SCIENCE CHINA
Information Sciences

• Supplementary File •

Supplementary Materials of Exploring Dispersion of Feature Distribution from Parameters for Network Pruning

Dongyue WU[1], Zilin GUO[1], Nong SANG[1] & Changxin GAO[1*]

[1]*National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artifcial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China*

Appendix A The Detailed Proof

In the main text, we directly give the final results of the estimations. Here we provide detailed proof of Equation 15-M and Equation 21-M of the main text. To differentiate, we use "Equation X-M" to denote the equations in the main text. The equations in this appendix are denoted as "Equation X".

Appendix A.1 Detailed Proof of the Estimated Expectation of Activation Maps with ReLU (Equation 15-M)

According to LOTUS introduced in Equation 3-M, we can get the following derivation:

$$\begin{aligned}\hat{\mathbb{E}}(A_j^i)_{relu} &= \int_{-\infty}^{+\infty} p_x(x)\cdot \mathrm{ReLU}(x)\ \mathrm{d}x \\ &= \int_{0}^{+\infty} \frac{x}{\gamma_j\sqrt{2\pi}}\ e^{-\frac{(x-\beta_j)^2}{2\gamma_j^2}}\ \mathrm{d}x.\end{aligned} \tag{A1}$$

Let $t=\frac{x-\beta_j}{\sqrt{2}\gamma_j}$, the above equation can be formulated as:

$$\begin{aligned}\hat{\mathbb{E}}(A_j^i)_{relu} &= \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} \frac{\beta_j}{\sqrt{\pi}}\ e^{-t^2}\ \mathrm{d}t + \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} \frac{t}{\sqrt{\pi}}\ e^{-t^2}\sqrt{2}\gamma_j\ \mathrm{d}t \\ &= \frac{\beta_j}{\sqrt{\pi}}\left[\int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{0} e^{-t^2}\ \mathrm{d}t + \int_{0}^{+\infty} e^{-t^2}\ \mathrm{d}t\right] + \frac{-\gamma_j}{\sqrt{2\pi}}\ e^{-t^2}\Big|_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} \\ &= \frac{\beta_j}{2}\left[\frac{2}{\sqrt{\pi}}\int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{0} e^{-t^2}\ \mathrm{d}t + \mathrm{erf}(+\infty)\right] + \frac{\gamma_j}{\sqrt{2\pi}}e^{\frac{-\beta_j^2}{2\gamma_j^2}}\end{aligned} \tag{A2}$$

Since $\mathrm{erf}(z)=\frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\ \mathrm{d}t$ and $\mathrm{erf}(+\infty)=1$, we get:

$$\hat{\mathbb{E}}(A_j^i)_{relu} = \frac{\beta_j}{2}[1-\mathrm{erf}(\frac{-\beta_j}{\sqrt{2}\gamma_j})] + \frac{\gamma_j}{\sqrt{2\pi}}e^{\frac{-\beta_j^2}{2\gamma_j^2}} \tag{A3}$$

Considering the Gauss error function is an odd function, we can formulate the estimated expectation as the follows:

$$\begin{aligned}\hat{\mathbb{E}}(A_j^i)_{relu} &= \frac{\beta_j}{2}\left[1+\mathrm{erf}(\frac{\beta_j}{\sqrt{2}\gamma_j})\right] + \frac{\gamma_j}{\sqrt{2\pi}}e^{\frac{-\beta_j^2}{2\gamma_j^2}} \\ &= \frac{\beta_j}{2}(1+A)+B,\end{aligned} \tag{A4}$$

where $A=\mathrm{erf}(\frac{\beta_j}{\sqrt{2}\gamma_j})$, $B=\frac{\gamma_j}{\sqrt{2\pi}}e^{\frac{-\beta_j^2}{2\gamma_j^2}}$. The final results of the Equation A4 is just the same as Equation 15-M.

* Corresponding author (email: cgao@hust.edu.cn)

Appendix A.2 Proof of the Estimated Variance of Activation Maps with ReLU (Equation 21-M)

Based on the Equation 16-M in the main text, we can derive the following results of the $\mathbb{V}(\hat{A_j^i})_{relu}$:

$$\begin{aligned}\hat{\mathbb{V}}(A_j^i)_{relu} &= \int_{-\infty}^{+\infty} p_X(x)[\mathrm{ReLU}(x) - \hat{\mathbb{E}}(A_j^i)]^2 \, \mathrm{d}x \\ &= \hat{\mathbb{E}}(A_j^i)^2 \int_{-\infty}^{0} e^{-\frac{(x-\beta_j)^2}{2\gamma_j^2}} \, \mathrm{d}x + \int_{0}^{+\infty} [x - \hat{\mathbb{E}}(A_j^i)]^2 \, e^{-\frac{(x-\beta_j)^2}{2\gamma_j^2}} \, \mathrm{d}x\end{aligned} \tag{A5}$$

With $t = \frac{x-\beta_j}{\sqrt{2}\gamma_j}$, we can derive the following equation:

$$\begin{aligned}\hat{\mathbb{V}}(A_j^i)_{relu} =& \frac{\hat{\mathbb{E}}(A_j^i)^2}{\sqrt{\pi}} \int_{-\infty}^{\frac{-\beta_j}{\sqrt{2}\gamma_j}} e^{-t^2} \, \mathrm{d}t \, + \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} \frac{[\sqrt{2}\gamma_j t + \beta_j - \hat{\mathbb{E}}(A_j^i)]^2}{\sqrt{\pi}} \, e^{-t^2} \, \mathrm{d}t \\ =& \frac{\hat{\mathbb{E}}(A_j^i)^2}{\sqrt{\pi}} \int_{-\infty}^{\frac{-\beta_j}{\sqrt{2}\gamma_j}} e^{-t^2} \mathrm{d}t + \frac{2\gamma_j^2}{\sqrt{\pi}} \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} t^2 e^{-t^2} \mathrm{d}t + \frac{2\sqrt{2}\gamma_j}{\sqrt{\pi}} \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} \left[\beta_j - \hat{\mathbb{E}}(A_j^i)\right] t e^{-t^2} \mathrm{d}t + \\ & \frac{1}{\sqrt{\pi}} \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} \left[\beta_j - \hat{\mathbb{E}}(A_j^i)\right]^2 \, e^{-t^2} \mathrm{d}t.\end{aligned} \tag{A6}$$

