Pie & AI: Strasbourg - Autoencoders and Variational Autoencoders

Strasbourg
05.11.2020

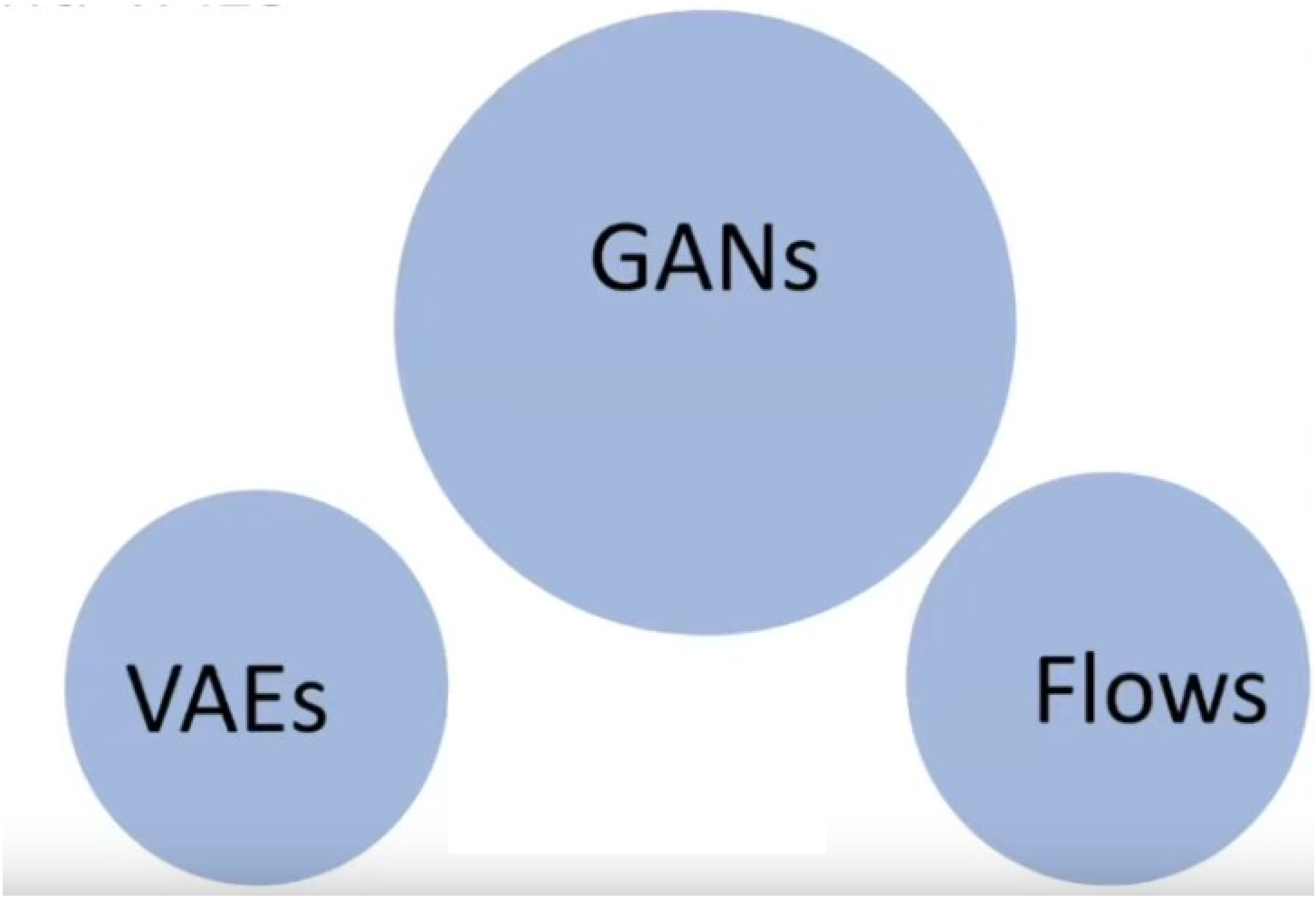
Titus Nicolae
Robert Maria

Richard Feynman: "*What I cannot create, I do not understand*"

Generative modeling: "*What I understand, I can **create***"

source

Most popular methods today



source

Overview

1) Dimensionality reduction

2) Autoencoders (AE)

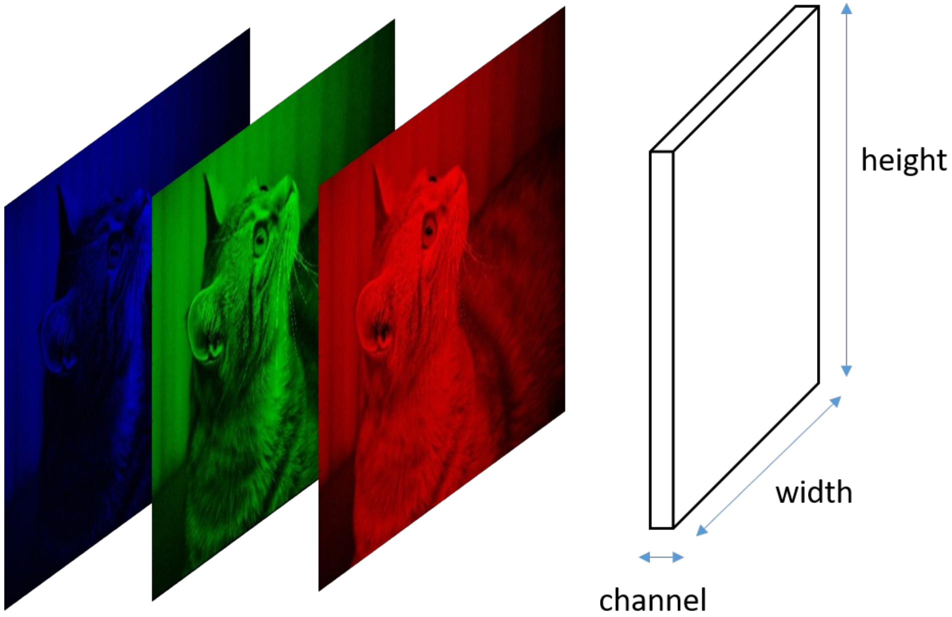
3) Variational Autoencoders (VAE)

4) Disentanglement

5) Beta-VAE

Dimensionality reduction

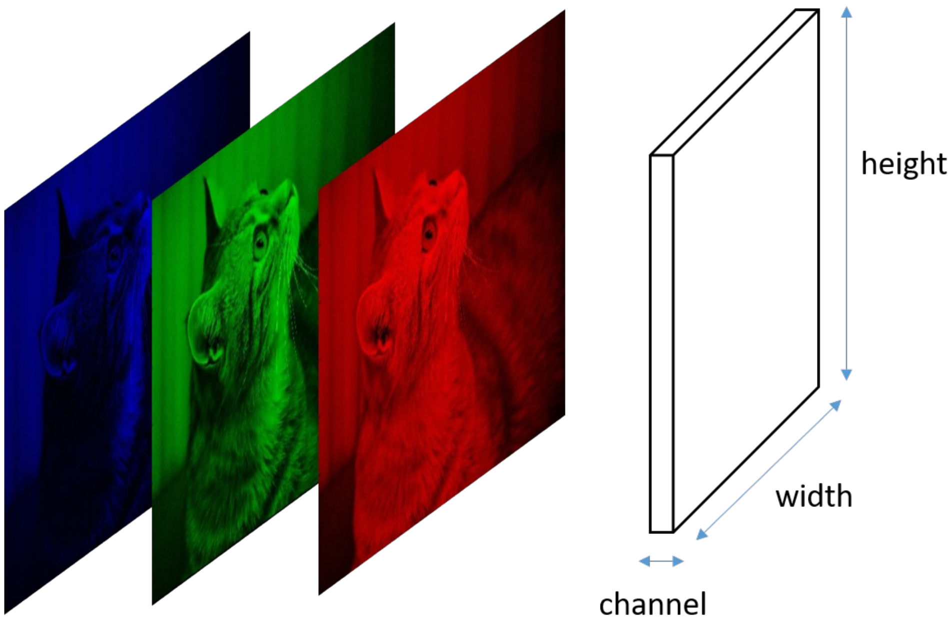
What is an image made of



source

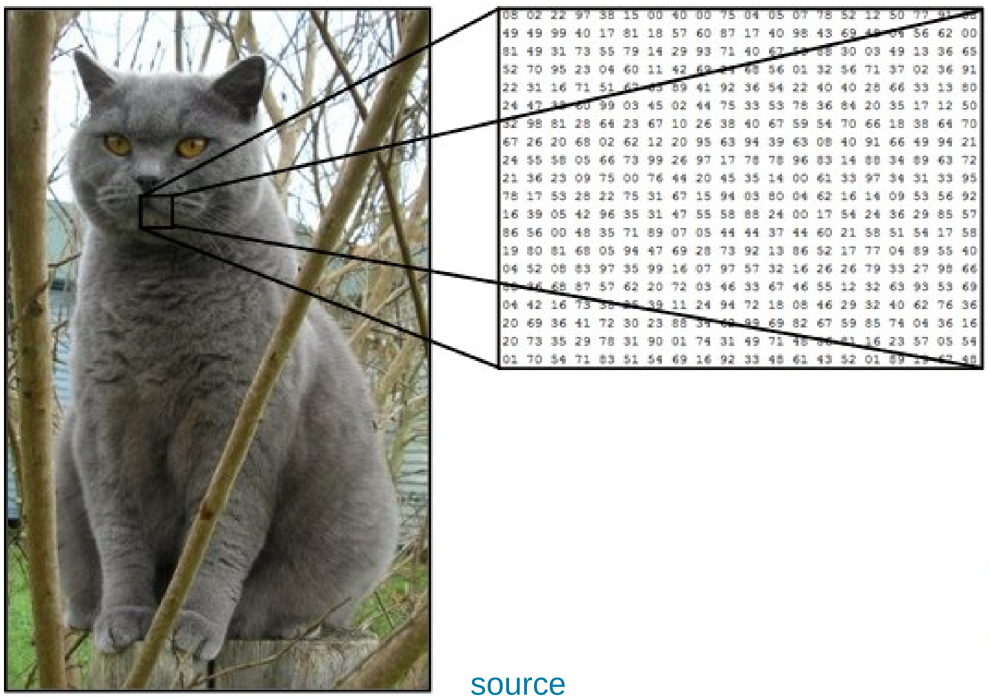
Dimensionality reduction

What is an image made of



source

What the computer sees



source

Can we reduce the dimensionality of the features (in this case pixels)?

