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

Data Augmentation, Dataset Distillation and Adversarial Attacks

Prof. Artur Jordão

Introduction

Data Augmentation

- The consensus is that high-capacity networks promote better predictive ability
  - However, these models are **data-hungry**, which means that they need a larger number of training samples to avoid poor generalization

- We can deal with data-hungry regimes by generating additional training data
  - Namely data augmentation

- Data augmentation techniques expand the training set by creating **replicas** of the training samples

Forms of Data Augmentation

Data Augmentation

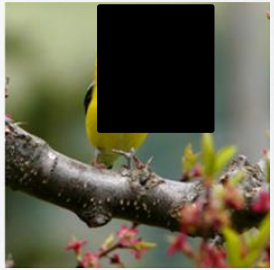
Clean Image

$y = \text{bird}$

Random Flip
(He et al., 2016)

$y = \text{bird}$

Cutout
(DeVries et al., 2017)



$y = \text{bird}$

MixUp
(Zhang et al., 2018)

$\boldsymbol{y = \alpha}$ ∗**bird+**
$\boldsymbol{(1-\alpha)}$ ∗**turtle**

CutMix
(Yun et al. , 2019)

$\boldsymbol{y = \alpha}$ ∗**bird+**
$\boldsymbol{(1-\alpha)}$ ∗**turtle**

PixMix
(Hendrycks et al. , 2022)

$y = \text{bird}$

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

DeVries et al. *Improved regularization of convolutional neural networks with cutout*. ArXiv, 2017

Zhang et al. *Mixup: Beyond Empirical Risk Minimization*. International Conference on Learning Representations (ICLR), 2018

Yun et al. *CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features*. International Conference on Computer Vision (ICCV), 2019

Hendrycks et al. *PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures*. Conference on Computer Vision and Pattern Recognition (CVPR), 2022

Stylized Augmentation

Data Augmentation

- Convolutional models prioritize texture identification over global shape (as opposed to humans)
  - When trained on ImageNet, these models exhibit a significant **bias towards texture**

Texture Image

81.4% Indian elephant
10.3% Indri
8.2% Black swan

Content Image

71.1% Tabby cat
17.3% Grey fox
3.3% Siamese cat

Texture-shape Cue Conflict

63.9% Indian elephant
26.4% Indri
9.6% Black swan

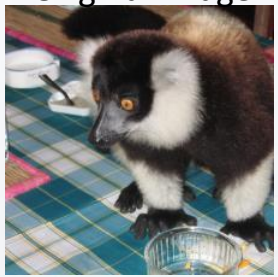
Geirhos et al. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*. International Conference on Learning Representations (ICLR), 2019 (Oral)

Stylized Augmentation

Data Augmentation

- Stylized augmentation removes texture information through style transfer

Original Image



Augmented Images

Geirhos et al. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*. International Conference on Learning Representations (ICLR), 2019 (Oral)

Data Augmentation Beyond Images

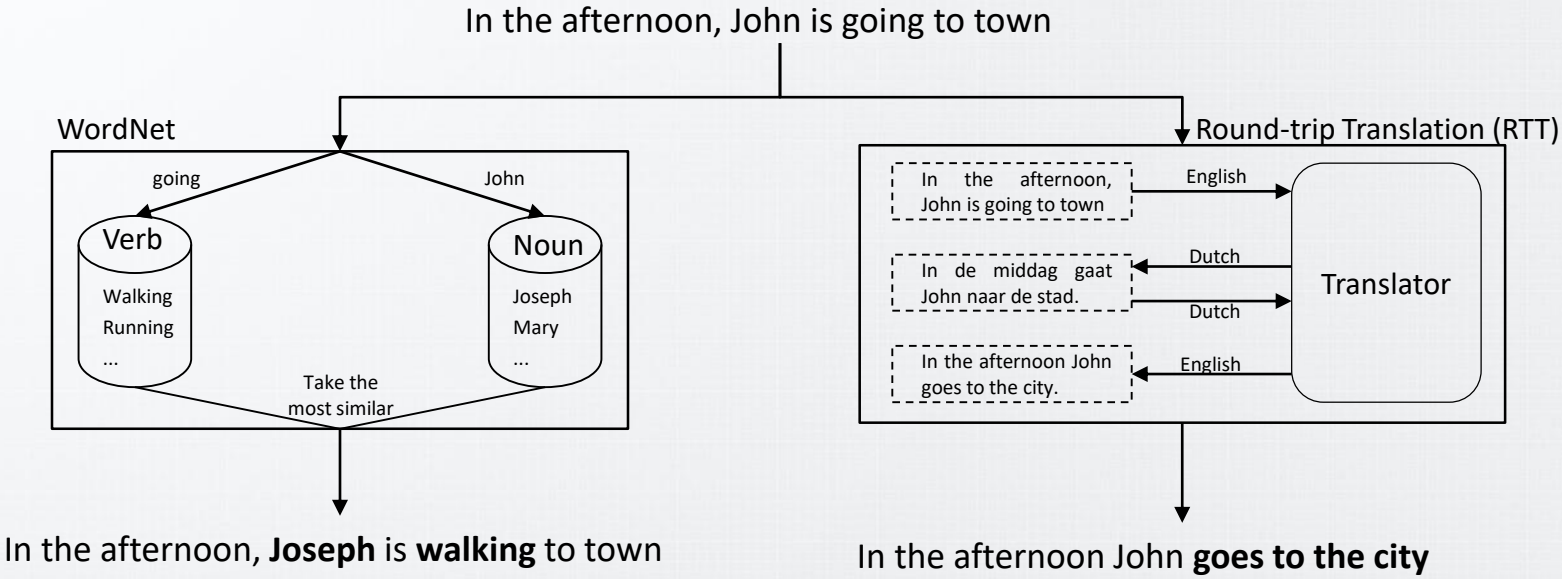
- Augmentation for time series (Abbaspourazad et al., 2024)
  - Crop
  - Gaussian noise
  - Time warp
  - Magnitude warp
  - Channel swap

