Dynamic Convolution: Attention over Convolution Kernels

Yinpeng Chen Xiyang Dai Mengchen Liu Dongdong Chen Lu Yuan Zicheng Liu
Microsoft
{yiche,xidai,mengcliu,dochen,luyuan,zliu}@microsoft.com

Abstract

Light-weight convolutional neural networks (CNNs) suffer performance degradation as their low computational budgets constrain both the depth (number of convolution layers) and the width (number of channels) of CNNs, resulting in limited representation capability. To address this issue, we present Dynamic Convolution, a new design that increases model complexity without increasing the network depth or width. Instead of using a single convolution kernel per layer, dynamic convolution aggregates multiple parallel convolution kernels dynamically based upon their attentions, which are input dependent. Assembling multiple kernels is not only computationally efficient due to the small kernel size, but also has more representation power since these kernels are aggregated in a non-linear way via attention. By simply using dynamic convolution for the state-of-the-art architecture MobileNetV3-Small, the top-1 accuracy of ImageNet classification is boosted by 2.9% with only 4% additional FLOPs and 2.9 AP gain is achieved on COCO keypoint detection.

1. Introduction

Interest in building light-weight and efficient neural networks has exploded recently. It not only enables new experiences on mobile devices, but also protects user’s privacy from sending personal information to the cloud. Recent works (e.g. MobileNet [11, 25, 10] and ShuffleNet [40, 22]) have shown that both efficient operator design (e.g. depthwise convolution, channel shuffle, squeeze-and-excitation [12], asymmetric convolution [5]) and architecture search ([27, 6, 2]) are important for designing efficient convolutional neural networks.

However, even the state-of-the-art efficient CNNs (e.g. MobileNetV3 [10]) suffer significant performance degradation when the computational constraint becomes extremely low. For instance, when the computational cost of MobileNetV3 reduces from 219M to 66M Multi-Adds, the top-1 accuracy of ImageNet classification drops from 75.2% to 67.4%. This is because the extremely low computational cost severely constrains both the network depth (number of layers) and width (number of channels), which are crucial for the network performance but proportional to the computational cost.

This paper proposes a new operator design, named dynamic convolution, to increase the representation capability with negligible extra FLOPs. Dynamic convolution uses a set of $K$ parallel convolution kernels $\{\tilde{W}_k, \tilde{b}_k\}$ instead of using a single convolution kernel per layer (see Figure 2). These convolution kernels are aggregated dynamically

\[
\tilde{W} = \sum_k \pi_k(x) \tilde{W}_k
\]

for each individual input $x$ (e.g. im-
age) via input dependent attention \( \pi_k(x) \). The biases are aggregated using the same attention \( b = \sum_k \pi_k(x) b_k \). Dynamic convolution is a non-linear function with more representation power than its static counterpart. Meanwhile, dynamic convolution is computationally efficient. It does not increase the depth or width of the network, as the parallel convolution kernels share the output channels by aggregation. It only introduces extra computational cost to compute attentions \( \{ \pi_k(x) \} \) and aggregate kernels, which is negligible compared to convolution. The key insight is that within reasonable cost of model size (as convolution kernels are small), dynamic kernel aggregation provides an efficient way (low extra FLOPs) to boost representation capability.

Dynamic convolutional neural networks (denoted as DY-CNNs) are more difficult to train, as they require joint optimization of all convolution kernels and the attention across multiple layers. We found two keys for efficient joint optimization: (a) constraining the attention output as \( \sum_k \pi_k(x) = 1 \) to facilitate the learning of attention model \( \pi_k(x) \), and (b) flattening attention (near-uniform) in early training epochs to facilitate the learning of convolution kernels \( \{ \tilde{W}_k, \tilde{b}_k \} \). We simply integrate these two keys by using softmax with a large temperature for kernel attention.

We demonstrate the effectiveness of dynamic convolution on both image classification (ImageNet) and keypoint detection (COCO). Without bells and whistles, simply replacing static convolution with dynamic convolution in MobileNetV2 and V3 achieves solid improvement with only a slight increase (4%) of computational cost (see Figure 1). For instance, with 100M Multi-Adds budget, our method gains 4.5% and 2.9% top-1 accuracy on image classification for MobileNetV2 and MobileNetV3, respectively.

