Universidade de São Paulo
Escola Politécnica - Engenharia de Computação e Sistemas Digitais

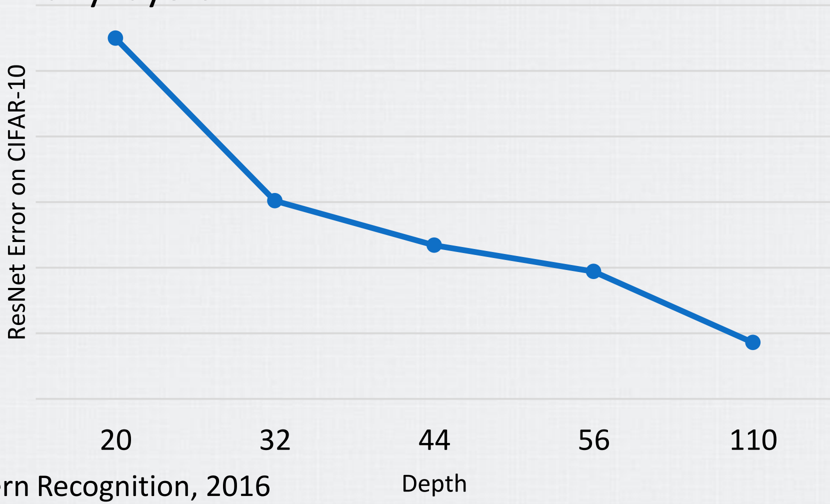
Deep Learning

Prof. Artur Jordão

Introduction

Deep Learning

- Deep learning is advancing machine learning toward human-level performance in many cognitive tasks
  - It is now the powerhouse for learning patterns from data
- Informal definition
  - A neural network architecture organized into many layers
  - For example, 1202 layers He et al. (2016)



He et al. *Deep Residual Learning for Image Recognition*. Computer Vision and Pattern Recognition, 2016

The Role of Depth and Width

Architectural Considerations

The Role of Depth and Width

- Two essential aspects of a neural network architecture are:
  - Width N
  - Depth L

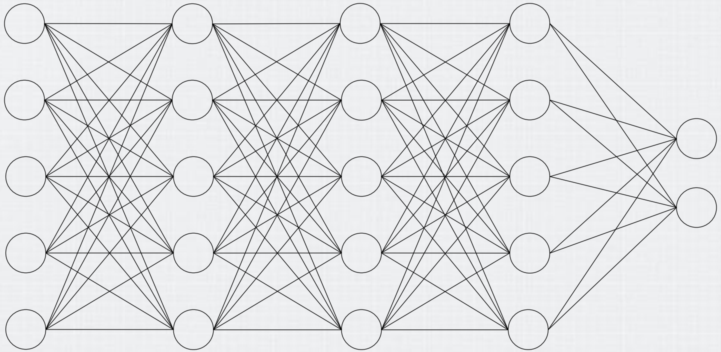
- Let $\mathcal{F}(x, \theta)$ be a neural network parametrized by θ that predicts $\hat{y}$ (i.e., $\mathcal{F}(x, \theta) = \hat{y}$)
  - We can rewrite $\mathcal{F}(\cdot,\cdot)$ in terms of a set of functions f
  - $\mathcal{F}(x, \theta) \Rightarrow f_L(f_{L-1}(\ldots, f_2\,(f_1(x, \theta_1), \theta_2), \ldots \theta_{L-1}), \theta_L)$

Width

The Role of Depth and Width

- The number of neurons in each layer $l \in L$
    - Often, we do not take into account the input and output layers because they depend on the data dimension and task (i.e., number of categories), respectively

- The width defines the **number of parameters**
    - The number of (learnable) weights



Width = $\{5, 5, 5\}$

Width

The Role of Depth and Width

- Typically, widths are in powers of 2

Architeture	Width per Layer
ResNet56 (He et al. , 2016)	16, 32, 64
ResNet50 (He et al., 2016)	64, 128, 256, 512, 1024, 2048
MobileNetV2 (Sandler et al., 2018)	16, 24, 32, 64, 96, 160, 320, 1280
Transformer (Vaswani et al., 2017)	8, 64, 128, 512, 1024, 2048, 4096

