Scratching Visual Transformer's Back with Uniform Attention – Supplementary Material –

Contents

1. Details of Experiments Setup

This section provides information on datasets and models used in the main paper with hyper-parameters of the training.

1.1. Datasets

ImageNet-1K. ImageNet-1K [12] is the popular large-scale classification benchmark dataset, and the license is custom for research and non-commercial. ImageNet-1K consists of 1.28M training and 50K validation images with 1K classes. We use the training and the validation sets to train and evaluate architectures, respectively.

ImageNet-V2. ImageNet-V2 [11] is new test data for the ImageNet benchmark. Each of the three test sets in ImageNet-V2 comprises 10,000 new images. After a decade of progress on the original ImageNet dataset, these test sets were collected. This ensures that the accuracy scores are not influenced by overfitting and that the new test data is independent of existing models.

ImageNet-ReaL. ImageNet-ReaL [2] develops a more reliable method for gathering annotations for the ImageNet validation set and is under the Apache 2.0 license. It re-evaluates the accuracy of previously proposed ImageNet classifiers using these new labels and finds their gains are smaller than those reported on the original labels. Therefore, this dataset is called the "Re-assessed Labels (ReaL)" dataset.

ADE20K. ADE20K [18, 19] is a semantic segmentation dataset, and the license is custom research-only and non-commercial. This contains over 20K scene-centric images that have been meticulously annotated with pixel-level objects and object parts labels. There are semantic categories, which encompass things like sky, road, grass, and discrete objects like person, car, and bed.

ImageNet-A. ImageNet-A [8] is a set of images labeled with ImageNet labels that were created by collecting new data and preserving just the images that ResNet-50 [7] models failed to categorize properly. This dataset is under the MIT license. The label space is identical to ImageNet-1K.

VQAv2. VQA [1] dataset contains open-ended questions about images. These questions require an understanding of vision, language, and commonsense knowledge to answer. VQAv2 [5] is the second version of the VQA dataset, which contains 204K COCO images.

1.2. Models

Dosovitskiy *et al*. [4] have proposed ViT-B. Touvron *et al*. [13] have proposed tiny and small ViT architectures named as ViT-Ti and ViT-S. The ViT architecture is similar to Transformer [14] but has patch embedding to make tokens of images. Specifically, ViT-Ti/-S/-B have 12 depth layers with 192, 384, and 768 dimensions, respectively. Heo *et al*. [9] have proposed a variant of ViT by reducing the spatial dimensions and increasing the channel dimensions. ViTs consist of a patch embedding layer, multi-head self-attention (`MSA`) blocks, multi-layer perceptron (`MLP`) blocks, and layer normalization (`LN`) layers. Our module is the modification of `MLP` block. Our module only requires 1 line modification at the end of the `MLP` layer.

1.3. Hyper-parameters

Touvron *et al*. [13] have proposed data-efficient training settings with strong regularization, such as MixUp [16], CutMix [15], and random erasing [17]. We adopt the training setting of DeiT [13] and denote ViT-Ti/-S/-B with `CB` module. We do not use repeated augmentation for ViT-Ti/-

(a) Training Loss (b) Eval Loss

(c) Top-1 Accuracy (d) Top-5 Accuracy

Figure S1. **Training curve of ViT-S with and without CB module.** CB increases the accuracies across epochs and decreases both training and evaluation losses more. CB improves the capacity of ViT.

S with ours. For ViT-B with ours, we increase the warmup epochs from 5 to 10 and the drop path. In distillation, we use the same hyper-parameters except for the drop path of ViT-B with ours to 0.2.

2. Additional Experiments

This section provides additional experiments we cannot report due to page limitations.

