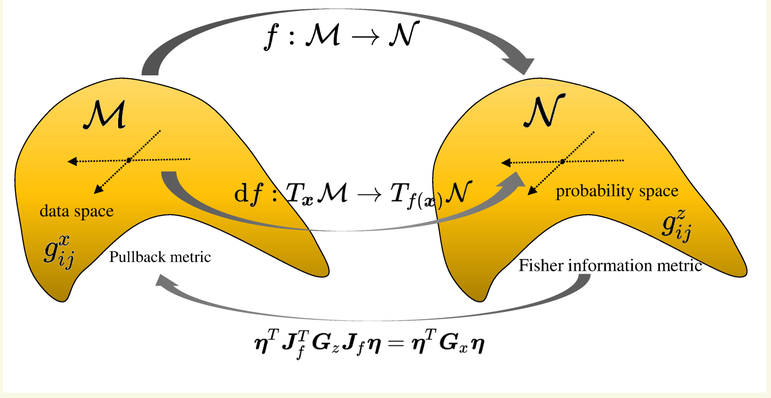
The Adversarial Attack and Detection under the Fisher Information Metric

Introduction / Motivation

- Neural networks are vulnerable to the adversarial attacks, making a severe challenge for safety-critical deep learning applications.
- Characterizing the robustness / vulnerability of deep learning models to a given sample is a typical question.
- We propose to measure the vulnerability of neural networks using the pullback from the output space to the data space.

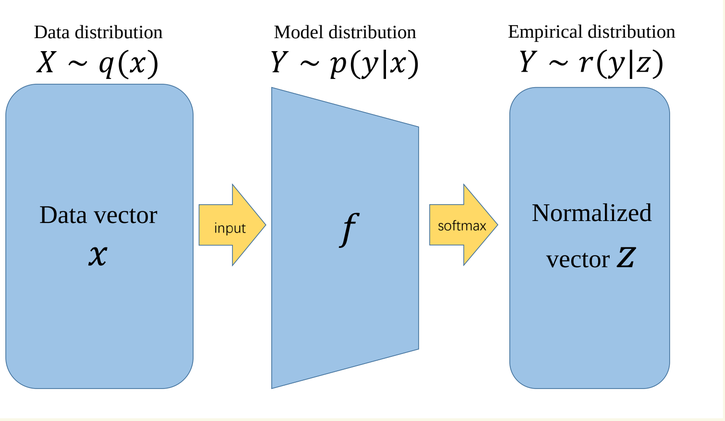


- This motivate us to perform both the adversarial attack and detection under the same framework.

Fisher Information Matrix

Previous ways to define the metric tensor G^x:

- Correlated by Jacobian $\mathcal{J}_f$: $G^x = \mathcal{J}_f^T G^z \mathcal{J}_f$
- Expectation over Hessian: $G^x_{ij} = -\mathbb{E}_{y|\boldsymbol{x}}[\frac{\partial^2}{\partial \boldsymbol{x}_i \partial \boldsymbol{x}_j} \log p(y|\boldsymbol{x})]$
- Hessian of KL w.r.t. the adversarial perturbation $\boldsymbol{\eta}$: $G^x_{ij} = \frac{\partial^2}{\partial \eta_i \partial \eta_j} D_{KL}(p(y|\boldsymbol{x}) || p(y|\boldsymbol{x}+\boldsymbol{\eta}))$



Our approach is to use the empirical distribution $r(y|\boldsymbol{x})$, this provides engineering benefits:

- $G^x_{ij} = \mathbb{E}_{r(y|\boldsymbol{x})}[(\frac{\partial}{\partial x_i} \log r(y|\boldsymbol{x}))(\frac{\partial}{\partial x_i} \log r(y|\boldsymbol{x}))^T]$
- Eigenvalue $\boldsymbol{\eta}^T G^x \boldsymbol{\eta} = \mathbb{E}_{r(y|\boldsymbol{x})}[(\boldsymbol{\eta}^T(\frac{\partial}{\partial x} \log p(y|\boldsymbol{x})))^2]$
- $G^x \boldsymbol{\eta} = \mathbb{E}_{r(y|\boldsymbol{x})}[(\boldsymbol{\eta}^T(\frac{\partial}{\partial x} \log p(y|\boldsymbol{x})))(\frac{\partial}{\partial x} \log p(y|\boldsymbol{x}))]$

Making it easier to calculate the eigen-decomposition without access to explicit G^x.

Formulation and Optimization Strategies

Objective function

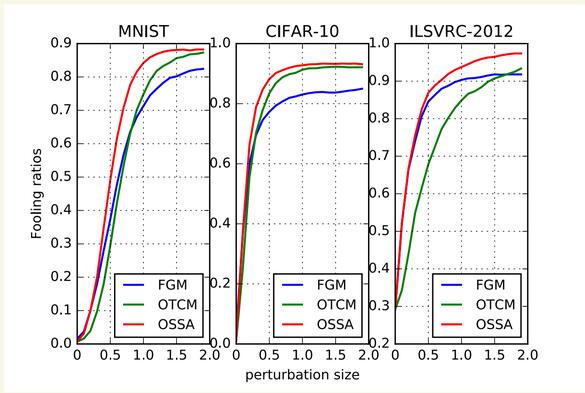
$$\max_{\boldsymbol{\eta}} \boldsymbol{\eta}^T G^x \boldsymbol{\eta} \qquad \text{s.t. } \|\boldsymbol{\eta}\|^2 = \epsilon,\ \mathcal{J}(y, \boldsymbol{x}) \leq \mathcal{J}(y, \boldsymbol{x}+\boldsymbol{\eta})$$

- An eigenvector gives two direction, but they are not equivalent.

