Multi-Dimensional Pruning: A Unified Framework for Model Compression

Jinyang Guo   Wanli Ouyang   Dong Xu
School of Electrical and Information Engineering, The University of Sydney
{jinyang.guo,wanli.ouyang,dong.xu}@sydney.edu.au

Abstract

In this work, we propose a unified model compression framework called Multi-Dimensional Pruning (MDP) to simultaneously compress the convolutional neural networks (CNNs) on multiple dimensions. In contrast to the existing model compression methods that only aim to reduce the redundancy along either the spatial/spatial-temporal dimension (e.g., spatial dimension for 2D CNNs, spatial and temporal dimensions for 3D CNNs) or the channel dimension, our newly proposed approach can simultaneously reduce the spatial/spatial-temporal and the channel redundancies for CNNs. Specifically, in order to reduce the redundancy along the spatial/spatial-temporal dimension, we downsample the input tensor of a convolutional layer, in which the scaling factor for the downsampling operation is adaptively selected by our approach. After the convolution operation, the output tensor is upsampled to the original size to ensure the unchanged input size for the subsequent CNN layers. To reduce the channel-wise redundancy, we introduce a gate for each channel of the output tensor as its importance score, in which the gate value is automatically learned. The channels with small importance scores will be removed after the model compression process. Our comprehensive experiments on four benchmark datasets demonstrate that our MDP framework outperforms the existing methods when pruning both 2D CNNs and 3D CNNs.

1. Introduction

With the popularity of Convolutional Neural Networks (CNNs) for various computer vision applications, several model compression technologies were developed (see Sec. 2 for more details) to deploy CNNs on resource constrained platforms. Among these techniques, the channel pruning methods [8, 23, 38] aim to reduce the redundancy along the channel dimension of CNNs. However, substantial redundancy also exists on the spatial/spatial-temporal dimension of CNNs (i.e., spatial dimension for 2D CNNs, spatial and temporal dimensions for 3D CNNs), which is not considered by the existing channel pruning methods.

Figure 1(a) shows the output feature maps of four consecutive frames generated from the output tensor of the first convolutional layer (the C3D model [31] is used for illustration). The hand position of the person is mostly identical in these frames. (b) The downsampled tensor, in which the information is almost the same as the original output tensor.
nels considered as unimportant in the first pruning stage may become important after reducing the STR of CNNs. However, we cannot recover these channels in the second pruning stage because they have already been removed in the first stage, which degrades the performance of the compressed model. Different channels in CNNs often pay attention to different parts of feature maps. Before reducing the STR, the channels with detailed information may be more important than those with overall information as the details in the high resolution feature maps can provide rich information to CNNs. Therefore, the channels with overall information will be removed. After reducing the STR, the resolution of the feature maps will be reduced and the details in these feature maps will be lost. In this case, the channels with less details might become more informative when compared with the high resolution case. As a result, we should keep the channels with overall information.

To address the aforementioned issue, we propose a new unified framework called Multi-Dimensional Pruning (MDP) to simultaneously reduce the spatial/spatial-temporal and the channel redundancies in CNNs in an end-to-end fashion. Specifically, our MDP framework consists of three stages: the searching stage, the pruning stage, and the fine-tuning stage. In the searching stage, we construct an over-parameterized network by expanding each convolutional layer to multiple parallel branches, in which different branches correspond to information processing at different spatial/spatial-temporal resolutions. The information from different branches will be aggregated based on their importance scores. We also introduce a gate for each channel to indicate its importance. The importance scores of the branches and the channels are automatically learned in the searching stage. In the pruning stage, we prune the branches and channels based on their importance scores. We finally fine-tune the compressed model to recover from the accuracy drop.

