Paper Review

MLP-Mixer: An all-MLP Architecture for Vision

YeongHyeon Park

Department of Electrical and Computer Engineering

SungKyunKwan University

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

*equal contribution

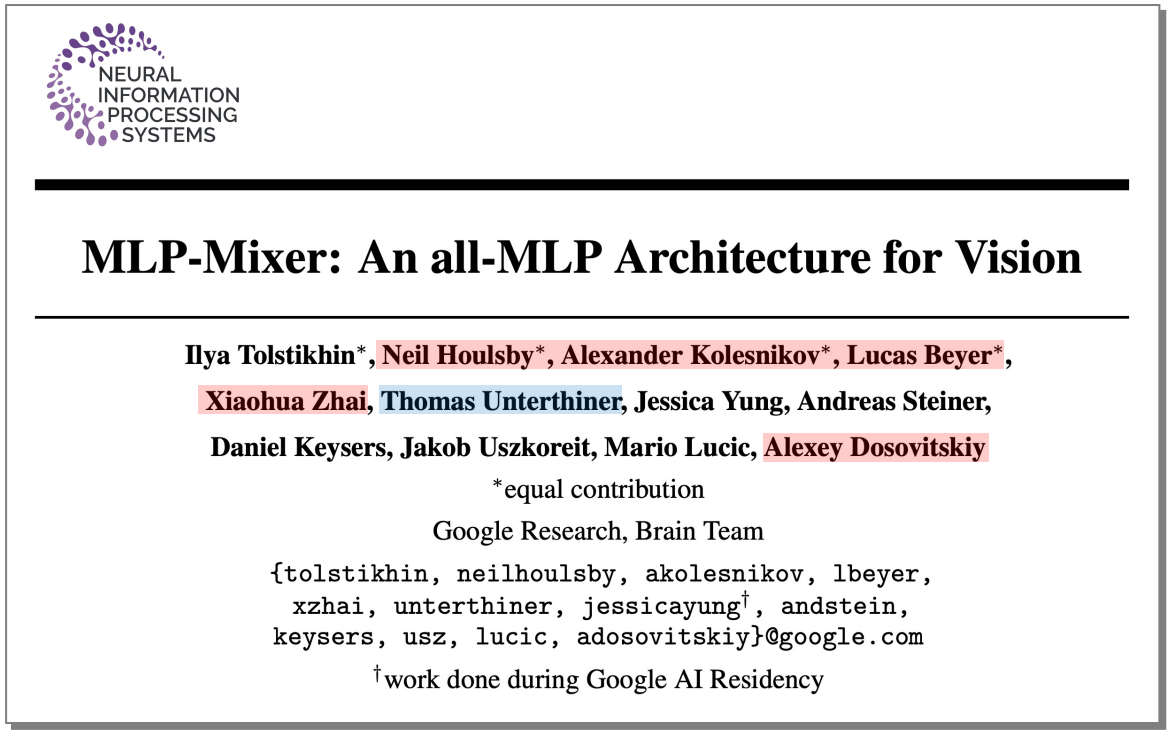
Google Research, Brain Team

{tolstikhin, neilhoulsby, akolesnikov, lbeyer,
xzhai, unterthiner, jessicayung†, andstein,
keysers, usz, lucic, adosovitskiy}@google.com

†work done during Google AI Residency

Tolstikhin, Ilya O., et al. "Mlp-mixer: An all-mlp architecture for vision." *Advances in Neural Information Processing Systems.*, 2021.

Related Work



NEURAL INFORMATION PROCESSING SYSTEMS

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**
*equal contribution
Google Research, Brain Team
{tolstikhin, neilhoulsby, akolesnikov, lbeyer,
xzhai, unterthiner, jessicayung†, andstein,
keysers, usz, lucic, adosovitskiy}@google.com
†work done during Google AI Residency

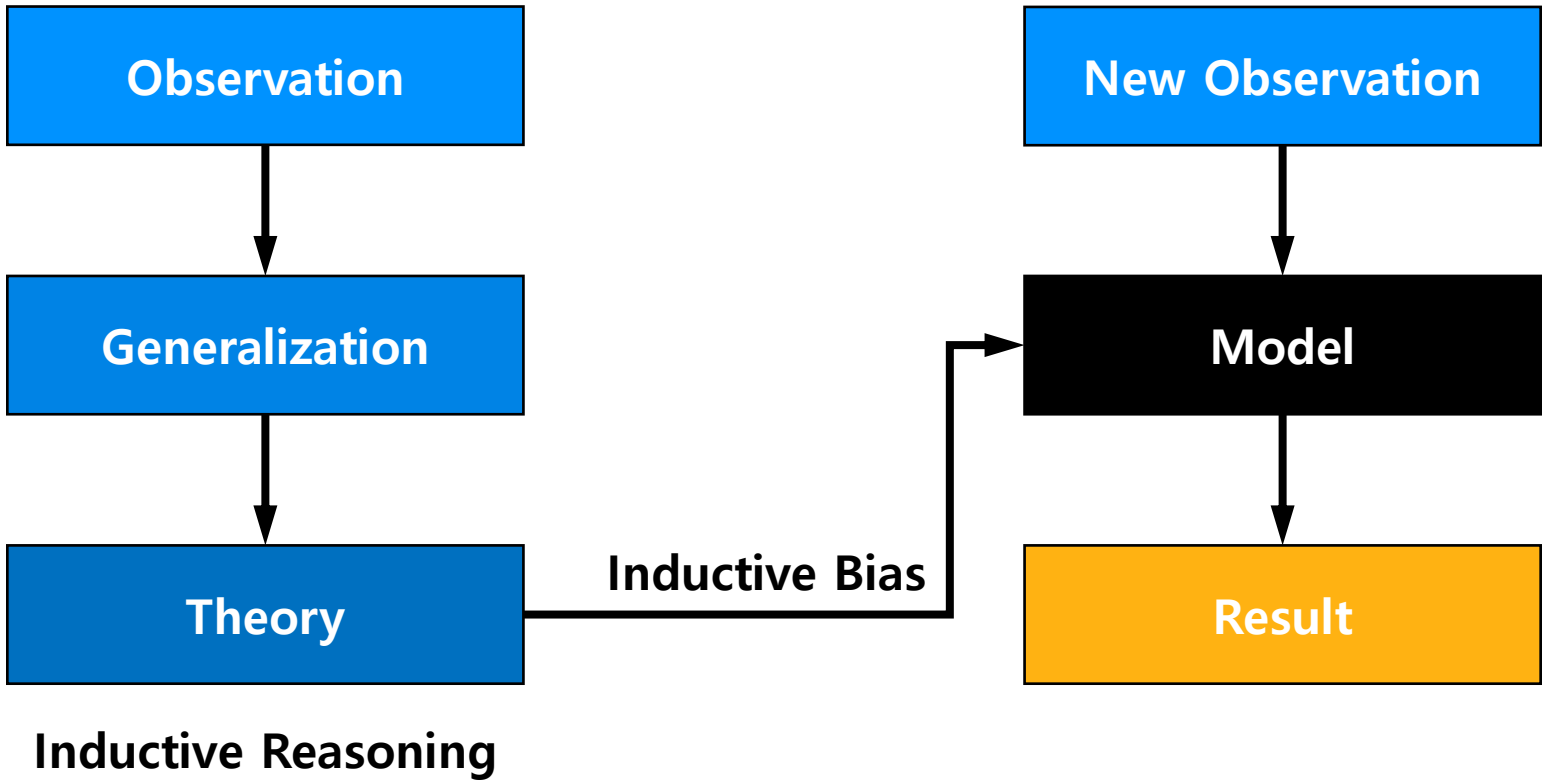
- **Publication**
  - Vision Transformer (ViT)
    - arXiv: October 2020
    - ICLR: May 2021
    - 3383 citations
  - MLP-Mixer
    - arXiv: May 2021 (after 7 months of ViT)
    - NeurIPS: December 2021
    - 265 citations

- **Authors**
  - Half (6/12) of the authors are ViT authors
  - Five authors are the first author of ViT
  - One author is the second author of ViT

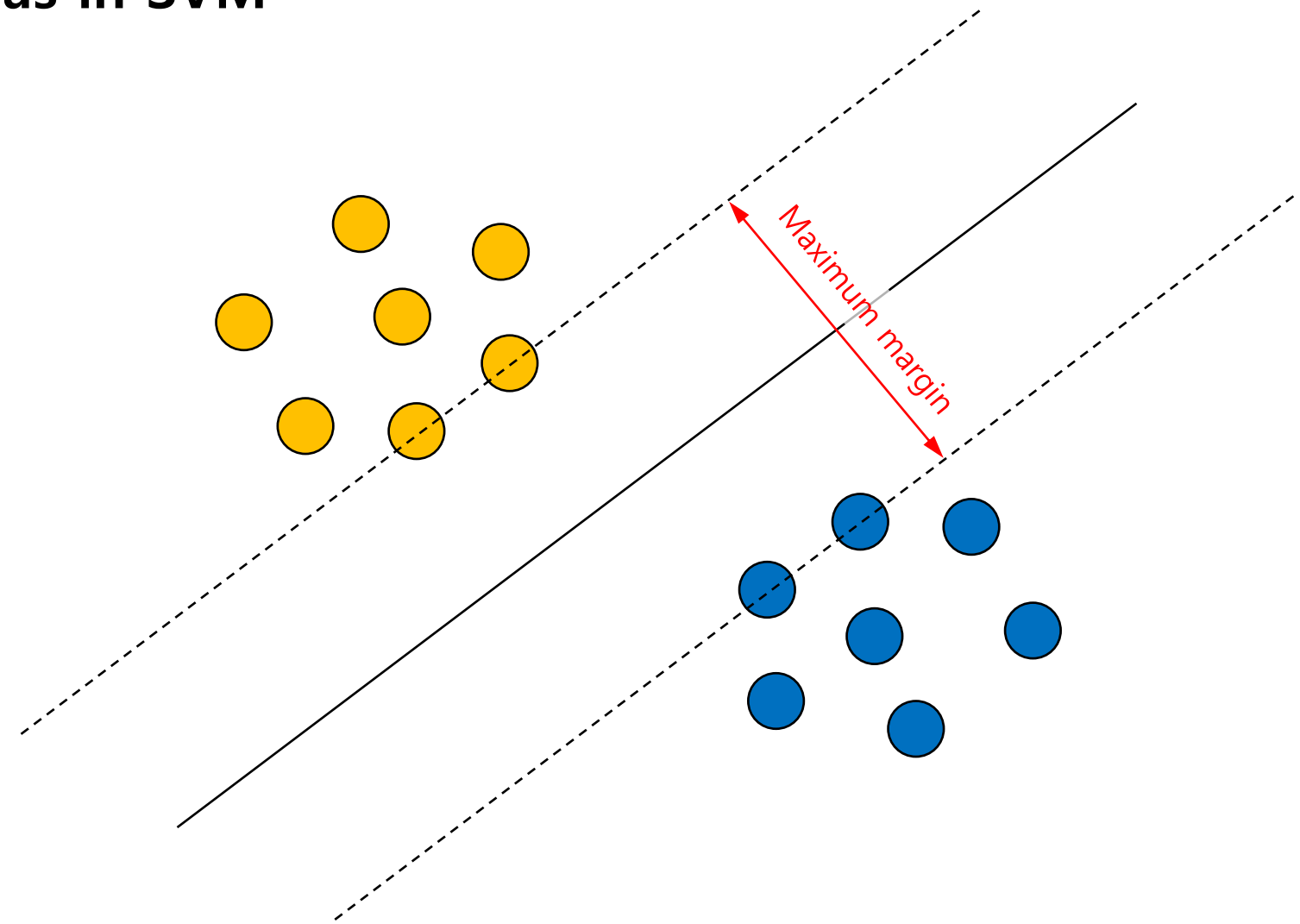
Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *International Conference on Learning Representations. 2021*. 3383 citations
Tolstikhin, Ilya O., et al. "Mlp-mixer: An all-mlp architecture for vision." *Advances in Neural Information Processing Systems.*. 2021. 265 citations

