ILLINOIS

Bioengineering Department
GRAINGER COLLEGE OF ENGINEERING

Investigation of Adversarial Robust Training for Establishing Human-Interpretable CNN-based Numerical Observers

Sourya Sengupta

BIOE 580

Deep Learning Classification Models

- Given a set of data and their corresponding labels, a deep neural network tries to optimize a set of parameters by minimizing the loss between the predicted labels and actual labels.

- Provided data x and their corresponding labels y, the goal is to estimate the weight parameters θ that specify the network.

- Standard training problem: $\theta^* = \min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} [\mathcal{L}_\theta(x, y)]$

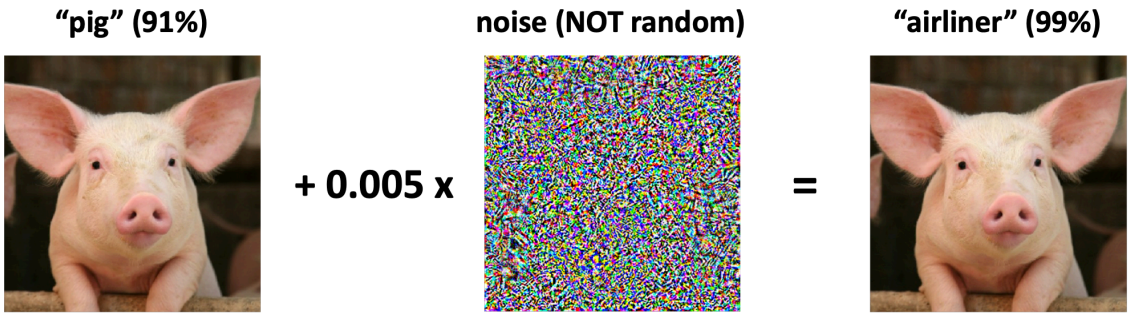
- For example, cross entropy for binary classification problems

Conventional networks are brittle

- Conventionally trained networks can be “brittle”.
- Deep networks can make predictions based on image features that are vastly different from what humans use, or even recognize.
- For example, one can readily construct pairs of images that appear completely different to a human but are nearly identical in terms of learned feature representations.
- This implies the existence of adversarial examples.

Adversarial examples

•Adversarial perturbations to input images:



•The left image was classified properly by a standard deep neural network but after introducing small perturbations the classifier predicts it as airliner.

•The adversarial perturbation looks like random noise but can result in an incorrect classification with high confidence.

Image Source: Madry et. al: A Brief Introduction to Adversarial Examples

Adversarial training

- To mitigate adversarial examples, new training strategies are actively being developed.

- The basic idea ("adversarial training") is to create and then incorporate adversarial examples into the training process.

- Adversarial training: $\theta^* = \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\Delta} \mathcal{L}_\theta(x+\delta, y) \right]$

- Here, Δ is a set of perturbations

- In this case, one repeatedly finds the worst-case input perturbations δ and then update the model parameters to reduce the loss on these perturbed inputs.

Adversarial training

- Adversarial training: $\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\Delta} \mathcal{L}_\theta(x+\delta, y) \right]$
- Common choice for $\Delta = \{\delta \in R^d \mid \left|\left|\delta\right|\right|_p \leq \epsilon\}$
  - Set of l_p bounded perturbations
- Here, ϵ is a tunable parameter
- Larger ϵ will result in a higher degree of robustness

Adversarial training

- **Important observation**: Robustly trained classifiers can learn feature representations that are recognizable to humans.

- Moreover, there is a tradeoff between robustness and accuracy.

Ilyas, Andrew, et al. "Adversarial examples are not bugs, they are features." *arXiv:1905.02175* (2019).

Numerical Observer

- An observer is a person or computer algorithm who infers some decision based on some available data.
- Numerical model observers are mathematical models designed to take decision depending upon different tasks.
- The ideal observer is a model that describes the performance of the optimum decision maker on a given decision task in bayesian sense. The ideal observer therefore provides a theoretical upper bound on task performance.
- But humans are sub-optimal observers, humans are substantially less efficient than the ideal observer.

Geisler, Wilson S. "Contributions of ideal observer theory to vision research." *Vision research* 51.7 (2011)

Burgess, Arthur E.. "Visual signal detection. II. Signal-location identification." *JOSA A* 1.8 (1984):

Numerical Observer

- Anthropomorphic numerical observers (ANO) were found to be close to human in terms of performance.
- Some widely known ANOs are difference-of-gaussian channelized hoteling observer (DOG-CHO), filtered channelized observer (FCO).
- Recently some works have investigated convolutional neural network's performance as ANO.
- But CNNs are brittle and do not really capture human-interpretable features !
- How to introduce human-interpretable component in CNN-based NOs?

Diaz, Ivan, et al. "Derivation of an observer model adapted to irregular signals based on convolution channels." *IEEE transactions on medical imaging* 34.7 (2015): 1428-1435.

Study Outline

- Goals: To investigate the use of adversarially trained CNNs as human-interpretable numerical observers.

- Dataset
  - Doubiso CLB images background with FDA breastmass tool generated signal
  - noise model: Poisson-gaussian
  - Image size : 128 X 128

Dataset

- Task : SKE/BKS binary Signal Detection Task, Detection-localization task.
- Signal Model: Simulate mass lesion with spicules. Generated using FDA BreastMass Tool.
- Length-scale of the signal fixed same value with that of the background doubiso model.
- Cone-beam projection was used.
- Background Models: Advanced CLB Doubiso model.
- Signal insertion method: background*(background + intensity*signal).
- Noise model: Mixed Poisson-Gaussian

Castella, Cyril, et al. "Mammographic texture synthesis: second-generation clustered lumpy backgrounds using a genetic algorithm." *Optics express* 16.11 (2008): 7595-7607.

