2 most hot topics on NIPS 2016?

2 most hot topics on NIPS 2016

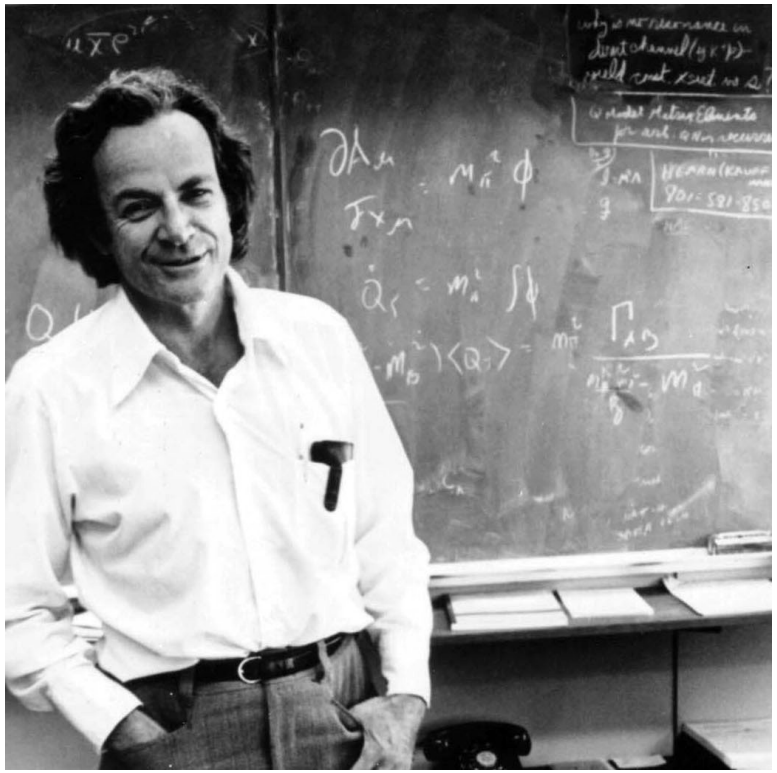
- Reinforcement Learning
- Generative models

Generative Models. Why?

Generative Models. Why?

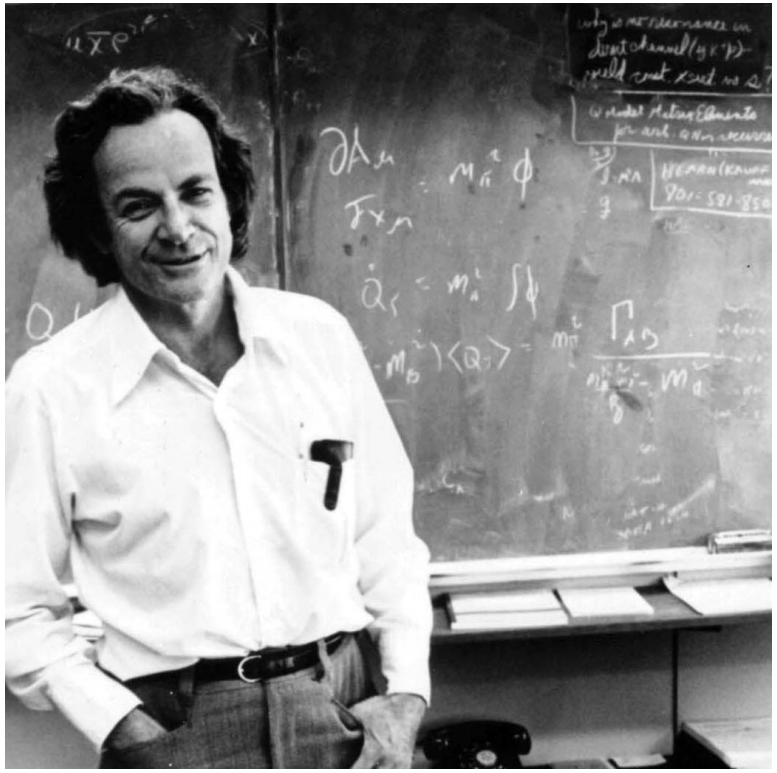
What I cannot create,
I do not understand.

Generative Models. Why?



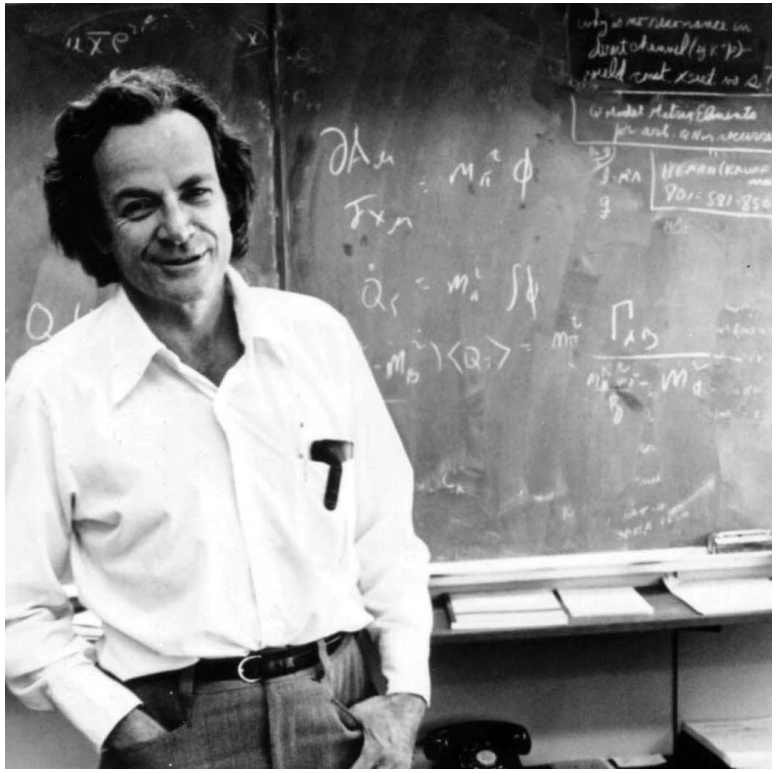
What I cannot create,
I do not understand.
Richard Feynman

