Are Generative Classifiers More Robust to Adversarial Attacks?

Julien Delavande Soël Megdoud

Paris-Saclay

Context

This study aims to compare the robustness of different generative classifiers to adversarial attacks, focusing on how factorization choices (generative versus discriminative) affect performance. Generative classifiers explicitly model the joint probability distribution $p(x|.)$, which allows them to incorporate prior knowledge and better understand data uncertainties. This capability could potentially make them more robust against adversarial perturbations compared to discriminative classifiers, which model the conditional probability $p(y|.)$ directly without considering the data generation process.

Deep Bayes: Conditional Deep LVM

Generative classifiers model the joint distribution $p(x, y)$ of data and their labels. We introduce a latent variable z to better model the joint distribution:

$$p(x|y) = \frac{\int p(x, z, y)dz}{\int\int p(x, z, y)dzdx}, \quad p(x, y) = \int p(x, z, y)dz.$$

The joint distribution $p(x, z, y)$ can then be factorized under six distinct assumptions about dependencies between x, z, and y:

- GFZ: $p(x, z, y) = p(z)p(y|z)p(x|z, y)$
- GFY: $p(x, z, y) = p_D(y)p(z|y)p(x|z, y)$
- GBZ: $p(x, z, y) = p(z)p(y|z)p(x|z)$
- GBY: $p(x, z, y) = p_D(y)p(z|y)p(x|z)$
- DFX: $p(x, z, y) = p_D(x)p(z|x)p(y|z, x)$
- DFZ: $p(x, z, y) = p(z)p(x|z)p(y|z, x)$

The initial character "G" to denote generative classifiers and "D" to denote discriminative classifiers.

These factorization choices reflect different assumptions about the underlying data generation process and directly influence the robustness and accuracy of the classifiers.

For example, in GFZ, z is affects both x and y and x depends both on z and y whereas in BGZ z is the only link between x and y, because they are conditionally independents.

Training with ELBO

Directly optimizing the log-likelihood $\log p(x, y)$ is often intractable. To address this, a variational approach is used, introducing a recognition model $q_\phi(z|x, y)$ to approximate the true posterior $p_\theta(z|x, y)$. The objective becomes maximizing the Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{VI}} = \mathbb{E}_{q(z|x,y)} \left[\log p(x, z, y) - \log q(z|x, y)\right].$$

This formulation facilitates efficient training of generative classifiers, leveraging probabilistic encoders and decoders to model complex data distributions.

Inference with Importance Sampling

During inference, computing $p(x, y)$ requires approximating an intractable integral:

$$p(x, y) = \int p(x, z, y)dz.$$

Importance sampling is employed to estimate this integral efficiently. Using an auxiliary distribution $q(z|x, y)$, the integral can be rewritten as:

$$\int p(x, z, y)dz = \int \frac{p(x, z, y)}{q(z|x, y)} q(z|x, y)dz.$$

By sampling $z_k \sim q(z|x, y)$, we get:

$$\int p(x^*, z, y^*)dz \approx \frac{1}{K} \sum_{k=1}^{K} \frac{p(x^*, z_k, y^*)}{q(z_k|x^*, y^*)}.$$

Finally, the probability of class $p(y^*|x^*)$ is approximated as:

$$p(y^*|x^*) \approx \text{softmax}_{c=1}^{C} \left(\log \frac{1}{K} \sum_{k=1}^{K} \frac{p(x^*, z_k^c, y^c)}{q(z_k^c|x^*, y^c)} \right).$$

Adversarial Attacks

We evaluated the robustness of the different classifier architectures against Fast Gradient Sign Method (FGSM) adversarial attacks. FGSM is an attack that perturbs input data along the gradient of the loss function with respect to the input. This method generates adversarial examples by maximizing the model's loss for the true label, using the formula:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y)),$$

where ϵ controls the magnitude of the perturbation.

FGSM highlights the vulnerabilities of classifiers by finding minimal, targeted changes to the input that can mislead predictions. It is computationally efficient, making it a popular baseline for evaluating adversarial robustness.

Experiments and Results

We tested the robustness of six generative and discriminative classifiers using FGSM attacks on the Fashion MNIST dataset. Fashion MNIST contains 10 classes of fashion article images, offering a more complex and realistic benchmark compared to MNIST.