M. Ruschin, A. Tingberg, M. Brath, et al., “Using simple mathematical functions to simulate pathological structures—input for digital mammography clinical trial,” Radiation protection dosimetry 114(1-3), 424–431 (2005)

Doubiso Dataset Example Image

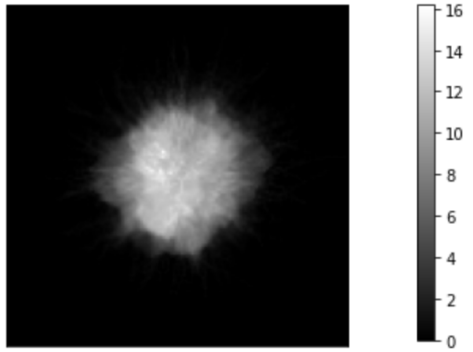


Fig: Signal Image

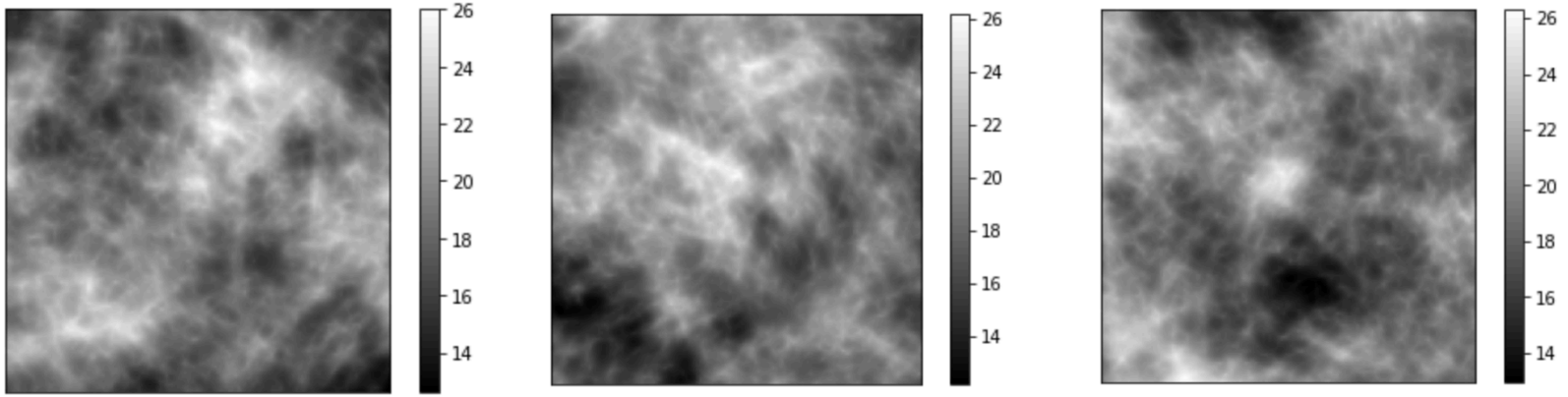
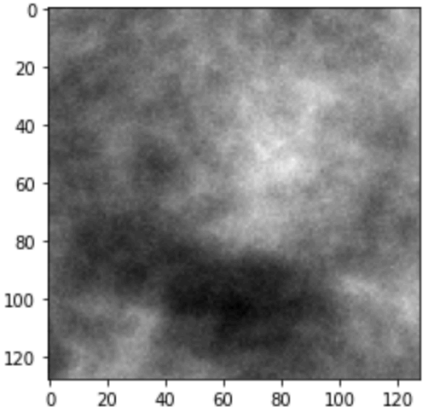
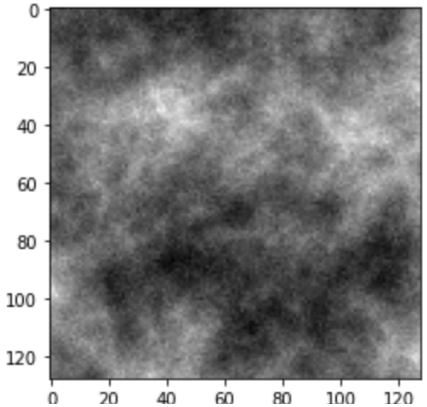


