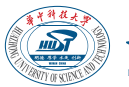
How Sparse Can We Prune A Deep Network: A Fundamental Limit Perspective

Qiaozhe Zhang, Ruijie Zhang, Jun Sun, Yingzhuang Liu

School of Electronic Information and Communication
Huazhong University of Science and Technology

November 13, 2024



華中科技大學
HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Motivation

- Deep neural networks (DNNs) rely heavily on the over-parametrization.
- **Network pruning** is commonly used to alleviate the storage and computational burden of DNNs.
- A **systematic** exploration of the **fundamental limits** of network pruning remains lacking.

Question

What's the fundamental limit of network pruning?

Answer (informal)

The limit of pruning ratio(ρ) approximates:

$$1 - \frac{1}{D}\sum_{i=1}^{D} \frac{r_i^2}{\|\mathbf{w}^r\|_2^2 + r_i^2}, \quad r_i = \sqrt{2\epsilon/\lambda_i}$$

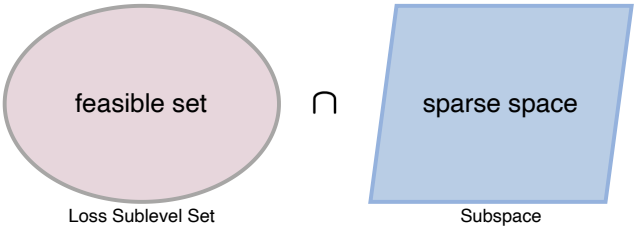
$\mathbf{w}^r$: removed weights; λ_i: sharpness of loss landscape.

Formulation

subspace feasible set

- **Pruning Objective:** $\min \|\mathbf{w}\|_0$ s.t. $\mathcal{L}(\mathbf{w}) \leq \mathcal{L}(\mathbf{w}^*) + \epsilon.$
  - $\mathbf{w} \in \mathbb{R}^D$ is the pruned weight and $\mathbf{w}^* \in \mathbb{R}^D$ is the original one.
  - sparse network: $\mathbf{w} = \mathbf{w}^* \odot \mathbf{m}$, where $\odot$ is element-wise multiplication and $\mathbf{m} \in \mathbb{R}^D$ only contains 0 and 1.
- **Pruning Ratio:** the proportion of the nonzeros weights $\rho = \|\mathbf{m}\|_0 / D$.

An Intuition



network pruning ⇒ intersection of sets!

Theorem (The Approximate Kinematics Formula)

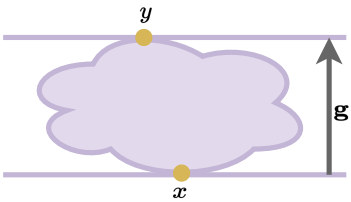
Let $\mathcal{C}$ be a convex conic hull of set S in $\mathbb{R}^D$, $\delta(\mathcal{C})$ donates the statistical dimension of $\mathcal{C}$, draw a random orthogonal basis $\mathbf{Q} \in \mathbb{R}^{D\times D}$. For a k-dimensional subspace S_k, it holds that:

$$\delta(\mathcal{C}) + k \lesssim D \Rightarrow \mathbb{P}\{\mathcal{C} \cap \mathbf{Q}S_k = \emptyset\} \approx 1$$
$$\delta(\mathcal{C}) + k \gtrsim D \Rightarrow \mathbb{P}\{\mathcal{C} \cap \mathbf{Q}S_k = \emptyset\} \approx 0$$

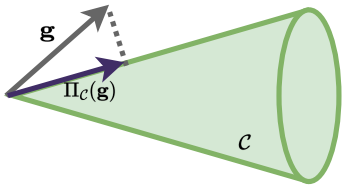
Key: high dim. ⇒ sharp phase transition!

Dimension of Sets

How to estimate the **dimension (complexity)** of the sets?



(a) gaussian width



(b) statistical dimension

$$\frac{1}{2}\mathbb{E}\sup_{\mathbf{x},\mathbf{y}\in S}\langle \mathbf{g}, \mathbf{x}-\mathbf{y}\rangle, \mathbf{g}\sim\mathcal{N}(\mathbf{0}, \mathbf{I}_{D\times D})$$

$$\mathbb{E}[\|\Pi_{\mathcal{C}}(\mathbf{g})\|_2^2], \mathbf{g}\sim\mathcal{N}(\mathbf{0}, \mathbf{I}_{D\times D})$$

statistical dimension = squared gaussian width (easier).

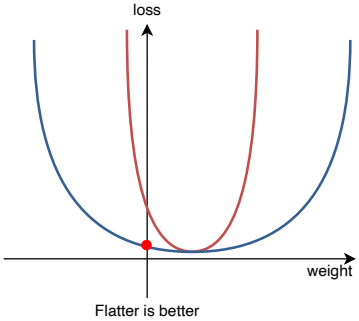
Theoretical Results

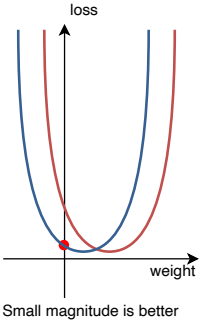
$$1-\frac{1}{D}\sum_{i=1}^{D}\frac{r_i^2}{\|\mathbf{w}^*-\mathbf{w}^k\|_2^2+r_i^2}\leq\rho\leq 1-\frac{1}{D}\sum_{i=1}^{D-k}\frac{\widetilde{r}_i^2}{\|\mathbf{w}^*-\mathbf{w}^k\|_2^2+\widetilde{r}_i^2}$$

Table: Gap Between Lower Bound and Upper Bound of Pruning Ratio.

