Denoising Diffusion Probabilistic Models

Summary written by Nicholas Hinke
April 26, 2022

Summary: As demonstrated by the recent surge of research papers coming from a wide variety of deep learning subdomains, the importance and prevalence of generative models has been growing rapidly. In recent years, four different classes of models in particular have been developed and shown to produce state-of-the-art results with this task in mind, namely: generative adversarial networks (GANs), variational autoencoders (VAEs), flows, and diffusion probabilistic models (otherwise known as "diffusion models") [7]. Despite their initial successful proposal several years prior, diffusion models, however, had not yet been shown to be able to generate high quality samples before this work from 2020 [7, 13]. For this reason, the authors of this paper sought to carefully and deliberately modify the original approach in order to achieve better results on par with their impressive competitors. Heavily inspired by other recent works regarding generative models [1, 2, 4, 8, 9, 10, 11, 12, 15], energy-based methods [3, 5, 14], and, of course, diffusion models [13], the authors successfully demonstrated how several nuanced observations regarding the nature of diffusion models allowed them to achieve incredible results that were sometimes superior to other state-of-the-art methods. Indeed, when conducting experiments with their developed models, the authors were able to achieve better FID scores (Fréchet inception distance) on the CIFAR10 dataset than almost any other model in published literature [7].

Approach: Initially developed in 2015, diffusion models fall under the umbrella of latent variable models [13]. Despite their apparent complex mathematical formulation, the working principle behind diffusion models is actually relatively simple. Given an input sample (*e.g.* image, text, natural audio *etc.*), Gaussian noise is repeatedly applied over a series of time steps during what is known as the "forward process". Subsequently, this destruction of data is undone through the "reverse process", which can essentially be defined as a Markov chain with learned Gaussian transitions [7].

Despite their seemingly rigid formulation, there are many choices one must make when designing a diffusion model that will affect its performance. Critically, the authors of this work made several novel choices that allowed their models to achieve better results than any other of this kind. Inspired by the relationship between diffusion models and denoising score matching, the authors designed the reverse process around the idea of a reparameterization that very closely resembles that of Langevin dynamics. Consequently, this choice–in addition to the decision to fix the learnable variances within the forward process as constants–led to a much simpler variational bound objective within the model. As a result, the models were much easier to train and implement, and ultimately performed better than any of their predecessors [7].

When performing experiments using their models, the authors discovered that an additional simplification on the training objective could even further improve the generated sample qualities. However, the improved outputs came at the cost of inferior codelengths [7]. Upon completion of their models, the authors evaluated their performances using Inception scores, FID scores, and negative log likelihoods on a variety of datasets, including CIFAR10, CelebA-HQ, and LSUN. Ultimately, although the diffusion models were bested by other state-of-the-art generative methods in most categories, it was clearly demonstrated the generated sample quality was quite competitive (or better) [7].

Strengths: Most notably, the resulting models as constructed by the authors were able to outperform many other state-of-the-art generative methods, thus demonstrating the true future potential of diffusion models. Additionally, due to their insightful connection to denoising score matching, the authors were able to further simplify the training process for a class of model that was already considered to be straightforward and efficient to train (especially when compared to some other generative models such as GANs) [7]. Finally, the authors likely influenced several future related advances in the field, as they made their trained models publicly available online [6].

Weaknesses: While it is implied within the paper that this approach to defining and training a diffusion model would also be applicable to other data modalities (other than images), very little discussion is provided regarding this possibility. Additionally, although discussed at several points throughout the work, the authors fail to provide much of the quantitative data that was ostensibly used to evaluate the performance of their models [7].

Reflections: As briefly mentioned above, further research regarding the applicability of this type of denoising diffusion model to other data modalities could prove beneficial to a variety of other deep learning sub-communities. Additionally, as addressed toward the end paper, further work involving the use of diffusion models for enhanced data compression could prove immensely valuable to every stakeholder within the ever-growing internet ecosystem [7].

References

[1] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2019.

[2] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *ArXiv*, abs/1605.08803, 2017.

[3] Y. Du and I. Mordatch. Implicit generation and modeling with energy based models. In *NeurIPS*, 2019.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[5] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. K. Duvenaud, M. Norouzi, and K. Swersky. Your classifier is secretly an energy based model and you should treat it like one. *ArXiv*, abs/1912.03263, 2020.

[6] J. Ho. Denoising diffusion probabilistic models. https://github.com/hojonathanho/diffusion, 2020.

[7] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

[8] N. Kalchbrenner, A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video pixel networks. *ArXiv*, abs/1610.00527, 2017.

[9] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ArXiv*, abs/1710.10196, 2018.

[10] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

[11] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

[12] R. J. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621, 2019.

[13] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *ArXiv*, abs/1503.03585, 2015.

[14] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *ArXiv*, abs/1907.05600, 2019.

[15] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *ArXiv*, abs/1601.06759, 2016.