# Autoenocders

## MCnet ML School

Najmeh Abiri (naab@itu.dk)

25 June 2020

IT University of Copenhagen

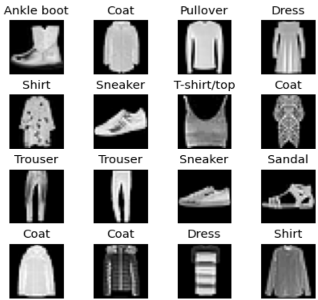
- No supervision needed (no labels). Model discover information from data.
- Why is it important?
  - Labels are expensive
  - Instead of mapping the input to labels, it draws inferences from datasets and extract features and patterns.
  - Sometimes we don't know the data.

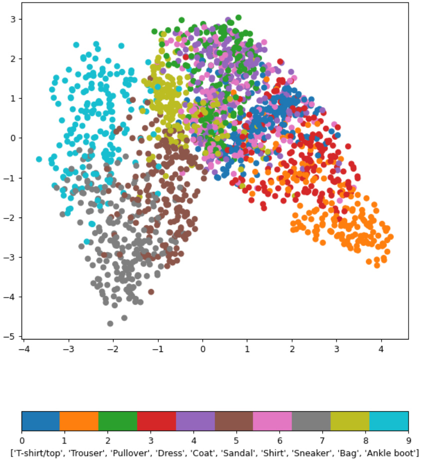
# Unsupervised learning and manifold learning

# Unsupervised learning and manifold learning

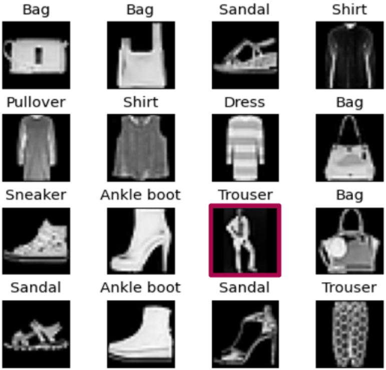
# Applications

- Clustering
- Anomaly detection
- Density estimation
- Reduce number of features



- Clustering
- Anomaly detection
- Density estimation
- Reduce number of features

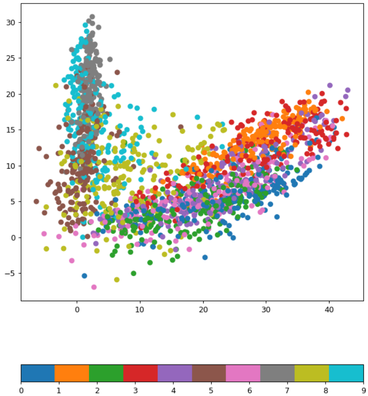


# Types

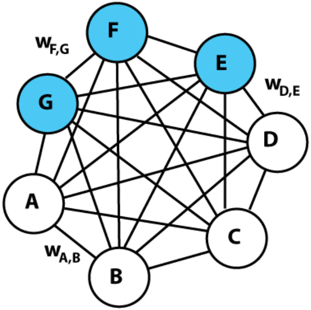
- ML unsupervised methods: PCA, k-means, mixture models, etc.

  Principle component analysis (PCA): orthogonal linear transformation of data x to latent factors z

$$z = xW$$



- NN unsupervised methods: Hopfield network, Boltzmann machines, autoencoders, and Generative adversarial networks (GAN).




[1]

[1] Boltzmann machines - Wikipedia

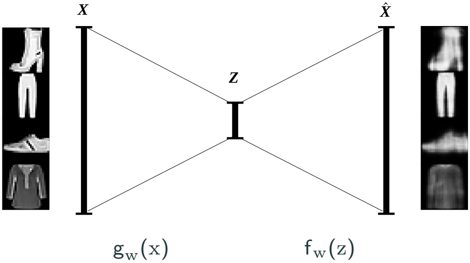
# Types

- NN unsupervised methods: Hopfield network, Boltzmann machines, autoencoders, and Generative adversarial networks (GAN).

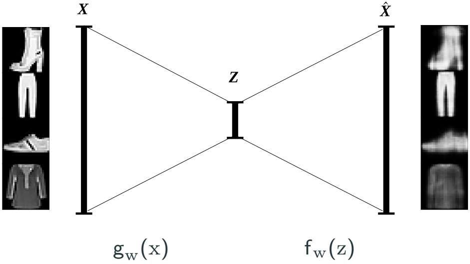
# Types

- Our focus is on two types of autoencoders: autoencoder AE and variational autoencoder VAE.

