Pruning Deep Neural Networks from a Sparsity Perspective

Enmao Diao[1]* Ganghua Wang[2]* Jiawei Zhang[2]
Yuhong Yang[2] Jie Ding[2] Vahid Tarokh[1]
[1]Duke University [2]University of Minnesota-Twin Cities *Equal Contribution

Overview

We connect the compressibility and performance of a neural network to its sparsity. In a highly over-parameterized network, one popular assumption is that the relatively small weights are considered redundant or non-influential and may be pruned without impacting the performance.

- We propose a new notion of sparsity for vectors named PQ Index (PQI), with a larger value indicating higher sparsity. We prove that PQI meets all six properties proposed by [1,2,3,4], which is the first measure of sparsity related to vector norms that satisfies all the properties shared by the Gini Index [1].
- We develop a new perspective on the compressibility of neural networks. We measure the sparsity of pruned models by PQI and postulate a hypothesis on the relationship between sparsity and compressibility of neural networks.
- Motivated by our proposed PQI and hypothesis, we further develop a Sparsity-informed Adaptive Pruning (SAP) algorithm that uses PQI to choose the pruning ratio adaptively.
- We conduct extensive experiments to measure the sparsity of pruned models and corroborate our hypothesis. SAP can compress more efficiently and robustly compared with iterative pruning algorithms such as the lottery ticket-based pruning methods.

Motivation

Existing approaches lack a quantifiable measure to estimate the compressibility of a sub-network . For a non-negative vector ***w*** = [w_1,…,w_d] , we have six properties that an ideal sparsity measure ***S(w)*** should have

(D1) Robin Hood. For any $w_i > w_j$ and $\alpha \in (0, (w_i - w_j)/2)$, we have $S([w_1, \ldots, w_i - \alpha, \ldots, w_j + \alpha, \ldots, w_d]) < S(w)$.

(D2) Scaling. $S(\alpha w) = S(w)$ for any $\alpha > 0$.

(D3) Rising Tide. $S(w + \alpha) < S(w)$ for any $\alpha > 0$ and w_i not all the same.

(D4) Cloning. $S(w) = S([w, w])$.

(P1) Bill Gates. For any $i = 1, \ldots, d$, there exists $\beta_i > 0$ such that for any $\alpha > 0$ we have

$$S([w_1, \ldots, w_i + \beta_i + \alpha, \ldots, w_d]) > S([w_1, \ldots, w_i + \beta_i, \ldots, w_d]).$$

(P2) Babies. $S([w_1, \ldots, w_d, 0]) > S(w)$ for any non-zero w.

Hypothesis

Regularization: Performance moderately improves, and Sparsity decrease.
Compression: Performance moderately degrades, and Sparsity increases.
Collapse: Performance significantly degrades, and Sparsity decreases.

