 7132–7141. Computer Vision Foundation / IEEE Computer Society (2018). <https://doi.org/10.1109/CVPR.2018.00745>, http://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html 3, 25

16. Huang, G., Chen, D., Li, T., Wu, F., van der Maaten, L., Weinberger, K.Q.: Multi-Scale Dense Networks for Resource Efficient Image Classification. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), <https://openreview.net/forum?id=Hk2aImxAB> 4
17. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto (2009) 8
18. Lavin, A., Gray, S.: Fast Algorithms for Convolutional Neural Networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 4013–4021. IEEE Computer Society (2016). <https://doi.org/10.1109/CVPR.2016.435>, <https://doi.org/10.1109/CVPR.2016.435> 2, 5
19. Li, F., Li, G., He, X., Cheng, J.: Dynamic Dual Gating Neural Networks. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 5310–5319 (2021), <https://api.semanticscholar.org/CorpusID:244477302> 1, 2, 4, 5, 9, 10, 11, 21, 25
20. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning Filters for Efficient ConvNets. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), <https://openreview.net/forum?id=rJqFGTslg> 1
21. Lin, M., Ji, R., Wang, Y., Zhang, Y., Zhang, B., Tian, Y., Shao, L.: HRank: Filter Pruning Using High-Rank Feature Map. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 1526–1535. Computer Vision Foundation / IEEE (2020). <https://doi.org/10.1109/CVPR42600.2020.00160>, https://openaccess.thecvf.com/content_CVPR_2020/html/Lin_HRank_Filter_Pruning_Using_High-Rank_Feature_Map_CVPR_2020_paper.html 4, 9, 11, 25
22. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11966–11976 (2022). <https://doi.org/10.1109/CVPR52688.2022.01167> 6, 7, 8
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 3431–3440. IEEE Computer Society (2015). <https://doi.org/10.1109/CVPR.2015.7298965>, <https://doi.org/10.1109/CVPR.2015.7298965> 1, 22
24. Luo, J., Wu, J., Lin, W.: ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 5068–5076. IEEE Computer Society (2017). <https://doi.org/10.1109/ICCV.2017.541>, <https://doi.org/10.1109/ICCV.2017.541> 1
25. Ma, N., Zhang, X., Zheng, H., Sun, J.: ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV. Lecture Notes in Computer Science, vol. 11218, pp. 122–138. Springer (2018). https://doi.org/10.1007/978-3-030-01264-9_8, https://doi.org/10.1007/978-3-030-01264-9_8 2
26. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.Z., DeVito, Z.,

- Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, December 8-14, 2019, Vancouver, BC, Canada. pp. 8024–8035 (2019), <https://proceedings.neurips.cc/paper/2019/hash/bdca288fee7f92f2bfa9f7012727740-Abstract.html> 21, 22
27. Raposo, D., Ritter, S., Richards, B.A., Lillicrap, T.P., Humphreys, P.C., Santoro, A.: Mixture-of-Depths: Dynamically allocating compute in transformer-based language models. *CoRR* **abs/2404.02258** (2024). <https://doi.org/10.48550/ARXIV.2404.02258>, <https://doi.org/10.48550/arXiv.2404.02258> 2, 3, 5, 14, 24
28. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, June 27-30, 2016. pp. 779–788. IEEE Computer Society (2016). <https://doi.org/10.1109/CVPR.2016.91>, <https://doi.org/10.1109/CVPR.2016.91> 1
29. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, December 7-12, 2015, Montreal, Quebec, Canada. pp. 91–99 (2015), <https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html> 1, 23
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **115**(3), 211–252 (2015). <https://doi.org/10.1007/s11263-015-0816-y>, <https://doi.org/10.1007/s11263-015-0816-y> 4, 8, 9
31. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18-22, 2018. pp. 4510–4520. Computer Vision Foundation / IEEE Computer Society (2018). <https://doi.org/10.1109/CVPR.2018.00474>, http://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html 3, 7, 8, 10, 21
32. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Bengio, Y., LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), <http://arxiv.org/abs/1409.1556> 8, 24
33. Sze, V., Chen, Y., Yang, T., Emer, J.S.: Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proc. IEEE* **105**(12), 2295–2329 (2017). <https://doi.org/10.1109/JPROC.2017.2761740>, <https://doi.org/10.1109/JPROC.2017.2761740> 2, 5
34. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, Boston, MA, USA, June 7-12, 2015. pp. 1–9. IEEE Computer Society (2015). <https://>