Here, we calculate each of the four terms separately. The first term can be formulated as follows:

$$\begin{aligned}\frac{\hat{\mathbb{E}}(A_j^i)^2}{\sqrt{\pi}} \int_{-\infty}^{\frac{-\beta_j}{\sqrt{2}\gamma_j}} e^{-t^2} \, \mathrm{d}t =& \frac{\hat{\mathbb{E}}(A_j^i)^2}{\sqrt{\pi}} \left(\int_{-\infty}^{0} e^{-t^2} \, \mathrm{d}t + \int_{0}^{\frac{-\beta_j}{\sqrt{2}\gamma_j}} e^{-t^2} \, \mathrm{d}t \right) \\ =& \frac{\hat{\mathbb{E}}(A_j^i)^2}{2} \left(\frac{2}{\sqrt{\pi}} \int_{-\infty}^{0} e^{-t^2} \, \mathrm{d}t + \frac{2}{\sqrt{\pi}} \int_{0}^{\frac{-\beta_j}{\sqrt{2}\gamma_j}} e^{-t^2} \, \mathrm{d}t \right) \\ =& \frac{\hat{\mathbb{E}}(A_j^i)^2}{2} \left[\frac{-2}{\sqrt{\pi}} \int_{0}^{-\infty} e^{-t^2} \, \mathrm{d}t + \mathrm{erf}(\frac{-\beta_j}{\sqrt{2}\gamma_j}) \right] \\ =& \frac{\hat{\mathbb{E}}(A_j^i)^2}{2} \left[\mathrm{erf}(+\infty) - \mathrm{erf}(\frac{\beta_j}{\sqrt{2}\gamma_j}) \right] \\ =& \frac{\hat{\mathbb{E}}(A_j^i)^2}{2} (1 - A).\end{aligned} \tag{A7}$$

Similarly, we can calculate the second, third, and fourth term, respectively. The results of them can be formulated as follows:

$$\begin{aligned}\frac{2\gamma_j^2}{\sqrt{\pi}} \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} t^2 e^{-t^2} \mathrm{d}t =& \frac{-\gamma_j^2}{\sqrt{\pi}} \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} t(-2t) e^{-t^2} \mathrm{d}t \\ =& \frac{-\gamma_j^2}{\sqrt{\pi}} \left(t e^{-t^2} \Big|_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} - \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} e^{-t^2} \mathrm{d}t \right) \\ =& \frac{-\gamma_j^2}{\sqrt{\pi}} (0 - \frac{(-\beta_j)}{\sqrt{2}\gamma_j} e^{\frac{-\beta_j^2}{2\gamma_j^2}}) + \frac{\gamma_j^2}{2} \cdot \frac{2}{\sqrt{\pi}} \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} e^{-t^2} \, \mathrm{d}t \\ =& \frac{-\beta_j \gamma_j}{\sqrt{2\pi}} e^{\frac{-\beta_j^2}{2\gamma_j^2}} + \frac{\gamma_j^2}{2} \left[\mathrm{erf}(+\infty) - \mathrm{erf}(-\frac{\beta_j}{\sqrt{2}\gamma_j}) \right] \\ =& -\beta_j B + \frac{\gamma_j^2}{2} (1 + A),\end{aligned} \tag{A8}$$

$$\begin{aligned}\frac{2\sqrt{2}\gamma_j}{\sqrt{\pi}} \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} \left[\beta_j - \hat{\mathbb{E}}(A_j^i)\right] t e^{-t^2} \mathrm{d}t =& -\sqrt{\frac{2}{\pi}} \gamma_j \left[\beta_j - \hat{\mathbb{E}}(A_j^i)\right] \int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty} -2t e^{-t^2} \mathrm{d}t \\ =& -\sqrt{\frac{2}{\pi}} \gamma_j \left[\beta_j - \hat{\mathbb{E}}(A_j^i)\right] e^{-t^2} \Big|_{\frac{-\beta_j^2}{2\gamma_j^2}}^{+\infty} \\ =& \frac{2\gamma_j}{\sqrt{2\pi}} \left[\beta_j - \hat{\mathbb{E}}(A_j^i)\right] e^{\frac{-\beta_j^2}{2\gamma_j^2}} \\ =& 2 \left[\beta_j - \hat{\mathbb{E}}(A_j^i)\right] B,\end{aligned} \tag{A9}$$

$$\begin{aligned}\frac{1}{\sqrt{\pi}}\int_{\frac{-\beta_j}{\sqrt{2}\gamma_j}}^{+\infty}\left[\beta_j-\hat{\mathbb{E}}(A_j^i)\right]^2 e^{-t^2}\mathrm{d}t=&\frac{\left[\beta_j-\hat{\mathbb{E}}(A_j^i)\right]^2}{\sqrt{\pi}}\left(\int_0^{+\infty}e^{-t^2}\mathrm{d}t-\int_0^{\frac{-\beta_j}{\sqrt{2}\gamma_j}}e^{-t^2}\mathrm{d}t\right)\\ =&\frac{\left[\beta_j-\hat{\mathbb{E}}(A_j^i)\right]^2}{2}\left[\mathrm{erf}(+\infty)-\mathrm{erf}(-\frac{\beta_j}{\sqrt{2}\gamma_j})\right]\\ =&\frac{1}{2}\left[\beta_j-\hat{\mathbb{E}}(A_j^i)\right]^2(1+A).\end{aligned}\tag{A10}$$

