# Adversarial Attack and Detection under the Fisher Information Metric

Presenter: Chenxiao Zhao[1]
Joint work with Tom Fletcher[2], Mixue Yu[1],
Yaxin Peng[3], Guixu Zhang[1], Chaomin Shen[1]

[1]Dept of Computer Science, East China Normal University, China
[2]Dept of Computer Science, Univ. of Virginia, USA
[3]Dept of Mathematics, Shanghai University, China

January 30, 2019

# What do we know about adversarial examples?

- Some imperceptible noise added on the input can alter the output prediction[1]


[1] I. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and harnessing adversarial examples". In: *ArXiv preprints arXiv:1412.6572* (2014).

# Characterizing the vulnerability of deep learning models

How to measure the vulnerability of a deep learning model?

- Worst case perturbation ⇒ adversarial training[2]
- Density[3] / model uncertainty / topological dimension[4] ⇒ adversarial detection

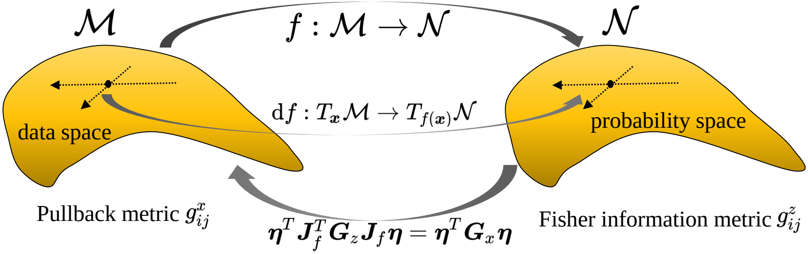
---

[2] A. Sinha, H. Namkoong, and J. Duchi. "Certifying some distributional robustness with principled adversarial training". In: *ArXiv preprints arXiv:1710.10571* (2017).

[3] J. Metzen et al. "On detecting adversarial perturbations". In: *ArXiv preprints arXiv:1702.04267* (2017).

[4] X. Ma et al. "Characterizing adversarial subspaces using local intrinsic dimensionality". In: *ArXiv preprints arXiv:1801.02613* (2018).

# The Fisher information metric approach



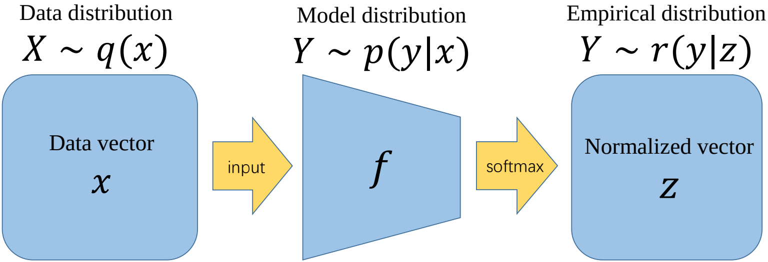
# Objective function

For adversarial attacks, the goal is to find a subtle perturbation $\eta$ for a given input, such that the output prediction varies from the the correct to the wrong output.

$$\max_{\boldsymbol{\eta}} \boldsymbol{\eta}^T \boldsymbol{G}^{\boldsymbol{x}} \boldsymbol{\eta} \qquad \text{s.t. } \|\boldsymbol{\eta}\|_2^2 = \epsilon$$

- The optimal solution for $\boldsymbol{\eta}$ is the greatest eigenvector of matrix $G^x$
- But how do we define the metric tensor $g^x$?

# FIM of the input samples



# Fisher information

**Definition (Fisher information)**

*Let $p(x|\theta)$ be a probability density function of random variable $X$ conditioned on parameter $\theta$. The Fisher information matrix of $\theta$, denoted as $G^{\theta}$, is defined as the variance of the expectation over the derivative of log-likelihood with respect to $\theta$:*

$$\boldsymbol{G}_{ij}^{\boldsymbol{\theta}} = \mathbb{E}_{x|\boldsymbol{\theta}}[(\frac{\partial}{\partial\theta_i}\log p(x|\boldsymbol{\theta}))(\frac{\partial}{\partial\theta_j}\log p(x|\theta))^T]$$

Many theoretical benefits in[5]

[5] S. Amari and H. Nagaoka. *Methods of Information Geometry*. Providence, RI: American Mathematical Society, 2007.