To the best of our knowledge, this is the first unified model compression framework that can simultaneously reduce the spatial and channel redundancy for 2D CNNs, and the spatial, temporal, and channel redundancy for 3D CNNs. When compared with the existing channel pruning methods [21, 38] or other methods that aim at reducing the STR [4, 33, 39], our MDP framework has several advantages: (1) The MDP framework is a unified model compression framework, which is suitable for both 2D CNNs and 3D CNNs. (2) The optimal combination of selected features from multiple dimensions (i.e., the selected branches for the spatial/spatial-temporal dimension, and the selected channels for the channel dimension) can be automatically learned in the searching stage, which can solve the sub-optimal problem in the alternative approach that sequentially uses different model compression algorithms in a step-by-step fashion.

The experiments on four benchmark datasets demonstrate the effectiveness of our proposed MDP framework for both image classification and video classification tasks.

2. Related Work

Channel pruning. Channel pruning technologies [23, 8, 38, 35, 24, 7, 22, 37] aim to reduce the channel-wise redundancy in CNNs. In [18], Lin et al. proposed to use adversarial learning to prune the redundant structures in CNNs. Guo et al. [3] pruned the channels by using the guidance from the classification loss and feature importance. However, the existing channel pruning methods ignore the STR in CNNs. Unlike these channel pruning methods, our MDP framework can additionally reduce the redundancy along the spatial/spatial-temporal dimension.

Spatial/spatial-temporal redundancy reduction. Recently, many methods [4, 12] were proposed to reduce the spatial redundancy when designing the architectures of CNNs. For example, Chen et al. [4] proposed to replace the vanilla convolution by the octave convolution. On the other hand, several network architectures [33, 39, 17] were proposed to reduce the temporal redundancy in 3D CNNs. TSN [33] uses the sparse temporal sampling strategy to reduce the computational cost for long-term temporal structure. ECO [39] mixes 2D and 3D networks to save computation. The goals of these methods [4, 12, 33, 39, 17] are to design new type of convolution operations or network architectures instead of compressing a given network. In contrast, we aim to compress a given model by jointly pruning the network along the spatial and channel dimensions for 2D CNNs, and along the spatial, temporal, and channel dimensions for 3D CNNs. Due to the newly designed convolution operations or network structures, these methods are not suitable for model compression as they need to train the models from scratch. Therefore, these approaches cannot transfer the information from the pre-trained model to the compressed model, which will degrade the performance of the compressed models. Moreover, these methods only focus on how to reduce the redundancy along one dimension (spatial or temporal). In contrast, our approach can jointly reduce the redundancies along multiple dimensions, which thus achieves better performance.

Multi-scale representation learning. Multi-scale representation learning [10, 30, 19] has been demonstrated to be effective for many computer vision tasks. For example, Elastic-Net [32] introduces the elastic module in CNNs to extract multi-scale feature representations. In [16], the features from multiple scales are concatenated to obtain the information from different scales. The goal of these multi-scale representation learning approaches is to capture the information of features at different scales, but our MDP framework aims to compress a given CNN to obtain a more efficient network.
Network architecture search. While our MDP method is also related to the network architecture search methods [20, 2], we aim to reduce the redundancies in CNNs by pruning a given model along different dimensions instead of searching the optimal network architecture as in [20, 2].

3. Multi-Dimensional Pruning

In this section, we take the process of compressing 3D CNNs as an example to introduce our MDP framework, which is a more general case. The algorithm for compressing 2D CNNs can be readily obtained.

3.1. Overview

Our MDP framework consists of three stages: the searching stage, the pruning stage, and the fine-tuning stage. In the searching stage, we firstly construct an over-parameterized network from any given original network to be pruned and then train this over-parameterized network by using the objective function introduced in Sec. 3.2.2. In the pruning stage, we prune the unimportant branches and channels in this over-parameterized network based on the importance scores learned in the searching stage. In the fine-tuning stage, we fine-tune the pruned network to recover from the accuracy drop.

