BUILDING EXPLAINABLE AND ROBUST NEURAL NETWORKS BY USING LIPSCHITZ CONSTRAINTS AND OPTIMAL TRANSPORT

M. Serrurier
IRIT/ANITI, Toulouse, France

Co-authors

- Louis Béthune
- Franck Mamalet
- Thibaut Boissin
- Jen-Michel Loubès
- Alberto González-Sanz

- Lipschitz constant of neural networks
- 1-Lipschitz neural networks
- Training 1-Lipschitz neural network with optimal transport
- Experimental results

LIPSCHITZ CONSTANT OF NEURAL NETWORKS

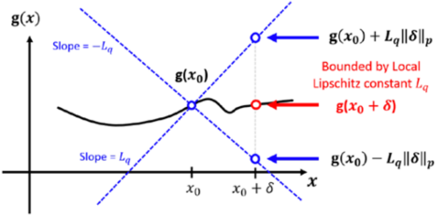
1-LIPSCHITZ FUNCTION

// DEFINITION

- $f : E \rightarrow F$ is k-Lipschitz iif:

  $$||f(x) - f(y)|| \leq k||x - y||$$

- Lipschitz constant : smallest value of k
  - 1D case : $k = max(f'(x))$



Intuition

how much the output of the fonction mary vary when I change the input

1-LIPSCHITZ FUNCTION
// NEURAL NETWORKS

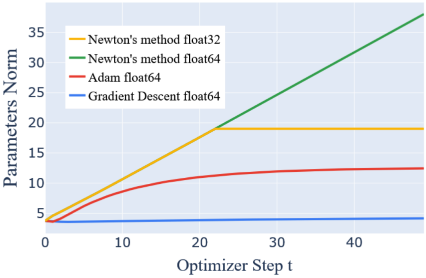
- Very hard to evaluate accurately (np-hard)
- Multilayer perceptron :

$$f(x) = \phi_k(W_k.(\phi_{k-1}(W_{k-1} \dots \phi_1(W_1.x)))$$

- Lipschitz constant upper-bound :

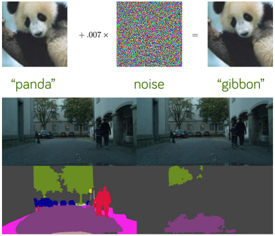
$$L(f) \leq L(\phi_k)*L(W_k)*L(\phi_{k-1})*L(W_{k-1})*\dots*L(\phi_1)*L(W_1.x).$$

- High constant value enforced by entropy minimization



HIGH LIPSCHITZ CONSTANT : CONSEQUENCES

// ADVERSARIAL ATTACK



Adversarial example

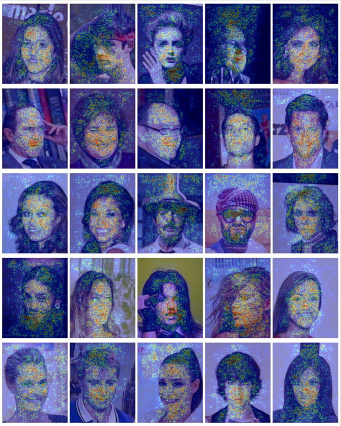
closest example with an opposite **decision** :

$$adv(f, \mathbf{x}) = \underset{\mathbf{z} \in \Omega | sign(f(\mathbf{z})) = -sign(f(\mathbf{x}))}{argmin} \| \mathbf{x} - \mathbf{z} \| .$$

Robustness :Average distance to the decision frontier w.r.t the input space

Counterfactual explanation

closest example in the opposite **class**



1-LIPSCHITZ NEURAL NETWORK

- Principles : all the layers have to be 1-lipschitz
- Dense Layer with kernel W

$$L(W) = ||W|| \leq ||W||_F \leq \max_{ij}(|W_{ij}|) * \sqrt{nm}$$

- Constraining Lipschitz constant :
  - WGAN : weight clipping (last term of the equation)
  - Weight normalization with Frobenius norm $||W||_F$
  - Spectral normalization with spectral norm $W_s = \frac{W}{||W||}$.