- doi.org/10.1109/CVPR.2015.7298594, <https://doi.org/10.1109/CVPR.2015.7298594> 1
35. Teerapittayanon, S., McDanel, B., Kung, H.T.: BranchyNet: Fast inference via early exiting from deep neural networks. In: 23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016. pp. 2464–2469. IEEE (2016). <https://doi.org/10.1109/ICPR.2016.7900006>, <https://doi.org/10.1109/ICPR.2016.7900006> 4
 36. Veit, A., Belongie, S.J.: Convolutional Networks with Adaptive Inference Graphs. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11205, pp. 3–18. Springer (2018). https://doi.org/10.1007/978-3-030-01246-5_1, https://doi.org/10.1007/978-3-030-01246-5_1 2, 5, 9, 10, 11, 25
 37. Verelst, T., Tuytelaars, T.: Dynamic Convolutions: Exploiting Spatial Sparsity for Faster Inference. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 2317–2326. Computer Vision Foundation / IEEE (2020). <https://doi.org/10.1109/CVPR42600.2020.00239>, https://openaccess.thecvf.com/content_CVPR_2020/html/Verelst_Dynamic_Convolutions_Exploiting_Spatial_Sparsity_for_Faster_Inference_CVPR_2020_paper.html 2, 5, 9, 11, 25
 38. Wang, X., Yu, F., Dou, Z., Darrell, T., Gonzalez, J.E.: SkipNet: Learning Dynamic Routing in Convolutional Networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII. Lecture Notes in Computer Science, vol. 11217, pp. 420–436. Springer (2018). https://doi.org/10.1007/978-3-030-01261-8_25, https://doi.org/10.1007/978-3-030-01261-8_25 5
 39. Wimmer, P., Mehnert, J., Condurache, A.: COPS: Controlled Pruning Before Training Starts. In: International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021. pp. 1–8. IEEE (2021). <https://doi.org/10.1109/IJCNN52387.2021.9533582>, <https://doi.org/10.1109/IJCNN52387.2021.9533582> 4
 40. Wimmer, P., Mehnert, J., Condurache, A.: Interspace Pruning: Using Adaptive Filter Representations to Improve Training of Sparse CNNs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 12517–12527. IEEE (2022). <https://doi.org/10.1109/CVPR52688.2022.01220>, <https://doi.org/10.1109/CVPR52688.2022.01220> 1
 41. Wu, Z., Nagarajan, T., Kumar, A., Rennie, S., Davis, L.S., Grauman, K., Feris, R.S.: BlockDrop: Dynamic Inference Paths in Residual Networks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 8817–8826. Computer Vision Foundation / IEEE Computer Society (2018). <https://doi.org/10.1109/CVPR.2018.00919>, http://openaccess.thecvf.com/content_cvpr_2018/html/Wu_BlockDrop_Dynamic_Inference_CVPR_2018_paper.html 2, 4, 5
 42. Xie, Z., Zhang, Z., Zhu, X., Huang, G., Lin, S.: Spatially Adaptive Inference with Stochastic Feature Sampling and Interpolation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I. Lecture Notes in Computer

