

---

# MLP-Mixer: An all-MLP Architecture for Vision

---

School of Industrial and Management Engineering, Korea University

Lee Kyung Yoo

# Contents

---

- ❖ Research Purpose
- ❖ MLP-Mixer
- ❖ Experiments
- ❖ Conclusion

# Research Purpose

---

## ❖ MLP-Mixer: An all-MLP Architecture for Vision (arXiv, 2021)

- Google Research, Brain Team에서 발표한 논문이며 2021년 07월 23일 기준으로 30회 인용

---

### 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, neilhoulby, akolesnikov, lbeyer,  
xzhai, unterthiner, jessicayung<sup>†</sup>, andstein,  
keyzers, usz, lucic, adosovitskiy}@google.com

<sup>†</sup>work done during Google AI Residency

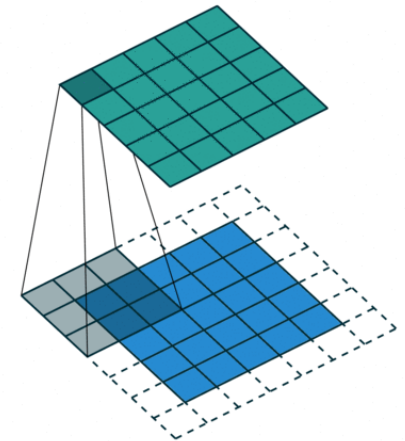
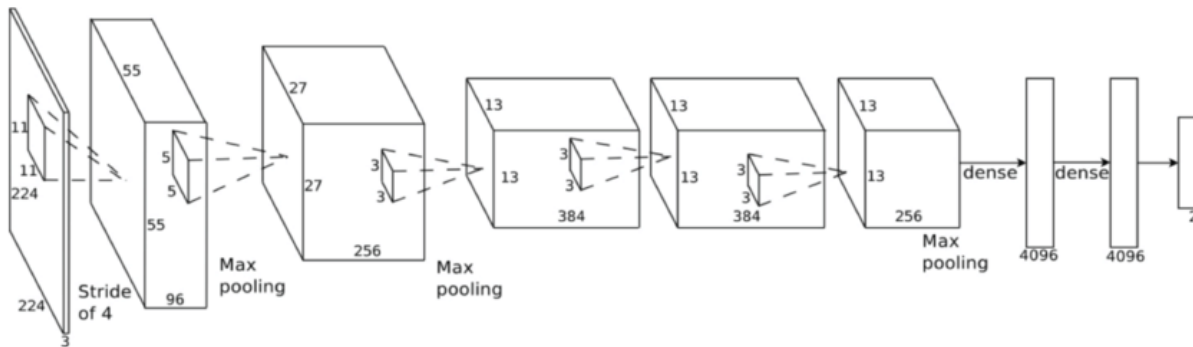
#### Abstract

Convolutional Neural Networks (CNNs) are the go-to model for computer vision. Recently, attention-based networks, such as the Vision Transformer, have also become popular. In this paper we show that while convolutions and attention are both sufficient for good performance, neither of them are necessary. We present *MLP-Mixer*, an architecture based exclusively on multi-layer perceptrons (MLPs). MLP-Mixer contains two types of layers: one with MLPs applied independently to image patches (i.e. “mixing” the per-location features), and one with MLPs applied across patches (i.e. “mixing” spatial information). When trained on large datasets, or with modern regularization schemes, MLP-Mixer attains competitive scores on image classification benchmarks, with pre-training and inference cost comparable to state-of-the-art models. We hope that these results spark further research beyond the realms of well established CNNs and Transformers. 

