MLP-Mixer: An all-MLP Architecture for Vision

Ilya Tolstikhin*, Neil Houlsby*, Alexander Kolesnikov*, Lucas Beyer*, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, Alexey Dosovitskiy
Google Research, Brain Team

Presented by Yilin Liu

Introduction

- A conceptually and technically simple alternative to ViT
- Key idea: mixing features

    (i) Along channels

    (ii) Across spatial locations

    Separately!

    Based purely on MLPs (1 x 1 convolutions)!

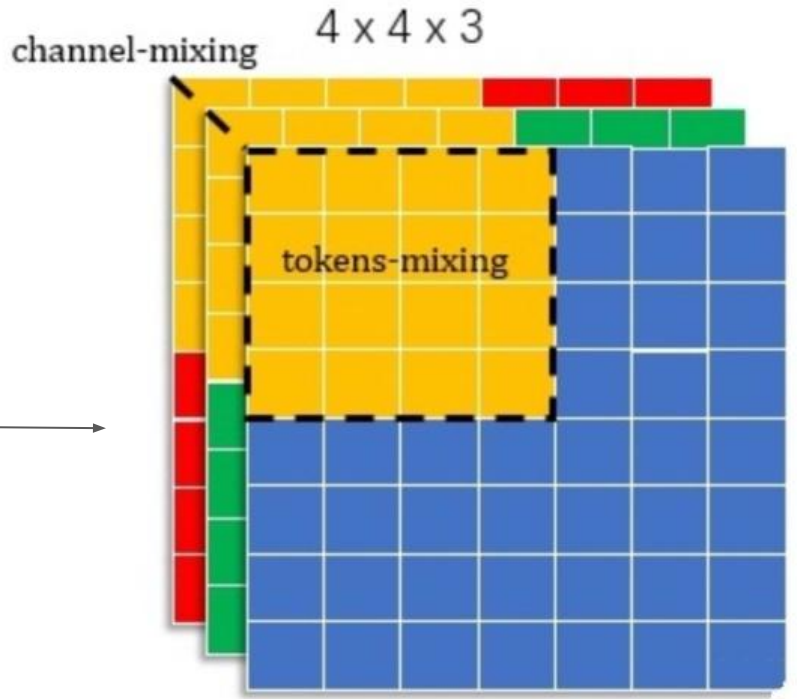
Introduction

- Key idea: mixing features

  (i) Along channels (within a token)

  (ii) Across spatial locations (between tokens)

  A regular convolutional filter (filter size > 1), performing both (i) and (ii)



Method

- Input image: H x W x 3

  → Patches: S x (P^2 x 3), where S = HW/P^2, each patch (P, P).

