Sparsity in Variational Autoencoders

First International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2019)

Barcelona 20-22 march 2019

Andrea Asperti

DISI - Department of Informatics: Science and Engineering
University of Bologna
Mura Anteo Zamboni 7, 40127, Bologna, ITALY
andrea.asperti@unibo.it

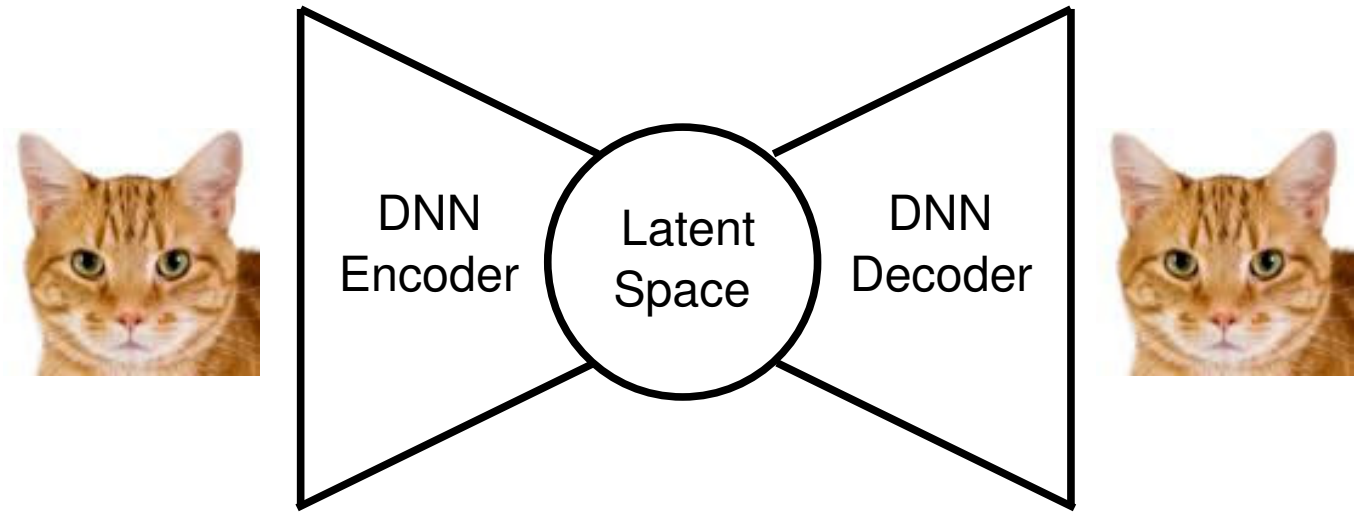
Content

- Variational Autoencoder
- Sparsity/overpruning

Investigation and discussion in the article.

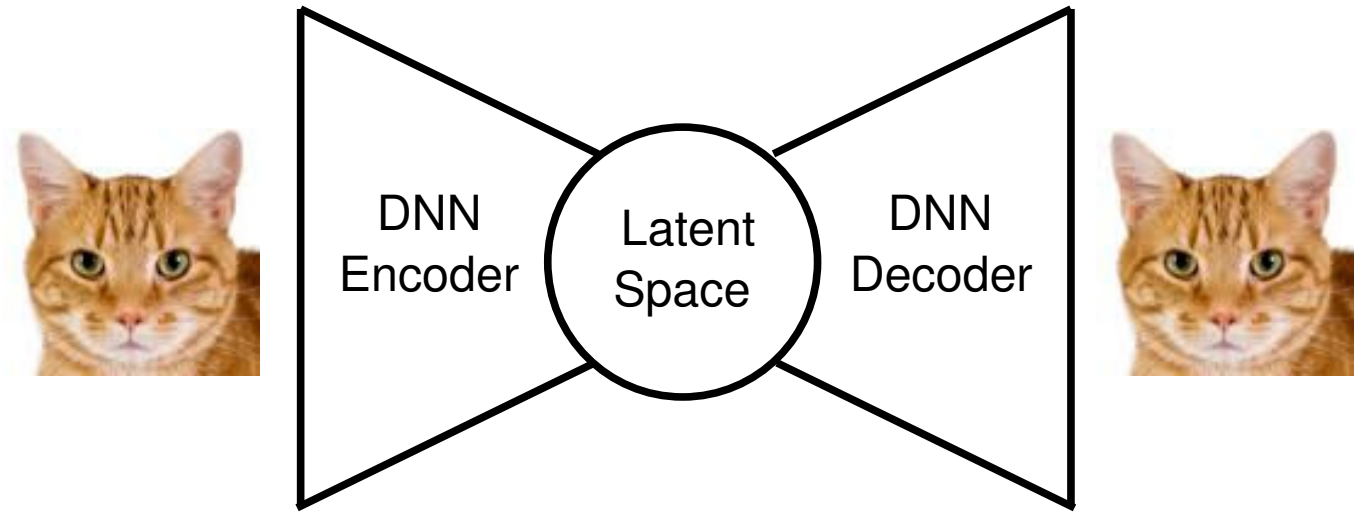
Deterministic autoencoder

An autoencoder is a net trained to reconstruct input data out of a learned internal representation (e.g. minimizing quadratic distance)



