Knowledge-Based Systems 317 (2025) 113393

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

FMDConv: Fast multi-attention dynamic convolution via speed-accuracy trade-off

Tianyu Zhang [a], Fan Wan [a], Haoran Duan [a], Kevin W. Tong [b], Jingjing Deng [a], Yang Long [a,*]

[a] Computer Science Department, Durham University, The Palatine Centre, University, Stockton Rd, Durham, DH1 3LE, County Durham, United Kingdom
[b] College of Automation & College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Guangdong Rd, Gulou, Nanjing, 210023, Jiangsu, China

ARTICLE INFO

Keywords:
Dynamic convolution
Attention mechanism
Speed-accuracy trade-off

ABSTRACT

Spatial convolution is fundamental in constructing deep Convolutional Neural Networks (CNNs) for visual recognition. While dynamic convolution enhances model accuracy by adaptively combining static kernels, it incurs significant computational overhead, limiting its deployment in resource-constrained environments such as federated edge computing. To address this, we propose Fast Multi-Attention Dynamic Convolution (FMDConv), which integrates input attention, temperature-degraded kernel attention, and output attention to optimize the speed-accuracy trade-off. FMDConv achieves a better balance between accuracy and efficiency by selectively enhancing feature extraction with lower complexity. Furthermore, we introduce two novel quantitative metrics, the Inverse Efficiency Score and Rate-Correct Score, to systematically evaluate this trade-off. Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet demonstrate that FMDConv reduces the computational cost by up to 49.8% on ResNet-18 and 42.2% on ResNet-50 compared to prior multi-attention dynamic convolution methods while maintaining competitive accuracy. These advantages make FMDConv highly suitable for real-world, resource-constrained applications.

1. Introduction

Convolutional Neural Networks (CNNs) [1–4] have become the dominant approach for various vision-based tasks, including object detection [5,6], semantic segmentation [7,8], and image classification [9,10]. While traditional networks like VGGNets, GoogLeNets, and ResNets rely on static convolutional kernels, these fixed-size kernels limit the ability to capture diverse contexts in images with varying scales and resolutions.

To address this, recent studies have explored dynamic convolution [11], where kernels adapt to different input characteristics. SENet [12] introduced dynamic channel weighting, while CondConv [13] further extended dynamic convolution by constructing unique kernels for individual images. DynamicConv [14] and ODConv [15] incorporated attention mechanisms into dynamic convolution, significantly improving the adaptability of convolutional operations and enhancing feature representation learning.

However, despite these advancements, existing methods struggle to effectively balance the trade-off between computational efficiency and accuracy. Many prior works focus primarily on improving accuracy but overlook the need for efficiency in real-world applications, particularly in edge computing and mobile environments. While ODConv [15] integrates multiple attention mecha nisms to enhance feature extraction, its complexity leads to substantial computational cost increases, making it less feasible for deployment in resource-limited scenarios.

A key challenge in deep learning is the speed-accuracy trade-off, which remains inadequately addressed. Existing studies primarily rely on empirical observations, using FLOPs or inference time as proxies for efficiency, but struggle to provide a systematic framework to jointly evaluate both efficiency and accuracy. The lack of quantifiable measures limits the development of efficiency-balanced deep learning models for broader applications. Moreover, most prior studies assess trade-offs based on empirical observations rather than theoretical formulation, leading to inconsistencies in evaluating efficiency-performance balance. The speed-accuracy trade-off, a well-established concept in psychology and neuroscience [16], offers valuable insights for computational science.

To address these limitations, we introduce Fast Multi-Attention Dynamic Convolution (FMDConv), as shown in Fig. 1, a novel lightweight convolutional block that selectively integrates multiple attention mecha

* Corresponding author.
E-mail address: yang.long@ieee.org (Y. Long).

https://doi.org/10.1016/j.knosys.2025.113393
Received 22 October 2024; Received in revised form 25 February 2025; Accepted 18 March 2025
Available online 7 April 2025

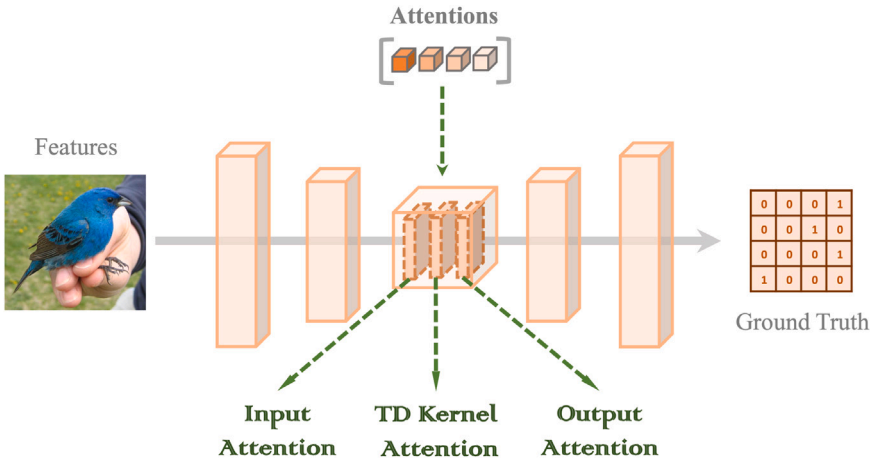


Fig. 1. Overview of the FMDConv framework. The diagram illustrates the three attention mechanisms (Input, TD Kernel, and Output Attention) in FMDConv, each targeting a distinct stage of feature extraction for optimal efficiency and accuracy.

nisms—input attention, temperature-degraded kernel attention, and output attention—to reduce computational complexity while maintaining competitive accuracy.

Additionally, we propose two novel quantitative metrics, the Inverse Efficiency Score (IES) and the Rate-Correct Score (RCS), to systematically evaluate the efficiency-accuracy trade-off in deep learning architectures. Unlike prior works that only measure FLOPs or inference time, our metrics provide a unified framework to jointly assess computational efficiency and model accuracy, enabling standardized comparisons across different approaches.

Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet demonstrate that FMDConv achieves state-of-the-art efficiency-accuracy trade-offs, reducing FLOPs by up to 49.8% on ResNet-18 and 42.2% on ResNet-50, while maintaining competitive accuracy. Our findings highlight FMDConv's potential for real-world, resource-constrained applications and underscore the necessity of standardized efficiency-performance evaluations in deep learning.

