Exercise Sheet 8

Neural Computation
University of Birmingham

Week 8 2025

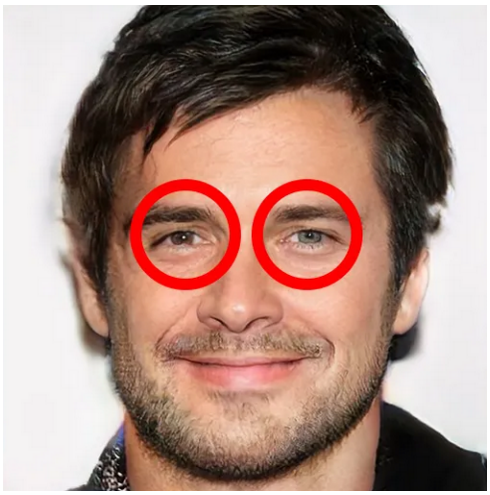
Exercise 1. Consider the following statements about DDPMs. Which of them are true/false? Explain your answer.

1. The forward process of the network is deterministic; meaning when the forward process is applied to an image, the same image at $t = 0$ will always lead to the same noise at $t = T$.
2. The backward process of the network is deterministic; meaning when it is applied to the same noise realisation at $t = T$ it will always lead to the same generated image at $t = 0$.
3. Similarly to a generator network in a GAN, the role of the UNet is to map noise to a clean image.
4. The UNet is deterministic; the same input will always produce the same output.
5. DDPMs are trained using backpropagation.
6. In DDPMs, the noise at $t = T$ corresponds to the latent code in an autoencoder or the noisy input of a GAN generator as it encodes the image in a compressed form.

Solution:

1. **False.** The forward process is stochastic because Gaussian noise is added at each step. Even with the same starting image, different noise samples will produce different outcomes at $t = T$.
2. **False.** The backward process is also stochastic. Although the UNet predicts noise components, sampling introduces randomness, so the same noise at $t = T$ can lead to different generated images.
3. **False.** The UNet does not directly map noise to a clean image like a GAN generator. Instead, it predicts the noise (or denoised estimate) at each reverse step.
4. **True.** The UNet is a deterministic function: given the same input and parameters, it produces the same output. However, the overall reverse process remains stochastic due to sampling.
5. **True.** DDPMs are trained using backpropagation to minimise a loss function, typically the mean squared error between predicted and true noise.
6. **False.** The noise at $t = T$ is not a compressed latent representation like in an autoencoder or GAN. It is pure Gaussian noise sampled from a simple distribution, not an encoding of the original image.

Exercise 2. A diffusion model is trained to generate high-resolution images of faces. It uses a UNet with two levels (two pooling layers and two upsampling layers). The model frequently produces outputs like the one below, where the colour of the left and right eye does not match. While this can occur in humans it is quite rare. What could be the reason? Suggest architectural changes to address this.



Solution:

The mismatch in eye colour likely arises because the UNet's receptive field (see week 5) is too small to capture long-range dependencies across the entire face. Each pixel prediction is influenced mainly by local context, so features that should be globally consistent (like eye colour) are treated independently.

Reason:

- A shallow UNet or insufficient downsampling limits the receptive field, preventing the model from integrating information from distant regions of the image.

Possible fixes:

- Increase network depth or add more downsampling layers to enlarge the receptive field.

Note that this is just one possible solution and if you can argue for other solutions this is fine as well.