Once for Both: Single Stage of Importance and Sparsity Search for Vision Transformer Compression

Supplementary Material

In supplementary material, we first provide a summary of all notations mentioned in the main body for a clear understanding of the paper, as shown in Table A.1. Then, we make a deep analysis of the adaptive one-hot loss function, including the theoretical justification of its effectiveness (Appendix A). In addition, we demonstrate the necessity of activating variance regularization with the tangent function from the lens of optimization space (Appendix B). Then, we introduce the implementation details of our experiments on different baseline models (Appendix C). Finally, we provide more experimental analyses (Appendix D, E, and F). Our code is available at https://github.com/HankYe/Once-for-Both.

A. Motivation behind Entropy and Variance Regularizations

In this section, we take a deep dive into the design of the adaptive one-hot loss function, which targets discretizing each p_i into a one-hot vector. Considering the inaccessible one-hot index in the search process (Sec. 3.3), the optimization objective of p_i should adapt to potential members in a one-hot vector group according to the search results of other submodules. For example, the sparsity target of one submodule can change from [0, 0, 1, 0] to [0, 1, 0, 0] if other pruned submodules contribute to a small computation reduction. Our method employs an adaptive one-hot loss function to learn members' invariant and unique properties within a one-hot vector group, fulfilling the optimization objectives.

First, we present two learnable properties of the one-hot vector set: entropy $\mathcal{H}(p_i)$ and variance $\sigma(p_i)$. Focusing on the i-th dimension in the normalized α using *softmax*, denoted as p_i, we establish a theorem revealing the unique relationship between $\mathcal{H}(p_i)$, $\sigma(p_i)$, and the set of one-hot vectors.

Theorem 1. *Suppose $p_i \in R^{1\times D}$ and $\sum_{k=1}^{D} p_{ik} = 1$, with $p_{ik} \geq 0, k = 1, 2, ..., D$. Then the following propositions are equivalent:*

$$(1)\ \ \mathcal{H}(p_i) = -\sum\nolimits_{k=1}^{D} p_{ik}\log p_{ik} = 0; \tag{A.1}$$

$$(2)\ \ \sigma\left(p_i\right) = \sum\nolimits_{k=1}^{D} \left(p_{ik} - \bar{p}\right)^2/D = \left(D-1\right)/D^2; \tag{A.2}$$

$$(3)\ \ p_i \in \left\{e_k\right\}, k = 1, 2, ..., D, \tag{A.3}$$

where e_k represents the D-dimensional one-hot vector with the k-th element set to one.

Proof. We prove the equivalence by demonstrating A.1 ⇔ A.3 and A.2 ⇔ A.3.

As for the former equivalence, given that $p_i \in \{e_k\}$, the entropy of p_i can be easily computed as zero, thus A.3 ⇒ A.1. Then we prove that A.1 ⇒ A.3. Since p_i is constrained by $\sum_{k=1}^{D} p_{ik} = 1$, we construct a Lagrange function $L(p_i, \lambda)$ as follows:

$$L(p_i, \lambda) = \mathcal{H}(p_i) + \lambda(1 - \sum\nolimits_{k=1}^{D} p_{ik}), \tag{A.4}$$

where λ is the Lagrange multiplier. Now, we analyze the extremum and monotonicity of $L(p_i, \lambda)$ by taking partial derivatives with respect to p_{ik} and λ, as shown in Eq. (A.5).

$$\frac{\partial L}{\partial p_{ik}} = -1 - \lambda - \log p_{ik}, \quad k = 1, 2, ..., D. \tag{A.5}$$

By setting each partial derivative to zero, we can obtain that: $\lambda = \log D - 1, p_{ik} = D^{-1}, k = 1, 2, ..., D$. If $0 < p_{ik} < D^{-1}$, then $\partial L/\partial p_{ik} = -\log(Dp_{ik}) > 0, k = 1, 2, ..., D$. Similarly, if $D^{-1} < p_{ik} < 1$, then $\partial L/\partial p_{ik} = -\log(Dp_{ik}) < 0, k = 1, 2, ..., D$. Consequently, $\mathcal{H}(p_i)$ is monotone increasing if $0 < p_{ik} < D^{-1}$ and monotone decreasing if $D^{-1} < p_{ik} < 1$ in the dimension of p_{ik}. Therefore, $\mathcal{H}(p_i)$ reaches maximum as $\log D$ when $p_i = D^{-1}\mathbf{1}_{1\times D}$, and reaches the minimum as zero when $p_{ik} \in \{0, 1\}, k = 1, 2, .., D$. In other words, $\mathcal{H}(p_i) = 0$ holds only if $p_i \in \{e_k\}, k = 1, 2, ..., D$. Therefore, $\mathcal{H}(p_i) = 0 \Rightarrow p_i \in \{e_k\}$, *i.e.*, A.1 ⇒ A.3. Finally, we can get the conclusion that A.1 is equivalent to A.3.