- Science, vol. 12346, pp. 531–548. Springer (2020). https://doi.org/10.1007/978-3-030-58452-8_31, https://doi.org/10.1007/978-3-030-58452-8_31 5
43. Yu, F., Koltun, V.: Multi-Scale Context Aggregation by Dilated Convolutions. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016), <http://arxiv.org/abs/1511.07122> 1
 44. Yu, R., Li, A., Chen, C., Lai, J., Morariu, V.I., Han, X., Gao, M., Lin, C., Davis, L.S.: NISP: Pruning Networks Using Neuron Importance Score Propagation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 9194–9203. Computer Vision Foundation / IEEE Computer Society (2018). <https://doi.org/10.1109/CVPR.2018.00958>, http://openaccess.thecvf.com/content_cvpr_2018/html/Yu_NISP_Pruning_Networks_CVPR_2018_paper.html 1
 45. Zhuang, Z., Tan, M., Zhuang, B., Liu, J., Guo, Y., Wu, Q., Huang, J., Zhu, J.: Discrimination-aware Channel Pruning for Deep Neural Networks. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. pp. 883–894 (2018), <https://proceedings.neurips.cc/paper/2018/hash/55a7cf9c71f1c9c495413f934dd1a158-Abstract.html> 1

A Implementation Details and Further Results on ImageNet

A.1 Implementation Details on ImageNet

We conducted our MoD experiments on ImageNet for 100 epochs, using PyTorch 1.10.1 [26] across four Nvidia GeForce 1080Ti GPUs, following standard training protocols.*** Our configuration included a batch size of 256, an initial learning rate of 0.1 decreasing by a factor of 0.1 every 30 epochs, momentum set at 0.9, and weight decay of 1×10^{-4} . This setup aligns with the practices of comparable studies, such as DGNNet [19].

A.2 Architecture Details of ResNet MoD Models

Table 5 outlines the layer configurations of the ResNet MoD models. The models incorporates alternating block patterns in each module.

Table 5: Layer Configurations of ResNet MoD Models

Model	Layer Configuration	Block Type
ResNet18-MoD	[2, 2, 2, 2]	BasicBlock
ResNet26-MoD	[2, 2, 3, 4]	BasicBlock
ResNet34-MoD	[3, 4, 6, 3]	BasicBlock
ResNet42-MoD	[3, 3, 6, 6]	BasicBlock
ResNet50-MoD	[3, 4, 6, 3]	BottleneckBlock
ResNet75-MoD	[3, 4, 14, 3]	BottleneckBlock
ResNet86-MoD	[3, 4, 18, 3]	BottleneckBlock
ResNet101-MoD	[3, 4, 23, 3]	BottleneckBlock
ResNet152-MoD	[3, 8, 36, 3]	BottleneckBlock

A.3 Architecture Details of MobileNetV2 MoD Models

The “MobileNetV2-MoD-L” represents a deeper configuration of the standard MobileNetV2. The standard MobileNetV2 [31] architecture utilizes a sequence of inverted residual blocks with a specific configuration pattern defined by the parameters t (expansion factor), c (number of channels), n (number of times the block is repeated), and s (stride). The “MobileNetV2-MoD-L” model modifies these parameters as shown in Table 6.

A.4 Comparative Performance of ResNet Models on ImageNet

Table 7 shows further results for MoD-enhanced ResNets, detailing improvements in computational efficiency, model compactness, and inference speeds.

*** Refer to PyTorch’s official training recipes at [PyTorch repository](#).

Table 6: Comparison of Standard MobileNetV2 and MobileNetV2-MoD-L Configurations

Layer	Standard Configuration	MoD-L Configuration
1	[1, 16, 1, 1]	[1, 16, 1, 1]
2	[6, 24, 2, 2]	[6, 32, 2, 2]
3	[6, 32, 3, 2]	[6, 64, 3, 2]
4	[6, 64, 4, 2]	[6, 96, 4, 2]
5	[6, 96, 3, 1]	[6, 128, 3, 1]
6	[6, 160, 3, 2]	[6, 160, 3, 2]
7	[6, 320, 1, 1]	[6, 320, 1, 1]

Table 7: Comparative performance of standard and MoD-enhanced ResNet models on the ImageNet dataset. The table evaluates Top-1 Accuracy, computational complexity (GMAC), model size (Params, in millions), and inference speed improvements on CPU and GPU.