Abbaspourazad et al. *Large-scale Training of Foundation Models for Wearable Biosignals*. International Conference on Learning Representations (ICLR), 2024

Data Augmentation Beyond Images

Data Augmentation

- Augmentation for natural language processing

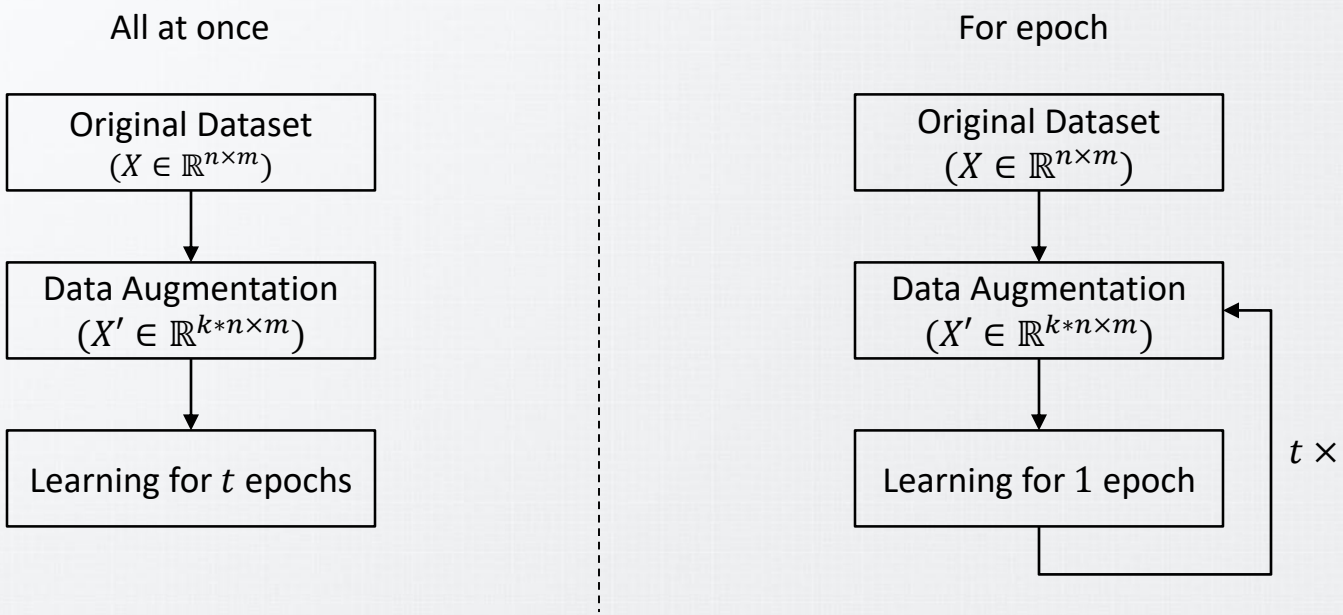


Marivate et al. *Improving Short Text Classification through Global Augmentation Methods*. International Cross-Domain Conference for Machine Learning and Knowledge Extraction, 2020

