CycleMLP

A MLP-like Architecture for Dense Prediction

Marco Benelli

University of Florence

February 14, 2022

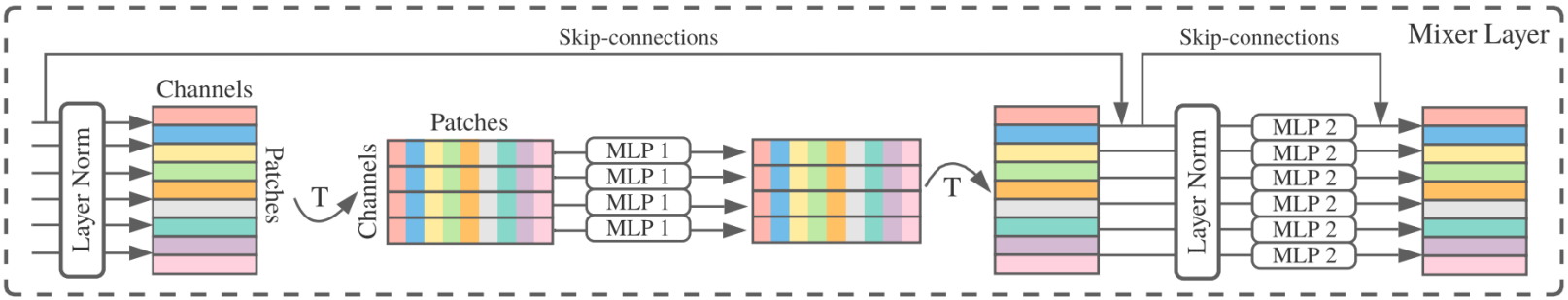
Outline

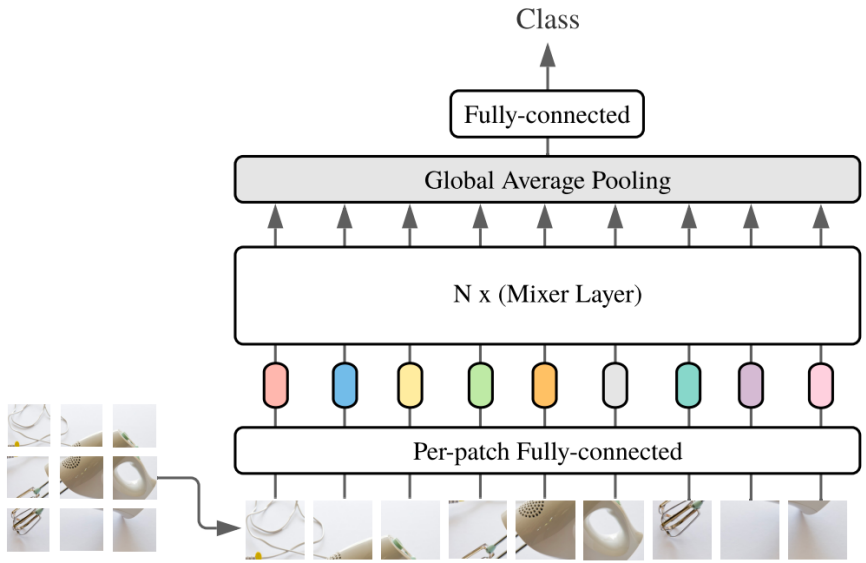
Paradigm Shifts

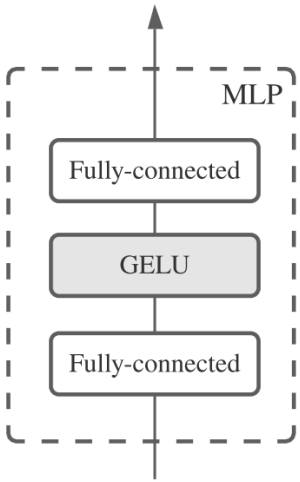
Recent paradigm shifts:

- 2012 AlexNet
- 2020 ViT
- 2021 MLP-Mixer

MLP-Mixer


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


MLP
Fully-connected
GELU
Fully-connected

Mixer Layer

$$\mathbf{U}_{*,i} = \mathbf{X}_{*,i} + \mathbf{W}_2\sigma(\mathbf{W}_1\,\text{LayerNorm}(\mathbf{X})_{*,i}), \quad \text{for } i = 1 \dots C$$