Our main contributions can be summarized as follows:

- We propose IES & RCS, the first standardized metrics for evaluating the speed-accuracy trade-off in dynamic convolution, moving beyond previous qualitative assessments.
- We conduct a comprehensive evaluation of attention mechanisms in dynamic convolution and identify input attention, temperature-degraded kernel attention, and output attention as the optimal structures for balancing efficiency and accuracy.
- We introduce Fast Multi-Attention Dynamic Convolution (FMDConv), a novel lightweight convolutional block that selectively integrates these attentions, significantly reducing computational cost while maintaining accuracy. Compared to ODConv, FMDConv achieves up to 49.8% and 42.2% FLOP reductions on ResNet-18 and ResNet-50, respectively, making it highly efficient for resource-constrained applications.
- Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet demonstrate the superiority of FMDConv, achieving a state-of-the-art efficiency-accuracy trade-off with reduced computational overhead.

2. Related works

Deep Convolutional Neural Networks (CNNs) Architecture. LeNet-5 is one of the first convolutional neural networks proposed by LeCun et al. [17] in the late 1990s, which was trained using backpropagation for a handwritten zip code recognition task with superior performance. AlexNet, invented by Krizhevsky *et al.* won the ImageNet 2012 challenge and significantly outperformed the second place by over 10% in both classification and localization tasks. It adopts convolutional layers, dropout regularization, and data augmentation strategies, and its success demonstrates the great potential of deep CNNs in visual recognition. By then, limited by computational resources, their work did not explore deeper architecture. Simonyan et al. from Oxford proposed a family of VGGNets [18], which introduces a systematic approach to designing deeper models. For example, the paper proposes using stacked small filters (*i.e.,* 2 stacked 3×3 kernels) to replace a single large filter (*i.e.,* 1 single 5×5 kernel), which can span the same receptive field with a deeper architecture but fewer computational operations. They showed that the number of layers positively correlates with the model accuracy of up to 16 layers, which outperforms AlexNet by a significant margin. However, they also observed that adding more layers (*i.e.,* 19) cannot further improve the performance. InceptionNet [19], also known as GoogLeNet, takes a novel approach to expand the architecture horizontally within the same layer by using multiple kernels at different scales, where the computed feature maps are then concatenated to form the input for the next layer. It also uses numerous auxiliary classifiers at different levels to improve training convergence. He et al. [20] proposed a ResNet that uses residual connection structure to learn the depth of the architecture dynamically via backpropagation. ResNet won several visual recognition challenges and was used as the core for AlphaGo.

Lightweight architecture design has attracted great attention. Howa rd et al. [21] introduced MobileNet, which is suitable for deploying on mobile and embedded devices. The model decomposes a standard convolutional filter into a depth-wise convolution and a point-wise convolution, greatly reducing computational operations. In addition, group convolution and channel shuffling used in ShuffleNet [22] are often used jointly with other design principles. In this paper, we consider both efficiency and performance from a trade-off perspective, and the proposed FMDConv can be used as a basic building block or a design paradigm in convolutional neural network architecture. These backbone networks, as we mentioned above, are mainly based on static convolution operators and have been widely adopted in many subsequent deep CNNs, which has stimulated further research.

Dynamic Convolution Neural Networks. The core idea of dynamic convolution originated from ConvCond, proposed by Yang et al. [13] in 2019. A static convolution applies the same kernel across the whole dataset; a dynamic convolution has a unique kernel for each image, which can be achieved by using parameterized convolutions that are conditioned on the input images. Chen et al. [14] proposed to use the attention mechanism over the kernel itself, which dynamically integrates multiple parallel convolution kernels into one that is conditioned on the layer input. The experimental results showed that such an integration strategy can improve expression capacity without

Table 1
Comparison of existing static and dynamic convolution methods, summarizing key techniques, application scenarios, and limitations.

Method	Key technique	Application scenarios	Limitations
VGG [18]	Stacked small convolutional kernels	General-purpose CNNs	High computational cost
ResNet [20]	Residual connections for deeper networks	Deep CNN architectures	Performance saturates at extreme depth
SENet [12]	Channel recalibration via SE module	Lightweight models	Does not learn spatial dependencies
CondConv [13]	Mixture of convolutional kernels	Efficient deep models	High memory overhead
DynamicConv [14]	SoftMax-weighted kernel aggregation	Image classification	High computational cost
ODConv [15]	Kernel, spatial, channel, and filter attention	Large-scale vision models	Higher computational cost

increasing the depth and width of the network. We refer the readers to the works of Han et al. [23], Sun et al. [24], Wu et al. [25], Zhao et al. [26], Huang et al. [27], and Zhang et al. [28] for existing works and open problems in constructing dynamic convolution. It is worth noting that most works endow convolution kernels with dynamic properties through the single dimension of the kernel space (*i.e.,* the number of convolution kernels). Li et al. [15] argued that integrating all dimensions (including kernel attention, output channel attention, input channel attention, and spatial attention) is capable of improving the learning capacity of the model and achieving better recognition performance. However, dynamic convolution usually involves additional operations that are computationally expensive. The DCD Network proposed by Li et al. [29] replaces dynamic attention over channel groups with channel fusion in a low-dimensional space, which requires fewer parameters and lower computational costs without sacrificing model accuracy. Inspired by ODConv [15] and DynamicConv [14] models, in this paper, we focus on trading off efficiency and accuracy via optimizing the integration of dynamic strategies through all possible dimensions of the kernel space.

Speed-Accuracy Trade-off. Huang [30] demonstrated various feasible approaches to trading accuracy for speed and memory usage in deep learning-based object detection frameworks. Similar works were proposed for different application domains, such as Riel et al. [31] for medical applications using axial Computed Tomography (CT) images, Javadi et al. [32] for humanoid robots, and Chaves et al. [33] for forensic surveillance. In summary, the speed-accuracy trade-off can be achieved via either a model pruning strategy (such as partial sequential pruning [34]) or specifically designed building blocks (such as COSFORMER [35], GSoP [36]). We notice that the speed-accuracy trade-off [16] is a well-established concept in psychology and neuroscience that refers to the tendency of individuals to balance the speed and accuracy of their responses in a given task. Motivated by [16], in this paper we propose two new metrics to measure the trade-off between speed and accuracy and develop FMDConv, a novel efficiency-accuracy-balanced building block for deep CNNs.

Table 1 provides a structured comparison of existing static and dynamic convolution methods, emphasizing computational complexity, key mechanisms, and practical applications.

3. Methodology

In this section, we provide a detailed description of the architecture and implementation of Fast Multi-Attention Dynamic Convolution (FMDConv). FMDConv introduces multiple attention mechanisms to enhance the efficiency and accuracy of convolutional neural networks by dynamically adjusting convolutional kernel weights as well as the input and output feature attentions.

