Name & Student No.:

- Abdel Kader, Schehat & 10064822

- Tarbouch, Johny & 10033994

1. VAE vs. Diffusion: Explain the differences and similarities between VAE and Diffusion models. Name at least two differences and two similarities.

- Similarities

1. Both VAEs and Diffusion Models involve latent representations. VAEs learns a compact latent space through an encoder-decoder structure, while Diffusion models work with latent representations at different noise levels at different time steps during the diffusion process.
2. Use probabilistic frameworks to model data. VAEs explicitly maximize the Evidence Lower Bound (ELBO) on the data likelihood, while Diffusion models learn a sequence of probability distributions transitioning from noise to data

- Differences

1. Generation Process

- VAEs: generates data by sampling from a learned latent space. The latent space is typically modeled as a normal distribution and the decoder maps samples from this latent space back to the data space. The generation process is direct and typically involves a **single forward pass through the decoder**
- Diffusion Models: generate data through an iterative process. They **start with noise and gradually denoise it over multiple time steps to produce a sample**. This process involves reversing the forward diffusion process (typically normal distribution) through a learned reverse process.

2. Training Objective

- VAEs: Maximize the ELBO, balancing **reconstruction error and KL divergence** (regularization) to ensure smooth and meaningful latent space representations
- Diffusion model: Minimize the reconstruction error by learning to **predict and reverse the noise added at each time step**, enabling the generation of clean data through step-by-step denoising

2. Dequantization: In your own words, explain how dequantization helps us with discrete data.

- It is a technique used to handle discrete data like integer pixel values from 0 to 255 by converting discrete data into a continuous representation. This is done by adding a small random noise like uniform noise to the discrete value (and clipping the values). This leads to generally better results when working with continuous density models like flow models.
- Benefits
    - Smoothens the Space. Prevents degenerate solutions where the model assigns all probability mass to discrete datapoints
    - Enables smoother gradient computations which leads to better training results

3. Flow: Explain the significance of the Coupling Network in Coupling Flow

- The coupling flow splits the input x into 2 parts x_A and x_B. Then defined is z_A = x_A and only x_A gets passed to the coupling network θ. The coupling network plays an important role in the coupling flow in making invertible, efficient and expressive transformations of probability distributions.

- The coupling network θ is deterministic and the transformed x_A gets passed to the coupling transform to compute z_B = θ(x_A) + x_B. By design, the transformation is easily inverted without requiring the Coupling Network itself to be inverted because z_A = x_A and x_B can be computed with x_B = z_B - θ(z_A)

4. Reverse Diffusion: In your own words, explain why we need intermediate steps in the reverse diffusion process, and we don't reconstruct the input from the noisy image directly.

- We do not reconstruct the image directly from pure noise because this task is too complex. By breaking it into smaller steps it makes the reverse process more stable and controllable. This also makes training easier, allowing us to compare results at different time steps.
- It is intractable to go from pure noise to the original image. Having intermediate steps aligns with the forward process. In the forward process small noise as a normal distribution was added incrementally at different time steps (controlled by variance scheduler). The reverse diffusion process should mimic this by reversing the noise for every time step, which makes it follow a normal distribution. This makes the reverse diffusion process tractable and makes training the model easier

5. Self-Supervised Learning: In your own words, describe the idea and purpose of SSL. Include how SSL can help for downstream tasks.

- The idea is to have a model learn useful representation of unlabeled data, instead generate the labels yourself. For example in images, rotate the image and then predict the rotation angle. These tasks are called pretext tasks and by this approach the model learns useful representation of the data
- The purpose is to rely less on labelled data, because they are expensive and time consuming. This is reason for the success of SSL like in LLMs because large amount of unlabeled data is easily available in the internet
- Downstream task: SSL is commonly used as a pre-training step. The model is first trained on large unlabeled data using a pretext task, which has its own predictor. Afterwards the model learned a representation and it can be fine-tuned on a smaller labeled dataset for a specific downstream task like classification or sentiment analysis. By replacing the predictor of the pretext task with a predictor for the downstream task e.g. add an FNN for classification.

6. Contrastive Learning: Explain when it can be applied and describe the idea of the triplet loss function.

- Application: it can be applied in a self-supervised learning setting to pretrain a model on large amount of unlabeled data to learn an embedding. Then fine-tune it for downstream tasks
- Triplet loss function operates on triplets of input points. It tries to minimize the distance between x and x^+ and maximizes the distance between x and x^-.
    - Anchor point: **x**
    - Positive point: $\mathbf{x}^+$, sampled from the same class as x
    - Negative point: $\mathbf{x}^-$, sampled from a different class as x
    - The triplet loss is displayed below. The left part in the subtraction is the distance between the positive pair and the right part the distance between the negative pair. Epsilon is a hyperparameter and acts as an upper bound, enforcing a margin between the pairs. Epsilon is necessary so that the negative pair does not dominate the loss function and disregards minimizing the distance between positive pairs

$$L(x, x^+, x^-, \theta) = \max(0, \left|\left|f_\theta(x) - f_\theta(x^+)\right|\right|_2^2 - \left|\left|f_\theta(x) - f_\theta(x^-)\right|\right|_2^2 + \epsilon)$$