Abstract

We propose a framework for compressing state-of-the-art Single Shot MultiBox Detector (SSD). The framework addresses compression in the following stages: Sparsity Induction, Filter Selection, and Filter Pruning. In the Sparsity Induction stage, the object detector model is sparsified via an improved global threshold. In Filter Selection & Pruning stage, we select and remove filters using sparsity statistics of filter weights in two consecutive convolutional layers. This results in the model with the size smaller than most existing compact architectures. We evaluate the performance of our framework with multiple datasets and compare over multiple methods. Experimental results show that our method achieves state-of-the-art compression of 6.7X and 4.9X on PASCAL VOC dataset on models SSD300 and SSD512 respectively. We further show that the method produces maximum compression of 26X with SSD512 on German Traffic Sign Detection Benchmark (GTSDB). Additionally, we also empirically show our method’s adaptability for classification based architecture VGG16 on datasets CIFAR and German Traffic Sign Recognition Benchmark (GTSRB) achieving a compression rate of 125X and 200X with the reduction in flops by 90.50% and 96.6% respectively with no loss of accuracy. In addition to this, our method does not require any special libraries or hardware support for the resulting compressed models.

1. Introduction

Object detection has been a major application that has seen tremendous progress with the advent of deep learning. There have been some techniques in object detection [11, 30] that have achieved state of the art accuracy. These advances enable object detection to be widely used for various applications ranging from pedestrian detection for autonomous driving to indoor object detection for robotics applications. One factor that however limits the wide-scale deployment of object detection has been that the models obtained through deep learning are rather large with millions of parameters. It is in this context that research in model compression for object detection is relevant. However, model compression [12, 17] has mostly been researched in the area of object classification. The task of model compression for object detection is more challenging than classification because for classification one relies on the whole image for obtaining the prediction while for detection one need to localize a specific image region that contains the object. Model compression might inaccurately affect specific image regions reducing the object detection performance. In this paper, we show that using a multi-layer pruning algorithm it is possible to obtain significant compression of the model while maintaining the accuracy in detection. The proposed method provides us with state of the art results for model compression for object detection.

In this paper, we propose a method for deep learning compression focusing on the Single Shot MultiBox Detector (SSD). However, it should be noted that the method is general and given a different loss function such as that of classification, the method is still applicable. Further by benchmarking on SSD object detection we can compare our method with other deep model compression techniques that have been proposed for object detection. Our method has the following steps, a sparsity constraint-based training step, a filter selection, and pruning step and finally a retraining step with the resulting selected filters. At the end of the filter selection and pruning step, we observe a drop in accuracy in object detection. However, the retraining step, in the end, ensures that we regain the accuracy in object detection.

Models resulting from our method do not rely on any special hardware/sparse library support. Furthermore, our method is complementary to all other pruning methods such as Binary/Quantized weights, etc.
Through this paper we make the following contributions:

1. We propose an algorithm for multi-layer pruning with sparsity constraint-based training, filter selection and pruning followed by a retraining step for obtaining a reduced set of weights that provide us with accurate object detection.

2. We investigate our methods over various configurations of SSD across major detection benchmarks such as the PASCAL VOC dataset and the German Traffic Sign Detection Benchmark (GTSDB). Our algorithm performs well for both these benchmark datasets.

3. We show that this framework is also applicable for object classification. In fact, for the classification task, we obtain a significant reduction of parameters of the order of 125X to 200X. These result in extremely compact models that can be used practically.

2. Related work

Deep neural networks are often over-parameterized with a large number of redundant parameters. Deep compression exploits these redundancies to reduce the overall model size. Multiple approaches are proposed which fall broadly into the following major categories.

**Pruning and parameter sharing:** Initial works that have proposed model compression for classification include optimal brain damage [25] and optimal brain surgeon [14]. Recently there have been some works including work on Deep Compression [12] that uses an iterative threshold based pruning and retraining. Other such methods that have proposed variants of pruning for classification include [36, 7, 20, 26]. They vary in the approach they adopt for pruning whether they use a hash function to group weights for pruning [7] or use sparsity constraint [24]. In our method, we adopt a single multi-layer pruning iteration instead of other techniques that advocate iterative layer pruning strategies. Further, an object detection the loss function objective includes a classification loss function and a localization objective function. This necessitates that any pruning adopted should not be biased by the classification objective alone. In classification pruning may not adversely affect the classification objective as the context present in an image may also help in satisfying the objective. This is not the case with detection. Our approach allows for careful pruning that satisfies both the objective functions.