3.2. The searching stage

3.2.1 Overview of the over-parameterized network

In the searching stage, we firstly construct an over-parameterized network from any given original network. The compressed model can be obtained by pruning this over-parameterized network. Figure 2 shows one convolutional layer (left) in the original network and its corresponding layer (right) in the over-parameterized network. We expand each convolutional layer into several parallel branches in the corresponding layer of the over-parameterized network. Since we only focus on one convolutional layer in most places in this section, we omit the index of this layer when introducing each operation except in Eq. (3), where the layer index is denoted as the superscript \( l \) for the l-th layer.

In each branch, we perform the following operations: (1) We downsample the input tensor of the convolutional layer by applying average pooling along the spatial and/or temporal dimension. In our implementation, we downsample the input tensor to 4 different resolutions along the spatial dimension with four scaling factors 1, 2, 3, and 4, where the scaling factor of 1 means there is no downsampling operation (i.e., identity mapping). We also downsample the input tensor with four different scaling factors (1, 2, 3, and 4) along the temporal dimension. Therefore, the total number of branches in each layer of the over-parameterized network is \( 4 \times 4 = 16 \) after performing the downsampling operation along the spatial and temporal dimensions. In Figure 2, we have three channels of the input tensor of this layer \( X \) (the blue part) and each channel is represented as one blue block. There are 4 frames along the temporal dimension in each channel, which is represented as the depth of each blue block. For \( \text{branch}_2 \) in Figure 2, we downsample \( X \) by...
using the scaling factor of 2 and 1 along the spatial dimension and the temporal dimension, respectively. We obtain the downsampled tensor $I^2$ in branch $2$, which is marked as green. In this case, the height/width of $I^2$ becomes half of the height/width of $I$ while the number of frames is 4 in both $I$ and $I^2$. Similar to the downsampling operation in branch $2$, we downsample $I$ by using the scaling factor of 4 and 2 along the spatial dimension and the temporal dimension in branch $4$, respectively. The downsampled input tensor of this branch $4$ is marked as yellow. In this case, the height/width of $I^4$ become a quarter of the height/width of $I$ and 4 frames in $I$ are downsampled as 2 frames.

The downsampled tensor of each branch is fed into the convolutional layer. The parameters of the convolutional layers in all branches of the over-parameterized network are copied from the original network. After the convolutional layer, we multiply each channel by a gate, which is e.g., $\Lambda_i$.

Suppose we have $B$ branches in the over-parameterized network at each layer, let us denote the input tensor at this layer as $X \in \mathbb{R}^{c_{in} \times d_{in} \times h_{in} \times w_{in}}$ where $c_{in}$ is the number of input channels for this layer, $d_{in}$ is the length of the input tensor along the temporal dimension, and $h_{in}$ and $w_{in}$ are the height and width of the input tensor, respectively. Similarly, the output tensor at this layer can be denoted as $Y \in \mathbb{R}^{c_{out} \times d_{out} \times h_{out} \times w_{out}}$, where $c_{out}$ is the number of output channels for this layer, $d_{out}$ is the length of the output tensor along the temporal dimension, and $h_{out}$ and $w_{out}$ are the height and the width of the output tensor, respectively. The convolutional layer connects the input tensor $X$ and the output tensor $Y$ by a given transformation (e.g., a convolution operation) $T$, where $Y = T(X)$.

Suppose we have $B$ branches for this layer in the over-parameterized network, the output tensor of this layer in the over-parameterized network can be written as follows:

$$Y = \frac{B}{\sum_{i=1}^{B} \mathcal{S}(\lambda_i) \cdot \mathcal{U}_i \{T_i[A_i(X)]\}},$$

where $A_i$, $T_i$, and $\mathcal{U}_i$ are the average pooling operation, the transformation function, and the upsampling operation in the $i$-th branch at this layer, respectively. $\lambda_i$ is the learnable parameter in the $i$-th branch and $\mathcal{S}(\cdot)$ is the softmax function, namely, $\mathcal{S}(\lambda_i) = \frac{\exp(\lambda_i)}{\sum_{k=1}^{B} \exp(\lambda_k)}$. Note $S(\lambda_i)$ represents the importance score of the $i$-th branch at this layer.

