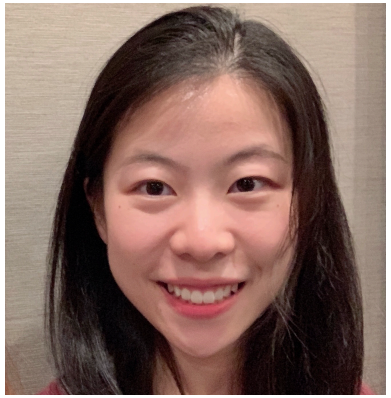
Neural networks with recurrent generative feedback

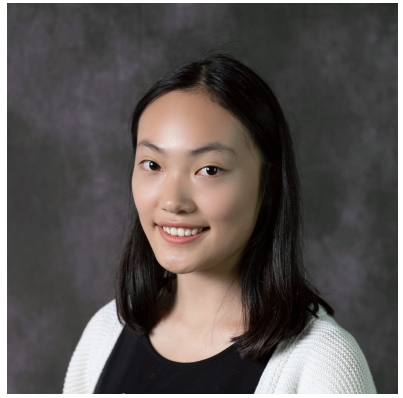
Yujia Huang, Caltech

yjhuang@caltech.edu

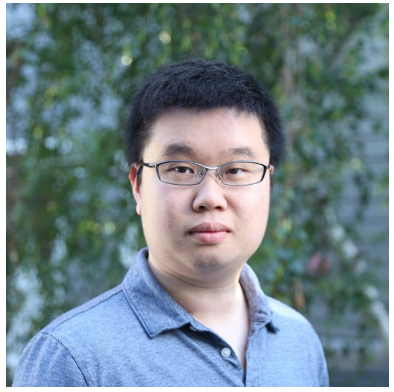


Yujia Huang, Caltech

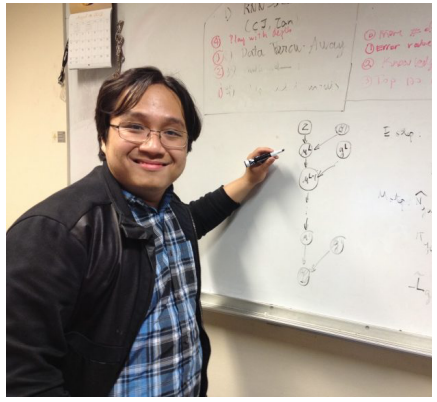
James Gornet, Caltech



Sihui Dai, Caltech



Zhiding Yu, NVIDIA

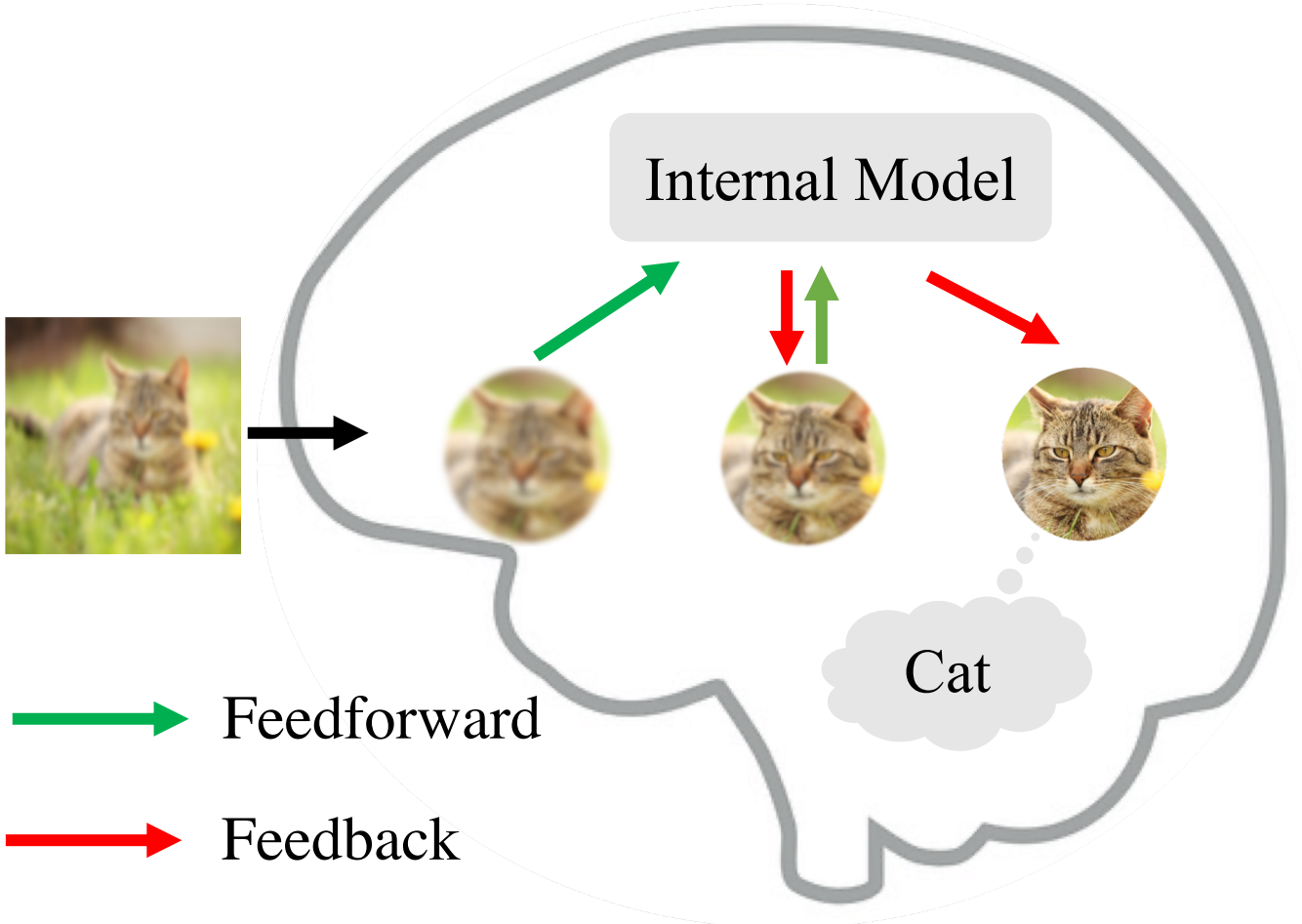


Tan Nguyen,
Rice University