As for the latter equivalence, A.2 ⇔ A.3, the proof is similar and thus omitted here.

□

Based on the analysis, the entropy and variance regularization of p_i can effectively drive it towards a one-hot vector, discretizing both p_{ik} and the sparsity score $\mathcal{V}$ as binary.

Having demonstrated the effectiveness of both regularization items, another question is why both regularization items should be employed. To answer this question, we analyze the respective contribution of the entropy and variance regularization items to the discretization process, making the following observations.

Theorem 2. *Let $\|\mathcal{H}(p_i) - 0\|$ and $\|\sigma(p_i) - (D-1)/D^2\|$ denote the regularization items for entropy and variance, respectively. Then, the following properties hold:*
(1) $\|\mathcal{H}(p_i) - 0\|$ works as ℓ_1 sparsity for p_i, guiding p_i towards a potential one-hot vector;
(2) $\|\sigma(p_i) - (D-1)/D^2\|$ works as ℓ_2 smoothness for p_i, facilitating a seamless transition from one potential target vector to another.

Symbol	Notation	Symbol	Notation
$\mathcal{N}$	supernet	m	bi-mask
$\mathcal{A}$	search space	$\lambda(t)$	time-varying weight coefficient
W	supernet weights	i	the submodule index
$\mathcal{F}_d$	decoder	j	the unit index
$\mathcal{D}$	dataset	α	architecture parameters
$\boldsymbol{f}$	importance criterion	p	the normalized architecture score
$\mathcal{S}$	importance score	Δ	search step
$\mathcal{V}$	sparsity score	M	submodule number in $\mathcal{N}$
$\mathcal{L}_{val}$	validation loss	σ	measured variance of p
$\mathcal{L}_{train}$	training loss	σ^t	target variance of p
$\mathcal{L}_{\mathcal{S}}$	regularization item for $\mathcal{S}$	ω	normalized variance of p
$\mathcal{L}_{\mathcal{V}}$	regularization item for $\mathcal{V}$	μ_1, μ_2	weight coefficients in $\mathcal{L}_{\mathcal{V}}$
$\mathcal{L}_{rec}$	reconstruction loss	μ_3	weight coefficient in $\mathcal{L}_{\mathcal{S}}$
$\mathcal{L}_m$	regularization item for m	$\bar{p}$	mean of p
$R(p)$	regularization item for p	η	scaling factor
g	computation cost	γ	masking ratio
τ	resource constraint	ΔT	pruning interval

Table A.1. Notation Summary.

Proof. We first derive the approximate order of two items to identify regularization types. Then we analyze the function of each regularization from the lens of the optimization space.

As for entropy regularization, since $\mathcal{H}(p_i) \geq 0$, the regularization can be simplified as $\mathcal{H}(p_i)$. Further, according to [30], $\mathcal{H}(p_i)$ can be regarded as the first-order entropy of the distribution p_i, as shown in Eq. (A.6).

$$\mathcal{H}\left(p_i\right) = \lim_{r \to 1} \mathcal{H}_r\left(p_i\right) = \lim_{r \to 1} \frac{1}{1-r} \log \left(\sum_{k=1}^{D} p_{ik}^r \right), \tag{A.6}$$

where $\mathcal{H}_r\left(p_i\right)$ is the generalized entropy measure, Rényi Entropy. Therefore, $\|\mathcal{H}(p_i) - 0\|$ can be regarded as ℓ_1 sparsity for the discrete distribution of p_i.

As for variance regularization, since $\sigma(p_i) \leq (D-1)/D^2$, the regularization can be simplified as $(D-1)/D^2 - \sigma(p_i)$. Further, we expand $\sigma(p_i)$ into the polynomial form as follows:

$$\begin{aligned} \frac{D-1}{D^2} - \sigma(p_i) &= \frac{D-1}{D^2} - \frac{\sum_{k=1}^{D} p_{ik}^2 - \frac{2}{D}\sum_{k=1}^{D} p_{ik} + \frac{1}{D}}{D} \\ &= \frac{D-1}{D^2} - \frac{\sum_{k=1}^{D} p_{ik}^2 - \frac{1}{D}}{D} \\ &= \frac{1}{D}\left(1 - \sum_{k=1}^{D} p_{ik}^2\right). \end{aligned} \tag{A.7}$$

From Eq. (A.7), minimizing $(D-1)/D^2 - \sigma(p_i)$ can be regarded as maximizing ℓ_2-norm of p_i. Therefore, the variance regularization cannot be viewed as the ℓ_2 sparsity. Instead, compared with entropy regularization, we argue that variance regularization works as ℓ_2 smoothness.