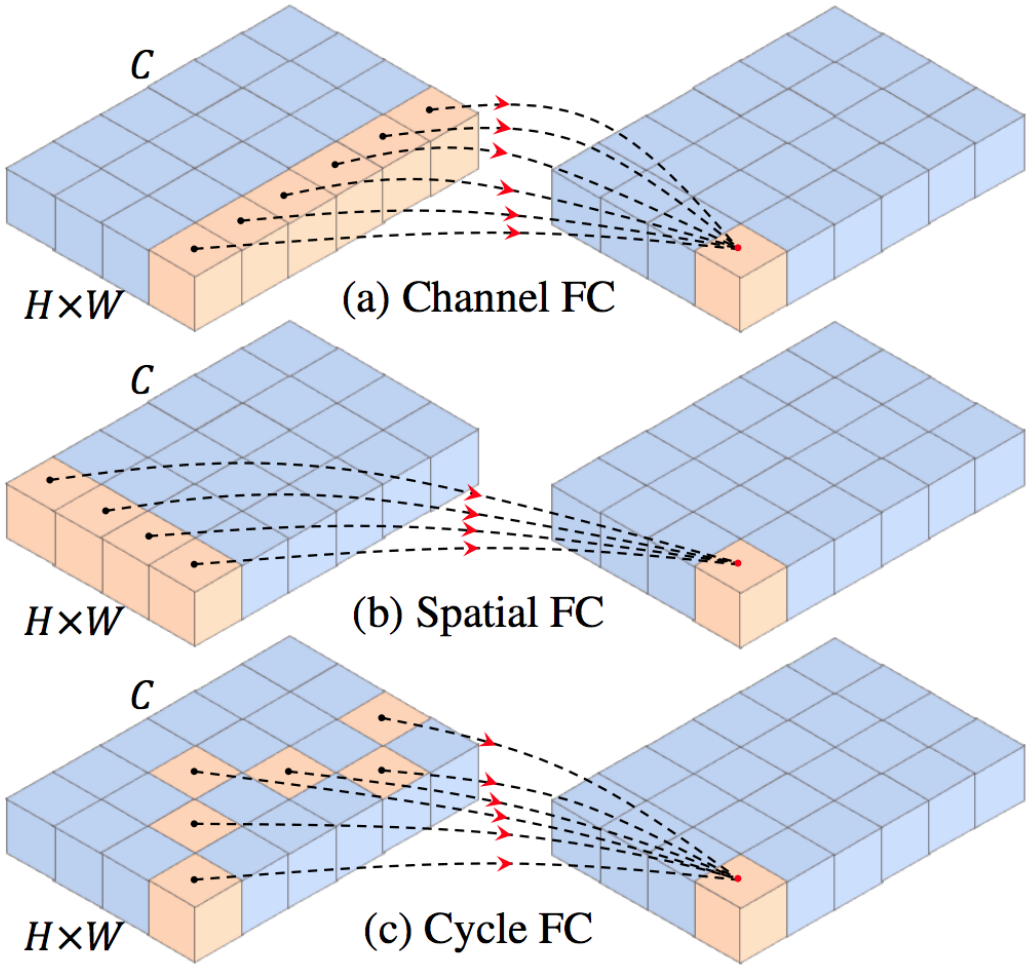
$$\mathbf{Y}_{j,*} = \mathbf{U}_{j,*} + \mathbf{W}_4\sigma(\mathbf{W}_3\,\text{LayerNorm}(\mathbf{U})_{j,*}), \quad \text{for } j = 1 \dots S$$

