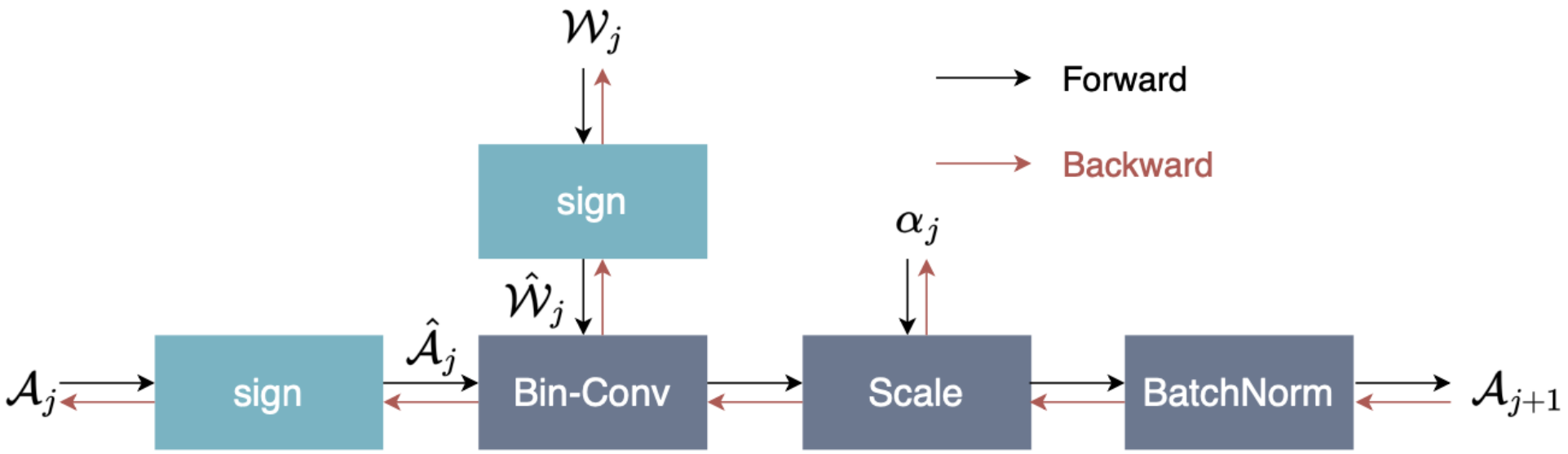
OvSW: Overcoming Silent Weights for Accurate Binary Neural Networks

Jingyang Xiang[1], Zuohui Chen[2], Siqi Li[1],

Qing Wu[3], Yong Liu[1]

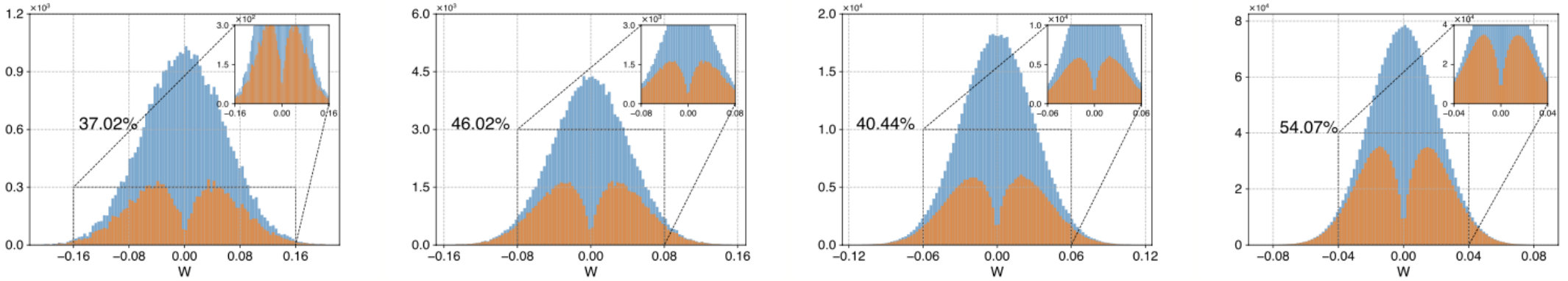
APRIL Lab, Zhejiang University, Hangzhou, China[1]
IVSN, Zhejiang University of Technology, Hangzhou, China[2]
College of Computer Science, Hangzhou Dianzi University[3]
https://github.com/JingyangXiang/OvSW

Overview: Binary Neural Network



Silent Weights

Histogram of the initialized weight distribution (blue) and the weights that never update signs throughout training (orange) for Vanilla BNNs. 37.02%, 46.02%, 40.44% and 54.07% represent the ratio of the corresponding orange area to the blue.



