TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2022

Diffusion Model Basics

Denoising Diffusion Probabilistic Models (DDPM)

Ho, Jain and Abbeel, June 2020

Markovian VAEs

A diffusion model (DDPM) is a Markovian VAE.

We model y with a latent variable $z = (z_0, z_1, \ldots, z_L)$.

The encoder is defined by $z_0 = y$ and $P_{\mathrm{enc}}(z_\ell | z_{\ell-1})$.

The prior is defined by $P_{\mathrm{pri}}(z_L)$ and $P_{\mathrm{pri}}(z_{\ell-1} | z_\ell)$ and subsumes the decoder as $P_{\mathrm{pri}}(z_0 | z_1)$.

Markovian VAEs

We model y with a latent variable $z = (z_0, z_1, \ldots, z_L)$.

The encoder is defined by $z_0 = y$ and $P_{\mathrm{enc}}(z_\ell | z_{\ell-1})$.

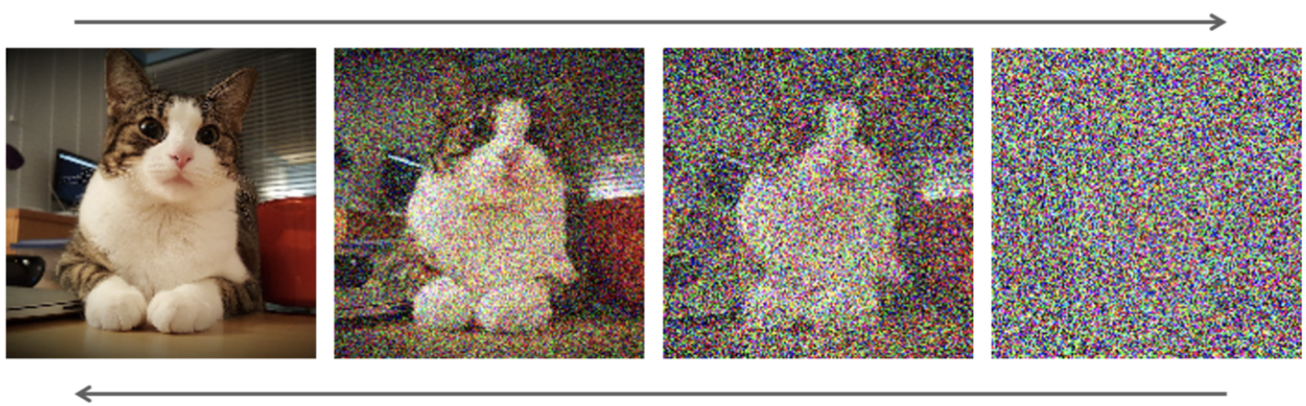
The prior is defined by $P_{\mathrm{pri}}(z_L)$ and $P_{\mathrm{pri}}(z_{\ell-1} | z_\ell)$.

We can generate y by sampling from the prior.

Denoising Diffusion Probabilistic Models (DDPM)

We model y with a latent variable $z = (z_0, z_1, \ldots, z_L)$ with $z_0 = y$.

In a DDPM we have that z_ℓ is the result of adding noise to the given image y.



DDPM SDE

The DDPM stochastic differential equation (SDE) provides the formal motivation for DDPM models.

For the DDPM SDE one can show analytically that the true reverse process probabilities $P(z_{\ell-1}|z_\ell)$ (as defined by the forward process) are Gaussians with a known variance.

This implies that in the SDE limit we can model any population **exactly** by a model in which $P(z_{\ell-1}|z_\ell)$ is taken to be Gaussian.

DDPM SDE

To formulate the DDPM SDE we will use the same σ at all levels.

$$\text{for } \ell \geq 1 \quad z_\ell = \sqrt{1-\sigma^2}\, z_{\ell-1} + \sigma\epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

This is designed so that if $z_{\ell-1}$ has unit variance in each dimension then z_ℓ also has unit variance in each dimension.

$z_0 = y$ is scaled so that each coordinate is in the interval $[0, 1]$ so that all z_ℓ have approximately unit variance.

DDPM SDE

$$\text{for } \ell \geq 1 \quad z_\ell = \sqrt{1-\sigma^2}\, z_{\ell-1} + \sigma\epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

Unit Variance is desired because in implementations the same prior network is used for all levels and it is then important that z_ℓ has the same scale and variance for all ℓ.

DDPM SDE

Because a sum of independent Gaussians is also a Gaussian, we can sample z_ℓ directly from z_0.

$$\text{define} \quad \alpha = \sqrt{1-\sigma^2}$$

$$z_\ell = \alpha^\ell z_0 + \sqrt{1-\alpha^{2\ell}}\,\epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

The variance of the noise term follows from the fact that z_ℓ has unit variance in each dimention when $z_{\ell-1}$ does.