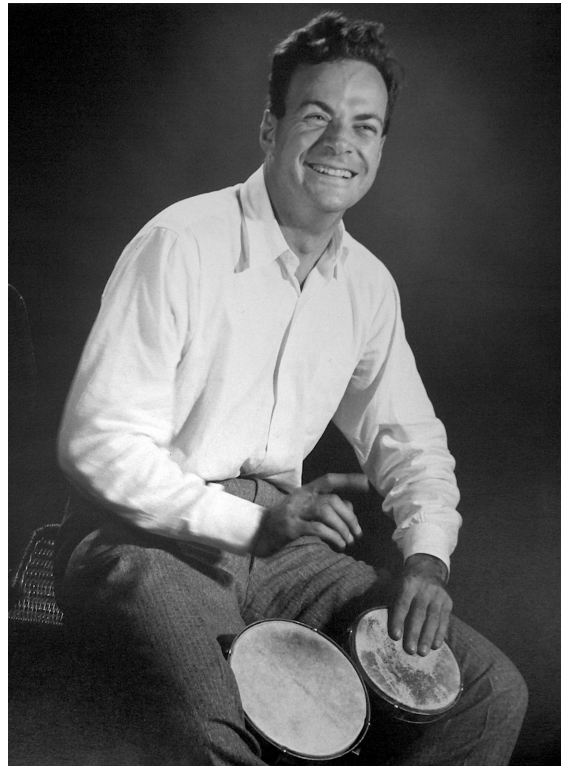
Generative Models. Why?



What I cannot create,
I do not understand.
Richard Feynman

Generative Models. Why?





AI Cake

Reinforcement Learning

Supervised Learning

Unsupervised Learning

Why?

- Most of the data is unsupervised
- Humans learn in an unsupervised way

Generative Models of Images

Density Estimators

- Train Inversible Transformation $\boldsymbol{x} = f(\boldsymbol{z})$, where $\boldsymbol{z}$ have fixed distribution e.g. $\mathcal{N}(0,1)$
- $p(\boldsymbol{x}) = p(\boldsymbol{z}) \det |\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}|^{-1}$
- *Examples: NICE, Real NVP…*

– Must have $dim(\boldsymbol{z}) = \dim(\boldsymbol{k})$

GANs

- Based on adversarial training
- *Examples: GAN, InfoGAN, StackGAN …*

– Tricky to train

– Tricky to evaluate

Variational Autoencoders

- Based on approximate Bayesian Inference
- *Examples: VAE, DRAW…*

– Currently cannot achieve good samples quality (blurry samples)

Generative Models of Images

Density Estimators

- Train Inversible Transformation $\boldsymbol{x} = f(\boldsymbol{z})$, where $\boldsymbol{z}$ have fixed distribution e.g. $\mathcal{N}(0,1)$
- $p(\boldsymbol{x}) = p(\boldsymbol{z}) \det |\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}|^{-1}$
- *Examples: NICE, Real NVP…*

– Must have $dim(\boldsymbol{z}) = \dim(\boldsymbol{k})$

GANs

- Based on adversarial training
- *Examples: GAN, InfoGAN, StackGAN …*

– Tricky to train

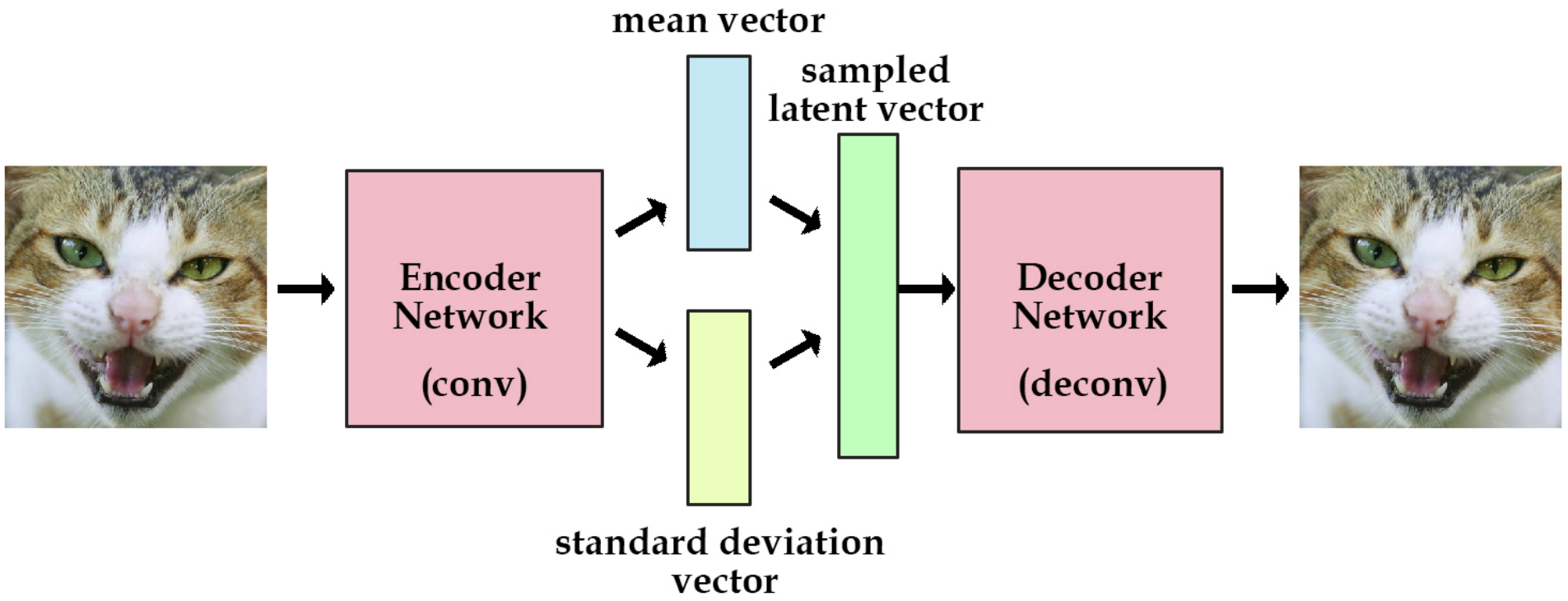
– Tricky to evaluate

Variational Autoencoders

- Based on approximate Bayesian Inference
- *Examples: VAE, DRAW…*

– **Currently** cannot achieve good samples quality (blurry samples)

VAE – Variational AutoEncoder



VAE = Variational Autoencoder

Why variational?