**Compact Architectures:** Another set of ideas advocates use of compact architectures. One such approach is that of using transferred convolutional filters motivated from works by Cohen and Welling [8], where they use same weights to analyze each part of the image, thereby reducing the number of parameters in the overall model. This approach would be more suited for classes of object detection that have relatively less variance between instances. Objects like people that have high variance cannot be approached using such approaches. Other works [28, 35, 40] focus on methods where they build CNN layer’s using multiple base filters. The second set of works look more towards binarization for reducing memory. Binary net [21], BinaryConnect [9], XNOR-Net [34] focus on training CNN’s end to end using binary weights. Compact architectures usually exhibit a high drop in accuracy and are not suitable for object detection.

**Knowledge Distillation:** Knowledge distillation was proposed by Cuarna et al. [4, 2] and later explored by Hinton et al. [17]. Recent works in this area by Balan et al. [3] used a parametric student model for a Monte Carlo teacher, which was extended to Net2Net [6] that accelerated the experimentation process by instantaneously transferring the knowledge from a previous network to each new deeper or wider network. Recently, Zagoruyko et al. [39] proposed Attention Transfer (AT). They transferred the attention maps that are summaries of the full activations. While the methods described above makes the model thinner and reduce computational cost, it can be only applied to classification tasks with softmax loss function. Our approach is equally applicable for detection as well as classification tasks with the limited trade-off between accuracy and model size.

**Object detection:** While there is a large body of works on compression of classification based architectures, very few works are done till date on compression of object detection based deep learning models. Most notable are works by Xie et al. [37] which focuses on compressing SSD by using de-convolution to analyze kernels and active neurons, thereby removing inactive neurons and redundant kernels. Alternatively, Chen et al. [5] proposed a trainable framework for multiclass object detection through knowledge distillation via the introduction of weighed cross entropy and teacher bounded regression for knowledge distillation. Finally, in the very recent work Anisimov et al. [1], focus on eliminating channels by sampling using multiple methods and fine-tuning the resulting model on the given task. In conclusion, works that focus on SSD, concentrate either only a few layers of the SSD architecture or use compact base networks which result in reduced performance. In our work, we focus on compressing original SSD with VGG16 as its base network.

3. Multi-layer Pruning Framework

In this section, we give a comprehensive description of our pruning approach.

3.1. Framework Overview

The core of our pruning framework is based on the principle: “Evaluate the importance of the filters and remove...”
the unimportant ones” in line with previous works [31].

Let $C_l$ denote the input tensor and parameters of $l$-th convolutional layer. Here $C_l \in \mathbb{R}^{c_l \times h_l \times w_l}$ has $c_l$ channels, $h_l$ rows and $w_l$ columns. The weight tensor $W_l \in \mathbb{R}^{c_{l-1} \times c_l \times k_l \times k_l}$ is a set of $c_l$ filters of $c_{l-1} \times k_l \times k_l$ size each. This convolutional layer produces the output tensor $O_l \in \mathbb{R}^{c_{l+1} \times h_{l+1} \times w_{l+1}}$, which is a set of $c_{l+1}$ feature maps.

Our goal is to remove the unimportant filters in $W_l$. Let $c'_l \times c_{l-1} \times k_l \times k_l$ be the new size of $W_l$, where $c'_l$ is the number of remaining filters after pruning of the unimportant ones. Since the number of filters is modified in layer $l$, the size of output tensor $O_l$ and equivalently the size of input tensor of next layer $I_{l+1}$ also reduces from $c_l \times h_{l+1} \times w_{l+1}$ to $c'_l \times h_{l+1} \times w_{l+1}$. Hence, the corresponding weights in $W_{l+1}$ also need to be removed, which would in turn reduce the size of $W_{l+1}$ from $c_{l+1} \times c_l \times k_{l+1} \times k_{l+1}$ to $c'_l \times c_{l} \times k_{l+1} \times k_{l+1}$.

