CycleMLP

A MLP-like Architecture for Dense Prediction

Marco Benelli

University of Florence

February 14, 2022

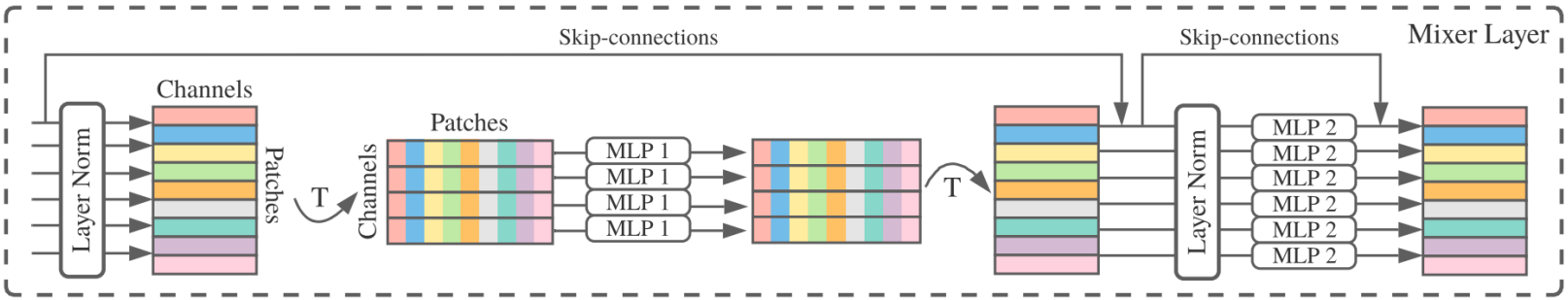
Outline

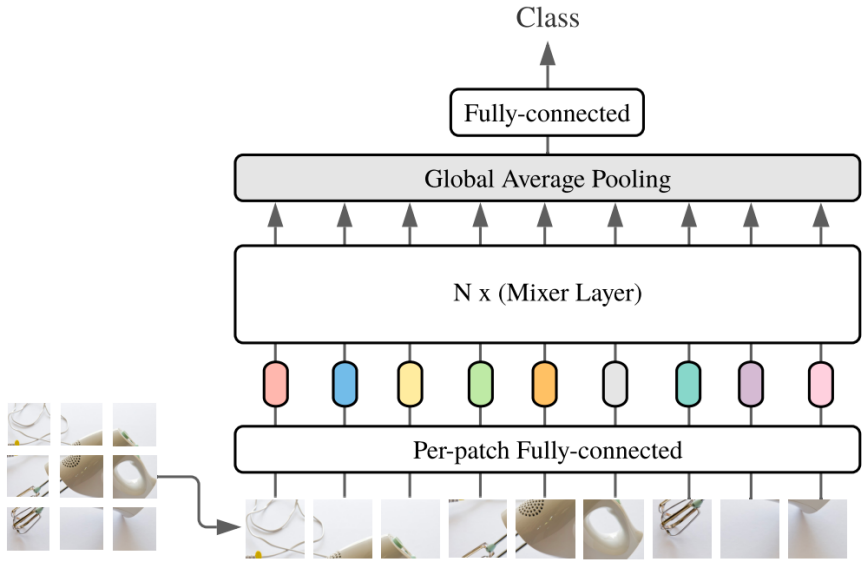
Paradigm Shifts

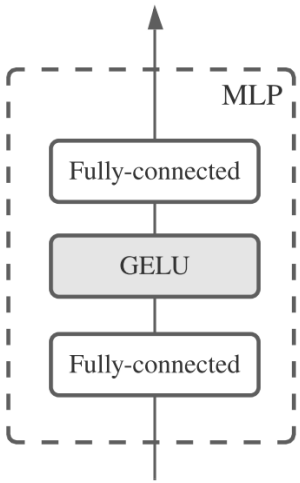
Recent paradigm shifts:

