# A Bound Proof

In this section, we provide proof for the upper bound of pruning process:

$$\begin{aligned}&\sup_{\|x\|_2\leq 1}\left\|f\left(x;W^{(1:L)}\right)-f\left(x;W^{(1:l-1)},\widetilde{W}^{(l)},W^{(l+1:L)}\right)\right\|_2\\&\leq\left\|W^{(l)}-\widetilde{W}^{(l)}\right\|_F\cdot\prod_{\substack{j\neq l\\ j\in[1,L]}}\left\|W^{(j)}\right\|_F\end{aligned}\tag{11}$$

Inspired by Neyshabur et al. (2015), we can write the network $f\left(x;W^{(1:L)}\right)$ with its layers:

$$f\left(x;W^{(1:L)}\right)=W_L\sigma\left(W_{L-1}\sigma\left(W_{L-2}\left(\ldots\sigma\left(W_1x\right)\right)\right)\right)\tag{12}$$

Then we can peel the highest layer $W_L$:

$$\begin{aligned}&\left\|f\left(x;W^{(1:L)}\right)-f\left(x;W^{(1:l-1)},\widetilde{W}^{(l)},W^{(l+1:L)}\right)\right\|_2\\&=\|W^{(d)}\left(\sigma\left(f\left(x;W^{(1:L-1)}\right)-\sigma\left(f\left(x;W^{(1:l-1)},\widetilde{W}^{(l)},W^{(l+1:L-1)}\right)\right)\right)\right)\|_2\end{aligned}\tag{13}$$

With Cauchy-Schwarz inequality Steele (2004), we have the upper bound of Equation 13 as:

$$\left\|W^{(L)}\right\|_F\cdot\|\sigma\left(f\left(x;W^{(1:L-1)}\right)-\sigma\left(f\left(x;W^{(1:l-1)},\widetilde{W}^{(l)},W^{(l+1:L-1)}\right)\right)\right)\|_2\tag{14}$$

Suppose $\sigma$ is the ReLU activation, so we use the 1-Lipschitzness of ReLU activation with respect to $\ell_2$ norm for the upper bound of Equation 14 as:

$$\left\|W^{(L)}\right\|_F\cdot\left\|f\left(x;W^{(1:L-1)}\right)-f\left(x;W^{(1:l-1)},\widetilde{W}^{(l)},W^{(l+1:L-1)}\right)\right\|_2\tag{15}$$

Then we continue peeling the second highest layer $W_{L-1}$ to lower layers and stop the peeling at the **pruned layer** $\widetilde{W}^{(l)}$. We have this upper bound for Equation 13:

$$\begin{aligned}&\left\|f\left(x;W^{(1:L)}\right)-f\left(x;W^{(1:l-1)},\widetilde{W}^{(l)},W^{(l+1:L)}\right)\right\|_2\\&\leq\left(\prod_{j>l}\left\|W^{(j)}\right\|_F\right)\cdot\left\|f\left(x;W^{(1:l)}\right)-f\left(x;W^{(1:l-1)},\widetilde{W}^{(l)}\right)\right\|_2\end{aligned}\tag{16}$$

We consider the effect of pruning for the term on the right side and use Cauchy-Schwarz inequality again:

$$\begin{aligned}\left\|f\left(x;W^{(1:l)}\right)-f\left(x;W^{(1:l-1)},\widetilde{W}^{(l)}\right)\right\|_2&=\left\|\left(W^{(l)}-\widetilde{W}^{(l)}\right)\sigma\left(f\left(x;W^{(1:l-1)}\right)\right)\right\|_2\\&\leq\left\|W^{(l)}-\tilde{W}^{(l)}\right\|_F\cdot\left\|\sigma\left(f\left(x;W^{(1:l-1)}\right)\right)\right\|_2\end{aligned}\tag{17}$$

For the activation term, based on $\sigma(\mathbf{0})=\mathbf{0}$, we have:

$$\left\|\sigma\left(f\left(x;W^{(1:l-1)}\right)\right)\right\|_2=\left\|\sigma\left(f\left(x;W^{(1:l-1)}\right)\right)-\sigma(\mathbf{0})\right\|_2\leq\left\|f\left(x;W^{(1:l-1)}\right)-\mathbf{0}\right\|_2=\left\|f\left(x;W^{(1:l-1)}\right)\right\|_2.\tag{18}$$

If we continue the peeling process as shown in equation 13, we can achieve equation 11.

