Maximum Likelihood Is Meaningless

Presenter: Kim Seung Hwan (overnap@khu.ac.kr)

Maximum Likelihood

- VAE optimizes $\log p_\theta(x)$ - so what is this?
- $p_\theta(x)$ is likelihood: the joint probability of the observed data
- Note that this is the function of the parameter θ
- In other word, it is $p(\theta|x)$; it means a function of θ when x is fixed

Maximum Likelihood

- Maximum likelihood is point estimation for best θ
- VAE is also a type of it
- What I would claim is that this is (almost) meaningless to generation

Maximum Likelihood Is Meaningless

- Imagine a very well-fitted likelihood model $p_\theta(x)$
- Following the typical modeling, let $p_\theta(x|z)$ be the decoder of the model
- The likelihood $p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz$
- The trivial solution for fitting θ is just do $p_\theta(x|z) = p_D(x)$ for all z
- This posterior-collapse-like solution is a global optimum

Maximum Likelihood Is Meaningless

- Let us see another case
- Let X $= \{x_0, x_1, \dots, x_n\}$ be the observed true data distribution (dataset)
- $\forall i,\ p_\theta(z_i) = \frac{1}{n}$ and $p_\theta(x_i|z_i) = np_D(x_i)$
- Then $p_\theta(x) = \prod p_\theta(x_i) = \prod \sum p_\theta(x_i|z_i)p_\theta(z_i) = \prod p_D(x_i) = p_D(x)$
- This autoencoder-like solution is also an optimum

Maximum Likelihood Is Meaningless

- We want to generate samples via latent distribution z
- Intuitively, we would want somewhere between two cases
- But maximum likelihood only forces reconstruction of the observed data
- Optimizing sole likelihood does not guarantee “sampling quality”
- Even in the situation of posterior collapse or autoencoder,
- Maximum likelihood might have done right job

ELBO

- Let us go back to the VAE
- Optimizing the ELBO or even exact likelihood is a prerequisite
- It just make for good reconstruction, but we also want sampling
- Reconstruction may not even be necessary…

$$-\log p_\theta(\mathbf{x}) \leq E_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[-\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))$$

ELBO

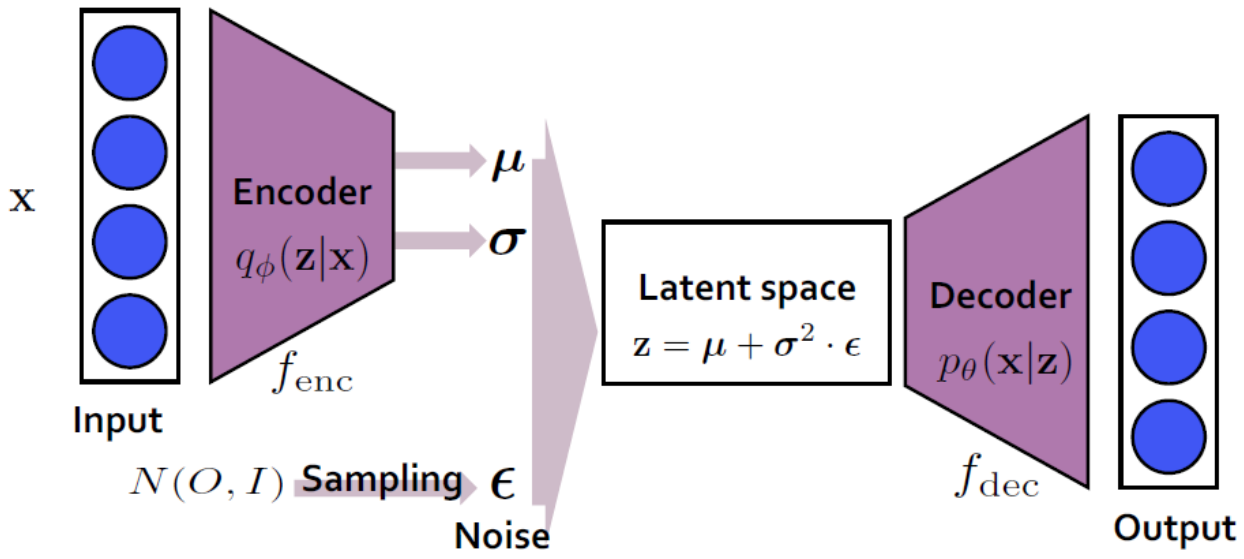
- VAE tried to obtain a generative model by fixing the prior $p(z)$ to gaussian
- But the chosen loss, the likelihood, is neither more or less than reconstruction loss
- It does not determine what will be sampled actually!
- I believe beautiful sampling quality is a side effect of VAE implementation