Dimensionality reduction

Yes, we can reduce the dimension of data

by selection

Just some of the existing features are kept

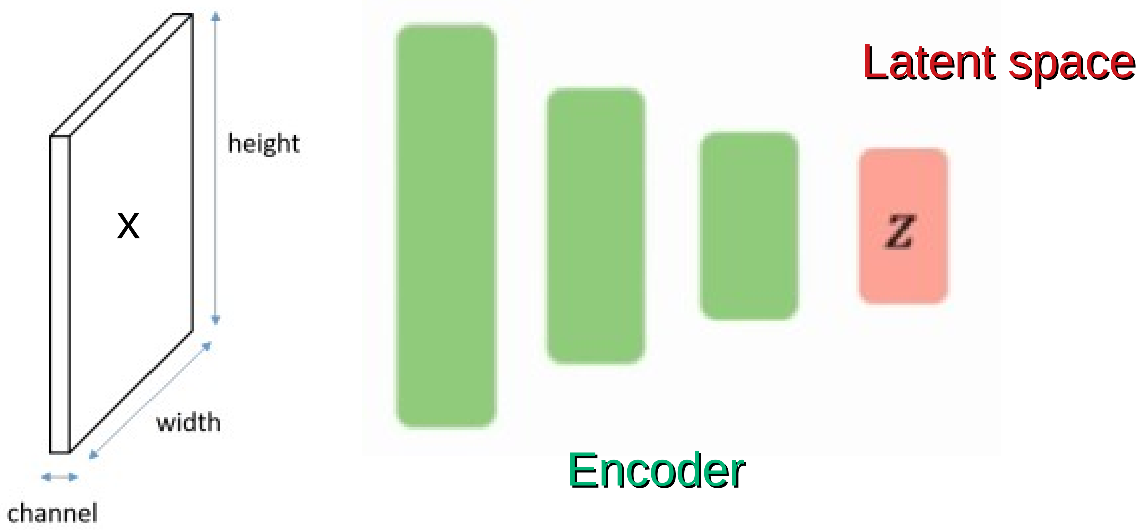
by extraction

New features are created

Reduce dimensionality in a semi-supervised manner with Autoencoders

Dimensionality reduction with Autoencoders

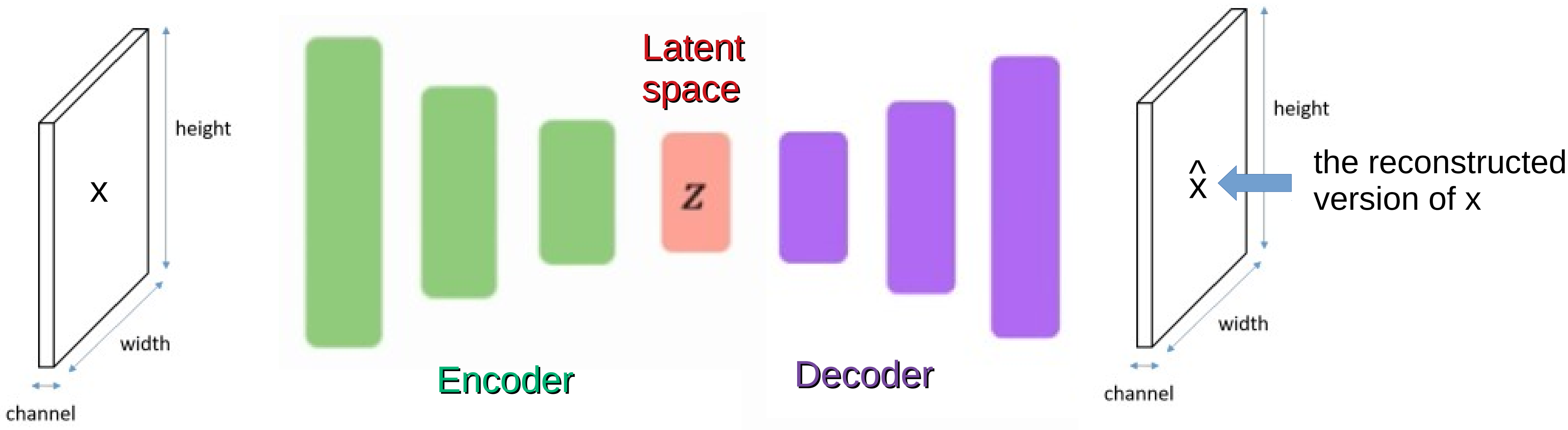
The Encoder of Autoencoder



Data reconstruction

The Encoder of Autoencoder

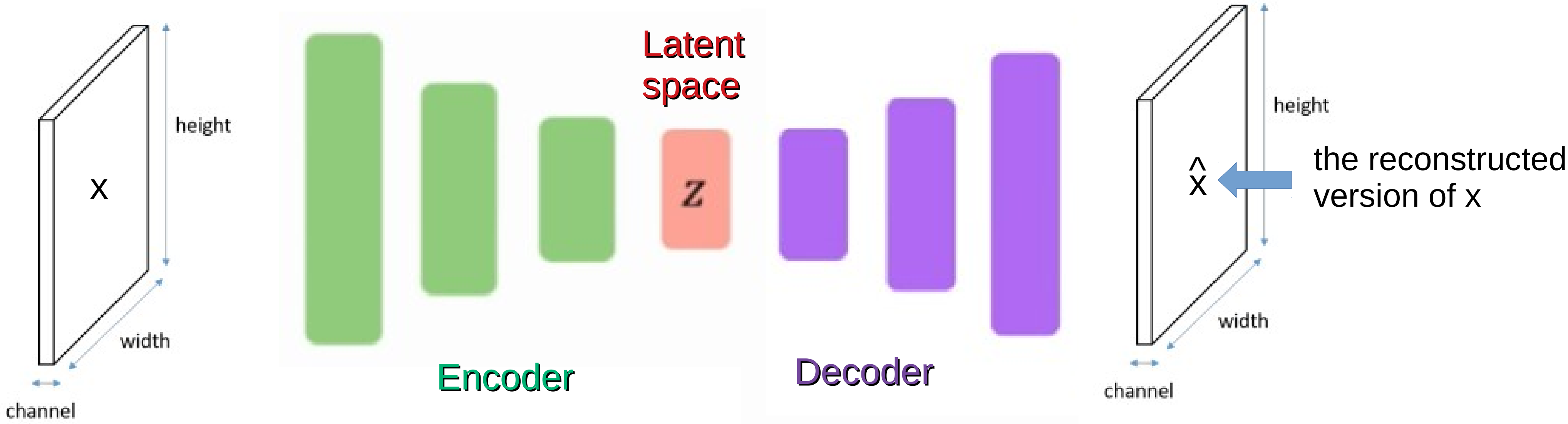
The Decoder of Autoencoder



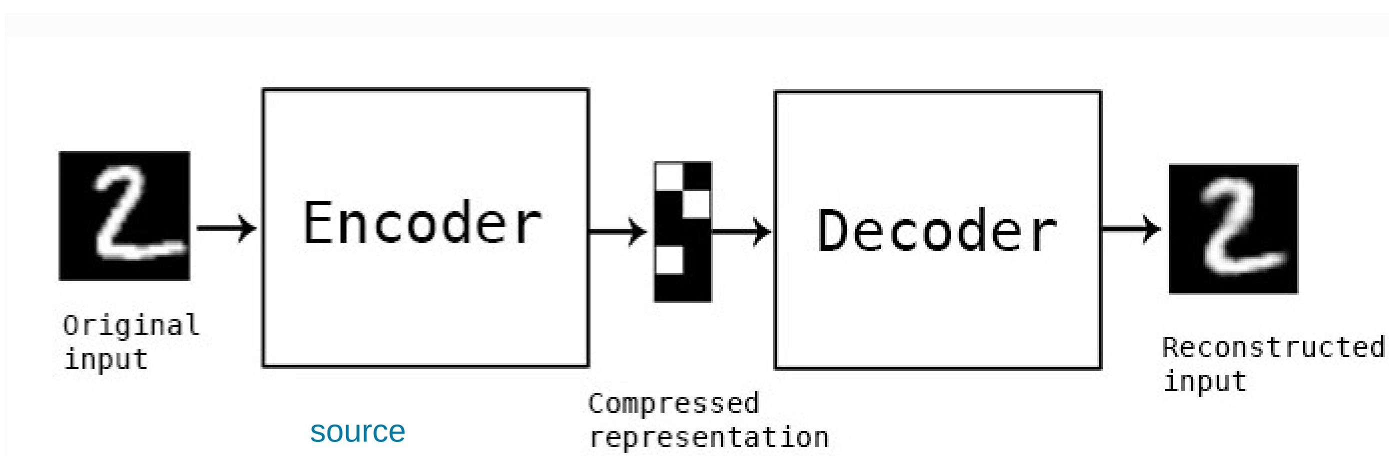
Data reconstruction

The Encoder of Autoencoder

The Decoder of Autoencoder



Example:



source

Autoencoders

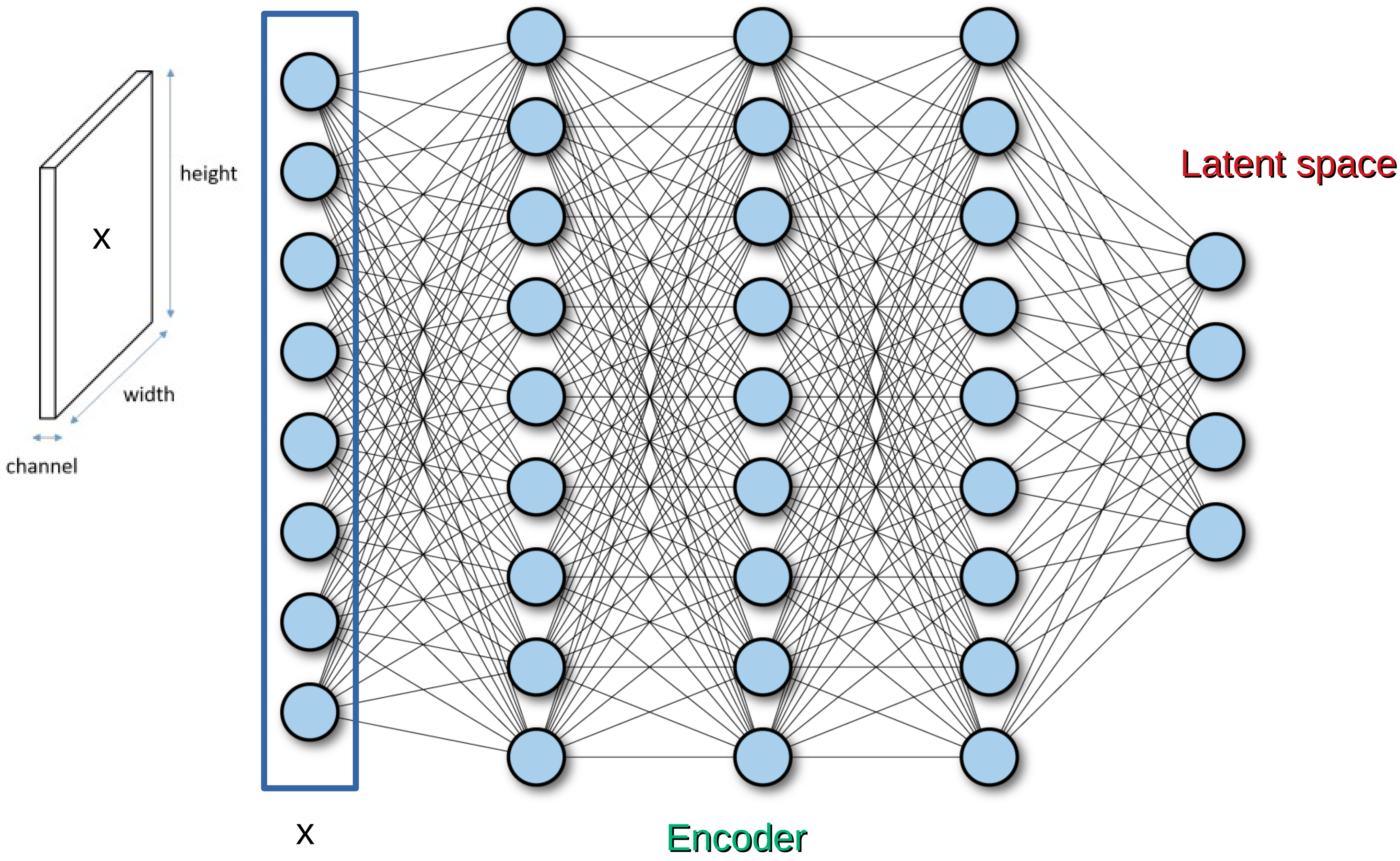
Autoencoders are a data compression algorithms where the compression and decompression functions are:

