Assignment 2

Learning latent space representations

and application to image generation

Alexandre Vérine, Constant Bourdrez
(Benjamin Negrevergne)

PSL University – ENS PSL

Outline

Reptile classification challenge

Task: classify pictures of crocodiles and alligators

Crocodile

Alligator

Reptile classification challenge

Task: classify pictures of crocodiles and alligators

Representation 1: features

$$\begin{bmatrix} \textit{beast\_color} & = & \textit{Light} \\ \textit{beast\_size} & = & \textit{Large} \end{bmatrix} \longrightarrow f_1 \longrightarrow \text{crocodile}$$

Reptile classification challenge

Task: classify pictures of crocodiles and alligators

Representation 1: features

$$\begin{bmatrix} beast\_color & = & Dark \\ beast\_size & = & Small \end{bmatrix} \longrightarrow f_1 \longrightarrow \text{alligator}$$

Reptile classification challenge

Task: classify pictures of crocodiles and alligators

- **Representation 1: features**

$$\begin{bmatrix} \textit{beast\_color} & = & \textit{Dark} \\ \textit{beast\_size} & = & \textit{Small} \end{bmatrix} \longrightarrow f_1 \longrightarrow \text{alligator}$$

- **Representation 2: Use raw pixel data**

$$\longrightarrow f_2 \longrightarrow \text{crocodile}$$

Reptile classification challenge

Task: classify pictures of crocodiles and alligators

- **Representation 1: features**

$$\begin{bmatrix} beast\_color & = & Dark \\ beast\_size & = & Small \end{bmatrix} \quad \rightarrow \quad f_1 \quad \rightarrow \quad \text{alligator}$$

- **Representation 2: Use raw pixel data**

$$\rightarrow \quad f_2 \quad \rightarrow \quad \text{alligator}$$

Reptile classification challenge

Task: classify pictures of crocodiles and alligators

■ **Representation 1: features**

$$\begin{bmatrix} \textit{beast\_color} & = & \textit{Dark} \\ \textit{beast\_size} & = & \textit{Small} \end{bmatrix} \longrightarrow f_1 \longrightarrow \text{alligator}$$

■ **Representation 2: Use raw pixel data**

$$\longrightarrow f_2 \longrightarrow \text{alligator}$$

Which **representation** of the input is easier to work with ? Why ?

What's the difference?

$$f_1 \quad : \quad \mathbb{R}^2 \quad \to \quad \mathbb{R}$$

Input space 1: $\mathbb{R}^2$

$$f_2 \quad : \quad \mathbb{R}^{w \times h \times 3} \quad \to \quad \mathbb{R}$$

Input space 2: $\mathbb{R}^{w \times h \times 3}$

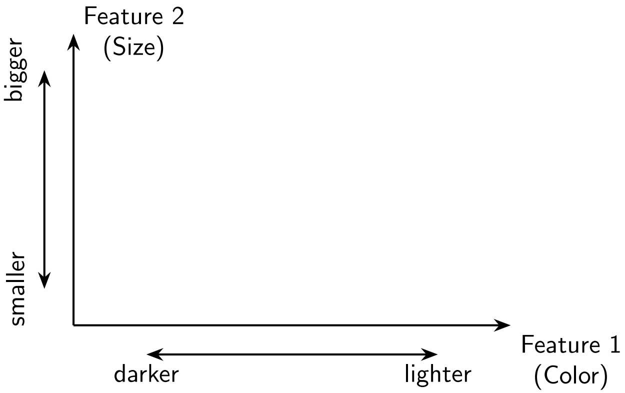
What's the difference?

$f_1 \quad : \quad \mathbb{R}^2 \quad \to \quad \mathbb{R}$

Input space 1: $\mathbb{R}^2$

$f_2 \quad : \quad \mathbb{R}^{w \times h \times 3} \quad \to \quad \mathbb{R}$

Input space 2: $\mathbb{R}^{w \times h \times 3}$



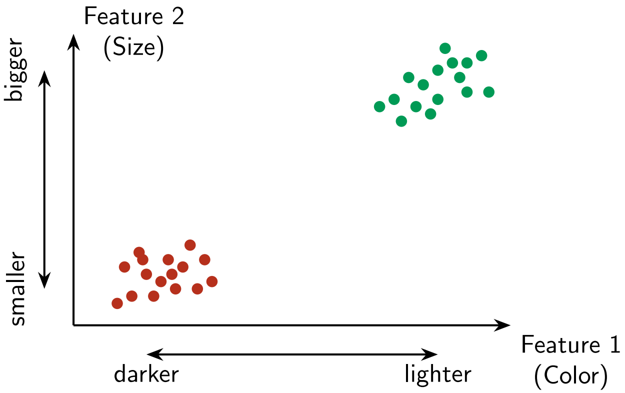
What's the difference?