Amount vs. Diversity

Data Augmentation

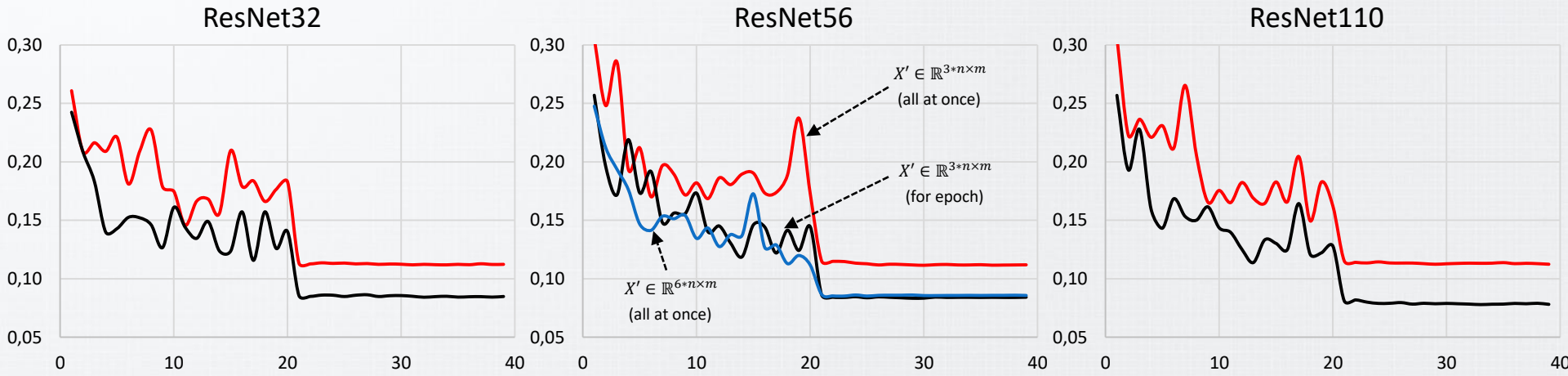
- Which strategy is expected to yield better results?



Amount vs. Diversity

Data Augmentation

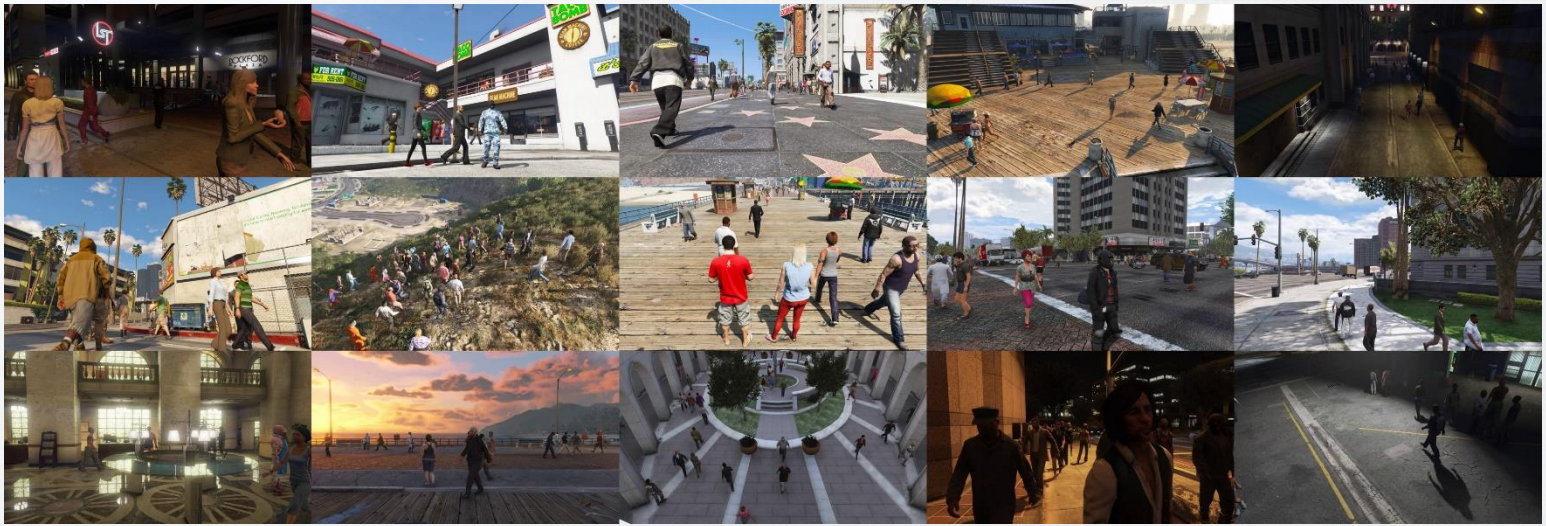
- Which strategy is expected to yield better results?