Method	Top1	GMAC	Params	Inference (ms)		SpeedUp	
	Acc (%)			(M)	CPU	GPU	CPU
ResNet34	73.92	3.68	21.29	98.83	1.27	—	—
ResNet42-MoD	72.03	2.29	17.72	64.01	0.88	1.54	1.45
ResNet34-MoD	71.44	2.06	12.93	60.79	0.83	1.63	1.53
ResNet18	70.37	1.82	11.18	53.86	0.75	—	—
ResNet26-MoD	69.53	1.36	11.4	42.16	0.58	1.28	1.28
ResNet18-MoD	64.05	0.89	5.46	33.63	0.47	1.60	1.58

A.5 Comparative Performance of ResNet MoD-lk Models on ImageNet

Table 8 presents a comparison of the performance of standard and MoD-enhanced ResNet models with a specific focus on the position of reintegration of processed channels. “MoD-lk” denotes models where processed channels are added to the last k channels of the original feature map. This experimental variation aims to assess the impact of consistent channel positioning on network performance. The results indicate that MoD-lk models perform comparably to their counterparts where processed channels are added to the first k channels, suggesting that maintaining a consistent position for processed information within the feature maps is beneficial for optimizing model performance.

B Implementation Details Semantic Segmentation

In the Cityscapes experiments, PyTorch 1.10.1 [26] and four Nvidia GeForce 1080Ti GPUs were used. Employing the MMSegmentation Framework [4], we utilized FCN [23] on the dataset [5], which consists of 2,975 training, 500 validation, and 1,525 testing images across 19 semantic classes. Training involved

Table 8: This table presents a comparison between standard MoD models and those enhanced with MoD-lk, where processed channels are integrated into the last kk channels. The evaluation covers Top-1 Accuracy, computational complexity (GMAC), model size (in millions of parameters), and inference speed improvements on both CPU and GPU. The aim is to analyze the impact of channel positioning on the performance of ResNet models on the ImageNet dataset.

Method	Top1	GMAC	Params	Inference (ms)		SpeedUp	
	Acc (%)			(M)	CPU	GPU	CPU
ResNet86-MoD	76.72	3.92	25.60	150.96	2.40	1.06	1.05
ResNet86-MoD-lk	76.64	3.92	25.60	150.96	2.40	1.06	1.05
ResNet75-MoD	76.27	3.48	23.10	128.90	2.19	1.25	1.15
ResNet75-MoD-lk	76.22	3.48	23.10	128.90	2.19	1.25	1.15
ResNet50-MoD	74.79	2.60	18.11	108.74	1.75	1.48	1.44
ResNet50-MoD-lk	74.57	2.60	18.11	108.74	1.75	1.48	1.44

resizing, random cropping, flipping, photometric distortion, normalization, and padding. Testing employed multi-scale flip augmentation and normalization.

The FCN model, with a ResNet50 backbone, used an Encoder-Decoder architecture with FCN-Head as the decode and auxiliary heads. The model used SyncBN and a dropout ratio of 0.1, with the auxiliary head contributing 40% to the total loss.

Optimization was via SGD (learning rate 0.01, momentum 0.9, weight decay 0.0005). A polynomial decay learning rate policy was applied (power 0.9, minimum learning rate $1e-4$), over 80,000 iterations with checkpoints and evaluations (focusing on mIoU) every 8,000 iterations.

C Implementation Details Object Detection

Model Configuration: We configure our Faster R-CNN [29] with a ResNet-50 backbone and a Feature Pyramid Network (FPN) neck for multi-scale feature extraction.

Data Preprocessing and Augmentation: Our preprocessing pipeline employs a sequence of transformations to prepare input images for object detection tasks. Initially, images are loaded and their corresponding annotations are retrieved. Subsequently, we resize the images to a resolution of (1000, 600), ensuring the preservation of their original aspect ratio. To augment the dataset and introduce variability, we apply random horizontal flips with a 50% probability. This augmentation strategy is applied uniformly across the training dataset, aiming to enhance model robustness and generalization capability. For validation, images undergo a similar resizing process without the application of random flips, maintaining consistency in evaluation conditions.

