Image Synthesis Using Denoising Diffusion Probabilistic Models

Olgeir Ingi Árnason - s212564

DTU Compute · Technical University of Denmark

Introduction

Denoising diffusion models are latent variable generative models which have recently shown great results in image generation even when compared to state-of-the-art GANs. Diffusion models learn to reverse a diffusion process that is a Markov chain which incrementally removes small amounts of noise to the data x until the data is completely destroyed. Diffusion models are generally trained using variational inference but in this project we used a simplified training objective proposed in [1].

Forward Process

As mentioned before, the forward process is a Markov chain $q(x_t|x_{t-1})$ that incrementally adds noise to the data until completely destroyed according to a variance schedule β. The forward process is parameterized and visualized by the following

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

Truncated forward process in 50 step increments (0,50,100,150,200).

The visualized forward process clearly shows the importance of the variance schedule β since too many steps will lead to redundant steps in the chain for example. Two methods of defining a variance schedule are generally used which is either a fixed linear schedule as done in [1] or a cosine schedule as done in [2]. For simplicity we choose a qualitatively evaluated fixed schedule of 200 steps. An important property of the forward process is that it admits a closed form solution for sampling at an arbitrary timestep [1]. By defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod^t \alpha_t$ the arbitrary timestep sample becomes

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

Reverse Process

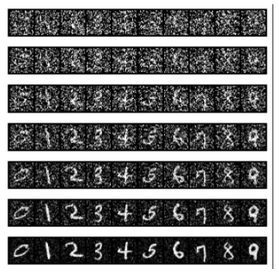
The learned reverse process is also a Markov chain with Gaussian transitions parameterized by

$$p_\theta(x_t|x_{t-1}, y) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, y), \Sigma_\theta(x_t, t, y)) \tag{3}$$

where t and y denote the timestep and the label of the image to be diffused respectively. Note that the label is not necessary and leads to conditional sampling while omitting the label leads to unconditional sampling, both of which were implemented in this project. Note that the reverse conditional probability is only tractable when conditioned on x_0.

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t\mathbf{I}) \tag{4}$$

The reverse process done by a fully trained conditional network looks like this



Truncated reverse process.

Training

As mentioned earlier the model is trained using variational inference which is equivalent to using the variational lower bound to optimize the log-likelihood

$$-\log p_\theta(x_0) \leq -\log p_\theta(x_0) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0)) \tag{5}$$

Which leads to

$$L_{LB} = E_{q(x_0:T)}[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}] \tag{6}$$

$$= E_q[D_{KL}(q(x_T|x_0)||p_\theta(x_T)) + \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1)] \tag{7}$$

$$= L_T + L_{T-1} + \cdots + L_0 \tag{8}$$

Because of the forward process property (2), we can optimize random terms of L_{LB}. This means that during training, t is drawn from a uniform distribution and the model is trained to predict x_{t-1}. Moreover, the forward process property (2) leads to the loss function

$$L_t = E_{x_0,\epsilon}[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)||\Sigma_\theta||_2^2}||\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}}x_0 + \sqrt{1-\bar{\alpha}}\epsilon_t, t)||^2] \tag{9}$$

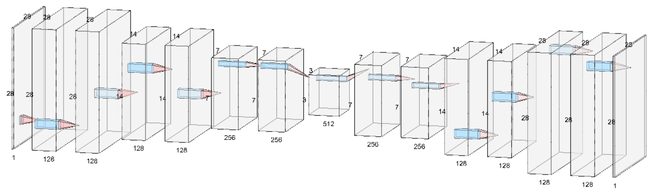
According to [1] training the model with a slightly simplified training objective yielded empirically better results, the final training objective is then

$$L_t = E_{t\in[1,T],x_0,\epsilon}[||\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}}x_0 + \sqrt{1-\bar{\alpha}}\epsilon_t, t)||^2] \tag{10}$$

This simplified objective allows us to not learn the variance of the reverse process Σ_θ and instead have it fixed. In [1] they found that fixing the variance to $\beta_t\mathbf{I}$ was beneficial to sample quality when sampling from a randomized initial state.

Model Architecture

Diffusion models are generally modeled as some type of a U-net or auto-encoder since the output should be of the same dimensions as the input. We used a U-net with slight modifications to incorporate the timestep and condition for the conditional U-net. General dimensions of convolutional blocks and the U-net as a whole can be seen in the figure below.



U-net architecture with skip connections and embeddings omitted.

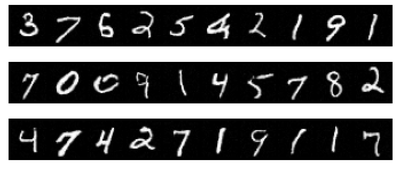
While the above figure illustrates the general dimensions of the U-net, it is missing the skip connections between convolutional blocks at the same depth in the U-net. Note that the bottom block in the U-net is the result of concatenating the downward process in the U-net, the diffusion timestep embedding and in the conditional case, the condition embedding. Overall, the model contains a total of 15 million parameters and takes approx. 20 minutes to train over 50 epochs.

Embeddings

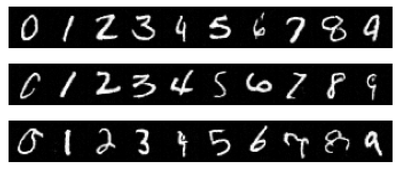
In order to train diffusion models, the randomly drawn timestep t must be embedded in the model. Our way of embedding the timestep in the model was to create a single layer fully connected neural net that takes the timestep as input and the output is a vector that is reshaped into a $n \cdot 3 \cdot 3$ tensor such that it can be concatenated at the bottom of the U-net. In the conditional case, we do the exact same and in that case the input channels for the first up convolution increase.

Results

Ideally we would present some objective model performance measures such as the negative log-likelihood or inception score for example but due to time constraints that will have to be left out for now. Until then we will have to let visual sample inspection suffice.



Unconditional samples.



Conditional samples.

References

[1] Pieter Abbeel Jonathan Ho, Ajay Jain. Denoising diffusion probabilistic models. 2020.

[2] Prafulla Dhariwal Alex Nichol. Improved denoising diffusion probabilistic models. 2021.