Synthetic Data

Data Augmentation

- The community has already recognized the potential of synthetic data for both benchmarking and augmenting training data
  - Due to high diversity, training solely on synthetic data leads to state-of-the-art performance (Fabbri et al., 2021)



Fabbri et al. *MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking?*. International Conference on Computer Vision (ICCV), 2021

Dataset Distillation and Adversarial Attacks

Dataset Distillation/Condensation

Dataset Distillation and Adversarial Attacks

- Assume a target dataset $T = \{(x_i, y_i)\}_{i=1}^{n}$

- Assume $\mathcal{L}_T(\mathcal{F}, \mathcal{Z})$ be the loss of a model $\mathcal{F}(\cdot, \theta)$ trained on a dataset $\mathcal{Z}$ and evaluate on a dataset T
  - $\mathcal{F}$ is parameterized by θ

Dataset Distillation/Condensation

Dataset Distillation and Adversarial Attacks

- Given some metric (PerM, ParM, DisM), the objective of dataset distillation is to distillate the knowledge of T into a small synthetic dataset $S = \{(s_j, y_j)\}_{j=1}^{m}$, where $m \ll n$ (Geng et al. 2023; Lei et al., 2024)

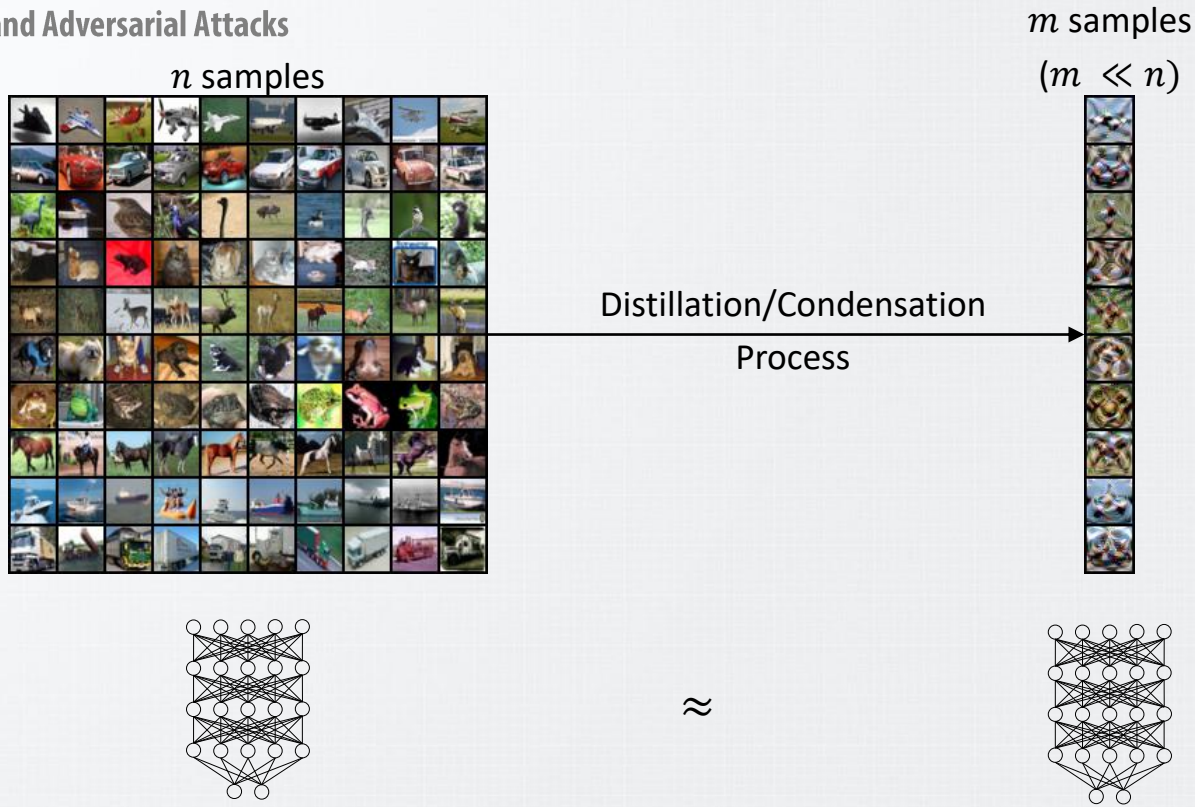
Metric	Formulation	Comment
Performance Matching (PerM)	$\mathcal{L}_T(\mathcal{F}, T) \approx \mathcal{L}_T(\mathcal{F}, S)$	It makes the model approximate some performance metric (i.e., loss)
Parameter Matching (ParM)	$\nabla_\theta \mathcal{L}(T, \theta) \approx \nabla_\theta \mathcal{L}(S, \theta)$	It aims to make the model approximate the original model in the parameter space
Distribution Matching (DisM)	$\left\| \frac{1}{S}\sum_{i=1}^{\|S\|} f_\theta(s_i) - \frac{1}{T}\sum_{i=1}^{\|T\|} f_\theta(x_i) \right\|^2$	The distribution of the synthetic samples is similar to that of real samples in the feature space