VAE = Variational Autoencoder

Why variational?

Variational inference!

Bayesian ML in a nutshell

Recall Bayes formula

$$P(a \mid b) =$$

Recall Bayes formula

$$P(a \mid b) = \frac{P(b \mid a) \cdot P(a)}{P(b)}$$

Recall Bayes formula

$$P(a \mid b) = \frac{P(b \mid a) \cdot P(a)}{P(b)} = \frac{P(b \mid a) \cdot P(a)}{\int_a P(b \mid a) \cdot P(a) da}$$

Recall Bayes formula

$$P(a \mid b) = \frac{P(b \mid a) \cdot P(a)}{\int_a P(b \mid a) \cdot P(a) da}$$

Recall Bayes formula

$$P(\mathbf{z} \mid \boldsymbol{x}) = \frac{P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z})}{\int_{\boldsymbol{z}} P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z}) d\boldsymbol{z}}$$

Holds when $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in \mathbb{R}^k$

Recall Bayes formula

$$P(\mathbf{z} \mid \boldsymbol{x}) = \frac{P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z})}{\int_{\boldsymbol{z}} P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z}) d\boldsymbol{z}}$$

Holds when $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in \mathbb{R}^k$

- $\boldsymbol{x}$ – observed data
- $\boldsymbol{z}$ – unobserved latent variables

Recall Bayes formula

$$P(\mathbf{z} \mid \boldsymbol{x}) = \frac{P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z})}{\int_{\boldsymbol{z}} P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z}) d\boldsymbol{z}}$$

- $P(\mathbf{z} \mid \boldsymbol{x})$ – posterior
- $P(\mathbf{x} \mid \boldsymbol{z})$ – likelihood
- $P(\boldsymbol{z})$ – prior

How to be Bayesian in 3(4) easy steps:

1. Make up your likelihood $P(\mathbf{x} \mid \boldsymbol{z})$
2. Make up your prior on $P(\boldsymbol{z})$ (regularization)
3. Infer $P(\mathbf{z} \mid \boldsymbol{x})$

How to be Bayesian in 3(4) easy steps:

1. Make up your likelihood $P(\mathbf{x} \mid \boldsymbol{z})$
2. Make up your prior on $P(\boldsymbol{z})$ (regularization)
3. Infer $P(\mathbf{z} \mid \boldsymbol{x})$
4. (optional) Find best model $P(\mathbf{x} \mid \boldsymbol{z}), P(\boldsymbol{z})$ maximizing $\int_{\boldsymbol{z}} P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z}) d\boldsymbol{z}$

How to be P easy s:

1. Ma $\boldsymbol{z}$
2. Ma ization)
3. In
4. $P(\boldsymbol{z})$

Recall Bayes formula

$$P(\mathbf{z} \mid \boldsymbol{x}) = \frac{P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z})}{\int_{\boldsymbol{z}} P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z}) d\boldsymbol{z}}$$

- What about denominator $\int_{\boldsymbol{z}} P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z}) d\boldsymbol{z}$?

Recall Bayes formula

$$P(\mathbf{z} \mid \boldsymbol{x}) = \frac{P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z})}{\int_{\boldsymbol{z}} P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z}) d\boldsymbol{z}}$$

- What about denominator $\int_{\boldsymbol{z}} P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z}) d\boldsymbol{z}$?
- We **don't** need it if we only need Maximum a-posteriori (MAP)

Recall Bayes formula

$$P(\mathbf{z} \mid \boldsymbol{x}) = \frac{P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z})}{\int_{\boldsymbol{z}} P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z}) d\boldsymbol{z}}$$

- What about denominator $\int_{\boldsymbol{z}} P(\mathbf{x} \mid \boldsymbol{z}) \cdot P(\boldsymbol{z}) d\boldsymbol{z}$?
- We **don't** need it if we only need Maximum a-posteriori (MAP)
- However, it's essential if we want to
  1. Sample from $P(\mathbf{z} \mid \boldsymbol{x})$
  2. Compute mean/variance/almost anything about $P(\mathbf{z} \mid \boldsymbol{x})$

How to compute denominator?

$$\int_{\boldsymbol{z}} P\left(\mathbf{x} \mid \boldsymbol{z}\right) \cdot P(\boldsymbol{z}) d\boldsymbol{z} \quad - ???$$

How to compute denominator?

$$\int_{\boldsymbol{z}} P\left(\mathbf{x} \mid \boldsymbol{z}\right) \cdot P(\boldsymbol{z}) d\boldsymbol{z} \quad - ???$$

- Analytically
  - \+ Exact
  - – Works only in some cases

How to compute denominator?

$$\int_{\boldsymbol{z}} P\left(\mathbf{x} \mid \boldsymbol{z}\right) \cdot P(\boldsymbol{z}) d\boldsymbol{z} \quad - ???$$

- Analytically
  - \+ Exact
  - – Works only in some cases
- Numerically
  - \+ Always possible
  - – Takes **AGES** if $k > 10$

How to compute denominator?

$$\int_{\mathbf{z}} P(\mathbf{x} \mid \mathbf{z}) \cdot P(\mathbf{z}) d\mathbf{z} \quad - ???$$

- Analytically
  - \+ Exact
  - – Works only in some cases
- Numerically
  - \+ Always possible
  - – Takes **AGES** if $k > 10$

We can’t compute posterior…

$$\int_{\mathbf{z}} P(\mathbf{x} \mid \mathbf{z}) \cdot P(\mathbf{z}) d\mathbf{z}$$

We can’t compute posterior… Approximate it!