He et al. *Deep Residual Learning for Image Recognition*. Computer Vision and Patttern Recognition (CVPR), 2016

Sandler et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. Computer Vision and Patttern Recognition (CVPR), 2018

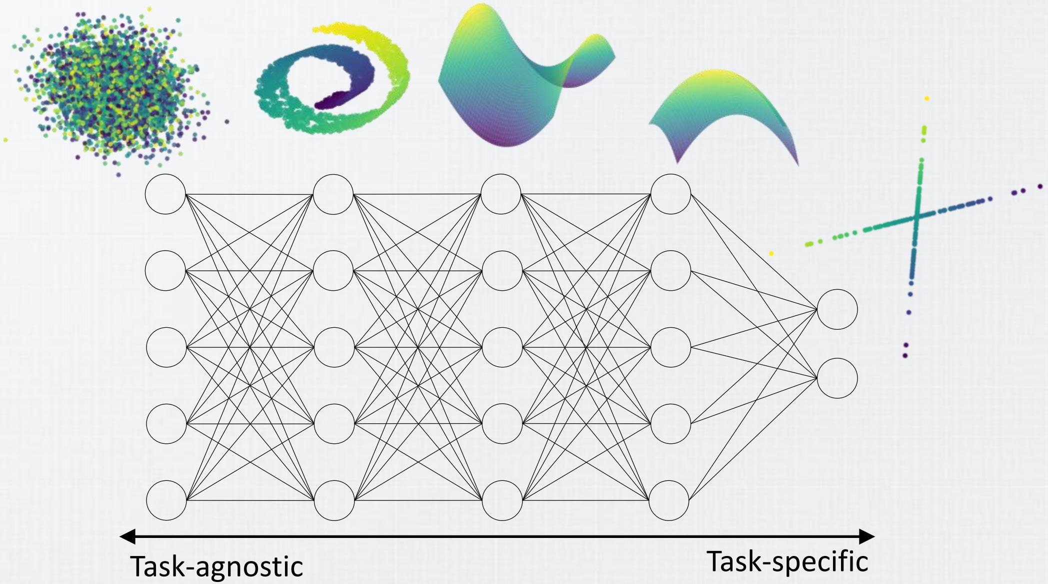
Vaswani et al. *Attention Is All You Need*. Neural Information Processing Systems (NeurIPS), 2017

