Improving Adversarial Robustness Through The Contrastive-guided Diffusion Process

Yidong Ouyang*, Liyan Xie*, Guang Cheng†

*School of Data Science, The Chinese University of Hong Kong, Shenzhen †Department of Statistics, University of California, Los Angeles

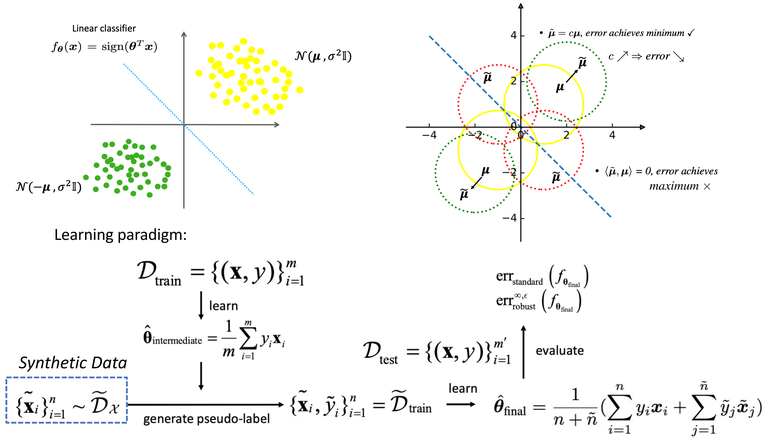
Overview

- Use synthetic data to improve the adversarial robustness
- Present contrastive-guided diffusion model to generate synthetic data
- Shed light on the optimal synthetic distribution for downstream tasks

Motivations

- [1] achieves the state-of-the-art performance on adversarial robustness on the CIFAR-10 dataset by leveraging 100M high-quality synthetic images generated by DDPM [2].
- DDPM is usually **1000 times** slower than GAN [3], which takes more than 99 GPU days on a RTX 4x2080Ti GPU cluster for generating 100M image data.
- **Goal**: improving the sample efficiency of synthetic data, especially for diffusion model.

Theoretical Insights



- (Sample Complexity) Under the parameter setting $\epsilon < \frac{1}{2}$, $\sigma = (nd)^{1/4}$, $\|\boldsymbol{\mu}\|^2 = d$, there exists a universal constant $\tilde{C}$ such that for $\epsilon^2\sqrt{d/n} \geq \tilde{C}$, where n is the number of labeled real data used to construct the intermediate classifier, and additional $\tilde{n}$ synthetic feature generated with mean vector $\pm c\boldsymbol{\mu}$ and pseudo labels, if

$$\tilde{n} \geq \frac{288n}{c}\epsilon^2\sqrt{\frac{d}{n}},$$

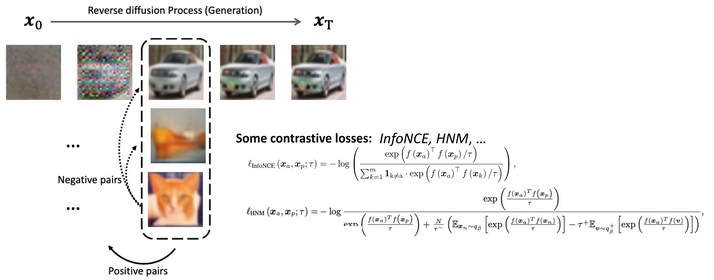
then

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}_{\text{final}}}[\ \text{err}_{\text{robust}}\left(f_{\hat{\boldsymbol{\theta}}_{\text{final}}}\right)] \leq 10^{-3}.$$

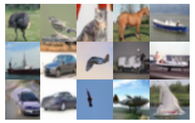
Main Algorithm

Algorithm 1 Generation in Contrasive-guided Diffusion Process (Constrastive-DP)

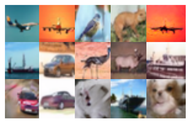
$\mathbb{X}_T = \{\boldsymbol{x}_T^{(i)}\}_{i=1}^N \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$
$t = T$
while $t \neq 1$ **do**
  for i = 1:N **do**
    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$
    Choosing $\boldsymbol{x}_p^{(i)}$ as the positive pair of $\boldsymbol{x}_t^{(i)}$
    $\Delta\boldsymbol{x}_t^{(i)} = \lambda \cdot \nabla_{\boldsymbol{x}_t^{(i)}} \ell_{\text{contrastive}}(\boldsymbol{x}_t^{(i)}, \boldsymbol{x}_p^{(i)}; \tau) + \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t^{(i)}, t\right)$
    $\boldsymbol{x}_{t-1}^{(i)} = \sqrt{\alpha_{t-1}}\left(\frac{\boldsymbol{x}_t^{(i)} - \sqrt{1-\alpha_t}\Delta\boldsymbol{x}_t^{(i)}}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}} \cdot \Delta\boldsymbol{x}_t^{(i)}$
    $t = t - 1$
  end for
end while
return $\mathbb{X}_0 = \{\boldsymbol{x}_0^{(i)}\}_{i=1}^N$



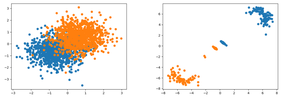
Visualizations

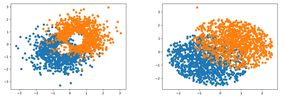


(a) DDIM



(b) Contrasrive-DP





CIFAR-10 Results

	50K		200K		1M	
	clean acc	rob acc	clean acc	rob acc	clean acc	rob acc
DDIM	**84.05%(0.06%)**	56.29%(0.15%)	84.86%(0.43%)	57.83%(0.28%)	85.73%(0.51%)	59.85%(0.26%)
Contrastive-DP	83.66%(0.21%)	**56.60%**(0.17%)	**85.71%(0.18%)**	**58.24%**(0.20%)	**86.30%**(0.09%)	**59.99%**(0.23%)
DDPM	**84.84%(0.37%)**	**56.30%(0.06%)**	85.23%(0.25%)	58.28%(0.09%)	**86.86%(0.04%)**	59.03%(0.16%)
Contrastive-DP	84.70%(0.43%)	56.18%(0.10%)	**85.61%(0.14%)**	**58.62%**(0.12%)	86.30%(0.10%)	**59.74%**(0.26%)

Ablation Study

	50K		200K		1M	
	clean acc	rob acc	clean acc	rob acc	clean acc	rob acc
DDIM (Separability)	79.93%	49.49%	85.09%	54.90%	84.87%	56.08%
Contrastive-DP (Gradient norm)	80.41%	49.47%	84.64%	55.17%	**86.36%**	57.11%
Contrastive-DP (Gradient norm-rob)	83.91%	**55.23%**	84.78%	55.42%	85.93%	**57.18%**
Contrastive-DP (Entropy)	**83.66%**	53.91%	**85.71%**	**55.79%**	86.30%	56.84%

Selected References

[1] Gowal, Sven et al. Improving Robustness using Generated Data. NeurIPS (2021).
[2] Ho, Jonathan et al. Denoising Diffusion Probabilistic Models. NeurIPS (2020).
[3] Song, Jiaming et al. Denoising Diffusion Implicit Models. ICLR (2021).