Feature Variance Ratio-Guided Channel Pruning for Deep Convolutional Network Acceleration: Supplementary Material

Junjie He[0000−0002−1009−945X], Bohua Chen[0000−0003−4467−1060], Yinzhang Ding[0000−0001−9375−9400], and Dongxiao Li*[0000−0002−5619−0419]

Zhejiang University, Hangzhou 310027, China
{he_junjie, chenbohua, dingyzh, lidx}@zju.edu.cn

1 Proof

Proposition 1 *Let x, y be two n-dimensional data vectors with elements $\{x_i\}$ and $\{y_i\}$, respectively. Assume their Pearson correlation coefficient is r_{xy}, i.e.,*

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}, \tag{1}$$

where $\bar{x} = \sum_i x_i/n, \bar{y} = \sum_i y_i/n$. Let $\varepsilon_i = y_i - x_i, i = 1, 2, ..., n$, be the residuals of data elements, forming a vector ε. Suppose that the variances of ε and y are σ_ε^2, σ_y^2 respectively, and $\sigma_y^2 > 0$, $\sigma_\varepsilon^2 \neq \sigma_y^2$. Then we have:

$$0 \leq 1 - r_{xy}^2 \leq \frac{\sigma_\varepsilon^2/\sigma_y^2}{(1 - \sigma_\varepsilon/\sigma_y)^2}. \tag{2}$$

Proof. Without loss of generality, we assume that $\bar{x} = \bar{y} = 0$. In convenient vector notation, we define

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i, \tag{3}$$

the *inner product* between x and y. Then we can write

$$r_{xy} = \frac{\langle x, y\rangle}{\sqrt{\langle x, x\rangle}\sqrt{\langle y, y\rangle}}. \tag{4}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} r_{xy}^2 &= \frac{|\langle x, y\rangle|^2}{\langle x, x\rangle\langle y, y\rangle} \\ &\leq 1, \end{aligned} \tag{5}$$

* Corresponding author

and

$$\begin{aligned} r_{xy}^2 &= \frac{|\langle x,y\rangle|^2}{\langle x,x\rangle\langle y,y\rangle} \\ &= \frac{|\langle y,y\rangle - \langle \varepsilon,y\rangle|^2}{(\langle y,y\rangle - 2\langle \varepsilon,y\rangle + \langle \varepsilon,\varepsilon\rangle)\langle y,y\rangle} \\ &\geq \frac{|\langle y,y\rangle|^2 - 2\langle \varepsilon,y\rangle\langle y,y\rangle}{(\langle y,y\rangle - 2\langle \varepsilon,y\rangle + \langle \varepsilon,\varepsilon\rangle)\langle y,y\rangle} \\ &= 1 - \frac{\langle \varepsilon,\varepsilon\rangle}{\langle y,y\rangle - 2\langle \varepsilon,y\rangle + \langle \varepsilon,\varepsilon\rangle} \\ &\geq 1 - \frac{\langle \varepsilon,\varepsilon\rangle}{\langle y,y\rangle - 2\sqrt{\langle y,y\rangle\langle \varepsilon,\varepsilon\rangle} + \langle \varepsilon,\varepsilon\rangle} \\ &= 1 - \frac{\frac{\langle \varepsilon,\varepsilon\rangle}{\langle y,y\rangle}}{(1-\sqrt{\frac{\langle \varepsilon,\varepsilon\rangle}{\langle y,y\rangle}})^2}. \end{aligned} \tag{6}$$

Combing the preceding inequalities and noting that $\sigma_y^2 = \langle y,y\rangle/n, \sigma_\varepsilon^2 = \langle \varepsilon,\varepsilon\rangle/n$, we conclude the proof.

□

2 Network-Equivalent Transformation Invariance

As illustrated in the paper, due to the positively homogeneous property of ReLU and normalization process of BN, there exist two network-equivalent transformations. Mathematically, they can be expressed as:

$$\mathrm{BN}(K_{j,:}^l * X^l) = \mathrm{BN}(\alpha_1 K_{j,:}^l * X^l), \tag{7}$$

and

$$K_{:,i}^l * \mathrm{ReLU}(\gamma_i^{l-1} Z_i^{l-1} + \beta_i^{l-1}) = \frac{1}{\alpha_2} K_{:,i}^l * \mathrm{ReLU}(\alpha_2\gamma_i^{l-1} Z_i^{l-1} + \alpha_2\beta_i^{l-1}), \tag{8}$$

$$Z_i^{l-1} = \frac{Y_i^{l-1} - \mathrm{mean}(Y_i^{l-1})}{\mathrm{std}(Y_i^{l-1})}. \tag{9}$$

Here $\alpha_1, \alpha_2 > 0$, and the superscript l represents the layer index of that variable.