2. Related Work

Efficient CNNs: Recently, designing efficient CNN architectures [14, 11, 25, 10, 40, 22] has been an active research area. SqueezeNet [14] reduces the number of parameters by using 1 \times 1 convolution extensively in the fire module. MobileNetV1 [11] substantially reduces FLOPs by decomposing a 3 \times 3 convolution into a depthwise convolution and a pointwise convolution. Based upon this, MobileNetV2 [25] introduces inverted residuals and linear bottlenecks. MobileNetV3 [10] applies squeeze-and-excitation [12] in the residual layer and employs a platform-aware neural architecture approach [27] to find the optimal network structures. ShuffleNet further reduces MAdds for 1 \times 1 convolution by channel shuffle operations. ShiftNet [31] replaces expensive spatial convolution by the shift operation and pointwise convolutions. Compared with existing methods, our dynamic convolution can be used to replace any static convolution kernels (e.g. 1 \times 1, 3 \times 3, depthwise convolution, group convolution) and is complementary to other advanced operators like squeeze-and-excitation.

Model Compression and Quantization: Model compression [7, 21, 9] and quantization [3, 41, 38, 36, 28] approaches are also important for learning efficient neural networks. They are complementary to our work, helping reduce the model size for our dynamic convolution method.

Dynamic Deep Neural Networks: Our method is related to recent works of dynamic neural networks [17, 20, 29, 32, 37, 13] that focus on skipping part of an existing model based on input image. D^2NN [20], SkipNet [29] and BlockDrop [32] learn an additional controller for skipping decision by using reinforcement learning. MSDNet [13] allows early-exit based on the current prediction confidence. Slimmable Nets [37] learns a single neural network executable at different width. Once-for-all [1] proposes a progressive shrinking algorithm to train one network that supports multiple sub-networks. The accuracy for these sub-networks is the same as independently trained networks. Compared with these works, our method has two major differences. Firstly, our method has dynamic convolution kernels but static network structure, while existing works have static convolution kernels but dynamic network structure. Secondly, our method does not require an additional controller. The attention is embedded in each layer, enabling end-to-end training. Compared to the concurrent work [35], our method is more efficient with better performance.

Neural Architecture Search: Recent research works in neural architecture search (NAS) are powerful on finding high-accuracy neural network architectures [42, 24, 43, 19, 34] as well as hardware-aware efficient network architectures [2, 27, 30]. The hardware-aware NAS methods incorporate hardware latency into the architecture search process, by making it differentiable. [6] proposed single path supernet to optimize all architectures in the search space simultaneously, and then perform evolutionary architecture search to handle computational constraints. Based upon NAS, MobileNetV3 [10] shows significant improvements over human-designed baselines (e.g. MobileNetV2 [25]). Our dynamic convolution method can be easily used in advanced architectures found by NAS. Later in this paper, we will show that dynamic convolution not only improves the performance for human-designed networks (e.g. MobileNetV2), but also boosts the performance for automatically searched architectures (e.g. MobileNetV3), with low extra FLOPs. In addition, our method provides a new and effective component to enrich the search space.

3. Dynamic Convolutional Neural Networks

We describe dynamic convolutional neural networks (DY-CNNs) in this section. The goal is to provide better trade-off between network performance and computational burden, within the scope of efficient neural networks. The two most popular strategies to boost the performance are making neural networks “deeper” or “wider”. How-
ever, they both incur heavy computation cost, thus are not friendly to efficient neural networks.

We propose dynamic convolution, which does not increase either the depth or the width of the network, but increase the model capability by aggregating multiple convolution kernels via attention. Note that these kernels are assembled differently for different input images, from where dynamic convolution gets its name. In this section, We firstly define the generic dynamic perceptron, and then apply it to convolution.

3.1. Preliminary: Dynamic Perceptron

**Definition:** Let us denote the traditional or static perceptron as $y = g(W^T x + b)$, where $W$ and $b$ are weight matrix and bias vector, and $g$ is an activation function (e.g. ReLU [23, 15]). We define the dynamic perceptron by aggregating multiple ($K$) linear functions $\{\tilde{W}^T_k x + \tilde{b}_k\}$ as follows:

$$y = g(\tilde{W}^T(x)x + \tilde{b}(x))$$

$$\tilde{W}(x) = \sum_{k=1}^{K} \pi_k(x)\tilde{W}_k, \tilde{b}(x) = \sum_{k=1}^{K} \pi_k(x)\tilde{b}_k$$

s.t. $0 \leq \pi_k(x) \leq 1, \sum_{k=1}^{K} \pi_k(x) = 1,$ \hspace{1cm} (1)

where $\pi_k$ is the attention weight for the $k^{th}$ linear function $\tilde{W}^T_k x + \tilde{b}_k$. Note that the aggregated weight $\tilde{W}(x)$ and bias $\tilde{b}(x)$ are functions of input and share the same attention.