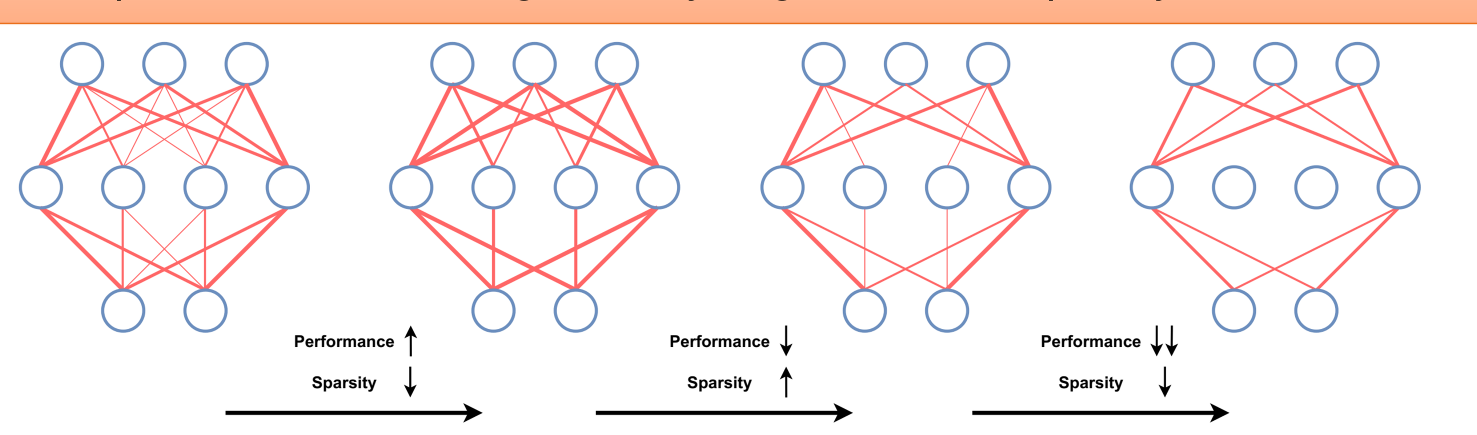

Figure 1. An illustration of our hypothesis on the relationship between sparsity and compressibility of neural networks. The width of connections denotes the magnitude of model parameters.

Paper

Code

Pruning with PQ Index

PQ Index (PQI)

- All the six properties (D1)-(D4) and (P1), (P2) are hold for our proposed PQ Index (PQI).

Definition 1 (PQ Index). *For any $0 < p < q$, the PQ Index of a non-zero vector $w \in \mathbb{R}^d$ is*

$$\mathrm{I}_{p,q}(w) = 1 - d^{\frac{1}{q}-\frac{1}{p}} \frac{\|w\|_p}{\|w\|_q}, \tag{1}$$

where $\|w\|_p = (\sum_{i=1}^d |w_i|^p)^{1/p}$ is the ℓ_p-norm of w for any $p > 0$. For simplicity, we will use $\mathrm{I}(w)$ and drop the dependency on p and q when the context is clear.

Theorem 1. *We have $0 \le \mathrm{I}_{p,q}(w) \le 1 - d^{\frac{1}{q}-\frac{1}{p}}$, and a larger $\mathrm{I}_{p,q}(w)$ indicates a sparser vector. Furthermore, $\mathrm{I}_{p,q}(w)$ satisfies all the six properties (D1)-(D4) and (P1), (P2).*

Remark 1 (Sanity check). *For the densest or most equal situation, we have $w_i = c$ for $i = 1, \ldots, d$, where c is a non-zero constant. It can be verified that $\mathrm{I}_{p,q}(w) = 0$. In contrast, the sparsest or most unequal case is that w_i's are all zeros except one of them, and corresponding $\mathrm{I}_{p,q}(w) = 1 - d^{\frac{1}{q}-\frac{1}{p}}$. Note that $\mathrm{I}(w)$ for an all-zero vector is not defined. From the perspective of the number of important elements, an all-zero vector is sparse; however, it is dense from the aspect of energy distribution.*

Remark 2 (Insights). *The form of $\mathrm{I}_{p,q}$ is not a random thought but inherently driven by properties (D1)-(D4). Why do we need the ratio of two norms? It is essentially decided by the requirement of (D2) Scaling. If $S(w)$ involves only a single norm, then $S(w)$ is not scale-invariant. However, since ℓ_r-norm is homogeneous for all $r > 0$, the ratio of two norms is inherently scale-invariant. Why is there an additional scaling constant $d^{\frac{1}{q}-\frac{1}{p}}$? This is necessary to satisfy (D4) Cloning. Inspired by the well-known Root Mean Squared Error (RMSE), we found out that the additional scaling constant is the correct term to help $\mathrm{I}_{p,q}$ be independent of the vector length. It is essentially appealing for comparing the sparsity of neural networks with different model parameters. Why do we require $p < q$? We find it plays a central role in meeting (D1) and (D3). The insight is that $\|w\|_p$ decreases faster than $\|w\|_q$ when a vector becomes sparser, thus guaranteeing a larger PQ Index.*

