-Supplementary Material-
Channel Propagation Networks for Refreshable Vision Transformer

Junhyeong Go
Ajou University
gojunhyeong6545@gmail.com

Jongbin Ryu *
Ajou University
jongbinryu@ajou.ac.kr

Table 1. Ablation studies on various design approaches for the 24-layer CP-DeiT. Except for the baseline model, the size of the initial embedding dimension is uniformly set to 288.

Methods	RC	Heads	# param. (M)	Top-1 Acc. (%)
DeiT-S24 [9]	-	6	43.3	77.4
(a)	8	4	47.4	77.3
(b)	12	4	45.3	77.8
(c)	16	4	46.2	**79.3**

Table 2. Performance on the ImageNet-1k with 100 epoch training.

Methods	#Param (M)	FLOPs (G)	Throughput (img / s)	Top-1 Acc.
Swin-T	28.3	4.5	871.4	77.6
CP-Swin-T	20.1	4.2	924.1	78.1

1. Further Ablation Study

Deeper Vision Transformer. The evaluation of the proposed Channel Propagation method is performed by various design approaches. Fig. 1 illustrates various design approaches for the ViT model, where each block incorporates MHSA and FFN operations. Tab. 1 shows the performance of the architectural designs of Fig. 1.

Comparable computational budget. Tab. 7 in the manuscript shows that the lightweight CP modules achieve better performance while reducing the computational cost. Furthermore, we redesign the competitive CP-Swin-T by tuning hyperparameters. As shown in Tab. 2, CP-Swin-T achieves lower FLOPs and faster inference speed than the baseline while maintaining higher performance.

Refreshable Channel Dimension. We perform the ablation study to compare the performance of the refreshable channels dimension and the initial embedding channel dimension of a network. Tab. 3 shows that our Channel Propagation consistently outperforms the baseline, and even with smaller initial channel sizes, larger refreshable channels lead to greater performance improvement.

*Corresponding author.

Table 3. Ablation studies on the dimensionality of Refreshable Channels(RC). The number of heads is set proportionally to the channel dimension. We utilize Swin-S [5] as the baseline network with 100 epoch training setup.

Init dim.	RC	Heads	#Param. (M)	Top-1 (%)	throughput (image / s)
96	0	{3, 6, 12, 24}	49	81.0	524
128	16	{4, 4, 8, 16}	37	81.5	482
100	20	{4, 4, 10, 20}	43	81.7	453
72	24	{4, 6, 8, 24}	51	81.9	449

Table 4. Experimental results on the CIFAR-100 with 300 epochs.

Methods	#Param (M)	FLOPs (G)	Top-1 Acc.
ConvNeXt-T [6]	28.6	1.1	79.2±0.19
ConvNeXt-S [6]	50.2	2.2	80.1±0.09
CP-ConvNeXt-T	27.5	1.7	81.2±0.22
CP-ConvNeXt-S	53.2	3.0	82.0±0.25

Channel propagation in modern CNN. We evaluate the performance of CP-ConvNeXt on the CIFAR-100. As shown in Tab 4, our CP improves the performance of the ConvNeXt. The experiments are conducted three times using different seeds. Also, we perform spatial entropy and frequency analysis on CNNs. As shown in Fig. 2, unlike ViTs, CNNs do not present feature redundancy. The reason behind this is the hierarchical structure of CNN, which leads to an incremental increase in the receptive field through convolution operations [8].

Efficiency of Channel Propagation. We compare ours with baseline networks without the CP. We only remove the CP from our CP-DeiT-S while keeping all others the same. Tab. 5 shows that our CP w/ concat significantly improves the performance with manageable additional resources. We also compare our concat method with a skip connection (residual operation). We replace the concat operation in

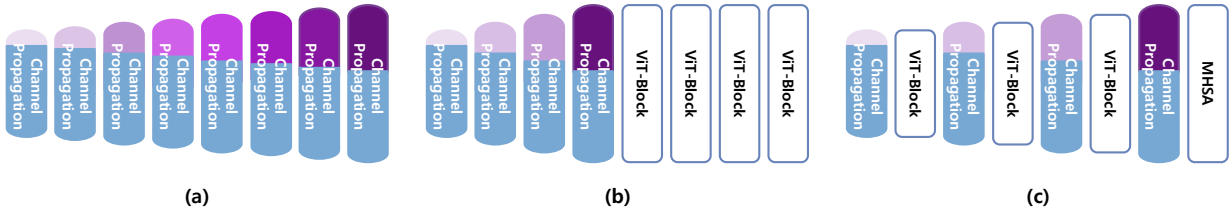

Figure 1. Three architectural designs. (a): Channel Propagation only, (b): Sequential, and (c): Crossover design.

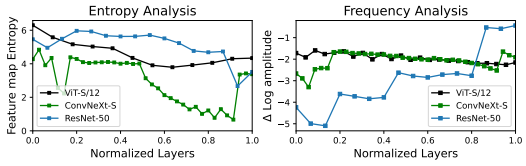

Figure 2. Entropy and frequency analysis on CNNs. We compare them using ViT, ResNet, and ConvNeXt.

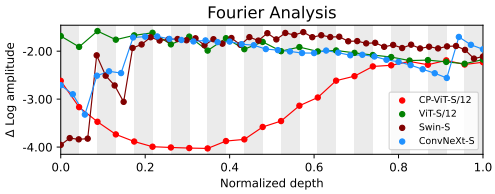

Figure 3. Fourier frequency analysis on the latest networks.

Table 5. Results for CP efficiency on CP-DeiT-S networks.

