Constructing Semantics-Aware Adversarial Examples with a Probabilistic Perspective

Andi Zhang [1] Mingtian Zhang [2] Damon Wischik [1]

[1]Department of Computer Science and Technologies, University of Cambridge
[2]Centre for Artificial Intelligence, University College London

Review: Adversarial Examples

Given a victim classifier $C : [0,1]^n \to \mathcal{Y}$, an image $x_{\text{ori}} \in [0,1]^n$ and a target label $y_{\text{tar}} \in \mathcal{Y}$, the optimization problem for finding an adversarial instance x_{adv} for x_{ori} can be formulated as follows [2]:

$$\min \mathcal{D}(x_{\text{ori}}, x_{\text{adv}}) \quad \text{s.t. } C(x_{\text{adv}}) = y_{\text{tar}} \text{ and } x_{\text{adv}} \in [0,1]^n \tag{1}$$

Here, $\mathcal{D}$ is a distance can be represented by $\mathcal{L}_1$, $\mathcal{L}_2$, or $\mathcal{L}_\infty$ norms.

However, solving (1) is challenging. [2] propose a relaxation:

$$\min c_1 \mathcal{D}(x_{\text{ori}}, x_{\text{adv}}) + c_2 f(x_{\text{adv}}, y_{\text{tar}}) \quad \text{s.t. } x_{\text{adv}} \in [0,1]^n \tag{2}$$

where c_1, c_2 are constants, and f is an objective function closely tied to the classifier C's prediction. For example, in [2], f could be the cross-entropy loss function.

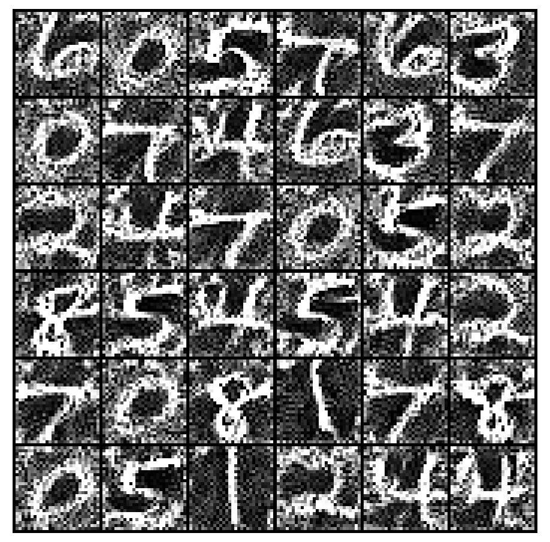
Adversarial Examples: A Probabilistic Perspective

By applying box-constrained Langevin dynamics (LD) [1] as an optimization method to (2), we get a Gibbs distribution:

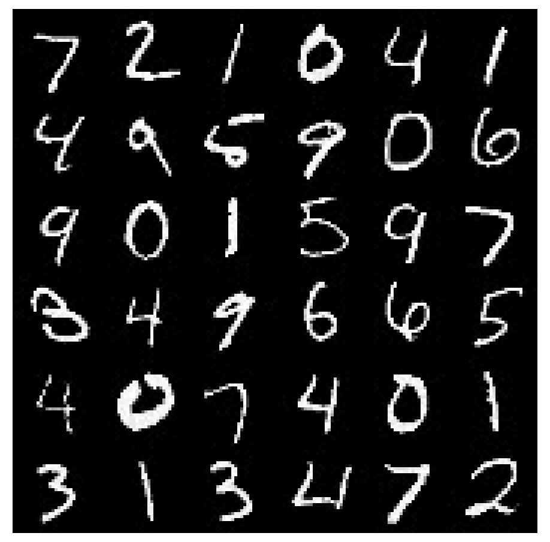
$$p_{\text{adv}}(x_{\text{adv}}|x_{\text{ori}}, y_{\text{tar}}) \propto p_{\text{vic}}(x_{\text{adv}}|y_{\text{tar}}) p_{\text{dis}}(x_{\text{adv}}|x_{\text{ori}}) \tag{3}$$

where $p_{\text{vic}}(x_{\text{adv}}|y_{\text{tar}}) \propto \exp(-c_2 f(x_{\text{adv}}, y_{\text{tar}}))$ and $p_{\text{dis}}(x_{\text{adv}}|x_{\text{ori}}) \propto \exp(-c_1 \mathcal{D}(x_{\text{ori}}, x_{\text{adv}}))$.

