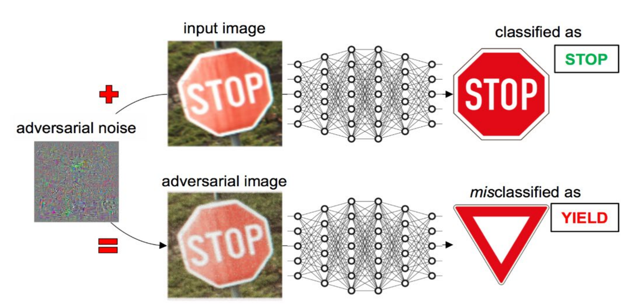
RANDOM ENSEMBLE OF LOCALLY OPTIMUM DETECTORS FOR DETECTION OF ADVERSARIAL EXAMPLES

Amish Goel, Pierre Moulin, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign

ADVERSARIAL MACHINE LEARNING

- Recent works have shown a significant vulnerability of machine learning based classifiers: an adversary can construct an input that resembles legitimate input but is incorrectly recognized by classifier.



Goal : Design a defense method against the adversarial attacks to linear classifiers.

Adversarial Model (h, ϵ, t):

- Adversary adds a perturbation along some specific direction (h) such that the input image is misclassified.
- Adversary is constrained by maximum distortion (ϵ)
- Adversary uses **Fast Gradient Sign Method** (FGSM) but can additionally choose target (t) and maximizes the probability of a particular target class. Overall, the adversarial output is given as,

$$\tilde{x} = x + \epsilon h$$

SYSTEM MODEL AND NOTATION

- Consider M-ary classifier. Let output probabilities for a sample x denoted by $P(y|x)$. Classifier's decision is $\Psi(x) = \underset{y \in Y}{\operatorname{argmax}} P[y|x]$

Detection Method :
View perturbation as a watermark and apply hypothesis testing to detect the adversary.

- Watermarks are weak signals added to content to trigger a positive response by watermark detector.
- Watermark detectors are used for protecting content against adversaries. Here, we are doing the opposite.
- $\delta(x)$: detector's output;
  $\delta(x) = 1$ if forgery, 0 otherwise.
- Events of interest:
  (1) (Undetectability) Undetected forgery: $\delta(\tilde{x}) = 0$
  (2) (Utility) Successful forgery: $\Psi(\tilde{x}) \neq y$

Adversary aims to achieve both goals, but for (1) it needs small ϵ, and for (2) it needs larger ϵ.

DEFENSE METHOD

- $p_\epsilon(x)$: PDF of adversarially perturbed examples; for $\epsilon = 0$, $p_0(x)$ denotes data distribution.

Assuming small ϵ, we use Locally Optimum (LO) testing to motivate the detector.

- Consider Neyman-Pearson (NP) hypothesis testing to maximize detection probability P_D given a false alarm rate constraint $P_F \leq \alpha$ and a target class t.
- NP test reduces to LO test as $\epsilon \to 0$, which is limiting form of a Likelihood Ratio Test (LRT):

$$T_t(x) = \frac{\frac{\partial}{\partial \epsilon} p_\epsilon(x; h_t)|_{\epsilon=0}}{p_0(x)}.$$

- This is the statistic for a specific target t. For unknown t we can use a LO version of the Generalized Likelihood Ratio Test (GLRT), estimating the most likely target giving statistic :

$$\delta(x) = \max_{t \in \mathcal{Y}} \frac{\frac{\partial}{\partial \epsilon} p_\epsilon(x; h_t)|_{\epsilon=0}}{p_0(x)} > \gamma.$$

Detector : Gaussian Mixture Model (GMM) and Random Ensemble (k, m ,L):

Need tractable model for learning the distribution $p_0(x)$ and substituting in GLRT

- Use GMM model for small image patches. Compute average statistic over a random ensemble of patches extracted from image.
- k : Number of components of the GMM model
- μ_c, Σ_c : Mean vectors and Covariance matrix for each component $c \in \{1,2,\ldots,k\}$
- S_l : Mask for l^{th} patch sampled from a random location on the image x.
- Our LO test statistic for S_l is then given by

$$T(x, S_l, t) = \sum_{c=1}^{k} p(c|S_l \cdot x) \left[(S_l \cdot h_t)^T \Sigma_c^{-1} (S_l \cdot x - \mu_c) \right]$$

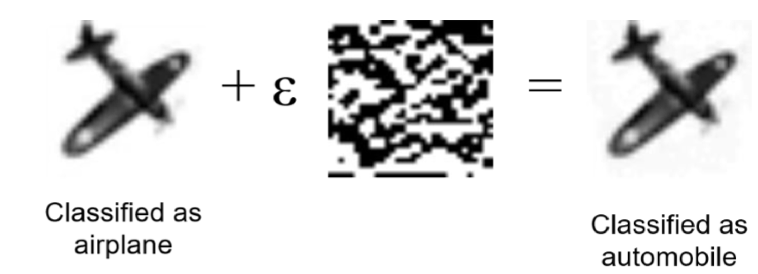
- Using L random patches, the overall statistic computed from the image for a target t is given by

$$T_t(x) = \frac{1}{L} \sum_{l=1}^{L} T(x, S_l, t).$$

- The overall detection statistic is given by:

$$\delta(x) = \max_{t \in Y} T_t(x) > \gamma$$

EXPERIMENTS AND RESULTS



- Detection performance for different values of k and $m = 16$, L = 30 is illustrated in the figure below.

