Variational Autoencoders

Kate Farrahi

ECS Southampton

May 5, 2022

Variational Autoencoder (VAE)

- VAEs architecturally similar to autoencoders (AEs).
- VAEs (vs AEs) significantly different in their goal and mathematical formulation.
- AEs map the input into a fixed vector.
- However, VAEs map the input into a distribution.
- VAEs are a combination of neural networks (AEs) and **graphical models.**

Graphical Models (Background)

- A graphical model is a probabilistic model for which a graph expresses the conditional dependence structure between random variables.
- Graphical models are commonly used in probability theory, statistics —particularly Bayesian statistics— and machine learning. [1]

[1] Definition taken from Wikipedia

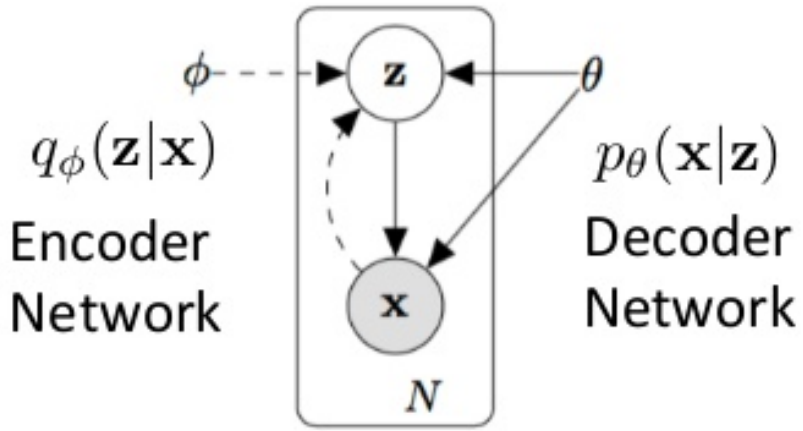
KL Divergence (Background)

- Kullback–Leibler divergence, $D_{\mathrm{KL}}(P \parallel Q)$: a measure of how one probability distribution Q is different from a second, reference probability distribution P. [2]
- A simple interpretation of the divergence of P from Q is the expected excess surprise from using Q as a model when the actual distribution is P.
- While it is a distance, it is not a metric, the most familiar type of distance: it is asymmetric in the two distributions.

[2] Definition taken from Wikipedia

Variational Autoencoders (VAEs)

Variational Autoencoder



Minimize: $D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})]$

Intractable: $p_\theta(\mathbf{z}|\mathbf{x}) = \dfrac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})}{p_\theta(\mathbf{x})}$

3

[3] Auto-Encoding Variational Bayes `https://arxiv.org/abs/1312.6114`

Variational Autoencoders (VAEs)

The distance loss just defined is expanded as

$$
\begin{aligned}
D_{KL}(q_\Phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) &= \int q_\Phi(\mathbf{z} \mid \mathbf{x}) \log \frac{q_\Phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z} \mid \mathbf{x})} \, d\mathbf{z} \\
&= \int q_\Phi(\mathbf{z} \mid \mathbf{x}) \log \frac{q_\Phi(\mathbf{z} \mid \mathbf{x}) p_\theta(\mathbf{x})}{p_\theta(\mathbf{z}, \mathbf{x})} \, d\mathbf{z} \\
&= \int q_\Phi(\mathbf{z} \mid \mathbf{x}) \left(\log(p_\theta(\mathbf{x})) + \log \frac{q_\Phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z}, \mathbf{x})} \right) d\mathbf{z} \\
&= \log(p_\theta(\mathbf{x})) + \int q_\Phi(\mathbf{z} \mid \mathbf{x}) \log \frac{q_\Phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z}, \mathbf{x})} \, d\mathbf{z} \\
&= \log(p_\theta(\mathbf{x})) + \int q_\Phi(\mathbf{z} \mid \mathbf{x}) \log \frac{q_\Phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z})} \, d\mathbf{z} \\
&= \log(p_\theta(\mathbf{x})) + E_{\mathbf{z} \sim q_\Phi(\mathbf{z} \mid \mathbf{x})} (\log \frac{q_\Phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z})} - \log(p_\theta(\mathbf{x} \mid \mathbf{z}))) \\
&= \log(p_\theta(\mathbf{x})) + D_{KL}(q_\Phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z})) - E_{\mathbf{z} \sim q_\Phi(\mathbf{z} \mid \mathbf{x})} (\log(p_\theta(\mathbf{x} \mid \mathbf{z})))
\end{aligned}
$$