**Attention:** the attention weights $\{\pi_k(x)\}$ are not fixed, but vary for each input $x$. They represent the optimal aggregation of linear models $\{\tilde{W}^T_k x + \tilde{b}_k\}$ for a given input. The aggregated model $\tilde{W}^T(x)x + \tilde{b}(x)$ is a non-linear function. Thus, dynamic perceptron has more representation power than its static counterpart.

**Computational Constraint:** compared with static perceptron, dynamic perceptron has the same number of output channels but bigger model size. It also introduces two additional computations: (a) computing the attention weights $\{\pi_k(x)\}$, and (b) aggregating parameters based upon attention $\sum_k \pi_k \tilde{W}_k$ and $\sum_k \pi_k \tilde{b}_k$. The additional computational cost should be significantly less than the cost of computing $W^T x + b$. Mathematically, the computational constraint can be represented as follows:

$$O(\tilde{W}^T x + \tilde{b}) \gg O(\sum_k \pi_k \tilde{W}_k) + O(\sum_k \pi_k \tilde{b}_k) + O(\pi(x))$$ \hspace{1cm} (2)

where $O(\cdot)$ measures the computational cost (e.g. FLOPs). Note that fully connected layer does not satisfy this, while convolution is a proper fit for this constraint.

3.2. Dynamic Convolution

In this subsection, we showcase a specific dynamic perceptron, dynamic convolution that satisfies the computational constraint (Eq. 2). Similar to dynamic perceptron, dynamic convolution (Figure 3) has $K$ convolution kernels that share the same kernel size and input/output dimensions. They are aggregated by using the attention weights $\{\pi_k\}$. Following the classic design in CNN, we use batch normalization and an activation function (e.g. ReLU) after the aggregated convolution to build a dynamic convolution layer.

**Attention:** we apply squeeze-and-excitation [12] to compute kernel attentions $\{\pi_k(x)\}$ (see Figure 3). The global spatial information is firstly squeezed by global average pooling. Then we use two fully connected layers (with a ReLU between them) and softmax to generate normalized attention weights for $K$ convolution kernels. The first fully connected layer reduces the dimension by 4. Different from SENet [12] which computes attentions over output channels, we compute attentions over convolution kernels. The computation cost for the attention is cheap. For an input feature map with dimension $H \times W \times C_{in}$, the attention requires $O(\pi(x)) = HC_{in}C_{out}D^2_k/4 + C_{in}K/4$ Multi-Adds. This is much less than the computational cost of convolution, i.e. $O(\tilde{W}^T x + \tilde{b}) = HC_{in}C_{out}D^2_k$ Multi-Adds, where $D_k$ is the kernel size, and $C_{out}$ is the number of output channels.

**Kernel Aggregation:** aggregating convolution kernels is computationally efficient due to the small kernel size. Aggregating $K$ convolution kernels with kernel size $D_k \times D_k$, $C_{in}$ input channels and $C_{out}$ output channels introduces $KC_{in}C_{out}D^2_k + KC_{out}$ extra Multi-Adds. Compared with the computational cost of convolution ($HC_{in}C_{out}D^2_k$), the extra cost is negligible if $K \ll HW$. Table 1 shows the computational cost of using dynamic convolution in MobileNetV2. For instance, when using MobileNetV2 ($\times 1.0$), dynamic convolution with $K = 4$ kernels only increases the computation cost by 4%. Note that even though dynamic convolution increases the model size, it does not increase
Table 1. Mult-Adds of static convolution and dynamic convolution in MobileNetV2 with four different width multipliers (×1.0, ×0.75, ×0.5, and ×0.35).