Denote $I^i$ as the pooled input tensor of the $i$-th branch at this layer after using the average pooling operation, i.e., $I^i = A_i(X)$. Denote $I_{j;:::;}^i$ as the $j$-th channel of $I^i$. In the $i$-th branch, the output after performing the transformation operation $T_i$ can be written as follows:

$$O_{k;:::;}^i = g_{i,k} \sum_{j=1}^{c_{in}} I_{j;:::;}^i \ast W_{k,j;:::;}^i$$

where $O^i$ is the output tensor before the upsampling operation in the $i$-th branch and $O_{j;:::;}^i$ is the $k$-th channel of $O^i$. $W^i \in \mathbb{R}^{c_{out} \times c_{in} \times d_{in} \times h_{in} \times w_{in}}$ is the weight tensor for the convolution operation where the subscript $kr$ denotes the kernel/filter. $g_{i,k}$ is the gate value for the $k$-th channel in the $i$-th branch. $\ast$ is the convolution operation. We omit the bias term and the activation function in Eq. (2) for better presentation.

**Objective function.** After independently constructing the over-parameterized network at each layer, we train the over-parameterized network by using the following objective function, which is inspired by [21, 2]. For better presentation, we additionally introduce the superscript $(l)$ to denote the corresponding symbols of the $l$-th layer in the following objective function.

$$\arg \min_{\Theta, \lambda_G} \mathcal{L} = \mathcal{L}_c + \alpha \mathcal{L}_{st} + \eta \mathcal{L}_{gate},$$

where

$$\mathcal{L}_{st} = \sum_{l=1}^{L} \sum_{i=1}^{B(l)} \mathcal{S}(\lambda_i^{(l)}) \cdot \mathcal{F}(T_i^{(l)}),$$

$$\mathcal{L}_{gate} = \sum_{l=1}^{L} \sum_{i=1}^{B(l)} \sum_{k=1}^{c_{out}} \|g_{i,k}^{(l)}\|_1.$$
where \( \lambda_i^{(l)} \) is the learnable parameter to obtain the importance score for the \( i \)-th branch at the \( l \)-th layer. \( G \) is the set containing the gate values for all channels from all layers, \( G = \{ g_{i,1}^{(1)}, g_{i,2}^{(1)}, \ldots, g_{i,k}^{(1)}, \ldots \} \) where \( g_{i,k}^{(1)} \) is the gate value for the \( k \)-th channel of the \( i \)-th branch at the \( l \)-th layer. \( F(T_i^{(1)}) \) denotes the number of floating point operations (FLOPs) of the transformation \( T_i^{(1)} \) at the \( l \)-th layer, which is widely used for computational complexity measurement. \( \alpha \) and \( \eta \) are two coefficients to balance different losses. \( \| \cdot \|_1 \) denotes the \( l_1 \) norm. We can control the compression ratio by adjusting the values of \( \alpha \) and \( \eta \). Specifically, higher values of \( \alpha \) and \( \eta \) will result in higher compression ratio.

### 3.3. The pruning stage

We perform the pruning process after the searching stage is finished. At each layer, we only select the branch with the largest importance score and prune other branches, and we also prune the channels with small gate values in the selected branch. In Figure 3, we show the pruning process for one convolutional layer. Here, we ignore the superscript \( .^{(l)} \) again for better presentation. Suppose the \( i \)-th branch in Figure 2 has the largest importance score \( S(\lambda_i) \). In this case, in the pruning stage, we keep the \( i \)-th branch and remove other branches at this layer. At the same time, let us assume the second channel of this branch has the largest importance score \( g_{i,2} \). Therefore, we preserve the second channel and remove the first channel. As shown in Figure 3, we finally arrive at the pruned network for this layer after the pruning stage.

### 3.4. The fine-tuning stage

We perform the fine-tuning process on the pruned network to recover from the accuracy drop. After the fine-tuning stage, the compressed model is obtained.