3.1. Dynamic convolution & omni-dynamic convolution

Traditional static convolution applies the same kernel across all input images, whereas dynamic convolution adjusts kernel parameters dynamically according to the input image. In essence, the convolutional kernel is a learned function conditioned on the input data. Mathematically, dynamic convolution can be defined as:

$$y = (\alpha^x_{w_1} W_1 + \cdots + \alpha^x_{w_i} W_i + \cdots + \alpha^x_{w_n} W_n) * x \tag{1}$$

where W_i is the weight of the ith convolutional kernel, $\alpha^x_{w_i}$ is the corresponding attention value based on input x, y represents the output feature map, and $*$ denotes the convolution operation.

Li et al. proposed to jointly use four different attentions in ODConv [15] that can be formally defined as follows:

$$y = (\alpha^x_{w_1} \odot \alpha^x_{f_1} \odot \alpha^x_{c_1} \odot \alpha^x_{s_1} + \cdots + \alpha^x_{w_i} \odot \alpha^x_{f_i} \odot \alpha^x_{c_i} \odot \alpha^x_{s_i} + \cdots + \alpha^x_{w_n} \odot \alpha^x_{f_n} \odot \alpha^x_{c_n} \odot \alpha^x_{s_n}) * x \tag{2}$$

where $\odot$ denotes the element-wise Hadamard product. Here, $\alpha^x_{w_i}$, $\alpha^x_{f_i}$, $\alpha^x_{c_i}$, $\alpha^x_{s_i}$ represent the kernel attention, output channel attention, input channel attention, and spatial attention, respectively, while x, and y denote the input and output feature maps, respectively.

3.2. Metric of speed-accuracy trade-off

Motivated by the well-established speed-accuracy trade-off concept in psychology [16], we introduce two novel metrics, Inverse Efficiency Score (IES) [37] and Rate-Correct Score (RCS) [38], to jointly measure computational overhead and model accuracy. Note that in cognitive psychology, reaction time (RT) refers to human subjects' response speed. Here, we adapt the concept to measure the computational overhead (e.g., training time, FLOPs) of deep learning models. We do not imply an exact one-to-one mapping but merely draw inspiration from the speed-accuracy trade-off phenomenon.

Inverse Efficiency Score (IES). The most commonly used measure for a speed-accuracy trade-off in experimental psychology is IES [37], which is typically defined as the mean correct reaction time (RT) divided by the proportion of correct classifications. In our case, we adopt the concept of the original IES and formulate the score as the ratio of the training time of an epoch to the Top-1 accuracy rate:

$$IES_{ij} = \frac{\overline{RT_{ij}}}{PC_{ij}} \tag{3}$$

where $\overline{RT_{ij}}$ is the mean training time of the model i on correct-classification trials with hyper-parameter set j, and PC_{ij} is the proportion of correct classifications of i for j. Although most (if not all) research using IES has only included RTs with correct trails, the original study suggests all RTs (including error trials) should be taken into account.

Rate-Correct Score (RCS). We also adopt an alternative speed-accuracy trade-off metric, the rate-correct score (RCS), that can be defined as:

$$RCS_{ij} = \frac{NC_{ij}}{\sum_{k=1}^{n_{ij}} RT_{ijk}} \tag{4}$$

where NC_{ij} is the number of correct classifications of the model i in condition of j, and the denominator reflects the total time the model i spent on training in condition of j (*i.e.,* the sum of RTs across all n_{ij} training of the model i in condition of j). RCS can be interpreted directly as the number of correct classifications per unit of time. Both methods compare accuracy and training time. The difference is that IES is only related to accuracy and running time, while RCS also considers the size of the training database.

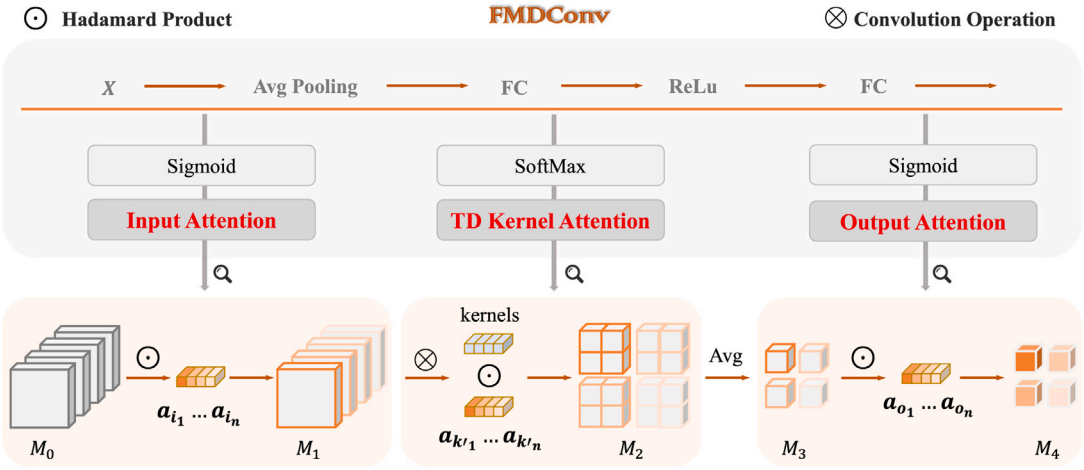


Fig. 2. The architecture of the Fast Multi-Attention Dynamic Convolution (FMDConv) block. It integrates three attention mechanisms: Input Attention, Temperature-Degraded (TD) Kernel Attention, and Output Attention. These attentions are computed via Sigmoid and SoftMax functions to adjust feature maps and convolution kernels dynamically.

3.3. Fast multi-attention dynamic convolution

In this section, we present the architecture of Fast Multi-Attention Dynamic Convolution (FMDConv). Unlike DynamicConv and ODConv, we introduce a multi-attention mechanism and optimize the integration of dynamic strategies across multiple kernel dimensions to enhance both efficiency and accuracy.

3.3.1. Architecture design

In this subsection, we present the proposed FMDConv block, namely Fast Multi-Attention Dynamic Convolution. Similar to Dynamic Convolution [14] and Omni-Dynamic Convolution [15], we adopt a multi-attention mechanism and calculate convolution from N learnable kernels with the same spatial size and channel dimension. However, based on those two proposed speed-accuracy trade-off metrics, we conclude that calculating kernel attention and spatial attention are computationally expensive, while both contribute very little to improving accuracy. Therefore, we replace these two attention mechanisms with temperature-degraded kernel attention originating from DynamicConv [14]. The overall architecture of the proposed FMDConv is illustrated in Fig. 2. The proposed FMDConv can be formulated as:

$$y = (\alpha_{i_1} \odot \alpha_{k'_1} \odot \alpha_{o_1} + \alpha_{i_2} \odot \alpha_{k'_2} \odot \alpha_{o_2} + \cdots + \alpha_{i_n} \odot \alpha_{k'_n} \odot \alpha_{o_n}) * x \tag{5}$$

where x, and y denote the input and output feature maps, respectively, while α_{i_n}, $\alpha_{k'_n}$ and α_{o_n} correspond to the input channel attention, temperature-degraded kernel attention, and output channel attention, respectively, as detailed in Algorithm 1.

