Generalisation in humans and deep neural networks

NIPS 2018

Presented by Zubia Mansoor

August 20, 2020

Overview

Motivation

Motivating Example

Consider a motivating scenario in radiology

- What happens when the model encounters something it hasn't seen before?
  - For instance, if the X-Ray copies are blurry and noisy
- Changes in the training and test distribution pose a serious challenge to deep learning vision systems

Goals of this paper

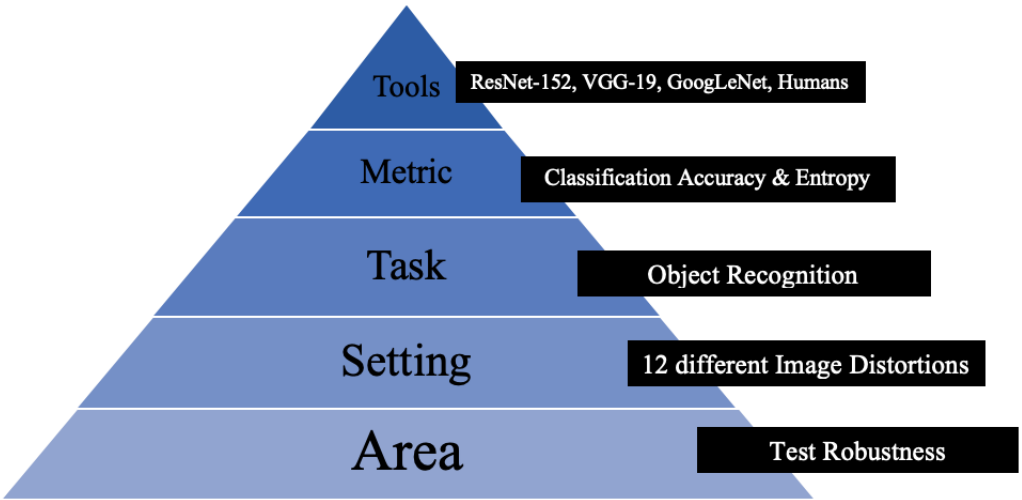
General

- Conduct a behavioral comparison of human and DNNs to test for robustness on image distortions

Specfic

- Test the performance of DNNs on the *exact distortion types they were trained on*
- Test the performance of DNNs on *previously unseen distortions*

Workflow


Tools
ResNet-152, VGG-19, GoogLeNet, Humans
Metric
Classification Accuracy & Entropy
Task
Object Recognition
Setting
12 different Image Distortions
Area
Test Robustness

Methods

Paradigm & Procedures

- Controlled lab experiment
- ImageNet data
- Design Challenge 1
  - Problem: DNNs are usually trained for fine-grained categories as compared to humans
  - Solution: Distill over 20,000 categories of ImageNet into 16 entry-level categories using WordNet
- Design Challenge 2
  - Problem: DNNs only use feedforward computation versus recurrent connections used by human brain
  - Solution: Each image is followed by a 200 ms presentation of a pink noise mask

Types of Distortion

Name	Data type
Colour vs. greyscale	Dichotomous
True vs. false colour	Dichotomous
Power equalisation	Dichotomous
Rotation	4 levels
Phase noise	7 levels
Uniform noise	8 levels
Low pass	8 levels
High pass	8 levels
Contrast	100% to 0%
Eidolon I,II,III	8 levels
Salt-and-Pepper	100% to 0%

Types of Distortion Contd.

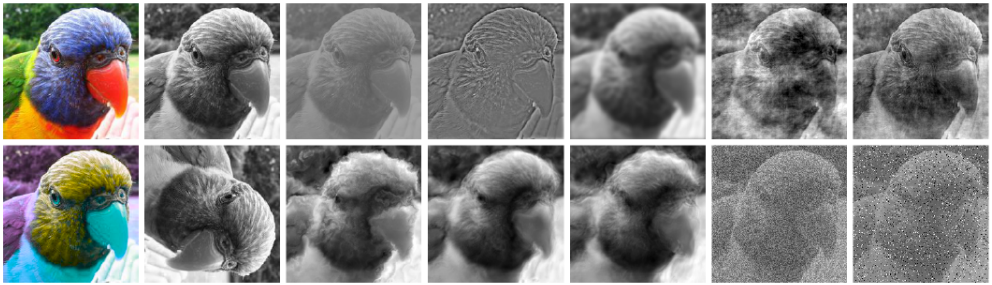


Figure: Sample image of class *bird* across all distortion types

Experiments

Experiment I: Tested on unseen distortions

Goals

- How well do DNNs generalise as compared to humans on previously unseen distortions?