Our proposed metric is:

$$\mathrm{SFVR}_{i_0} = \sum_{j=1}^{N} \mathrm{FVR}_{j,i_0} = \sum_{j=1}^{N} \frac{\sigma_{M_{j,i_0}}^2}{\sigma_{Y_j}^2}. \tag{10}$$

For the first transformation, which scales a filter by α_1, we have the following transformed results:

$$\widehat{M}_{j,i}^l = \widehat{K}_{j,i}^l * \widehat{X}_i^l = \alpha_1 K_{j,i}^l * X_i^l = \alpha_1 M_{j,i}^l, \quad i = 1, \ldots, C, \tag{11}$$

and

$$\widehat{Y}_j^l = \sum_{i=1}^{C} \widehat{M}_{j,i}^l = \alpha_1 \sum_{i=1}^{C} M_{j,i}^l = \alpha_1 Y_j^l. \tag{12}$$

The effect of α_1 is canceled by the division. The calculated SFVR of each channel in the l-th layer is thus unaffected. For the later layers, the calculated results are also unchanged since the effect of α_1 has been normalized by the BN transform in the l-th layer.

For the second transformation, we have:

$$\widehat{M}_{j,i}^l = \widehat{K}_{j,i}^l * \widehat{X}_i^l = \frac{1}{\alpha_2} K_{j,i}^l * \alpha_2 X_i^l = M_{j,i}^l, \quad j = 1, \ldots, N, \tag{13}$$

and

$$\widehat{Y}_j^l = Y_j^l, \quad j = 1, \ldots, N. \tag{14}$$

The calculation of SFVR remains the same.

Concluding above, we arrive at our result that SFVR is invariant to the two equivalent transformations.

3 General Networks without Batch Normalization

For the general networks without BN, we can still use SFVR to measure the importance of corresponding channels in the network. Just in this case we cannot leverage the statistics of BN to estimate the variances of output feature maps and therefore require spending a little extra computation cost on them.

In fact, for the network without BN, the magnitude of parameters can also vary independently of the channel importance:

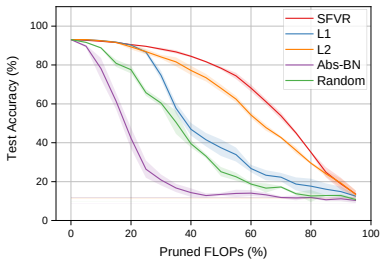
$$K_{:,i}^l * f(K_{i,:}^{l-1} * X^{l-1} + b_i^{l-1}) = \frac{1}{\alpha} K_{:,i}^l * f(\alpha K_{i,:}^{l-1} * X^{l-1} + \alpha b_i^{l-1}), \tag{15}$$

where α is a positive scalar, b_i^{l-1} is the i-th bias term in the (l-1)-th layer, and $f(\cdot)$ is a positively homogeneous activation like ReLU$(\cdot)$. The parameter magnitudes are less relevant to the identification of channel importance, either. In this case, however, since the scale and bias of convolutional output feature map Y_j^l can be relearned by the channel weights $K_{:,j}^{l+1}$ and bias term b_j^l, we can still use the Pearson correlation coefficient to characterize the essential information loss of Y_j^l resulted from pruning and then derive SFVR from Proposition 1 to describe the channel importance. SFVR is also invariant to the transformation 15. Note that for the activation that is not positively homogeneous, SFVR is also meaningful from the feature-correlation perspective.

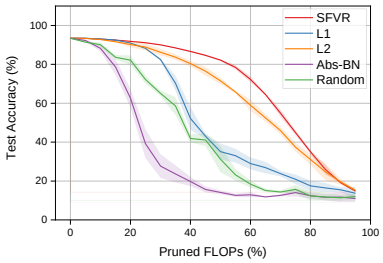
4 Comparison of Channel Importance Metrics on Deeper Networks

Fig. 1 presents the single-shot pruning results of different metrics on deeper networks. Consistently, our proposed SFVR metric outperforms the conventional

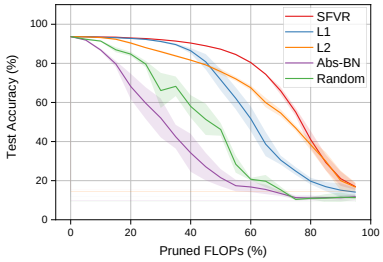
magnitude-based ones, better in identifying redundant channels in modern networks. All these experiments are repeated 3 times with different random seeds. ± standard derivation is reported with the shaded region.