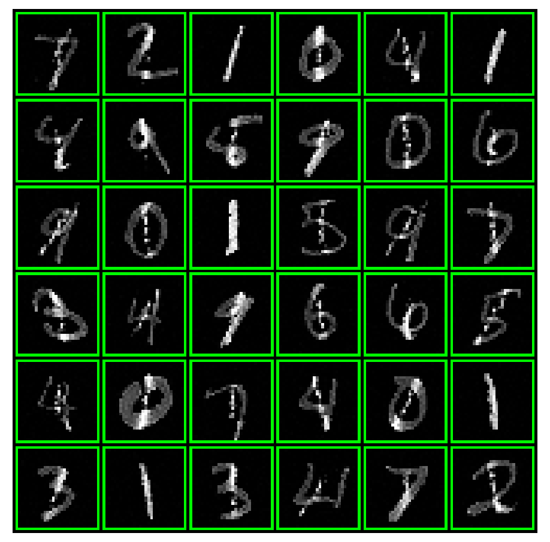
Victim Distribution. By LD, we can sample from p_{vic} of the adversarially trained victim classifier. The samples (right) display clear digit structures, indicating the robust classifier's semantic understanding and resistance to deception.



Distance Distribution. Let $\mathcal{D}$ be the squared $\mathcal{L}_2$ norm, defined as $\mathcal{D}(a,b) = \|a-b\|_2^2$ - a common choice in adversarial attacks. The resulting distance distribution p_{dis} is a Gaussian distribution, with samples (for a suitable constant c_1) shown on the right.



Adversarial Distribution. As illustrated in (3), the adversarial distribution is the product of p_{vic} and p_{dis}. The samples drawn from p_{adv} (right) use appropriate values of c_1, c_2 and target class $y_{\text{tar}} = 1$. Green borders indicate successful attacks.



So What? What Can This Perspective Bring to Us?

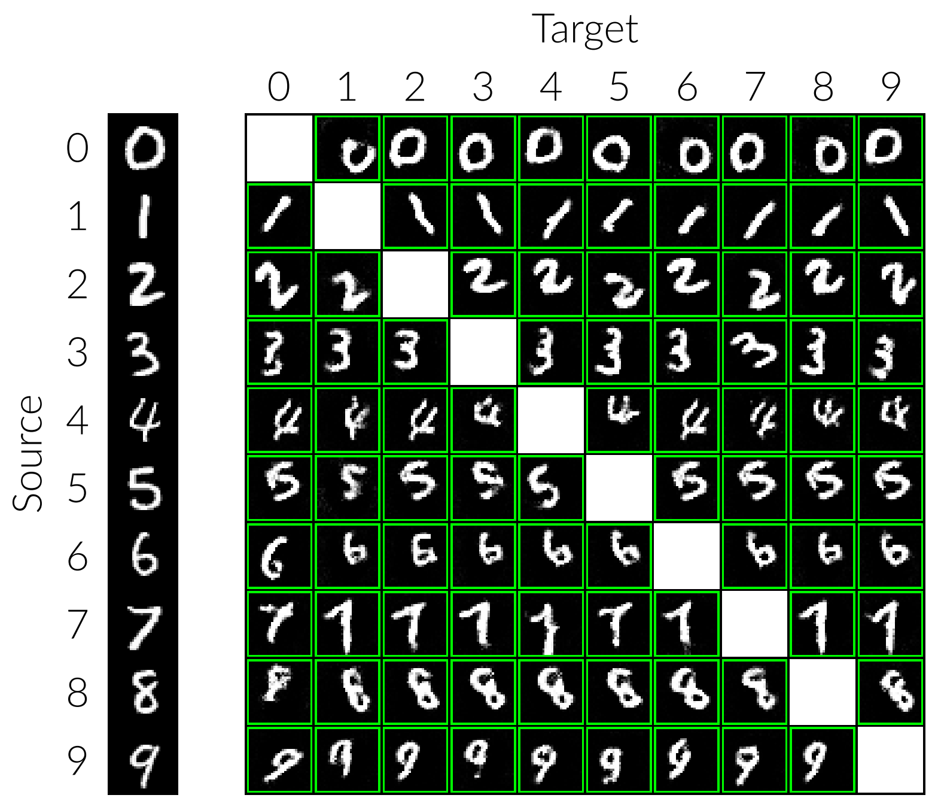
The probabilistic perspective enables us to replace traditional geometry-based distance distributions with those fitted by modern probabilistic generative models (PGMs). Using PGMs as distance distributions offers two key advantages:

Semantic Injection: We can incorporate subjective semantic understanding by training a PGM on semantic-preserving transformations of the original input data x_{ori}.

Model Adaptation: We can leverage pretrained PGMs to create semantically similar distributions around a single image x_{ori} through fine-tuning.

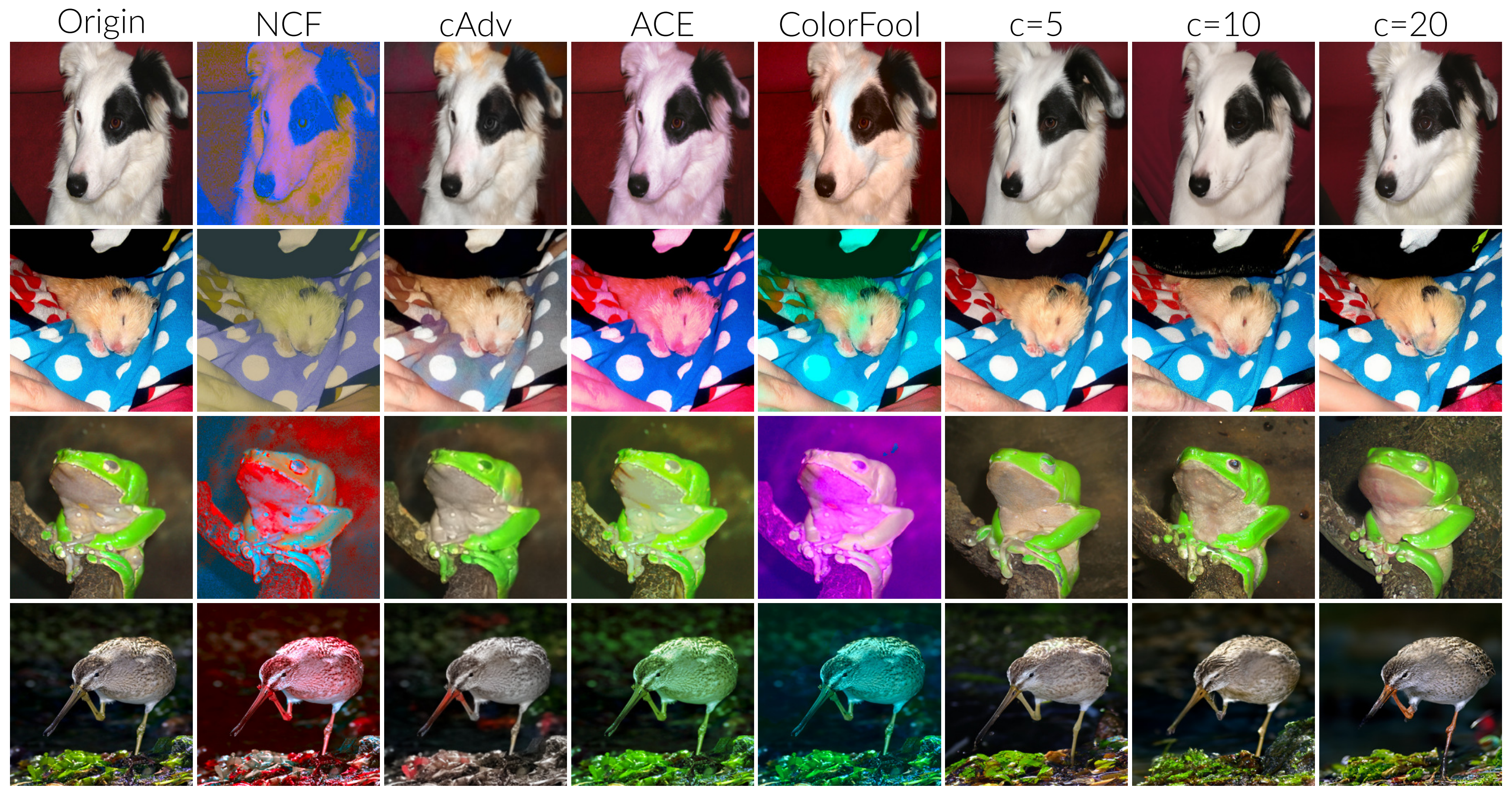
Semantics-Invariant Data Augmentation

Let $\mathcal{T}$ be transformations that preserve x_{ori}'s semantics. Training a PGM on $t_i(x_{\text{ori}})$ where $t_i \in \mathcal{T}$ defines p_{dis}. This allows encoding semantic beliefs into p_{dis} – e.g., including **rotation, scaling and distortion** in $\mathcal{T}$ if they're deemed semantics-preserving. The adversarial examples based on this p_{dis} is shown on the right.



Fine-Tuning Pretrained PGMs

By finetuning a pretrained PGM on x_{ori}, we obtain p_{dis} around x_{ori}. Below are samples drawn from p_{adv} compared with other methods.



Concrete PGM Implementations

We introduce two concrete PGM implementations for p_{dis}:

Energy-based Model: Let E_θ denote the energy in the EBM that is trained / finetuned around x_{ori}, then the adversarial distribution is

$$p_{\text{adv}}(x_{\text{adv}}|x_{\text{ori}}, y_{\text{tar}}) \propto e^{-cf(x_{\text{adv}}, y_{\text{tar}})} e^{-E_\theta(x_{\text{adv}})}$$

with the corresponding Stein score:

$$\nabla_{x_{\text{adv}}} \log p_{\text{adv}}(x_{\text{adv}}|\ldots) = -c \nabla_{x_{\text{adv}}} f(x_{\text{adv}}, y_{\text{tar}}) - \nabla_{x_{\text{adv}}} E_\theta(x_{\text{adv}}) \tag{4}$$

Samples from p_{adv} can then be drawn using Langevin dynamics (LD).

Diffusion Model: Let θ be the parameters of a diffusion model that is trained / finetuned around x_{ori}, then