- 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

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

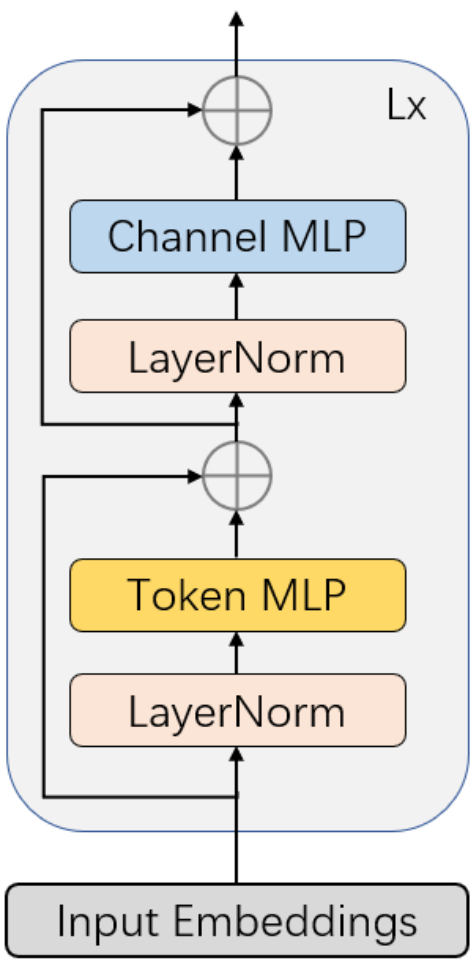
$$\mathbf{Y}_{j,*} = \mathbf{U}_{j,*} + \mathbf{W}_4\sigma(\mathbf{W}_3\,\text{LayerNorm}(\mathbf{U})_{j,*}), \quad \text{for } j = 1 \dots S$$

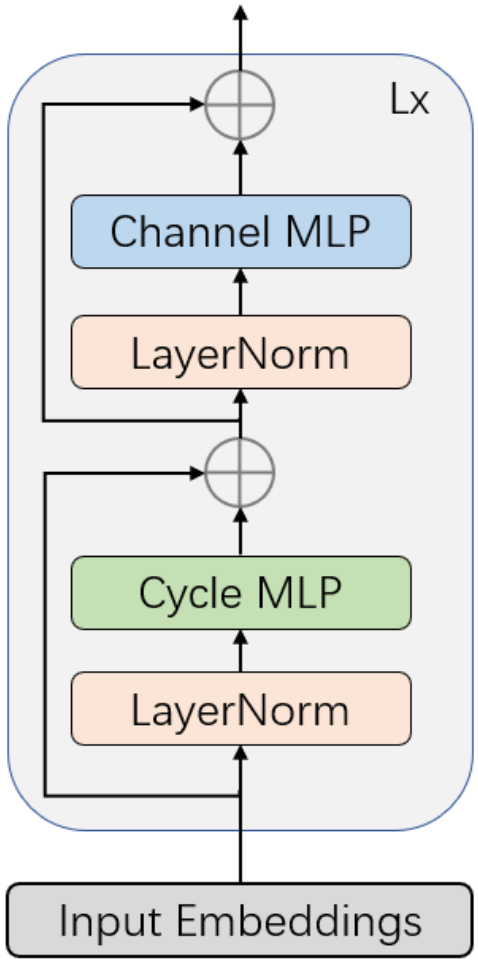
Challenges

MLP-like models are facing these challenges:

