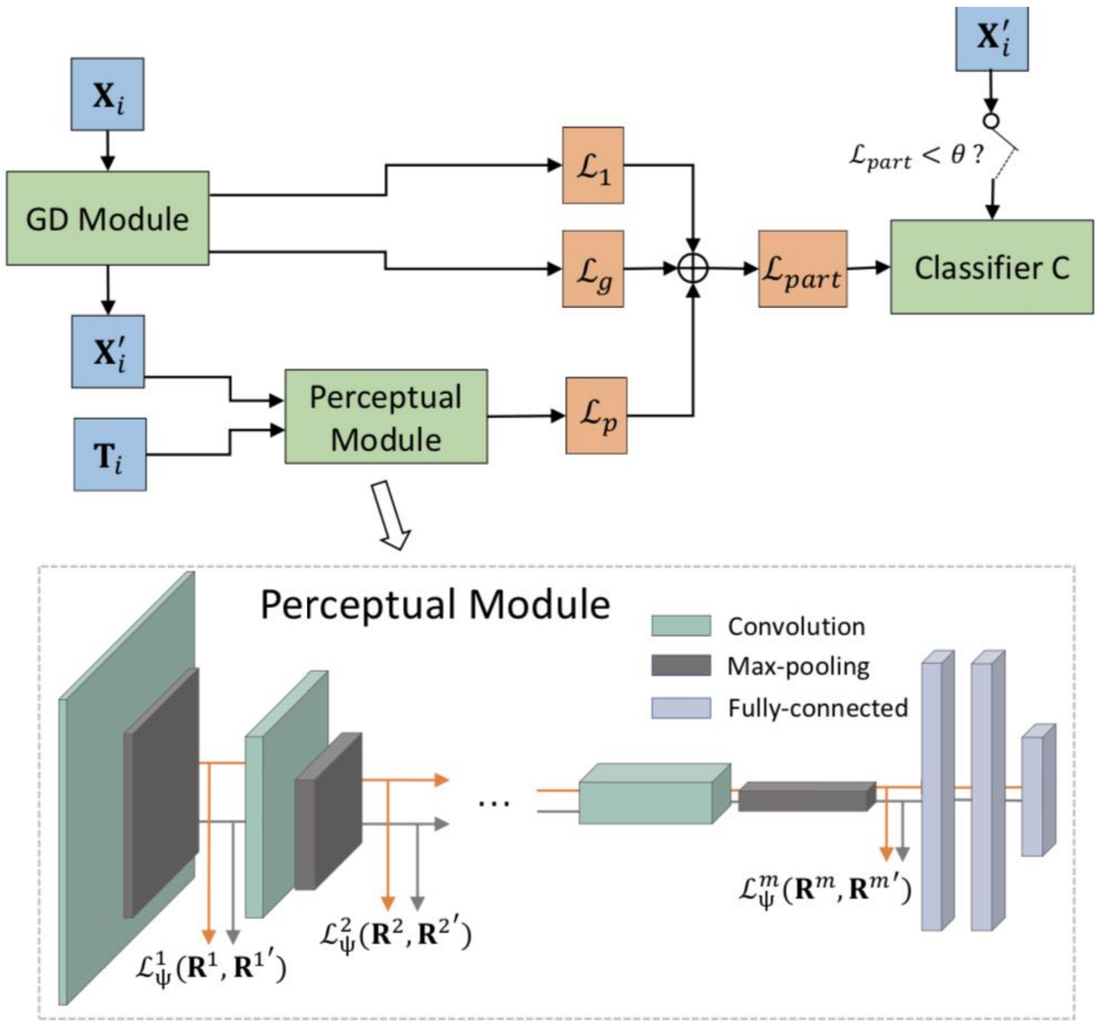
HLR: Generating Adversarial Examples by High-Level Representations

Yuying Hao[1], Tuanhui Li[2], Li Li[2], Yong Jiang[2⋆], and Xuanye Cheng[3]
[1]Tsinghua-Berkely Shenzhen Institute, Tsinghua University, Shenzhen, China
[2]Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
[3]SenseTime Research, SenseTime, Shenzhen, China
Contact Author: (haoyy17@mails.tsinghua.edu.cn)

ICANN19
28th International Conference on Artificial Neural Networks

Model structure



The overall model structure of the HLR attack method. $\mathbf{X}_i$ is the benign image, $\mathbf{X}'_i$ is generated image and $\mathbf{T}_i$ is reference image. GD module produces losses L_1 and L_g, perceptual module generates loss L_p. Classifier C is the state-of-the-art classifier.