Theorem 2 (PQI-bound on pruning). *Let M_r denote the set of r indices of w with the largest magnitudes, and η_r be the smallest value such that $\sum_{i \notin M_r} |w_i|^p \le \eta_r \sum_{i \in M_r} |w_i|^p$. Then, we have*

$$r \ge d(1 + \eta_r)^{-q/(q-p)} [1 - \mathrm{I}(w)]^{\frac{qp}{q-p}}. \tag{2}$$

Sparsity-informed Adaptive Pruning (SAP)

- We propose SAP to adaptively determine the number of pruned parameters at each pruning iteration based on the PQI-bound and lottery ticket hypothesis [5].
- After arriving at w_t, our proposed SAP will compute the PQ Index, denoted by $I(w_t)$, and the lower bound of the number of retrained model parameters, denoted by r_t, as follows

$$\mathrm{I}(w_t) = 1 - d_t^{\frac{1}{q}-\frac{1}{p}} \frac{\|w_t\|_p}{\|w_t\|_q}, \qquad r_t = d_t(1 + \eta_r)^{-q/(q-p)} [1 - \mathrm{I}(w_t)]^{\frac{qp}{q-p}}.$$

- Then, we compute the number of pruned model parameters as follows

$$c_t = \lfloor d_t \cdot \min(\gamma(1 - \tfrac{r_t}{d_t}), \beta) \rfloor.$$

Algorithm 1 Sparsity-informed Adaptive Pruning (SAP)

Input: model parameters w, mask m, norm $0 < p < q$, compression hyper-parameter η_r, scaling factor γ, maximum pruning ratio β, number of epochs E, and number of pruning iterations T.
Randomly generate model parameters w_{init}
Initialize mask m_0 with all ones
for each pruning iteration $t = 0, 1, 2, \ldots T$ **do**
  Initialize model parameters $\tilde{w}_t = w_{\text{init}} \odot m_t$
  Compute the number of model parameters $d_t = |m_t|$
  Train the model parameters $\tilde{w}_t$ with m_t for E epochs and arrive at w_t
  Compute PQ Index $\mathrm{I}(w_t) = 1 - d_t^{\frac{1}{q}-\frac{1}{p}} \frac{\|w_t\|_p}{\|w_t\|_q}$
  Compute the lower bound of the number of retained model parameters
  $r_t = d_t(1 + \eta_r)^{-q/(q-p)} [1 - \mathrm{I}(w_t)]^{\frac{qp}{q-p}}$
  Compute the number of pruned model parameters
  $c_t = \lfloor d_t \cdot \min(\gamma(1 - \frac{r_t}{d_t}), \beta) \rfloor$
  Prune c_t model parameters with the smallest magnitude based on w_t and m_t
  Create new mask m_{t+1}
end
Output: The pruned model parameters w_T and mask m_T.

Experiments

Retrained and Pruned models

- Obtained from the models after (a) retraining and (b) directly from those after pruning.
- The dynamics of the sparsity corroborate our hypothesis.