$f_1 \quad : \quad \mathbb{R}^2 \quad \to \quad \mathbb{R}$

Input space 1: $\mathbb{R}^2$

$f_2 \quad : \quad \mathbb{R}^{w \times h \times 3} \quad \to \quad \mathbb{R}$

Input space 2: $\mathbb{R}^{w \times h \times 3}$



Data points **are linearly separable** in $\mathbb{R}^2$

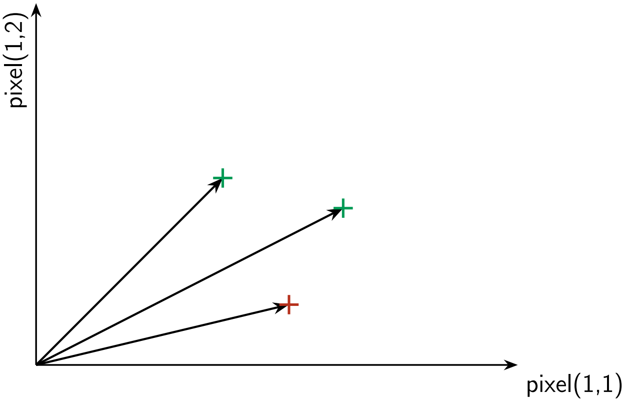
What's the difference?

$$f_1 \quad : \quad \mathbb{R}^2 \quad \to \quad \mathbb{R}$$

Input space 1: $\mathbb{R}^2$

$$f_2 \quad : \quad \mathbb{R}^{w \times h \times 3} \quad \to \quad \mathbb{R}$$

Input space 2: $\mathbb{R}^{w \times h \times 3}$



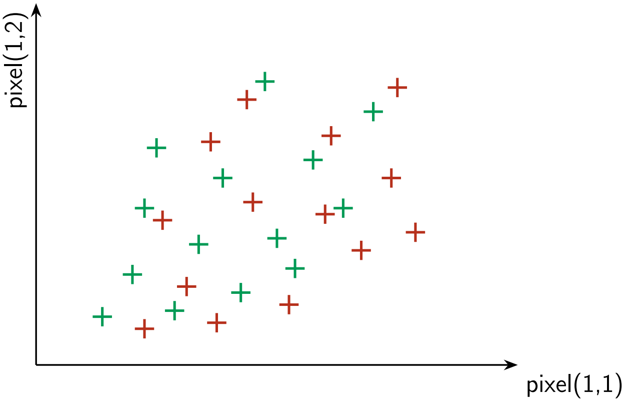
What's the difference?

$f_1 \quad : \quad \mathbb{R}^2 \quad \rightarrow \quad \mathbb{R}$

Input space 1: $\mathbb{R}^2$

$f_2 \quad : \quad \mathbb{R}^{w \times h \times 3} \quad \rightarrow \quad \mathbb{R}$

Input space 2: $\mathbb{R}^{w \times h \times 3}$



Data points **are not linearly separable** in $\mathbb{R}^{w \times h \times 3}$, Why?

Pro/cons

- **Representation 1: hand-crafted features**
  + Can be processed with simple (linear) models
    - Individual features are highly discriminant
    - Input data points are (almost) linearly separable
  − Requires expertise and manual labor to be built
  − No extra information in case of ties

Pro/cons

- **Representation 1: hand-crafted features**
  - \+ Can be processed with simple (linear) models
    - Individual features are highly discriminant
    - Input data points are (almost) linearly separable
  - – Requires expertise and manual labor to be built
  - – No extra information in case of ties

- **Representation 2: raw pixel data**
  - \+ Contains all the information available
  - – Input data points are not (nearly) linearly separable

    Features are individually non-discriminant
  - – Difficult to process with simple models