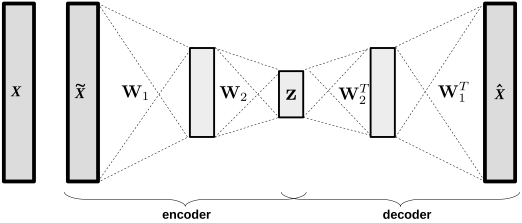
# Autoencoders (AE)



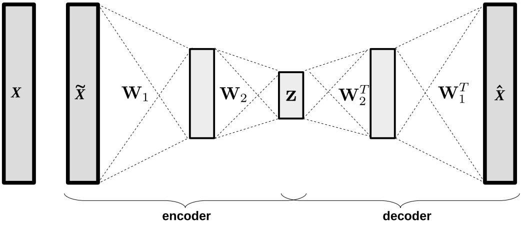

- A compression (coding) forces the network to learn meaningful latent representations z.
- Encoder (deterministic mapping) $\rightarrow z = g_w(x)$ and decoder $\rightarrow \hat{x} = f_w(z)$. Typically is an affine mapping followed by a nonlinearity.
- Decoding reconstructs the input. Network learns by minimizing reconstruction loss.




- Data denoising
- Dimensionality reduction
- Outlier detection
- Imputation




- History of autoencoder in Deep Learning book [Goodfellow et al., 2016]
- Why lower latent dimension? When Z is of the same dimensionality as X (or larger), a simple AE can achieve perfect reconstruction simply by learning an identity mapping Z=X.

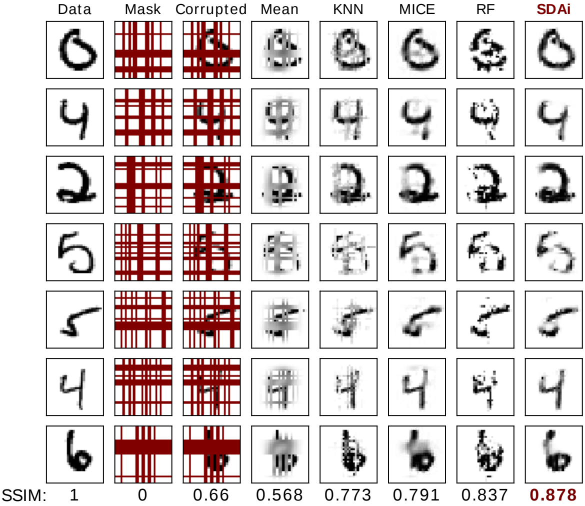



- Stacked denoising autoencoder [Vincent et al., 2010]
- Another way to make sure that AE learns useful representations is denoising AE.
- How train AE: minimize

$$\underset{\mathrm{w}}{\mathsf{argmin}} \underset{\mathrm{p(x)}}{\mathbb{E}} [\mathrm{L}(\mathrm{x}, \hat{\mathrm{x}})]$$

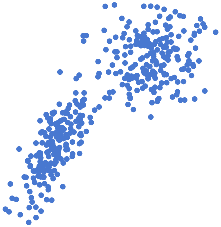
$$\mathrm{L}(\mathrm{x}, \hat{\mathrm{x}}) = -\log \mathrm{p}(\mathrm{x}|\hat{\mathrm{x}})$$

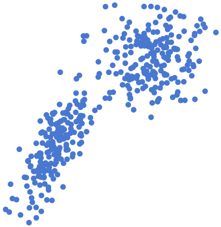
# Imputation with denoising autoencoder

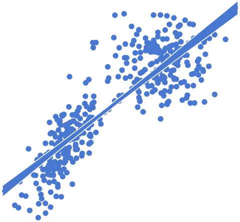


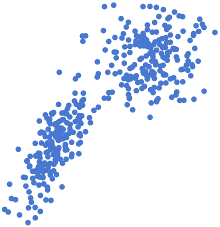

SSIM $\in [-1, 1]$ measures the perceptual difference between two similar images [Abiri et al., 2019]

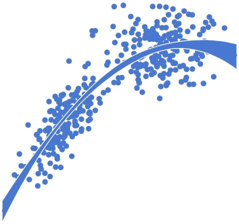
# First exercise!

