Understanding Invariance via Feedforward Inversion of Discriminatively Trained Classifiers

Piotr Teterwak[1], Chiyuan Zhang[2], Dilip Krishnan[2], Michael C. Mozer[2]

1 Boston University

2 Google Research

What is preserved in the logit (pre-softmax) outputs?

- A trained neural network can filter all information from the input other than class label

What is preserved in the logit (pre-softmax) outputs?

- A trained neural network can filter all information from the input other than class label
- Prior work has shown that *some* information can be preserved in the logits, the most successful of which is Dosovitskiy and Brox(2016, NeurIPS)

What is preserved in the logit (pre-softmax) outputs?

- A trained neural network can filter all information from the input other than class label
- Prior work has shown that *some* information can be preserved in the logits, the most successful of which is Dosovitskiy and Brox(2016, NeurIPS)

Original

Reconstruction

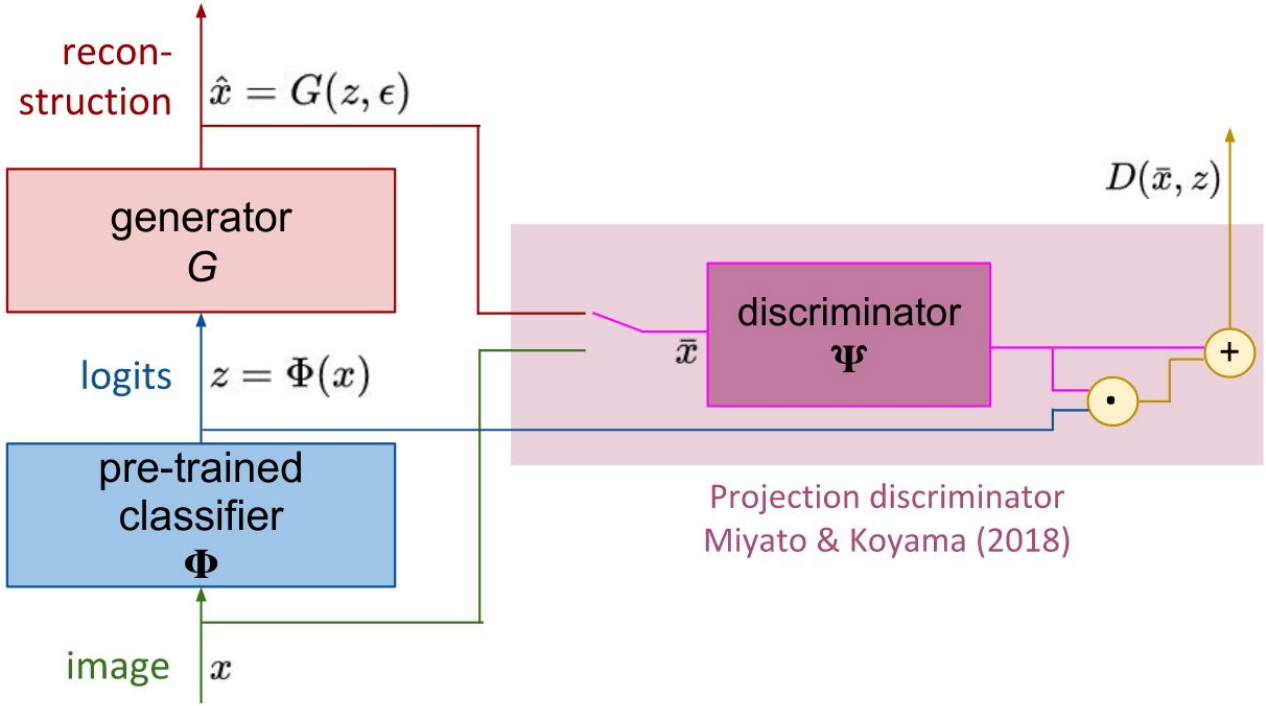
What is preserved in (pre-softmax) outputs of neural nets?

- The field was overdue to revisit this question with:
  - Modern classifiers
  - Better methods for Image synthesis

What is preserved in (pre-softmax) outputs of neural nets?

- The field was overdue to revisit this question with:
  - Modern classifiers
  - Better methods for Image synthesis
- We show:
  - Surprising fidelity in reconstruction
  - Information preserved in logits depends on optimization method and architecture
  - Robust models contain information in the logits which supports better reconstructions of both shapes and textures relative to standard models

Method


recon-
struction
$\hat{x} = G(z, \epsilon)$
generator
G
logits
$z = \Phi(x)$
pre-trained
classifier
Φ
image
x
$D(\bar{x}, z)$
$\bar{x}$
discriminator
Ψ
+
Projection discriminator
Miyato & Koyama (2018)