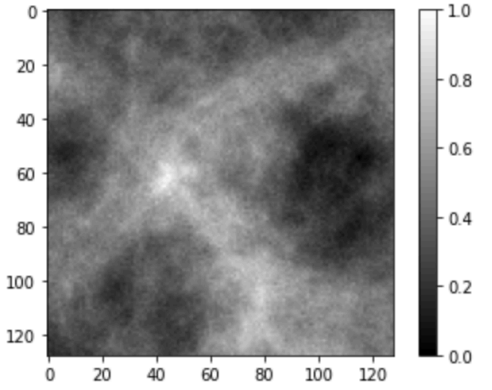
Fig: background images examples

SA/SP Example Image

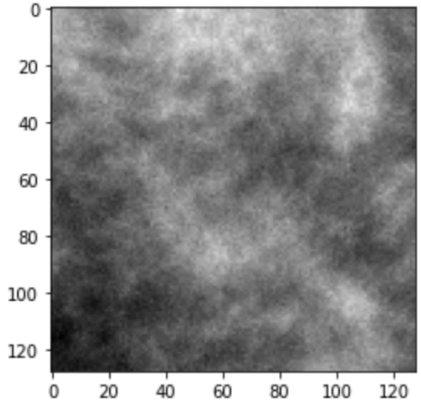
SP images

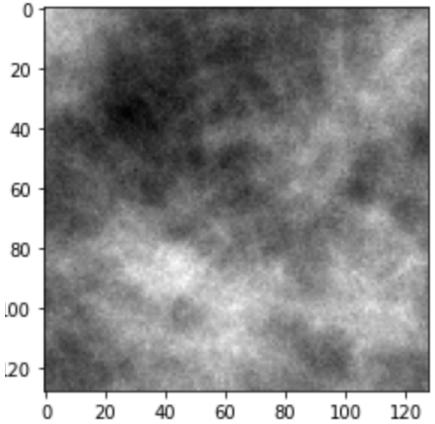


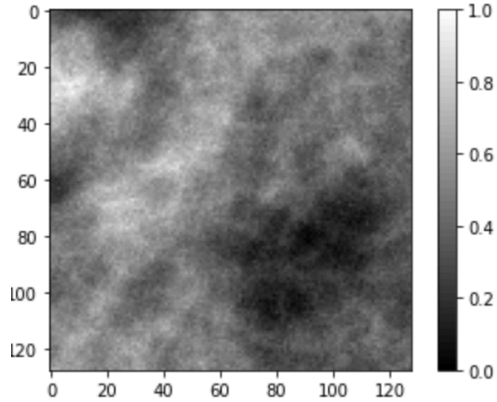




SA images







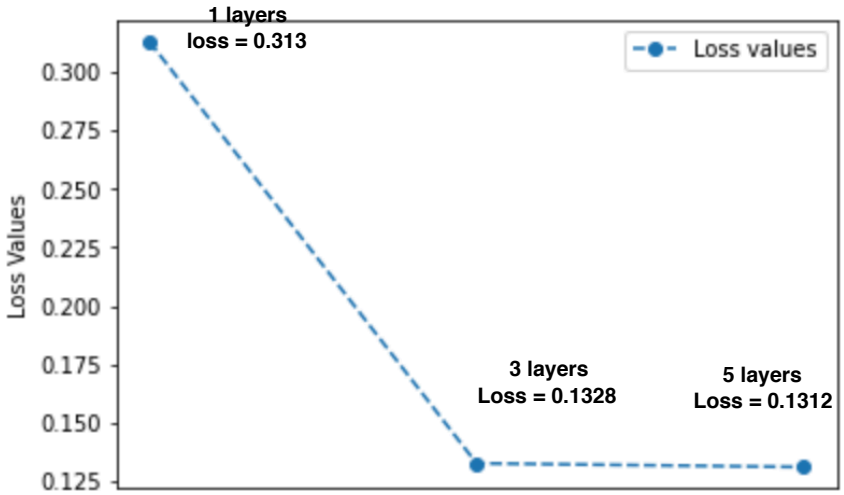
Studies

- Study 1 : Investigate CNN-IO architecture.
- Study 2 : Impact of epsilon value of robust model training : L_2 perturbation set.
- Study 3 : Impact of inner optimization steps of robust model training : L_2 perturbation set.
- Study 4 : Comparing performances with other anthropomorphic model observers: DOG-CHO, FCO.
- Study 5 : Exploring feature visualization by standard and robust network.

CNN-IO performance for the task

CNN-IO performance for the task

- Number of training images 400k
- There was no significant increase in AUC value after 300k training images. (200k SP + 200k SA)
- Online data augmentation was done.
- CNN with 1, 3 and 5 layers were explored and after 3 layers there was no significant increase in AUC value.
- 3 layer CNN can be termed as CNN-IO here.
- Test data- 10k images (5k + 5k)
- Validation data - 10k images (5k + 5k)
- Stopping Rule: If there is no decrease in validation loss for 5 consecutive epochs, the training stops and the weights corresponding to lowest validation loss are saved.
- CNN architecture : Conv+Activation—> Conv+Activation—> Conv+Activation—> Maxpool —> dense



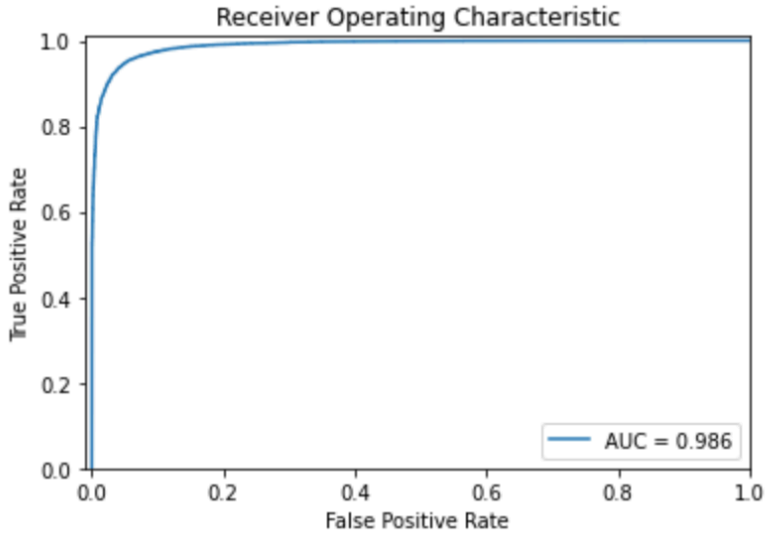


Fig: Results with Different No of CNN layers

Impact of epsilon value of robust model training in the classifiers' performance:

Impact of epsilon value of robust model training in the classifiers' performance:

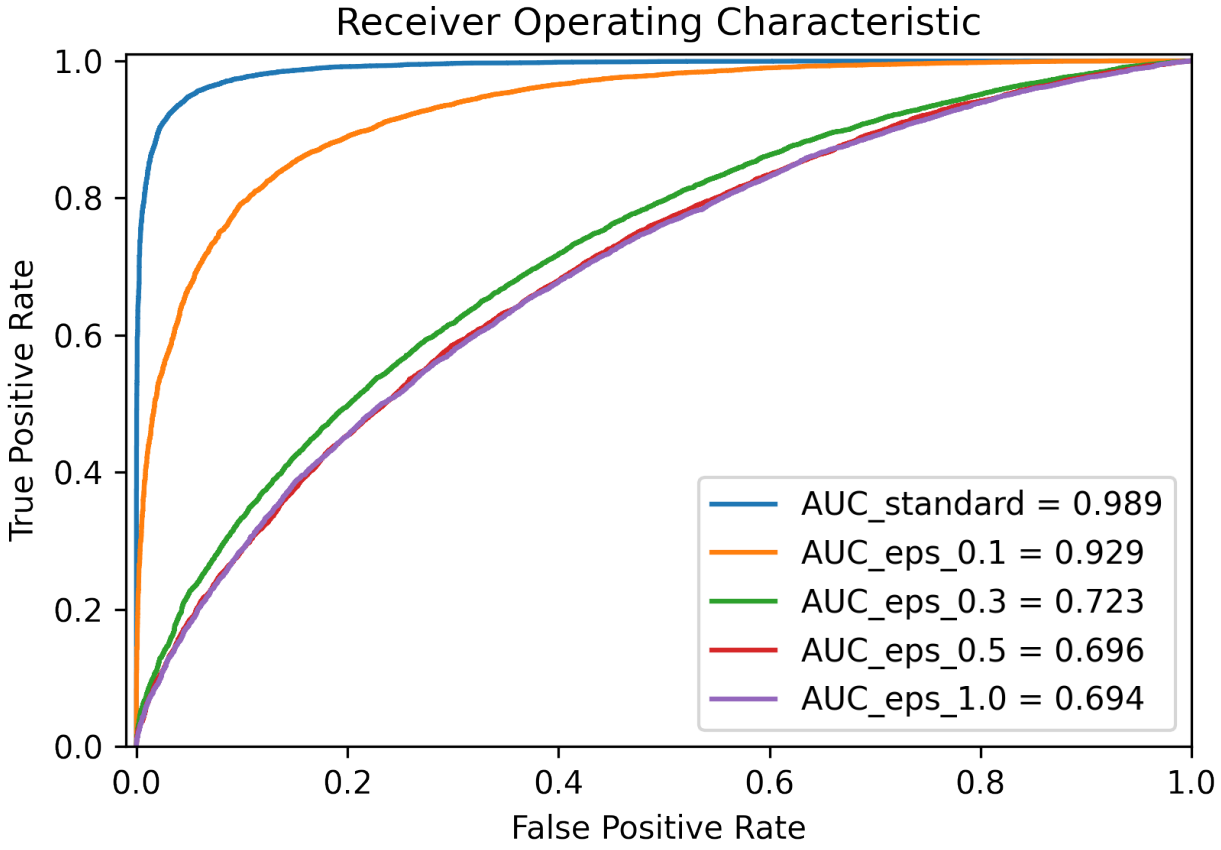
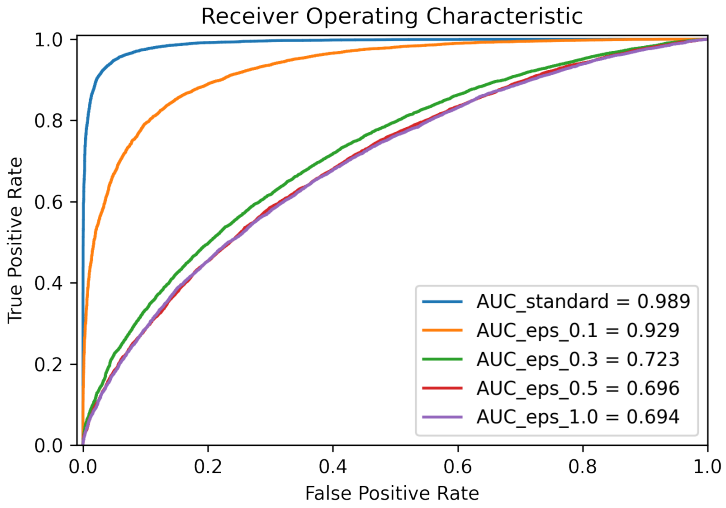


Fig: ROC Curves for standard and robust models- different epsilon values and fixed iterations = 5

Impact of epsilon value of robust model training in the classifiers' performance:

- Same CNN architecture which was used for CNN-IO
- Same training, validation, testing dataset
- Here the number of iterations' value was fixed at 5.
- The epsilon value was set to 0.1, 0.3, 0.5, 1.0.
- The ROC curves are plotted for these 4 robust models and also the standard model.
- It can be seen that AUC values are decreasing with increasing #epsilon.