As shown in Fig. 2, in input attention and output attention, the input information will first be squeezed by global average pooling, followed by a fully connected layer, a ReLu activation layer, and another fully connected layer to calculate the sample-dependent information. To compute the final attention values $\{\alpha_c, \alpha'_k, \alpha_f\}$, we again apply different non-linear activation operators individually on the extracted sample-dependent feature for input attention, temperature-degraded kernel attention, and output attention. For input and output attention activation functions, sigmoid is used, while we use SoftMax for kernel attention instead.

3.3.2. Temperature of SoftMax activation function

The SoftMax activation function is commonly used for kernel attention, but in the early stages of training, the uniform output of SoftMax can lead to slow convergence. To address this, we introduce a temperature degradation mechanism where the initial temperature is set to 40 and decreases by 3 after each epoch until the temperature reaches 1. The formula for temperature-degraded SoftMax is given by:

$$\sigma(z_i) = \frac{e^{z_i/T}}{\sum_{j=1}^{K} e^{z_j/T}} \tag{6}$$

where $\sigma(z_i)$ denotes the output probability for the ith kernel, z_i is the ith element of the input vector z, K is the total number of kernels,

Algorithm 1 Fast Multi-Attention Dynamic Convolution (FMDConv2d)

Input: Input tensor x, temperature T, kernel size k, number of kernels K, input channels C_{in}, output channels C_{out}
Output: Output tensor after dynamic convolution

Initialization:
Initialize attention layers Attention, Attention2, and convolution kernels W with Kaiming initialization.
Initialize bias terms b if applicable.

Step 1: Input Attention Computation
Compute input attention $A_{\text{input}} = \sigma(\text{Conv}(x))$ ▷ σ: Sigmoid activation
Compute output attention $A_{\text{output}} = \sigma(\text{Conv}(x))$

Step 2: Kernel Attention Computation
Compute kernel attention $A_{\text{kernel}} = \text{Softmax}\left(\frac{\text{Attention2}(x)}{T}\right)$ ▷ Temperature-scaled softmax

Step 3: Dynamic Convolution Calculation
Multiply input x by input attention: $x = x \times A_{\text{input}}$
Reshape input tensor x to shape $(1, C_{\text{in}} \times \text{batch size}, H, W)$
Reshape convolution kernel weights W to shape $(K, C_{\text{out}} \times C_{\text{in}}, k, k)$
Compute aggregate weight matrix: $W_{\text{agg}} = A_{\text{kernel}} \times W$
if bias is not None **then**
    Compute aggregate bias $b_{\text{agg}} = A_{\text{kernel}} \times b$
    Perform convolution: Output = $\text{Conv2d}(x, W_{\text{agg}}, b_{\text{agg}})$
else
    Perform convolution: Output = $\text{Conv2d}(x, W_{\text{agg}})$
end if
Reshape output to batch size and apply output attention: Output = Output $\times A_{\text{output}}$

Return: Output tensor after applying multi-attention dynamic convolution.

and T is the temperature parameter that controls the sharpness of the SoftMax distribution. We set the initial temperature to start at 40 and subtract 3 after each epoch of training until T. When $T = 1$, the formula is identical to normal SoftMax. Setting the temperature to decrease by 3 every epoch can greatly mitigate the slow start issue of SoftMax at the early epoch, which produces near one-hot output at the beginning of the training. When the temperature $T = 1$, this function reduces to the standard SoftMax. Our experiments show that setting the initial temperature to 40 improves Top-1 accuracy by 2.95% on the CIFAR-100 dataset.

4. Experiments

4.1. Benchmark and experiment setting

To evaluate the proposed FMDConv, we adopted ResNet-18 as the backbone due to its low computational and memory requirements, making it suitable for resource-constrained platforms. The experiments were conducted on widely used benchmarks, including CIFAR-10, CIFAR-100, and ImageNet (ILSVRC 2012). The CIFAR datasets consist of 50,000 training images and 10,000 testing images, while the ILSVRC2012 dataset contains 1,281,167 training images and 50,000 validation images across 1000 categories. Compared to CIFAR datasets, ImageNet offers higher resolution, and more diverse image categories, and presents a greater challenge for classification tasks.

In this study, we initially tuned the hyperparameters based on the training set performance, as our model demonstrated robustness across a wide range of hyperparameter settings (as presented in Section 4.4). However, following the reviewer's suggestion and adhering to best practices, we recognize the importance of using a validation set for hyperparameter optimization. In future work, we plan to incorporate a separate validation set to ensure hyperparameter tuning is conducted independently of the test set, thereby further enhancing the robustness and generalizability of our results.

For comparative analysis, we evaluated our method against existing dynamic convolution approaches, such as ODConv [15], DynamicConv, and CondConv, to comprehensively demonstrate the advantages of FMDConv.

In our training regimen for both CIFAR-10 and CIFAR-100, the initial learning rate was set to 0.1, with a decay factor of 20 applied every 30 epochs over 70 epochs of training. We used a weight decay of 1e-4, a dropout rate of 0.1, and a reduction rate of 0.0625. The training batch size was 32, and the test batch size was 70. For ImageNet, we adopted a different strategy, initializing the learning rate at 0.1 and reducing it by a factor of 30 every 30 epochs, for a total of 100 epochs.

All experiments were conducted on a system with an NVIDIA GeForce RTX 3080 GPU (10 GB GDDR6X memory), 32 GB Corsair VENGEANCE RGB PRO DDR4 RAM, and an Intel® Core i9-12900K CPU. The software environment included PyTorch 1.12.1, CUDA 11.3, and Python 3.9.