$$-\log p_\theta(\mathbf{x}) \leq E_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[-\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))$$

ELBO

- Specifically, examples of elements of the implementation are:

- Encoder is a diagonal covariance Gaussian model (this forces independence)
- Decoder is a isotropic & constant covariance Gaussian model (to be desc. later)
- Using CNN
- etc.



ELBO

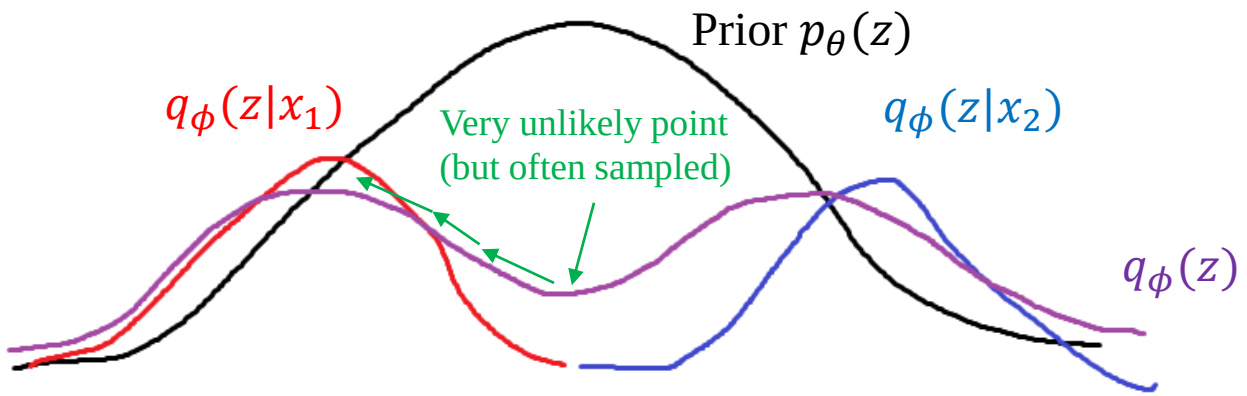
- So, to summarize my claim:

- Maximum likelihood only does reconstruction
- Thus, VAE is a sample-able reconstruction model through variational inference
- As for sampling quality, inductive bias does “everything.”