Challenges

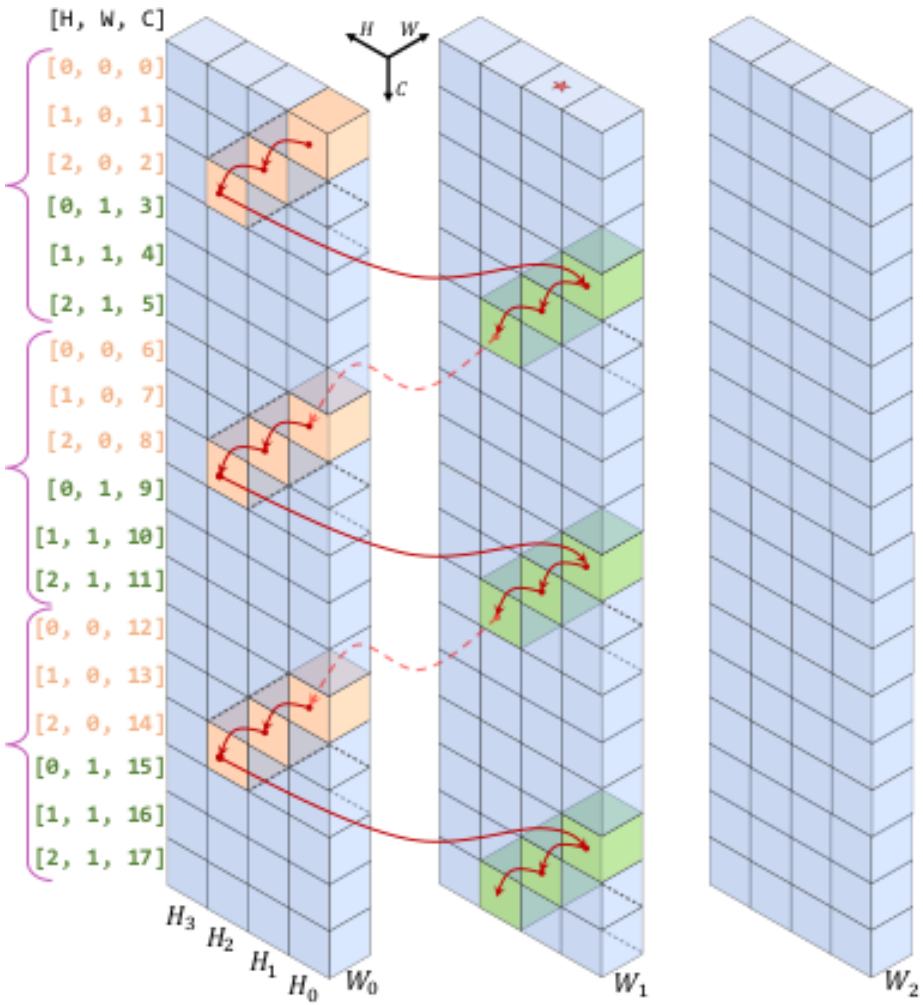
MLP-like models are facing these challenges:

- non-hierarchical architectures
- flexible input scales
- quadratic costs

Cycle Fully-Connected Layer



Stepsize Example



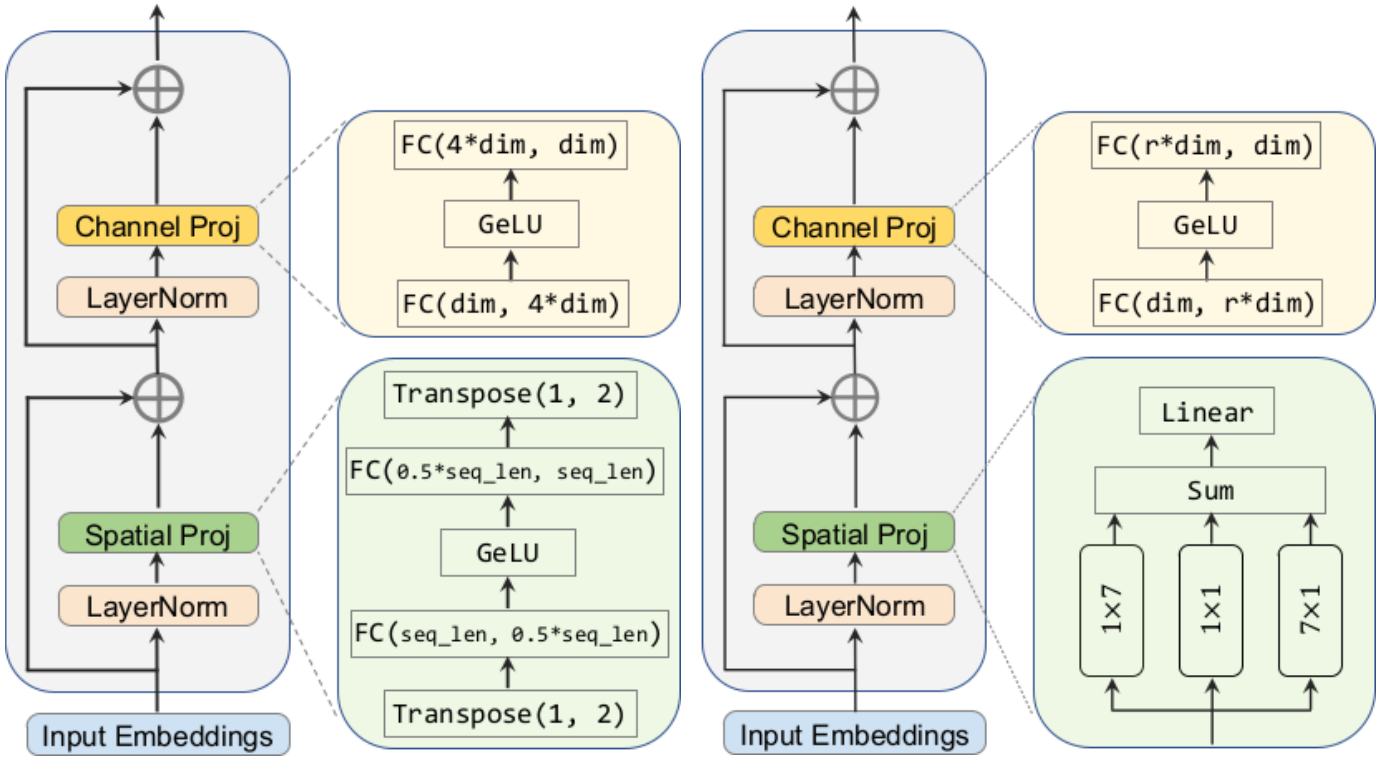
Cycle Fully-Connected Layer

$$\text{CycleFC}(\mathbf{X})_{i,j,:} = \sum_{c=0}^{C_{\text{in}}} \mathbf{X}_{i+\delta_{\text{i}}(c),j+\delta_{\text{j}}(c),c} \cdot \mathbf{W}_{c,:} + \mathbf{b}$$

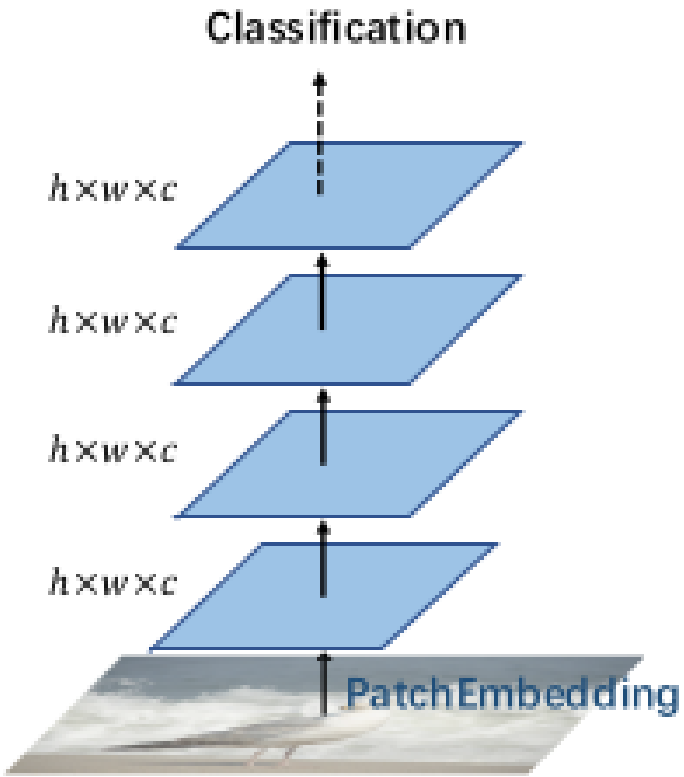
$$\delta_{\text{i}}(c) = (c \bmod S_H) - \left\lfloor \frac{S_H}{2} \right\rfloor$$

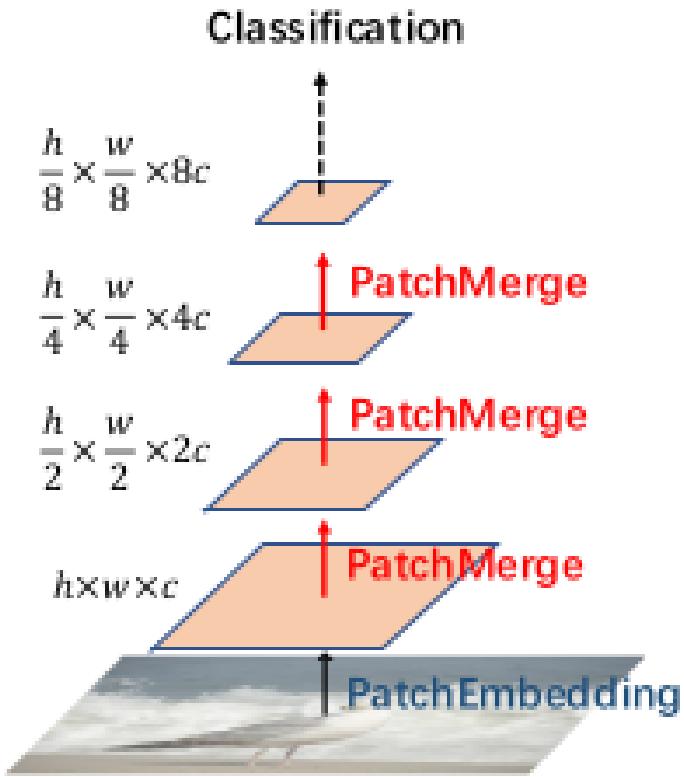
$$\delta_{\text{j}}(c) = \left(\left\lfloor \frac{c}{S_H} \right\rfloor \bmod S_W \right) - \left\lfloor \frac{S_W}{2} \right\rfloor$$