Specifically, we visualize the distributions of entropy and variance regularization in Fig. A.1. Considering the normalization constraint, we focus on the two-dimensional setting of p_i to simplify the analysis. From the figure, we observe that variance regularization distribution is flatter than entropy regularization in the region neighboring the maximum point. With the increase in the dimensionality of p_i, the entropy regularization becomes stronger (sharper peak), while the variance regularization becomes weaker (flatter peak). Consequently, the variance regularization effectively smooths the optimization space. □

Note that during optimization, entropy regularization is sensitive to the initialization of p_i, as the gradient continually drives the maximum p_{ik} towards one. This behavior is independent of the sparsity constraint and leads to fixing the potential one-hot index throughout the pruning process. This issue is evident in the results of lines 1 and 5 in Table 5d and Fig. A.2. In lines 1 and 5 of Table 5d, the searched models have the same size and are the largest among all searched models. The bi-mask score learning process in Fig. A.2, which is sampled from one submodule in DeiT-S compressed by entropy regularization alone, shows that the bi-mask scores in different segments are continually increased or decreased. Therefore, **entropy regularization mainly contributes to the score sparsity but is constrained by the initial score distribution**.

In contrast, variance regularization is agnostic to the one-hot index and operates within a much flatter optimization space than entropy regularization. This characteristic allows variance regularization to adaptively adjust the target one-hot vector based on the search results of other submodules or units. Fig. A.3 visualizes the bi-mask score learning process from the same submodule as in Fig. A.2, employing the same compression target. The scores in the box initially increase and gradually decrease after the pruning of other units, indicating a switch in the target one-hot vector from

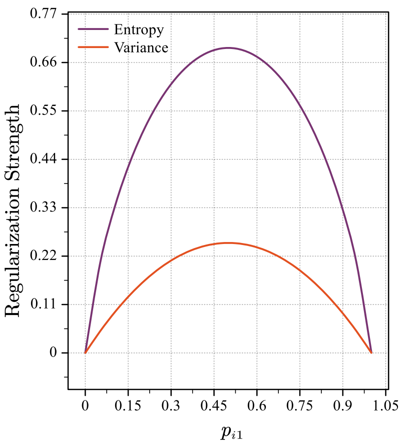

Figure A.1. Visualization of entropy and variance regularization distributions under the two-dimension p_i setting for simplicity.

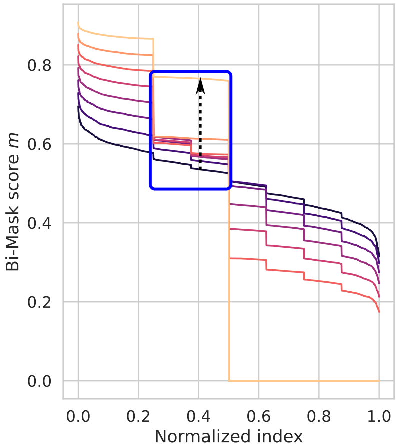

Figure A.2. Visualization of the learning process of one bi-mask sampled from the compression of DeiT-S with the target sparsity of 1BFLOPs and without variance regularization.

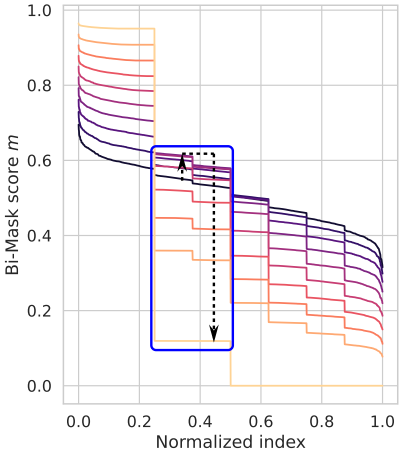

Figure A.3. Visualization of the learning process of one bi-mask sampled from the compression of DeiT-S with the target sparsity of 1BFLOPs and without entropy regularization.

[0, 0, 1, 0, 0, 0, 0] to [1, 0, 0, 0, 0, 0, 0]. Additionally, the results in lines 2 and 4 of Table 5d demonstrate that models compressed solely with variance regularization achieve the smallest model size (4.0MParams, 0.8BFLOPs) while satisfying the sparsity constraint. Hence, **variance regularization primarily contributes to a smoother optimization space, enabling easier adjustment of the target one-hot vector to align with the pruning process of other units and the desired sparsity constraint.**

In summary, we demonstrate the effectiveness and necessity of both entropy and variance regularization items from the lens of equivalent properties, regularization types, and optimization contributions.

