Title:
Exploiting the Knowledge in a Discriminator of Generative Adversarial Networks

Author
Yuejiang Liu, Parth Kothari, Alexandre Alahi

Affiliation
VITA, EPFL

Abstract:
Recent works have shown the power of the generator in Generative Adversarial Networks (GANs) for producing samples of high quality, whereas the value of the discriminator is far less explored, which is typically discarded once training completes. In this work, we exploit the knowledge learned by the discriminator regarding the data distribution for improved sample generation and prediction robustness. We first present a collaborative sampling scheme that allows both the generator and the discriminator to jointly contribute to the sample generation process through a gradient feedback loop with respect to the generated samples. We then introduce a discriminator shaping method via post adversarial training, which further improves the performance of our sampling scheme. We finally show that a well-shaped discriminator using our method has a high potential for detecting out-of-distribution samples as well as defending against adversarial attacks. We validate our method through experiments on synthetic datasets and natural images.