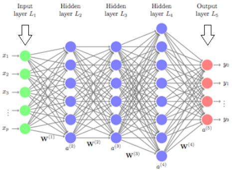
Collaborative Channel Pruning for Deep Networks

11th June 2019

Tencent
AI Lab

Background



Source:https://orbograph.com/deep-learning-how-will-it-change-healthcare/

Source:http://mypcsupport.ca/portable-devices/

Tencent AI Lab

Model compression method

- Compact network design;
- Network quantization;
- Channel or filter pruning;

Here we focus on channel pruning.

Background

Some criterion for channel pruning

- Magnitude-based pruning of weights.e.g. ℓ_1-norm (Li et al.,2016) and ℓ_2-norm (He et al.,2018a);
- Average percentage of zeros (Luo et al., 2017);
- First-order information (Molchanov et al., 2017);

Tencent AI Lab

Background

Some criterion for channel pruning

- Magnitude-based pruning of weights.e.g. ℓ_1-norm (Li et al.,2016) and ℓ_2-norm (He et al.,2018a);
- Average percentage of zeros (Luo et al., 2017);
- First-order information (Molchanov et al., 2017);

These measures **consider channels independently** to determine pruned channels.

Tencent AI Lab

Motivation

We focus on exploiting the **inter-channel dependency** to determine pruned channels.

Problems:

- Criterion to represent the inter-channel dependency?
- Effects on loss function?

Tencent AI Lab

Method

We analyze the impact via second-order Taylor expansion:

$$\mathcal{L}(\beta, \mathbf{W}) \approx \mathcal{L}(\mathbf{W}) + \mathbf{g}^T\mathbf{v} + \frac{1}{2}\mathbf{v}^T\mathbf{H}\mathbf{v}, \tag{1}$$

An efficient way to approximate $\mathbf{H}$.

- For least-square loss, $\mathbf{H} \approx \mathbf{g}^T\mathbf{g}$;
- For cross-entropy loss, $\mathbf{H} \approx \mathbf{g}^T\Sigma\mathbf{g}$;

where $\Sigma = diag\left(\left(\mathbf{y} \oslash \left(f\left(\mathbf{w}, \mathbf{x}\right) \odot f\left(\mathbf{w}, \mathbf{x}\right)\right)\right)\right)$.

Tencent AI Lab

Method

We reformulate Eq.1 to a linearly constrained binary quadratic problem[1]:

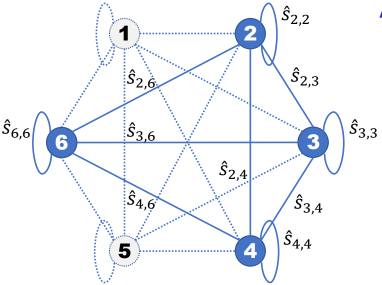
$$\begin{aligned} &\min\ \boldsymbol{\beta}^T \hat{\mathbf{S}} \boldsymbol{\beta} \\ &\text{s.t. } \mathbf{1}^T \boldsymbol{\beta} = p,\ \boldsymbol{\beta} \in \{0,1\}^{c_o}\ . \end{aligned} \tag{2}$$

The pairwise correlation matrix $\hat{\mathbf{S}}$ reflects the inter-channel dependency.

Tencent AI Lab

[1]More details can be found in our paper

Method



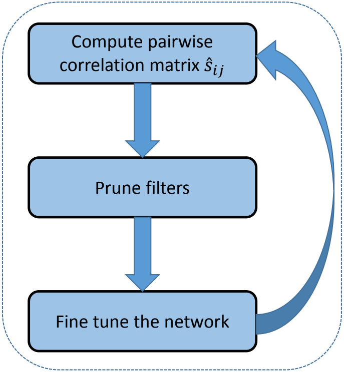
A graph perspective:

- Nodes denote channels
- Edges are assigned with the corresponding weight $\hat{s}_{ij}$.
- Find a sub-graph such the sum of included weights is minimized.

Tencent AI Lab

Method

Algorithm



Algorithm 1 Collaborative Channel Pruning

Input: Training set $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$
Input: Pre-trained network $\boldsymbol{\theta}_0 = \{\mathbf{W}_0^{(l)}\}_{l=1}^{L}$
Output: Channel pruned network $\boldsymbol{\theta} = \{(\boldsymbol{\beta}^{(l)}, \mathbf{W}^{(l)})\}_{l=1}^{L}$

```
initialize {u_i} and {s_ij} for all layers
for n = 1, ..., N do
    compute outputs and gradients for (x_n, y_n)
    update {u_i} and {s_ij} for all layers
end for
for l = 1, ..., L do
    compute pairwise correlation matrix Ŝ
    solve (22) to obtain binary mask β^(l)
end for
fine-tune the model with binary masks {β^(l)}
```

Tencent
AI Lab

Results

Table 1: Comparison on the classification accuracy drop and reduction in FLOPs of ResNet-56 on the CIFAR-10 data set.

Method	Baseline	Pruned	
	Acc.	Acc. ↓	FLOPs
Channel Pruning (He et al.,2017)	92.80%	1.00%	50.0%
AMC (He et al., 2018b)	92.80%	0.90%	50.0%
Pruning Filters (Li et al., 2016)	93.04%	-0.02%	27.6%
Soft Pruning (He et al., 2018a)	93.59%	0.24%	52.6%
DCP (Zhuang et al., 2018)	93.80%	0.31%	50.0%
DCP-Adapt (Zhuang et al., 2018)	93.80%	-0.01%	47.0%
CCP	93.50%	**0.08%**	**52.6%**
CCP-AC		**-0.19%**	**47.0%**

Tencent
AI Lab

Results

Table 2: Comparison on the top-1/5 classification accuracy drop, and reduction of ResNet-50 in FLOPs on the ILSVRC-12 data set.

Method	Baseline		Pruned		
	Top-1	Top-5	Top-1 $\downarrow$	Top-5 $\downarrow$	FLOPs
Channel Pruning	-	92.20%	-	1.40%	50.0%
ThiNet	72.88%	91.14%	1.87%	1.12%	55.6%
Soft Pruning	76.15%	92.87%	1.54%	0.81%	41.8%
DCP	76.01%	92.93%	1.06%	0.61%	55.6%
Neural Importance	-	-	0.89%	-	44.0%
CCP	76.15%	92.87%	**0.65%**	**0.25%**	**48.8%**
CCP			0.94%	0.45%	54.1%
CCP-AC			**0.83%**	**0.33%**	**54.1%**

Tencent AI Lab

Thanks for your attention!

Tencent
AI Lab