2.1. Training Curve

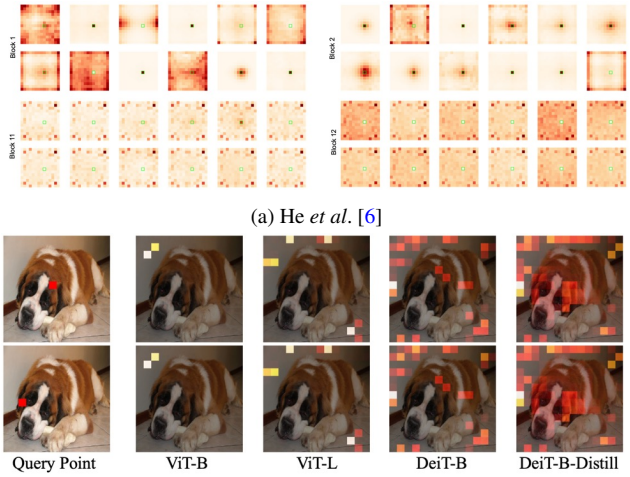
We draw the training curve to see if CB improves the capacity and convergence of ViTs. As shown in Fig. S1, CB increases the top-1/-5 accuracies across epochs and decreases both training and evaluation losses more. The curves show that CB improves the capacity of ViTs.

2.2. Distilled Performance

We follow the specification of ViT-Ti/-S from DeiT [13]. As shown in Table S1, our modules CB and CB_S improve performance compared to the distilled ViTs consistently. Table S2 shows the results of the robustness benchmark in ViT-S⚗. CB_S increases 0.5, 1.1, and 2.2 of Occ, ImageNet-A, and FGSM, respectively. CB does 0.5, 2.1, and 4.7, respectively.

2.3. Architecture generalizability.

To show further applicability of our module, We compare PVT, LocalViT-Ti, PVTv2, and Swin trained w/ and w/o ours on ImageNet-1K; we set the default epochs to 120.[1] PVT uses a pyramid structure as CNN backbones, LocalViT-Ti puts the convolution layer (local operation), PVTv2 is the improved version of the PVT placing convolutional layer, and Swin employs a hierarchical structure and local attention. Table S3 shows that our module consistently improves performance.



(a) He *et al*. [6]

(b) Ma *et al*. [10]

Figure S2. **Visualization from He *et al*. [6] and Ma *et al*. [10].** (a) He *et al*. [6] visualized the attention of DeiT-B 12 heads in shallow and deeper layers. (b) Ma *et al*. [10] visualized the attention of ViTs given the query point.

2.4. Discussion on Attention Visualization

The visualization of the attention map is not the first attempt. As shown in Fig. S2, He *et al*. [6] and Ma *et al*. [10] visualized the attention across layers or architectures. He *et al*. [6] reported that the deeper layers attend the dense global regions, and shallow layers attend the sparse local regions in DeiT-B. Ma *et al*. [10] reported that the attention weights are sparse in DeiT-B compared to DeiT-B-Distill. Despite the fact that they analyze the same architecture, DeiT-B, He *et al*. and Ma *et al*. argued the different statements on sparsity. They use different criteria depending on what they want to compare in *relative ways*. Thereby, it has been open discussion in the community about sparsity characteristics of attention maps in Transformers due to those subjective visualization-based analyses.

Based on the visualization, the statements of He *et al*. and Ma *et al*. are conditional on the reference of an attention visualization. These conditional statements can be changed by choosing a different reference; thus, our observation is not contrary to theirs. To compare more general ways, we employ an objective measure, *i.e*., entropy measure; high entropy implies dense interactions and vice versa. It is our contribution. Our entropy analysis in Sec. 3 supports the

[1] We set 150 epochs for LocalViT.

Architecture	# Params [M]	FLOPs [G]	Acc@1 [%]	Acc@5 [%]	IN-V2 [%]	IN-ReaL [%]
ViT-Ti⚗	5.9	1.3	74.5	91.9	62.4	82.1
+ CB	5.9	1.3	74.7	92.3	62.5	82.3
+ CB_S	5.9	1.3	**75.3**	**92.5**	**63.4**	**82.8**
ViT-S⚗	22.4	4.6	81.2	95.4	69.8	86.9
+ CB	22.4	4.6	81.3	**95.6**	70.2	87.0
+ CB_S	22.4	4.6	**81.6**	**95.6**	**70.9**	**87.3**
ViT-B⚗	87.3	17.6	83.4	96.4	72.2	88.1
+ CB	87.3	17.6	83.5	**96.5**	72.3	88.1
+ CB_S	87.3	17.6	**83.6**	**96.5**	**73.4**	**88.3**

Table S1. **ImageNet-1K performance.** We train vision transformer architectures [4, 13] with CB and CB_S and evaluate the accuracy on ImageNet-1K [3], ImageNet-V2 [11], and ImageNet-ReaL [2]. **Bold** is the best number at each row. Our module improves all the metrics incurring negligible extra computational costs.

