LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Deep Learning

Chapter 10: Variational Autoencoder (VAE)

Mina Rezaei

Department of Statistics – LMU Munich

Winter Semester 2020

VARIATIONAL AUTOENCODER (VAE)

Independently proposed by:

- Kingma and Welling, *Auto-Encoding Variational Bayes*, ICLR 2014
- Rezende, Mohamed and Wierstra, *Stochastic back-propagation and variational inference in deep latent Gaussian models.* ICML 2014

Instead of mapping the input into a fixed vector, we want to map it into a distribution. Let's label this distribution as p_{θ}, parameterized by θ. The relationship between the data input x and the latent encoding vector z can be fully defined by:

- Prior $p_{\theta}(\mathbf{z})$
- Liklihood $p_{\theta}(\mathbf{x}|\mathbf{z})$
- Posterior $p_{\theta}(\mathbf{z}|\mathbf{x})$

VAE-PARAMETER FITTING

Assuming that we know the real parameter θ^* for this distribution.
To generate a sample from a real data point $x^{(i)}$:

- sample a $x^{(i)}$ from a prior distribution $p_{\theta^*}(z)$
- a value $x^{(i)}$ is generated from a conditional distribution $p_{\theta^*}(x|z = z^{(i)})$

The optimal parameter θ^* is the one that maximizes the probability of generating real data samples:

$$\theta^* = \arg\max_\theta \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(x^{(i)})$$

we use the log probabilities to convert the product on RHS to a sum:

$$\theta^* = \arg\max_\theta \sum_{i=1}^{n} log p_{\boldsymbol{\theta}}(x^{(i)})$$

VAE-PARAMETER FITTING

Now let's update the equation to better demonstrate the data generation process so as to involve the encoding vector:

$$p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = \int p_{\boldsymbol{\theta}}(x^{(i)}|z)p_{\boldsymbol{\theta}}(z)dz$$

VAE-PARAMETER FITTING: VARIATIONAL LOWER BOUND

- $\log p_{\theta}(\mathbf{x})$ is intractable.
- But we can compute a **variational lower bound**:

$$\log p_{\theta}(\mathbf{x}) = \log \int p_{\theta}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int q_{\phi}(\mathbf{z}|\mathbf{x}) \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

VAE-PARAMETER FITTING: VARIATIONAL LOWER BOUND

- $\log p_\theta(\mathbf{x})$ is intractable.
- But we can compute a **variational lower bound**:

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int q_\phi(\mathbf{z}|\mathbf{x}) \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

$$\geq \int q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

Jensen's inequality

Let f be a concave function and $\mathbf{x}$ an integrable random variable. Then it holds: $f(\mathbb{E}[\mathbf{x}]) \geq \mathbb{E}[f(\mathbf{x})]$.

VAE-PARAMETER FITTING: VARIATIONAL LOWER BOUND

- $\log p_{\theta}(\mathbf{x})$ is intractable.
- But we can compute a **variational lower bound**:

$$\log p_{\theta}(\mathbf{x}) = \log \int p_{\theta}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int q_{\phi}(\mathbf{z}|\mathbf{x}) \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

$$\geq \int q_{\phi}(\mathbf{z}|\mathbf{x}) \log \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

$$= \int q_{\phi}(\mathbf{z}|\mathbf{x}) \big(\log p_{\theta}(\mathbf{x}|\mathbf{z}) + \log \frac{p_{\theta}(\mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{x})} \big) d\mathbf{z}$$

VAE-PARAMETER FITTING: VARIATIONAL LOWER BOUND

- $\log p_{\theta}(\mathbf{x})$ is intractable.
- But we can compute a **variational lower bound**:

$$
\begin{aligned}
\log p_{\theta}(\mathbf{x}) &= \log \int p_{\theta}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int q_{\phi}(\mathbf{z}|\mathbf{x}) \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{x})} d\mathbf{z} \\
&\geq \int q_{\phi}(\mathbf{z}|\mathbf{x}) \log \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{x})} d\mathbf{z} \\
&= \int q_{\phi}(\mathbf{z}|\mathbf{x}) \big(\log p_{\theta}(\mathbf{x}|\mathbf{z}) + \log \frac{p_{\theta}(\mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{x})} \big) d\mathbf{z} \\
&= \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} \big[\log p_{\theta}(\mathbf{x}|\mathbf{z}) \big] - KL\big[q_{\phi}(\mathbf{z}|\mathbf{x}) || p_{\theta}(\mathbf{z}) \big]
\end{aligned}
$$