Results

original

original

original

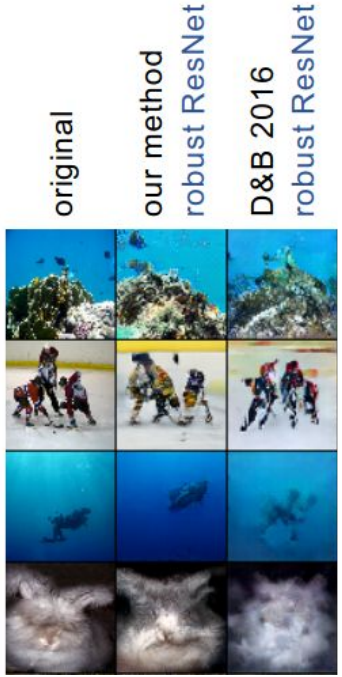
Results


original
our method
robust ResNet


original
our method
robust ResNet


original
our method
robust ResNet

Results


original
our method
robust ResNet
D&B 2016
robust ResNet


original
our method
robust ResNet
D&B 2016
robust ResNet


original
our method
robust ResNet
D&B 2016
robust ResNet

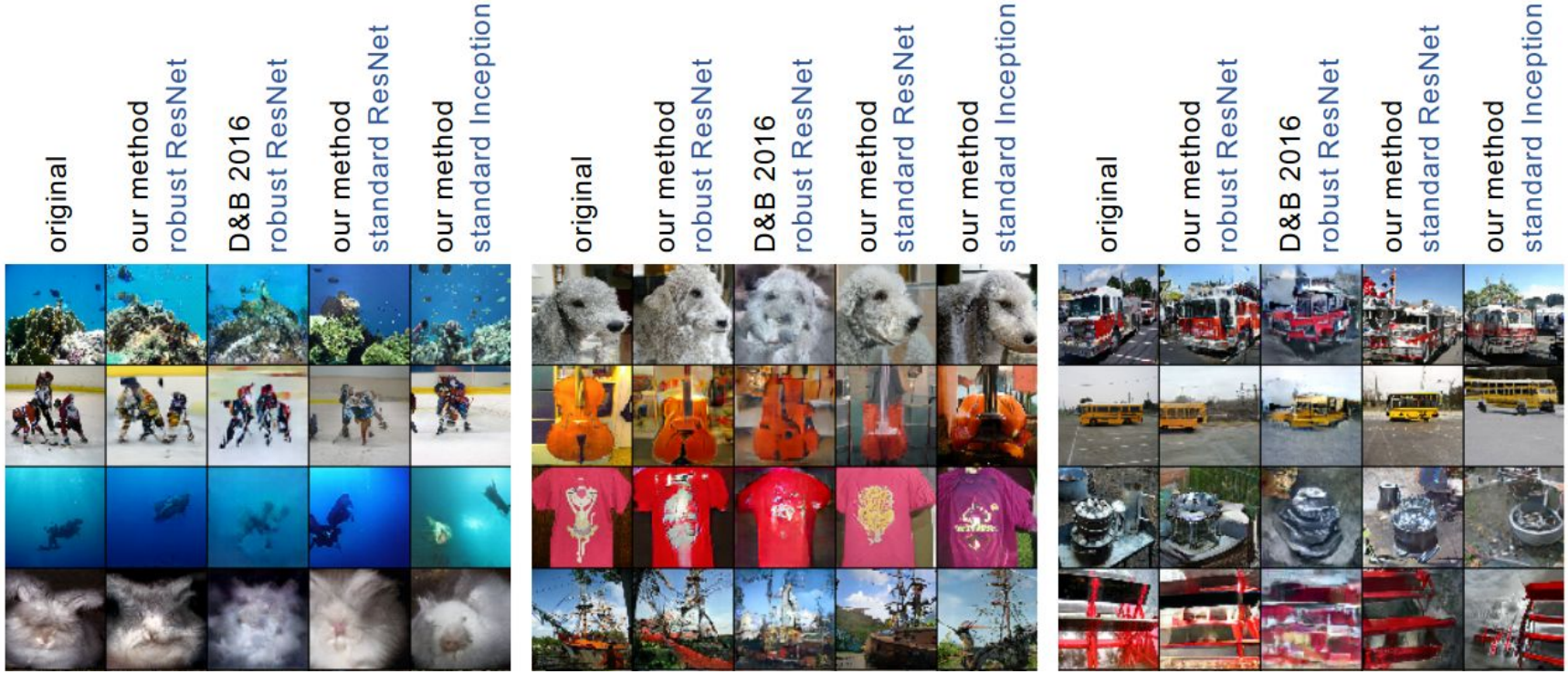
Results


original
our method robust ResNet
D&B 2016 robust ResNet
our method standard ResNet


original
our method robust ResNet
D&B 2016 robust ResNet
our method standard ResNet


original
our method robust ResNet
D&B 2016 robust ResNet
our method standard ResNet

Results


original
our method robust ResNet
D&B 2016 robust ResNet
our method standard ResNet
our method standard Inception
original
our method robust ResNet
D&B 2016 robust ResNet
our method standard ResNet
our method standard Inception
original
our method robust ResNet
D&B 2016 robust ResNet
our method standard ResNet
our method standard Inception

Invariances: Constant Logit, Varying GAN Noise Input

And More!

- Manipulating logits with scales/shifts/perturbations
- Logit interpolations between images
- Reconstructions of adversarially attacked images
- Reconstructions of incorrectly classified images
- Reconstructions of OOD data
- Find the paper here: https://arxiv.org/abs/2103.07470

Thank you!