Impact of Adversarial Perturbation Magnitude:

We evaluated the success rate of adversarial attacks for different perturbation magnitudes (ϵ). As shown in Figure 1, the success rate of FGSM increases with larger perturbations. However, generative classifiers (GFZ, GFY, GBZ) demonstrate significantly better robustness compared to discriminative classifiers.

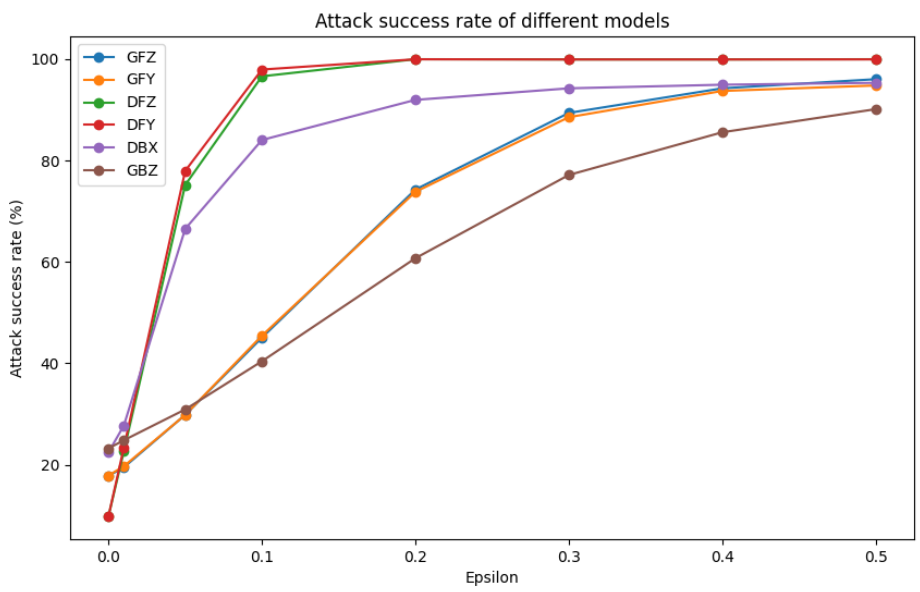

Figure 1. Success rate of FGSM attacks for generative and discriminative classifiers.

Visualizing Adversarial Examples:

The impact of adversarial perturbations is clearly visible in Figure 2. For small perturbations ($\epsilon = 0.1$), the class remains visually discernible, whereas at higher perturbations ($\epsilon = 0.3$), the image quality deteriorates significantly. Generative models maintain better predictions under these conditions.

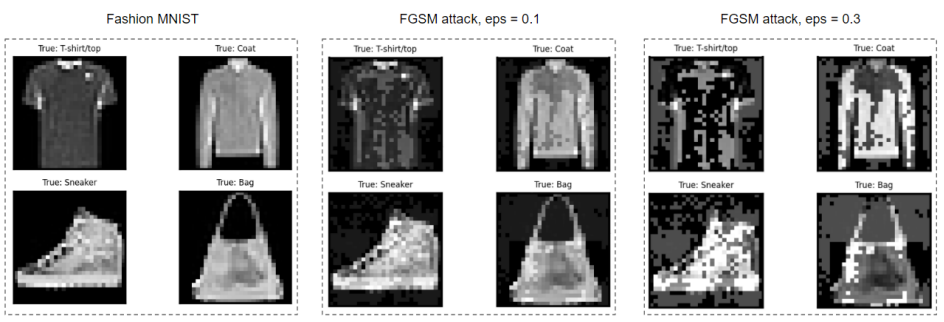

Figure 2. Examples of FGSM attacks on Fashion MNIST for varying ϵ.

Classifier Robustness:

As the perturbation magnitude increases, the success rate of adversarial attacks rises sharply for discriminative models, while generative classifiers exhibit a slower degradation in performance. This aligns with the hypothesis that generative models, which explicitly model the data distribution, are more robust to adversarial perturbations.

Conclusion: Generative classifiers demonstrate superior robustness against FGSM attacks on realistic datasets like Fashion MNIST, particularly for higher perturbations. This highlights the advantages of explicitly modeling the data generation process in adversarially robust learning.

julien.delavande@ens-paris-saclay.fr

soel.megdoud@ens-paris-saclay.fr