The Principle of Diversity: Training Stronger Vision Transformers Calls for Reducing All Levels of Redundancy

Tianlong Chen¹, Zhenyu Zhang¹, Yu Cheng², Ahmed Awadallah², Zhangyang Wang¹

¹University of Texas at Austin, ²Microsoft Research

{tianlong.chen, zhenyu.zhang, atlaswang}@utexas.edu, {yu.cheng, hassanam}@microsoft.com

Abstract

Vision transformers (ViTs) have gained increasing popularity as they are commonly believed to own higher modeling capacity and representation flexibility, than traditional convolutional networks. However, it is questionable whether such potential has been fully unleashed in practice, as the learned ViTs often suffer from over-smoothening, yielding likely redundant models. Recent works made preliminary attempts to identify and alleviate such redundancy, e.g., via regularizing embedding similarity or re-injecting convolution-like structures. However, a “head-to-toe assessment” regarding the extent of redundancy in ViTs, and how much we could gain by thoroughly mitigating such, has been absent for this field. This paper, for the first time, systematically studies the ubiquitous existence of redundancy at all three levels: patch embedding, attention map, and weight space. In view of them, we advocate a principle of diversity for training ViTs, by presenting corresponding regularizers that encourage the representation diversity and coverage at each of those levels, that enabling capturing more discriminative information. Extensive experiments on ImageNet with a number of ViT backbones validate the effectiveness of our proposals, largely eliminating the observed ViT redundancy and significantly boosting the model generalization. For example, our diversified DeiT obtains 0.70% ~ 1.76% accuracy boosts on ImageNet with highly reduced similarity. Our codes are fully available in https://github.com/VITA-Group/Diverse-ViT.

1. Introduction

Transformer [57], as the de facto neural architecture in natural language processing (NLP) [4, 19], recently revolutionizes modern computer vision applications such as image classification [21, 26, 53], object detection [5, 17, 74, 80], and image generation [10, 31, 46]. Rather than relying on convolution-like inductive bias, vision transformers [21] (ViTs) leverage the self-attention [57] to aggregate image patches across all spatial positions and model their global-range relationships, which are believed to improve model expressiveness and representation flexibility.

Despite their promising potentials, the ViT training still suffers from considerable instability, especially when going deeper [23, 55]. One of the major reasons [23] is that the global information aggregation among all patches encourages their representations to become overly similar, causing substantially degraded discrimination ability. This phenomenon, known as over-smoothening, suggests a high degree of “redundancy” or ineffective usage of the ViT expressiveness and flexibility, and has been studied by a few prior arts [23, 55, 75, 76]. Several initial attempts strive to fill the gap from different aspects. For example [23] proposes contrastive-based regularization to diversity patch embed-
Aiming to address the ubiquitous existence of redundancy at all three levels: patch embedding, attention map, and weight space, for current state-of-the-art (SOTA) ViTs. That is even the case for those equipped with strong data augmentations (i.e., DeiT [54]) or sophisticated attention mechanisms (i.e., Swin [43]), e.g. as shown in Figure 1. In view of such collapse, we advocate a principle of diversity for training ViTs, by proposing corresponding regularizers that encourage the representation diversity and coverage at each of those levels, that unleashes the true discriminative power and representation flexibility of ViTs. We find each level's regularizers to provide generalization gains, and applying them altogether consistently yields superior performance. Our contributions lie in the following aspects:

• We provide the first comprehensive investigation of redundancy in ViTs by demonstrating its ubiquitous existence in all three levels of patch embeddings, attention, and weights, across SOTA ViT models.

• For each of the three levels, we present diversity regularizers for training ViTs, which demonstrate complementary effects in eliminating redundancy, encouraging diversity, and enhancing generalization.

• We conduct extensive experiments with vanilla ViT, DeiT, and Swin transformer backbones on the ImageNet datasets, showing consistent and significant performance boost gains by addressing the tri-level redundancy issues with our proposed regularizers. Specifically, our proposals improve DeiT and Swin, by $0.70\% \sim 1.76\%$ and $0.15\% \sim 0.32\%$ accuracy.