Hence, the estimated $\hat{\mathbb{V}}(A_j^i)_{relu}$ can be formulated as follows:

$$\begin{aligned}\hat{\mathbb{V}}(A_j^i)_{relu}=&\frac{\hat{\mathbb{E}}(A_j^i)^2}{2}(1-A)-\beta_jB+\frac{\gamma_j^2}{2}(1+A)+2\left[\beta_j-\hat{\mathbb{E}}(A_j^i)\right]B+\frac{1}{2}\left[\beta_j-\hat{\mathbb{E}}(A_j^i)\right]^2(1+A)\\ =&\frac{\hat{\mathbb{E}}(A_j^i)^2}{2}(1-A)+\left[\beta_j-2\hat{\mathbb{E}}(A_j^i)\right]B+\frac{1}{2}\left(\left[\beta_j-\hat{\mathbb{E}}(A_j^i)\right]^2+\gamma_j^2\right)(1+A),\end{aligned}\tag{A11}$$

which is the same with Equation 21-M in the main text.

Appendix B Importance Function for More Architectures

Please note that although we only provide the derivation of $\hat{\mathbb{E}}(A_j^i)$ and $\hat{\mathbb{V}}(A_j^i)$ for Conv-BN-ReLU as an example in the main text, our method can also generalize on networks using other activation functions, like $tanh(\cdot)$ and $ReLU6(\cdot)$, and also other modules with different architectures. In this section, we show more derived results of estimations for more activation functions and modules.

Appendix B.1 Importance Function for Conv-BN-Tanh Modules

Our method can also performs well on networks using $tanh(\cdot)$ as activation function. Based on Equation 15-M, the estimated mean value of the activation map can be formulated as:

$$\begin{aligned}\hat{\mathbb{E}}(A_j^i)_{tanh}=&\int_{-\infty}^{+\infty}\frac{tanh(x)}{\gamma_j\sqrt{2\pi}}\,e^{-\frac{(x-\beta_j)^2}{2\gamma_j^2}}\,\mathrm{d}x\\ =&\,tanh(\beta_j).\end{aligned}\tag{B1}$$

The estimation of the activation map's variance are as follows:

$$\begin{aligned}\hat{\mathbb{V}}(A_j^i)_{tanh}=&\int_{-\infty}^{+\infty}\frac{[tanh(x)-\hat{\mathbb{E}}(A_j^i)]^2}{\gamma_j\sqrt{2\pi}}\,e^{-\frac{(x-\beta_j)^2}{2\gamma_j^2}}\,\mathrm{d}x\\ =&\,tanh^2(\beta_j^2).\end{aligned}\tag{B2}$$

Appendix B.2 Importance Function for Conv-BN-ReLU6 Modules in the MobileNetV2

Here we provide the estimation of $\hat{\mathbb{E}}(A_j^i)_{re6}$ and $\hat{\mathbb{V}}(A_j^i)_{re6}$ for the Conv-Norm-ReLU6 module adopted in the MobileNetV2 [1] using the similar methods shown in the Appendix Appendix A. The ReLU6 function is defined as:

$$\mathrm{ReLU6}(x)=\begin{cases}0, & x<0,\\ x, & 0\leqslant x<6,\\ 6, & x\geqslant 6.\end{cases}\tag{B3}$$

Combining the ReLU6$(\cdot)$ with Equation 15-M, we can obtain the following equation:

$$\begin{aligned}\hat{\mathbb{E}}(A_j^i)_{re6}=&\int_{-\infty}^{+\infty}p_x(x)\cdot\mathrm{ReLU6}(x)\,\mathrm{d}x\\ =&\int_{-\infty}^{0}0\cdot p_x(x)\,\mathrm{d}x+\int_0^6 x\cdot p_x(x)\,\mathrm{d}x+\int_6^{-\infty}6\cdot p_x(x)\,\mathrm{d}x\\ =&\frac{\gamma_j}{\sqrt{2\pi}}\left[e^{\frac{-\beta_j^2}{2\gamma_j^2}}-e^{\frac{-(\beta_j-6)^2}{2\gamma_j^2}}\right]+\frac{\beta_j}{2}\left[\mathrm{erf}(\frac{6-\beta_j}{\sqrt{2}\gamma_j})+\mathrm{erf}(\frac{\beta_j}{\sqrt{2}\gamma_j})\right]+3\left[1+\mathrm{erf}(\frac{\beta_j-6}{\sqrt{2}\gamma_j})\right]\\ =&D+\frac{\beta_j}{2}F+3G,\end{aligned}\tag{B4}$$

where $D=\frac{\gamma_j}{\sqrt{2\pi}}\left[e^{\frac{-\beta_j^2}{2\gamma_j^2}}-e^{\frac{-(\beta_j-6)^2}{2\gamma_j^2}}\right]$, $F=\mathrm{erf}(\frac{6-\beta_j}{\sqrt{2}\gamma_j})+\mathrm{erf}(\frac{\beta_j}{\sqrt{2}\gamma_j})$, and $G=1+\mathrm{erf}(\frac{\beta_j-6}{\sqrt{2}\gamma_j})$.