$$\begin{aligned} p_{\text{adv}}(x_0|x_{\text{ori}}, y_{\text{tar}}) &\propto p_{\text{vic}}(x_0|y_{\text{tar}}) \int p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) dx_{1\ldots T} \\ &\approx \int p(x_T) \prod_{t=1}^{T} p_{\text{vic}}(\hat{x}_{0|t}|y_{\text{tar}})^{1/T} p_\theta(x_{t-1}|x_t) dx_{1\ldots T} \end{aligned}$$

where $\hat{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t))$ is estimated by Tweedie's formula as we cannot obtain x_0 at every sampling step. Theorem 2 in the paper shows that $p_{\text{vic}}(\hat{x}_{0|t}|y_{\text{tar}})^{1/T} p_\theta(x_{t-1}|x_t)$ is Gaussian. Algorithm 2 in the paper illustrates how to draw samples from this p_{adv}.

FAQs

Just Another Diffusion + Adv. Attack Work? No. While combining these methods may seem intuitive, previous works have largely approached this combination in an ad hoc manner. In contrast, our work is derived from the classical optimization problem defined in (2) and provides a principled probabilistic perspective on adversarial examples. For example, as shown in (4), our method naturally leads to the addition of generative and adversarial gradients through mathematical derivation, rather than through intuitive design.

How to Defend it? Our method circumvents robust classifiers since they are typically trained against conventional adversarial examples, not our semantic-preserving ones. However, defenders could adapt by incorporating our generated examples into their training process, creating a new form of adversarial training.

References

[1] A. Lamperski. Projected stochastic gradient langevin algorithms for constrained sampling and non-convex learning. In *Conference on Learning Theory*, pages 2891–2937. PMLR, 2021.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.