1-LIPSCHITZ NEURAL NETWORK
// CONSTRAINTS CONVOLUTION

Normalizing kernel is not enough

$$||Y_1 - Y_2||^2 = ||\bar{Y}_1 - \bar{Y}_2||^2 \leq ||\bar{W}||^2.||\bar{X}_1 - \bar{X}_2||^2 \leq \Lambda^2.||W||^2.||X_1 - X_2||^2$$

Λ depends on the duplication of pixels. We use the following upper bound :

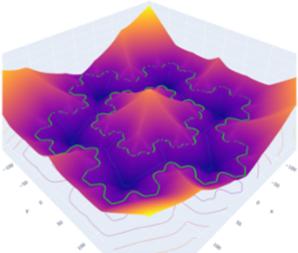
$$\Lambda = \sqrt{\frac{(k.w - \bar{k}.(\bar{k}+1)).(k.h - \bar{k}.(\bar{k}+1))}{h.w}}$$

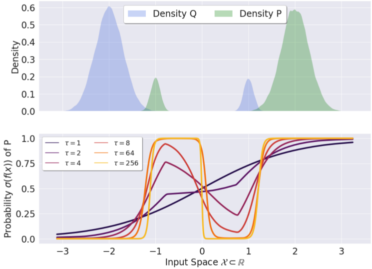
1-LIPSCHITZ NEURAL NETWORK
// OTHER LAYERS

- ReLU, sigmoid, Tanh: already 1-lipschitz
- LeakyReLU : 1-lipschitz if $\alpha < 1.0$
- PReLU (Parametric Rectified Linear Unit): need a constraint on scaling factor (=>PReLUlip)
- Pooling : scaling factor or l2 norm pooling
- BatchNormalization: Not lipschitz
- Dropout: Not Lipschitz

1-lipschitz classifier are too limited ?

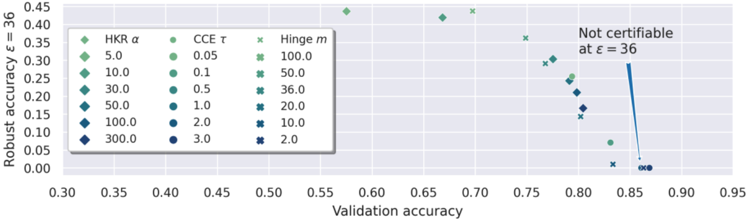
- 1-lipschitz classifiers can approximate as precisely as possible any arbitrarily complex decision frontier
- Constraining the Lipschitz constant change the optimal value of the loss function



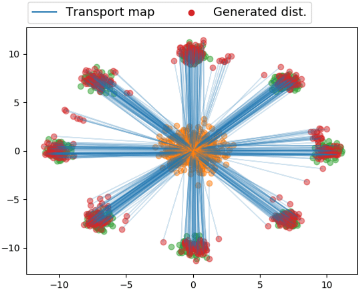


Consequence

- Tuning the loss (or equivalently the lipschitz constraint) change the accuracy/robustness tradeoff
- Cross entropy losses provide poor robustness certificates



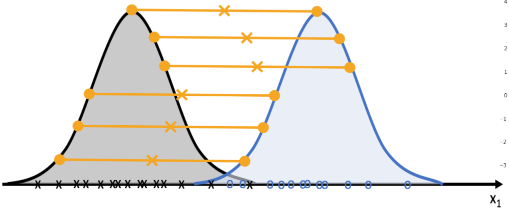
OPTIMAL TRANSPORT
// WASSERSTEIN DISTANCE

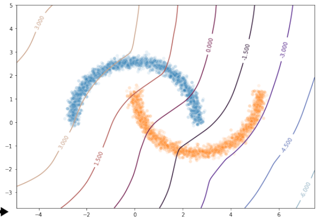


- Primal formulation: $\mathcal{W}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \underset{x,z\sim\pi}{\mathbb{E}} \parallel \mathbf{x}-\mathbf{z} \parallel$
- Dual formulation
  $\mathcal{W}(\mu,\nu) = \sup_{f\in Lip_1(\Omega)} \underset{\mathbf{x}\sim\mu}{\mathbb{E}}[f(\mathbf{x})] - \underset{\mathbf{x}\sim\nu}{\mathbb{E}}[f(\mathbf{x})]$
- f can be a represented by a 1-lipschitz neural network