Impact of inner optimization steps of robust model training in the classifiers' performance:

Impact of inner optimization steps of robust model training in the classifiers' performance:

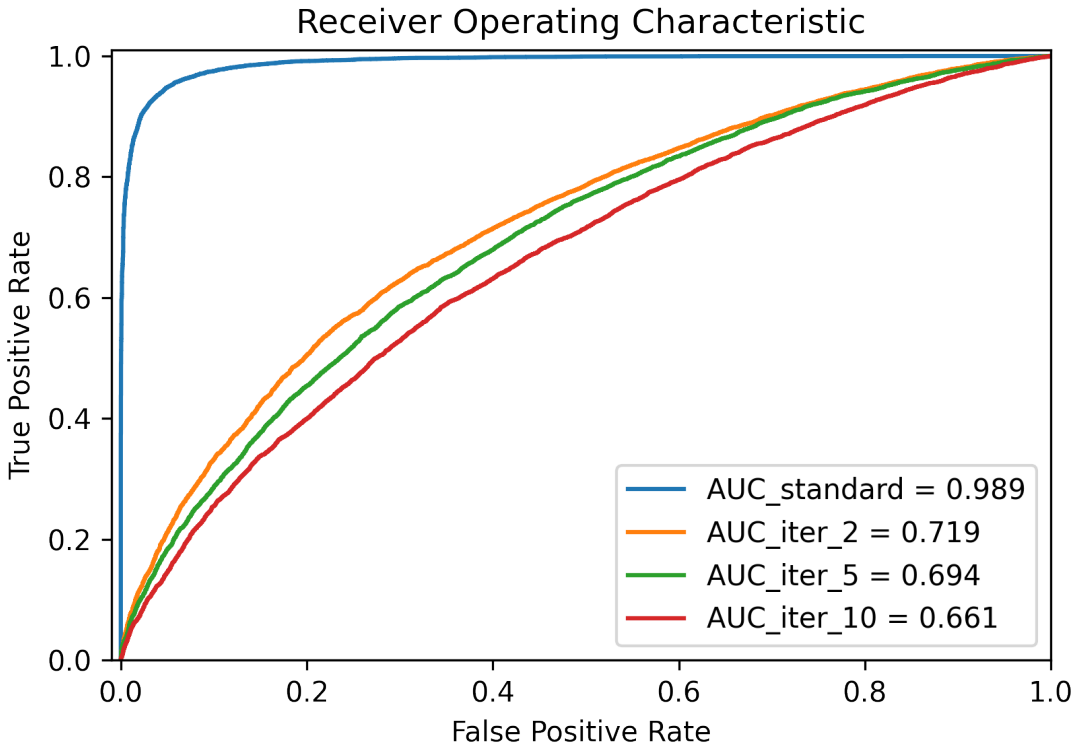
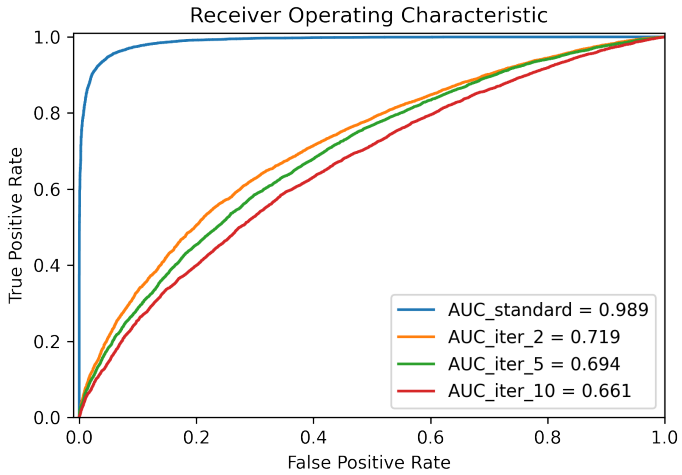


Fig: ROC Curves for standard and robust models for different iterations but fixed epsilon = 1

Impact of inner optimization steps of robust model training in the classifiers' performance:

- Same CNN architecture which was used for CNN-IO
- Same training, validation, testing dataset
- Here the epsilon value was fixed at 1.
- The number of iterations (of the inner optimization process) were set to 2, 5 and 10.
- The ROC curves are plotted for these 3 robust models and also the standard model.
- It can be seen that AUC values are decreasing with increasing #iterations.



Comparing performances with other anthropomorphic model observers: DOG-CHO, FCO

Comparing performances with other anthropomorphic model observers: DOG-CHO, FCO

- For comparison two anthropomorphic numerical observers were selected.
- DOG-CHO and FCO.
- The DOG-CHO is a widely used ANO and designed mainly for circularly symmetric signal.
- FCO is designed for signals with irregular shape.
- The templates for both are shown in the next slide.
- AUC values are computed and shown in slide #19. ROC curves for two other robust models with nearly equal AUC values are also shown together.

