Introduction to Machine Learning

Part IV

Judith Katzy
Hamburg, September 2024

Outline

Understanding Neural Network Behaviour

- Model/architecture
  - How is the model working? What is learned by a particular layer?
    - Example: filter visualization in lecture III

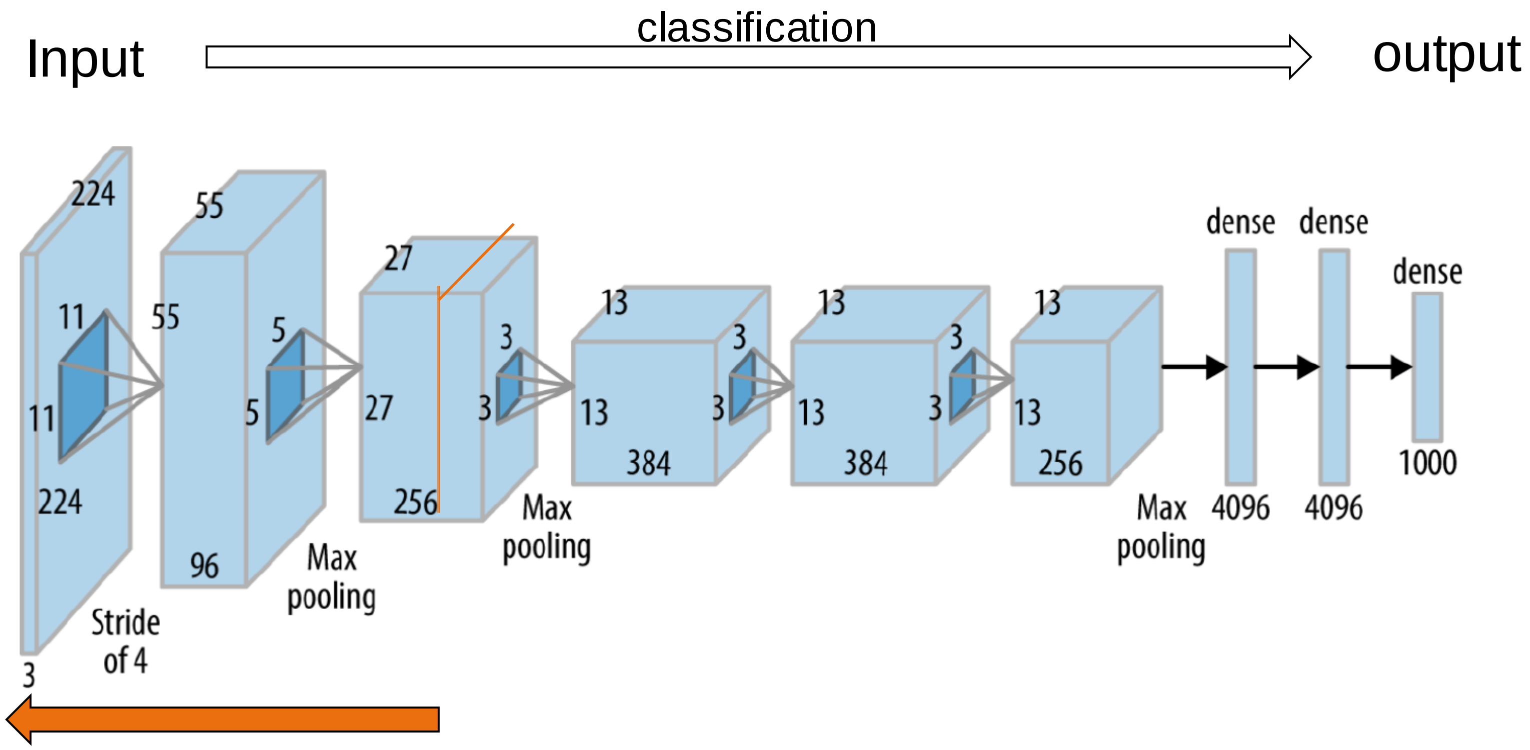
- Data:
  - Which part of the data is most important for the task?

- Predictions
  - Whis is a certain class (or value) predicted and can we quantify what contributes how much to the prediction?

Use CNN as example to illustrate the method, adaptable for other networks

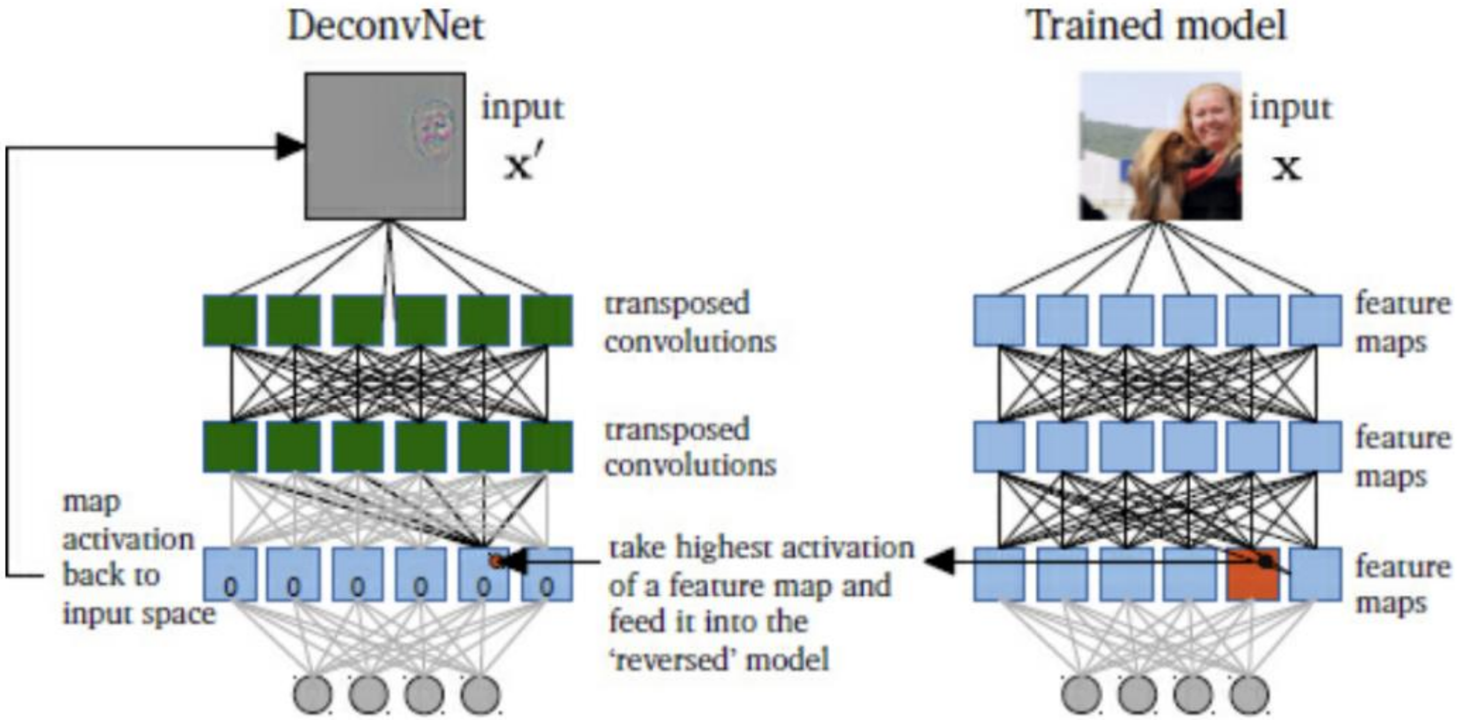
Data inspection

AlexNet: image classification



From activation in feature map back to image

Inspect which input pattern in the pixel space caused a given activation in the feature maps.