Depth

The Role of Depth and Width

- The number of layers composing the network

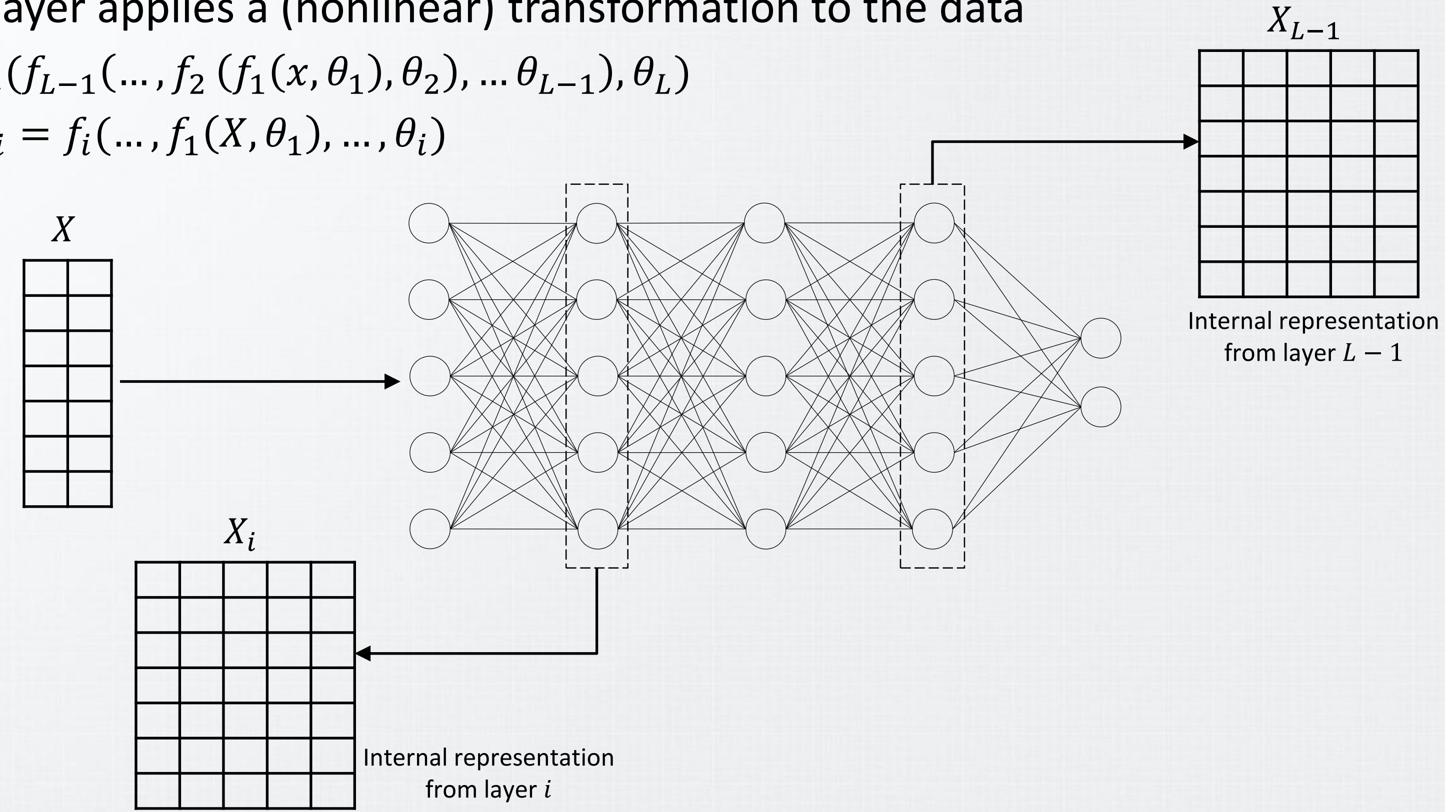
- Each layer applies a (nonlinear) transformation to the data
  - $f_L(f_{L-1}(\dots, f_2\,(f_1(x, \theta_1), \theta_2), \dots \theta_{L-1}), \theta_L)$



Internal Representations

The Role of Depth and Width

- Each layer applies a (nonlinear) transformation to the data
  - $f_L(f_{L-1}(\dots, f_2\,(f_1(x, \theta_1), \theta_2), \dots \theta_{L-1}), \theta_L)$
  - $X_i = f_i(\dots, f_1(X, \theta_1), \dots, \theta_i)$



Width vs. Depth

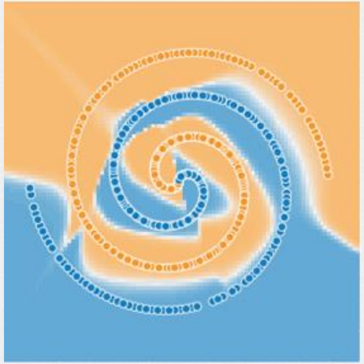
The Role of Depth and Width

- *We need to go deeper* paradigm (Tan et al., 2019; Han et al., 2020)

- Deep-learning models can extract a rich variety of features from data

Single-layer 2 neurons

Single-layer 8 neuron



Two-layer, 4 neuron each

Tan and Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. International Conference on Machine Learning (ICML), 2019
Han et al. *Model Rubik's Cube: Twisting Resolution, Depth and Width for TinyNets*. Neural Information Processing Systems (NeurIPS), 2020

Deep Learning Basics

Capacity

- Model's ability to fit a wide variety of functions
- Given sufficient capacity, a model can approximate any continuous function arbitrarily closely
- The total number of hidden units is a measure of the network's capacity
    - Therefore, we can control the capacity by adjusting the number of layers and number of neurons per layer

Capacity

- The consensus is that large (deep and wide) networks lead to better predictive ability and generalization (Tan and Le, 2019; Han et al., 2020)
  - Large architectures have a higher capacity

- Compromise between low- and high-capacity models
  - Low-capacity models may struggle to fit the training set
  - High-capacity models can overfit by memorizing properties of the training set that do not serve them well on the test set