# Research Purpose

## ❖ MLP-Mixer: An all-MLP Architecture for Vision (arXiv, 2021)

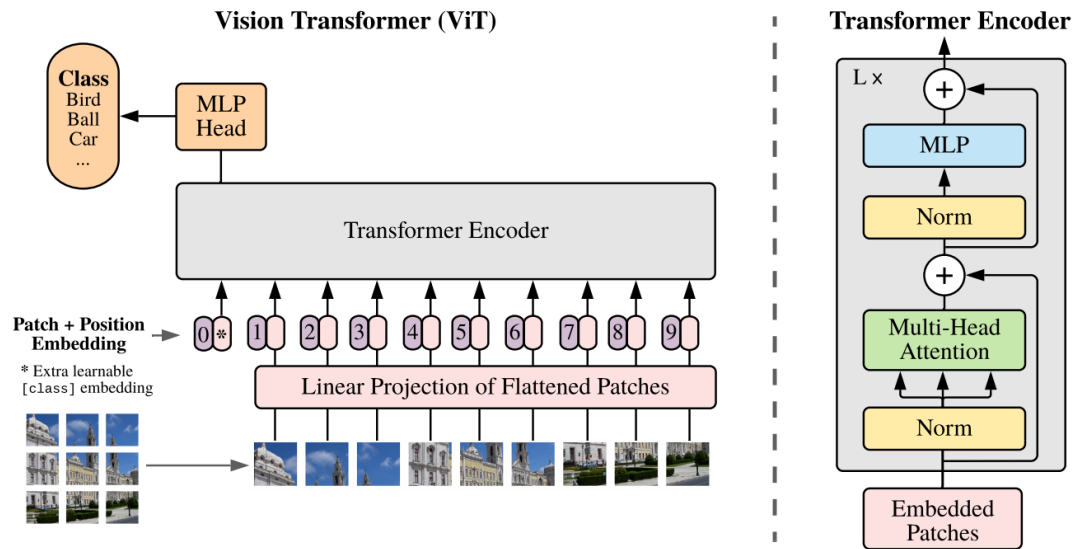
- Computer Vision 분야에서 de-facto standard로 자리잡은 Convolutional Neural Network(CNN)
  - Convolution, Pooling, Fully connected layer 등으로 구성
  - 전체가 아닌 정해진 size의 값들과 연산함으로써 local feature를 학습
  - 동일한 weight를 가진 filter를 sliding window로 연산



# Research Purpose

## ❖ MLP-Mixer: An all-MLP Architecture for Vision (arXiv, 2021)

- Computer Vision 분야에서 새롭게 SOTA를 달성한 Vision Transformer(ViT)
  - Linear projection of flattened patches, Transformer encoder 등으로 구성
  - Locally connected가 아닌 Fully connected
  - Input data에 따라 서로 다른 weight를 가짐



Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

# Research Purpose

## ❖ MLP-Mixer: An all-MLP Architecture for Vision (arXiv, 2021)

- Differences between CNN/ ViT/ MLP
  - **Inductive bias: CNN > ViT, MLP**
  - Model complexity: CNN, ViT > MLP

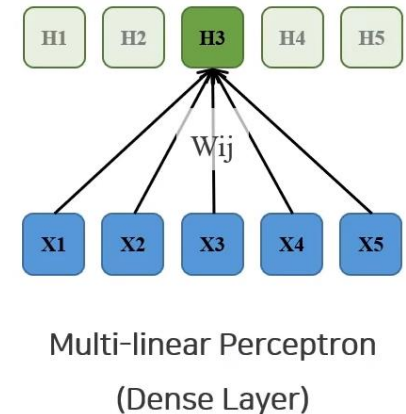
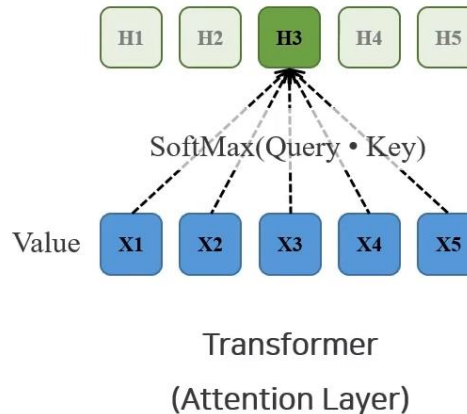
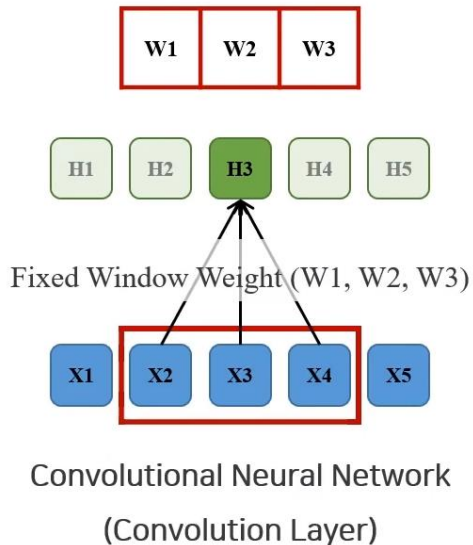
Component	Entities	Relations	Rel. inductive bias	Invariance
Fully connected	Units	All-to-all	Weak	-
Convolutional	Grid elements	Local	Locality	Spatial translation
Recurrent	Timesteps	Sequential	Sequentiality	Time translation
Graph network	Nodes	Edges	Arbitrary	Node, edge permutations