# FIM of the input samples

For adversarial attacks, the input $x$ is the only changeable variable. With some exchange of variables we obtain

$$\boldsymbol{G}_{ij}^{\boldsymbol{x}} = \mathbb{E}_{y|\boldsymbol{x}}[(\frac{\partial}{\partial x_i}\log p(y|\boldsymbol{x}))(\frac{\partial}{\partial x_j}\log p(y|\boldsymbol{x}))^T]$$

What is $p(y|\boldsymbol{x})$ here?

- True model distribution $p(y|\boldsymbol{x})$ (like Gaussian or sth)
- Empirical distribution $r(y|f(\boldsymbol{x}))$ (the output of the model)

# FIM of the input samples

How to compute the matrix $\boldsymbol{G}^{\boldsymbol{x}}$?

- Using the Jacobian $\boldsymbol{J}_f$ of the network $f : \mathcal{X} \to \mathcal{Z}$.[6]

$$
\begin{aligned}
G^{x} &= \boldsymbol{J}_f^T \mathbb{E}_{y|f(\boldsymbol{x})}[(\frac{\partial}{\partial \boldsymbol{z}} r(y|\boldsymbol{z}))(\frac{\partial}{\partial \boldsymbol{z}} r(y|\boldsymbol{z}))^T]\boldsymbol{J}_f \\
&= \boldsymbol{J}_f^T \boldsymbol{G}^{\boldsymbol{z}} \boldsymbol{J}_f
\end{aligned}
$$

- Given $\eta$ as the adversarial perturbation, a general approach is to compute the Hessian of the KL divergence.[7]

$$
\boldsymbol{G}_{ij}^{\boldsymbol{x}} = \mathbb{E}_{y|f(\boldsymbol{x})}[\frac{\partial^2}{\partial \eta_i \partial \eta_j} D_{KL}(p(y|\boldsymbol{x})||p(y|\boldsymbol{x}+\boldsymbol{\eta}))]
$$

[6] Hyeyoung Park, S-I Amari, and Kenji Fukumizu. "Adaptive natural gradient learning algorithms for various stochastic models". In: *Neural Networks* 13.7 (2000), pp. 755–764.

[7] Takeru Miyato et al. "Virtual adversarial training: a regularization method for supervised and semi-supervised learning". In: *IEEE transactions on pattern analysis and machine intelligence* (2018).

# FIM of the input samples

- How can we calculate FIM more efficiently?
- We use empirical distribution to compute the FIM with its original form[8]

$$
\begin{aligned}
\boldsymbol{G}_{ij}^{\boldsymbol{x}} &= \mathbb{E}_{y|\boldsymbol{z}}[(\frac{\partial}{\partial x_i}\log r(y|f(\boldsymbol{x})))(\frac{\partial}{\partial x_j}\log r(y|f(\boldsymbol{x})))^T] \\
&= \sum_{k=1}^{n} r_k(y|\boldsymbol{z})[(\frac{\partial}{\partial x_i}\log r_k(y|f(\boldsymbol{x})))(\frac{\partial}{\partial x_j}\log r_k(y|f(\boldsymbol{x})))^T]
\end{aligned}
$$

[8] James Martens. "New insights and perspectives on the natural gradient method". In: *arXiv preprint arXiv:1412.1193* (2014).

# Why empirical distribution?

What are the advantages for using the empirical distribution instead of true model distribution?