Deterministic autoencoder

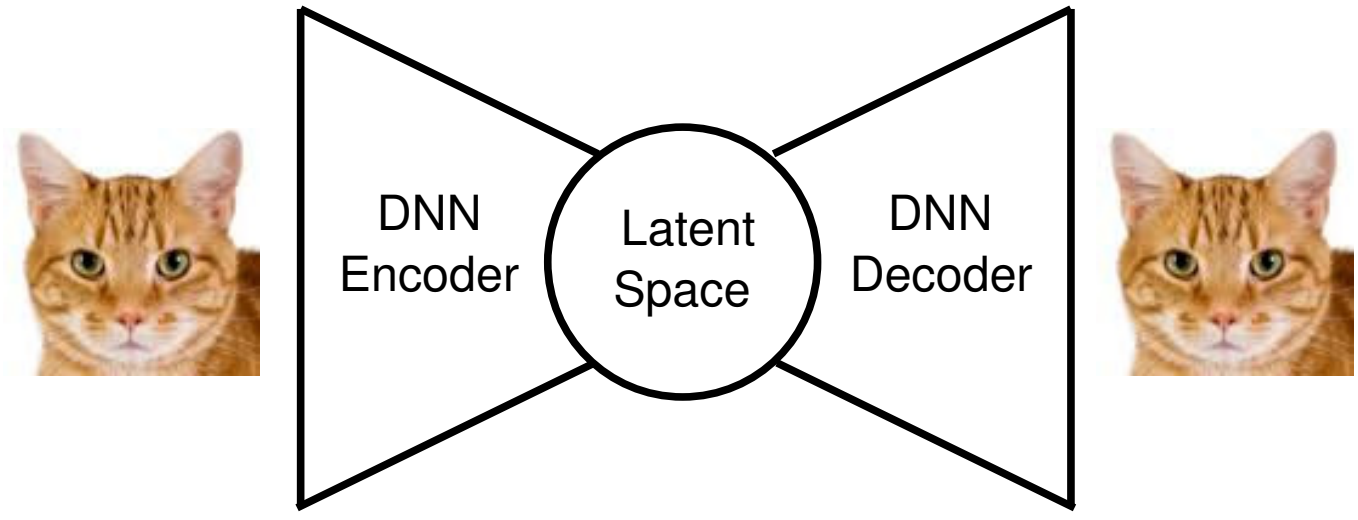
An autoencoder is a net trained to reconstruct input data out of a learned internal representation (e.g. minimizing quadratic distance)



Can we use the decoder to **generate** data by **sampling** in the latent space?

Deterministic autoencoder

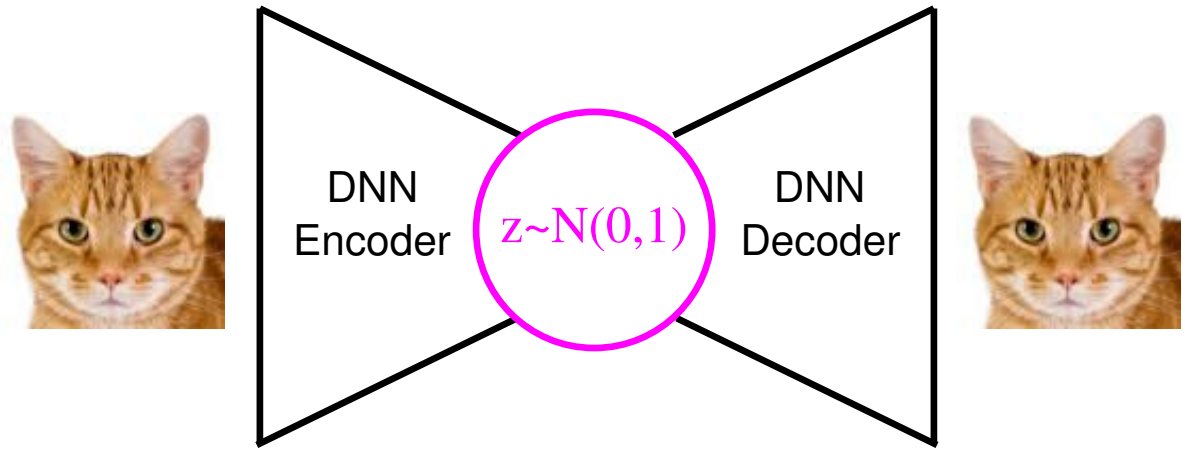
An autoencoder is a net trained to reconstruct input data out of a learned internal representation (e.g. minimizing quadratic distance)



Can we use the decoder to **generate** data by **sampling** in the latent space?
No, since we do not know the distribution of latent variables.

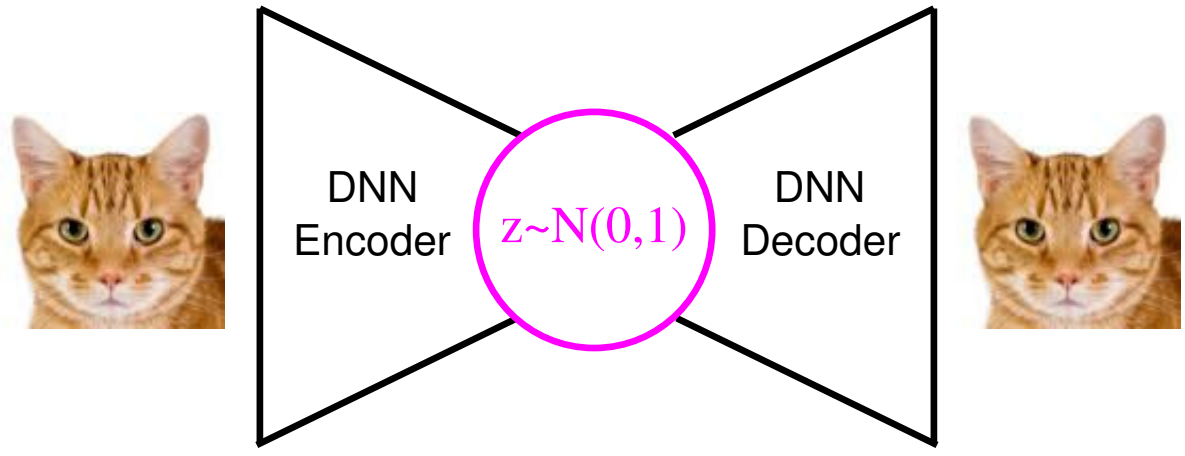
Variational autoencoder

In a Variational Autoencoder (VAE) we try **to force** latent variables to have a known distribution (e.g. a Normal distribution)