Warm Up

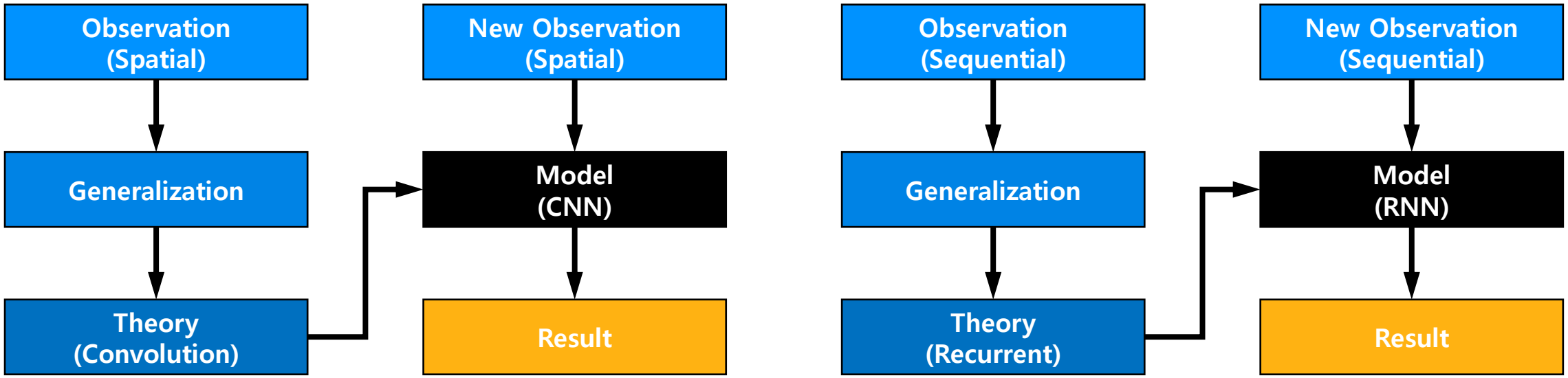
Inductive Reasoning & Inductive Bias



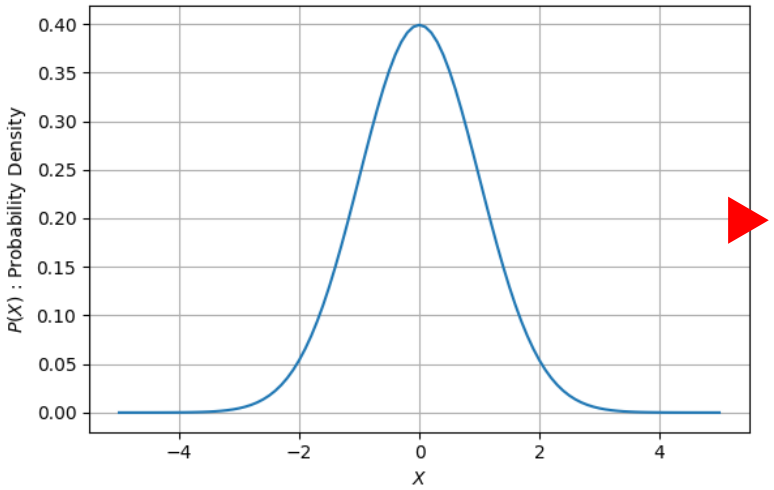
Inductive Bias in SVM



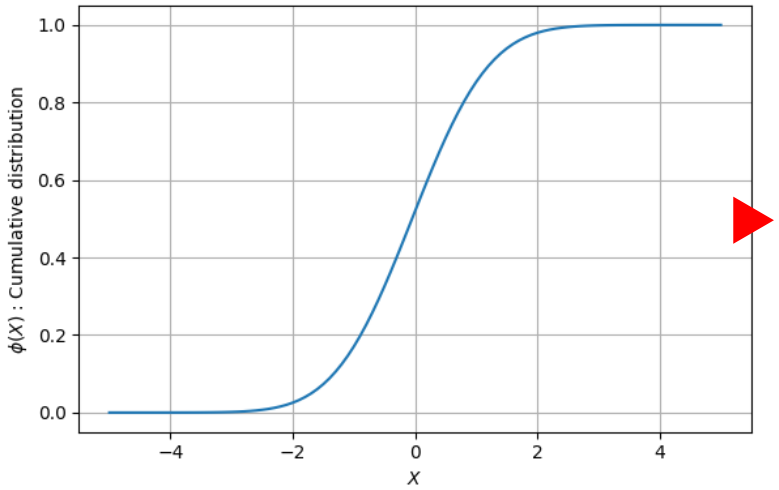
Inductive Bias in Deep Learning



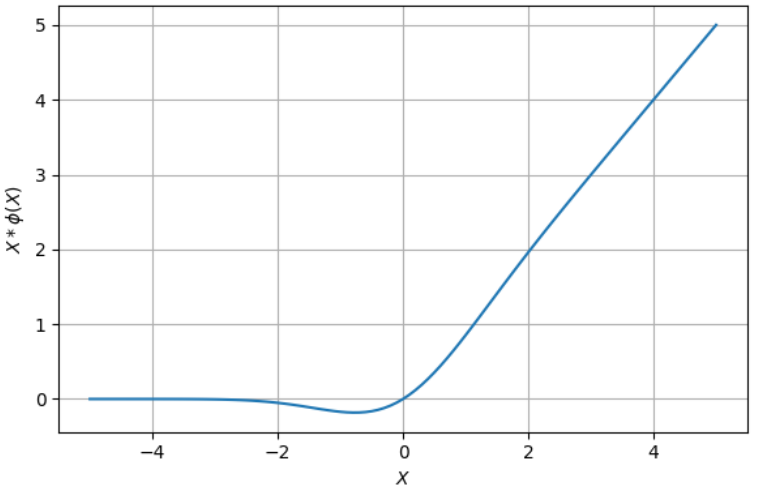
Gaussian Error Linear Unit (GELU)



a) PDF $P(X)$ of Gaussian Distribution



b) CDF $\phi(X)$ of Gaussian Distribution



c) $X\ \phi(X) = GELU(X)$

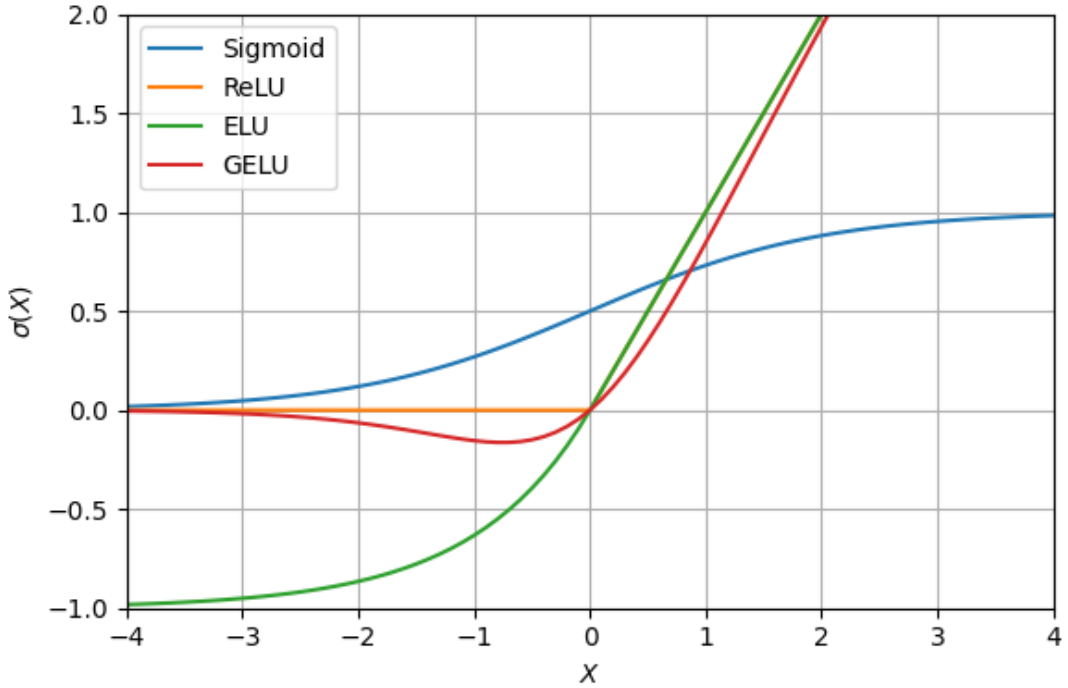
a) Assume that the distribution of the input X as a Gaussian (X will be normalized in every layer).

b) Take the form of a CDF $\phi(X)$ to make the ability of deterministic decision (sigmoid form).

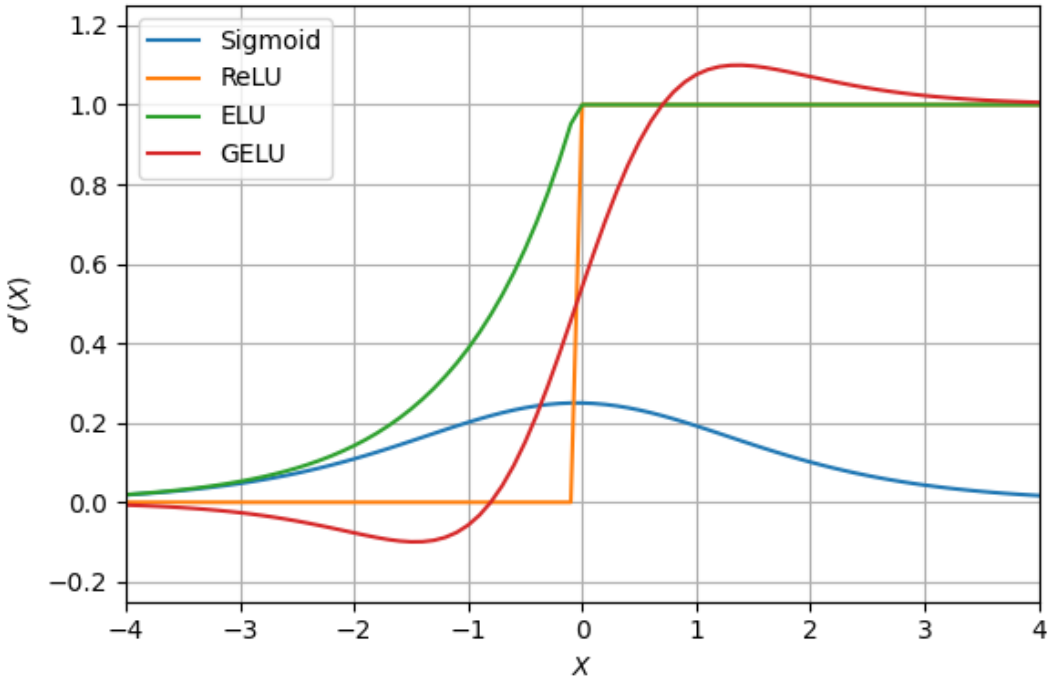
c) Multiply X to the CDF $\phi(X)$ to ease the limit of zero gradients for most X.