### 3.5. Compressing 2D CNNs

Compressing 2D CNNs by using our MDP framework is a special case of compressing 3D CNNs. For 2D CNNs, we only apply the average pooling operation for downsampling along the spatial dimension. Similar to the process of compressing 3D CNNs, we downsample the input tensor by using the scaling factors of 1, 2, 3, 4. In this case, we have 4 branches for each layer in the over-parameterized network. At the same time, we only upsample the output tensor of each branch along the spatial dimension. After constructing the over-parameterized network, the following stages are the same as those for pruning 3D CNNs.

### 3.6. Comparison with other methods

Our work is related to multi-scale representation learning approaches [16, 32] and channel pruning methods [21, 18]. However, our work is different from these methods in both motivation and formulation. The multi-scale representation learning approaches aim to capture the multi-scale information to improve the accuracy of CNNs. Therefore, multiple branches are selected and preserved in the final model in [16, 32]. In contrast to these two methods, our work aims to compress CNNs and only one branch is preserved in the compressed model. Although the channel pruning methods in [21, 18] also prune the channels based on the learnable importance scores, the spatial-temporal redundancy is not explored in their works [21, 18].

### 4. Experiments

In order to demonstrate the effectiveness our MDP framework, we compare our MDP approach with several state-of-the-art model compression methods, including ThiNet [25], Channel Pruning (CP) [8], Slimming [21], Width-multiplier (WM) [9], DCP [38], GAL [18], Filter Pruning (FP) [24], Taylor Pruning (TP) [24], and Regularization-based pruning (RBP) [36] on four benchmark datasets: CIFAR-10 [13], ImageNet [26], UCF-101 [29], and HMDB51 [14].

The percentage of the number of floating point operations (\#FLOPs(\%)) in this section refers to the ratio of the FLOPs from the pruned network over that from the original network, which is a commonly used criterion for computational complexity measurement.

**Datasets.** CIFAR-10 and ImageNet are used to evaluate the effectiveness of our proposed method when pruning 2D CNNs for the image classification task. CIFAR-10 consists of 50k training images and 10k testing images from 10 classes. ImageNet is a large dataset, which contains over 1 million training images and 50k testing images from 1000 categories. On the other hand, UCF-101 and HMDB51 are used to evaluate the performance of our MDP method when pruning 3D CNNs for the video classification task. Specifically, UCF-101 consists of 13,320 videos from 101 classes, while HMDB51 contains 6,766 videos from 51 classes.

**Implementation details.** Based on the original network, we apply our MDP approach to compress the model along
multiple dimensions. We adjust the #FLOPs by choosing different values of $\alpha$ and $\eta$. For image classification, we use the SGD optimizer with nesterov for optimization at the searching stage. On CIFAR-10, the initial learning rate, the batch size, and the momentum are set to 0.1, 256, and 0.9, respectively. The settings on ImageNet are the same as CIFAR-10 except that the initial learning rate is set to 0.9, respectively. Similar results are also reported in the DCP work [38]. One possible explanation is that the overfitting problem on small-scale datasets like CIFAR-10 can be partially solved by compressing the models.

### 4.2. Results on ImageNet

In order to compare our proposed approach with other state-of-the-art methods on large-scale datasets, we compress ResNet-50 [6] and MobileNet-V2 [27] on ImageNet. We follow the setting in [8] to compare the Top-5 accuracy with other methods and the results are shown in Table 2.

To investigate the advantages by simultaneously reducing the redundancies along multiple dimensions, we also report the results from an alternative approach by using the step-by-step pruning strategy, which prunes the channels by using the DCP method in the first step and then reduces the spatial redundancy by using our MDP approach at the second step. The result is referred as DCP+SP in Table 2.

