Improving Vision Transformers by Revisiting High-frequency Components (Appendix)

Jiawang Bai[1], Li Yuan[2,5,✉], Shu-Tao Xia[1,5,✉], Shuicheng Yan[4], Zhifeng Li[3,✉], and Wei Liu[3,✉]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]School of ECE at Peking University [3]Data Platform, Tencent
[4] Sea AI Lab [5]Peng Cheng Laboratory
bjw19@mails.tsinghua.edu.cn; yuanli-ece@pku.edu.cn;
xiast@sz.tsinghua.edu.cn; yansc@sea.com;
michaelzfli@tencent.com; wl2223@columbia.edu

A More Related Works

Deep neural networks (DNNs) have been showing state-of-the-art performances in various vision tasks [6,19,16,3,4,1,14,25], however, previous works mainly investigated DNNs in the spatial domain, and also it is still an open problem to understand their mechanisms. Recently, many researchers take the Fourier transformation as a tool to improve DNNs' performance or analyze their behaviors [17,23,18]. One recent work [17] investigates the CNN models from a frequency perspective. It reveals that CNN can exploit the high-frequency image components that are not perceivable to human, and demonstrates that it helps the generalization behaviors of CNN models. The method in [23] pre-processes the inputs in the frequency domain to better preserve image information and achieve improved accuracy in various vision tasks. Inspired by properties of the Fourier transformation [10,11,13], Xu *et al*. [24] proposed a novel Fourier-based data augmentation strategy called amplitude mix to help domain generalization. Also, there are some works studying ViT models from the frequency domain. Park *et al*. demonstrated that [12] multi-head self-attentions in ViT models are low-pass filters, but convolutional layers are high-pass filters, and proposed a novel architecture to combine these two operations. Theoretically, Wang *et al*. explained that [18] cascading self-attention blocks in ViT models is equivalent to repeatedly applying a low-pass filter. Besides, most recent works [8,21] solved the self-supervised visual pre-training through the lens of the frequency domain.

B Theoretical Analysis

Based on definitions of the low-pass filtering $\mathcal{M}_l^S$ and high-pass filtering $\mathcal{M}_h^S$ in Section 3, following [18], we present Theorem 1 for the attention map $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ produced by the softmax function in ViT models. It is proven based on the property of the matrix $\boldsymbol{A}$ and the details can be found in [18]. Theorem 1 reveals that the self-attention mechanism diminishes the high-frequency information with depth increasing, which further confirms our hypothesis.

Theorem 1. $\lim_{k\to\infty}\frac{||\mathcal{M}_h^{n-1}(\boldsymbol{A}^k\boldsymbol{v})||_2}{||\mathcal{M}_l^1(\boldsymbol{A}^k\boldsymbol{v})||_2}=0$ *for any vector* $\boldsymbol{v}\in\mathbb{R}^n$*, where* $\boldsymbol{A}^k$ *denotes the product of* $\boldsymbol{A}$ *with itself* k *times.*

C Algorithm Pipeline

Algorithm 1 Training a ViT model with HAT for one epoch

Input: Training set $X=\{(\boldsymbol{x},\boldsymbol{y})\}$, model weights $\boldsymbol{\theta}$, learning rate τ, maximum perturbation strength ϵ, number of PGD steps K, PGD step size η

for minibatch $B\subset X$ **do**
  $\boldsymbol{\delta}_0\leftarrow\boldsymbol{0}$
  for $t=1...K$ **do**
    ▷ Compute gradients of model weights and perturbations simultaneously
    if $t=1$ **then**
      $\nabla_{\boldsymbol{\theta}},\nabla_{\boldsymbol{\delta}}\leftarrow\nabla L\big(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}_0,\boldsymbol{y}\big)$
    else
      $\nabla_{\boldsymbol{\theta}},\nabla_{\boldsymbol{\delta}}\leftarrow\nabla\frac{1}{K-1}\left[\alpha L(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}_{t-1},\boldsymbol{y})+\beta L_{kl}(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}_{t-1},\boldsymbol{x})\right]$
    end if
    $\boldsymbol{g}_t\leftarrow\boldsymbol{g}_{t-1}+\nabla_{\boldsymbol{\theta}}$ ▷ Accumulate gradients of model weights
    $\boldsymbol{\delta}_t\leftarrow\mathrm{clip}(\boldsymbol{\delta}_{t-1}+\eta\cdot\mathrm{sign}(\nabla_{\boldsymbol{\delta}}),-\epsilon,\epsilon)$ ▷ Update and clip perturbations
  end for
  $\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}-\tau\cdot\boldsymbol{g}_K$ ▷ Update model weights
end for

