Deep Unsupervised Learning

Simon Jenni
(slides by Paolo Favaro)

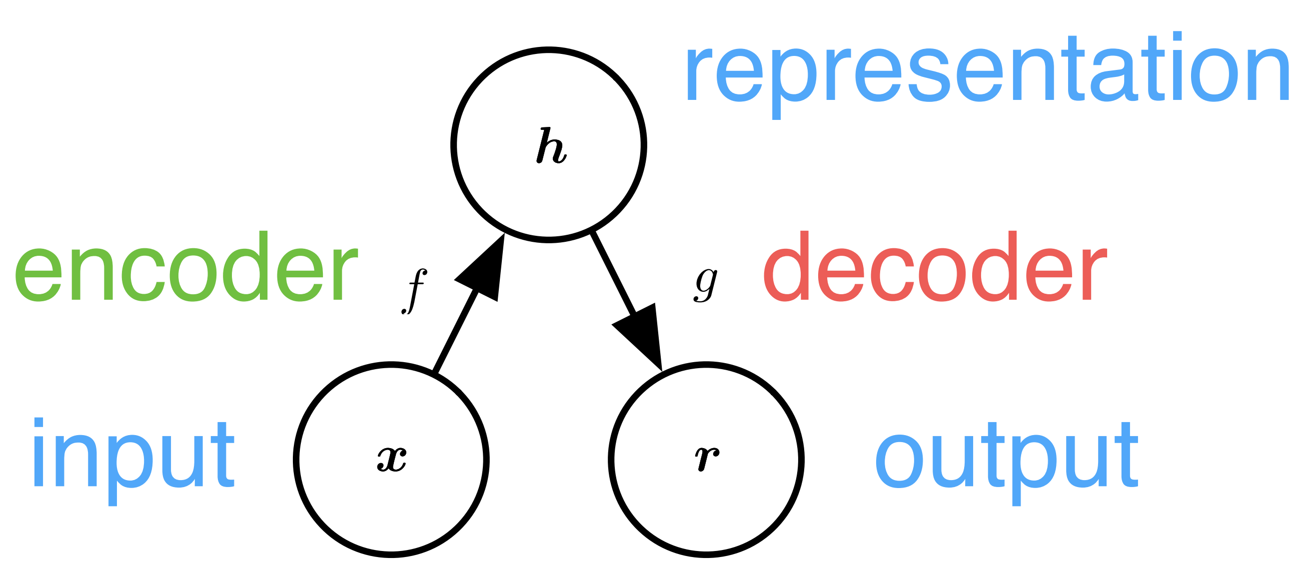
UNIVERSITÄT BERN

Contents

- Autoencoders
- Variational autoencoders
- Based on **Chapter 14** of Deep Learning by Goodfellow, Bengio, Courville

Autoencoders

- A network that replicates the input
- Internally it builds a **representation** of the input (e.g., as a vector)
- The network before the internal representation is the **encoder** and the network following it is the **decoder**



Avoiding the Trivial Identity

- **Undercomplete** autoencoders
  - **h** has lower dimension than **x**
  - **f** or **g** has low capacity (e.g., a shallow **g**)
  - Discard information up to **h**
- **Overcomplete** autoencoders
  - **h** has higher dimension than **x**
  - Must be regularized

Undercomplete Autoencoders

- A way to obtain a useful representation ***h*** is to constrain it to have a smaller dimension than ***x***

- In this case the AE is called **undercomplete**

- We force the AE to focus on the most important attributes of the training data

- A linear decoder and MSE loss L learns to map the representation to the same subspace as PCA

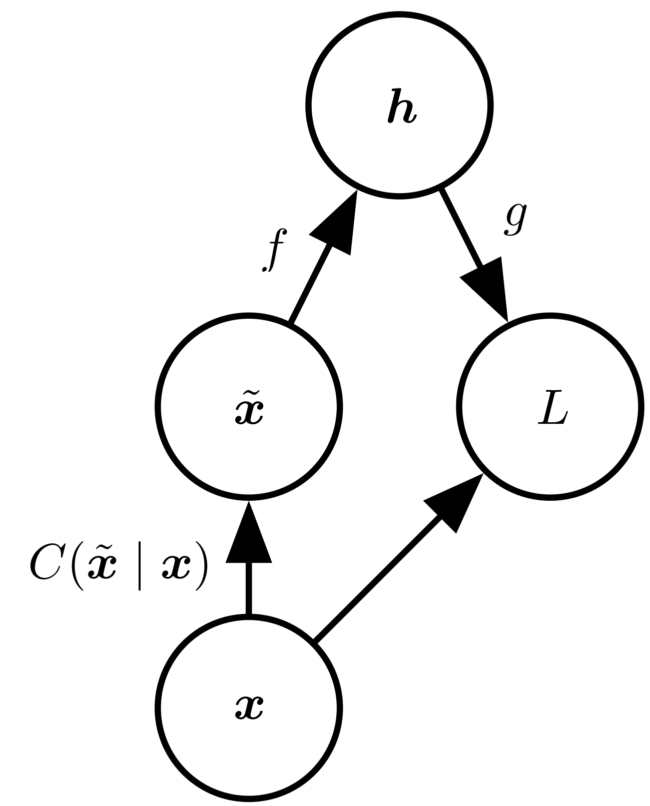
Overcomplete Autoencoders

- Choosing the representation size and the capacity of the encoder and decoder depends on the complexity of the data distribution

- A useful strategy to avoid trivial mappings is to introduce **regularization** (e.g., sparsity of the representation, smoothness of the representation, robustness to noise or missing data)

Denoising Autoencoders

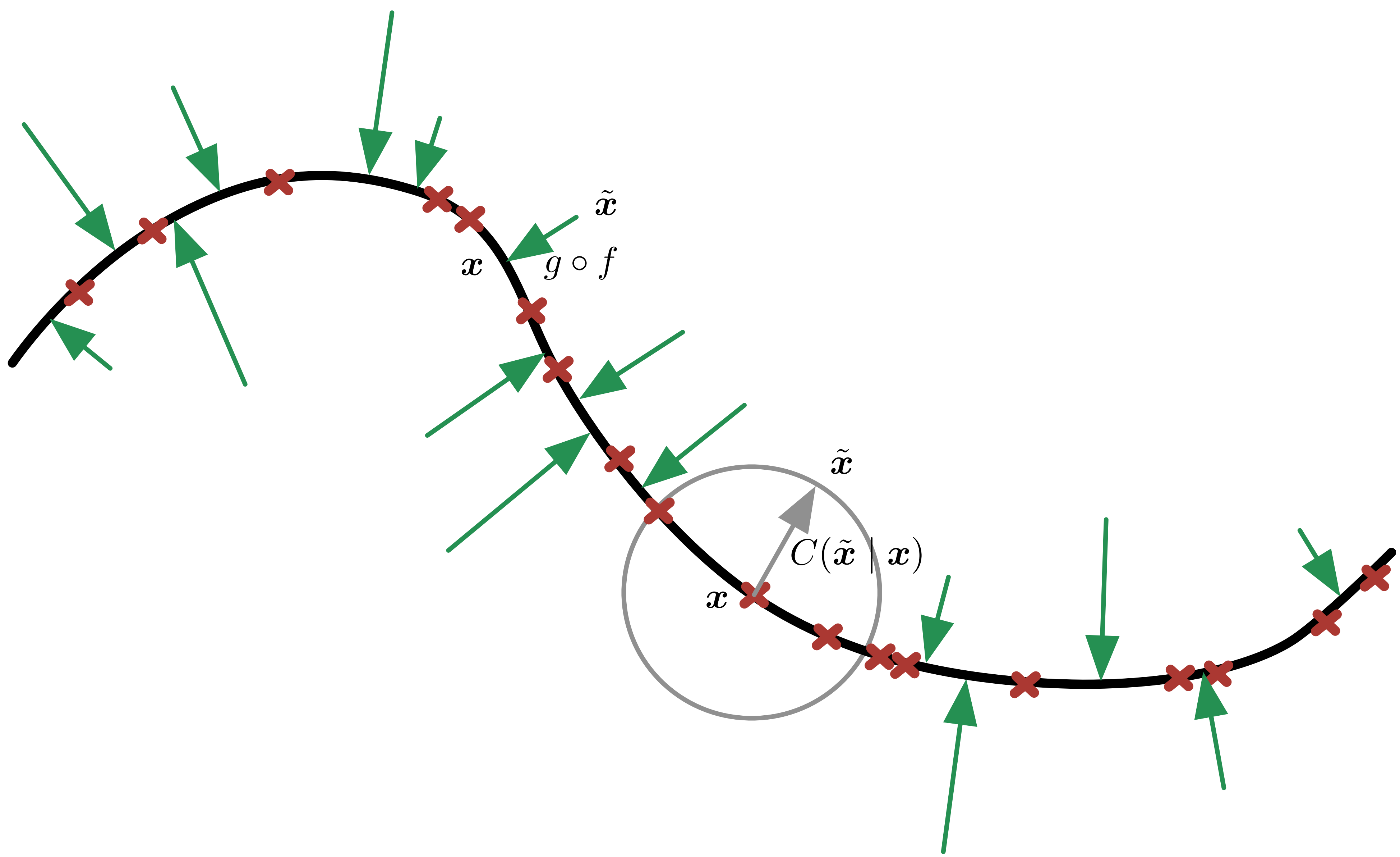
- A **denoising autoencoder** (DAE) maps noisy data to the original uncorrupted data



- Let $C(\tilde{x}|x)$ be a corruption process
- Training set $(\tilde{x}, x)$
- Optimize

$$-E_{x\sim\hat{p}_{\text{data}}} E_{\tilde{x}\sim C(\tilde{x}|x)} \log p_{\text{decoder}}(x|h = f(\tilde{x}))$$

Denoising Autoencoders



Applications

- Dimensionality reduction (representation learning)
  - More effective than PCA
  - Low-dim representations useful in other tasks (e.g., classification)
- Information retrieval (matching query to entries in a database)
  - More efficient and accurate search