1) data-specific

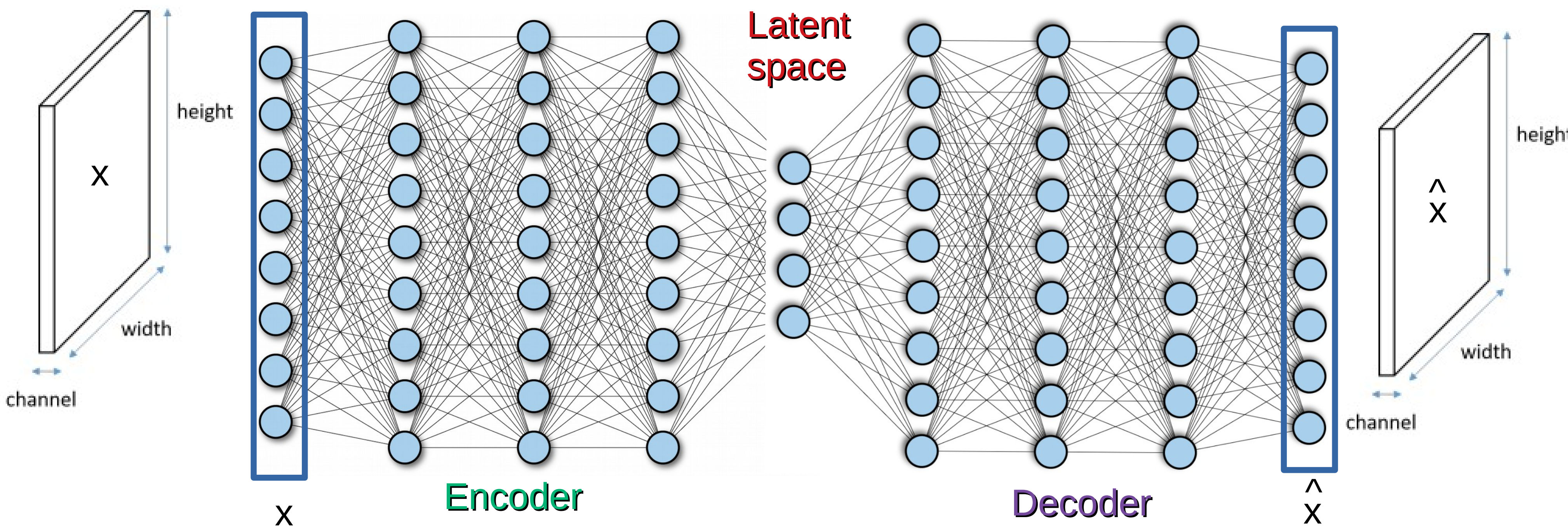
2) lossy

3) learned automatically

Deep Fully Connected Autoencoder

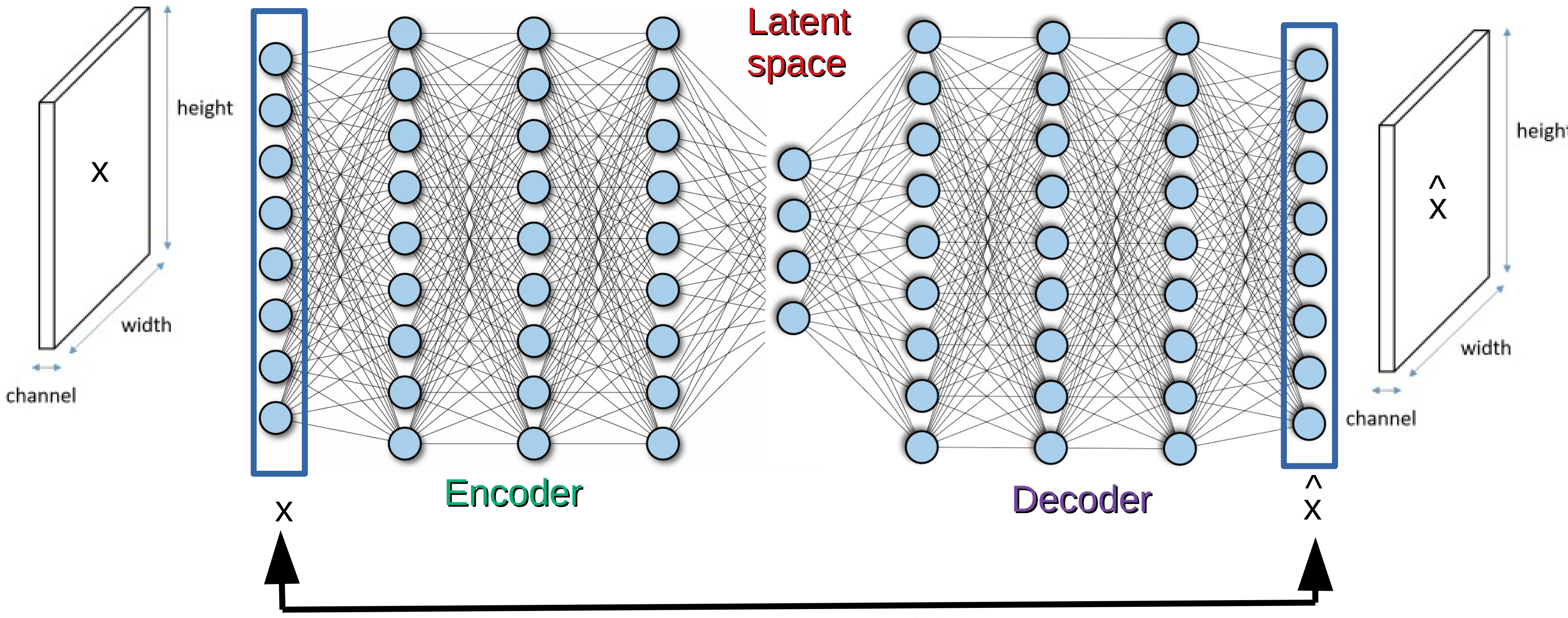


Deep Fully Connected Autoencoder



$$\hat{X} = D(E(x))$$

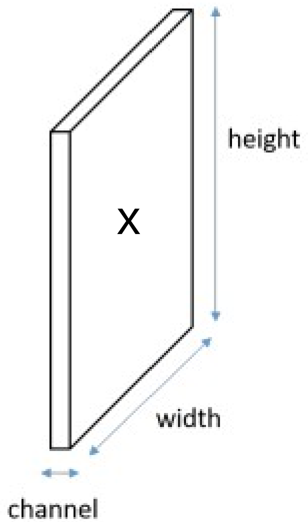
How do we train it?

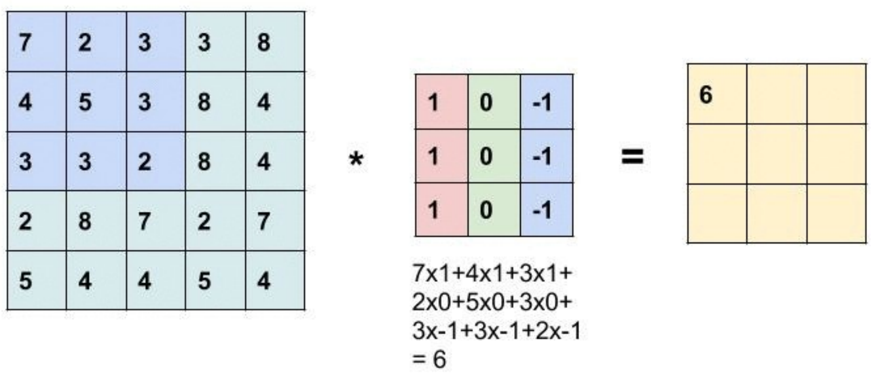


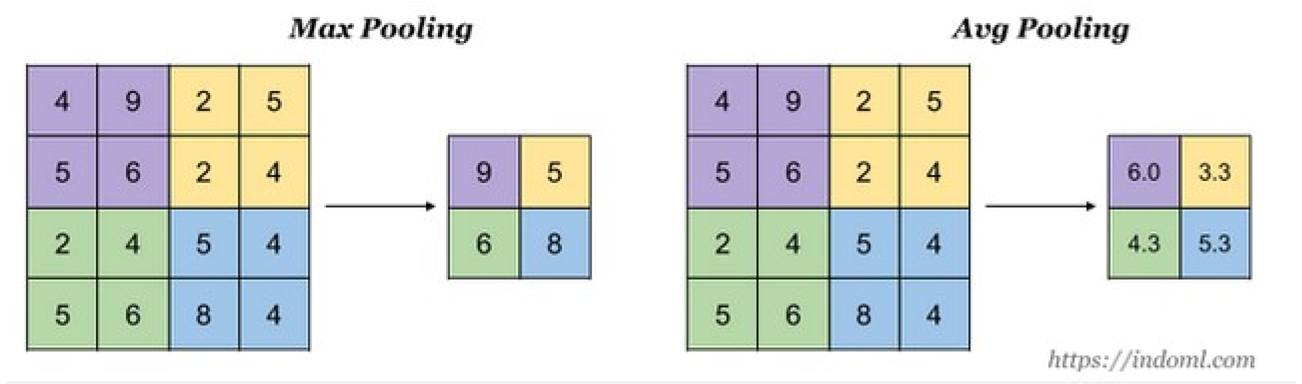
$$\mathcal{L}(x, \hat{x}) = \|x - \hat{x}\|^2$$

Deep Convolutional Autoencoder

Convolution layers

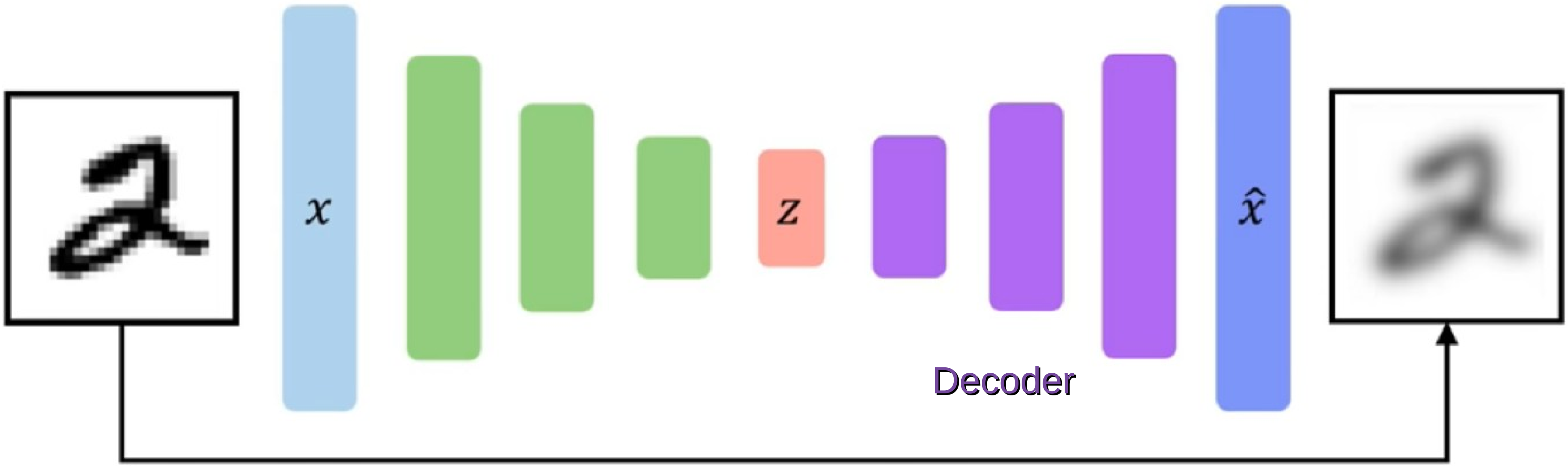






How do we train it?

Deep Convolutional Autoencoder



$$\mathcal{L}(x, \hat{x}) = \|x - \hat{x}\|^2$$

source

Encoder convolutional layers, pooling

Decoder convolutional layers, transposed convolutions or upsampling layers

What are they good for?