Pro/cons

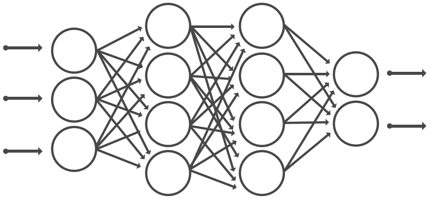
- **Representation 1: hand-crafted features**
  - \+ Can be processed with simple (linear) models
    - Individual features are highly discriminant
    - Input data points are (almost) linearly separable
  - – Requires expertise and manual labor to be built
  - – No extra information in case of ties

- **Representation 2: raw pixel data**
  - \+ Contains all the information available
  - – Input data points are not (nearly) linearly separable

    Features are individually non-discriminant
  - – Difficult to process with simple models

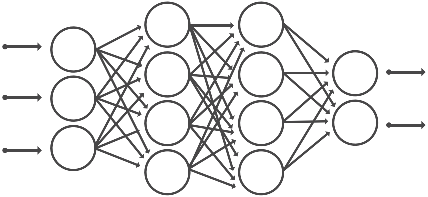
Can we build high level features automatically ?

Deep neural nets



$$f = f_1 \circ f_2 \circ \ldots \circ f_{n-1} \circ f_n$$

Deep neural nets



$$f = f_1 \circ f_2 \circ \ldots \circ f_{n-1} \circ f_n$$

- $f_1 : \mathcal{X} \to \mathcal{Z}_1$
- $f_i : \mathcal{Z}_{i-1} \to \mathcal{Z}_i$
- $f_n : \mathcal{Z}_{n-1} \to \mathcal{Y}$

Deep neural nets

$$f = \underbrace{f_1 \circ f_2 \circ \ldots \circ f_{n-1}}_{g} \circ \underbrace{f_n}_{h}$$

Remarks

- h is a linear classifier: $\mathcal{Z}_{n-1} \rightarrow \mathcal{Y}$
  - ▶ Data points **must be** linearly separable in $\mathcal{Z}_{n-1}$
- Data points x **are not** linearly separable in the input space $\mathcal{X}$

Deep neural nets

$$f = \underbrace{f_1 \circ f_2 \circ \ldots \circ f_{n-1}}_{g} \circ \underbrace{f_n}_{h}$$

Remarks

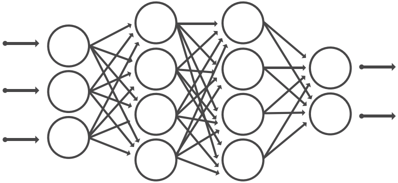
- h is a linear classifier: $\mathcal{Z}_{n-1} \to \mathcal{Y}$
  ▶ Data points **must be** linearly separable in $\mathcal{Z}_{n-1}$
- Data points x **are not** linearly separable in the input space $\mathcal{X}$

What is g?

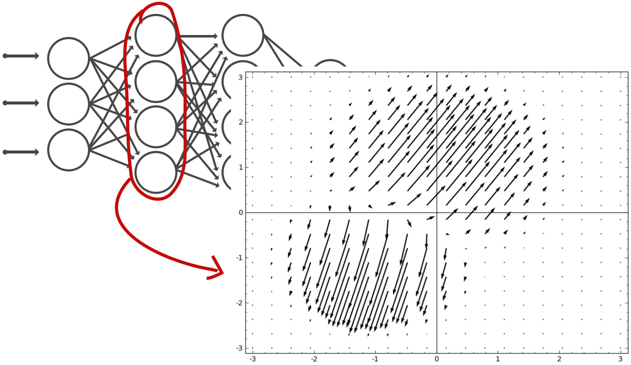
- a function $g : \mathcal{X} \to \mathcal{Z}_{n-1}$
- such that data points are linearly separable in $\mathcal{Z}_{n-1}$

▶ a representation of the point in $\mathcal{X}$ that is adequate for the task at hand