Generative Models

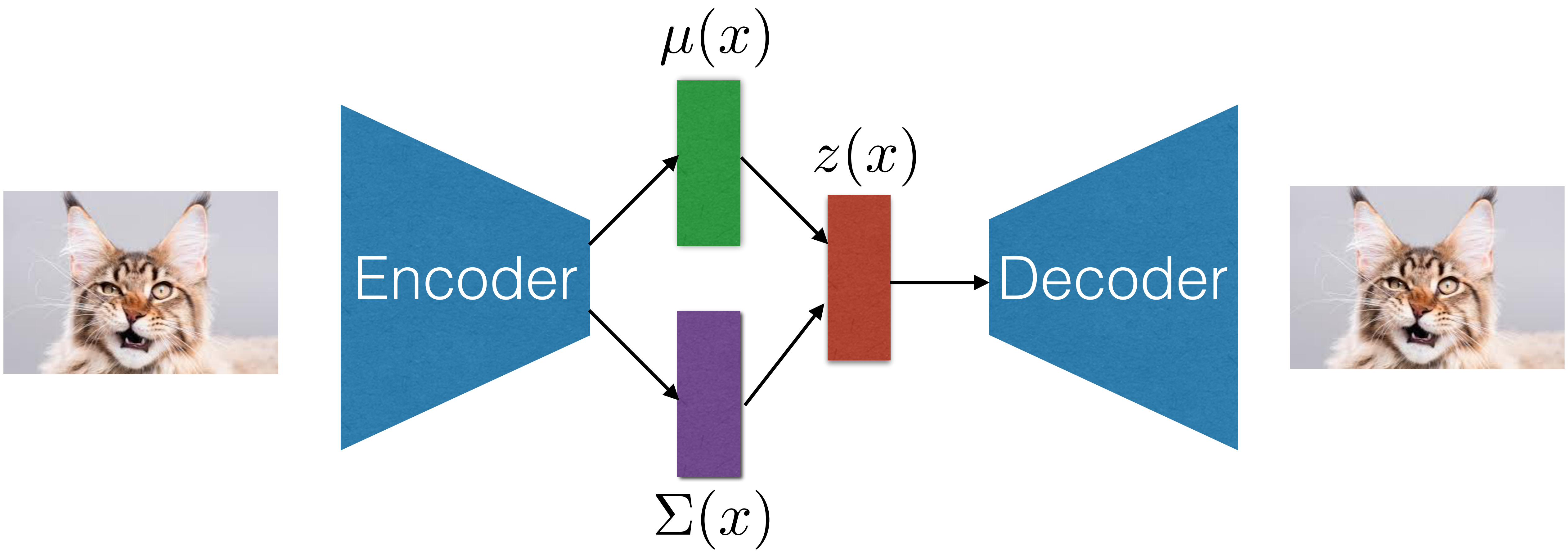
- Autoencoders can be generalized to generative models
- An important example is the **variational autoencoder**
    - Encode the moments of a Gaussian distribution

$$f(x) = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$$

    - Sample the distribution to generate data

$$\epsilon \sim \mathcal{N}(0, I_d) \qquad g(\mu + \sigma\epsilon)$$

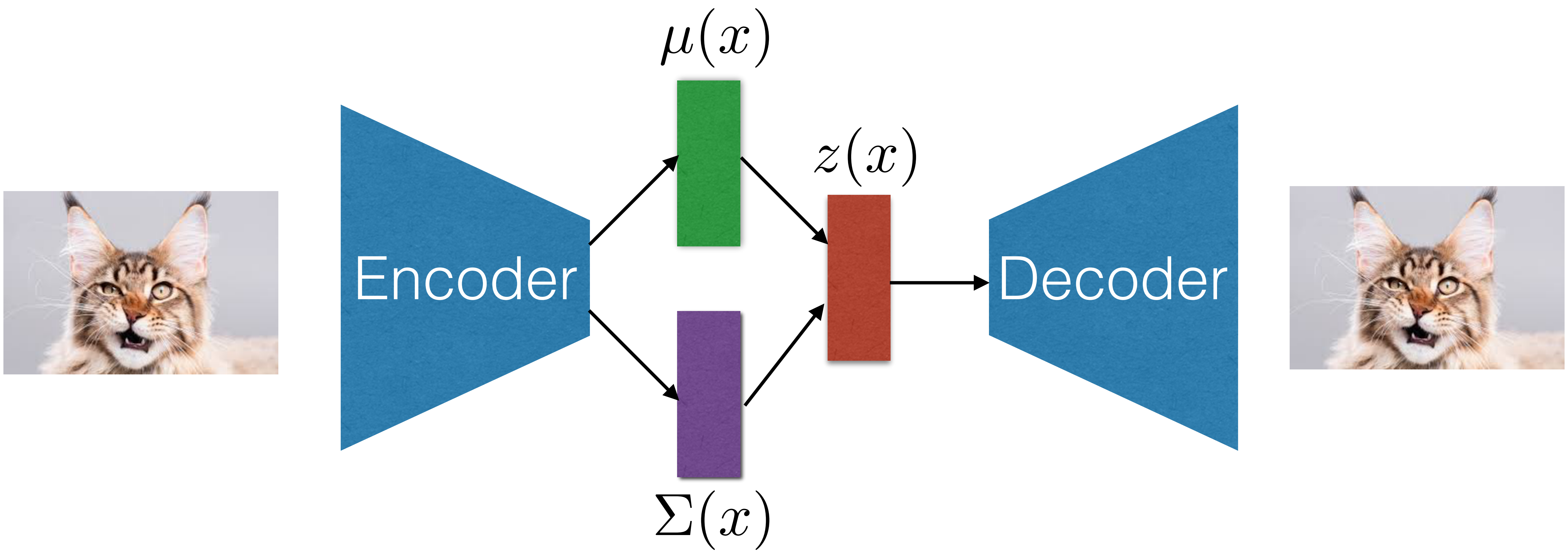
Variational Autoencoders



Reparametrization-Trick:

$$z(x) = \mu(x) + \Sigma(x)^{1/2}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

Variational Autoencoders



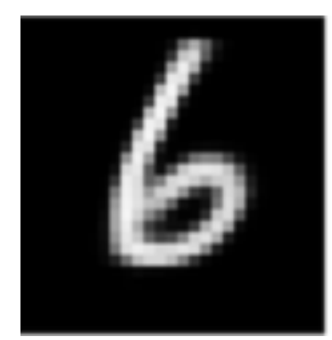
Cost function:

$$E_{\mathbf{z}(\mathbf{x})\sim Q}[\log p(\mathbf{x}|\mathbf{z})] - D_{KL}(Q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$$

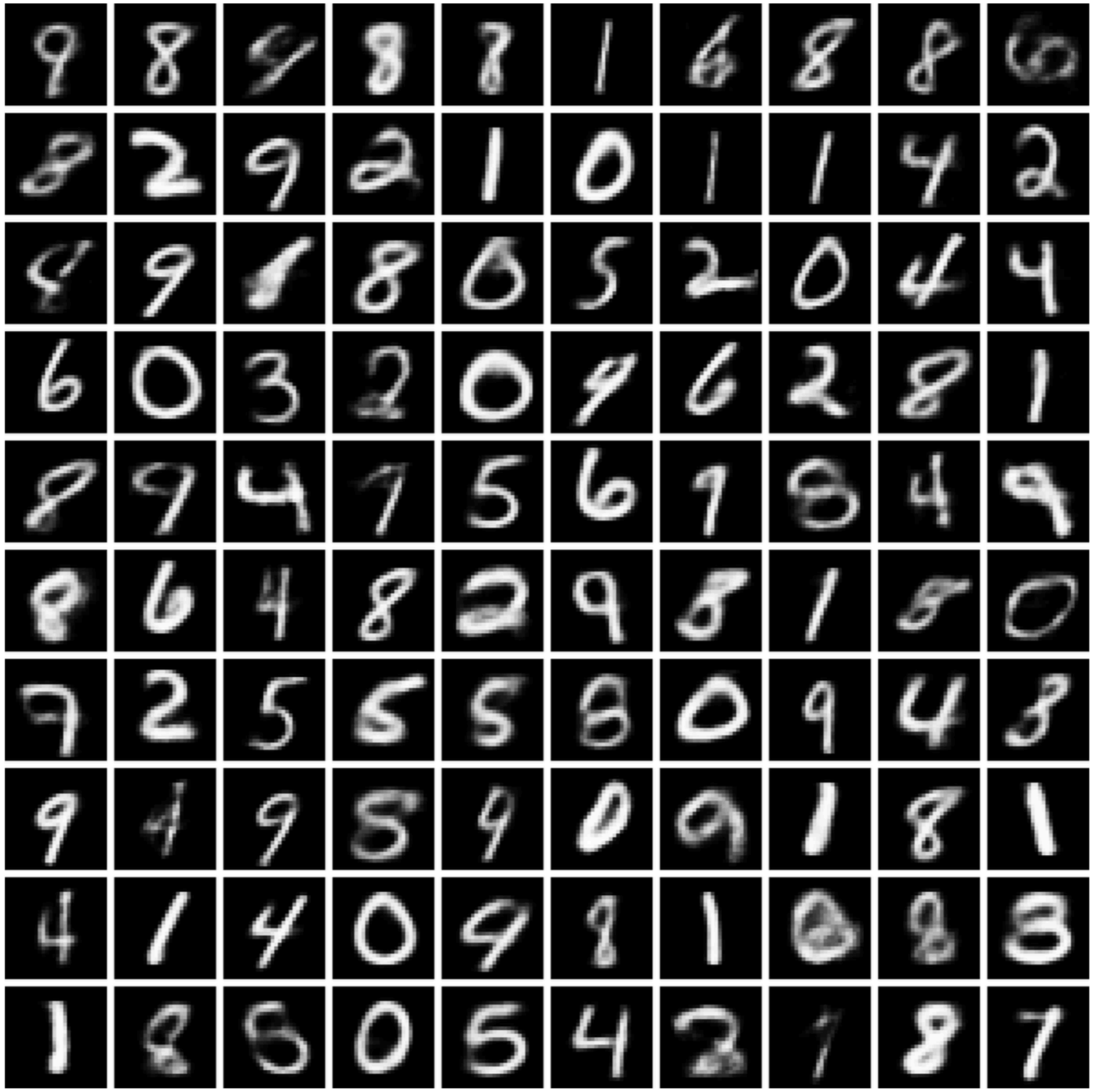
Using the Model

- After training we obtain the parameters θ and we use the decoder after sampling **z** from the Normal distribution