<table>
<thead>
<tr>
<th></th>
<th>×1.0</th>
<th>×0.75</th>
<th>×0.5</th>
<th>×0.35</th>
</tr>
</thead>
<tbody>
<tr>
<td>static</td>
<td>300.0M</td>
<td>209.0M</td>
<td>97.0M</td>
<td>59.2M</td>
</tr>
<tr>
<td>K=2</td>
<td>309.3M</td>
<td>215.6M</td>
<td>100.5M</td>
<td>61.5M</td>
</tr>
<tr>
<td>K=4</td>
<td>312.9M</td>
<td>217.5M</td>
<td>101.4M</td>
<td>62.5M</td>
</tr>
<tr>
<td>K=6</td>
<td>316.3M</td>
<td>219.5M</td>
<td>102.3M</td>
<td>62.5M</td>
</tr>
<tr>
<td>K=8</td>
<td>319.8M</td>
<td>221.4M</td>
<td>103.2M</td>
<td>62.9M</td>
</tr>
</tbody>
</table>

the output dimension of each layer. The amount of the increase is acceptable as convolution kernels are small.

**From CNNs to DY-CNNs:** dynamic convolution can be easily used as a drop-in replacement for any convolution (e.g. 1 × 1 conv, 3 × 3 conv, group convolution, depthwise convolution) in any CNN architecture. It is also complementary to other operators (like squeeze-and-excitation [12]) and activation functions (e.g. ReLU6, h-swish [10]). In the rest of the paper, we use prefix **DY-** for the networks that use dynamic convolution. For example, **DY-MobileNetV2** refers to using dynamic convolution in MobileNetV2. We also use weight \( \tilde{W}_k \) to denote a convolution kernel and ignore bias \( b_k \), for the sake of brevity.

### 4. Two Insights of Training Deep DY-CNNs

Training deep DY-CNNs is challenging, as it requires joint optimization of all convolution kernels \( \{ \tilde{W}_k \} \) and attention model \( \pi_k(x) \) across multiple layers. In this section, we discuss two insights for more efficient joint optimization, which are crucial especially to deep DY-CNNs.

#### 4.1. Insight 1: Sum the Attention to One

The first insight is: **constraining the attention output can facilitate the learning of attention model \( \pi_k(x) \).** Specifically, we have the constraint \( \sum_k \pi_k(x) = 1 \) in Eq. (1) to keep the aggregated kernel \( \tilde{W} = \sum_k \pi_k \tilde{W}_k \) within the convex hull of \( \{ \tilde{W}_k \} \) in the kernel space. Figure 4 shows an example with 3 convolution kernels. The constraint \( 0 \leq \pi_k(x) \leq 1 \) only keeps the aggregated kernel within the two pyramids. The sum-to-one constraint further compresses the kernel space to a triangle. It compresses the red line that comes from the origin into a dot by normalizing the attention sum. This normalization significantly simplifies the learning of \( \pi_k(x) \), when it is jointly optimized with \( \{ \tilde{W}_k \} \) in a deep network. **Softmax** is a natural choice of \( \pi_k(x) = 1 \).

#### 4.2. Insight 2: Near-uniform Attention in Early Training Epochs

The second insight is: **near-uniform attention can facilitate the learning of all kernels \( \{ \tilde{W}_k \} \) during early training epochs.** This is because near-uniform attention enables more convolution kernels to be optimized simultaneously.

**Softmax** does **NOT** work well on this due to its near one-hot output. It only allows a small subset of kernels across layers to be optimized. Figure 5-(Left) shows that the training converges slowly when using softmax (**blue** curves) to compute attention. Here, **DY-MobileNetV2** with width multiplier ×0.5 is used. The final top-1 accuracy (64.8%) is even worse than its static counterpart (65.4%). This inefficiency is related to the number of dynamic convolution layers. To validate this, we reduce the number of dynamic convolution layers by 3 (only use dynamic convolution for the last 1 × 1 convolution in each bottleneck residual block) and expect faster convergence in training. The training and validation errors are shown in Figure 5-(Right) (**blue** curves). As we expected, the training converges faster with higher top-1 accuracy (65.9%) at the end.

We address this inefficiency in training deeper DY-CNNs by using a large temperature in softmax to flatten attention as follows:

\[
\pi_k = \frac{\exp(z_k/\tau)}{\sum_j \exp(z_j/\tau)},
\]

where \( z_k \) is the output of the second FC layer in attention branch (see Figure 3), and \( \tau \) is the temperature. The original softmax is a special case (\( \tau = 1 \)). As \( \tau \) increases, the output is less sparse. When using a large temperature \( \tau = 30 \), the training becomes significantly more efficient (see the **red** curves in Figure 5-(Left)). As a result, the top-1 accuracy boosts to 69.4%. The larger temperature is also helpful when stacking fewer dynamic convolution layers (see **red** curves in Figure 5-(Right)).