In our multilayer pruning framework, we propose architecture such as SSD in multiple phases. In each phase, we target a different set of consecutive layers. First, we move, which would in turn reduce the size of input tensor of next layer $I_{l+1}$ also reduced, which would in turn reduce the size of $W_{l+1}$ from $c_{l+1} \times c_{l} \times k_{l+1} \times k_{l+1}$ to $c_{l+1} \times c_{l} \times k_{l+1} \times k_{l+1}$.

**Figure 1:** Illustration of filter selection and pruning of convolutional filters in our pruning framework. After introducing sparsity in network weights by training with L1 regularization and thresholding, we examine the filter weights $W_l$ and $W_{l+1}$ of two consecutive convolutional layers $l$ and $l+1$ respectively. Based on sparsity in these weights, we select the filters in $W_l$ to be pruned. Note that corresponding weights in $W_{l+1}$ also get removed. Removing these weights reduces the number of output channels in $l$-th layer from $c_l$ to $c'_l$.

movable from the model. We repeat step 2 for each layer in $L$. A more detailed explanation is presented in Section 3.4.

3. **Pruning:** Filters selected in step 2 for layer $l$ and corresponding weights in layer $l+1$ associated with the output of these filters are removed from model $\Theta_{L,1}$. This is repeated for each layer in $L$ sequentially.

4. **Retraining:** Finally, we retrain the pruned network using the original loss (without L1 regularization) to recover the performance drop due to sparsity induction in step 1.

**3.2. Set-wise optimized solution**

Instead of finding a layer-wise optimized solution (layer-wise thresholding), we are finding the set-wise optimized solution (global thresholding for the whole layer-set) which results in high pruning with less performance drop because it preserves dependency across layers in the layer-set $L$.

**3.3. Sparsity Induction**

Let $D = \{(x_0, y_0), (x_1, y_2), \ldots, (x_n, y_n)\}$ be the training set, where $x_i$ and $y_i$ are input and target label. The parameters $\Theta$ of the original model are optimized to minimize
the cost $C_D(\Theta)$.

$$\Theta = \text{argmin}\{C_D(\Theta)\}$$

(1)

The form of this cost function depends on the task to be solved by the original network. For instance, we use multi-box loss function for SSD.

Let $L = \{l_{\text{start}}, \ldots, l_{\text{end}}\}$ be the set of consecutive layers to be pruned. In order to induce sparsity in parameters of layers in $L$, we add L1-norm of parameters of these layers to the original cost function and train the network initialised with $\Theta$.

$$\Theta_{L1} = \text{argmin} \left\{ C_D(\Theta) + \alpha \sum_{l=l_{\text{start}}}^{l_{\text{end}}} ||W_l||_1 \right\}$$

(2)

We choose the regularisation factor $\alpha$ such that the performance of the model with new parameters on a validation set $D_{\text{val}}$ is close to the original performance, i.e. $P_{D_{\text{val}}}(\Theta_{L1}) \geq P_{D_{\text{val}}}(\Theta) - \epsilon_1$. Here, $\epsilon_1$ is the tolerance constant that allows us to control the degree of sparsity in the parameters of layers in $L$. $P_{\text{D}_{\text{val}}}(\Theta)$ is the performance (e.g. accuracy, AP etc.) of model with parameters $\Theta$ on the validation set $D_{\text{val}}$.

After obtaining $\Theta_{L1}$, we set all the weights in $L$ with absolute values smaller than a threshold $t$ to zero. This gives us the parameters $\Theta_{L1}^{t\text{h}}$. We search for optimal $t$ in a range proportional to standard deviation of weights in $L$, such that $P_{D_{\text{val}}}(\Theta_{L1}^{t\text{h}}) \geq P_{D_{\text{val}}}(\Theta_{L1}) - \epsilon_2$. The constant $\epsilon_2$ provides us the additional control on the number of zero weights in $L$. At this point, our network parameters are ready for filter pruning in layers in $L$, which is described in next section.