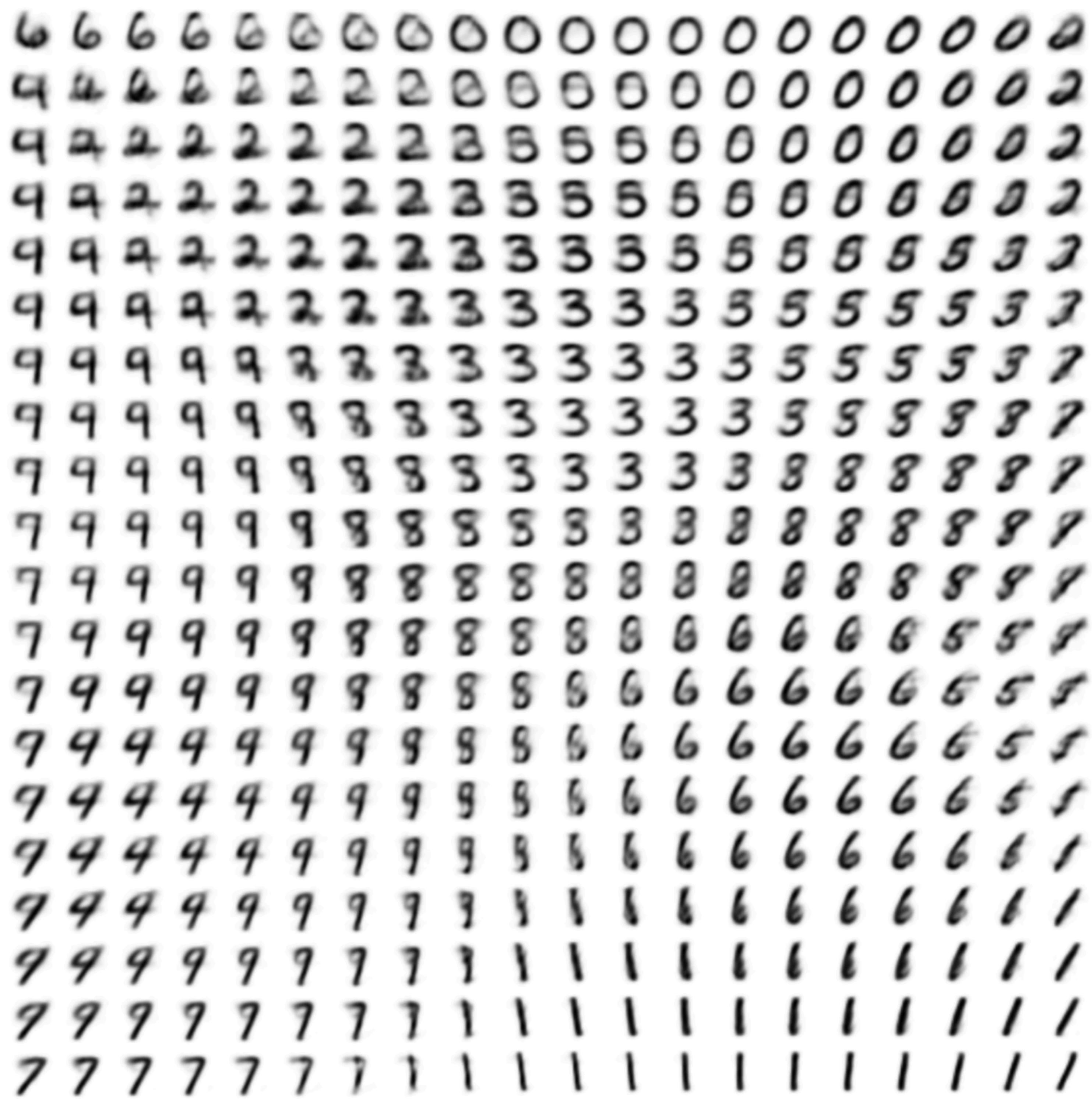
$$\mathbf{z} \sim \mathcal{N}(0, I_d) \longrightarrow f(\mathbf{z};\theta) \longrightarrow$$



VAE on MNIST



VAE on MNIST



Limitations

- VAEs produce samples that typically have a lower quality than those of other generative models
- A possible limitation is the Gaussian assumption for $p(\mathbf{x}|\mathbf{z})$
- Indeed it is not true that real images can be obtained by adding Gaussian noise to a real image
- This might the reason why generated samples tend to be blurry

UNIVERSITÄT
BERN

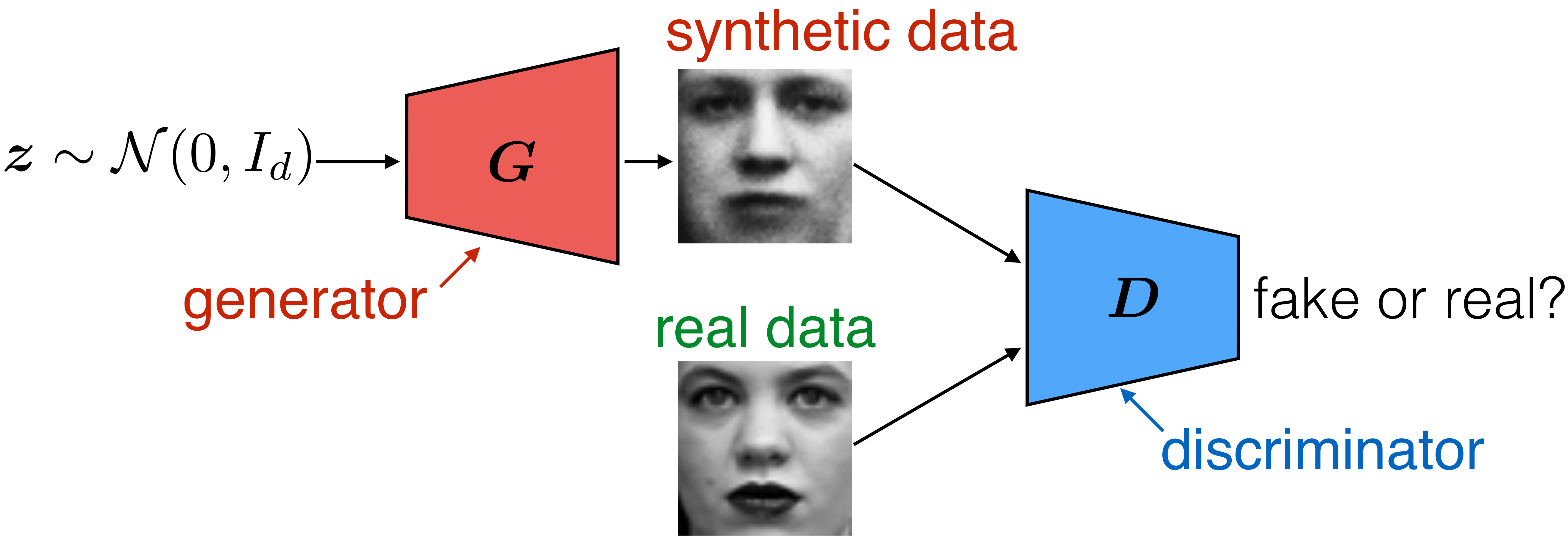
Generative Adversarial Networks

Simon Jenni
(slides by Paolo Favaro)