VAE-PARAMETER FITTING: VARIATIONAL LOWER BOUND

- $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ is intractable.
- But we can compute a **variational lower bound**:

$$
\begin{aligned}
\log p_{\boldsymbol{\theta}}(\mathbf{x}) &= \log \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} d\mathbf{z} \\
&\geq \int q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} d\mathbf{z} \\
&= \int q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \big(\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) + \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \big) d\mathbf{z} \\
&= \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) \big] - KL\big[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) || p_{\boldsymbol{\theta}}(\mathbf{z}) \big]}_{:= ELBO(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x})}
\end{aligned}
$$

VAE-PARAMETER FITTING: VARIATIONAL LOWER BOUND

$$ELBO(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\big] - KL\big[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z})\big]$$

VAE-PARAMETER FITTING: VARIATIONAL LOWER BOUND

$$ELBO(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\big] - KL\big[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z})\big]$$

- Also known as **Evidence Lower BOund (ELBO)**.

VAE-PARAMETER FITTING: VARIATIONAL LOWER BOUND

$$ELBO(\theta, \phi, \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - KL\big[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\big]$$

- Also known as **Evidence Lower BOund (ELBO)**.
- First term resembles reconstruction loss.

VAE-PARAMETER FITTING: VARIATIONAL LOWER BOUND

$$ELBO(\theta, \phi, \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - KL\big[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\big]$$

- Also known as **Evidence Lower BOund (ELBO)**.
- First term resembles reconstruction loss.
- Second term penalizes encoder for deviating from prior.

VAE-PARAMETER FITTING: VARIATIONAL LOWER BOUND

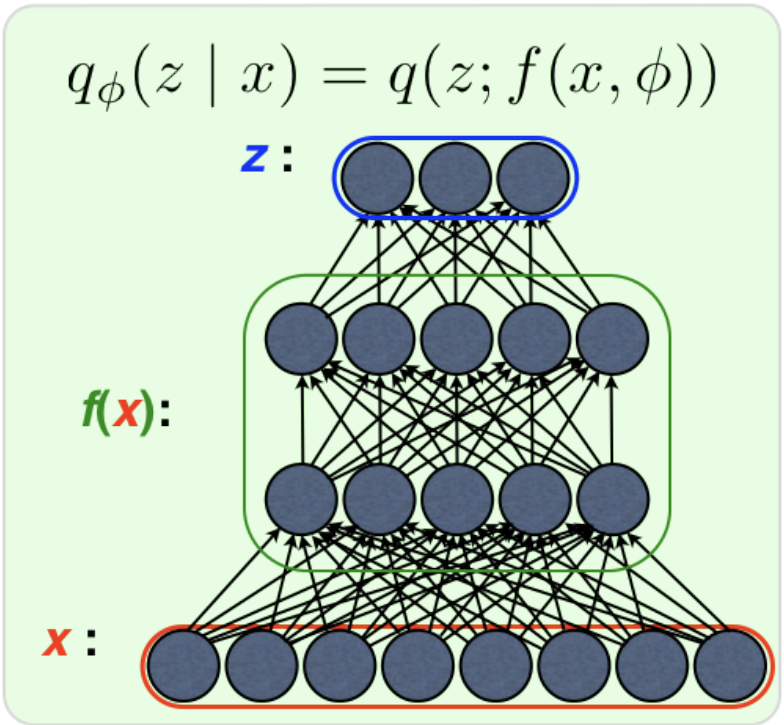
$$ELBO(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\big] - KL\big[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z})\big]$$