DDPM SDE

We select the endpoint L such that z_L is essentially all noise.

$$z_L = \alpha^L z_0 + \sqrt{1 - \alpha^{2L}}\,\epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

For some limit δ we select L to be the least integer satisfying

$$\alpha^L = \sqrt{1 - \sigma^2}^{\,L} < \delta \tag{1}$$

The stochastic differential equation is defined by simultaneously taking $\sigma \to 0$ and $L \to \infty$ while satisfying (**??**).

DDPM SDE

$$z_t = e^{-t} z_0 + \sqrt{1 - e^{-2t}}\ \epsilon_1$$

$$\Delta z = -z_t \Delta t + \sqrt{2\Delta t} \epsilon_2$$

A first observation is that for infinitisimal Δt we have that Δt is infinitesmal compared to $\sqrt{2\Delta t}$. **Hence the distribution** $P(\Delta z | z_t, \Delta t)$ **is Gaussian with variance** $\sigma = \sqrt{2\Delta t}$ **and with mean infinitesimal with respect to the variance.**

DDPM SDE

$$z_\ell = \sqrt{1-\sigma^2}\, z_{\ell-1} + \sigma\,\epsilon$$

As we take $\sigma \to 0$ we can rewrite $\sqrt{1-\sigma^2}$ using the first order Taylor expansion of the square root function.

$$z_\ell = \left(1 - \frac{1}{2}\sigma^2\right)\, z_{\ell-1} + \sigma\,\epsilon$$

$$\Delta z = -\frac{1}{2}\sigma^2\, z_{\ell-1} + \sigma\,\epsilon$$

DDPM SDE

$$\Delta z = -\frac{1}{2}\sigma^2 \; z_{\ell-1} + \sigma \; \epsilon$$

Here we are taking $\sigma \to 0$ which means that the second order term $(1/2)\sigma^2 z_\ell$ is infintesimal compared to the noise term $\sigma\epsilon$. This is the hallmark of a stochastic differential equation.

Here Δz is distributed as a Gaussian with variance σ and an infinitesimal mean (relative to its variance). The mean is infinitesimal compared to the variance, the mean accumulutes over the (very large) sequence $z_0, \ldots, z_L$.

Rewriting the ELBO

We will derive the structure of the prior by optimizing the ELBO loss.

The following is a standard reformulation of the ELBO loss that is valid for all Markovian VAEs.

$$H(z_0) \;\leq\; E_{y,z}\; - \ln \frac{p_{\mathrm{pri}}(z_0, \ldots, z_L)}{p_{\mathrm{enc}}(z_1, \ldots, z_L | z_0)}$$

$$=\; E_{y,z} - \ln p_{\mathrm{pri}}(z_L) - \sum_{\ell \geq 1} \frac{\ln p_{\mathrm{pri}}(z_{\ell-1} | z_\ell)}{\ln p_{\mathrm{enc}}(z_\ell | z_{\ell-1})}$$

Rewriting the ELBO

$$
\begin{aligned}
H(z_0) \;&\leq\; E_{y,z} - \ln p_{\mathrm{pri}}(z_L) - \sum_{1 \leq \ell \leq L} \ln \frac{p_{\mathrm{pri}}(z_{\ell-1}|z_\ell)}{p_{\mathrm{enc}}(z_\ell|z_{\ell-1})} \\
&=\; E_{y,z} - \ln p_{\mathrm{pri}}(z_L) - \sum_{1 \leq \ell \leq L} \ln \frac{p_{\mathrm{pri}}(z_{\ell-1}|z_\ell)}{p_{\mathrm{enc}}(z_\ell|z_{\ell-1}, z_0)} \\
&=\; E_{y,z} - \ln p_{\mathrm{pri}}(z_L) - \sum_{1 \leq \ell \leq L} \ln \frac{p_{\mathrm{pri}}(z_{\ell-1}|z_\ell)p(z_{\ell-1}|z_0)}{p_{\mathrm{enc}}(z_\ell, z_{\ell-1}|z_0)} \\
&=\; E_{y,z} - \ln p_{\mathrm{pri}}(z_L) - \sum_{1 \leq \ell \leq L} \ln \frac{p_{\mathrm{pri}}(z_{\ell-1}|z_\ell)p_{\mathrm{enc}}(z_{\ell-1}|z_0)}{p_{\mathrm{enc}}(z_{\ell-1}|z_\ell, z_0)p_{\mathrm{enc}}(z_\ell|z_0)}
\end{aligned}
$$