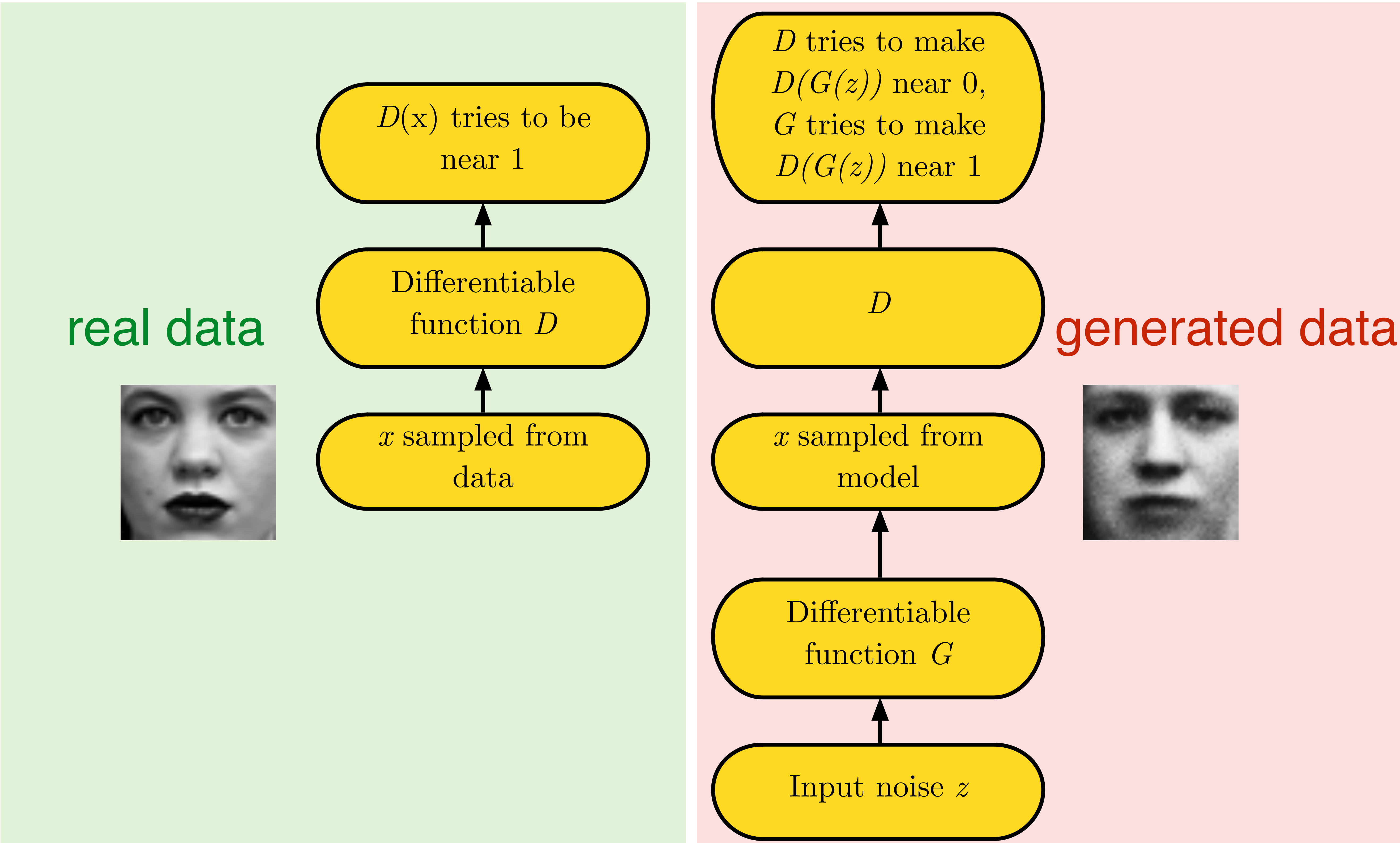
Contents

- Generative Adversarial Networks
  - Generative modeling, Principles of Adversarial Learning, Issues: Vanishing gradient and Mode Collapse
- Based on the tutorial paper
  - NIPS 2016 Tutorial: GAN by Goodfellow, 2016
- Other resources
  - How GAN and its Variants Work: An Overview of GAN by Hong, Hwang, You and Yoon, 2018

The GAN Framework

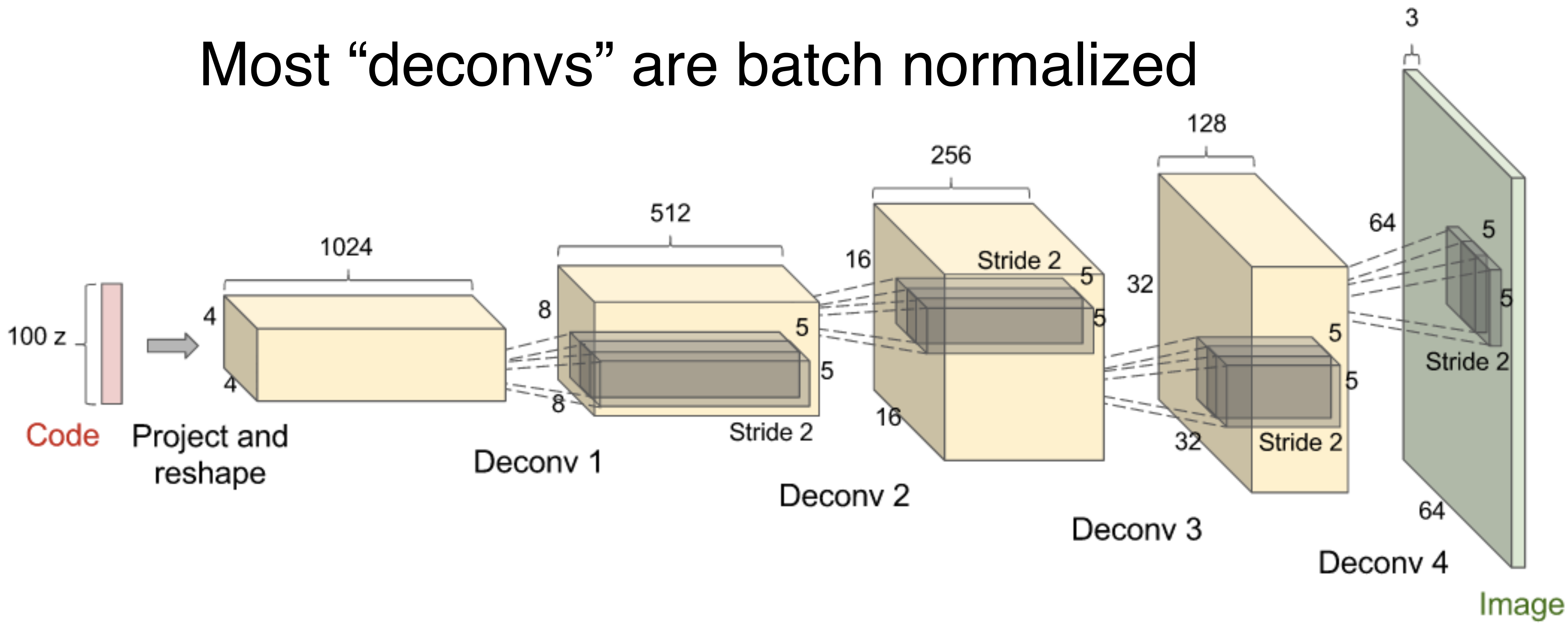


The GAN Framework



Example of a Generator: DCGAN Architecture

Most “deconvs” are batch normalized



(Radford et al 2015)

Loss Function

$$\min_G \max_D \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}} \log D(\boldsymbol{x}) + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}} \log\left(1 - D(G(\boldsymbol{z}))\right)$$

- G tries to minimize the cost
  - The minimum is achieved when D(**x**) = 0 and D(G(**z**)) = 1
    That is, when D thinks that **x** is fake and G(**z**) is real
- D tries to maximize the cost
  - The maximum is achieved when D(**x**) = 1 and D(G(**z**)) = 0
    That is, when D thinks that **x** is real and G(**z**) is fake

Training

- Use SGD-like algorithm of choice (e.g., Adam) on two minibatches simultaneously:
    - A minibatch of training examples
    - A minibatch of generated samples
- Optional: run *k* steps of one player (Discriminator) for every step of the other player (Generator)

Heuristic Cost

- To avoid the vanishing gradient of G optimize instead

$$J^{(D)} = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}} \log D(\boldsymbol{x}) - \frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log\left(1 - D\left(G(\boldsymbol{z})\right)\right)$$
$$J^{(G)} = -\frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log D\left(G(\boldsymbol{z})\right)$$

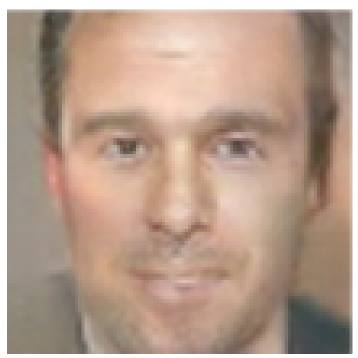
obtained by flipping the two probabilities in the loss

- This moves the cost in G also to the high gradient region (when D converges)