Tan and Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. International Conference on Machine Learning (ICML), 2019

Han et al. *Model Rubik's Cube: Twisting Resolution, Depth and Width for TinyNets*. Neural Information Processing Systems (NeurIPS), 2020

Overparametrized Regime

- The number of parameters overwhelms the number of training examples

- Because deep learning models are often highly overparameterized, the empirical risk easily reaches zero
    - **Deep neural networks are capable of memorizing randomly labeled data** (Maini et al., 2023)

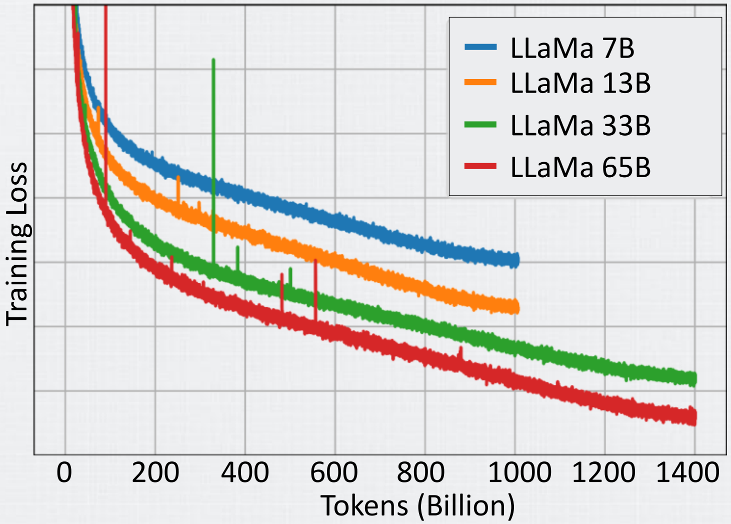
Dataset	Models	#Parameters
CIFAR-10 (50K Samples)	ResNet56	861,770
	ResNet110	1,742,762
	NASNet	3,354,858
ImageNet (1.2M Samples)	ResNet50	25,636,712
	ResNet152	60,419,944
	Visual Transformer	$\sim 632 \times 10^6$

Maini et al. *Can Neural Network Memorization Be Localized?*. In International Conference on Machine Learning (ICML), 2023

Data Hungry Regime

Deep Learning Basics

- Deep and Wide networks are data hungry
  - Such networks need a lot of data to learn effectively
  - For example, Vision Transformer (ViT) attains excellent results when pre-trained on JFT-300M Dataset (300 million images)



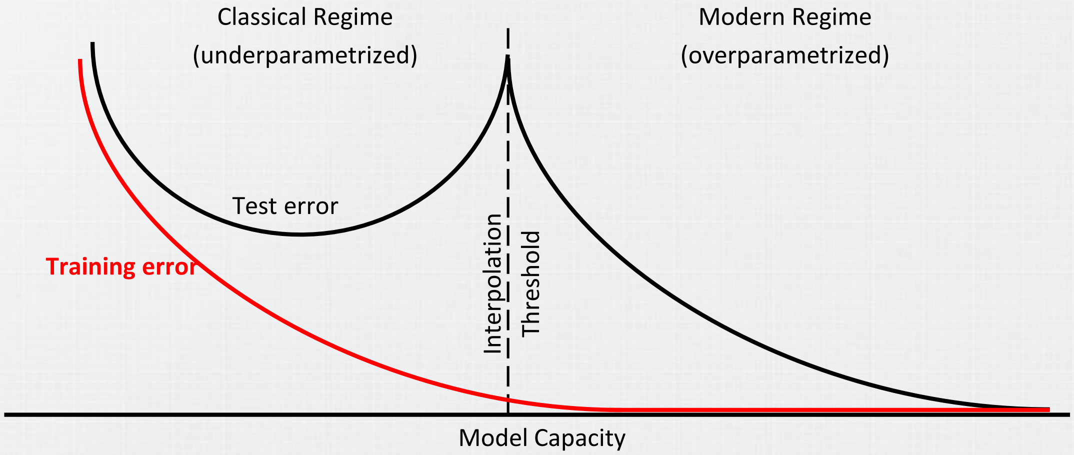
Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. International Conference on Learning Representations (ICLR), 2021