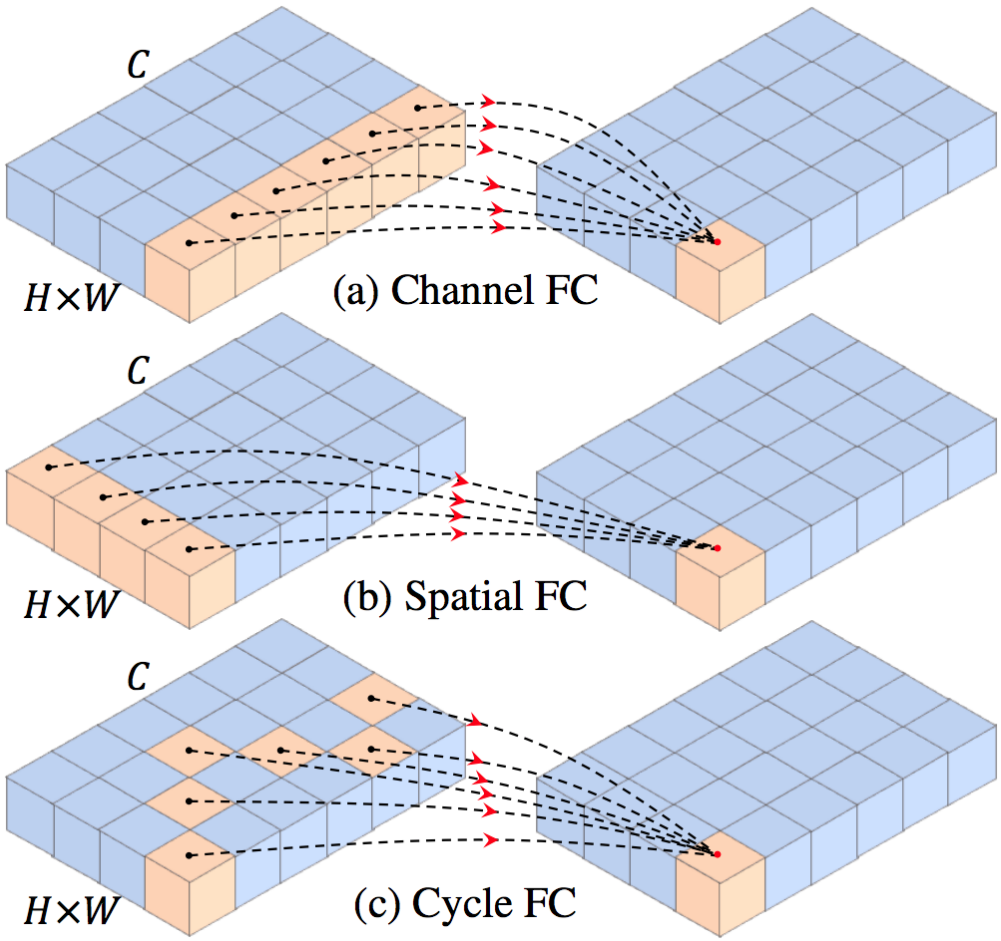
- non-hierarchical architectures
- flexible input scales
- quadratic costs