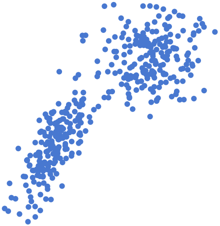


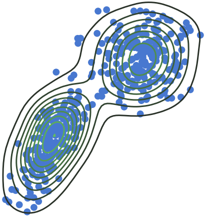




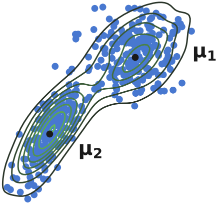








- Learning uncertainty




Goal: Learn $p_{data}(x)$

- Learning underlying distribution of the data
- Combine neural networks with Statistical modeling
- Subset of unsupervised learning (No labels)
- Generative models $(p(x, z))$ learn data distribution by learning latent variables z.

# Why study generative models?

istock.com-georgeclerk

- Outlier detection
- Trained with missing data
- Combined with reinforcement learning

# Why study generative models?


istock.com-georgeclerk

- Outlier detection
- Trained with missing data
- Combined with reinforcement learning

# Why study generative models?

- Outlier detection
- Trained with missing data
- Combined with reinforcement learning

Deep learning meets Bayesian approach:

- Generative Adversarial Networks (GANs)
  This person does not exist
- Variational Autoencoders (VAEs)

# Inference problem

- x data
- z latent variables (part of the model)
- Bayesian inference: Parameter uncertainty in machine learning

$$p(x, z) = p(z|x)p(x)$$

$$p(z|x) = \frac{p(x, z)}{p(x)} = \frac{p(x|z)p(z)}{p(x)}$$

# Inference problem

- x data
- z latent variables (part of the model)
- Bayesian inference: Parameter uncertainty in machine learning

$$\mathrm{p}(\mathrm{x}, \mathrm{z}) = \mathrm{p}(\mathrm{z}|\mathrm{x})\mathrm{p}(\mathrm{x})$$

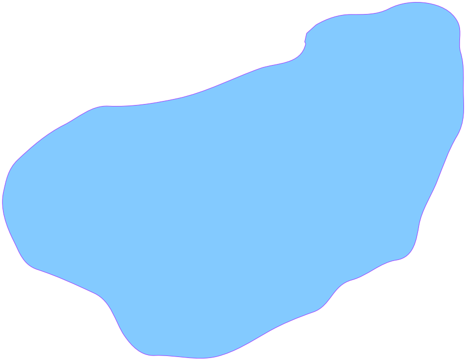
$$\mathrm{p}(\mathrm{z}|\mathrm{x}) = \frac{\mathrm{p}(\mathrm{x}, \mathrm{z})}{\mathrm{p}(\mathrm{x})} = \frac{\mathrm{p}(\mathrm{x}|\mathrm{z})\mathrm{p}(\mathrm{z})}{\mathrm{p}(\mathrm{x})}$$

- Marginal density of the observation is intractable

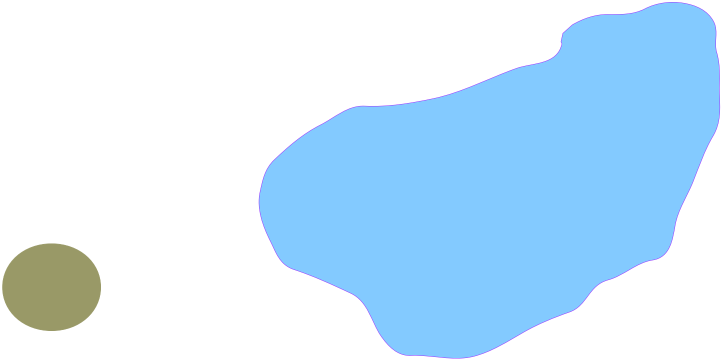
$$\mathrm{p}(\mathrm{x}) = \int \mathrm{p}(\mathrm{x}, \mathrm{z})\mathrm{d}\mathrm{z}$$

- Optimization-based approaches: Variational Inference
- Sampling approaches: MCMC

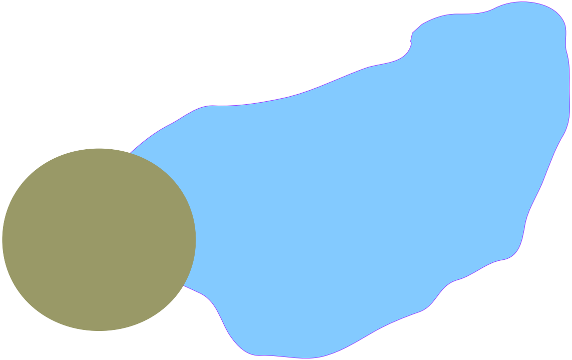
Approximate the posterior $p(z|x)$ with a simpler distribution $q(z|x)$.