Comparison with some other activation functions

Activation Functions



Derivative of Activation Functions



MLP-Mixer

Vision Transformer

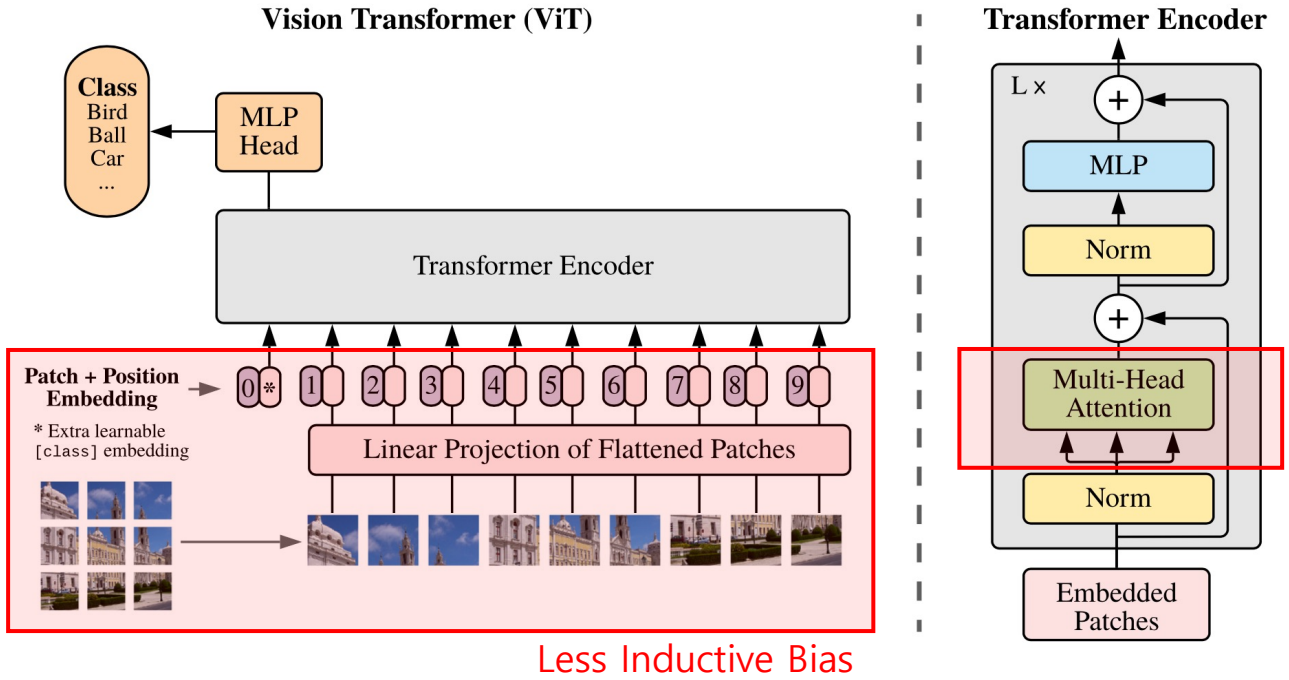


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *International Conference on Learning Representations. 2021.* 3383 citations

Multi-Head Attention

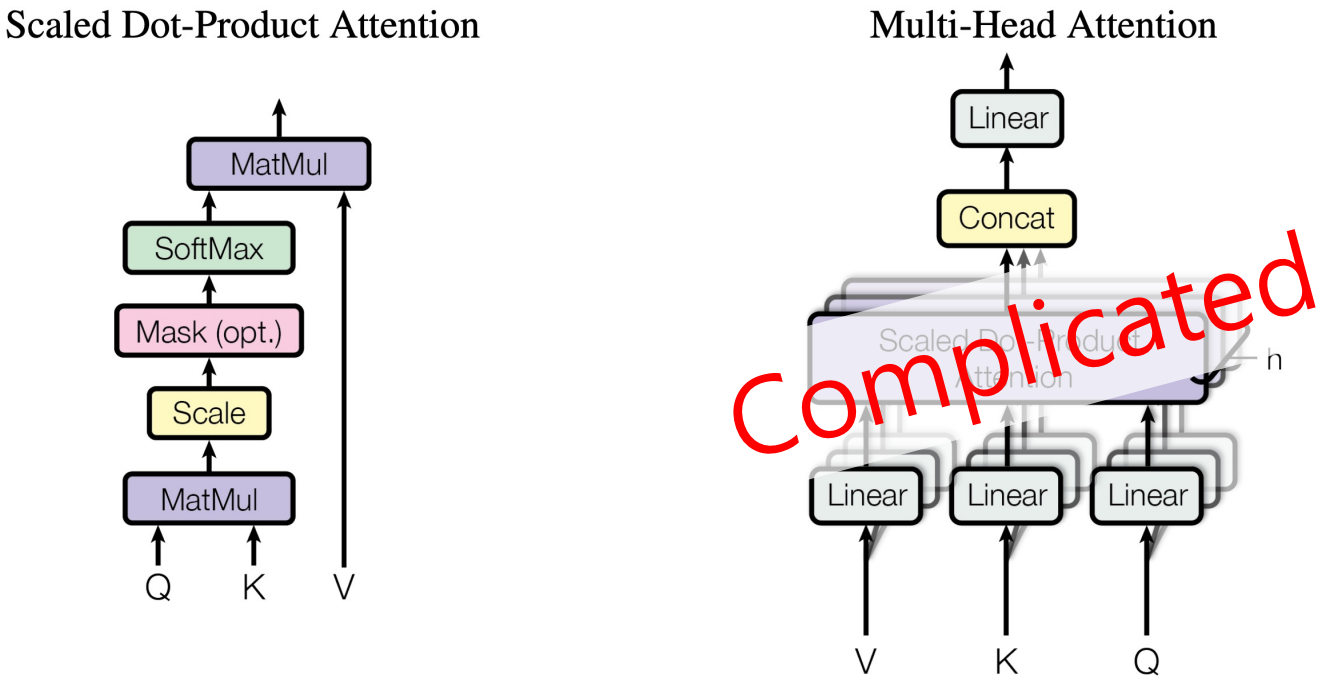


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems.* 2017.

Summary of MLP-Mixer

Purpose

- Eliminate and beyond the Convolution and Attention.
- Also eliminate positional embedding.
- Main purpose is not to demonstrate state-of-the-art results.

Contributions

- Positional embedding & invariance
  - Eliminate positional embedding by per-patch Fully-connected (ViT alternative).
  - Also, the channel mixing operation provides positional invariance (CNN alternative).
- Computational complexity
  - The computational complexity of the network is linearly increased.
- Classification performance
  - Higher classification performance than ViT when training with a large-scale dataset (JFT-300M).

Limitations

- Requires large-scale dataset
  - Because of lower inductive bias.
  - With a medium-scale dataset (smaller than 100M), MLP-Mixer shows…
    - lower classification performance than ViT.
    - Lower cost-effectiveness than ViT.
- Like the ViT, the Mixer cannot handle various input sizes that fully convolutional network (FCN) can.

Model	Operation	Detail	Big-O
MLP-Mixer	Mixing operation	Matrix multiplication (Fully-connected)	$O(D)$
ViT	Multi-head attention	Matrix multiplication (Key x Query) after matrix multiplication (Key, Query, Value)	$O(D \times D)$

MLP-Mixer

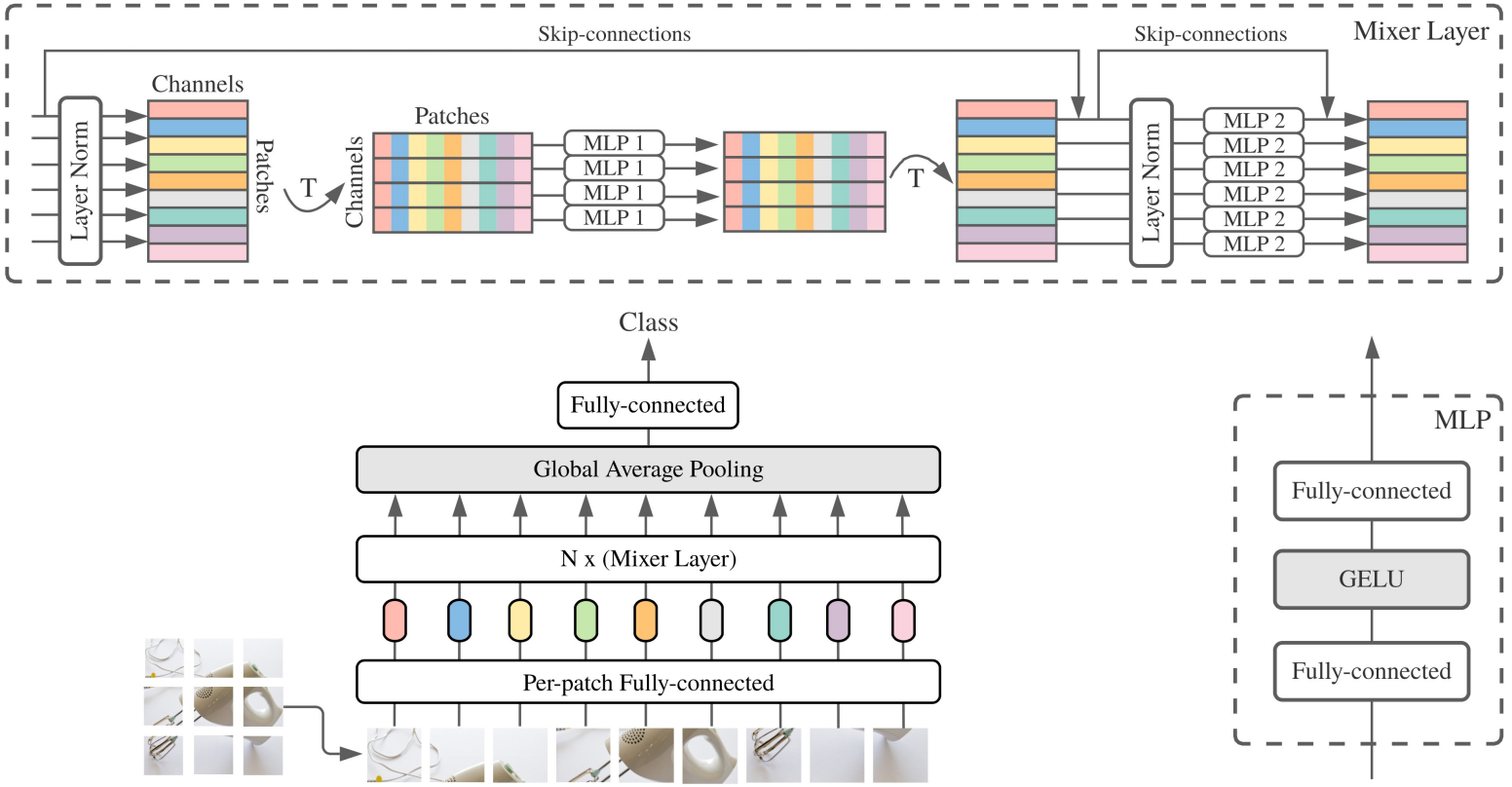
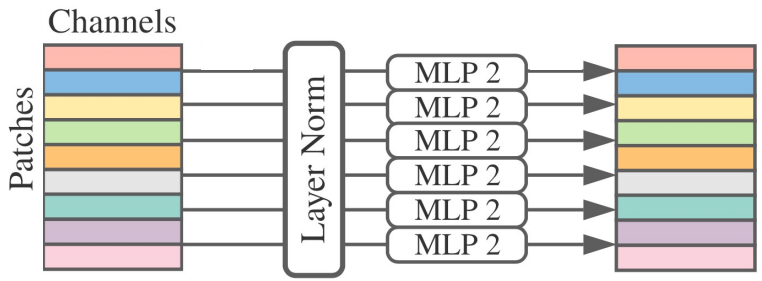


Figure 1: MLP-Mixer

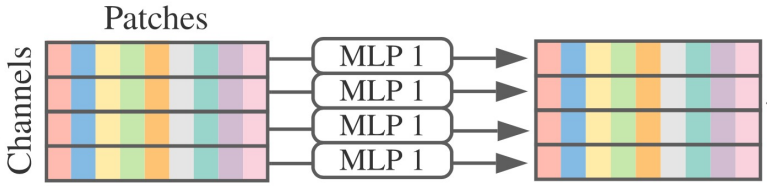
Mixer Operators

Channel Mixing MLP



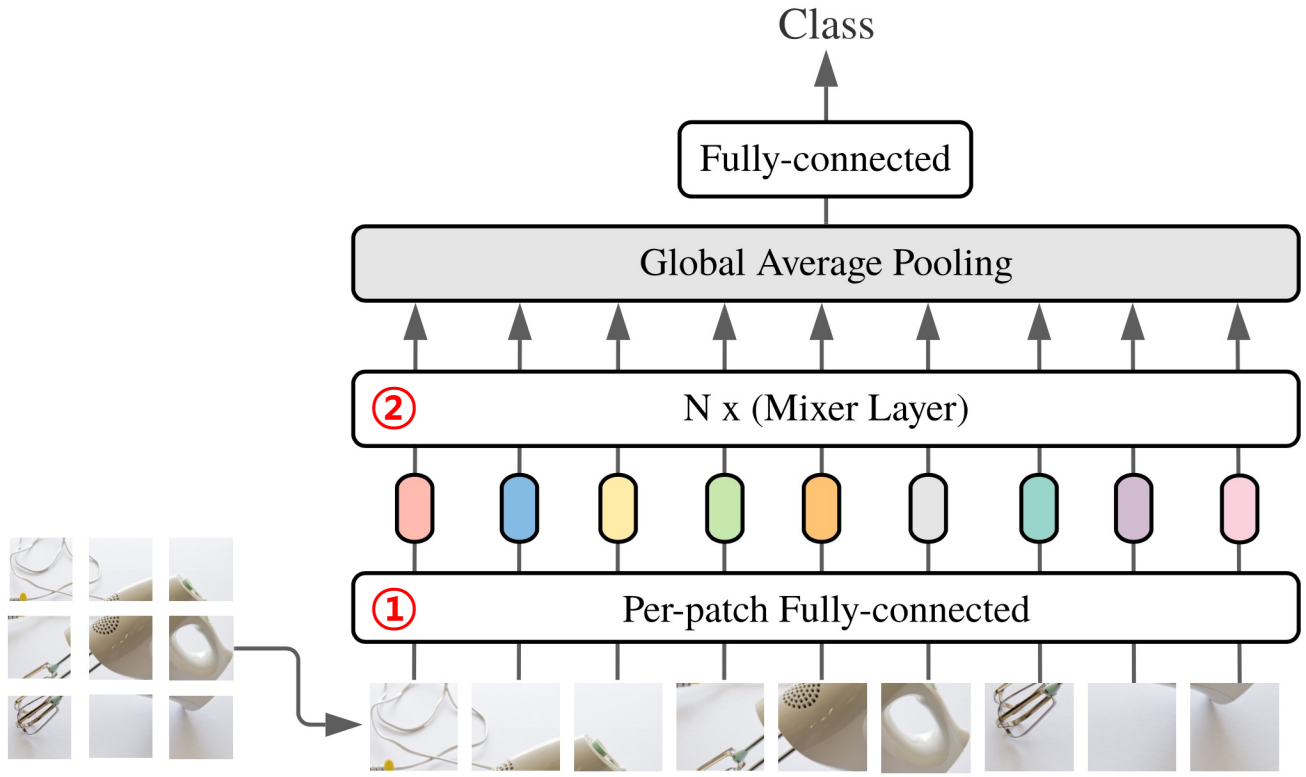
- Per-location operations
- Communication between channel
- Channel specific & spatial agnostic

