Diffusion Models in Image Segmentation

Milan Szabo

ELTE

March 2023

Introduction

- First proposed by Sohl-Dickstein et al. 2015
- Inspired by non-equilibrium statistical physics
- They can achieve remarkable results in generative modelling
- They are highly flexible and tractable
- The research is still in an early phase
- They are now able to beat GANs in sample diversity and quality
- Used by DALL-E 2, Stable Diffusion, Google Imagen and GLIDE

Background: Based on Ho, Jain, and Abbeel 2020

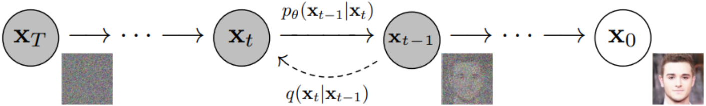


Figure: Diffusion Process from Ho, Jain, and Abbeel 2020

Background: Foward and Reverse Process

- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$
- $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I)$
- $\alpha_t = 1 - \beta_t, \overline{\alpha}_t = \prod_{s=1}^{t} \alpha_t$
- The reverse process depends on the data distribution, so we have to estimate it.
- $p_\theta(x_t|x_{t-1}) = \mathcal{N}(x_t; \mu_\theta(x_{t-1}, t), \Sigma_\theta(x_{t-1}, t))$

Background: Model architecture

- U-Net with self-attention between the 16×16 residual bocks.
- Weight normalization is replaced by group normalization.
- t is added as the Transformer sinusoidal position embedding to each residual block

Dhariwal and Nichol 2021: Diffusion Models Beat GANs on Image Synthesis

- They use an improved sampling process proposed by Song, Meng, and Ermon 2020
- The reverse steps of DDPMs have to be performed sequentially, which requires large T values.
- Their algorithm turns any $\varepsilon_\theta(x_t, x_0)$ model to a deterministic mapping from latents to images.
- Beneficial when the sampling steps are less than 50

Dhariwal and Nichol 2021: Architecture Improvements

- Increasing depth versus width, holding model size relatively constant. This increases performance, but also training time.
- 64 channels per attention head
- Using attention at 32×32, 16×16, and 8×8 resolutions rather than only at 16×16
- Using the BigGAN residual block for upsampling and downsampling the activations
- Adaptive Group Normalization: $AdaGN(h, y) = y_s GroupNorm(h) + y_b$, where $y = [y_s, y_b]$ is the linear projection of the time step and class embedding, h is the activation of the residual blocks

Dhariwal and Nichol 2021: Training the Variance

- Fixing the variance to a constant is not ideal
- New mixed objective
- v is the output of a neural network, $\beta_t, \tilde{\beta}_t$ are as defined before

$$\Sigma_\theta(x_t, t) = \exp(v \log \beta_t + (1 - v) \log \tilde{\beta}_t)$$

Dharival and Nichol 2021: Classifier Guidance

- Exploit a classifier to improve a diffusion generator
- Train a classifier on noisy images
- Use gradient $\nabla_{x_t} \log p_\theta(y|x_t, t)$
- Two stage diffusion models: low-resolution diffusion model, upsampling diffusion model. Both improve FID
- Trade of distribution coverage for sample quality

Dhariwal and Nichol 2021: Results

- Guidance and upsampling improve FID on a different axis
- Guidance trades off sample diversity for quality
- Upsampling improves precision while keeping a high recall
- They achieve the best FIDs by using guidance at a lower resolution before upsampling to a higher resolution
- This way they obtain better sample quality than state-of-the-art GANs

Rombach et al. 2022: High Resolution Image Synthesis with Latent Diffusion Models

- Slow inference speed
- High training costs
- Train DM-s in a compressed latent space
- Almost no reduction in synthesis quality
- Improved conditioning mechanism
- Unconditional guidance

Rombach et al. 2022: Image compression

- VQ-GAN with the quantization absorbed by the encoder
- Slight regularization towards the standard normal
- Different compression rates
- Keeps the 2D structure of the original image
- Trained on ImageNet

Rombach et al. 2022: Results

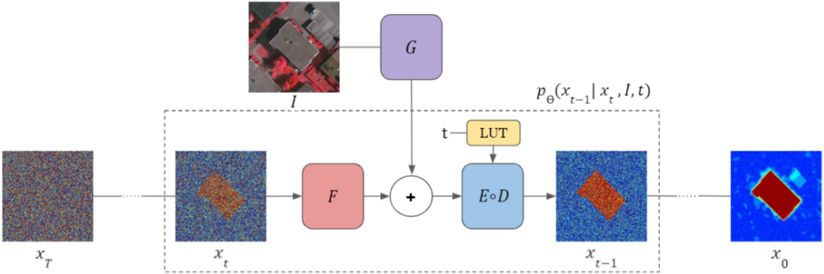
- They improve the sampling efficiency of diffusion models without degrading their quality
- Diffusion models can be favourable in some scenarios since their previous limitations were their inefficiency.

Possible Directions

- Main disadvantage remains having to take multiple steps
- Despite advances in this regard, GANs are still faster.