Comparison Of MLP Blocks



Hierarchy


Classification
$h \times w \times c$
$h \times w \times c$
$h \times w \times c$
$h \times w \times c$
PatchEmbedding


Classification
$\frac{h}{8} \times \frac{w}{8} \times 8c$
$\frac{h}{4} \times \frac{w}{4} \times 4c$
$\frac{h}{2} \times \frac{w}{2} \times 2c$
$h \times w \times c$
PatchMerge
PatchMerge
PatchMerge
PatchEmbedding

Instantiation

	Output Size	Layer Name	B1
Stage 1	$\frac{H}{4} \times \frac{W}{4}$	Overlapping Patch Embedding	$C_1 = 64$
		CycleMLP Block	$E_1 = 4$ $L_1 = 2$
Stage 2	$\frac{H}{8} \times \frac{W}{8}$	Overlapping Patch Embedding	$C_2 = 128$
		CycleMLP Block	$E_2 = 4$ $L_2 = 2$
Stage 3	$\frac{H}{16} \times \frac{W}{16}$	Overlapping Patch Embedding	$C_3 = 320$
		CycleMLP Block	$E_3 = 4$ $L_3 = 4$
Stage 4	$\frac{H}{32} \times \frac{W}{32}$	Overlapping Patch Embedding	$C_4 = 512$
		CycleMLP Block	$E_4 = 4$ $L_4 = 2$