Similarly, we can also get the $\hat{\mathbb{V}}(A_j^i)_{re6}$ based on Equation 16-M when the activation function is ReLU6(·):

$$\begin{aligned}\hat{\mathbb{V}}(A_j^i)_{re6} &= \int_{-\infty}^{+\infty} p_X(x)[f(x) - \hat{\mathbb{E}}(A_j^i)_{re6}]^2 \ \mathrm{d}x \\ &= \int_{-\infty}^{+\infty} p_X(x)[\mathrm{ReLU6}(x) - \hat{\mathbb{E}}(A_j^i)_{re6}]^2 \ \mathrm{d}x \\ &= \int_{-\infty}^{0} \hat{\mathbb{E}}(A_j^i)_{re6}^2 \ p_x(x) \ \mathrm{d}x + \int_0^6 (x - \hat{\mathbb{E}}(A_j^i)_{re6})^2 p_x(x) \ \mathrm{d}x + \int_6^{-\infty} (6 - \hat{\mathbb{E}}(A_j^i)_{re6})^2 p_x(x) \ \mathrm{d}x \\ &= [\beta_j - 2\hat{\mathbb{E}}(A_j^i)_{re6}]D - \frac{6\gamma_j}{\sqrt{2\pi}} e^{-\frac{(\beta_j-6)^2}{2\gamma_j^2}} + \frac{\hat{\mathbb{E}}(A_j^i)_{re6}^2}{2}(1-A) + \frac{1}{2}\left[(\hat{\mathbb{E}}(A_j^i)_{re6} - \beta_j)^2 + \gamma_j^2\right] F + \frac{\left[\hat{\mathbb{E}}(A_j^i)_{re6} - 6\right]^2}{2} G.\end{aligned} \tag{B5}$$

Appendix B.3 Importance Function for Vision Transformers using Layer Norm

We also implement our pruning method on vision transformers. Since the commonly used Layer Norm (LN) layers do not contain the channel-wise running mean and running variance, we have to ignore the $\phi(P(O_j^i))$. Thus, the importance score of the LN can be calculated as the following:

$$\Phi(l^i, j)_{LN} = \phi(P(A_j^i)) = \frac{\hat{\mathbb{V}}(A_j^i)}{\hat{\mathbb{E}}(A_j^i)}. \tag{B6}$$

The Transformer Block usually employs a 'pre-norm' architecture and without activation right behind the LN. Thus, in this case, we can regard the activation function as $f(x) = x$. For simplicity, we approximate the mean and variance using the scale factor and bias of each channel. Hence, the important score of a LN layer can be approximated as:

$$\Phi(l^i, j)_{LN} \approx \frac{\gamma_j^2}{\beta_j}. \tag{B7}$$

Appendix B.4 Importance Function for Modules using Group Norm and Instance Norm

Since Group Norm and Instance Norm also do not tracking the running mean and running variance by default, their importance score can be calculated just the same as LN. *However, it does not means that we can not get the $\phi(P(O_j^i))$ for all cases.* If you are using PyTorch, simply set `track_running_stats=True` when defining the `InstanceNorm` layer. Then, after training, we can still get the running mean and running variance. Hence, the Conv-BN-ReLU imporance score in the main text (Equation 6 and 22) is still effective for IN.

In practice, with only minimal modification, one can achieve the same effect as the `track_running_stats=Ture` option in Instance Norm layer *by adding code during training to record the running mean and variance after LayerNorm or GroupNorm.* Since this operation involves only extremely fast averaging and addition, its computational cost is negligible and has virtually no impact on training speed. With this approach, transformer blocks using LayerNorm can also employ the same importance function as the Conv–BN–ReLU modules.

Appendix B.5 Importance Function for BasicBlock and BottleNeck of ResNet

We employ our importance function for Conv-BN-ReLU introduced in the main text for the first Conv-BN-ReLU module in the BasicBlock. For the second Conv-BN module, the importance function differs due to the presence of a shortcut in the BasicBlock. The final output of the BasicBlock is the activated sum of the output from the Conv-BN module and the original input of the BasicBlock. Thus, we can not simply adopt the same importance score of a Conv-BN-ReLU module. For simplicity, we can regard the Conv-BN module as a Conv-BN-Identity module. Hence, the importance function for the second Conv-BN module, the importance function is defined as:

$$\Phi(l^i, j)_{2nd}^{Res56} = \frac{(\hat{\sigma}_j^i)^2}{\hat{\mu}_j^i} \frac{\gamma_j^2}{\beta_j}. \tag{B8}$$

Similarly, for the third Conv-BN module in BottleNecks, we employ the same importance function as defined in Equation B8. Consequently, applying the aforementioned method allows us to perform filter pruning on ResNet-56 and ResNet-50 using our proposed DDisp.

Appendix C Detailed Importance Function Used in Ablation Studies

Appendix C.1 Importance Function of D-Mean

To compare with the contribution-based importance, we introduce the D-Mean, taking the estimated expectation of mean value as the importance score. It can be formulated as follows:

$$\Phi(l^i, j)_{D-Mean} = \hat{\mu}_j^i \cdot \hat{\mathbb{E}}(A_j^i). \tag{C1}$$

As the D-Mean importance score is composed of the mean value of the intermediate feature and activation feature, this method is apparently based on the contribution perspective.