4.2. Speed-accuracy trade-off evaluation

We first evaluate the speed-accuracy trade-off of four attention mechanisms used in Omni-Dynamic Convolution with the proposed metrics (IES and RCS) on an image classification task. Table 2 illustrates the accuracy and time consumption of channel attention, kernel attention, spatial attention, and filter attention, respectively, on the CIFAR-10 dataset. The best two results are highlighted in bold. Our findings reveal that channel attention and filter attention achieve better accuracy (improvements of 1.13% and 1.12% on Top-1 accuracy, respectively) with a relatively small increase (12.53 s and 13.00 s extra per training epoch) in time consumption. However, kernel attention and spatial attention lead to significant time consumption with limited improvement in accuracy. We further conducted experiments on kernel attention with two and four kernels, where the kernel attention improves Top-1 accuracy by 0.45% with an extra 97.71 s of time cost per training epoch when the kernel number is two and by 0.9% with an extra 158.53 s with four kernels. We also used RCS as the key index to measure the effectiveness of these four attention mechanisms. We found that channel attention and filter attention outperform kernel attention and spatial attention with RCS scores of 915.32 and 908.05, compared to kernel attention and spatial attention with RCS scores of 373.76 and 481.63, respectively.

We conclude that spatial attention and kernel attention have little impact on the Top-1 accuracy of the CIFAR-10 dataset while significantly increasing time consumption.

In Fig. 3(a), we present the experimental effects of the four attention mechanisms on CIFAR-10. The horizontal axis represents the time of each training epoch, and the vertical axis represents the percentage of Top-1 accuracy. The bubbles in the upper-left corner indicate higher Top-1 accuracy rates with less time and better results. Fig. 3(b) illustrates the IES, which is the ratio of time to accuracy. A smaller ratio indicates better results with high accuracy and less time. Fig. 3(c) displays the RCS of each attention mechanism, which is the number of correct classifications per unit of time. A higher number indicates better results, implying that more correct images can be classified within a certain period.

4.3. Fast multi-attention dynamic convolution

We evaluate the performance of our FMDConv model on CIFAR-10, CIFAR-100, and ImageNet. The results of our experiments, presented in Table 3, show that our FMDConv model achieves the highest Top-1 accuracy of 94.21% and Top-5 accuracy of 99.85%, with a time per epoch of 114.7 s in CIFAR-10. In addition, FMDConv outperforms the baseline regarding IES and RCS scores by a significant margin, with 121.75 in IES points and 241.08 in RCS scores.

Similarly, on CIFAR-100, the FMDConv model achieves a top-1 accuracy of 74.99% and a top-5 accuracy of 93.61%, with a time per epoch of 115.16 s in CIFAR-100. Table 4 illustrates that our approach achieves 153.16 in IES points and 390.71 in RCS score.

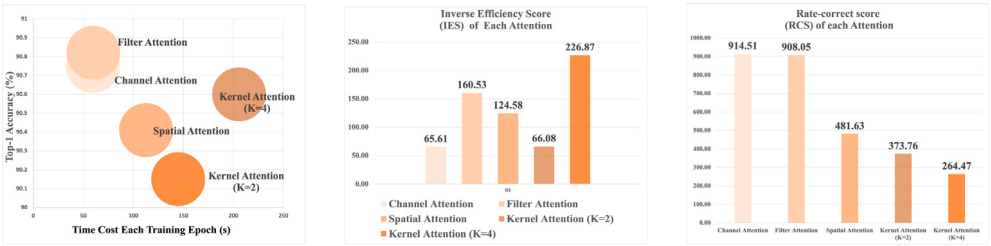
In our ImageNet training experiments, when we applied identical parameters, including learning rate, batch size, and number of epochs, our approach exhibited a substantial reduction in training time, nearly halving it in comparison to ODConv. Specifically, Table 5 compares the performance of different dynamic convolution models on the ImageNet validation set using ResNet18 as the backbone, trained for 100 epochs. Our proposed method, FMDConv (×4), achieved the highest Top-1 accuracy of 73.21% and Top-5 accuracy of 90.88%, outperforming CondConv, DynamicConv, and ODConv. In terms of efficiency, FMDConv showed clear advantages, with a time per epoch of 620.34 s, which is nearly half that of ODConv (1236.14 s). Additionally, FMDConv achieved the lowest Inverse Efficiency Score (IES) of 847.34 and the highest Rate-Correct Score (RCS) of 1511.98, demonstrating that it offers the best trade-off between speed and accuracy.

Based on the experimental results presented in Table 6, we compare the performance of various dynamic convolution models on the ImageNet validation set using ResNet50 as the backbone, trained for 100 epochs. Our proposed method, FMDConv (×4), achieved the best Top-1 accuracy of 78.34% and Top-5 accuracy of 93.57%, marginally outperforming ODConv (×4) in both metrics. Notably, FMDConv significantly reduced the computational overhead compared to ODConv, with a much lower Time per Epoch of 1028.57 s compared to 1780.35 s for ODConv. Additionally, FMDConv achieved superior efficiency, reflected in the lowest Inverse Efficiency Score (IES) of 1099.25 and a higher Rate-Correct Score (RCS) of 975.79, indicating that it provides better performance per unit of time.

Based on the experimental results shown in Table 7, we evaluate the performance of various dynamic convolution approaches on the ImageNet validation set, using MobileNetV2 (×0.5) as the backbone and training for 100 epochs. Our proposed FMDConv (×4) delivered the best

Table 2
Comparison of results of each attention on the Cifar-10 validation set with the ResNet18 backbones trained for 100 epochs. We set r = 0.1. The best results are bold.

Kernel number	Channel attention	Kernel attention	Spatial attention	Filter attention	Top-1%	Top-5%	Time per epoch/s	IES	RCS
1	–	–	–	–	89.7	99.65	47.01	–	–
1	✓	–	–	–	**90.83**	**99.68**	**59.54**	**65.65**	**915.32**
2	–	✓	–	–	90.15	99.72	144.72	160.53	373.76
4	–	✓	–	–	90.6	99.71	205.54	226.87	264.47
1	–	–	✓	–	90.41	99.76	112.63	124.58	481.63
1	–	–	–	✓	**90.82**	**99.82**	**60.01**	**66.08**	**908.05**



(a) Attentions Comparison (b) IES (c) RCS

Fig. 3. (a) Attentions Comparison on the CIFAR-10; (b) Inverse Efficiency Score (IES); (c) Rate-correct Score (RCS).

Table 3
Comparison of results on the Cifar-10 validation set with the ResNet18 backbones trained for 70 epochs. We set r = 0.1. The best results are bold.

Model	Top-1 accuracy	Top-5 accuracy	Time cost each epoch	IES	RCS
CondConv	81.19	98.96	100.50 s	123.78	484.72
DynamicConv	85.19	99.33	104.05 s	122.14	491.24
ODConv	93.82	99.82	223.61 s	238.34	251.74
Ours	**94.21**	**99.85**	114.7 s	**121.75**	**492.82**

Table 4
Comparison of results on the Cifar-100 validation set with the ResNet18 backbones trained for 100 epochs. We set r = 0.1. The best results are bold.