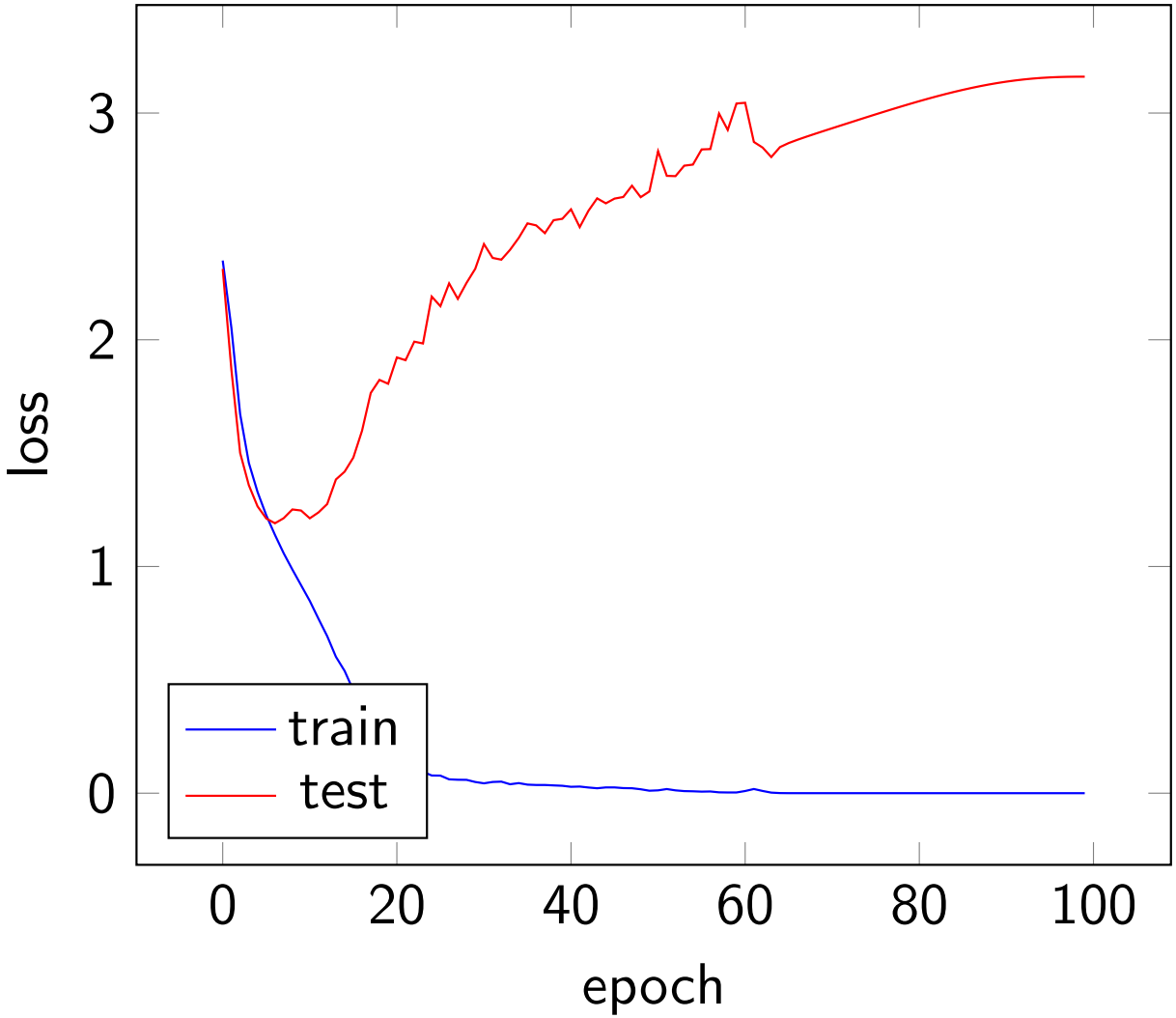
Experimental Setup

- optimizer AdamW
- $\lambda = 5 \times 10^{-2}$
- cosine annealing learning rate schedule
- $\eta_{\text{max}} = 1 \times 10^{-3}$
- $T_{\text{max}} = 100$
- batch size $= 256$

