Investigation of DNN Model Robustness Using Heterogeneous Datasets

Wen-Hung Liao, Yen-Ting Huang

Dept. of Computer Science, National Chengchi University, TAIWAN

Pervasive Artificial Intelligence Research (PAIR) Labs, TAIWAN

ICPR 2020

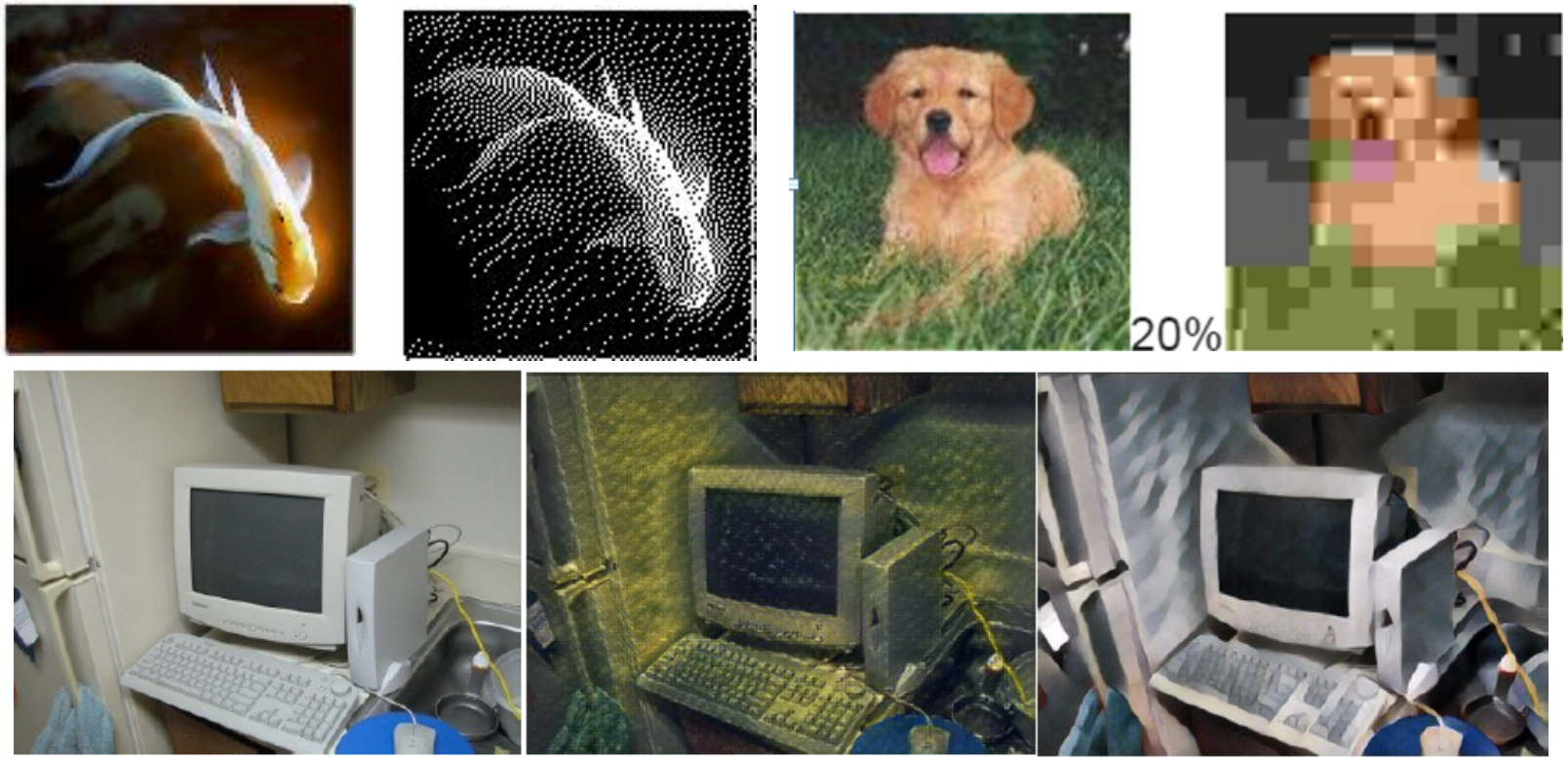
Objective

- This research aims to investigate some fundamental properties of the deep learning framework for computer vision tasks, including:
  - How to train the network when heterogeneous data (multiple distinct representations of the same source) are concerned?
  - What is the optimal information fusion strategy?
  - What actually happens inside the network when different fusion strategies are applied? (XAI)
  - Implications on devising defenses against adversarial attacks

ICPR 2020

Heterogeneous Data

- Definition: Multiple distinct representations of the same source
- Examples: RGB vs. halftone, original vs. compressed, original vs. stylized image