(a) Single-shot pruning without fine-tuning on PreResNet-32 on CIFAR-10



(b) Single-shot pruning without fine-tuning on PreResNet-44 on CIFAR-10

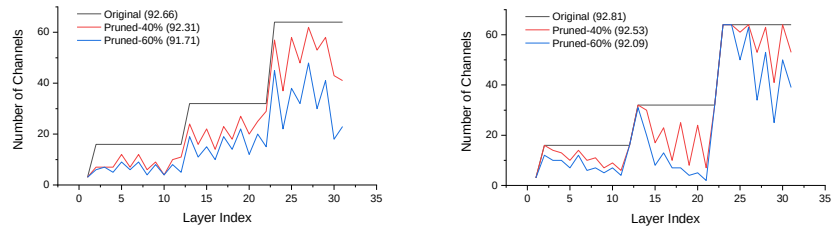


(c) Single-shot pruning without fine-tuning on PreResNet-110 on CIFAR-10

Fig. 1. Comparison of channel importance metrics on deeper networks.

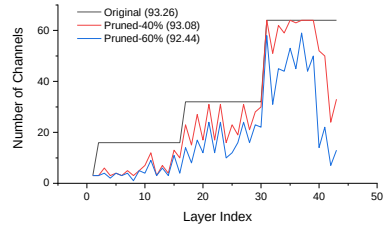
5 Pruned Architectures

Fig. 2 shows the FVRCP-pruned architectures of many popular networks, including ResNets, pre-activation ResNets, DenseNets, and MobileNets.

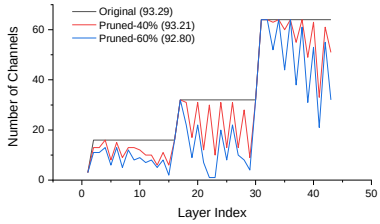


(a) ResNet-32

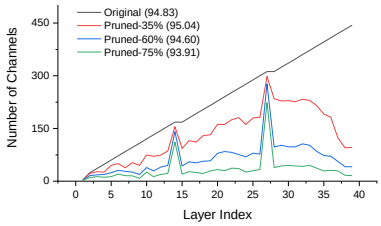
(b) PreResNet-32



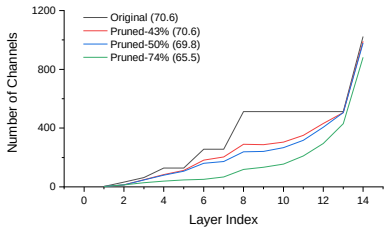
(c) ResNet-44



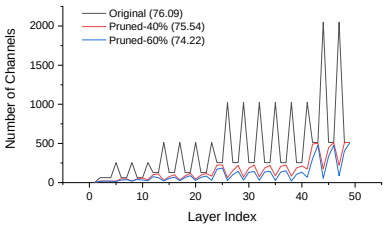
(d) PreResNet-44



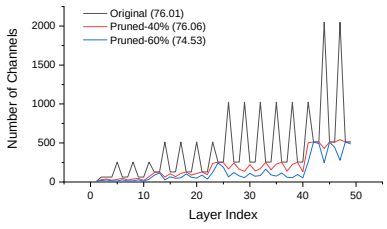
(e) DenseNet-40



(f) MobileNet V1



(g) ResNet-50

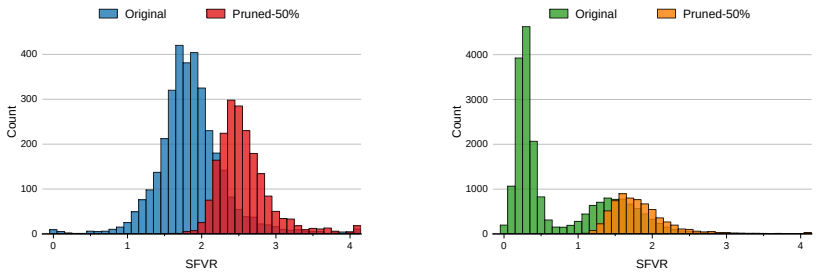


(h) PreResNet-50

Fig. 2. Illustration of FVRCP-pruned architectures. The FLOPs pruning ratio and corresponding (top-1) accuracy are reported in the legend.

6 Distribution of SFVR

Fig. 3 shows the distribution of SFVR in the original and pruned ResNet-18 and ResNet-50 models. After pruning, the distribution of SFVR concentrates on a larger value, which implies that the representation of feature maps has become more compact.



(a) ResNet-18 (b) ResNet-50

Fig. 3. Distribution of SFVR in the original and pruned (50% FLOPs reduction by FVRCP) ResNet-18 and ResNet-50 models.