At this point, it is possible to rewrite the equation as

$$
\log(p_\theta(\mathbf{x})) - D_{KL}(q_\Phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) = E_{\mathbf{z} \sim q_\Phi(\mathbf{z} \mid \mathbf{x})} (\log(p_\theta(\mathbf{x} \mid \mathbf{z}))) - D_{KL}(q_\Phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z}))
$$

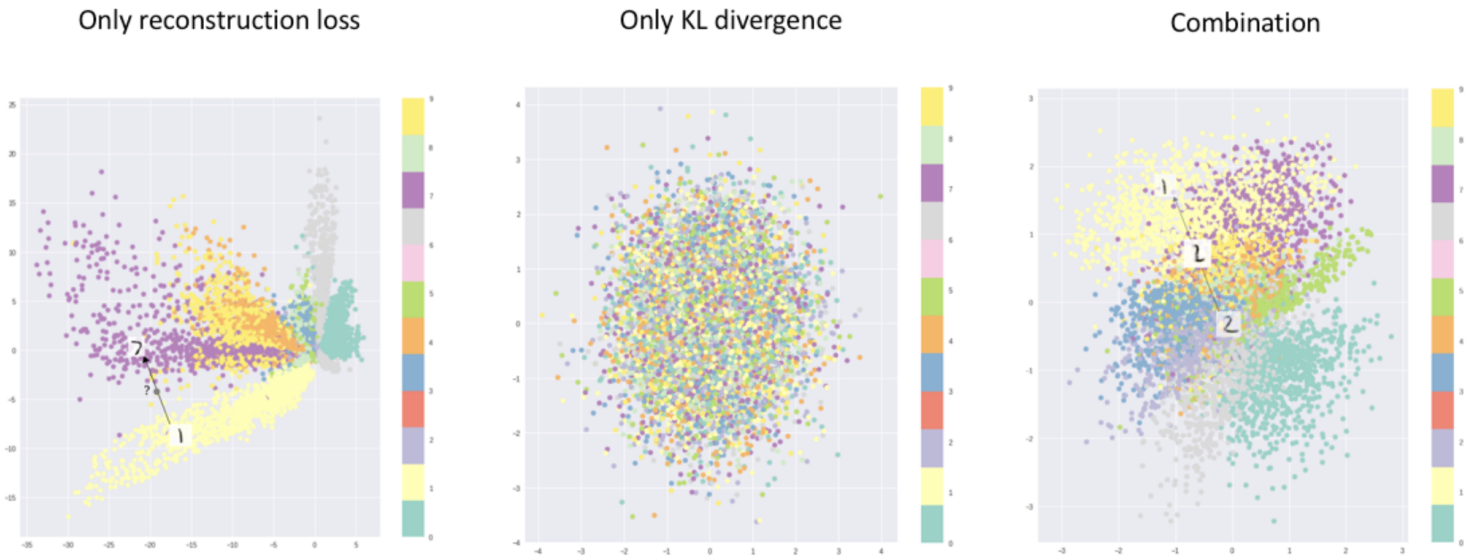
Evidence Lower Bound (ELBO) Loss

$$L_{VAE}(\theta, \phi) = -\mathbb{E}_{z \sim q_\phi(z|x)} log(p_\theta(x|z)) + D_{KL}(q_\phi(z|x)||p_\theta(z))$$

- We are trying to minimize the ELBO loss with respect to the model parameters.

Why Autoencoder?

