Spatial-Channel Token Distillation for Vision MLPs

Yanxi Li[1,2] Xinghao Chen[2] Minjing Dong[1,2] Yehui Tang[2,3] Yunhe Wang[2*] Chang Xu[1*]

[1] School of Computer Science, University of Sydney, Australia
[2] Huawei Noah's Ark Lab
[3] School of Artificial Intelligence, Peking University, China

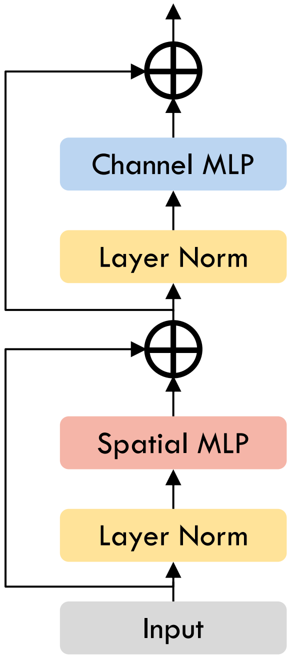
[*] Corresponding Authors

Vision MLPs

Vision MLPs typically split an image into small patches and mix features along two dimensions: 1) the **spatial MLP** layers mix feature across different spatial locations and share weights among channels, and 2) the **channel MLP** layers mix features across channels at a given spatial location and share weights among locations:

$$\boldsymbol{U}^{(l)} = \mathrm{MLP}_S^{(l)}(\mathrm{LN}(\boldsymbol{Z}^{(l-1)})) + \boldsymbol{Z}^{(l-1)},$$
$$\boldsymbol{Z}^{(l)} = \mathrm{MLP}_C^{(l)}(\mathrm{LN}(\boldsymbol{U}^{(l)})) + \boldsymbol{U}^{(l)},$$

where $l = 1, \dots, L$ are L blocks.



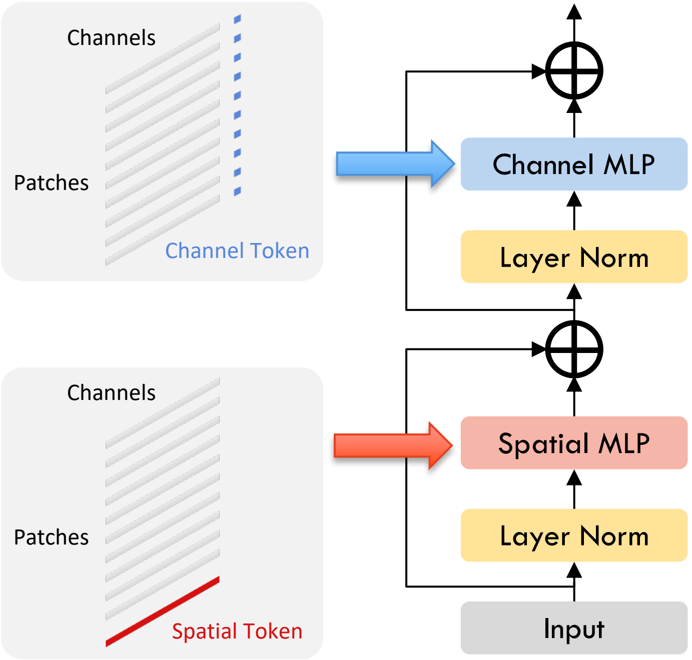
Vision MLPs

- Vision models with pure MLPs are hard to train: MLP-Mixer requires costly pre-training on large-scale datasets, such as ImageNet-21K and JFT-300M.
- One possible way to solve this problem is to design complex architectures: ResMLP, CycleMLP.
- In this work, we seek for another solution: knowledge distillation.

Spatial-channel Token Distillation

Based on the spatial-channel paradigm of Vision MLPs, we propose a novel **Spatial-channel Token Distillation** (STD) mechanism:

$$T_S^{(k)} = \mathrm{MLP}_S^{(k)}(\mathrm{LN}([\boldsymbol{Z}^{(l)}||T_S^{(k-1)}])) + T_S^{(k-1)}$$
$$T_C^{(k)} = \mathrm{MLP}_C^{(k)}(\mathrm{LN}([\boldsymbol{Z}^{(l)}||T_C^{(k-1)}])) + T_C^{(k-1)}$$



Mutual Information Regularization

We design a **Mutual Information Regularization** (MIR) term to disentangle the spatial and channel information. The MI is a measure of dependence between random variables based on the Shannon entropy:

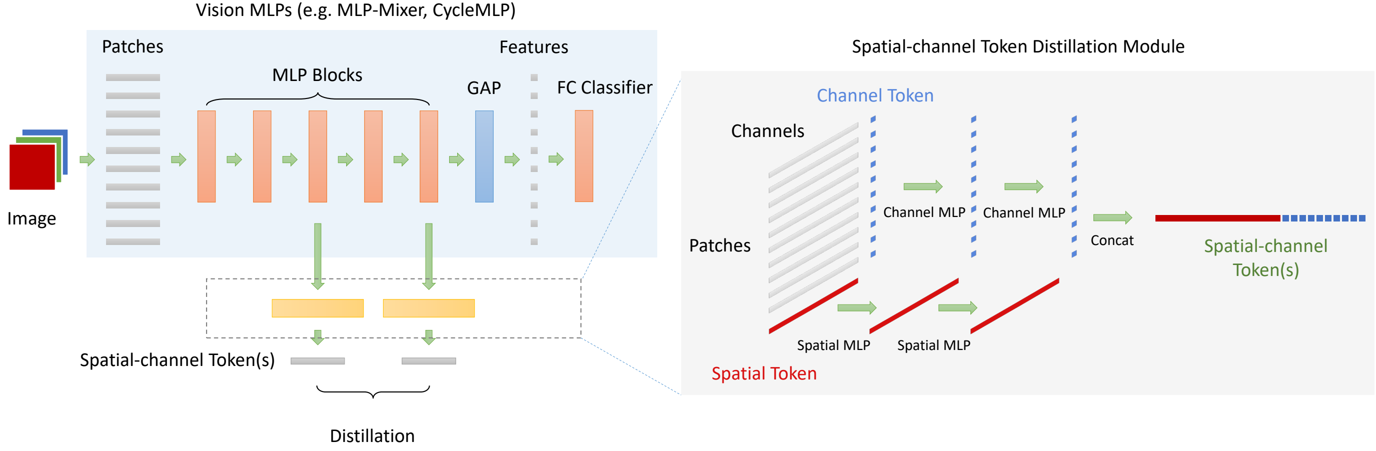
$$I(X;Y) := D_{KL}(\mathbb{P}_{XY} || \mathbb{P}_X \otimes \mathbb{P}_Y),$$

