About generative aspects of Variational Autoencoders

LOD'19
The Fifth International Conference on Machine Learning, Optimization, and Data Science
September 10-13, 2019

Certosa di Pontignano, Siena, Tuscany, Italy

Andrea Asperti

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

Generative Models

Generative models are meant to learn rich data distributions, allowing sampling of new data.

Two main classes of generative models

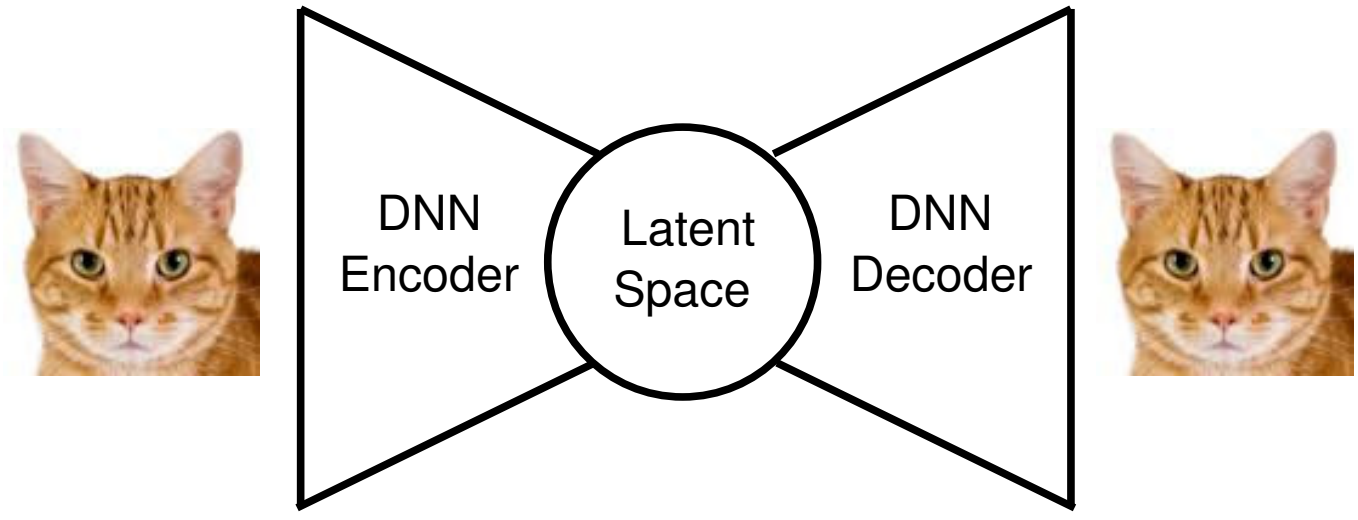
- Generative Adversarial Networks (GANs)
- Variational Autoencoders (VAEs)

At the current state of the art, GANs give better results.

What is the problem with VAEs?

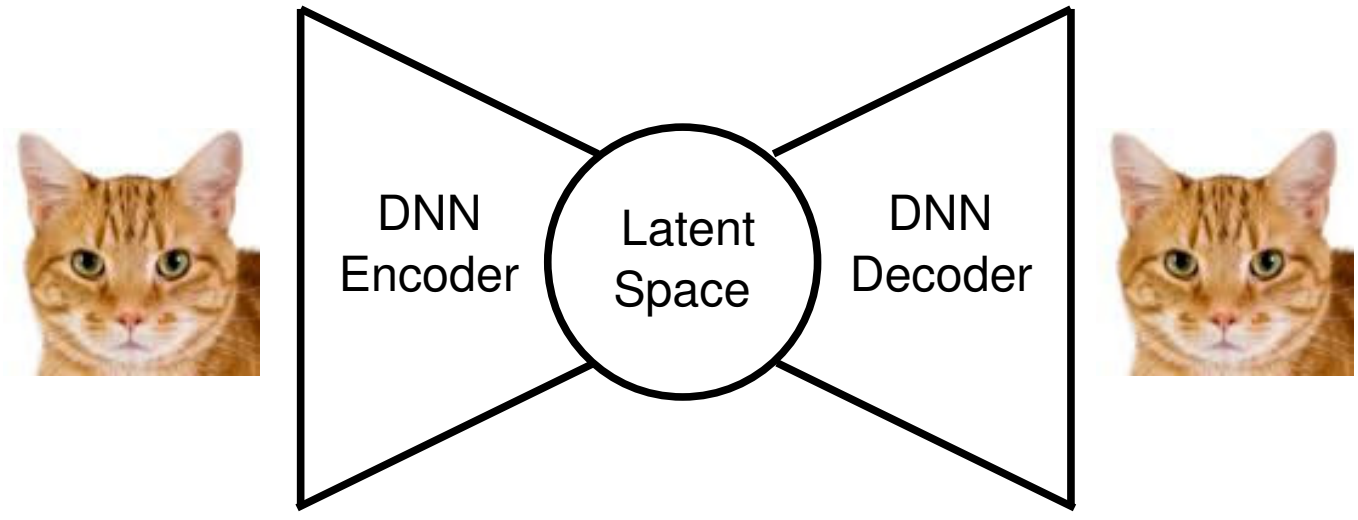
Deterministic autoencoder

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



Deterministic autoencoder

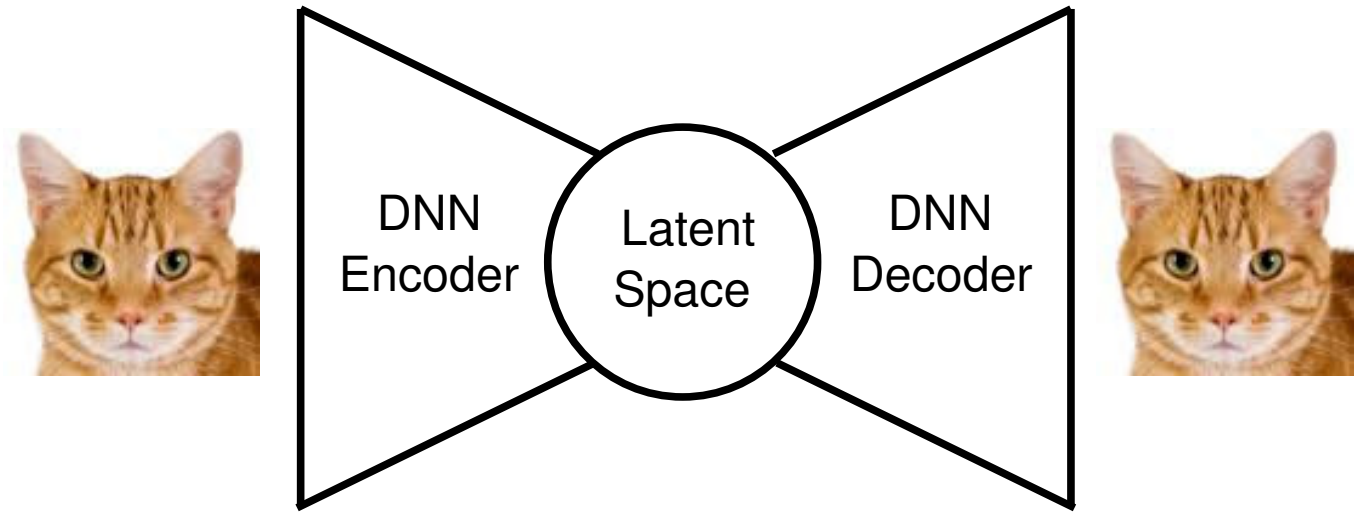
An autoencoder is a net trained to reconstruct input data out of a learned internal representation (e.g. minimizing quadratic distance)



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

Deterministic autoencoder

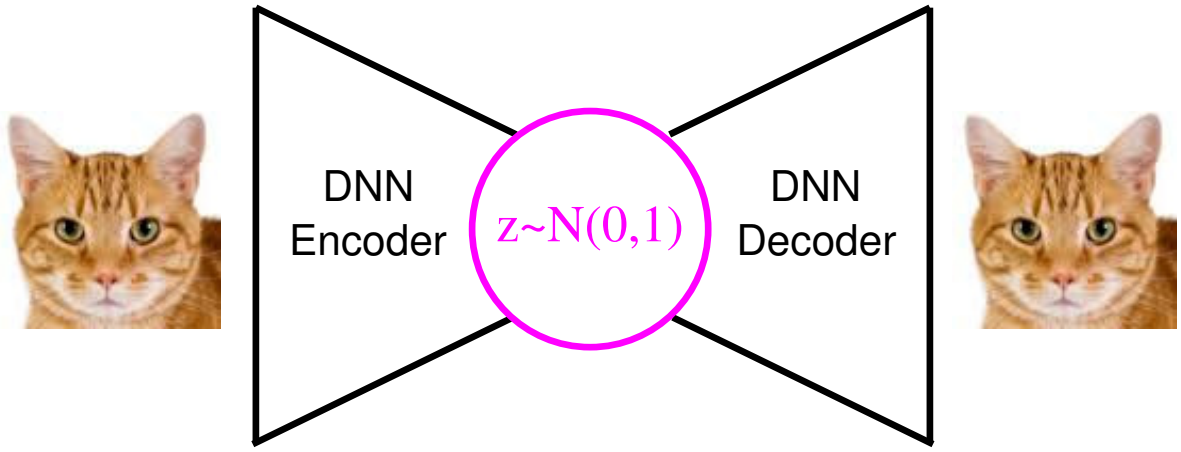
An autoencoder is a net trained to reconstruct input data out of a learned internal representation (e.g. minimizing quadratic distance)



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.

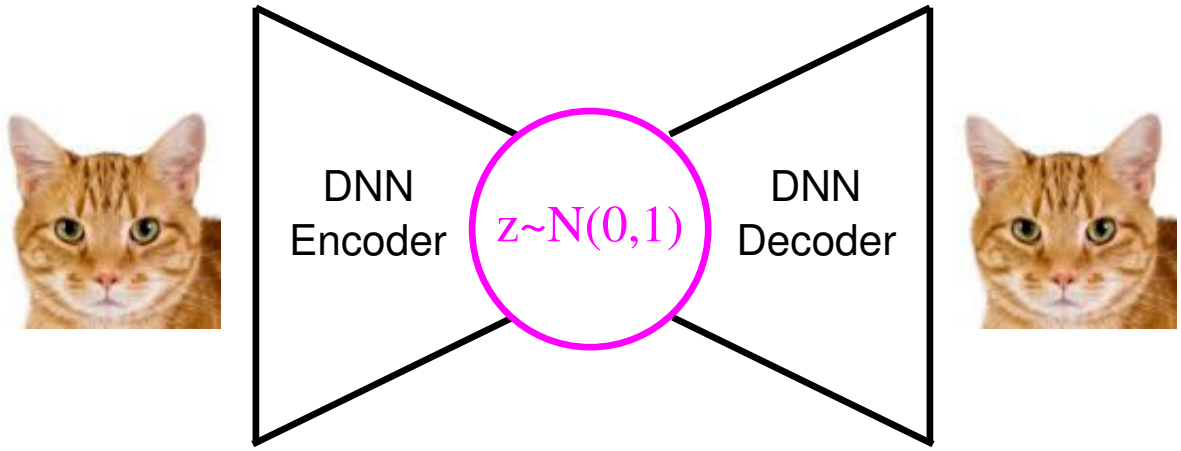
Variational autoencoder

In a Variational Autoencoder (VAE) [9, 10, 6] we try **to force** latent variables to have a known distribution (e.g. a Normal distribution)



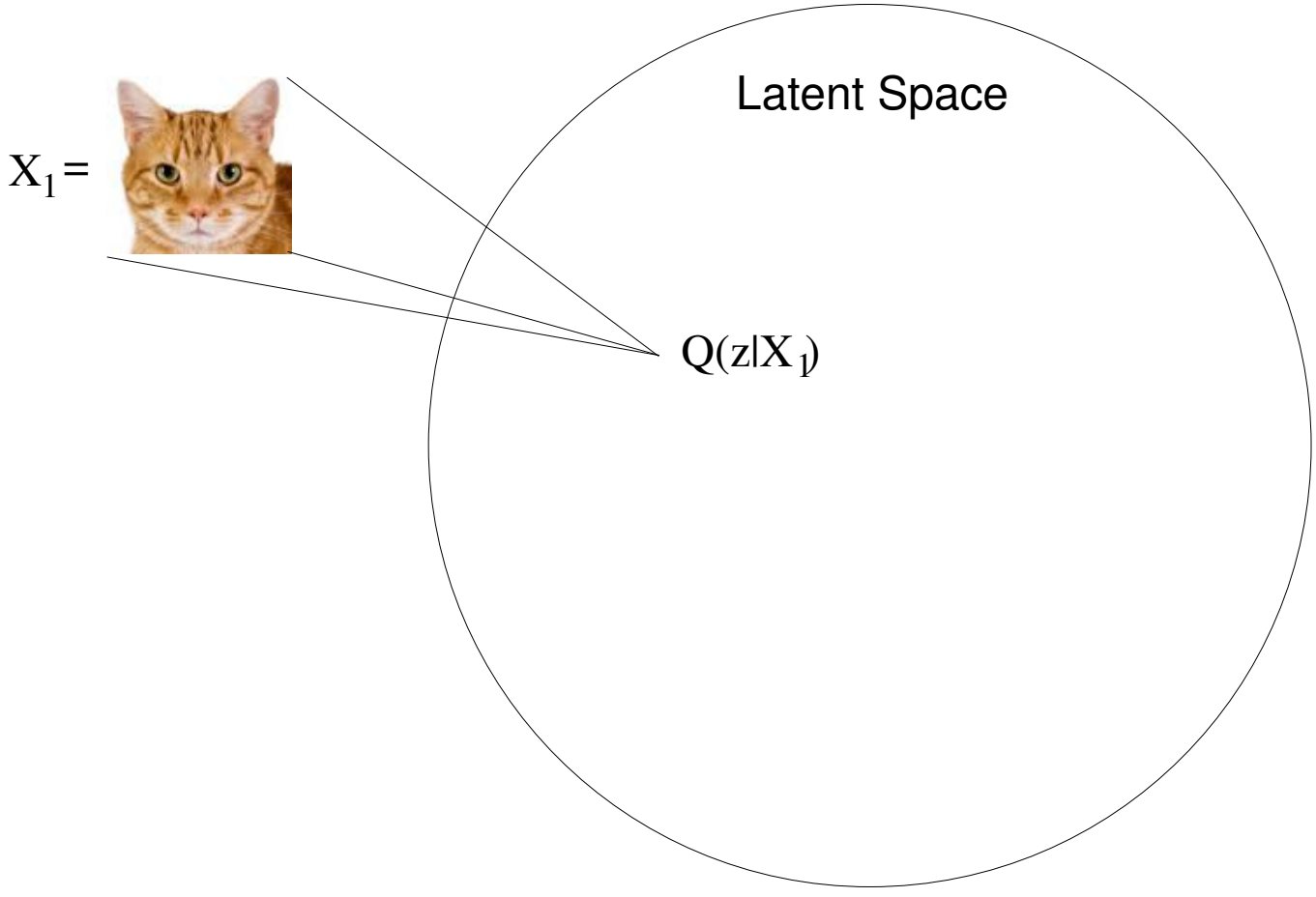
Variational autoencoder

In a Variational Autoencoder (VAE) [9, 10, 6] we try **to force** latent variables to have a known distribution (e.g. a Normal distribution)

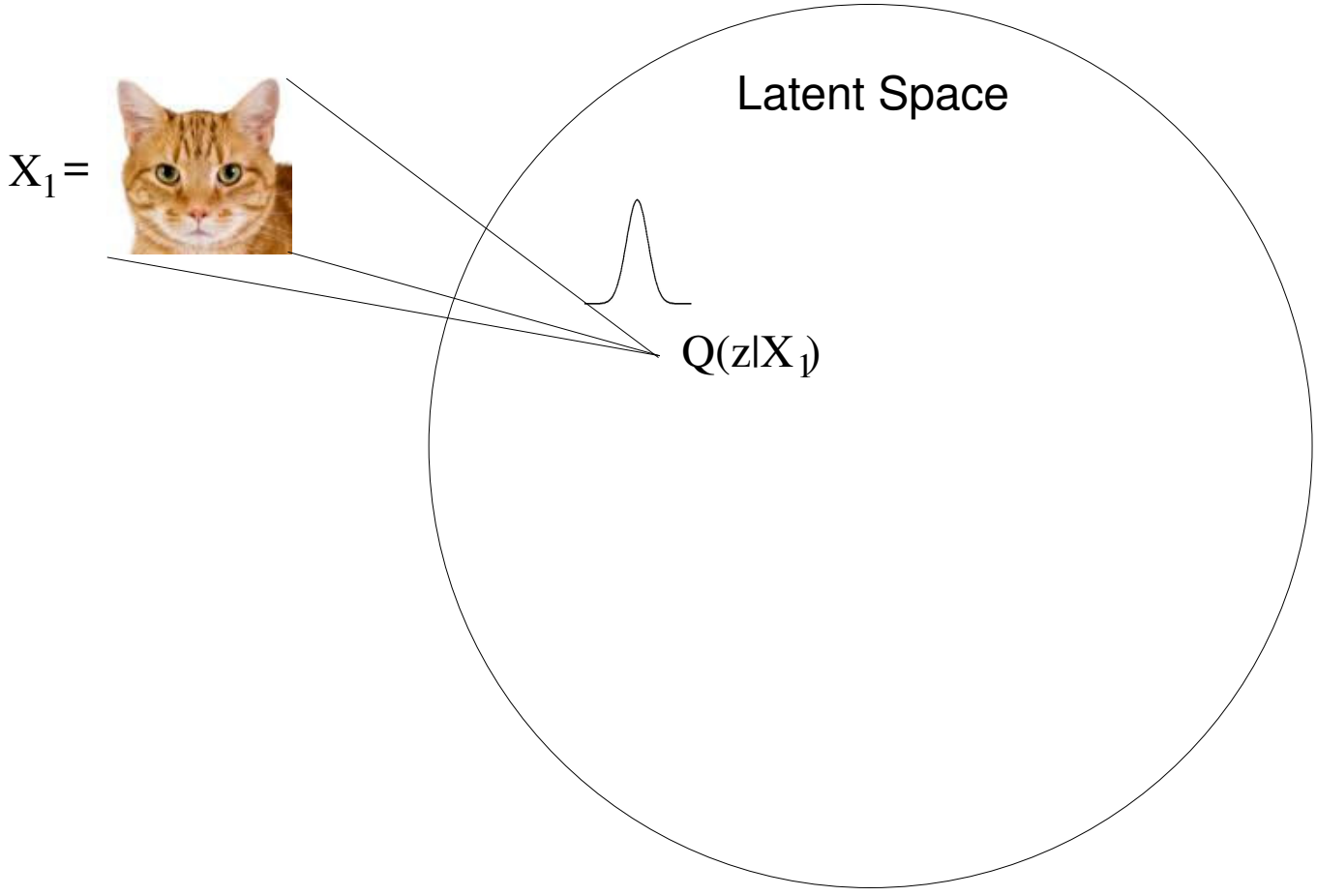


How can we do it? Is this actually working?

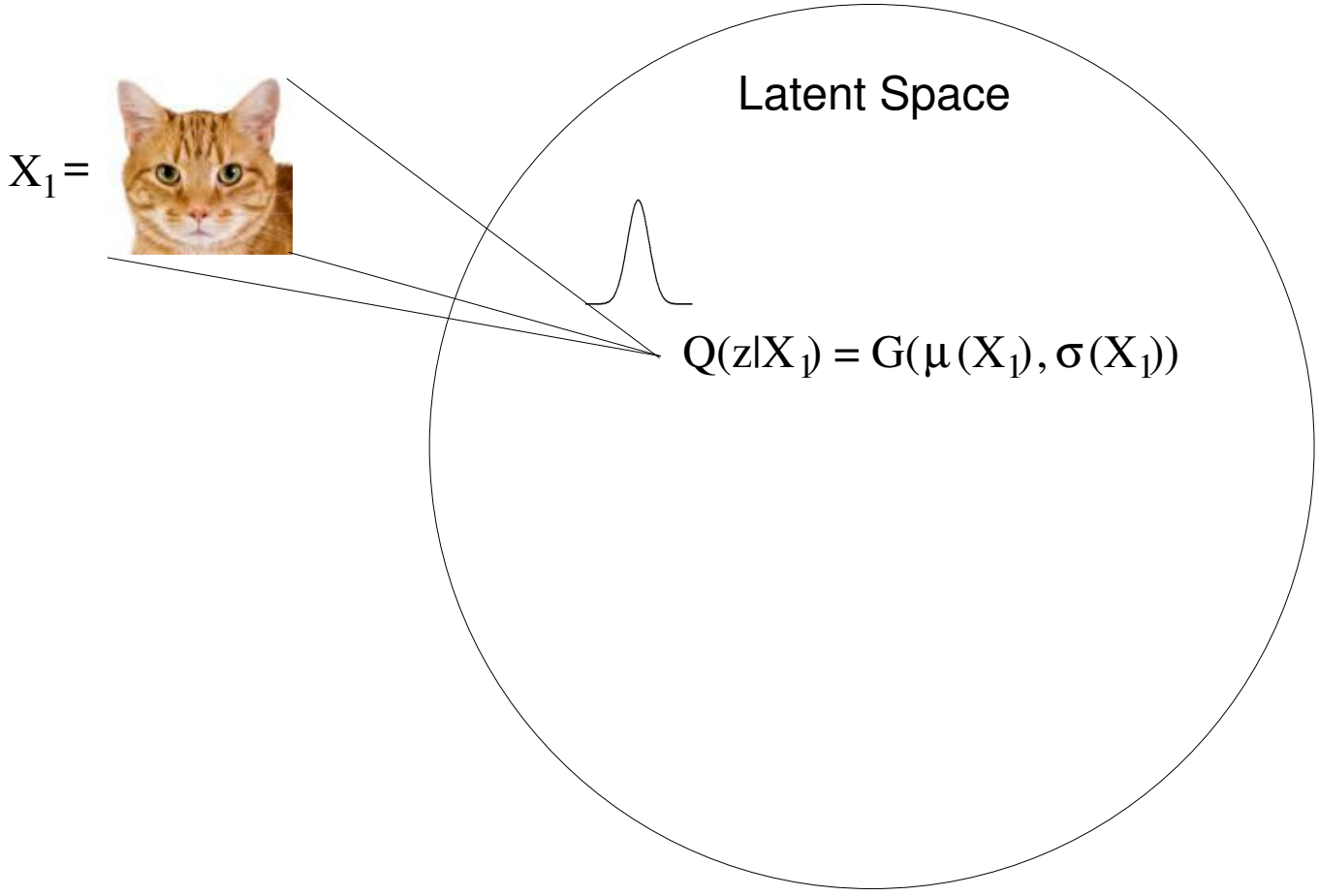
The encoding distribution $Q(z|X)$



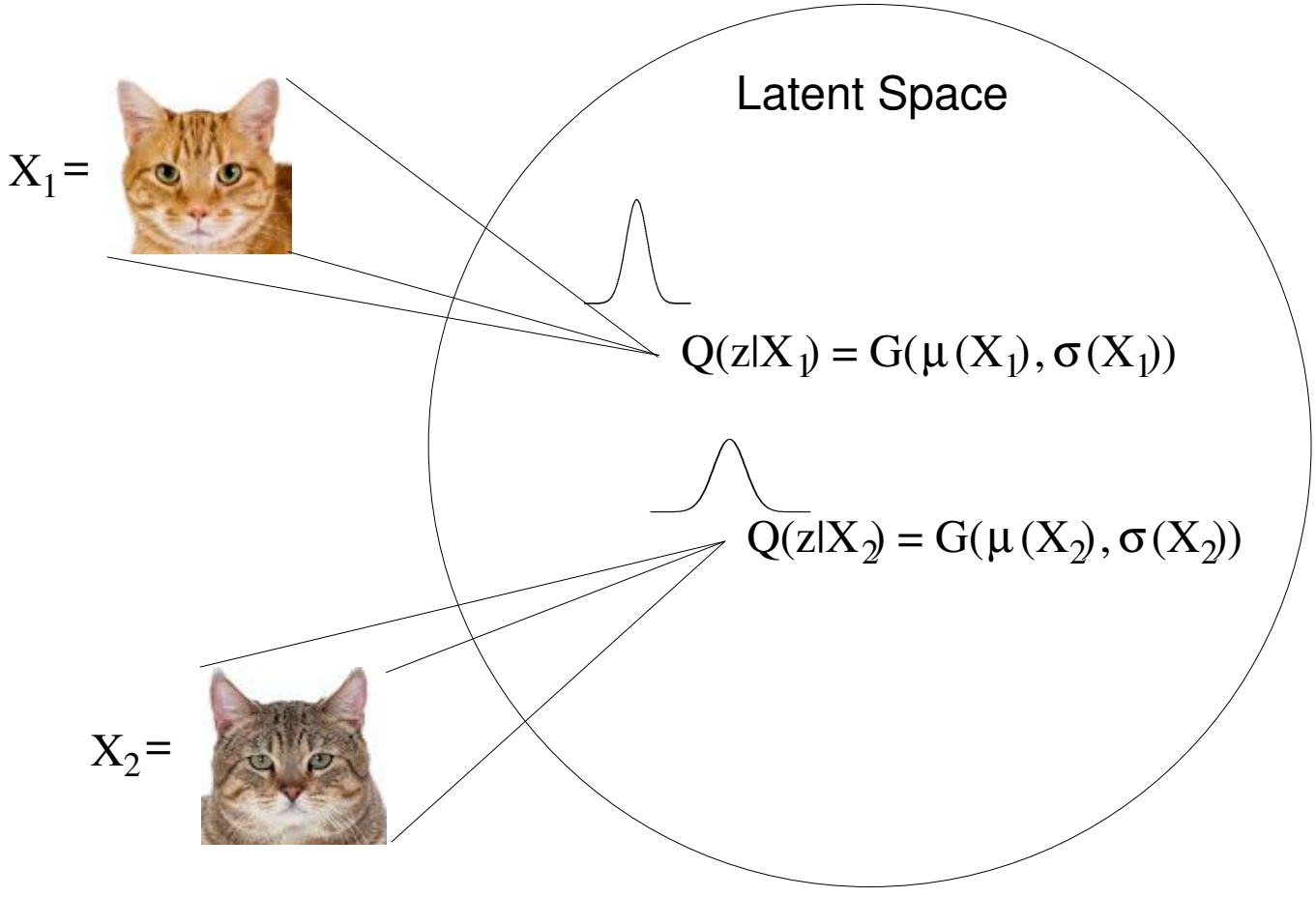
Estimate relevant statistics for $Q(z|X)$



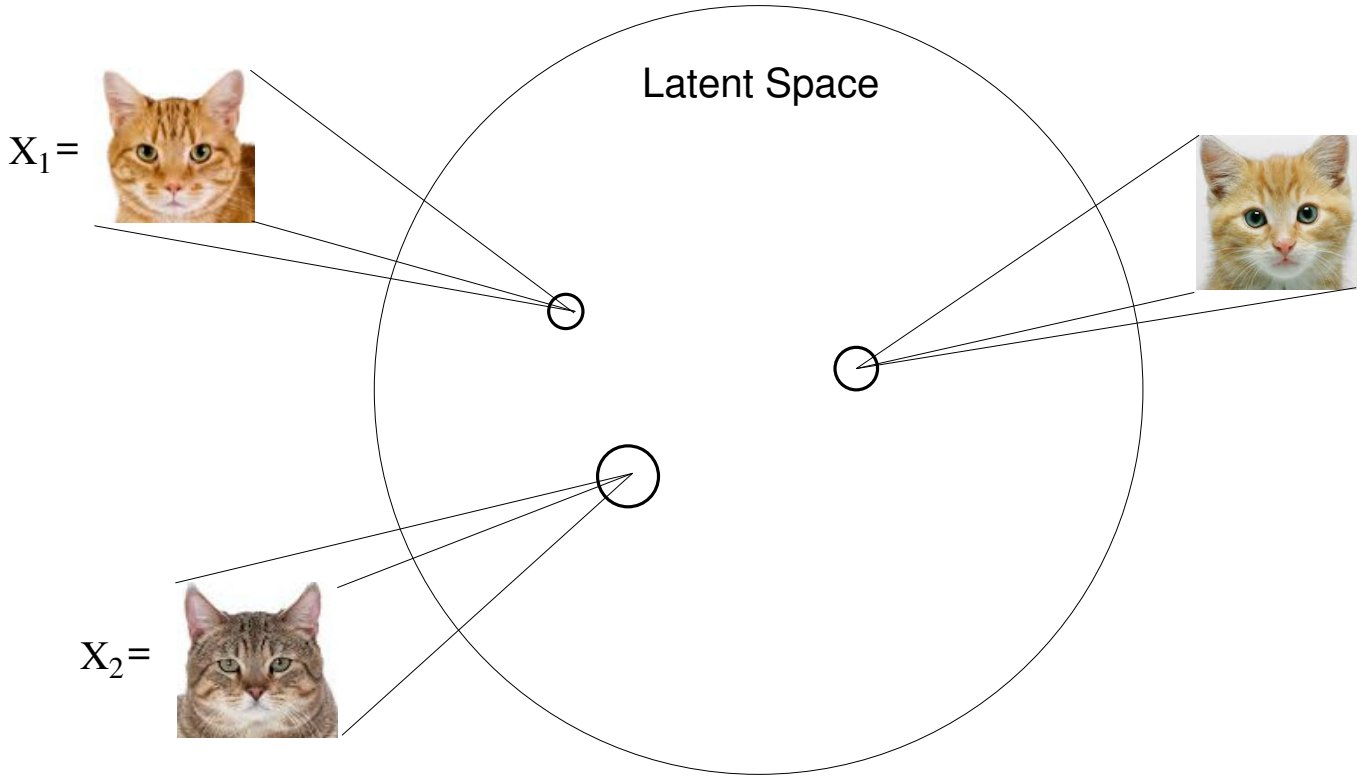
Estimate relevant statistics for $Q(z|X)$



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

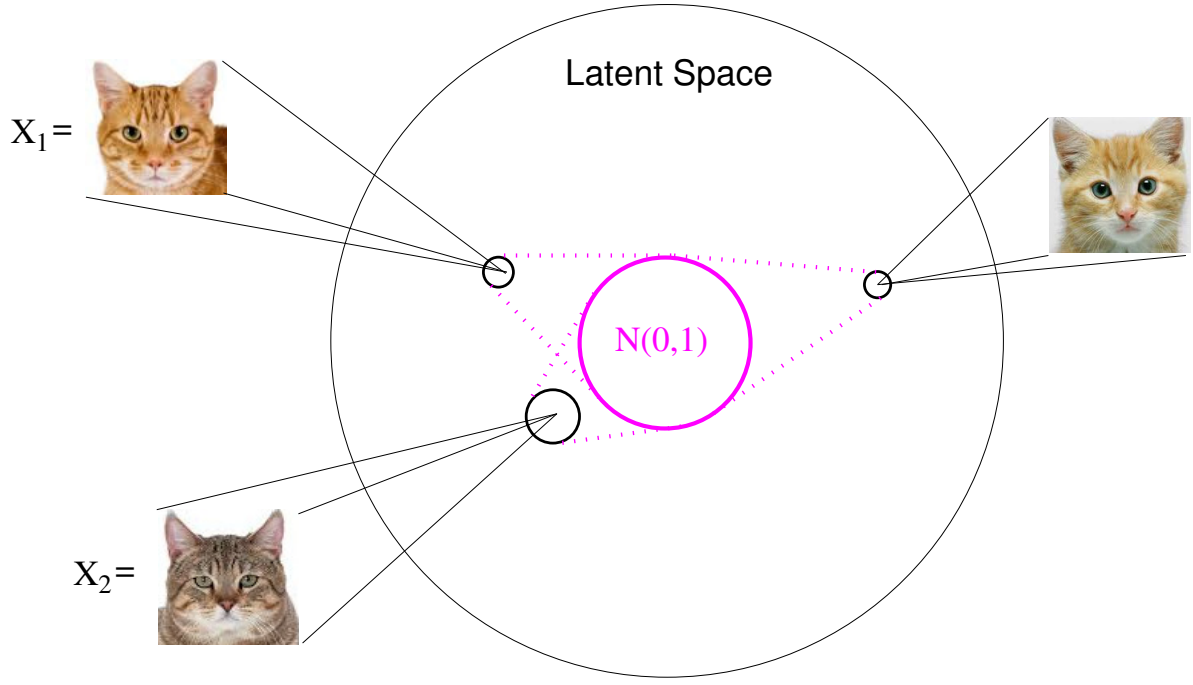


Estimate relevant statistics for $Q(z|X)$



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

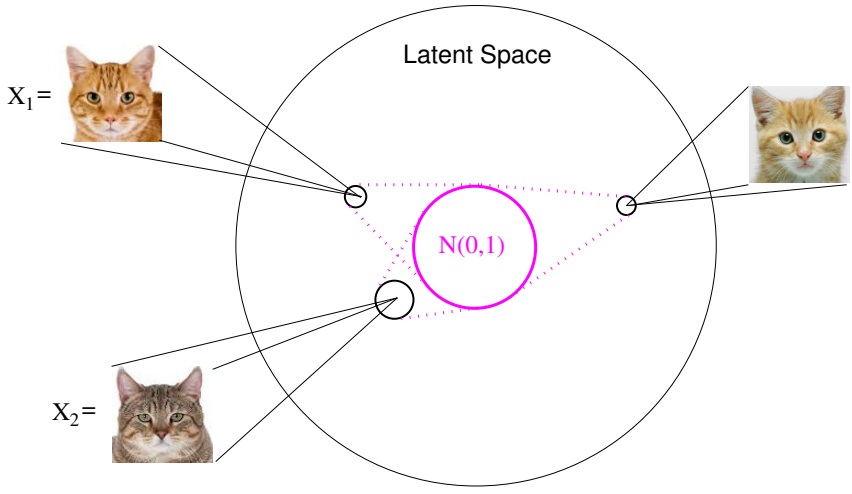
Kullback-Leibler regularization



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

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

The marginal posterior

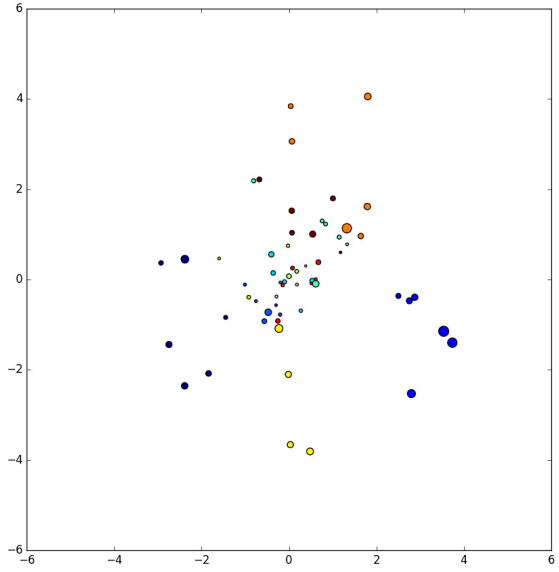


The actual distribution of latent variables is the marginal (aka cumulative) distribution $Q(z)$, hopefully resembling the prior $P(z) = N(0, 1)$

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

MNIST case

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



It does indeed have a Guassian shape... Why?