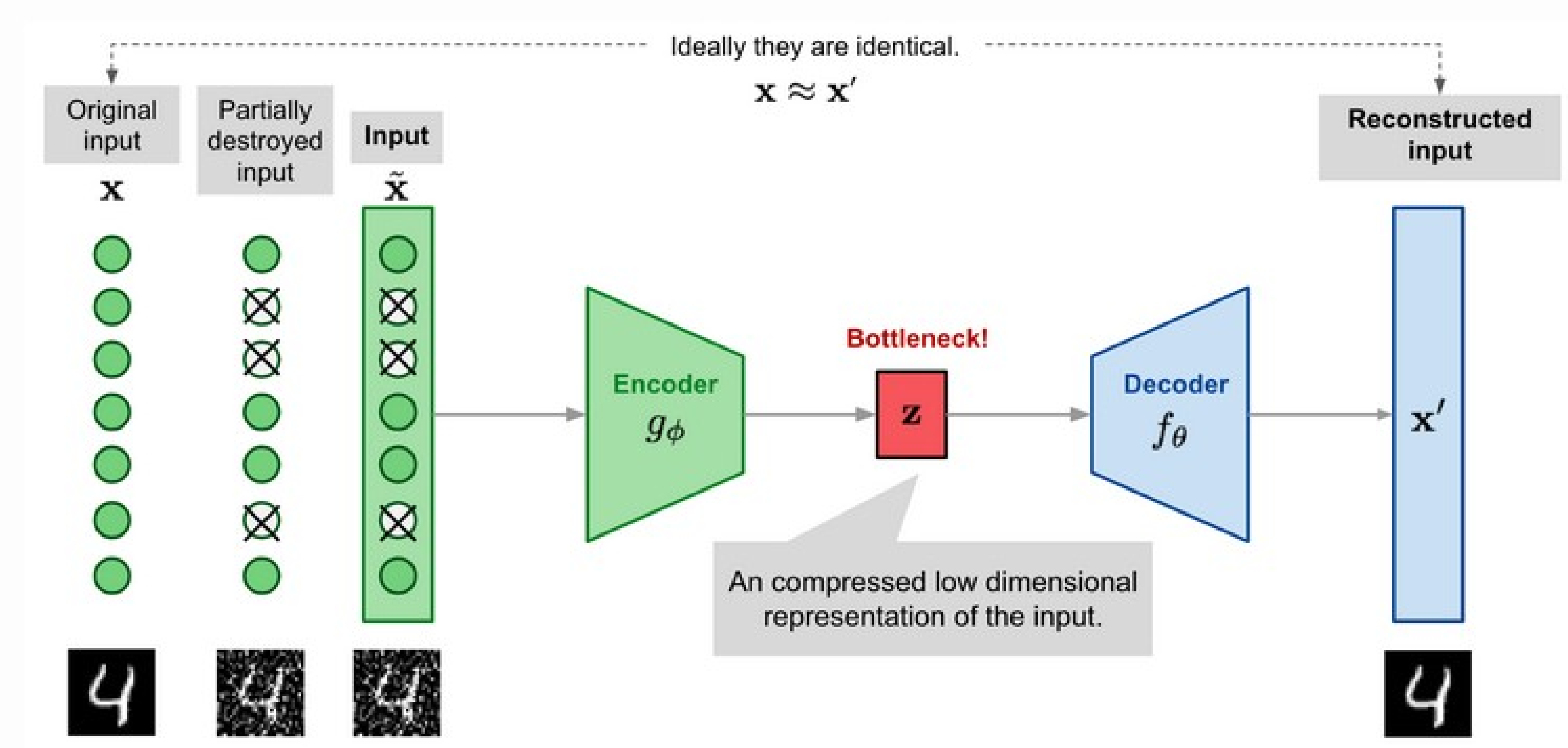
Not best choice for data compression

Data denoising (denoising autoencoders)

Dimensionality reduction for data visualization

Good starting point to understand Variational Autoencoders (VAEs)

Denoising Autoencoder



source

Overview

1) Dimensionality reduction

2) Autoencoders (AE)

3) Variational Autoencoders (VAE)

4) Disentanglement

5) Beta-VAE

Goal

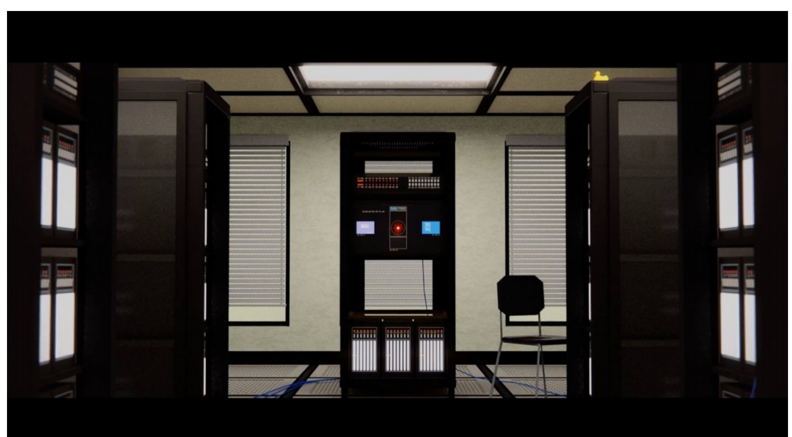

Shutdown
You don't have permission to shut down this computer.
OK
I'm sorry Dave, I can't let you do that

Goal

Building Machines That Learn and Think Like People

Brenden M. Lake,[1] Tomer D. Ullman,[2,4] Joshua B. Tenenbaum,[2,4] and Samuel J. Gershman[3,4]
[1]Center for Data Science, New York University
[2]Department of Brain and Cognitive Sciences, MIT
[3]Department of Psychology and Center for Brain Science, Harvard University
[4]Center for Brains Minds and Machines

source



Goal


Shutdown
You don't have permission to shut down this computer.
OK
I'm sorry Dave, I can't let you do that

Goal

Observe
the world

Goal

Observe the world

Interact with the world

Goal

Observe the world

Interact with the world

Internal representation of the world

Goal

Understand the laws of the world (physics, psychology, etc.)

Observe the world

Language

Interact with the world

Learn and pattern recognition

Internal representation of the world

Understand compositionality, causality

Imagination

Challenges

Generalization

Humans generalize from a single example?

Training example ("water bear")

Challenges

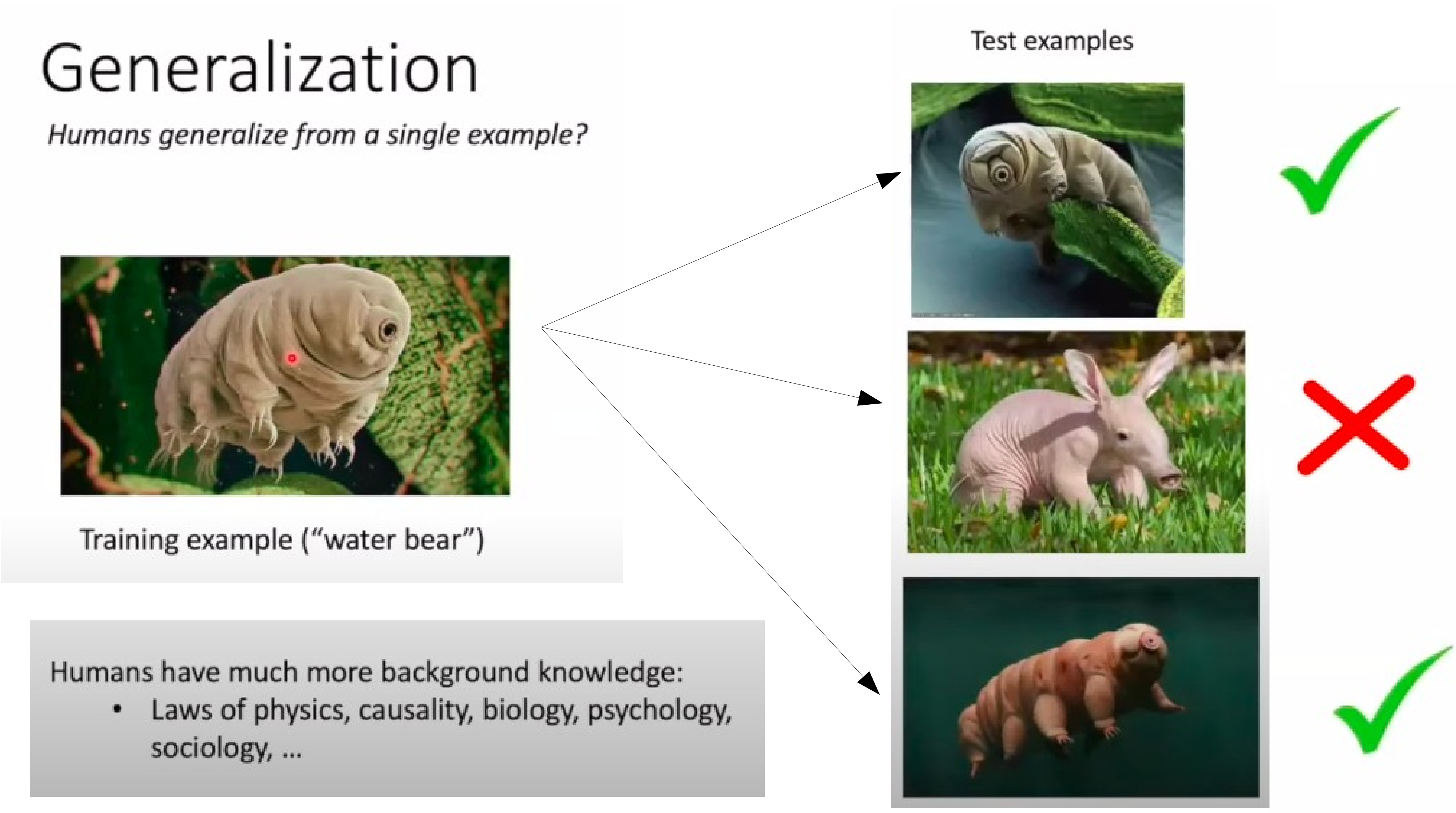
Generalization

Humans generalize from a single example?

Training example ("water bear")

Test examples

Challenges



source

Representation learning

Can you perform aritmetic on Roman numerals?

$$\boxed{XXXVII + XLII = ?}$$

Representation learning

Can you perform aritmetic on Roman numerals?

XXXVII + XLII = ?

Or you perform faster on Arabic numerals?

37 + 42 = ?

Representation learning

Can you perform aritmetic on Roman numerals?

XXXVII + XLII = ?

Or you perform faster on Arabic numerals?

37 + 42 = ?

“Many AI tasks can be solved by designing the right set of features to extract for that task”

source

Representation learning

Can you perform aritmetic on Roman numerals?