Step-by-step

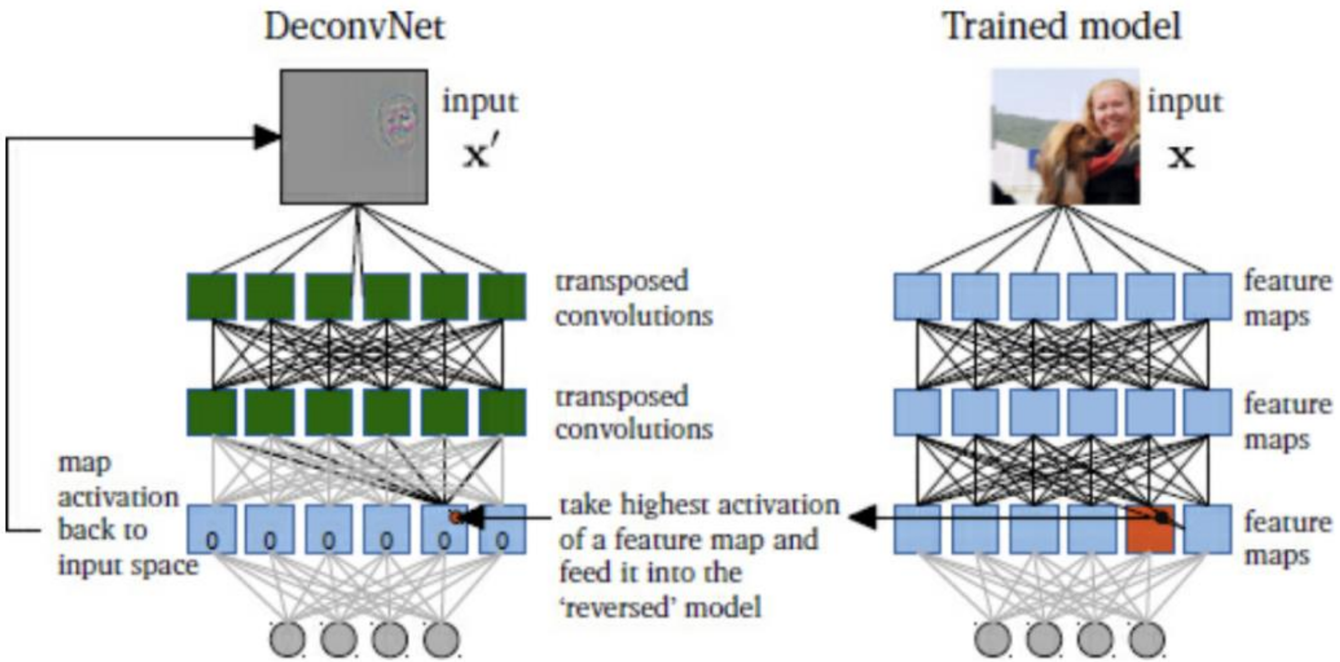
Present single image to trained network

➢ Chose image with high activation in one feature or just take maximal activation in one layer

Chose activated feature of interest, set all other features in that layer to 0

Reconstruct iteratively the information in each layer until the input layer is reached

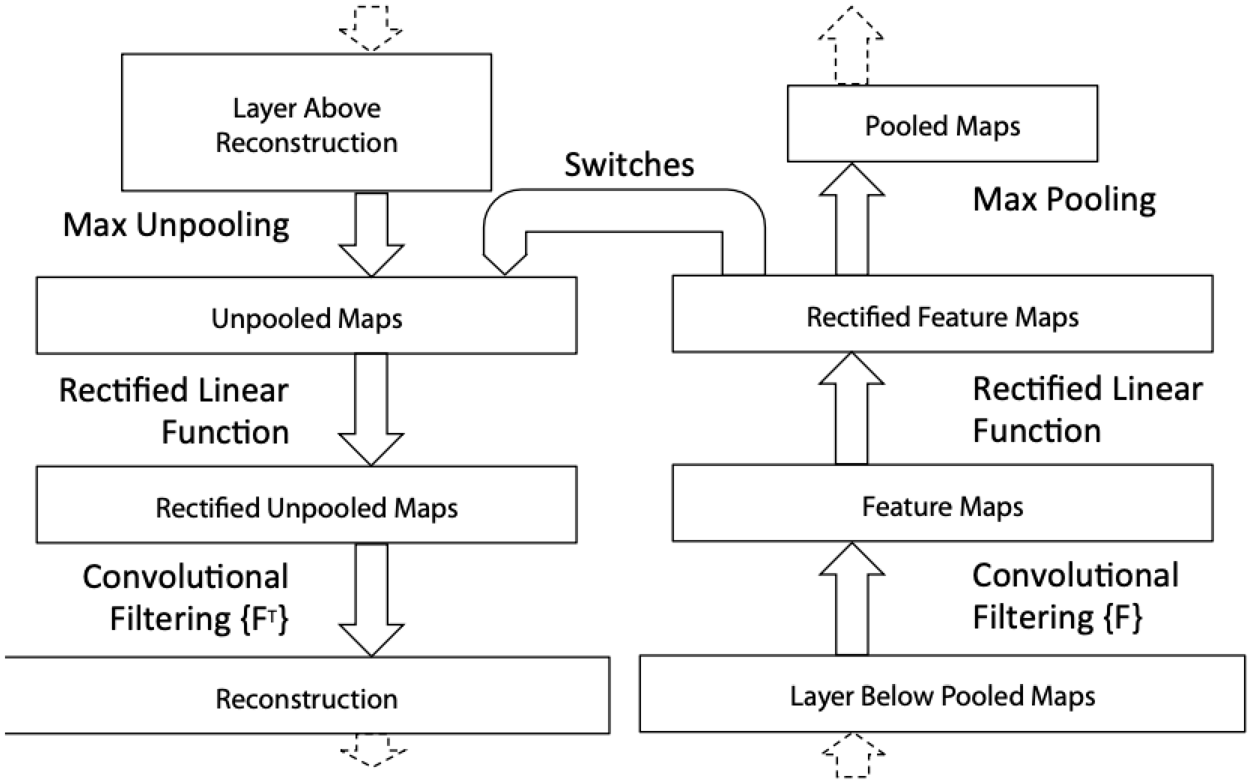
➢ Display activations projected down to pixel space and parts of input images



DeconvNet

Pass layer to an attached DeconvNet

classification



(pooling stage)

(activation stage)

(convolution stage)