Parameter flip from vanilla BNNs. Massive weights don't flip!

Theoretical Analysis

$$\hat{\mathcal{W}}_j = \hat{\mathcal{W}}'_j, \alpha_j = \alpha'_j, \mathrm{BN}_j = \mathrm{BN}'_j, \forall j, \Longrightarrow \mathcal{A}_{j+1} = \mathrm{BN}\left(\left(\hat{\mathcal{A}}_j \circledast \hat{\mathcal{W}}_j\right) \odot \alpha_j\right) = \mathrm{BN}'\left(\left(\hat{\mathcal{A}}_j \circledast \hat{\mathcal{W}}'_j\right) \odot \alpha'_j\right) = \mathcal{A}'_{j+1}. \tag{1}$$

$$\frac{\partial \mathcal{L}}{\partial \mathcal{W}_j} = \frac{\partial \mathcal{L}}{\partial \mathcal{A}_{j+1}} \frac{\partial \mathcal{A}_{j+1}}{\partial \hat{\mathcal{W}}_j} \frac{\partial \hat{\mathcal{W}}_j}{\partial \mathcal{W}_j} = \frac{\partial \mathcal{L}'}{\partial \mathcal{A}'_{j+1}} \frac{\partial \mathcal{A}'_{j+1}}{\partial \hat{\mathcal{W}}'_j} \frac{\hat{\mathcal{W}}'_j}{\partial \mathcal{W}'_j} = \frac{\partial \mathcal{L}'}{\partial \mathcal{W}'_j}. \tag{2}$$

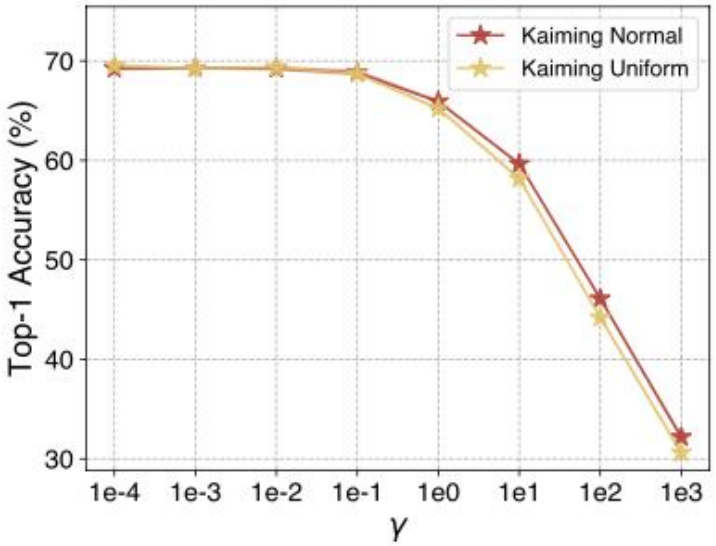
$$\mathcal{W}'_j(t+1) = \mathcal{W}'_j(t) - \beta(t) \frac{\partial \mathcal{L}'(t)}{\partial \mathcal{W}'_j(t)} = \gamma \mathcal{W}_j(t) - \beta(t) \frac{\partial \mathcal{L}(t)}{\partial \mathcal{W}_j(t)} = (\gamma - 1)\mathcal{W}_j(t) + \mathcal{W}_j(t+1), \tag{3}$$

$$\lim_{\gamma \to \infty} \mathcal{W}'_j(t+1) = \lim_{\gamma \to \infty} [(\gamma - 1)\mathcal{W}_j(t) + \mathcal{W}_j(t+1)] = (\gamma - 1)\mathcal{W}_j(t). \tag{4}$$