Architecture	Occ [%]	ImageNet-A [%]	FGSM [%]
ViT-S⚗	74.6	21.5	11.8
+ CB_S	75.1	22.6	13.0
+ CB	**75.1**	**23.6**	**15.5**

Table S2. **Robustness evaluation.** We evaluate ViT-S⚗ with CB and CB_S on center occlusion (Occ), ImageNet-A, and fast sign gradient method (FGSM) attack. Ours shows improved robustness across the board against ViT-S⚗.

Model	Hierarchy	Local	ACC@1[%]
PVT-S	✓		76.7
+ Ours	✓		**77.3**
LocalViT-Ti		✓	69.4
+ Ours		✓	**72.5**
PVTv2-B1	✓	✓	76.4
+ Ours	✓	✓	**76.5**
Swin-Ti	✓	✓	79.3
+ Ours	✓	✓	**79.5**

Table S3. **ImageNet-1K results with hierarchical ViTs.** We further report the results of training on ImageNet-1K.

Architecture	Accuracy
ViT-S	74.73
+ CB	**75.43**

Table S4. **Accuracy on CIFAR-100.** We train ViT-S from scratch on CIFAR-100 dataset. CB increases the accuracy by 0.7%p.

visualization (Fig. 10), where CB lowers the entropy and helps MSA attend to more informative signals.

2.5. Classification on CIFAR-100

We train ViT-S from scratch on CIFAR-100. CIFAR-100 consists of 50,000 training and 10,000 validation images with 100 classes. Table S4 shows the accuracy on CIFAR-100. CB improves the performance by 0.7%p more.

2.6. Discussion on Position of CB

We conclude the position of CB to the end of the MLP block. We provide our intuition and discussion about Table 3-(b), which shows performance depends on CB position in the MLP block.

Gradient signals. We think that the gradient signals are dependent on the position of CB. For simplicity, we assume a single layer composed of the MSA and MLP block. Let the MLP layer consist as follows: $<$ Front $>$ $-$FClayer$-$ $<$ Mid $>$ $-$FClayer$-$ $<$ End $>$.

- Case 1, Front: If CB is located at Front, the subsequent weights in the corresponding MLP block cannot receive the gradient signals during training.
- Case 2, End: If CB is located at End, the preceding weights in the MLP block are updated by the gradient signals by uniform attention.

Why is the improvement of Mid and End similar? There is no non-linear function (e.g., GELU) between Mid and End positions. Since uniform attention is the addition of a globally averaged token, the output is identical wherever CB is located at Mid and End. Therefore, the accuracy of both positions is similar. Nonetheless, CB at End achieves a bit higher top-5 accuracy than CB at Mid. As aforementioned, we suspect the End position provides the gradient induced by uniform attention to weights of the MLP block.

2.7. Utilizing the Class Token

Since the class token evolves by interacting with entire tokens for tasks, we think that the class token could be uti-

Module	Position		FLOPs [M]	Acc@1 [%]	Acc@5 [%]
	MLP	MSA			
ViT-S	✗	✗	1260	79.9	95.0
+CB	✓	✗	+0.9	**80.5**	**95.3**
	✗	✓	+0.9	80.1	95.0
	✓	✓	+1.8	80.1	95.0
+$\texttt{CB}_{\texttt{gate}}$	✓	✗	+0.9	**80.4**	**95.1**
	✗	✓	+0.9	80.0	95.0
	✓	✓	+1.8	80.0	95.0
+$\texttt{CB}_{\texttt{hybrid}}$	✓	✗	+1.8	**80.5**	95.0
	✗	✓	+1.8	80.4	**95.3**
	✓	✓	+3.6	-	-

Table S5. **Performance of ViT-S with CB, $\texttt{CB}_{\texttt{gate}}$, and $\texttt{CB}_{\texttt{hybrid}}$.** We train ViT-S with CB, $\texttt{CB}_{\texttt{gate}}$, and $\texttt{CB}_{\texttt{hybrid}}$ on ImageNet-1K training set and evaluate top-1/-5 accuracies on the validation set. We vary the position of our modules at MLP, MSA, and both. **Bold** is the best number at each row.

