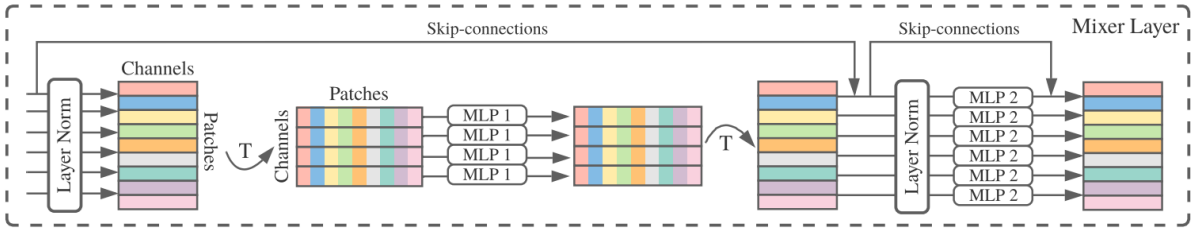
ICML
International Conference
On Machine Learning

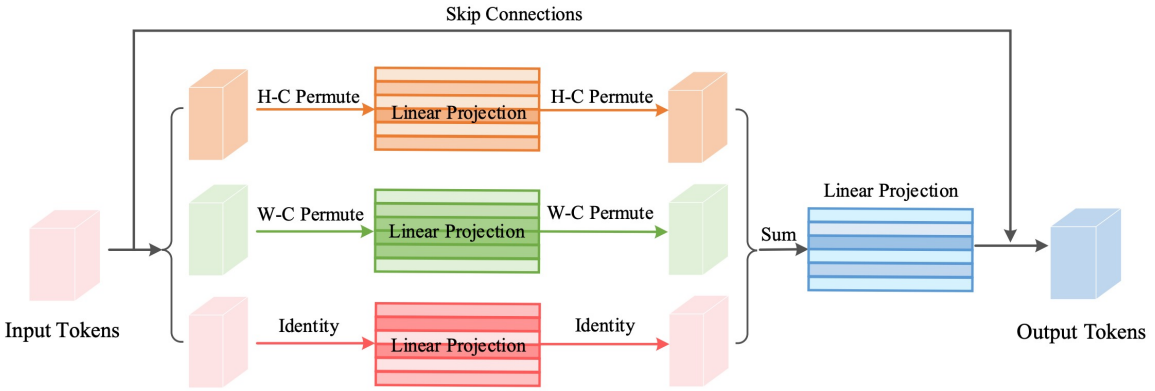
DynaMixer: A Vision MLP Architecture with Dynamic Mixing