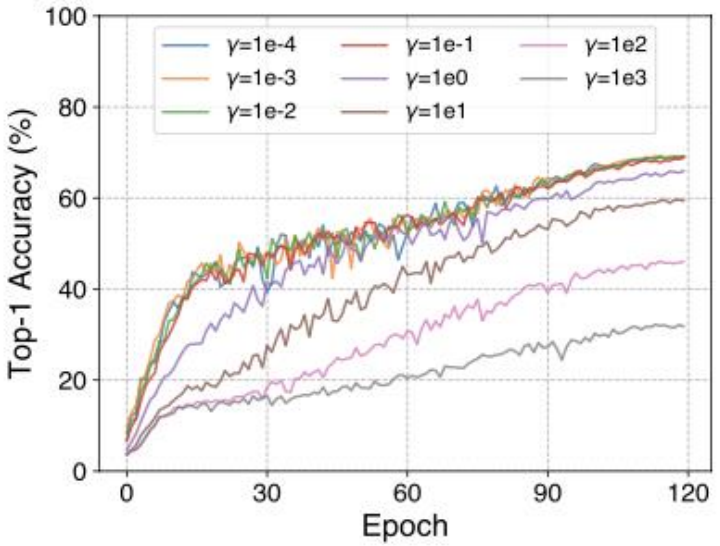
the independence of the BNNs gradient from the latent weight distribution

Validation

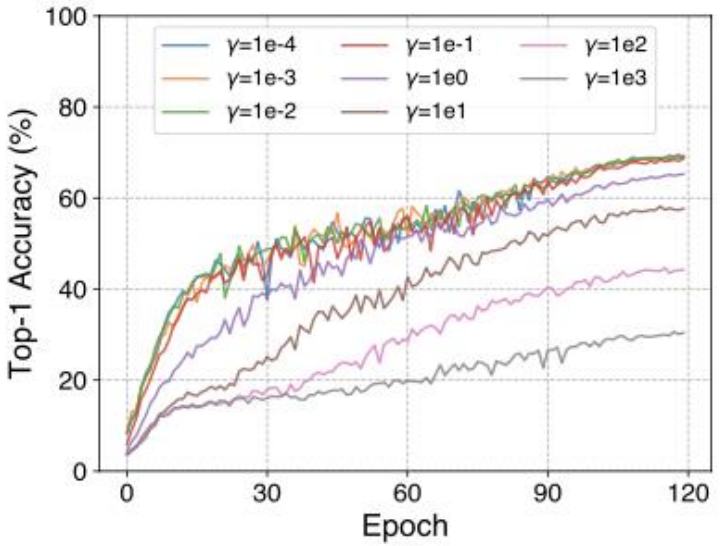
Accuracy & Distribution.



Convergence & Kaiming Normal.



Convergence & Kaiming Uniform.



$$\mathcal{W}_j \sim \text{Normal}\left(0, \lambda^2 \text{std}^2\right)$$

$$\mathcal{W}_j \sim \text{Uniform}\left(-\lambda\text{bound}, \lambda\text{bound}\right),$$

Overcome Silent Weights

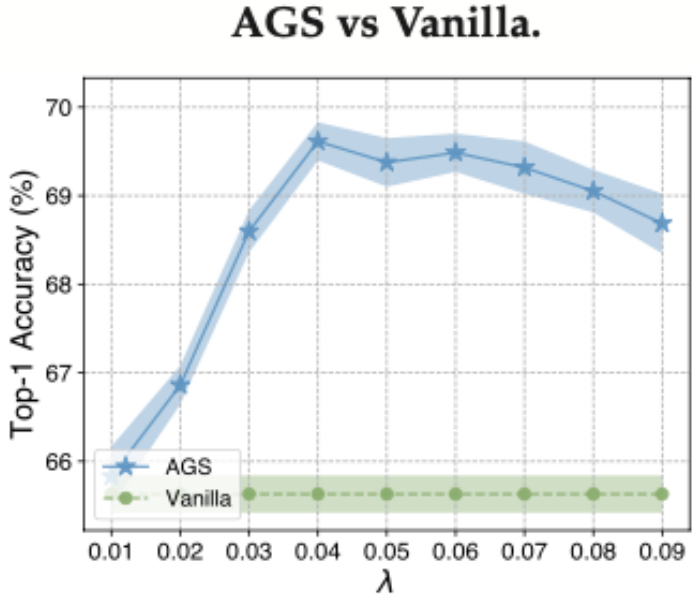
- Adaptive Gradient Scaling (AGS)
- Silence Awareness Decaying (SAD)