where $D_{KL}(\cdot \, || \, \cdot)$ is the KL-divergence.

We use **Mutual Information Neural Estimation** (MINE, Belghazi et al., 2018) to efficiently estimates the MI:

$$I_\Theta(X;Y) = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XY}} [\psi_\theta] - \log \left(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y} \left[e^{\psi_\theta} \right] \right).$$

The Overall Pipeline



By inserting different number of tokens to different positions, STD is suitable for:

- Single-teacher and multi-teacher distillation;
- Last-layer and intermediate-layer distillation.

Multi-teacher Distillation

	Teachers			Student
Archtecture	ResNet-50	ResNet-101	Swin-B/224	Top-1 Acc. (%)
Params (M)	25.58	44.57	87.77	
FLOPs (G)	4.36	8.09	15.14	
Selection	✓	✗	✗	81.47
	✗	✗	✓	81.91
	✓	✓	✗	81.96

Distilling with two ResNet teachers can improve the performance of the student model and can reach competitive performance to distilling with a single large Swin Transformer, even though the Swin Transformer has more parameters and FLOPs than the sum of the two ResNet teachers.

Intermediate-layer Distillation

Teachers			Student
Architecture	ResNet-50	ResNet-101	Top-1 Acc. (%)
Position	Last Inter	Last Last	81.96 82.09

By moving one pair of spatial-channel distillation tokens to the intermediate layer of the student model, we can distill the shallow layers with the shallow ResNet-50 teacher and the deep layers with the deep ResNet-101 teacher. This further improves the performance of the student model.

Comparison with SOTAs

Model	Params (M)	FLOPs (G)	Top-1 Acc. (%)
CNN			
ResNet-18 (He et al., 2016)	12.5	1.8	69.8
ResNet-50 (He et al., 2016)	22.0	4.1	78.9
RSB-ResNet-18 (Wightman et al., 2021)	12.5	1.8	71.5
RSB-ResNet-50 (Wightman et al., 2021)	22.0	4.1	80.4
Transformer-based			
ViT-B/16/384 (Dosovitskiy et al., 2021)	86.0	-	77.9
ViT-L/16/384 (Dosovitskiy et al., 2021)	307.0	-	76.5
DeiT-Ti (Touvron et al., 2021b)	6.0	-	74.5
DeiT-S (Touvron et al., 2021b)	22.0	-	81.2
DeiT-B (Touvron et al., 2021b)	87.0	-	83.4
MLP-like			
Mixer-S16 (Tolstikhin et al., 2021)	18.5	3.8	72.9
+ JFT-300M	18.5	3.8	73.8 (+0.9)
+ DeiT Distillation (Touvron et al., 2021b)	20.0	3.8	74.2 (+1.3)
+ **STD (ours)**	22.2	4.3	**75.7** (+2.8)
Mixer-B16 (Tolstikhin et al., 2021)	59.9	12.7	76.4
+ JFT-300M	59.9	12.7	80.0 (+3.6)
+ ImageNet-21K	59.9	12.7	80.6 (+4.2)
+ **STD (ours)**	66.7	13.7	80.0 (+3.6)
ResMLP-S24 (Touvron et al., 2021a)	30.0	6.0	79.4
+ **STD (ours)**	32.5	6.2	**80.0** (+0.6)
ResMLP-B24 (Touvron et al., 2021a)	115.7	23.0	81.0
+ **STD (ours)**	122.6	24.1	**82.4** (+1.4)
CycleMLP-B1 (Chen et al., 2021)	15.2	2.1	78.9
+ **STD (ours)**	18.4	2.2	**80.0** (+1.1)
CycleMLP-B2 (Chen et al., 2021)	26.8	3.9	81.6
+ DeiT Distillation (Touvron et al., 2021b)	28.6	3.9	81.9 (+0.3)
+ **STD (ours)**	30.1	4.0	**82.1** (+0.5)

We compare Vision MLPs distilled by our STD to CNNs, Transformers, and MLPs with similar number of parameters and FLOPs on the ImageNet-1K dataset. We also report the results of Vision MLPs with large-scale pre-training and DeiT distillation.

Conclusion

- We propose a novel **Spatial-channel Token Distillation** (STD) mechanism specially designed for vision MLPs:
  - adding **distillation tokens** into both the spatial and channel dimension of MLP blocks to improve the spatial and channel mixing,
  - utilizing a **mutual information regularization** to disentangle the spatial and channel information.
- STD is suitable for:
  - **last-layer** and **intermediate-layer** distillation,
  - **single-teacher** and **multi-teacher** distillation.

Thank you!