Representation Linearity

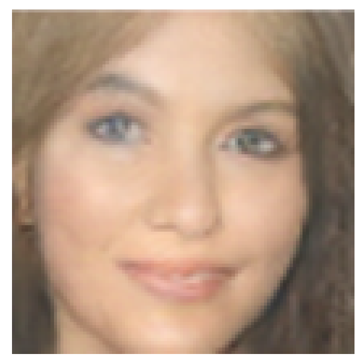
Man with glasses

-



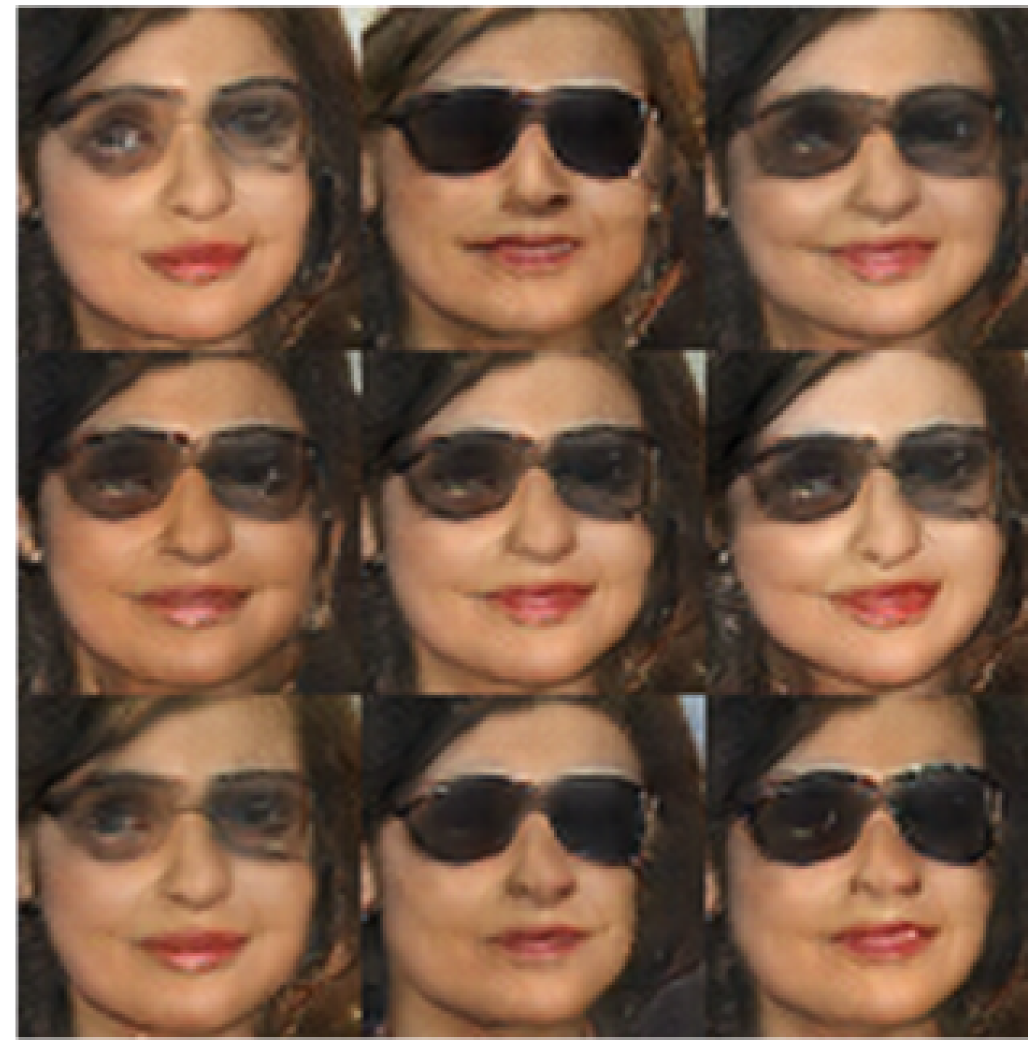
Man

+



Woman

=



Woman with Glasses

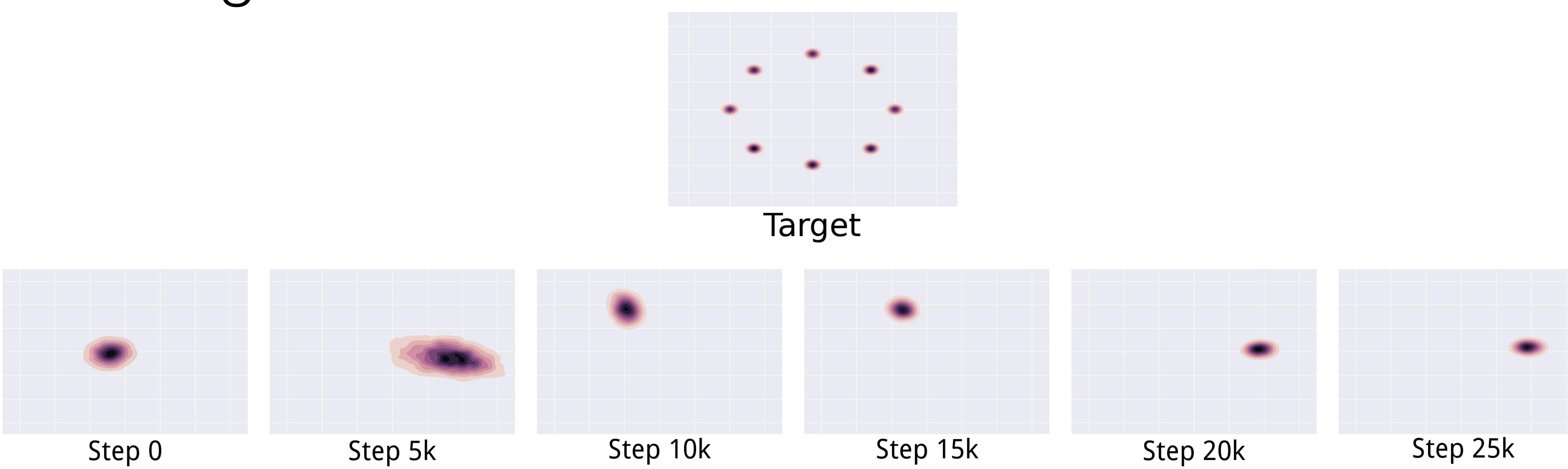
(Radford et al, 2015)

Tricks to Train GANs

- Use labels
- Use Spectral Normalization
- Use existing working network architectures
- Use the Adam optimizer

Mode Collapse

- Although our target is multimodal, the generator converges to only one or only some of the modes
- The generator learns to map multiple z to the same image x

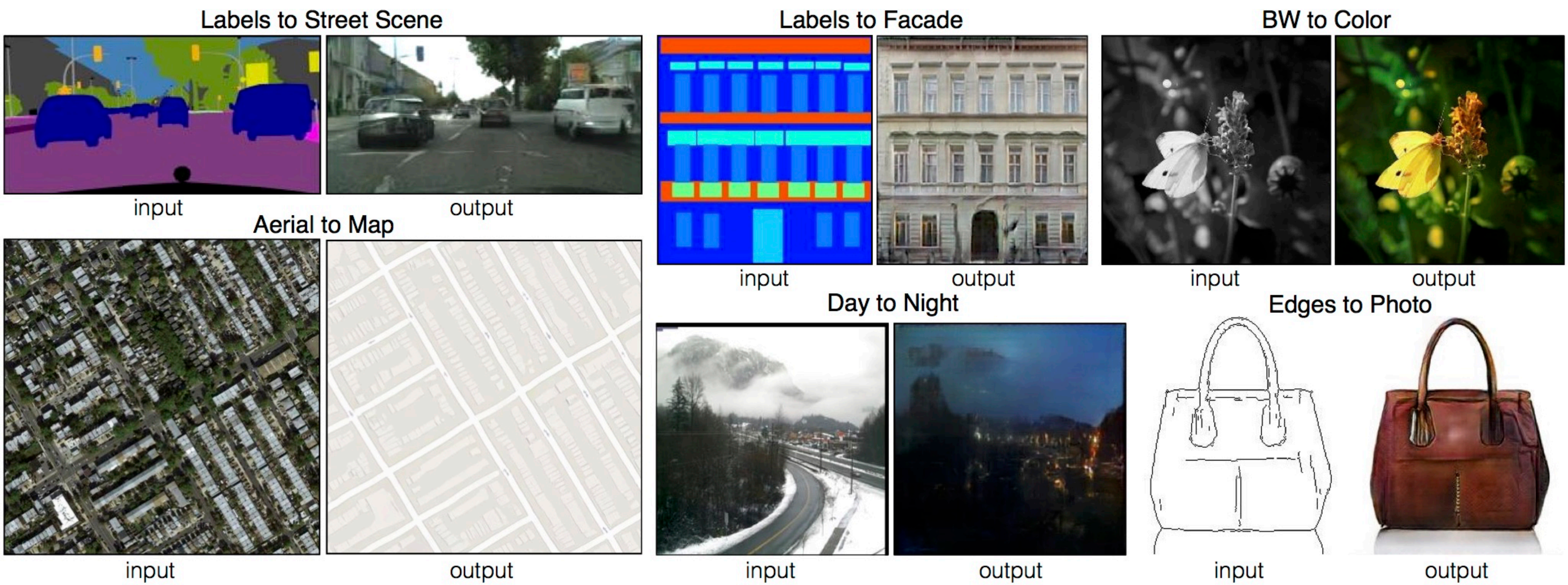


(Metz et al 2016)

Learning the Loss

- We might not know the best loss function for a task beforehand
- One solution is then to **learn** it
- **Generative adversarial networks** (GAN) provide one such framework

Learning the Loss



Current SOTA in GANs

(Karras et al, 2018)

Current SOTA in GANs

(Brock et al, 2018)

UNIVERSITÄT BERN

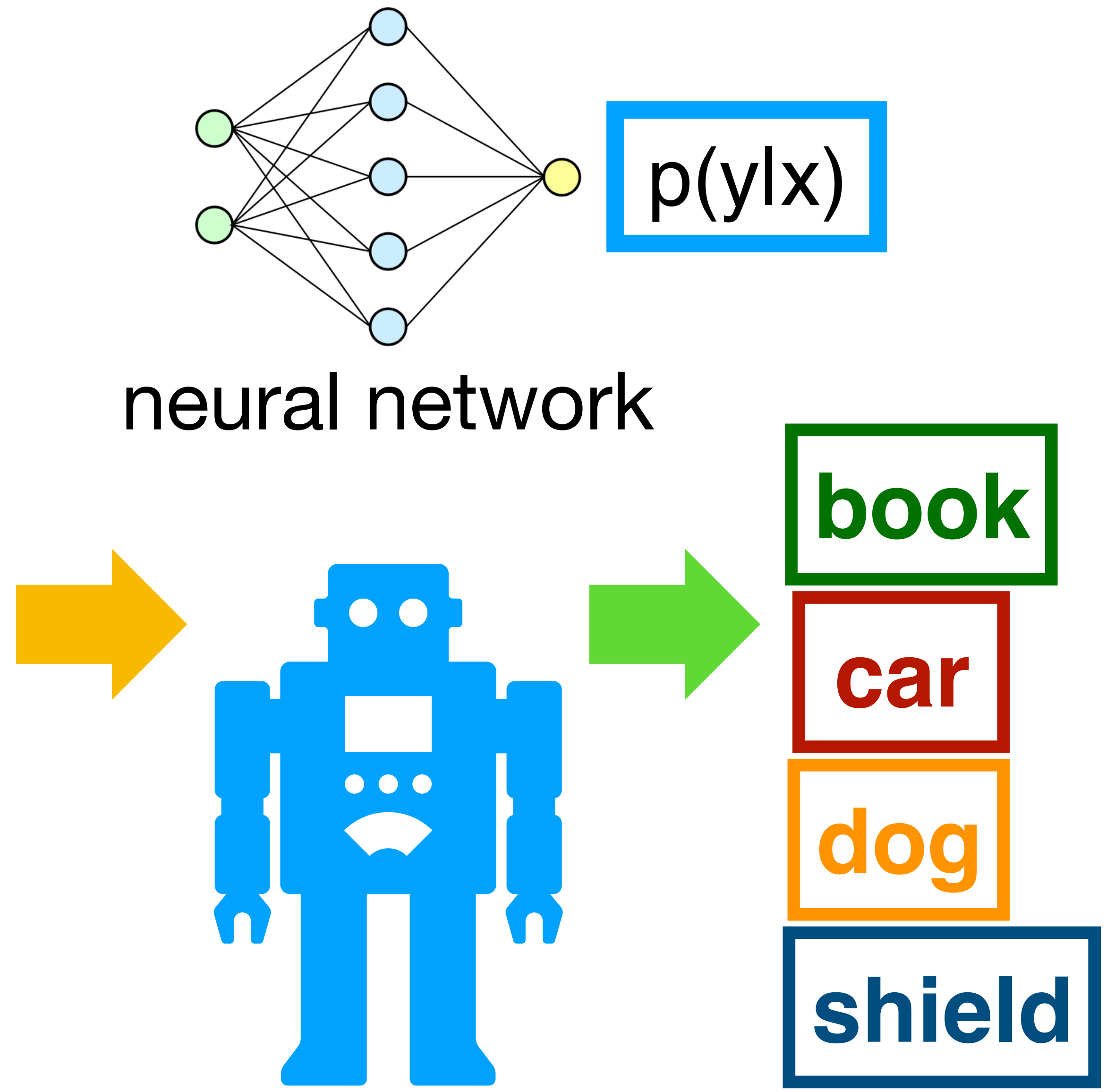
Self-Supervised Learning

Simon Jenni
(slides by Paolo Favaro)

