DEEL

DEpendable & Explainable Learning

IVADO

IRT SAINT EXUPÉRY

CRIAQ
CONSORTIUM DE RECHERCHE ET D'INNOVATION EN AÉROSPATIALE AU QUÉBEC

ANITI
Université Fédérale
Toulouse Midi-Pyrénées

iid
Institut
intelligence
et données

DEEL

DEpendable & Explainable Learning

Other applications of LipNet

DEpendable & Explainable Learning

Semantic Segmentation

What does mean robustness in Semantic Segmentation

- Adversarial attacks and defences in classification
  - Single budget, single decision on each image => single robustness success/failure per image
  - Robust Accuracy is an average on a dataset
  - huge literature in the domain
- Adversarial attach in semantic segmentation:
  - Single budget but multiple decision => multiple robustness robustness success/failure
  - Robust Pixel Accuracy can be evaluated on a single image
  - Literature is less important and often metrics are not comparable

- PhD in DEEL and CALM chair (T. Massena SNCF)
  - Paper under review “Fast and Flexible Robustness Certificates for Semantic Segmentation”

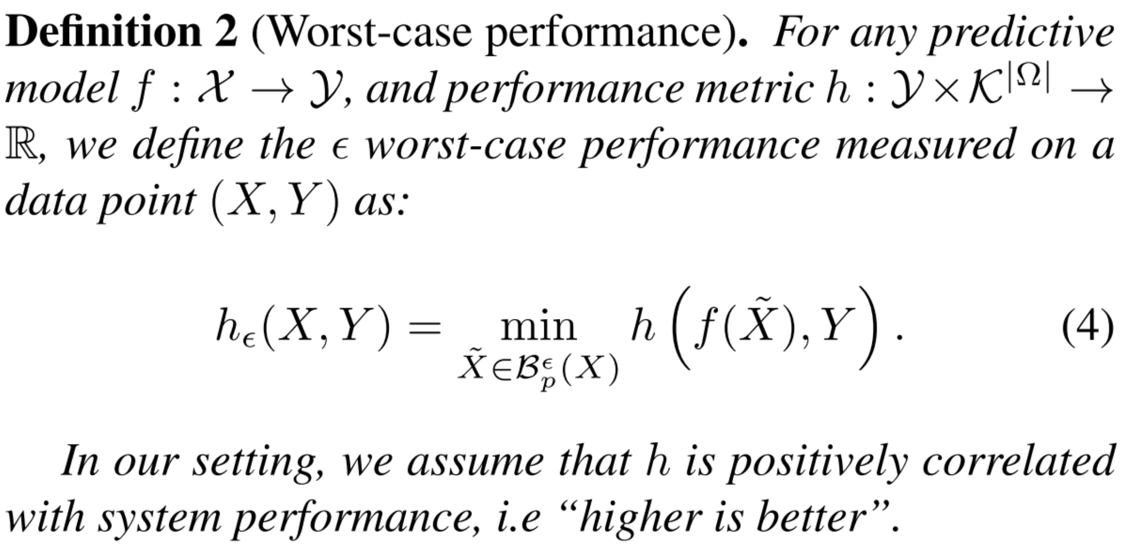
What does mean robustness in Semantic Segmentation

- Unifying Semantic Segmentation Robustness Metrics

Notation X input image, Y annotation (class for each pixel), $f(X)$ output of the segmentation (logits), $\hat{Y}$ prediction of the segmentation (class for each pixel), $\mathcal{B}_\epsilon(X)$ the robustness ball

Q1 — Given an adversarial budget ϵ, what is the worst performance I could reach?

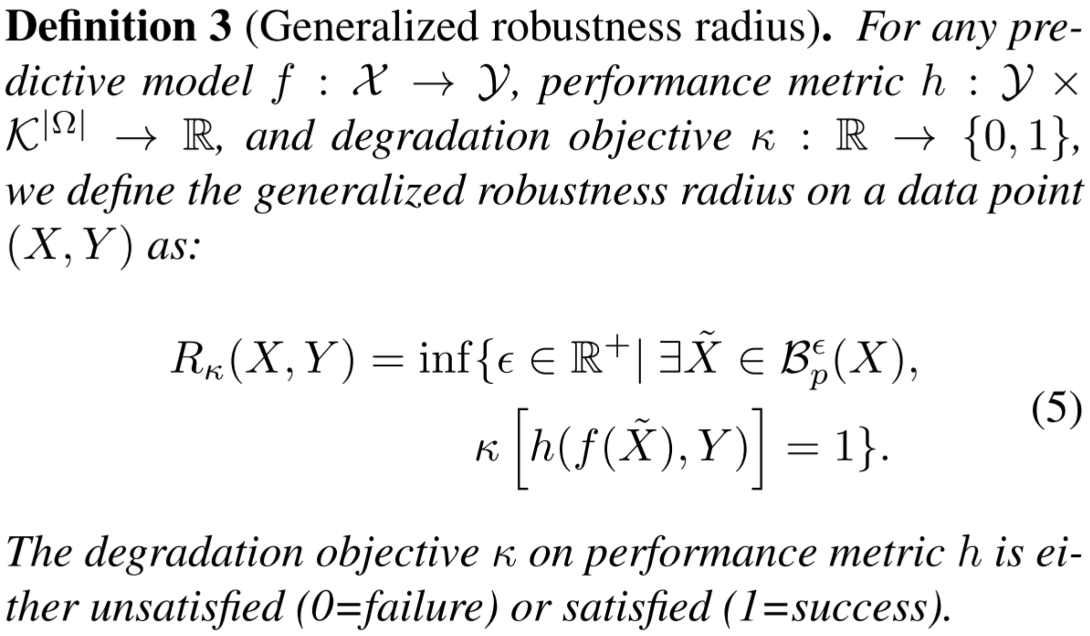
Q2 — If I want to degrade the performance metric to satisfy a degradation objective κ, what adversarial budget do I need?



Definition 2 (Worst-case performance). *For any predictive model $f : \mathcal{X} \rightarrow \mathcal{Y}$, and performance metric $h : \mathcal{Y} \times \mathcal{K}^{|\Omega|} \rightarrow \mathbb{R}$, we define the ϵ worst-case performance measured on a data point (X, Y) as:*

$$h_\epsilon(X, Y) = \min_{\tilde{X} \in \mathcal{B}_p^\epsilon(X)} h\left(f(\tilde{X}), Y\right). \qquad (4)$$

In our setting, we assume that h is positively correlated with system performance, i.e "higher is better".