**Temperature annealing**, i.e. reducing \( \tau \) from 30 to 1 linearly in the first 10 epochs, can further improve the top-1 accuracy (from 69.4% to 69.9%). These results support that near-uniform attention in early training epochs is crucial.

### 4.3. Relation to Concurrent Work

These two insights are the key differences between our method and the concurrent work (CondConv [35]), which uses sigmoid to compute kernel attention. Even though sigmoid provides near-uniform attention in early training
epochs, it has significantly larger kernel space (two pyramids in Figure 4) than our method (the shaded triangle in Figure 4). Thus, learning attention model $\pi_k(x)$ becomes more difficult. As a result, our method has less kernels per layer, smaller model size, less computations but achieves higher accuracy (see Table 2).

### 5. Experiments: ImageNet Classification

In this section, we present experimental results of dynamic convolution along with comprehensive ablations on ImageNet [4] classification. ImageNet has 1000 classes, including 1,281,167 images for training and 50,000 images for validation.

#### 5.1. Implementation Details

We evaluate dynamic convolution on three architectures (ResNet [8], MobileNetV2 [25], and MobileNetV3 [10]), by using dynamic convolution for all convolution layers except the first layer. Each layer has $K = 4$ convolution kernels. The batch size is 256. We use different training setups for the three architectures as follows:

**Training setup for DY-ResNet:** The initial learning rate is 0.1 and drops by 0.1 at epoch 30, 60 and 90. The weight decay is 1e-4. All models are trained using SGD optimizer with 0.9 momentum for 100 epochs. We use dropout rate 0.1 before the last layer of DY-ResNet-18.

**Training setup for DY-MobileNetV2:** The initial learning rate is 0.05 and is scheduled to arrive at zero within a single cosine cycle. The weight decay is 4e-5. All models are trained using SGD optimizer with 0.9 momentum for 300 epochs. To prevent overfitting, we use label smoothing and dropout before the last layer for larger width multipliers ($\times 1.0$ and $\times 0.75$). The dropout rate is 0.2 and 0.1 for $\times 1.0$ and $\times 0.75$, respectively. Mixup [39] is used for $\times 1.0$.

**Training setup for DY-MobileNetV3:** The initial learning rate is 0.1 and is scheduled to arrive at zero within a single cosine cycle. The weight decay is 3e-5. We use SGD optimizer with 0.9 momentum for 300 epochs and dropout rate of 0.2 before the last layer.

### 5.2. Inspecting DY-CNN

We inspect if DY-CNN is dynamic, using DY-MobileNetV2 $\times 0.5$, which has $K = 4$ kernels per layer and is trained by using $\tau = 30$. Two properties are expected if it is dynamic: (a) the convolution kernels are diverse per layer, and (b) the attention is input dependent. We examine these two properties by contradiction. Firstly, if the convolution kernels are not diverse, the performances will be stable if different attentions are used. Thus, we vary the kernel aggregation per layer in three different ways: averaging $\sum W_k/K$, choosing the convolution kernel with the maximum attention $\tilde{W}_{\text{argmax}}(\pi_k)$, and random shuffling attention over kernels per image $\sum \pi_j(x)\tilde{W}_{k, j}$. Compared with using the original attention, the performances of these variations are significantly degraded (shown in Table 3). When choosing the convolution kernel with the maximum attention, the top-1 accuracy (0.1) is as low as randomly choosing a class. The significant instability confirms the diversity of convolution kernels. In addition, we shuffle attentions across images to check if the attention is input dependent. The poor accuracy (27.3%) indicates that it is crucial for each image to use its own attention.

Furthermore, we inspect the attention across layers and find that attentions are flat at low levels and sparse at high levels. This is helpful to explain why variations in Table 3...
have poor accuracy. For instance, averaging kernels with sparse attention at high levels or picking one convolution kernel (with the maximum attention) at low levels (where attention is flat) is problematic. Table 4 shows how attention affects the performance across layers. We group layers by their input resolutions, and switch on/off attention for these groups. If attention is switched off for a resolution, its effective at higher layers with lower resolutions. Table 4 shows how attention affects the performance across layers. W e group layers by their input resolutions, and switch on/off attention for these groups. If attention is switched off for a resolution, its effective at higher layers with lower resolutions.