Geng et al. *A Survey on Dataset Distillation: Approaches, Applications and Future Directions*. International Joint Conference on Artificial Intelligence (IJCAI), 2023

Lei et al. *A Comprehensive Survey of Dataset Distillation*. Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2024

Dataset Distillation/Condensation

Dataset Distillation and Adversarial Attacks



He et al. *Multisize Dataset Condensation*. International Conference on Learning Representations (ICLR), 2024

Lei et al. *A Comprehensive Survey of Dataset Distillation*. Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2024

Dataset Distillation/Condensation

Dataset Distillation and Adversarial Attacks

- Overview

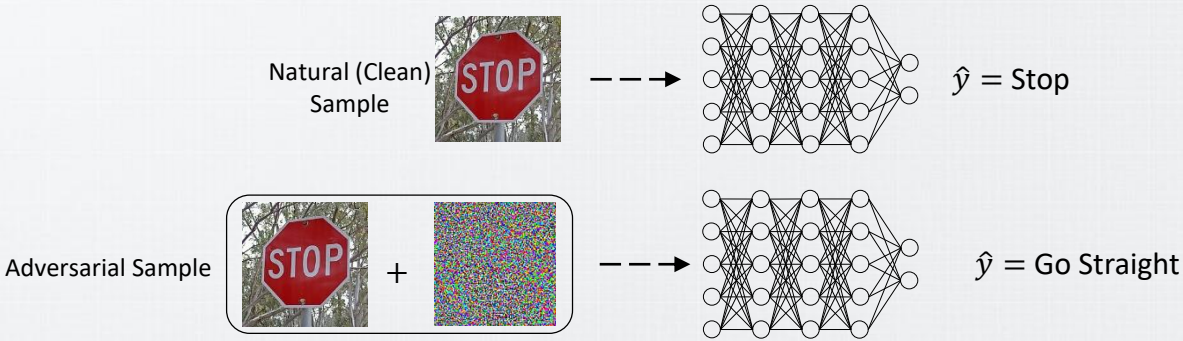
Dataset Distillation Framework
Input: Original dataset T
Output: Synthetic dataset S
Initialize S ▷ Random, real, or core-set
while not converge **do**
Get a network θ ▷ Random or from some cache (i.e., epoch)
Update θ ▷ Via S or T, for some steps
Update S via $\mathcal{L}(S, T)$ ▷ PerM, ParM, DisM (or their variants)
end

Adversarial Attacks

Dataset Distillation and Adversarial Attacks

- Adversarial attacks involve making small modifications to an image (or sample) to force a network to wrong its prediction
  - Such modifications are often imperceptible to a human

- The success of adversarial images is due to the excessive overparameterization and capacity of deep networks (Tran et al. 2018; Kaya et al., 2019)



Tran et al. *Spectral signatures in backdoor attacks*. Neural Information Processing Systems (NeurIPS), 2018

Kaya et al. *Shallow-deep networks: Understanding and mitigating network overthinking*. International Conference on Machine Learning (ICML), 2019