Ziyu Wang, Wenhao Jiang, Yiming Zhu, Li Yuan, Yibing Song, Wei Liu

Existing MLP-like Vision Models

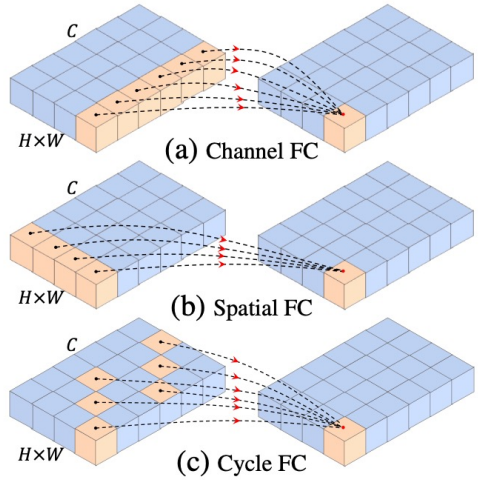


MLP-Mixer

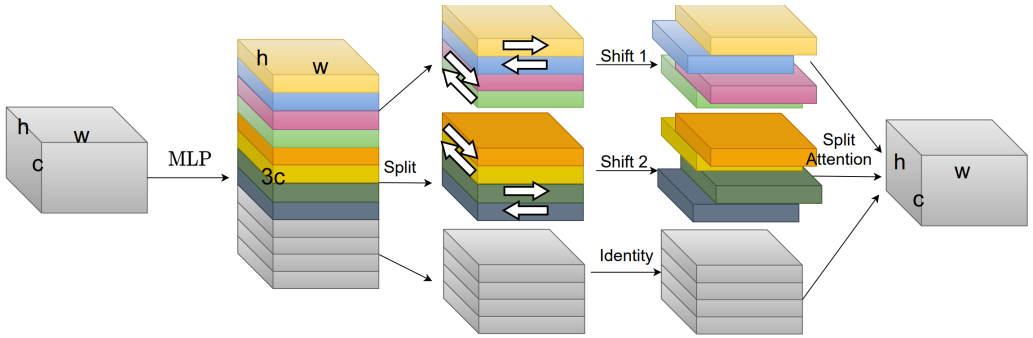


ViP

Cross-space MLP



CycleMLP



S^2-MLP

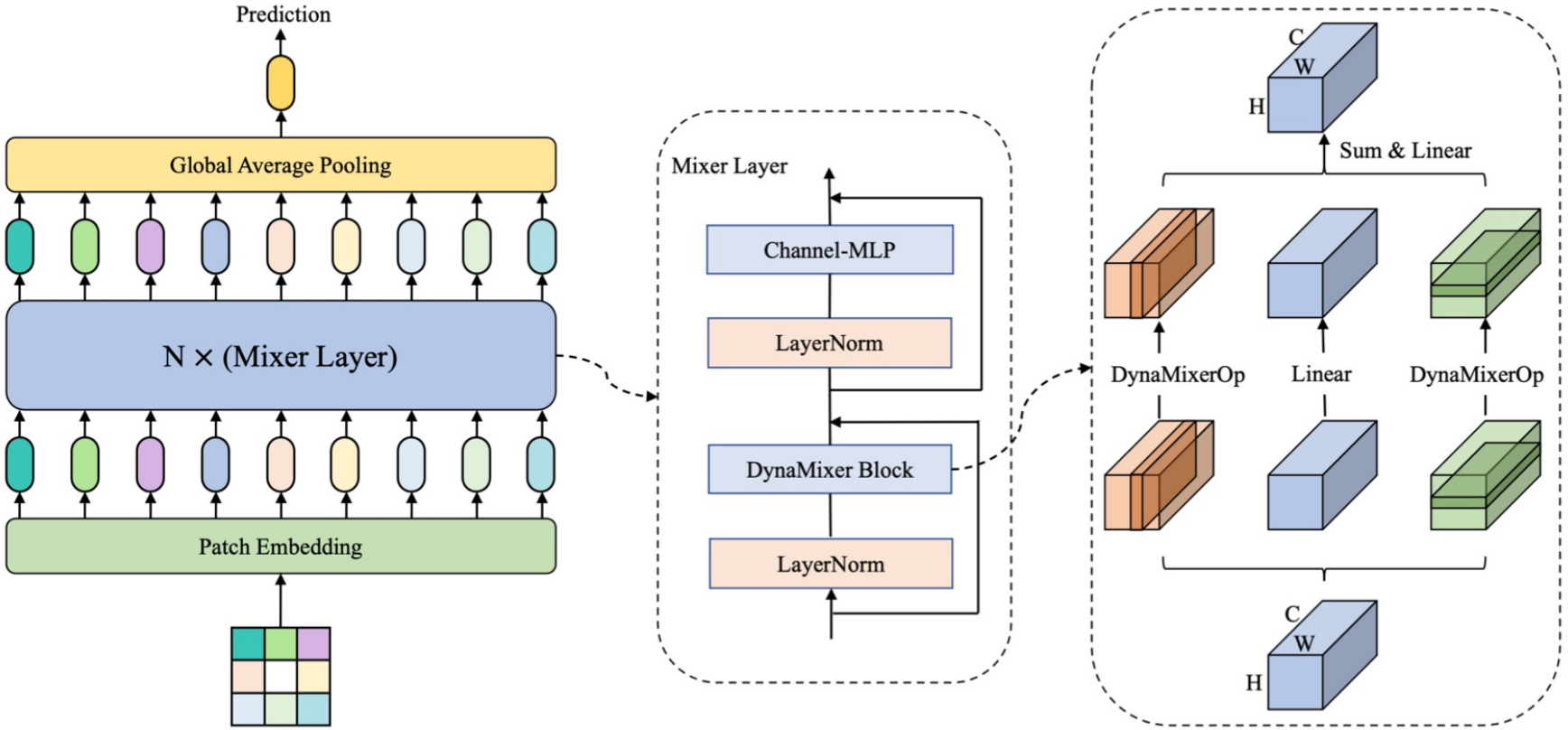
Cross-channel MLP

Motivation & Solution

- Existing MLP-like models fuse tokens through static fusion operations, lacking adaptability to the contents of the tokens to be mixed. Thus, customary information fusion procedures are not effective enough.