Comparing performances with other anthropomorphic numerical observers: DOG-CHO, FCO

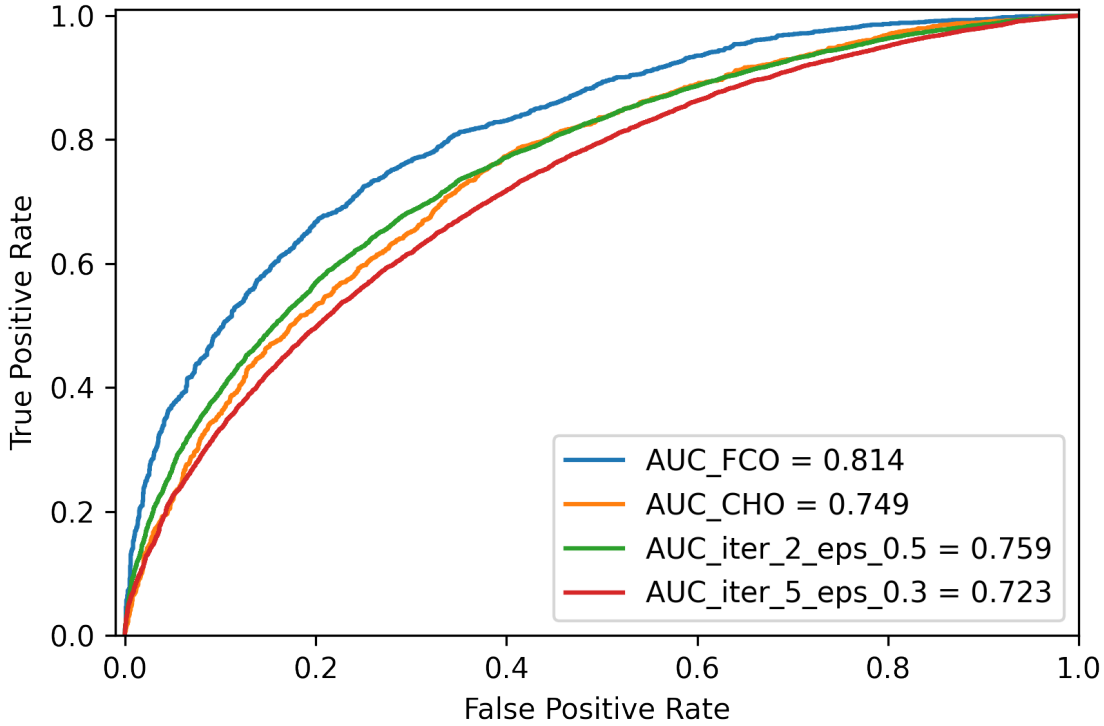


Fig: ROC Curves for DOG-CHO and FCO. Two robust model curves having nearly similar performances are also shown here

Exploring feature visualization by standard and robust network: Gradient Maps and Classification Images

Exploring feature visualization by standard and robust network: Gradient Maps

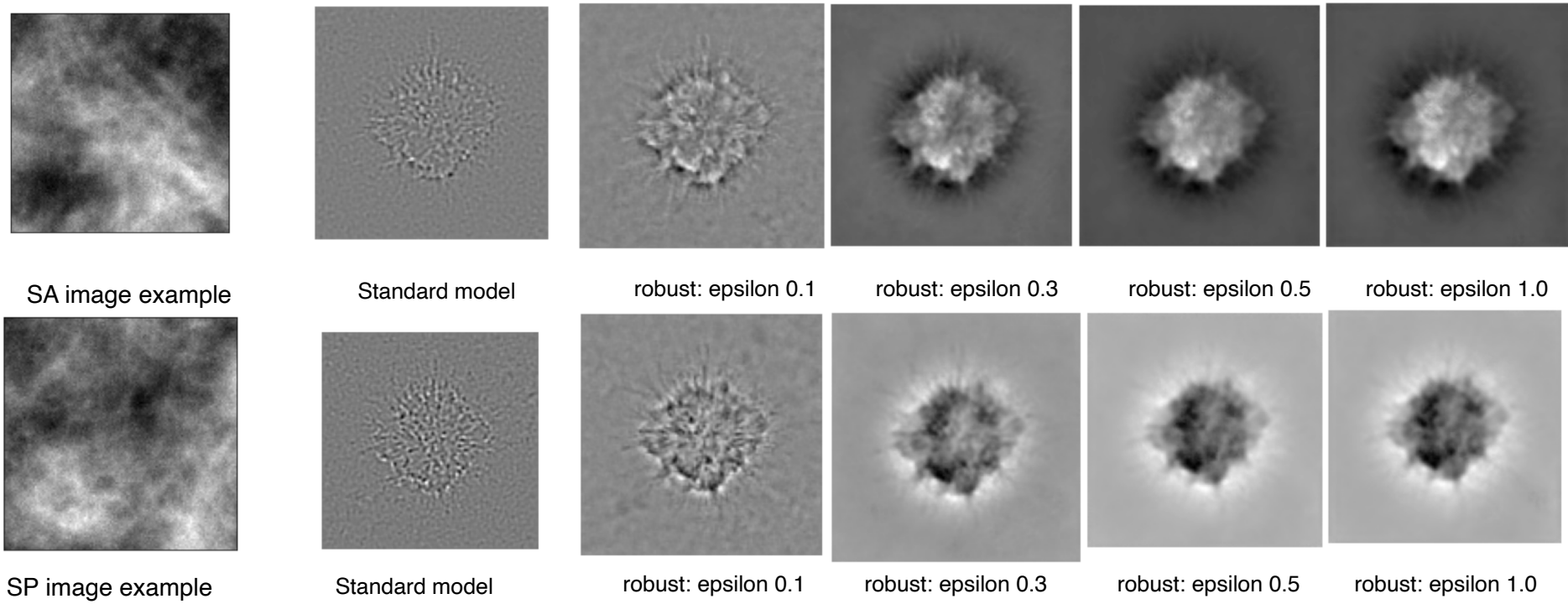
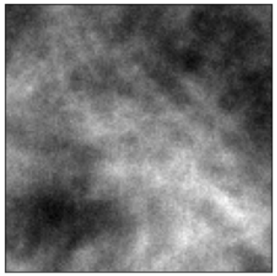