Doris Y. Tsao, Caltech

Anima Anandkumar,
Caltech/NVIDIA

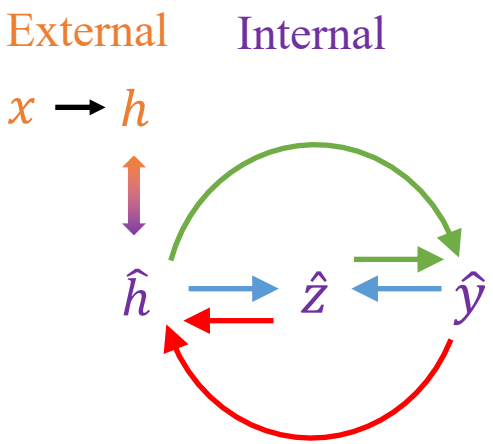
Intuition



Self-Consistency

Given a joint distribution $p(h, y, z; \theta)$ parameterized by θ, $(\hat{h}, \hat{y}, \hat{z})$ are self-consistent if they satisfy the following constraints:

$$\hat{y} = \arg\max_{y} p(y|\hat{h}, \hat{z}),$$
$$\hat{h} = \arg\max_{h} p(h|\hat{y}, \hat{z}),$$
$$\hat{z} = \arg\max_{z} p(z|\hat{h}, \hat{y})$$