$$\int_{\mathbf{z}} P(\mathbf{x} \mid \mathbf{z}) \cdot P(\mathbf{z}) d\mathbf{z}$$

$$q(\boldsymbol{z}) \approx p(\boldsymbol{z} \mid \boldsymbol{x})$$

- $q(\boldsymbol{z})$ – easy to work with
- $p(\boldsymbol{z} \mid \boldsymbol{x})$ – true posterior

Two ideas how to approximate $P(\mathbf{z} \mid \boldsymbol{x})$

- MCMC – Markov Chain Monte Carlo
  - *Idea:* Sample a lot of stuff that looks like $P(\mathbf{z} \mid \boldsymbol{x})$ somehow
  - *Examples:* Metropolis-Hastings, Gibbs sampling
- VI – Variational Inference
  - *Idea:* Find the best approximation in some parametric family $q(\boldsymbol{z}) = q_{\phi}(\boldsymbol{z})$
  - *Examples:* Mean-Field, SGVB

Two ideas how to approximate $P(\mathbf{z} \mid \boldsymbol{x})$

- MCMC – Markov Chain Monte Carlo
  - *Idea:* Sample a lot of stuff that looks like $P(\mathbf{z} \mid \boldsymbol{x})$ somehow
  - *Examples:* Metropolis-Hastings, Gibbs sampling
- VI – Variational Inference
  - *Idea:* Find the best approximation in some parametric family $q(\boldsymbol{z}) = q_{\phi}(\boldsymbol{z})$
  - *Examples:* Mean-Field, SGVB

Two ideas how to approximate $P(\mathbf{z} \mid \boldsymbol{x})$

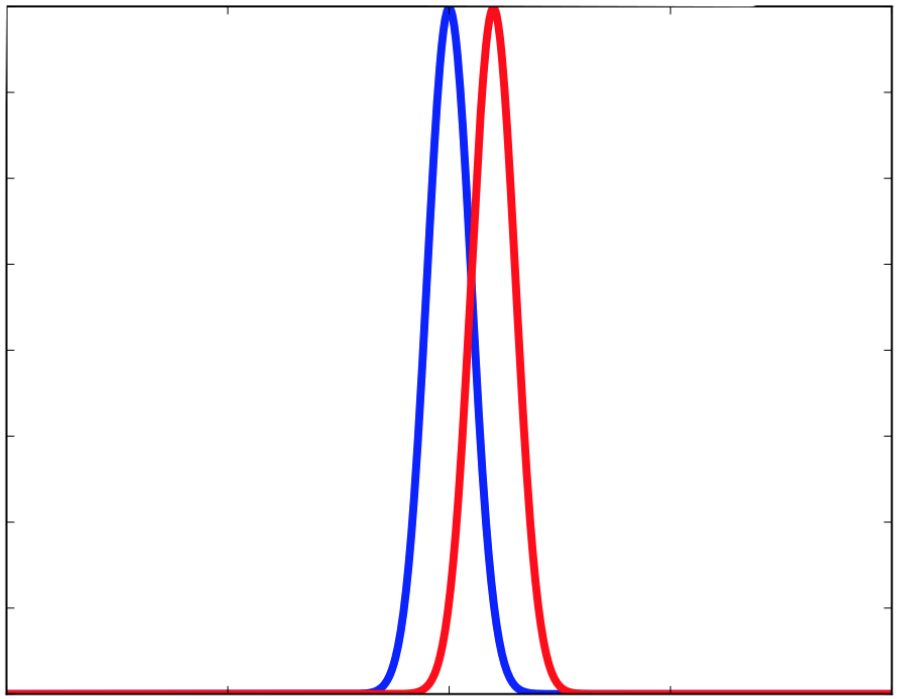
- MCMC – Markov Chain Monte Carlo
  - *Idea:* Sample a lot of stuff that looks like $P(\mathbf{z} \mid \boldsymbol{x})$ somehow
  - *Examples:* Metropolis-Hastings, Gibbs sampling
- VI – Variational Inference
  - *Idea:* Find the best approximation in some parametric family $q(\boldsymbol{z}) = q_{\phi}(\boldsymbol{z})$
  - *Examples:* Mean-Field, SGVB

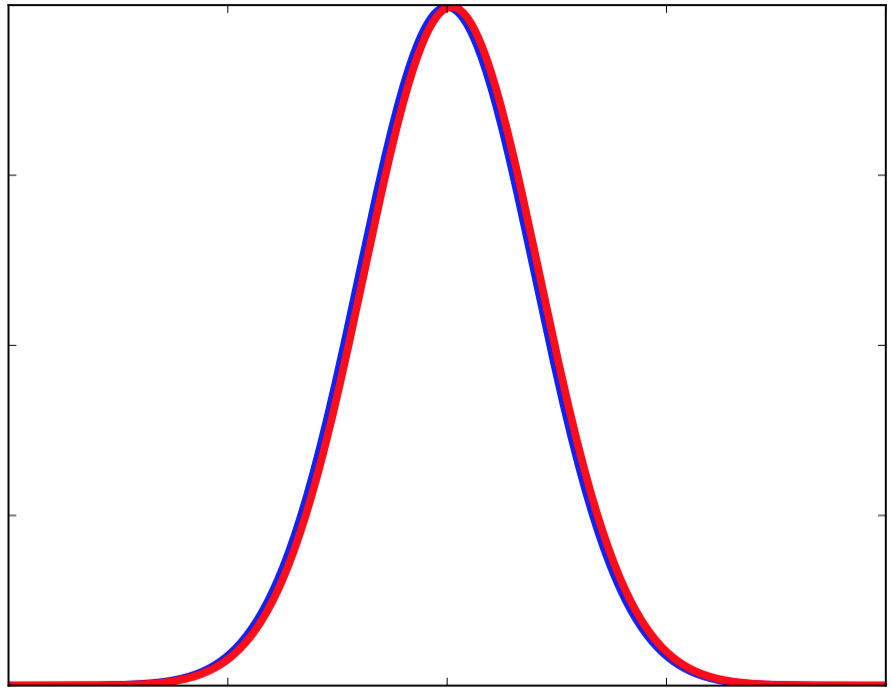
Two ideas how to approximate $P(\mathbf{z} \mid \boldsymbol{x})$

- MCMC – Markov Chain Monte Carlo
  - *Idea:* Sample a lot of stuff that looks like $P(\mathbf{z} \mid \boldsymbol{x})$ *somehow*
  - *Examples:* Metropolis-Hastings, Gibbs sampling
- VI – Variational Inference
  - *Idea:* Find *the best* approximation in some parametric family $q(\boldsymbol{z}) = q_{\phi}(\boldsymbol{z})$
  - *Examples:* Mean-Field, SGVB

How to evaluate quality of approximation?

$$q(\boldsymbol{z}) \approx p(\boldsymbol{z} \mid \boldsymbol{x})$$





How to evaluate quality of approximation?