B. Motivation behind Tangent Activation for Variance Regularization

As mentioned earlier, applying ℓ_1 regularization to the discrete variable p_i and continuous variable $\mathcal{S}$ enhances model sparsity while maintaining model performance. This observation is also supported by the results in Table 5d (lines 1, 5, 6, and 7).

Now, let's delve into the explanation for why we use tangent to map the variance regularization. This choice is primarily motivated by the over-smoothness present in the high-dimensional optimization space during the early search process. To better understand this, we analyze the regularization strength in the high-dimensional optimization space. In particular, we focus on scenarios where the dimensionality D is much larger than 2 ($D >> 2$). In such cases, the maximum strength of $(D-1)/D^2 - \sigma(p_i)$, which equals $(D-1)/D^2$, is very close to the minimum strength of zero. As a result, the gradient of the variance regularization becomes extremely

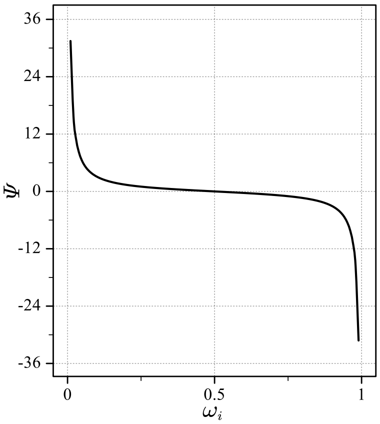

Figure A.4. Distribution of tangent-activated variance regularization.

small compared to the gradient of the entropy regularization. Consequently, the impact of variance regularization on searching for the potential target one-hot vector becomes minimal, as the gradient of variance regularization is overwhelmed by that of entropy regularization.

To solve this problem, we propose utilizing tangent activation to produce a large gradient during the search for the potential target one-hot vector under the performance and sparsity objectives. Specifically, we design the activation as follows:

$$\Psi(p_i) = \tan\left(\frac{\pi}{2} - \pi\omega_i\right), \tag{A.8}$$

where $\omega_i = \sigma_i/\sigma_i^t \in [0, 1]$ as mentioned in the main body. The distribution of $\Psi(p_i)$ is presented in Fig. A.4, where Ψ rapidly decreases when ω_i is close to zero, *i.e.*, σ_i is close to zero. In other words, if the variance of p_i is small, referring to the initial distribution of p_i or the distribution after pruning small-score units, we will assign a larger gradient for

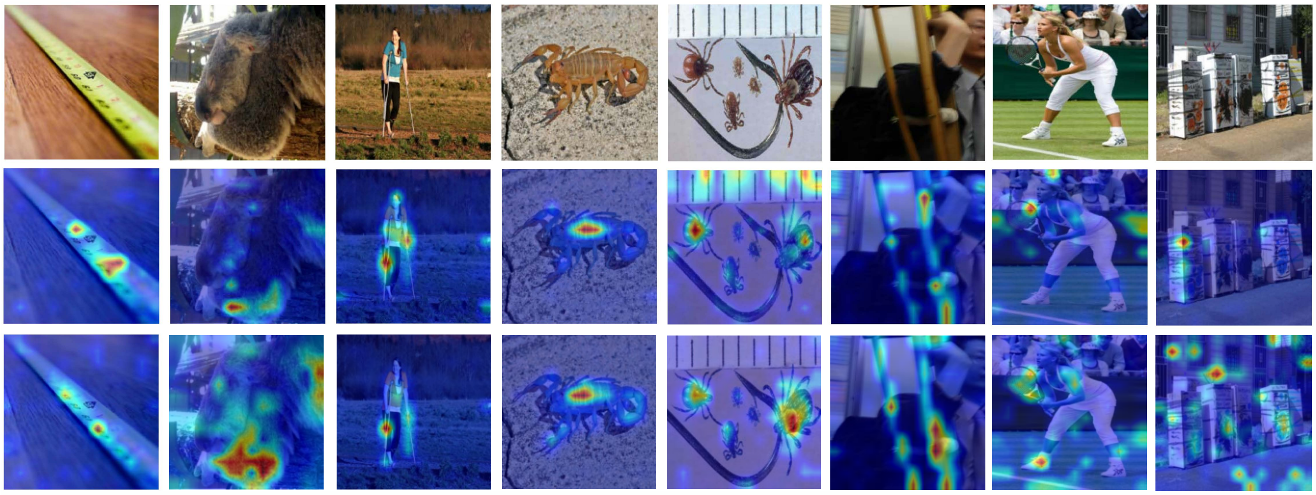

Figure A.5. Attention maps from different models for several images sampled from ImageNet-1K. The first row is the original images, the second row represents visualizations from DeiT-S, and the third row denotes the results from DeiT-B_3.6BFLOPs.

variance regularization than entropy one. This prioritization allows for faster optimization of p_i towards the potential target one-hot vector under the performance and sparsity objective. By doing so, we prevent entropy regularization from dominating the optimization process and ensure the target one-hot vector can be dynamically adjusted.