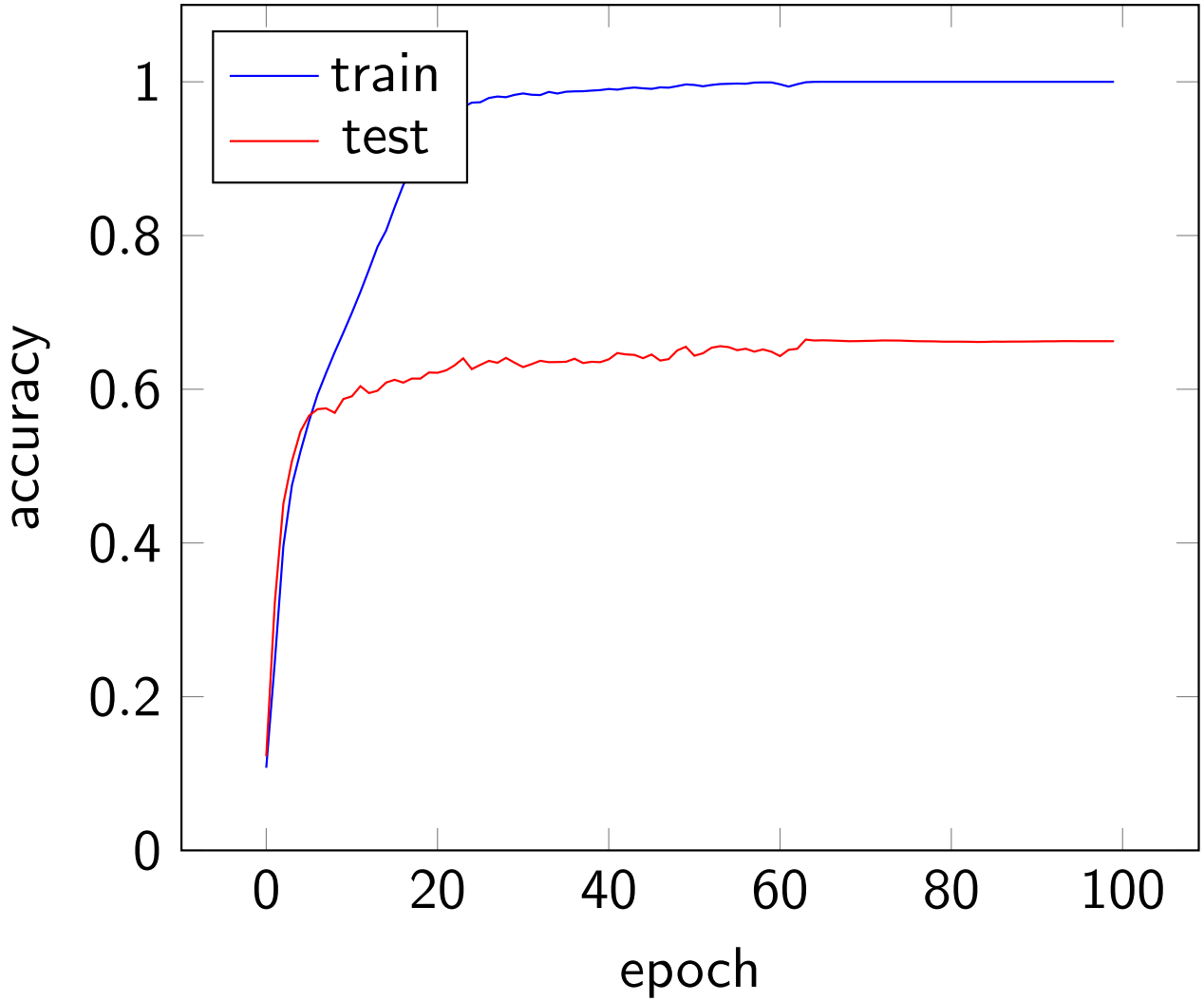
Experiments

Model	STL10	CIFAR10
ResNet	64.9%	77.1%
ViT	44.4%	53.4%
MLP-Mixer	51.4%	55.5%
CycleMLP	49.8%	66.5%

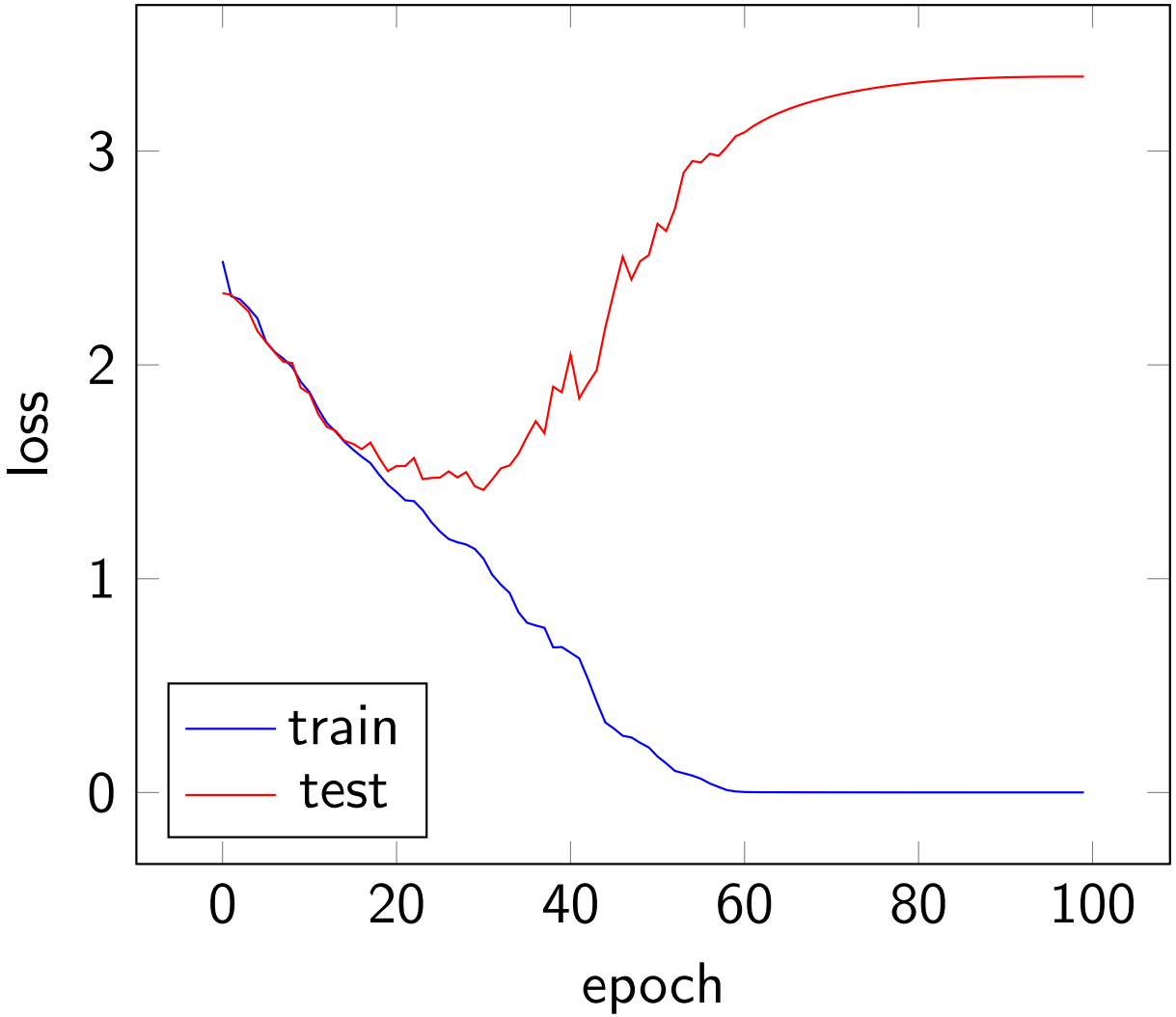
Loss Plot (CIFAR10)