Kullback-Leibler divergence:

$$KL(p_1 \parallel p_2) = \mathbb{E}_{z\sim p_1} \log \frac{p_1(\boldsymbol{z})}{p_2(\boldsymbol{z})} = \int_{\boldsymbol{z}} p_1(\boldsymbol{z}) \log \frac{p_1(\boldsymbol{z})}{p_2(\boldsymbol{z})} d\boldsymbol{z}$$

- $KL(p_1 \parallel p_2) \geq 0$
- $KL(p_1 \parallel p_2) = 0 \Leftrightarrow p_1 = p_2$
- $KL(p_1 \parallel p_2) \neq KL(p_2 \parallel p_1)$

How to evaluate quality of approximation?

Kullback-Leibler divergence:

$$KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) = \mathbb{E}_q \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}$$

ELBO: Evidence Lower Bound

$\log p(\boldsymbol{x})$

ELBO: Evidence Lower Bound

$\log p(\boldsymbol{x}) = \mathbb{E}_{z\sim q} \log p(\boldsymbol{x})$ // we integrate over z here. $\int_{\boldsymbol{z}} q(\boldsymbol{z}) d\boldsymbol{z} = 1$

ELBO: Evidence Lower Bound

$$\log p(\boldsymbol{x}) = \mathbb{E}_{z\sim q} \log p(\boldsymbol{x}) =$$

$$= \mathbb{E}_{z\sim q} \log \left[\frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right] \qquad // \; p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{x})}$$

ELBO: Evidence Lower Bound

$$\log p(\boldsymbol{x}) = \mathbb{E}_{z\sim q} \log p(\boldsymbol{x}) =$$

$$= \mathbb{E}_{z\sim q} \log \left[\frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right] =$$

$$= \mathbb{E}_{z\sim q} \left[\log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} + \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right]$$

ELBO: Evidence Lower Bound

$$
\begin{aligned}
\log p(\boldsymbol{x}) = \; & \mathbb{E}_{z\sim q} \log p(\boldsymbol{x}) = \\
= \; & \mathbb{E}_{z\sim q} \log \left[\frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right] = \\
= \; & \mathbb{E}_{z\sim q} \left[\log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} + \log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})} \right] = \\
= \; & KL(q(\boldsymbol{z}) \;\|\; p(\boldsymbol{z}|\boldsymbol{x})) + \mathbb{E}_{z\sim q}[\log p(\boldsymbol{x},\boldsymbol{z}) - \log q(\boldsymbol{z})]
\end{aligned}
$$

ELBO: Evidence Lower Bound

$$
\begin{aligned}
\log p(\boldsymbol{x}) = \; & \mathbb{E}_{z\sim q} \log p(\boldsymbol{x}) = \\
& = \mathbb{E}_{z\sim q} \log \left[\frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right] = \\
& = \mathbb{E}_{z\sim q} \left[\log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} + \log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})} \right] = \\
& = KL(q(\boldsymbol{z}) \;\|\; p(\boldsymbol{z}|\boldsymbol{x})) + \mathbb{E}_{z\sim q}[\log p(\boldsymbol{x},\boldsymbol{z}) - \log q(\boldsymbol{z})] \\
& \qquad \geq 0 \qquad\qquad\qquad\qquad \mathcal{L}(q)
\end{aligned}
$$

ELBO: Evidence Lower Bound

$$\log p(\boldsymbol{x}) = KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) + \mathcal{L}(q)$$

$$\mathcal{L}(q) \to \max_q \quad \Leftrightarrow KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) \to \min_q$$

ELBO: Evidence Lower Bound

$$\log p(\boldsymbol{x}) = KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) + \mathcal{L}(q)$$

$$\mathcal{L}(q) \to \max_q \quad \Leftrightarrow KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) \to \min_q$$

Let's maximize ELBO instead of minimizing KL!

- Now we don't need to know how $p(\boldsymbol{z}|\boldsymbol{x})$ looks like

ELBO: Evidence Lower Bound

$$\log p(\boldsymbol{x}) = KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) + \mathcal{L}(q)$$

$$\mathcal{L}(q) \to \max_q \quad \Leftrightarrow KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) \to \min_q$$

Let's maximize ELBO instead of minimizing KL!

- Now we don't need to know how $p(\boldsymbol{z}|\boldsymbol{x})$ looks like
- Note 1: we can maximize $\mathcal{L}$ wrt $q(\boldsymbol{z})$ and $p(\boldsymbol{x}|\boldsymbol{z})$ finding best model and making inference simultaneously!