Model	Top-1 accuracy	Top-5 accuracy	Time cost each epoch	IES	RCS
CondConv	66.80	85.24	108.50 s	162.42	369.40
DynamicConv	67.21	86.89	105.98 s	157.68	380.50
ODConv	72.63	92.13	222.1 s	305.79	196.21
Ours	**74.99**	**93.61**	115.16 s	**153.57**	**390.71**

Table 5
Comparison of results on the ImageNet validation set with the ResNet18 backbones trained for 100 epochs. We set r = 0.0625. The best results are bold.

Model	Top-1 (%)	Top-5 (%)	Time per epoch (s)	IES	RCS
CondConv (×8)	71.99	90.27	625.68 s	869.12	1474.10
DynamicConv (×4)	72.76	90.79	618.45 s	849.98	1507.28
ODConv (×4)	73.09	90.86	1236.14 s	1691.26	757.52
Ours(×4)	**73.21**	**90.88**	**620.34 s**	**847.34**	**1511.98**

Table 6
Comparison of results on the ImageNet validation set with the ResNet50 backbones trained for 100 epochs. We set r = 0.0625. The best results are bold.

Model	Top-1 (%)	Top-5 (%)	Time per epoch (s)	IES	RCS
CondConv (×8)	75.20	93.12	990.12	1316.65	973.05
DynamicConv (×4)	75.82	93.16	1008.54	1330.18	963.16
ODConv (×4)	78.32	93.56	1780.35	2273.17	563.60
Ours(×4)	**78.34**	**93.57**	1028.57	**1099.25**	**975.79**

performance with a Top-1 accuracy of 70.23% and a Top-5 accuracy of 92.07%, slightly surpassing ODConv (×4) in both measures. Notably, FMDConv significantly improved computational efficiency, requiring only 87.37 s per epoch, which is a notable reduction compared to ODConv's 119.21 s. Moreover, FMDConv demonstrated enhanced overall efficiency with the lowest Inverse Efficiency Score (IES) of 124.41 and the highest Rate-Correct Score (RCS) of 10298.31, reflecting its superior balance between speed and accuracy when compared to the other models.

To further demonstrate the effectiveness of our FMDConv model, we visualize Grad-CAM++ results for different attention mechanisms using ResNet-18, as shown in Fig. 4. The experimental results indicate that input attention enhances the network's focus on discriminative regions before convolution operations, thereby improving the effectiveness of early feature extraction. Temperature-degraded kernel attention dynamically adjusts convolutional kernel weights, reinforcing fine-grained structural information while suppressing irrelevant background noise, leading to more stable feature representations. Output

Table 7
Comparison of results on the ImageNet validation set with the MobileNetv2(x0.5) backbones trained for 100 epochs. We set r = 0.0625. The best results are bold.

Model	Top-1 (%)	Top-5 (%)	Time per epoch (s)	IES	RCS
CondConv (×8)	66.41	90.32	83.27	125.39	10 217.64
DynamicConv (×4)	68.75	91.37	85.62	124.54	10 287.34
ODConv (×4)	70.21	91.95	119.21	169.79	7545.57
Ours(×4)	**70.23**	**92.07**	87.37	**124.41**	**10298.31**

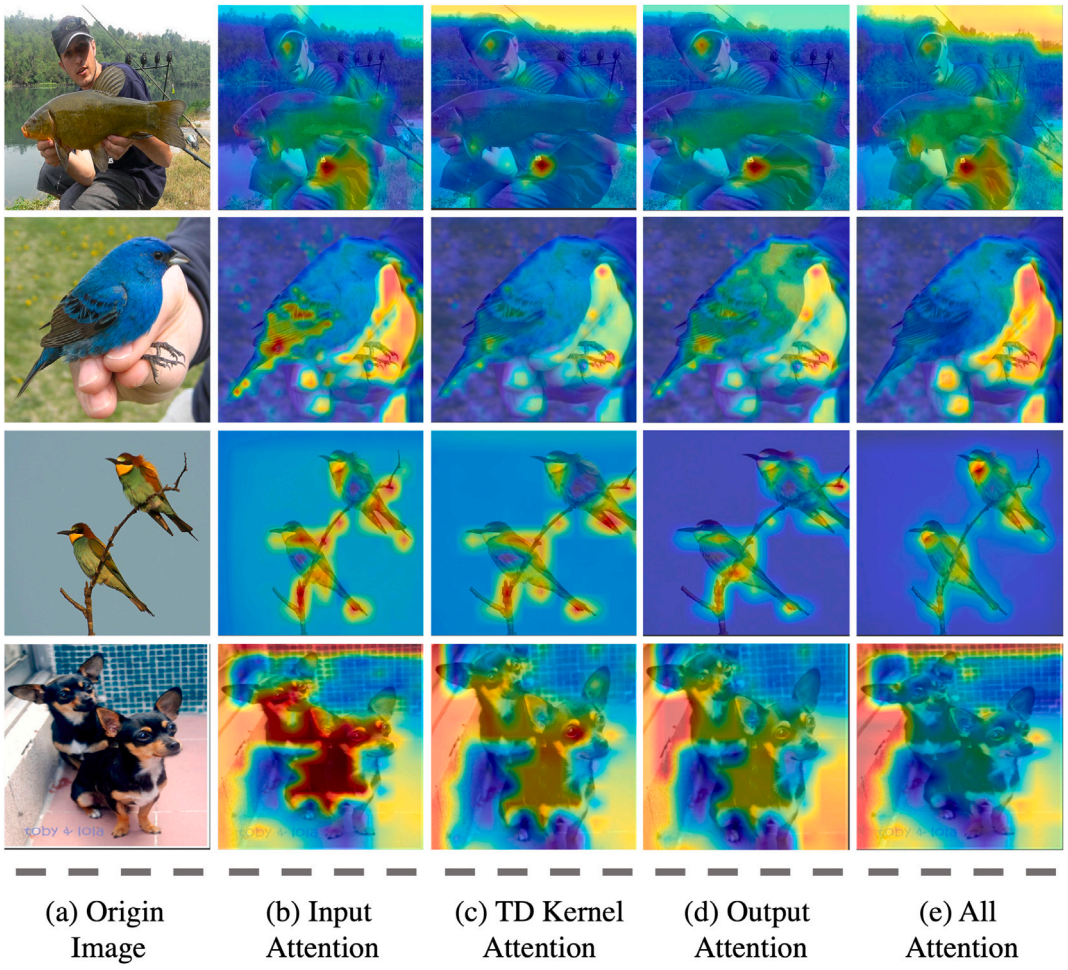


Fig. 4. Grad-CAM++ visualization results for multiple attention mechanisms on ImageNet. (a) Original Images, (b) Feature Maps with Input Attention, (c) Feature Maps with Temperature-Degraded (TD) Kernel Attention, (d) Feature Maps with Output Attention, and (e) Feature Maps with All Combined Attentions.