Definition 3 (Generalized robustness radius). *For any predictive model $f : \mathcal{X} \rightarrow \mathcal{Y}$, performance metric $h : \mathcal{Y} \times \mathcal{K}^{|\Omega|} \rightarrow \mathbb{R}$, and degradation objective $\kappa : \mathbb{R} \rightarrow \{0, 1\}$, we define the generalized robustness radius on a data point (X, Y) as:*

$$\begin{aligned} R_\kappa(X, Y) = \inf\{\epsilon \in \mathbb{R}^+ |\ \exists \tilde{X} \in \mathcal{B}_p^\epsilon(X), \\ \kappa\left[h(f(\tilde{X}), Y)\right] = 1\}. \end{aligned} \qquad (5)$$

The degradation objective κ on performance metric h is either unsatisfied (0=failure) or satisfied (1=success).

What does mean robustness in Semantic Segmentation

- Application to the Pixel Accuracy Metric

Notation X input image, Y annotation (class for each pixel), $f(X)$ output of the segmentation (logits), $\hat{Y}$ prediction of the segmentation (class for each pixel), $\mathcal{B}_\epsilon(X)$ the robustness ball

Pixel Accuracy:

$$h(f(X), Y) = \frac{1}{|S|} \sum_{\omega \in S} \mathbb{1}_{\hat{Y}_\omega = Y_\omega}$$

Q1 — What is the maximal degradation of pixel accuracy that can be achieved given an adversarial budget ϵ?

Definition 2 (Worst-case performance). *For any predictive model $f : \mathcal{X} \to \mathcal{Y}$ and performance metric $h : \mathcal{Y} \times \mathcal{K}^{|\Omega|} \to$ $\mathbb{R}$, we define* *sured on a data point (X, Y) as:*

Robust Pixel Accuracy (RPA)

$$h_\epsilon(X, Y) = \min_{\tilde{X} \in \mathcal{B}_p^\epsilon(X)} h\left(f(\tilde{X}), Y\right). \tag{4}$$

In our setting, we assume that h is positively correlated with system performance, i.e "higher is better".

Q2 — What is the maximum attack level ϵ under which the pixel accuracy is guaranteed to remain above or equal to γ?

Pixel accuracy threshold

$$\kappa[h(f(\tilde{X}), Y)] = \mathbb{1}_{h(f(\tilde{X}), Y) \leq \gamma}.$$

Definition 3 (Generalized robustness radius). *For any predictive model $f : \mathcal{X} \to \mathcal{Y}$, performance metric $h : \mathcal{Y} \times \mathcal{K}^{|\Omega|} \to \mathbb{R}$, and degradation objective $\kappa : \mathbb{R} \to \{0, 1\}$, we define the generalized robustness radius on a data point (X, Y) as:*

$$\begin{aligned} R_\kappa(X, Y) = \inf\{\epsilon \in \mathbb{R}^+ |\ \exists \tilde{X} \in \mathcal{B}_p^\epsilon(X), \\ \kappa\left[h(f(\tilde{X}), Y)\right] = 1\}. \end{aligned} \tag{5}$$

Other examples of metrics (FNR, Stability, IoU) are included in the paper

Could we use Lipschitz constant to compute certificates?

- Each pixel output ω can provide a robustness radius (similar to classification)

$$R^{\omega}(X,Y) := \mathbb{1}_{\hat{Y}_{\omega}=Y_{\omega}} .2^{\frac{1-p}{p}} .\mathcal{M}_X^{\omega}(f)/L,$$

$$\text{with} \quad \mathcal{M}_X^{\omega}(f) \quad = \quad f^{\text{top1}}(X)_{\omega} \; - \; f^{\text{top2}}(X)_{\omega}.$$

But with a shared budget ϵ

Could we provide a Certified Robust Pixel Accuracy (CRPA)?

Given the Lipschitz constant of the network L, we can provide a lower bound of CRA

$$\begin{aligned} h_{\epsilon}(X) &= \min_{\delta \in \mathcal{B}_p^{\epsilon}(0)} h(f(X+\delta), Y) \\ &\geq \min_{\alpha \in \mathcal{B}_p^{L\epsilon}(0)} h(f(X)+\alpha, Y). \end{aligned}$$

We can reformulate the CRPA as a knapsack problem of the maximum number of pixels that can be attacked under a budget ϵ

$$\begin{aligned} N_{\text{PA}}(X,\epsilon) = \max & \sum_{\omega \in S} p_{\omega} \\ \text{s.t.} & \sum_{\omega \in S} L^p c_{\omega} p_{\omega} \leq (L\epsilon)^p \end{aligned}$$

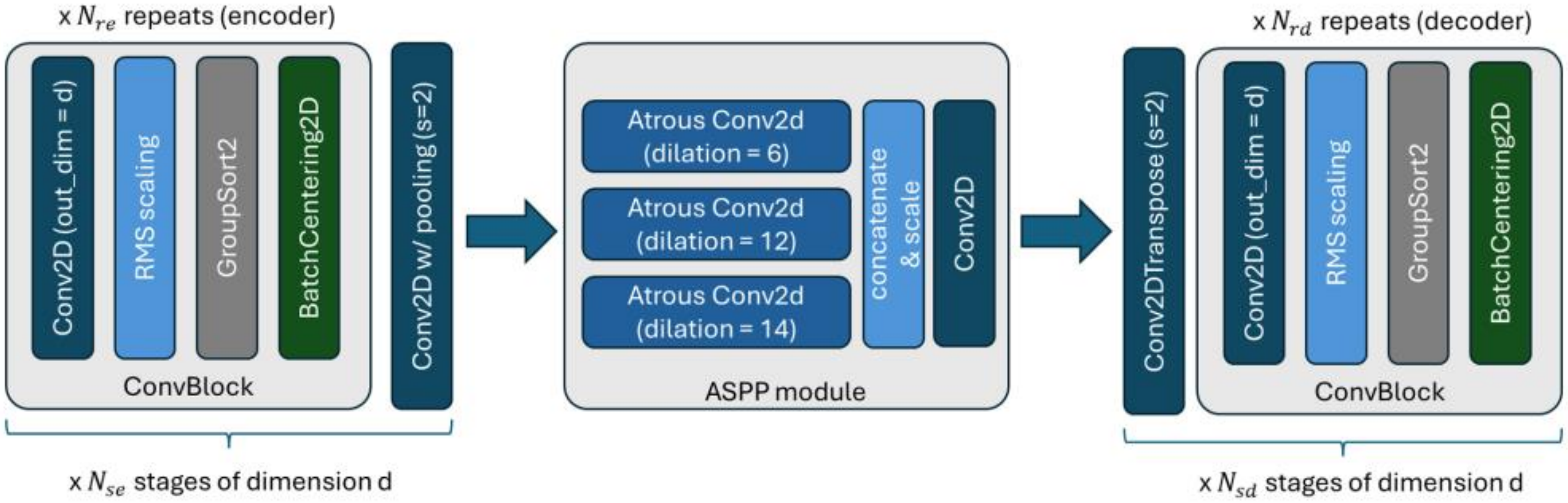
$$CRPA_{\epsilon}(X) = 1 - \frac{N_{PA}}{|S|}$$