Token Mixing MLP

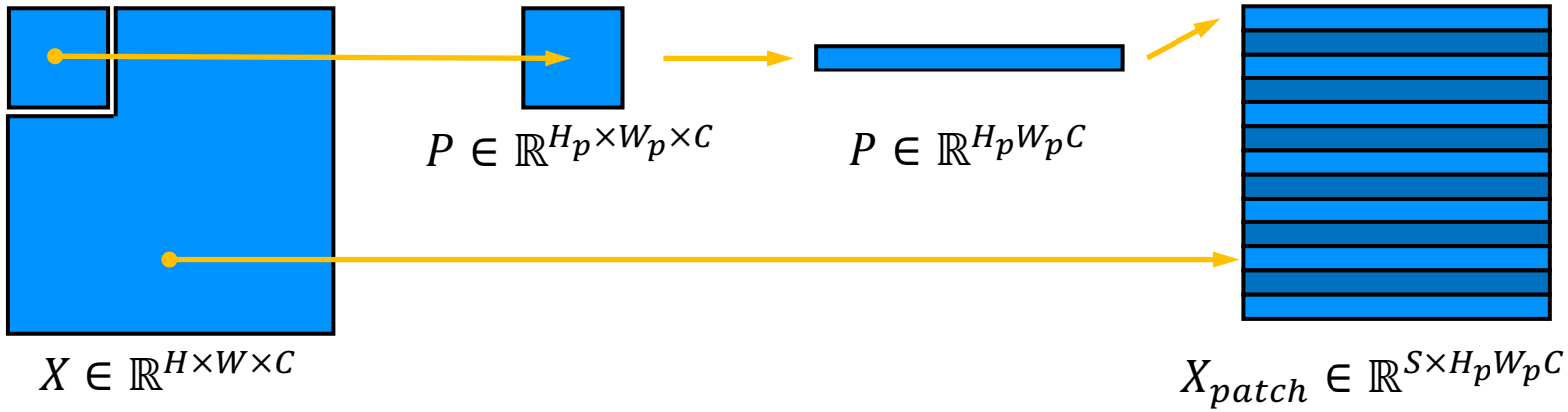


- Cross-location operations
- Communication between spatial location (token)
- Spatial specific & channel agnostic

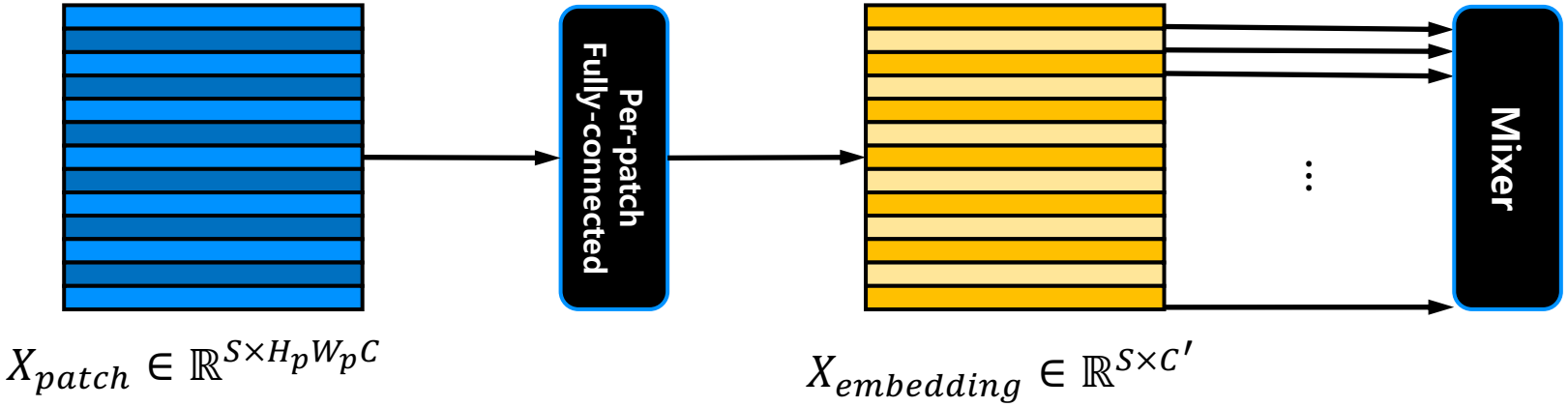
Mixer Details

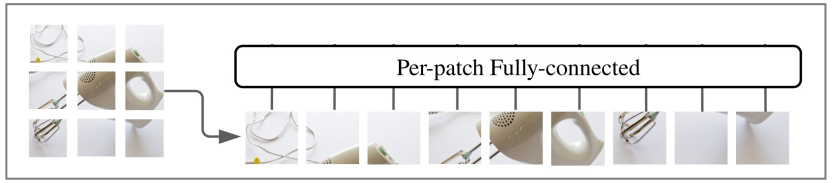


Step 1: Per-patch Linear Embeddings



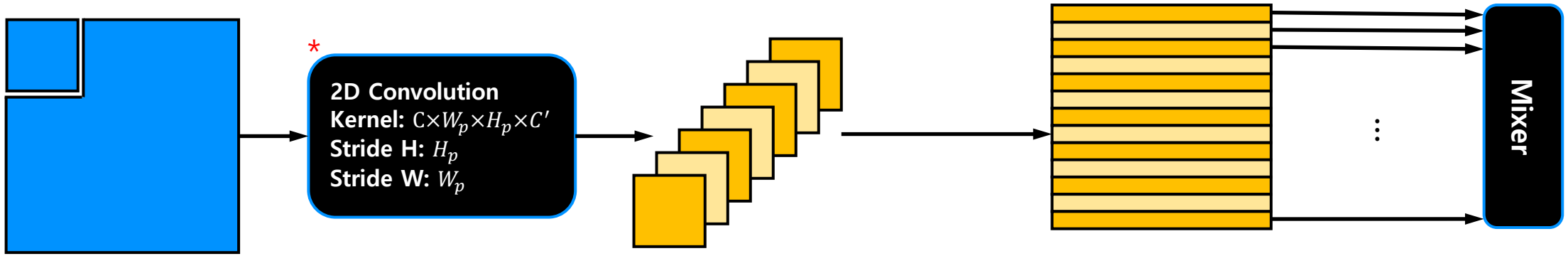
Number of Patch: $S = HW/H_pW_p$





Trick

Step 1: Per-patch Linear Embeddings



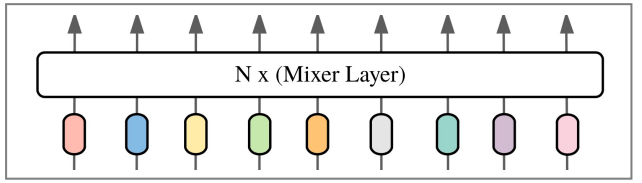
$X \in \mathbb{R}^{H \times W \times C}$

$X_{patch} \in \mathbb{R}^{(H/H_p) \times (W/W_p) \times C'}$

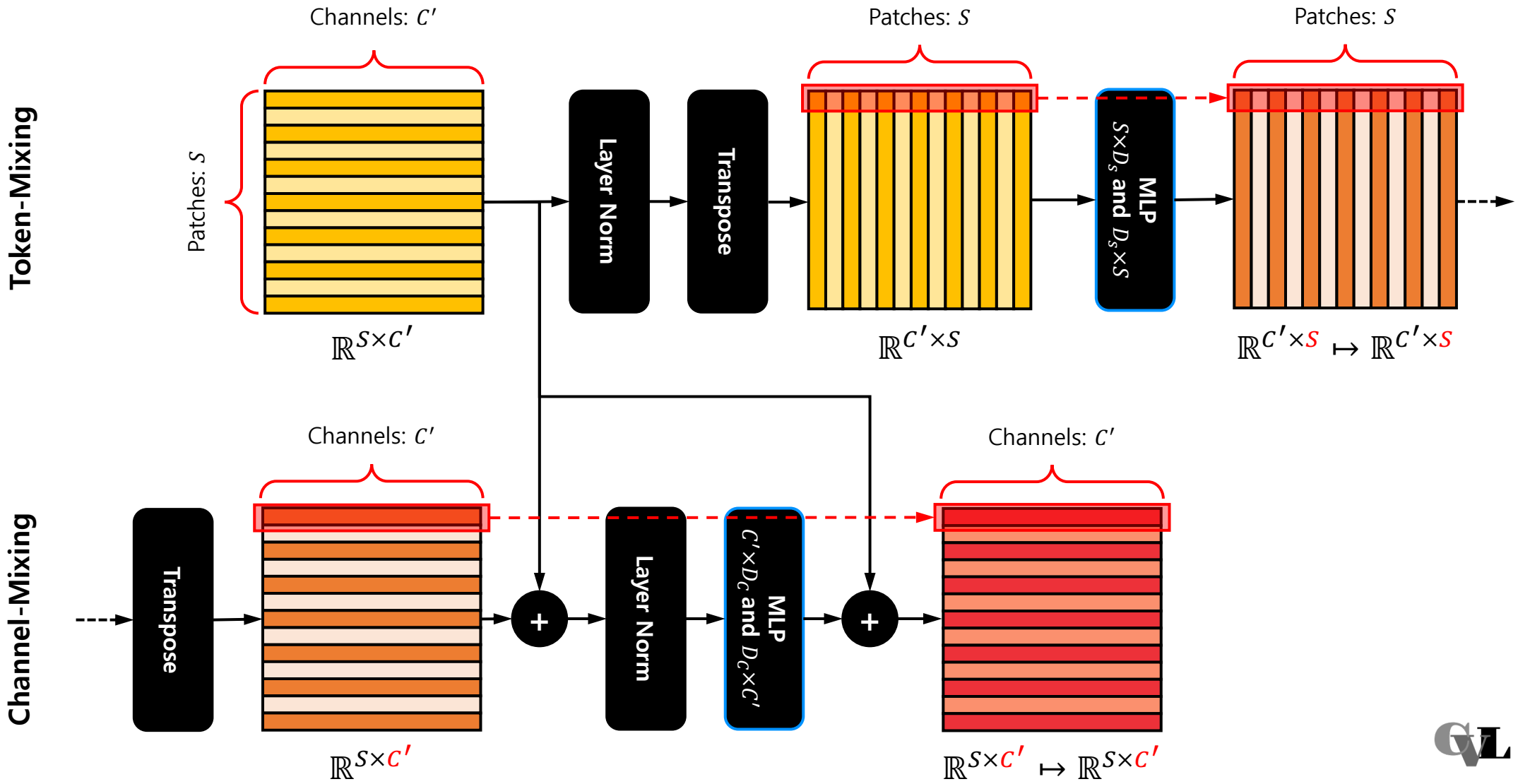
$X_{embedding} \in \mathbb{R}^{S \times C'}$

Number of Patch: $S = \left(\frac{H}{H_p}\right) \times \left(\frac{W}{W_p}\right) = HW/H_pW_p$

* Replace Per-patch Fully-connected with 2D convolution.