Table 1: Various relational inductive biases in standard deep learning components. See also Section 2.

# Research Purpose

## ❖ MLP-Mixer: An all-MLP Architecture for Vision (arXiv, 2021)

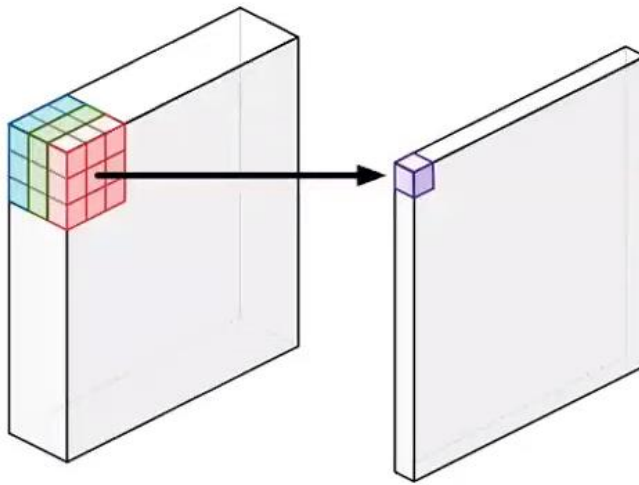
- Differences between CNN/ ViT/ MLP
  - Inductive bias: CNN > ViT, MLP
  - **Model complexity: CNN, ViT > MLP**



# Research Purpose

## ❖ MLP-Mixer: An all-MLP Architecture for Vision (arXiv, 2021)

- We DON'T NEED NEITHER!
- Convolutions와 attention을 사용하지 않은 비교적 간단한 구조 제안
- 현대 deep vision architectures를 공통적으로 구성하고 있는 레이어를 대체 하자는 것이 main idea
  - (1) Mix features between spatial locations
  - (2) Mix features at a given spatial location