KP problem can be easily solved

$$N_{\text{SUP}}(X,\epsilon,S,R^{\omega}) = \\ \sup\left\{ n \in \mathbb{N} \;\middle|\; \sum_{k=1}^{n} R^{\pi x(k)}(X,Y)^p \leq \epsilon^p \right\}.$$

How to learn LipNet for semantic segmentation

- In semantic segmentation most efficient architecture use transformers (not 1-Lip)
- Unet is a well known architecture and can be transformed into LipNet
- A more recent Convolutional architecture called DeepLabV3 has better performance on CityScape
- We provide a LipNet variant of DeepLabV3 (based on Orthogonium library)



LipNet performances for semantic segmentation

- LipNet DeepLab V3 performances

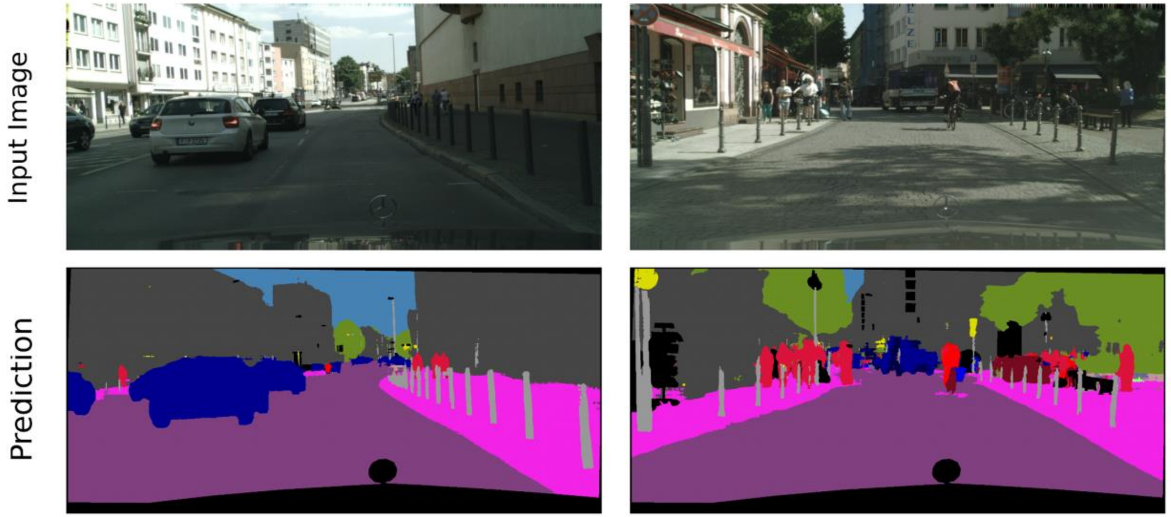
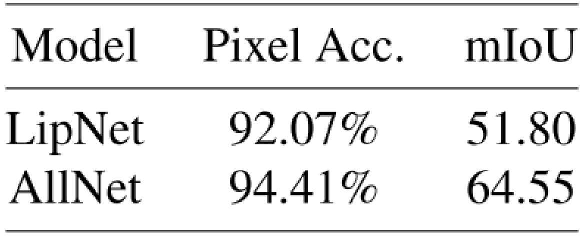
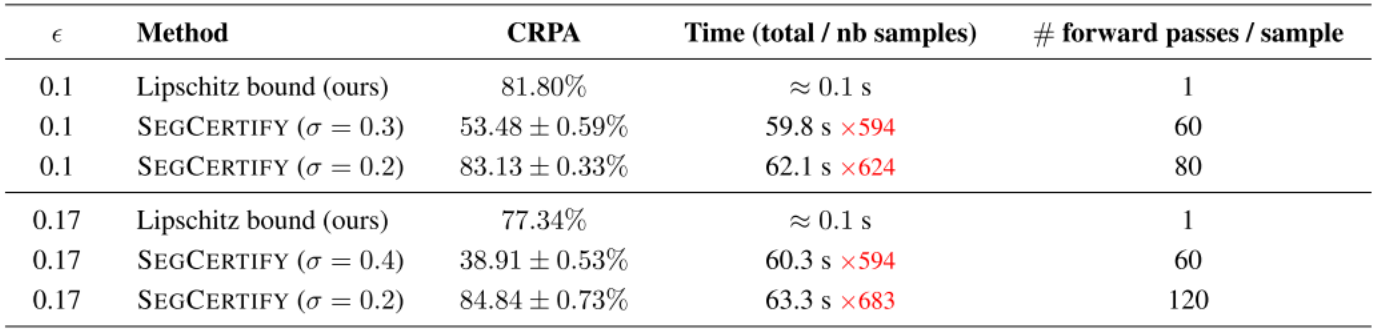


Figure 4. Visualization of test set segmentation results using our Lipschitz constrained neural networks trained using the cosine similarity.



Model	Pixel Acc.	mIoU
LipNet	92.07%	51.80
AllNet	94.41%	64.55

CRPA comparison



ϵ	Method	CRPA	Time (total / nb samples)	# forward passes / sample
0.1	Lipschitz bound (ours)	81.80%	≈ 0.1 s	1
0.1	SEGCERTIFY ($\sigma = 0.3$)	$53.48 \pm 0.59\%$	59.8 s ×594	60
0.1	SEGCERTIFY ($\sigma = 0.2$)	$83.13 \pm 0.33\%$	62.1 s ×624	80
0.17	Lipschitz bound (ours)	77.34%	≈ 0.1 s	1
0.17	SEGCERTIFY ($\sigma = 0.4$)	$38.91 \pm 0.53\%$	60.3 s ×594	60
0.17	SEGCERTIFY ($\sigma = 0.2$)	$84.84 \pm 0.73\%$	63.3 s ×683	120

Table 1. CRPA values across methods on the Cityscapes dataset [11] using 1024×1024 images. We choose $\alpha = 0.001$ as the failure probability of SEGCERTIFY and tune $\sigma \in \{0.15, 0.2, 0.25, 0.3, 0.4, 0.5\}$ for each run. Finally, given the very long computation time of smoothing based methods, evaluations are run on only 100 images of the dataset, as done in [12]. We report the mean and standard deviation of results across 5 runs that use the best performing σ value. We also report the mean runtime for each evaluation divided by the number of samples. The results using Lipschitz bounds are obtained on the whole test set.