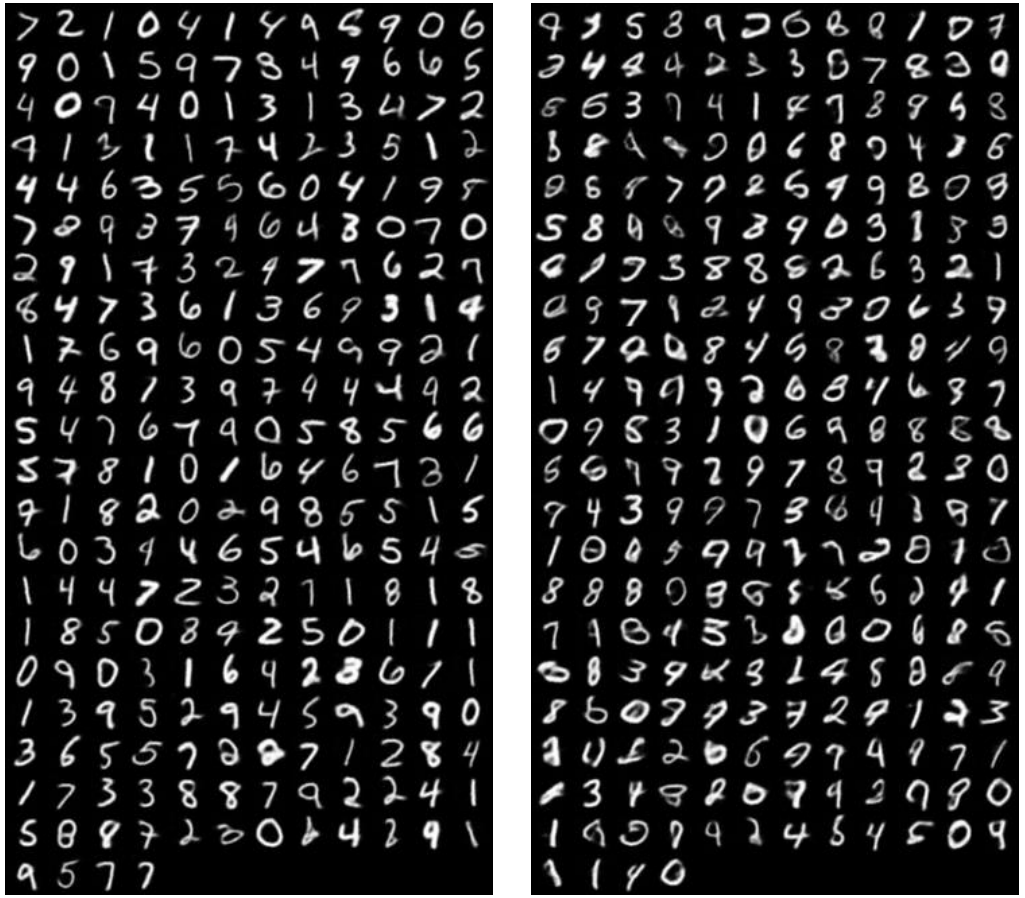
$$-\log p_\theta(\mathbf{x}) \leq E_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[-\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))$$

Experiments

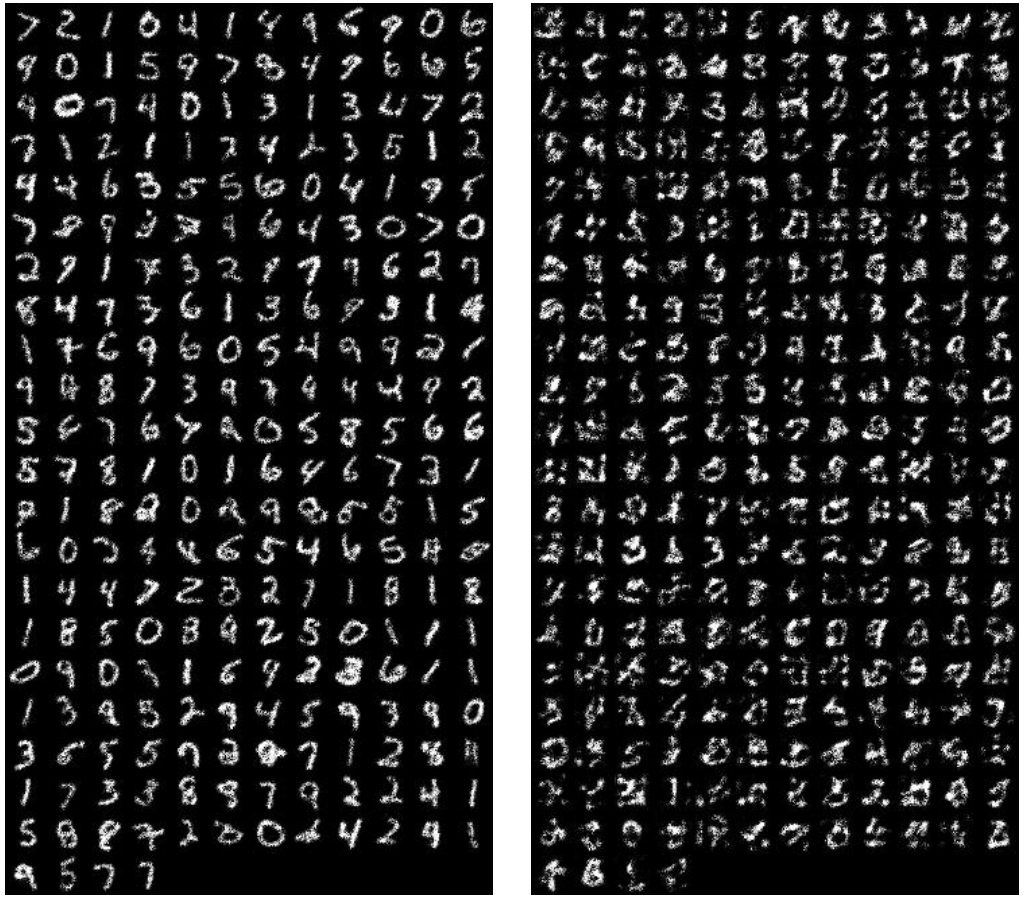
- As I said in my last presentation, I experimented with:
- Increasing the sampling quality via diffusion-like method
- I implemented a correct ELBO VAE as needed for sigma optimization
- However, the performance of this model itself was not good!