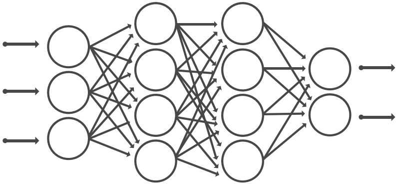
Learning deep representations



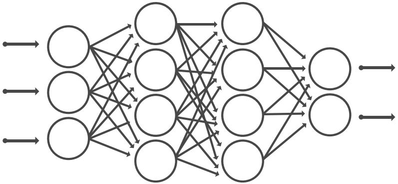
Learning deep representations

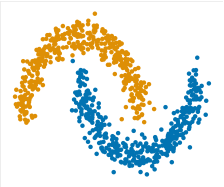


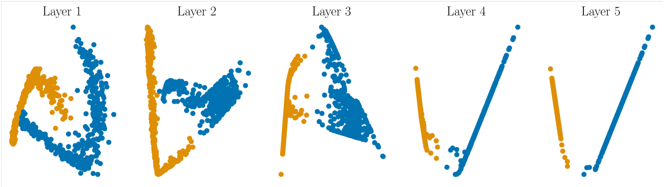
Learning deep representations



Learning deep representations







Deep representations: first lessons

- DNN learn how to projet inputs into a latent space
- The structure of the latent space is useful for the task at hand.
- Given an input $x \in \mathcal{X}$ we say that $g(x)$ is an **embedding** of x, i.e. continuous vector representations of input data (image, text, graph ...)

Outline

1. What is a good representation ?
2. Learning representations with Deep Learning
3. Intriguing properties of learned representations
4. Synthetic data generation

Outline

Image embeddings

- D: a database with 80 000 pictures.
- $f = g \circ h$: a classifier trained on object recognition
  g non-linear function, h linear classifier
- x a random picture from the internet

$$x_1 = \arg\min_{x' \in D} ||g(x) - g(x')|| \quad \text{and} \quad x_2 = \arg\min_{x' \in D \setminus \{x_1\}} ||g(x) - g(x')||$$

Image embeddings

- D: a database with 80 000 pictures.
- $f = g \circ h$: a classifier trained on object recognition
  g non-linear function, h linear classifier
- x a random picture from the internet

$$x_1 = \underset{x' \in D}{\arg\min} ||g(x) - g(x')|| \quad \text{and} \quad x_2 = \underset{x' \in D \setminus \{x_1\}}{\arg\min} ||g(x) - g(x')||$$

x

Image embeddings

- D: a database with 80 000 pictures.
- $f = g \circ h$: a classifier trained on object recognition
  g non-linear function, h linear classifier
- x a random picture from the internet

$$x_1 = \arg\min_{x' \in D} ||g(x) - g(x')|| \quad \text{and} \quad x_2 = \arg\min_{x' \in D \setminus \{x_1\}} ||g(x) - g(x')||$$

x

x_1

x_2

Query by image (2)

x

Query by image (2)

x

x_1

x_2

Query by image (2)

X

Query by image (2)

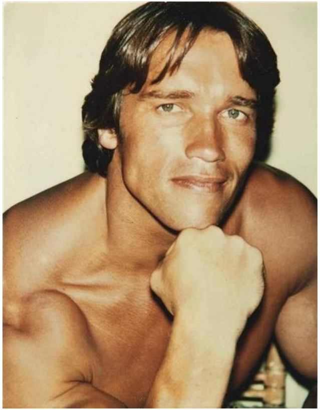
x

x_1

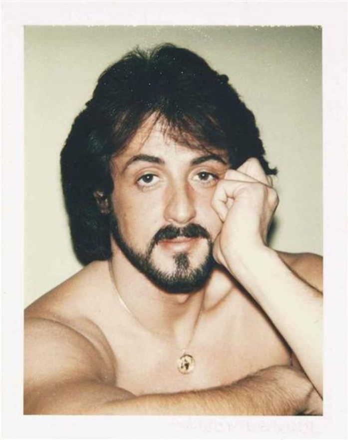
x_2

Query by image (2)

x



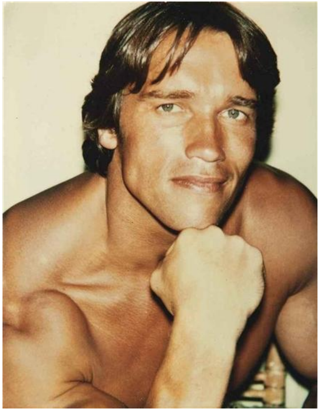
x_1



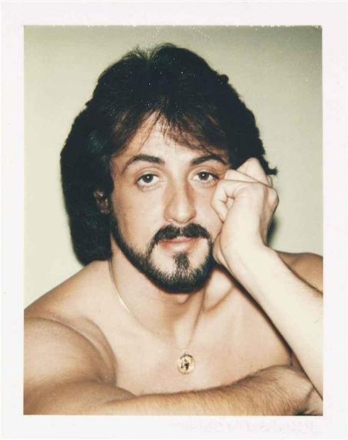
x_2

Query by image (2)

x



x_1



x_2

Result: The distance in the latent space seems to be meaningful!