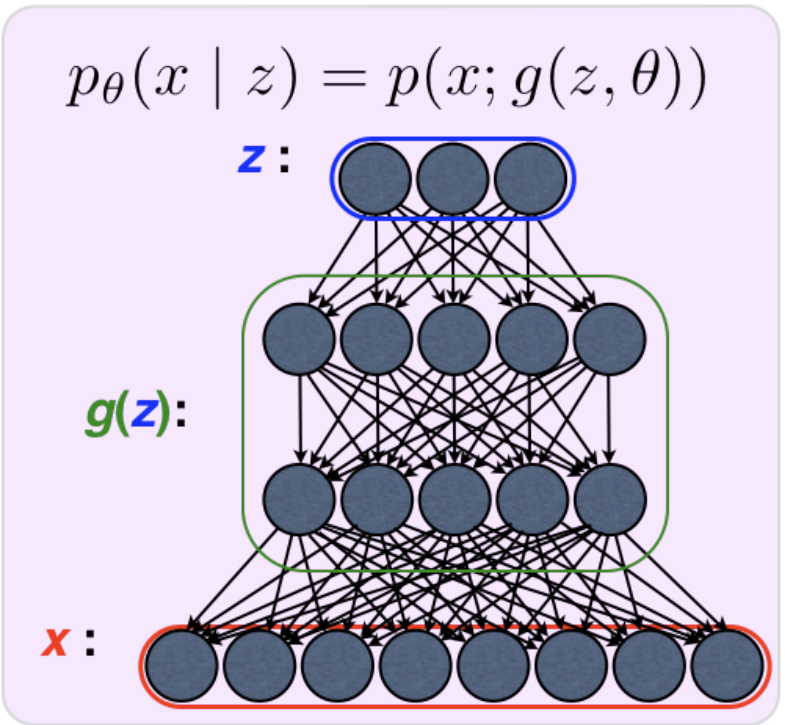
- Also known as **Evidence Lower BOund (ELBO)**.
- First term resembles reconstruction loss.
- Second term penalizes encoder for deviating from prior.
- It can be shown that
  $ELBO(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}) - KL\big[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\big]$
  $\Rightarrow$ by maximizing the ELBO we maximize $p_{\boldsymbol{\theta}}(\mathbf{x})$ and minimize $KL\big[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\big]$.

VAE-MODEL DEFINITION

Idea:

- Set $p_\theta(\mathbf{z})$ to some simple distribution.
- Parametrize inference model and generative model with neural networks $f(\mathbf{x}, \phi)$ and $g(\mathbf{z}, \theta)$.

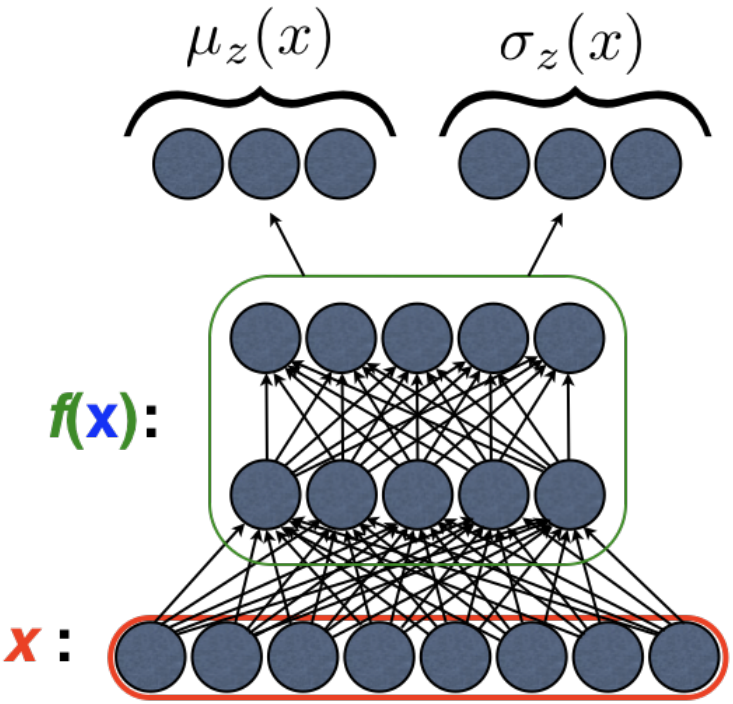




VAE-MODEL DEFINITION

Usually:

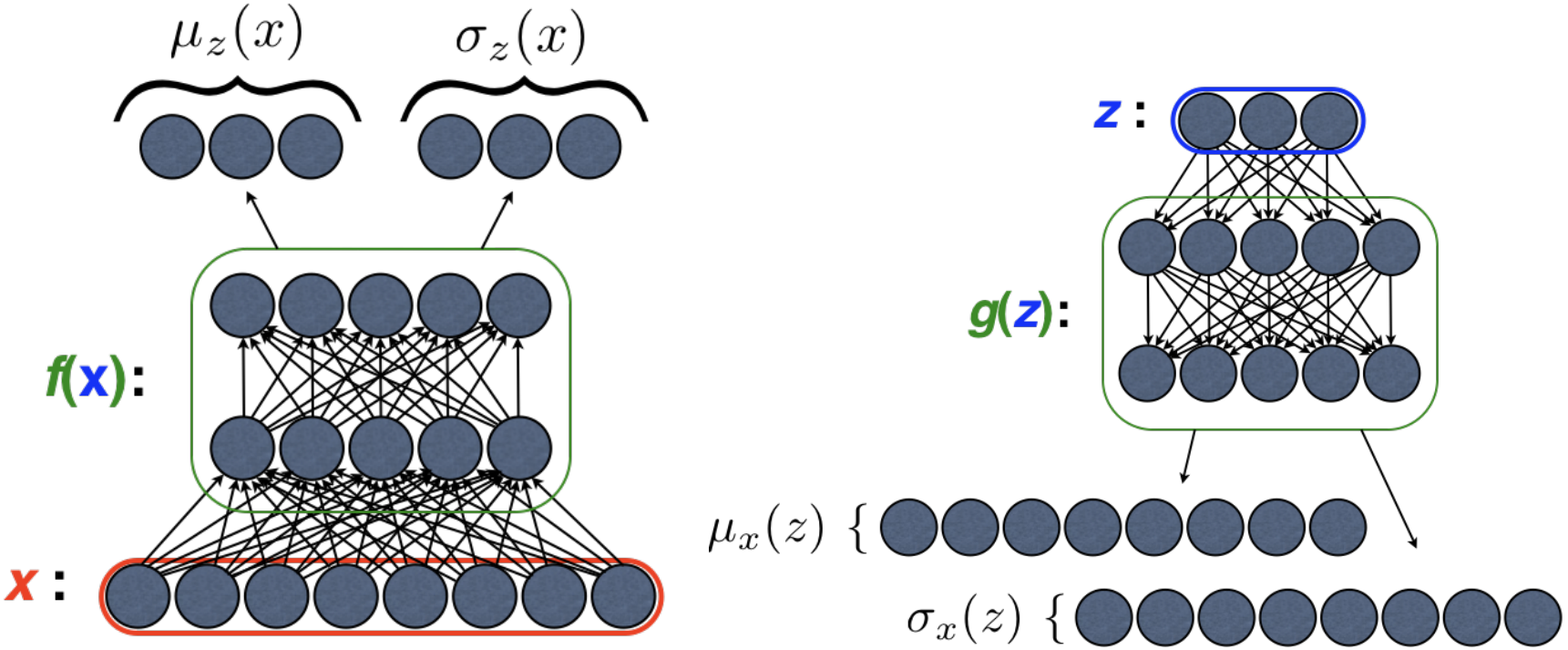
- $f(\mathbf{x}, \phi) = (\boldsymbol{\mu}_{\mathbf{z}}(\mathbf{x}), \boldsymbol{\sigma}_{\mathbf{z}}(\mathbf{x}))$ and $q_{\phi}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_{\mathbf{z}}(\mathbf{x}), \boldsymbol{\sigma}^2_{\mathbf{z}}(\mathbf{x}))$



VAE-MODEL DEFINITION

Usually:

- $f(\mathbf{x}, \phi) = (\mu_{\mathbf{z}}(\mathbf{x}), \sigma_{\mathbf{z}}(\mathbf{x}))$ and $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu_{\mathbf{z}}(\mathbf{x}), \sigma^2_{\mathbf{z}}(\mathbf{x}))$
- $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, 1)$
- $g(\mathbf{z}, \theta) = (\mu_{\mathbf{x}}(\mathbf{z}), \sigma_{\mathbf{x}}(\mathbf{z}))$ and $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mu_{\mathbf{x}}(\mathbf{z}), \sigma^2_{\mathbf{x}}(\mathbf{z}))$



VAE-PARAMETER FITTING: REPARAMETERIZATION TRICK

- Goal: Learn parameters ϕ and θ by maximizing

$$ELBO(\theta, \phi, \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - KL\big[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\big]$$

  based on gradient ascent.

- Idea: Approximate first term

$$\begin{aligned}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] &= \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{z};\boldsymbol{\mu}_\mathbf{z}(\mathbf{x}),\boldsymbol{\sigma}_\mathbf{z}(\mathbf{x}))}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(\mathbf{x}|\mathbf{z}^{(l)})\end{aligned}$$

- Problem: Given this average, how should one take derivatives w.r.t. ϕ?

VAE-PARAMETER FITTING: REPARAMETERIZATION TRICK

Recall: Linear transformation of a normal random variable

Let $\boldsymbol{\epsilon}$ be standard normally distributed, i.e. $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{1})$. Then for $\mathbf{z} = \boldsymbol{\epsilon} \cdot \boldsymbol{\sigma} + \boldsymbol{\mu}$ it holds: $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$

VAE-PARAMETER FITTING: REPARAMETERIZATION TRICK

Recall: Linear transformation of a normal random variable

Let ϵ be standard normally distributed, i.e. $\epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{1})$. Then for $\mathbf{z} = \epsilon \cdot \boldsymbol{\sigma} + \boldsymbol{\mu}$ it holds: $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$

Back to our problem:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] = \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{z};\boldsymbol{\mu}_\mathbf{z}(\mathbf{x}),\boldsymbol{\sigma}_\mathbf{z}(\mathbf{x}))}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big]$$
$$\approx \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}|\mathbf{z}^{(l)})$$

Solution: Define $\mathbf{z} = \epsilon \cdot \boldsymbol{\sigma}_\mathbf{z}(\mathbf{x}) + \boldsymbol{\mu}_\mathbf{z}(\mathbf{x})$ where $\epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{1})$.