Discriminator in W-GAN

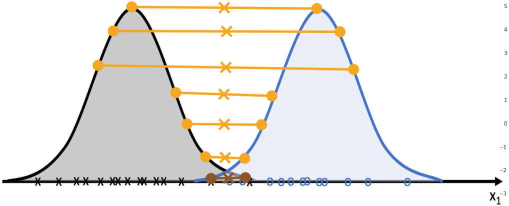
- Weak classifier
- Most robust classifier

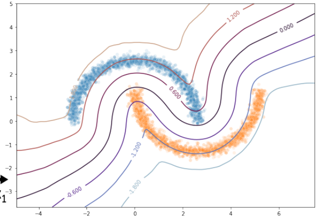




Hinge regularized Wasserstein classifier

$$\inf_{f \in Lip_1(\Omega)} \mathop{\mathbb{E}}_{\mathbf{x} \sim P_-} [f(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x} \sim P_+} [f(\mathbf{x})] + \lambda \mathop{\mathbb{E}}_{\mathbf{x}} (1 - Yf(\mathbf{x}))_+$$





HINGE KR LOSS
// THEORITICAL PROPERTIES

- Existence of the solution
- Can achieve 100% accuracy when classes are separable
- Hinge regularized Wasserstein is still an optimal transport problem

$$\sup_{f \in \text{Lip}_1(\Omega)} -\mathcal{L}_\lambda^{hKR}(f) = \inf_{\pi \in \Pi_\lambda^p(P_+, P_-)} \int_{\Omega \times \Omega} |\mathbf{x} - \mathbf{z}| d\pi + \pi_{\mathbf{x}}(\Omega) + \pi_{\mathbf{z}}(\Omega) - 1$$

- $||\nabla_x f^*(\mathbf{x})|| = 1$ almost everywhere for the optimal f^*

- Provable robustness

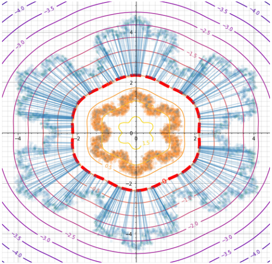
$$||x - adv(\hat{f}, x)|| \geq \hat{f}(x)$$

- even better :

$$||x - adv(f^*, x)|| = f^*(x)$$

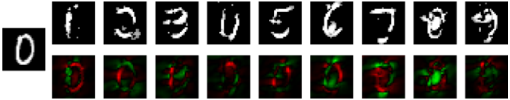
- Adversarial attack :
  - Follow the transportation path
  - In the direction of $\nabla_x \hat{f}(\mathbf{x})$

HINGE KR LOSS



$$||x - adv(f^*, x)|| = f^*(x)$$

- Adversarial attacks become conterfactual explanations
- Saliency maps represent the direction of the explanation



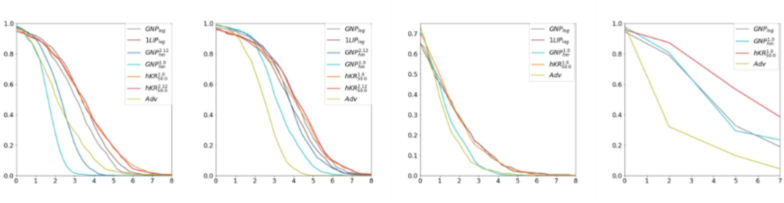
BEYOND 1-LIPSCHITZ NEURAL NETWORKS
// ORTHOGONAL KERNELS

- $||\nabla_x f^*(\mathbf{x})|| = 1$ almost everywhere i.e. f* is piecewise linear with slope equal to 1 almost everywhere
- How to achieve that :
  - Use gradient preserving activation function
    - Max min
    - Group sort
    - Full-sort
  - Ortonormalize the eigen vectors of the kernel of each layers (all singular values equal to one)
    - Bjork algorithm during inference
    - Can be time consuming

BEYOND 1-LIPSCHITZ NEURAL NETWORKS
// DEEL LIP LAYER LIBRARY

- Keras/tensorflow and pythorch implementation
- Open source
- Keras extention :
  - k-lischitz layers
  - activation functions
  - weight initializers
  - monitoring tools
- layer exportation (optimization for inference)