2. Related Works

Vision transformer. Transformer [57] emerges from NLP applications with prevailing successes, motivating its adaptation to the computer vision scenarios [21]. ViTs encodes an image into a sequence of patches and feeds them to transformer encoders. Such self-attention based models get rid of common inductive biases, e.g., locality in convolutional neural networks (CNNs), and the global interactions among embedding grant ViTs a stronger learning capacity. ViT’s empirical successes in various computer vision tasks include image classification [11, 21, 24], object detection [5, 17, 74, 80], segmentation [60, 63, 73], enhancement [9, 68], image generation [10, 31, 46], video and vision-language understanding [15, 36–38, 44, 51, 52, 71, 77, 78].

However, the global information aggregation of ViTs also results in over-smoothed and redundant representations [23, 76]. That makes the effective learning capacity of ViTs “collapsed”, which prohibits ViTs from practically achieving higher capacity and sophisticated representations. Existing works have taken two angles: (i) (re-)injecting locality via convolution-like structures and fusing global and local contexts, for self-attention [2, 27, 32, 43, 55, 56, 64, 65, 67, 69, 76]; (ii) adopting patch-wise contrastive or mixing loss to boost diversity, for patch embeddings. Aiming to connect and expand those isolated efforts, our work is the first to target the full-scale redundancy in ViTs at embedding, attention, and weight levels. Note that our tri-level diversification framework is compatible with existing approaches by plugging them in the corresponding level. Detailed investigations are presented in Section 4.3.

Diversity regularization. Diversity constraints are designed to learn discriminative patterns for improved feature coverage and generalization [12, 23, 41]. Representative regularizers include cosine similarity-based [23], the margin or distance-based [13, 23, 33, 42, 49, 50], the hyperspherical uniformity-based [39–41], and the orthogonality-based [1, 3, 12, 25, 30, 34, 48, 58, 72]. Most of them are applied to CNNs, with [23] making the recent attempt in ViTs.

3. Methodology

3.1. Examining the Tri-Level Redundancy in ViTs

Preliminaries. Revisit that transformer architectures [21, 57] usually contain the multi-head self-attention modules (MHA) and feed-forward networks (FFN). In MHA, keys, queries, and values are linearly transformed for computing the attention heads, and then all the heads are aggregated by another linear transformation. FFNs are also built on two linear transformations with activations, as shown in Fig. 2.

Here, we use $W_{\text{MHA}}$ and $W_{\text{FFN}}$ to denote the weights in MHA and FFN modules, respectively. $A$ represents the attention map (or the affinity matrix). It is calculated by $A = \mathrm{softmax}(\alpha K^T)$, where $Q$ is the query matrix, $K$ is the key matrix, and $\alpha$ is a scale (typically $\frac{1}{\sqrt{d}}$ and $d$ is the dimension of the keys and the queries). Let $e_l^i = [e_{i,\text{class}}^l, e_{i,1}^l, \cdots, e_{i,n}^l]$ be the feature embedding of layer $l$ ($1 \leq l \leq L$), where $n$ is the total number of image patches. Without loss of generality, we take image recognition as an example. Then, the vision transformer is optimized by minimizing a classification loss $\mathcal{L}(C(e_{\text{class}}^l), y)$, where $C$ is the classification head and $y$ is the label of input samples.

Redundancy in patch embeddings. We investigate the redundancy of the feature embedding by calculating the token-wise cosine similarity. It is depicted as follows:

$$R_{\text{cosine}}^Q(h) := \frac{1}{n(n-1)} \sum_{i \neq j} \frac{|h_i^T h_j|}{\|h_i\|_2 \|h_j\|_2}, \quad (1)$$

$$R_{\text{cosine}}^D(h^1, h^2) := \frac{1}{n} \sum_i \frac{|h_i^1 h_i^2|}{\|h_i^1\|_2 \|h_i^2\|_2}, \quad (2)$$
where \( h \) is the feature embedding \( e = [e_{\text{class}}, e_1, \cdots, e_n] \) (superscript \( l \) is omitted for simplicity), and \( n \) is the total number of tokens.

Notably, \( R_{\cosine}^s(h) \) and \( R_{\cosine}^d(h^l_1, h^l_2) \) denote the cosine similarity of the feature embedding within the same layer and across two different layers \( l_1, l_2 \), respectively. The larger cosine similarity suggests more redundancy. Intuitively, the within-layer redundancy hinders ViT from capturing different tokens’ features; and the cross-layer redundancy hurts the learning capacity of ViTs since highly correlated representations actually collapse the effective depth of ViT to fewer or even single transformer layer.