Word embeddings

We can train (monolingual) *word embeddings*,
i.e. representations trained to predict well words that appear in its context (ref here)

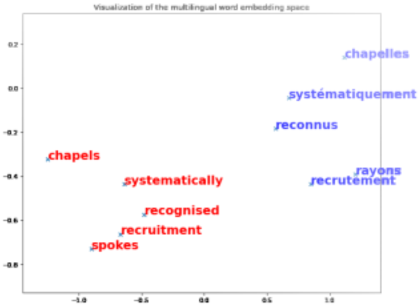


Image generated using pretrained embeddings available here. *En avant guiGuan team*, 2021.

Word embeddings

We can train (monolingual) *word embeddings*,
i.e. representations trained to predict well words that appear in its context (ref here)

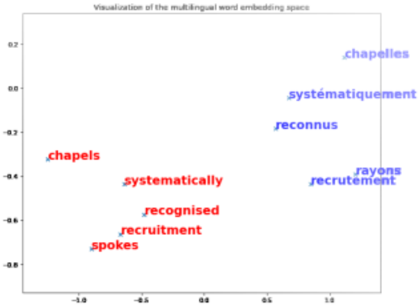
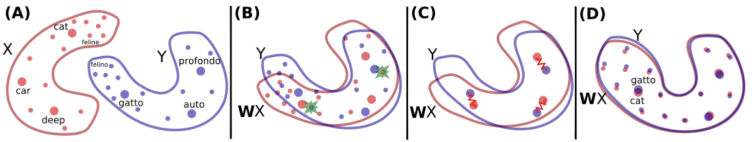


Image generated using pretrained embeddings available here. *En avant guiGuan team*, 2021.

- The structure between word embeddings is preserved across languages

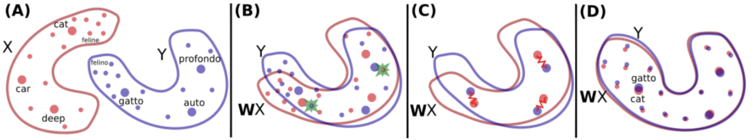
Word to word translation using word embeddings



Exploit linear transformations and rotations to translate a word.

$$Y = \mathbf{W}X$$

Word to word translation using word embeddings



Exploit linear transformations and rotations to translate a word.

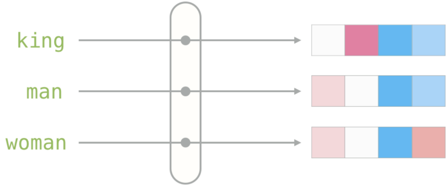
$$Y = \mathbf{W}X$$

Learn $\mathbf{W}$

- with a parallel corpus (e.g. supervised dataset FR-EN)
- without a parallel corpus using a minmax and a discriminator (GAN-like)

Text manipulation with word embeddings

The embedding space is geometric!

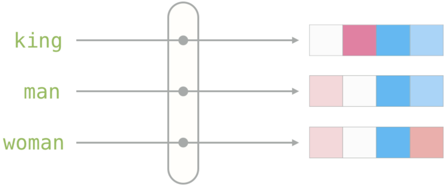


The embeddings space is geometric:

$$v(king) - v(man) + v(woman) = \ ?$$

Text manipulation with word embeddings

The embedding space is geometric!



The embeddings space is geometric:

$$v(king) - v(man) + v(woman) = \text{ queen}$$

Visualising the data manifold

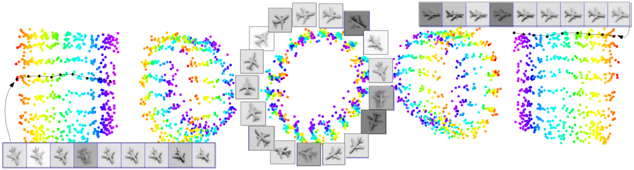
Dimensionality reduction by learning an invariant mapping.
Hadsell, R., Chopra, S., LeCun, Y. CVPR (2006)
Learns a mapping g that maps inputs with **few controlled variations** to a **low dimensional space**