x: image

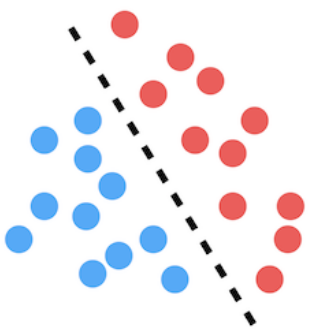
h: encoded feature

z: latent variable

y: label

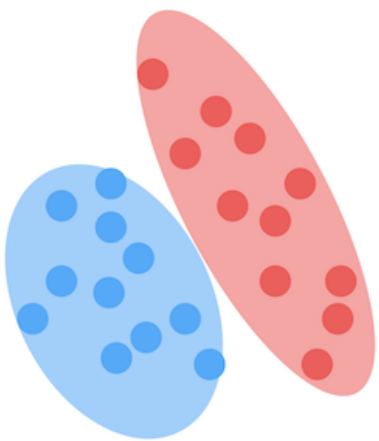
Generative Classifier

Logistic Regression



$p(y|x)$

Gaussian Naïve Classifier

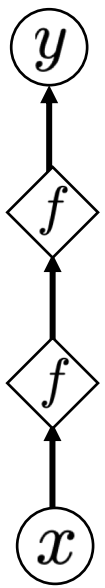


$p(x, y) \rightarrow p(y|x)$

A. Ng, and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Neurips 2002.

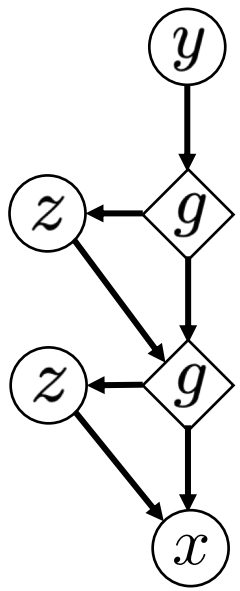
Deconvolutional generative model (DGM)

CNN



$p(y|x)$

DGM



$p(x, y, z) \rightarrow p(y|x, z)$

$$y \sim p(y)$$

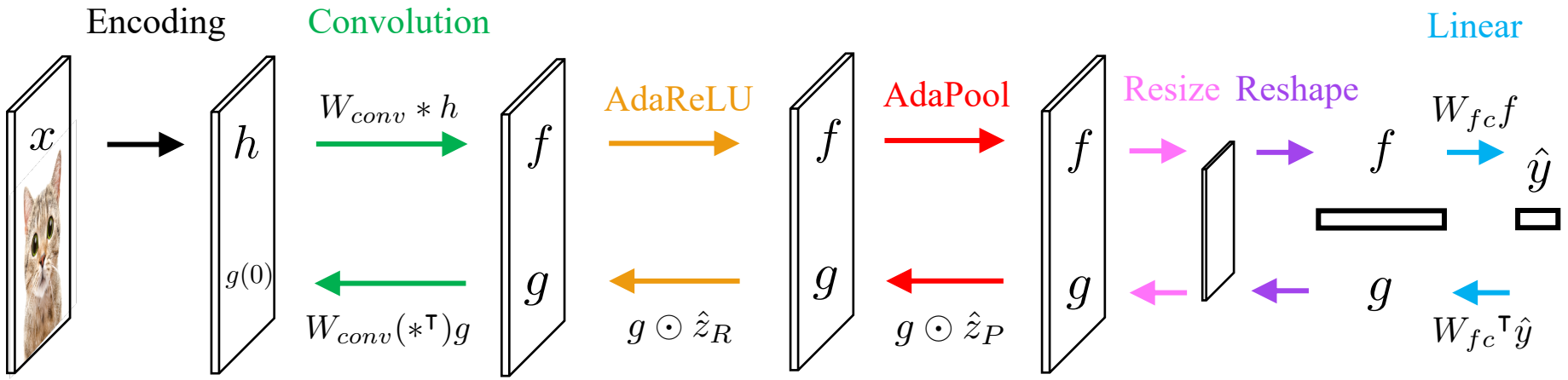
$$z_P^{(i)} \sim \text{Ber}\left(\frac{e^{b \cdot g^{(i)}}}{e^{b \cdot g^{(i)}} + 1}\right)$$