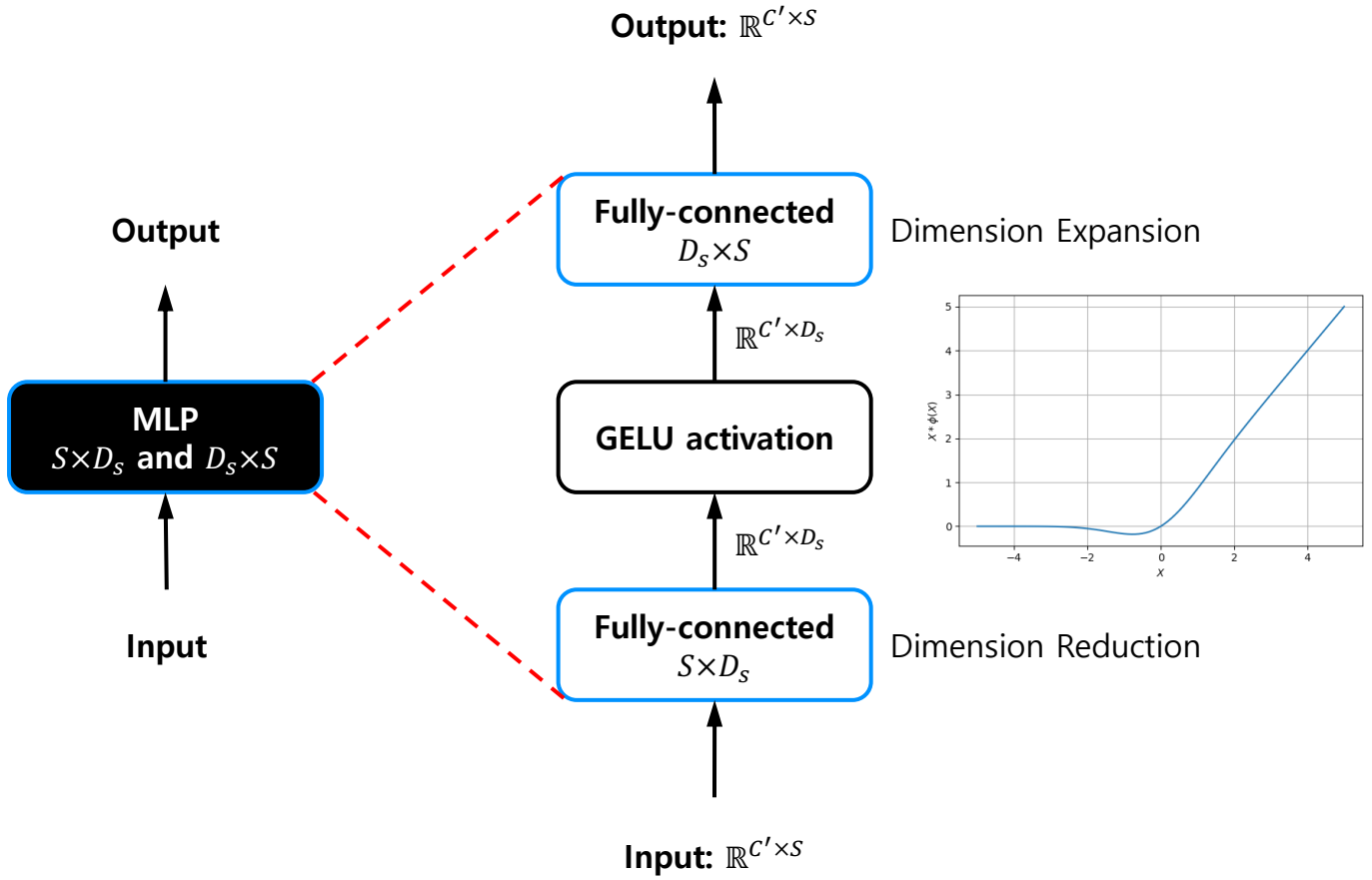


Step 2: Mixer Layer



Note: no dimension change after passing through the Mixer layer

Step 2: Mixer Layer (MLP)



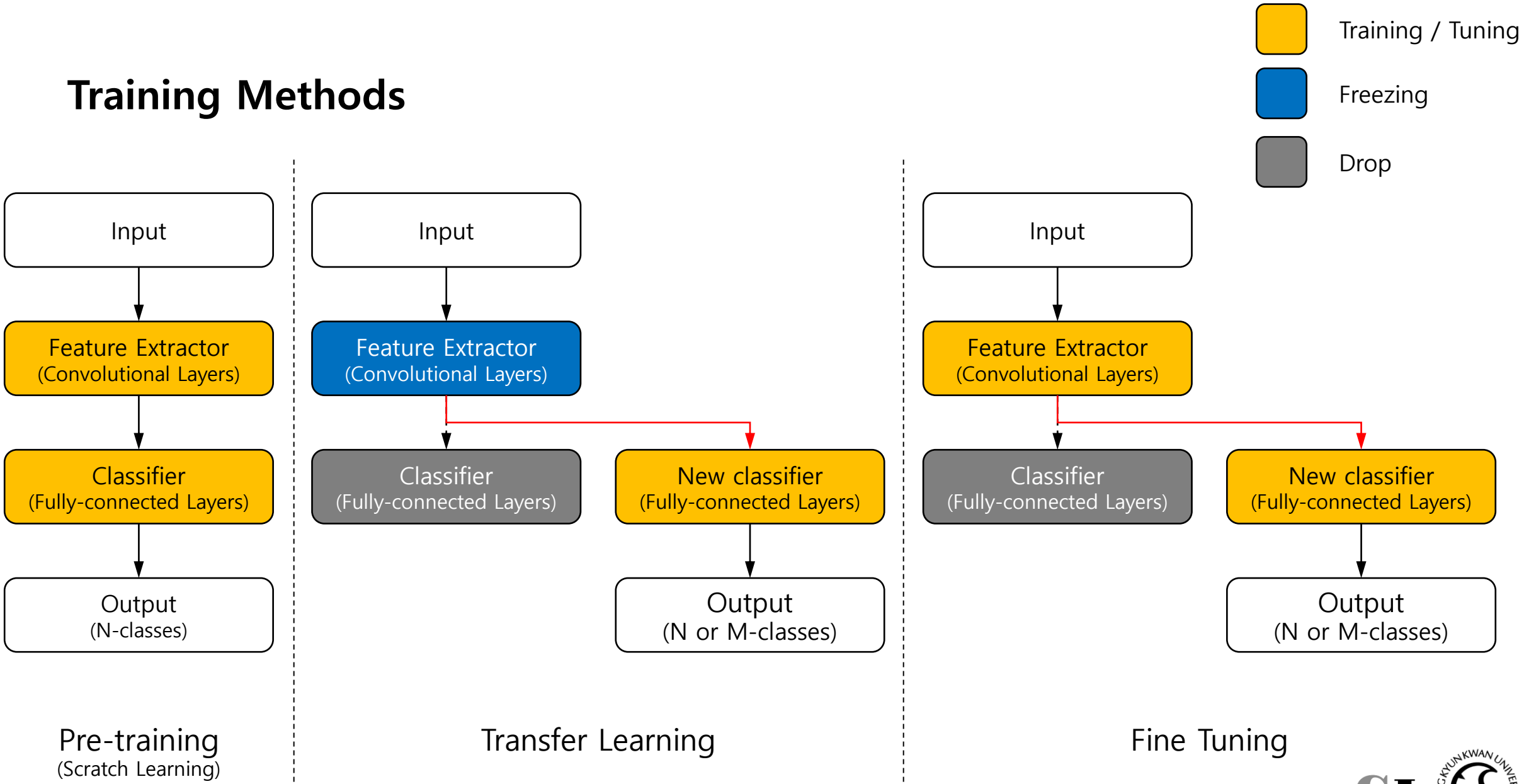
Mixer architectures

Table 1: Specifications of the Mixer architectures. The "B", "L", and "H" (base, large, and huge) model scales follow Dosovitskiy et al. [14]. A brief notation "B/16" means the model of base scale with patches of resolution 16×16. The number of parameters is reported for an input resolution of 224 and does not include the weights of the classifier head.

Specification	S/32	S/16	B/32	B/16	L/32	L/16	H/14
Number of layers	8	8	12	12	24	24	32
Patch resolution $P \times P$	32×32	16×16	32×32	16×16	32×32	16×16	14×14
Hidden size C	512	512	768	768	1024	1024	1280
Sequence length S	49	196	49	196	49	196	256
MLP dimension D_C	2048	2048	3072	3072	4096	4096	5120
MLP dimension D_S	256	256	384	384	512	512	640
Parameters (M)	19	18	60	59	206	207	431
		Small		Base		Large	Huge

Experiments

Training Methods



- **ImNet**: ImageNet (1k classes, 1.3M images)
- **Real**: Reassessed Labels (cleaned-up
- **Avg 5**: average performance of five datasets
  ImageNet
  CIFAR-10/100 (10/100 classes, 50k images)
  Oxford Pets (36 classes, 3.7k images)
  Oxford Flowers (102 classes, 2k images)
- **VTAB-1k**: Visual Task Adaptation

Transfer Learning

Freeze Feature Extractor

The larger the data set, **the better the cost-effectiveness** of MLP-Mixer.

Table 2: Transfer performance, inference throughput, and training cost. The rows are sorted by inference throughput (fifth column). Mixer has comparable transfer accuracy to state-of-the-art models with similar cost. The Mixer models are fine-tuned at resolution 448. Mixer performance numbers are averaged over three fine-tuning runs and standard deviations are smaller than 0.1.

	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) Medium-scale dataset						
CNNs · 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) Large-scale dataset						
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 [34]	90.0	91.12	—	—	—	20.48k
ALIGN [21]	88.64	—	—	79.99	15	14.82k

Training cost

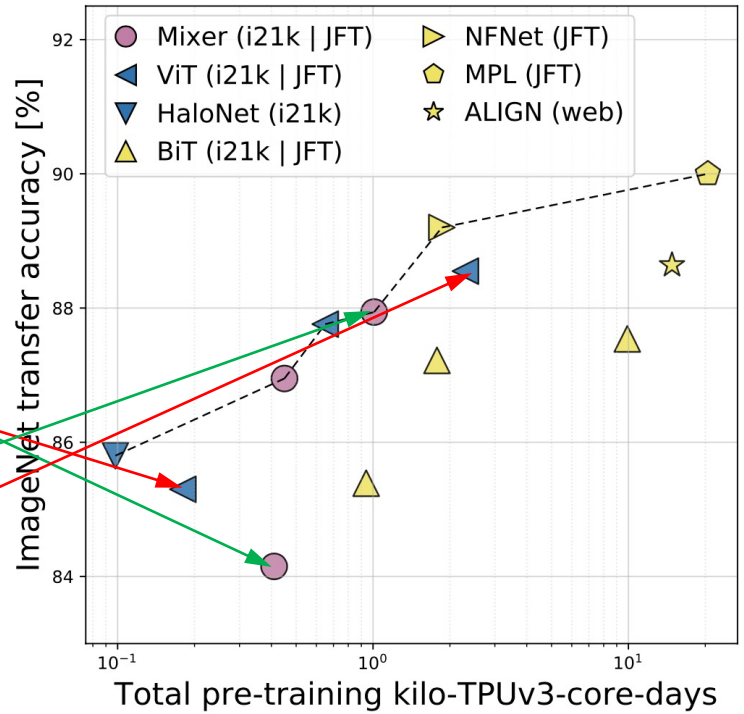
GVL SUNG KYUN KWAN UNIVERSITY 1398 성균관대학교

Transfer Learning

Freeze Feature Extractor

Table 2: Transfer performance, inference throughput, and training cost. The rows are sorted by inference throughput (fifth column). Mixer has comparable transfer accuracy to state-of-the-art models with similar cost. The Mixer models are fine-tuned at resolution 448. Mixer performance numbers are averaged over three fine-tuning runs and standard deviations are smaller than 0.1.

