



HAP: Hessian-Aware Pruning and Optimal Neural Implant

Shixing Yu, Zhewei Yao, Amir Gholami, Zhen Dong, Sehoon Kim

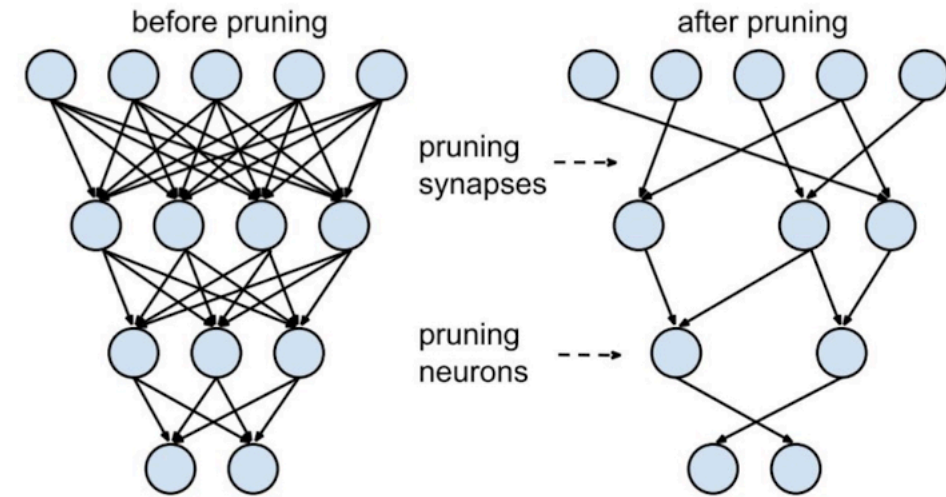
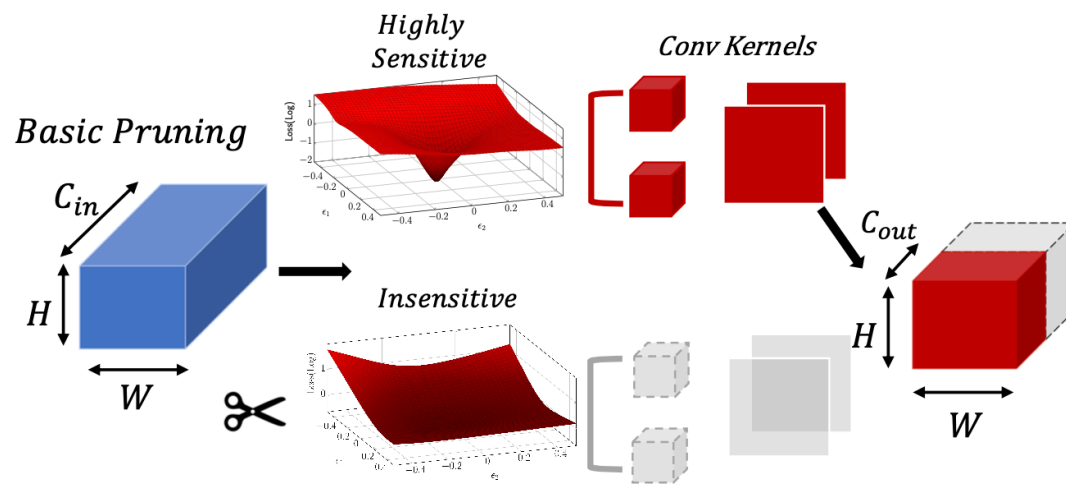
Michael Mahoney, Kurt Keutzer



Background

Two main stream:

- Structured pruning: channel wise
- Unstructured pruning: neuron level



Second-order based pruning

- Taylor expansion

$$L(w) = \frac{1}{N} \sum_{i=1}^N l(x_i, y_i, w)$$

$$\Delta L = L(w + \Delta w) - L(w)$$

$$\Delta L = \frac{1}{2} \Delta w^T H \Delta w$$


PSD

Hessian-aware pruning

- With Taylor Expansion

$$\begin{aligned} \min_{\Delta w} \frac{1}{2} \Delta w^T H \Delta w &= \frac{1}{2} \begin{pmatrix} \Delta w_p \\ \Delta w_l \end{pmatrix}^T \begin{pmatrix} H_{p,p} & H_{p,l} \\ H_{l,p} & H_{l,l} \end{pmatrix} \begin{pmatrix} \Delta w_p \\ \Delta w_l \end{pmatrix} \\ \text{s.t. } \Delta w_p + w_p &= 0. \end{aligned}$$

Hessian-aware pruning

- Lagrange Multiplier

$$\mathcal{L} = \frac{1}{2} \Delta w^T H \Delta w + \lambda^T (\Delta w_p + w_p),$$



$$H_{p,p} \Delta w_p + H_{p,l} \Delta w_l + \lambda = 0$$

$$H_{l,p} \Delta w_p + H_{l,l} \Delta w_l = 0$$

Hessian-aware pruning

- Equations: $\Delta w_p + w_p = 0$ $\Delta w_l = -H_{l,l}^{-1} H_{l,p} w_p$
- Linear algebra

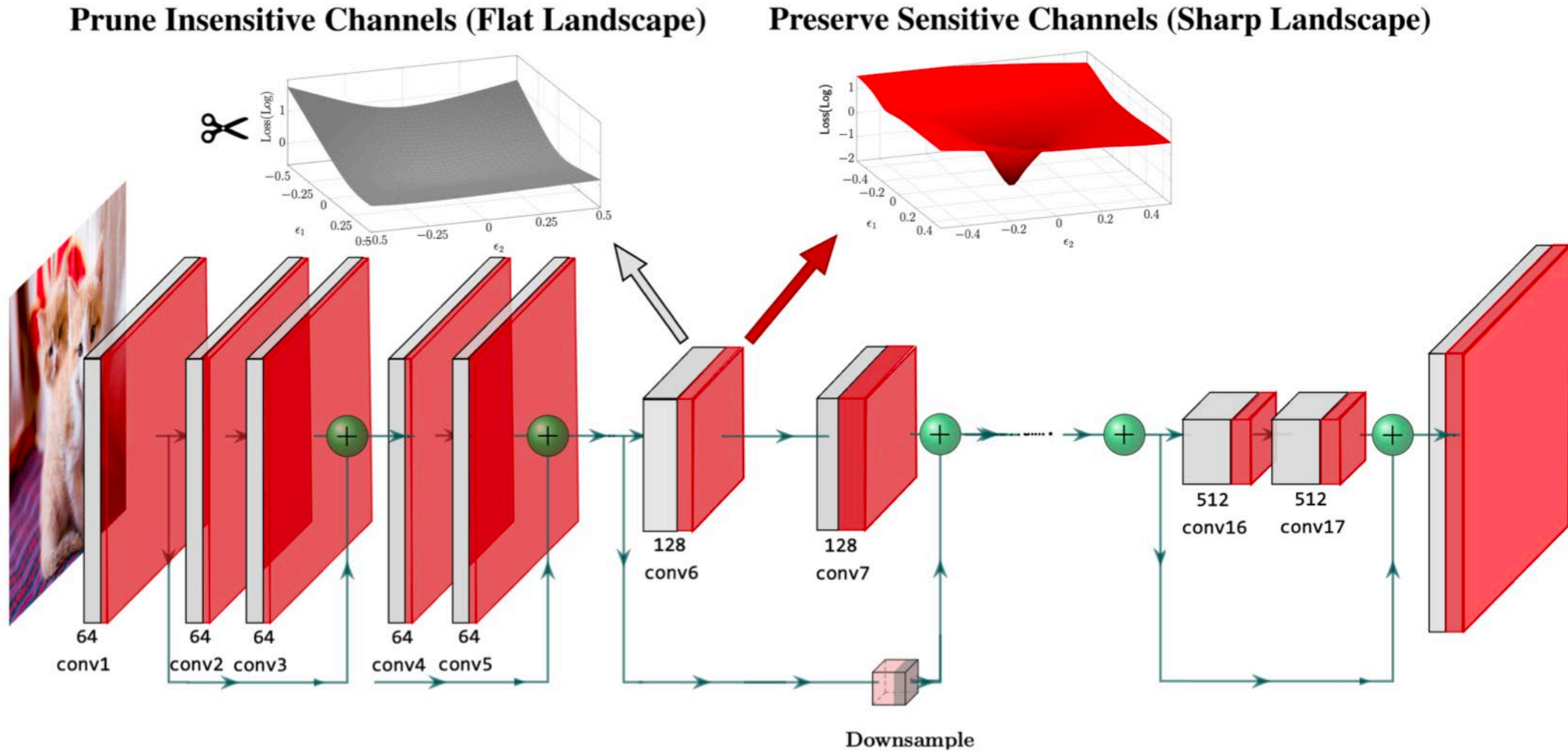
$$\begin{aligned}\frac{1}{2} \Delta w^T H \Delta w &= \frac{1}{2} \begin{pmatrix} \Delta w_p \\ \Delta w_l \end{pmatrix}^T \begin{pmatrix} H_{p,p} & H_{p,l} \\ H_{l,p} & H_{l,l} \end{pmatrix} \begin{pmatrix} \Delta w_p \\ \Delta w_l \end{pmatrix} \\ &= \frac{1}{2} (\Delta w_p^T H_{p,p} \Delta w_p + 2 \Delta w_p^T H_{p,l} \Delta w_l \\ &\quad + \Delta w_l^T H_{l,l} \Delta w_l) \\ &= \frac{1}{2} w_p^T (H_{p,p} - H_{p,l} H_{l,l}^{-1} H_{l,p}) w_p.\end{aligned}$$