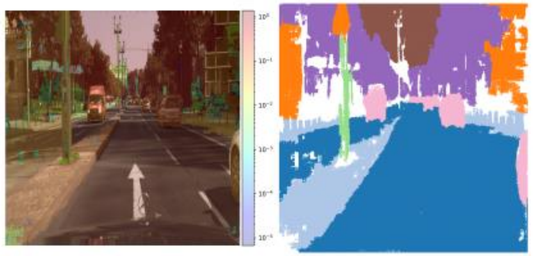


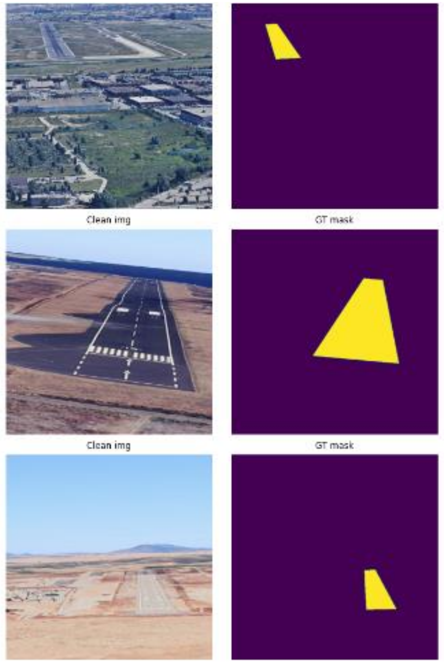
Figure 1. (Left) The ϵ budget required to attack dense segmentations to make all but $N_{\min}$ pixels change. (Right) We display only the groups of predictions where $\epsilon \geq 0.1$, non-robust pixel groups are in white.

LARD-V1: LipNet for Semantic Segmentation

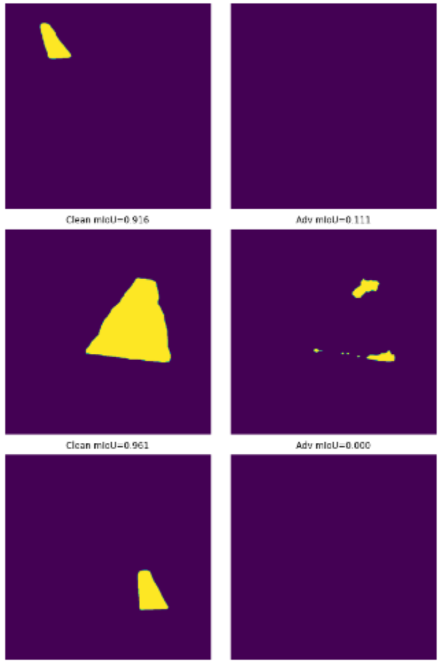
Architecture FCN (tested also with Unet)
Adversarial attack: vanishing objective
$(\epsilon = 1.0)$

Model	Clean IoU	Robust IoU
Standard model	0.89	0.26
Lipschitz slightly robust	0.82	0.57
Lipschitz strongly robust	0.79	0.65

Image GT mask

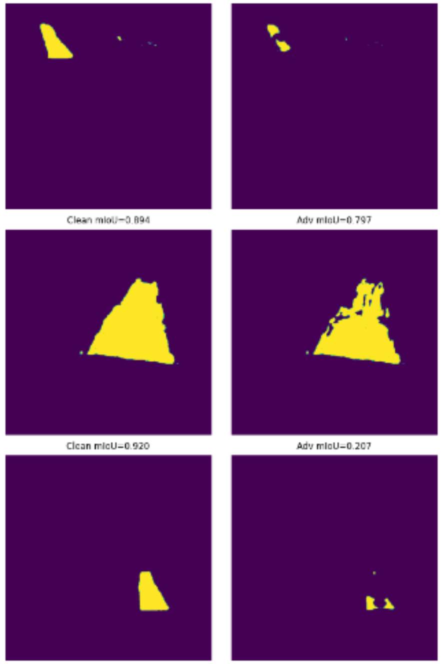


Standard model



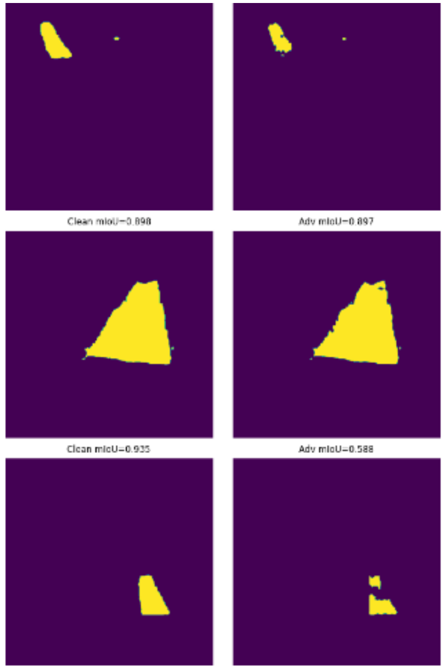
Clean prediction Robust prediction

Lipschitz sligthly robust



Clean prediction Robust prediction

Lipschitz strongly robust

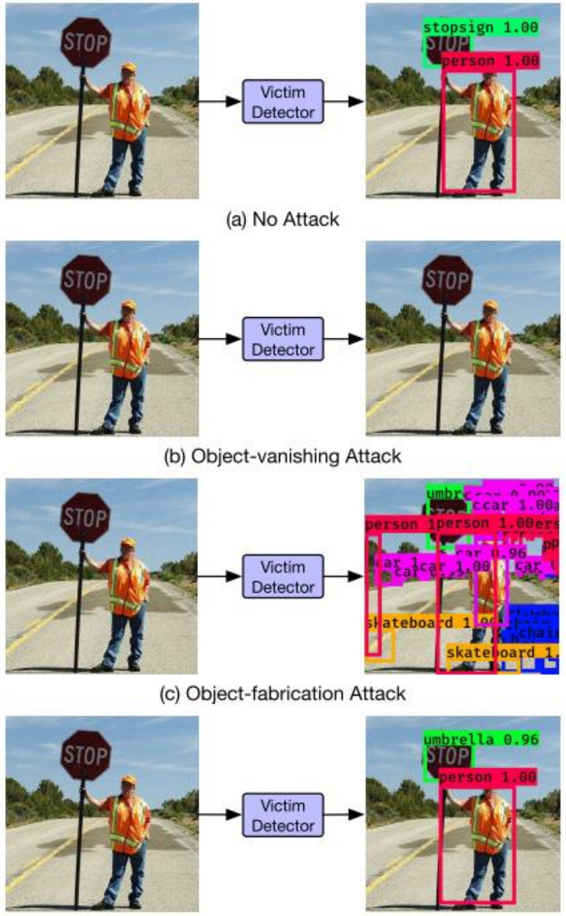


Clean prediction Robust prediction

Object Detection

What does mean robustness in Object detection

- Adversarial attach in object detection: several types of attack



(a) No Attack

(b) Object-vanishing Attack

(c) Object-fabrication Attack

(d) Object-mislabeling Attack (“stop sign” → “umbrella”)

