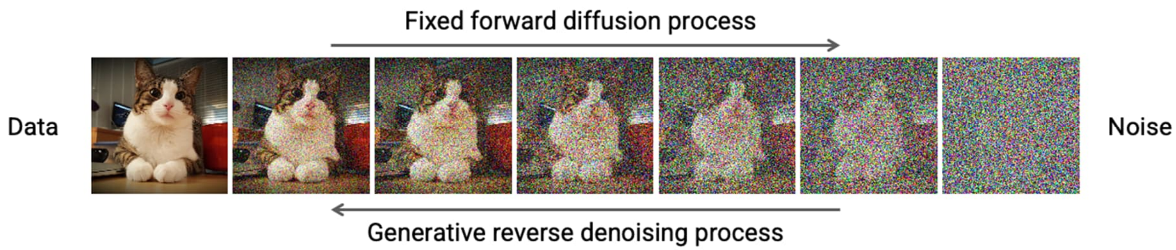
Restricted Boltzmann machines and generative diffusion models III

Eric Paquet

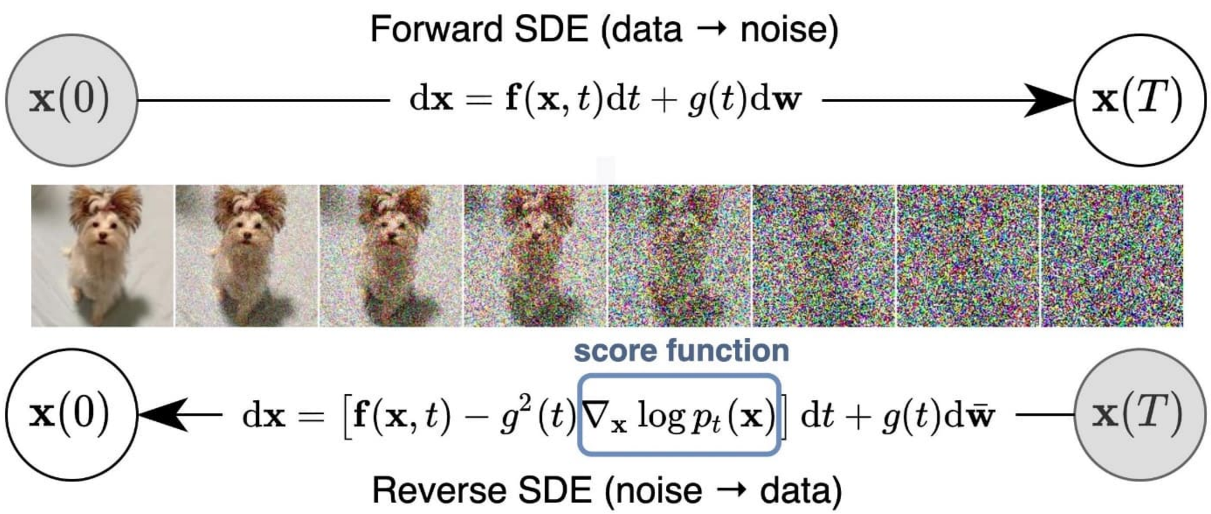
National Research Council

28 November 2024

Denoising diffusion probabilistic models:



Score-based generative modelling with stochastic differential equations:



Forward process; noising process; Wiener process; drift; diffusion coefficient:

$$d\mathrm{x} = f(\mathrm{x},t)dt + g(t)d\mathrm{w} \tag{1}$$

Reverse process; denoising process; score function; exact results:

$$d\mathrm{x} = \left[f(\mathrm{x},t) - g(t)^2 \nabla_{\mathrm{x}} \ln p(\mathrm{x})\right]dt + g(t)d\bar{\mathrm{w}} \tag{2}$$

Variance preserving stochastic differential equation:

$$f(\mathrm{x},t) = -\frac{1}{2}\beta(t)\mathrm{x} \tag{3}$$

$$g(t) = \sqrt{\beta(t)} \tag{4}$$

Noise scheduling: linear, sinusoidal, signal-to-noise ratio e.g.

$$\beta(t) = \beta_{\mathrm{min}} + \frac{t}{T}\left(\beta_{\mathrm{max}} - \beta_{\mathrm{min}}\right) \tag{5}$$

https://arxiv.org/abs/2011.13456

Restricted Boltzmann machine; energy:

$$E(\mathrm{v},\mathrm{h}) = -\mathrm{v}^\top \mathrm{W}\mathrm{h} - \mathrm{b}^\top \mathrm{v} - \mathrm{c}^\top \mathrm{h} \qquad (6)$$