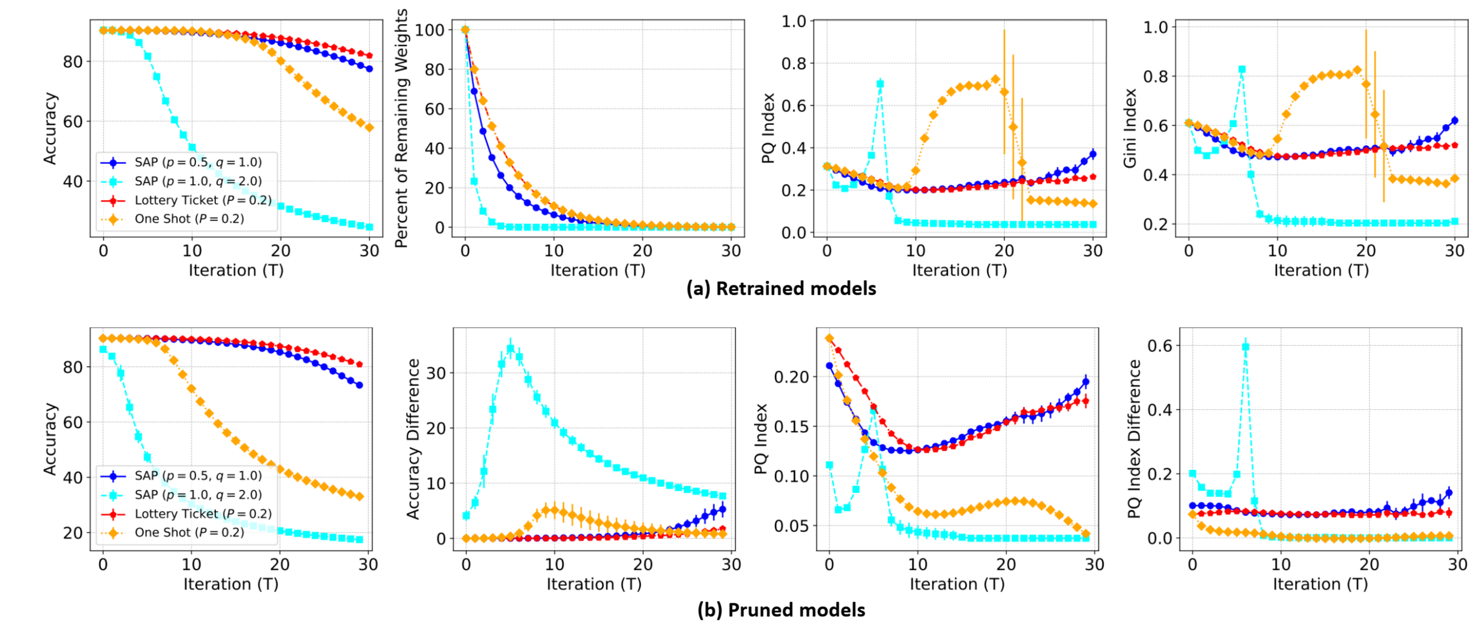

Figure 2. Results of (a) retrained and (b) pruned models at each pruning iteration for 'Global Pruning' with CIFAR10 and CNN.

Pruning scopes

- SAP with 'Global Pruning' performs worse than 'One Shot' and 'Lottery Ticket' when the percent of remaining weights is small.
- SAP with 'Neuron-wise Pruning' and 'Layer-wise Pruning' perform better than 'One Shot' and 'Lottery Ticket.'
- SAP Aligns with the intuition that the initial layers of CNN are more important to maintain the performance [6].