Convolution-based architecture

**In CNN,**

Pooling = (1)에 해당

1\*1 convolutions = (2) 에 해당

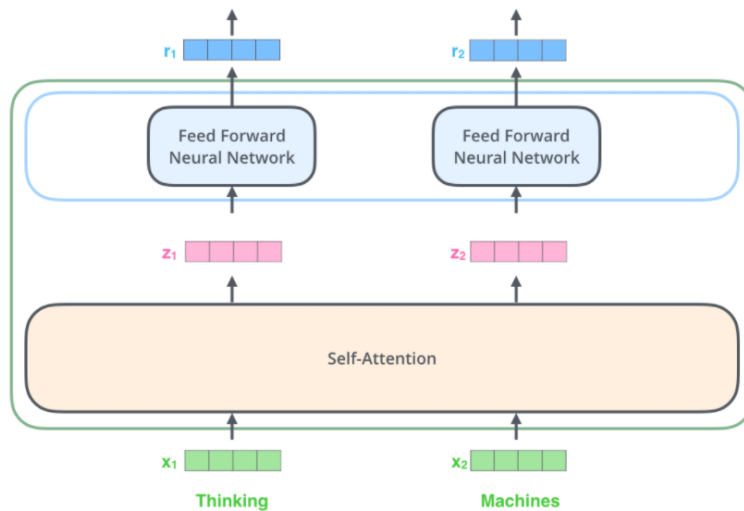
N\*N convolutions = (1),(2) 에 해당



# Research Purpose

## ❖ MLP-Mixer: An all-MLP Architecture for Vision (arXiv, 2021)

- We DON'T NEED NEITHER!
- Convolutions와 attention을 사용하지 않은 비교적 간단한 구조 제안
- 현대 deep vision architectures를 공통적으로 구성하고 있는 레이어를 대체 하자는 것이 main idea
  - (1) Mix features between spatial locations
  - (2) Mix features at a given spatial location



Attention-based architecture

**In Transformer,**

Self-attention = (1),(2) 에 해당

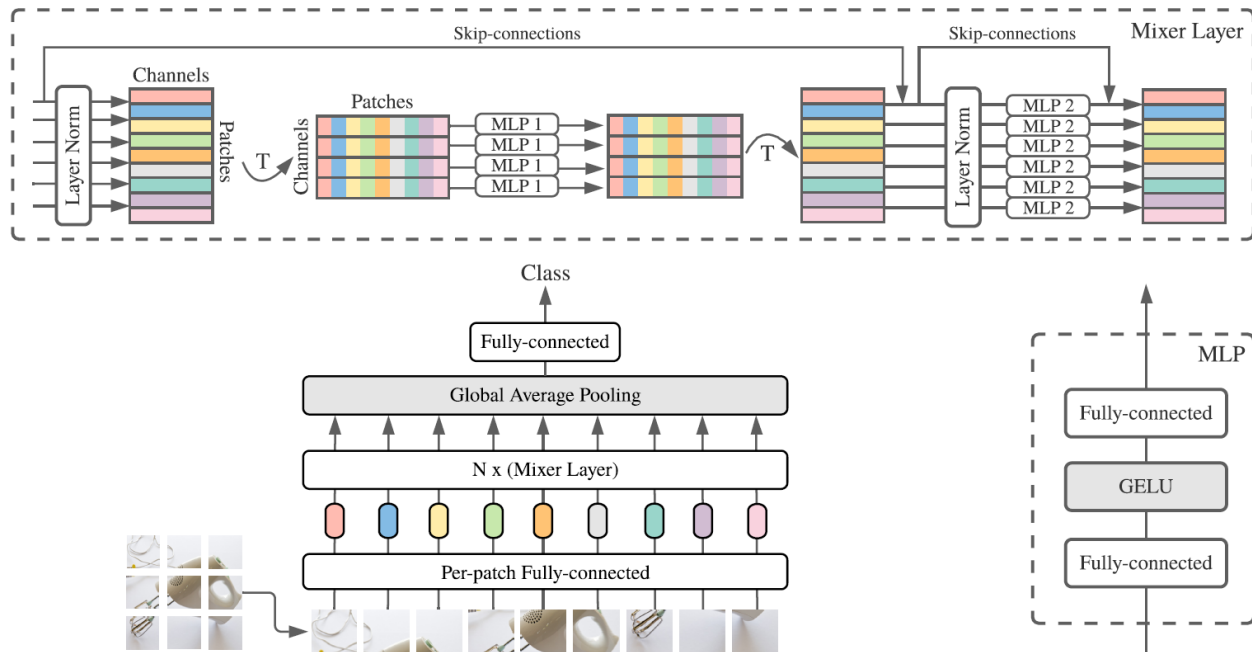
FFNN = (2)에 해당

# MLP-Mixer

## - Overview of MLP-Mixer

### ❖ Architecture

- 앞의 (1)과 (2)를 완전히 분리시켜 2가지 종류의 MLP로 구성
  - (1) **Token-mixing MLP**: Image patch들 간의 mixing을 이용하여 cross-location operations 수행
  - (2) **Channel-mixing MLP**: 하나의 Image Patch 내의 channel들 간의 mixing을 이용하여 per-location operations 수행