Training Configuration: The model is trained on the combined train sets of VOC2007 and VOC2012 and evaluated on the VOC2007 val set. Training employs a batch size of 2, using an AspectRatioBatchSampler for GPU efficiency. We adopt SGD with momentum and weight decay, adjusting the learning rate as per a predefined schedule. The mean Average Precision (mAP) metric, calculated using the ‘11points’ interpolation method, serves as the evaluation metric.

Evaluation: The evaluation on the VOC2007 val set employs the standard VOC mAP metric, adhering to the “11points” method. This setup mirrors the training configuration but without data augmentation, ensuring deterministic inference.

D Experiments on CIFAR

Our evaluation of the MoD approach was conducted on the CIFAR-10/100 datasets, comprising 50,000 training and 10,000 test color images of 32x32 pixels. We utilized a range of CNN architectures for our experiments, including ResNet18/34/50 [11] and VGG16/19-BN [32].

To ensure robustness and reproducibility, each model was trained and evaluated five times using different random seeds, impacting network initialization, data ordering, and augmentation processes. We present the mean test accuracy and its standard deviation for these trials. The data split comprised 90% for training and 10% for validation, with the best-performing model on the validation set chosen for the final evaluation.

E Comparative Analysis of MoD in CNNs and Transformers

In the main body of this paper, we detailed the application of the MoD approach to CNNs. This section aims to outline how this approach differs from the Mixture-of-Depths application in Transformers [27].

Token vs. Channel Processing:

- **Transformers:** In Transformers, MoD operates at the token level. Tokens represent the units of data processed throughout the model’s architecture, typically as subwords or whole words. They are processed throughout the entire Transformer architecture.
- **CNNs:** Conversely, MoD in CNNs treats channels within feature maps as “tokens”. This novel approach differs from Transformers because channels in traditional CNNs are not treated as tokens, and their significance and composition vary from one convolutional layer to the next.

Token/Channel Selection:

- **Transformers:** Selection is based on a linear projection that assigns a scalar value to each token.

- **CNNs:** CNNs use a mini neural network (incorporating Adaptive Average Pooling 2D, a two-layer fully connected network with Sigmoid activation) inspired by Squeeze-and-Excitation blocks [15], specifically tailored for image-based tasks.

Architecture Modification:

- **Transformers:** The architecture of Transformer blocks remains unchanged with the use of MoD; only the quantity of processed tokens varies. Transformer models are designed to handle a variable number of tokens.
- **CNNs:** In CNNs, varying the number of channels in convolutional layers is impractical. MoD thus requires adjustments to the convolutional layers themselves, including a reduction in the number of channels in the convolution kernels to match the reduced number of input feature map channels.

Reintegration of Processed Information:

- **Transformers:** Processed tokens are added back to their original counterparts, a method made effective through the use of positional encoding.
- **CNNs:** Unlike in Transformers, neither replacing nor adding back processed channels to their original positions has proven effective in CNNs. More effective is the addition of processed channels to a fixed set of channels, such as the first k channels, to maintain consistency in locating processed information within the network.

E.1 Effect of Channel Parameter c on Accuracy and Inference Times

We evaluated the impact of varying the channel parameter c on the Top-1 accuracy and inference times of the ResNet50-MoD model on the ImageNet dataset. The parameter c controls the number of channels processed in the MoD approach. As shown in Table 10, using $c = 64$ yields the best balance between accuracy and inference times. Note that $c > 64$ is not possible since the number of channels in the first Conv-Block of ResNets is fixed at 64.

E.2 Consistency of Performance across Random Seeds

We provide standard deviations for the ResNet experiments on ImageNet in Table 11. Other state-of-the-art pruning and dynamic computation methods, such as DGNet [19], Batch-Shaping [1], ConvNet-AIG [36], HRANK [21], FPGM [13], and DynConv [37], do not report standard deviations, preventing direct comparison. Nevertheless, the consistent results across different MoD configurations demonstrate that the MoD approach does not introduce additional variance, as evidenced by the low standard deviations.