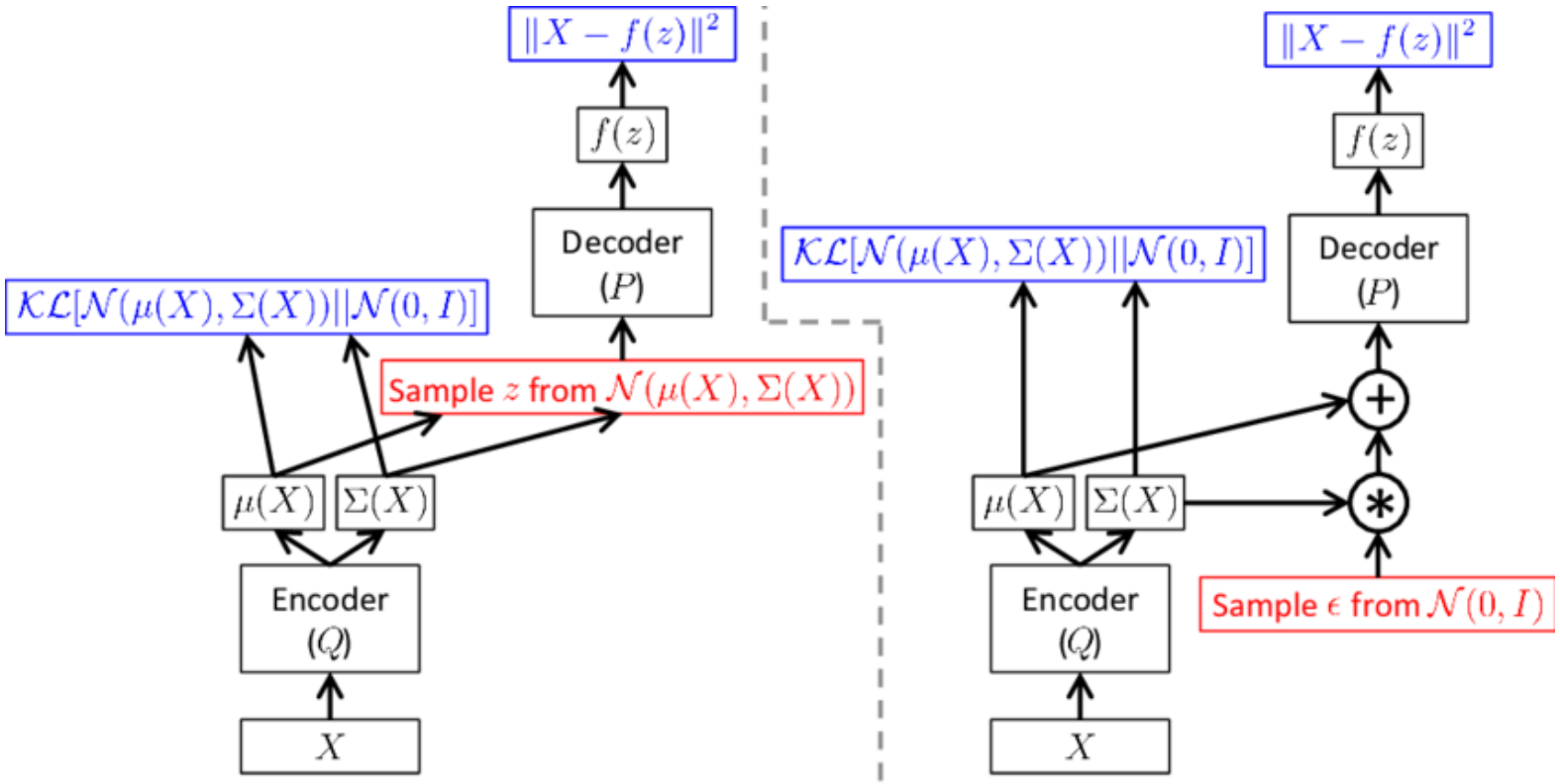
- The reconstruction term, forces each image to be unique and spread out.
- The KL term will push all the images towards the same prior.