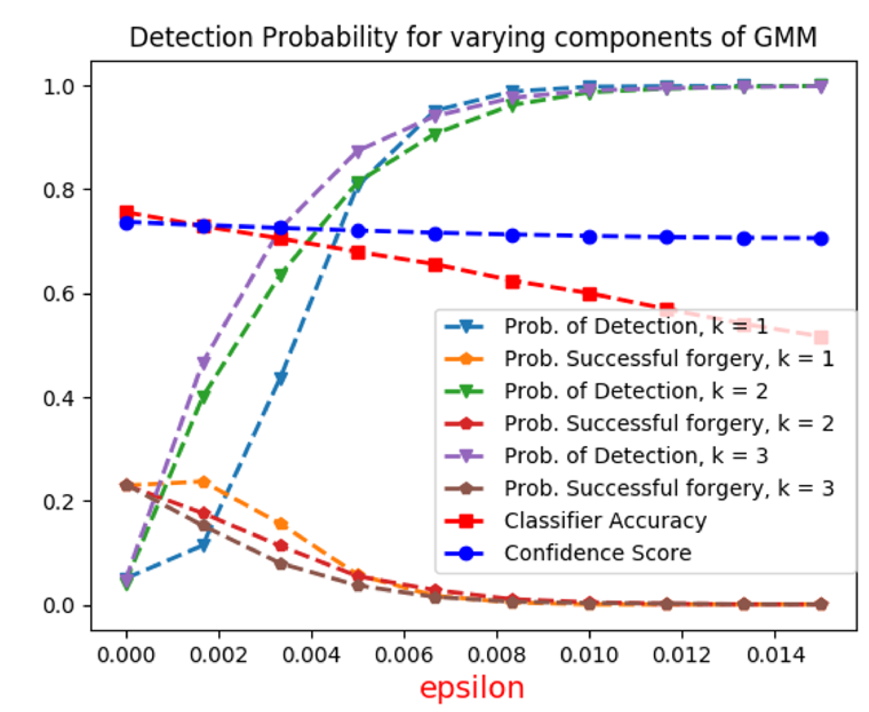


Fig. 1 *Detection performance for various values of k, the number of GMM components. The red and blue curves show change in accuracy and confidence of the classifier. Observe that for smaller ϵ, detectors with $k > 1$ have much higher detection rate, than for $k = 1$ (Gaussian)*

- We also experiment with the patch size m and the number of patches L and illustrate the detection performance using Receiver Operating Characteristics.
- For smaller patch sizes, we would need to sample more patches in order to have enough information about the image. As a heuristic, for an image of size $I \times I$, and patch $m \times m$, we randomly sample about 10% of total $(I - m + 1)^2$ possible patches.

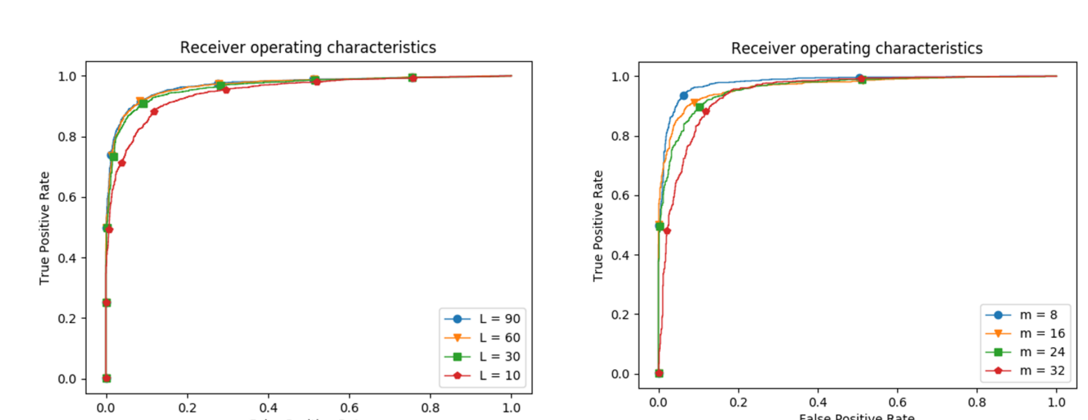


Fig. 2 *ROCs for different values of L and m. For Left-fig., we fix $m = 16, k = 3$. Here, we observe that $L = 10$ discards too much data, while $L \geq 30$ may cause redundancy. For Right-fig., we fix $k = 3$ and simultaneous vary m, L as $m \in \{8,16,24,32\}$ and $L \in \{60,30,7,1\}$. Plot indicates higher detection performance for smaller m, likely due to more accurate estimation of GMM parameters.*

EXPERIMENTS AND RESULTS

- Used CIFAR10 dataset which consists of 60000 color images of size 32×32 divided into 10 classes. Pixel values are normalized to lie in the interval [0,1].
- Trained Logistic classifier for binary classification – airplane vs automobile, gives error rate of 75% and prediction confidence of 77%.

CONCLUSION AND FUTURE WORK

- Proposed detection scheme works well in weak perturbation scenarios.
- Detector has several tunable hyperparameters and evaluates a randomized statistic. This potentially provides more robustness against a white box adversary.
- We are currently studying how much an attacker can gain if he knows the patches in advance (full white box attack).

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

[3] P. Moulin and A. Goel, "Locally optimal detection of adversarial inputs to image classifiers," in Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference, July 2017, pp. 459–464.

[4] P. Moulin and V. Veeravalli, Statistical Inference for Engineers and Data Scientists. Cambridge University Press, 2018.

[5] https://www.pluribus-one.it/sec-ml/wild-patterns

ILLINOIS