Appendix C.2 Importance Function of DDisp w/o Act

The DDisp w/o Act method ignores the dispersion of the activation features. The importance score is defined as:

$$\Phi(l^i, j)_{w/o\ Act} = \frac{(\hat{\sigma}_j^i)^2}{\hat{\mu}_j^i}. \tag{C2}$$

Appendix C.3 Importance Function of DDisp w/o Int

The DDisp w/o Int method ignores the dispersion of the activation features. The importance score is defined as:

$$\Phi(l^i, j)_{w/o\ Int} = \frac{\hat{\mathbb{V}}(A_j^i)}{\hat{\mathbb{E}}(A_j^i)}. \tag{C3}$$

Appendix D More Experimental Results and Details

Table D1 Detailed results of the ablation study on CIFAR-100 for different importance function and FLOPs reduction configurations. DDisp-real denote that using the real statistics that collected based on real features, which is the upper bound.

Model	Importance Function	Pruned Accuracy under Different FLOPs Speedup Ratio						
		1.5×	**3.0×**	**6.0×**	**8.0×**	**10×**	**12×**	**16×**
ResNet-56 (72.53)	Random	70.85 (-1.68)	67.84 (-4.69)	60.35 (-12.18)	54.88 (-17.65)	51.59 (-20.94)	46.54 (-25.99)	39.29 (-33.24)
	L1-Norm	72.41 (-0.12)	**69.95 (-2.58)**	59.16 (-13.37)	54.90 (-17.63)	53.29 (-19.24)	48.20 (-24.33)	44.38 (-28.15)
	Only-Scale Factor	71.99 (-0.54)	67.88 (-4.65)	58.32 (-14.21)	57.59 (-14.94)	53.27 (-19.26)	52.22 (-20.31)	48.42 (-24.11)
	Only-Running Mean	72.06 (-0.47)	68.39 (-4.14)	57.28 (-15.25)	46.36 (-26.17)	44.39 (-28.14)	35.71 (-36.82)	19.47 (-53.06)
	D-Mean	72.78 (+0.25)	69.13 (-3.40)	61.66 (-10.87)	56.04 (-16.49)	54.62 (-17.91)	51.59 (-20.94)	48.60 (-23.93)
	Only-Running Variance	72.52 (-0.01)	68.15 (-4.38)	60.45 (12.08)	57.97 (-14.56)	53.54 (-18.99)	52.28 (-20.25)	50.25 (-22.28)
	D-Variance	72.65 (+0.12)	68.13 (-4.40)	60.97 (-11.56)	57.99 (-14.54)	53.90 (-18.63)	52.33 (-20.20)	50.28 (-22.25)
	D-CV	72.71 (+0.18)	68.26 (-4.27)	61.31 (-11.22)	58.47 (-14.06)	56.58 (-15.95)	52.87 (-19.66)	51.40 (-21.13)
	DDisp w/o Act	72.81 (+0.28)	68.92 (-3.61)	63.16 (-9.37)	60.70 (-11.83)	56.32 (-16.21)	53.45 (-19.08)	51.84 (-20.69)
	DDisp w/o Int	**72.84 (+0.31)**	68.89 (-3.64)	63.08 (-9.45)	60.19 (-12.34)	56.83 (-15.70)	50.78 (-21.75)	47.65 (-24.88)
	DDisp	72.83 (+0.30)	69.12 (-3.41)	63.26 (-9.27)	60.80 (-11.73)	57.10 (-15.43)	54.92 (-17.61)	52.99 (-19.54)
	DDisp-real	72.30 (-0.23)	69.53 (-3.00)	**64.54 (-7.99)**	**62.65 (-9.88)**	**60.57 (-11.96)**	**57.70 (-14.83)**	**53.49 (-19.04)**

Appendix D.1 Ablation Study on Relative and Absolute Dispersion Scores

We conduct the ablation study on how to measure the dispersion score on ResNet-56. We introduce the D-Variance and D-CV to show the difference between utilizing only the variance, CV, and our dispersion score which is defined as the production of variance and CV. The importance function of D-Variance is defined as:

$$\Phi(l^i, j)_{var} = \hat{\mathbb{V}}(O_j^i)\hat{\mathbb{V}}(A_j^i). \tag{D1}$$

The importance function of D-CV is defined as:

$$\Phi(l^i, j)_{cv} = \frac{\sqrt{\hat{\mathbb{V}}(O_j^i)}}{\hat{\mathbb{E}}(O_j^i)} \frac{\sqrt{\hat{\mathbb{V}}(A_j^i)}}{\hat{\mathbb{E}}(A_j^i)}. \tag{D2}$$

According to the results presented in Table D1, while the performance for small FLOPs speedup ratios is comparable among D-Variance, D-CV, and DDisp, the performance improvements of DDisp over D-Variance and D-CV become pronounced for large FLOPs speedup ratios. These findings suggest that the comprehensive dispersion score of DDisp is more effective than the D-Variance, which only considers absolute variation, and the D-CV, which relies on relative variation alone.

Appendix D.2 Upper-bound Using the Real Statistics

To explore the potential of our proposed method, we employ real statistics to calculate the importance score on ResNet-56. This method is denoted as DDisp-real. To be specific, the importance score of the DDisp is formulated as:

$$\Phi(l^i, j)_{real} = \frac{\mathbb{V}(O_j^i)}{\mathbb{E}(O_j^i)} \frac{\mathbb{V}(A_j^i)}{\mathbb{E}(A_j^i)}. \tag{D3}$$

According to the comparison between DDisp-real and DDisp on ResNet-56 shown in Table D1, using the real statistics can further improve the accuracy performance, especially on high FLOPs Speedup Ratios (10x-16x). Moreover, this phenomenon indicates that better statistics can lead to higher accuracy performance. If better estimations were available, our method would approach the performance upper limit more closely. We hope that our work can attract more research interest in exploring a more accurate method to analyze and predict the feature distribution of a given well-trained network, which is the gate to more explainable deep learning approaches.