5.3. Ablation Studies

We perform a number of ablations on DY-MobileNetV2 and DY-MobileNetV3. The default setup includes using $K = 4$ kernels per layer and $\tau = 30$.

The number of convolution kernels ($K$): the hyperparameter $K$ controls the model complexity. Figure 6 shows the classification accuracy and computational cost for dynamic convolution with different $K$. We compare DY-MobileNetV2 with MobileNetV2 on different depth/width multipliers. Firstly, the dynamic convolution outperforms its static counterpart for all depth/width multipliers, even with small $K = 2$. This demonstrates the strength of our method. In addition, the accuracy stops increasing once $K$ is larger than 4. This is because as $K$ increases, even though the model has more representation power, it is more difficult to optimize all convolution kernels and attention simultaneously and the network is more prone to over-fitting.

**Dynamic convolution at different layers:** Table 5 shows the classification accuracy for using dynamic convolution at three different layers ($1 \times 1$ conv, $3 \times 3$ depthwise conv, $1 \times 1$ conv) per bottleneck residual block in MobileNetV2. The accuracy increases as more dynamic convolution layers are used. Using dynamic convolution for all three layers yields the best accuracy. If only one layer is allowed to use dynamic convolution, using it for the last $1 \times 1$ convolution yields the best performance.

---

**Table 4.** Inspecting DY-CNN by enabling/disabling attention at different input resolutions. DY-MobileNetV2 $\times 0.5$ is used. Each resolution has two options: $\checkmark$ indicates enabling attention $\sum \pi_k(x) \tilde{W}_k$ for each layer in that resolution, while $\checkmark$ indicates disabling attention and using average kernel $\sum \tilde{W}_k / K$ for each layer in the corresponding resolutions. The attention is more effective at higher layers with lower resolutions.

<table>
<thead>
<tr>
<th>Input Resolution</th>
<th>Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$112^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$56^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$28^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$14^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$7^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$- - - - - \checkmark$</td>
<td>57.3</td>
<td>79.9</td>
</tr>
<tr>
<td>$- - - \checkmark \checkmark$</td>
<td>67.0</td>
<td>87.2</td>
</tr>
<tr>
<td>$- \checkmark \checkmark \checkmark$</td>
<td>67.5</td>
<td>87.4</td>
</tr>
<tr>
<td>$\checkmark \checkmark \checkmark \checkmark$</td>
<td>69.1</td>
<td>88.4</td>
</tr>
<tr>
<td>$\checkmark \checkmark \checkmark - -$</td>
<td>69.4</td>
<td>88.6</td>
</tr>
<tr>
<td>$\checkmark \checkmark - - -$</td>
<td>50.9</td>
<td>76.2</td>
</tr>
<tr>
<td>$\checkmark - - - -$</td>
<td>42.5</td>
<td>68.4</td>
</tr>
<tr>
<td>$\checkmark - - -$</td>
<td>41.2</td>
<td>67.0</td>
</tr>
<tr>
<td>$- - - - -$</td>
<td>37.9</td>
<td>63.5</td>
</tr>
<tr>
<td>$- - - -$</td>
<td>36.0</td>
<td>61.5</td>
</tr>
</tbody>
</table>

---

**Table 5.** Dynamic convolution at different layers in MobileNetV2 $\times 0.5$. C1, C2 and C3 indicate the $1 \times 1$ convolution that expands output channels, the $3 \times 3$ depthwise convolution and the $1 \times 1$ convolution that shrinks output channels per block respectively. C1=1 indicates using static convolution, while C1=4 indicates using dynamic convolution with 4 kernels. The numbers in brackets denote the improvement over the baseline.

