Improving Vision Transformers on Small Datasets by Increasing Input Information Density in Frequency Domain

Xiangyu Chen[1], Ying Qin[2]*, Wenju Xu[3]*, Andrés M. Bur[4], Cuncong Zhong[1], Guanghui Wang[5]
[1]Department of EECS, University of Kansas, KS, USA
[2]Institute of Information Science, Beijing Jiaotong University, Beijing, China
[3]OPPO US Research Center, InnoPeak Technology Inc, Palo Alto, CA, USA
[4]Department of Otolaryngology–Head and Neck Surgery, University of Kansas, Kansas City, Kansas, USA
[5]Department of CS, Ryerson University, Toronto ON, Canada

Abstract

Vision Transformer has attracted a lot of attention recently due to its superior ability to catch long-term dependencies. However, it is very hard to train the Transformer from scratch on small-scale datasets. To tackle this problem, we propose to solve the problem from the perspective of information density in the frequency domain. Observing that the major difference between languages (where Transformer originally comes from) and images is the information density, to feed vision Transformers with a high-density input as languages, multi-scale DCT based dense tokens are created and a heatmap based frequency channel selection strategy is proposed to reduce redundant frequency channels when forming single scale dense DCT input. Experimental results demonstrate that the proposed strategy significantly improves the performance of vision Transformers trained on small datasets for image classification.

1. Introduction

Transformer, a dominant architecture in natural language processing (NLP), has become a more and more powerful competitor architecture of Convolutional neural networks (CNNs) since ViT [1] defeated ResNet on ImageNet1k in 2020. However, it shows relatively inferior performance on small datasets [4]. As mentioned in MAE [2], the major difference between languages and images is that languages have higher information density. This study aims to increase the input information density of images and improves the image interpretability for deep learning algorithms. To this end, an input channel-wise heatmap-based frequency channel selection is proposed to reduce the DCT input from 192 channels to 24 channels. Since most of the selected channels are of low-frequency that hold most information of the entire image, the input information density increases largely by keeping most information while reducing the DCT channels from 192 to 24, resulting 87.5% reduction of the spatial area of the DCT input tensor.

The contributions of this work can be summarized below: (i) We first introduce the concept of input information density in deep learning algorithms and propose a DCT-based dense token for vision Transformers to close the gap between languages and images input. The framework to feed DCT input to Transformers requires neither convolutional layer nor linear layer on the input and increases the model's explainability. (ii) We propose a simple yet effective channel-wise heatmap-based strategy to select useful DCT-based frequency channels for the input. The channel-wise heatmap clearly shows the sparsity of images, following which customized image-wise / category-wise channel selection can be developed to further reduce the large redundancy caused by the compromise of static channel selection.

2. Method

The proposed preprocessing pipeline is illustrated in Fig. 1, including resizing, block DCT, and channel selection module guided by the input channel-wise heatmap. Following a similar process, we generate frequency representations from 8 and 6 DCT transformation and then concatenate them to a $54 \times 56 \times 56$ multi-scale frequency representation. Finally, we pass this integrated representation to an attention module. In the following, we present our input channel-wise heatmap-based frequency channel selection strategy and some basis for input information density.

To evaluate how each input contributes to the final high-level features, we mask out all channels except the candidate channel and then calculate the GradCAM heatmap, *i.e.*, channel-wise heatmap. Specifically, to calculate the channel-wise heatmap, we first train a ResNet50 with all

*Equal Contributions

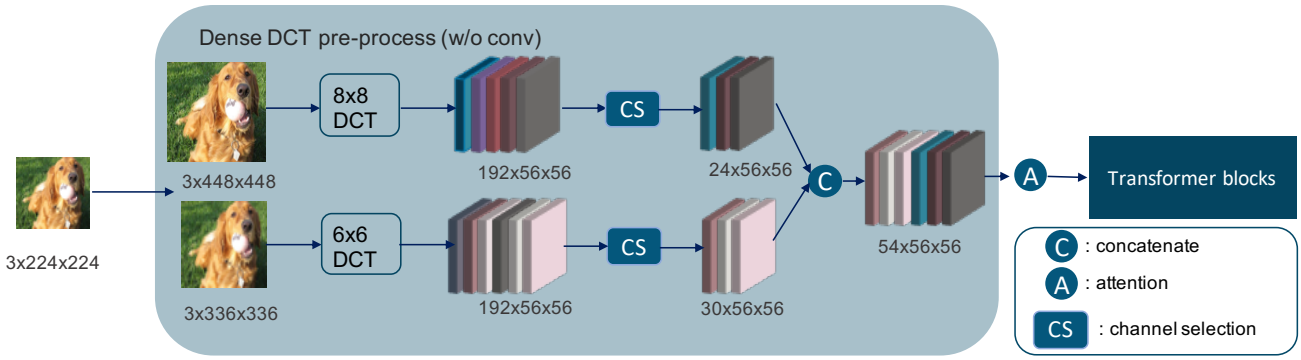

Figure 1. Dense DCT based vision Transformer.

Backbone	Param.#	CFR10	CFR100
ResNet50 [3]	25M	92.77	73.24
Deit-S [6]	22M	81.79	59.11
T2T-ViT-14 [8]	22M	87.58	67.22
Swin-Tiny [5]	28M	81.91	62.30
DenseDCT_Swin-T	28M	84.76	67.62
		(+2.85)	(+5.32)
Focal-Tiny [7]	29M	88.31	71.58
DenseDCT_Focal-Tiny	29M	90.26	**74.44**
		(+1.95)	(+2.86)

Table 1. Top-1 classification accuracy on CIFAR-10/100 dataset.

frequency channels. Then, calculate the heatmaps with GradCAM based on the learned model by setting the DCT input tensor to zeros except for the candidate frequency channel. Finally, we average the channel-wise heatmaps from all training images and average across all positions for each channel. To obtain the final selected channels, we apply a thresholding approach to the heatmap values.

3. Experiments

CIFAR-10 and CIFAR-100 have a small image size of 32×32, compared with ImageNet, holding less original information. We first test our dense DCT token with Swin Transformer [5] and Focal Transformer [7]. From Tab. 1, we can see an obvious performance gap between ResNet50 [3] and vision Transformers, including Swin [5] and Focal Transformer [7] on CIFAR-10 and CIFAR-100. While with the proposed dense DCT input, the gap is significantly reduced and the performance is close to ResNet. For Swin Transformer, our dense DCT input achieves 2.86% improvement on CIFAR-10 compared with RGB input. For CIFAR-100, the performance is improved by 5.32%.

4. Conclusion

In this paper, we have proposed a practical strategy to increase the performance of vision Transformers on small datasets by increasing the input information density. We have designed a simple yet effective channel selection method to choose functional input frequency channels by analyzing the input channel-wise heatmap. We further integrate a multi-scale frequency features from both 6×6 and 8×8 DCTs to obtain more useful frequency components and add an attention module on the input DCT representation directly to increase input SNR.

References

[1] Alexey Dosovitskiy, Lucas Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 1

[2] Kaiming He, Xinlei Chen, et al. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, pages 770–778, 2016. 2

[4] Yahui Liu, Enver Sangineto, et al. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[5] Ze Liu, Yutong Lin, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ICCV*, pages 10012–10022, 2021. 2

[6] Hugo Touvron, Matthieu Cord, et al. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357. 2021. 2

[7] Jianwei Yang, Chunyuan Li, et al. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 2

[8] Li Yuan, Yunpeng Chen, et al. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In *Proceedings of ICCV*, pages 558–567, 2021. 2