**Redundancy in attentions.** We consider \( R_{\cosine}(A) \) to measure the cosine similarity of attention maps within the same layer. Similarly, \( R_{\text{MSE}}(A) := \frac{1}{n(n-1)} \sum_{i \neq j} ||A_i - A_j||^2_2 \) can also be used for the redundancy quantification. In contrast to these two metrics which show the similarity across attention heads, we further use the standard deviation statistics to indicate the element-wise variance within an attention head.

**Redundancy in model weights.** If the parameter space is highly redundant, then the weight matrix will fall approximately into a low-rank parameters subspace. Thus, we use the reconstruction error to depict the weight redundancy:

\[
R_{\text{PCA}}(W) := ||W - \hat{W}||^2_2
\]

(3)

where \( \hat{W} \) is the reconstructed weight matrix by the principal component analysis (PCA) with the top-\( k \) principal components. Given a fixed reconstruction error, the larger \( k \) implies better diversity. In other words, given \( k \), the larger reconstruction error means less weight redundancy. \([12,41]\) also dissect the weight redundancy from the view of rank.

### 3.2. Eliminating the Tri-Level Redundancy in ViTs

To mitigate the observed redundancy, we introduce three groups of regularization to encourage the diversity of \( i \) learned feature embeddings; \( ii \) attention maps; \( iii \) model weights in the training of vision transformers.

**Patch embedding diversity.** To diversify patch feature embeddings, we use the cosine angle regularization \( R_{\cosine}^s(e) \) and \( R_{\cosine}^d(e^l_1, e^l_2) \) to constrain within-layer and cross-layer embedding, respectively. Similar methods are leveraged to obtain diversified representations in vision \([23]\), language \([22]\), and graph \([8]\) scenarios. Meanwhile, we adopt the contrastive regularization \( R_{\text{contrastive}}^s(e^l_1, e^l_2) \) to boost cross-layer embedding diversity, which is presented as follows:

\[
R_{\text{contrastive}}^s(e^l_1, e^l_2) := - \frac{1}{n} \sum_{i=1}^n \log \frac{1}{\exp(e^T_i e^l_j) + \exp(e^T_i (\frac{1}{n-1} \sum_{j \neq i} e^l_j))},
\]

(4)

where \( l_1 \) and \( l_2 \) are two different layer indexes. Note that the contrastive regularizer is not applicable for the within-layer embedding diversification since the lack of positive pairs.

\( \triangleright \text{Rationale.} \) As pointed out by \([23,45]\), the cosine angle regularization can function like minimizing the upper bound of the largest eigenvalue of patch embedding \( e \), hence bringing improvements of expressiveness \([23]\) and diversity to learned representations. For contrastive regularization, it pulls embeddings corresponding to the same patch together and simultaneously pushes apart embeddings belonging to different patches, reducing the feature correlation between different layers. As a result, it enables to learn separable patch embedding and maintain tolerance to semantically similar patches \([59,61]\), improving the representation qualities and the ViT performance.
Attention diversity. In the same way, the cosine regularization $\mathcal{R}_{\text{ cosine}}(A)$ can be applied to remove the redundancy of attention, where $A = [A_1, A_2, \ldots, A_H]$ and $H$ is the number of attention heads within one layer. Inspired by the orthogonality regularization’s empirical effectiveness in vision [12, 35, 47] and language tasks [72], we investigate it under the context of ViTs. We adopt the canonical soft orthogonal regularization (SO) [3] as follows:

$$\mathcal{R}_{\text{SO}}(A) := \|A^\top A - I\|^2_F,$$  

where $\| \cdot \|_F$ is the Frobenius norm and $I$ is the identity matrix sharing the same size as $A^\top A$.

We also try an alternative Conditional number orthogonality regularizer (CNO) [12] as follows:

$$\mathcal{R}_{\text{CNO}}(A) = \|\lambda_1(A^\top A) - \lambda_2(A^\top A)\|^2.$$

It enforces the orthogonality via directly regularizing the conditional number $\kappa = \lambda_1/\lambda_2$ to 1, where $\lambda_1$ and $\lambda_2$ are the largest and smallest eigenvalues of the target matrix $A^\top A$. To make it computationally more tractable and stable, we alternatively constrain the difference between $\lambda_1$ and $\lambda_2$.

Rationale. These regularizations (i.e., SO and CNO) encourage diverse attention maps by constraining them to be orthogonal with each other, which actually upper-bounds the Lipschitz constant of learned function mappings [72], leading to robust and informative representations. As illustrated in [72], introducing an orthogonal diversity regularizer to the attention map also stabilizes the transformer training and boosts its generalization on NLP tasks.