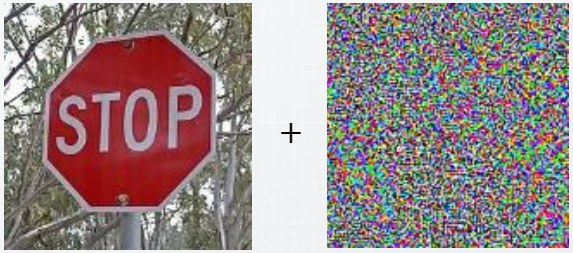
Adversarial Attacks

Dataset Distillation and Adversarial Attacks

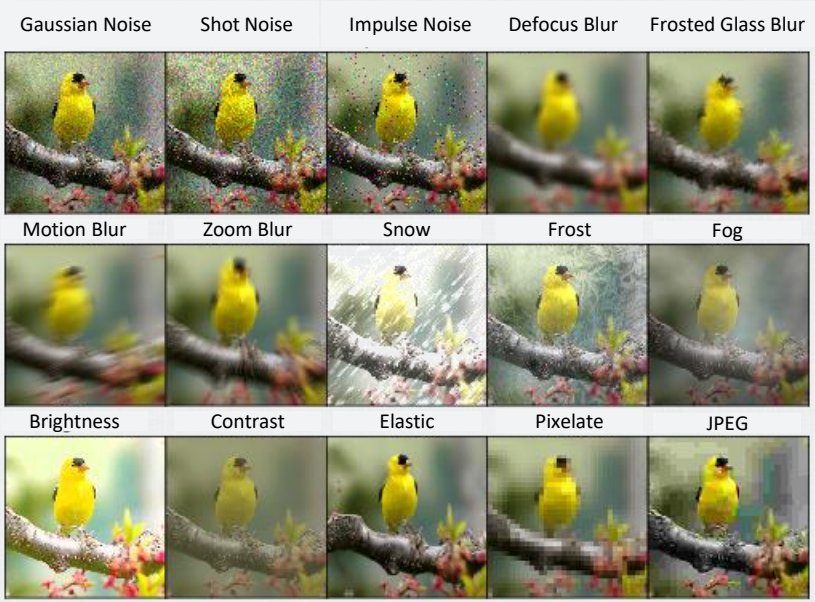
- **F**ast **G**radient **S**ign **M**ethod (FGSM)
  - $x' = x + \epsilon * sign(\nabla_\theta \mathcal{L}(\theta, x, y))$
  - ϵ controls the severity of the attack

- Corruptions (Hendrycks et al., 2019)

FGSM



Common-Corruptions



Hendrycks et al. Benchmarking neural network robustness to common corruptions and perturbations, International Conference on Learning Representations (ICLR), 2019

Backdoor Attacks

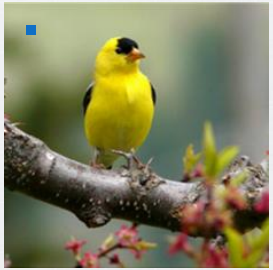
Dataset Distillation and Adversarial Attacks

- Backdoor attacks involve making models perform normally on clean data but give a **specific prediction** on trigger samples

- Formulation (θ omitted for simplicity)
  - $\mathcal{L}_{total} = \sum_i^{|X|} \mathcal{L}_{clean}(\mathcal{F}(x_i), y_i) + \sum_j^{|\hat{X}|} \mathcal{L}_{adv}(\mathcal{F}(\hat{x}_j), \boldsymbol{y_j})$

Clean Image (x)

Poisoned Image ($\hat{x}$)



Poisoned Image ($\hat{x}$)