D Implementation Details

D.1 Combining HAT with Knowledge Distillation

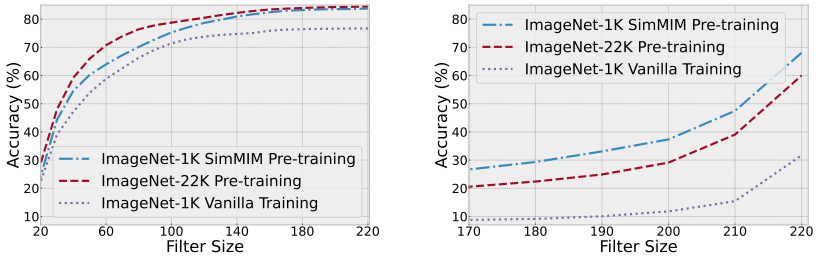
We combine HAT with knowledge distillation by optimizing the loss function, which is obtained by replacing the cross-entropy loss in Eq. (4) with the distillation loss used in DeiT [15], as follows:

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\left[L_{dist}\big(\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{y},\boldsymbol{y}_t\big)+\max_{||\boldsymbol{\delta}||_\infty\leqslant\epsilon}\Big(\alpha L_{dist}\big(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta},\boldsymbol{y},\boldsymbol{y}_t\big)+\beta L_{kl}\big(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta},\boldsymbol{x}\big)\Big)\right],$$

where $L_{dist}\big(\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{y},\boldsymbol{y}_t\big)=\frac{1}{2}\mathrm{CE}(f_{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{y})+\frac{1}{2}\mathrm{CE}(f_{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{y}_t)$ is the distillation loss and $\boldsymbol{y}_t$ is the hard decision of the teacher model. We also keep the optimization and all hyper-parameters unchanged.

D.2 Object Detection and Instance Segmentation

We take three variants of Swin Transformer trained without and with our HAT as pre-trained models to evaluate the performance in object detection and instance



(a) Low-pass Filtering (b) High-pass Filtering

Fig. I: Comparison of vanilla training and two pre-training methods on low- and high-pass filtered validation set with different filter sizes. The top-1 accuracy of ImageNet-1K Vanilla Training, ImageNet-22K Pre-training, and ImageNet-1K SimMIM Pre-training on the ImageNet validation set is 76.7%, 84.5%, and 83.8%, respectively.

segmentation. The experiments are conducted on COCO 2017 [7], which contains 118K training, 5K validation, and 20K test-dev images. We use the Cascade Mask R-CNN object detection framework [2,5] with multi-scale training (resizing the input such that the shorter side is between 480 and 800 while the longer side is at most 1,333), AdamW optimizer (initial learning rate of 0.0001, weight decay of 0.05, and batch size of 16), and 3x schedule (36 epochs). Our implementation is based on Swin Transformer and more details can be found in [9].

D.3 Semantic Segmentation

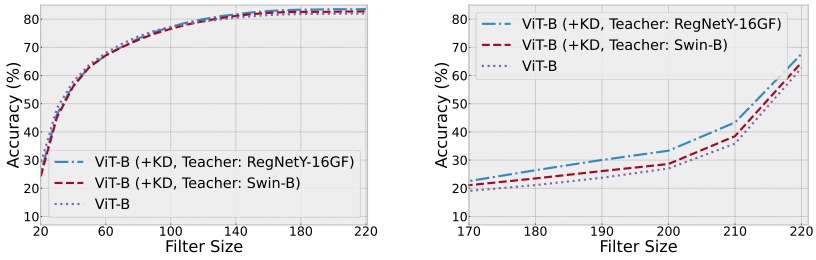
We also use Swin Transformer to evaluate the performance in semantic segmentation. We report results on the widely-used segmentation benchmark ADE20K [26], where ADE20K contains 25K images in total, including 20K images for training, 2K images for validation, and 3K images for test. And the UperNet [20] is selected as the segmentation framework. In training, we follow the setup of the original paper [9]. Specifically, we utilize the AdamW optimizer with an initial learning rate of 6×10^{-5} and a weight decay of 0.01, and we set the linear learning schedule with a minimum learning rate of 5×10^{-6}. Models are trained on 8 GPUs with 2 images per GPU for 160K iterations. In inference, we perform multi-scale test with interpolation rates of $[0.75, 1.0, 1.25, 1.5, 1.75]$.