Weight diversity. Similarly, the orthogonality regularization, e.g., $\mathcal{R}_{\text{CNO}}(W)$, can be easily plugged in and promote the diversity in ViT’s weight space. Compared to orthogonality, hyperspherical uniformity is another more general diversity regularization demonstrated in [41]. Although it has been explored in CNNs, its study in ViTs has been absent so far. We study the minimum hyperspherical separation (MHS) regularizer, which maximizes the separation distance (or the smallest pairwise distance) as follows:

$$\max_{\{\hat{w}_1, \ldots, \hat{w}_m\} \in S^{t-1}} \{\mathcal{R}_{\text{MHS}}(\hat{W}) := \min_{i\neq j} \rho(\hat{w}_i, \hat{w}_j)\},$$

where $\hat{w}_i := \frac{w_i}{\|w_i\|}$ is the ith weight vector projected onto a unit hypersphere $S^{t-1} := \{\hat{w} \in \mathbb{R}^t \|\hat{w}\| = 1\}$, $\rho(\cdot, \cdot)$ is the geodesic distance on the unit hypersphere. As indicated in Equation 7, it is formulated as a max-min optimization and we solve it with alternative gradient ascent/descent.

Furthermore, we examine another maximum gram determinant (MGD) regularizer $\mathcal{R}_{\text{MGD}}(\hat{W})$ as follows:

$$\max_{\{\hat{w}_1, \ldots, \hat{w}_m\} \in S^{t-1}} \log \det(G := (K(\hat{w}_i, \hat{w}_j))_{i,j=1}^m),$$

where $\det(G)$ is the determinant of the kernel gram matrix $G \in \mathbb{R}^{m \times m}$ and $K(\mathbf{u}, \mathbf{v}) := \exp(-\sum_{i=1}^t \epsilon^2 (u_i - v_i)^2)$ denotes the kernel function with a scale $\epsilon > 0$. By maximizing the $\det(G)$ of weights $\hat{W}$, MGD forces weight vectors to uniformly disperse over the hypersphere.

Rationale. As demonstrated in [39–41], the hyperspherical uniformity regularizations (i.e., MHS and MGD) characterize the diversity of vectors on a unit hypersphere, which encodes a strong inductive bias with relational information. We believe it to benefit ViT training from two perspectives [41]: (i) eliminating weight redundancy and improving the representational capacity; (ii) learning better optimization and generalization by reducing the spurious local minima, evidenced in [39–41, 66].

4. Experiment

Implementation details. We conduct extensive experiments on the ImageNet-1k [18] dataset with ViT [21], DeiT [53] and Swin transformers [43]. All the hyperparameters of our introduced diversity regularizations are carefully tuned by a grid search, and the best configurations are provided in Section A2. Tesla V100-SXM2-32GB GPUs are used as our computing resources. Specifically, each experiment is run with 8 V100s for 1 $\sim$ 4 days.

For vanilla ViT models, we consider two architectures with 12 layers, i.e., ViT-Small and ViT-Base, which contains 6 and 12 heads for the multi-head self-attention block in each layer, respectively. We train each model for 300 epochs with a batch size of 4096. An AdamW optimizer is adopted with 0.3 weight decay, and the learning rate starts from $4 \times 10^{-3}$ with 4 epochs for warm-up and decays by a cosine annealing schedule. We maintain all training settings the same as the original ones in [21].

For DeiT architectures, we choose DeiT-Small, DeiT-Small24, and DeiT-Base. Specifically, both DeiT-Small and DeiT-Base contain 12 layers, while DeiT-Small24 has 24 layers. In each layer, DeiT-Small and DeiT-Small24 have 6 heads for the self-attention module, and DeiT-Base has 12 heads. Following [53], we train the models for 300 epochs with the batch size of 1024. We use an AdamW optimizer of 0.05 weight decay. The initial learning rate is $1 \times 10^{-3}$ with 5 epochs of warm-up, which reduces by a cosine annealing schedule. More details about data augmentation and other training tricks can be found in [53].