Jayalath et al. *LLaMA: Open and Efficient Foundation Language Models*. ArXiv, 2023

Deep Double Descent

Deep Learning Basics

- Double descent (Nakkiran et al., 2020) is a hallmark phenomenon of deep learning
  - It shows the rift between classical learning theory and the generalization capabilities of deep learning systems
  - Double descent is not universal (Jayalath et al. 2023)



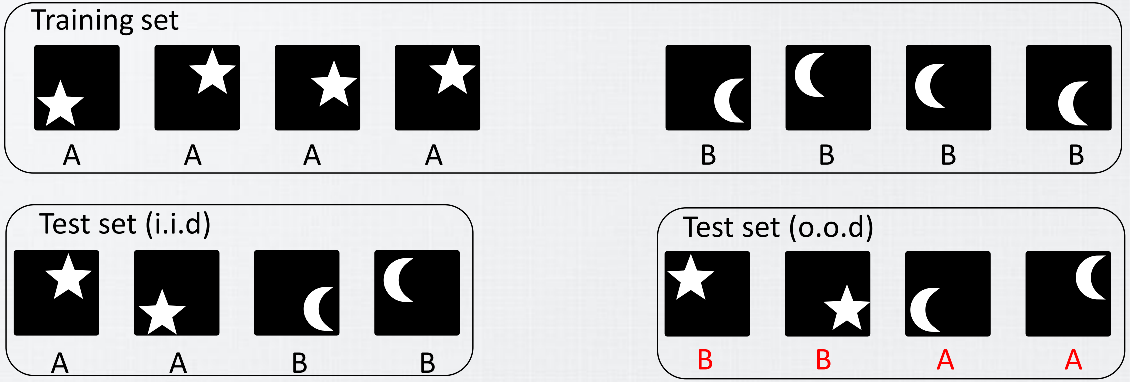
Nakkiran et al. *Deep Double Descent: Where Bigger Models and More Data Hurt*. International Conference on Learning Representations (ICLR), 2020

Jayalath et al. *No Double Descent In Self-supervised Learning*. International Conference on Learning Representations (ICLR) Tiny Papers Track, 2023

Shortcut Learning

Deep Learning Basics

- Shortcuts are features that perform well on standard benchmarks but fail to transfer to more challenging test conditions
  - They arise from dataset shortcut opportunities and discriminative feature learning that fail to generalize as intended
  - Good performance on both training and **I**ndependent and **I**dentically **D**istributed (i.i.d) test sets, but poor generalization to **O**ut-**O**f-**D**istribution (o.o.d.) inputs



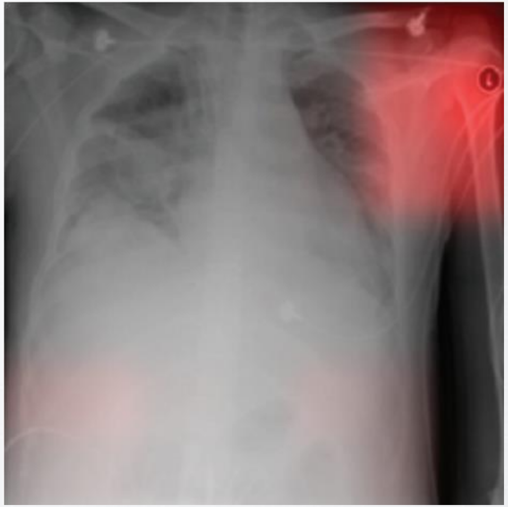
Geirhos et al. *Shortcut Learning in Deep Neural Networks*. Nature Machine Intelligence, 2020

Hermann et al. *On the Foundations of Shortcut Learning*. International Conference on Learning Representations (ICLR), 2024

Shortcut Learning

Deep Learning Basics

- Shortcut “cheat” features



Article: Super Bowl 50

Paragraph: “Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV.” Question: “What is the name of the quarterback who was 38 in Super Bowl XXXIII?”