	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 [34]	90.0	91.12	—	—	—	20.48k
● ALIGN [21]	88.64	—	—	79.99	15	14.82k

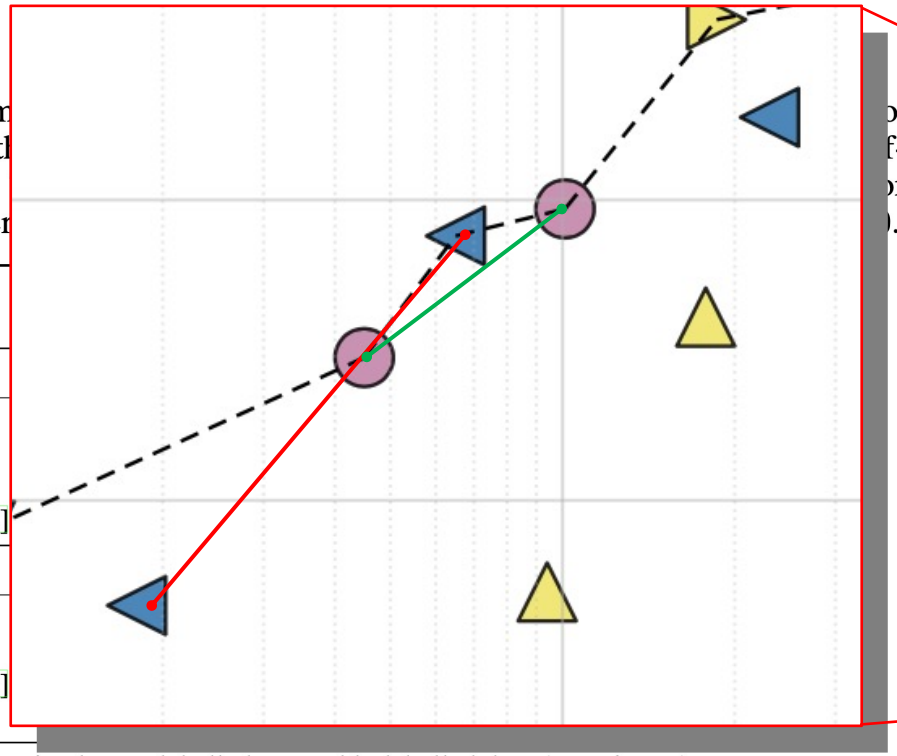


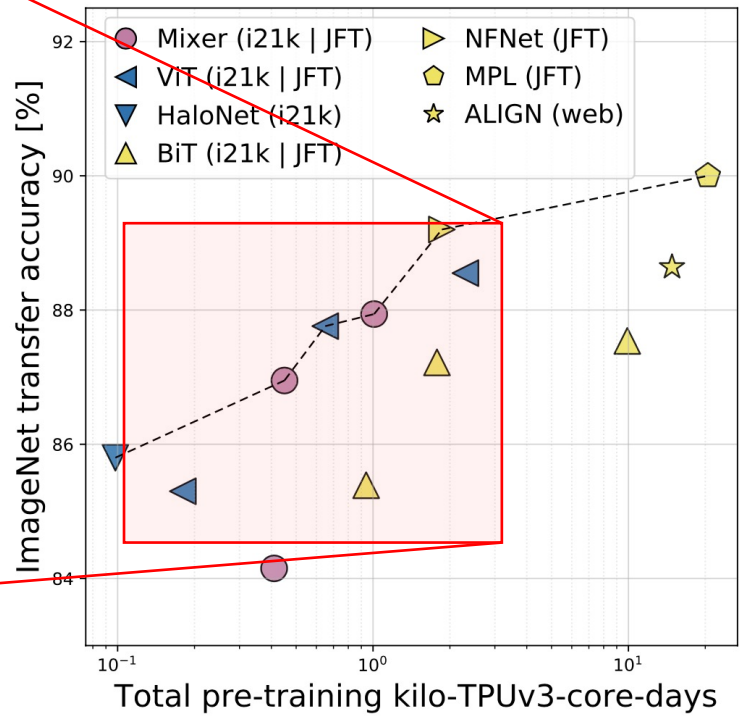
Transfer Learning

Freeze Feature Extractor

Table 2: Transfer perform … orted by inference throughput (fifth … f-the-art models with similar cost. … ormance numbers are averaged over … .1.

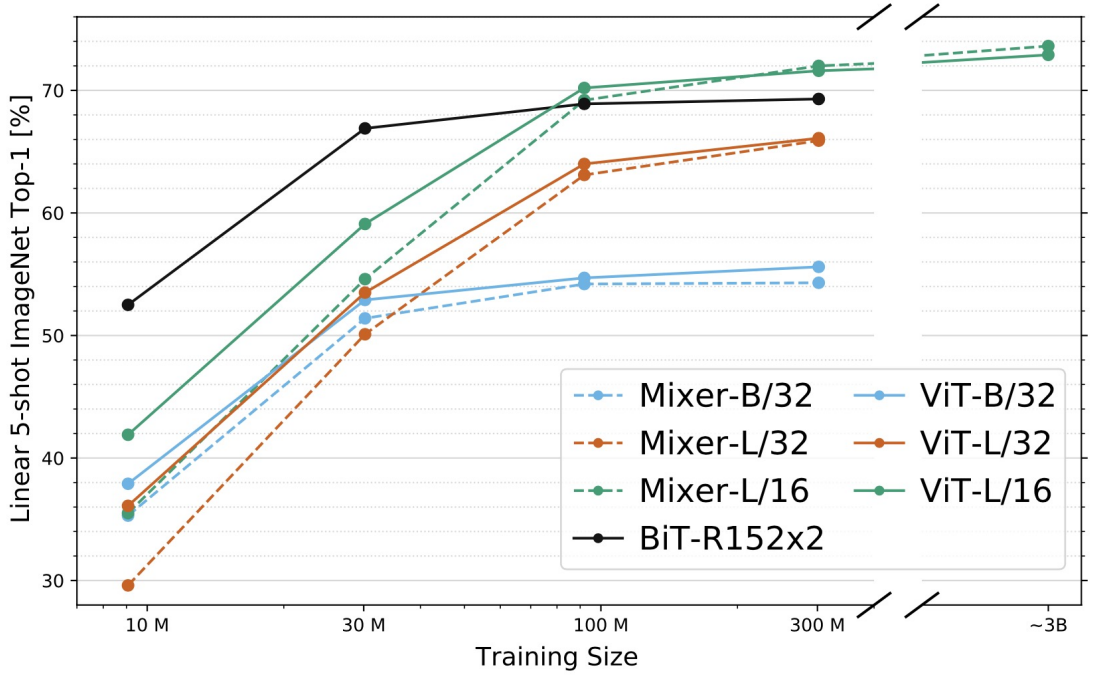
HaloNet [51]						
Mixer-L/16						
ViT-L/16 [14]						
BiT-R152x4 [22]						
NFNet-F4+ [7]						
Mixer-H/14						
BiT-R152x4 [22]						
ViT-H/14 [14]						
Pre-trained on unlabelled or weakly labelled data (proprietary)						
MPL [34]	90.0	91.12	—	—	—	20.48k
ALIGN [21]	88.64	—	—	79.99	15	14.82k



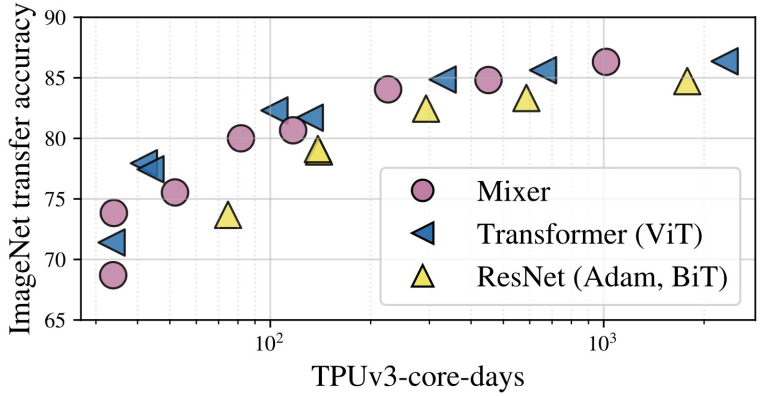


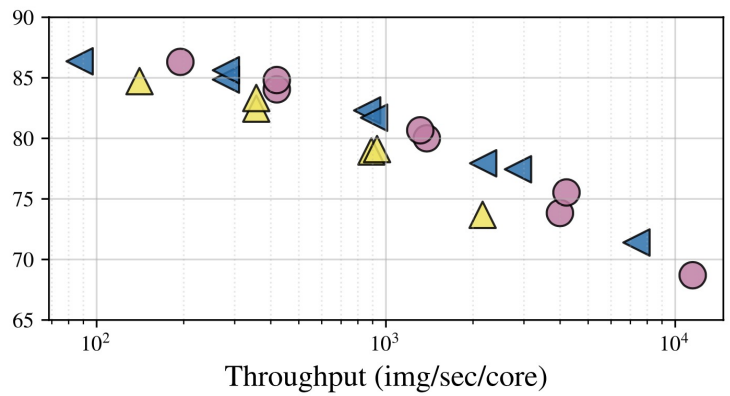
Transfer Learning

Freeze Feature Extractor



- When a medium-scale dataset is used for pre-training, MLP-Mixer shows lower cost-effectiveness.
- When the size of the dataset grows beyond a 300M-scale, MLP-Mixer shows the highest classification performance.





- **ImageNet:** 1k classes, 1.3M images
- **ImageNet-21k:** 21k classes, 14M images

Fine Tuning

Tune all parameters

Table 3: Performance of Mixer and other models from the literature across various model and pre-training dataset scales. "Avg. 5" denotes the average performance across five downstream tasks. Mixer and ViT models are averaged over three fine-tuning runs, standard deviations are smaller than 0.15. (‡) Extrapolated from the numbers reported for the same models pre-trained on JFT-300M without extra regularization. (☎) Numbers provided by authors of Dosovitskiy et al. [14] through personal communication. Rows are sorted by throughput.

	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

Shuffle Invariant

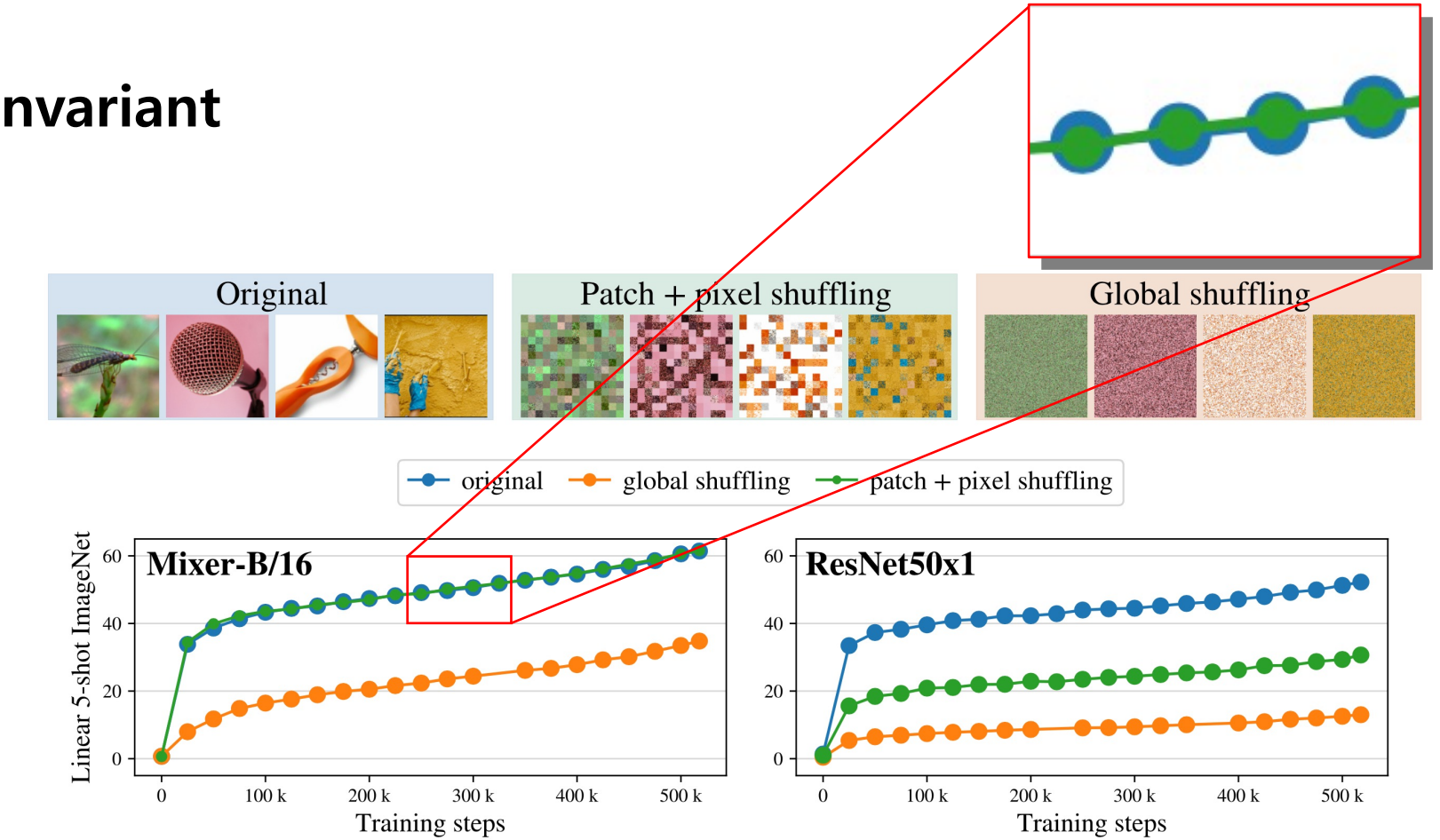


Figure 4

Token mixing and channel mixing are guaranteed to be invariant for patching and pixel shuffling, respectively.