Experiment II: Tested on trained distortions

Goals

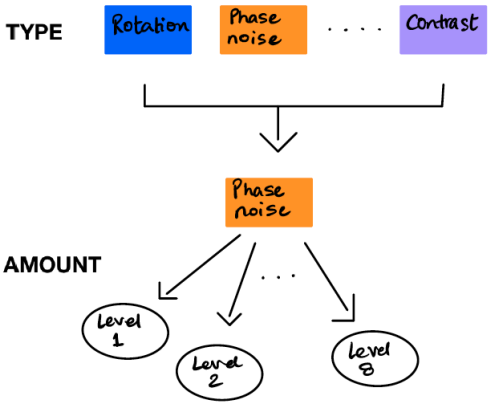
- How well do DNNs generalise as compared to humans when trained directly on the same distortions?
- To what extent does training on one distortion type helps the network cope with other distortion types?

More on the training process...

- Experiment I
  - Standard pre-trained DNNs
- Experiment II
  - Trained on a subset of the standard ImageNet dataset (16-class-ImageNet)
  - Weight each sample in the loss function to correct for imbalanced classes
  - ResNet 50- like architecture

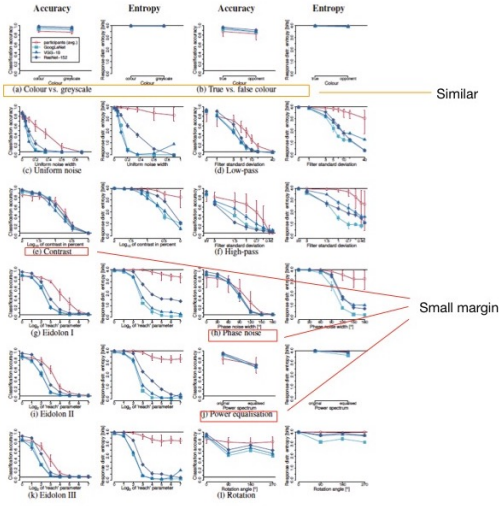
More on the training process...(1)

- Experiment II
  - Type and amount of distortion selected uniformly



Results

Results: Experiment I



Findings

- Humans and DNNs exhibit *similar performance* on colour-related distortions
- Human appear to be more robust *by a small margin* for low-contrast, power equalisation and phase noise
- Human appear to be more robust *by a large margin* for uniform noise, low-pass, high-pass, rotation and all eidolon experiments

Results: Experiment II

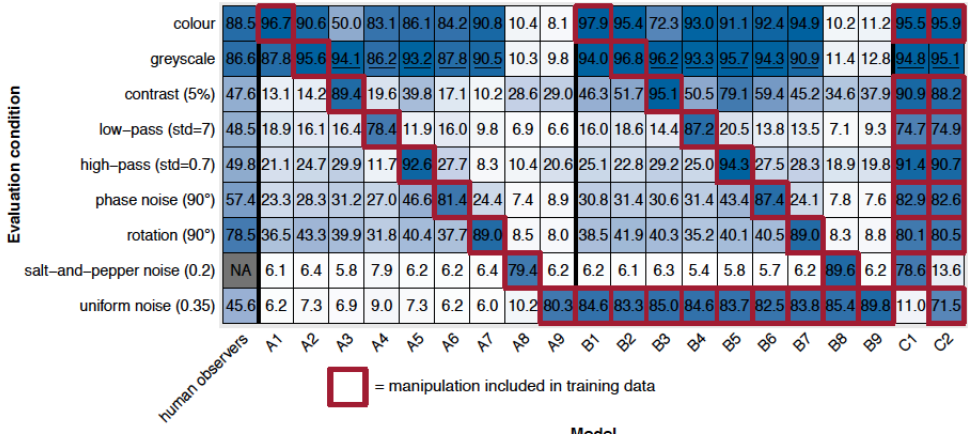


Figure: Classification Accuracy (in %)

Findings

- **Models A1-A9**: Training on a certain distortion improves the performance greater for that same distortion but only slightly for other distortions
- **Models B1-B9**: Training on a specific distortion combined with uniform noise improves more compared to models A1 to A9
- **Models C1-C2**: Training either without uniform or salt-and-pepper noise leads to poor and closer to chance level performance

Summary

Summary

- Human visual system appears to be more robust than DNNs for the most part
- Diverging classification error-patterns between humans and DNNs as signal weakens
- DNNs surpass human performance only when trained on the exact distortions type they are later tested on
- Benchmark dataset of 83,000 human pyschophysical trials

Questions?