ELBO: Evidence Lower Bound

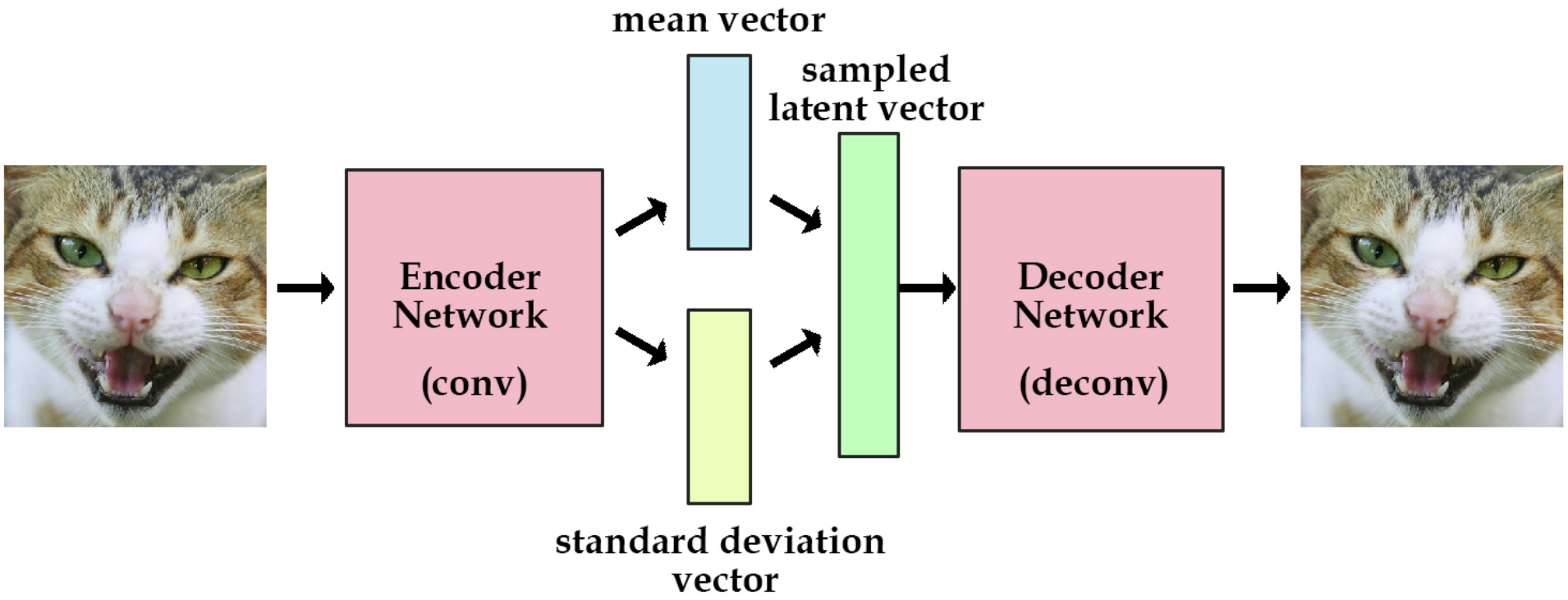
$$\log p(\boldsymbol{x}) = KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) + \mathcal{L}(q)$$

$$\mathcal{L}(q) \to \max_q \quad \Leftrightarrow KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) \to \min_q$$

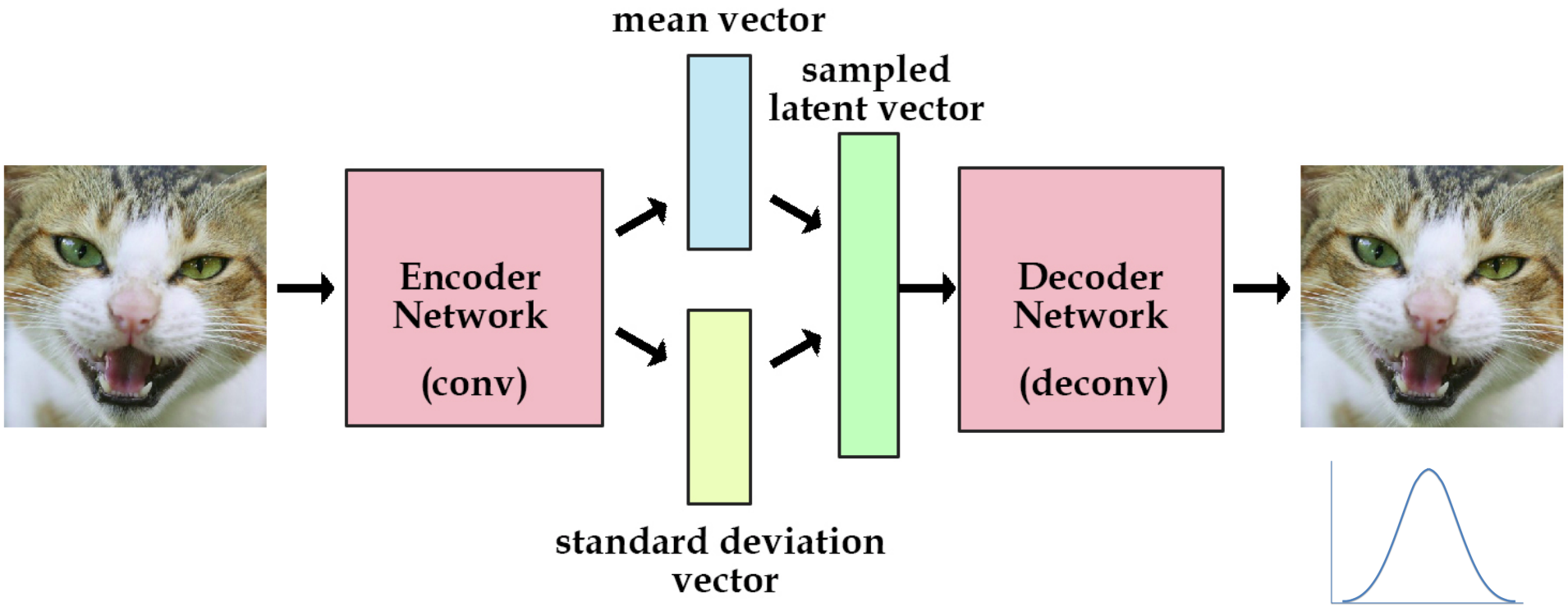
Let's maximize ELBO instead of minimizing KL!

