Attack Agnostic Statistical Method for Adversarial Detection

Sambuddha Saha, Aashish Kumar, Pratyush Sahay, George Jose, Srinivas Kruthiventi, Harikrishna Muralidhara - Harman International India Pvt. Ltd., Bangalore

INTRODUCTION

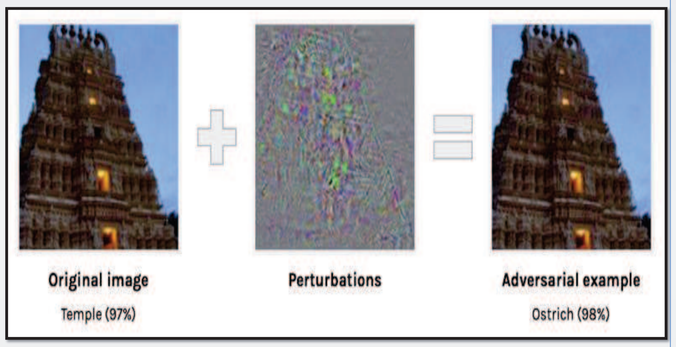


Figure 1: Small perturbations added to original image to fool the classifier

BACKGROUND

Adversarial Attacks:

- FGSM:[1] $\Delta x = \epsilon . sign(\nabla_x C(x,y))$
- Madry :[2] $x^{t+1} = \prod_{x+\mathcal{S}}\left(x^t + \epsilon . sign(\nabla_x C(\theta, x, y))\right)$
- Carlini Wagner L2:[3]

$$min \parallel \frac{1}{2}(tanh(w)+1) - x \parallel_2^2 + c.f\left(\frac{1}{2}(tanh(w))+1\right)$$

$$with\ f\ defined\ as$$

$$f(x') = max\left(max\{Z(x')_i : i \neq t\} - Z(x')_{t,} - \kappa\right)$$

Statistical Metrics:

- MMD:[4]

$$MMD_b[F, X_1, X_2] = sup_{f \in F}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_{1i}) - \frac{1}{m}\sum_{i=1}^{m} f(x_{2i})\right)$$

- ED: [5]

$$D^2(F,G) = 2E \parallel X - Y \parallel - E \parallel X - X' \parallel - E \parallel Y - Y' \parallel$$

HYPOTHESIS

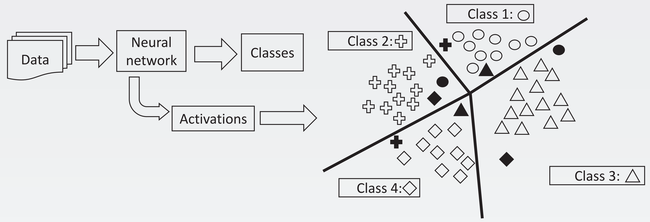


Figure 2: Hypothesis: Original and adversarial images belong to different distributions

- Distribution of original and adversarial activations for a particular class are different
- Statistical Distances like MMD, ED can distinguish between original and adversarial activations distribution
- Statistical distances are agnostic to the type of attack

