Concept-based Adversarial Attack: a Probabilistic Perspective

Andi Zhang [a,b] **Xuan Ding** [c] **Steven McDonagh** [d] **Samuel Kaski** [b,e]

[a] University of Warwick [b] University of Manchester
[c] CUHK (Shenzhen) [d] University of Edinburgh [e] Aalto University

ICLR

The AI Hub in Generative Models

Review: Adversarial Attack

Adversarial Attack has two aims:

- **Deceiving victim classifiers**
- **Deceiving human**

Given an image $x_{\text{ori}} \in [0,1]^n$ and a target label $y_{\text{tar}} \in \mathcal{Y}$, the optimization problem [1] to find an adversarial example x_{adv}:

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

where $\mathcal{D}$ is a distance function (commonly chosen from ℓ_1, ℓ_2, or ℓ_∞), and f is an objective function tied to the victim classifier's prediction (e.g., the cross-entropy loss).

Probabilistic Adversarial Attack

By applying the **box-constrained Langevin dynamics** as an optimization method to (1), we derive a probabilistic perspective of adversial attack [2]:

$$p_{\text{adv}}(x_{\text{adv}}|x_{\text{ori}}, y_{\text{tar}}) \propto p_{\text{vic}}(x_{\text{adv}}|y_{\text{tar}})\, p_{\text{dis}}(x_{\text{adv}}|x_{\text{ori}}) \qquad (2)$$

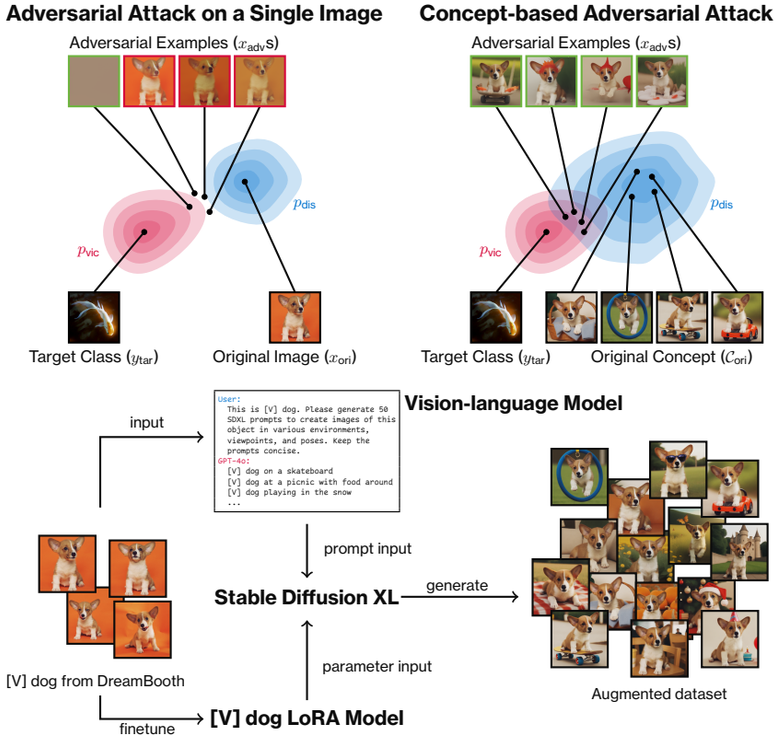
where $p_{\text{vic}}(x_{\text{adv}}|y_{\text{tar}}) \propto \exp(-c_2\, f(x_{\text{adv}}, y_{\text{tar}}))$ is the victim distribution and $p_{\text{dis}}(x_{\text{adv}}|x_{\text{ori}}) \propto \exp(-c_1\, \mathcal{D}(x_{\text{ori}}, x_{\text{adv}}))$ is the distance distribution.

If $\mathcal{D}$ is ℓ_1, then p_{dis} is a Laplace distribution; if $\mathcal{D}$ is squared ℓ_2, then p_{dis} is a Gaussian distribution.

We suggest replacing geometry-induced distributions with **PGM-learned** distributions for p_{dis} [2].

Concept-based Adversarial Attack

The probabilistic perspective naturally extends $p_{\text{dis}}(\cdot|x_{\text{ori}})$ to $p_{\text{dis}}(\cdot|\mathcal{C}_{\text{ori}})$.



Generated Adversarial Examples

Green: successfully fools the victim (surrogate) classifier; **Red**: fails.

Reference

[1] Szegedy, Christian, et al. "Intriguing properties of neural networks." ICLR (2014).
[2] Zhang, Andi, Mingtian Zhang, and Damon Wischik. "Constructing semantics-aware adversarial examples with a probabilistic perspective." NeurIPS (2024).

MANCHESTER 1824
The University of Manchester

THE UNIVERSITY OF EDINBURGH

A!
Aalto University

UNIVERSITY OF WARWICK