VAE-PARAMETER FITTING: REPARAMETERIZATION TRICK

Recall: Linear transformation of a normal random variable

Let ϵ be standard normally distributed, i.e. $\epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{1})$. Then for $\mathbf{z} = \epsilon \cdot \sigma + \mu$ it holds: $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \mu, \sigma^2)$

Back to our problem:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] = \mathbb{E}_{\epsilon\sim\mathcal{N}(\epsilon;\mathbf{0},\mathbf{1})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] \approx \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}|\mathbf{z}^{(l)})$$

Solution: Define $\mathbf{z} = \epsilon \cdot \sigma_{\mathbf{z}}(\mathbf{x}) + \mu_{\mathbf{z}}(\mathbf{x})$ where $\epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{1})$.

VAE-PARAMETER FITTING: REPARAMETERIZATION TRICK

Recall: Linear transformation of a normal random variable

Let ϵ be standard normally distributed, i.e. $\epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{1})$. Then for $\mathbf{z} = \epsilon \cdot \boldsymbol{\sigma} + \boldsymbol{\mu}$ it holds: $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$

Back to our problem:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] = \mathbb{E}_{\epsilon\sim\mathcal{N}(\epsilon;\mathbf{0},\mathbf{1})}\big[\log p_\theta(\mathbf{x}|\mathbf{z} = \epsilon \cdot \boldsymbol{\sigma}_\mathbf{z}(\mathbf{x}) + \boldsymbol{\mu}_\mathbf{z}(\mathbf{x}))\big]$$
$$\approx \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}|\mathbf{z}^{(l)})$$

Solution: Define $\mathbf{z} = \epsilon \cdot \boldsymbol{\sigma}_\mathbf{z}(\mathbf{x}) + \boldsymbol{\mu}_\mathbf{z}(\mathbf{x})$ where $\epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{1})$.

VAE-PARAMETER FITTING: REPARAMETERIZATION TRICK

Recall: Linear transformation of a normal random variable

Let ϵ be standard normally distributed, i.e. $\epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{1})$. Then for $\mathbf{z} = \epsilon \cdot \sigma + \mu$ it holds: $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \mu, \sigma^2)$

Back to our problem:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] = \mathbb{E}_{\epsilon\sim\mathcal{N}(\epsilon;\mathbf{0},\mathbf{1})}\big[\log p_\theta(\mathbf{x}|\mathbf{z} = \epsilon \cdot \sigma_\mathbf{z}(\mathbf{x}) + \mu_\mathbf{z}(\mathbf{x}))\big]$$

$$\approx \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}|\mathbf{z}^{(l)} = \epsilon^{(l)} \cdot \sigma_\mathbf{z}(\mathbf{x}) + \mu_\mathbf{z}(\mathbf{x}))$$

Solution: Define $\mathbf{z} = \epsilon \cdot \sigma_\mathbf{z}(\mathbf{x}) + \mu_\mathbf{z}(\mathbf{x})$ where $\epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{1})$.

VAE-TRAINING WITH BACKPROPAGATION

Due to the reparameterization trick, we can simultaneously train both the generative model $p_{\theta}(\mathbf{x}|\mathbf{z})$ and the inference model $q_{\phi}(\mathbf{z}|\mathbf{x})$ by maximizing the ELBO based on gradient ascent and backpropagation.

