LipSim: A Provably Robust Perceptual Similarity Metric

A Defense Mechanism for Perceptual Metrics against Adversarial Attacks

Sara Ghazanfari, Alexandre Araujo,
Prashanth Krishnamurthy, Farshad Khorrami, Siddharth Garg

New York University

April 18, 2024

Similarity Metrics

Low-level Metrics

- Point-wise metrics Including ℓ_p norms.
- Fail to capture the high-level structure, and the semantic concept.

Perceptual Similarity Metrics

- Neural networks are used as feature extractors.
- Low-level metrics are employed in the embeddings of images in the new space.
  - **LPIPS** *(R Zhang)*: a convolutional neural network
  - **DreamSim** *(S Fu)*: an ensemble of ViT-based models

Perceptual metrics align better with human perception.

R Zhang, The unreasonable effectiveness of deep features as a perceptual metric (2018) S Fu, DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data (2023)

Training Perceptual Metrics – 2AFC dataset

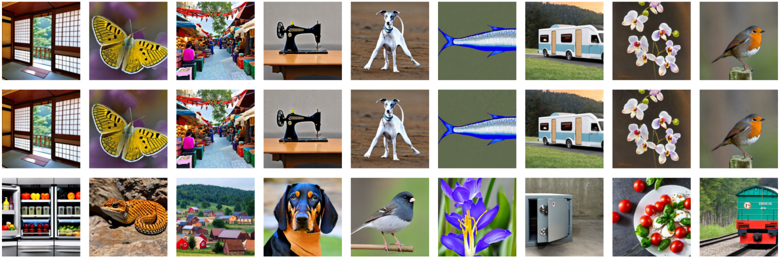
Two-alternative forced choice (2AFC) dataset

- BAPPS *(R Zhang)* dataset.
- NIGHT *(S Fu)* dataset.

R Zhang, The unreasonable effectiveness of deep features as a perceptual metric (2018) S Fu, DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data (2023)

Motivation

❶
❷
❸



d(❶, ❷)	0.64	0.59	0.50	0.76	0.65	0.64	0.62	0.65	0.73
d(❶, ❸)	0.68	0.63	0.54	0.75	0.66	0.64	0.66	0.62	0.75

Perceptual Similarity Metrics are not robust to adversarial attacks!

Method

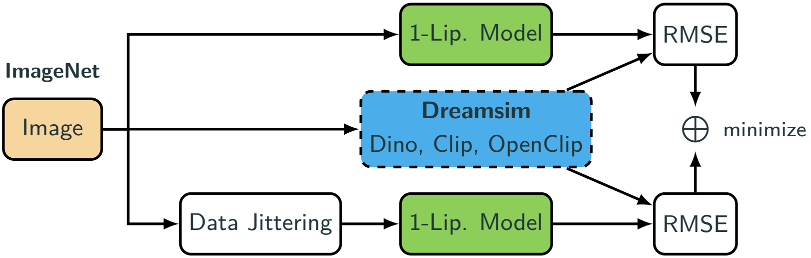
Definition (L_f-Lipschitz function)

Let f be a Lipschitz function with L_f Lipschitz constant in terms of ℓ_2 norm, then we can bound the output of the function by:

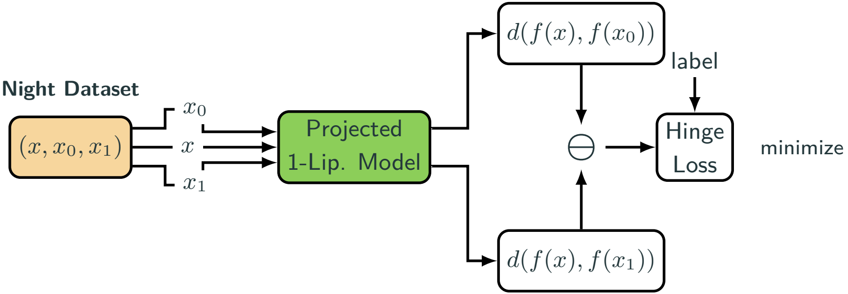
$$\|f(x) - f(x+\delta)\|_2 \leq L_f \|\delta\|_2$$

The Lipschitz constant of neural networks quantifies how much their outputs can change when inputs are perturbed.