Experiments



Typical beta-VAE, recon (left) sample (right)



Exact ELBO VAE, sampling is very poor

Experiments

- Why would this happen?
- Maybe the Gaussian decoder is not well suited to the image dataset?
- But the typical beta-VAE is pretty much the same

- Original VAE loss:

$$E_{z\sim q_\phi(z|x)}[-\log p_\theta(x|z)] + D_{KL}(q_\phi(z|x)||p_\theta(z))$$

- Beta-VAE loss :

$$E_{z\sim q_\phi(z|x)}[MSE] + \beta D_{KL}(q_\phi(z|x)||p_\theta(z))$$

- Implemented correct ELBO :

$$E_{z\sim q_\phi(z|x)}\left[\frac{MSE}{2\sigma^2(z)} + \frac{1}{2}\log 2\pi\sigma^2(z)\right] + D_{KL}(q_\phi(z|x)||p_\theta(z))$$

Experiments

- Beta-VAE is a special case where the variance of the decoder is constant
- In the first place, for a sufficiently flexible function $\mu(z)$,
- $N(\mu(z), \sigma^2(z))$ will have the same (or wider) expressivity

- I just increased the expressivity of the model
- Then the model became good at reconstruction only
- The most likely suspect is the model is overfitting to the loss function and data
- This is consistent with the argument that likelihood only induces reconstruction

Experiments

- This is hard to see as a local optimum
- It has very low loss value

- The variance of the decoder approached zero
- i.e., the model could perfectly reconstruct
- Under the latent distribution allowed by variational inference, ($p(z) \approx q(z)$)
- The loss goes zero if only perfect matching is possible
- The result is empirical proof of it

Don’t blame the ELBO

- This is the paper cited in my last presentation
- My claim should blame a likelihood itself 😂
- We do not have to be bound by ELBO or mathematical justification
- What matters is the inductive bias
- The success of other variants such as WAE seems to have originated here

“Unexpectedly, we show that spurious local maxima may arise even in the optimization of exact marginal likelihood, and such local maxima are linked with a collapsed posterior”

Lucas, James, et al. "Don't blame the elbo! a linear vae perspective on posterior collapse." Advances in Neural Information Processing Systems 32 (2019).

Don’t blame the ELBO

- There are few previous works on this learnable decoder variance of VAE
- They believed that an optimal variance exists

- But I think it is correct that the variance goes zero if it is expressive enough
- Rather, depending on how we define sampling quality,
- We are looking for sub-optimal variance
- I will try to explain previous studies through my claim

More Experiments

Thank you for listening

Presenter: Kim Seung Hwan (overnap@khu.ac.kr)