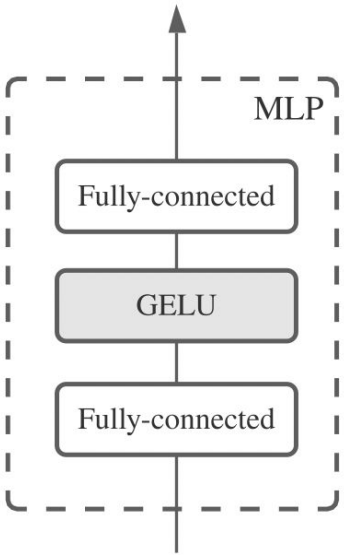
  → Linear projection: S x C

- During Training

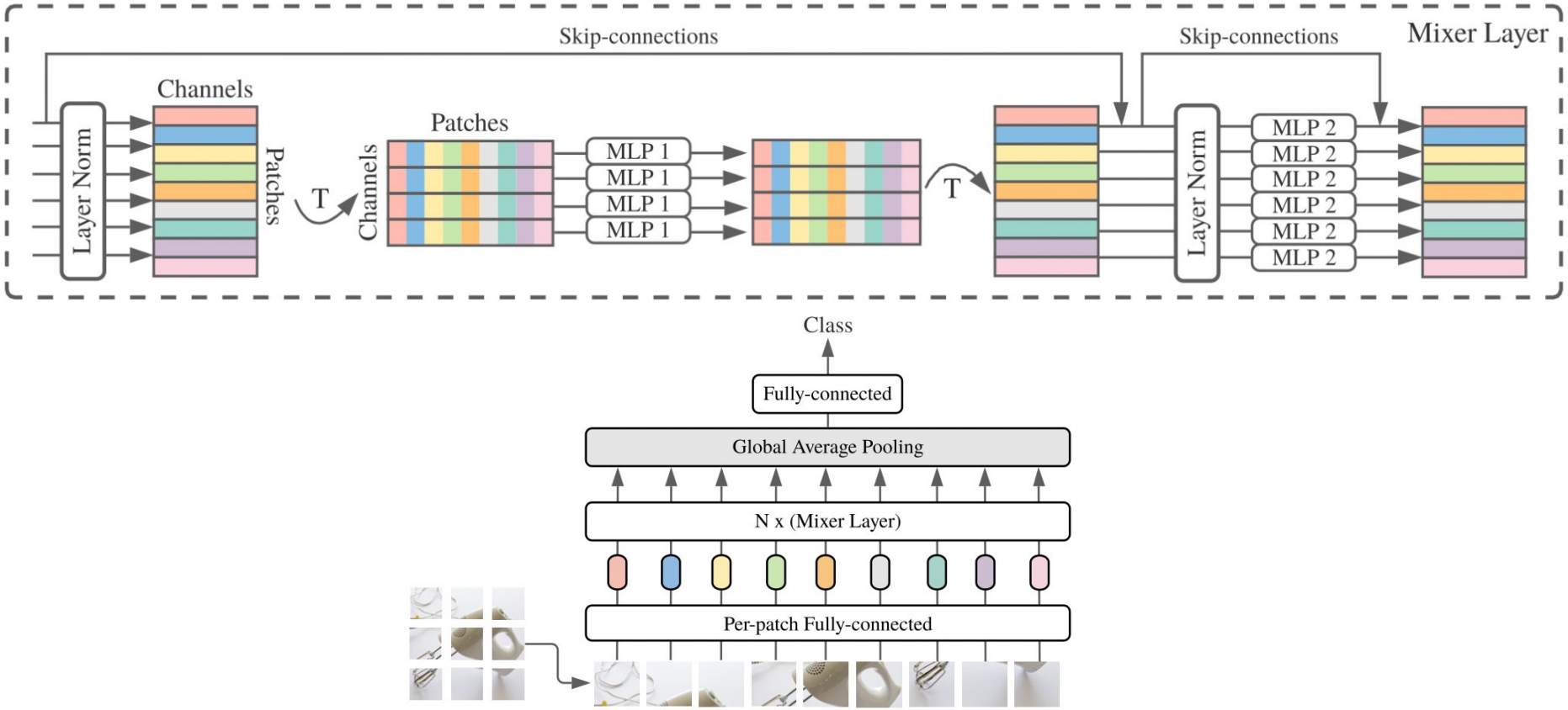
  → B x S x C, reshape to B x C x S

  → ‘token-mixing’ MLP, reshape to B x S x C

  → ‘channel-mixing’ MLP



Method


Skip-connections
Skip-connections
Mixer Layer
Channels
Layer Norm
Patches
T
Patches
Channels
MLP 1
MLP 1
MLP 1
MLP 1
T
Layer Norm
MLP 2
MLP 2
MLP 2
MLP 2
MLP 2
MLP 2
Class
Fully-connected
Global Average Pooling
N x (Mixer Layer)
Per-patch Fully-connected

Experiments

- Pretrained on medium- to large-scale datasets
  - ILSVRC2021 ImageNet
  - ImageNet-21k (14M data, 21k classes)
  - JFT-300M (300M data, 18k classes)
- Small and mid-sized downstream classification tasks
  - ILSVRC2012 “ImageNet” (1.3M data, 1k classes)
  - CIFAR-10/100 (50k data, 10/100 classes)
  - Oxford-IIIT Pets (3.7k data, 36 classes)
  - Oxford Flowers-102 (2k data, 102 classes)
  - Visual Task Adaptation Benchmark (VTAB-1k, 19 diverse datasets each with 1k data)
- Evaluation metrics:
  - Accuracy on the downstream task
  - Total computational cost of pre-training
  - Test-time throughput