The proposed HLR attack

➢ **Loss for GD Module**

$$\mathcal{L}_1 = ||\mathbf{X}'_i - \mathbf{X}_i||_1$$

$$\mathcal{L}_g = E_{G(\mathbf{X}_i)\sim p_{G(\mathbf{X})}}[D(G(\mathbf{X}_i))] - E_{\mathbf{X}_i\sim p_{\mathbf{X}}}[D(\mathbf{X}_i)]$$
$$= E_{\mathbf{X}'_i\sim p_{\mathbf{X}'}}[D(\mathbf{X}'_i)] - E_{\mathbf{X}_i\sim p_{\mathbf{X}}}[D(\mathbf{X}_i)]$$
$$= E_{\mathbf{X}'_i\sim p_{\mathbf{X}'}}[D(\mathbf{X}'_i)]$$

We denote the generator framework as G, the distribution of the images produced by generator as $p_{G(X)}$ and discriminator framework as D, discriminator output as $D(\cdot)$

➢ **Loss for Perceptual Module**

$$\mathcal{L}^l_p = \frac{1}{N^l W^l H^l}||\psi(\mathbf{R}^l) - \psi(\mathbf{R}^{l\prime})||^2$$

$$\mathcal{L}_p = \frac{1}{m}\sum_{l=1}^{m}\mathcal{L}^l_p$$

where $\mathbf{R}^l$ is the input response of a single reference image $\mathbf{T}_i$ in layer l and $\mathbf{R}^{l\prime}$ is the input response of a single generated image $\mathbf{X}'_i$.

➢ **Loss for Classifier C**

$$\mathcal{L}_{KL}(p||q) = \sum p(C(\mathbf{T}_i))log\frac{p(C(\mathbf{T}_i))}{q(C(\mathbf{X}'_i))}$$

where $C(\cdot)$ denotes the output of classifier C, $p(C(T_i))$ is the distribution of $C(T_i)$, $q(C(X'_i))$ is the distribution of $C(X'_i)$.

Contact us by Wechat

Experiments

➢ **Comparison to SOTA on CIFAR-10**

fooling rates / g-model \ d-model		Model1	Model2	Model3
VGG-16	NOISE	7.12	7.75	5.76
	PGD [14]	21.19	23.50	19.17
	CW-L2 [3]	32.56	37.20	32.82
	FGSM [4]	24.29	26.24	23.55
	VIRTUAL [16]	30.97	32.27	28.46
	Ours	**43.62**	**40.78**	**55.74**
VGG-19	NOISE	1.97	3.49	0.75
	PGD [14]	8.26	13.24	6.90
	CW-L2 [3]	21.95	29.43	21.88
	FGSM [4]	24.07	27.44	23.85
	VIRTUAL [16]	28.95	35.14	28.26
	Ours	**43.33**	**37.35**	**29.09**
ResNet-18	NOISE	6.62	6.84	5.98
	PGD [14]	12.68	29.85	11.88
	CW-L2 [3]	30.26	**34.90**	14.46
	FGSM [4]	18.83	20.22	17.6
	VIRTUAL [16]	27.48	31.06	25.51
	Ours	**38.01**	33.87	**30.85**

➢ **Comparison to SOTA on ImageNet**

fooling rates / g-model \ d-model		VGG-16	VGG-19	ResNet-152
VGG-16	UAP [17]	78.30	73.10	63.40
	FFF [19]	47.10	41.98	–[1]
	OPT [13]	100.00	–	**78.00**
	FGA [13]	99.00	–	–
	SA [7]	52.00	60.00	–
	NAG [22]	73.25	**77.50**	54.38
	Ours	**100.00**	66.05	67.8
VGG-19	UAP [17]	73.50	77.80	58.00
	FFF [19]	38.20	–	43.62
	OPT [13]	–	–	–
	FGA [13]	–	–	–
	SA [7]	48.00	60.00	–
	NAG [22]	80.56	83.78	65.43
	Ours	**83.30**	**100.00**	**82.20**
ResNet-152	UAP [17]	47.00	45.50	84.00
	FFF [19]	–	–	–
	OPT [13]	81.00	–	100.00
	FGA [13]	80.00	–	96.00
	SA [7]	–	–	–
	NAG [22]	52.17	53.18	87.24
	Ours	**84.00**	**80.80**	**100.00**

Contribution

(1) We propose a novel network structure to generate adversarial examples.

(2) We introduce a perceptual loss to evaluate the high-level representations differences between benign images and the adversarial examples.