Chow et al, Adversarial Objectness Gradient Attacks on Real-time Object Detection Systems, 2020

Vanishing attack: for instance by reducing the objectness/confidence score, or by modifying the bbox

Fabrication attack: for instance by increasing the confidence score at a given position

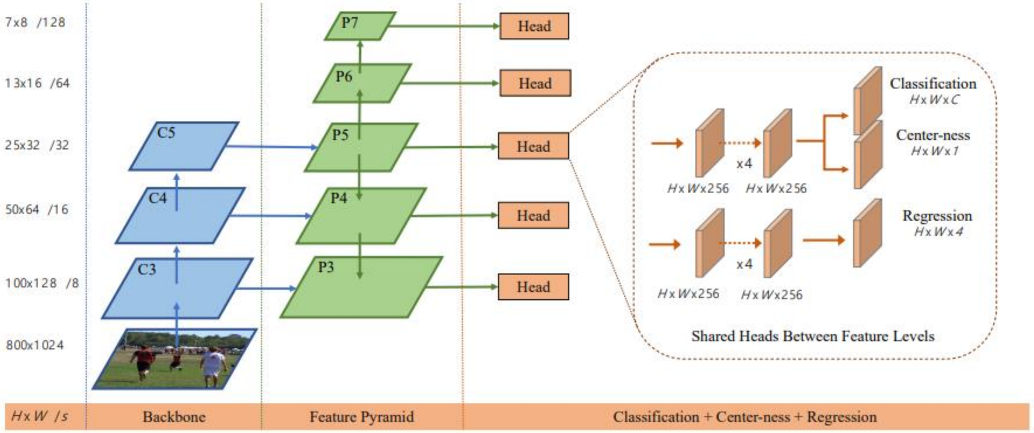
Mislabeling attack: attacking the classification head

Could we compute certificates

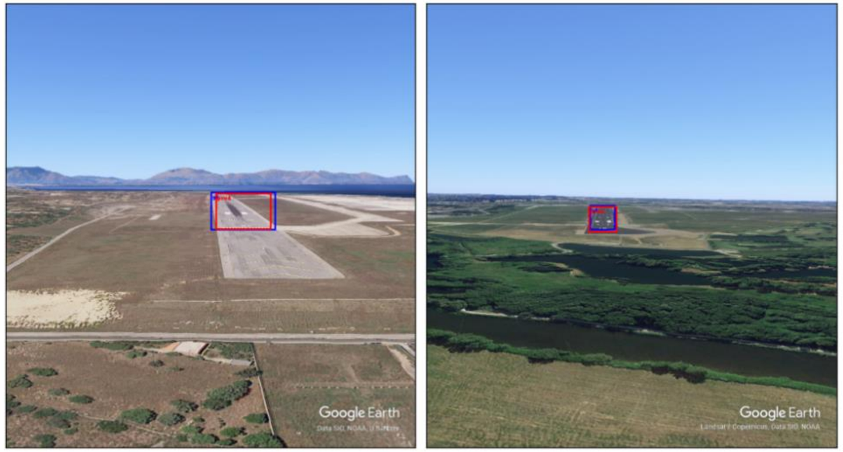
- Object detection has several post-processing (depending on the model)
  - Threshold on objectness score
  - Non Maximum suppression with a IoU threshold
  - mAP computation: IoU with GT, AUC computation
- So the global performance doesn’t rely only on the NN certificates
- Work in progress:
  - Certificates on Objectness only
  - Certificates on classification head
- Empirical studies

How to learn LipNet for Object Detection

- In Object Detection most efficient architecture use transformers, or complex architecture (Yolo)
- Several simpler but efficient convolutional architecture exist:
  - FCOS: FCOS: Fully Convolutional One-stage Object Detection is an anchor-free (**One stage detector**:)
  - **Architecture**:
    - **Backbone (Blue)**: ResNet18 with Lipschitz layers
    - **Feature extractor (Green)**: FPN (*Feature Pyramid Network*) with Lipschitz layers
    - **Heads (Orange)**: Non-Lipschitz for regression (x,y)



FCOS architecture. (Tian et al, 2019) "*FCOS: Fully Convolutional One-Stage Object Detection*"



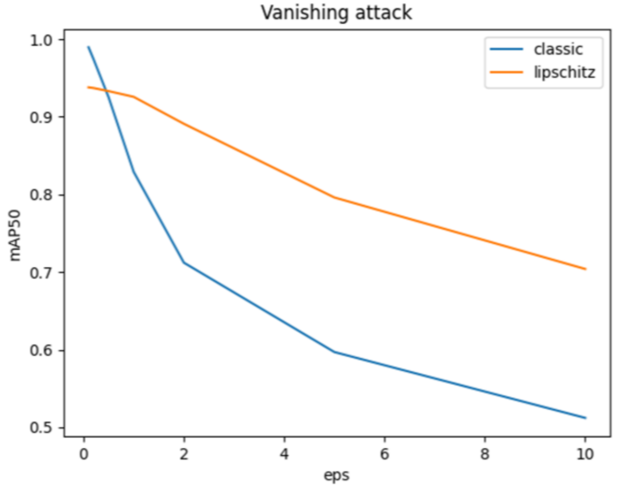
Early results on LARD synthetic test set
(**Blue**: ground truth / **Red**: prediction)