### 3.4. Filter pruning

In this step, we determine filter importance corresponding to layer $l \in L$. This is executed by examining sparsity statistics of $W_l$ and $W_{l+1}$ in $\Theta_{L1}^{t\text{h}}$. Let $F_i \in \mathbb{R}^{c_l \times k_l \times k_l}$ be the $i$-th of $c_l$ filters in $W_l$. Note that in the next layer, the output feature map of filter $F_i$ is connected to a slice of tensor $W_{l+1}$ of size $c_{l+1} \times k_{l+1} \times k_{l+1}$, we call this slice of weights $G_i$. Essentially, our filter pruning strategy is to remove the $i$-th filter if either one of the tensors $F_i$ and $G_i$ has a very large fraction of zero weights (see Figure 1). We implement this strategy using three thresholds $s_F$, $s_F'$ and $s_G$ with $s_F > s_F' > s_G$. For each $i \in \{1, \ldots, c_l\}$, we evaluate following conditions and remove the $i$-th if atleast one of them holds true:

1. sparsity statistics in $F_i$ is higher than $s_F$
2. sparsity statistics in $F_i$ is higher than $s_F'$ and sparsity statistics in $G_i$ is higher than $s_G$.

Intuitively, the first condition selects the filters that have a very high level of sparsity ($> s_F$). These filters can be safely removed because their output activation maps are very weak. The second condition further selects those filters whose output activations are stronger than those selected by the first condition ($s_F > \text{sparsity level} > s_F'$) but have an overall low contribution because of high sparsity in the weights connected to them in the next layer (sparsity level $> s_G$). Our method for computing sparsity statistics is unlike previous works [20] where the method is data-driven and measures sparsity on layers activation (e.g. output of ReLU) rather than its filter weights. To compute the sparsity level, we use the concept of zero row. A row in a filter $k_l \times k_l$ is considered as a zero row (size $1 \times k_l$) if all elements in the row are zero. Sparsity level is simply the percentage of zero rows in a filter.

$$\text{Sparsity level}(F_i) = \frac{\text{Number of zero rows}}{c_{l-1} \times k_l}$$

$$\text{Sparsity level}(G_i) = \frac{\text{Number of zero rows}}{c_{l+1} \times k_{l+1}}$$

Let $\Theta_c$ be the parameters of the network after filter pruning. We empirically found that taking the values of parameters in $\Theta_c$ from $\Theta_{L1}$ rather than from $\Theta_{L1}^{t\text{h}}$ results in im-

---

**Algorithm 1 Algorithm for filter pruning**

**Inputs:** Parameters of network after sparsity induction, $\Theta_{L1}$ and $\Theta_{L1}^{t\text{h}}$; set of layers to be pruned $L$.

**Output:** New compressed model $M_c$ with weights $\Theta_c$.

Filterindex = empty-list();

\[ \Theta_c = \Theta_{L1} \]

for layer $l$ in $L$ do

for filter $i$ in $\{1, \ldots, c_l\}$ do

Extract $F_i$ and $G_i$ from $\Theta_{L1}^{t\text{h}}$;

splevelF = Sparsity-level($F_i$);

splevelG = Sparsity-level($G_i$);

if splevelF $\geq s_F'$ then

if splevelF $\geq s_F$ then

filterindex.add($i$)

else

if splevelG $\geq s_G$ then

filterindex.add($i$)

end if

end if

end for

for filter $i$ in filterindex do

$\Theta_c = \text{Remove } F_i, G_i \text{ from } W_l, W_{l+1} \text{ in } \Theta_{L1}$;

$\Theta_{t\text{h}}$ = Remove $F_i, G_i$ from $W_l, W_{l+1}$ in $\Theta_{L1}^{t\text{h}}$;

end for

end for

$M_c = \text{Redefine model with the remaining parameters } \Theta_c$;

Initialize $M_c$ with $\Theta_c$;

Return Model $M_c$ and $\Theta_c$. 

proved performance. The reason for improved performance is due to the restoration of values of zero weights that did not get pruned. \( \theta_{L_1} \) only serve as a guide for the selection of filters to be pruned. Algorithm 1 describes the full procedure of filter pruning in detail.