Methods	#Param	FLOPs	Top-1 Acc.	Throughput
Baseline w/o CP	23.4	4.8	76.6	1160
CP w/ skip-conn.	23.7	5.9	81.8	1034
CP w/ concat	24.6	6.1	82.5	1015

Eq.5 with the skip connection in the CP block. Therefore, instead of incrementally increasing the channel dimension from 288 to 480 with the concat, we use skip-connection with the CP block in a network with fixed 384 channel dimensions. Tab. 5 shows our CP-DeiT-S with concat performs better than CP-DeiT-S with the skip-connection.

Further Frequency Analysis. We perform frequency analysis on more models. As shown in Fig. 3, both hierarchical models, Swin-S [5] and ConvNeXt-S [6], learn diverse frequency levels. Despite being based on a plain ViT, the proposed Channel Propagation method can capture a wide range of frequency levels as the hierarchical models.

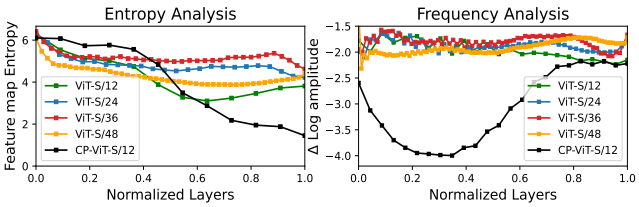

Figure 4. Analysis of the **Left**) Entropy, **Right**) Frequency. Except for our CP-ViT, all ViT networks do not show diversified entropy and frequency levels for different layers of each network.

Feature similarity in deeper ViT. The model's high feature similarity poses a challenge in capturing the distinct features of different tokens during training, which ultimately reduces the network's capacity. We perform experiments on ViTs with depths of 12, 24, 36, and 48 layers, where we analyze feature map entropy and frequency. As depicted in Fig. 4, deeper networks tend to capture similar features more often, which aligns with a decrease in performance as shown in Tab. 6. We utilize the training recipe from CaiT [10] to implement different drop path ratios for each depth.

Skip connection in U-Net. The skip connection in U-Net helps bridge the gap between the encoder and decoder, allowing for the recovery of detailed information. While Channel Propagation tackles redundancy by fusing information between adjacent layers, U-Net still encounters redundancy problems as it mainly binds distant features between the encoder and decoder [4].

2. Details of the analysis.

We follow [8] for the frequency analysis, where we use the feature maps in the 2D frequency domain using the Discrete Fourier Transform. For ViT, we make use of the output from the MHSA and FFN layers. On the other hand, with CNNs, we rely on the output from the convolutional blocks. Initially, the magnitude of the feature map frequency is extracted, and then log scaling is applied. The frequency

Table 6. Performance comparison of deeper networks on the ImageNet-1K dataset.

Methods	#Param (M)	FLOPs (G)	Top-1 Acc. (%)
ViT-S/12	22	4.6	79.8
ViT-S/24	43	9.2	80.2
ViT-S/36	65	13.7	79.8
ViT-S/48	86	18.3	76.9

Table 7. Architecture configuration of CP-Swin-T. The size of the refreshable channel is configured by 30 in this configuration.

	Spatial size	Channel size	# Heads
Stage-1	56×56	$100 \rightarrow 130$ $130 \rightarrow 160$	4
Stage-2	28×28	$160 \rightarrow 190$ $190 \rightarrow 220$	4
Stage-3	14×14	$220 \rightarrow 250$ $250 \rightarrow 280$ $280 \rightarrow 310$ $310 \rightarrow 340$ $340 \rightarrow 370$ $370 \rightarrow 400$	5
Stage-4	7×7	$400 \rightarrow 430$ $430 \rightarrow 460$	10

features are divided into lowest- and highest-frequency elements. We compute the difference in logarithmic amplitude of the two elements.The analysis presents the mean values obtained from 32 images.

3. Architectural Design Choice

We illustrate the specific design choices for CP-Swin-T in Tab. 7. For the design of CP-Swin-T, we employ a refreshable channel of size 30. In other words, a layer-specific new channel is added for each depth. The channel size represents the size of the input and output channels. Specifically, the arrow on the left indicates the size of the input channel, while the arrow on the right indicates the output channel size. The number of heads is adopted based on the results regarding the channel size.

4. Hyperparameter details

We provide the details of the hyperparameters used in our experiments as shown in Tab. 8.

Table 8. Training recipe for ViT models.

Networks	DeiT [9] & PiT [2]	Swin [5]
Epochs	300	300
Batch size	1024	1024
Optimizer	AdamW [7]	AdamW
Learning rate	1e-3	1e-3
Learning rate decay	cosine	cosine
Weight decay	0.05	0.05
Warmup epochs	5	20
Stoch. Depth[1] [3]	0.1	0.3
Gradient clip.	✗	✓
Rand Aug. [1]	9/0.5	9/0.5
Mixup [12]	0.8	0.8
Cutmix [11]	1.0	1.0
Color jitter	0.4	0.4
Decay epochs	30	30
EMA	✗	✓

References

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 3

[2] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11936–11945, 2021. 3

[3] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016. 3

[4] Wei Hao Khoong. Busu-net: An ensemble u-net framework for medical image segmentation. *arXiv preprint arXiv:2003.01581*, 2020. 2

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1, 2, 3

[6] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1, 2

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[8] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 1, 2

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 1, 3

[10] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021. 2, 3

[1] To train a 24-depth model, we set the stochastic depth ratio to 0.2, following [10].

[11] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[12] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3