Chaitanya Kohli (S25090)

Q Derive KL Div of 2 gaussian distributions.

$$KL(P\|Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx.$$

$$p(x) = N(\mu_1, \Sigma_1)$$

$$q(x) = N(\mu_2, \Sigma_2)$$

$$\int N(\mu_1, \Sigma_1) \log \left(\frac{N(\mu_1, \Sigma_1)}{N(\mu_1, \Sigma_2)} \right)$$

$$\int N(\mu_1, \Sigma_1) \log \left(\frac{(2\pi)^{d/2} |\Sigma_2|^{1/2} \exp\left\{ -\left(\frac{1}{2} (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \right) \right\}}{(2\pi)^{d/2} |\Sigma_1|^{1/2} \exp\left\{ -\left(\frac{1}{2} (x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2) \right) \right\}} \right)$$

$$\int N(\mu_1, \Sigma_1) \left\{ \log \left(\frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} \right) + \left[\begin{array}{l} -\frac{1}{2} (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \\ +\frac{1}{2} (x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2) \end{array} \right] \right\} dx$$

$$\underset{p \sim N(\mu_1, \Sigma_1)}{E} \left\{ \frac{1}{2} \log \left(\frac{|\Sigma_2|}{|\Sigma_1|} \right) + \left[\begin{array}{l} -\frac{1}{2} (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \\ +\frac{1}{2} (x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2) \end{array} \right] \right\}$$

$$x^T A x = \mathrm{Tr}(x^T A x) = \mathrm{Tr}(A x x^T)$$

$$\underset{p \sim N(\mu_1, \Sigma_1)}{E} \left\{ \frac{1}{2} \log \left(\frac{|\Sigma_2|}{|\Sigma_1|} \right) + \left[\begin{array}{l} -\frac{1}{2} \mathrm{Tr}\left(\Sigma_1^{-1} (x-\mu_1)^T (x-\mu_1) \right) \\ +\frac{1}{2} (x-\mu_1 + (\mu_1-\mu_2))^T \Sigma_2^{-1} (x-\mu_1+\mu_1-\mu_2) \end{array} \right] \right\}$$

$$\underset{p}{E}\left[\frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)\right] - \frac{1}{2}\underset{p}{E}\left[tr\left(\Sigma_1^{-1}(x-\mu_1)^T(x-\mu_1)\right)\right]$$

$$+ \frac{1}{2}\underset{p}{E}\left[(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1) + 2(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

$$\underset{p}{E}\left[\frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)\right] - \frac{1}{2}tr\left(\Sigma_1^{-1}\underset{p}{E}(x-\mu_1)^T(x-\mu_1)\right)$$

$$+\frac{1}{2}tr\left[\Sigma_2^{-1}Ep(x-\mu_1)^T(x-\mu_1)\right] + 2\,tr\left(\Sigma_2^{-1}\underset{p}{E}(x-\mu_1)(\mu_1-\mu_2)\right)$$

$$+\frac{1}{2}\underset{p}{E}\left[(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

