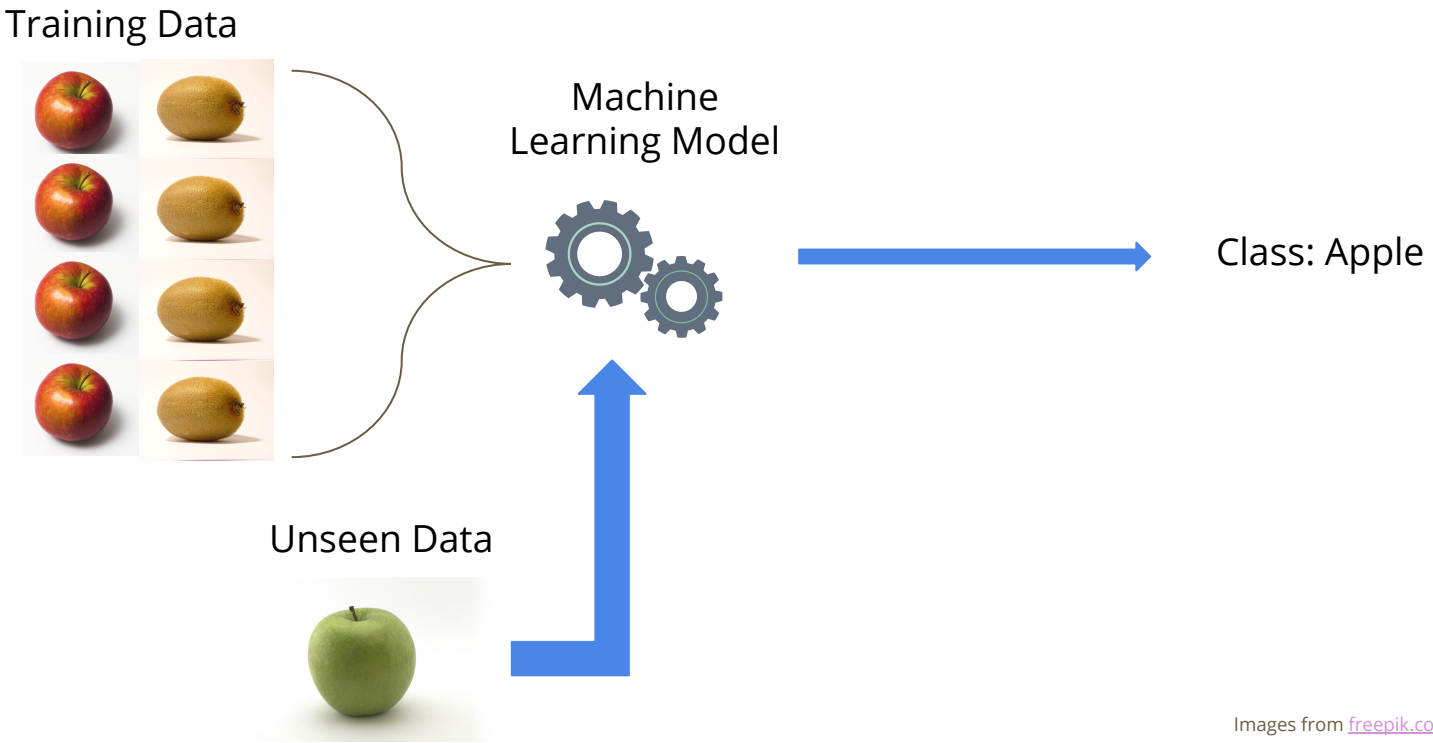
Generalization Comparison of Deep Neural Networks via Output Sensitivity