Amit et al. 2021

- SegDiff: Image Segmentation with Diffusion Probabilistic Models
- First to employ diffusion models to image segmentation



G
I
$p_\theta(x_{t-1} | x_t, I, t)$
t
LUT
F
+
$E \circ D$
x_T
x_t
x_{t-1}
x_0

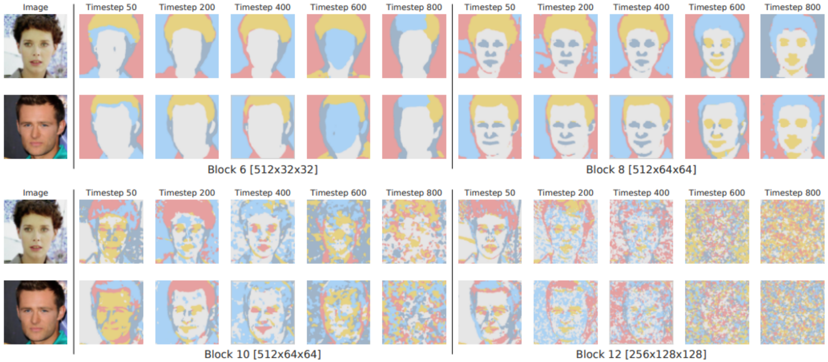
Amit et al. 2021: Architecture

- G: Conv. layer followed by a Residual in Residual Dense Block with a residual connection around it followed by a conv. layer with leaky RELU and a conv. output layer.
- F: Convolutional layer
- U-Net: $16 \times 16, 8 \times 8$ are followed by a multiheaded attention layer. The bottleneck contains two residual blocks with an attention layer in between. The residual block receives the time embedding through a linear layer.
- They obtain state-of-the-art segmentation results on a wide variety of benchmarks, including street view images, aerial images, and microscopy.

Baranchuk et al. 2021

- Label-Efficient Semantic Segmentation with Diffusion Models
- Extract intermediate representations from the 18 U-Net decoder blocks
- An MLP predicts the pixel's semantic label from the features on a specific diffusion step.

Baranchuk et al. 2021



Baranchuk et al. 2021

- Train Diffusion model unsupervised
- Train an ensemble of MLPs
- Decide the pixel label by majority voting
- The paper shows that DDPMs can serve as excellent representation learners, the representations are straightforward to compute compared to GANs.
- It requires a trained, high-quality diffusion model for the dataset. Which will most likely be available for a wide range of datasets in the near future.

Pinaya et al. 2022

- Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models
- They train a DDPM on the latent space obtained by a VQ-VAE on healthy data
- They use the L_{t-1} to verify the distance of the reverse step and the expected Gaussian transition
- High L_{t-1} values indicate an anomaly

Pinaya et al. 2022

- Calculate the mean L_{t-1} values t-s in the range of $[400, 600]$ for a validation data.
- Use the 97.5 percentile on the validation dataset as a threshold to calculate the latent mask.
- Denoise the masked region and decode with the VQ-VAE.
- To clean areas that the DDPM did not specify as anomalous, they upsample the latent mask, smooth it using a Gaussian filter, and multiply it with the residuals.
- Finally areas with high residual values are identified as anomalous.
- Competitive results with faster inference times.
- DDPM-based methods have the potential to be further improved.

Experiments

- Trained two models on TinyImageNet
- One in latent space
- One in pixel space
- Sampling:
  - 1000 diffusion steps using vanilla sampling
  - 50 and 200 steps using DDIM

ADM model

- 128-channel UNet
- 2 residual blocks per level
- Attention at resolutions 16 and 8
- Linear noise schedule
- 1000 diffusion steps for the reverse process backpropagating every 250 steps

LDM model

- Downsampling autoencoder:
  - $64 \times 64 \times 3$ to $32 \times 32 \times 4$
  - LPIPS loss with convolutional autoencoder
- Diffusion model:
  - Attention blocks at resolution $8, 16, 32$ with 8 heads per block
  - Two residual blocks per level
  - 160 base channels and $1, 2, 4, 4$ channel multiplication

Results

Model	Recall	Precision	IS	FID	sFID
ADM,DDIM, 50 steps	0.0	0.0082	1.23	495.88	165.26
ADM,DDIM, 200 steps	0.0	0.0082	1.30	443.75	156.41
ADM,DDPM, 1000 steps	*0.149*	*0.5287*	*5.42*	*77.30*	*35.79*
LDM,DDIM, 50 steps	0.3789	**0.5723**	7.25	55.39	33.15
LDM,DDIM, 200 steps	0.3669	0.5395	7.74	50.14	31.59
LDM,DDPM, 1000 steps	**0.3808**	0.5315	**7.90**	**49.03**	**30.99**

Table: Sample quality metrics for the trained models

References I

Amit, Tomer et al. (2021). "Segdiff: Image segmentation with diffusion probabilistic models". In: *arXiv preprint arXiv:2112.00390*.

Baranchuk, Dmitry et al. (2021). "Label-efficient semantic segmentation with diffusion models". In: *arXiv preprint arXiv:2112.03126*.

Dhariwal, Prafulla and Alexander Nichol (2021). "Diffusion models beat gans on image synthesis". In: *Advances in Neural Information Processing Systems* 34, pp. 8780–8794.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33, pp. 6840–6851.

References II

- Pinaya, Walter HL et al. (2022). "Fast unsupervised brain anomaly detection and segmentation with diffusion models". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*. Springer, pp. 705–714.
- Rombach, Robin et al. (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695.
- Sohl-Dickstein, Jascha et al. (2015). "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning*. PMLR, pp. 2256–2265.
- Song, Jiaming, Chenlin Meng, and Stefano Ermon (2020). "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502*.