Generative adversarial networks

Sheng Zhong

February 25, 2019

The most interesting result from this paper is its theoretic results on the final outcome of adversarial training - that the generative model produces fakes perfectly and the discriminator can only choose randomly. This has critical implications for safety and reliability of AI classifiers. It means that **we cannot hope to build a robust classifier**, because someone could always train a generator adversarially for it. For example in the field of self-driving vehicles where classifying stop signs is a key task. With access to the discriminator model (such as through publications or a leak), a malicious party could reliably fool the car and cause catastrophic failure. There is economic incentive for one company to fool the cars of other companies as a way of hurting competition, so we can expect active efforts on achieving this on an industrial level.

Thankfully, while the theoretic results with arbitrarily complex generative and discriminator models are impressive and scary, practically there are many limitations. For example, training does not work if the discriminator is too good [1]. This is because the best discriminators have gradient 0 except at the decision boundary (it's a step function), so when you fail far from the decision boundary, the generator has no idea how to improve. Current work to address this saturation problem (as linked in the footnote) is to intentionally cripple the discriminator.

One idea for improvement is to change what the discriminator is optimizing for. Instead of optimizing for perfect distinction between data and generated, optimize it for producing a gradient for the generator to go towards generator real-seeming data. This discourages binary classification (either 0 or 1 output), and requires the discriminator to tell the generator why it thinks some data is fake, guaranteeing useful gradients. How to formulate an objective function to achieve this is not clear.

[1] https://arxiv.org/pdf/1611.04076v1.pdf