25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION,
Milan, Italy
10-15 January 2021

OPTIMAL TRANSPORT AS A DEFENSE AGAINST ADVERSARIAL ATTACKS

Quentin Bouniot, Romaric Audigier, Angélique Loesch

Université Paris-Saclay, CEA, List,
F-91120, Palaiseau, France

ADVERSARIAL EXAMPLES AND PERTURBATION

Inputs:

Original

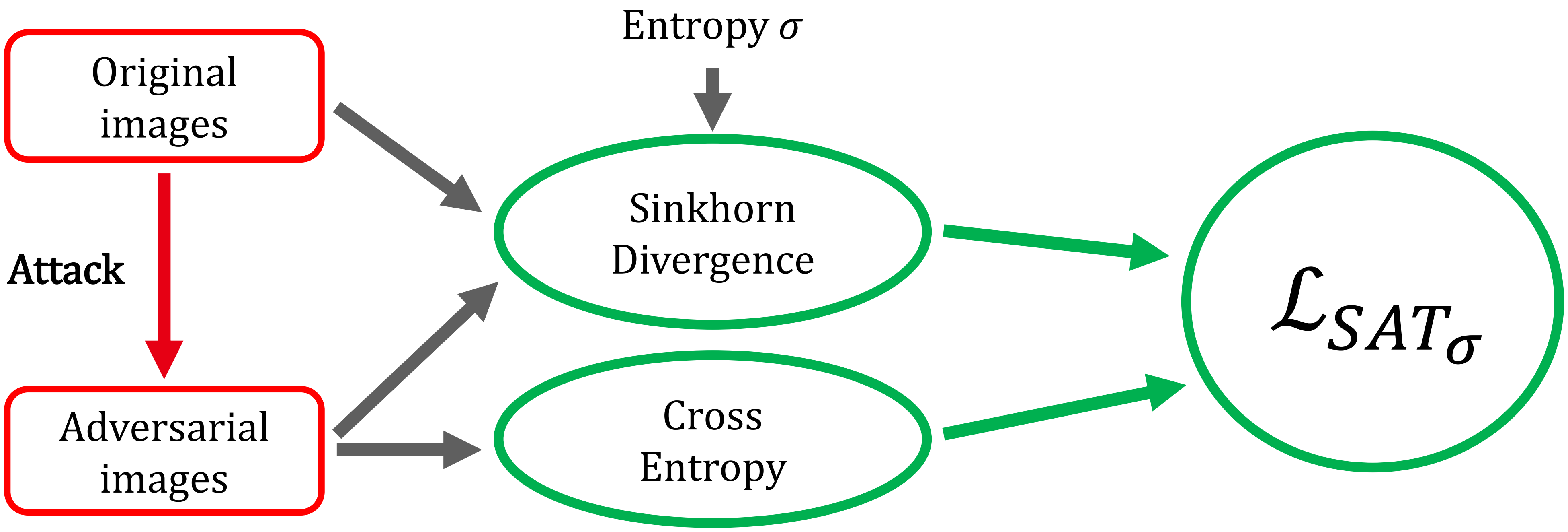
$\epsilon = 16$

$\epsilon = 30$

Predictions: frog ✓ deer ✗ deer ✗

- **Adversarial example:**
  - Human-imperceptible perturbation for a given image to mislead a model.
  - Most effective defenses based on adversarial training align *original* and *adversarial* representations.
- **Problems:**
  - Defenses are *partially* aligning moments of distributions.
  - Current evaluation use a *fixed* perturbation size ϵ that can *differ* between papers.

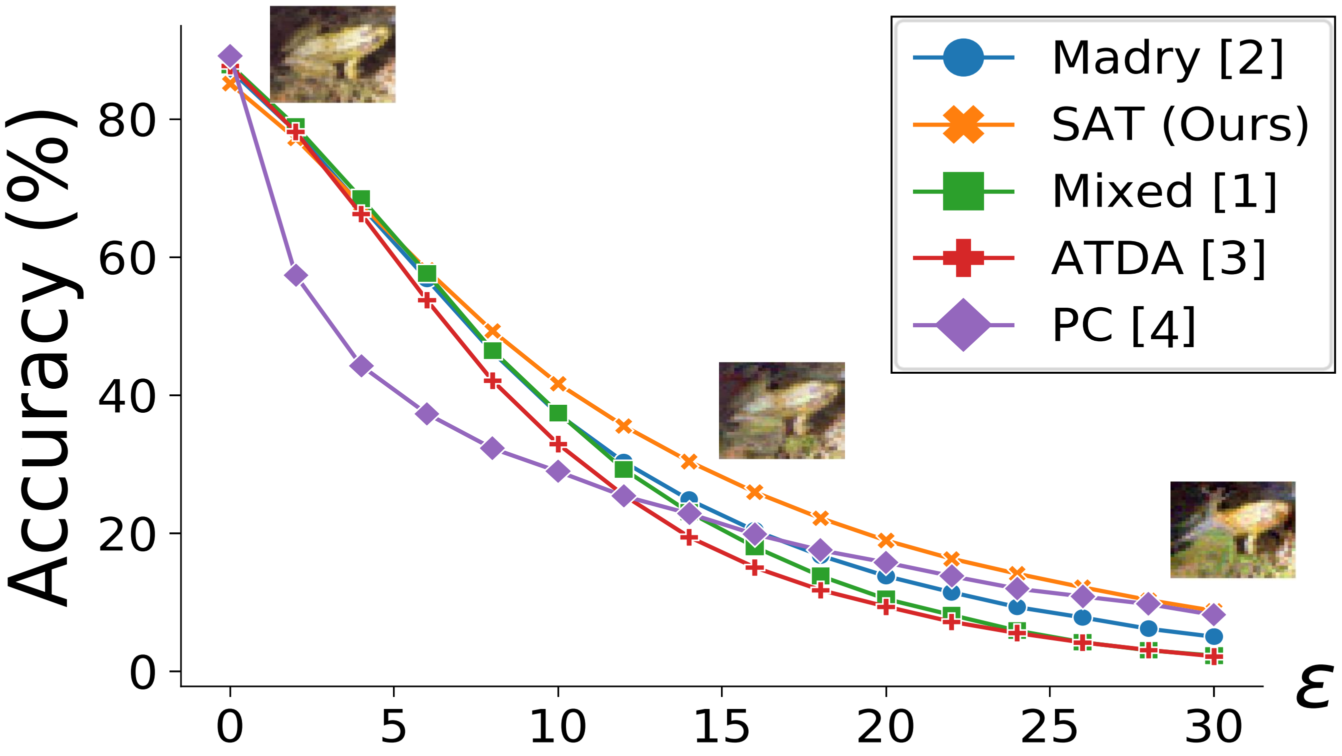
SINKHORN ADVERSARIAL TRAINING (SAT)