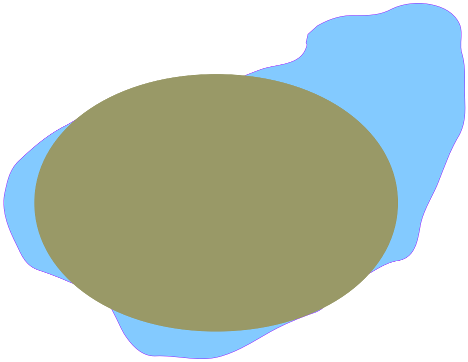
Approximate the posterior $p(z|x)$ with a simpler distribution $q(z|x)$.



Approximate the posterior p(z|x) with a simpler distribution q(z|x).



Approximate the posterior p(z|x) with a simpler distribution q(z|x).



- MFVI: Mean Field Variational Inference (1987) [Peterson, 1987]

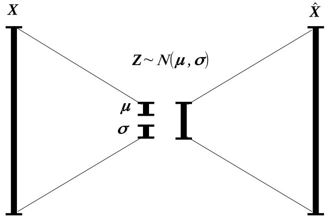
$$q(z) = \prod_i q(z_i)$$

$\vdots$

- Amortized Variational Inference (1995) [Dayan et al., 1995, Gershman and Goodman, 2014]

$$z_i = f(x_i)$$

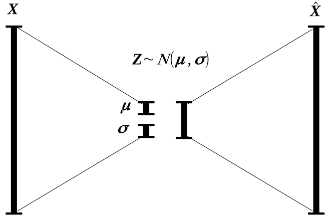
- Variational Autoencoders (2014) [Kingma and Welling, 2013, Rezende et al., 2014]




Inference model (encoder): $q_\phi(z|x)$ Generative model (decoder) :$p_\theta(x, z)$

- Prior: $p(z) = \mathcal{N}(0, \mathbb{I})$
- Variational posterior: $q_\phi(z|x) = \prod_i \mathcal{N}(z_i|\mu_i, \sigma_i^2)$
- Kullback-Leibler Divergence score:

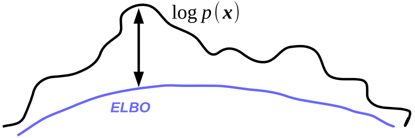
$$\mathcal{D}_{KL}(q_\phi(z|x) \| p_\theta(z|x))$$

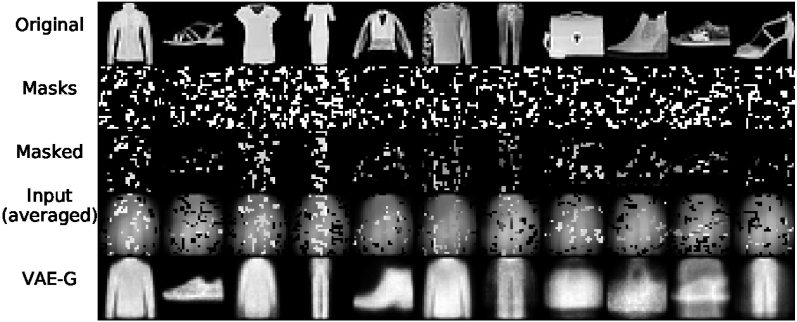



Inference model (encoder): $q_\phi(z|x)$ Generative model (decoder) :$p_\theta(x, z)$

- Evidence lower bound objective (ELBO):

$$\mathfrak{L} = \mathop{\mathbb{E}}_{q_\phi(z|x)}[\log(p_\theta(x|z))] - \mathcal{D}_{KL}(q_\phi(z|x) \| p_\theta(z))$$






Fashion-MNIST with 10 classes and 70000 instances [Abiri and Ohlsson, 2020]

# References

Najmeh Abiri and Mattias Ohlsson. Variational auto-encoders with student's t-prior. arXiv preprint arXiv:2004.02581, 2020.

Najmeh Abiri, Björn Linse, Patrik Edén, and Mattias Ohlsson. Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems. Neurocomputing, 365:137–146, 2019.

Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. Neural computation, 7(5):889–904, 1995.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In Proceedings of the annual meeting of the cognitive science society, volume 36, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016. `http://www.deeplearningbook.org`.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Carsten Peterson. A mean field theory learning algorithm for neural networks. Complex systems, 1:995–1019, 1987.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:1401.4082, 2014.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 11(12), 2010.