<table>
<thead>
<tr>
<th>Network</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNetV2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>65.4</td>
<td>86.4</td>
</tr>
<tr>
<td>DY-MobileNetV2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>67.4 (2.0)</td>
<td>87.5 (1.1)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>67.4 (2.0)</td>
<td>87.3 (0.9)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>68.2 (2.8)</td>
<td>87.9 (1.5)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>68.7 (3.3)</td>
<td>88.0 (1.6)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>68.4 (3.0)</td>
<td>87.9 (1.5)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>68.6 (3.2)</td>
<td>88.0 (1.6)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>69.4 (4.0)</td>
<td>88.6 (2.2)</td>
</tr>
</tbody>
</table>
Table 6. Softmax Temperature: large temperature in early training epochs is important. Temperature annealing refers to reducing \( \tau \) from 30 to 1 linearly in the first 10 epochs. The numbers in brackets denote the performance improvement over the baseline.

|  \begin{array}{c|c|c|c} 
  Network & Temperature & Top-1 & Top-5 \\
  \hline 
  MobileNetV2 & \tau = \infty & 86.4 & 86.4 \\
  \tau = 1 & 64.7\% \text{(-0.6)} & 85.5\% \text{(-0.9)} & \\
  \tau = 5 & 65.7\% \text{(-0.3)} & 85.8\% \text{(-0.6)} & \\
  \tau = 10 & 67.4\% \text{(-2.1)} & 87.4\% \text{(+1.0)} & \\
  \tau = 20 & 69.4\% \text{(+4.0)} & 86.5\% \text{(+2.1)} & \\
  \tau = 30 & 69.4\% \text{(+4.0)} & 88.6\% \text{(+2.2)} & \\
  \tau = 40 & 69.4\% \text{(+4.0)} & 88.4\% \text{(+2.0)} & \\
  \tau annealing & 69.5\% \text{(-4.5)} & 89.0\% \text{(+2.6)} & \\
  \hline 
  DY-MobileNetV2 \\
  \tau = \infty & 86.4 & 86.4 \\
  \tau = 1 & 64.7\% \text{(-0.6)} & 85.5\% \text{(-0.9)} & \\
  \tau = 5 & 65.7\% \text{(-0.3)} & 85.8\% \text{(-0.6)} & \\
  \tau = 10 & 67.4\% \text{(-2.1)} & 87.4\% \text{(+1.0)} & \\
  \tau = 20 & 69.4\% \text{(+4.0)} & 86.5\% \text{(+2.1)} & \\
  \tau = 30 & 69.4\% \text{(+4.0)} & 88.6\% \text{(+2.2)} & \\
  \tau = 40 & 69.4\% \text{(+4.0)} & 88.4\% \text{(+2.0)} & \\
  \tau annealing & 69.5\% \text{(-4.5)} & 89.0\% \text{(+2.6)} & \\
 \end{array} |

Table 7. Dynamic convolution vs Squeeze-and-Excitation (SE) [12] on MobileNetV3-Small. The numbers in brackets denote the performance improvement over the baseline. Compared with static convolution with SE, dynamic convolution without SE gains 2.2% top-1 accuracy.

|  \begin{array}{c|c|c} 
  Network & Top-1 & Top-5 \\
  \hline 
  MobileNetV3-Small & 67.4 & 86.4 \\
  MobileNetV3-Small w/o SE & 65.4\% \text{(-2.0)} & 85.2\% \text{(-1.2)} & \\
  DY-MobileNetV3-Small & 70.3\% \text{(-2.9)} & 88.7\% \text{(+2.3)} & \\
  DY-MobileNetV3-Small w/o SE & 69.0\% \text{(-2.2)} & 88.4\% \text{(+2.0)} & \\
 \end{array} |

Table 8. ImageNet [4] classification results of DY-CNNs. The numbers in brackets denote the performance improvement over the baseline.

|  \begin{array}{c|c|c|c|c|c} 
  Network & \#Param & MAdds & Top-1 & Top-5 \\
  \hline 
  MobileNetV2 \times 1.0 & 3.5M & 308.0M & 72.0 & 91.0 \\
  DY-MobileNetV2 \times 1.0 & 11.1M & 312.9M & 75.2\% \text{(3.2)} & 92.1\% \text{(1.1)} & \\
  MobileNetV2 \times 0.75 & 2.6M & 209.0M & 69.8 & 89.6 \\
  DY-MobileNetV2 \times 0.75 & 7.0M & 217.5M & 73.7\% \text{(3.9)} & 91.3\% \text{(1.7)} & \\
  MobileNetV2 \times 0.5 & 2.0M & 97.0M & 65.4 & 86.4 \\
  DY-MobileNetV2 \times 0.5 & 4.0M & 101.4M & 69.9\% \text{(4.0)} & 89.0\% \text{(2.6)} & \\
  MobileNetV2 \times 0.35 & 1.7M & 59.2M & 60.3 & 82.9 \\
  DY-MobileNetV2 \times 0.35 & 2.8M & 62.0M & 65.9\% \text{(6.6)} & 86.4\% \text{(4.5)} & \\
  MobileNetV3-Small & 2.9M & 66.0M & 67.4 & 86.4 \\
  DY-MobileNetV3-Small & 4.8M & 68.5M & 70.3\% \text{(2.9)} & 88.7\% \text{(2.3)} & \\
  ResNet-18 & 11.1M & 1.81G & 70.4 & 89.7 \\
  DY-ResNet-18 & 12.7M & 1.85G & 72.7\% \text{(2.3)} & 90.7\% \text{(1.0)} & \\
  ResNet-10 & 5.2M & 0.89G & 63.5 & 85.9 \\
  DY-ResNet-10 & 16.6M & 0.91G & 67.5\% \text{(4.2)} & 87.6\% \text{(2.6)} & \\
 \end{array} |