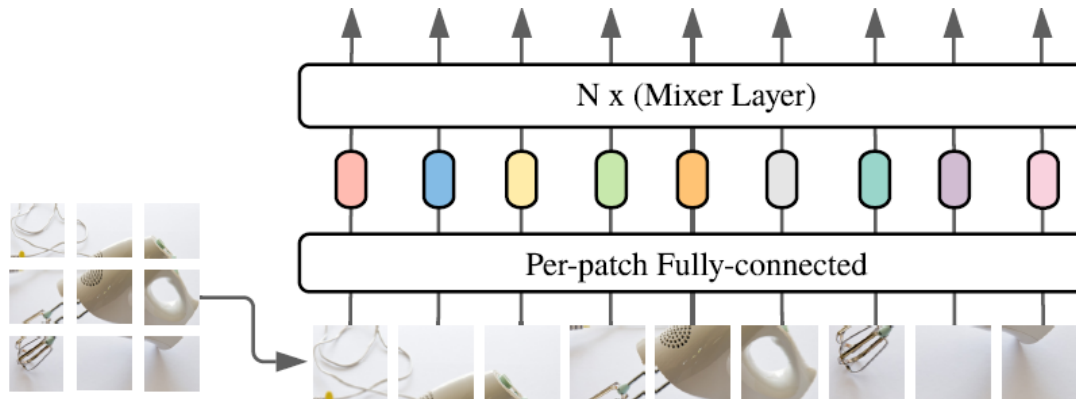


# MLP-Mixer

- Input ~ Per-patch Fully-connected

## ❖ Architecture

- S개의 Image patch들을 input sequence로 받고 각각의 patch들은 hidden dim C를 가짐
- 기존 image가 (H, W)의 resolution을 지니고 있다면, 각 image patch는 (P, P)의 resolution을 갖고 총 image patch의 수는  $S = HW/P^2$  로 계산됨
- 모든 image patch는 동일한 projection matrix에 의하여 linear projected



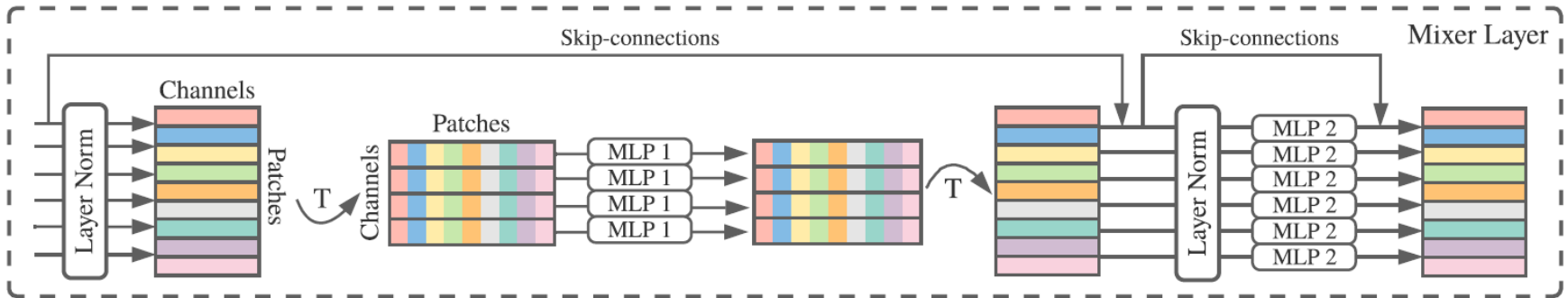
$$\mathbf{X} \in \mathbb{R}^{S \times C}$$

# MLP-Mixer

## - Mixer Layer

### ❖ Architecture

- 입력값  $X$ 를 전치하여 각 column에 Token-mixing MLP 적용
- 이후 다시 전치하여  $X$ 의 각 row에 channel-mixing MLP 적용
- 각 MLP는 2개의 fully-connected layer와 GELU로 구성되어 있으며 layer norm과 residual connection을 함께 사용



$$U_{*,i} = X_{*,i} + W_2 \sigma(W_1 \text{LayerNorm}(X)_{*,i}), \quad \text{for } i = 1 \dots C,$$