MLP-Mixer v. CycleMLP





Cycle FC



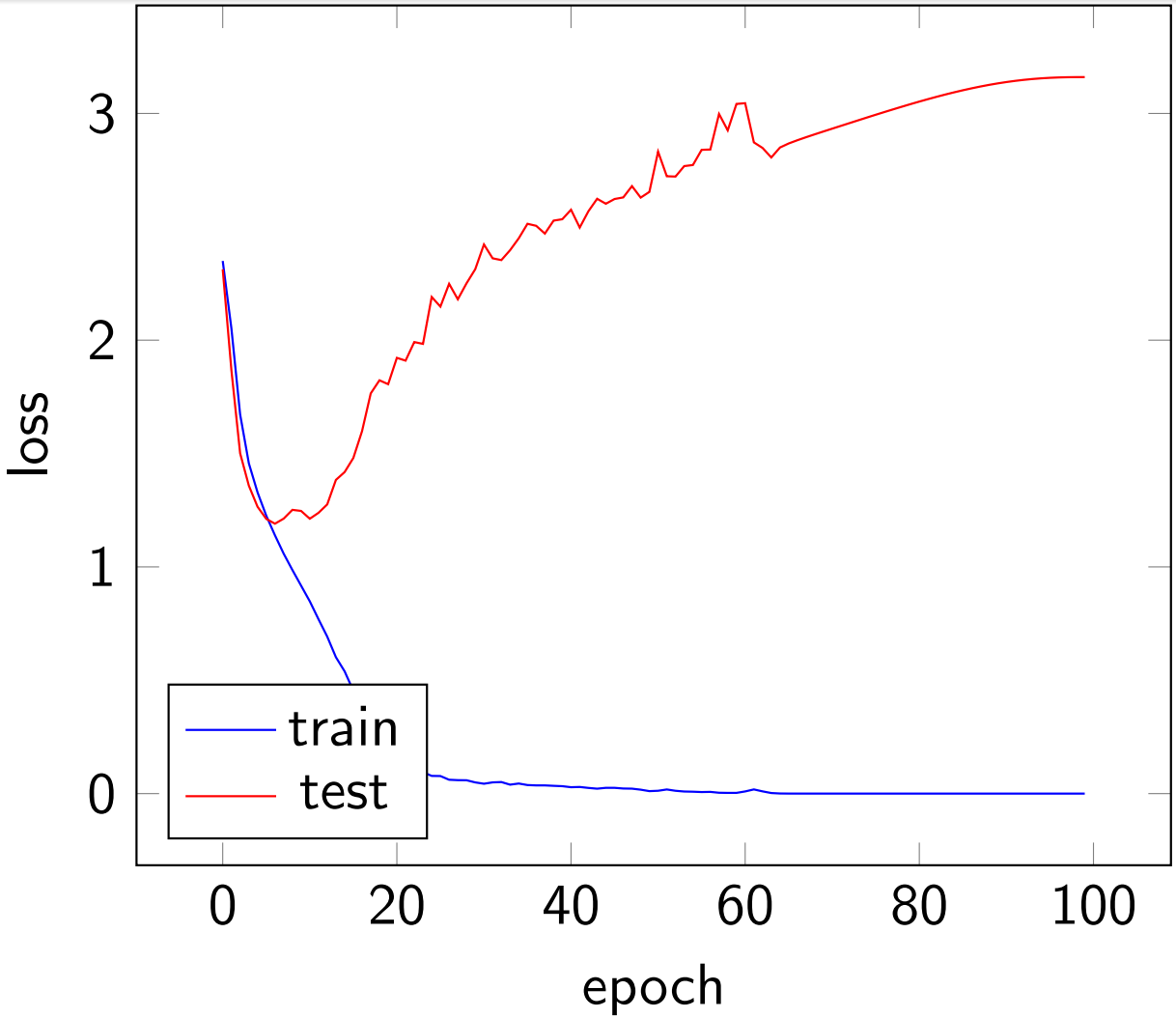
Experimental Setup

- 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 $= 256$