Contents

- Self-supervised learning, principles, overview of the literature, recent developments, transfer learning
- Based on recent works in the literature (citations provided throughout the slides)

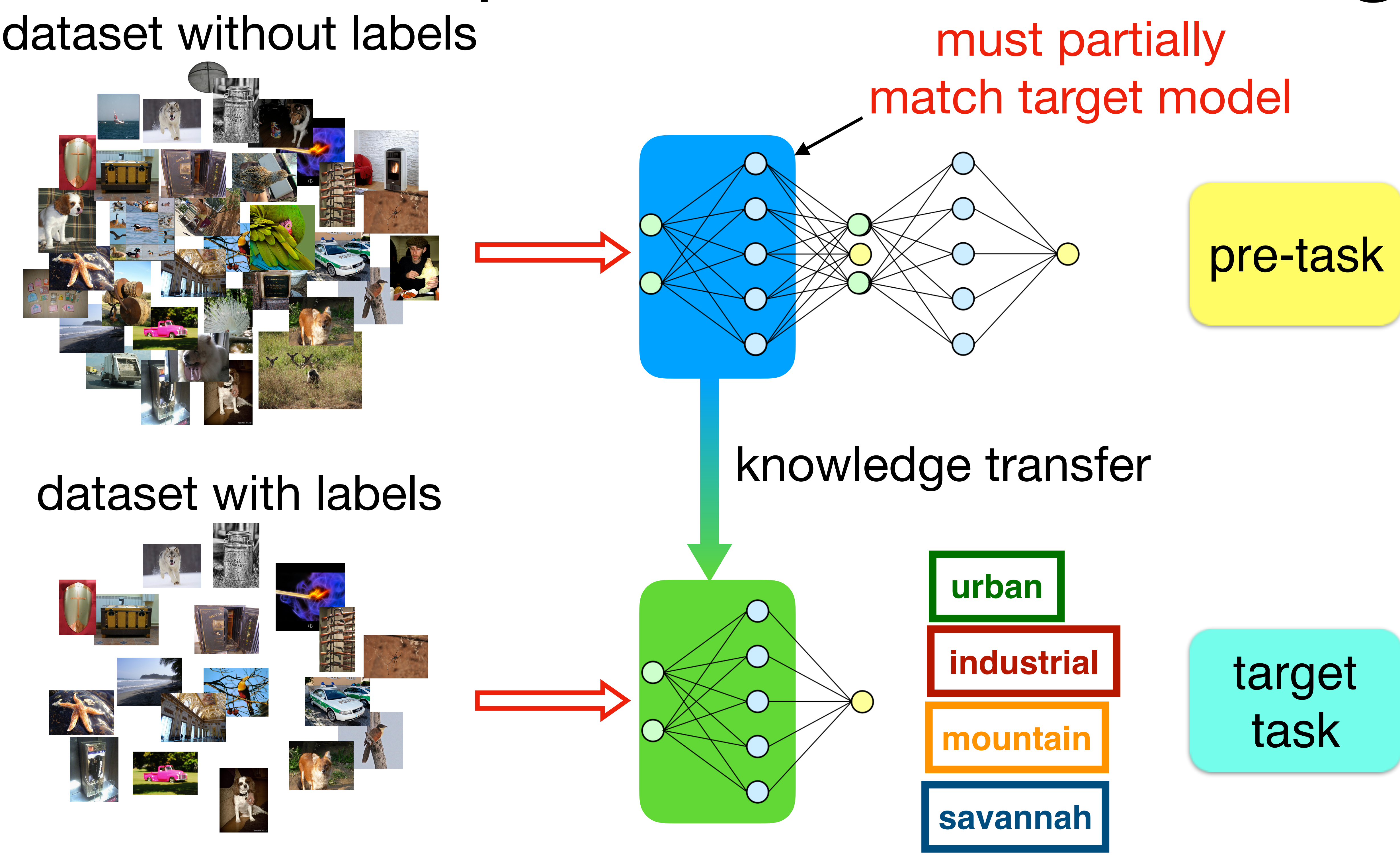
Supervised Learning



data x

attributes y

Self-Supervised Learning



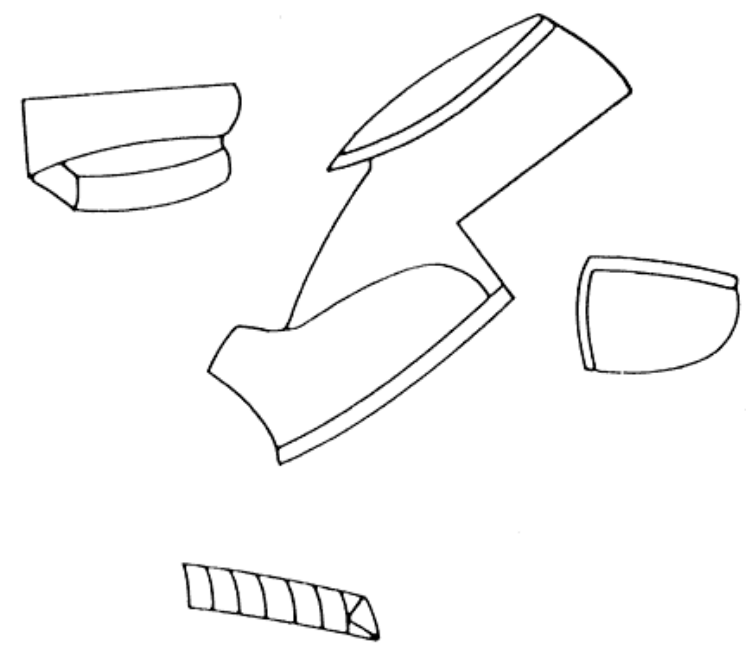
Self-Supervised Learning

- **Example #1: A Puzzle (hide the original spatial placement)**

- **Task**: rearrange parts to form a familiar object

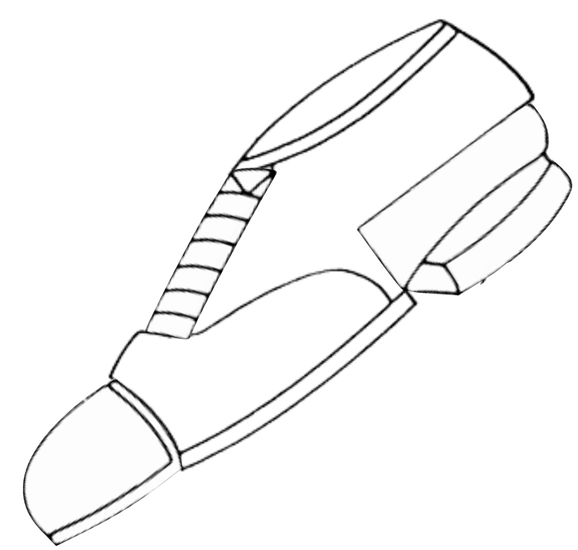
- **No additional information** is made available to us in addition to the photo

- What knowledge do we need to be able to solve the puzzle?



Self-Supervised Learning

- **Example #1: A Puzzle (hide the original spatial placement)**
- **Task**: rearrange parts to form a familiar object
- **No additional information** is made available to us in addition to the photo
- What knowledge do we need to be able to solve the puzzle?
- We need to know **how objects are made**

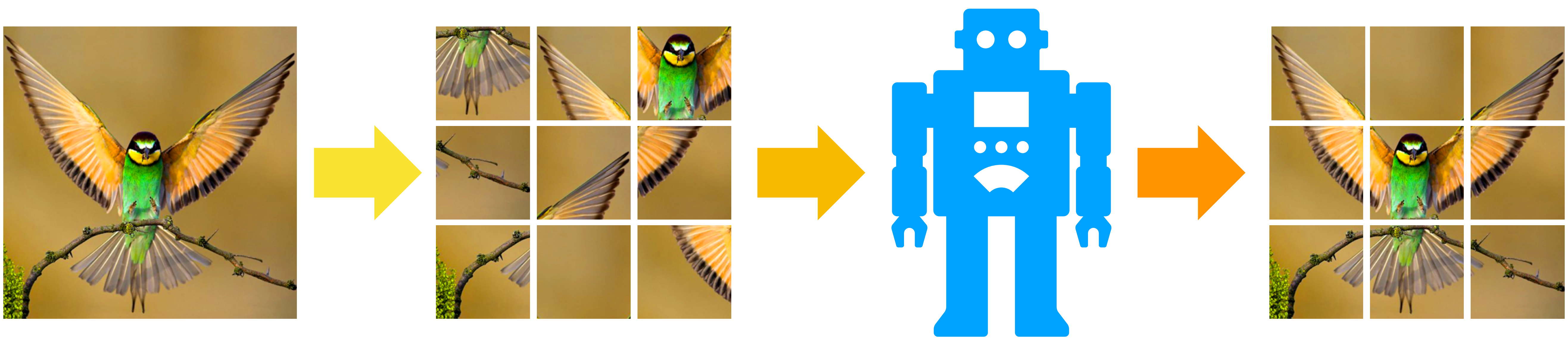


Self-Supervision by Hiding Data

- **Example #1**

- Predict the hidden part of the data (the ordering of the puzzle tiles)

$$x_i = \begin{bmatrix} r \\ g \\ b \\ u \\ v \end{bmatrix}$$



1. image **2. puzzle** **3. learning** **4. ordered**

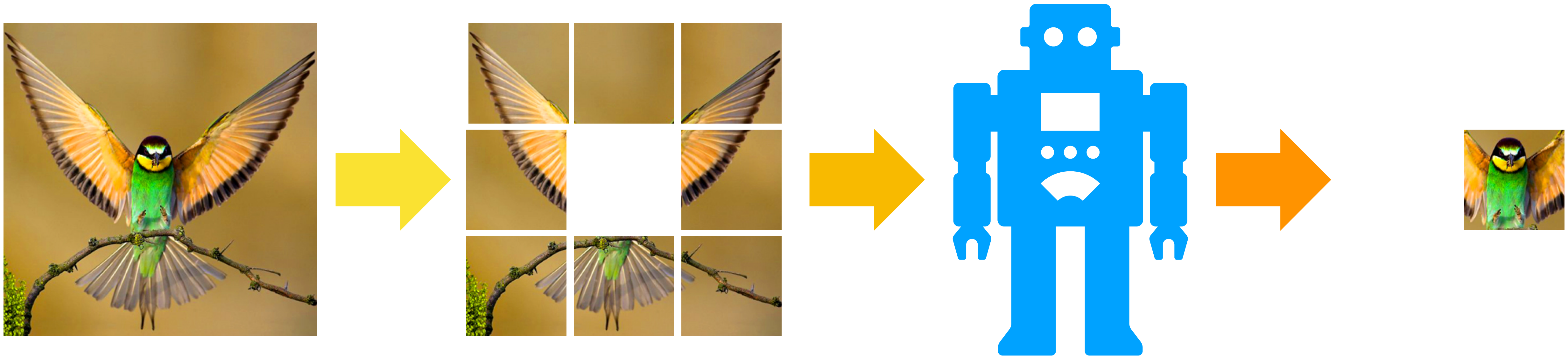
Unsupervised learning of visual representations by solving jigsaw puzzles
M. Noroozi and P. Favaro, ECCV 2016