Finally, we retrain the pruned network without L1 regularization to restore the performance drop due to sparsity induction.

4. Experiments

We apply our multilayer pruning framework to compress Single Shot MultiBox Detector (SSD). We experiment on both smaller and larger variant of SSD, namely SSD300 and SSD512. In our initial experiments on German Traffic Sign Detection Benchmark (GTSDB) [18], we achieved a compression of SSD512 by a factor of 26X. To further test the effectiveness of our multi-layer pruning framework, we run experiments on PASCAL VOC [10] where we compress SSD300 by a factor of 7X. We also compare our pruned SSD models with other existing light-weight SSD variants, in terms of model size and parameters. We perform additional experiments on the COCO object detection set [29]. On COCO, we compress SSD300 by a factor of \( \sim 3X \) with no loss of AP@0.5. Please refer to the appendix for more details on our experiments on COCO.

To test the effectiveness of our method and its task independence across the convolutional neural network, we perform experiments for pruning of image classification architecture namely VGG16 on datasets German Traffic Sign Recognition Benchmark (GTSRB) [18] and CIFAR10. On CIFAR10, we compress the VGG16 by a factor 125X with 90.5% reduction in flops.

4.1. Pruning of SSD on PASCAL VOC

SSD consists of a large number of convolutional layers, where multiple convolutional layers contribute to the detections at various scales. The complex dependency of output detections on convolutional layers makes it challenging to prune. Previous works [37] only attempt to prune the base network of SSD. In our experiments, we prune the full network including the detection layers. However, we observed that pruning the entire network together in a single step massively degrades the performance. Therefore, we apply our multilayer pruning strategy and target the network in chunks of layers in multiple phases.

In our experiments, we use SSD with base network VGG16. For pruning of SSD512, we employ our method in 4 phases: conv1\_1 - conv4\_3, conv4\_3 - conv7, conv7 - conv8\_2 and conv8\_2 - conv12\_2. We also include the conv layers that perform the localization and classification in these groups of layers.

In each phase, we introduce sparsity in the model and prune the filters of conv layers as discussed in Section 3.

Figure 2: Weight distributions of convolutional layers conv3\_3 and conv4\_3 of SSD network without (first row) and with (second row) L1 regularization in phase 1.

Figure 3: Number of non-zero parameters in SSD convolutional layers before (blue) and after (orange) different phases of our sparsity induction method on PASCAL VOC dataset.

For SSD300 architecture, we prune the model in 4 phases: conv1\_1 - conv4\_3, conv4\_3 - conv7, conv7 -
Table 1: Layer-wise pruning results in SSD300 and SSD512 models on PASCAL VOC dataset. CONVx_DN denotes the N-th layer that is connected to detection layer (conf and loc layer).

<table>
<thead>
<tr>
<th>Method</th>
<th>data</th>
<th>layer</th>
<th>model</th>
<th>scale</th>
<th>pos</th>
<th>feat</th>
<th>conv</th>
<th>rpn</th>
<th>total</th>
<th>cls</th>
<th>loc</th>
<th>mean AP</th>
<th>mean</th>
<th>FP</th>
<th>MParms</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD512-orig</td>
<td>800.12</td>
<td>73.56</td>
<td>50</td>
<td>76.24</td>
<td>57.07</td>
<td>57.07</td>
<td>56.91</td>
<td>58.92</td>
<td>52.54</td>
<td>57.47</td>
<td>57.47</td>
<td>57.47</td>
<td>74.47</td>
<td>74.47</td>
<td>8.67</td>
</tr>
<tr>
<td>SSD512-pruned-6det</td>
<td>800.12</td>
<td>77.70</td>
<td>63.21</td>
<td>76.72</td>
<td>57.07</td>
<td>57.07</td>
<td>56.91</td>
<td>58.92</td>
<td>52.54</td>
<td>57.47</td>
<td>57.47</td>
<td>57.47</td>
<td>74.47</td>
<td>74.47</td>
<td>8.67</td>
</tr>
<tr>
<td>SSD512-pruned-9det</td>
<td>800.12</td>
<td>77.70</td>
<td>63.21</td>
<td>76.72</td>
<td>57.07</td>
<td>57.07</td>
<td>56.91</td>
<td>58.92</td>
<td>52.54</td>
<td>57.47</td>
<td>57.47</td>
<td>57.47</td>
<td>74.47</td>
<td>74.47</td>
<td>8.67</td>
</tr>
<tr>
<td>SSD500-orig</td>
<td>1024.20</td>
<td>76.72</td>
<td>57.07</td>
<td>56.91</td>
<td>58.92</td>
<td>52.54</td>
<td>57.47</td>
<td>74.47</td>
<td>74.47</td>
<td>8.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSD500-pruned-6det</td>
<td>1024.20</td>
<td>76.72</td>
<td>57.07</td>
<td>56.91</td>
<td>58.92</td>
<td>52.54</td>
<td>57.47</td>
<td>74.47</td>
<td>74.47</td>
<td>8.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSD500-pruned-9det</td>
<td>1024.20</td>
<td>76.72</td>
<td>57.07</td>
<td>56.91</td>
<td>58.92</td>
<td>52.54</td>
<td>57.47</td>
<td>74.47</td>
<td>74.47</td>
<td>8.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: AP for each class with original SSD and pruned SSD models on VOC2007 test dataset. Training data, 07+12 is the union of the VOC2007 and VOC2012 trainval dataset.