Probability associated with the visible units; marginalisation over the hidden units:

$$p(\mathrm{v}) = \frac{1}{Z}\sum_{\mathrm{h}} e^{-E(\mathrm{v},\mathrm{h})} \equiv \boxed{\frac{e^{-F(\mathrm{v})}}{Z}}$$
$$F = \sum_{\mathrm{v}} F(\mathrm{v}) \qquad (7)$$

Free energy associated with the visible units:

$$F_\theta(\mathrm{v}) = -\mathrm{b}^\top \mathrm{v} - \sum_{j=1}^{H} \ln\left(1 + e^{c_j + \mathrm{W}_j^\top \mathrm{v}}\right), \quad \theta = \{\mathrm{b}, \mathrm{c}, \mathrm{W}\} \qquad (8)$$

The connection is made through the score function associated with the model:

$$\nabla_{\mathrm{v}} \ln p_{\theta}(\mathrm{v}) = -\nabla_{\mathrm{v}} F_{\theta}(\mathrm{v}) \tag{9}$$

$$\frac{\partial F_{\theta}(\mathrm{v})}{\partial v_i} = \frac{v_i - b_i}{\sigma_i^2} - \sum_j \frac{W_{ij}}{\sigma_i^2}\sigma\left(c_j + \sum_k \frac{v_k}{\sigma_k^2} W_{kj}\right) \tag{10}$$

$$v_i \in \{0,1\} \Rightarrow \nabla ? \tag{11}$$

The gradient may be evaluated with the Gumbel trick.

Alternatively, with the mean field approximation:

$$v_i \to \langle v_i \rangle \tag{12}$$

$$v_i \in \{0,1\} \to \langle v_i \rangle \in [0,1] \tag{13}$$

$$\langle v_i \rangle = 1p(v_i = 1) + 0p(v_i = 0) \Rightarrow$$

$$\boxed{\langle v_i \rangle = p(v_i = 1)} \qquad (14)$$

Logit; activation function; real number:

$$a_i = \text{logit}(\langle v_i \rangle) = \ln\left(\frac{\langle v_i \rangle}{1 - \langle v_i \rangle}\right) \in \mathbb{R} : \quad]0,1[\to \mathbb{R} \qquad (15)$$

This transformation is invertible:

$$\langle v_i \rangle = \sigma(a_i) = \frac{1}{1 + \exp(-a_i)} \qquad (16)$$

Free energy of the activation function:

$$v_i \to \langle v_i \rangle = p\left(v_i = 1\right) = \sigma(a_i) \Rightarrow$$

$$\boxed{F_\theta(\mathrm{a}) = -\mathrm{b}^\top \sigma(\mathrm{a}) - \ln\left(1 + e^{c_j + \mathrm{W}_j^\top \sigma(\mathrm{a})}\right)} \quad (17)$$

Score of the activation function associated with the model:

$$\boxed{\mathrm{s}_\theta(\mathrm{a}) = -\nabla_\mathrm{a} F_\theta(\mathrm{a})} \quad (18)$$

$$\mathrm{s}_\theta(\mathrm{a}) = \left(\mathrm{b} + \sum_{j=1}^{H} \sigma(c_j + \mathrm{W}_j^\top \sigma(\mathrm{a}))\mathrm{W}_j\right) \odot \sigma(\mathrm{a}) \odot (1 - \sigma(\mathrm{a})) \quad (19)$$

Denoising score matching; non-tractable; expectation:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,a_0,a_t}\left[\lambda(t)\left\| s_\theta(a_t,t) - \nabla_{a_t} \ln p(a_t)\right\|_2^2\right] \quad (20)$$

Denoising score matching; conditional score; closed-form; expectation:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,a_0,a_t}\left[\lambda(t)\left\| s_\theta(a_t,t) - \nabla_{a_t} \ln p(a_t \mid a_0)\right\|_2^2\right] \quad (21)$$

