Stochasticity and Smoothness: Investigating and Strengthening Diffusion Purification

Anonymous submission

Abstract

Recent works have introduced diffusion-based adversarial purification (diffusion purification), claiming the ability to eliminate adversarial noise by applying significant Gaussian noise during the forward process of diffusion. Our extensive and thorough evaluation of different purification-classification pipelines revealed that the reverse process also significantly impacts purification robustness. Through empirical studies utilizing noise immobilization and visualization of the loss landscape under varying settings of model stochasticity, we pinpointed the stochasticity embedded in both the forward and reverse processes as the primary factor driving diffusion purification. Based on our insights on the stochasticity-embedded models, we introduce a novel concept: robustness loss under model stochasticity, and highlight two distinct pathways to achieve robustness: stochasticity and smoothness. To combine these two pathways for better robustness, we proposed a novel approach termed classifier-guided adversarial training, synergizing stochasticity and smoothness-induced robustness. Extensive experimental results validate its efficacy, particularly in scenarios with few diffusion purification steps, and affirm the compatibility of these two forms of robustness.

Introduction

Deep learning has demonstrated remarkable success across various domains, including computer vision (He et al. 2016), natural language processing (OpenAI 2023), and speech recognition (Radford et al. 2022). However, within this thriving landscape, the persistent specter of adversarial attacks casts a shadow over the reliability of these models.

Adversarial attacks involve injecting imperceptible perturbations into legitimate inputs, coaxing models into generating erroneous outputs with unwavering confidence (Goodfellow, Shlens, and Szegedy 2015; Szegedy et al. 2014). In response, great efforts have been made for adversarial defense (Zhang et al. 2019; Samangouei, Kabkab, and Chellappa 2018; Shafahi et al. 2019; Wang et al. 2023; Nie et al. 2022). Adversarial training (Madry et al. 2018), among these methods, stands as a prevalent in practice for enhancing the adversarial robustness of DNNs.

Recent research addressed the concept of purifying adversarial noise with diffusion models (Nie et al. 2022; Wang et al. 2022; Wu, Ye, and Gu 2022; Xiao et al. 2022), proposing that the introduction of Gaussian noise during the forward process can effectively neutralize adversarial noise. However, the investigation of this concept has been impeded by GPU memory constraints encountered in the reverse process of diffusion model back-propagation. To overcome this limitation, we harness the gradient checkpointing (Chen et al. 2016).

Tests on various sampling methods reveal an intriguing phenomenon: despite utilizing identical models and achieving similar generation outcomes, different sampling methods lead to distinct levels of purification robustness. Further investigation through the noise immobilization experiment highlights a pivotal divergence between DDIM and DDPM—stochasticity within the reverse process. This discrepancy underscores the influence of stochastic attributes in both forward and reverse processes on overall robustness.

We further evaluated stochasticity's impact via the loss landscape. Our findings revealed an interesting dynamic: while diffusion models demonstrate vulnerability in malicious perturbation at distinct noise settings, the introduction of stochasticity which leads to simultaneously attacking across various noise settings proved infeasible. This observation underscores the pivotal roles that randomness and expectation play in bolstering model robustness.

To delve deeper into the interplay between model robustness and stochasticity, we introduced a novel concept termed "robustness loss under model stochasticity". Decomposing this loss into clean loss and robustness loss and employing Gaussian noise mix-up to control model stochasticity, we evaluated robustness across different stochasticity levels. This experiment concluded that robustness stems from both stochasticity and smoothness. While the former can be achieved through deliberate stochasticity design in models, the latter can be attained through adversarial training.

Incorporating both robustness through stochasticity and smoothness, we proposed classifier-guided adversarial training. This involves integrating a classifier into the diffusion training process to generate meaningful adversarial perturbations that reinforce diffusion models. Empirical results demonstrated a substantial enhancement in the robustness of diffusion models, particularly evident in low diffusion purification step scenarios where the robustness gained through stochasticity is impaired. Upon revisiting our previous experiments, we reaffirmed the synergistic effect of stochasticity and smoothness in reducing model vulnerability.