Table 3: Comparison of our pruned SSD models with other existing light-weight variants of SSD on VOC 07 test dataset, conv8.2 and conv8.2 - conv11.2. We test two pruned networks on VOC 2007 test set: SSD300-pruned-3det and SSD300-pruned-6det with three and six detection layers respectively as shown in Table 1. AP for each class with original SSD and pruned SSD models on VOC2007 test dataset is shown in Table 2. Table 3 shows the comparison of performance and model parameters (in Millions) of our pruned SSD models with other existing light-weight SSD variants. We also compare SSD300-pruned-6det with compressed F-RCNN models of [16] and found that our pruned model has more than 34X lesser parameters than theirs as shown in Table 4(a).

4.1.1 Baselines

To the best of our knowledge, this is the first work that compresses all layers of SSD including the base network and the detection layers. Therefore, we have created a baseline to show the efficacy of our method. Our baseline RRF-x-6det is obtained by randomly removing x% filters from all layers of SSD300 up to CONV7. Please refer to Table 4(b) for comparison with the baseline.
4.1.2 Hyper-Parameters Selection and Analysis

Regularization constant $\alpha$ is chosen in such a way that performance loss $\epsilon_1 \leq 5$ at a later stage we can not recover the performance loss due to the pruning of the important connections. Recall that $\theta_{L1}$ and $\theta^t_{L1}$ denote the model parameters obtained after applying regularization and global thresholding (with threshold $t$) respectively. The threshold $t$ is selected such that there is a further performance drop $\epsilon_2 \in [5, 7]$ after thresholding. Note that $\theta^t_{L1}$ is only used for selecting unimportant filters while the actual filter pruning is done on the model with parameters $\theta_{L1}$. This allows us to afford a relatively high-performance drop after thresholding and explains the high value of $\epsilon_2$. Experimentally, we found that $\epsilon_2 \geq 10$ results in too much sparsity in $\theta^t_{L1}$ and consequently affects the filter selection procedure by selecting important filters for pruning. This results in an irrecoverable loss in performance.

We use grid search for selecting hyperparameters $s_F, s'_F, s_G, G$ in the range $[0.70, 0.99]$ and found $s_F = 0.9, s'_F = 0.85, s_G = 0.95$ to work best for $\epsilon_1$ and $\epsilon_2$ in the ranges mentioned above.

4.1.3 Ablation study on pruned SSD300

SSD300-pruned-1det and SSD300-pruned-6det have a significant difference in mAP because PASCAL VOC dataset has high variation in object size. It is clear from Table 5, that at least three detection layers are required for a decent mAP. Please note that our pruned models have very fewer parameters in $4^{th}$, $5^{th}$ and $6^{th}$ detection layer but play a significant role in detecting larger objects.

