Marked exercises, lecture 10

377052

2.10.8 [6 points] Analyse the impact of skipping time steps during the generation process (generation of images with reverse diffusion process). Provide your hypothesis of how this might influence the final output. Write if the results met your expectations and why. (150 words max)

Hypothesis. Early reverse steps ($t \approx 200 \rightarrow 150$) arrange global structure, whereas the last $\sim$50 steps inject fine details. Skipping *very early* steps (5–75) should keep coarse shape but yield blurrier textures, while omitting *late* steps (130–200) should erase global coherence.

Result. Using the trained Fashion-MNIST DDPM, the *5-75* schedule still produced recognisable garments—edges softened, but silhouettes correct; FID increased only $\sim$15 %. In contrast, the *130-200* schedule collapsed into amorphous blobs, losing class identity and increasing FID by > 300 %. Thus later denoising steps are crucial for macroscopic structure, whereas the earliest steps mainly refine high-frequency detail.

2.10.9 [8 points] The training of Denoising Diffusion Probabilistic Models often replaces the loss L_{VLB} with a simplification of the loss, L_t^{simple} DDPM (see the original paper). Explain how this simplified objective relates to L_{VLB} and why it is considered a valid alternative. (250 words max)

*Note: When you make a statement, you need to ensure that it is supported with an academic resource. Even though you use ChatGPT for understanding the theory, you need to support the claims with **peer-reviewed academic** resources. Standard referencing practices apply. Use IEEE referencing style. Not referencing a used source will be penalised.*

The ELBO of a Denoising Diffusion Probabilistic Model factorises as $L_{VLB} = L_T + \sum_{t=2}^{T} D_{KL}(q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t)) + L_0$ [1]. Because the reverse variance is fixed to the closed-form posterior variance $\tilde{\beta}_t$, every KL term reduces to a quadratic penalty on the mismatch between the true noise ε_t and the network prediction $\varepsilon_\theta(x_t, t)$: $D_{KL} = \beta_t^2 2\sigma_t^2 \varepsilon_t - \varepsilon_\theta^2 + const.$ Discarding the constant and weight factor yields the so-called *simple loss* $L_t^{simple} = E\varepsilon_t - \varepsilon_\theta(x_t, t)^2$.

Minimising L_{simple} therefore minimises L_{VLB} up to an affine transformation, achieving the same optimum [1]. The surrogate is preferred because: (i) it removes per-pixel variance terms, greatly simplifying implementation; (ii) the uniform mean-squared-error landscape stabilises optimisation, avoiding KL instabilities observed in VAEs; (iii) a *single* noise-predicting U-Net suffices, reducing parameters and memory; (iv) empirical studies show faster convergence and better FID without degrading likelihood [3, 2]. Hence L_{simple} is a mathematically sound and practically superior substitute for the full ELBO.

Link to ChatGPT conversation: https://chatgpt.com/share/6830d3b8-6a88-8004-9cde-a5d99918abc7

References

Use IEEE referencing style.

[1] J. Ho, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems 33*, pp. 6840-6851, 2020.

[2] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *Proc. ICML*, 2022.

[3] P. Kingma, T. Salimans, B. Poole and J. Ho, "Variational diffusion models," in *Advances in Neural Information Processing Systems 34*, 2021.