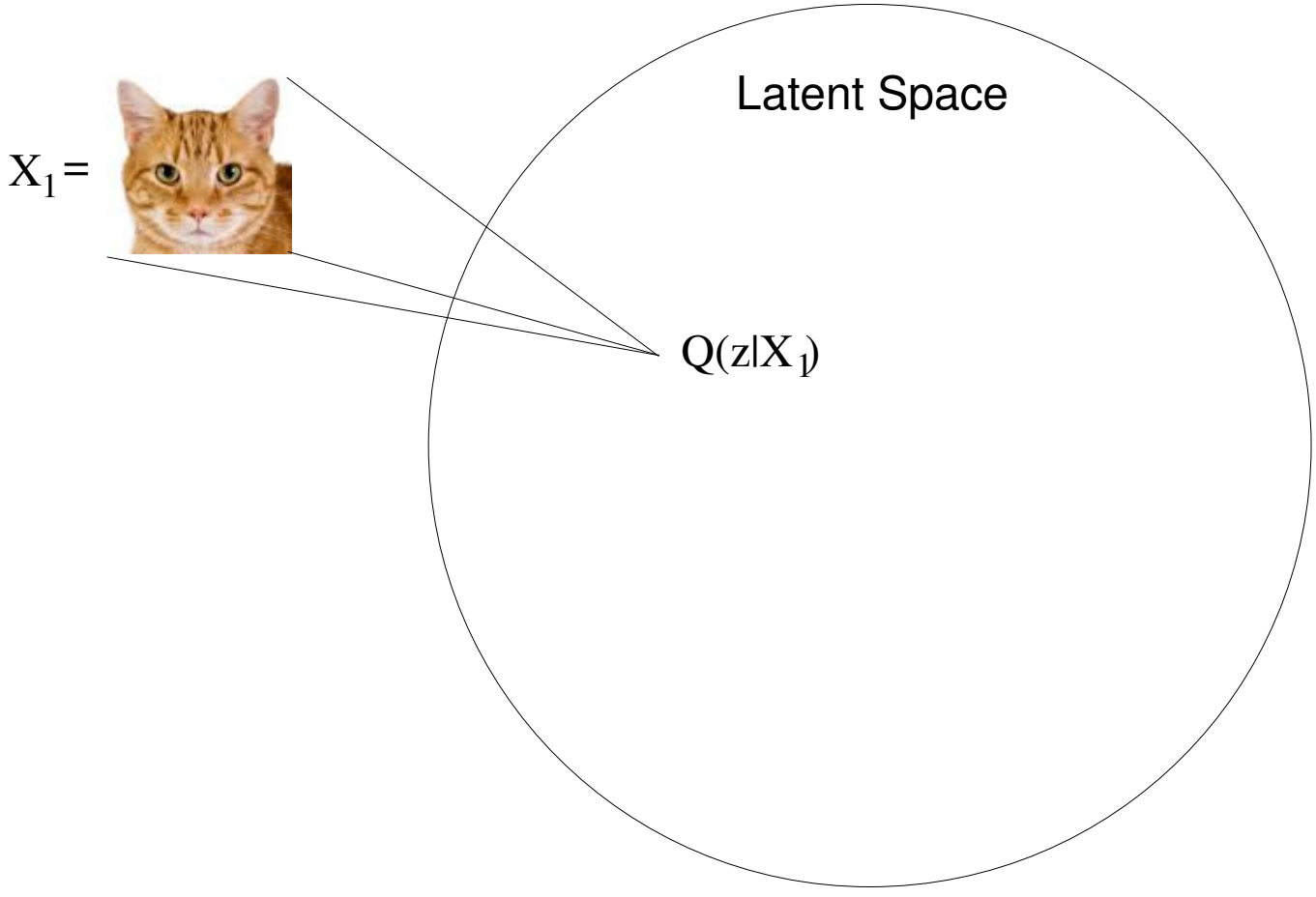
Variational autoencoder

In a Variational Autoencoder (VAE) we try **to force** latent variables to have a known distribution (e.g. a Normal distribution)

