Novel Research Proposal: Narrowing Information Bottleneck GAN (NIB-GAN)

Motivation

GANs are extensively used in generating adversarial examples; however, their latent representations often encode entangled and noisy features, reducing both interpretability and effectiveness of adversarial attacks. Existing interpretability and semantic disentanglement approaches rely heavily on random sampling or additional hyperparameters, creating unstable or ambiguous results.

Problem Statement

Current adversarial GANs suffer from:

- High entanglement in latent space, reducing semantic interpretability.
- Limited controllability of targeted adversarial manipulations.
- High reliance on randomness in feature attribution, resulting in unreliable adversarial attacks.

Proposed Method

Introduce a **Narrowing Information Bottleneck (NIB)** approach tailored specifically for GAN architectures to enhance adversarial image generation interpretability. Inspired by NIBT, the method deterministically narrows latent space information flow during adversarial example generation without additional hyperparameters.

- **Latent Space Control:**
  Introduce a deterministic scalar parameter to regulate information flow in the latent encoding, gradually narrowing the bottleneck to isolate key semantic dimensions crucial for adversarial effectiveness.
- **Negative Feature Attribution:**
  Identify and explicitly handle negative feature contributions, thereby improving semantic targeting and reducing irrelevant or harmful latent features in adversarial image generation.
- **Semantic Feature Isolation:**
  Directly associate latent dimension attribution with adversarial impact, avoiding indirect optimization or sampling methods, and enhancing interpretability.

Research Contributions

1. **Theoretical Formulation:**
   Develop and formally prove the theoretical underpinnings of a deterministic Narrowing Information Bottleneck adapted specifically for GAN-based adversarial image generation.
2. **Semantic Controllability Enhancement:**
   Demonstrate improved control over semantic dimensions in the latent space, providing transparent manipulation paths for generating targeted adversarial examples.
3. **Adversarial Robustness Evaluation:**
   Empirically validate NIB-GAN on standard benchmarks (e.g., CIFAR-10, ImageNet, Fashion-MNIST), showcasing significant improvements in semantic interpretability, adversarial effectiveness, and computational efficiency compared to state-of-the-art GAN interpretability methods.

Experimental Design

- **Datasets:** CIFAR-10, ImageNet, Fashion-MNIST.
- **Metrics:**
  - Adversarial robustness (fooling rate, adversarial accuracy).
  - Interpretability metrics (confidence drop/increase, latent dimension attribution clarity).
  - Computational efficiency (processing speed/FPS).

Expected Results

- Clearer semantic disentanglement and higher interpretability of GAN-generated adversarial examples.
- Reduced computational overhead compared to traditional random-sampling-based bottleneck approaches and Improved targeted adversarial manipulation through explicit negative attribution identification.

Implications

This research will significantly advance the understanding and control of GAN-based adversarial image generation, ensuring more interpretable, robust, and computationally

efficient adversarial attacks, critical for improving model defenses and interpretability frameworks.