XXXVII + XLII = ?

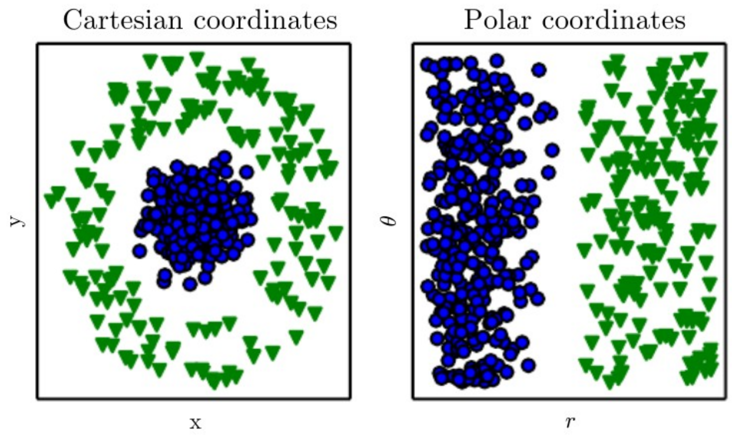
Or you perform faster on Arabic numerals?

37 + 42 = ?

“Many AI tasks can be solved by designing the right set of features to extract for that task”

source

Example of different representations



Which of the two can be correctly classified by a linear classifier?

Representation learning

Can you perform aritmetic on Roman numerals?

XXXVII + XLII = ?

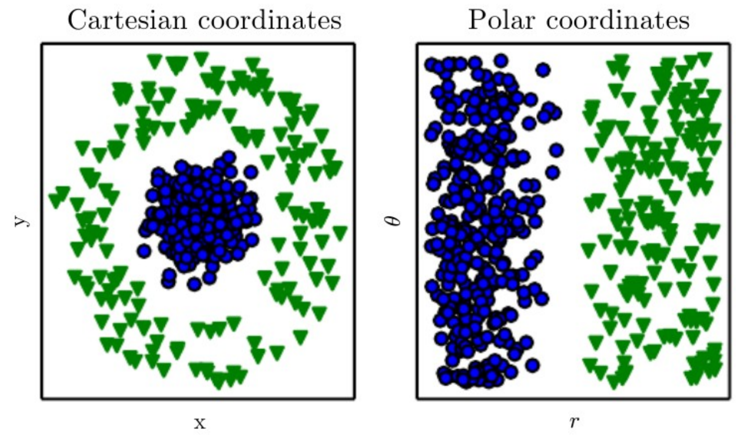
Or you perform faster on Arabic numerals?

37 + 42 = ?

“Many AI tasks can be solved by designing the right set of features to extract for that task”

source

Example of different representations



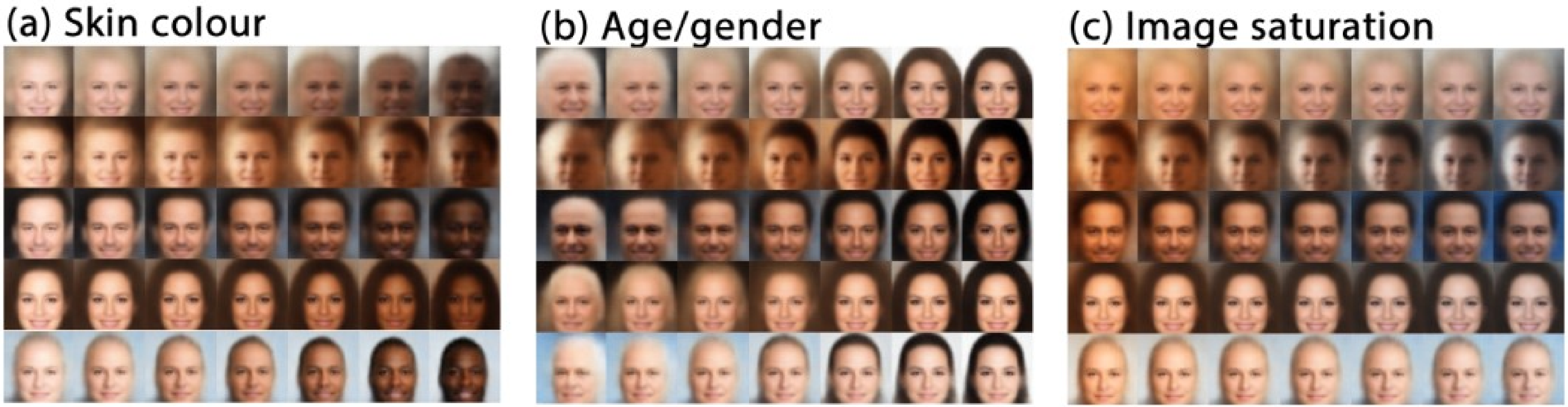
Which of the two can be correctly classified by a linear classifier?

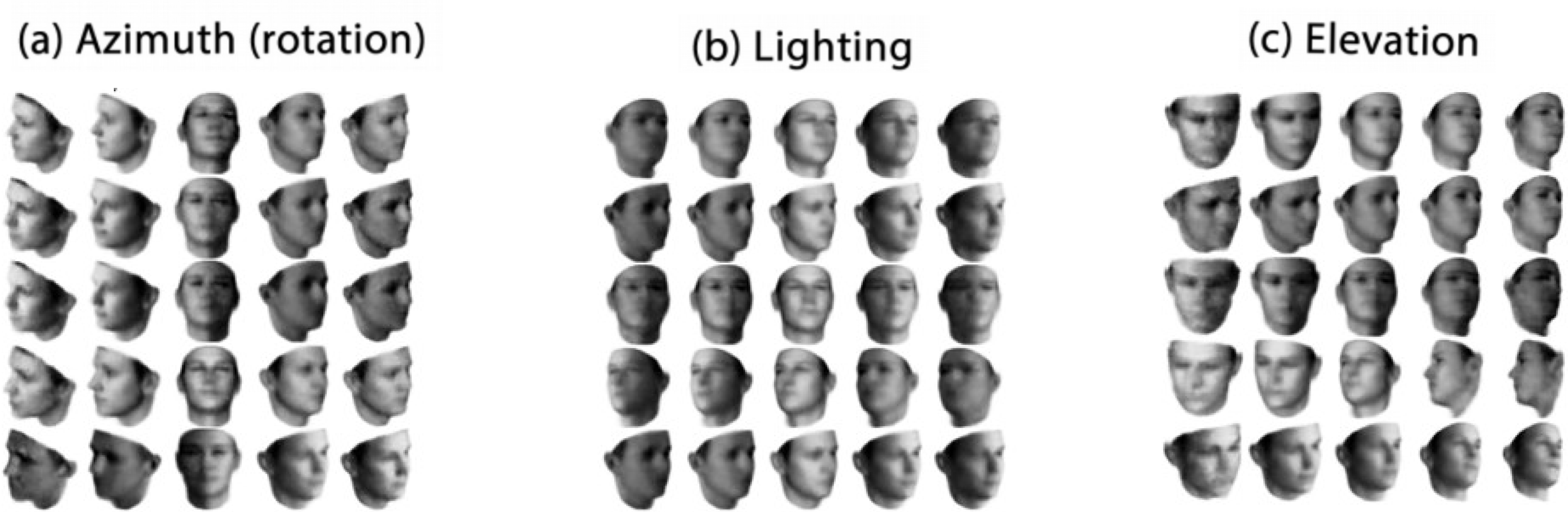
Good representation capture posterior belief about explanatory causes, disentangle these underlying vactors of variations.

source

Disentanglement

Disentanglement





Challenges

Unsupervised learning of a disentangled posterior distribution over the underlying generative factors of sensory data

Challenges

Unsupervised learning of a disentangled posterior distribution over the underlying generative factors of sensory data

Why?

Challenges

Unsupervised learning of a disentangled posterior distribution over the underlying generative factors of sensory data

Why?

Knowledge of one factor can generalize to novel configurations of other factors

Faster learning / learning from few examples

Challenges

Challenges

We may not have direct access to the generative factors

Challenges

We may not have direct access to the generative factors

source

Challenges

We may not have direct access to the generative factors

source

Little or no supervision for discovering the factors

How to compare models which perform disentanglement?