For Swin transformer, we start from the official Swin-Small and Swin-Base pre-trained models and then fine-tune them for another 30 epochs, in which we use a constant $1 \times 10^{-5}$ learning rate, $1 \times 10^{-8}$ weight decay and a batch size of 1024 [23]. We keep other training details the same as [43] and compare the fine-tuning performance with/without diversity regularizations.
4.1. Redundancy in Current ViTs

In this section, we conduct a thorough investigation to reveal the broadly existing redundancy in ViT’s patch embedding, attention map and weight spaces. Specifically, 11 current SOTA ViTs and our diversified DeiT are examined on ImageNet, which can be grouped into four categories according to their proposed approaches on ViT [21]: (i) improving training techniques such as data augmentation in DeiT [53] and flatness-aware regularizer in SAM-ViT [14]; (ii) introducing convolution layers or crafting convolution-like operations like CvT [64] and Refiner [76] (termed as CNN + ViT); (iii) designing hierarchical structures to capture multi-scale information such as Swin [21], Pyramid-ViT [62], and Cross-ViT [7]; (iv) exploring finer-level features or encoding local contexts into featuring like T2T [69], TNT [27], and VOLO [70]. From Figure 3, we observe that:

1. At the patch embedding level, most of current SOTA ViTs exclude SAM-ViT have amplified embedding redundancy compared to vanilla ViT. It seems to suggest that the flatness-aware regularizer can enable ViT to generate more diverse embeddings. Furthermore, our diverse DeiT presents a consistent and substantial reduction on both layer-wise and overall embedding similarity, outperforming all other ViT variants.

2. At the attention level, our diverse DeiT again achieves superior diverse attention maps with the least redundancy. In addition, self-attention with shifted windows in Swin [21] and convolution-like aggregation in Refiner [76] effectively reduce the attention map correlation. Also, flatness-aware SAM-ViT brings more diversity across attention heads.

3. At the weight level, with similar parameter counts, Cross-ViT [7] with a dual-branch structure for extracting multi-scale features has the largest weight reconstruction error, implying more weight diversity. Besides, our diverse DeiT still consistently outperforms its vanilla counterpart as shown in Figure 3 (right), which validates our proposal’s effectiveness in eliminating the weight space redundancy.
4.2. Enhanced ViTs by Introducing Diversity

Superior generalization of our proposal. In this section, we demonstrate that improving diversity in ViT training achieves superior generalization. Specifically, our experiments consider six representative transformer backbones, i.e., vanilla ViT-Small/Base, DeiT-Small/Base and Swin-Small/Base. As shown in Table 1, several consistent observations can be drawn: ① Compared to ViT, DeiT\(^1\) and Swin baselines, our diversified variants obtain ∼ 2%, 0.70% ∼ 1.76%, and 0.15% ∼ 0.32% accuracy boosts respectively, which evidence the effectiveness of our tri-level diversity regularization. ② Vanilla ViTs tend to benefit more from our diversity-aware training, by 3 ∼ 7 times improvements on DeiT and Swin. A possible explanation is that the data augmentation in DeiT and self-attention with shifted windows in Swin have already injected a certain level of diversity, as suggested by the reduced attention and weight correlation in Figure 3. ③ Compared with existing methods on the same ViT backbone, our diverse ViTs gain 0.18% and 0.48% accuracy improvements over DiversePatch [23] and DLA [76] respectively, which indicate the necessity of reducing all levels of redundancy. ④ Deep ViTs receive more benefits from diversity regularized training. Particularly, DeiT-Small24 achieves a 1.76% accuracy increase, while its 12-layer variants have 0.83% accuracy gains. It’s within expectation since deeper transformers usually suffer harsher over-smoothing and representation redundancy [20, 79], leaving more potential advancement for diversification. Similar conclusions can be observed in [23].

Effectively reduced redundancy. In order to further validate the superiority of our diverse ViTs, we provide extensive qualitative and quantitative visualizations for all levels of patch embedding, attention maps, and model weights. ⑤ DiversePatch is the diversification on the token embedding level in [23]. ⑥ DLA is the diversification on the attention level in [76].

Table 1. Top-1 accuracies (%) of ViTs with or without all-level diversification on ImageNet. ↑ indicates the performance gains from our diversity regularization compared with its vanilla version. All ViT and DeiT transformers are trained from scratch for 300 epochs with the default setup in [21] and [53]. Swin Transformers are fine-tuned from released checkpoints from [43] for 30 epochs.