From Table 2, we have the following observations: (1) For ResNet-50, our MDP method outperforms other existing approaches, which indicates that it is beneficial to compress the model by using our MDP approach. (2) For MobileNet-V2, our proposed method surpasses other state-of-the-art approaches by more than 2.4% when #FLOPs are comparable, which is a significant improvement on the ImageNet dataset. (3) When comparing the DCP+SP approach with the DCP method, the DCP+SP approach performs better, which indicates that it is beneficial to additionally reduce the redundancy along the spatial dimension. (4) Our MDP framework outperforms the DCP+SP approach by 0.2%, which demonstrates the effectiveness of our MDP method to jointly prune models along multiple dimensions.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>94.57</td>
<td>94.02</td>
<td>-</td>
<td>93.30</td>
<td>93.24</td>
<td>93.49</td>
<td>94.29</td>
<td></td>
</tr>
<tr>
<td>MobileNet-V2</td>
<td>71.29</td>
<td>73.53</td>
<td>73.53</td>
<td>71.29</td>
<td>73.53</td>
<td>73.53</td>
<td>73.53</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparison of Top-1 accuracies from different model compression methods for compressing VGGNet, ResNet-56, and MobileNet-V2 on the CIFAR-10 dataset. When using the CP method [8], we directly quote the results in the original work [8] for compressing ResNet-56. The other results are copied from the work in [38].

<table>
<thead>
<tr>
<th>Model</th>
<th>Top-5 Acc. (%)</th>
<th>Top-5 Acc. (%)</th>
<th>#FLOPs (%)</th>
<th>#FLOPs (%)</th>
<th>#FLOPs (%)</th>
<th>#FLOPs (%)</th>
<th>#FLOPs (%)</th>
<th>MDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>92.66</td>
<td>90.82</td>
<td>44.44</td>
<td>44.44</td>
<td>44.44</td>
<td>45.06</td>
<td>44.29</td>
<td></td>
</tr>
<tr>
<td>MobileNet-V2</td>
<td>88.86</td>
<td>86.84</td>
<td>55.25</td>
<td>55.25</td>
<td>55.25</td>
<td>56.55</td>
<td>56.85</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Comparison of Top-5 accuracies from different model compression methods for compressing ResNet-50 and MobileNet-V2 on ImageNet. For the state-of-the-art works, we directly quote the results from [38].
We conduct more experiments to investigate the effectiveness of our MDP method for pruning 3D CNNs. **Effect of channel pruning.** To investigate the effectiveness for pruning the channels in our MDP method, we perform the experiment to only reduce the spatial redundancy without pruning the channels, which is referred to as MDP w/o CP in Table 5. From Table 5, our MDP approach outperforms MDP w/o CP method by 1.43%, which indicates that it is effective to prune the channels in our MDP approach. **Effect of spatial pruning.** To investigate the effectiveness for reducing the spatial redundancy in our MDP framework, we perform the experiment to only prune the channels without removing the spatial redundancy, which is referred to as MDP w/o SP in Table 5. From Table 5, our MDP method surpasses the MDP w/o SP approach by 0.9%, which indicates that it is effective to reduce the redundancy along the spatial dimension in our MDP approach.

The experimental results in Table 5 show that either the channel pruning or the spatial pruning alone does not perform better than our MDP method, which jointly performs both of them. Therefore, it is beneficial to jointly prune the models along multiple dimensions by using our MDP method.