Module	Position			FLOPS [M]	Acc@1 [%]	Acc@5 [%]
	Front	Mid	End			
ViT-S	✗	✗	✗	1260	79.9	95.0
+CB	✓	✗	✗	+0.9	79.9	94.8
	✗	✓	✗	+3.6	**80.5**	95.2
	✗	✗	✓	+0.9	**80.5**	**95.3**
+$\texttt{CB}_{\texttt{gate}}$	✓	✗	✗	+0.9	80.3	94.9
	✗	✓	✗	+3.6	80.2	**95.1**
	✗	✗	✓	+0.9	**80.4**	**95.1**
+$\texttt{CB}_{\texttt{hybrid}}$	✓	✗	✗	+1.8	**80.5**	95.0
	✗	✓	✗	+7.3	80.1	**95.1**
	✗	✗	✓	+1.8	80.3	95.0

Table S6. **Performance of ViT-S with different positions in MLP block.** MLP has following schematic: $<$ Front $>$ $-$FClayer$-$ $<$ Mid $>$ $-$FClayer$-$ $<$ End $>$. We insert CB, $\texttt{CB}_{\texttt{gate}}$, and $\texttt{CB}_{\texttt{hybrid}}$ at Front, Mid, and End and evaluation on ImageNet-1K. **Bold** is the best number at each row.

lized to complement spatial interactions of attention. We propose two additional baselines employing the class token.

The first one is the multiplication of the class token with each visual token, similar to the gating mechanism. We denote the first as $\texttt{CB}_{\texttt{gate}}$ and formalize it as follows: $\texttt{CB}_{\texttt{gate}}(\mathbf{x}_i) = \mathbf{x}_i(\mathbf{x}_0 + \mathbf{1})$ for every token i, where $\mathbf{x}_0$ is the class token and $\mathbf{1}$ is one vector. The second one is the combination of the class and average token denoted as $\texttt{CB}_{\texttt{hybrid}}$: $\texttt{CB}_{\texttt{hybrid}}(\mathbf{x}_i) = \mathbf{x}_i\mathbf{x}_0 + \texttt{CB}(\mathbf{x}_i)$ for every token i. These modules are also parameter-free and computation efficient.

Firstly, we analyze the positions of MLP and MSA; we locate the modules at the end of blocks. Table S5 lists FLOPs and validation accuracy of MLP and MSA. Both $\texttt{CB}_{\texttt{gate}}$ and $\texttt{CB}_{\texttt{hybrid}}$ improve the top-1 accuracy regardless of positions except for the failure case of $\texttt{CB}_{\texttt{hybrid}}$ at both MLP and MSA layers. These modules have the best top-1 accuracy at MLP, consistent with our CB module.

We investigate different positions in an MLP layer with $\texttt{CB}_{\texttt{gate}}$ and $\texttt{CB}_{\texttt{hybrid}}$. Table S6 lists FLOPs and validation accuracies of Front, Mid, and End. The best accuracy occurs at End for our CB module and $\texttt{CB}_{\texttt{gate}}$ and Front for $\texttt{CB}_{\texttt{hybrid}}$. At the best positions of respective modules, our CB module achieves 0.1%p higher top-1 accuracy than $\texttt{CB}_{\texttt{gate}}$ and demands half of the FLOPs than $\texttt{CB}_{\texttt{hybrid}}$.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1

[2] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 1, 3

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 3

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017. 1

[6] Haoyu He, Jing Liu, Zizheng Pan, Jianfei Cai, Jing Zhang, Dacheng Tao, and Bohan Zhuang. Pruning self-attentions into convolutional layers in single path. *arXiv preprint arXiv:2111.11802*, 2021. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[8] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 1

[9] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *ICCV*, 2021. 1

[10] Xu Ma, Huan Wang, Can Qin, Kunpeng Li, Xingchen Zhao, Jie Fu, and Yun Fu. A close look at spatial modeling: From attention to convolution. *arXiv preprint arXiv:2212.12552*, 2022. 2

[11] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 1, 3

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 1

[13] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training

data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1, 2, 3

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[15] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 1

[16] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1

[17] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 1

[18] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 1

[19] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127(3):302–321, 2019. 1