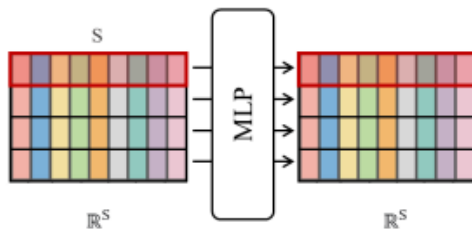
$$Y_{j,*} = U_{j,*} + W_4 \sigma(W_3 \text{LayerNorm}(U)_{j,*}), \quad \text{for } j = 1 \dots S.$$

# MLP-Mixer

## - Mixer Layer

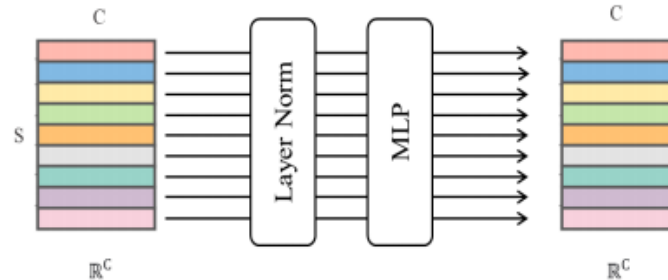
### ❖ Architecture

- $D_s, D_c$  는 token, channel mixing MLP 각각의 hidden layer width
- $D_s$  는 input patch들의 개수와 독립적으로 선택이 가능
  - ViT는 input patch들의 개수와 관계가 있기 때문에 연산량이 quadratic인 반면 MLP-Mixer는 input patch들의 개수에 linear함



**Token-mixing MLP**

$$\mathbb{R}^S \mapsto \mathbb{R}^S$$



**Channel-mixing MLP**

$$\mathbb{R}^C \mapsto \mathbb{R}^C$$

# Experiments

## ❖ Settings

- MLP-Mixer 모델의 퍼포먼스는 데이터셋 스케일별로 pre-trained 진행
- Model size는 Small, Base, Large, Huge로 나뉘어짐
- Model size 옆의 숫자는  $/16 =$  patches of resolution  $16 \times 16$ 와 같은 의미를 가짐

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 $P \times P$	$32 \times 32$	$16 \times 16$	$32 \times 32$	$16 \times 16$	$32 \times 32$	$16 \times 16$	$14 \times 14$
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

# Experiments

## ❖ Metrics for computational cost and quality

- For computational cost,
  - Total pre-training time on TPU-v3 accelerators
  - Pre-training에서의 total computational cost
- For model quality,
  - Top-1 downstream accuracy after fine-tuning