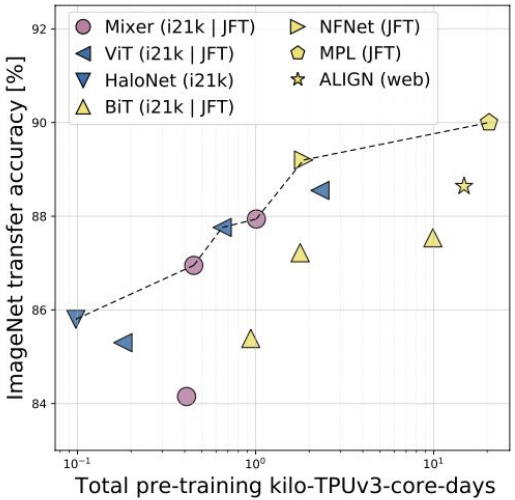
Results

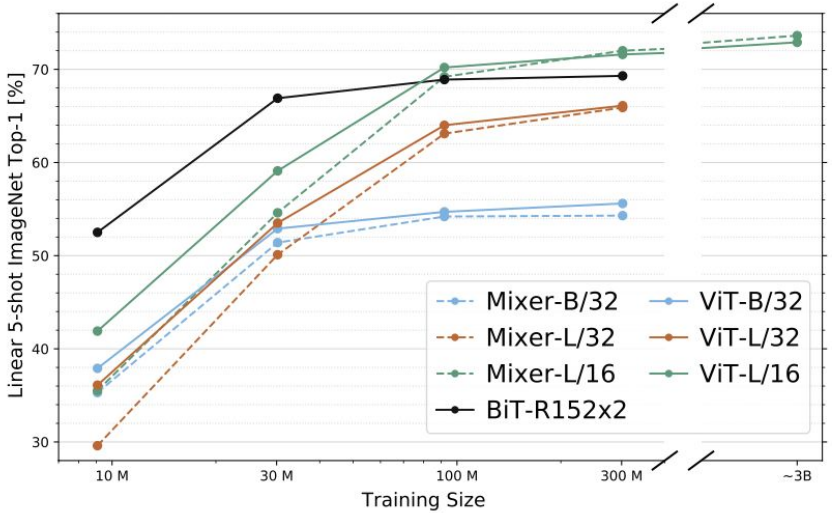
- Mixer overall strong, but inferior to other models
- Size of the upstream dataset increases, performance improved

	ImNet top-1	ReaL top-1	Avg 5 top-1	VTAB-1k 19 tasks	Throughput img/sec/core	TPUv3 core-days
Pre-trained on ImageNet-21k (public)						
HaloNet [51]	85.8	—	—	—	120	0.10k
Mixer-L/16	84.15	87.86	93.91	74.95	105	0.41k
ViT-L/16 [14]	85.30	88.62	94.39	72.72	32	0.18k
BiT-R152x4 [22]	85.39	—	94.04	70.64	26	0.94k
Pre-trained on JFT-300M (proprietary)						
NFNet-F4+ [7]	89.2	—	—	—	46	1.86k
Mixer-H/14	87.94	90.18	95.71	75.33	40	1.01k
BiT-R152x4 [22]	87.54	90.54	95.33	76.29	26	9.90k
ViT-H/14 [14]	88.55	90.72	95.97	77.63	15	2.30k
Pre-trained on unlabelled or weakly labelled data (proprietary)						
MPL [35]	90.0	91.12	—	—	—	20.48k
ALIGN [21]	88.64	—	—	79.99	15	14.82k

Results

- Total pre-training cost correlates with downstream accuracy
- Accuracy-compute trade-off, Mixer (pink) is competitive with CNNs (yellow)
- The size of the dataset matters more to Mixers; overfits when small; surpasses ViTs when large





Results - Ablation studies

- Increase the model size (depth, width, etc.) when pre-training
  - Mixers overfits more easily than ViTs, especially when the dataset is small
  - Although Mixers are weak when the scales are small; they become strong when the scales increased.