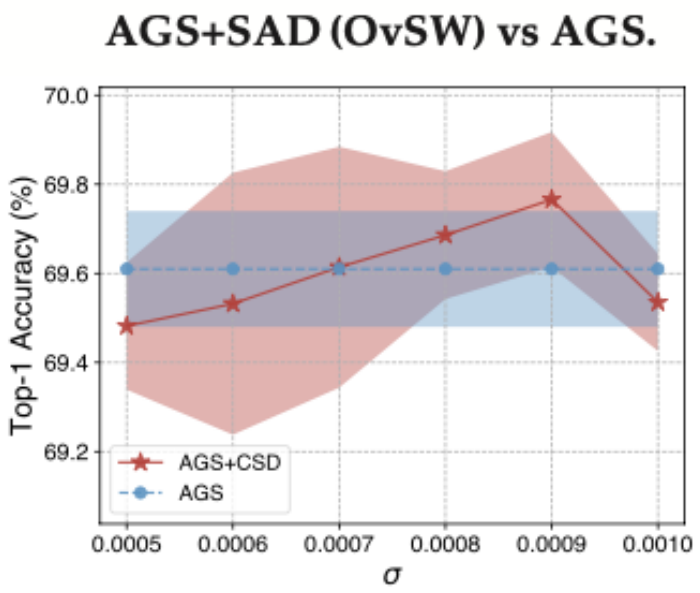
$$\textbf{AGS} \quad \overline{\mathcal{G}}_j^{k,l,m,n} = \begin{cases} \lambda \frac{\|\mathcal{W}_j^k\|_F}{\|\mathcal{G}_j^k\|_F} \mathcal{G}_j^{k,l,m,n} & \text{if } \frac{\|\mathcal{G}_j^k\|_F}{\|\mathcal{W}_j^k\|_F} < \lambda, \\ \mathcal{G}_j^{k,l,m,n} & \text{otherwise.} \end{cases} \qquad \|\mathcal{W}_j^k\|_F = \sqrt{\sum_{l=1}^{C_{\text{in}}^j} \sum_{m=1}^{K_{\text{h}}^j} \sum_{n=1}^{K_{\text{w}}^j} \mathcal{W}_j^{k,l,m,n}}.$$

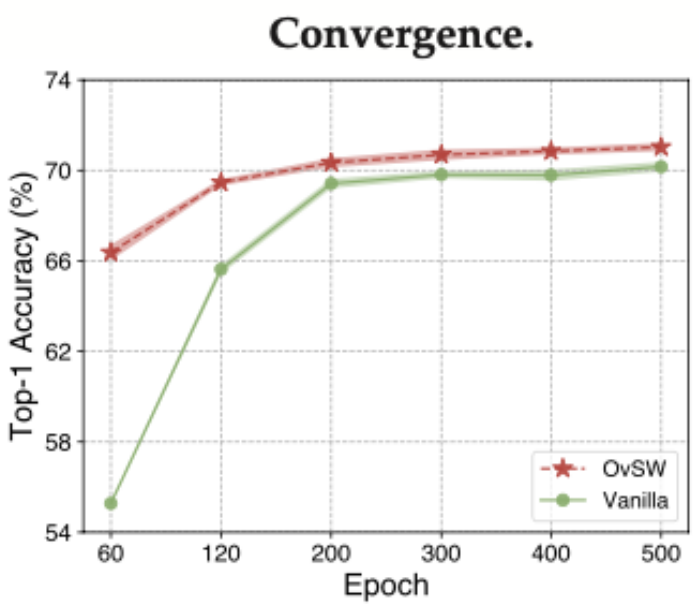
$$\textbf{SAD} \quad \mathcal{S}_j(t) = m \cdot \mathcal{S}_j(t-1) + (1-m) \cdot \frac{\left|\text{sign}\left(\mathcal{W}_j(t)\right) - \text{sign}\left(\mathcal{W}_j(t-1)\right)\right|_{abs}}{2},$$

$$\overline{\overline{\mathcal{G}}}_j^{k,l,m,n}(t) = \begin{cases} \overline{\mathcal{G}}_j^{k,l,m,n}(t) + \gamma \mathcal{W}_j^{k,l,m,n}(t), & \text{if } \mathcal{S}_j^{k,l,m,n}(t) < \sigma, \\ \overline{\mathcal{G}}_j^{k,l,m,n}(t), & \text{otherwise,} \end{cases}$$