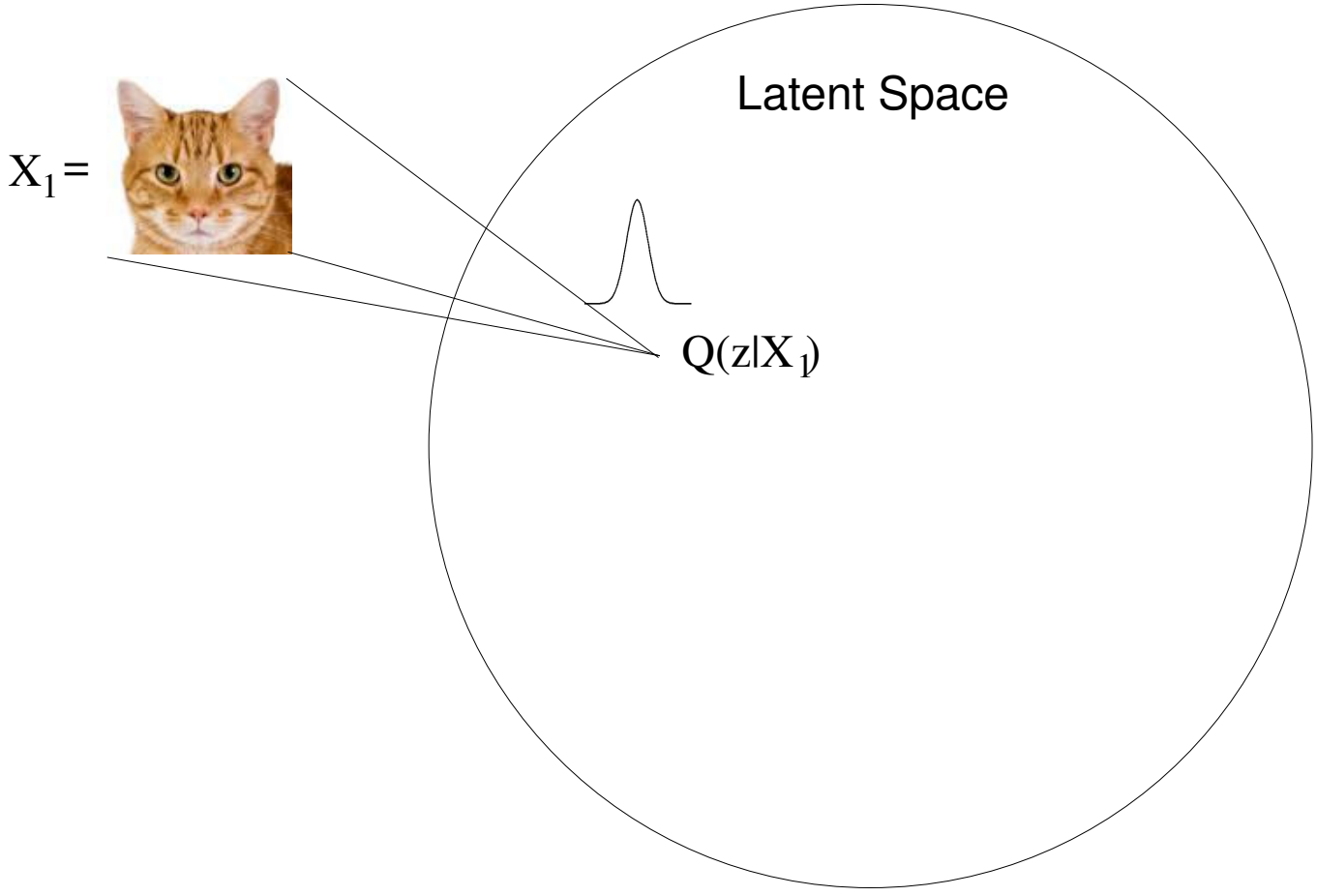


How can we do it?