Appendix D.3 Ablation Study of Importance Criteria Only Based on BN Parameters

Our readers may concern that the performance improvement of DDisp is brought by the introduction of the BN parameters instead of our importance design and statistic estimation methods. Hence, we introduce three importance criteria Only-Scale Factor, Only-Running Mean, and Only-Running Variance to provide a comprehensive comparison between our DDisp and these criteria that only dependent on the BN parameters. The results of these criteria are shown in the Table D1. Comparing the performance of criteria with DDisp, we can conclude that DDisp gains significant performance advantage. Moreover, It is worth noting that although they perform well on 1.5x FLOPs Speedup Ratio, their performance on 8.0-16x FLOPs Speedup Ratios falls far behind the performance of DDisp. This demonstrates that using the running mean, running variance, or the scale factors without appropriate design and mathematical foundation can not achieve the competitive results of DDisp.

Appendix E Visualization of the Estimated and Real Statistics on ResNet-56

In this section, we show the detailed visualization of the comparison between our derived estimations and the real statistics on ResNet-56 [2]. As the largest channel number of ResNet-56 is 64, we do not average the channel to a certain number like what we did on VGG-16, all the estimations and real statistics are normalized in each layer for visualization. Visualization results in Figure E1-E9 demonstrate that, in terms of relative magnitudes, our estimations for the input features exhibit decent accuracy when compared to real values. Although the estimated values for the output features show minor deviations from the true values, the estimated final dispersion importance scores maintain a high level of consistency with those calculated using real statistics.

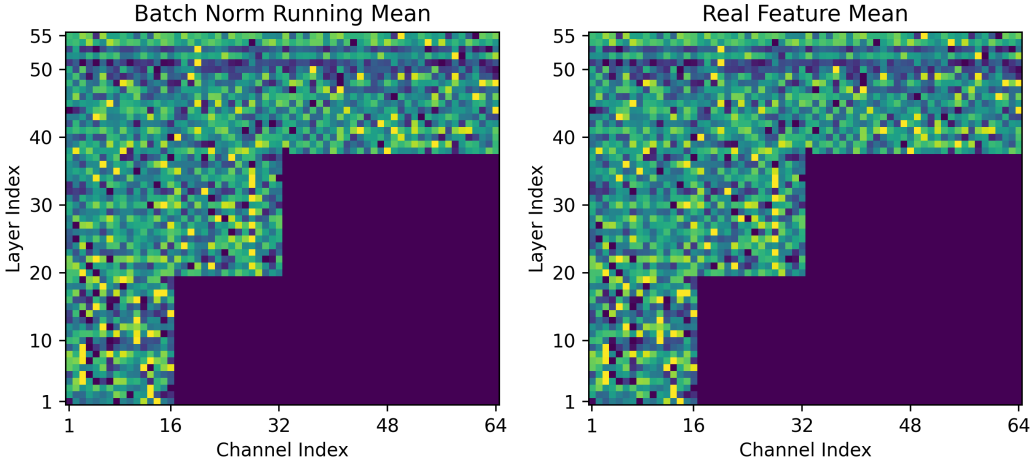


Figure E1 Visualization of our derived estimations towards the mean value and the real mean of the input feature distribution of BN layers on ResNet-56.

References

1 Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4510–4520

2 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778

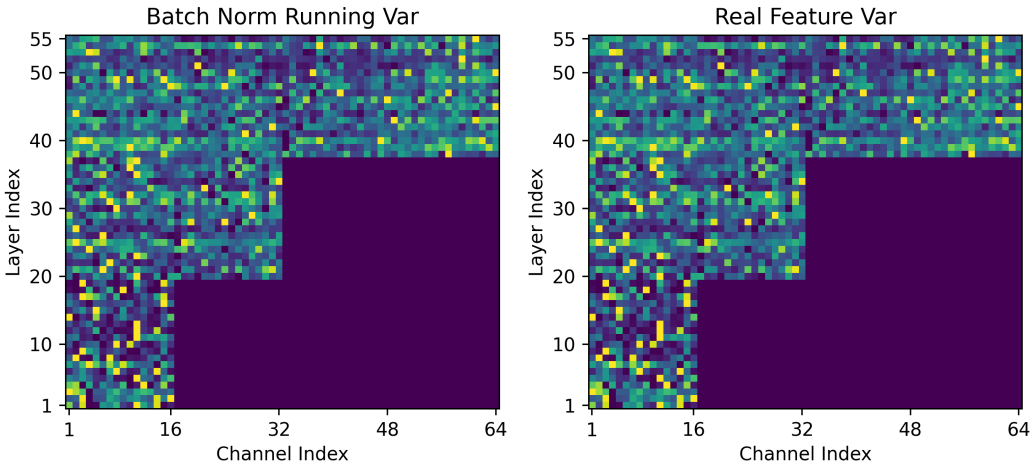


Figure E2 Visualization of our derived estimations towards the variance and the real variance of the input feature distribution of BN layers on ResNet-56.