- We propose **DynaMixer** to dynamically generate mixing matrices by leveraging the contents of all the tokens to be mixed, thus enhancing the expression power of MLP operations.

Overall Architecture of DynaMixer



DynaMixer Operation

Input: $X \in R^{N \times D}$, output: $Y \in R^{N \times D}$

Regular Token Mixing Operation:

$$Y = PX, P \in R^{N \times N}$$

Generating mixing matrix from input:

$$P_i = softmax\left(flat(X)W^{(i)}\right), flat(x) \in R^{1 \times ND}, W^{(i)} \in R^{ND \times N}$$

Dimension reduction:

$$\hat{X} = XW_d, W_d \in R^{D \times d}, d \ll D$$

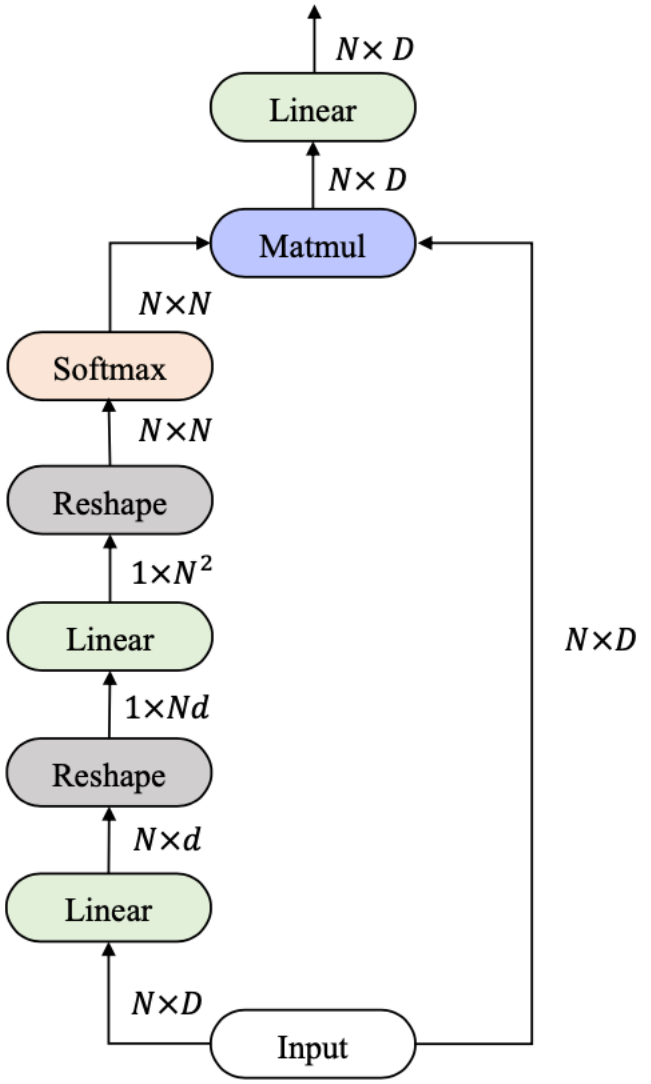
$$P_i = softmax(flat(\hat{X})W^{(i)})$$

Multi-segment fusion:

$$\hat{X}^{(s)} = XW_d^{(s)}$$

$$P_i^{(s)} = softmax(flat(\hat{X}^{(s)})W^{(s,i)})$$

$$Y = [P^{(0)}X^{(0)}, \ldots, P^{(s)}X^{(s)}]W_o$$



Ablation study

Model	D/S	# parameters	Top-1 (%)
DynaMixer-S	D	26M	82.2
DynaMixer-S	96	26M	82.4
DynaMixer-S	48	26M	82.5
DynaMixer-S	24	26M	82.7

The results with different segment S.

Model	d	# parameters	Top-1 (%)
DynaMixer-S	1	26M	82.4
DynaMixer-S	2	26M	82.7
DynaMixer-S	4	27M	82.6
DynaMixer-S	8	29M	82.7

The effects of reduced dimensionality.

Model	generating methods	# parameters	Top-1 (%)
DynaMixer-S	Synthesize(Random)	25M	81.5
DynaMixer-S	Synthesizer(Dense)	32M	81.4
DynaMixer-S	DynaMixer-S	26M	82.7

The effects of weight generation methods.

Model	# parameters	Top-1 (%)
DynaMixer-S	26M	82.7
- column mixing	23M	79.2
- row mixing	23M	79.4
- channel fusion	23M	82.1
- reweighting	24M	81.9
+ sharing DynaMixer op	23M	82.2

The effects of different components.