Step 1: Lipschitz-based Student-Teacher training of embeddings



Step 2: Lipschitz finetunning on Night Dataset



LipSim

Let $f : \mathcal{X} \to \mathbb{R}^k$ such that:

$$f(x) = \pi_{B_2(0,1)} \circ \phi^{(l)} \circ \cdots \circ \phi^{(1)}(x)$$

where l is the number of layers, $\pi_{B_2(0,1)}$ is a projection on the unit ℓ_2 ball, *i.e.*, $\pi_{B_2(0,1)}(x) = \arg\min_{z \in B_2(0,1)} \|x - z\|_2$.

Let f be the feature extractor function, the LipSim distance metric $d(x_1, x_2)$ is defined as:

$$d(x_1, x_2) = 1 - S_c(f(x_1), f(x_2))$$

Certified Robustness for 2AFC datasets.

$$h(x) = \begin{cases} 1, & d(x, x_1) \leq d(x, x_0) \\ 0, & d(x, x_1) > d(x, x_0) \end{cases}$$

Soft Classifier

Let us define a soft classifier $H : \mathcal{X} \to \mathbb{R}^2$ with respect to some feature extractor f as follows:

$$H(x) = [d(x, x_1), d(x, x_0)]$$

It is clear that $h(x) = \arg\max_{i \in \{0,1\}} H_i(x)$ where H_i represent the i-th value of the output of H.

Theorem

Let $H : \mathcal{X} \to \mathbb{R}^2$ be the soft classifier as defined earlier. Let $\delta \in \mathcal{X}$ and $\varepsilon \in \mathbb{R}^+$ such that $\|\delta\|_2 \leq \varepsilon$. Assume that the feature extractor $f : \mathcal{X} \to \mathbb{R}^k$ is 1-Lipschitz and that for all x, $\|f(x)\|_2 = 1$, then we have the following result:

$$M_{H,x} \geq \varepsilon \|f(x_0) - f(x_1)\|_2 \quad \Longrightarrow \quad M_{H,x+\delta} \geq 0$$

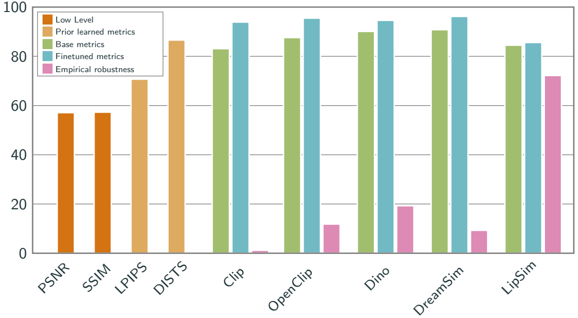
Based on Theorem, and assuming $x_1 \neq x_0$, the certified radius for the classier h at point x can be computed as follows:

$$R(h, x) = \frac{M_{H,x}}{\|f(x_0) - f(x_1)\|_2}$$

Experiments

Experiments – Empirical Score

Metric/ Embedding	Natural Score	ℓ_2-APGD			
		0.5	1.0	2.0	3.0
CLIP	93.91	29.93	8.44	1.20	0.27
OpenCLIP	95.45	72.31	42.32	11.84	3.28
DINO	94.52	81.91	59.04	19.29	6.35
DreamSim	**96.16**	46.27	16.66	0.93	0.93
LipSim (ours)	85.58	**82.89**	**79.82**	**72.20**	**61.84**

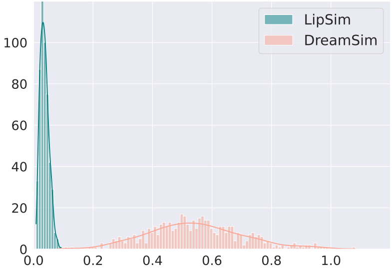


Direct Attack to LipSim

- Direct ℓ_2-PGD attack ($\epsilon = 1.0$) to LipSim and DreamSim by employing the following MSE loss is used during the optimization:

$$\max_{\delta : \|\delta\|_2 \leq \varepsilon} \mathcal{L}_{\mathsf{MSE}} \left[f(x+\delta), f(x) \right]$$

- The histogram of $d(x, x+\delta)$



Thanks for Your Attention!