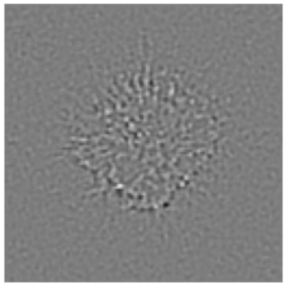
Fig: Gradient Maps for a SP and a SA image- Standard model and robust model with different epsilon values. (fixed inner optimization steps = 5)

- Take home point : Gradient Maps for robust model are human-interpretable than standard models

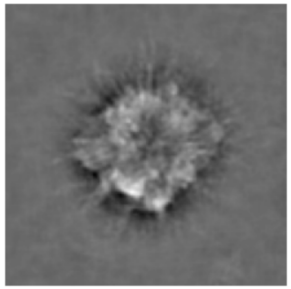
Exploring feature visualization by standard and robust network: Gradient Maps



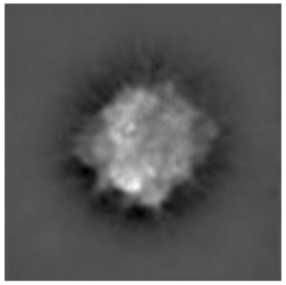
SA image example



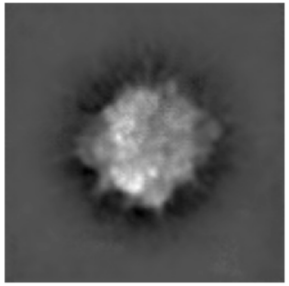
Standard model



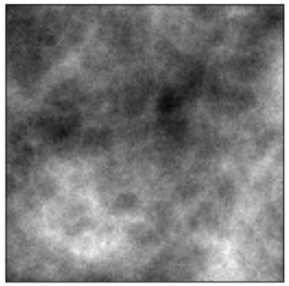
robust: iter 2



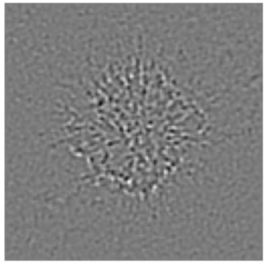
robust: iter 5



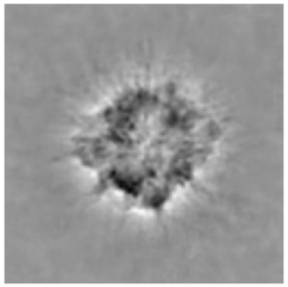
robust: iter 10



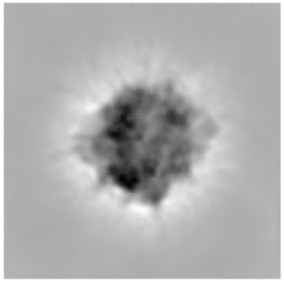
SP image example



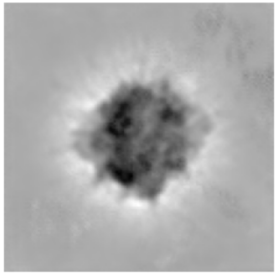
Standard model



robust: iter 2



robust: iter 5



robust: iter 10

Fig: Gradient Maps for a SP and a SA image- Standard model and robust model with different inner optimizatio steps (fixed epsilon = 0.5)

- Take home point : Gradient Maps for robust model are human-interpretable than standard models

Classification Image Computing Steps

- Take one image (say, signal-present image : size 128X128)
- Each time generate a white noise field of size 128X128
- Add the noise to the image
- Give the image input to the trained CNN model
- Calculate the probability outcome
- Do this for 10k noise fields and save the probability values.
- So we will have a 128X128X10000 and corresponding 10000 probability values
- Calculate the correlation at each pixel and visualize it.
- Hypothesis is: the visualization will be different between robust CNN and standard CNN.

Ringach, Dario, and Robert Shapley. "Reverse correlation in neurophysiology." *Cognitive Science* 28.2 (2004): 147-166.

Exploring feature visualization by standard and robust network: Classification Image