<table>
<thead>
<tr>
<th>Settings &amp; Methods</th>
<th>Size</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>(CNN)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-152 [28]</td>
<td>230M</td>
<td>78.10</td>
</tr>
<tr>
<td>DenseNet-201 [29]</td>
<td>77M</td>
<td>77.60</td>
</tr>
<tr>
<td>(CNN + ViT)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CVT-21 [64]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeiT-Small [23]</td>
<td>32M</td>
<td>82.50</td>
</tr>
<tr>
<td>Refiner [76]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ViT + DiversePatch(*))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViT-Small [21]</td>
<td>22M</td>
<td>76.54</td>
</tr>
<tr>
<td>ViT-Small + Ours</td>
<td>22M</td>
<td>76.60 (↑2.04)</td>
</tr>
<tr>
<td>ViT-Base [21]</td>
<td>86M</td>
<td>77.90</td>
</tr>
<tr>
<td>ViT-Base + Ours</td>
<td>86M</td>
<td>79.96 (↑2.06)</td>
</tr>
<tr>
<td>(Vanilla ViT 12 layers)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeiT-Small [53]</td>
<td>22M</td>
<td>79.78</td>
</tr>
<tr>
<td>DeiT-Small + Ours</td>
<td>22M</td>
<td>80.61 (↑0.83)</td>
</tr>
<tr>
<td>DeiT-Base [53]</td>
<td>86M</td>
<td>80.98</td>
</tr>
<tr>
<td>DeiT-Base + Ours</td>
<td>86M</td>
<td>81.68 (↑1.70)</td>
</tr>
<tr>
<td>(DeiT 12 layers)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swin-Small [43]</td>
<td>50M</td>
<td>83.18</td>
</tr>
<tr>
<td>Swin-Small + Ours</td>
<td>50M</td>
<td>83.33 (↑0.15)</td>
</tr>
<tr>
<td>Swin-Base [43]</td>
<td>88M</td>
<td>83.40</td>
</tr>
<tr>
<td>Swin-Base + Ours</td>
<td>88M</td>
<td>83.72 (↑0.32)</td>
</tr>
<tr>
<td>(DeiT 24 layers)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeiT-Small[24] [53]</td>
<td>43M</td>
<td>80.03</td>
</tr>
<tr>
<td>DeiT-Small24 + Ours</td>
<td>43M</td>
<td>81.79 (↑1.76)</td>
</tr>
</tbody>
</table>

\(*\) DiversePatch is the diversification on the token embedding level in [23].
\(*\) DLA is the diversification on the attention level in [76].
Figure 6. Comparison of the attention map (i.e., affinity matrix $A$) between DeiT baseline (1st row) and our diversified variant (2nd row).

Figure 7. Attention probabilities for DeiT baseline (1st row) and our diversified variant (2nd row) with 24 layers (columns) and 6 heads (rows), visualized by [16]. Attention maps are averaged over 100 test images from ImageNet. The black square is the query pixel. Zoom-in for better visibility.

Figure 8. The weight reconstruction errors over the number of selected principle components. The dash and solid lines represent our diverse ViT and its vanilla version, respectively. The smaller error suggest a larger redundancy in the original weight space.

1. **The patch embedding level.** As presented in Figure 3 (right) and Figure A10, our diverse DeiT obtains significantly reduced similarity for both within-layer and cross-layer embedding, compared to the baseline DeiT. Figure 4 visualizes the all patch embedding of 196 tokens for randomly selected images from ImageNet. Our methods show a much diverse embedding distribution and improved discriminative power. Specifically, the patch ① and ② from the left image are belong two distinct objects (i.e., human hands and dog toys) and also have totally different visual contents. However, the baseline ViT produces highly correlated patch embedding (almost fully overlapped points in Figure 4) and fails to distinguish them. Fortunately, these embedding redundancy and semantic ambiguity are well addressed by our diversity-aware ViT training. Similar observations can be drawn from patch pairs (③,④), (③,⑤), (⑥,⑩), and (⑧,⑨).

2. **The attention map level.** From Figure 3 (right), Figure A11, and Figure A12, our proposals consistently reach an enhanced diversity in terms of within-layer/cross-layer attention heads and the standard deviation within attention maps. Furthermore, we offer illustrative visualizations to show the improved representation flexibility of ViTs. Figure 5 is generated by an advanced ViT interpretable approach [6]. The heatmaps from our methods display more diverse and accurate focuses. Specifically, ours concentrate attention on more focus points that spread across the whole
Table 2. Ablation on the different combination of diversity regularizers. Experiments are conducted on DeiT-Small with ImageNet, and top-1 testing accuracy (%) is reported.