- mAP@50 0.870
- mAP@[50:95] 0.399

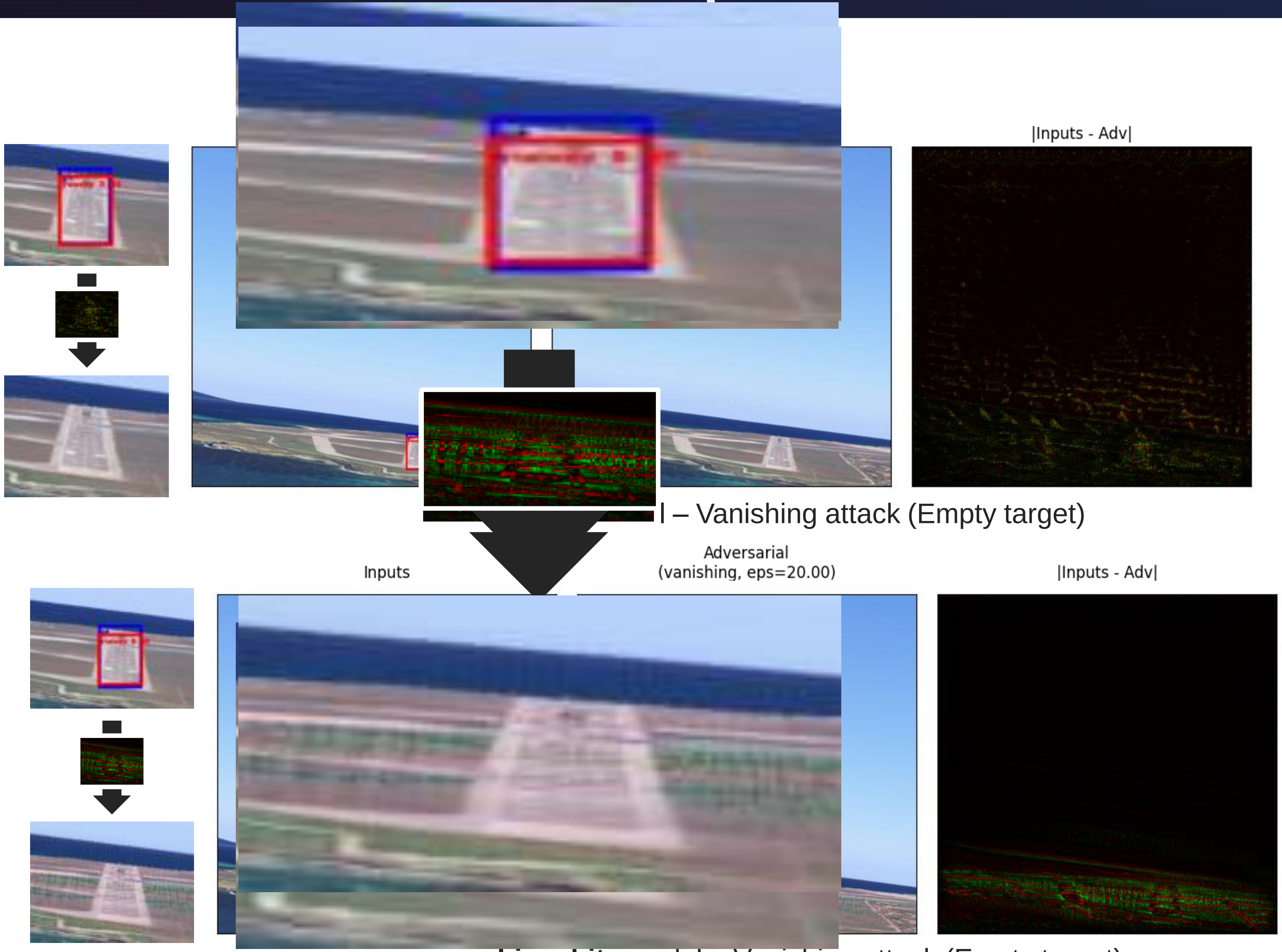
Lard-V1 Object Detection With LipNet

Vanishing attack

Objective: Find minimal perturbation (of L2 norm ϵ) to trick the model into detecting *no more targets*.



Robustness of lipschitz vs classic equivalent models wrt L2 norm *vanishing adversarial attacks*, evaluated using *mAP50 metric* (the higher the better).

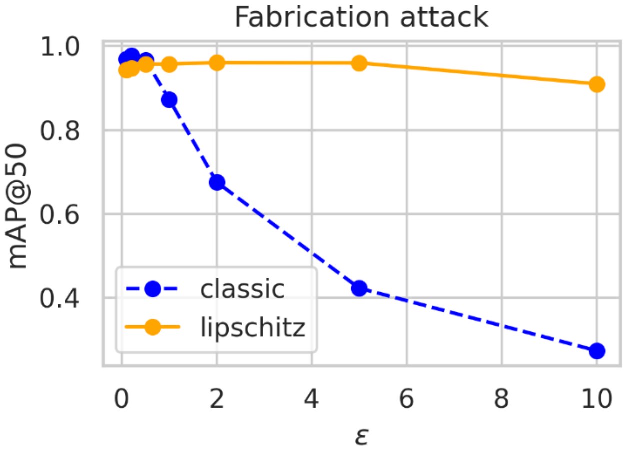


l – Vanishing attack (Empty target)

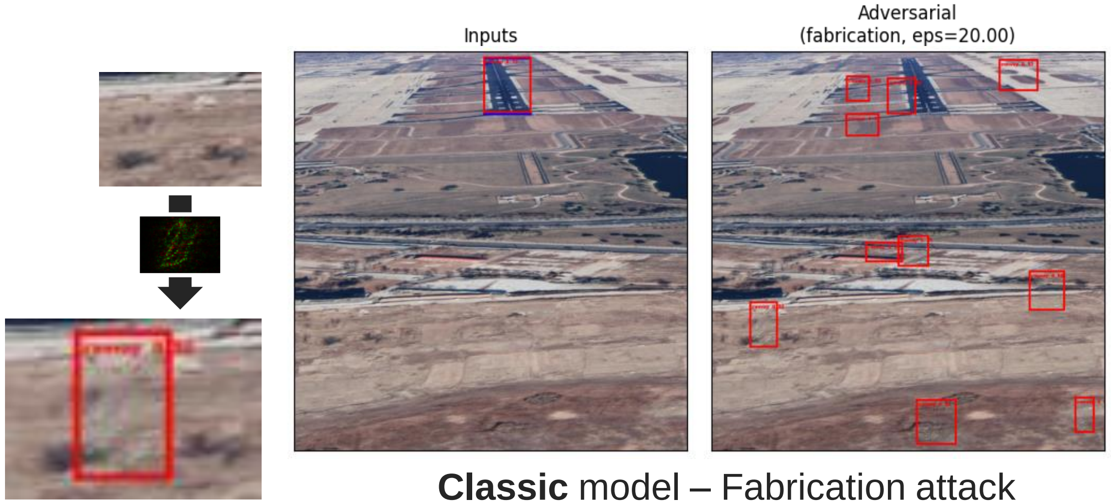
Lipschitz model – Vanishing attack (Empty target)

