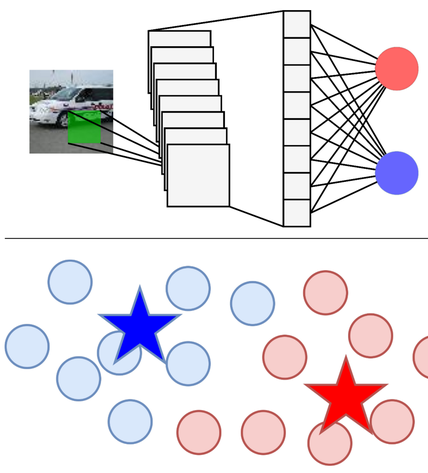
PORSCHE

Prototype-based neural network layers as a defense against adversarial attacks

Motivation



Neural networks
Able to model complex tasks, but susceptible to adversarial attacks

Learning Vector Quantization (LVQ)
Robust and interpretable, but hard to train without feature extraction

Prototype-based neural networks
Uses robust and interpretable layers within a neural network, taking advantage of the high capacity of neural networks and their ability to be trained end-to-end

LVQ as fully connected layer

- Replace final classification layer with a LVQ model
- Use existing techniques such as adversarial training and a High Level Guided Denoiser (HLGD) to further improve robustness

$$\mathbf{d}(\mathbf{v}) = -2\mathbf{W}\mathbf{v} + \mathbf{b}(\mathbf{v}, \mathbf{W})$$

$$\mathbf{b}(\mathbf{v}, \mathbf{W}) = \left(\|\mathbf{v}\|_2^2 + \|\mathbf{w}_1\|_2^2, \|\mathbf{v}\|_2^2 + \|\mathbf{w}_2\|_2^2, ..., \|\mathbf{v}\|_2^2 + \|\mathbf{w}_{N_w}\|_2^2 \right)^T$$

Eq. 1: The output of a LVQ classification model defined as a addition of a matrix multiplication with a bias term.

Implementation

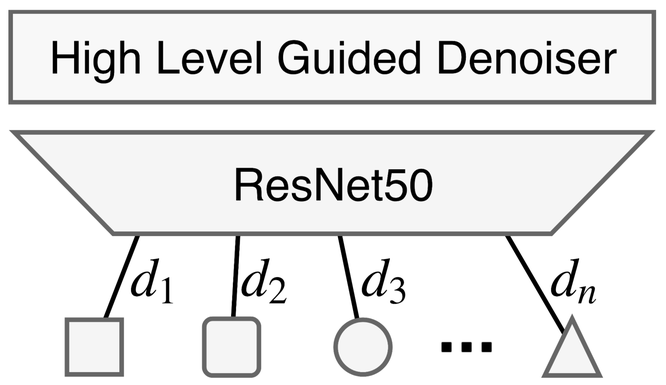


Fig. 1: General overview of the architecture used during the adversarial vision challenge. The output of the final convolutional layer in the ResNet is compared to all prototypes in the LVQ model, using the label of the closest prototype to classify the input.