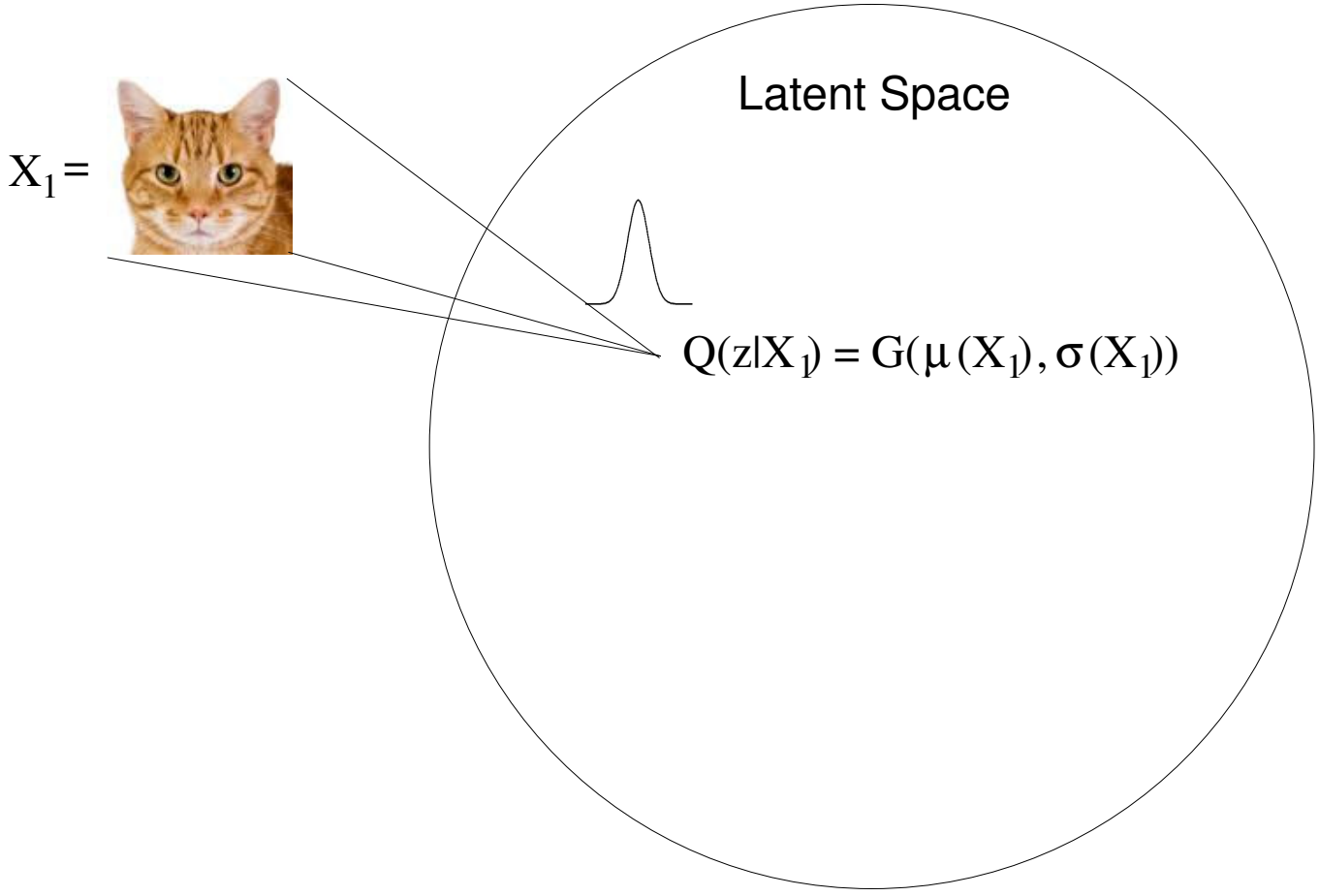
The encoding distribution $Q(z|X)$



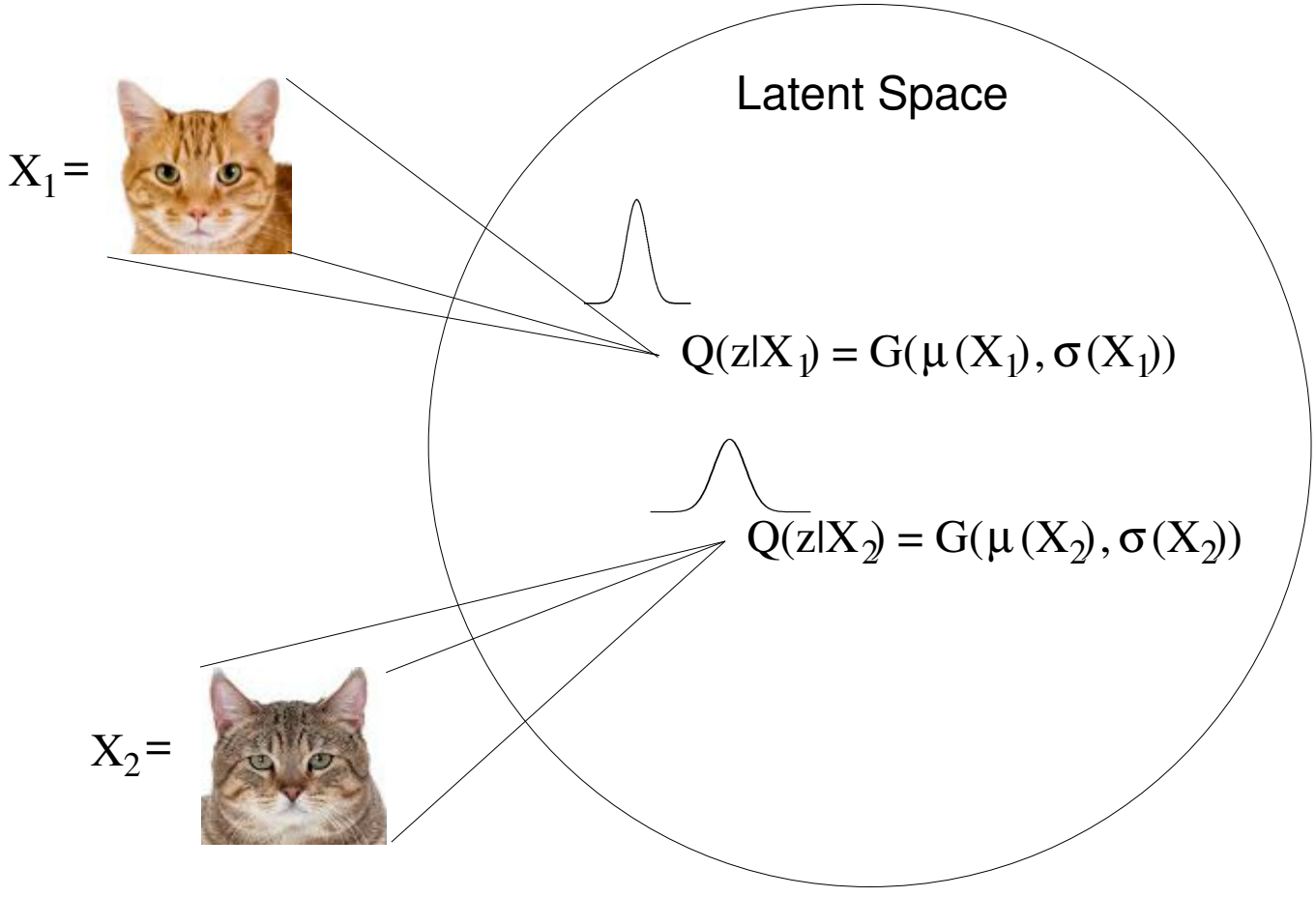
Estimate relevant statistics for $Q(z|X)$



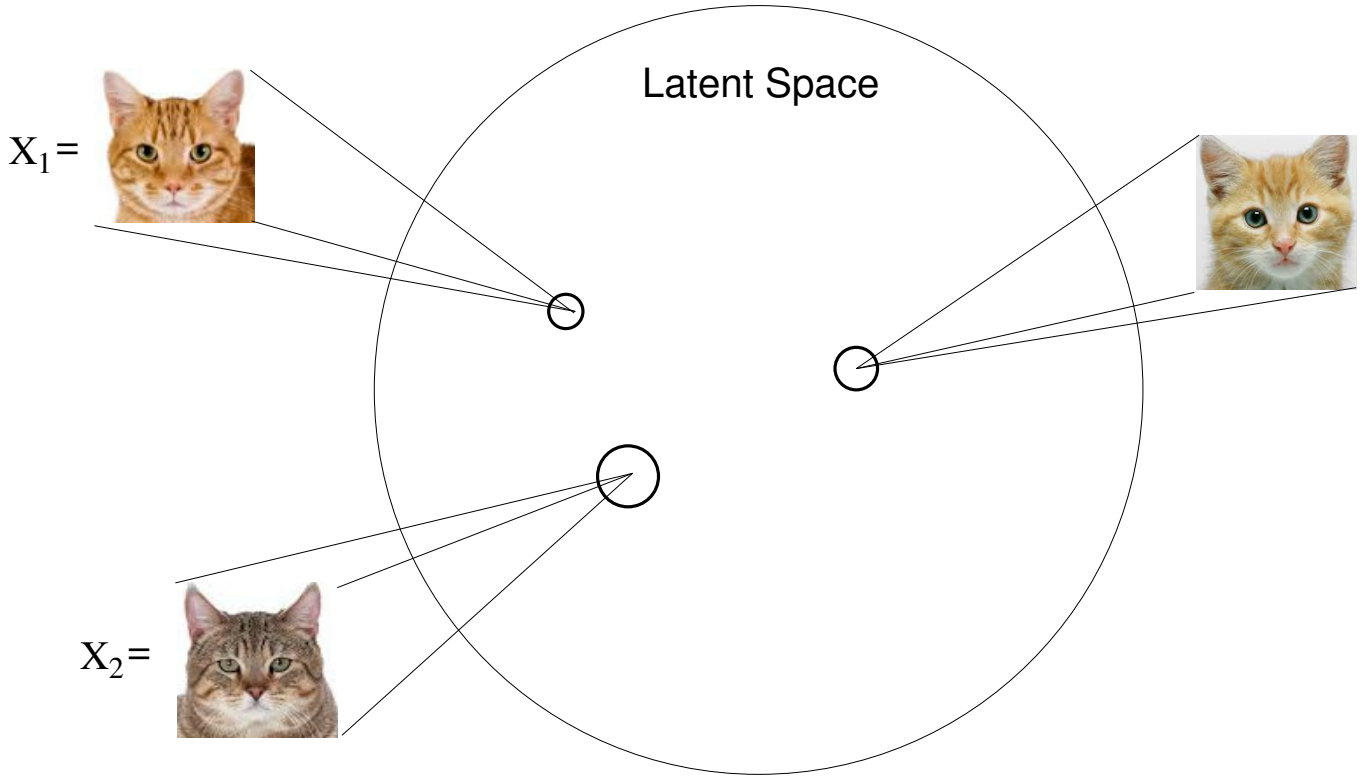
Estimate relevant statistics for $Q(z|X)$