Self-Supervision by Hiding Data

$$x_i = \begin{bmatrix} r \\ g \\ b \\ u \\ v \end{bmatrix}$$

- **Example #2**

- Predict the hidden part of the data (an image tile)



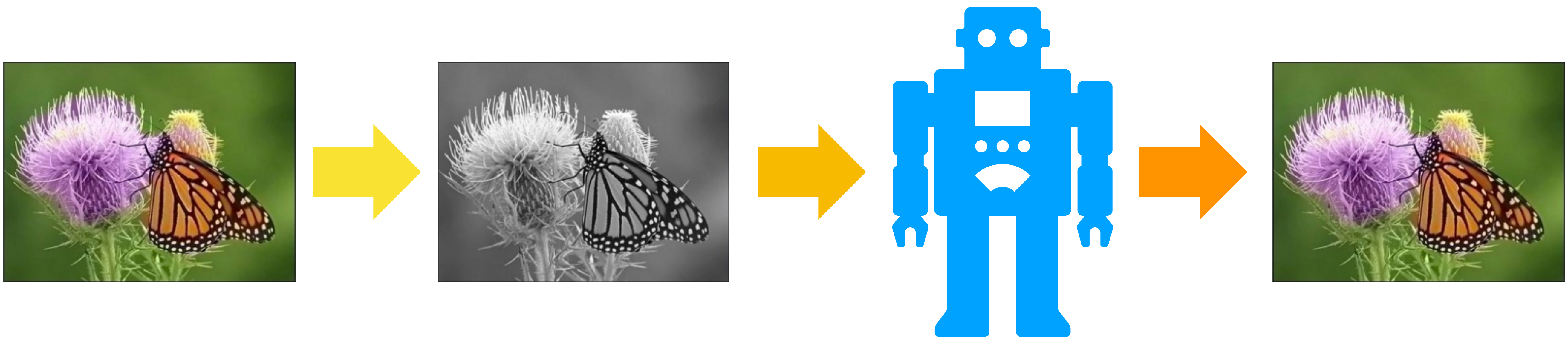
1. image **2. context** **3. learning** **4. missing tile**

Context encoders: Feature learning by inpainting
D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros, CVPR 2016

Self-Supervision by Hiding Data

- **Example #3**

- Predict the hidden part of the data (colors)



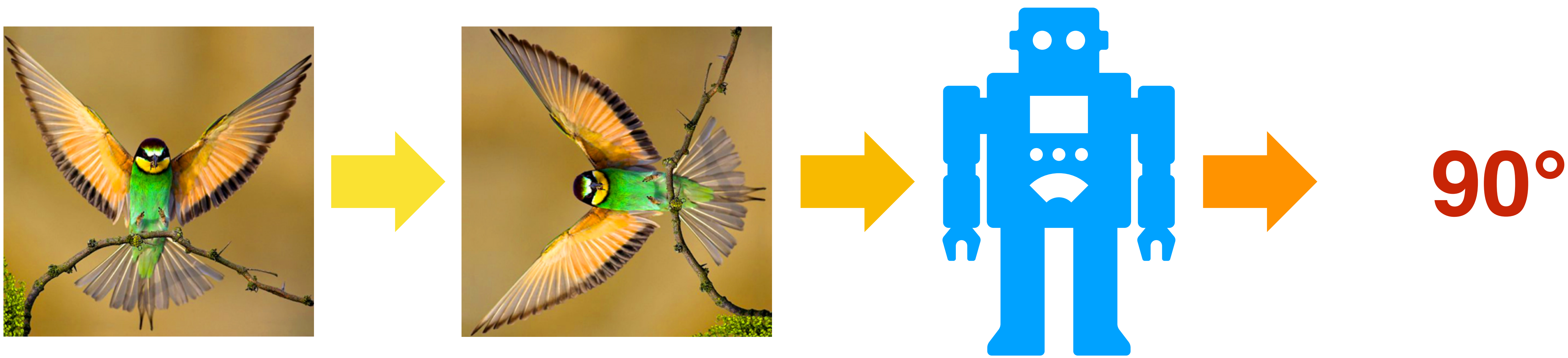
1. image **2. de-color** **3. learning** **4. color**

Learning representations for automatic colorization
G. Larsson, M. Maire, and G. Shakhnarovich, ECCV 2016

Self-Supervision by Hiding Data

- **Example #4**

- Predict the hidden part of the data (orientation)



1. image **2. rotate randomly** **3. learning** **4. color**

Learning representations for automatic colorization
G. Larsson, M. Maire, and G. Shakhnarovich, ECCV 2016

Deep Recurrent Neural Networks

Simon Jenni
(slides by Paolo Favaro)

Sequence Modeling

- While convolutional neural networks are used for data on a grid, **recurrent neural networks (RNN)** are used for **sequential** data

- Based on parameter sharing (across time)

Sequence Modeling

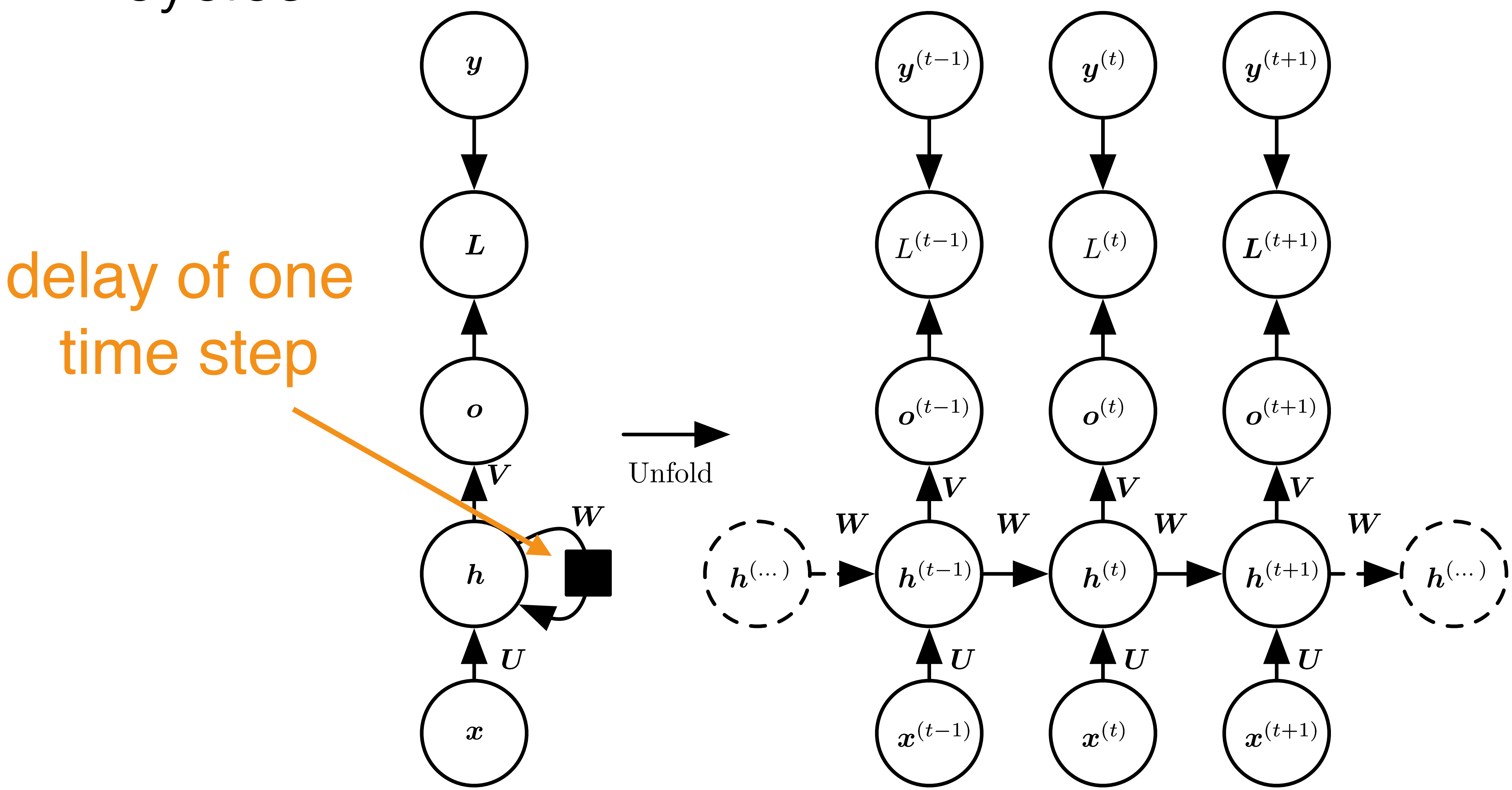
- RNNs define **dynamical systems** described by