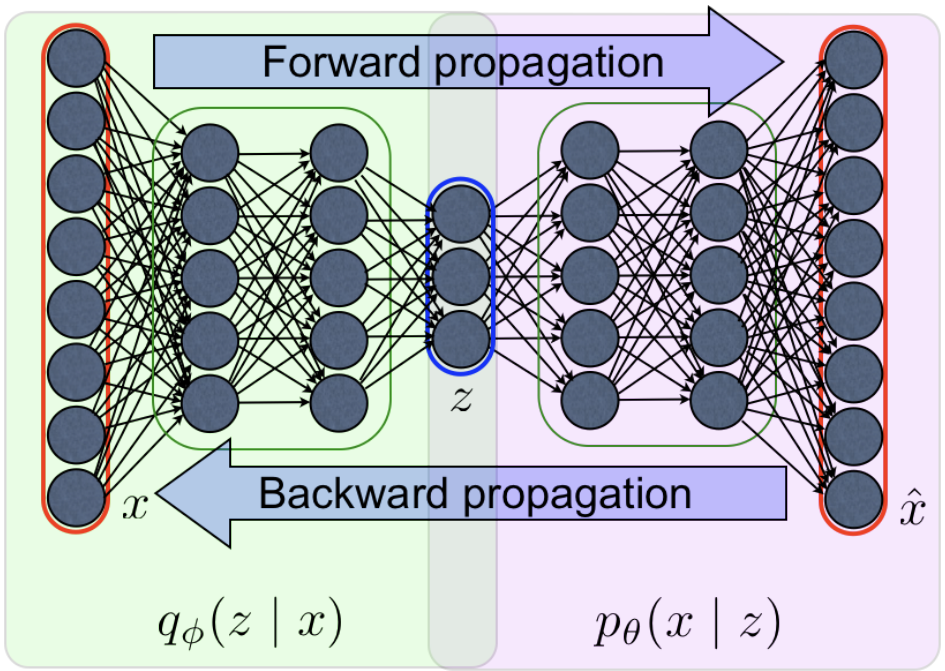
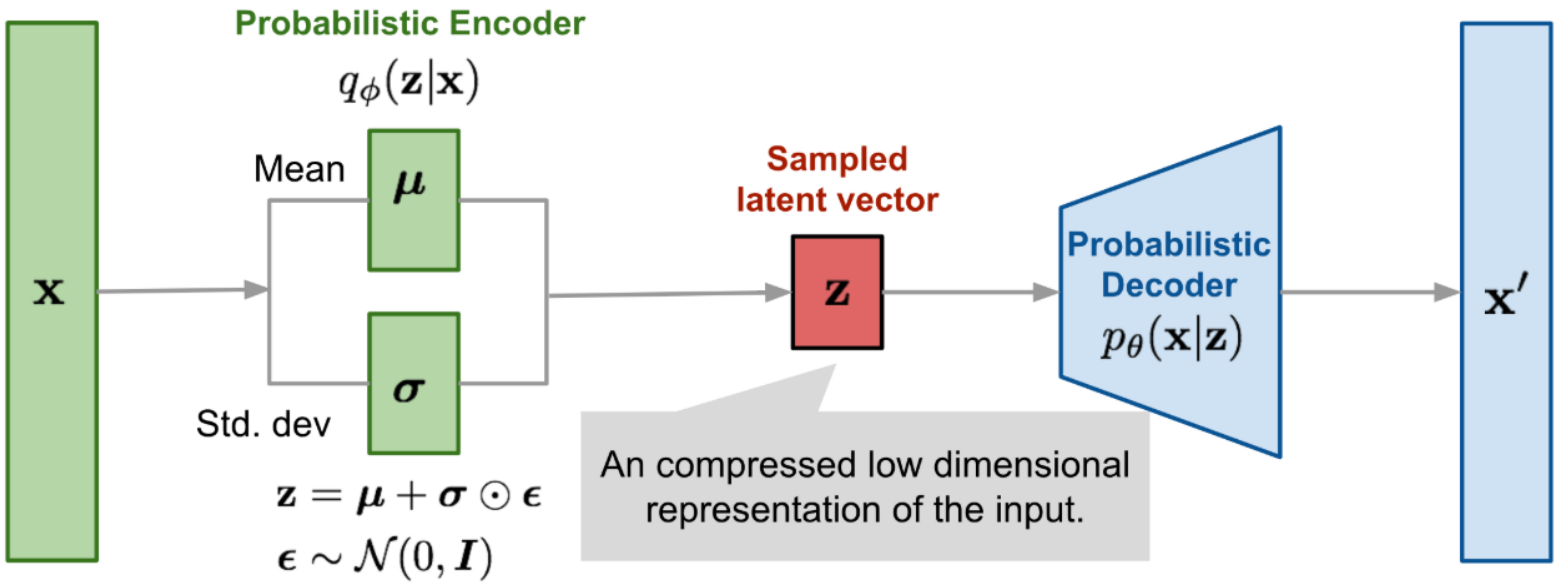


ILLUSTRATION OF FORWARD PASS



credit: Lillian Weng

Figure: Illustration of the forward pass in a VAE. For a given $\mathbf{x}$, the encoder network outputs the mean(s) $\boldsymbol{\mu}_{\mathbf{z}}(\mathbf{x})$ and standard deviation(s) $\boldsymbol{\sigma}_{\mathbf{z}}(\mathbf{x})$ of $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_{\mathbf{z}}(\mathbf{x}), \boldsymbol{\sigma}^2_{\mathbf{z}}(\mathbf{x}))$. From this, a vector $\mathbf{z}$ is sampled and fed to the decoder network which gives us $\mathbf{x}'$. If $\mathbf{x}$ is an image, $\mathbf{x}'$ is commonly interpreted as the reconstructed image.