Mahsa Forouzesh, Farnood Salehi, Patrick Thiran

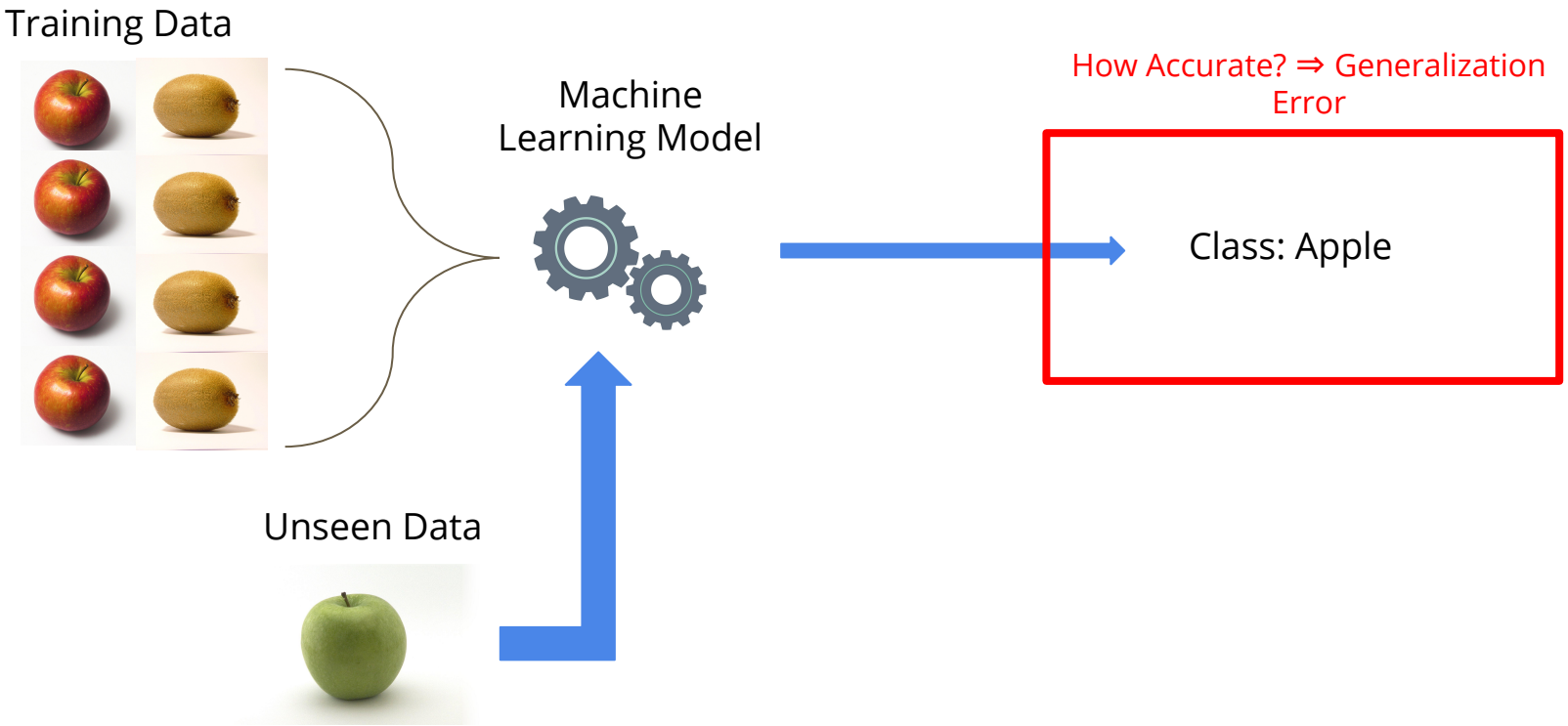
EPFL

Supervised Learning



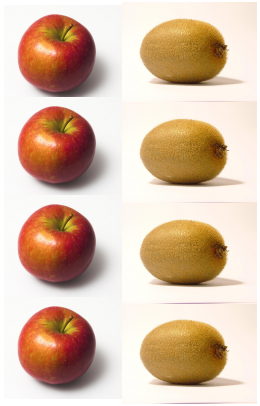
Images from freepik.com and freeimages.com

Supervised Learning



Supervised Learning

Training Data



Machine Learning Model

Can we compare the generalization performance of different models?