$$h^{(t)} = f\left(h^{(t-1)}, x^{(t)}; \theta\right)$$

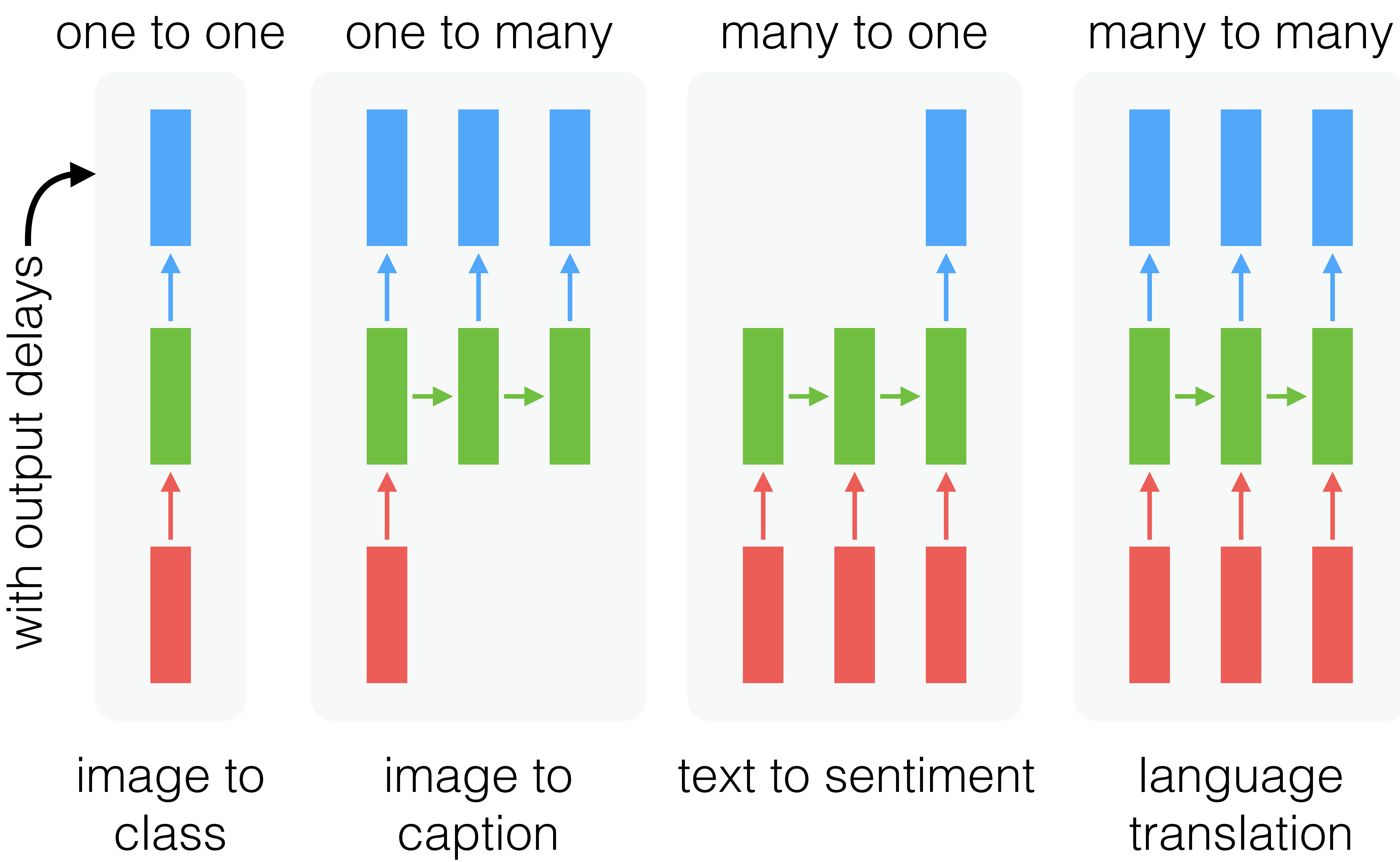
where **h** is the state of the system, **x** is the input, and θ are the parameters of the network

Recurrent Neural Networks

- We can describe RNNs with a graph containing cycles



RNN Sequence Mappings

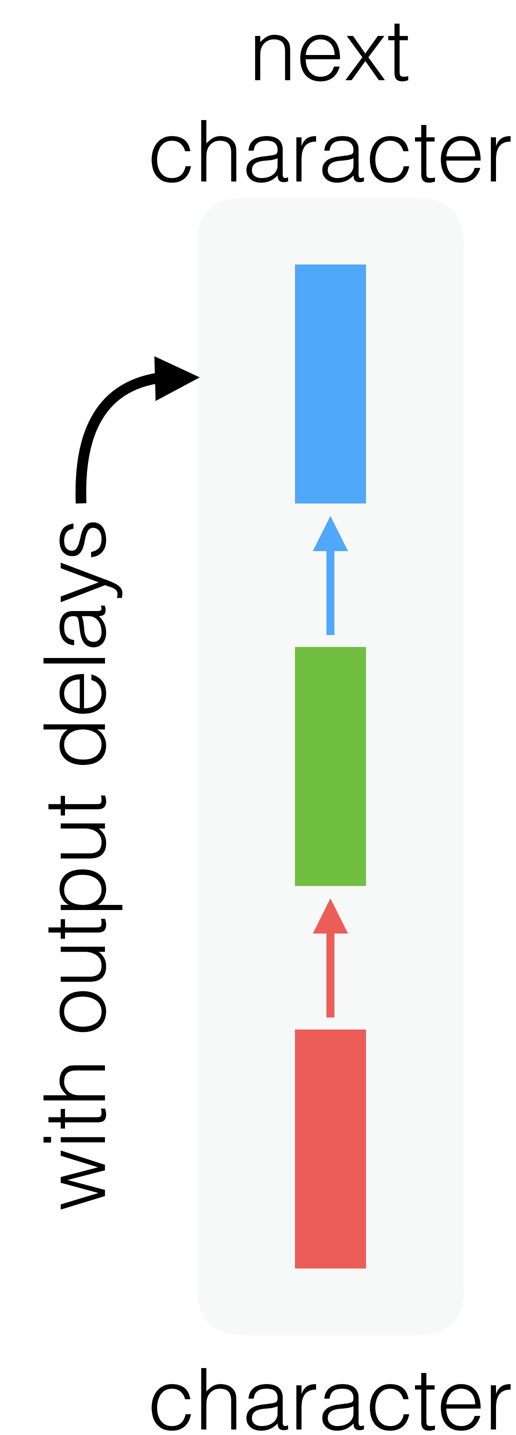


RNN Example I

- Train with:

The Sonnets by William Shakespeare

From fairest creatures we desire increase,
That thereby beauty's rose might never die,
But as the riper should by time decease,
His tender heir might bear his memory:
But thou, contracted to thine own bright eyes,
Feed'st thy light's flame with self-substantial fuel,
Making a famine where abundance lies,
Thyself thy foe, to thy sweet self too cruel:
Thou that art now the world's fresh ornament,
And only herald to the gaudy spring,
Within thine own bud buriest thy content,
And tender churl mak'st waste in niggarding:
Pity the world, or else this glutton be,
To eat the world's due, by the grave and thee.



RNN Example I

```
tyntd-iafhatawiaoihrdemot  lytdws  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng
```

train more

```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

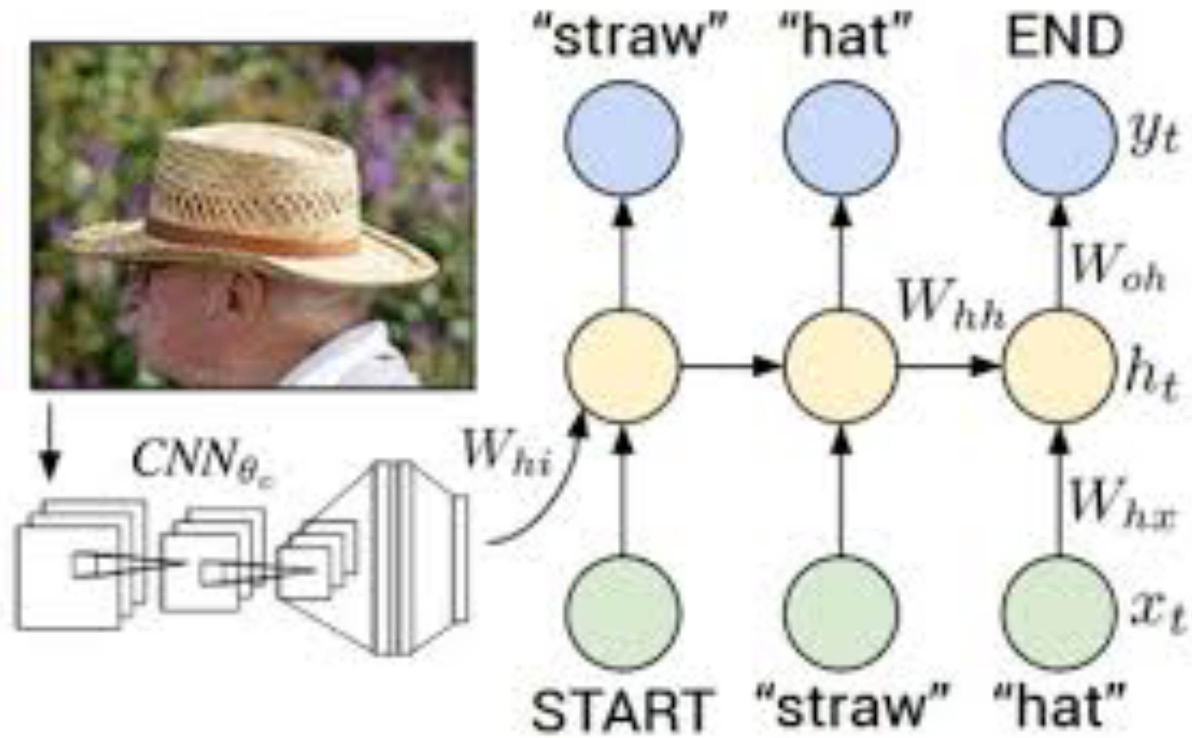
train more

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overelical and ofter.
```

train more

```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had oftened him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

RNN Example II




man in black shirt is playing guitar.


construction worker in orange safety vest is working on road.


two young girls are playing with lego toy.


boy is doing backflip on wakeboard.