Disadvantages for OBS

- ☹️ Computation of Hessian Inverse
- ☹️ Per-parameter sensitivity measurement
- ☹️ Unstructured – hardware unfriendly

$$\frac{1}{2} \frac{w_p^2}{H_{p,p}^{-1}}$$

Methodology



OBD to solve

- Computational cost of Hessian Inverse

$$\begin{aligned}\frac{1}{2}\Delta w^T H \Delta w &= \frac{1}{2}w_p^T (H_{p,p} - H_{p,l}H_{l,l}^{-1}H_{l,p})w_p \\ &\approx \frac{1}{2}w_p^T \text{Diag}(H_{p,p})w_p,\end{aligned}$$

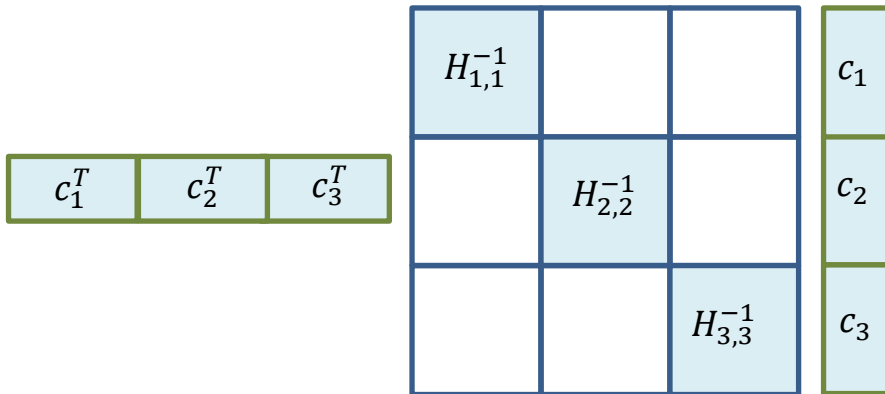
To solve

- Computational cost of Hessian Inverse
- 😊 Group the parameters
- Unstructured – hardware unfriendly

c_1^T	c_2^T	c_3^T	$H_{1,1}^{-1}$	$H_{1,2}^{-1}$	$H_{1,3}^{-1}$	c_1
			$H_{2,1}^{-1}$	$H_{2,2}^{-1}$	$H_{2,3}^{-1}$	c_2
			$H_{3,1}^{-1}$	$H_{3,2}^{-1}$	$H_{3,3}^{-1}$	c_3

