# Distributional Detection

* Li and Li [1] use PCA to detect natural images from adversarial examples. They collect PCA coefficients for hidden layers and train a classifier to distinguish clean vs adversarial examples.
* Lu et al. [2] made a hypothesis that adversarial examples produce different patterns of ReLU activations in (the late stages of) networks. Based on this hypothesis, they insert RBF-SVM classifier before the last fc layer.A picture containing diagram

  Description automatically generated
* Metzen et al. [3] propose to augment the network with a binary adversarial detector at every inner layer of network.

Diagram, schematic

Description automatically generated

* Gong et al. [4] simply train a binary classifier to distinguish clean inputs and adversarial inputs
* Grosse et al. [5] augment the learning model with an additional outlier class. The model detects the adversarial examples by classifying it as an outlier.
* Feinman et al. [6] present a defense they call kernel density estimation. They use a Gaussian Mixture Model to model outputs from the final hidden layer of a neural network, and argue that adversarial examples belong to a different distribution than that of natural images.
* Feinman et al. [6] propose a second detection method called Bayesian neural network uncertainty that measures the uncertainty of the neural network on the given input. In a network with dropout, adversarial examples will get unstable classification results every time dropout is randomized. Whereas clean examples should get consistent predictions even though different neurons are dropout.
* Magnet [7] trains an autoencoder on normal images. Large reconstruction error is expected for adversarial images because they are not used in training. Then it compares the Jensen-Shannon divergence between the classification result for the test input and the classification result for the test input after autoencoder’s reconstruction.

## References

[1] Li, Xin, and Fuxin Li. "Adversarial examples detection in deep networks with convolutional filter statistics." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[2] Lu, Jiajun, Theerasit Issaranon, and David Forsyth. "Safetynet: Detecting and rejecting adversarial examples robustly." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[3] Metzen, Jan Hendrik, et al. "On detecting adversarial perturbations." arXiv preprint arXiv:1702.04267 (2017).

[4] Gong, Zhitao, Wenlu Wang, and Wei-Shinn Ku. "Adversarial and clean data are not twins." *arXiv preprint arXiv:1704.04960*(2017).

[5] Grosse, Kathrin, et al. "On the (statistical) detection of adversarial examples." *arXiv preprint arXiv:1702.06280*(2017).

[6] Feinman, Reuben, et al. "Detecting adversarial samples from artifacts." *arXiv preprint arXiv:1703.00410* (2017).

[7] Meng, Dongyu, and Hao Chen. "Magnet: a two-pronged defense against adversarial examples." *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 2017.