Comparisons with SOTA

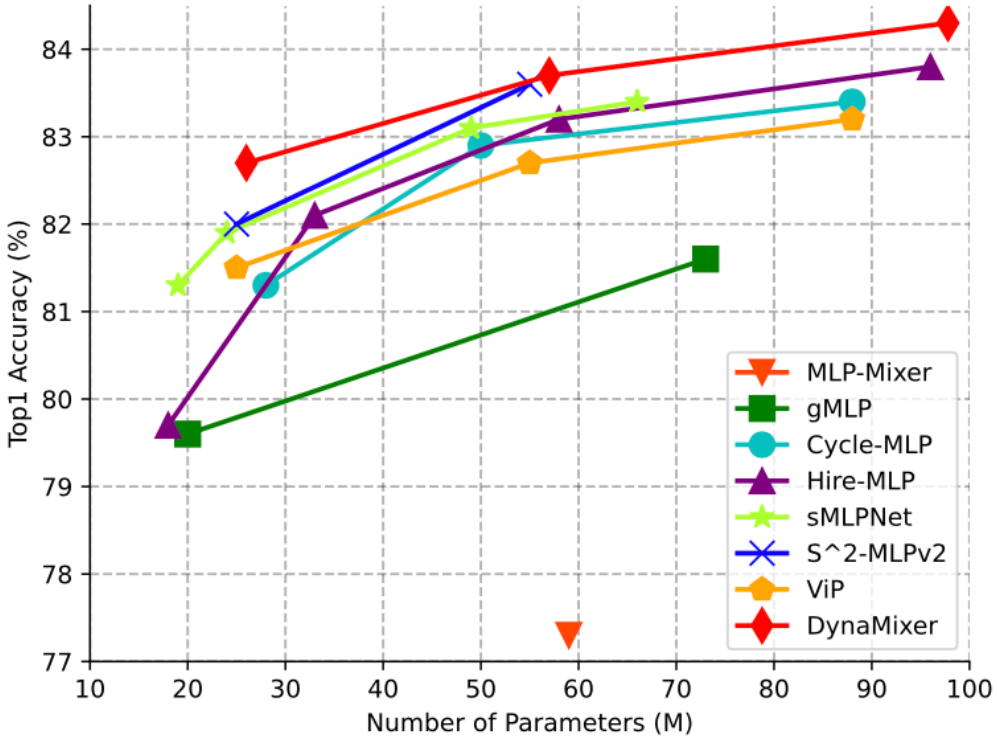
Model	Family	Scale	Param	FLOPs	Top-1 (%)
ResNet18 (He et al., 2016)	CNN	224^2	12M	1.8G	69.8
EffNet-B3 (Tan & Le, 2019)	CNN	300^2	12M	1.8G	81.6
PVT-T (Wang et al., 2021a)	Trans	224^2	13M	1.9G	75.1
GFNet-H-Ti (Rao et al., 2021)	FFT	224^2	15M	2.0G	80.1
ResNet50 (He et al., 2016)	CNN	224^2	26M	4.1G	78.5
RegNetY-4G (Radosavovic et al., 2020)	CNN	224^2	21M	4.0G	80.0
DeiT-S (Touvron et al., 2021b)	Trans	224^2	22M	4.6G	79.8
BoT-S1-50 (Srinivas et al., 2021)	Hybrid	224^2	21M	4.3G	79.1
PVT-S (Wang et al., 2021a)	Trans	224^2	25M	3.8G	79.8
T2T-14 (Yuan et al., 2021a)	Trans	224^2	22M	4.8G	81.5
Swin-T (Liu et al., 2021b)	Trans	224^2	29M	4.5G	81.3
GFNet-H-S (Rao et al., 2021)	FFT	224^2	32M	4.5G	81.5
CrossFormer-T (Wang et al., 2021b)	Trans	224^2	27M	2.9G	81.5
CrossFormer-S (Wang et al., 2021b)	Trans	224^2	30M	4.9G	82.5
DynaMixer-S	MLP	224^2	26M	7.3G	**82.7**
ResNet101 (He et al., 2016)	CNN	224^2	45M	7.9G	79.8
RegNetY-8G (Radosavovic et al., 2020)	CNN	224^2	39M	8.0G	81.7
BoT-S1-59 (Srinivas et al., 2021)	Hybrid	224^2	34M	7.3G	81.7
PVT-M (Wang et al., 2021a)	Trans	224^2	44M	6.7G	81.2
GFNet-H-B (Rao et al., 2021)	FFT	224^2	54M	8.4G	82.9
Swin-S (Liu et al., 2021b)	Trans	224^2	50M	8.7G	83.0
PVT-L (Wang et al., 2021a)	Trans	224^2	61M	9.8G	81.7
CrossFormer-B (Wang et al., 2021b)	Trans	224^2	52M	9.2G	83.4
T2T-24 (Yuan et al., 2021a)	Trans	224^2	64M	13.8G	82.1
DynaMixer-M	MLP	224^2	57M	17.0G	**83.7**
CrossFormer-L (Wang et al., 2021b)	Trans	224^2	92M	16.1G	84.0
Swin-B (Liu et al., 2021b)	Trans	224^2	88M	15.4G	83.3
DeiT-B (Touvron et al., 2021b)	Trans	224^2	86M	17.5G	81.8
DeiT-B (Touvron et al., 2021b)	Trans	384^2	86M	55.4G	83.1
DynaMixer-L	MLP	224^2	97M	27.4G	**84.3**

ImageNet top-1 accuracy comparisons with other MLP-like models.