Out of Distribution Generalization

Dataset Distillation and Adversarial Attacks

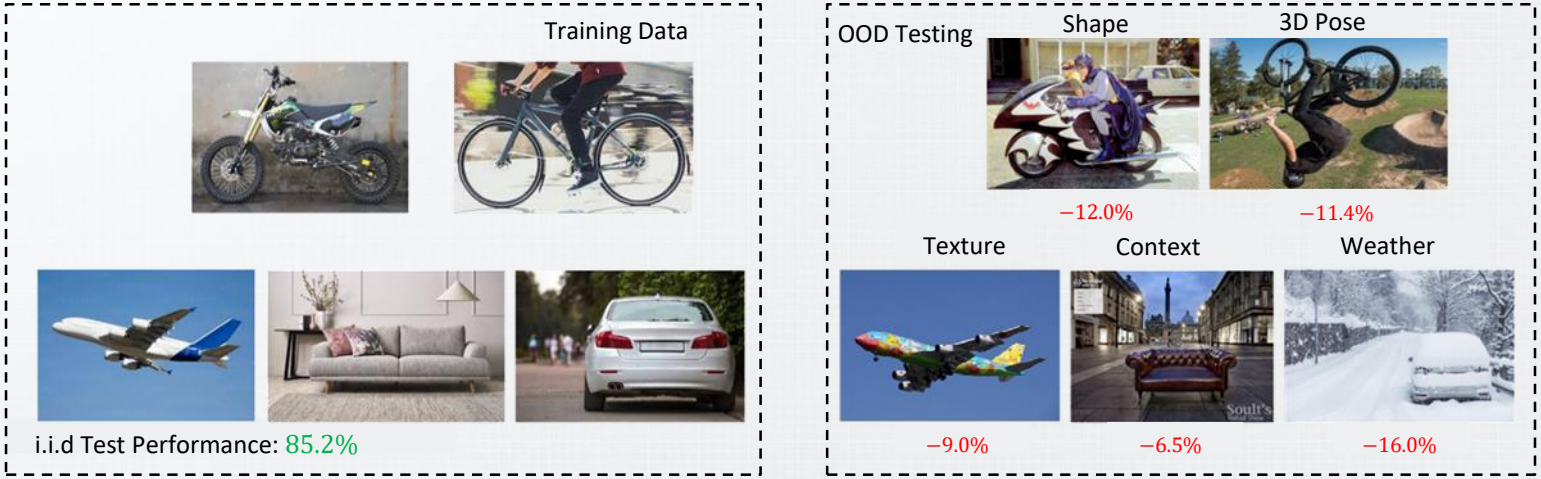
- **I**ndependent and **I**dentically **D**istributed (i.i.d) samples
  - Training and testing samples are drawn independently from the same distribution

- We usually develop and test deep learning models under the implicit assumption that the *training and test data are i.i.d*
  - Traditional machine-learning paradigms share the same assumption

Out of Distribution Generalization

Dataset Distillation and Adversarial Attacks

- **O**ut-**o**f-**D**istribution (OOD) samples
  - Data draw from different distribution (unseen distributions)
  - OOD generalization: The ability to perform tasks beyond the training distribution

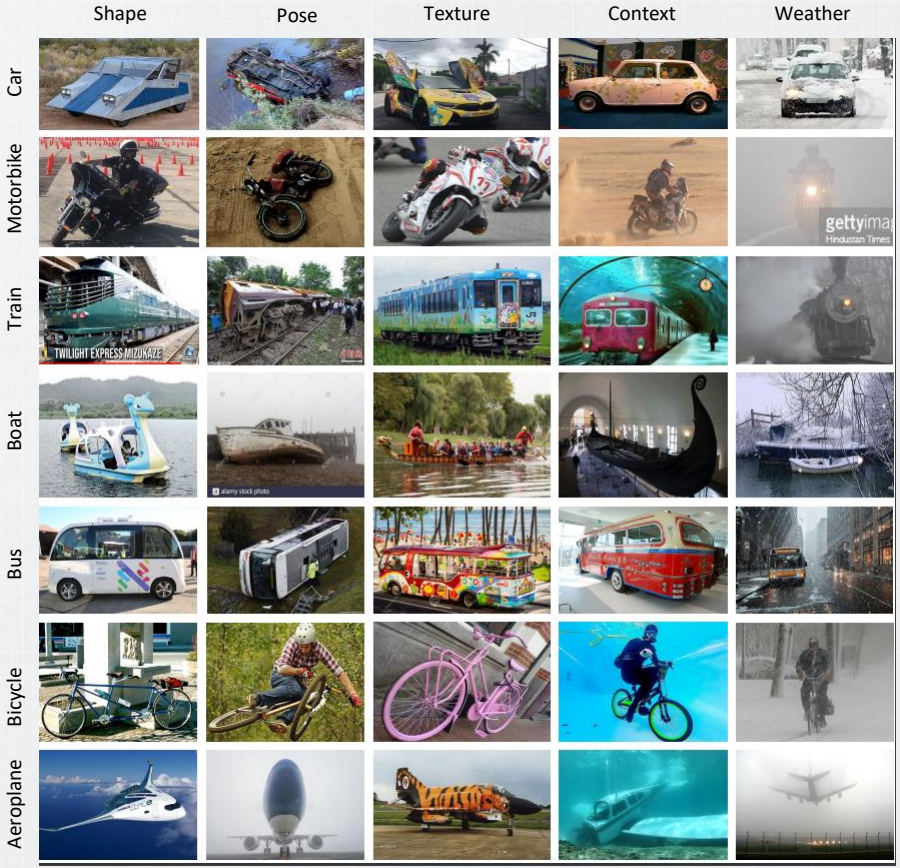


Adapted from Zhao et al., (2022)

Zhao et al. *OOD-CV : A Benchmark for Robustness to Out-of-Distribution Shifts of Individual Nuisances in Natural Images*. European Conference on Computer Vision (ECCV), 2022

