Measuring the Tendencies of CNNs to Learn Surface Level Regularities

J. Jo, Y. Bengio

Université de Montréal

arXiv:1711.11561

Reviewed by : Bill Zhang
University of Virginia
https://qdata.github.io/deep2Read/

Outline

Introduction

Basic Premise and Motivation

- CNNs have achieved record-breaking object recognition on CIFAR-10, SVHN, and ImageNet datasets; de facto machine learning model for visual tasks
  - Good generalization Performance
- CNNs also sensitive to adversarial examples which humans can identify easily, but CNNs predict incorrectly usually with high confidence
  - Doubt that CNNs learn high-level abstraction; possibility that CNN overfits to superficial cues present in both training and testing sets

Introduction

Perturbation Map

- Create a perturbation map $F : X \to X'$ which satisfies the following:
    - Preserves object recognizability: For any $x \in X$ and its perturbation $x' \in X'$, recognizability should be preserved in human perspective
    - Qualitatively different image statistics: With property 1, guarantees preservation of high level abstraction but different superficial cues
    - Existence of non-trivial generalization gap: A model trained on unperturbed training set or perturbed training set tested on unperturbed and perturbed testing set should yield different accuracy results
- Use radial and random Fourier masks

Introduction

Claims

- Claim 1: CNNs are generalizing extremely well to an unseen test set
- Claim 2: General sensitivity to adversarial examples show that deep CNNs are not truly capturing abstractions in the dataset
- Hypothesis: The current incarnation of deep neural networks exhibit a tendency to learn surface statistical regularities as opposed to higher level abstractions in the dataset.
  - Sufficient for image recognition due to strong statistical properties of natural images, but only in a narrow distributional sense

Fourier Filtering

Overview and Equations

- Although natural images have high variance in pixel space, they tend to have most of their Fourier frequencies concentrated in low to mid-range frequencies
  - It is possible to filter frequencies out while preserving most of original image
- Consider the following sets:
  - (X, Y), the original dataset
  - (X_{radial}, Y), low frequency filtered version
  - (X_{random}, Y), randomly filtered version
- With $X \in \mathbb{R}^{HxW}$, the 2D DFT of an image is:

$$F(X)[k, l] := \frac{1}{\sqrt{HW}} \Sigma_{h=0}^{H-1} \Sigma_{w=0}^{W-1} X[w, h] e^{-j2\pi(\frac{wk}{W} + \frac{lh}{H})}$$

where k ranges from 0 to W-1 and l ranges from 0 to H-1

Fourier Filtering

Equations

- Consider a shifted DFT where DC component is in the center of the image; masks applied to all 3 color channels
- Radial Filtering: Parameterized by radius r, W and H even

$$M_r[i,j] := \begin{cases} 1 & \text{if } ||(i,j) - (W/2, H/2)||_{l_2} \leq r \\ 0 & \text{otherwise} \end{cases}$$

$$X_{radial} := F^{-1}(F(X) \circ M_r)$$

- Random Filtering: Parameterized by probability p

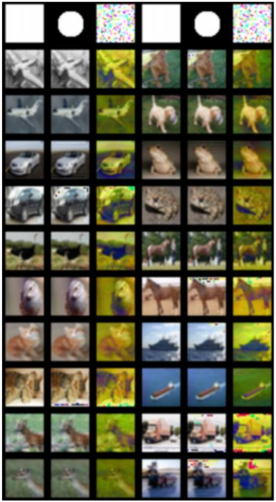
$$M_p[c,i,j] := \begin{cases} 0 & \text{with probability } p \\ 1 & \text{otherwise} \end{cases}$$

$$X_{random} := F^{-1}(F(X) \circ M_p)$$

Fourier Filtering

Post-Filtered Images

- Filtered images remain recognizable to humans; SVHN on left, CIFAR-10 on right



Experiments

Procedures

- Train on one of $(X_{train}^{unfiltered}, X_{train}^{radial}, X_{train}^{random})$ and test accuracy on all test sets $(X_{test}^{unfiltered}, X_{test}^{radial}, X_{test}^{random})$
- Define generalization gap as the maximum difference in accuracy among testing sets
- Used a Preact Resnet with Bottleneck architecture of depth 92 and 200
- Also trained on fully augmented training set

$$X_{train}^{augmented} := X_{train}^{unfiltered} \cup X_{train}^{radial} \cup X_{train}^{random}$$

Experiments
SVHN Results

- Train ResNet for 40 epochs using Nesterov Momentum; learning rate = 0.01, momentum = 0.9
- Batch size of 128; learning rate divided by 10 at epochs 20 and 30

Train/Test	Unfilt.	Radial	Random	Gen. Gap
Unfilt.	1.95%	8.41%	2.68%	6.46%
Radial	3.50%	5.07%	5.67%	2.17%
Random	4.01%	11.90%	2.04%	7.89%
Augmented	2.11%	5.06%	2.15%	2.95%

(a) Preact-ResNet-Bottleneck-92 SVHN Generalization

Train/Test	Unfilt.	Radial	Random	Gen. Gap
Unfilt.	1.88%	8.31%	2.42%	6.43%
Radial	3.56%	4.90%	4.77%	1.34%
Random	2.95%	9.85%	1.96%	7.89%
Augmented	1.94%	4.87%	2.06%	2.93%

(b) Preact-ResNet-Bottleneck-200 SVHN Generalization

Experiments

CIFAR-10 Results

- Train ResNet for 100 epochs using Nesterov Momentum; learning rate $= 0.01$ (boosted up to 0.1 after 400 updates), momentum $= 0.9$
- Batch size of 128; learning rate divided by 10 at epochs 50 and 75
- Augmented training for 120 epochs, decay at 60 and 80

Train/Test	Unfilt.	Radial	Random	Gen. Gap
Unfilt.	5.54%	25.75%	12.31%	20.21%
Radial	6.91%	7.91%	18.45%	11.54%
Random	7.12%	35.03%	6.76%	28.27%
Augmented	5.85%	7.89%	6.74%	2.04%

(a) Preact-ResNet-Bottleneck-92 CIFAR-10 Generalization

Train/Test	Unfilt.	Radial	Random	Gen. Gap
Unfilt.	5.22%	23.37%	11.26%	18.15%
Radial	6.35%	7.07%	17.09%	10.74%
Random	6.47%	34.19%	5.9%	28.29%
Augmented	5.37%	7.25%	6.3%	1.88%

(b) Preact-ResNet-Bottleneck-200 CIFAR-10 Generalization

Experiments

Discussion

- All trained models generalized well to unfiltered set; suggests that Fourier filtering produced datasets perceptually not far off from unfiltered sets
- Model trained on unfiltered set tend to latch onto image statistics of the training set, yielding a non-trivial generalization gap; no training set generalized to all other sets
- Although augmented set did reduce generalization gap, it does not mean augmented set is sufficent for all adversarial examples

Summary

- CNNs generalize well but are also sensitive to adversarial examples; models may be learning superficial cues rather than high-level abstraction
- Can use Fourier filtering as a perturbation map to show how models fail to recognize perceptually similar images due to different image statistics
- No training dataset generalized well to all of the datasets; model trained on augmented set, although effective at closing the generalization gap, may not be sensitive to other perturbation maps

References

- https://arxiv.org/pdf/1711.11561.pdf