β-VAE

Learning an interpretable factorised representation of the independent data generative factors of the world without supervision

β-VAE

D = {X, V, W} = a set of images x and 2 generative factors

V = conditionally independent factors

W = conditionally dependent factors

β-VAE

D = {X, V, W} = a set of images x and 2 generative factors

V = conditionally independent factors

W = conditionally dependent factors

Assume images x are generated by $p(\mathbf{x}|\mathbf{v},\mathbf{w}) = \mathbf{Sim}(\mathbf{v},\mathbf{w})$

$\mathbf{Sim}(\mathbf{v},\mathbf{w})$ = true world simulator

β-VAE

D = {X, V, W} = a set of images x and 2 generative factors

V = conditionally independent factors

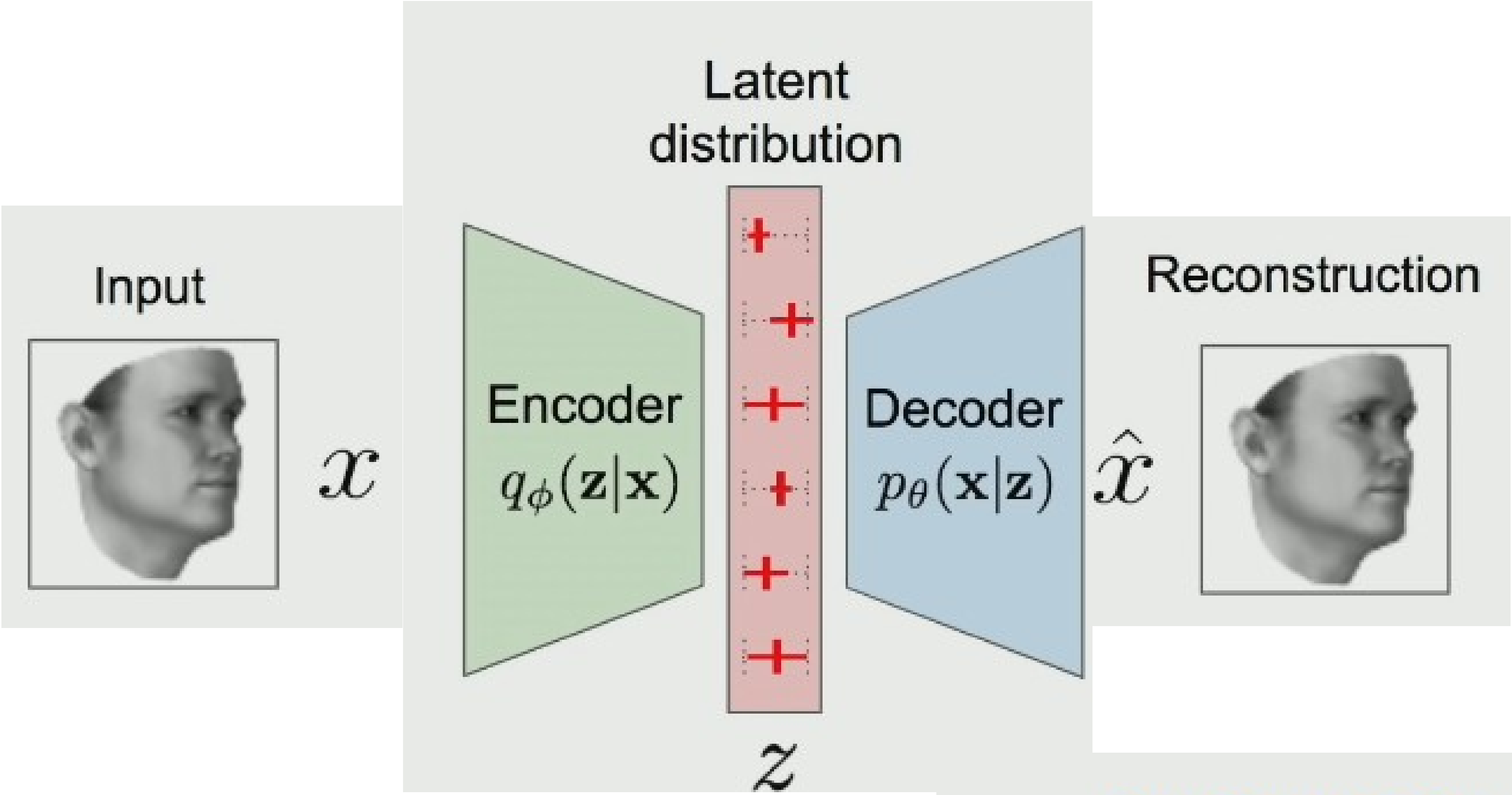
W = conditionally dependent factors

Assume images x are generated by p(**x**|**v**,**w**) = **Sim**(**v**,**w**)

Sim(**v**,**w**) = true world simulator

Initial objective: $\max_{\theta} \mathbb{E}_{p_\theta(\mathbf{z})}[p_\theta(\mathbf{x}|\mathbf{z})]$

β-VAE



$$\mathcal{L}_{\mathrm{VAE}}(\theta,\phi) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{Reconstruction cost}} - \overbrace{KL\left[q_\phi(\mathbf{z}|\mathbf{x}) \;||\; p(\mathbf{z})\right]}^{\text{Capacity constraint}}$$

Second objective: $q_\phi(\mathbf{z}|\mathbf{x})$ captures **v** in a disentangled manner

β-VAE

What about **w**?

They remain entangled in a separate subset of **z** that is not used in representing **v**

β-VAE

What about **w**?

They remain entangled in a separate subset of **z** that is not used in representing **v**

Introduce a constraint over $q_\phi(\mathbf{z}|\mathbf{x})$

Match it to a prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I)$

β-VAE

What about **w**?

They remain entangled in a separate subset of **z** that is not used in representing **v**

Introduce a constraint over $q_\phi(\mathbf{z}|\mathbf{x})$

Match it to a prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I)$

Constraint optimization problem

$$\max_{\phi,\theta} \mathbb{E}_{x\sim\mathbf{D}} \left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right]$$

$$\text{subject to } D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) < \epsilon$$

β-VAE

$$\max_{\phi,\theta} \mathbb{E}_{x\sim\mathbf{D}} \left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] \text{ subject to } D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) < \epsilon$$

β-VAE

$$\max_{\phi,\theta} \mathbb{E}_{x\sim\mathbf{D}} \left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] \text{ subject to } D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) < \epsilon$$

Re-write as:

$$\mathcal{F}(\theta,\phi,\beta;\mathbf{x},\mathbf{z}) \geq \mathcal{L}(\theta,\phi;\mathbf{x},\mathbf{z},\beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

β-VAE

$$\max_{\phi,\theta} \mathbb{E}_{x\sim\mathbf{D}} \left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] \text{ subject to } D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) < \epsilon$$

Re-write as:

$$\mathcal{F}(\theta,\phi,\beta;\mathbf{x},\mathbf{z}) \geq \mathcal{L}(\theta,\phi;\mathbf{x},\mathbf{z},\beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

$\mathcal{L}(\theta,\phi;\mathbf{x},\mathbf{z},\beta)$ = Lagrangian under KKT conditions

$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$ = reconstruction loss

$D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$ = regularization term

β = regularization coefficient

β-VAE

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

β
- controls the capacity of the latent
- puts independence pressure

β = 1 becomes VAE

β-VAE

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

β
- controls the capacity of the latent
- puts independence pressure

β = 1 becomes VAE

Disentanglement representation emerge when the right balance is found between information preservation and latent channel capacity restriction.

β-VAE

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

β controls the capacity of the latent

β puts independence pressure

β = 1 becomes VAE

Disentanglement representation emerge when the right balance is found between information preservation and latent channel capacity restriction.

Related to the information bottleneck principle:

$$\max[I(Z;Y) - \beta I(X;Z)]$$

New evaluation metric

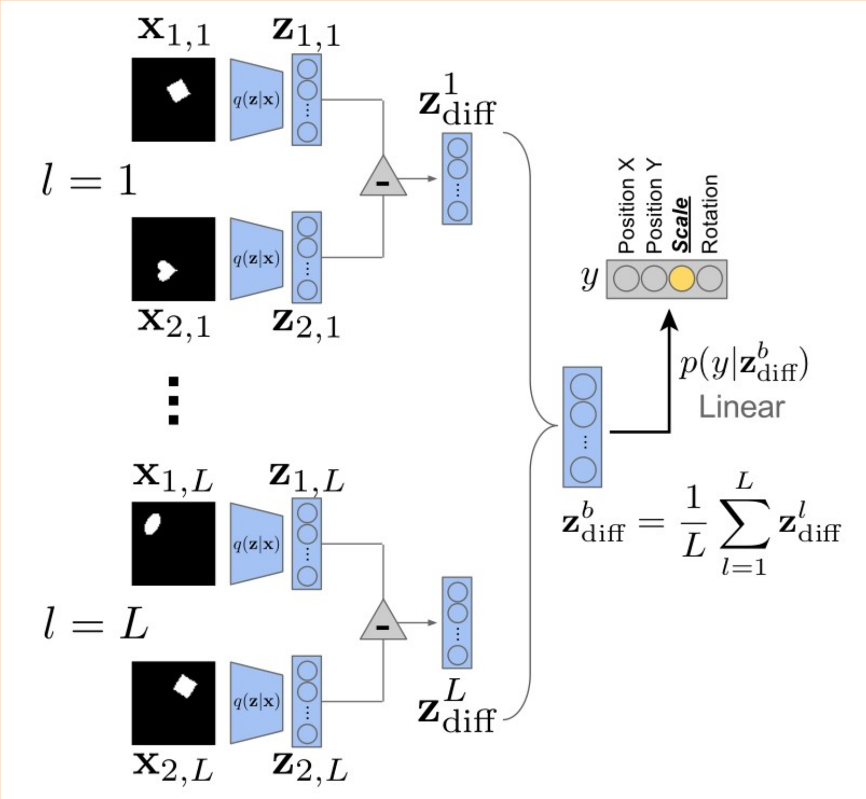
$\beta > 1$ leads to poorer reconstructions due to the loss of high freq details

cannot use log-likelihood of the data under the learnt model

New evaluation metric

β > 1 leads to poorer reconstructions due to the loss of high freq details
cannot use log-likelihood of the data under the learnt model