# B Module Definition

**Attention-related Module (QKV-M and PRJ-M).** Self-attention is an important operation in Swin Transformer. In every transformer block, there are two layers related to self-attention, that is, ATT-QKV and ATT-PRJ in Figure 3. ATT-QKV means the "Query, Key, and Value" matrix for self-attention, and ATT-PRJ indicates the projection layer for self-attention. Assuming $\mathbf{z}^l$ is the feature map of $l_{th}$ layer, two attention-related layers are:

$$\mathbf{z}^{l+1} = \text{ATT-QKV}\left(\mathbf{z}^{l}\right), \quad \mathbf{z}^{l+2} = \text{ATT-PRJ}\left(\mathbf{z}^{l+1}\right) \tag{19}$$

where $\mathbf{z}^{l}$ is the input feature map of ATT-QKV, and $\mathbf{z}^{l+1}$ is the input feature map of ATT-PRJ. After these two layers, there is a residual connection between the output of the ATT-PRJ layer and the feature map before the LN layer:

$$\hat{\mathbf{z}}^{l+2} = \mathbf{z}^{l+2} + \mathbf{z}^{l-1} \tag{20}$$

Assume there are $J$ transformer blocks in the whole network. For the transformer block shown in Figure 3(a), assume the range of the layer index is $[l,\ l+5]$. For the $j_{th}$ block after this block, the layer index range is $[l+p_j,\ l+p_j+5]$, where $p_j$ means the number of layers between these two blocks. Then we can define QKV-M and PRJ-M as:

$$\begin{aligned} \text{QKV-M} : \left\{\mathbf{W}^{l+1}, ..., \mathbf{W}^{l+p_j+1}, ...\right\} \quad \text{for} \quad j \in [1, J]. \\ \text{PRJ-M} : \left\{\mathbf{W}^{l+2}, ..., \mathbf{W}^{l+p_j+2}, ...\right\} \quad \text{for} \quad j \in [1, J]. \end{aligned} \tag{21}$$

**Multilayer Perceptron-related Module (MLP-M).** Another important part of the Swin Transformer is the multilayer perceptron. There are two MLP layers in every Swin Transformer block. We name these two layers MLP-FC1 and MLP-FC2:

$$\mathbf{z}^{l+4} = \text{MLP-FC1}\left(\mathbf{z}^{l+3}\right), \quad \mathbf{z}^{l+5} = \text{MLP-FC2}\left(\mathbf{z}^{l+4}\right) \tag{22}$$

where $\mathbf{z}^{l+3}$ is the input feature map of MLP-FC1, and $\mathbf{z}^{l+4}$ is the input feature map of MLP-FC2. Similarly, after two MLP layers, there is a residual connection between the output of the MLP-FC2 layer and the feature map before the LN layer:

$$\hat{\mathbf{z}}^{l+5} = \mathbf{z}^{l+5} + \hat{\mathbf{z}}^{l+2} \tag{23}$$

Considering $J$ blocks in the whole network, and $j \in [1, J]$, we can define the MLP-M as:

$$\text{MLP-M} : \left\{\mathbf{W}^{l+4}, \mathbf{W}^{l+5}, ..., \mathbf{W}^{l+p_j+4}, \mathbf{W}^{l+p_j+5}, ...\right\}. \tag{24}$$

**Auxiliary Module (AUX-M).** The auxiliary module consists of the auxiliary layers inside and outside the Swin Transformer block. Inside the Swin Transformer block, as shown in Figure 3(a), there are two LN layers. Some other auxiliary layers outside the Swin Transformer blocks, including a patch embedding layer and several patch merging layers, are also included in AUX-M. We do not prune these layers due to their important contribution to the network and their relatively small parameter count.

## C Model as Module

Despite defining multiple modules in our experiments and evaluating weight importance within each module, it is intriguing to view the entire model as a single module. In this setting, we keep the auxiliary layers unpruned and view all other layers, including ATT-QKV, ATT-PRJ, MLP-FC1, and MLP-FC2, as a module. This means all the weights are compared based on our novel weight metric. The results are shown in Tab. 4, demonstrating that our weight metric can achieve reasonably good performance, even when considering all functional layers as a unified module.

| Model | Multiple Modules (%) | Single Module (%) | Difference (%) |
|---|---|---|---|
| Swin-B-DIMAP3 | 83.28 | 83.17 | 0.11 |
| Swin-S-DIMAP3 | 82.63 | 82.31 | 0.32 |
| Swin-T-DIMAP3 | 80.35 | 80.12 | 0.23 |

Table 4: Comparison of assigning functional layers as multiple modules and a single module.