**Effect of $\alpha$ and $\eta$.** We conduct more experiments to investigate the performance when choosing different values of $\alpha$ and $\eta$. For fair comparison, we adjust $\alpha$ and $\eta$ in Eq. (3) to obtain the compressed models with similar #FLOPs and compare their performance. The results are shown in Table 6. From Table 6, we have several observations: (1) The models with similar #FLOPs can be obtained by increasing one of the coefficients and decreasing the other one. (2)
Table 6: Accuracies of our MDP method by using different values of $\alpha$ and $\eta$ when pruning ResNet-56 on CIFAR-10.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\eta$</th>
<th>#FLOPs (%)</th>
<th>Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2e^{-15}$</td>
<td>$1e^{-10}$</td>
<td>43.44</td>
<td>94.18</td>
</tr>
<tr>
<td>$2e^{-14}$</td>
<td>$1e^{-11}$</td>
<td>43.53</td>
<td>94.26</td>
</tr>
<tr>
<td>$2e^{-13}$</td>
<td>$1e^{-12}$</td>
<td>45.11</td>
<td>94.29</td>
</tr>
<tr>
<td>$2e^{-12}$</td>
<td>$1e^{-13}$</td>
<td>45.47</td>
<td>94.23</td>
</tr>
<tr>
<td>$2e^{-11}$</td>
<td>$1e^{-14}$</td>
<td>45.03</td>
<td>94.22</td>
</tr>
</tbody>
</table>

With similar #FLOPs, the performance of the compressed model obtained by using different values of $\alpha$ and $\eta$ are comparable, which shows the performance of our MDP approach is not sensitive to the coefficients $\alpha$ and $\eta$.

### 4.5.2 Ablation study for 3D CNNs

To investigate the effectiveness of different components in our MDP method for compressing 3D CNNs, we perform the experiments to compress C3D on UCF-101 and report the video-level accuracies. In 3D CNNs, the spatial-temporal redundancy exists along both the spatial and the temporal dimensions.

We firstly perform the experiment to only reduce the spatial-temporal redundancy without pruning the channels in the C3D model, which is referred to as MDP w/o CP in Table 7. We also perform the experiment to only prune the channels without reducing the spatial-temporal redundancy in the C3D model, which is referred to as MDP w/o STP. From Table 7, our MDP method outperforms the alternative approach MDP w/o CP, which indicates that it is effective to prune the channels for compressing 3D CNNs in our MDP method. Our MDP method also surpasses the alternative approach MDP w/o STP by 2.51%, which demonstrates the effectiveness of our MDP approach for reducing the spatial-temporal redundancy when compressing 3D CNNs.

### 4.6. Branch selection analysis

In Figure 4, we report the index of selected branch (i.e., the selected scaling factor for the downsampling operation) for each convolutional layer in the compressed ResNet-50 model on the ImageNet dataset. From Figure 4, it is interesting to see that the branches with higher downsampling scaling factor (e.g., the index of selected branch is 4) tend to appear in the shallower layers (close to the input of CNNs), while the branches with lower downsamping scaling factor (e.g., the index of selected branch is 1) tend to appear in the deeper layers (close to the output of CNNs). We hypothesize that the feature maps in the shallower layers have higher resolutions, which indicates more redundancy along the spatial dimension. On the other hand, the feature maps in the deeper layers have lower resolutions along the spatial dimension, which indicates there is less redundancy along this dimension. It is also worth mentioning that we do not reduce the resolution (i.e., the index of selected branch is 1) for most of the branches from layer 29 to layer 48, which suggests that the spatial redundancy in the deeper layers of the ResNet-50 model can be neglected possibly because of low spatial resolutions in these deep layers.

### 5. Conclusion

In this work, we have proposed a unified model compression framework called Multi-Dimensional Pruning (MDP) to compress 2D CNNs along the channel and spatial dimensions, and 3D CNNs along the channel, spatial, and temporal dimensions. In contrast to the existing model compression approaches that only reduce the redundancy along one certain dimension, our proposed framework can simultaneously reduce the redundancies along multiple dimensions and thus significantly accelerate CNNs when they are deployed on resource constrained platforms. Comprehensive experiments on four benchmark datasets demonstrate the effectiveness of our MDP method for pruning both 2D CNNs and 3D CNNs.

**Acknowledgement.** This work was supported by the Australian Research Council (ARC) Future Fellowship under Grant FT180100116. Wanli Ouyang was supported by the ARC Discovery Project under Grant DP200103223.
References


[35] Ruichi Yu, Ang Li, Chen-Fu Chen, Jui-Hsin Lai, Vlad I Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and


