CM 763: Paper Review

November 15, 2021

A Brief Comment on Our Review

The paper we are reviewing relies on proofs that are both technical and lengthy. The authors use 9 lemmas spanning order statistics, random matrix theory and probability theory in order to prove the main results. Furthermore, the authors provide 4 pruning techniques (with corresponding proofs). However these proofs are all quite similar. Consequently, we will not present the technical details of the proof, since we feel this would be too long and too complicated for a (relatively) short review. Instead, we hope to:

1. Clarify the terminology used by the authors.
2. Explain how this terminology intuitively describes pruning technique.
3. Explain why the assumptions made in the magnitude based pruning technique are required.
4. Briefly comment on the validity of these assumptions.

Clarifying the Objects of Study

In [5] the authors define the following standard objects and notation:

- $L+1$ denotes the depth the neural network (alternatively there are L hidden layers), with $l \in \{1, ..., L\}$ indexing hidden layers (the input layer is denoted with $l = 0$).
- d_l denotes the number of nodes in the hidden layer $l \in \{1, ..., L\}$.
- Let $v \in \mathbb{R}^{d_l}$. The L_0 norm of v is denoted $||v||_0$, and specifies the number of non-zero entries in v. The L_2 norm of v is denotes $||v||_2$, which denotes the Euclidean "standard" norm $||v||_2 = \sqrt{v^T v}$.
- Let $A \in \mathbb{R}^{m \times n}$. The vectorization of A is defined as follows:

  $$\text{vec}(A) = [A_{1,1}, ..., A_{1,n}, ..., A_{m,1}, ..., A_{m,n}]^T \in \mathbb{R}^{m \cdot n}$$

  where we use " $\cdot$ " to dinstinguish $\mathbb{R}^{m \cdot n}$ (a set of vectors with dimension $m \cdot n$) from $\mathbb{R}^{m \times n}$ (the set of linear operators V such that $V : \mathbb{R}^m \to \mathbb{R}^n$).
- Let $A \in \mathbb{R}^{m \times n}$, then the induced 2-norm of A is defined as

  $$||A||_2 = \sigma_{\max}(A)$$

  where $\sigma_{\max}(A)$ denotes the largest singular value of A.

- If $A, B \in \mathbb{R}^{m\times n}$ then $A \circ B$ denotes the Hadamard product so that $[A \circ B]_{ij} = A_{ij}B_{ij}$; $i \in \{1, ..., m\}$, $j \in \{1, ..., n\}$.
- $\mathcal{U}[a, b]$ denotes the uniform distribution on $[a, b]$ and $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate normal distribution.
- $\sigma_l : \mathbb{R} \to \mathbb{R}$, $l \in \{1, ..., L\}$ is the activation function associated with layer l. $W_l^* \in \mathbb{R}^{d_l \times d_{l-1}}$, $l \in \{0, ..., L\}$ denotes the weight matrix of this layer.
- The target network is defined as $F(x) = W_l^* \sigma_{l-1}(W_{l-1}^* \sigma_{l-1}(\cdots W_2^* \sigma_1(W_1^* x)))$ and is just the functional representation of a neural network. It is assumed activation functions act componentwise on their inputs.

The authors in [5] also define the some more idiosyncratic objects:

- The number of weights in the l^{th} layer is denoted $D_l := d_l d_{l-1}$.
- A mask matrix $M \in [0, 1]^{m\times n}$ is just an $m \times n$ valued matrix whose entries are identically 0 or 1.
- A pruned weight matrix corresponding to the weight matrix W_l^* of a fully connected network is given by $W_l = M_l \circ W_l^*$ where M_l is a mask matrix.
- A pruning $f(x)$ of the target network $F(x)$ is defined as $f(x) = W_l \sigma_{l-1}(W_{l-1}\sigma_{l-1}(\cdots W_2 \sigma_1(W_1 x)))$.
- Let W_l be a pruned weight matrix for layer l. The compression ratio of layer l is defined as

$$\gamma_l := \frac{||\text{vec}(W_l)||_0}{D_l}$$

- $f(x)$ is said to be ϵ-close to $F(x)$ iff

$$\sup_{x \in \mathcal{B}_{d_0}} ||f(x) - F(x)|| < \epsilon$$

where $\mathcal{B}_{d_0} := \left\{x \in \mathbb{R}^{d_0} \mid ||x||_2 \leq 1\right\}$ denotes the d_0 unit-ball with the Euclidean norm.

Intuition Behind the Objects

In general the objects defined above are either defined to formally express a pruned network, or to provide some measure of efficacy for a given pruned network. In the former category are mask matrices, pruned weight matrices and prunings of the target network. The latter consists of ϵ-closeness and the compression ratio.

