

In the thesis, we explore the prospects of creating adversarial examples using various generative models. We design two algorithms to create unrestricted adversarial examples by perturbing the vectors of latent representation and exploiting the target classifier's decision boundary properties. The first algorithm uses linear interpolation combined with bisection to extract candidate samples near the decision boundary of the targeted classifier. The second algorithm applies the idea behind the FGSM algorithm on vectors of latent representation and uses additional information from gradients to obtain better candidate samples. In an empirical study on MNIST, SVHN and CIFAR10 datasets, we show that the candidate samples contain adversarial examples, samples that look like some class to humans but are classified as a different class by machines. Additionally, we show that standard defence techniques are vulnerable to our attacks.