Coping with Hybrid Input

- Devise and train deep neural networks for the recognition of different representation of images.
  - Original vs. Low-Resolution Images
  - Original vs. Compressed Images
  - Original (intensity-based) vs. Halftone Images (density-based)

Training with Heterogeneous Data

- Hybrid training: train the model with both data sets simultaneously (data augmentation perspective)
- Feature concatenation: train the network with two branches and merge the extracted features (feature fusion perspective)
- The Order of Training: investigate the effect of continual learning by training the model with a pre-arranged order (incremental learning perspective)
- Finding: hybrid training yields better performance consistently

Training Results - Resolution and Compression Ratio

Original vs. down-sampled images

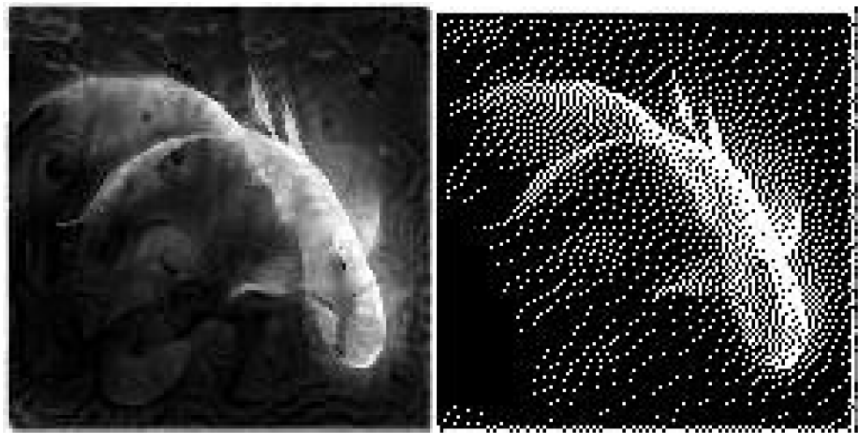
Training	Evaluation	Accuracy
Original	Original	0.61
	4x	0.17
	9x	0.08
4x down-sampled	Original	0.02
	4x	0.57
	9x	0.34
9x down-sampled	Original	0.01
	4x	0.28
	9x	0.51
Hybrid - Original + 4x down-sampled	Original	0.62
	4x	0.60
	9x	0.35
Hybrid - Original + 9x down-sampled	Original	0.57
	4x	0.47
	9x	0.53

Original vs. compressed images

Training	Evaluation	Accuracy
Original	Original	0.61
	60% compressed	0.36
	80% compressed	0.03
60% compressed	Original	0.47
	60% compressed	0.52
	80% compressed	0.05
80% compressed	Original	0.17
	60% compressed	0.24
	80% compressed	0.32
Hybrid - Original + 60% compressed	Original	0.58
	60% compressed	0.53
	80% compressed	0.09
Hybrid - Original + 80% compressed	Original	0.53
	60% compressed	0.48
	80% compressed	0.39

ICPR 2020

Training Results - Original vs. Halftone Images



Training	Evaluation	Accuracy	
		Top-1	Top-5
Grayscale	Grayscale	0.61	0.77
	FS Halftone	0.01	0.03
Floyd-Steinberg (FS) Halftone	Grayscale	0.42	0.64
	FS Halftone	0.61	0.80
Feature Concatenation	Grayscale	0.45	0.69
	FS Halftone	0.48	0.71
Hybrid Training	Grayscale	0.66	0.85
	FS Halftone	0.59	0.82

ICPR 2020

Training Results – Order and Hybrid Training

The Effect of Training Order

Training Order	Evaluation	Accuracy
4x then Original	Original	0.60
	4x	0.22
	9x	0.07
Original then 4x	Original	0.02*
	4x	0.55
	9x	0.36
9x then Original	Original	0.56
	4x	0.29
	9x	0.11
Original then 9x	Original	0.01*
	4x	0.27
	9x	0.51

*:catastrophic forgetting

Hybrid Training

Training	Evaluation	Accuracy
Hybrid	Original	0.59
	4x	0.57
	9x	0.54
	60% compressed	0.55
	80% compressed	0.40

Summary

- Strong dependence of classifier performance on training data
- Compressed and down-sampled images maintain the overall structure but remove detailed textures, resulting in poor performance (consistent with ICLR 2019 paper)
- Similar results observed in halftone representation, color-depth reduced images obtained with mean-shift segmentation or bilateral filtering.
- Remedy: increase shape bias by adding heterogeneous data sets to the training samples.
- Order of training is crucial in hybrid training.
- Cannot augment endlessly, when to stop?

ICPR 2020

ICPR 2020