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Task	Recognise pneumonia	Answer question
Problem	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Geirhos et al. *Shortcut Learning in Deep Neural Networks*. Nature Machine Intelligence, 2020

Spurious Correlations and Features

- Patterns that predict the target in the training data but are irrelevant to the true labeling function (Kirichenko et al. 2023)

- Core features (non-spurious)
  - Features that are truly discriminative to the task

- Deep models can largely rely on simple spurious features to make predictions
  - For example, backgrounds

- Spurious correlations can negatively impact predictive ability
  - On samples where the spurious correlations break

Kirichenko et al. *Last Layer Re-training is Sufficient for Robustness to Spurious Correlations*. International Conference on Learning Representations (ICLR), 2023

Spurious Correlations and Features

- Target: Bird type

- Spurious feature: Background type


Landbird on Land
(73%)


Landbird on Water
(4%)


Waterbird on Water
(22%)


Waterbird on Land
(1%)

Kirichenko et al. *Last Layer Re-training is Sufficient for Robustness to Spurious Correlations*. International Conference on Learning Representations (ICLR), 2023

Shortcuts vs. Spurious

- Shortcuts refer to features easily latched onto by a model

- Spurious refer to features that arise unintentionally in a **poorly constructed dataset**

Kirichenko et al. *Last Layer Re-training is Sufficient for Robustness to Spurious Correlations*. International Conference on Learning Representations (ICLR), 2023

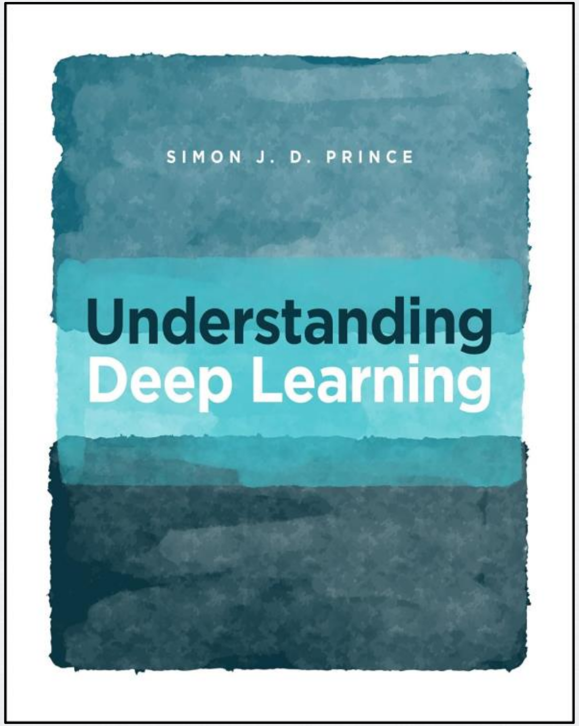
Historical Trends in Deep Learning

- Deep learning has become more useful as a function of available training
- Deep learning models have grown in size over time
  - Among the factors are the improvements in computer infrastructure (both hardware and software)
- Deep learning has solved increasingly complicated applications with increasing precision over time
  - Protein structure prediction
  - Image captioning and generating
  - Estimating residential solar potential
  - Discovering faster matrix multiplication

Bibliography

Bibliography

- Understanding Deep Learning
    - Chapter 4
        - 4.5 Shallow vs. deep neural networks
        - 4.6 Summary



Bibliography

- Nakkiran et al. *Deep Double Descent: Where Bigger Models and More Data Hurt*. International Conference on Learning Representations (ICLR), 2020

- d'Ascoli et al. *Double Trouble in Double Descent: Bias and Variance(s) in the Lazy Regime.* International Conference on Machine Learning (ICML), 2020

- Jayalath et al. *No Double Descent in Self-supervised Learning*. International Conference on Learning Representations (ICLR), 2023

International Conference On Learning Representations

International Conference On Machine Learning

Bibliography

- Tan and Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. International Conference on Machine Learning (ICML), 2019

- Han et al. *Model Rubik's Cube: Twisting Resolution, Depth and Width for TinyNets*. Neural Information Processing Systems (NeurIPS), 2020

- Liu et al. *Efficient Training of Visual Transformers with Small Datasets*. Neural Information Processing Systems (NeurIPS), 2021

NEURAL INFORMATION PROCESSING SYSTEMS

ICML International Conference On Machine Learning