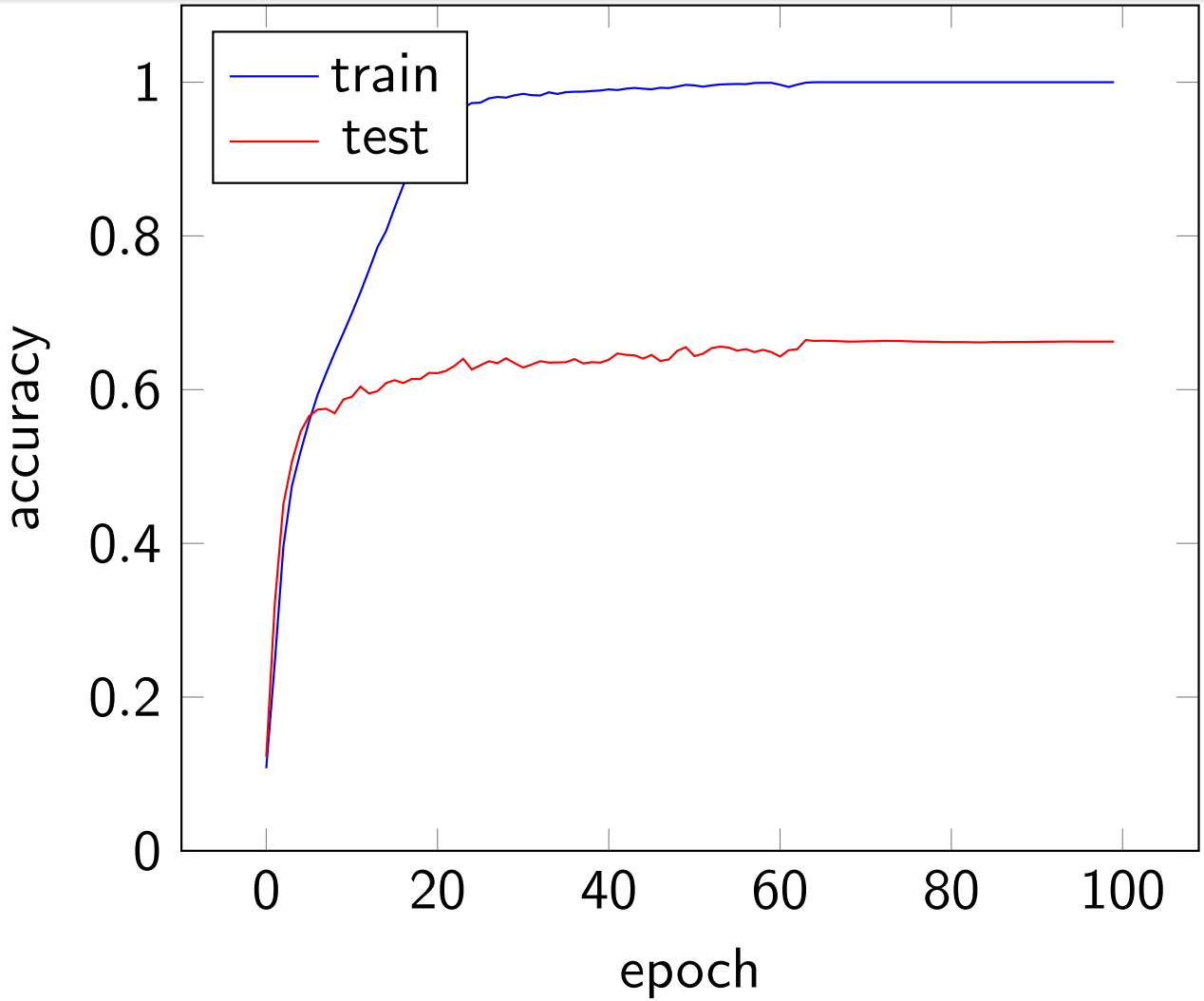
Experiments

Model	STL10	CIFAR10	ImageNet-1K
ResNet	64.9%	77.1%	
RegNet	51.5%	60.6%	
ViT	44.4%		
MLP-Mixer	51.4%		
CycleMLP	49.8%	66.5%	