Robustness to fabrication attack

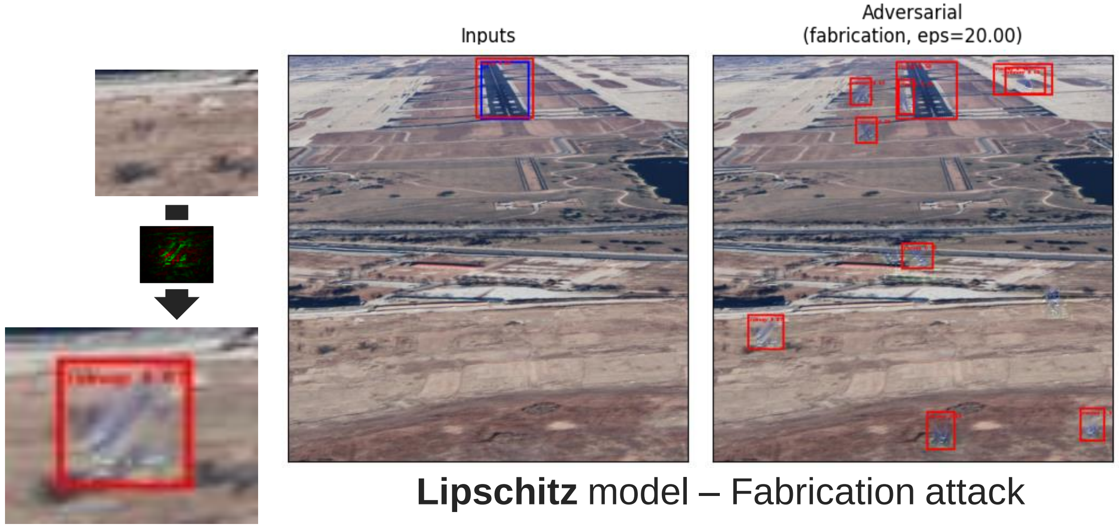
- **Objective:** Find minimal perturbation (of L2 norm ϵ) to trick the model into *detecting false targets* (randomly defined).



Robustness of classic vs Lipschitz equivalent models wrt L2 norm of *fabrication adversarial attacks*, evaluated using ***mAP@50*** *metric* (the higher the better).



Classic model – Fabrication attack



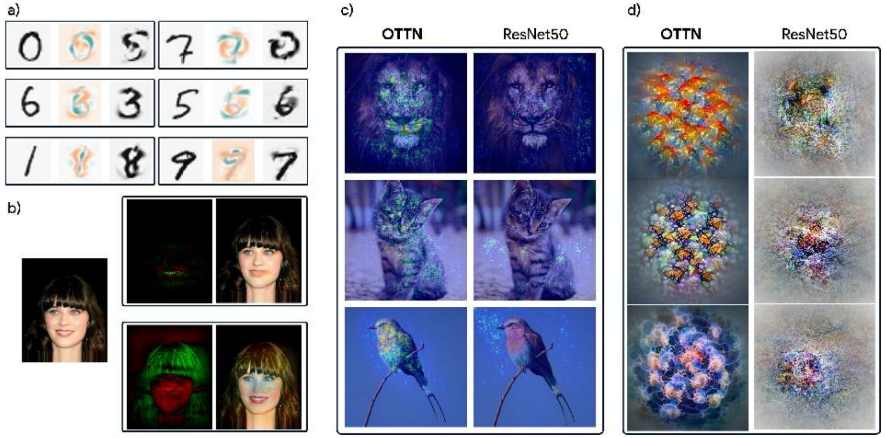
Lipschitz model – Fabrication attack

For very large perturbations, attacks are able to trick both models but the **modifications are only visible on Lipschitz model.**

Extensions and properties of 1-Lipschitz NN

OTNN are explainable by design:

Follow gradient to generate counterfactuals, XAI methods work well, , align to human explainability

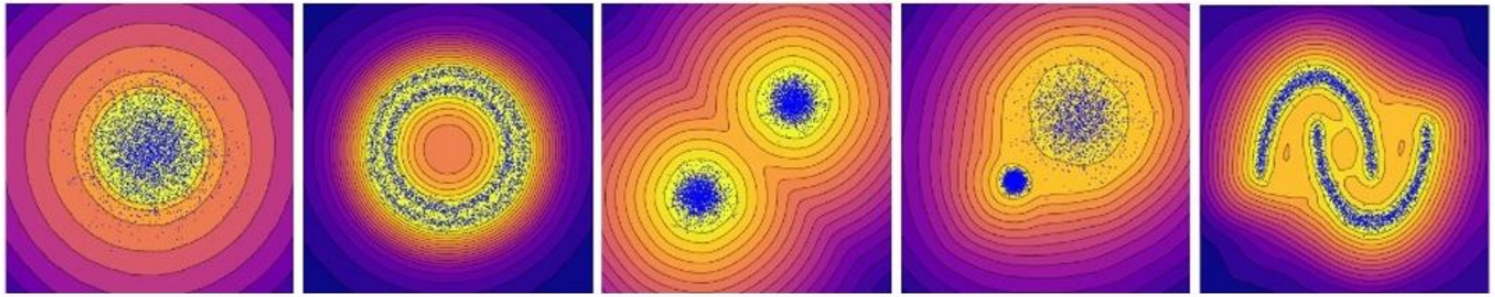


[SER23] “On the explainable properties of 1-Lipschitz Neural Networks: An Optimal Transport Perspective”, Mathieu Serrurier, et al. (https://arxiv.org/abs/2206.06854)

+ DEMO

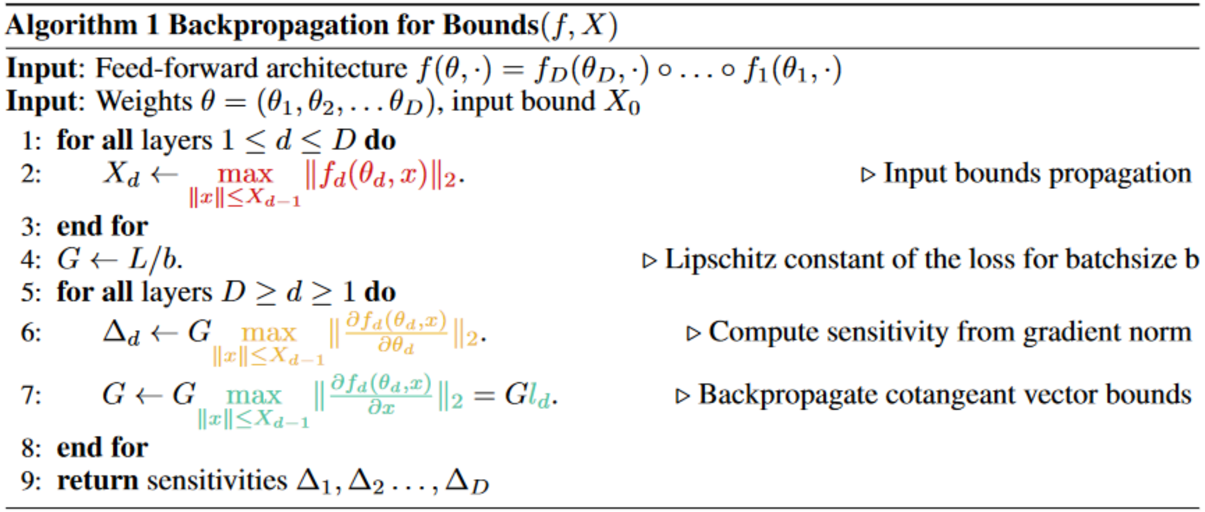
Robust one class classification and anomaly detection