Estimate relevant statistics for $Q(z|X)$

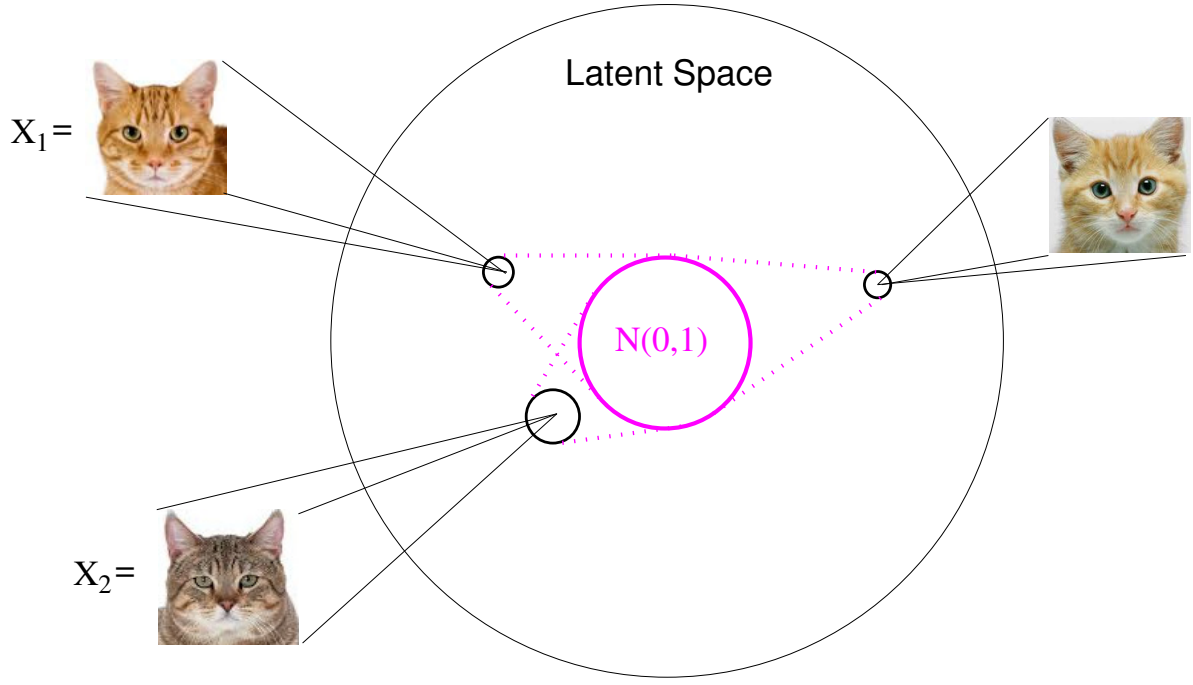


Estimate relevant statistics for $Q(z|X)$



We estimate the variance $\sigma(X)$ around $\mu(X)$ by **gaussian sampling** at **training time**.

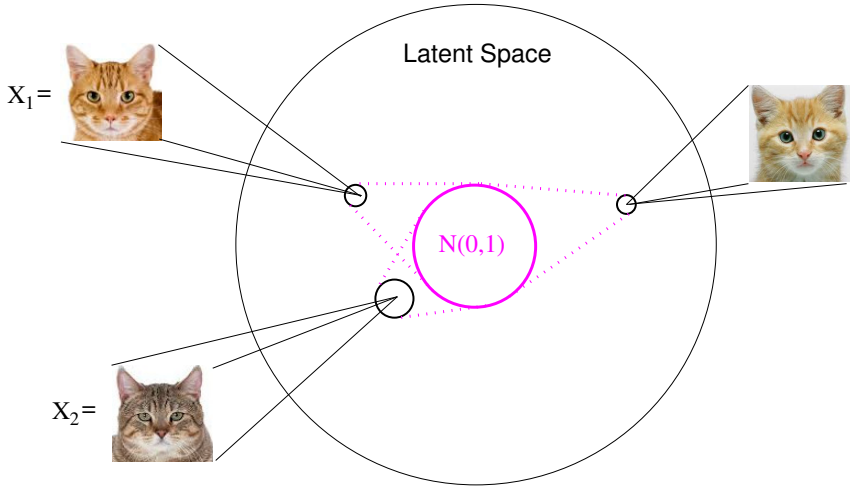
Kullback-Leibler regularization



minimize the Kullback-Leibler distance between **each** $Q(z|X)$ and a normal distribution:

$$KL(Q(z|X)||N(0,1))$$

The marginal posterior



The actual distribution of latent variables is the marginal distribution $Q(z)$, hopefully resembling the prior $P(z) = N(0, 1)$