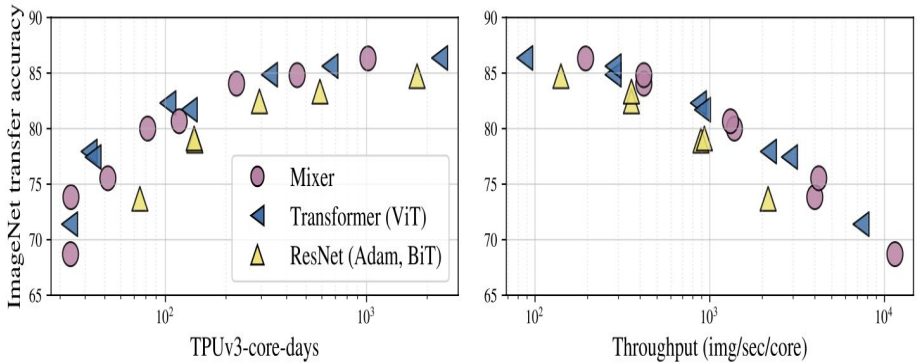
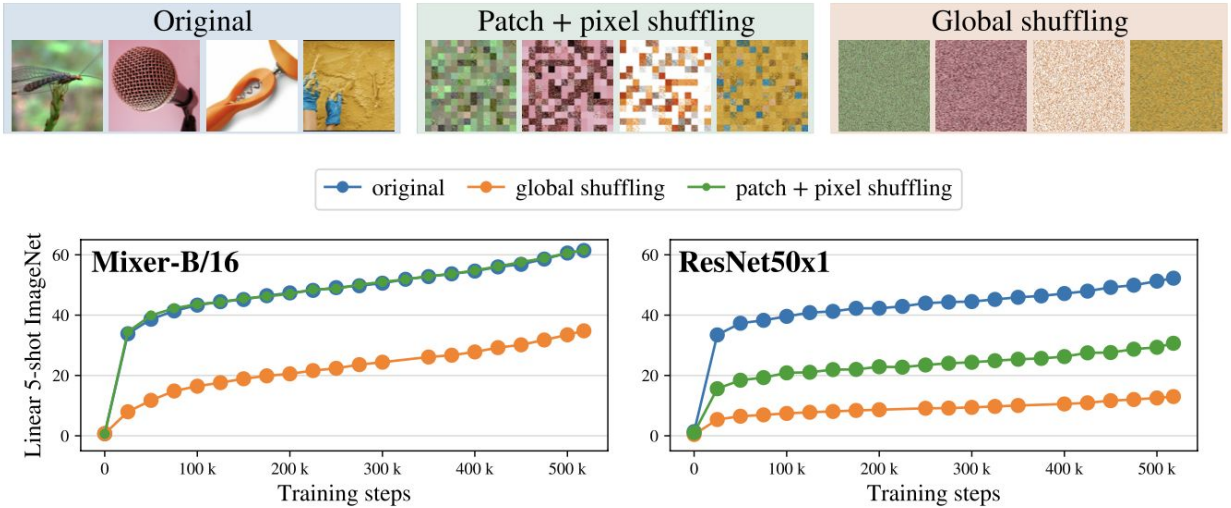


	Image size	Pre-Train Epochs	ImNet top-1	ReaL top-1	Avg. 5 top-1	Throughput (img/sec/core)	TPUv3 core-days
Pre-trained on ImageNet (with extra regularization)							
Mixer-B/16	224	300	76.44	82.36	88.33	1384	0.01k(‡)
ViT-B/16 (☎)	224	300	79.67	84.97	90.79	861	0.02k(‡)
Mixer-L/16	224	300	71.76	77.08	87.25	419	0.04k(‡)
ViT-L/16 (☎)	224	300	76.11	80.93	89.66	280	0.05k(‡)
Pre-trained on ImageNet-21k (with extra regularization)							
Mixer-B/16	224	300	80.64	85.80	92.50	1384	0.15k(‡)
ViT-B/16 (☎)	224	300	84.59	88.93	94.16	861	0.18k(‡)
Mixer-L/16	224	300	82.89	87.54	93.63	419	0.41k(‡)
ViT-L/16 (☎)	224	300	84.46	88.35	94.49	280	0.55k(‡)
Mixer-L/16	448	300	83.91	87.75	93.86	105	0.41k(‡)
Pre-trained on JFT-300M							
Mixer-S/32	224	5	68.70	75.83	87.13	11489	0.01k
Mixer-B/32	224	7	75.53	81.94	90.99	4208	0.05k
Mixer-S/16	224	5	73.83	80.60	89.50	3994	0.03k
BiT-R50x1	224	7	73.69	81.92	—	2159	0.08k
Mixer-B/16	224	7	80.00	85.56	92.60	1384	0.08k
Mixer-L/32	224	7	80.67	85.62	93.24	1314	0.12k
BiT-R152x1	224	7	79.12	86.12	—	932	0.14k
BiT-R50x2	224	7	78.92	86.06	—	890	0.14k
BiT-R152x2	224	14	83.34	88.90	—	356	0.58k
Mixer-L/16	224	7	84.05	88.14	94.51	419	0.23k
Mixer-L/16	224	14	84.82	88.48	94.77	419	0.45k
ViT-L/16	224	14	85.63	89.16	95.21	280	0.65k
Mixer-H/14	224	14	86.32	89.14	95.49	194	1.01k
BiT-R200x3	224	14	84.73	89.58	—	141	1.78k
Mixer-L/16	448	14	86.78	89.72	95.13	105	0.45k
ViT-H/14	224	14	86.65	89.56	95.57	87	2.30k
ViT-L/16 [14]	512	14	87.76	90.54	95.63	32	0.65k