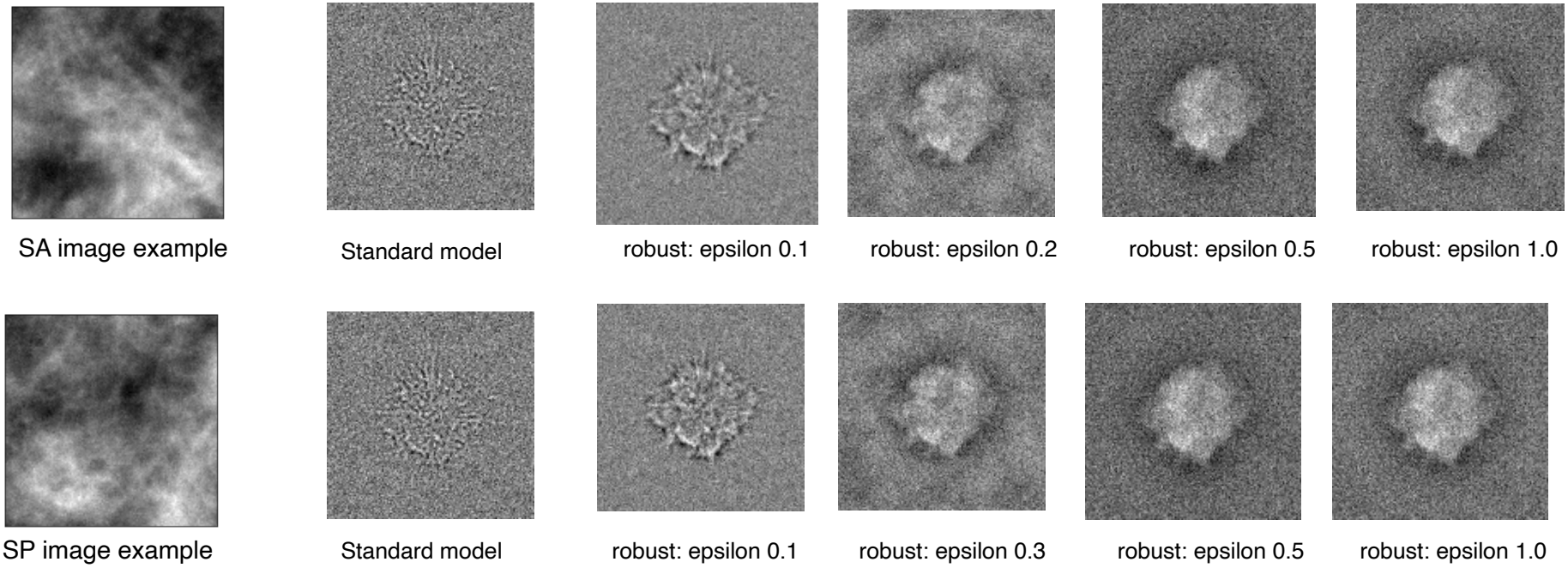
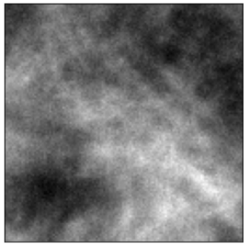


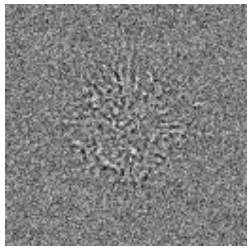
Fig: Classification images for a SP and a SA image- Standard model and robust model with different epsilon values (fixed inner optimization steps = 5)

- Take home point : Classification images for robust model are human-interpretable than standard

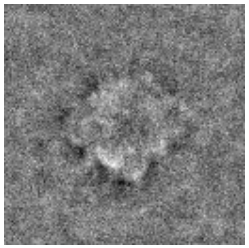
Exploring feature visualization by standard and robust network: Classification Image



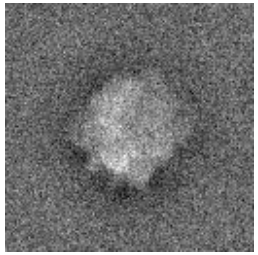
SA image example



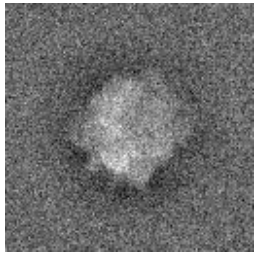
Standard model



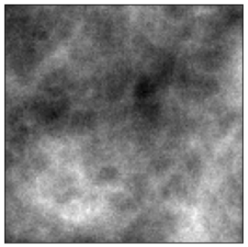
robust: iter 2



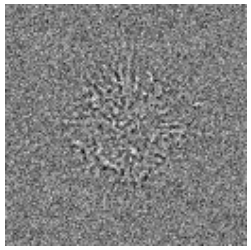
robust: iter 5



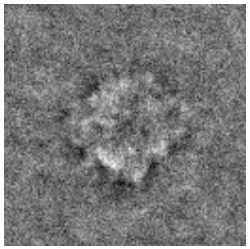
robust: iter 10



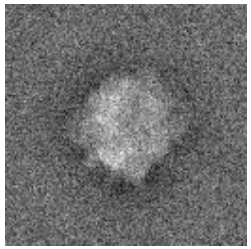
SP image example



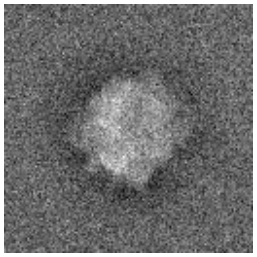
Standard model



robust: iter 2



robust: iter 5



robust: iter 10

Fig: Classification images for a SP and a SA image- Standard model and robust model with different inner optimization steps (fixed epsilon = 1.0)

- Take home point : Classification images for robust model are human-interpretable than standard

Conclusion

- The AUC value decreases with increasing epsilon and fixed inner optimization steps.
- The AUC value decreases with increasing inner optimizing steps and fixed epsilon.
- The gradient maps and classification images are human-interpretable than that of standard models.
- As baseline methods widely used DOG-CHO and FCO were computed. Robust models, having similar detection performances with DOG-CHO and FCO were also investigated.
- Adversarially trained CNNs hold great potential of being an anthropomorphic NO.

Thank you