Model	Date	Param	FLOPs	Top-1 (%)
Mixer-B/16 (Tolstikhin et al., 2021)	MAY, 2021	59M	12.7G	76.4
Mixer-B/16† (Tolstikhin et al., 2021)		59M	12.7G	77.3
ResMLP-S12 (Touvron et al., 2021a)		15M	3.0G	76.6
ResMLP-S24 (Touvron et al., 2021a)	MAY, 2021	30M	6.0G	79.4
ResMLP-B24 (Touvron et al., 2021a)		116M	23.0G	81.0
gMLP-Ti (Liu et al., 2021a)		6M	1.4G	72.3
gMLP-S (Liu et al., 2021a)	MAY, 2021	20M	4.5G	79.6
gMLP-B (Liu et al., 2021a)		73M	15.8G	81.6
ViP-Small/14 (Hou et al., 2021)		30M	6.9G	80.7
ViP-Small/7 (Hou et al., 2021)	JUN, 2021	25M	6.9G	81.5
ViP-Medium/7 (Hou et al., 2021)		55M	16.3G	82.7
ViP-Large/7 (Hou et al., 2021)		88M	24.4G	83.2
AS-MLP-T (Lian et al., 2021)		28M	4.4G	81.3
AS-MLP-S (Lian et al., 2021)	JUL, 2021	50M	8.5G	83.1
AS-MLP-B (Lian et al., 2021)		88M	15.2G	83.3
S^2-MLPv2-Small/7 (Yu et al., 2021b)	AUG, 2021	25M	6.9G	82.0
S^2-MLPv2-Medium/7 (Yu et al., 2021b)		55M	16.3G	83.6
RaftMLP-36 (Tatsunami & Taki, 2021)	AUG, 2021	44M	9.0G	76.9
RaftMLP-12 (Tatsunami & Taki, 2021)		58M	12.0G	78.0
sMLPNet-T (Tang et al., 2021)	SEP, 2021	24M	5.0G	81.9
sMLPNet-S (Tang et al., 2021)		49M	10.3G	83.1
sMLPNet-B (Tang et al., 2021)		66M	14.0G	83.3
ConvMLP-S (Li et al., 2021)		9M	2.4G	76.8
ConvMLP-M (Li et al., 2021)	SEP, 2021	17M	3.9G	79.0
ConvMLP-L (Li et al., 2021)		43M	9.9G	80.2
CycleMLP-T (Chen et al., 2021b)		28M	4.4G	81.3
CycleMLP-S (Chen et al., 2021b)	NOV, 2021	50M	8.5G	82.9
CycleMLP-B (Chen et al., 2021b)		88M	15.2G	83.4
Hire-MLP-Ti (Guo et al., 2021)		18M	2.1G	79.7
Hire-MLP-S (Guo et al., 2021)	NOV, 2021	33M	4.2G	82.1
Hire-MLP-B (Guo et al., 2021)		58M	8.1G	83.2
Hire-MLP-L (Guo et al., 2021)		96M	13.4G	83.8
DynaMixer-S		26M	7.3G	82.7
DynaMixer-M		57M	17.0G	83.7
DynaMixer-L		97M	27.4G	84.3

ImageNet top-1 accuracy comparisons with comparisons with SOTA classification models.

Efficiency comparison



ImageNet accuracy v.s. model capacity.

Model	# Param	FLOPs	Throughput image/s	Top-1 (%)
ResMLP S24	30M	6.0G	680	79.4
ResMLP B24	116M	23.0G	215	81.0
ViP-M/7	55M	16.3G	390	82.7
DynaMixer-S	**26M**	**7.3G**	**551**	82.7
ViP-L/7	88M	24.4G	276	83.2
DynaMixer-M	**57M**	**17.0G**	**326**	83.7

The number of parameters, FLOPs and throughput of different models.

Conclusion

- We propose the DynaMixer model to dynamically generate the mixing matrices by leveraging the contents of all the tokens to be mixed.
- To reduce the time complexity and meanwhile improve the robustness, we exerted a dimensionality reduction technique and a multi-segment fusion mechanism.
- Overall, our DynaMixer model has achieved state-of-the-art performance on the ImageNet recognition task among MLP-like models and is comparable with state-of-the-art transformer models.

Source code:

https://github.com/ziyuwwang/DynaMixer