<table>
<thead>
<tr>
<th>Ablation on</th>
<th>None</th>
<th>Weight</th>
<th>Attention</th>
<th>Embedding</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeiT-Small</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>79.78</td>
</tr>
<tr>
<td>Diversity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.12</td>
</tr>
<tr>
<td>Regularization</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.13</td>
</tr>
<tr>
<td>All levels diversity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.53</td>
</tr>
</tbody>
</table>

objects like heads, legs, and main bodies. Figure 6 visualizes the attention affinity matrices $A$. We find that the attention maps from our diverse ViT training become less uniform compared to the baseline ViT, and show stronger local patterns, which provides another possible interpretation of our beneficial diversity regularization. In the end, Figure 7 dissects the fine-grained attention behaviors of certain query pixels, in which our approach shows more heterogeneous attention responses, especially in the ViT’s later layers.

The weight level. As shown in Fig. 3 (right) and Fig. 8, given a fixed number of selected principal components, the diversified ViTs have significantly larger reconstruction errors across almost all transformer layers, implying eliminated weight redundancy and improved representative power. Consistent observations are demonstrated in Fig. 9, where our diverse ViT can widely span its weight principle components on the 2-dim space.

4.3. Ablation Study

Multi-level v.s. single-level diversity. To verify the effects of diversity regularizers at different levels, we conduct an incremental evaluation on DeiT-Small with ImageNet. Achieved results are included in Table 2. We observe that: (i) Diversifying patch embedding brings the most accuracy benefits ($\sim 0.4\%$) and the other two levels contribute similarly ($\sim 0.1\%$). It is worthy of mentioning that if adopting within-layer or cross-layer embedding diversity regularization only gains $\sim 0.1\%$ accuracy while combining them leads to extra performance boosts. (ii) Reducing tri-level redundancy in patch embedding, attention maps, and weights establishes superior performance, which validates the effectiveness of multi-level diversity compared to single-level diversity. (iii) The previous useful data-level diversification, i.e., mixing loss from [23], can be easily plugged in our training framework and plays a complementary role in improving ViT’s generalization ability. More details about the mixing loss are referred to Section A1.

How to choose different diversity regularizations. As discussed in Section 3, there are various regularizer op-

Table 3. Ablation on different categories of diversity regularization. Experiments are conducted on DeiT-Small with ImageNet, and top-1 testing accuracy (%) is reported.

<table>
<thead>
<tr>
<th>DeiT-Small (79.78)</th>
<th>Weight</th>
<th>Attention</th>
<th>Embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity Regularization</td>
<td>Consine</td>
<td>N.A.</td>
<td>79.95</td>
</tr>
<tr>
<td>Uniformity Regularization</td>
<td>MHS</td>
<td>80.05</td>
<td>N.A.</td>
</tr>
<tr>
<td>Orthogonality Regularization</td>
<td>ConDO</td>
<td>80.01</td>
<td>79.90</td>
</tr>
<tr>
<td>All Reg.</td>
<td>79.80</td>
<td>80.03</td>
<td>80.09</td>
</tr>
</tbody>
</table>

Table 3, first, we find that constraining layerwise attention maps in a contrastive manner might be too aggressive to hurt the performance. While patch embedding is more amenable to diversification, all examined regularizers at the embedding level boost the ViT’s performance. Secondly, compared to strict orthogonality regularization, the general hypersphere uniformity is more preferred by the weight level of ViTs. Lastly, at the attention level, SO outperforms all other choices. A possible reason is that orthogonal attention maps not only enjoy an unleashed representation flexibility but also stabilize and smoothen ViT training due to reduced Lipschitz constant, as suggested by [72].

5. Conclusion and Broader Impact

In this paper, we, for the first time, systematically reveal the broad existence of redundancy at all token embedding, attention map, and weight levels in vision transformers, which limits the ViT’s expressiveness and flexibility. We address this issue following the principle of diversity during ViT training. Comprehensive experiments on ImageNet across diverse ViT backbones demonstrate that equipped diversity regularizers effectively eliminate the redundant representation and lead to superior generalization.

For the limitation of this work, we only focus on vision transformers rather than general transformers in natural language processing, which we leave for future works. Meanwhile, although our paper is scientific in nature, it might amplify the existing societal risk of applying ViTs since we have no control of anyone who can get access to our improved training algorithms. A potential solution is to issue licenses and limit the abuse.
References