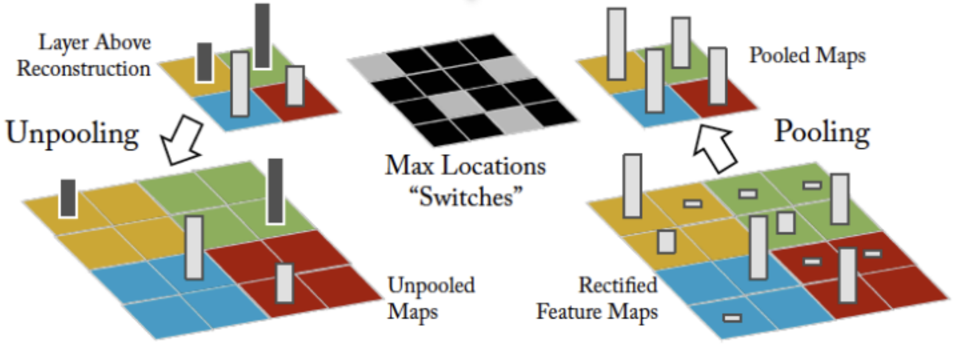
image

Unpooling

Store Location of Maximum value
from convolutional network

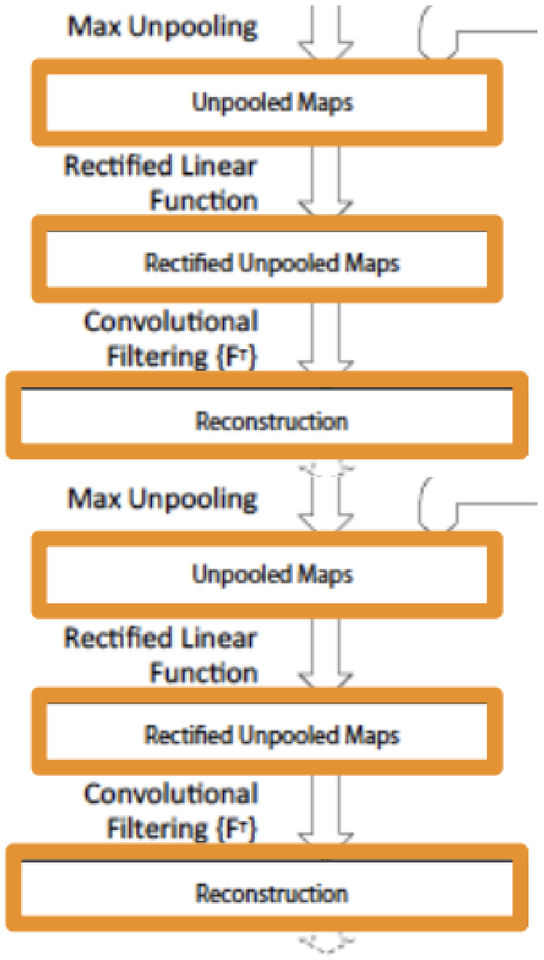
Deconvolution

Convolution

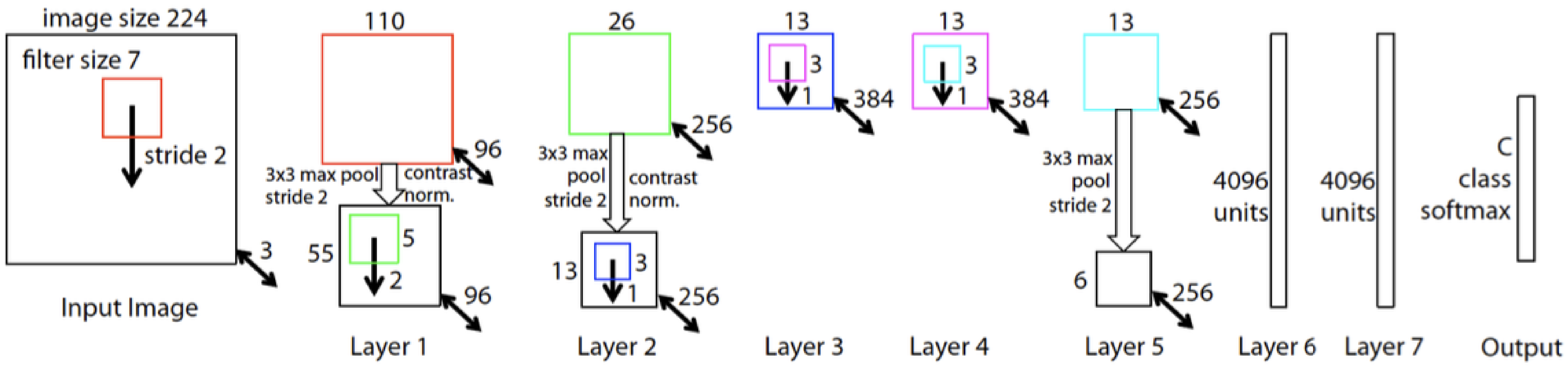


Reverse filtering operation

Deconvolute



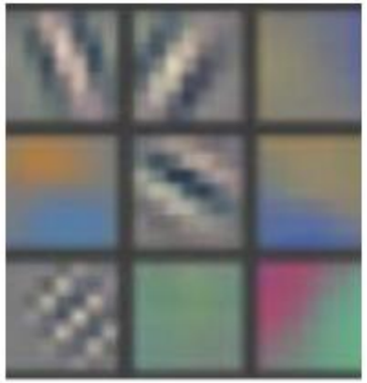
Example network



Results

Layer 1

Top 9 activations in layer 1



“visualisation”

Corresponding parts of input images (“patches”)

Only focusses on disriminating Features

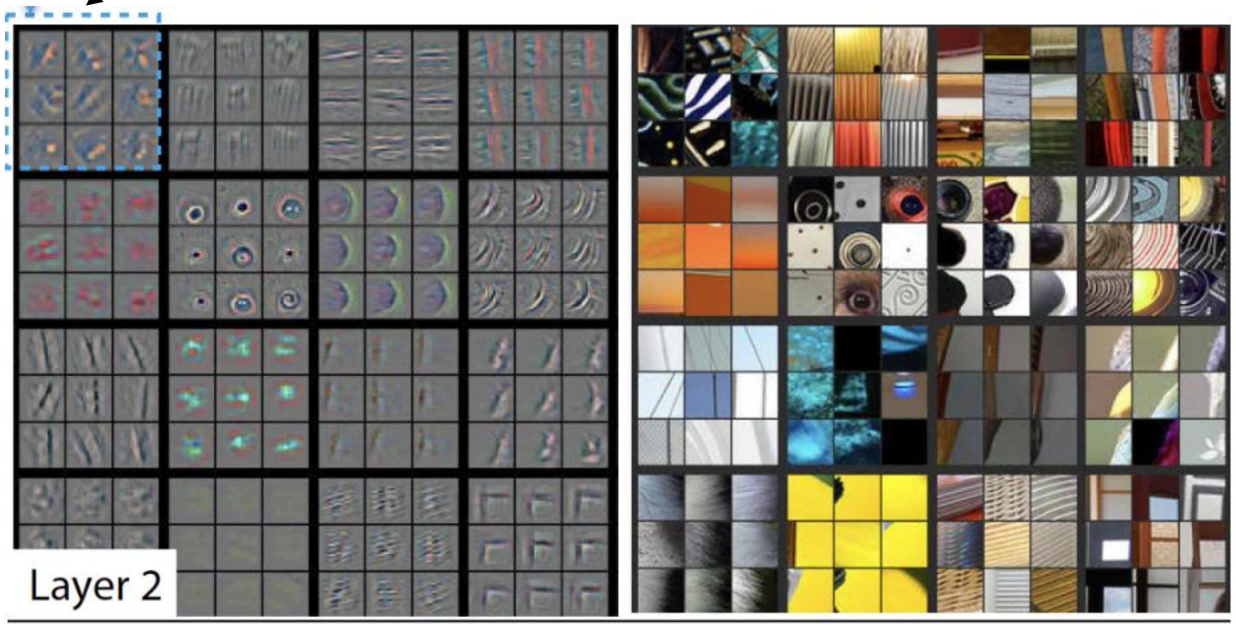
Results

Layer 2

Top 9 acitvations
In each feature map

Corresponding parts of input image
(patches)

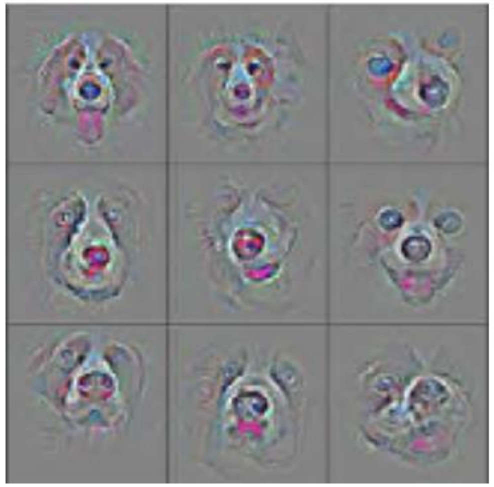
16 randomly selected
Feature maps



Results

Layer 4

Features triggering activations

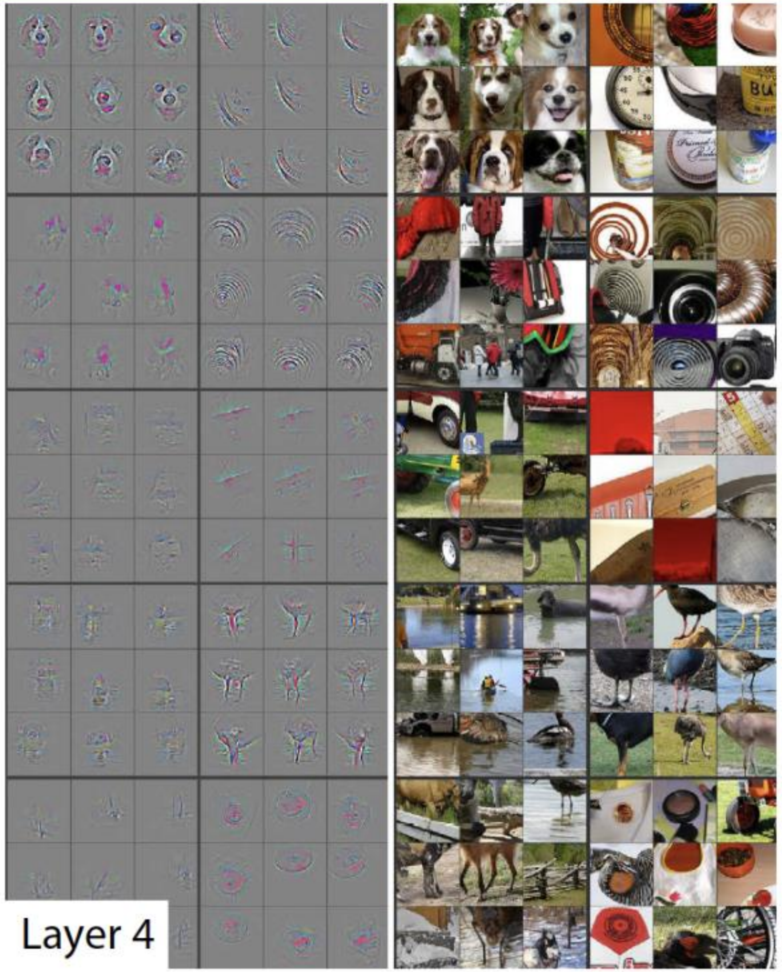


Corresponding parts of input images (“patches”)

Filters in later layers learn more specific and class related information
> here eyes and nose of the dog

Results

Different pictures in layer 4



Filters in later layers learn more specific and class related information

Results

Layer 5

Corresponding image patches

Little in common between images……?

