DiffDefense: Defending against Adversarial Attacks via Diffusion Models

Hondamunige Prasanna Silva, Lorenzo Seidenari, Alberto Del Bimbo
University of Florence, Italy

Summary

- Machine learning models are **susceptible** to small input variations.
- Adversarial attacks involve the meticulous crafting of input data by adding **imperceptible** perturbations causing **incorrect predictions**.
- Generative models are a powerful resource for **defending** against attacks, thanks to their **reconstruction** capabilities.
- Highlighting the need for models to be robust and capable of generalizing effectively even in the presence of maliciously crafted data.

Adversarial Attacks

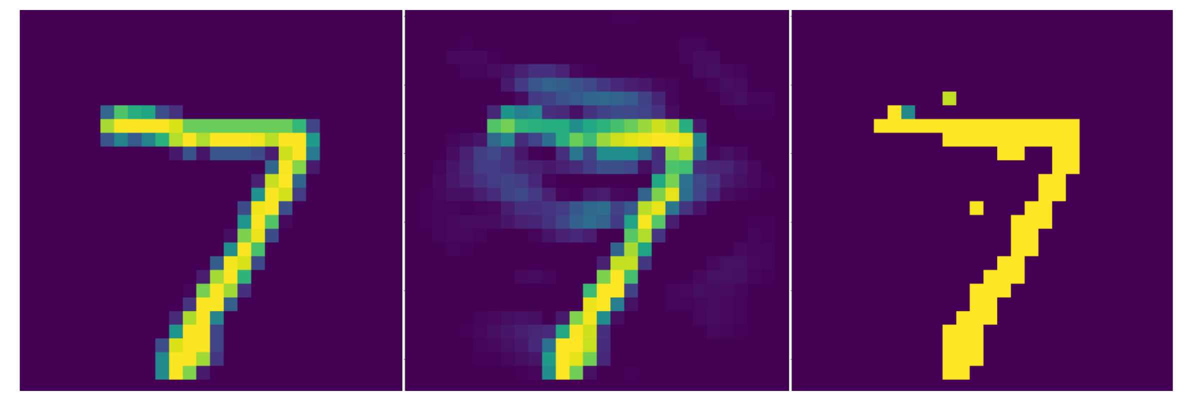
Given a pattern x, a classifier $C(x)$ and a label y, a classical adversarial attack consists in crafting a noise η such that

$$C(x+\eta) \neq y$$

Attacks such as the Fast Gradient Sign Method (FGSM) exploit the training loss gradient

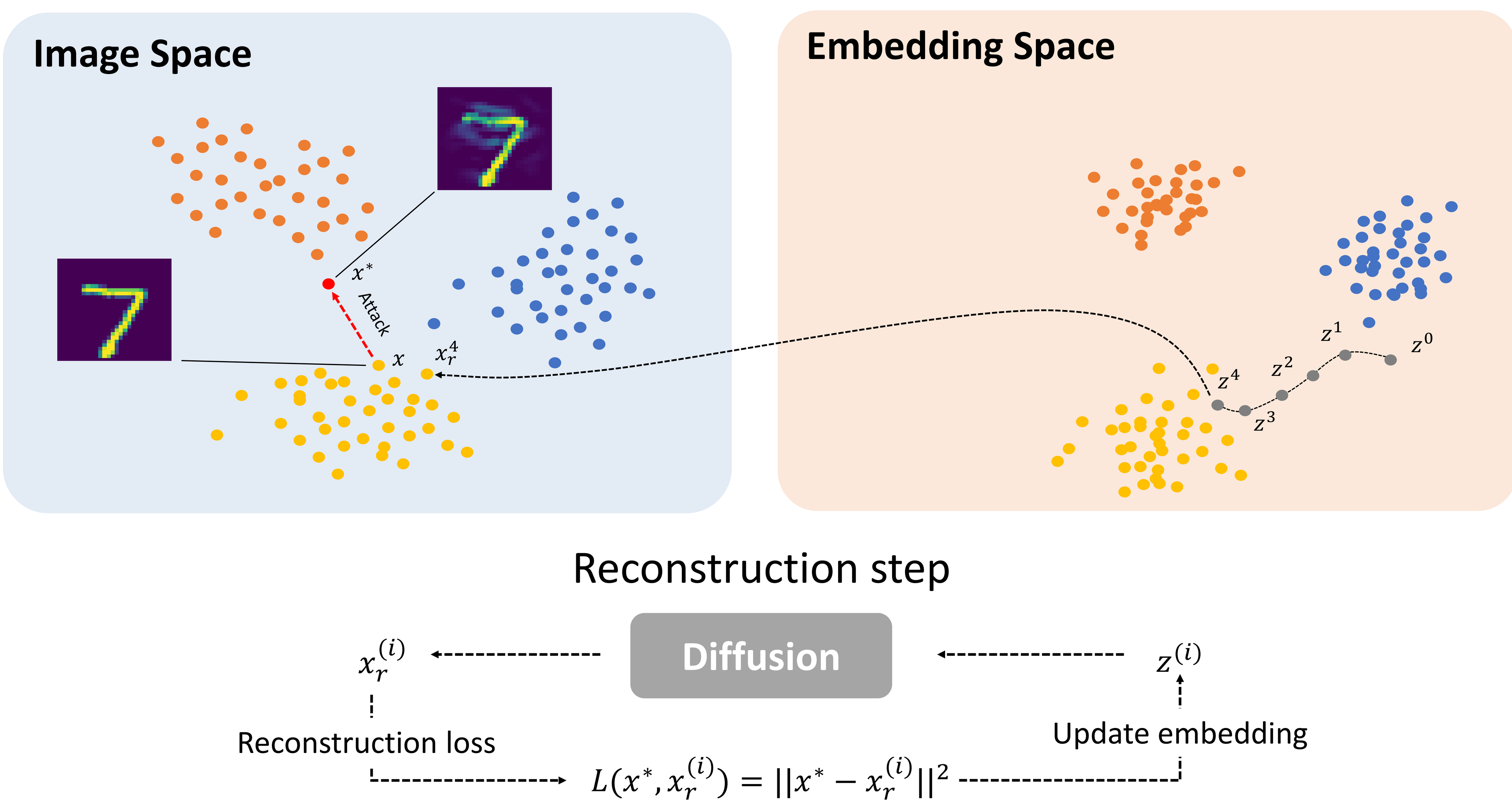
$$\eta = \epsilon \cdot \mathrm{sign}(\nabla_x \mathcal{L}(\theta, x, y))$$

