TTIC 31230 Fundamentals of Deep Learning, Fall 2022
Quiz 3

Problem 1. Consider a diffusion encoding process defined by

$$z_0 = y$$

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

Take α constant for all ℓ with $0 < \alpha < 1$ with $1 \leq \ell \leq L$. We assume that α^L is sufficiently small that z_L is independent of y.

Now consider training a deterministic decoder (aka denoiser) $\hat{y}_\Phi(z_\ell, \ell)$ for recovering the image y from z_ℓ using the following loss.

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\ E_{y,\ell,z_\ell}\ ||y - z_\Phi(z_\ell, \ell)||^2 \tag{1}$$

This is done in training the decoder in some successful diffusion models.

Although the model $z_\Phi(z_\ell, \ell)$ is trained to predict y it is used to generate $z_{\ell-1}$ from z_ℓ in generation. The decoding process is sometimes made to be deterministic with

$$z_{\ell-1} = z_\Phi(z_\ell, \ell) \tag{2}$$

(a) Assuming universality for Φ, and assuming that z_L is distributed as $\mathcal{N}(0, I)$ independent of y, write $z_{\Phi^*}(z_L, L)$ as an expression not involving optimization (not involving argmin).

Solution: Since z_L is independent of y, the minimizer is just the mean image.

$$\hat{y}_{\Phi^*}(z_L, L) = E\ y$$

(b) Does your answer to (a) have implications for the diversity of generated images in a model that uses both (**??**) and (**??**). Explain your answer.

Solution: Part (a) gives that z_{L-1} is deterministically the mean image independent of the noise z_L. If all other stages in the decoding are deterministic then the sampler should always get the same image.

(c) Why might an image generator based on (**??**) and (**??**) be diverse in practice?

Solution: The universality assumption does not hold in practice and hence we should not expect $z_{\ell-1}(z_L, L)$ to be truly constant independent of z_L. This variation can be amplified through the decoding process.

Problem 2. In a progressive VAE we are interested in modeling the distribution $p_\Phi(z_{\ell-1}|z_\ell, \ell)$. Current diffusion models use

$$z_{\ell-1} = z_\Phi(z_\ell, \ell) + \sigma\delta \quad \delta \sim \mathcal{N}(0, I)$$

However, if want to reduce the number of layers we have more noise between layers and the true conditional distribution $p(z_{\ell-1}|z_\ell, \ell)$ will not be Gaussian.

This problem asks you to formulate a conditional Gaussian VAE model for the conditional distribution $p_\Phi(z_{\ell-1}|z_\ell, \ell)$. Here z_ℓ and ℓ are given and you are to introduce a latent variable with an encoder, decoder and prior, for modeling $p_\Phi(z_{\ell-1}|z_\ell, \ell)$. In a Gaussian VAE we have that, without loss of generality, the prior can be taken to be $\mathcal{N}(0, I)$. In a Gaussian VAE for images the latent variable typically has smaller dimension than the images.

Letting ϵ denote the latent variable of the Gaussian VAE, and taking the prior to be $\mathcal{N}(0, 1)$, sampling from the prior, followed by sampling from the decoder, can be written as

$$z_{\ell-1} = z_\Phi(\epsilon, z_\ell, \ell) + \sigma_\Phi(\epsilon, z_\ell, \ell) \odot \delta \quad \epsilon \sim \mathcal{N}(0, I), \ \ \delta \sim \mathcal{N}(0, I)$$

Here $\odot$ denotes Hadamard product, or diminsionwise product, with $(x \odot y)[i] = x[i]y[i]$.

Write a similar equation for the Gaussian VAE encoder generating the latent variable ϵ from z_ℓ and ℓ and give the objective function for jointly training the encoder and the decoder.

Solution: For encoding the latent variable ϵ we introduce a model

$$\epsilon = \epsilon_\Psi(z_{\ell-1}, z_\ell, \ell) + \sigma_\Psi(z_{\ell-1}, z_\ell, \ell) \odot \gamma \quad \gamma \sim \mathcal{N}(0, I)$$

The training objective is then

$$\Phi^*, \Psi^* = \underset{\Phi, \Psi}{\mathrm{argmin}} \ E_{y, \ell, z_{\ell-1}, z_\ell, \epsilon} \ KL(p_\Psi(\epsilon|z_{\ell-1}, z_\ell, \ell), \mathcal{N}(0, I)) - \ln p_\Phi(z_{\ell-1}|\epsilon, z_\ell, \ell)$$

There are various equivalent ways of writing this which also get full credit.