Out of Distribution Generalization

Dataset Distillation and Adversarial Attacks



Zhao et al. *OOD-CV : A Benchmark for Robustness to Out-of-Distribution Shifts of Individual Nuisances in Natural Images*. European Conference on Computer Vision (ECCV), 2022

Adversarial Defenses and Robustness

Dataset Distillation and Adversarial Attacks

- Since adversarial images exist in real-world settings, increasing the adversarial **robustness** of convolutional networks plays a role in safety- and security-critical applications

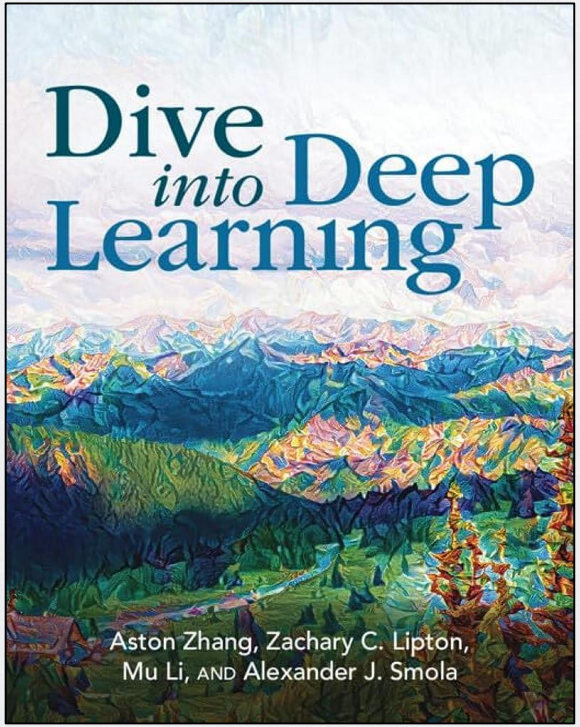
- Defense Mechanisms
  - Adversarial training
  - Regularization
  - Transfer learning
  - Data augmentation
  - Pruning

Jordao et al. *On the Effect of Pruning on Adversarial Robustness*. International Conference on Computer Vision (ICCV), 2021

Bibliography

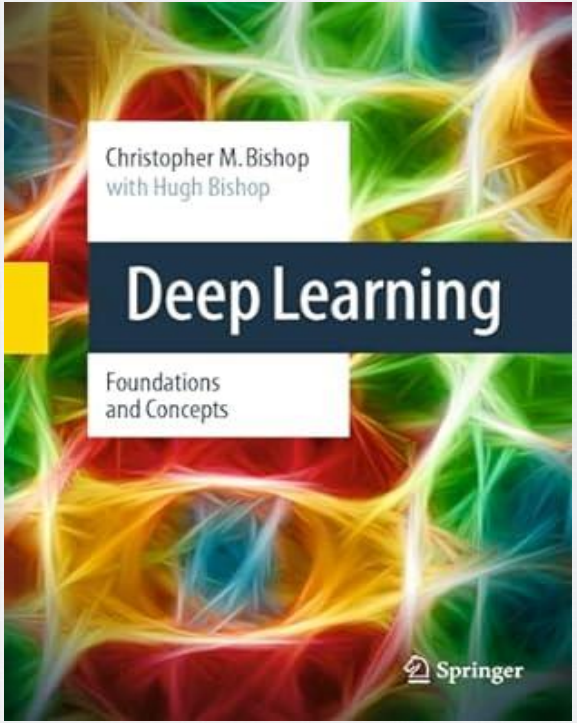
Bibliography

- Dive into Deep Learning
  - Chapter 14
    - 14.1 Augmentation



Bibliography

- Deep Learning: Foundations and Concepts
  - Chapter 10
    - 10.3.4 Adversarial attacks
    - 10.5.3 Synthetic images



Bibliography

- He et al. *Masked Autoencoders Are Scalable Vision Learners*. Computer Vision and Pattern Recognition (CVPR), 2022

- Hendrycks et al. *PIXMIX: Dreamlike Pictures Comprehensively Improve Safety Measures.* Computer Vision and Pattern Recognition (CVPR), 2022

- Li et al. Shape-texture Debiased Neural Network Training. International Conference on Learning Representations (ICLR), 2021

- SyntheticData4ML Workshop
  - https://www.vanderschaar-lab.com/syntheticdata4ml/

- Synthetic Data for Computer Vision
  - https://syndata4cv.github.io/