Rewriting the ELBO

$$
\begin{aligned}
H(z_0) \;&\leq\; E_{y,z} - \ln p_{\mathrm{pri}}(z_L) - \sum_{1\leq\ell\leq L} \ln \frac{p_{\mathrm{pri}}(z_{\ell-1}|z_\ell)}{p_{\mathrm{enc}}(z_{\ell-1}|z_\ell,\, z_0)} - \ln \frac{p_{\mathrm{pri}}(z_{\ell-1}|z_0)}{p_{\mathrm{enc}}(z_\ell|z_0)} \\
&=\; E_{y,z} - \ln \frac{p_{\mathrm{pri}}(z_L)}{p_{\mathrm{enc}}(z_L|z_0)} - \sum_{2\leq\ell\leq L} \ln \frac{p_{\mathrm{pri}}(z_{\ell-1}|z_\ell)}{p_{\mathrm{enc}}(z_{\ell-1}|z_\ell,\, z_0)} - \ln p_{\mathrm{pri}}(z_0|z_1) \\
&=\; E_{y,z} \left\{ \begin{array}{l} KL(p_{\mathrm{enc}}(z_L|z_0), p_{\mathrm{pri}}(z_L)) \\ \\ + \sum_{2\leq\ell\leq L} KL(p_{\mathrm{enc}}(z_{\ell-1}|z_\ell,\, z_0), p_{\mathrm{pri}}(z_{\ell-1}|z_\ell)) \\ \\ - \ln p_{\mathrm{pri}}(z_0|z_1) \end{array} \right.
\end{aligned}
$$

Using The Gaussian Model

The ELBO loss is optimized by

$$\mathrm{pri}^*(z_{\ell-1}|z_\ell) = \operatorname*{argmin}_{\mathrm{pri}} \; E_{y,z} \; \ln \frac{p_{\mathrm{enc}}(z_{\ell-1}|z_\ell, y)}{p_{\mathrm{pri}}(z_{\ell-1}|z_\ell)}$$

In the DDPM SDE both distributions are Gaussians with variance σ. The KL divergence then becomes.

$$\mathrm{pri}^* = \operatorname*{argmin}_{\mathrm{pri}} \; \sum_\ell \; E_{y,z} \; \frac{||\mu_{\mathrm{pri}}(z_{\ell-1}|z_\ell) - \mu_{\mathrm{enc}}(z_{\ell-1}|z_\ell, y)||^2}{2\sigma^2}$$

Setting the Variance to σ in the prior.

$$\mathrm{pri}^* = \underset{\mathrm{pri}}{\mathrm{argmin}} \sum_\ell E_{y,z} \frac{||\mu_{\mathrm{pri}}(z_{\ell-1}|z_\ell) - \mu_{\mathrm{enc}}(z_{\ell-1}|z_\ell, y)||^2}{2\sigma^2}$$

$$\mathrm{pri}^* = \underset{\mathrm{pri}}{\mathrm{argmin}} \sum_\ell E_{z_0,\ell,z_{\ell-1},z_\ell} \frac{||\mu_{\mathrm{pri}}(\ell, z_\ell) - z_{\ell-1}||^2}{2\sigma^2}$$

The natural thing now is to train $\mu_{\mathrm{pri}}(\ell, z_\ell)$ to predict $z_{\ell-1}$.

Reducing the Prior's Dependence on ℓ.

We have already reduced the prior's dependence on ℓ by making z_ℓ have unit variance for all ℓ.

But additional dependence on ℓ can still be removed.

First we solve for $z_{\ell-1}$ in terms of z_ℓ and ϵ.

$$z_\ell = \sqrt{1-\sigma_\ell^2}\, z_{\ell-1} + \sigma\epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

$$z_{\ell-1} = \frac{1}{\sqrt{1-\sigma^2}}(z_\ell - \sigma\epsilon)$$

$$\mu_{\mathrm{pri}}(\ell, z_\ell) = \frac{1}{\sqrt{1-\sigma^2}}\left(z_\ell - \sigma\; \epsilon_{\mathrm{pri}}(\ell, z_\ell)\right)$$

Here $\epsilon_{\mathrm{pri}}(\ell, z_\ell)$ is a trained network whose target value has the same behavior at all levels of ℓ.

An ϵ-Prior

$$\mu_{\mathrm{pri}}(\ell, z_\ell) = \frac{1}{\sqrt{1-\sigma^2}} \left(z_\ell - \sigma\ \epsilon_{\mathrm{pri}}(\ell, z_\ell)\right)$$

However, SGD on the loss $\left|\left|z_{\ell-1} - \mu_{\mathrm{pri}}(\ell, z_\ell)\right|\right|^2$ now scales the SGD gradients on Φ differently for different ℓ.

We effectively have different learning rates for different ℓ.