LATENT VARIABLES LEARNED BY A VAE

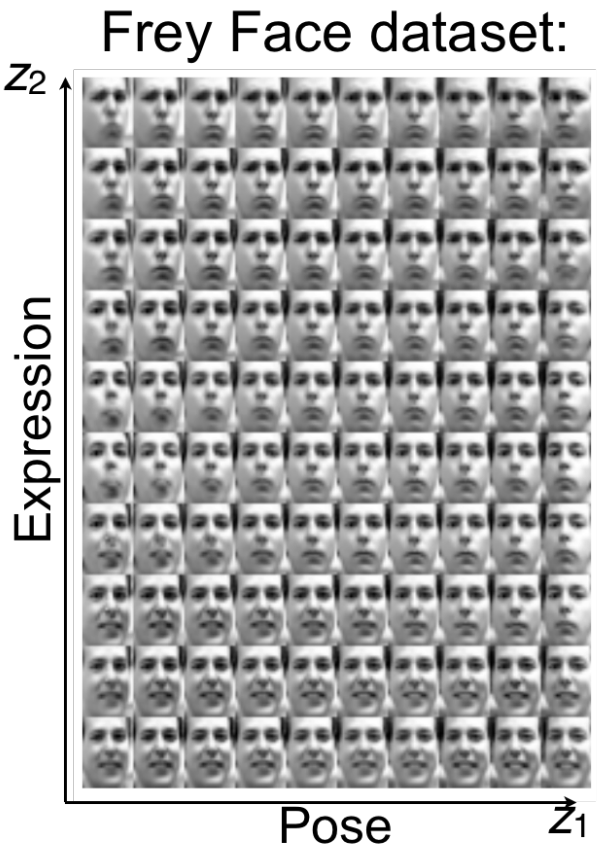


Figure: Images generated by a VAE are superimposed on the latent vectors that generated them. In the two-dimensional latent space of the VAE, the first dimension encodes the position of the face and the second dimension encodes the expression. Therefore, starting at any point in the latent space, if we move along either axis, the corresponding property will change in the generated image. (Goodfellow et al., 2016)