- **Sinkhorn Adversarial Training (SAT):**
  - Our defense is based on recent theory of **Optimal Transport** [5] to consider the *whole* distributions and reflect *geometric properties*.

[5] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouve, and G. Peyré, "*Interpolating Between Optimal Transport and MMD using Sinkhorn Divergences*," in Proceedings of Machine Learning Research (PMLR), 2019.

EXPERIMENTAL RESULTS I



- A *fixed* perturbation size does not fully compare robustness.
- Our **SAT** is globally more robust than other SOTA defenses.

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "*Explaining and harnessing adversarial examples*,"in International Conference on Learning Representations (ICLR), 2014.
[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "*Towards deep learning models resistant to adversarial attacks*",in International Conference on Learning Representations (ICLR), 2018.
[3] C. Song, K. He, L. Wang, and J. E. Hopcroft, "*Improving the generalization of adversarial training with domain adaptation*," in International Conference on Learning Representations (ICLR), 2019.
[4] A. Mustafa, S. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao, "*Adversarial defense by restricting the hidden space of deep neural networks*," in International Conference on Computer Vision (ICCV), 2019.

AREA UNDER ACCURACY CURVE

- **Area Under Accuracy Curve (AUAC):**
  - A new metric for robustness:

$$AUAC_{\epsilon_{max}}(f) = \frac{1}{\epsilon_{max}} \int_{\epsilon=0}^{\epsilon_{max}} Acc(f, \epsilon, \mathbf{D}^{ts}) d\epsilon$$

$Acc(f, \epsilon, \mathbf{D}^{ts})$ is the accuracy of f on the test set $\mathbf{D}^{\mathbf{ts}}$ with perturbations of size up to ϵ.

- **AUAC quantifies more completely robustness to adversarial attacks.**
  - Takes into account a wide range of perturbation sizes.

EXPERIMENTAL RESULTS II

Dataset	Archi.	Model	AUAC (%)	
			$\epsilon_{max} = 16$	$\epsilon_{max} = 30$
CIFAR-10	Resnet20	Non-defended	5.79	3.09
		Madry [2]	44.18	26.53
		Mixed [1]	40.68	22.73
		ATDA [3]	35.58	21.63
		SAT (Ours)	**44.26**	**29.69**
	Resnet110	PC [4]	37.89	26.47
	WideResnet28-10	Non-defended	8.8	4.69
		Madry [2]	49.37	31.54
		Mixed [1]	49.27	30.01
		ATDA [3]	46.19	27.94
		SAT (Ours)	**51.93**	**35.12**
CIFAR-100	WideResnet28-10	Non-defended	6.03	3.22
		Madry [2]	27.27	16.14
		Mixed [1]	27.80	16.13
		ATDA [3]	28.59	17.11
		SAT (Ours)	**29.69**	**19.83**

Original

$\epsilon = 16$

$\epsilon = 30$

- **Our SAT is the most robust adversarial defense.**
- **Evaluation also depends on the *attack* considered (see our paper for more examples).**

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "*Explaining and harnessing adversarial examples*," in International Conference on Learning Representations (ICLR), 2014.
[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "*Towards deep learning models resistant to adversarial attacks*", in International Conference on Learning Representations (ICLR), 2018.
[3] C. Song, K. He, L. Wang, and J. E. Hopcroft, "*Improving the generalization of adversarial training with domain adaptation*," in International Conference on Learning Representations (ICLR), 2019.
[4] A. Mustafa, S. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao, "*Adversarial defense by restricting the hidden space of deep neural networks*," in International Conference on Computer Vision (ICCV), 2019.

TAKE HOME MESSAGE

- We propose **Sinkhorn Adversarial Training (SAT)**, a defense that *fully* aligns distributions of *original* and *adversarial* representations by using Optimal Transport.
- We propose the **Area Under Accuracy Curve (AUAC)**, a metric of robustness for a *fair* and *exhaustive* evaluation of defenses.
- Our proposed defense is **globally more robust** than previous methods.

Contact:

https://qbouniot.github.io

quentin.bouniot@cea.fr

@QBouniot

25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION,

Milan, Italy

10-15 January 2021

Quentin Bouniot, Romaric Audigier, Angélique Loesch

Thank you for listening !

Commissariat à l'énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

Établissement public à caractère industriel et commercial | RCS Paris B 775 685 019