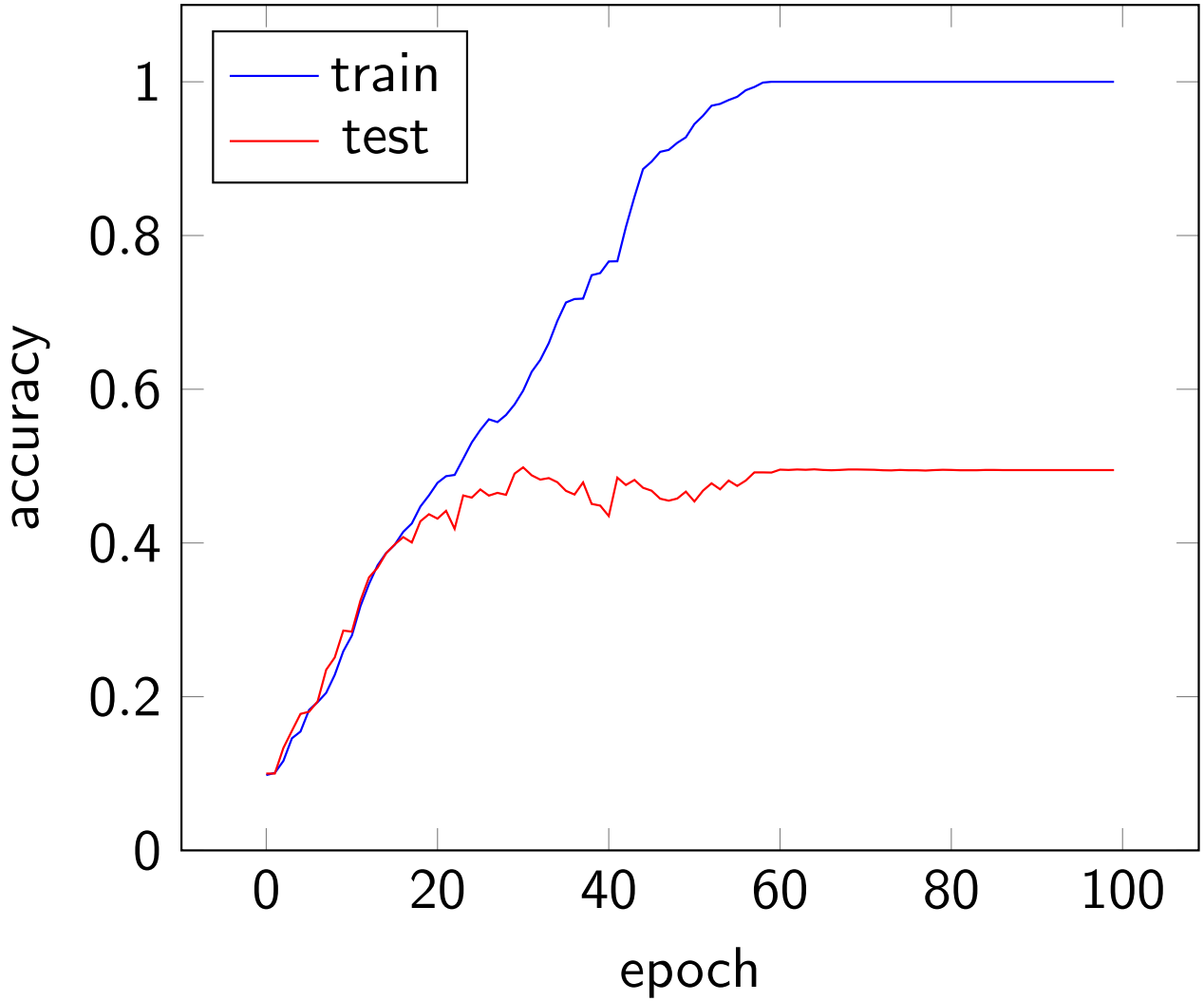
Accuracy Plot (CIFAR10)



Loss Plot (STL10)



Accuracy Plot (STL10)



ImageNet-1K Comparison

Model	Accuracy
ResNet	69.8%
ViT	77.9%
MLP-Mixer	61.4%
CycleMLP	79.1%

Summary

- CycleMLP is built upon the Cycle FC.
- Cycle FC is capable of dealing with variable input scales.
- The computational cost of Cycle FC is $O(HWC^2)$.