To solve

- Cross-channel independent
- 😊 Group the parameters
- Unstructured – hardware unfriendly



$$\begin{aligned}\frac{1}{2}\Delta w^T H \Delta w &= \frac{1}{2}w_p^T (H_{p,p} - H_{p,l}H_{l,l}^{-1}H_{l,p})w_p. \\ &= \frac{1}{2}w_p^T H_{p,p}w_p.\end{aligned}$$

Methodology

To solve

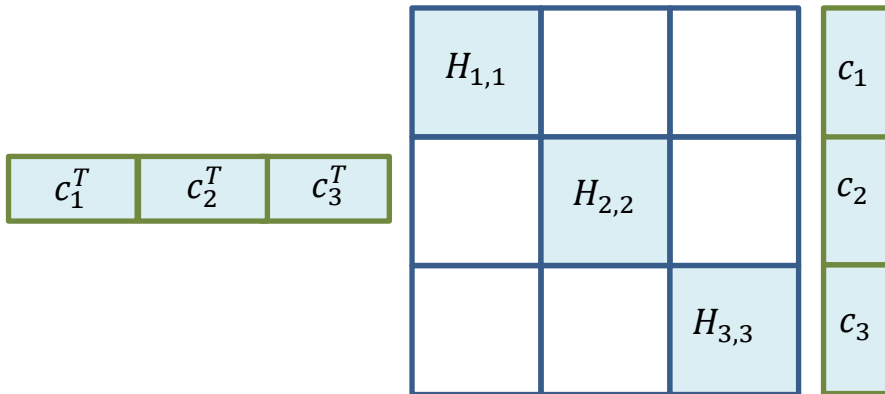


Cross-channel independent



Group the parameters

- Unstructured – hardware unfriendly



$$\begin{aligned}\frac{1}{2}\Delta w^T H \Delta w &= \frac{1}{2}w_p^T (H_{p,p} - H_{p,l}H_{l,l}^{-1}H_{l,p})w_p. \\ &= \frac{1}{2}w_p^T H_{p,p}w_p.\end{aligned}$$

Methodology

To solve



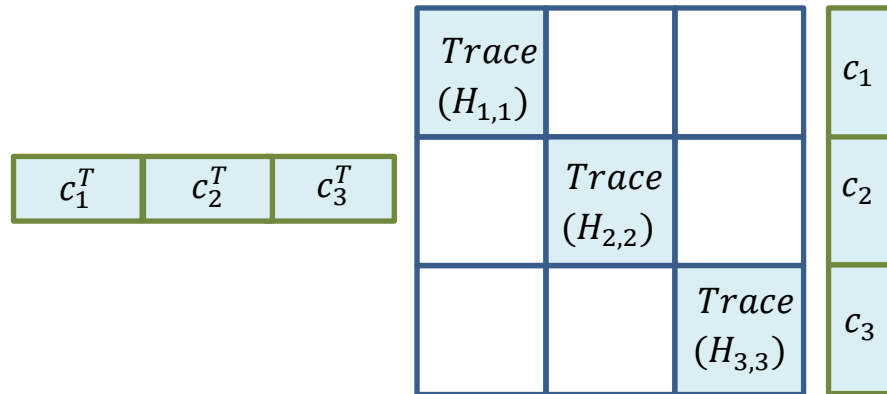
Cross-channel independent



Group the parameters



Structured pruning – Channel wise



$$\begin{aligned}\frac{1}{2}\Delta w^T H \Delta w &= \frac{1}{2}w_p^T (H_{p,p} - H_{p,l}H_{l,l}^{-1}H_{l,p})w_p. \\ &= \frac{1}{2}w_p^T H_{p,p}w_p. \\ &\approx \frac{1}{2}w_p^T \frac{\text{Trace}(H_{p,p})}{p}w_p \\ &= \frac{\text{Trace}(H_{p,p})}{2p} \|w_p\|_2^2,\end{aligned}$$