- Now we don't need to know how $p(\boldsymbol{z}|\boldsymbol{x})$ looks like
- Note 1: we can maximize $\mathcal{L}$ wrt $q(\boldsymbol{z})$ and $p(\boldsymbol{x}|\boldsymbol{z})$ finding best model and making inference simultaneously!
- Note 2: $q(\boldsymbol{z})$ can be parameterized as a function of $\boldsymbol{x}$: $q_\phi(\boldsymbol{z}|\boldsymbol{x})$

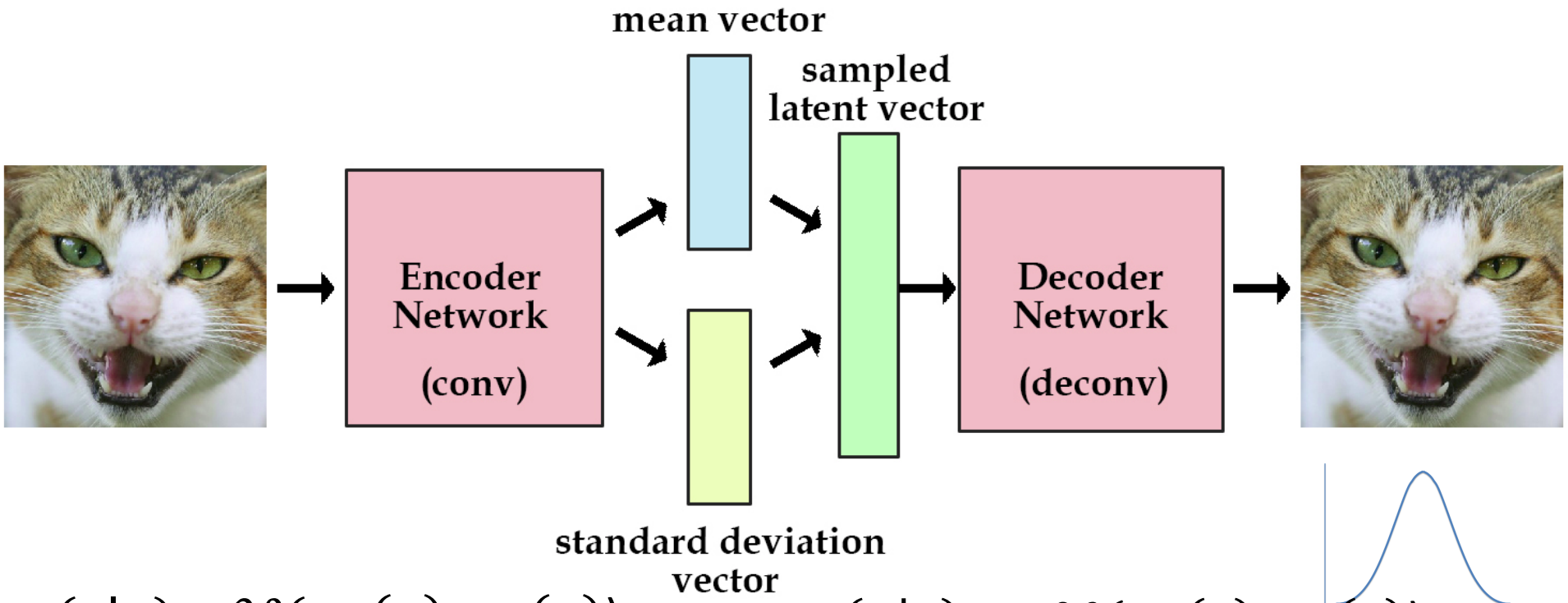
Variational Autoencoder



Variational Autoencoder



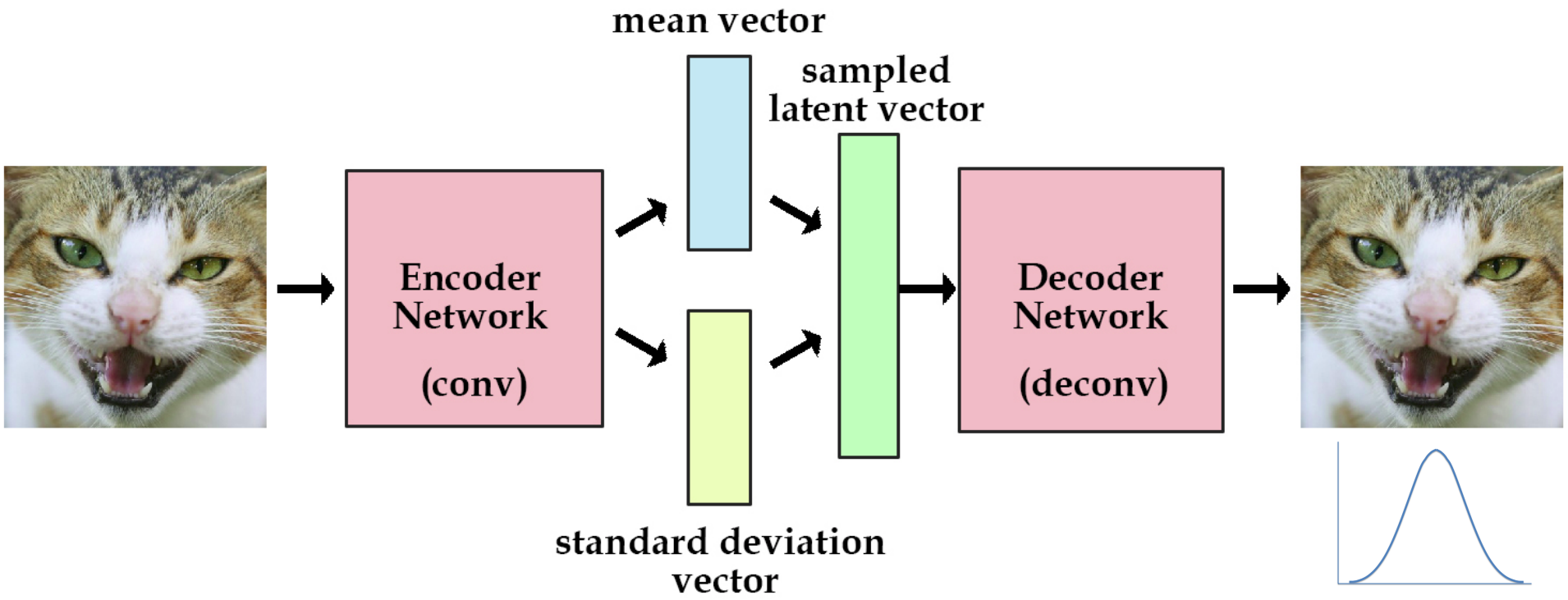
Variational Autoencoder



$$q_\varphi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\mu_\varphi(\boldsymbol{x}), \sigma_\varphi(\boldsymbol{x}))$$