Saliency Maps

Saliency maps

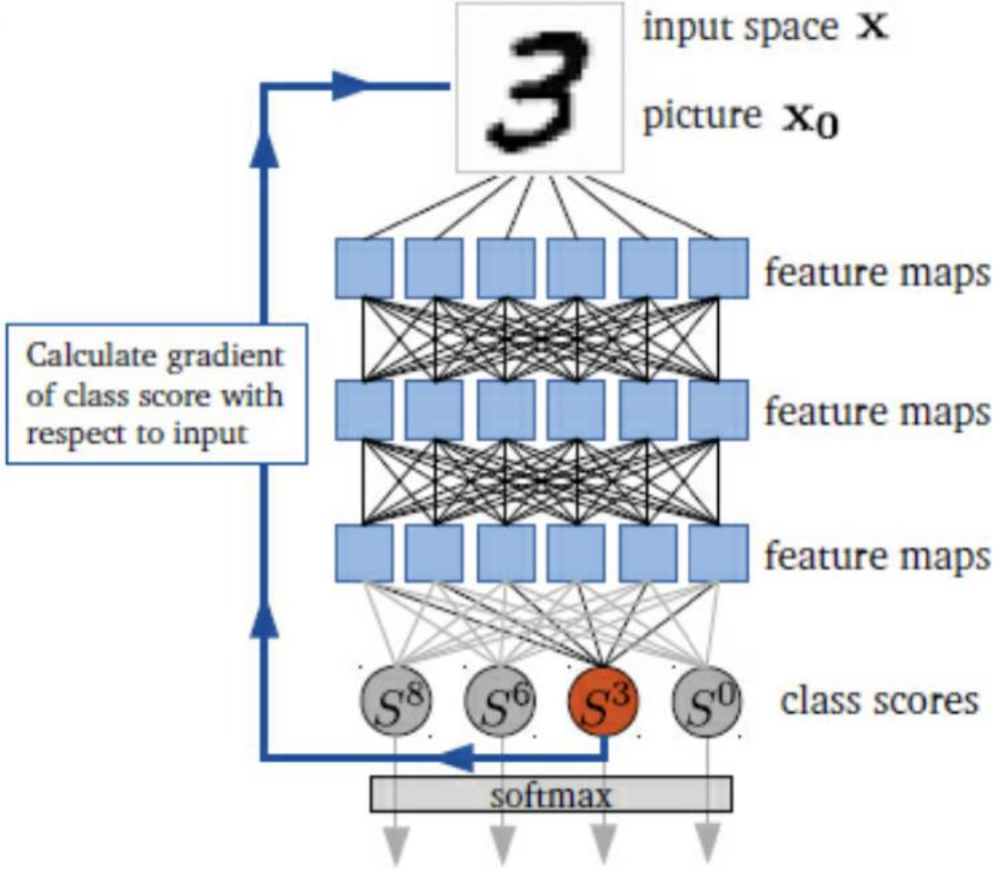
- Given a trained model f_θ
  > interpret model for given sample(image) $\mathbf{x}_0$: $f_\theta(\mathbf{x}_0)$
- > What caused the network prediction?

Example: a Mnist image of 3 is presented to the network, the network classifies it as 8

Why?

- Given a trained network predicting class c with score S for given input **x**: $S^c(\mathbf{x}_0,\theta)$

  > S^c = value before applying softmax to get output f_θ

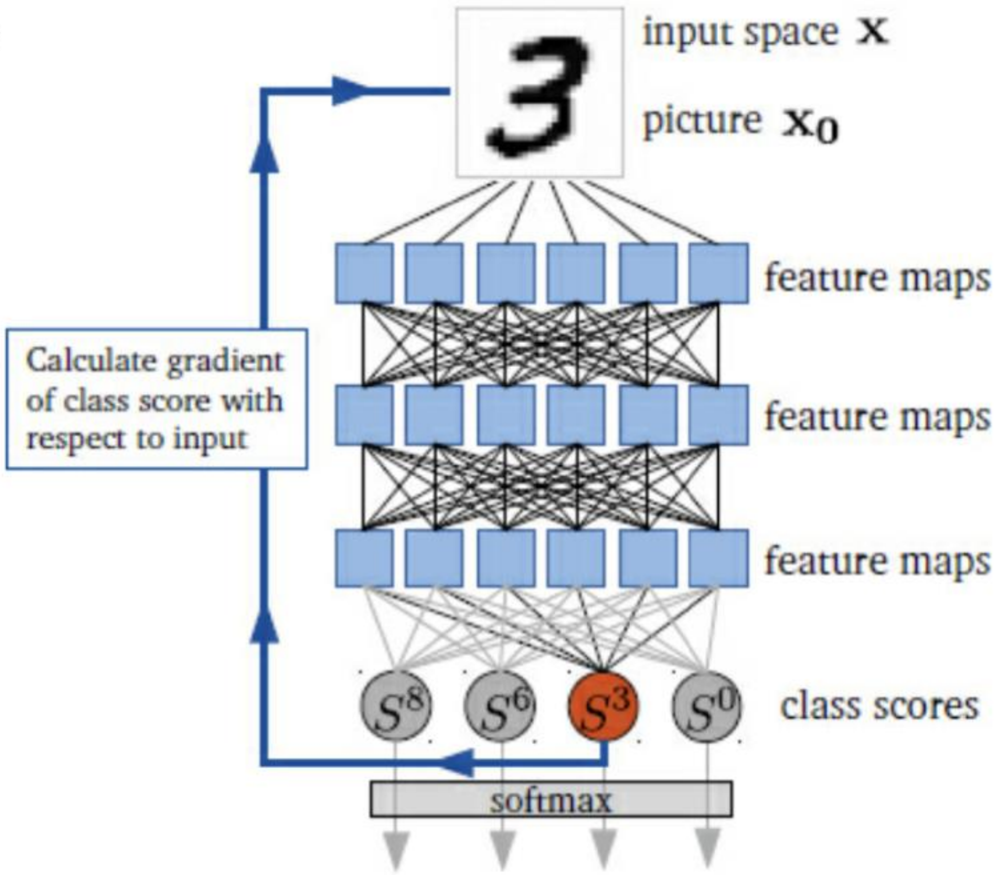


Gradients with respect to pixels

$$\mathrm{g}^c = \left. \frac{\partial S^c(\mathrm{x}, \theta)}{\partial \mathrm{x}} \right|_{\mathrm{x}=\mathrm{x}_0}$$

The larger g^c the stronger the sensitivity of the model to this input pixel value

