Review of Goodfellow et al. (2020)

Matthew Evans

April 30, 2025

Overview

Goodfellow et al. [1] address the limitations of traditional generative methods, which relied on costly approximations and often produced unconvincing outputs. They propose a two-network adversarial framework: a generator creates samples and a discriminator evaluates their authenticity. Through this competition, the generator progressively learns to produce high-quality, realistic data without complex inference procedures.

Approach

Prior Work

Prior to the authors' work, deep generative models had limited impact due to challenges in computing the maximum likelihood estimate, and their inability to take advantage of piecewise linear units (e.g., `ReLU`) which mitigate the vanishing-gradient problem in deep networks.

- **Deep Graphical Models** such as *deep belief networks* and *autoregressive decoders* specify an explicit joint probability, computed via MLE. Computing this exact likelihood is computationally expensive. GANs sidestep costly MLE through pure backpropagation.
- **Deep Undirected Graphical Models** such as *Boltzmann machines* define the joint probability $p(x)$ in terms of an unnormalized energy function and need Markov Chains Monte Carlo (MCMC) to approximate both gradients and sampling resulting in high computational cost. GANs eliminate MCMC entirely; sampling involves a single forward pass through G.
- **Generative Auto-encoders** such as variational autoencoders (VAEs), like GANs, use a generator with a second network–in this case, a recognition model–however unlike GANs, VAEs reconstruct data using a pixel-wise Gaussian log-likelihood which must be averaged over all plausible outputs, resulting in blurry or overly smooth generated samples. GANs eliminate this issue by using learned adversarial loss rather than a reconstruction loss.

Novelty

The authors' approach sidesteps the historical challenge of computing approximates to MLE and instead frames the problem of learning a generator as a adversarial minimax game with two competing learning models: the *generator* G and the *discriminator* D. Under this scheme, G learns how to generate new samples which are indistinguishable from those in the training data, while D provides feedback to G by learning to discriminate between samples from the training data and samples generated by G. In this way, D coaches G into producing every closer approximations to the true sample distribution.

More formally, given a random noise vector $z \sim p_z(z)$, $G(z; \theta_g)$ aims to approximate the true data distribution $p_{data}(x)$. Concurrently, $D(x; \theta_d) \in [0, 1]$ is trained to output the probability that x comes from the $p_{data}(x)$ rather than G. These models compete according to the minimax objective

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$

seeking equilibrium, at which G perfectly matches p_{data} and $D = \frac{1}{2}$.

Theoretical Results

The authors give proof of results under optimal conditions. These results, while idealized and theoretical, help explain the success of the model.

Proposition 1. For G fixed, the optimal discriminator D is

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}$$

This result guarantees that the minimax game converges precisely when $p_g = p_{data}$.

Theorem 1. The global minimum of the virtual training criterion $C(G)$ is achieved if and only if $p_g = p_{data}$. At that point, $C(G)$ achieves the value of $-\log 4$.

The virtual training criterion, which is given by

$$C(G) = \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}}[\log D^*(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D^*(G(z)))]$$

is a function of the generator alone that–when minimized–drives p_g toward p_{data}.

This result shows that, in the idealized infinite-capacity setting, the adversarial training objective has a unique global minimum achieved if and only if p_g exactly matches the true data distribution p_{data}, thereby providing a guarantee that the minimax game converges to the correct solution rather than just some arbitrary saddle point.

Proposition 2. If G and D have enough capacity, and at each step of Algorithm 1, the discriminator is allowed to reach its optimum given G, and p_g is updated so as to improve the criterion

$$\mathbb{E}_{x \sim p_{data}}[\log D_G^*(x)] + \mathbb{E}_{x \sim p_g}[log(1 - D_G^*(x))]$$

then p_g converges to p_{data}.

This result shows that by keeping the discriminator optimal at each step, the simple alternating by alternating updates between G and D will indeed drive the generator's distribution $p_g$to converge to the true data distribution p_{data}, thereby providing a concrete convergence guarantee for the training procedure.

Considerations

Strengths

- **Inference free training:** Training relies solely on backpropagation; no approximate inference networks or Markov chains are needed.
- **Flexibility:** Any differentiable function can be used in the generator or discriminator networks allowing for a variety of model designs.
- **Distribution representation fidelity:** GANs are capable of modeling un-smooth and degenerate distributions that MCMC-based methods struggle with.

Weaknesses

- **No explicit PDF:** There is no closed form of the learned distribution $p_g(x)$, so likelihoods must be estimated indirectly (e.g., with Parzen windows).
- **Training instability:** The update ratio of the competing generator and discriminator models must be properly tuned, otherwise training may collapse resulting in the generator converging to a single output for every noise input z, losing generative diversity. Similarly, the discriminator must be kept near its optimum so as to provide useful learning signals (gradients) to the generator.

Measures of Success

The authors evaluate their model by estimating the probability of the test set data under the learned distribution p_g by fitting a Gaussian Parzen window to the samples generated with G and reporting the log-likelihood under this distribution. The authors note that while this metric is imperfect, it is the best means of quantitative comparison available to them. Their results indicate (then) state of the art performance on the MNIST dataset and are within the margin of error for state of the art on the Toronto Face Database (TFD) dataset. The authors also provide visualizations of generated samples from both MNIST and TFD as indication of their generative framework's potential.

Impact

Eliminating the need for computationally costly approximate inference techniques in deep generative models, *Generative Adversarial Nets* sparked a new research subfield in which many variants have been proposed such as conditional GANs, image-to-image translation, and style transfer models. While the state-of-the-art has moved beyond GANs for tasks such as image generation, their introduction kick-started an era of learning to generate high-quality data through training a neural generator via an adversarial loss.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, October 2020.