Once the potential target one-hot vector is found, the optimization process should prioritize entropy regularization. This is because entropy regularization, as an approximate ℓ_1 sparsity measure, can promote sparsity in p_i while maintaining model performance. Therefore, the gradient of variance regularization can be suppressed to minimize interference from other one-hot objectives. When the distribution of p_i is close to a one-hot vector, meaning ω_i is close to one, the significant gradient of Ψ can facilitate the discretization of p_i in the same optimization direction as entropy regularization. In this situation, the disturbance caused by variance regularization from other one-hot objectives is typically negligible.

Based on the above analysis, the main contribution of tangent activation is providing a large gradient to adjust the potential target one-hot vector of p_i that satisfies the compression requirement every time the small-score units in p_i are pruned. Therefore, as validated in lines 1, 6, 5, and 7 of Table 5d, the variance regularization $\Psi(p_i)$ can drive the model to approach the target sparsity more closely and more efficiently.

C. Implementation Details

OFB adopts the searching-and-retraining scheme as previous works do. All experiments are conducted with 8 V100 GPUs. In the search process, we use the pre-trained models released from official implementation on ImageNet-1K as the supernet $\mathcal{N}$. The decoder $\mathcal{F}_d$ consists of one convolutional layer and a pixel-shuffle layer as SimMIM [39] does. We search for 100 epochs on DeiT-S and Swin-Ti, and 200 on DeiT-B, with 20 epochs for warming up. The other learning schedules and the augmentation strategy follow the official settings in the respective papers. The learning schedules of α, $\mathcal{S}$ and $\mathcal{F}_d$ shares the same setting as W, except that β_1 is set as 0.5 for the optimizer of $\{\alpha, \mathcal{S}\}$. The default values of μ_1, μ_2, μ_3, and η are set as 5e-1, 1e2, 2e-5, and 2e-1, respectively. The unit pruning is initiated at every one-third interval (ΔT) within each epoch. In the retraining process, we follow the default training strategy reported in official papers except for mixup [45] and cutmix [44], which impair the retraining performance in our setting, and the learning rate is set as 6e-4 for both types of models. The masking ratio linearly increases from 1% to 25% of the input patches for DeiTs and that of the downsampled patches for Swin-Ti during the search stage.

D. Additional Attention Map Visualizations

We take DeiT-B_3.6BFLOPs as an example to compare the attention maps with DeiT-S, which shares the same depth as DeiT-S and has higher performance but with smaller FLOPs and parameters. We adopt the method introduced in [3] to visualize the attention maps from the output layer. The results are shown in Fig. A.5. From the figure, it can be observed that the compressed model focuses more on the extraction of class-specific contextual information, meanwhile suppresses some useless information, *e.g.*, the background features in the picture of the fifth column. This indicates that OFB can effectively evaluate the prunability of units in different submodules and finally preserve useful and important units to perform high compression performance.

Case (w/o rt.)	Top-1 (%)	FLOPs (B)	#Param (M)
Uniform init	63.5	0.6	3.0
Random init	**72.8**	**1.1**	**5.3**

Table A.2. Inductive bias analysis.

Model	Top-1 (%)	Top-5 (%)	FLOPs (B)
ResNet50	76.2	92.9	4.1
DepGraph [13]	75.8	-	2.0
OFB	**75.8**	**92.6**	**1.6**

Table A.3. Generalization Performance on ResNet-50.

E. Inductive Bias Analysis

We further explore the impact of inductive bias on the search performance. As shown in Table. A.2, with the same computation constraint, despite the smaller model size, the uniformly-initialized search space performs poorly in model performance, while the randomly-initialized one can achieve better tradeoff between model performance and compression budget, demonstrating the negative impact of inductive bias in the initialization of model parameters.

F. Generalization Ability on CNNs

To test the generalization ability of OFB, we apply it to compressing ResNet-50 on ImageNet. As shown in Table A.3, compared with baseline and SOTA models, OFB achieves comparable performance with higher compression ratio, which further demonstrates the superiority of OFB in generalization ability.