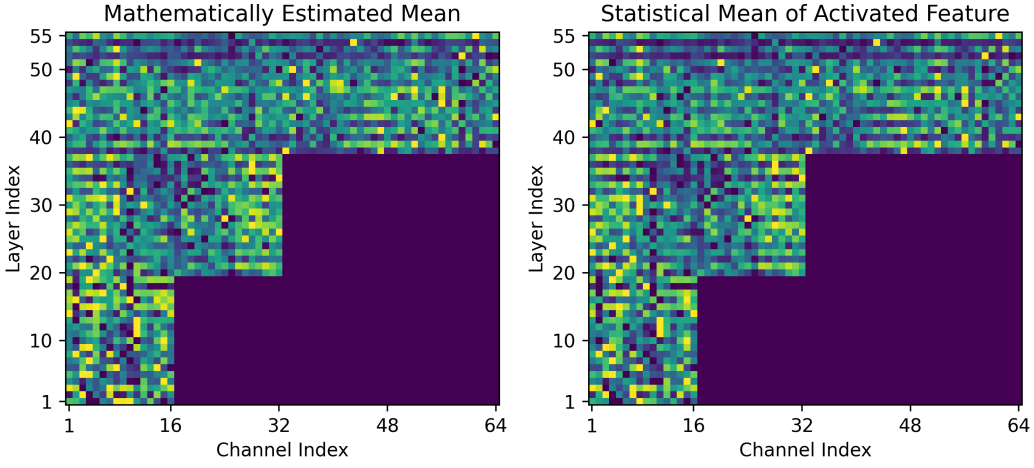


Figure E3 Visualization of our derived estimations towards the mean value and the real mean of the activation distribution on ResNet-56.

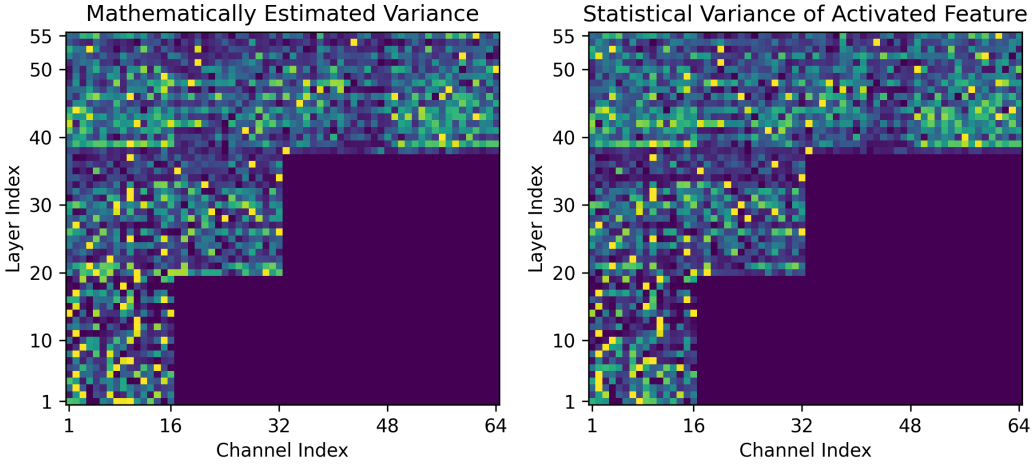


Figure E4 Visualization of our derived estimations towards the variance and the real variance of the activation distribution on ResNet-56.

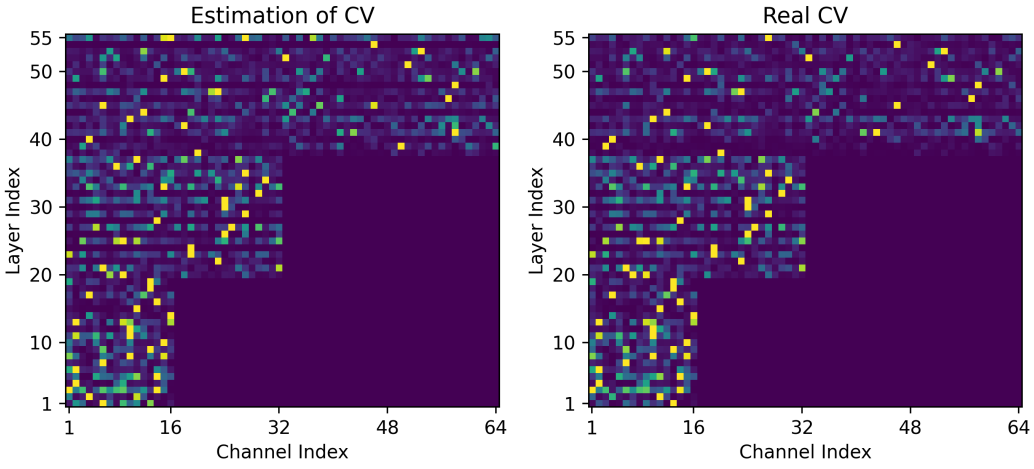


Figure E5 Visualization of our derived estimations towards the coefficient of variation (CV) and the real CV of the intermediate feature distribution of ResNet-56.

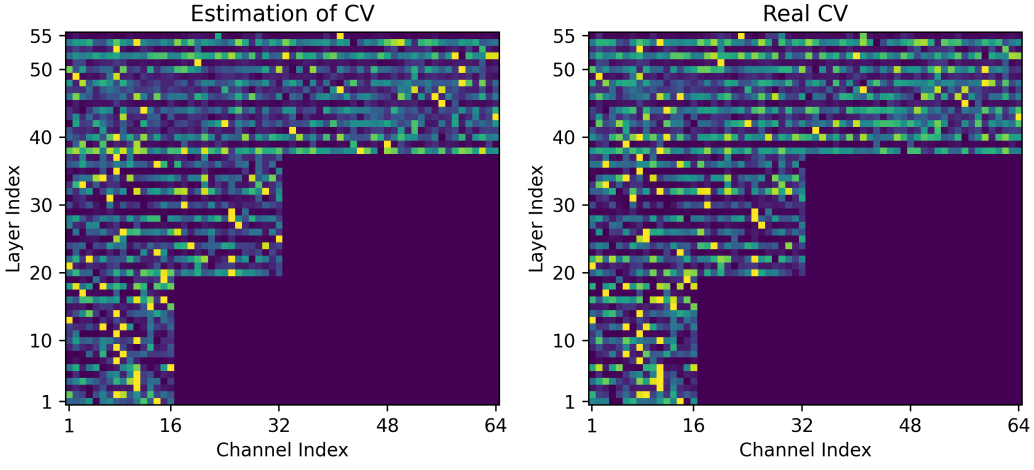


Figure E6 Visualization of our derived estimations towards the coefficient of variation (CV) and the real CV of the activation distribution of ResNet-56.