Hutchinson

$$\text{Tr}(H) = \mathbb{E}[v^T H v]$$

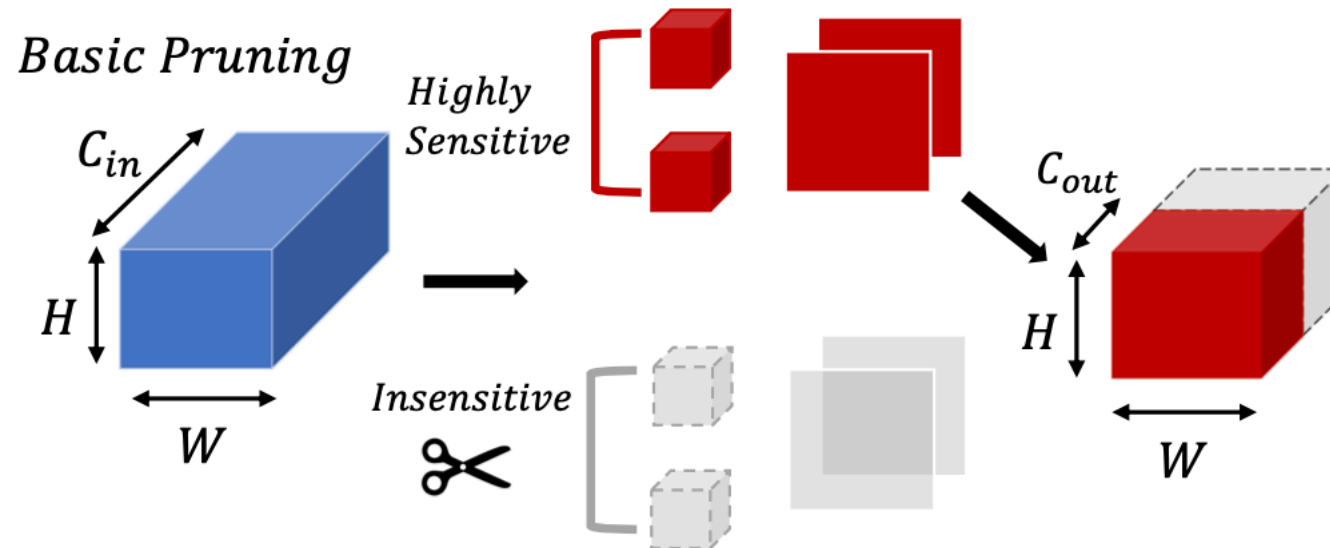
Methodology

HAP Results on CIFAR10: ResNet-56

Method	Baseline acc.	Final acc.	Acc. ↓	FLOPs (%)
GAL-0.8[38]	93.26	90.36	2.90	39.8
HRank[36]	93.26	90.72	2.54	25.9
HAP	93.88	91.57	2.31	21.0

