Review of Ho et al. (2020)

Matthew Evans

May 5, 2025

Overview

In Ho et al. (2020), the authors [1] tackle the problem of designing generative models that can both assign meaningful probabilities to data and produce high-fidelity samples. They introduce *diffusion probabilistic models*, which learn to reverse a simple noising process that gradually corrupts data with Gaussian noise. By training a neural network to undo each step of this corruption, the model effectively "denoises" random noise into realistic data. This method unifies likelihood-based training with stable, high-fidelity sample generation, and matches or exceeds the performance of leading approaches on standard image benchmarks.

Approach

Prior Work

Prior to diffusion probabilistic models, generative modeling split into two main tracks: methods that offered tractable likelihoods (but sometimes lower sample fidelity) and methods that achieved high-quality samples (often without exact densities). Likelihood-based approaches—such as normalizing flows and variational autoencoders—allowed explicit probability evaluation, while adversarial and autoregressive focused on sample realism.

- **Normalizing Flows.** Flows (or normalizing flows) are generative models that transform a simple base density (e.g. a standard Gaussian) into a complex data distribution via a sequence of invertible, differentiable mappings, allowing exact likelihood computation through the change-of-variables formula. Diffusion models improve on flows by not only admitting a tractable variational likelihood but also delivering state-of-the-art sample quality by generating data in a flexible, coarse-to-fine manner rather than via a fixed invertible mapping.

- **Variational Auto-encoders.** VAE models learn a latent representation by training an encoder $p_\phi(z|x)$ and decoder $p_\theta(x|z)$ to maximize a variational lower bound on the data likelihood, but its samples can be overly smooth or blurry due to the approximate posterior and simple decoders. Diffusion models instead learn a multi-step denoising chain that progressively transforms noise into data, yielding sharper, higher-fidelity samples while still providing a tractable likelihood bound.

- **Autogregressive Models.** Autoregressive models factorize the joint distribution as a product of conditionals, generating each element in a fixed sequence (e.g., pixels one at a time) . Diffusion probabilistic models instead learn to reverse a continuous noise process in many small steps, unveiling data in a coarse-to-fine manner across all dimensions simultaneously—generalizing autoregressive bit-orderings and delivering high-fidelity samples with tractable likelihoods.

- **Energy-based Models.** Energy-based models parameterize an unnormalized density $p(x) \propto e^{-E_\theta(x)}$ and rely on costly MCMC or annealed importance sampling for sampling and likelihood estimation. Diffusion models instead use a fixed Gaussian noising process and a learnable Gaussian reverse chain, yielding exact variational likelihoods, low-variance training, and high-fidelity samples without expensive MCMC.

Novelty

The authors' diffusion probabilistic models learn an explicit, likelihood-based reversal of a fixed Gaussian noising process, yielding a single Markov chain that admits exact variational likelihood evaluation and delivers state-of-the-art sample fidelity. By unifying tractable likelihoods with high-quality generation, this approach overcomes the limitations of VAEs, flows, GANs, and autoregressive models—rivaling or exceeding leading methods (e.g., FID 3.17 and Inception Score 9.46 on unconditional CIFAR-10) without adversarial training and matching top results on datasets such as LSUN.

The authors' key innovation is to parameterize the reverse diffusion process so that each denoising step is a simple Gaussian whose mean can be expressed either as the posterior Gaussian mean of the forward process or, more effectively, as a predicted noise component. This ε-prediction formulation yields a weighted MSE training objective (L_{simple}) that both reduces variance and directly connects to denoising score matching, while sharing nearly all implementation details with standard score-based generators.

ε-Prediction Reverse Process Parameterization

Instead of directly predicting the Gaussian mean, the network predicts the noise ε added at each forward step. This reparameterization simplifies the training loss to

$$L_{\text{simple}} = \mathbb{E}_{t,x_0,\varepsilon}\left\|\varepsilon - \varepsilon_\theta\big(\sqrt{\bar{\alpha}_t}\,x_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon,\ t\big)\right\|^2,$$

which empirically yields the best sample fidelity and is straightforward to implement.

Progressive Lossy Decoding Interpretation

Viewing the reverse diffusion as a *progressive decompression* scheme reveals that early steps recover coarse image structure and later steps refine details—analogous to bit-plane autoregressive decoders but generalized via Gaussian noise. This perspective explains the strong inductive bias of diffusion models toward natural images.

Equivalence to Score Matching & Langevin Dynamics

By analyzing the variational bound, the authors prove that training the diffusion reverse chain exactly matches learning a finite-time annealed Langevin sampler via denoising score matching across noise scales.

Considerations

Strengths

- **Tractable likelihood evaluation.** Diffusion models admit straightforward log-likelihood computation via a variational bound (unlike GANs).
- **State-of-the-art sample quality.** They produce high-fidelity images (e.g., FID 3.17 on CIFAR-10) without adversarial training.
- **Theoretical unification.** The model establishes an exact equivalence to denoising score matching and annealed Langevin dynamics.
- **Simple ε-prediction parameterization.** Predicting the added noise yields a low-variance, easy-to-implement training loss that empirically maximizes sample fidelity.

Weaknesses

- **High sampling computational cost.** Requires $T = 1000$ sequential neural network evaluations per sample, making generation orders of magnitude slower than one-shot methods.
- **Inferior log-likelihood performance.** Despite strong sample quality, diffusion models yield higher (worse) lossless codelengths than leading likelihood-based models.

- **Inefficient bit allocation.** Over half of the model's lossless codelength encodes imperceptible distortions.
- **Heuristic variance schedule.** The forward diffusion variances β are manually fixed instead of learned, requiring dataset-specific tuning and potentially limiting adaptability.

Measures of Success

Quantitative Results The models are evaluated on standard image benchmarks using Inception Score (IS), Fréchet Inception Distance (FID), and negative log-likelihood (NLL) in bits/dimension. On unconditional CIFAR-10, the authors report an IS of 9.46 and an FID of 3.17, with NLL ≤ 3.75 bits/dim on test data—surpassing many prior unconditional approaches without adversarial training. On 256×256 LSUN Church and Bedroom, FIDs of 7.89 and 4.90 are achieved, respectively.

Qualitative Results High-resolution sample grids on CIFAR-10, LSUN, and CelebA-HQ display sharp detail, diverse scene compositions, and realistic textures, affirming the quantitative metrics. Latent-space interpolations on CelebA-HQ reveal smooth attribute transitions—pose, expression, lighting—and progressive decoding visualizations illustrate the emergence of coarse-to-fine structure as the reverse diffusion unfolds.

Impact

Since this paper's release, diffusion probabilistic models have sparked a major shift in generative modeling. Researchers introduced *implicit samplers* (DDIM[2], PNDM[3]) to speed up generation, and *classifier-guided*[4] and *classifier-free guidance*[5] techniques for controllable synthesis. The framework evolved into *latent diffusion models*—most notably Stability AI's *Stable Diffusion*—powering text-to-image systems alongside OpenAI's *GLIDE*, *DALL-E 2*, and Google's *Imagen*. Extensions into *video* (e.g., *Sora*), *audio*[6], and other domains have established diffusion-based methods as a versatile, high-quality paradigm across vision, language, and beyond.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *CoRR*, abs/2010.02502, 2020.

[3] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds, 2022.

[4] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021.

[5] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *CoRR*, abs/2102.09672, 2021.

[6] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023.