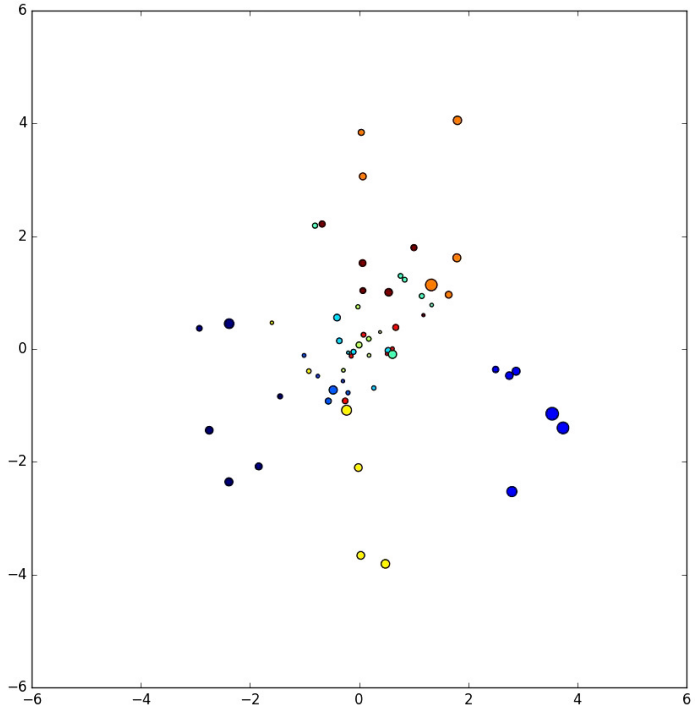
$$Q(z) = \sum_X Q(z|X) \approx N(0, 1)$$

In fact, averaging on all data ...

$$
\begin{aligned}
&\mathbb{E}_{X\sim P_{data}}\, KL(Q(z|X)||P(z)) \\
&= -\,\mathbb{E}_{X\sim P_{data}}\, \mathcal{H}(Q(z|X)) + \mathbb{E}_{X\sim P_{data}}\, \mathcal{H}(Q(z|X), P(z)) \\
&= -\,\mathbb{E}_{X\sim P_{data}}\, \mathcal{H}(Q(z|X)) + \mathbb{E}_{X\sim P_{data}}\, \mathbb{E}_{z\sim Q(z|X)}\, logP(z) \\
&= -\,\mathbb{E}_{X\sim P_{data}}\, \mathcal{H}(Q(z|X)) + \mathbb{E}_{z\sim Q(z)}\, logP(z) \\
&= -\underbrace{\mathbb{E}_{X\sim P_{data}}\, \mathcal{H}(Q(z|X))}_{\text{Entropy of } Q(z|X)} + \underbrace{\mathcal{H}(Q(z), P(z))}_{\text{Cross-entropy of } Q(X) \text{ vs } P(z)}
\end{aligned}
$$

MNIST case

Disposition in the latent space of 100 MNIST digits after 10 epochs of training



Sparsity

What happens if we augment the number of dimensions?

Sparsity

What happens if we augment the number of dimensions?

Typically, the representation gets **sparse**:
only a limited number of variables is actually used.

Sparsity

What happens if we augment the number of dimensions?

Typically, the representation gets **sparse**:
only a limited number of variables is actually used.

WHY?

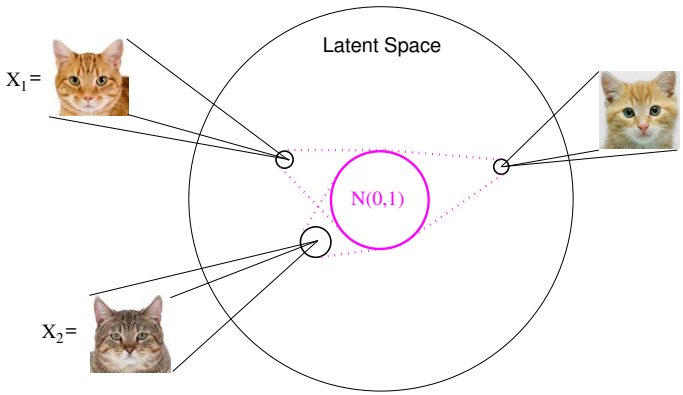
Sparsity

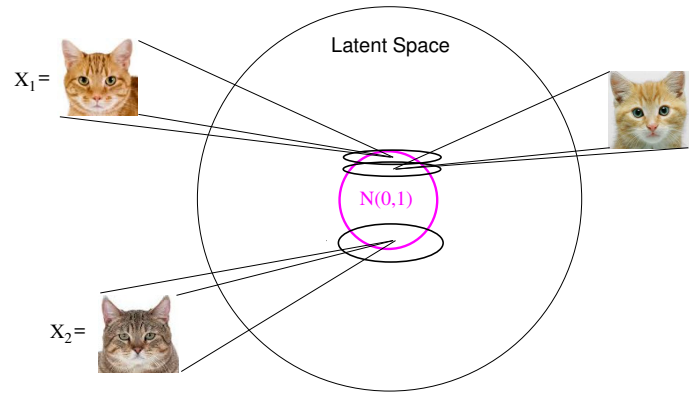
What happens if we augment the number of dimensions?

Typically, the representation gets **sparse**:
only a limited number of variables is actually used.