Saliency map = complete set of gradients

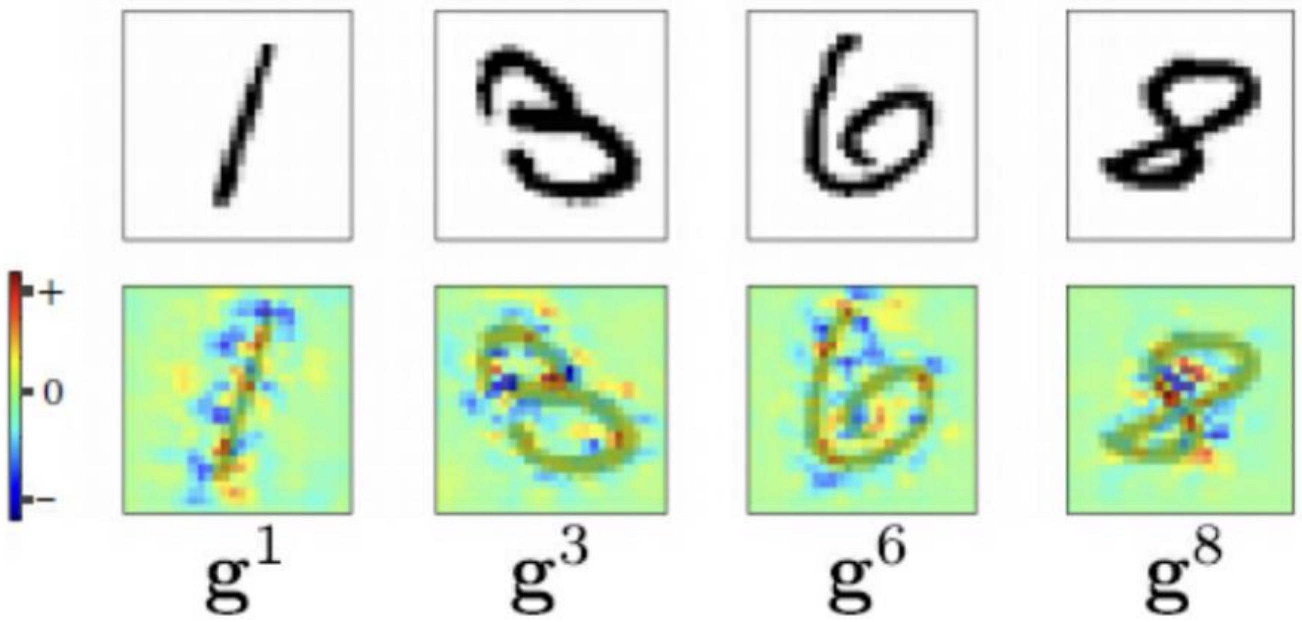


Saliency map

Visualisation

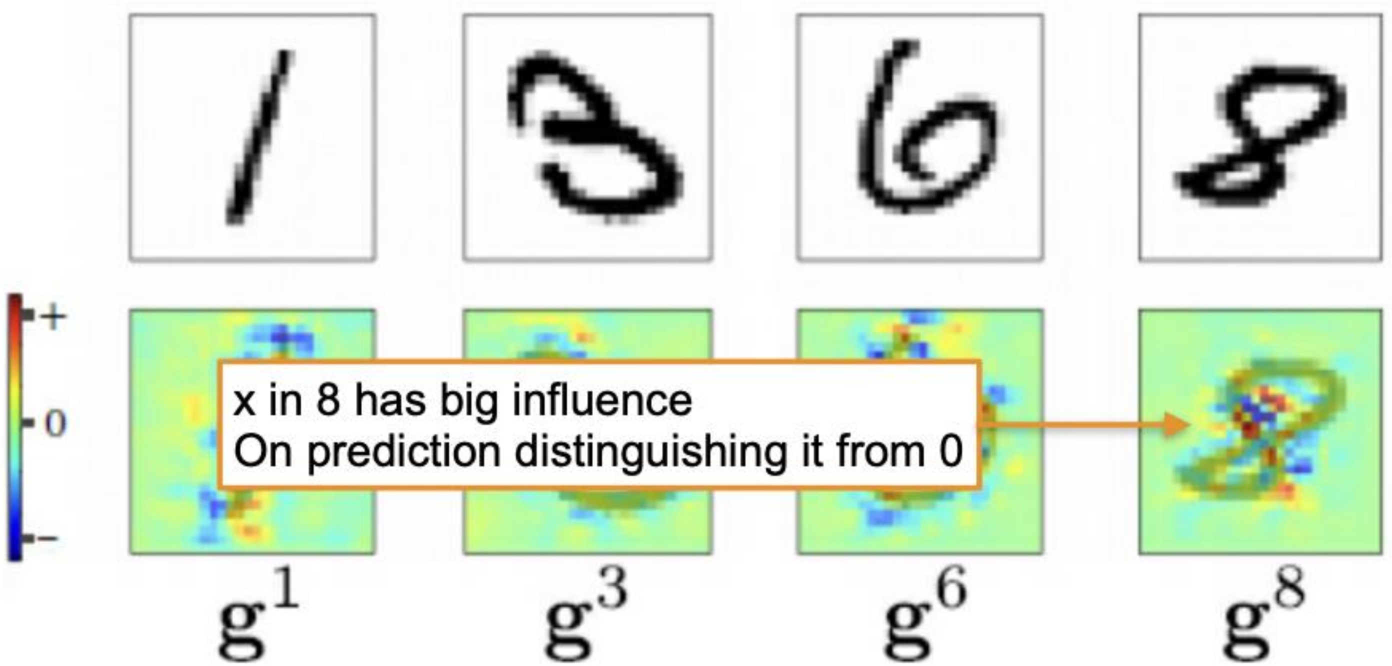
input

Saliency map for true classes (input overlayed)



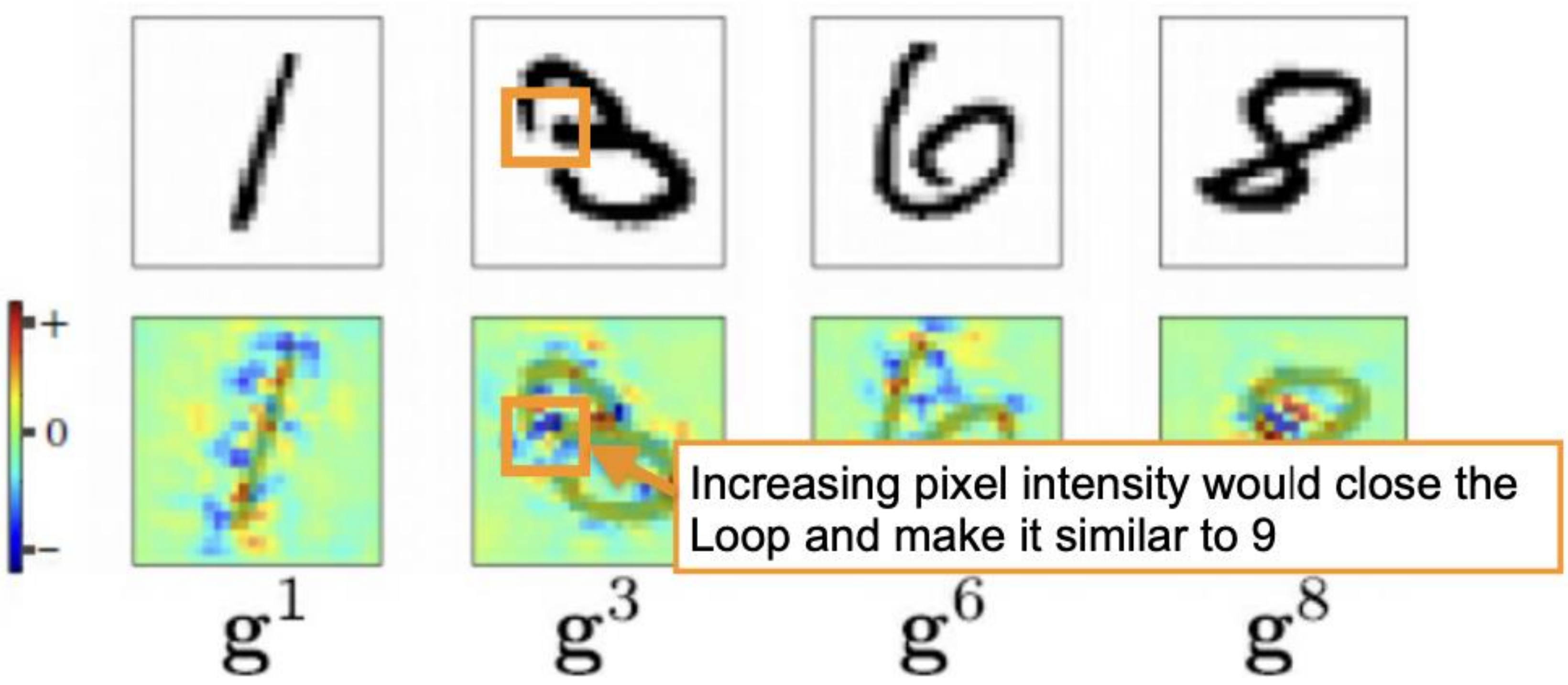
Saliency maps

Visualisation of pixel importance



Saliency maps

Visualisation of pixel importance



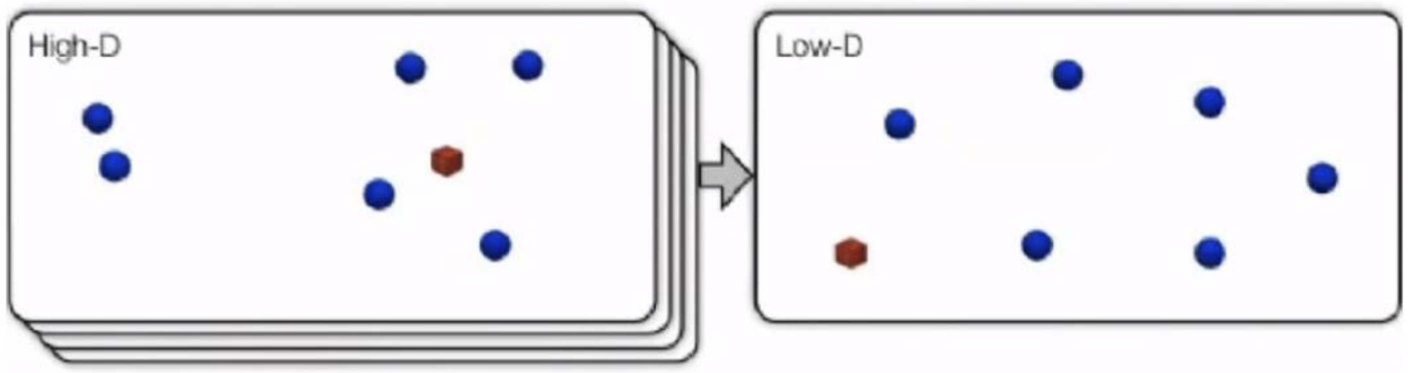
Visualisation of high dimensional data

or

How to visualize that two objects (pictures) are similar?

