Reflective Summary on Variational Autoencoders

The video on Variational Autoencoders (VAEs) helped me understand how these models actually work under the hood. A normal autoencoder just compresses and decompresses data, but a VAE does this in a probabilistic way. Instead of mapping every input to one fixed point, it learns a distribution in the latent space with a mean *(μ)* and a variance *(σ²)*. Then it samples from that distribution to reconstruct the original input. This makes the model better at generating new examples that look real, not just memorized copies.

The main idea is that the encoder gives parameters of *q(z|x)*, which is an approximation of the true posterior *p(z|x).* Since we can’t calculate that directly, we use something called the Evidence Lower Bound (ELBO) as the loss function:

*L = E[log p(x|z)] – Dₖₗ(q(z|x) ‖ p(z))*

The first part, *E[log p(x|z)]*, is the reconstruction loss. It checks how well the decoder can rebuild the input. The second part, the KL divergence, makes sure that the distribution we’re learning *(q)* doesn’t go too far away from a standard normal distribution *(N(0, I))*. So basically, one part wants the model to copy well, and the other part wants it to stay general.

Since there’s a random sampling step, they use the “reparameterization trick” so it’s still possible to train with backpropagation. It’s written as:

*z = μ + σ ⊙ ε, where ε ~ N(0, I)*

That way, randomness comes from ε but gradients still pass through μ and σ.

Then there’s β-VAE, which is just a small tweak to this loss. The new formula is:

*L = E[log p(x|z)] – β Dₖₗ(q(z|x) ‖ p(z))*

The difference is this *β* term in front of the KL divergence. When *β > 1*, the model focuses more on the regularization term. This helps make the latent features more “disentangled,” meaning each dimension in the latent space represents something different and independent. But the downside is it can reconstruct worse because it’s being too strict. When *β = 1*, it’s just a normal VAE.

Honestly, what I liked about the lecture was how it showed that VAEs are not just about compression but about understanding the structure of data. The whole balance between reconstruction and regularization kind of feels like trying to memorize versus actually learning patterns. β-VAE makes this trade-off clearer. Overall, I think the video explained it in a simple way, and I finally understood why VAEs are so important in generative AI.