Recap. of MLP-Mixer

Purpose

- Eliminate and beyond the Convolution and Attention.
- Also eliminate positional embedding.
- Main purpose is not to demonstrate state-of-the-art results.

Contributions

- Positional embedding & invariance
  - Eliminate positional embedding by per-patch Fully-connected (ViT alternative).
  - Also, the channel mixing operation provides positional invariance (CNN alternative).
- Computational complexity
  - The computational complexity of the network is linearly increased.
- Classification performance
  - Higher classification performance than ViT when training with a large-scale dataset (JFT-300M).

Limitations

- Requires large-scale dataset
  - Because of lower inductive bias.
  - With a medium-scale dataset (smaller than 100M), MLP-Mixer shows…
    - lower classification performance than ViT.
    - Lower cost-effectiveness than ViT.
- Like the ViT, the Mixer cannot handle various input sizes that fully convolutional network (FCN) can.

Model	Operation	Detail	Big-O
MLP-Mixer	Mixing operation	Matrix multiplication (Fully-connected)	$O(D)$
ViT	Multi-head attention	Matrix multiplication (Key x Query) after matrix multiplication (Key, Query, Value)	$O(D \times D)$

Appendix-A

Experimental Materials

Source Codes

- **Official Source Code:** https://github.com/google-research/vision_transformer

- **Non-official Source Code:** https://github.com/YeongHyeon/MLP-Mixer-TF2

ReaL dataset

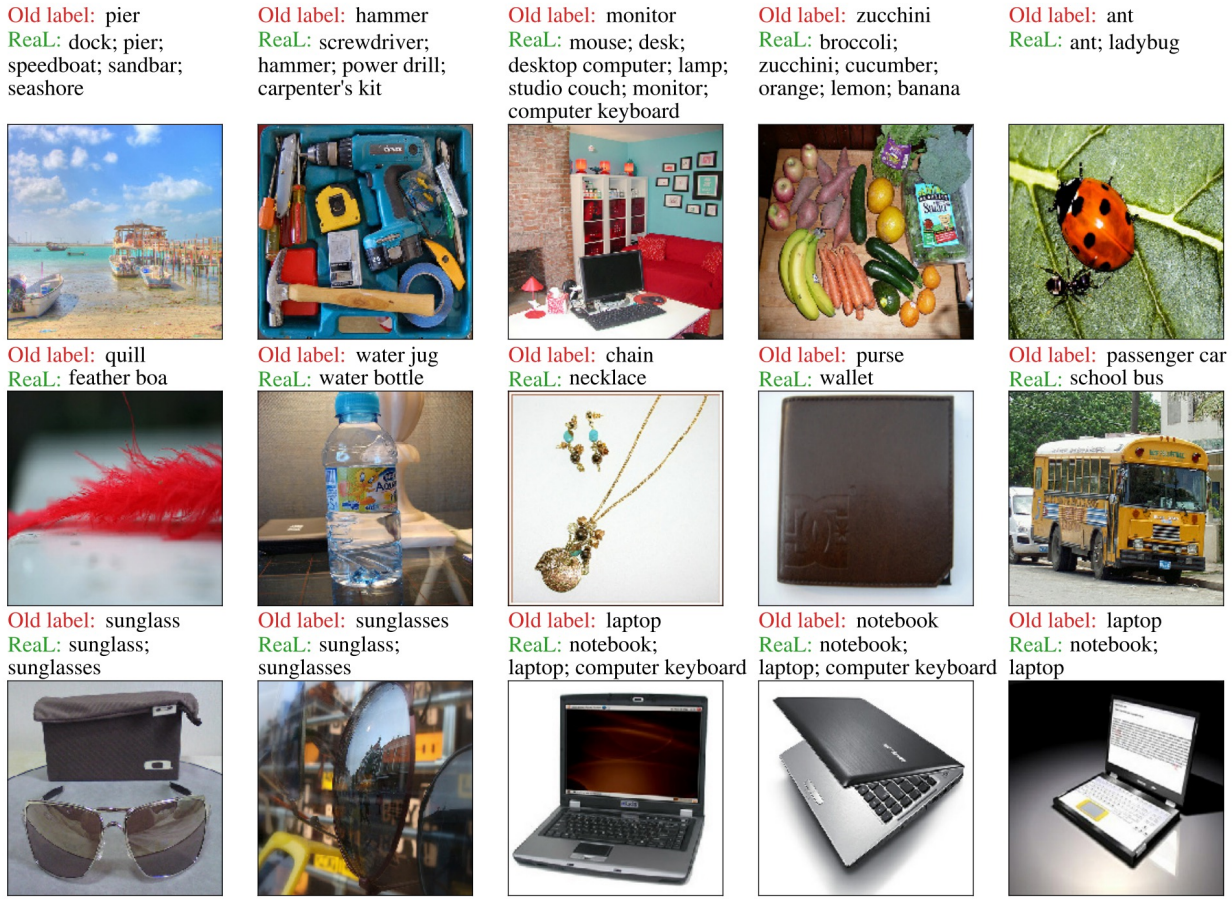


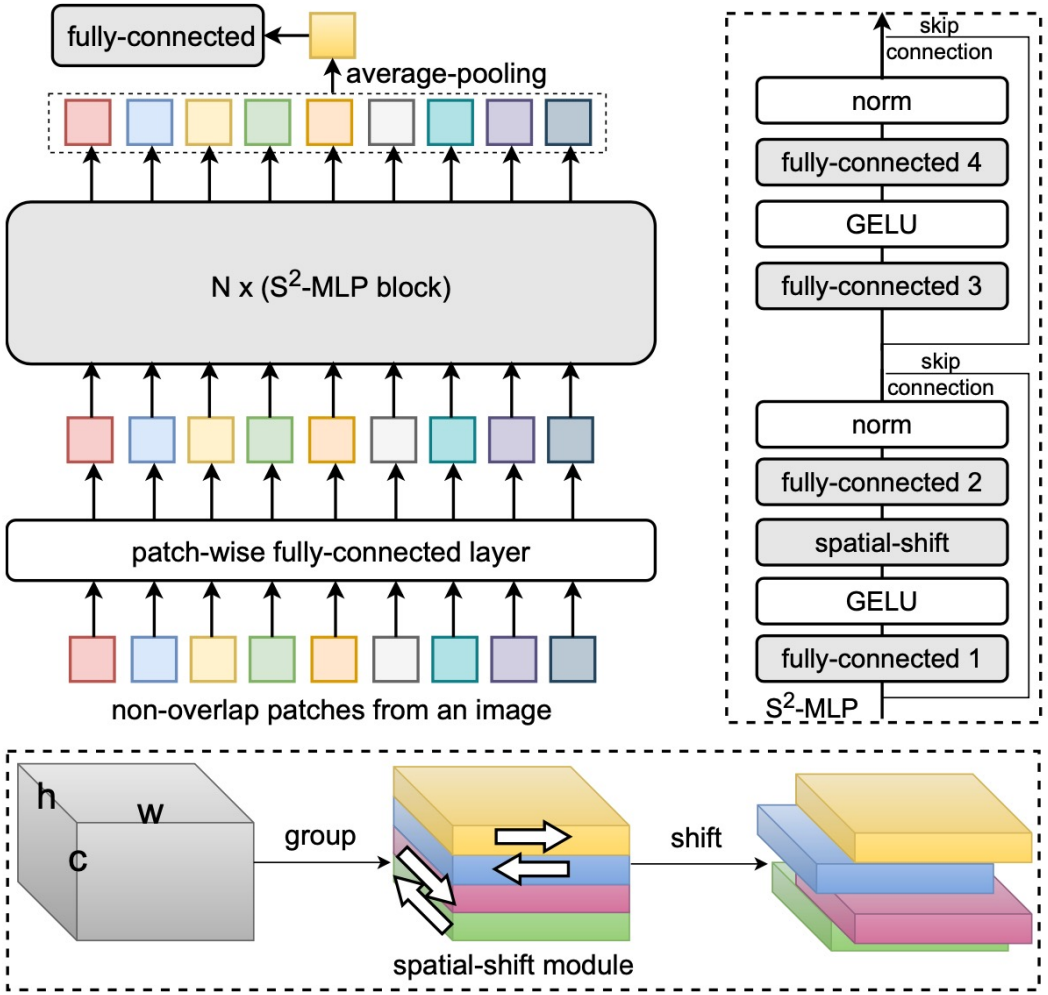
Figure 2: Example failures of the ImageNet labeling procedure. Red: original ImageNet label, green: proposed ReaL labels. **Top row**: ImageNet currently assigns a single label per image, yet these often contain several equally prominent objects. **Middle row**: Even when a single object is present, ImageNet labels present systematic inaccuracies due to their labeling procedure. **Bottom row**: ImageNet classes contain a few unresolvable distinctions.

Beyer, Lucas, et al. "Are we done with imagenet?." *arXiv preprint arXiv:2006.07159*. 2020.

Appendix-B

Mixer Variants

S^2-MLP



- S^2-MLP enables to process the various sized images via removing token mixing.
- S^2-MLP shows improved performance compared to MLP-Mixer.

Yu, Tan, et al. "S2-mlp: Spatial-shift mlp architecture for vision." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022.

S^2-MLP

Model	Resolution	Top-1 (%)	Top5 (%)	Params (M)	FLOPs (B)
CNN-based					
ResNet50 [14]	224 × 224	76.2	92.9	25.6	4.1
ResNet152 [14]	224 × 224	78.3	94.1	60.2	11.5
RegNetY-8GF [31]	224 × 224	79.0	–	39.2	8.0
RegNetY-16GF [31]	224 × 224	80.4	–	83.6	15.9
EfficientNet-B3 [35]	300 × 300	81.6	95.7	12	1.8
EfficientNet-B5 [35]	456 × 456	84.0	96.8	30	9.9
Transformer-based					
ViT-B/16 [9]	384 × 384	77.9		86.4	55.5
ViT-B/16* [9, 36]	224 × 224	79.7	–	86.4	17.6
DeiT-B/16 [38]	224 × 224	81.8	–	86.4	17.6
PiT-B/16 [16]	224 × 224	82.0	–	73.8	12.5
PVT-Large [42]	224 × 224	82.3	–	61.4	9.8
CPVT-B [7]	224 × 224	82.3	–	88	17.6
TNT-B [13]	224 × 224	82.8	96.3	65.6	14.1
T2T-ViT$_t$-24 [46]	224 × 224	82.6	–	65.1	15.0
CaiT-S32 [40]	224 × 224	83.3	–	68	13.9
Swin-B [28]	224 × 224	83.3	–	88	15.4
Nest-B [48]	224 × 224	83.8	–	68	17.9
Container [11]	224 × 224	82.7	–	22.1	8.1
MLP-based ($c = 768$, $N = 12$)					
Mixer-B/16 [36]	224 × 224	76.4	–	59	11.6
FF [30]	224 × 224	74.9	–	59	11.6
S^2-MLP-wide (ours)	224 × 224	80.0	94.8	71	14.0
MLP-based ($c = 384$, $N = 36$)					
ResMLP-36 [37]	224 × 224	79.7	–	45	8.9
S^2-MLP-deep (ours)	224 × 224	80.7	95.4	51	10.5

Table 3. Results on ImageNet-1K without extra data. ViT-B/16* denotes the ViT-B/16 model in MLP-mixer [36] with extra regularization.