$$z_R^{(i)} \sim \text{Ber}\left(\frac{e^{b \cdot g^{(i)}}}{e^{b \cdot g^{(i)}} + 1}\right)$$

$$x \sim \mathcal{N}(g(0), \text{diag}(\sigma^2))$$

T. Nguyen, N. Ho, A. Patel, A. Anandkumar, M. I. Jordan, and R. G. Baraniuk. A bayesian perspective of convolutional neural networks through a deconvolutional generative model. arXiv:1811.02657, 2018.

Inference in the DGM



Theorem (Informal): The generative classifier derived from the DGM is a CNN with AdaReLU and AdaPool, i.e. $\hat{y} = \mathrm{CNN}(h)$.

$$\sigma_{\mathrm{AdaReLU}}(f) = \begin{cases} \sigma_{\mathrm{ReLU}}(f), & \text{if } g \geq 0 \\ \sigma_{\mathrm{ReLU}}(-f), & \text{if } g < 0 \end{cases}$$

$$\sigma_{\mathrm{AdaPool}}(f) = \begin{cases} \sigma_{\mathrm{MaxPool}}(f), & \text{if } g \geq 0 \\ -\sigma_{\mathrm{MaxPool}}(-f), & \text{if } g < 0 \end{cases}$$

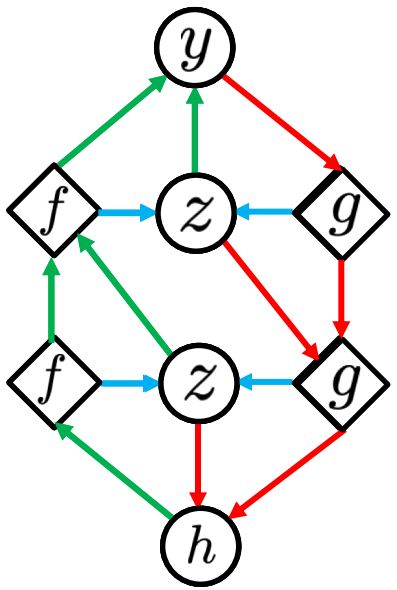
Iterative inference

Self-consistency

$$\hat{y} = \arg\max_{y} p(y|\hat{h}, \hat{z}),$$

$$\hat{h} = \arg\max_{h} p(h|\hat{y}, \hat{z}),$$

$$\hat{z} = \arg\max_{z} p(z|\hat{h}, \hat{y})$$



Feedforward

Feedback

Inference of z

y Label

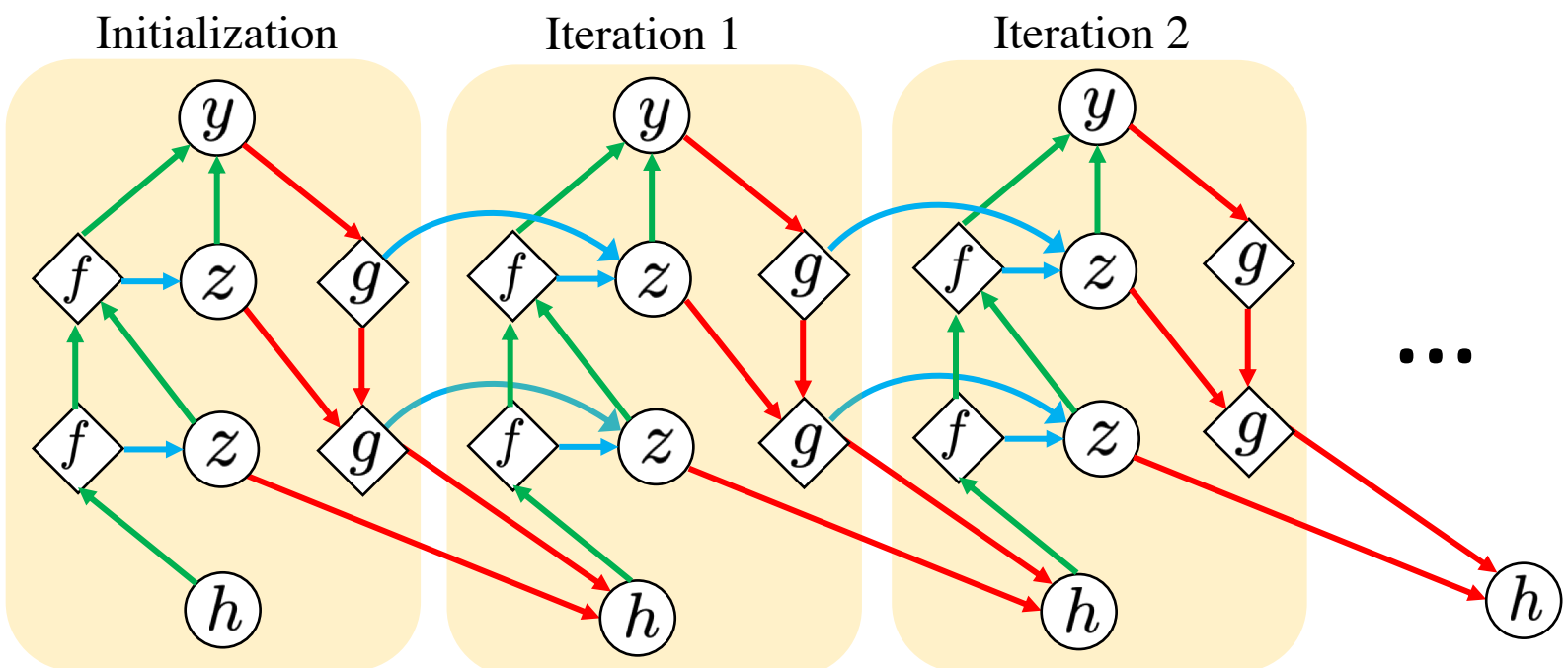
z Latent variables

h Image features