- all pictures are pictures of planes with different poses
- 9 different elevations and 18 different azimuth (orientation).
- input pictures are projected into a low 3-dimensional space

Visualising the data manifold

Dimensionality reduction by learning an invariant mapping.
Hadsell, R., Chopra, S., LeCun, Y. CVPR (2006)
Learns a mapping g that maps inputs with **few controlled variations** to a **low dimensional space**

- all pictures are pictures of planes with different poses
- 9 different elevations and 18 different azimuth (orientation).
- input pictures are projected into a low 3-dimensional space

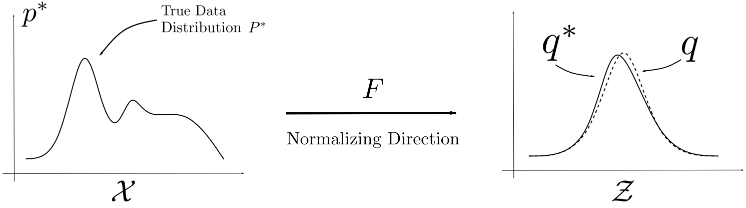


Result:

- Most input data lies on a well defined subspace of the output space
- There is a clear relation between spacial coordinates and features (elevation, azimuth)

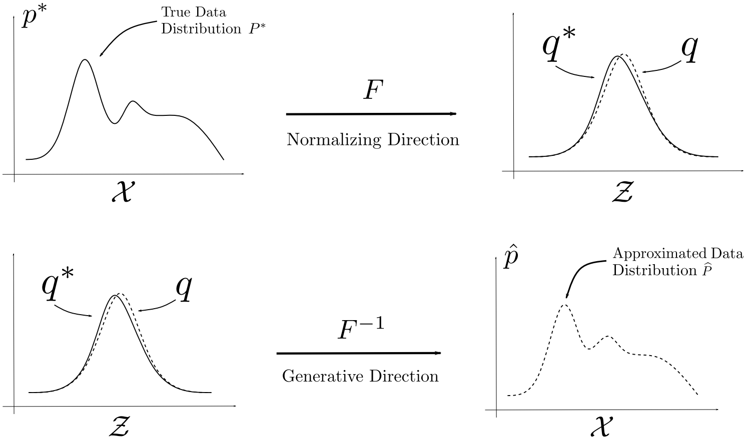
Continuous Representation

What if we enforce continuity and a specific structure in the latent space?



Continuous Representation

What if we enforce continuity and a specific structure in the latent space?



We can generate new data points by sampling in the latent space!

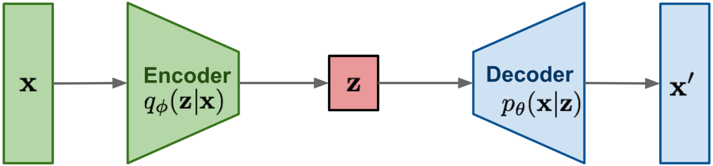
Visualising the latent space

Interpolating between two points in the latent space:

and more examples in this blog post: StyleGAN2 latent space exploration

Outline

Data generation with VAE

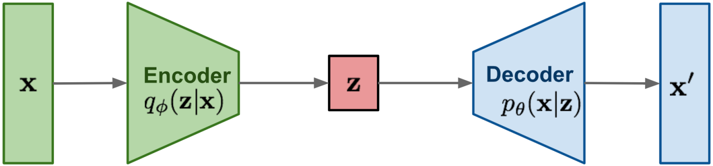


VAE, image from https://lilianweng.github.io

Main differences with AE

- Use a probabilistic encoder
- Regularize the latent space during training

Data generation with VAE



VAE, image from https://lilianweng.github.io

Main differences with AE

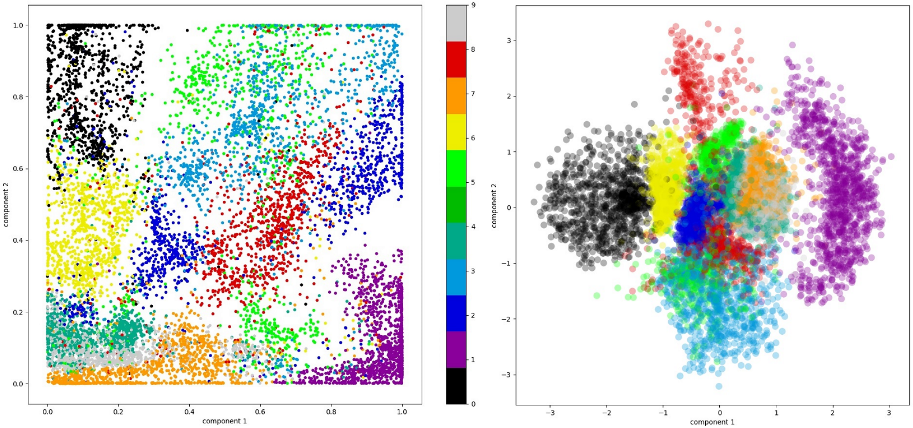
- Use a probabilistic encoder
- Regularize the latent space during training

Training procedure

1. take a training data point x, obtain μ_x and σ_x from the encoder
2. sample $z \sim \mathcal{N}(\mu_x, \sigma_x)$, decode z into x'
3. compute the loss and update parameters

$$loss(x, x') = \|x - x'\| + KL(\mathcal{N}(\mu_x, \sigma_x), \mathcal{N}(0, I))$$

AE vs. VAE



1.0
0.8
0.6
0.4
0.2
0.0
component 2
0.0
0.2
0.4
0.6
0.8
1.0
component 1
9
8
7
6
5
4
3
2
1
0
3
2
1
0
−1
−2
−3
component 2
−3
−2
−1
0
1
2
3
component 1

Data generation with VAE

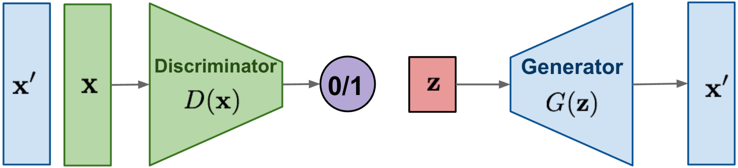
Pros:

- The latent space is continuous and well structured
- Easy to sample from the latent space

Cons:

- Generated images are often blurry
- Difficult to train
- Need to tune the weight of the KL term

Generative Adversarial Networks



GAN, image from https://lilianweng.github.io

Main difference with VAE

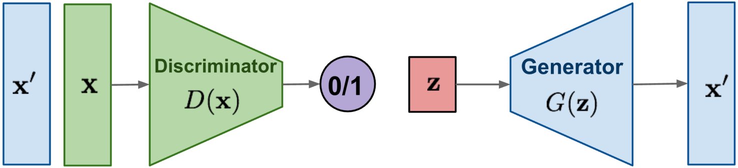
- VAE reconstructs inputs; GANs instead match the data distribution: train a generator $G(z)$ so its samples look real.
- Representation view: G maps a low-dimensional code z to data; a discriminator $D(x)$ provides the learning signal so $G(z)$ and real x share the same statistics.
- Adversarial training (min–max): G tries to fool D, and D learns to tell real from fake.

Training procedure

G and D are trained simultaneously to solve the following min-max problem:

$$\min_G \max_D \mathbb{E}_{x_r \sim P}[\log D(x)] + \mathbb{E}_{x_g \sim \hat{P}_G}[\log 1 - D(x)]$$

Generative Adversarial Networks



GAN, image from https://lilianweng.github.io

References:

- Generative Adversarial Nets
  https://arxiv.org/pdf/1406.2661
- Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks https://arxiv.org/abs/1511.06434

Goal of Assignment 2

1. Train a GAN on MNIST.
2. The structure of the **Generator is fixed**.
3. Use different one or two possible improvements to improve the data generation.

Requirements Assignment 2

1. Train a Generator (decoder) model to generate MNIST digits.
2. Write a script generate.py that generate 10000 samples in the folder samples (use mine).
3. Based on these 10k samples, you will be evaluated on FID, Precision and Recall.
   Precision/Recall

Base repo of the Assignment 2: Link

Requirements Assignment 2

Program of next week session:

- Review of different GAN implement for trading off quality of the samples with diversity.
- How to compute FID, Precision and Recall to measure the quality and the diversity of generated samples.
- How to run training on GPU cluster.
- Group session to help you start on Assignment 2.

To do list for next week:

- Read the GAN paperGenerative Adversarial Nets