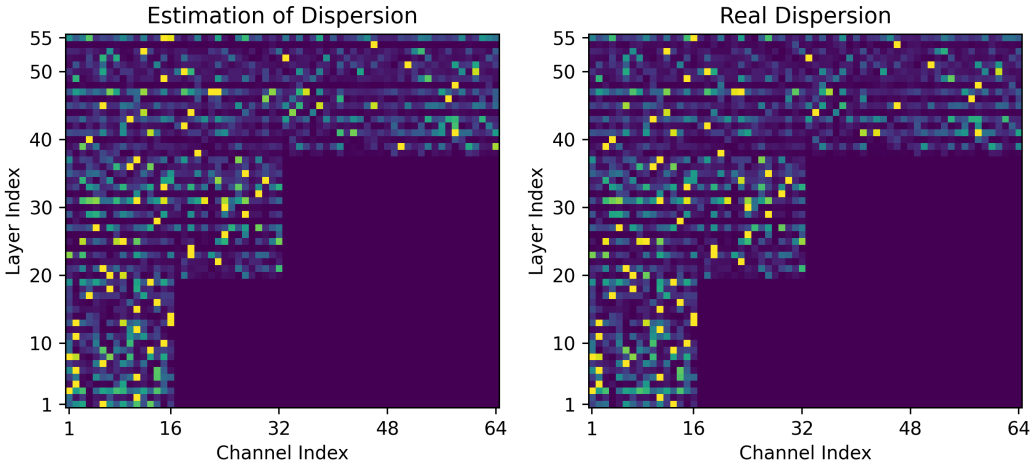


Figure E7 Visualization of the estimated and real Dispersion Score of the intermediate feature distribution of ResNet-56.

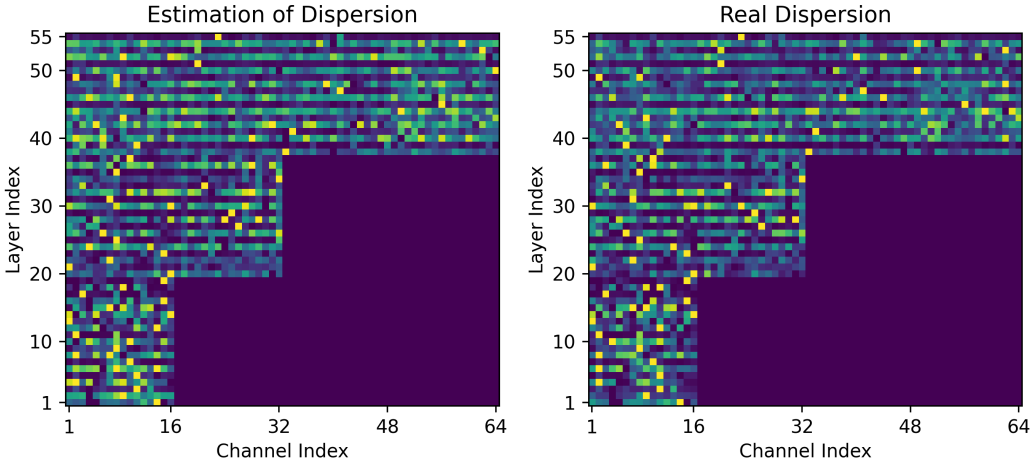


Figure E8 Visualization of the estimated and real Dispersion Score of the activation distribution of ResNet-56.

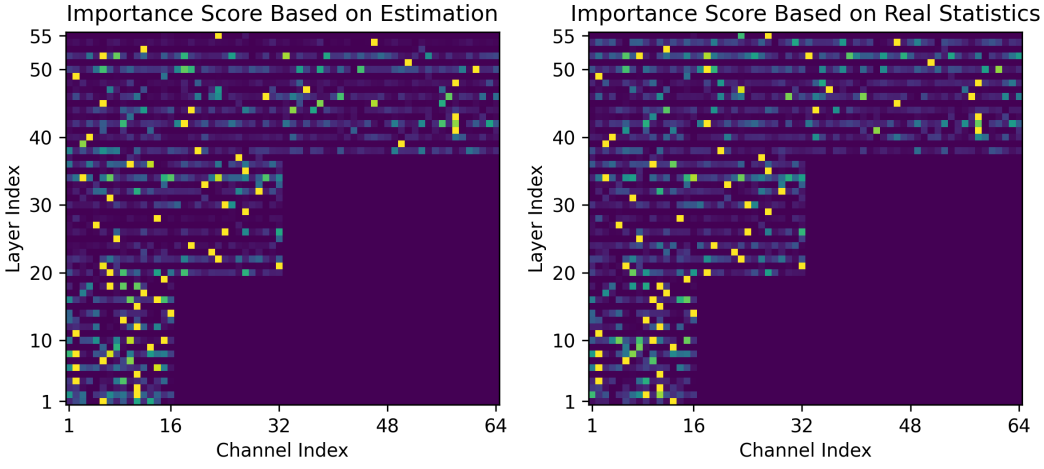


Figure E9 Visualization of the estimated and real Importance Score of the activation distribution of ResNet-56.