ConvMLP: Hierarchical Convolutional MLPs

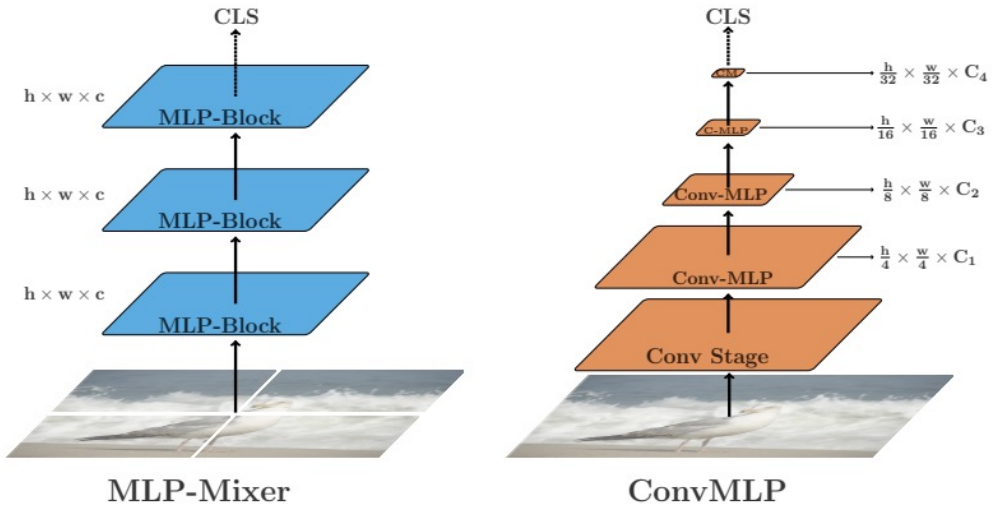


Figure 1: Comparing MLP-Mixer to ConvMLP. ConvMLP adopts a simple hierarchical multi-stage co-design of convolutions and MLPs and achieves both more suitable representations as well as better accuracy vs computation trade-offs for visual recognition tasks including classification, detection and segmentation.

Li, Jiachen, et al. "Convmlp: Hierarchical convolutional mlps for vision." *arXiv preprint*. 2021.

ConvMLP: Hierarchical Convolutional MLPs

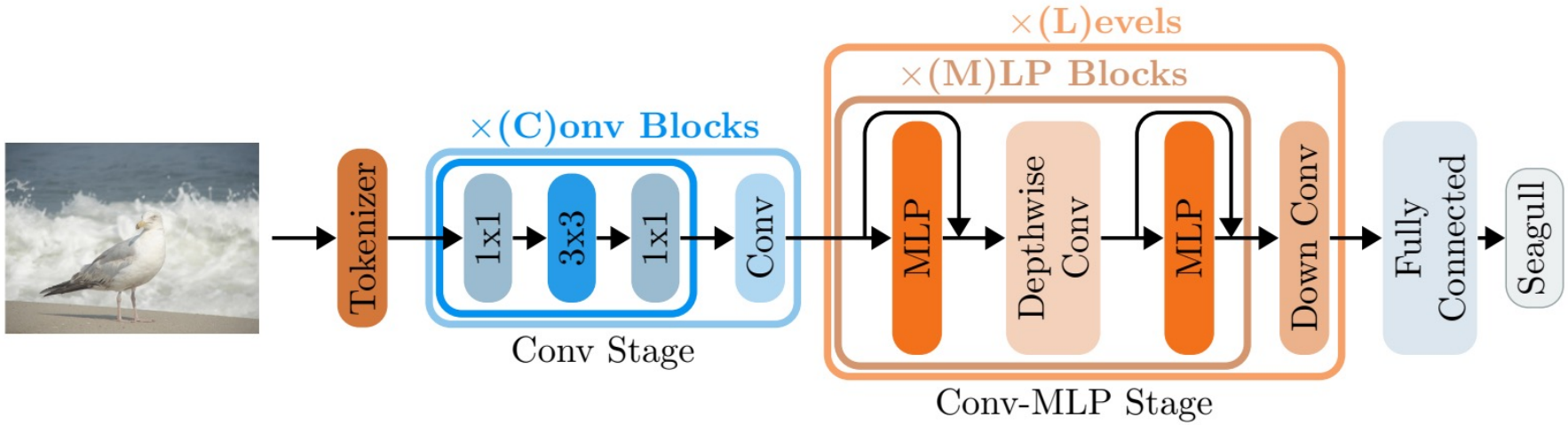


Figure 2: Overview of ConvMLP framework. The Conv Stage consists of C convolutional blocks with 1×1 and 3×3 kernel sizes. The MLP-Conv Stage consists of Channelwise MLPs, with skip layers, and a depthwise convolution. This is repeated M times before a down convolution is utilized to express a level L. A level down samples an image $\mathcal{L} : h \times w \times c \mapsto \frac{h}{2^L} \times \frac{w}{2^L} \times 2^L c$

Li, Jiachen, et al. "Convmlp: Hierarchical convolutional mlps for vision." *arXiv preprint*. 2021.

Appendix-C

Mixer Applications

Artwork Style Recognition

WikiArt (including 27 artwork styles)

- Abstract Expressionism (2782 images)
- Action Painting (92 images)
- Analytical Cubism (110 images)

  …
- Ukiyo-e (1167 images).

Table 2. ViT performance in artwork style recognition.

Optimizer	Accuracy
Adam	39.89%
Adamax	39.42%
Optimistic Adam	39.71%
SGD	39.28%
MGD	39.31%
RMSProp	38.97%

Table 4. MLP Mixer performance in artwork style recognition.

Model	Accuracy
MLP Mixer	39.59%

Iliadis, Lazaros Alexios, et al. "Artwork Style Recognition Using Vision Transformers and MLP Mixer." *MDPI Technologies*. 2021.

Point Cloud Reconstruction

Table 3. Point cloud reconstruction results.

Method	ShapeNet-Part [90]				ScanNet [9]				ICL-NUIM [20]			
	Cf.(↓)	Acc.(↑)	Cp.(↑)	F1(↑)	Cf.(↓)	Acc.(↑)	Cp.(↑)	F1(↑)	Cf.(↓)	Acc.(↑)	Cp.(↑)	F1(↑)
PointNet [54]	1.33	63.2	38.8	48.2	3.05	37.5	27.8	32.6	2.98	46.9	33.2	38.1
PointNet++ [55]	1.25	65.1	39.0	50.1	2.97	38.3	29.5	33.4	2.88	48.8	35.8	39.9
PointRecon [6]	1.19	**81.0**	40.4	53.4	2.86	40.4	30.2	34.1	2.78	54.1	38.1	43.6
PointTrans [99]	1.12	75.9	40.9	52.7	2.79	41.1	32.1	35.6	2.57	51.1	36.4	41.6
PointMixer (ours)	**1.11**	77.1	**41.5**	**53.7**	**2.74**	**42.1**	**33.5**	**37.8**	**2.43**	**56.5**	**38.2**	**44.7**

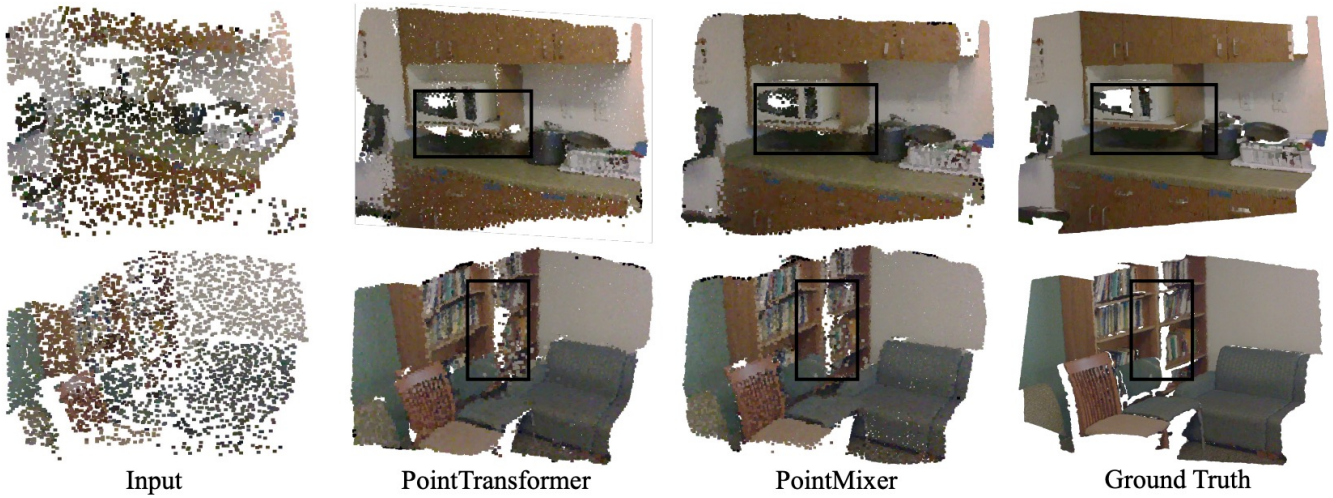


Fig. 7. Qualitative results in point reconstruction on ScanNet dataset [9].

Pose Estimation

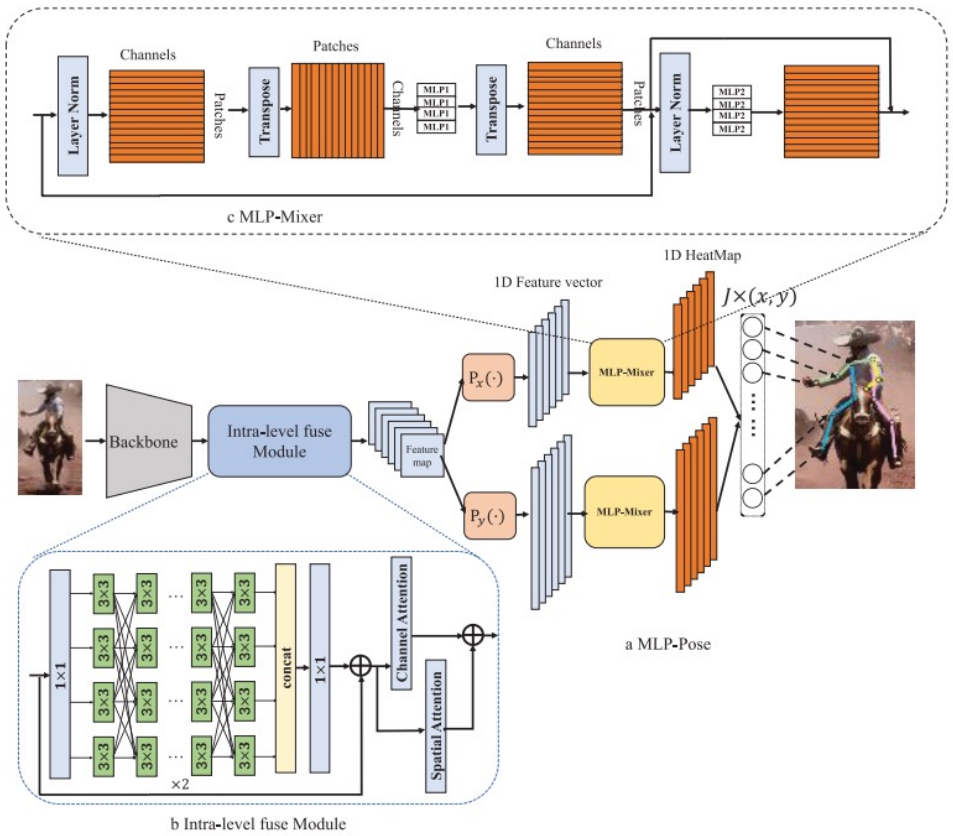


Figure 1. (a) The MLP-Pose model architecture. (b) Intra-level fuse Module, which can adaptively learn the relationship of features in the layer. (c) MLP-Mixer for learning the global relationship among features.

Figure 2 Qualitative COCO human pose estimation results of MLP-Pose. Left: result of single person. Right: result of multi-person.

Xiong, Songkai, et al. "MLP-Pose: Human Pose Estimation by MLP-Mixer." *IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS).* 2021.

Cross-View Image Translation

Figure 3: Qualitative results of different methods on (a) Dayton and (b) CVUSA datasets.

Ren, Bin, Hao Tang, and Nicu Sebe. "Cascaded Cross MLP-Mixer GANs for Cross-View Image Translation." *arXiv preprint.* 2021.