The constant ϵ is set to avoid excessive corruption of the attacked pattern e.g.: 10^{-2}



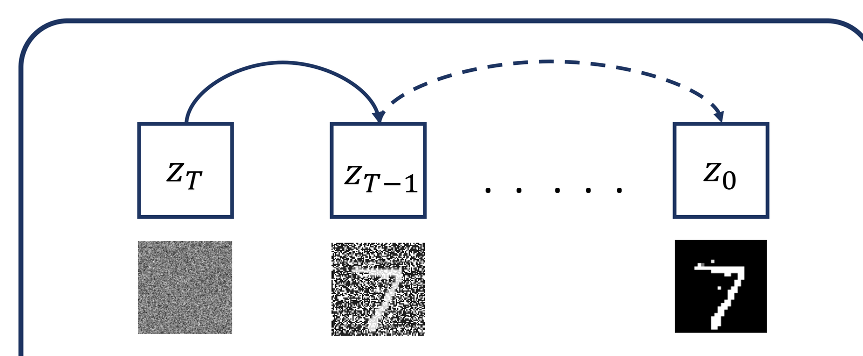
Original Attacked DiffDefense

DiffDefense: defending from attacks using diffusion models



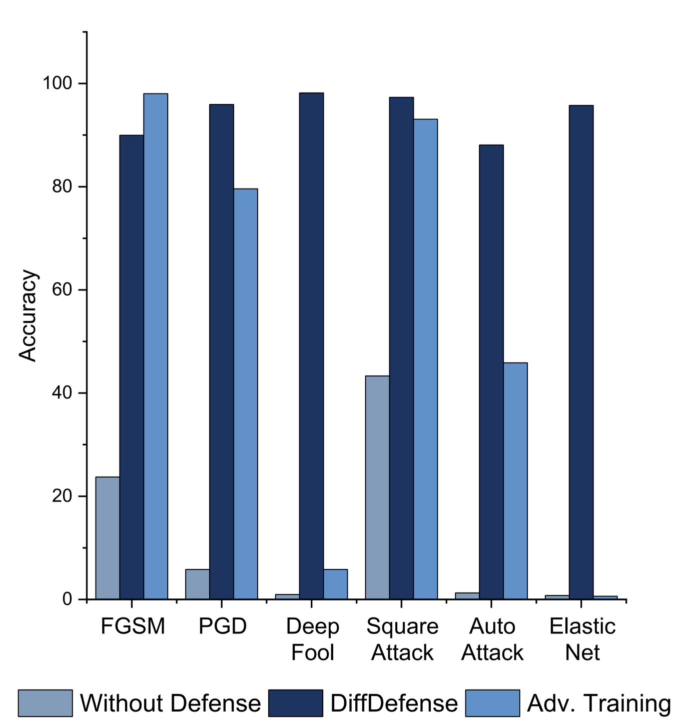
As a Loss $\mathcal{L}(x_r^{(i)}, x^*)$ we used Mean Square Error. T^* are the diffusion steps and L are the gradient descent iterations, both treated as hyperparameters. $\Delta = 0.1$ is a decay rate.