- Easy to compute, provided that one is already calculating the gradient

$$\boldsymbol{G}^{\boldsymbol{x}} = \sum_{i=1}^{n} r_i(y|f(\boldsymbol{x}))[(\frac{\partial}{\partial \boldsymbol{x}} \log r_i(y|f(\boldsymbol{x})))(\frac{\partial}{\partial \boldsymbol{x}} \log r_i(y|f(\boldsymbol{x})))^T]$$

- More optimization tricks to accelerate the computing process

$$\boldsymbol{\eta}^T \boldsymbol{G}^{\boldsymbol{x}} \boldsymbol{\eta} = \mathbb{E}_{y|f(\boldsymbol{x})}[(\boldsymbol{\eta}^T(\frac{\partial}{\partial \boldsymbol{x}} \log r(y|f(\boldsymbol{x}))))^2]$$

# Fisher information matrix on large datasets

Problems on large datasets

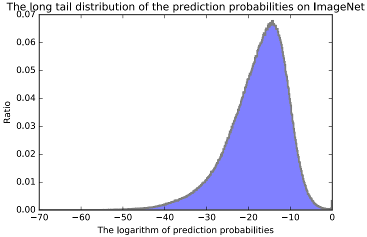
- Avoid the direct access to the explicit form of the matrix
  **Solution:**

$$\boldsymbol{\eta} \leftarrow \boldsymbol{G}^{\boldsymbol{x}}\boldsymbol{\eta} = \mathbb{E}_{y|f(\boldsymbol{x})}[((\frac{\partial}{\partial \boldsymbol{x}} \log r(y|f(\boldsymbol{x})))^T \boldsymbol{\eta})(\frac{\partial}{\partial \boldsymbol{x}} \log r(y|f(\boldsymbol{x})))]$$

- For datasets with large number of classes, e.g., ImageNet, compute the expectation more efficiently
  **Solution:** Monte Carlo sampling from $r(y|f(\boldsymbol{x}))$

# Fisher information matrix on large datasets

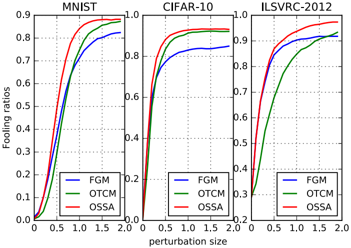
Output log-probabilities for a ResNet model.



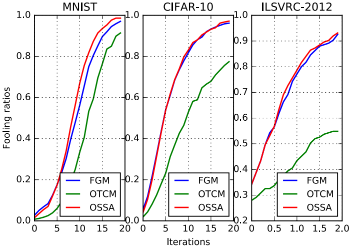

Empirically, about $\frac{1}{5}$ times of sampling, with alias method[9].

[9] G. Marsaglia, W. W. Tsang, and J. Wang. "Fast generation of discrete random variables". In: *Journal of Statistical Software* 11.3 (2004), pp. 17–24.

# Fisher information matrix on large datasets



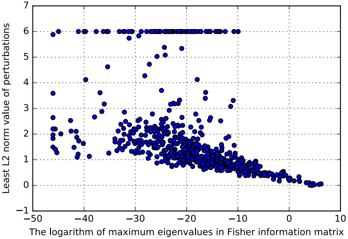

(a) One-step attack



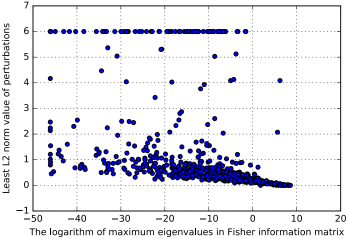

(b) Iterative attack

# Empirical evidence

Visualizing the vulnerability measured by the eigenvalues of FIM



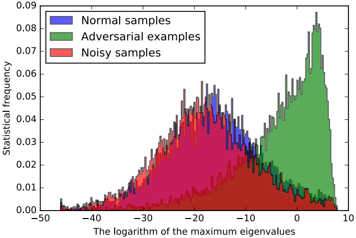

(c) MNIST



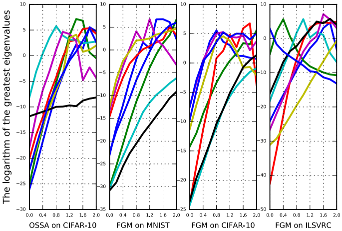

(d) CIFAR-10

# Empirical evidence

Why is it practical to distinguish the adversarial examples via the eigenvalues of Fisher information matrix?