LATENT VARIABLES LEARNED BY A VAE

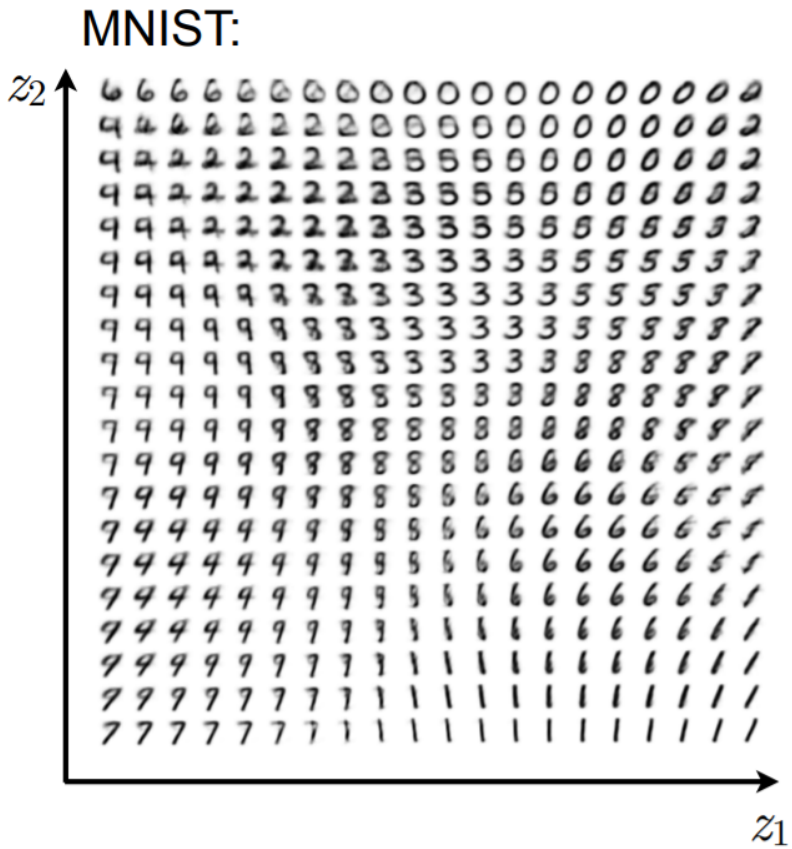
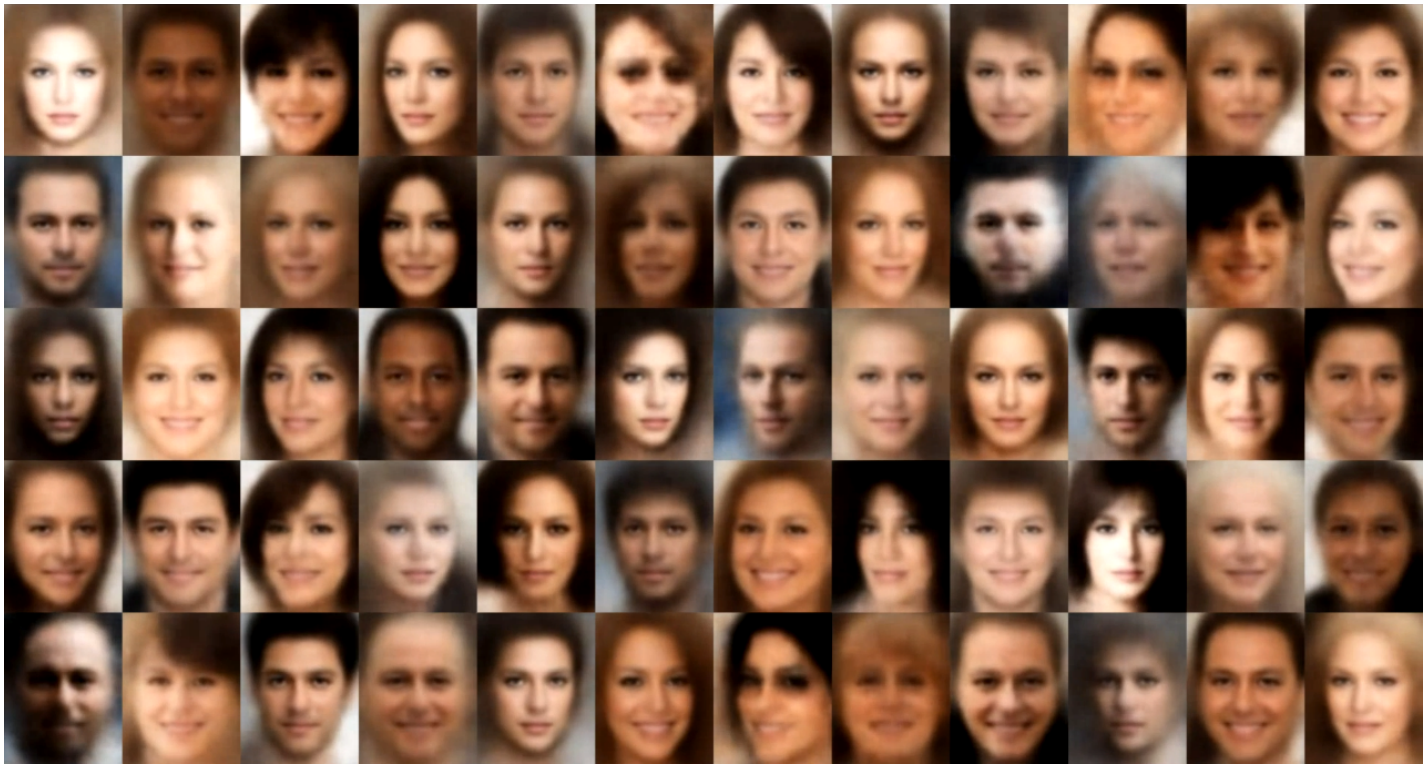


Figure: Images generated by a VAE are superimposed on the latent vectors that generated them. The two-dimensional latent space of the VAE captures much of the variation present in MNIST. Different regions in the latent space correspond to different digits in the generated images. (Goodfellow et al., 2016)

SAMPLES FROM A VANILLA VAE



Credit : Wojciech Mormul

Figure: Samples generated by a VAE that was trained on images of people's faces.

REFERENCES

Lilian Weng (2018)

From Autoencoder to Beta-VAE

https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html