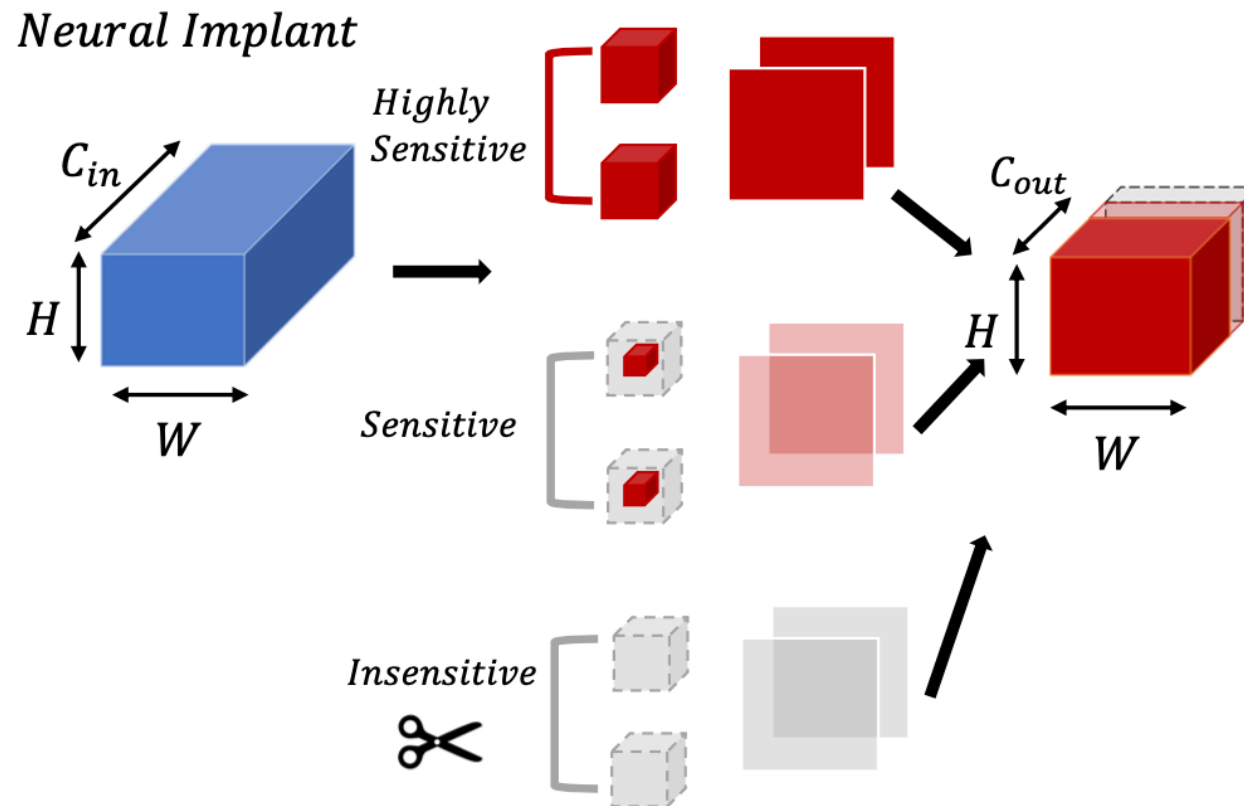
Hessian-aware Neural Implant

- Performance loss for HAP



Hessian-aware Neural Implant

- Performance loss for HAP



Experimental Results

HAP + Implant Results on CIFAR10: ResNet56

Table 4: ResNet56 on CIFAR10.

Method	Base-acc.	Final-acc.	Acc. ↓	FLOPs (%)
CP[2]	92.80	91.80	1.00	50.0
AMC[7]	92.80	91.90	0.90	50.0
FPGM[12]	93.59	93.26	0.33	47.4
LFPC[13]	93.59	93.24	0.35	47.1
HAP+IMPLANT	93.88	93.55	0.33	40.7
GAL-0.8[3]	93.26	90.36	2.90	39.8
HRank[5]	93.26	90.72	2.54	25.9
HAP	93.88	91.57	2.31	21.0
HAP+IMPLANT	93.88	92.92	0.96	23.9

Summary

- Hessian-aware pruning surpasses another second-order based method and other SOTA methods on CIFAR-10 and ImageNet.
- Neural Implant preserves the information loss caused by pruning out the sensitive layers.