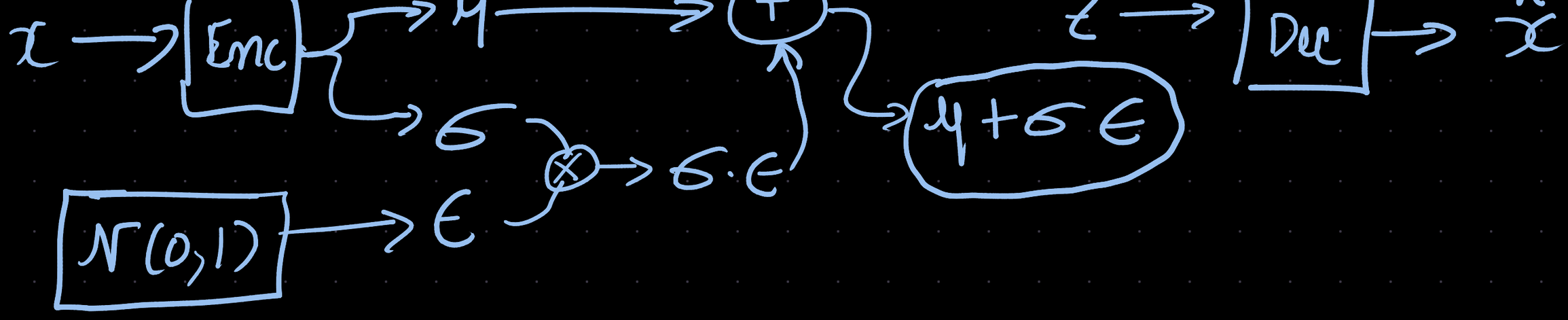
$$= \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) - \frac{1}{2}tr\left(\Sigma_1^{-1}\Sigma_1\right) + \frac{1}{2}tr\left(\Sigma_2^{-1}\Sigma_1\right)$$

$$+\frac{1}{2}(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)$$

Q why is sampling not differentiable

Sampling involves bitwise AND, OR etc operations to keep the sampling cryptographically secured. The bit wise operations are non-differentiable making sampling a non-differentiable operation.

Q How reparameterization helps in this case?



$$z = \mu + \sigma \epsilon$$

$$\frac{\partial z}{\partial \mu} = 1 \quad , \quad \frac{\partial z}{\partial \sigma} = \epsilon$$

Q Spectral Normalization

$$W = \frac{W}{\sigma(W)}$$

$\sigma(W) \rightarrow$ Largest singular value.

It represents maximum amount that the matrix W can stretch a vector.

Spectral Normalization helps enforce Lipchitz Condition

$$\| f(x) - f(y) \| \leq K \| x - y \|$$

This ensures gradient stability.

Q Explain

a) Deconvolution Layer

It is used to upsample the feature maps to higher Spatial dimension.

Also known as Transposed Convolution.
Used in VAE, AE, U-nets, in Generator of GAN etc.

e) Strided Convolutional Layer:
Used for downsampling & reducing spatial dimension.

f) Fraction Convolutional Layer:
Another name for Deconv Layer / Transposed Convolutional Layer.

Q PGGAN
Progressive Growing GANs for improved stability and variation.
The paper generated the first high resolution (1024×1024) image of human faces, solving instability seen in previous GAN.

1) Progressive Growing
Initially small Generator & Discriminator are used to train (4×4). Then add layers progressively Doubling the output (4 -> 8 -> 16 -> .. -> 1024)

this enables network to learn coarse structures, then finer details as output grows. This makes training stable & faster.

2) Fade-in mechanism:

Snapping on an entire layer which increases spatial dimension shocks the pre-trained model & ruins the parameters learned.

α parameter is used to fade-in the new layers.

It begins from $\alpha = 0$, completly ignore the new layers outputs.

α linearly increases our activations and NN outputs a weighted average of old lower resolution pathways & "new" resolution output.

3) Mode Collapse. (Generator produces only few variations of same image)

PGGAN authors introduced a layer at end of the Dis. that computes standard deviation of features across the minibatch.

Mode Collapse occurs if Generator realizes it can easily cheat Discriminator with only One image

calculate std. deviation of each feature map of batch take the average of std. deviation across this batch. And append a feature map of avg. std deviation to the original feature map.

If std. deviation is zero the Discriminator can easily identify fake batches. &
If std. deviation is high the Discriminator treats it as a normal feature of real data.

4) Pixel wise Normalization.

Batch Normalization causes GANs to be unstable.

Pixel Norm → It normalizes the feature vector in each pixel location to unit length after every conv layer.

$$b_{x,y} = a_{x,y} \Big/ \sqrt{\frac{1}{N} \sum a_{xy}^2 + \epsilon}$$

This prevents signal escalation without destroying the semantic info in magnitude.