Class: Apple

Unseen Data

Notation

- Input: x Output: y
- Neural Network: f_θ
- Cross Entropy Loss: $L(f_\theta(x), y)$
- Output Sensitivity: $S = \mathrm{Var}(\overline{f_\theta(x + \varepsilon_x) - f_\theta(x)})$

Loss vs. Sensitivity

- Cross Entropy vs. Mean Square Error: $L \approx \sqrt{L_{\mathrm{MSE}}/2}$ (Loss)
- Bias-Variance Decomposition: $L_{\mathrm{MSE}} = \varepsilon_{\mathrm{bias}} + \varepsilon_{\mathrm{variance}}$
- Variance vs. Sensitivity: $\varepsilon_{\mathrm{variance}} \approx S \cdot C + \Sigma$ (Sensitivity, Constant, ~0)

Loss vs. Sensitivity

$$L \approx \sqrt{\frac{1}{2}\left(S \cdot C + \Sigma\right)}$$

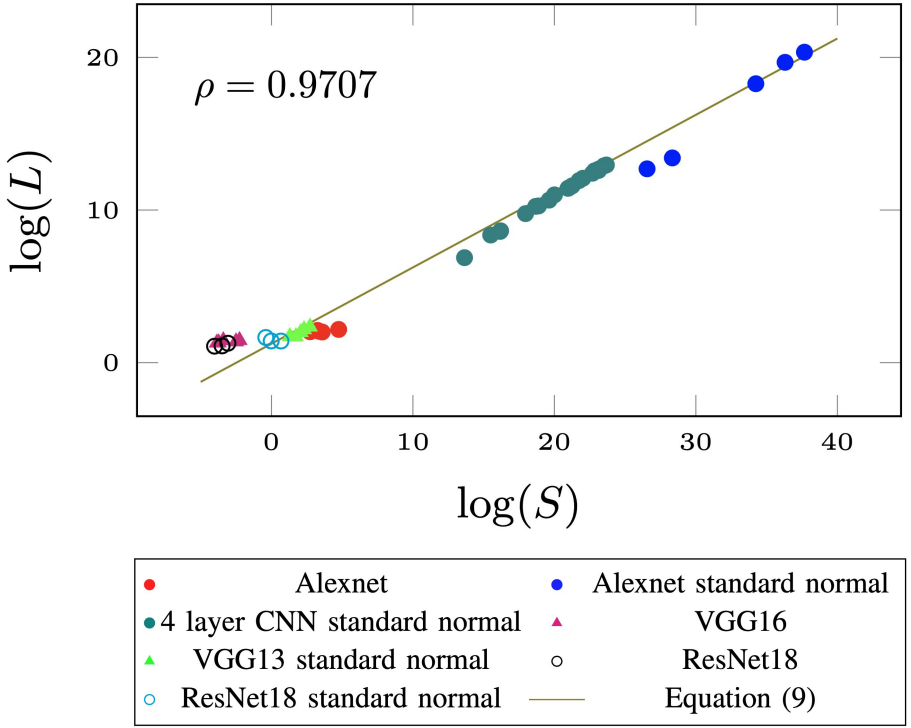
Loss vs. Sensitivity

$$L \approx \sqrt{\frac{1}{2}\left(S \cdot C + \Sigma\right)}$$

Requires labels

Does not require labels

Loss vs. Sensitivity

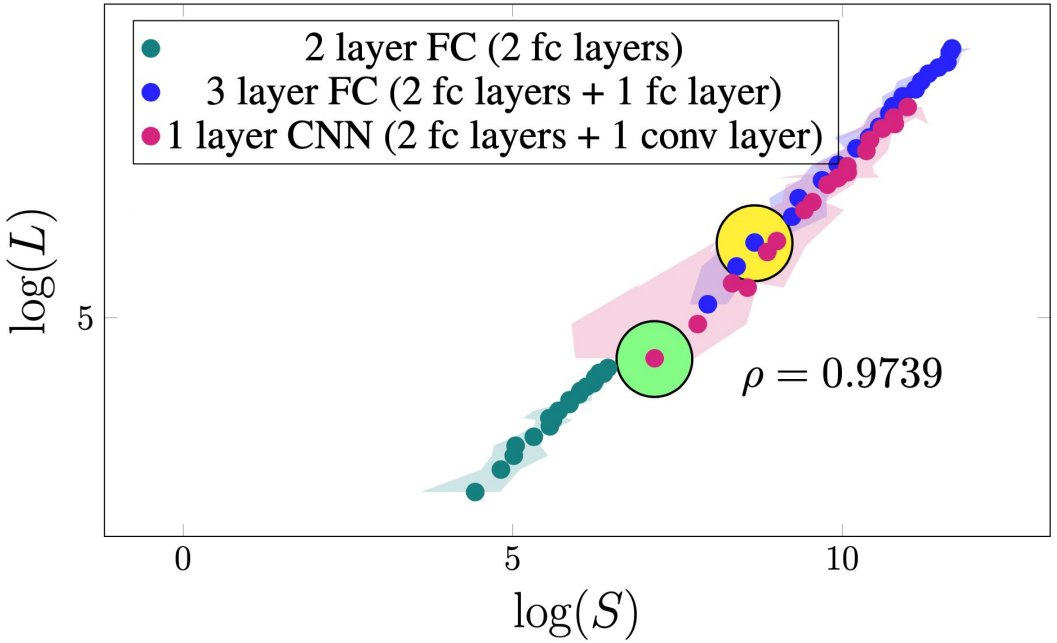


Sensitivity as a Proxy for Loss

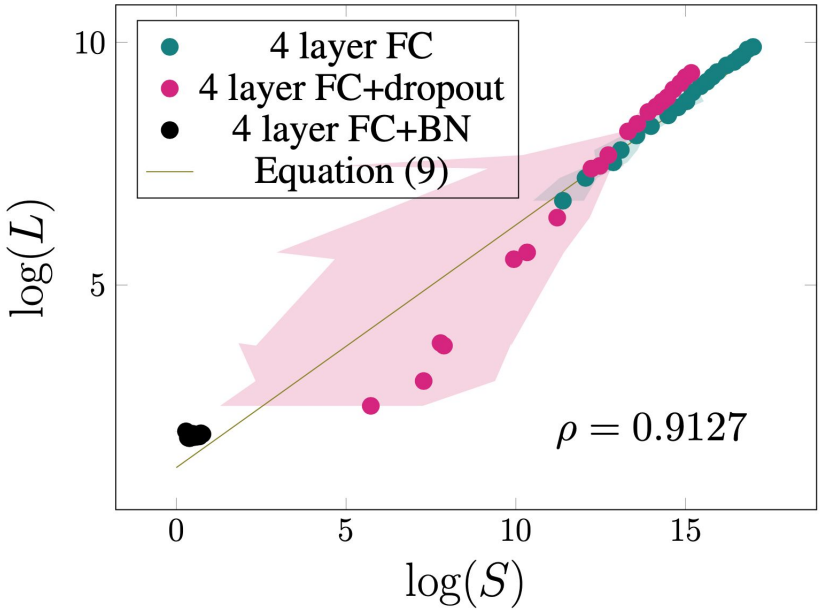
Convolutional

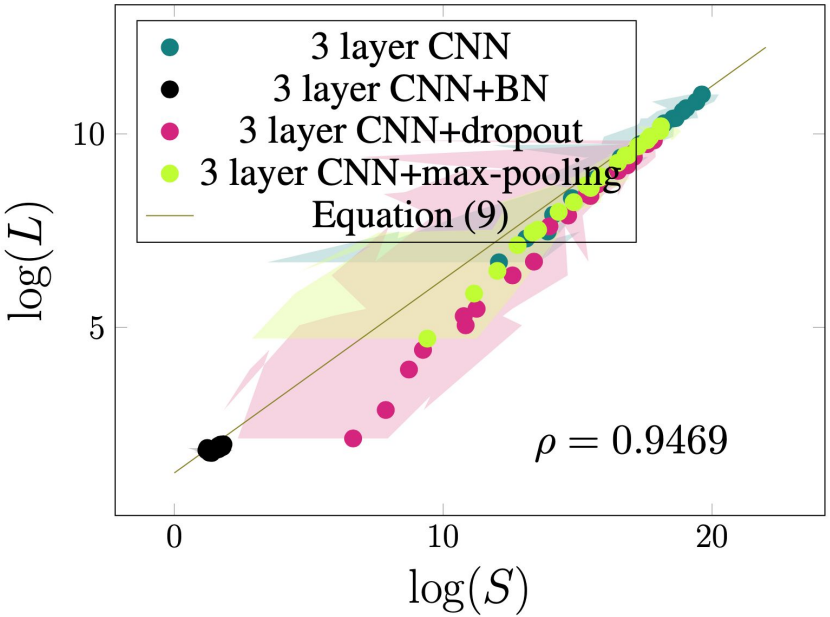
vs.