(e) statistical histogram of Fisher information matrix eigenvalues




(f) increasing of eigenvalues along the perturbation direction

# Adversarial detection

**Key idea**: using an auxiliary classifier to distinguish the adversarial examples with the eigenvalues of FIM serving as characteristics.

**Other practical techniques**

- The logarithm of the eigenvalues as the features
- Use Lanczos algorithm to calculate a group of eigenvalues[10]
- The positive training set is composed of both normal samples and noisy samples[11]

[10] D. Calvetti, L. Reichel, and D. C. Sorensen. "An implicit restarted Lanczos method for large symmetric eigenvalue problems". In: *Electronic Transactions on Numerical Analysis* 2 (1994), pp. 1–21.

[11] A. Fawzi, M. Seyed D. Moosavi, and P. Frossard. "Robustness of classifiers: From adversarial to random noise". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1632–1640.

# Evaluations

Table: The AUC scores of detecting adversarial attacks using random forest classifiers and eigenvalues of FIM as characteristics

[12][13]

| | **MNIST** | | | | |
|---|---|---|---|---|---|
| AUC (%) | FGM | OTCM | Opt | BIM | OSSA |
| KD | 78.12 | 95.46 | 95.15 | 98.61 | 84.24 |
| BU | 32.37 | 91.55 | 71.30 | 25.46 | 74.21 |
| KD+BU | 82.43 | 95.78 | 95.35 | 98.81 | 85.97 |
| Ours | **96.11** | **98.47** | **95.67** | **99.10** | **93.13** |

[12] R. Feinman et al. "Detecting adversarial samples from artifacts". In: *ArXiv preprints arXiv:1703.00410* (2017).

[13] Y. Liu et al. "Delving into transferable adversarial examples and black-box attacks". In: *ArXiv preprints arXiv:1611.02770* (2016).

# Evaluations

Table: The AUC scores of detecting adversarial attacks using random forest classifiers and eigenvalues of FIM as characteristics

| | **CIFAR-10** | | | | |
|---|---|---|---|---|---|
| AUC (%) | FGM | OTCM | Opt | BIM | OSSA |
| KD | 64.92 | 92.13 | 91.35 | 98.70 | 88.89 |
| BU | 70.40 | 91.93 | 91.39 | 97.32 | 87.44 |
| KD+BU | 76.40 | **94.45** | 93.77 | 98.90 | 93.54 |
| Ours | **80.18** | 93.68 | **99.45** | **99.43** | **98.01** |

# Generalization ability and bad case analysis

Generalizes well on $\ell_2$ norm attacks but failed to generalize to $\ell_0$

| AUC (%) | Tested on | | | | | |
|---|---|---|---|---|---|---|
| **Trained on** | FGM | OTCM | Opt | BIM | OSSA | **JSMA** |
| FGSM | 93.44 | 90.19 | 90.45 | 91.06 | 89.97 | 75.35 |
| OTCM | 98.55 | 98.96 | 98.26 | 97.78 | 98.57 | 70.12 |
| Opt | 95.18 | 95.30 | 96.90 | 97.15 | 96.11 | 68.78 |
| BIM | 98.10 | 96.00 | 97.09 | 98.57 | 96.35 | 57.86 |
| OSSA | 91.17 | 91.47 | 89.77 | 89.47 | 89.67 | 65.40 |
| **JSMA** | 40.99 | 58.46 | 50.11 | 60.23 | 50.18 | 49.88 |

# Thank you!

51174506043@stu.ecnu.edu.cn