EXPERIMENTAL SETUP

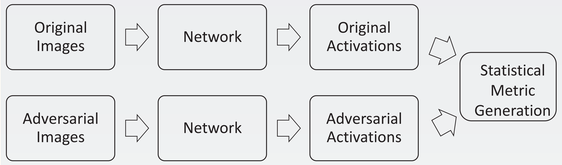


Figure 3: Brief Overview of methodology

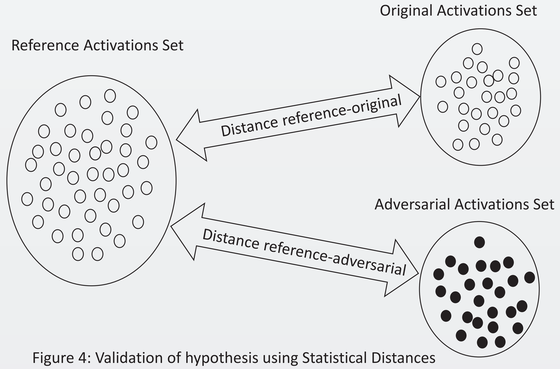
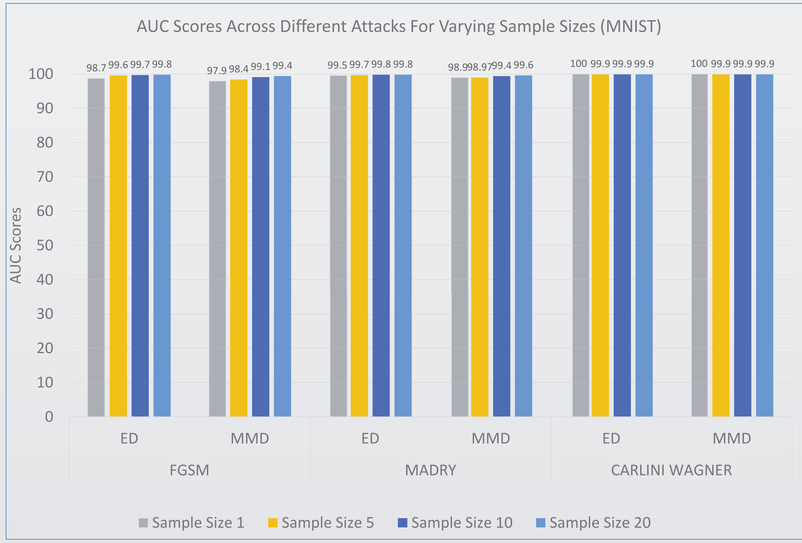
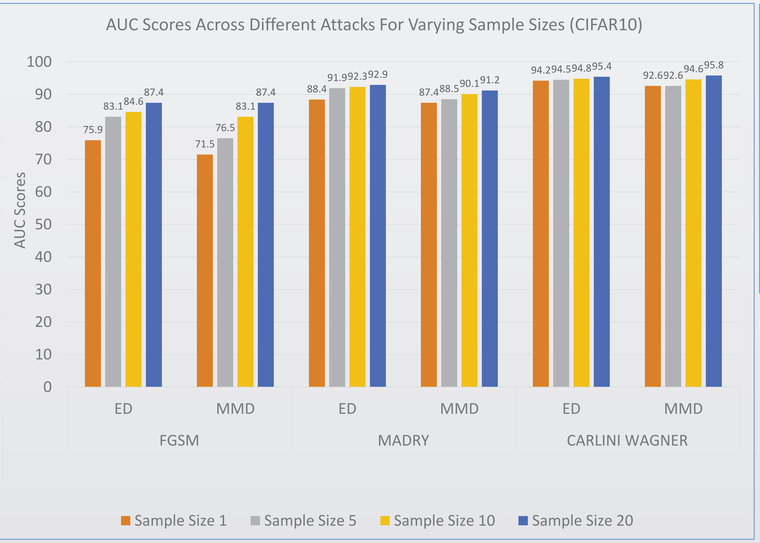


Figure 4: Validation of hypothesis using Statistical Distances

- Distance pairs computed for each class
- AUC scores computed from these distance pairs
- Experiment done for varying sample sizes of test set, different attacks and degree of perturbations for each attack

RESULTS





CONCLUSION

- The approach is sample efficient
  - High AUC scores obtained using ED and MMD
  - AUC scores increase with increase in sample size of the test set (works well with sample size 1)
- The approach is attack agnostic
  - AUC scores vary negligibly w.r.t
    - Attack type
    - Degree of perturbations

REFERENCES

1. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
2. Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.
3. Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).
4. Gretton, Arthur, et al. "A kernel two-sample test." *Journal of Machine Learning Research* 13.Mar (2012): 723-773.
5. Székely, Gábor J., and Maria L. Rizzo. "Energy statistics: A class of statistics based on distances." *Journal of statistical planning and inference* 143.8 (2013): 1249-1272.