Problem: you want to be able to detect anomalies, but you don’t necessarily have sample of anomalous data



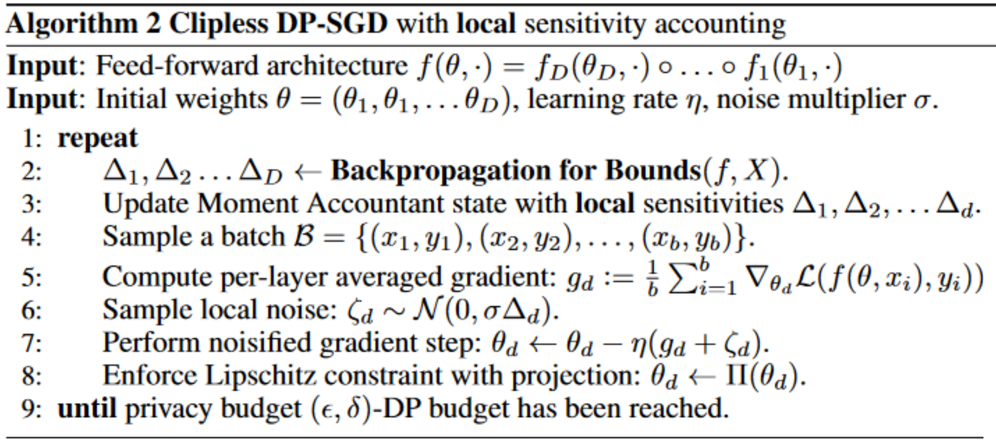
[BETH23] « Robust One-Class Classification with Signed Distance Function using 1-Lipschitz Neural Networks », Louis Bethune, et al. ICML’23 (https://arxiv.org/abs/2303.01978)

1-LIPSCHITZ NN FOR DIFFERENTIAL PRIVACY

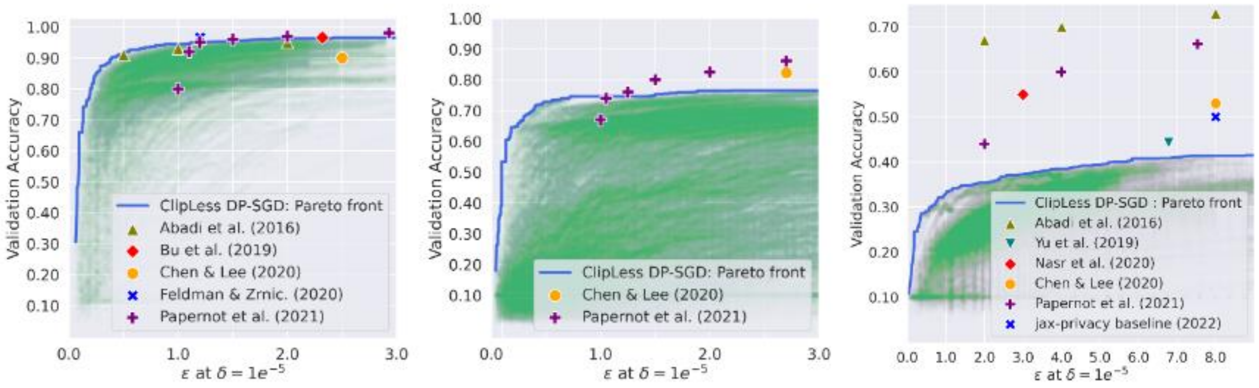


```
Algorithm 1 Backpropagation for Bounds$(f, X)$
Input: Feed-forward architecture $f(\theta,\cdot) = f_D(\theta_D,\cdot) \circ \ldots \circ f_1(\theta_1,\cdot)$
Input: Weights $\theta = (\theta_1, \theta_2, \ldots \theta_D)$, input bound $X_0$
for all layers $1 \leq d \leq D$ do
    $X_d \leftarrow \max_{\|x\| \leq X_{d-1}} \|f_d(\theta_d, x)\|_2.$    ▷ Input bounds propagation
end for
$G \leftarrow L/b.$    ▷ Lipschitz constant of the loss for batchsize b
for all layers $D \geq d \geq 1$ do
    $\Delta_d \leftarrow G \max_{\|x\| \leq X_{d-1}} \|\frac{\partial f_d(\theta_d, x)}{\partial \theta_d}\|_2.$    ▷ Compute sensitivity from gradient norm
    $G \leftarrow G \max_{\|x\| \leq X_{d-1}} \|\frac{\partial f_d(\theta_d, x)}{\partial x}\|_2 = G l_d.$    ▷ Backpropagate cotangeant vector bounds
end for
return sensitivities $\Delta_1, \Delta_2 \ldots, \Delta_D$
```

Upper bound of $||\nabla_\theta f||$ can be computed for 1-Lipschitz or GNP NN



```
Algorithm 2 Clipless DP-SGD with local sensitivity accounting
Input: Feed-forward architecture $f(\theta,\cdot) = f_D(\theta_D,\cdot) \circ \ldots \circ f_1(\theta_1,\cdot)$
Input: Initial weights $\theta = (\theta_1, \theta_1, \ldots \theta_D)$, learning rate $\eta$, noise multiplier $\sigma$.
repeat
    $\Delta_1, \Delta_2 \ldots \Delta_D \leftarrow$ Backpropagation for Bounds$(f, X)$.
    Update Moment Accountant state with local sensitivities $\Delta_1, \Delta_2, \ldots \Delta_d$.
    Sample a batch $\mathcal{B} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_b, y_b)\}$.
    Compute per-layer averaged gradient: $g_d := \frac{1}{b} \sum_{i=1}^{b} \nabla_{\theta_d} \mathcal{L}(f(\theta, x_i), y_i))$
    Sample local noise: $\zeta_d \sim \mathcal{N}(0, \sigma\Delta_d)$.
    Perform noisified gradient step: $\theta_d \leftarrow \theta_d - \eta(g_d + \zeta_d)$.
    Enforce Lipschitz constraint with projection: $\theta_d \leftarrow \Pi(\theta_d)$.
until privacy budget $(\epsilon, \delta)$-DP budget has been reached.
```



(a) MNIST.

(b) F-MNIST.

(c) CIFAR-10.

DEEL
France-Quebec

Thank you for your attention

IVADO · IRT SAINT EXUPÉRY · CRIAQ · ANITI Artificial & Natural Intelligence Toulouse Institute · Université Fédérale Toulouse Midi-Pyrénées · Institut intelligence et données