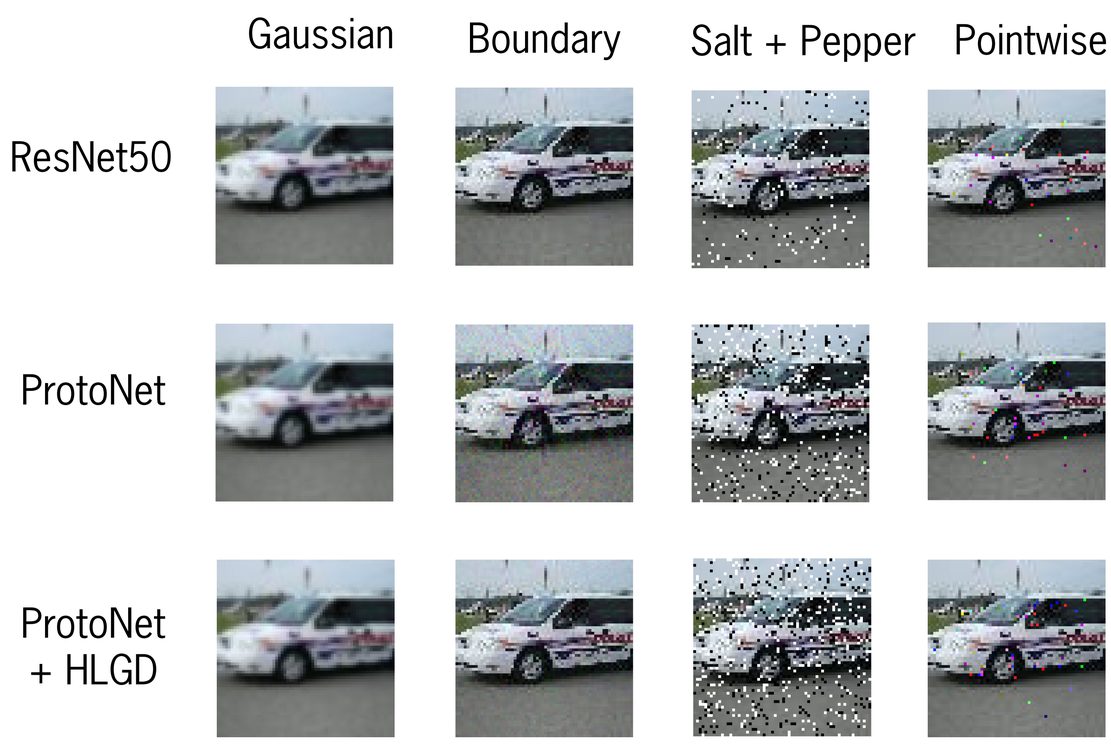


Fig. 2: An overview of the adversarial examples generated for each neural network by four different attacks.

Results

- Using an LVQ classification layer instead of a traditional dense layer provides a direct increase in robustness
- Smart regularization forces trained prototypes to be prototypical and interpretable

	Acc.	#Param.	Gaussian	Boundary	S.P.	Pointwise
ResNet50	60,6	74M	5,98	0,24	22,5	5,0
ProtoNet	63,2	48M	6,64	0,28	31,0	6,5
ProtoNet + HLGD	63,8	55M	6,63	0,33	48,0	9,0

Fig. 3: A summary of the networks robustness. For each attack the median perturbation required to fool the network is given. For the Gaussian and boundary attack the perturbation is measured using the L2-norm, and for the salt + pepper and pointwise attack the L0-norm is used.

Future work

- Replace classic convolution operation with prototype convolution
- Prototype convolution operations can be realized as summation of multiple convolution operations

$$\mathbf{x} \circledast \mathbf{k} = \mathbf{x}^2 * \mathbf{1} - 2\mathbf{x} * \mathbf{k} \oplus \|\mathbf{k}\|_2^2$$

Eq. 2: The prototype convolution defined in terms of standard convolution operations.

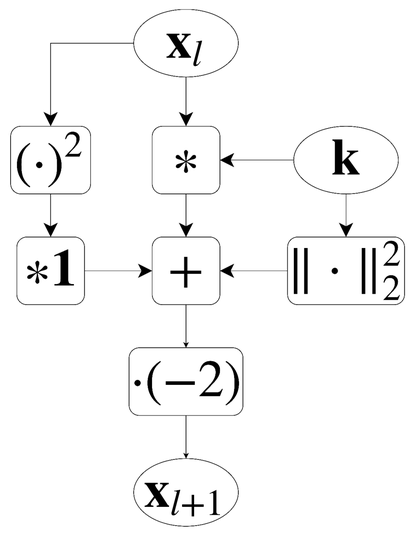


Fig. 4: The prototype convolution layer shown as a computation graph, including the gradient bypass

Further information

anysma

References:
1. Kohonen T. Learning Vector Quantization. 1988.
2. He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition.* 2016
3. Brendel W, Rauber J, Bethge M. *Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models.* 2017.
4. Liao F, Liang M, Dong Y, Pang T, Hu X, Zhu J. *Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser.* 2017.
5. Szegedy C, Zaremba W, Sutskever I, et al. *Intriguing properties of neural networks.* 2013.
6. Saralajew S, Holdijk L, Rees M, Villmann T. *Prototype-based Neural Network Layers: Incorporating Vector Quantization,* 2018.

PORSCHE

Lars Holdijk, Maike Rees, Thomas Villmann, Sascha Saralajew

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES