Variational Autoencoders

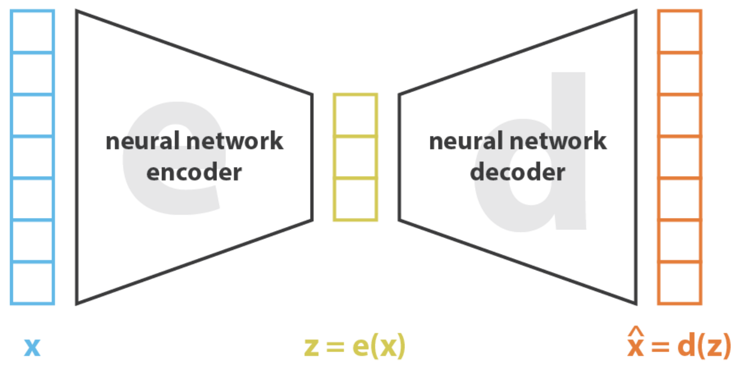
Classic Autoencoders (AE)

Encoder:
- From input to bottleneck layer
- Dimensionality is reduced

Decoder:
- From bottleneck layer to output
- Dimensionality is increased

Training:
- **Unsupervised**:
  - Input is unlabelled data
  - Loss is a **reconstruction loss** between input and output
- Regularization might be used to promote sparse encodings

Applications:
- Dimensionality reduction
- Compression (not very effective)
- Denoising
- Anomaly detection



$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

Architecture and loss of a classic AE [2]

Variational Autoencoders (VAE)

Same general structure, but:

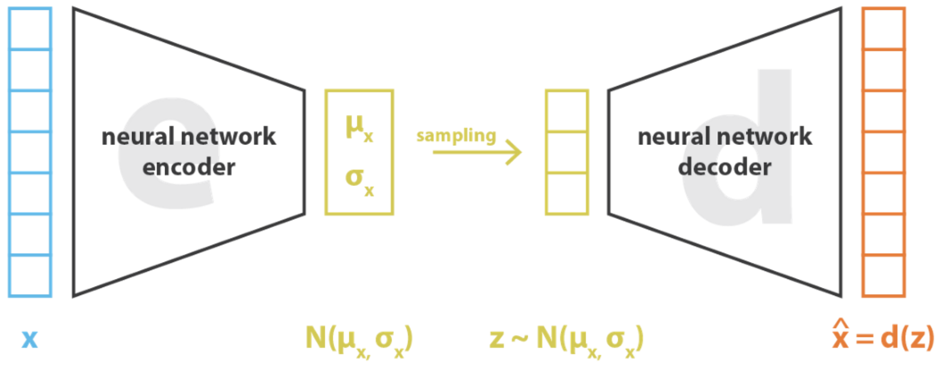
- **Encoded layer is not deterministic**
  - It's a gaussian distribution $N(\mu_x, \sigma_x)$
  - The encoder samples from N: $z \sim N(\mu_x, \sigma_x)$

Training:

- Loss is the same as classic AE, plus a regularization term, called **Kullback-Leibler divergence (KL)** [5]
  - KL penalizes a distribution the further it is from a normal distribution $N(0, 1)$
  - Without it the network would collapse to sparse punctual distributions in the latent space

Properties of the bottleneck layer:

- Provides a distribution instead of a deterministic value
- Due to KL loss the encoded space tends to be:
  - **Continuous**
  - **Complete**



$$\text{loss} = \| x - \hat{x} \|^2 + \text{KL}[\, N(\mu_x, \sigma_x), N(0, I) \,] = \| x - d(z) \|^2 + \text{KL}[\, N(\mu_x, \sigma_x), N(0, I) \,]$$

Architecture and loss of a VAE [2]

KL and regularity of the latent space

KL encourages the latent variables to behave like a normal distribution in order to obtain a more regular latent space.

Continuity

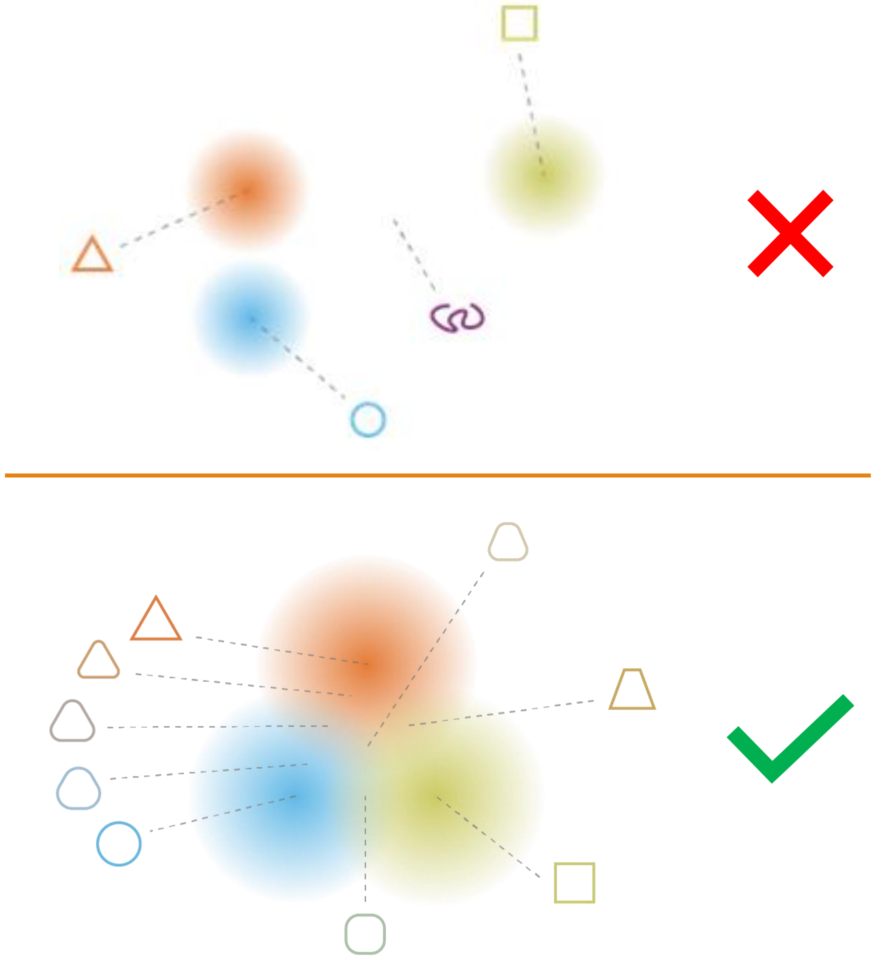
- Encoded variables that are close to each other in the latent space map to outputs that are close in the output space.
- *Counterexample*: the triangle and the circle in the upper figure should not be close.

Completeness

- Any point from the latent space is mapped to a meaningful output.
- *Counterexample*: the point between the main shapes should not map to a squiggly line in the upper figure.

Continuity + completeness

- Overall a **smoother gradient**
- Sampling from the latent space produces meaningful and coherent outputs



Irregular and regular latent spaces with their output mappings [2]

Reparametrisation

The **sampling** operation between the encoder and the decoder represents a **problem for backpropagation**.
We cannot perform partial derivatives over a stochastic operation.

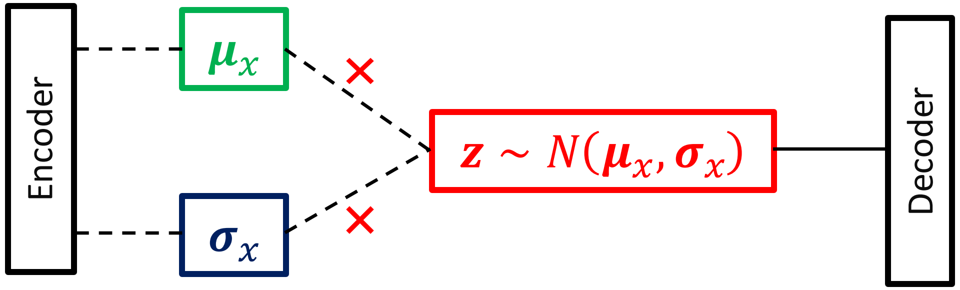
Reparametrisation

Previously: $\boldsymbol{z} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x)$

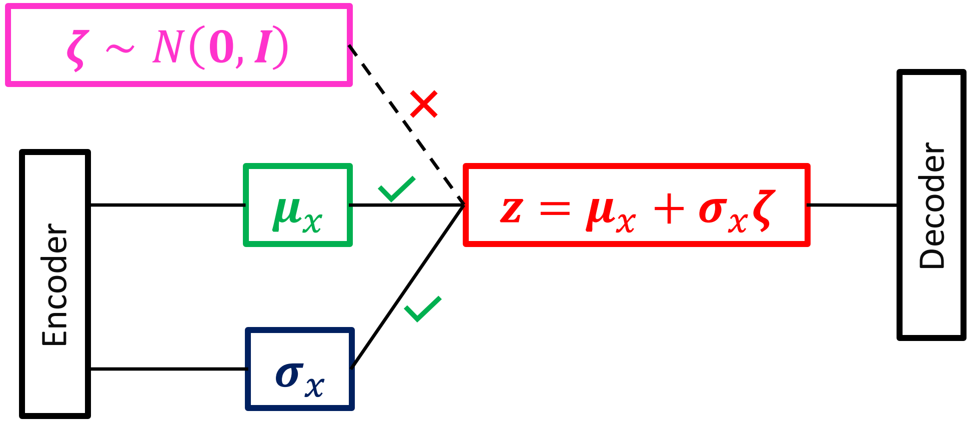
Now: $\boldsymbol{z} = \boldsymbol{\mu}_x + \boldsymbol{\sigma}_x \boldsymbol{\zeta}$ with $\boldsymbol{\zeta} \sim N(\boldsymbol{0}, \boldsymbol{I})$

This ensures that backpropagation can flow uninterrupted from the decoder to the encoder.

Note: It's still not possible to perform backpropagation in the branch with $\boldsymbol{\zeta}$, but we do not need to do that, so it is not a problem.



Sampling prevents backpropagation [2]



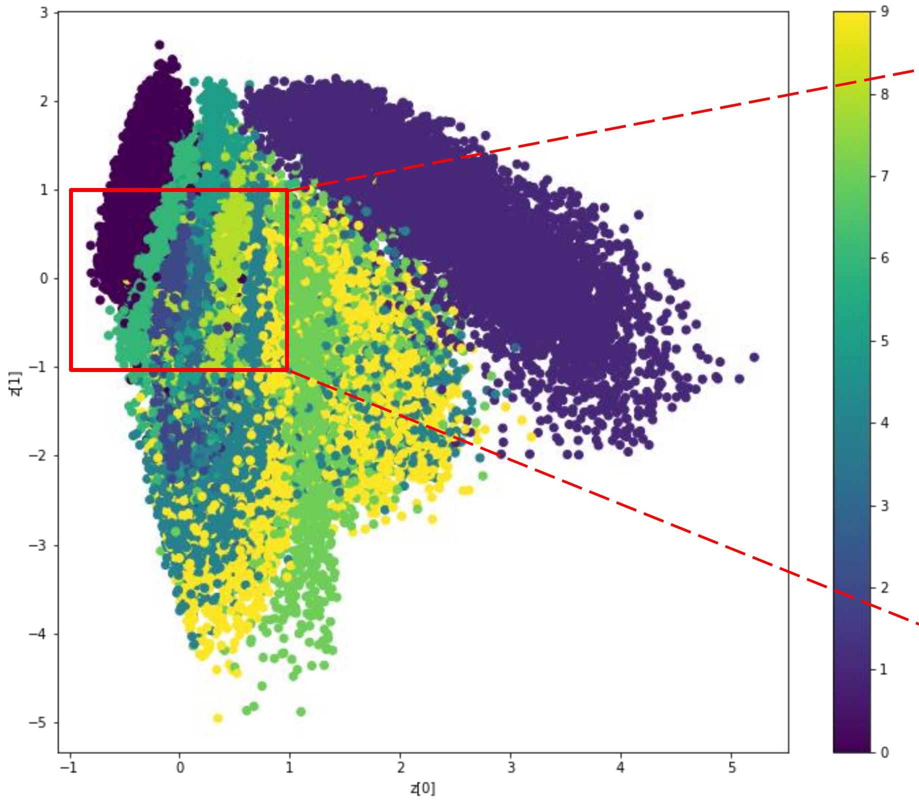
Reparametrisation allows backpropagation [2]

Latent space example

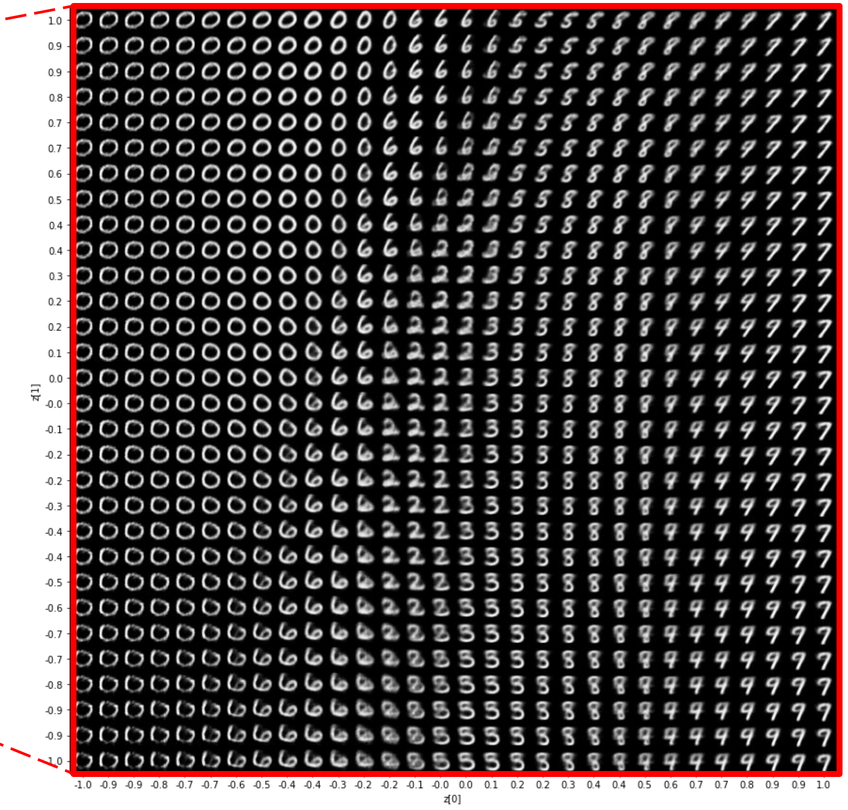
Examples of sample space of a VAE

- Trained on the MNIST handwritten digits dataset
- With a latent dimension of 2 (i.e. the latent subspace is a plane)

Notice the effect of both **continuity** and **completeness** in this image.



μ value position in the latent space, by digit [3]



A sampling from the latent space [3]

References and resources

[1] Arden Dertat, “Applied Deep Learning - Part 3: Autoencoders”, Towards Data Science, https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

[2] Joseph Rocca, “Understanding Variational Autoencoders”, Towards Data Science, https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

[3] “Convolutional Variational Autoencoder”, Keras tutorials, https://keras.io/examples/generative/vae/

[4] Alexander Amini, “MIT 6.S191: Deep Generative Modeling”, MIT 6.S191 lessons, https://www.youtube.com/watch?v=BUNl0To1IVw&t=536s

[5] “Kullback-Leibler divergence”, Wikipedia, https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

[6] Giorgio Bonvicini, "Varational Auto Encoders", https://github.com/GioBonvi/MachineLearning/tree/main/VAE/