CIFAR10	**FC5**	**FC12**	**Alexnet**	**VGG16**
$\Delta(\%)$	0.17±0.05	0.05±0.03	0.02±0.01	0.01±0.00
ResNet	**18 on CIFAR100**	**50 on CIFAR100**	**18 on TinyImagenet**	**50 on TinyImagenet**
$\Delta(\%)$	0.12±0.05	0.11±0.09	0.09±0.01	0.27±0.22

Key Factors of Fundamental Limit





Pruning Ratio vs Sharpness(Flatness):

$$\rho_L \leq 1 - \frac{2\epsilon D}{\|\mathbf{w}^* - \mathbf{w}^k\|_2^2 \mathrm{Tr}(\mathbf{H}) + 2\epsilon D}$$

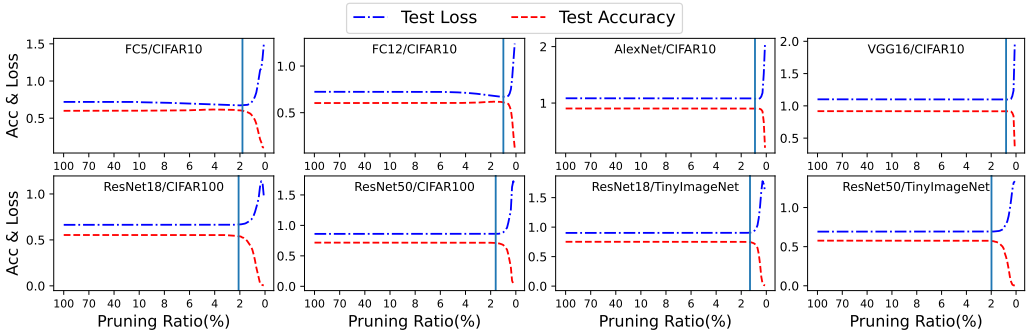
ρ_L: lower bound of pruning ratio, $\mathrm{Tr}(\mathbf{H})$: sharpness

Experiments

It's difficult to obtain the global minima of $\min \|\mathbf{w}\|_0 \quad \text{s.t.} \quad \mathcal{L}(\mathbf{w}) \leq \mathcal{L}(\mathbf{w}^*) + \epsilon.$

$$\downarrow$$

$$\min \ \mathcal{L}(\mathbf{w}) + \lambda \|\mathbf{w}\|_1$$

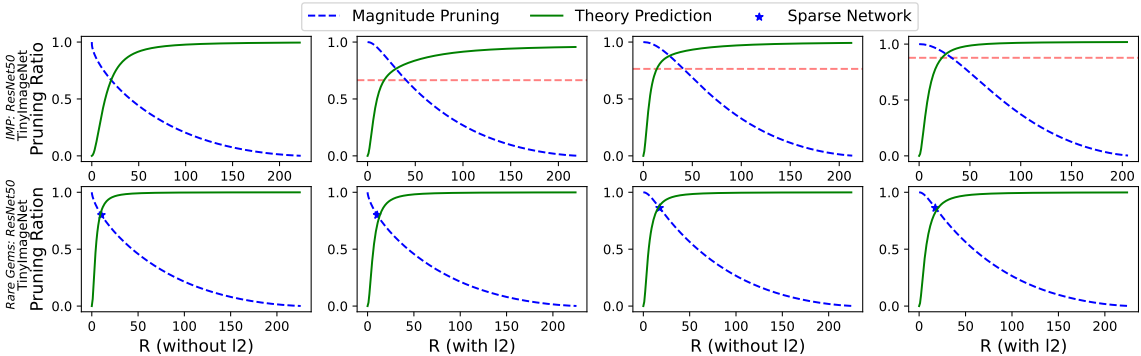


Experiments

Table: Gap between Theoretical and Actual Values of Pruning Limit

Dataset	Model	Theo. Value(%)	Actual Value(%)	Δ(%)
CIFAR10	FC5	2.1±0.25	1.7±0.12	-0.40±0.35
	FC12	1.0±0.30	0.8±0.06	-0.24±0.33
	AlexNet	0.9±0.00	0.8±0.08	-0.14±0.08
	VGG16	0.8±0.06	0.8±0.08	0.04±0.08
CIFAR100	ResNet18	1.5±0.05	2.0±0.13	0.54±0.15
	ResNet50	1.9±0.05	2.1±0.16	0.28±0.19
TinyImagenet	ResNet18	3.9±0.82	4.3±0.38	0.46±0.71
	ResNet50	2.6±0.24	2.9±0.33	0.36±0.10

Interpretation of Pruning Heuristics



- Adaptive pruning ratio;
- The role of regularization.

Open Directions

- The fundamental limits of iterative pruning;
- When will l_1 regularization yield the same pruning limits as l_0 regularization;
- A new regularization tool;
- New pruning paradigm or algorithm.