Table 9. Light-weight head structures for keypoint detection. We use MobileNetV2’s bottleneck residual block [25] (denoted as bneck). Each row is corresponding to a stage, which starts with a bilinear upsampling operator to scale up the feature map by 2. \#out denotes the number of output channels, and \( n \) denotes the number of bottleneck residual blocks.

**5.4. Main Results**

Table 8 shows the comparison between dynamic convolution and its static counterpart in three CNN architectures (MobileNetV2, MobileNetV3 and ResNet). \( K = 4 \) kernels are used in each dynamic convolution layer and temperature annealing is used in the training. Although we focus on efficient CNNs, we evaluate dynamic convolution on two shallow ResNets (ResNet-10 and ResNet-18) to show its effectiveness on \( 3 \times 3 \) convolution, which is only used for the first layer in MobileNet V2 and V3. Without bells and whistles, dynamic convolution outperforms its static counterpart by a clear margin for all three architectures, with small extra computational cost (~ 4%). DY-ResNet and DY-MobileNetV2 gains more than 2.3% and 3.2% top-1 accuracy, respectively. DY-MobileNetV3-Small is 2.9% more accurate than the state-of-the-art MobileNetV3-Small.

6. **DY-CNNs for Human Pose Estimation**

We use COCO 2017 dataset [18] to evaluate dynamic convolution on single-person keypoint detection. Our models are trained on train2017, including 57K images and 150K person instances labeled with 17 key-points. We evaluate our method on val2017 containing 5000 images and use the mean average precision (AP) over 10 object key point similarity (OKS) thresholds as the metric.

**Implementation Details:** We implement two types of networks to evaluate dynamic convolution. Type-A follows SimpleBaseline [33] by using deconvolution in head. We use MobileNetV2 and V3 as a drop-in replacement for the backbone feature extractor and compare static convolution and dynamic convolution in the backbone alone. Type-B still uses MobileNetV2 and V3 as backbone. But it uses upsampling and MobileNetV2’s bottleneck residual block in head. We compare dynamic convolution with its static counterpart in both backbone and head. The details of head...
Table 10. Keypoint detection results on COCO validation set. All models are trained from scratch. The top half uses dynamic convolution in the backbone and uses deconvolution in the head (Type A). The bottom half use MobileNetV2’s bottleneck residual blocks in the head and use dynamic convolution in both the backbone and the head (Type B). Each dynamic convolution layer includes $K = 4$ kernels. The numbers in brackets denote the performance improvement over the baseline.

Table 11. Keypoint detection results of using dynamic convolution in backbone and head separately. We use MobileNetV2 $\times 0.5$ as backbone and use the light-weight head structure discussed in Table 9. The numbers in brackets denote the performance improvement over the baseline. Dynamic convolution can improve AP at both the backbone and the head.

7. Conclusion

In this paper, we introduce dynamic convolution, which aggregates multiple convolution kernels dynamically based upon their attentions for each input. Compared to its static counterpart (single convolution kernel per layer), it significantly improves the representation capability with negligible extra computation cost, thus is more friendly to efficient CNNs. Our dynamic convolution can be easily integrated into existing CNN architectures. By simply replacing each convolution kernel in MobileNet (V2 and V3) with dynamic convolution, we achieve solid improvement for both image classification and human pose estimation. We hope dynamic convolution becomes a useful component for efficient network architectures.
References


11038