Training the ϵ-Prior

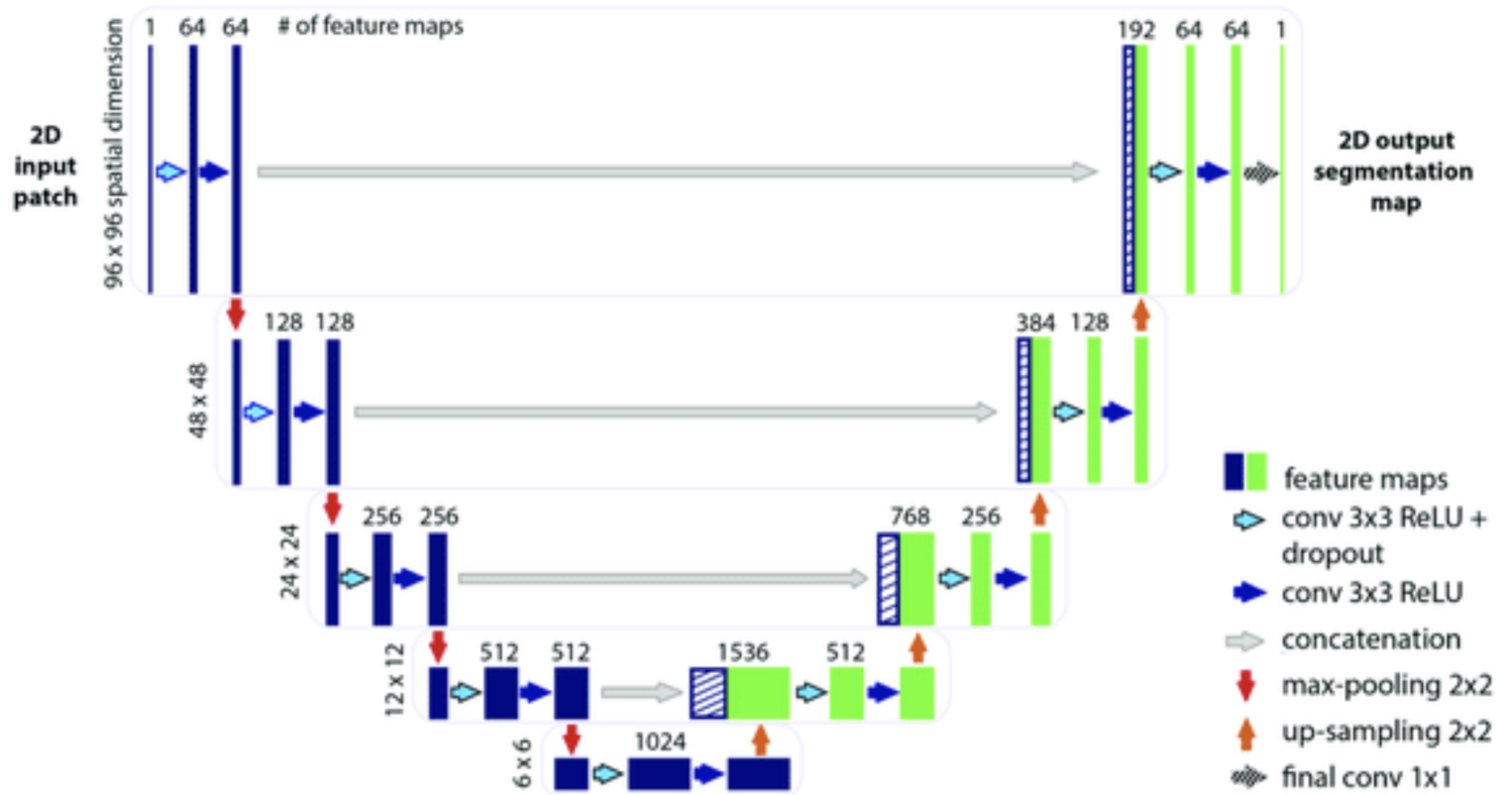
To make the scale of the SGD gradients independent of ℓ we use the following loss.

$$\Phi^* = \operatorname*{argmin}_{\Phi} \left\{ \begin{array}{l} E_{z_0,\ell,z_{\ell-1},\epsilon\sim\mathcal{N}(0,I)} \\ \\ \qquad \left|\left|\epsilon - \epsilon_{\mathrm{pri}}\left(\ell, z_\ell(z_{\ell-1}), \ell\right)\right)\right|\right|^2 \end{array} \right.$$

We now repeatedly sample z_0, ℓ, $z_{\ell-1}$ and ϵ and do gradient updates on Φ.

ϵ-Prior Architecture

The ϵ-decoder is a U-Net.



Generating Faces

But this is "mearly" a face generator. DALLE and DALLE-2 do text-conditioned image generation. Also, here we are using $L = 1000$.

END