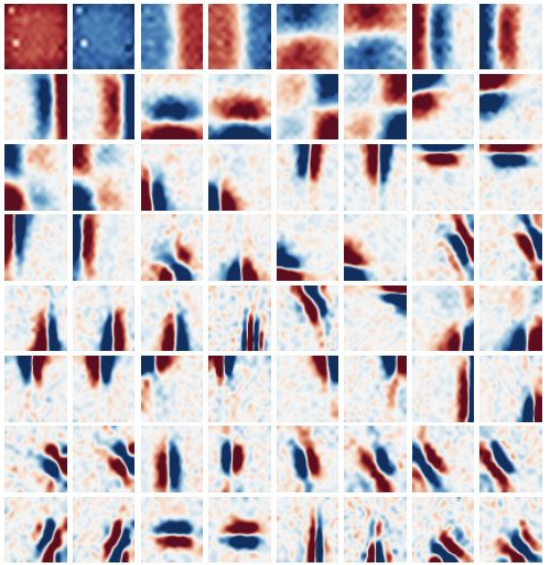
Results - Ablation studies

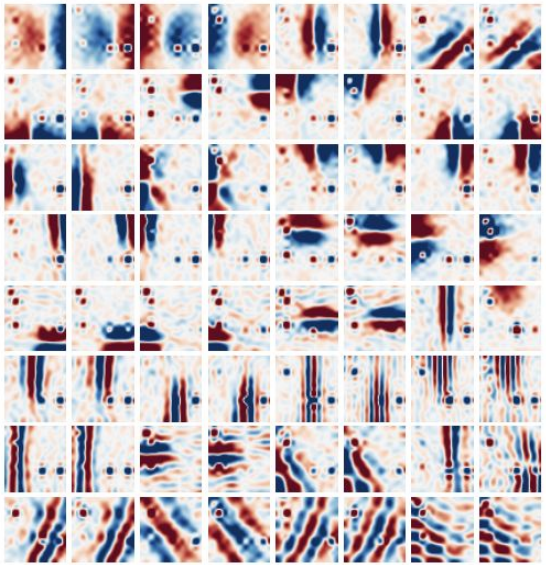
- Compare the inductive biases of Mixer and CNNs
  - Mixers invariant to the order of patches and pixels within the patches (blue and green are matched)
  - ResNet's performance drops significantly
  - When globally permuting the pixels, Mixer drops less (45%) compared to ResNet (75%)

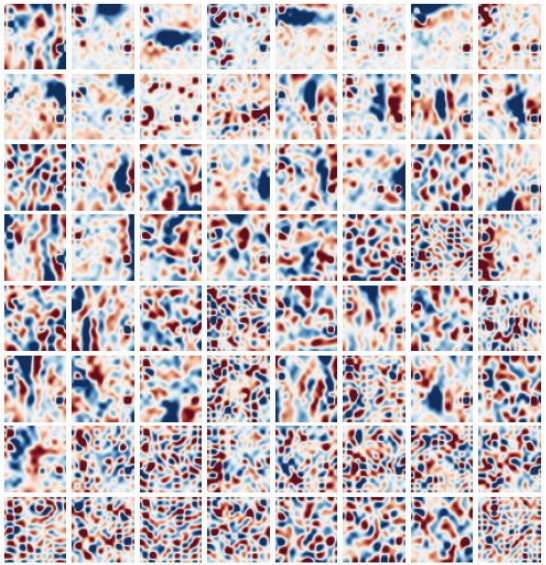


Visualizations

- Some learned features operate on the entire image, others on smaller regions
- Deeper layers have no clearly identifiable structure
- Pairs of feature detectors with opposite phases, Similar to CNNs







Conclusion

- The simple architecture is as good as the state-of-the-arts in terms of the trade-off between accuracy and computational resources
- Studying the differences of learned features between CNNs and Transformers
- Understand the role of the inductive biases hidden in these features in generalization