Table 9: Performance metrics comparison on the CIFAR10 and CIFAR100 datasets using standard models and their MoD variants. This table illustrates the trade-off between efficiency and inference time, demonstrating that the MoD models can achieve comparable performance to the standard models at faster inference times or improved performance at comparable inference times. FLOPS are in millions of multiply-accumulate operations (MMAC), parameters in millions (M), and inference times in milliseconds (ms).

Model	Set	Test Acc. (%)	FLOPS (MMAC)	Params (M)	Inference (ms)		SpeedUp	
					CPU	GPU	CPU	GPU
ResNet18	C10	94.04 ± 0.08	557	11.17	15.67	0.24	-	-
ResNet18-MoD	C10	92.37 ± 0.18	255	4.95	8.32	0.14	1.88	1.73
ResNet34	C10	93.69 ± 0.27	1016	21.28	30.52	0.42	-	-
ResNet34-MoD	C10	93.83 ± 0.20	633	12.42	18.10	0.27	1.69	1.54
ResNet50	C10	93.31 ± 0.33	1310	23.52	48.47	0.83	-	-
ResNet50-MoD	C10	93.24 ± 0.24	808	16.07	33.76	0.58	1.44	1.41
ResNet18	C100	76.47 ± 0.18	557	11.22	14.99	0.23	-	-
ResNet18-MoD	C100	72.73 ± 0.21	255	4.99	8.89	0.13	1.69	1.75
ResNet34	C100	77.07 ± 0.41	1160	21.33	30.67	0.42	-	-
ResNet34-MoD	C100	76.86 ± 0.23	633	12.47	18.66	0.27	1.64	1.55
ResNet50	C100	76.17 ± 0.63	1310	23.71	46.64	0.81	-	-
ResNet50-MoD	C100	76.76 ± 0.61	808	16.26	34.81	0.58	1.34	1.41
VGG16-BN	C10	93.27 ± 0.11	315	15.25	8.80	0.14	-	-
VGG16-BN-MoD	C10	91.79 ± 0.14	155	9.83	5.47	0.16	1.61	0.87
VGG19-BN	C10	93.21 ± 0.07	400	20.57	11.55	0.18	-	-
VGG19-BN-MoD	C10	91.82 ± 0.18	155	9.91	5.81	0.22	1.99	0.81
VGG16-BN	C100	72.48 ± 0.32	315	15.30	9.27	0.14	-	-
VGG16-BN-MoD	C100	69.23 ± 0.25	155	9.88	6.41	0.15	1.45	0.95
VGG19-BN	C100	71.34 ± 0.12	400	20.61	11.38	0.18	-	-
VGG19-BN-MoD	C100	69.19 ± 0.11	155	9.96	5.75	0.23	1.98	0.80

Table 10: Top-1 accuracy and inference times in ms on ImageNet for different values of c , comparing results when processed channels are added to the first k channels (S) versus their original positions (OP) for R50-MoD.

c	Top-1 (S)	Top-1 (OP)	CUDA (S)	CPU (S)
2	72.51	55.36	2.23	132.64
4	73.74	60.78	1.99	118.85
8	74.43	63.05	1.90	114.85
16	74.74	65.35	1.97	114.06
32	74.68	70.80	1.83	111.30
64	74.79	70.31	1.75	108.74

Table 11: Top-1 accuracy and standard deviation on ImageNet.

Method	Top-1 Acc. (%) \pm Std. Dev.
R152-MoD	77.81 ± 0.05
R101	77.81 ± 0.07
R101-MoD	77.08 ± 0.08
R86-MoD	76.72 ± 0.04
R75-MoD	76.27 ± 0.07
R50	76.25 ± 0.19
R50-MoD	74.79 ± 0.08