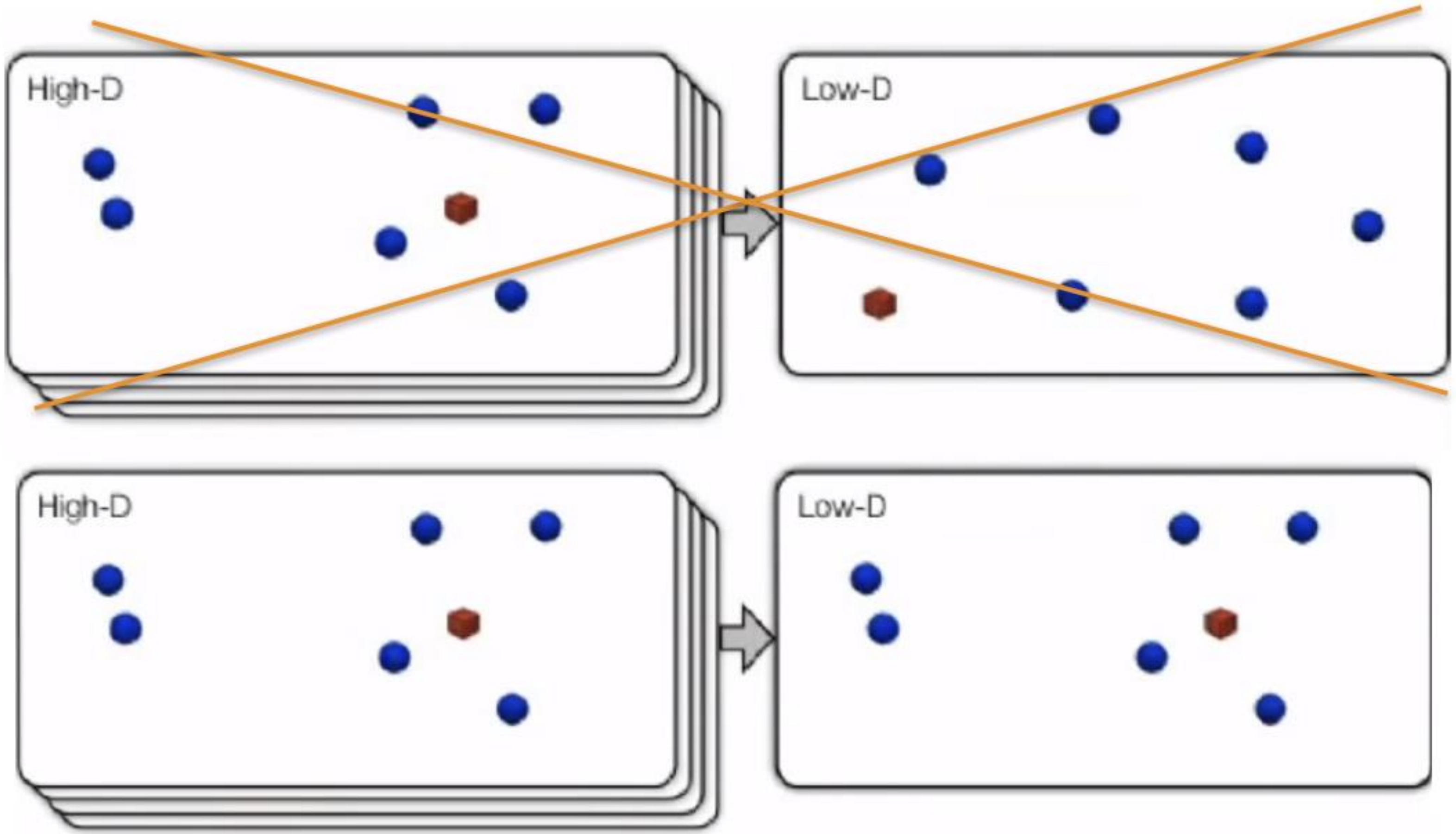
Basic idea

Reproduce distances in higher dimensional space as closely as possible in low dimensional “map”

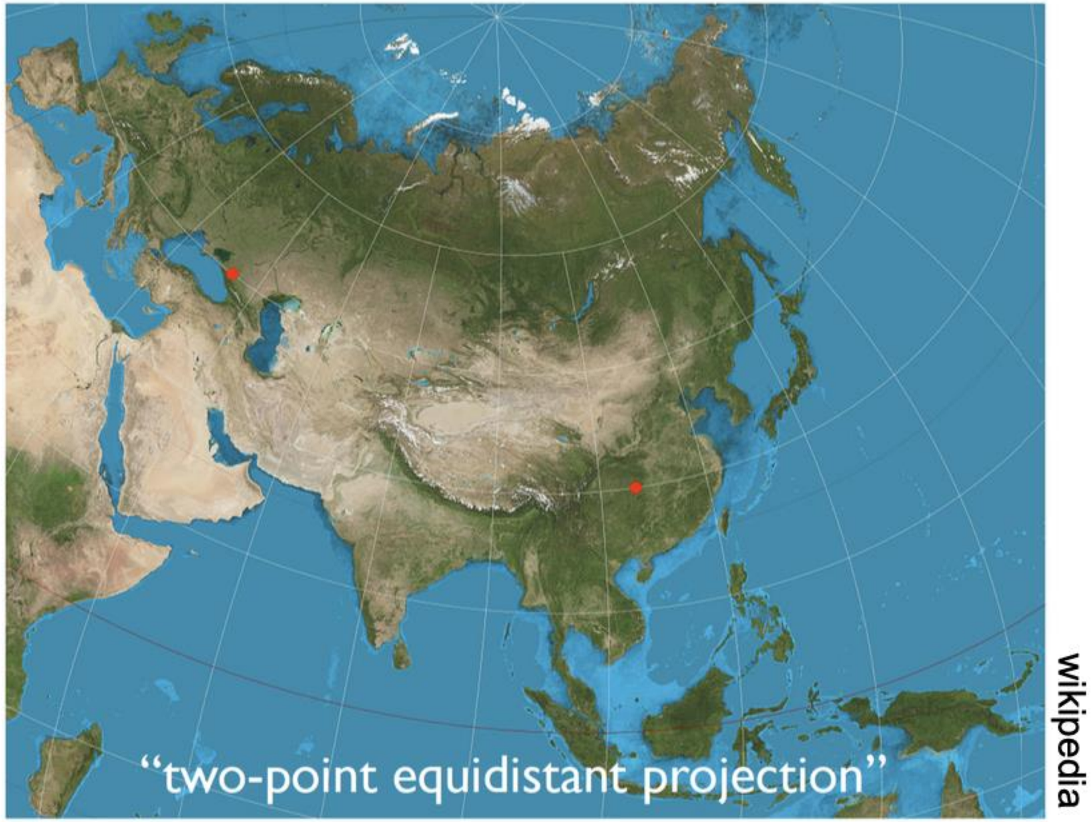


Basic idea

Reproduce distances in higher dimensional space as closely as possible in low dimensional "map"

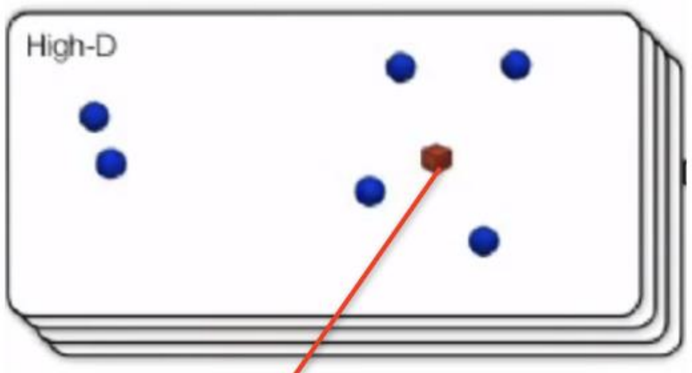


Map makers dilemma



T-distributed Stochastic Neighbour Embdding

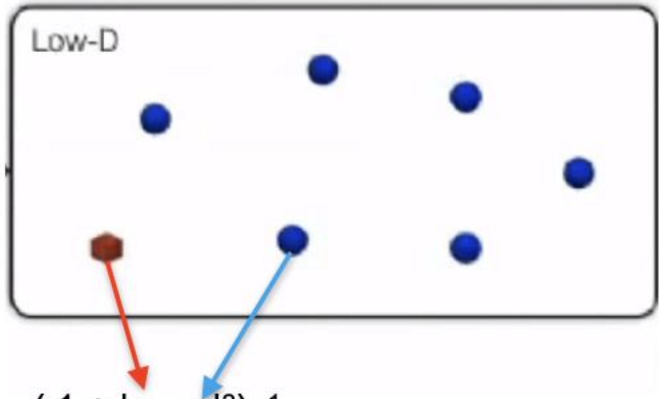
- Define pairwise probability distribution that depends on distance in high dimensional space
  - Probability higher for close neighbours



$$p_{ij} = \frac{\exp(- |\, x_i - x_j|^2 \,/\, 2\,\sigma^2)}{\Sigma_k \, \Sigma_{l!=k} \exp\,(- |x_k - x_l|^2 \,/\, 2\sigma^2)}$$