The new metric: **the accuracy of a low capacity classifier**

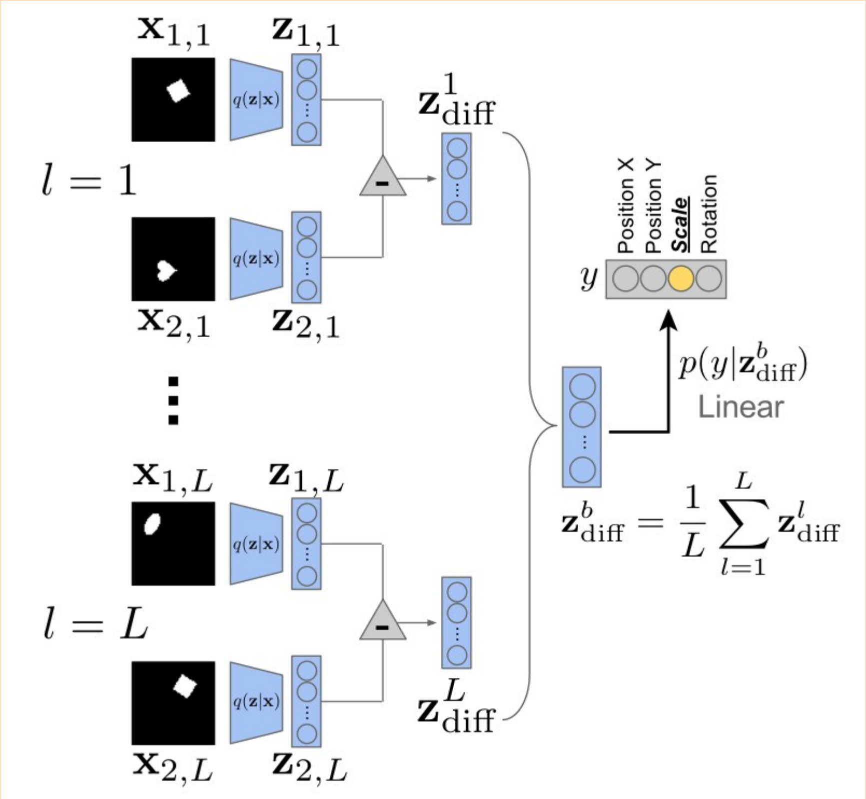


New evaluation metric

β > 1 leads to poorer reconstructions due to the loss of high freq details

cannot use log-likelihood of the data under the learnt model

The new metric: **the accuracy of a low capacity classifier**



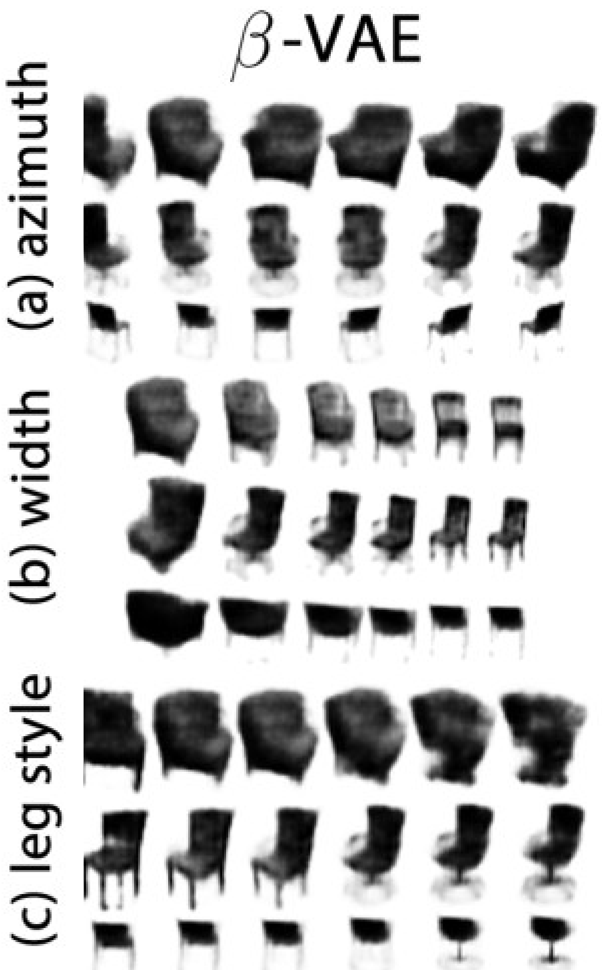
Cons: needs access to the world simulator **Sim**(**v**,**w**)

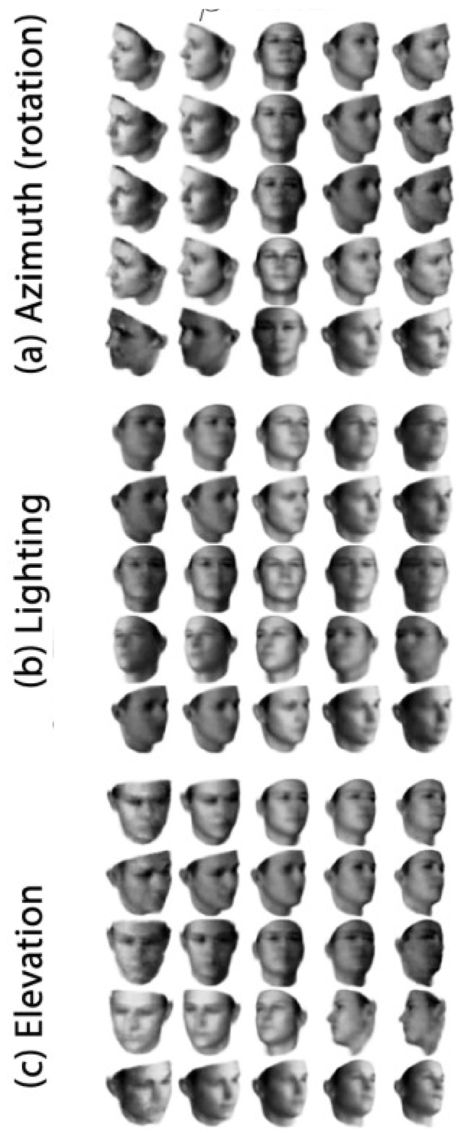
Results

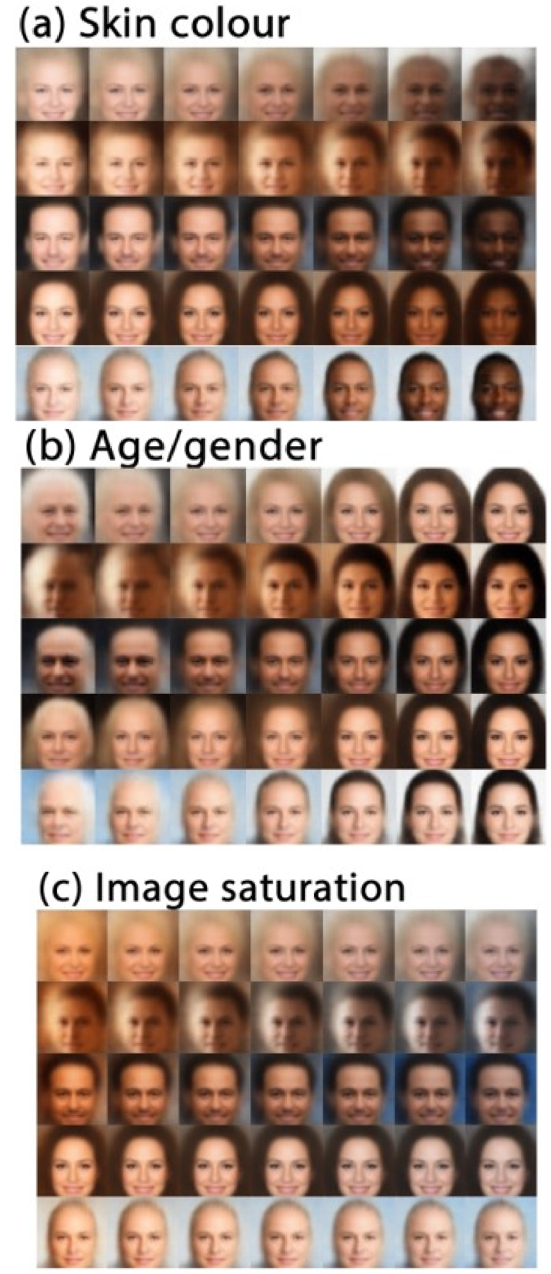
Qualitative

Results

Qualitative







Results

Quantitative

Results

Quantitative

Model	Disentanglement metric score
Ground truth	*100%*
Raw pixels	$45.75 \pm 0.8\%$
PCA	$84.9 \pm 0.4\%$
ICA	$42.03 \pm 10.6\%$
DC-IGN	$\mathbf{99.3} \pm \mathbf{0.1}\%$
InfoGAN	$73.5 \pm 0.9\%$
VAE untrained	$44.14 \pm 2.5\%$
VAE	$61.58 \pm 0.5\%$
β-VAE	$\mathbf{99.23} \pm \mathbf{0.1}\%$

On a synthetic dataset of 2D shapes (heart, oval, square)

With indep generative factors: positionX, positionY, scale and rotation