Intuitively, the mask matrix can be thought of as a "switch" which allows or blocks certain components of the previous layer to modify the input of the activation's output. This can be seen with a toy example. Suppose that for some layer l of a FCN, the weight function is given as follows

$$W_l^* = \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \end{bmatrix}$$

Now consider the mask matrix

$$M_l = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

This mask induces a pruned weight matrix W_l

$$W_l = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \circ \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 5 \\ 7 & 0 \end{bmatrix}$$

Now suppose the input from the previous layer is given by $x = [x_1, x_2]^T$, then

$$z := W_l x = \begin{bmatrix} x_1 + 2x_2 \\ 5x_2 \\ 7x_1 \end{bmatrix} \neq \begin{bmatrix} x_1 + 2x_2 \\ 4x_1 + 5x_2 \\ 7x_1 + 8x_2 \end{bmatrix} = W_l^* x =: z^*$$

Hence we see that the activation input is different between the pruned and unpruned weight matrices. But crucially, the i^{th} component x_i of the weight matrix input x has no influence of the j^{th} component z_j of the activation input z if $[M_l]_{ji} = 0$, and the influence of the x_i on z_j is unaltered when $[M_l]_{ji} = 0$. This is what we mean by saying M_l switches — or prunes — connections in a weight matrix. Its effect is tantamount to severing a connection between the inputs of a weight matrix (or, outputs of the previous layer's activation function) and the inputs of the layer's activation function.

Since $W_l = M \circ W_l^*$, W_l is naturally interpreted as the weight matrix of a network that has had its connections pruned by M_l. Finally, $f(x)$ has the same same general architecture as $F(x)$ — its activation functions, number of layers and nodes[1] are the same as $F(x)$. The only difference between $f(x)$ and $F(x)$ is that the former uses pruned weight vectors, while the latter uses the original, unpruned weight vectors. In this way it is clear that $f(x)$ just represents a subnetwork of $F(x)$ which has had its connections pruned as specified by the matrices $M_l,\ l \in \{1, ..., L\}$.

The discussion above makes it clear that $f(x), M_l, W_l$ are defined as a means of formally representing a pruned subnetwork of some FCN. However, we ultimately wish to describe the efficacy of the pruned network. Of particular imporance are how much "smaller" the pruned subnetwork is compared to its corresponding FCN, and the discrepancies between the predictions of the pruned and FC network.

The former notion is captured by the compression ratio γ_l. In particular this ratio can be interpreted as the percentage of connections that are left in the neural network after pruning. So if $\gamma_l = 0.18$, only 18% of the original connections between layers l and $l-1$ remain. This can be seen in the extreme cases where a masking matrix M_l for some FC weight matrix W_l^* has entries that are identically 0 or 1. In the first case, it is clear that every entry of M_l, $\text{vec}(M_l)$ are zero, so that $||\text{vec}(W_l)||_0 = 0$ and $\gamma_l = ||\text{vec}(W_l)||_0 / D_l = 0$. Meanwhile if M_l has all entries equal to 1 then $||\text{vec}(M_l \circ W_l^*)||_0 = d_l d_{l-1} = D_k$ so $\gamma_l = 1$. Every other case falls in between these two extremes, since $0 \leq ||\text{vec}(W_l)||_0 \leq D_k$ clearly.

[1] If every $M_{ji} = 0$ for some $i \in d_{l-1}$, then it is clear the value of x_i^l will make no contribution to the vector z. However, strictly speaking the i^{th}node of layer $l-1$ is still part of the network — it's just vestigial.

Returning to our toy example, we see that

$$W_l = \begin{bmatrix} 1 & 2 \\ 0 & 5 \\ 7 & 0 \end{bmatrix} \Rightarrow ||\text{vec}(W_l)||_0 = 3 \Rightarrow \gamma_l = \frac{||\text{vec}(W_l)||_0}{d_l \cdot d_{l-1}} = \frac{3}{3 \cdot 2} = 0.5$$

which again describes the percentage of connections which have not been pruned.

Finally ϵ-closeness can be interpreted as a guaranteed accuracy. By selecting $x \in \mathcal{B}_{d_0}$ that causes the largest discrepancy between the values of $f(\cdot)$ and $F(\cdot)$, $f(\cdot)$ and $F(\cdot)$ are ϵ-close if even in the worst case scenario the discrepancy between the $f(\cdot)$ and $F(\cdot)$ is less than ϵ.

With this intuition in mind, the goal of a pruning algorithm can be cast in a more technical form: given a FCN $F(\cdot)$, find a set of masking matrices $\{M_1, ..., M_l\}$ so that for a given ϵ the corresponding pruned network $f(\cdot)$ has the smallest possible compression ratios $\{\gamma_1, ..., \gamma_l\}$ while remaining ϵ-close to $F(\cdot)$.

The Main Result: Magnitude Based Pruning

Specifying a pruned subnetwork $f(\cdot)$ is tantamount to determining a set of masking matrices. One approach for selecting M considered by the authors is so-called "magnitude based pruning". This method begins by ordering the entries of a FC weight matrix W_l^* by magnitude

$$|W_l^*|_{i_1,j_1} \leq \cdots \leq |W_l^*|_{i_{d_l},j_{d_l}}$$

where i_n, j_n refer to the index of the entry in W_l^* with the n^{th} smallest magnitude ($n \in \{1, ..., D_k\}$). Then M_l is set by first specifying the desired compression ratio γ_l, setting the entries of M_l corresponding to the smallest $\lfloor \gamma_l D_l \rfloor$ components of W_l^* to zero, and the rest to one.