Why is KL-divergence working?

Many different answers ... relatively complex theory.

In this article, we investigate the marginal posterior distribution as a Gaussian Mixture Model (GMM) (one gaussian for each data point).

The normalization idea

- For a neural network, it is relatively easy to perform an **affine transformation** of the latent space

- The transformation **can be compensated in the next layer** of the network, keeping the loss invariant.

(same idea behind batch-normalization layers)

- This means we may assume the network is able to keep a **fixed ratio** ρ between the variance and the mean value of each latent variable.

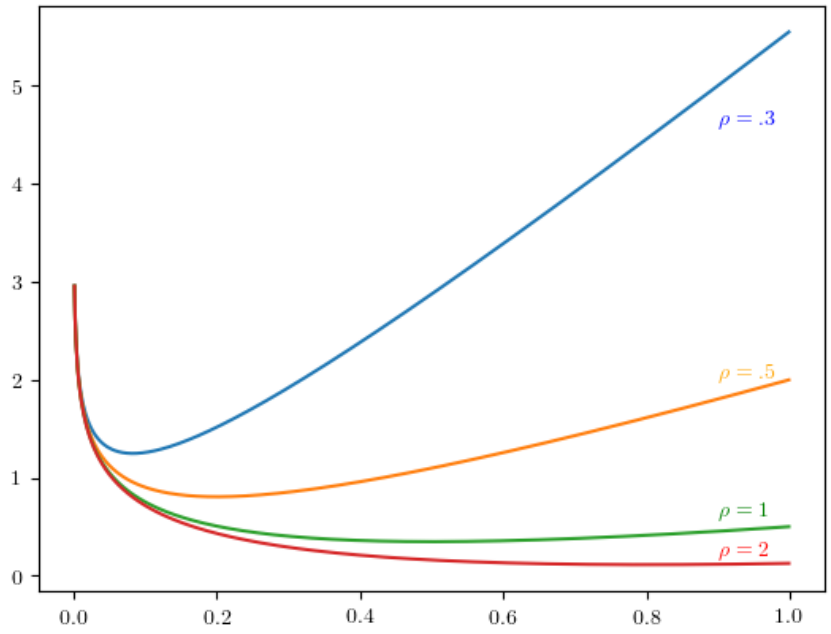
Pushing ρ in KL-divergence

Pushing ρ in the closed form of the KL-divergence, we get the expression

$$\frac{1}{2}(\sigma^2(X)\frac{1+\rho^2}{\rho^2} - log(\sigma^2(X)) - 1)$$

which has a minimum when

$$\sigma^2(X) + \mu^2(X) = 1$$



Corollaries

- The variance law:
  averaging on all X, we expect that for each latent variable z

$$\widehat{\sigma_z^2(X)} + \sigma_z^2 = 1$$

  (supposing $\widehat{\mu_z(X)} = 0$)
- By effect of the KL divergence the two first moments of the distribution of each latent variable should agree with those of a Normal $N(0, 1)$ distribution
- **What about the other moments?** Hard to guess.

Conclusion

For several years the cause of the mediocre performance of VAEs has been imputed to the so called overpruning phenomenon [2, 11, 12].

Recent research suggests the problem is due to the difformity between the latent distribution and the normal prior [4, 5, 1, 7].

Our contribution: we may reasonably expect the KL-divergence will force the two first moments of the distribution to agree with those of a Normal distribution, but we may hardly presume the same for the other moments.

Essential bibliography (1)

Andrea Asperti.

Variational Autoencoders and the Variable Collapse Phenomenon
Sensors & Transducers V.234, N.3, pages 1-8, 2018.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov.

Importance weighted autoencoders.
CoRR, abs/1509.00519, 2015.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner.

Understanding disentangling in β-vae.
2018.

Bin Dai, Yu Wang, John Aston, Gang Hua, and David P. Wipf.

Connections with robust PCA and the role of emergent sparsity in variational autoencoder models.
Journal of Machine Learning Research, 19, 2018.

Bin Dai and David P. Wipf.

Diagnosing and enhancing vae models.
In *Seventh International Conference on Learning Representations (ICLR 2019), May 6-9, New Orleans*, 2019.

Carl Doersch.

Tutorial on variational autoencoders.
CoRR, abs/1606.05908, 2016.

Essential bibliography (2)

Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael J. Black, Bernhard Schölkopf
From Variational to Deterministic Autoencoders
CoRR, abs/1903.12436.

Diederik P. Kingma, Tim Salimans, and Max Welling.
Improving variational inference with inverse autoregressive flow.
CoRR, *abs/1606.04934, 2016.*

Diederik P. Kingma and Max Welling.
Auto-encoding variational bayes.
CoRR, *abs/1312.6114, 2013.*

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra.
Stochastic backpropagation and approximate inference in deep generative models.
In Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, *volume 32 of* JMLR Workshop and Conference Proceedings, *pages 1278–1286. JMLR.org, 2014.*

Serena Yeung, Anitha Kannan, and Yann Dauphin.
Epitomic variational autoencoder.
2017.

Serena Yeung, Anitha Kannan, Yann Dauphin, and Li Fei-Fei.
Tackling over-pruning in variational autoencoders.
CoRR, *abs/1706.03643, 2017.*