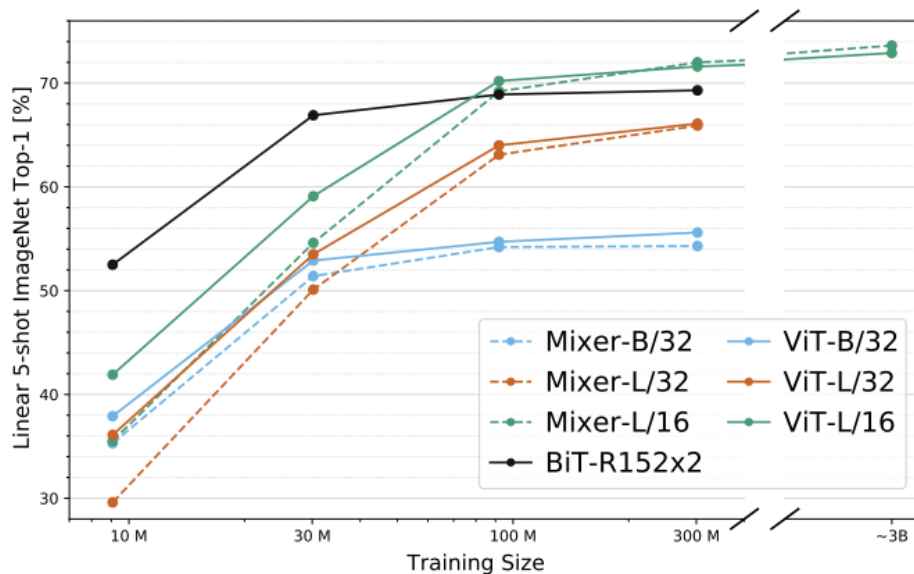
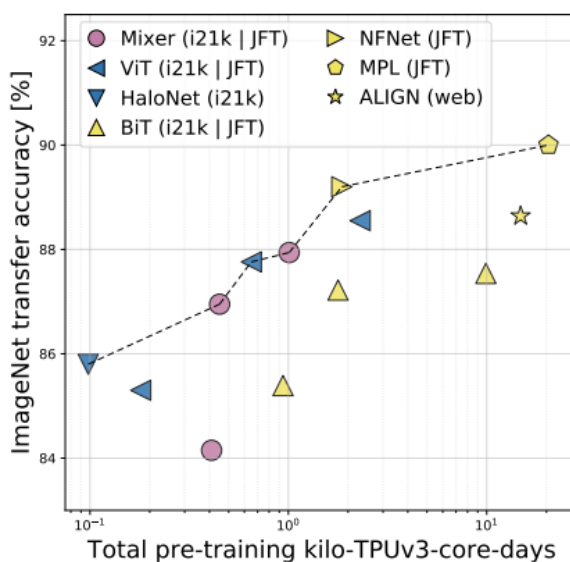
	ImNet top-1	ReaL top-1	Avg 5 top-1	VTAB-1k 19 tasks	Throughput img/sec/core	TPUv3 core-days
Pre-trained on ImageNet-21k (public)						
● HaloNet [51]	85.8	—	—	—	120	0.10k
● Mixer-L/16	84.15	87.86	93.91	74.95	105	0.41k
● ViT-L/16 [14]	85.30	88.62	94.39	72.72	32	0.18k
● BiT-R152x4 [22]	85.39	—	94.04	70.64	26	0.94k
Pre-trained on JFT-300M (proprietary)						
● NEfNet-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

- MLP-based Mixer model
- Convolution-based model
- Attention-based model

# Experiments

## ❖ Trade-off between accuracy and computational resources

- MLP-Mixer가 computational costs 대비 SOTA에 견줄 만한 accuracy를 보임
- 데이터 양이 충분해질 수록 더 좋은 성능을 보임( $\because$  inductive bias)
- 데이터가 적으면 overfitting 됨





# Conclusion

## ❖ Conclusion

- 충분히 많은 양의 데이터를 가지고 학습을 수행하면, MLP-Mixer가 image classification task에 대해 SOTA는 아니지만 그에 견줄 만한 성능을 보일 수 있음
- Trade-off between accuracy and computational resources 를 고려한다면, 향후 convolution 및 self-attention based model 영역을 넘어 연구해볼 수 있는 분야로 발전 가능

## ❖ Opinion

- Well, not “actually” conv free
  - Yann LeCun이 언급한 바와 같이 완전히 CNN과 분리된 개념이라고 보기에는 부족한 느낌
  - 그럼에도 불구하고, ViT에 초점을 맞추어 많은 연구가 진행되고 있는 시점에 주의를 환기할 만한 좋은 논문이라 생각



**Yann LeCun** @ylecun · 5월 7일

Well, not “actually” conv free.

1st layer: “Per-patch fully-connected” == “conv layer with 16x16 kernels and 16x16 stride”

other layers: “MLP-Mixer” == “conv layer with 1x1 kernels”



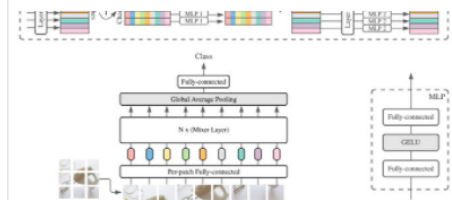
**Neil Hounsby** @neilhounsby · 5월 5일

New paper from Brain Zurich and Berlin!

We try a conv and attention free vision architecture:  
MLP-Mixer (arxiv.org/abs/2105.01601)

Simple is good, so we went as minimalist as possible (just MLPs!) to see whether modern training methods & data is sufficient...

[이 스레드 보기](#)



*Thank You*