<table>
<thead>
<tr>
<th></th>
<th>#Params</th>
<th>Model Size</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pruned-1det</td>
<td>2.7M</td>
<td>10.8 MB</td>
<td>25.6</td>
</tr>
<tr>
<td>Pruned-2det</td>
<td>3.42M</td>
<td>13.7 MB</td>
<td>66.7</td>
</tr>
<tr>
<td>Pruned-3det</td>
<td>3.7M</td>
<td>15 MB</td>
<td>72.15</td>
</tr>
<tr>
<td>Pruned-4det</td>
<td>3.8M</td>
<td>15.2 MB</td>
<td>74.5</td>
</tr>
<tr>
<td>Pruned-5det</td>
<td>3.85M</td>
<td>15.4 MB</td>
<td>74.94</td>
</tr>
<tr>
<td>Pruned-6det</td>
<td>3.9M</td>
<td>15.7 MB</td>
<td>75.07</td>
</tr>
</tbody>
</table>

Table 5: Ablation study on pruned SSD300 detection layers on PASCAL VOC07.

4.2. Pruning of SSD on GTSDB

We first trained SSD300 and SSD512 on GTSDB and obtained 38.96% and 82.44% mAP on test set respectively. The reason for the considerable difference between the performance of SSD300 and SSD512 is because the bounding boxes in GTSDB are very small on average. We only used SSD512 for our pruning experiments. For SSD512, we use the same set of layers in each phase as mentioned for the VOC 07+12 dataset in Section 4.1. Figure 4 shows the number of non-zero parameters in convolutional layers before and after different phases of our sparsity induction method (GTSDB).

4.3. Pruning of VGG16 on CIFAR10 and GTSRB

Unlike SSD, VGG16 also has fully connected layers in addition to convolutional layers. We prune the network in two phases. We prune all the convolutional layers in the first phase and the second phase we prune all the fully connected layers. In each phase, we first introduce sparsity in parameters and then prune them as discussed in Section 3. For pruning of fully connected layers, we follow Pan et al. [33], where we remove a neuron from a layer if either of the following two conditions is true:
<table>
<thead>
<tr>
<th>Model</th>
<th>AP</th>
<th>Size</th>
<th>Total Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD300-original</td>
<td>38.61</td>
<td>96.1MB</td>
<td>24.0M</td>
</tr>
<tr>
<td>SSD512-original</td>
<td>88.67</td>
<td>98.7 MB</td>
<td>24.7M</td>
</tr>
<tr>
<td>SSD512-pruned-1det</td>
<td>97.95</td>
<td>3.8 MB (26X)</td>
<td>938.8K (3.8%)</td>
</tr>
</tbody>
</table>

Table 7: AP for each class with original SSD and pruned SSD models on GTSDB.

Table 8: Layer-wise pruning results in VGG16 model on CIFAR10 and GTSRB dataset.

Table 9: Comparison of our pruned model (VGG16-Pruned) with [27, 32] on CIFAR10 dataset. [27]* prunes a slightly modified version of VGG16 described in [38].

1. all incoming connections’ weights are zero
2. all outgoing connections’ weights are zero

Figure 5 shows the number of non-zero parameters in conv layers after sparsity induction.

On CIFAR10, our compressed model is 125X smaller than the original model. Table 8 shows the layer-wise pruning results and performance of pruned models on CIFAR10 and GTSRB dataset.

Table 9 shows the comparison of our method with the weight sum method by Li et al. [27] and the method by Neklyudov et al. [32]. Our method prunes 99.2% of parameters on CIFAR10, significantly larger than 64.0% pruned by the weight sum method. Furthermore, our method reduces the FLOPs by 90.5% compared to 34.2% pruned by the weight sum method.

5. Conclusion

In this paper, we have proposed a Multi-layer pruning method. The main idea is to induce sparsity in a first procedure by modifying the training loss function. The second step prunes filter that after being trained with the sparsity-inducing loss-function are zero or nearly zero. This reduced architecture when retrained is observed to provide accurate results for both object detection and classification. Through this method, we have obtained state of the art results for deep model compression for object detection using SSD based object detection on real-world standard benchmark datasets. In the future, we would be interested in exploring this approach for other visual recognition methods and further evaluate other related loss functions that can aid pruning of the filters.
References