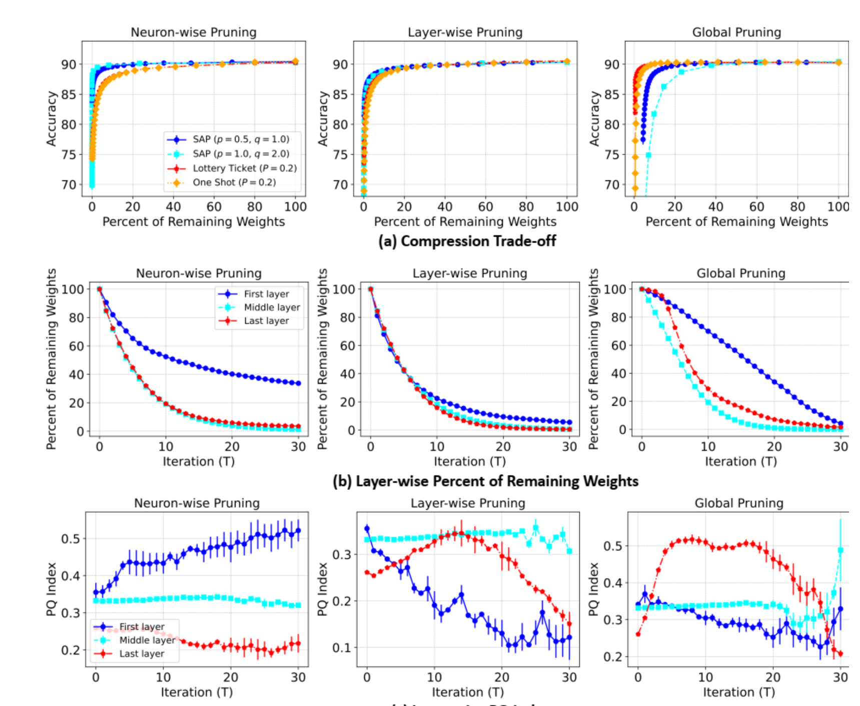

Figure 3. Results of various pruning scopes regarding (a) compression trade-off, (b) layer-wise percent of remaining weights, and (c) layer-wise PQ Index for CIFAR10 and CNN.

Ablation studies

- The dynamics of the sparsity measure of SAP with various ablation studies also corroborate our hypothesis.

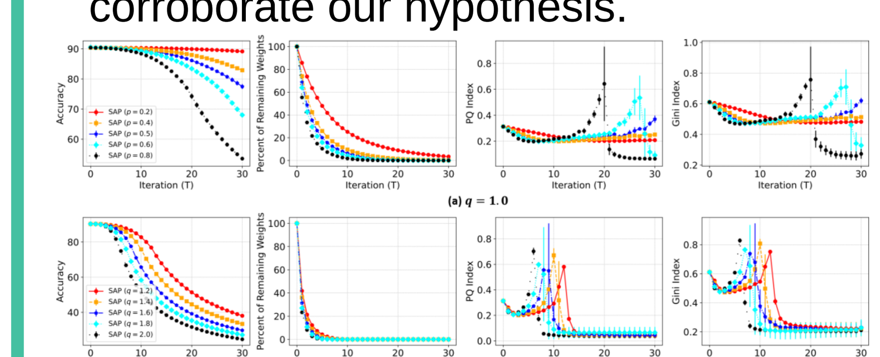

Figure 4. Ablation studies of p and q for global pruning with CIFAR10 and CNN.

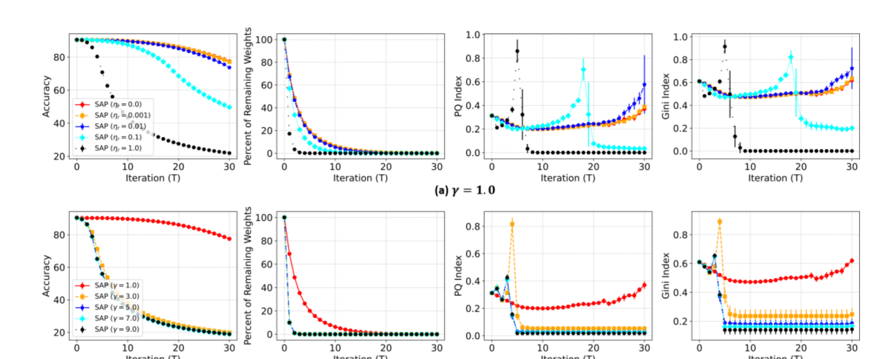

Figure 5. Ablation studies of η_r and γ for global pruning with CIFAR10 and CNN.

References

[1] Gini, Corrado. Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]. Tipogr. di P. Cuppini, 1912.

[2] Dalton, Hugh. "The measurement of the inequality of incomes." The Economic Journal 30.119 (1920): 348-361.

[3] Rickard, Scott, and Maurice Fallon. "The Gini index of speech." Proceedings of the 38th Conference on Information Science and Systems (CISS'04). 2004.

[4] Hurley, Niall, and Scott Rickard. "Comparing measures of sparsity." IEEE Transactions on Information Theory 55.10 (2009): 4723-4741.

[5] Frankle, Jonathan, and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." arXiv preprint arXiv:1803.03635 (2018).

[6] Gale, Trevor, Erich Elsen, and Sara Hooker. "The state of sparsity in deep neural networks." arXiv preprint arXiv:1902.09574 (2019).