Table 8
Top-1 and Top-5 accuracy comparison of FMDConv with different initial temperatures on CIFAR-100. The best results are bold.

Temperature	Top-1 accuracy (%)	Top-5 accuracy (%)
1	72.04	92.29
10	72.19	92.46
22	73.53	92.97
31	73.65	92.82
34	73.65	93.47
37	74.45	93.26
40	**74.99**	**93.61**
43	73.64	92.87
46	73.58	93.10
49	73.10	92.62

attention further refines the extracted features by emphasizing class-relevant regions and reducing redundant activations, making the final feature maps more distinct.

When all three attention mechanisms work together, the network's focus significantly improves, optimizing both spatial selectivity and class discriminability, as illustrated in Fig. 4(e). Unlike ODConv, which applies attention across all dimensions at a higher computational cost, FMDConv leverages a more lightweight multi-attention mechanism to achieve superior feature selectivity. In conclusion, our FMDConv model performs better in terms of model accuracy and training time cost than the baseline ODConv model on both CIFAR and ImageNet datasets. This improvement is mainly due to the introduction of the multi-attention mechanism in our FMDConv model.

4.4. Ablation studies

We conducted ablation studies on the CIFAR-100 dataset to evaluate the impact of various factors on the performance of the proposed FMDConv.

Effect of Initial Temperature. We introduced a temperature mechanism in the kernel attention module to enhance the convergence rate of dynamic convolution. As shown in Table 8, we compared the top-1

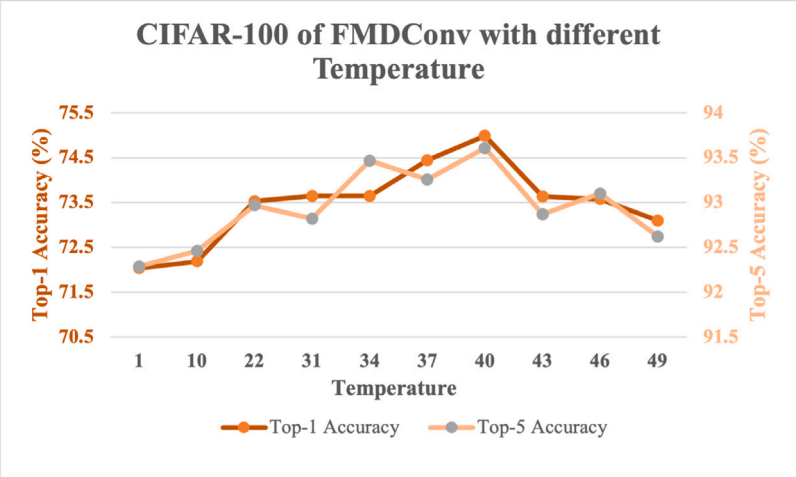


Fig. 5. Top-1 and Top-5 accuracy comparison for FMDConv under different initial temperatures on CIFAR-100.

Table 9
Comparison of results of different kernel numbers on the Cifar-100 validation set with the ResNet18 backbones trained for 100 epochs. We set r = 0.1. The best results are bold.

Kernel number	Params	Top-1 accuracy (%)	Top-5 accuracy (%)	Time cost each epoch (s)
K=1	12.23M	74.13	93.06	106.11
K=2	23.22M	73.80	93.01	110.64
K=4	45.20M	**74.99**	**93.61**	115.16
K=6	67.18M	74.23	93.05	117.07
K=8	89.16M	74.67	93.09	120.11
K=16	177.09M	73.13	92.64	132.26

Table 10
Comparison of results of different start learning rates on the Cifar-100 validation set with the ResNet18 backbones trained for 70 epochs. We set r = 0.1. The best results are bold.

Start LR	Top-1 accuracy (%)	Top-5 accuracy (%)	Time per epoch (s)
1/4	73.73	93.11	112.99
1/8	73.78	92.75	113.22
1/10	**74.99**	**93.61**	115.16
1/16	74.50	93.12	115.59

and top-5 accuracy rates under different initial temperatures to identify the optimal value. When the initial temperature decreased from 40, the model achieved its highest Top-1 accuracy, improving by 2.75% compared to the baseline temperature of $T = 1$, as shown in Fig. 5. Similarly, the Top-5 accuracy increased by 1.32%. The accuracy initially increases with temperature until it peaks at 40, after which it begins to decline. The best results are highlighted in bold.

Effect of the Number of Convolution Kernels. We evaluated the impact of varying the number of convolution kernels on classification accuracy. As shown in Table 9, when the initial temperature is fixed at $T = 40$, the Top-1 accuracy reaches a maximum of 74.99% with four kernels (K = 4). Beyond K=4, increasing the number of kernels does not lead to further improvements, as the accuracy stabilizes or decreases slightly. The corresponding Top-5 accuracy follows a similar trend, reaching 93.61%.

Effect of Learning Rate. In this set of experiments, we tested different initial learning rates (1/2, 1/4, 1/8, 1/10, and 1/16) to evaluate their influence on accuracy. As shown in Table 10, a learning rate of 1/10 produced the best Top-1 accuracy (74.99%) and Top-5 accuracy (93.61%) with minimal additional time per epoch. Lower learning rates did not yield significant improvements and, in some cases, resulted in a slight reduction in accuracy.

5. Limitation

While FMDConv shows clear improvements in balancing speed and accuracy across various benchmarks, its deployment in highly resource-constrained environments may still face challenges due to the additional computational overhead introduced by the multi-attention mechanisms. Additionally, the model's performance has been validated primarily on image classification tasks with ResNet architectures, leaving its efficacy on more complex tasks (such as object detection or segmentation) and other network architectures relatively unexplored.

6. Conclusion

In this paper, we introduced FMDConv, a novel dynamic convolution block designed to balance speed and accuracy. By leveraging three optimal attention mechanisms, FMDConv demonstrated superior performance on image classification benchmarks, making it suitable for resource-constrained environments. Our proposed metrics, the Inverse Efficiency Score (IES) and the Rate-Correct Score (RCS), effectively quantify the trade-offs between efficiency and accuracy. Future work will focus on extending FMDConv to more complex tasks, such as object detection, while further optimizing computational efficiency and exploring adaptive hyperparameter strategies for broader applicability.