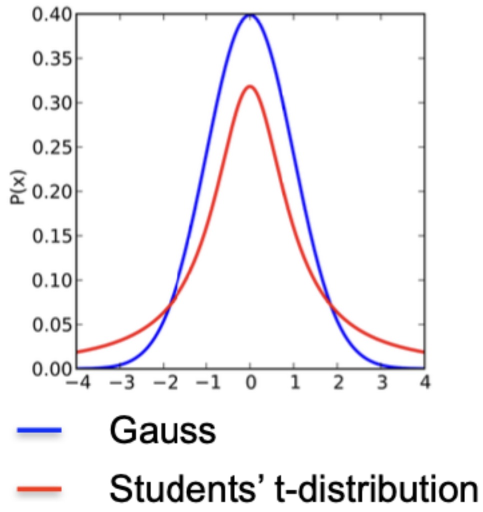
Normalisation over all pairs of points

tSNE similarities in low dimensional space



Pairwise density: $$q_{ij} = \frac{(1 + |y_i - y_j|^2)^{-1}}{\Sigma_k \Sigma_{k!=l} (1 + |y_k - y_l|^2)^{-1}}$$

Students't distribution
"Cauchy" distribution



q comparably larger at long distances:
> low-dim has less space for points -> need to give them more room
> allows points in low-dim space to spread out for intermediate distances

tSNE find optimal mapping

Find q-distribution similar to p-distribution

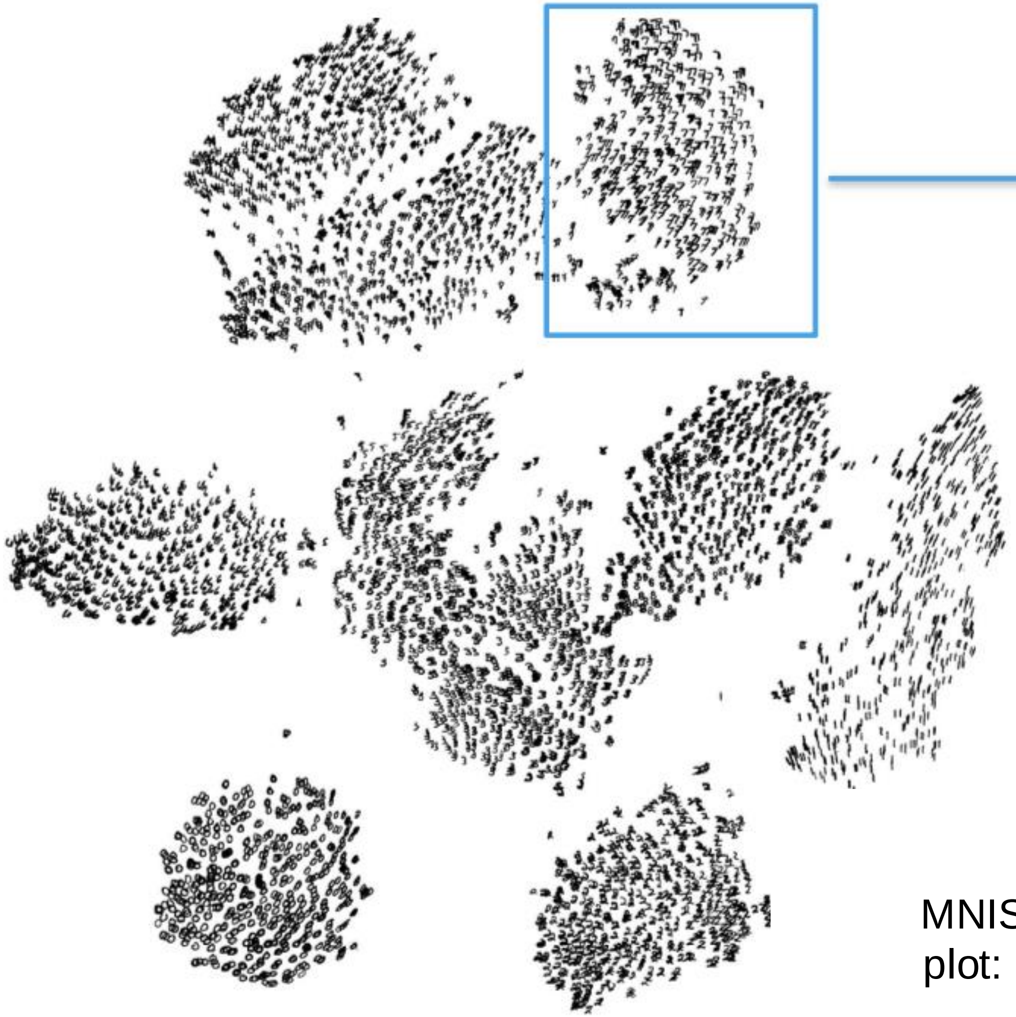
Performance measure: Kullback-Leibler (KL) divergence

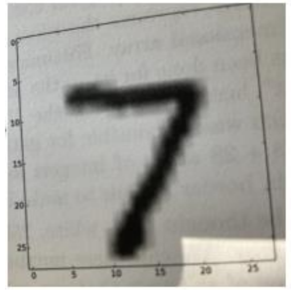
$$KL(P||Q) = \Sigma_i \Sigma_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Move points in low dimensional space around such that KL is minimized