E More Frequency Analysis

E.1 Large-scale Pre-training

We analyze supervised (i.e., ImageNet-22K) and self-supervised (i.e., SimMIM [22]) pre-training in Figure I. It shows that both are helpful for exploiting high-frequency components, especially for SimMIM.

Besides, we directly use `HAT` in fine-tuning under the SimMIM pre-training setting. Specifically, we perform adversarial training in the first 80 epochs for



(a) Low-pass Filtering (b) High-pass Filtering

Fig. II: Comparison of ViT-B and ViT-B (+KD) with different teacher models on low- and high-pass filtered validation set with different filter sizes. The top-1 accuracy of ViT-B, ViT-B (+KD, Teacher: Swin-B), and ViT-B (+KD, Teacher: RegNetY-16GF) on the ImageNet validation set is 82.0%, 82.8%, and 83.6%, respectively.

ViT-b and normal training in the rest 20 epochs, and keep other settings unchanged. Training with `HAT` results in an 83.9% top-1 accuracy, which is better than an 83.8% top-1 accuracy reported in its paper and an 83.6% top-1 accuracy in our reproduction without `HAT`.

E.2 Knowledge Distillation

Here we also analyze the knowledge distillation with CNN (*i.e.*, RegNetY-16GF) and ViT (*i.e.*, Swin-B) teacher in Figure II. It indicates that, compared to the ViT teacher, the CNN teacher is more helpful for ViT-B to capture the high-frequency components.

References

1. AlQuraishi, M.: Alphafold at casp13. Bioinformatics **35**(22), 4862–4865 (2019)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: CVPR (2018)
3. Deng, Z., Peng, X., Li, Z., Qiao, Y.: Mutual component convolutional neural networks for heterogeneous face recognition. IEEE Transactions on Image Processing **28**(6), 3102–3114 (2019)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1) (2019)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
7. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
8. Liu, H., Jiang, X., Li, X., Guo, A., Jiang, D., Ren, B.: The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training. arXiv preprint arXiv:2204.08227 (2022)
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: CVPR (2021)
10. Oppenheim, A., Lim, J., Kopec, G., Pohlig, S.: Phase in speech and pictures. In: ICASSP (1979)
11. Oppenheim, A.V., Lim, J.S.: The importance of phase in signals. Proceedings of the IEEE **69**(5), 529–541 (1981)
12. Park, N., Kim, S.: How do vision transformers work? In: ICLR (2022)
13. Piotrowski, L.N., Campbell, F.W.: A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. Perception **11**(3), 337–346 (1982)
14. Qiu, H., Gong, D., Li, Z., Liu, W., Tao, D.: End2end occluded face recognition by masking corrupted features. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
15. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
16. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: CVPR (2018)
17. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In: CVPR (2020)
18. Wang, P., Zheng, W., Chen, T., Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In: ICLR (2022)
19. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative deep feature learning approach for face recognition. In: ECCV (2016)
20. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018)
21. Xie, J., Li, W., Zhan, X., Liu, Z., Ong, Y.S., Loy, C.C.: Masked frequency modeling for self-supervised visual pre-training. arXiv preprint arXiv:2206.07706 (2022)
22. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: CVPR (2022)

23. Xu, K., Qin, M., Sun, F., Wang, Y., Chen, Y.K., Ren, F.: Learning in the frequency domain. In: CVPR (2020)
24. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: CVPR (2021)
25. Yang, X., Jia, X., Gong, D., Yan, D.M., Li, Z., Liu, W.: Larnet: Lie algebra residual network for face recognition. In: International Conference on Machine Learning. pp. 11738–11750. PMLR (2021)
26. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**(3), 302–321 (2019)