CRediT authorship contribution statement

Tianyu Zhang: Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Fan Wan:** Writing – original draft, Resources, Project administration. **Haoran Duan:** Formal analysis, Data curation, Conceptualization. **Kevin W. Tong:** Investigation, Funding acquisition. **Jingjing Deng:** Writing – review & editing, Software, Methodology. **Yang Long:** Visualization, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Royal Society International Exchanges Scheme-Towards Collaborative Cloud-Edge Deep Learning Deployment under Grant IEC/NSFC/223523; National Edge AI Hub for Real Data: Edge Intelligence for Cyber-disturbances and Data Quality EP/Y028813/1; and UK Medical Research Council (MRC) Innovation Fellowship under Grant MR/S003916/2.

Data availability

Data will be made available on request.

References

[1] J. Liu, M. Gong, Y. Gao, Y. Lu, H. Li, Bidirectional interaction of CNN and transformer for image inpainting, Knowl.-Based Syst. (2024) 112046.
[2] Z. Yang, F. Emmert-Streib, Optimal performance of binary relevance CNN in targeted multi-label text classification, Knowl.-Based Syst. 284 (2024) 111286.
[3] Y. Zhou, J. Li, J. Chi, W. Tang, Y. Zheng, Set-CNN: A text convolutional neural network based on semantic extension for short text classification, Knowl.-Based Syst. 257 (2022) 109948.
[4] S. Chen, T. Shu, H. Zhao, Y.Y. Tang, MASK-CNN-transformer for real-time multi-label weather recognition, Knowl.-Based Syst. 278 (2023) 110881.
[5] Y. Amit, P. Felzenszwalb, R. Girshick, Object detection, Comput. Vision: A Ref. Guid. (2020) 1–9.
[6] N. Hoanh, T.V. Pham, Focus-attention approach in optimizing DETR for object detection from high-resolution images, Knowl.-Based Syst. 296 (2024) 111939.
[7] Y. Guo, Y. Liu, T. Georgiou, M.S. Lew, A review of semantic segmentation using deep neural networks, Int. J. Multimed. Inf. Retr. 7 (2018) 87–93.
[8] W. Yue, Z. Zhou, Y. Cao, et al., Cross-modal domain generalization semantic segmentation based on fusion features, Knowl.-Based Syst. 302 (2024) 112356.
[9] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural Comput. 29 (9) (2017) 2352–2449.
[10] A. Rahman, L. He, H. Wang, Activation function optimization scheme for image classification, Knowl.-Based Syst. (2024) 112502.
[11] S. Huang, T. Fu, X. Han, J. Fan, H. Song, D. Xiao, G. Ma, J. Yang, Domain base dynamic convolution and distance map guidance for anterior mediastinal lesion segmentation, Knowl.-Based Syst. 296 (2024) 111881.
[12] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[13] B. Yang, G. Bender, Q.V. Le, J. Ngiam, Condconv: Conditionally parameterized convolutions for efficient inference, Adv. Neural Inf. Process. Syst. 32 (2019).
[14] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, Z. Liu, Dynamic convolution: Attention over convolution kernels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11030–11039.
[15] C. Li, A. Zhou, A. Yao, Omni-dimensional dynamic convolution, 2022, arXiv preprint arXiv:2209.07947.
[16] A.V. Reed, Speed-accuracy trade-off in recognition memory, Science 181 (4099) (1973) 574–576.
[17] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[21] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.
[22] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.
[23] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, Y. Wang, Dynamic neural networks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2021) 7436–7456.
[24] J. Sun, D. Li, Object tracking with channel group regularization and smooth constraints using improved dynamic convolution kernels in ITS, Multimedia Tools Appl. (2022) 1–25.
[25] H. Wu, S. Wu, Y. Wu, S. Pan, AttCluster-MDGCNs: multiscale dynamic graph convolution networks with an attention cluster for skeletal-based action, Multimedia Tools Appl. 81 (13) (2022) 18855–18874.
[26] M. Zhao, Z. Hu, S. Li, S. Bi, Z. Sun, Mask attention-guided graph convolution layer for weakly supervised temporal action detection, Multimedia Tools Appl. (2022) 1–18.
[27] S. Huang, M. Zhang, Y. Ke, X. Bi, Y. Kong, Image steganalysis based on attention augmented convolution, Multimedia Tools Appl. 81 (14) (2022) 19471–19490.
[28] Y.-F. Zhang, T. Xia, Y. Liu, 3D convolution network and siamese-attention mechanism for expression recognition, Multimedia Tools Appl. 78 (2019) 30355–30371.
[29] Y. Li, Y. Chen, X. Dai, M. Liu, D. Chen, Y. Yu, L. Yuan, Z. Liu, M. Chen, N. Vasconcelos, Revisiting dynamic convolution via matrix decomposition, 2021, arXiv preprint arXiv:2103.08756.
[30] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al., Speed/accuracy trade-offs for modern convolutional object detectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7310–7311.
[31] R. Castro-Zunti, K.J. Chae, Y. Choi, G.Y. Jin, S.-b. Ko, Assessing the speed-accuracy trade-offs of popular convolutional neural networks for single-crop rib fracture classification, Comput. Med. Imaging Graph. 91 (2021) 101937.
[32] M. Javadi, S.M. Azar, S. Azami, S.S. Ghidary, S. Sadeghnejad, J. Baltes, Humanoid robot detection using deep learning: a speed-accuracy tradeoff, in: RoboCup 2017: Robot World Cup XXI 11, Springer, 2018, pp. 338–349.
[33] D. Chaves, E. Fidalgo, E. Alegre, P. Blanco, Improving speed-accuracy trade-off in face detectors for forensic tools by image resizing, V Jornadas Nac. de InvestigaciÉn Ciberseguridad (JNIC) (2019) 1–2.
[34] X. Li, Y. Zhou, Z. Pan, J. Feng, Partial order pruning: for best speed/accuracy trade-off in neural architecture search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9145–9153.
[35] Z. Qin, W. Sun, H. Deng, D. Li, Y. Wei, B. Lv, J. Yan, L. Kong, Y. Zhong, Cosformer: Rethinking softmax in attention, 2022, arXiv preprint arXiv:2202.08791.
[36] Z. Gao, J. Xie, Q. Wang, P. Li, Global second-order pooling convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3024–3033.
[37] J.T. Townsend, F.G. Ashby, et al., Stochastic Modeling of Elementary Psychological Processes, CUP Archive, 1983.
[38] D.J. Woltz, C.A. Was, Availability of related long-term memory during and after attention focus in working memory, Mem. Cogn. 34 (2006) 668–684.