EXPERIMENTAL RESULTS



(a) MNIST dense (b) MNIST CNN (c) CIFAR (d) CelebA

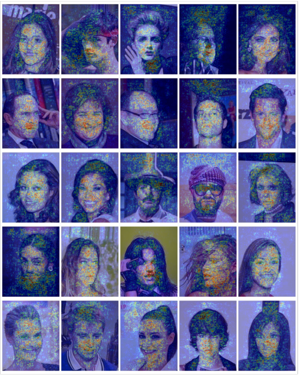
Figure 4: Accuracy (Y-axis) w.r.t. of l_2 norm of $FGSM$, l_2PGD, deepfool and l_2 Carlini and Wagner combined attacks on 500 images of the test set

EXPERIMENTAL RESULTS
// ILLUSTRATION WITH MUSTACHE

(a) Fooling CelebA images classical network

(b) Fooling images 1-lipschitz network (binary crossentropy)

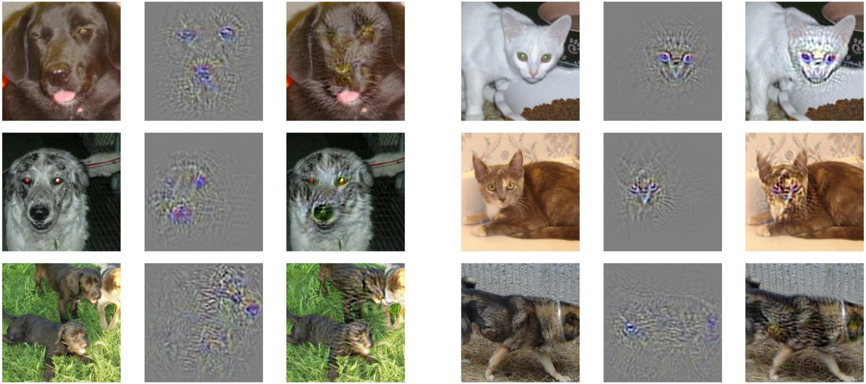
EXPERIMENTAL RESULTS
// EXPLANABILITY : SALIENCY MAPS



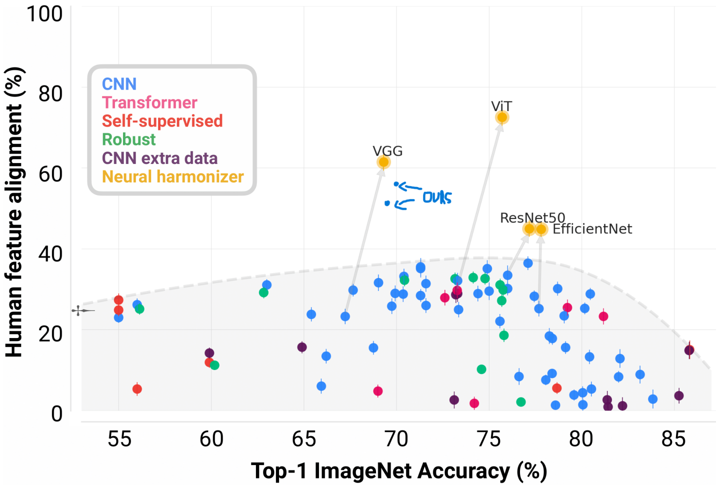
EXPERIMENTAL RESULTS

// EXPLANABILITY : COUNTERFACTUALS

EXPERIMENTAL RESULTS
// EXPLANABILITY : COUNTERFACTUALS



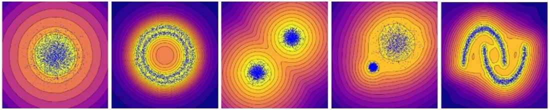
EXPERIMENTAL RESULTS
// HUMAN EVALUATION OF SALIENCY MAP



CONCLUSIONS

CONCLUSIONS

- Training 1-Lipschitz network with optimal transport loss
  - New interpretation of classification problem
  - Improve robustness structurally
  - Meets certification requirement
  - Interpretable
  - State of the art accuracy on large problems (70% imagenet)
- Deep lip : accessible and optimized librairy to train and use 1-Lipschitz networks
- Future works
  - One-class classification
  - Outlier detection



CONCLUSIONS
// LIMITATIONS AND OPEN PROBLEMS

- Limited to $l2$ norm optimal transport
- Prone to overfitting
- Counterfactuals less convicing for large multiclass problems
- problems
  - How to consider other norms ?
  - Specific architecture ? 1-Lip transformers ?
  - Optimization

THANK YOU FOR YOUR ATTENTION, QUESTIONS ?