Given adversarial image x^*
$x_T^1 \sim N(0, I)$
for $i = 1, 2, \ldots, L$ **do**
  for $t = T^*, T^*-1, \ldots, 0$ *steps* **do** — **Reverse process**
    $n \sim N(0, I)$
    $z_{t-1}^{(i)} = \frac{1}{\sqrt{\alpha_t}} - \left(z_t^{(i)} - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_\theta \left(z_t^{(i)}, t \right) \right) + \sigma_t n$
  end for
  $\eta^i = \eta^{(i-1)} \Delta^{\frac{1}{\lceil L*0.8 \rceil}}$
  $x_r^{(i)} = z_0^{(i)}$
  $z_T^{(i+1)} = z_T^{(i)} - \eta^i \nabla_z \mathcal{L}(x_r^{(i)}, x^*)$ — **Gradient Descent**
end for



Diffusion Process

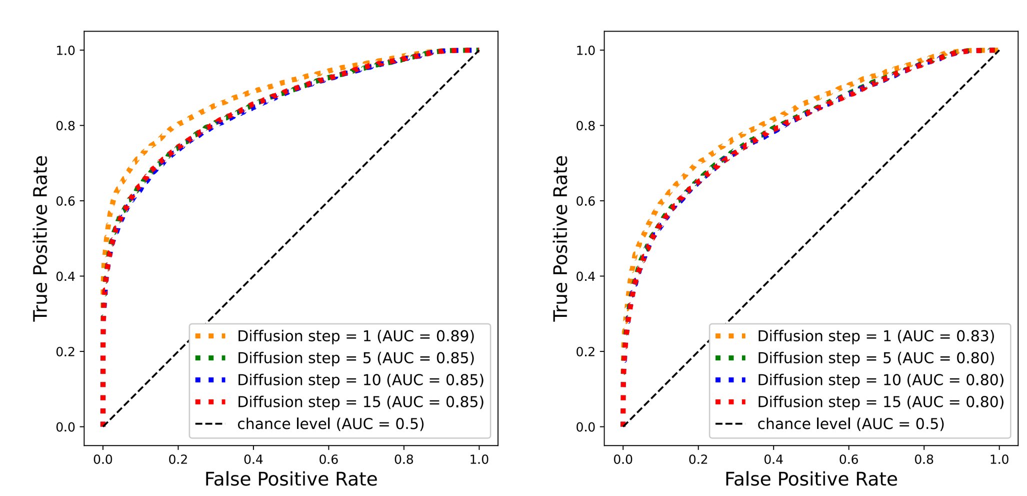
DiffDefense performance against white and black box attacks



- We test DiffDefense on several state-of-the-art attack methods.
- Except for the easier FGSM attack our defense **always improves accuracy.**
- Effective on stronger **Elastic Net** attack and **Deep Fool** methods.

Attack Detection

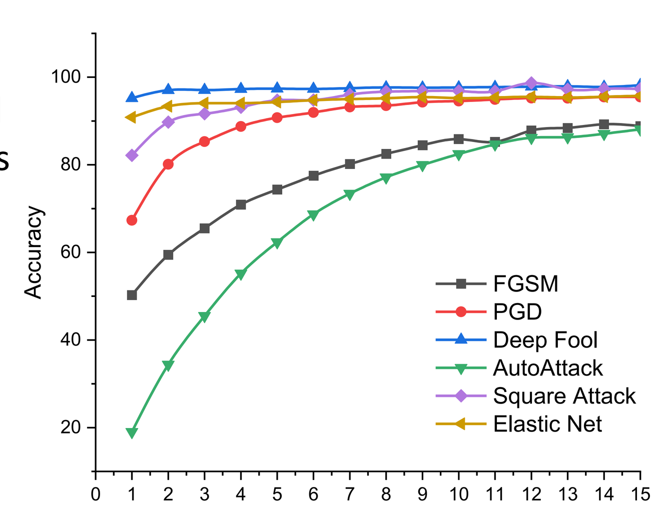
- Non-perturbed images are reconstructed with significantly **smaller reconstruction errors** after equal iterations.
- Magnitude of reconstruction error serves as an indicator for **detecting** the presence of **adversarial attacks**.
- The findings have important security implications, highlighting the potential use of reconstruction error as a signal for **identifying** and **responding** to potential adversarial attacks on images.



Attack detection ROC curves for DiffDefense (left Deep Fool, right Elastic Net). In our experiments FGSM, PGD, AutoAttack, Square Attack yielded an AUC ∈ [.99, 1]

Diffusion Steps

- Observed that the proposed method did **not require** the same number of steps as the **Diffusion Model**.
- Concluding that the proposed method achieved **convergence with much fewer steps** compared to the Diffusion Model.



DiffDefense vs DefenseGAN

Method	T	R	Time (s)	Accuracy
DefenseGan	25	10	0,086	79,98%
	10	1	0,273	50,11%
	100	10	0,338	89,11%
	200	10	0,675	91,55%
Ours	**5**	**1**	**0,28**	**87,78%**
	5	**5**	**0,28**	**89,95%**

- Our method achieved convergence with **fewer iteration steps.**
- Our method required a **smaller set of embeddings** for convergence.
- Our method took **less time to converge** compared to the *GAN-based* method.

Conclusion

- **Promising Path Forward**. Our findings suggest that Diffusion-based adversarial defense through **reconstruction** holds promise for developing secure AI systems.
- **Future Improvement**. Future research can focus on using better **solvers** to enhance further accuracy and speed in defending against adversarial attacks.
- **Future Improvement.** Refining this method on **RGB images**

Paper!

Code!