With this context, we state the main result of the paper:

Theorem 1 of [5]: *Suppose that F is a FC target network and that:*

(i) *σ_l is Lipschitz continuous with constant K_l, $\forall l \in \{1, ..., L\}$*

(ii) *$d := \min\{d_1, ..., d_{L-1}\} \geq \max\{d_0, d_L\}$*

(iii) *The entries in W_k^* are iid following*

$$[W_k^*]_{i,j} \sim \mathcal{U}\left[-\frac{K}{\sqrt{\max\{d_l, d_{l-1}\}}}, \frac{K}{\sqrt{\max\{d_l, d_{l-1}\}}}\right]; \ i \in \{1, ..., d_l\}, i \in \{1, ..., d_{l-1}\}$$

where K is a fixed positive constant.

Let $\epsilon, \delta > 0, \alpha \in (0, 1)$ so that

$$d \geq \max\left\{C_1^{\frac{1}{\alpha}}, \left(\frac{C_2}{\epsilon}\right)^{\frac{1}{\alpha}}, \left(\frac{C_3}{\delta}\right)^{\frac{1}{\alpha}}, C_4 + C_5 \log\left(\frac{1}{\delta}\right)\right\}$$

For $C_1, ..., C_5$ depending on values of l, K_l. Then with probability at least $1 - \delta$, the subnetwork f of F with mask $M = \{M_1, ..., M_L \mid M_l \in \{0, 1\}^{d_l \times d_{l-1}}\}$ pruning the smallest $\lfloor D_l^{1-\alpha} \rfloor$ entries of W_l^*, $l \in \{1, ..., L\}$

based on magnitude is ϵ-close to F, i.e.

$$\sup_{x \in \mathcal{B}_{d_0}} ||f(x) - F(x)||_2 \leq \epsilon$$

We note that α effectively specifies $\gamma_l = D_l^{-\alpha}$, so although it is not directly stated in the theorem, the compression ratio is present.

Unpacking and Justifying the Assumptions:

As previously stated we will not provide a full characterization of the proof. Instead, we clarify why the authors made the three assumptions that appear in the theorem.

First, we comment on the Lipschitz continuity of σ_k. For our purposes function $g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous iff

$$||g(x) - g(y)||_2 \leq K||x - y||_2$$

for any $x, y \in \mathbb{R}^n$ (see [7] for more details on Lipschitz continuity). The intuition when $g : \mathbb{R} \to \mathbb{R}$ is that g is Lipschitz just when its value never changes quicker than the function $h(x) = Kx$. The reason the activation functions are Lipschitz continuous is just because the proof of the main theorem relies on random matrix theory and, in particular, the probability that the 2-norms of random matrices exceed certain bounds. By ensuring that

$$||\sigma(z_l^*) - \sigma(z_l)||_2 \leq K_l||W_l^* x_l - W_l x_l||_2$$

the authors are able to extend the results of bounding the difference between vectors under random matrices, to bounding the difference between vectors under non-linear transformations (see section 4 of [5]). So, the results of random matrix theory can be harnessed to prove the main results.

The second assumption is largely used to provide tighter bounds on the probabilities given in the relevant theorem. As such its role is mostly related to technical details (again see section 4 of [5]).

Finally, the third assumption is by far the most important assumption. As previously mentionned, the main proof relies on a few theorems imported from random matrix theory [5]. In particular, the main theorem heavily relies on [2, 4], of which the former only applies to independent, mean zero random matrices with finite fourth moments and the latter which applies to (sub-Gaussian) iid random matrices.

The first two assumptions are easy to justify. For the first assumption, we note that in practice most activation functions are Lipschitz continuous. For instance, the sigmoid, ReLU, tanh and softmax functions are all Lipschitz continuous with $K = 1$ (see [6]). So this is clearly a reasonable assumption. The second assumption is even easier to justify. One can simply set all $d \geq \max\{d_0, d_l\}$ when designing the architecture of the FCN $F(\cdot)$. Hence the first two assumptions will almost always hold in practice.

Unlike the first two assumptions, the third assumption is not evidently the case. Assuming that $\{W_l^* \mid l \in \{1, ..., L\}\}$ are i.i.d. random is a priori strong assumption with no particularly strong justification. Given this difficulty, the authors essentially argue that because (1) there is no way to effectively measure how close weights are to independence, (2) some literature — namely [3, 1]— suggests that weights in a trained network do not deviate strongly from their initial values. However, in our view it is clear that this constitutes only a partial, if not promising, justification for assumption (iii). Consequently the most important assumption for the main theorem is also the least secure.

References

[1] Yu Bai and Jason D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks, 2020.

[2] H. A. David and H. N. Nagaraja. *Order Statistics.* American Cancer Society, 2004.

[3] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021.

[4] Rafal Latala. Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282, 2005.

[5] Xin Qian and Diego Klabjan. A probabilistic approach to neural network pruning, 2021.

[6] Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 3839 – 3848. Curran Associates Inc., 2018.

[7] Micheal O Searcoid. *Metric Spaces.* Springer Undergraduate Mathematics Series. Springer, 1 edition, 2006.