Robust Pruning at Initialization

S. Hayou, J.F. Ton, A. Doucet, Y.W. Teh

Department of Statistics, University of Oxford

Overparameterized Models

- Millions/Billions of parameters.
- Can we reduce the size of these models without major drop in the performance?
- Pruning after Training: train, prune, repeat...
  Very slow and requires excessive computational power.
- Pruning at Initialization?

Neural Networks Pruning

- Pruning: apply a binary mask δ to the weights. The pruned model is given by

$$y^l(x) = \mathcal{F}_l(\delta^l \circ W^l, y^{l-1}(x)) + B^l$$

- Sensitivity based pruning (SNIP, Lee et al. 2018): prune the weights at initialization based on $|W \frac{\partial \mathcal{L}}{\partial W}|$. Inspired from

$$\mathcal{L}_W \approx \mathcal{L}_{W=0} + W \frac{\partial \mathcal{L}}{\partial W}$$

Ordered, Chaotic, and EOC Initializations

Assume $W_{ij}^l \sim \mathcal{N}(0, \sigma_w^2/N_{l-1})$, $B_i^l \sim \mathcal{N}(0, \sigma_b^2)$.

- $q^l(x) = \mathrm{var}(y_1^l(x)) \overset{l\to\infty}{\to} q$
- $C_l(x, x') = \mathrm{corr}(y_1^l(x), y_1^l(x')) \overset{l\to\infty}{\to} ??$

Depending on the choice of (σ_b, σ_w):

- **Ordered phase** where $C_l(x, x') \to 1$ exponentially quickly [Schoenholz et al., 2017]
- **Chaotic phase** where $C_l(x, x') \to c < 1$ exponentially quickly [Schoenholz et al., 2017]
- **Edge of Chaos (EOC)** where $C_l(x, x') \to 1$ polynomial rate [Hayou et al., 2019]

Sensitivity Based Pruning (SBP)

• Critical sparsity: sparsity level s_{cr} such that one layer at least is fully pruned. s_{cr} is random.

Proposition (Initialization is crucial for SBP, Informal)

Assume $W^l \in \mathbb{R}^{N \times N}$, and let L be the depth.

- *If $(\sigma_b, \sigma_w) \in$ Ordered phase*

$$\mathbb{E}[s_{cr}] = \mathcal{O}\left(\frac{\log(LN^2)}{L} + \frac{1}{\sqrt{LN^2}}\right)$$

- *$(\sigma_b, \sigma_w) \in$ EOC, then the upper bound no longer holds.*

• On the Ordered phase, $\lim_{L\to\infty} \mathbb{E}[s_{cr}] = 0$.
• Similar results can be proven for the Chaotic phase.

Sensitivity Based Pruning (SBP)

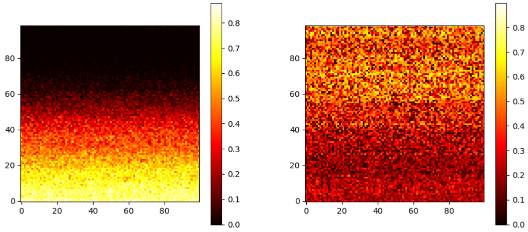


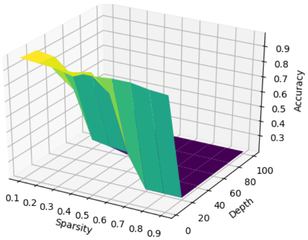
Figure: Percentage of weights kept after SBP. 100x100 FFNN, $s = 70\%$, Chaotic phase(left), EOC(right).

Training the Sparse Network

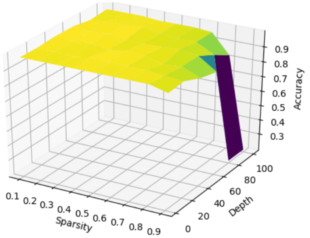
- After pruning, it might be difficult to train the sparse network...
- Putting the pruned network back on the EOC

$$y^l(x) = \rho_l \mathcal{F}_l(\delta^l \circ W^l, y^{l-1}(x)) + B^l$$

Training the Sparse Architecture



Init Ordered Phase



Init EOC + ReScaling

Training the Sparse Architecture

• Our algorithm SBP-SR yields SOTA (one shot pruning algorithms) performance for Deep ResNets.

Table: Classification accuracies on Tiny ImageNet for Resnet with varying depths

	ALGORITHM	85%	90%	95%
RESNET32	SBP-SR	**57.25 ± 0.09**	**55.67 ± 0.21**	50.63±0.21
	SNIP	56.92± 0.33	54.99±0.37	49.48±0.48
	GRASP	**57.25±0.11**	55.53±0.11	**51.34±0.29**
RESNET50	SBP-SR	**59.8±0.18**	**57.74±0.06**	**53.97±0.27**
	SNIP	58.91±0.23	56.15±0.31	51.19±0.47
	GRASP	58.46±0.29	57.48±0.35	52.5±0.41
RESNET104	SBP-SR	**62.84±0.13**	**61.96±0.11**	**57.9±0.31**
	SNIP	59.94±0.34	58.14±0.28	54.9±0.42
	GRASP	61.1±0.41	60.14±0.38	56.36±0.51

Paper

For more details, check our paper

Robust Pruning at Initialization. ICLR 2021. *S. Hayou, J.F. Ton, A. Doucet, Y.W. Teh*.

References

S.S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. *5th International Conference on Learning Representations*, 2017.

S. Hayou, A. Doucet, and J. Rousseau. On the impact of the activation function on deep neural networks training. *ICML*, 2019.