Fully-connected Layers



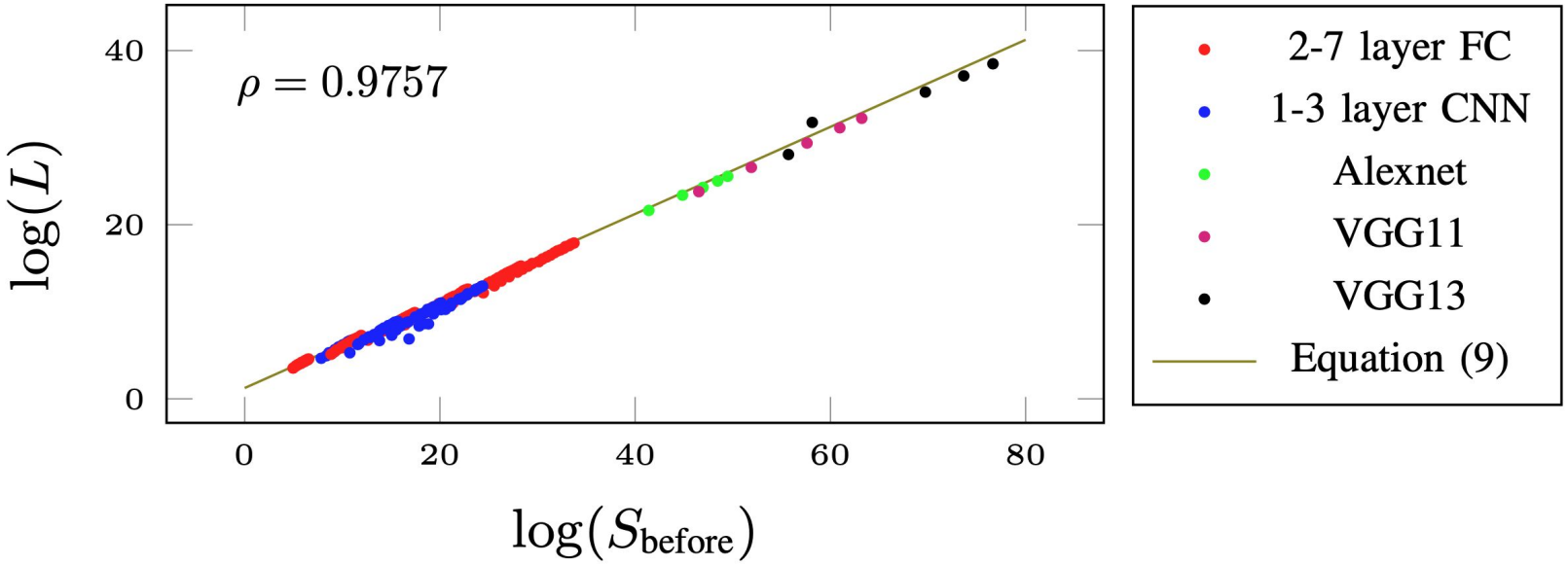
Sensitivity as a Proxy for Loss





Sensitivity of Untrained Networks

Output sensitivity computed *before* training: S_{before}



Future Directions

- When is the variance not the dominant term in the loss decomposition?
- Finding a link between the classification error and sensitivity
  - Bias-variance decomposition for the classification error?
  - Applying network calibration methods
- Beyond pixel-wise linear input perturbations and positive homogenous activation functions

Thank You!