[4]

[4] Figure taken from https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

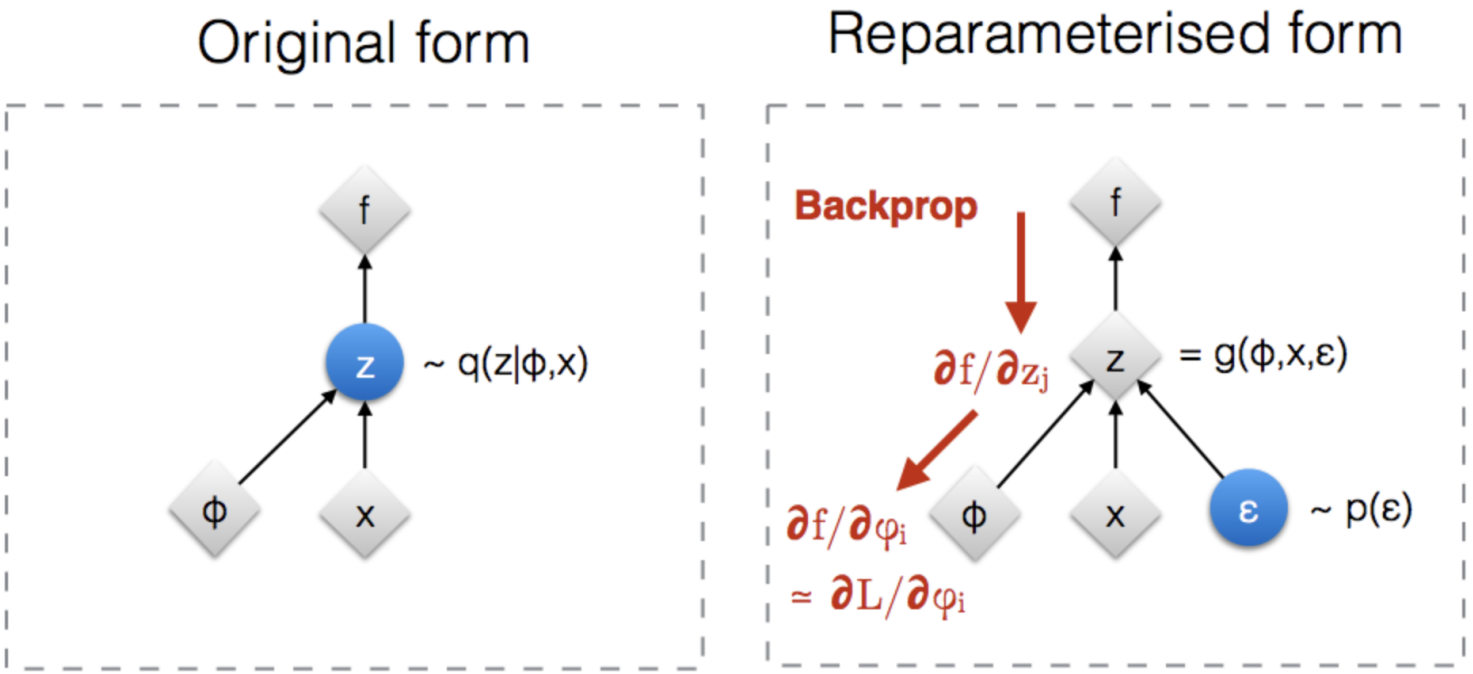
Training Procedure



[5]

[5] Figure taken from Carl Doersch tutorial

Reparametrization Trick Visualisation



: Deterministic node

: Random node

[Kingma, 2013]
[Bengio, 2013]
[Kingma and Welling 2014]
[Rezende et al 2014]

VAE Models and Performance

- VAEs can be used with any kind of data
  - the distributions and network architecture just needs to be set accordingly
  - e.g. it's common to use convolutions in the encoder and transpose convolutions in (Gaussian) decoder for image data
- VAEs have nice learning dynamics; they tend to be easy to optimise with stable convergence
- VAEs have a reputation for producing blurry reconstructions of images
  - Not fully understood why, but most likely related to a side effect of maximum-likelihood training
- VAEs tend to only utilise a small subset of the dimensions of $\boldsymbol{z}$

Reconstructions Example