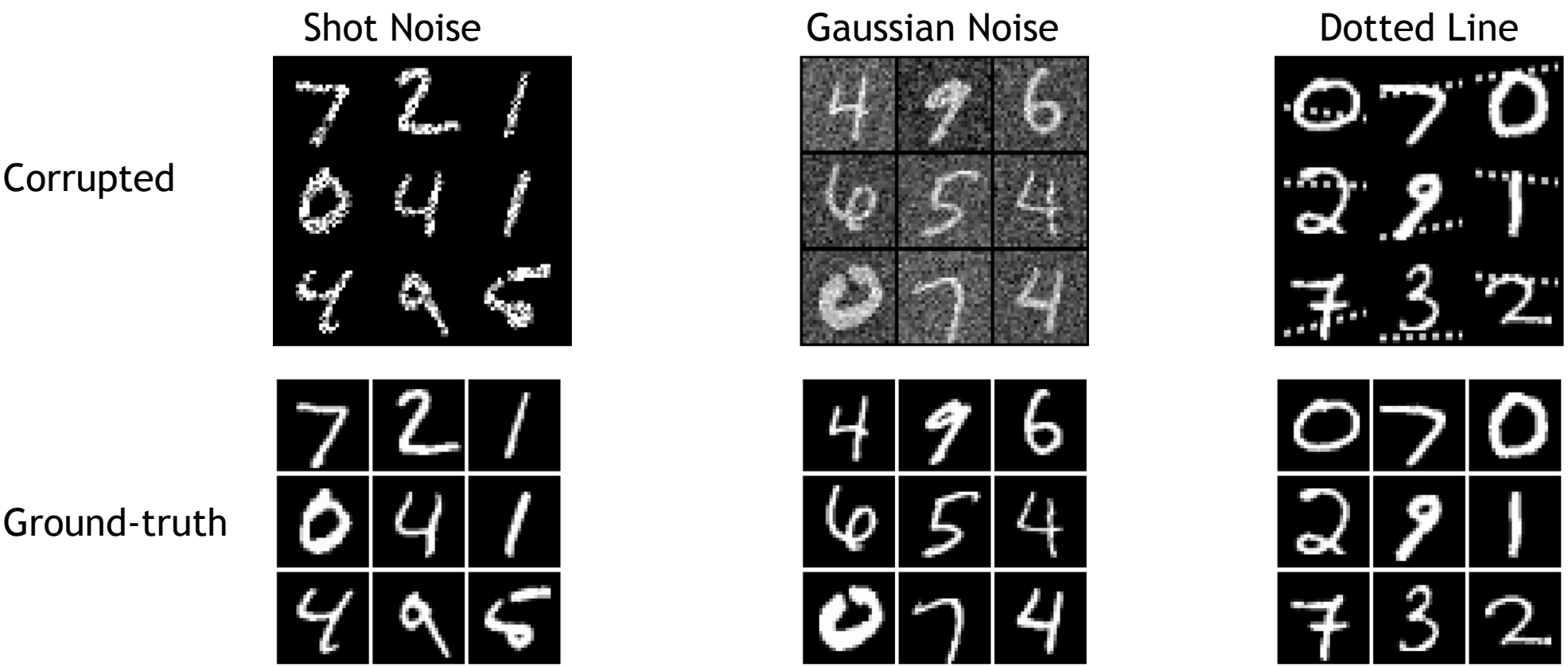
f Feedforward layer

g Feedback layer

CNN-F

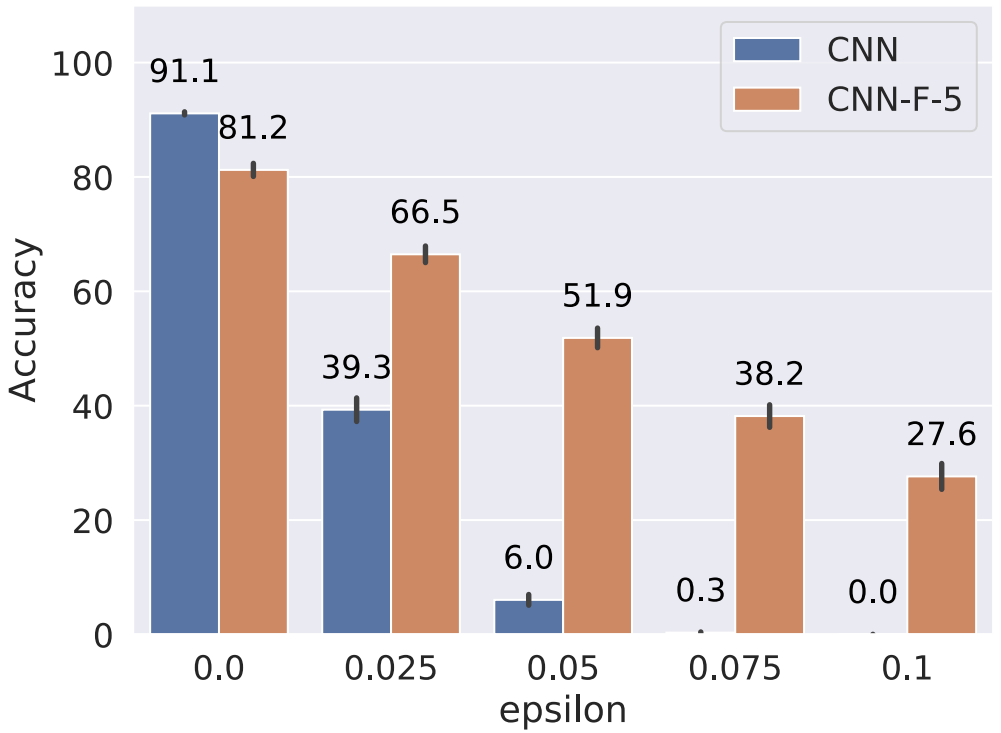


Sanity check: CNN-F repairs distorted images

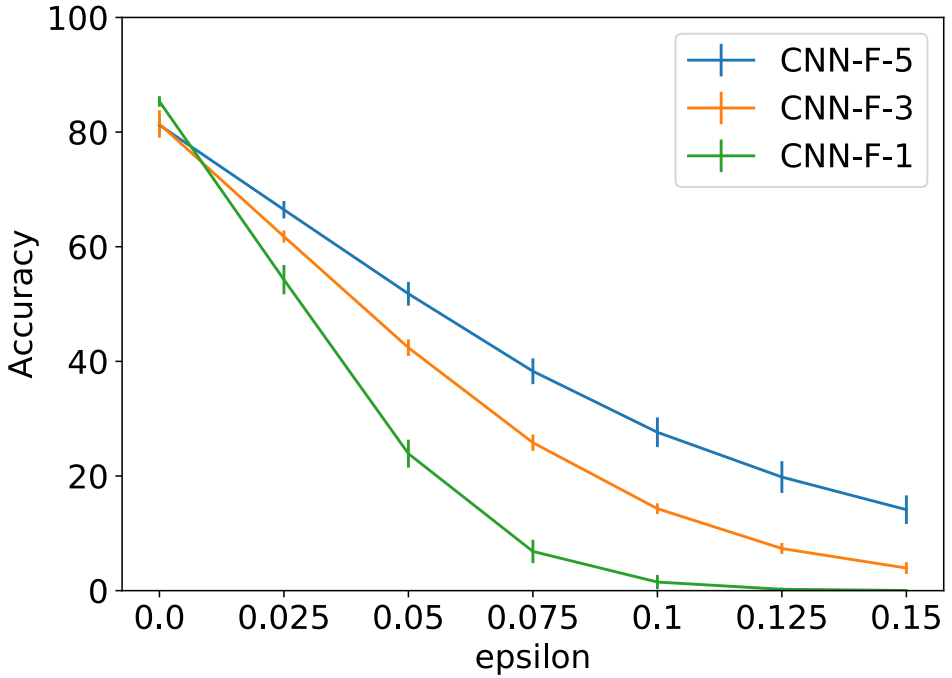


CNN-F improves adversarial robustness

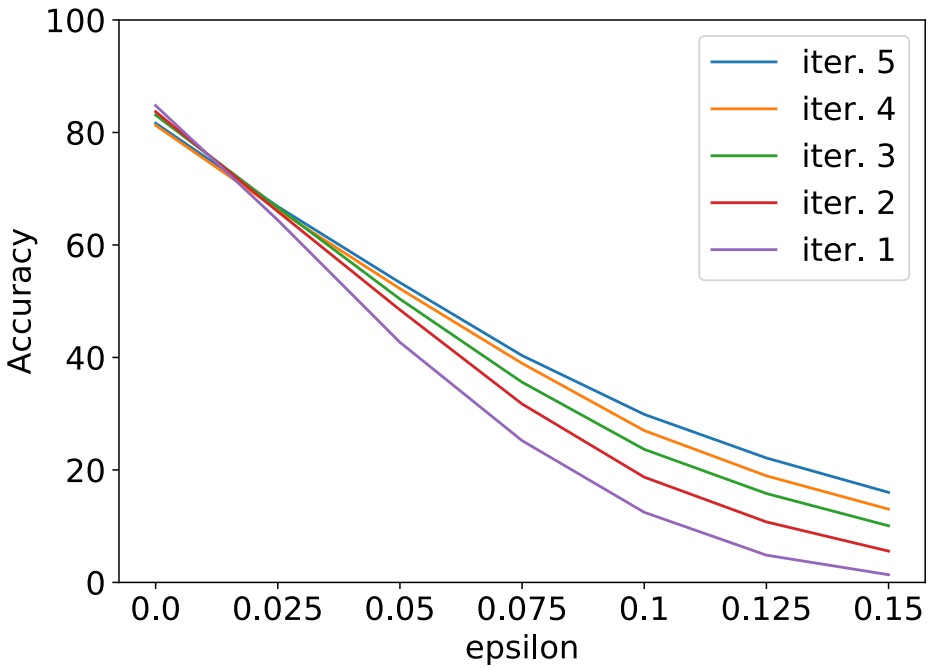
- Standard training on Fashion-MNIST.
- Attack with PGD-40.
- CNN-F has higher adversarial robustness than CNN.