- Use gradient descent for dKL/dy_i

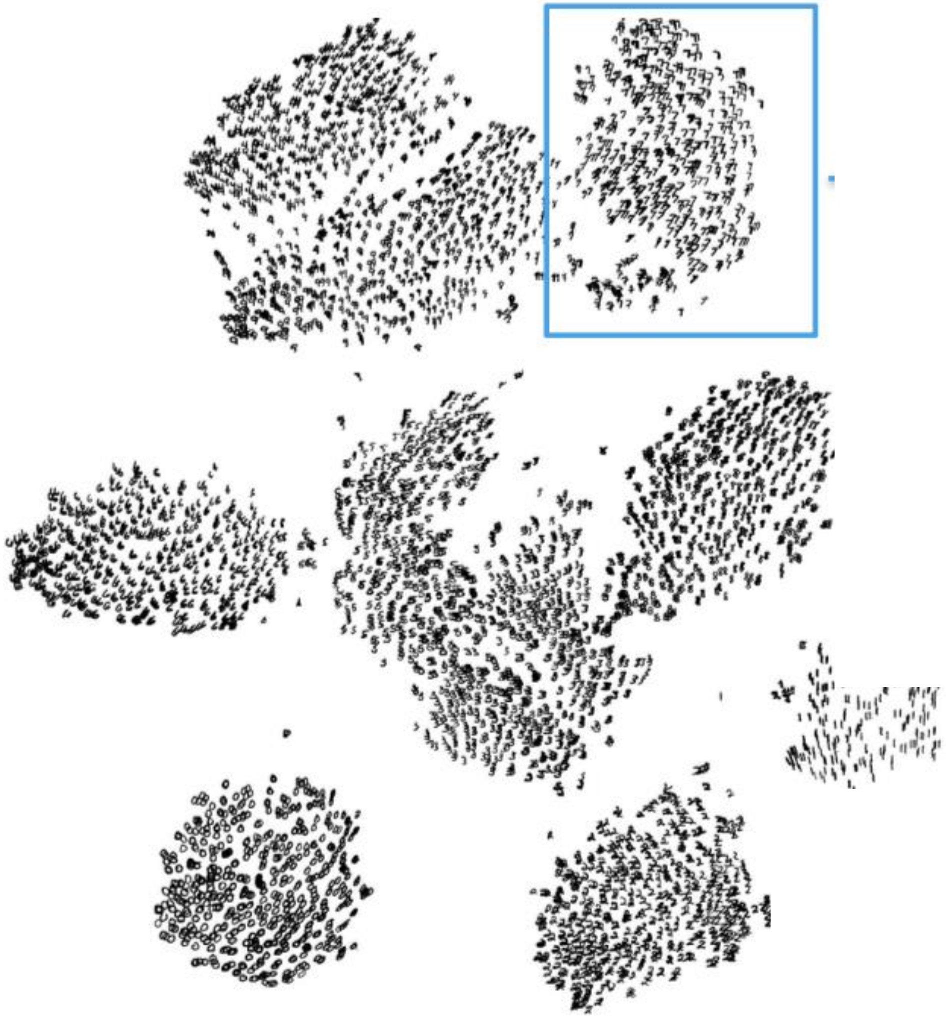
tSNE on MNIST

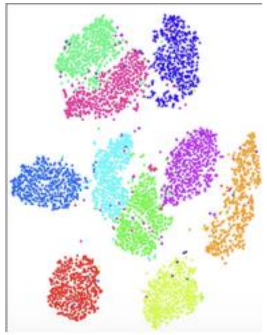


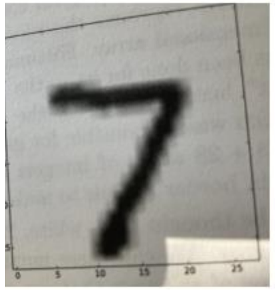


MNIST: high dimensional data 28x28
plot: 2 dimensional

tSNE on MNIST







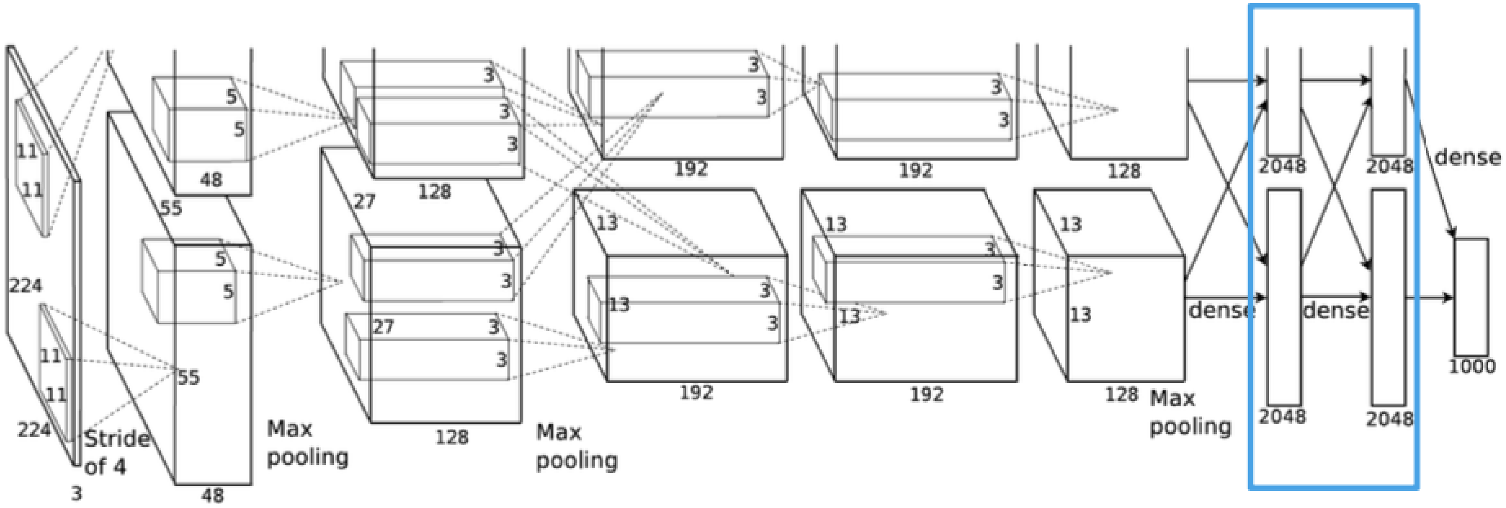
Color each image according to the truth label
From 0-9

Visualisation of CNN nodes with tSNE

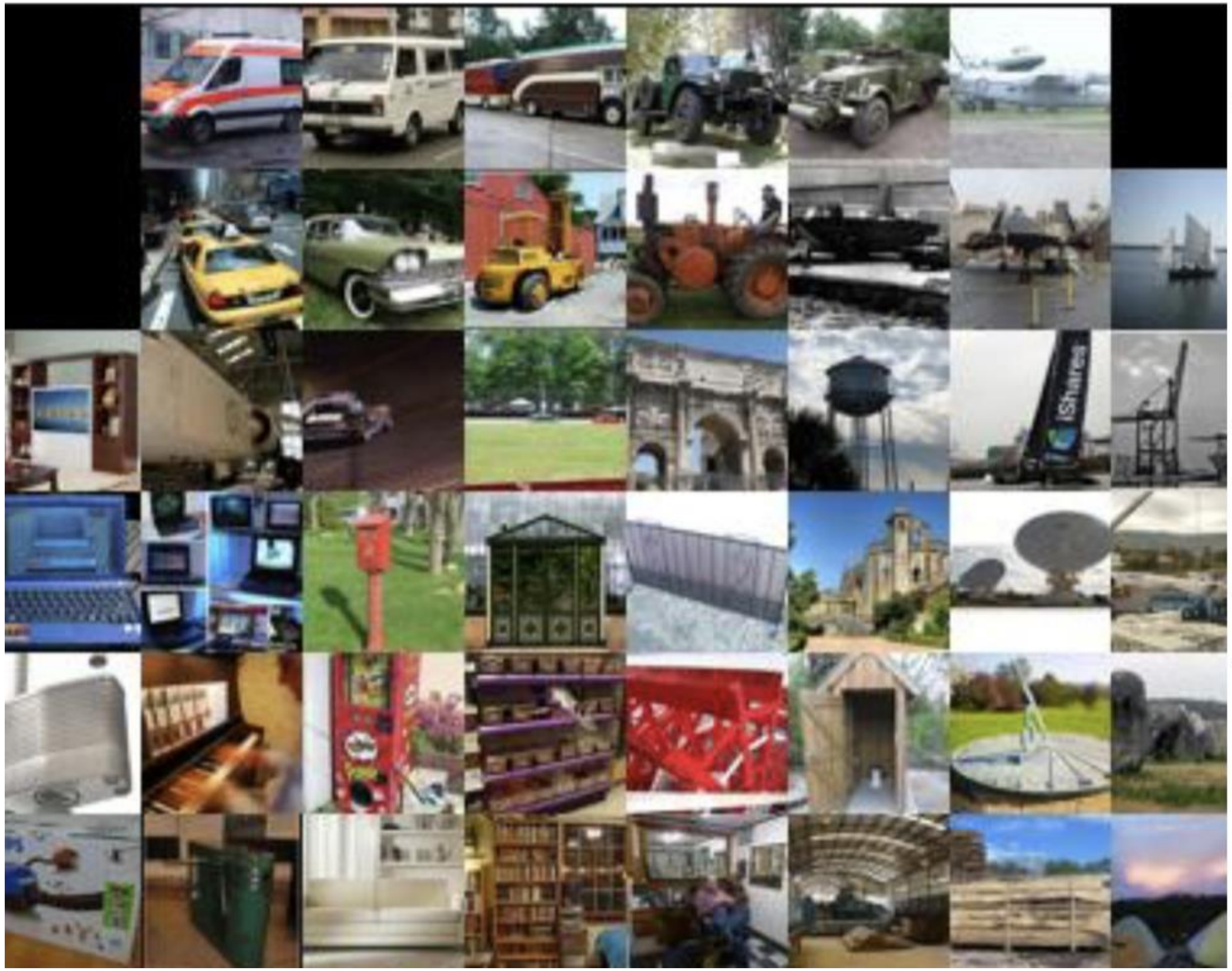
> Take 4096 nodes of multi-layer CNN classifying ImageNet pictures (2012 competition)

> map 4096 nodes down to 2-dim using tSNE

4096 noes



Nodes of CNN classifying image net pictures

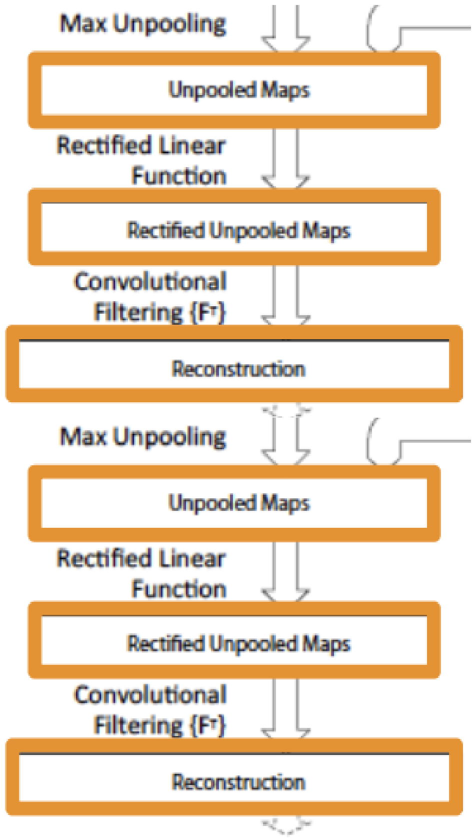


https://cs.stanford.edu/people/karpathy/cnnembed/

Back-up

Reverse filtering operation

Deconvolute



...

- Unpool
- Rectivy output of unpooling
- Apply transpose filter matrix to undo filtering

Repeat until input layer is reached