Experiments





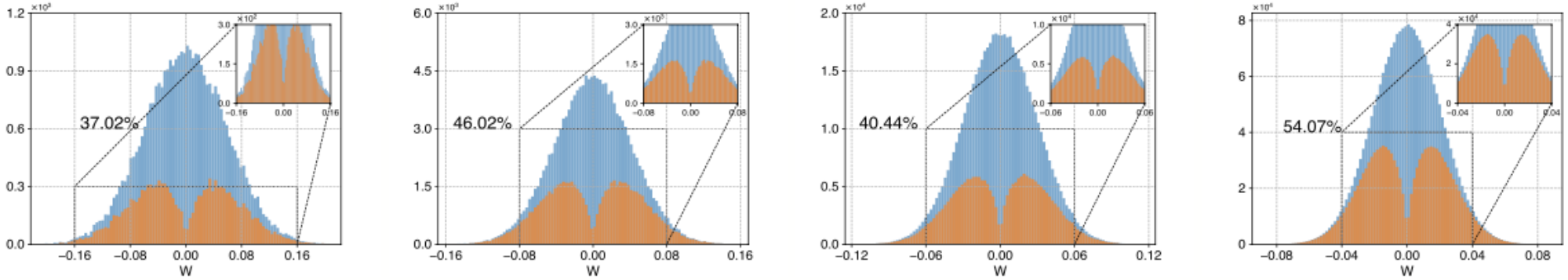


ImageNet

Model	Method	Bit-width (W/A)	Top-1 Acc.(%)	Top-5 Acc.(%)
ResNet18	Full-precision	32/32	69.6	89.2
	XNOR [42]	1/1	51.2	73.2
	BiReal [34]	1/1	56.4	79.5
	IR-Net [40]	1/1	58.1	80.0
	RBNN [30]	1/1	59.9	81.9
	SiMaN [29]	1/1	60.1	82.3
	FDA-BNN [52]	1/1	60.2	82.3
	ReCU [53]	1/1	61.0	82.6
	OvSW (Ours)	1/1	**61.6**	**83.1**
	ReActNet [33]	1/1	65.9	86.1
	ReCU [53]	1/1	66.4	86.5
	OvSW* (Ours)	1/1	**66.6**	**86.7**
ResNet34	Full-precision	32/32	73.3	91.3
	XNOR++ [3]	1/1	57.1	79.9
	BiReal [34]	1/1	62.2	83.9
	IR-Net [40]	1/1	62.9	84.1
	RBNN [30]	1/1	63.1	84.4
	SiMaN [29]	1/1	63.9	84.8
	CMIM [45]	1/1	65.0	85.7
	ReCU [53]	1/1	65.1	85.8
	OvSW (Ours)	1/1	**65.5**	**86.1**

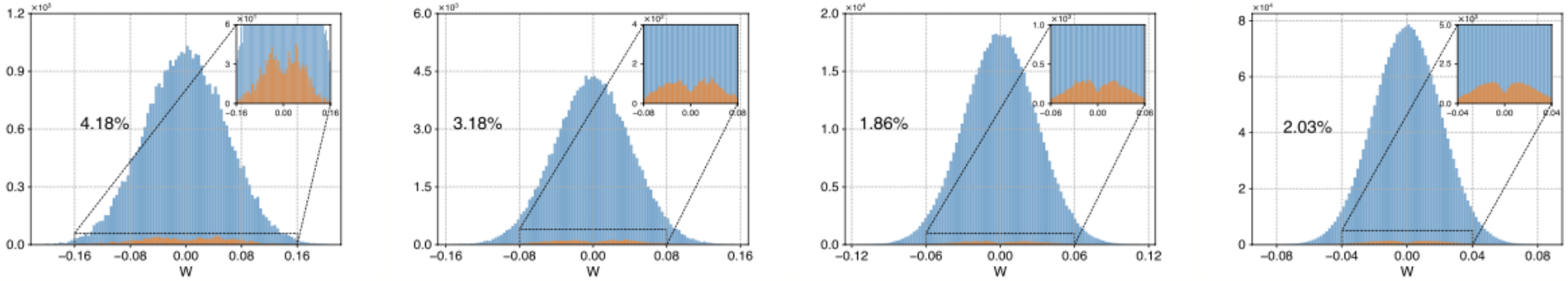
Flip Efficiency

Histogram of the initialized weight distribution (blue) and the weights that never update signs throughout training (orange) for Vanilla BNNs. 37.02%, 46.02%, 40.44% and 54.07% represent the ratio of the corresponding orange area to the blue.



Parameter flip from vanilla BNNs. Massive weights don't flip!

Histogram of the initialized weight distribution (blue) and the weights that never update signs throughout training (orange) for OvSW. 4.18%, 3.18%, 1.86% and 2.03% represent the ratio of the corresponding orange area to the blue.



OvSW successfully flip silent weights!

OvSW: Overcoming Silent Weights for Accurate Binary Neural Networks

https://github.com/JingyangXiang/OvSW

Thank you!