CNN-F improves adversarial robustness



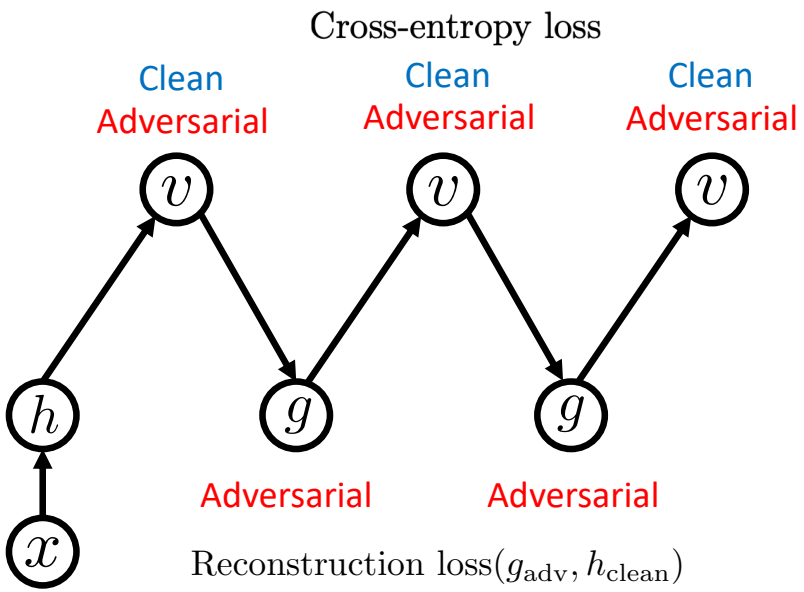
CNN-F trained with different iterations.



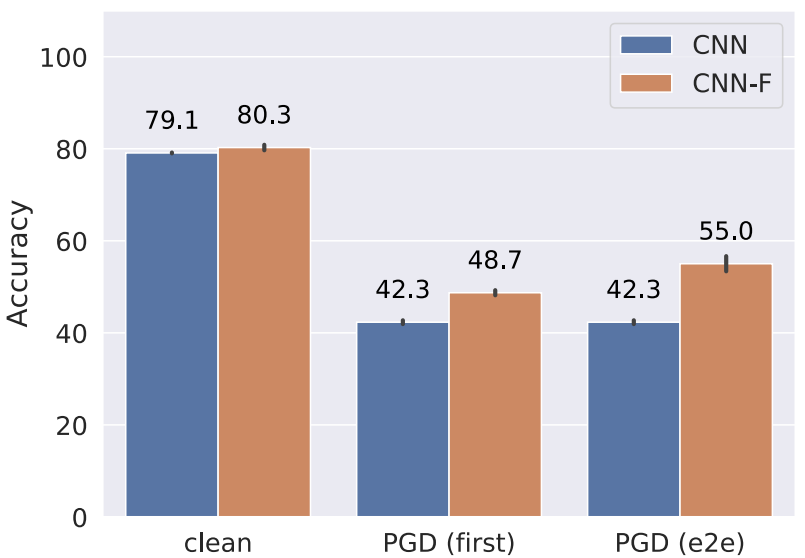
CNN-F tested with different iterations.

More iterations are needed for *harder* images.

CNN-F with adversarial training



- Dataset: CIFAR-10
- Architecture: WideResNet-40-2
- Evaluated against various adversarial attacks
- CNN-F improves clean and adversarial accuracy of CNN



Thank You!