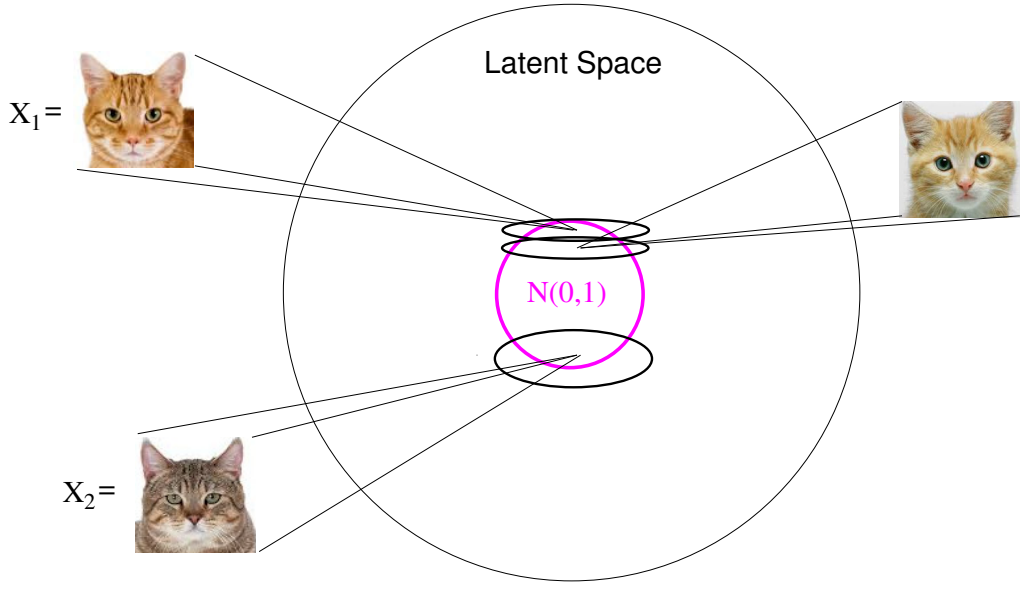
WHY?

KL pushes $Q(z|X) \rightsquigarrow N(0,1)$ ... along some axes it may prevail!





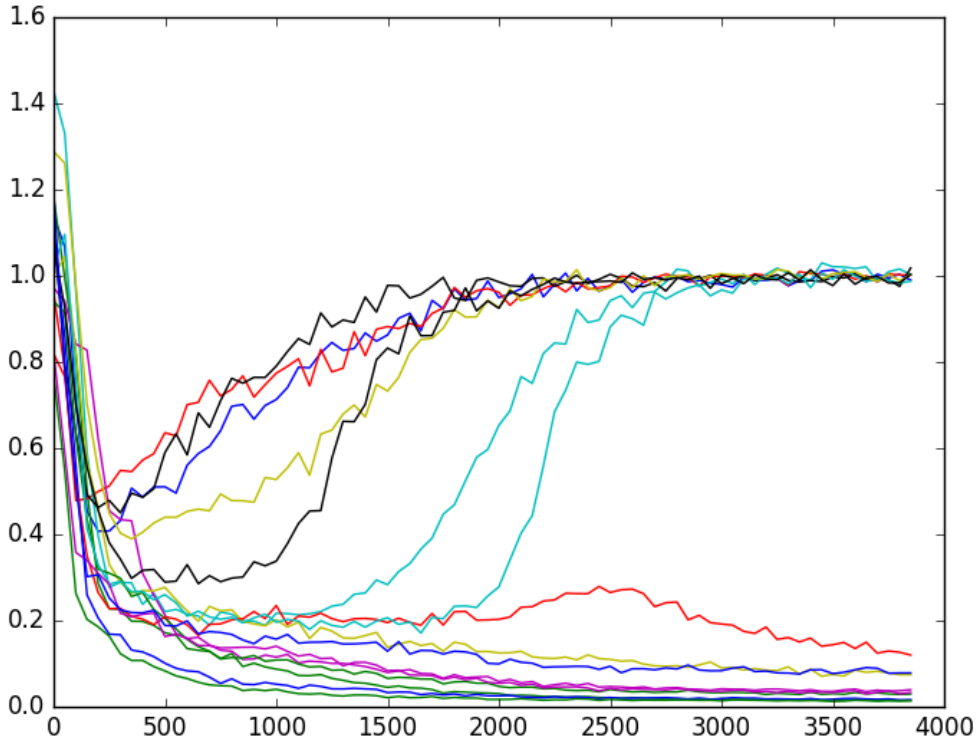
Inactive variables



Inactive latent variables are characterized by:

- a variance close to 0: the mean value $\mu(X)$ is almost always 0
- an average variance $\sigma^2(X)$ close to 1.

Mnist case



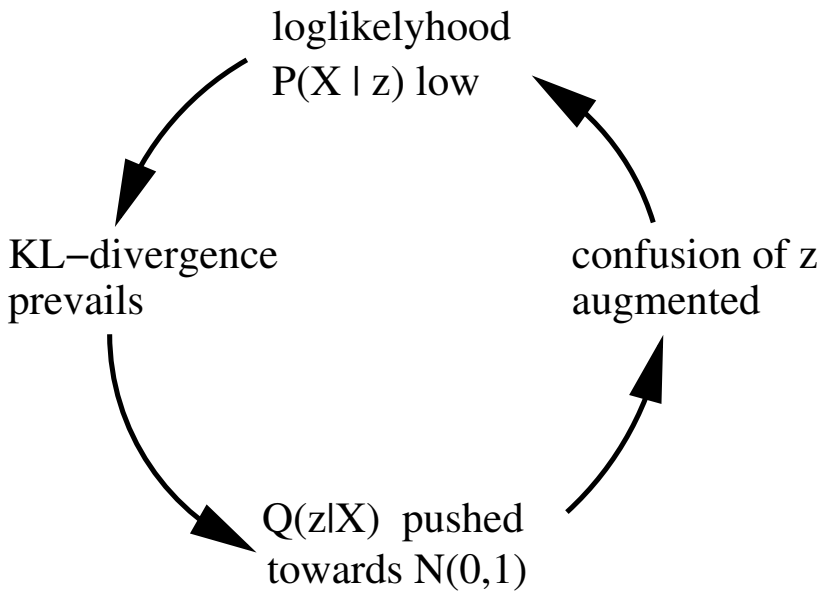
Evolution of the variance along training for a latent space composed by 16 variables.

On the x-axis we have numbers of minibatches, each one of size 128.

A vicious cycle

The loss function (ELBO):

$$\mathbb{E}_{z\sim Q(z|X)} \underbrace{logP(X|z)}_{loglikelyhood} - \underbrace{KL(Q(z|X)||P(z))}_{KL-regularizer}$$



Conclusions

An interesting a controversial aspects of VAE:

- A problem (overpruning): under-use of the model capacity
- A feature: self-regularization, robustness, disentanglement, ...

Discussion in the article!