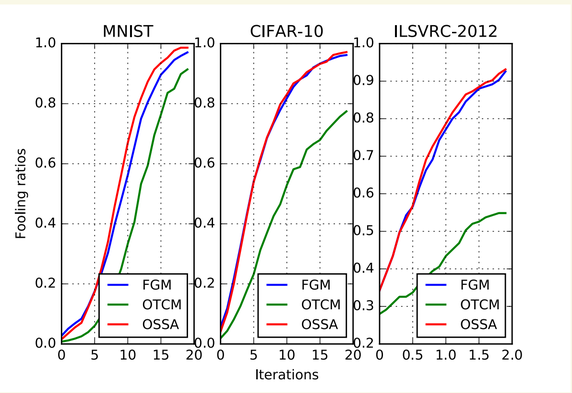
Optimal solution: the greatest eigenvector

- Only the greatest eigenvalues: Power iteration
- A group of eigenvectors and : Lanczos algorithm
- On large datasets: Monte-Carlo sampling from $r(y|\boldsymbol{x})$

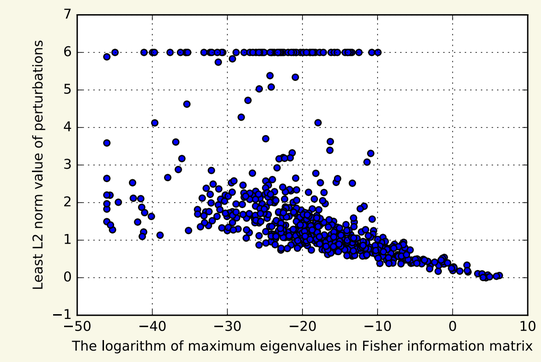
Empirical Evaluations



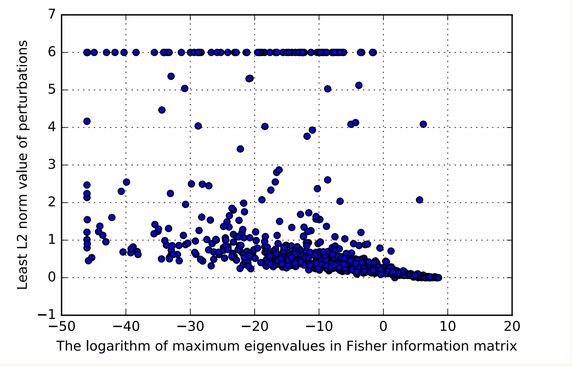
(a) One-step attack



(b) Iterative attack



(c) MNIST



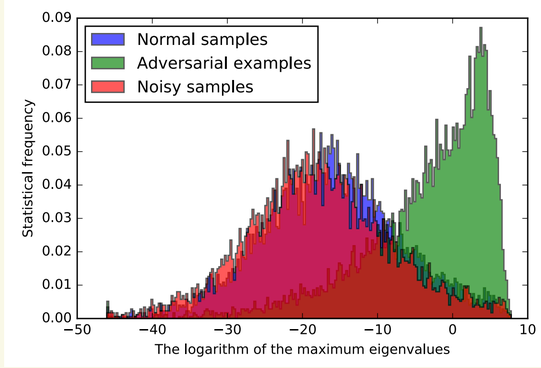
(d) CIFAR-10

Figure: (a,b) Comparison of fooling rates with gradient based attacks. (c,d) The relationship between least ℓ_2 perturbation size (obtained via binary search of OSSA) and log-eigenvalues

Adversarial Detection

Key idea: Use an auxiliary classifier to distinguish the adversarial examples with the eigenvalues serving as characteristics.

- Adding Gaussian noise on the input does not change the eigenvalues of the FIM, while the the adversarial examples are more likely to have larger eigenvalues.



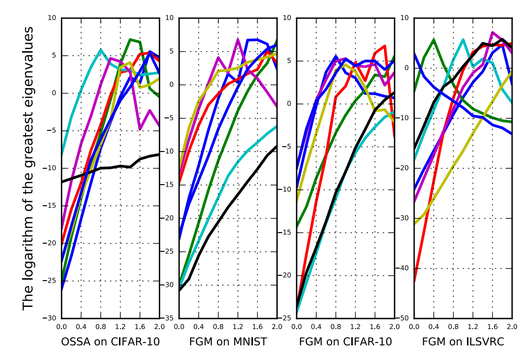


Figure: (left) The distribution of the eigenvalues of normal samples and adversarial examples. (right) The increasing of random samples' eigenvalues along the direction of adversarial perturbations.

Table: The AUC scores of detecting adversarial attacks using random forest with the eigenvalues as eigenvalues. The comparison is made between our proposed method, kernel density estimation (KD) and Bayesian uncertainty (BU).

	MNIST / CIFAR-10				
AUC (%)	FGM	OTCM	Opt	BIM	OSSA
KD	78.12	95.46	95.15	98.61	84.24
	64.92	92.13	91.35	98.70	88.89
BU	32.37	91.55	71.30	25.46	74.21
	70.40	91.93	91.39	97.32	87.44
KD+BU	82.43	95.78	95.35	98.81	85.97
	76.40	**94.45**	93.77	98.90	93.54
Ours	**96.11**	**98.47**	**95.67**	**99.10**	**93.13**
	80.18	93.68	**99.45**	**99.43**	**98.01**

Chenxiao Zhao[1] **P. Thomas Fletcher**[2] **Mixue Yu**[1] **Yaxin Peng**[3] **Guixu Zhang**[1] **Chaomin Shen**[1]

51174506043@stu.ecnu.edu.cn cmshen@cs.ecnu.edu.cn

[1]Dept of Computer Science, East China Normal University, China
[2]Dept of Electrical and Computer Engineering, and Dept of Computer Science, University of Virginia, Virginia, USA
[3]Dept of Mathematics, Shanghai University, China