Noising process; generation of noisy data; variance preserving SDE; location-scale; reparametrisation trick:

$$da = -\frac{1}{2}\beta(t) a\,dt + \sqrt{\beta(t)}\,dw_t \quad (22)$$

$$\mathrm{a}_t \simeq \overbrace{e^{-\frac{1}{2}\int_0^t \beta(s)ds}\mathrm{a}_0}^{\mu(\mathrm{a}_0,t)} + \sigma_t \zeta, \quad \zeta \sim \mathcal{N}(0,\mathrm{I}) \Rightarrow$$

$$\mathrm{a}_t \sim p(\mathrm{a}_t \mid \mathrm{a}_0) = \mathcal{N}\left(\mathrm{a}_t; \mu(\mathrm{a}_0,t), \boldsymbol{\Sigma}\left(t\right)\right) \tag{23}$$

$$\mu(\mathrm{a}_0,t) = e^{-\frac{1}{2}\int_0^t \beta(s)ds}\mathrm{a}_0 \tag{24}$$

$$\boldsymbol{\Sigma}(t) = \sigma_t^2 \mathrm{I}, \quad \sigma_t^2 = \int_0^t e^{-\int_s^t \beta(r)dr}\beta(s)ds \simeq 1 - e^{-\int_0^t \beta(s)ds} \tag{25}$$

$$\nabla_{\mathrm{a}_t} \ln p(\mathrm{a}_t \mid \mathrm{a}_0) = -\boldsymbol{\Sigma}^{-1}\left(t\right)\left(\mathrm{a}_t - \mu(\mathrm{a}_0,t)\right) \tag{26}$$

$$\nabla_{\mathrm{a}_t} \ln p(\mathrm{a}_t \mid \mathrm{a}_0) = -\frac{1}{\sigma_t^2}\left(\mathrm{a}_t - e^{-\frac{1}{2}\int_0^t \beta(s)ds}\mathrm{a}_0\right) \tag{27}$$

Score matching:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t,\mathrm{a}_0,\mathrm{a}_t}\left[\lambda(t)\left\|\mathrm{s}_{\boldsymbol{\theta}}(\mathrm{a}_t,t) + \frac{1}{\sigma_t^2}\left(\mathrm{a}_t - e^{-\frac{1}{2}\int_0^t \beta(s)ds}\mathrm{a}_0\right)\right\|_2^2\right] \quad (28)$$

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t,\mathrm{a}_0,\mathrm{a}_t}\left[\lambda(t)\left\|\left(\mathrm{b} + \sum_{j=1}^{H}\sigma(\mathrm{c}_j + \mathrm{W}_j^{\top}\sigma(\mathrm{a}_t))\mathrm{W}_j\right) \odot \sigma(\mathrm{a}_t) \odot (1-\sigma(\mathrm{a}_t)) + \frac{1}{\sigma_t^2}\left(\mathrm{a}_t - e^{-\frac{1}{2}\int_0^t \beta(s)ds}\mathrm{a}_0\right)\right\|_2^2\right] \quad (29)$$

Stochastic optimisation:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) \quad (30)$$

The activation function associated with the data may be evaluated from a mini-batch; sampling with replacement:

$$a_{0,i}^{\mathcal{B}} = \text{logit}\left(\frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}} v_{0,i}^{(j)}\right), \quad \mathcal{B} \ll \tag{31}$$

Pros: simplicity; deterministic, differentiable

Cons: ignore stochasticity, which may lead to a biased gradient estimate; underestimation of variance

Gumbel trick

Pros: unbiased gradient estimate

Cons: high variance in gradient estimates; temperature sensitivity

Generation of new samples from random noise; reverse process; Euler – Maruyama integration:

$$a_{t-1} = a_t + \left[f(a_t, t) - g(t)^2 s_\theta(a_t, t) \right] \Delta t + g(t) \sqrt{\Delta t}\, \zeta$$

$$\zeta \sim \mathcal{N}(0, I) \qquad (32)$$

$$a_T \longrightarrow a_{T-1} \longrightarrow \ldots \longrightarrow a_0 \qquad (33)$$

Inverse of the activation function:

$$\boxed{p(v_i = 1) = \langle v_i \rangle = \sigma(a_{0,i}), \quad \mathbb{R} \rightarrow \left]0,1\right[} \qquad (34)$$

Sampling:

$$\left(p(v_i = 1) < u \right) \Rightarrow v_i = 1, \quad v_i = 0$$
$$u \sim \mathcal{U}\left(0, 1\right)$$
$$\left]0, 1\right[\longrightarrow \left\{0, 1\right\} \tag{35}$$