Deriving the VAE Loss Function
https://learnopencv.com/variational-autoencoder-in-tensorflow/#vae-loss

Encoder ~ (Φ = weights,bias)

1. Inputs X → Outputs the (mean and sd) of an approximate posterior distribution q(Z|X) = N(mean,sd) (variational inference needed) Because true distribution p(Z|X) is very complex
    a. Usually p(Z|X) = bayes theorem → p(X|Z)p(Z) / p(X)
    b. The denominator p(X) aka probability of observing datapoint is usually calculated by integrating $\int p(x|z)p(z)dz$ but this TAKES A LONG TIME bc integral evaluates over all possible configurations of z

Sampling ~ (z)

1. Sample from normal approximation of q(Z|X)
2. Reparameterization Trick:
    a. Direct sampling of q(Z|X) is not differentiable (can't compute gradients for backprop) since we draw from a probability distribution , aka it's no longer deterministic.
    b. So instead, we separate the sampling step from the fixed parameters of the distribution by introducing a simple and fixed random variable ε ~ N(0, 1). We then use ε to transform into samples from the desired distribution. This transformation is differentiable, allowing us to compute gradients through it.
    c. So instead, we sample from a simple and fixed distribution, and then apply a transformation to map samples to the desired distribution Q(Z|X).
    d. Z = μ(X) + σ(X) * ε
3. Now, ε is drawn from a fixed distribution (standard normal), and we can compute gradients through the transformation step.

Decoder ~ (Θ = weights,bias)

1. Samples points from the latent space and tries to reconstruct original data from it.

Deriving the Loss Function Which Updates Encoder and Decoder

ELBO = (- VAE loss)

1. Either minimize the VAE Loss
   VAE Loss = (- log likelihood) + KL Divergence between P(z|x) and Q(z|x)
2. Or maximize ELBO
    a. ELBO = E[log p(x|z)] - KL[q(z|x) || p(z)] *instead of p(x|z) which is intractable
    b. E[log p(x|z)] = log likelihood

We want to maximize Log Likelihood (likelihood of observing x | z)

KL Divergence

$$\mathbb{KL}(q_\lambda(z \mid x) \,||\, p(z \mid x)) =$$
$$\mathbf{E}_q[\log q_\lambda(z \mid x)] - \mathbf{E}_q[\log p(x, z)] + \log p(x)$$

1. We want to minimize this KL difference between approx distributions q(z|x) and true p(z|x) but lol fuck look what we see …. p(x) again which is nondifferentiable
2. So then we introduce ELBO… things cancel out and we can get p(x) on one side

$$ELBO(\lambda) = \mathbf{E}_q[\log p(x, z)] - \mathbf{E}_q[\log q_\lambda(z \mid x)].$$

Notice that we can combine this with the Kullback-Leibler divergence and rewrite the evidence as

$$\log p(x) = ELBO(\lambda) + \mathbb{KL}(q_\lambda(z \mid x) \,||\, p(z \mid x))$$

By Jensen's inequality

- Which states for a convex function f(x), the following inequality holds: E[f(x)] >= f(E[x])
- So this means the KLD >= 0 , and then minimizing KL means maximizing ELBO

Maximizing ELBO = minimizing (-) ELBO = minimizing (-) log-likelihood + KLD over models parameters

AKA, all we need to do is MAXIMIZE ELBO

$$ELBO_i(\lambda) = \mathbb{E}q_\lambda(z \mid x_i)[\log p(x_i \mid z)] - \mathbb{KL}(q_\lambda(z \mid x_i) \,||\, p(z)).$$

Find encoder - decoder parameters to maximize ELBO)

3. Get weights and biases of encoder params with q(z|x)
4. And weights and biases of decoder params with p(x|z)

and include the inference and generative network parameters as:

$$ELBO_i(\theta, \phi) = \mathbb{E}q_\theta(z \mid x_i)[\log p_\phi(x_i \mid z)] - \mathbb{KL}(q_\theta(z \mid x_i) \,||\, p(z)).$$

Summary:

- $p(z|x)$ = “posterior”, given data → construct latent space z [Encoder]
  $q(z|x)$ = approximated posterior
  $p(z)$ = “prior” we assume $N(0,1)$
  $p(x|z)$ = likelihood, given z → reconstruct data [Decoder]

1. Given joint model $p(x,z) = p(x|z)\,p(z)$
2. We want to encode true posterior $p(z|x)$
   a. $p(z|x)$ = bayes theorem but denominator is intractable
   b. Approximate $p(z|x)$ with $q(z|x)$ where $q \sim N(0,1)$
   c. Measure KLD [$q(z|x)$ || $p(z|x)$] → again $p(z|x)$ is intractable
3. ELBO: Gives a good $q(z|x)$
   a. $\log p(x)$ = ELBO + KLD [$q(z|x)$ || $p(z|x)$]
   b. Jensens Inequality allows us to rearrange and maximize ELBO
   c. ELBO = $p(x|z)$ - KLD [$q(z|x)$ || $p(z)$]