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\mu_\theta(\boldsymbol{z}), \sigma_\theta(\boldsymbol{z}))$$

$$\mathcal{L}(q) = \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\varphi(\boldsymbol{z})\right] \to \max_{\varphi,\theta}$$



$$q_\varphi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\mu_\varphi(\boldsymbol{z}), \sigma_\varphi(\boldsymbol{z}))$$

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\mu_\theta(\boldsymbol{x}), \sigma_\theta(\boldsymbol{x}))$$

$$\mathcal{L}(q) = \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_\varphi(\boldsymbol{z}|\boldsymbol{x})\right] \to \max_{\varphi,\theta}$$

$$\mathcal{L}(q) = \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) + \log p_\theta(\boldsymbol{z}) - \log q_\varphi(\boldsymbol{z}|\boldsymbol{x})\right]$$

$$q_\varphi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\mu_\varphi(\boldsymbol{z}), \sigma_\varphi(\boldsymbol{z}))$$

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\mu_\theta(\boldsymbol{x}), \sigma_\theta(\boldsymbol{x}))$$

$$\mathcal{L}(q) = \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\varphi(\boldsymbol{z}|\boldsymbol{x})\right] \to \max_{\varphi,\theta}$$

$$\mathcal{L}(q) = \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) + \log p_\theta(\boldsymbol{z}) - \log q_\varphi(\boldsymbol{z}|\boldsymbol{x})\right] =$$
$$= \mathbb{E}_{z\sim q}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathbb{E}_{z\sim q}\left[\log q_\varphi(\boldsymbol{z}|\boldsymbol{x}) - \log p_\theta(\boldsymbol{z})\right]$$

$q_\varphi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\mu_\varphi(\boldsymbol{z}), \sigma_\varphi(\boldsymbol{z}))$

$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\mu_\theta(\boldsymbol{x}), \sigma_\theta(\boldsymbol{x}))$

$$\mathcal{L}(q) = \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\varphi(\boldsymbol{z}|\boldsymbol{x})\right] \to \max_{\varphi,\theta}$$

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) + \log p_\theta(\boldsymbol{z}) - \log q_\varphi(\boldsymbol{z}|\boldsymbol{x})\right] = \\ &= \mathbb{E}_{z\sim q}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathbb{E}_{z\sim q}\left[\log q_\varphi(\boldsymbol{z}|\boldsymbol{x}) - \log p_\theta(\boldsymbol{z})\right] = \\ &= \mathbb{E}_{z\sim q}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - KL\left(q_\varphi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})\right)\end{aligned}$$

$q_\varphi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\mu_\varphi(\boldsymbol{z}), \sigma_\varphi(\boldsymbol{z}))$

$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\mu_\theta(\boldsymbol{x}), \sigma_\theta(\boldsymbol{x}))$

$$\mathcal{L}(q) = \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\varphi(\boldsymbol{z}|\boldsymbol{x})\right] \to \max_{\varphi,\theta}$$

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) + \log p_\theta(\boldsymbol{z}) - \log q_\varphi(\boldsymbol{z}|\boldsymbol{x})\right] = \\ &= \mathbb{E}_{z\sim q}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathbb{E}_{z\sim q}\left[\log q_\varphi(\boldsymbol{z}|\boldsymbol{x}) - \log p_\theta(\boldsymbol{z})\right] = \\ &= \mathbb{E}_{z\sim q}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - KL\left(q_\varphi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})\right) \to \max_{\varphi,\theta}\end{aligned}$$

$$q_\varphi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\mu_\varphi(\boldsymbol{z}), \sigma_\varphi(\boldsymbol{z}))$$

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\mu_\theta(\boldsymbol{x}), \sigma_\theta(\boldsymbol{x}))$$

$$\mathcal{L}(q) = \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\varphi(\boldsymbol{z}|\boldsymbol{x})\right] \to \max_{\varphi,\theta}$$

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_{z\sim q}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) + \log p_\theta(\boldsymbol{z}) - \log q_\varphi(\boldsymbol{z}|\boldsymbol{x})\right] = \\ &= \mathbb{E}_{z\sim q}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathbb{E}_{z\sim q}\left[\log q_\varphi(\boldsymbol{z}|\boldsymbol{x}) - \log p_\theta(\boldsymbol{z})\right] = \\ &= \mathbb{E}_{z\sim q}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - KL\left(q_\varphi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})\right) \to \max_{\varphi,\theta}\end{aligned}$$

Reconstruction error *Regularization term*

$$q_\varphi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\mu_\varphi(\boldsymbol{z}), \sigma_\varphi(\boldsymbol{z}))$$

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\mu_\theta(\boldsymbol{x}), \sigma_\theta(\boldsymbol{x}))$$

$$\mathcal{L}(\varphi, \theta) = \mathbb{E}_{z \sim q}\left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_\varphi(\boldsymbol{z})\right] \to \max_{\varphi, \theta}$$

Training algorithm:

1. Take batch of data **x**. Sample batch of gaussian noise ε.
2. Pass **x** through encoder to get $\mu_\varphi(\boldsymbol{x}), \sigma_\varphi(\boldsymbol{x})$.
3. $\boldsymbol{z} = \sigma_\varphi(\boldsymbol{x})\boldsymbol{\varepsilon} + \mu_\varphi(\boldsymbol{x})$
4. Pass $\boldsymbol{z}$ through decoder to get $\mu_\theta(\boldsymbol{z}), \sigma_\theta(\boldsymbol{z})$
5. Compute and backpropagate estimate of loss $\mathcal{L}(\varphi, \theta)$: $\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_\varphi(\boldsymbol{z})$
6. Make gradient updates for decoder and encoder with your favourite SGD-like algorithm

Thank you for your attention!