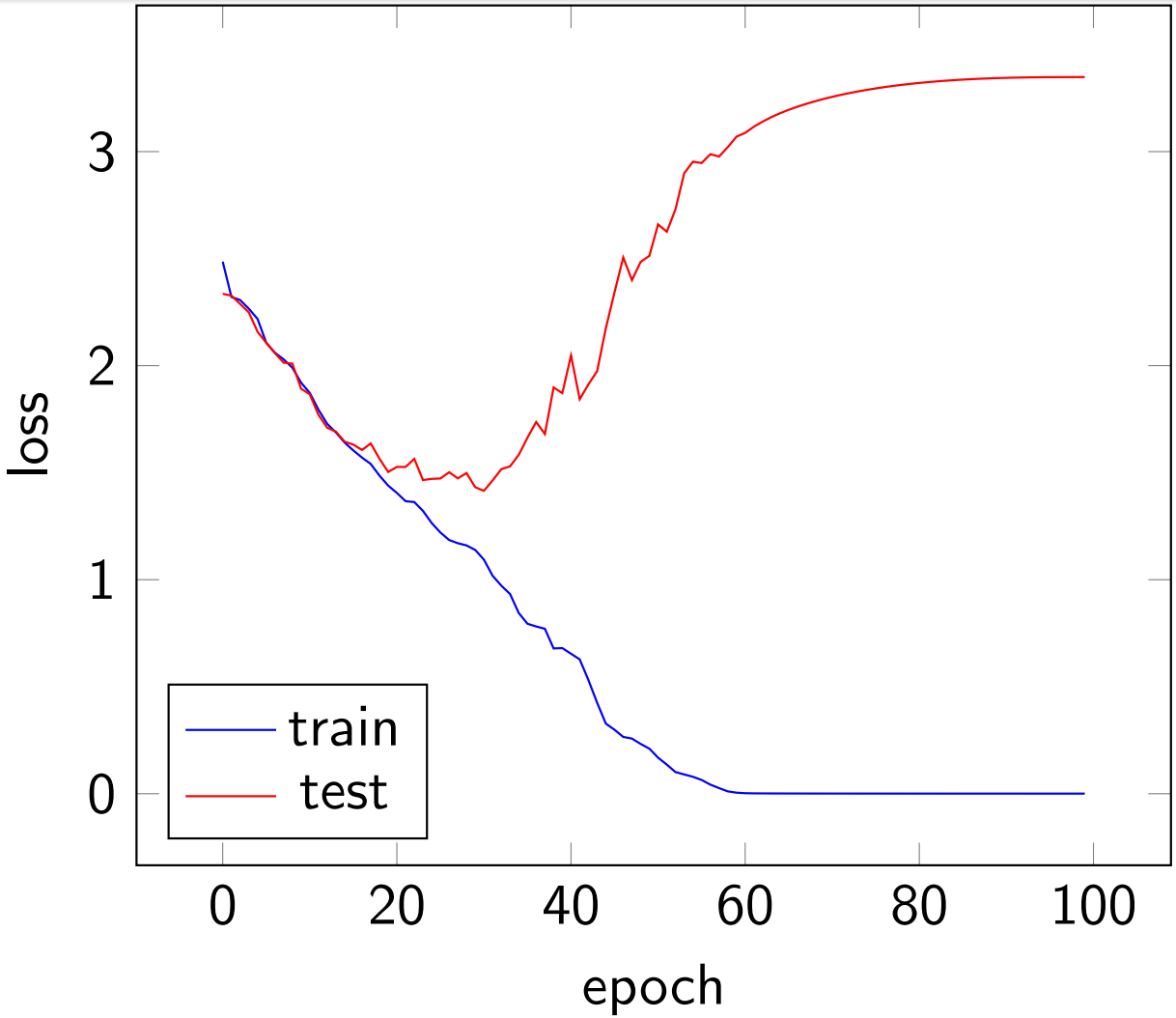
Loss Plot



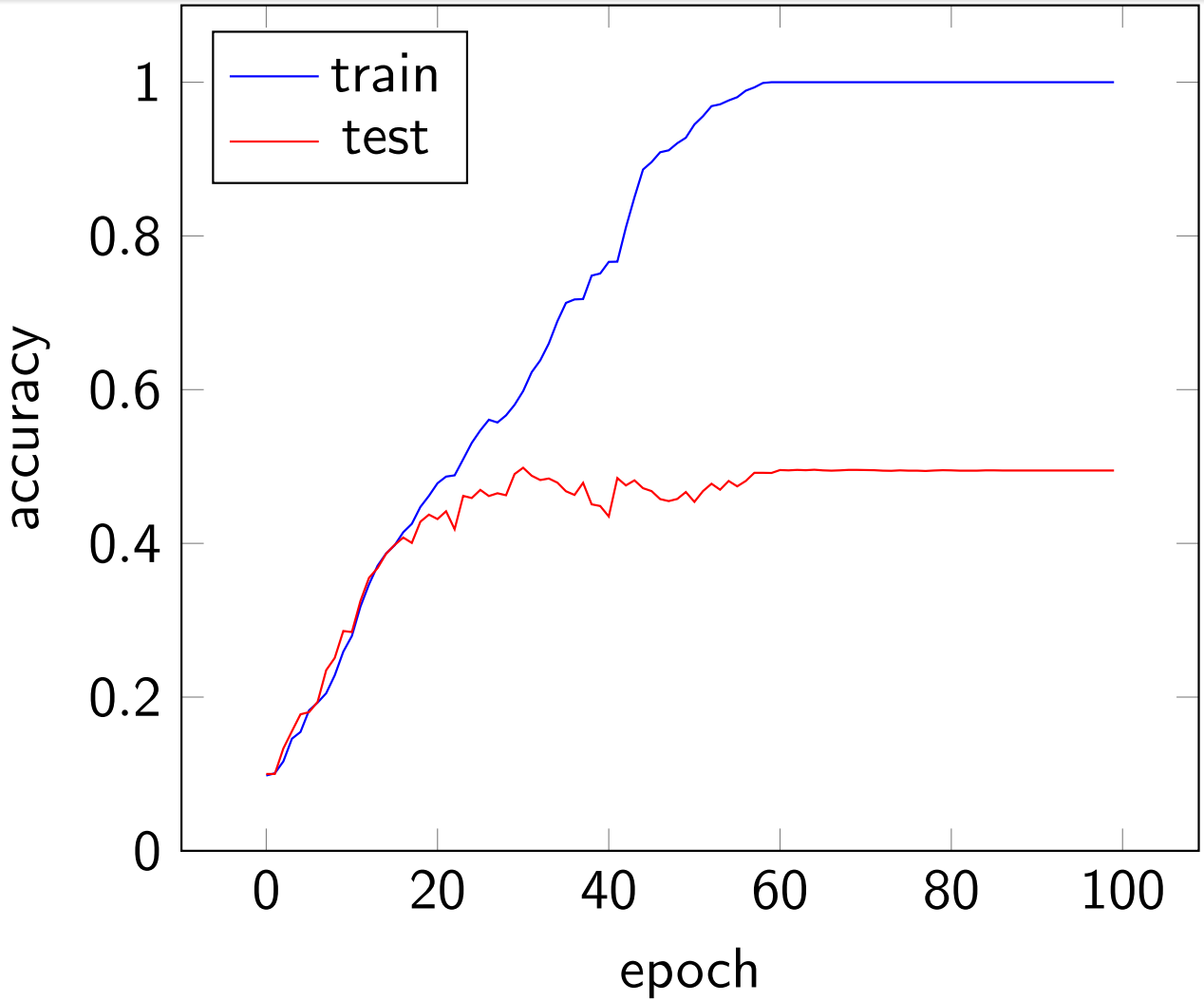
Accuracy Plot



Loss Plot



Accuracy Plot



Summary

- The first main message of your talk in one or two lines.
- The second main message of your talk in one or two lines.
- Perhaps a third message, but not more than that.