

Deep Learning & Computer Vision (and more)

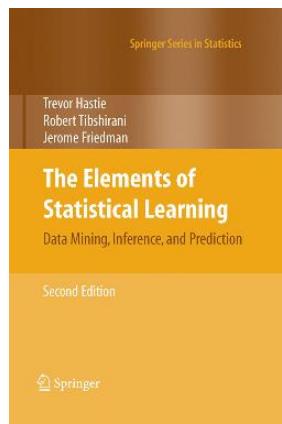
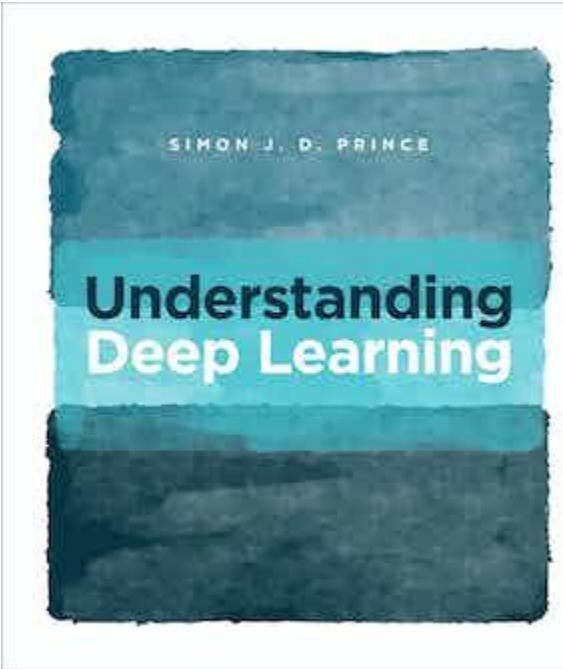
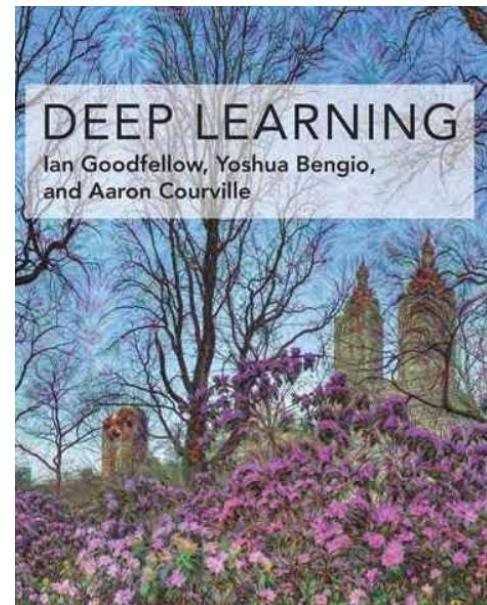
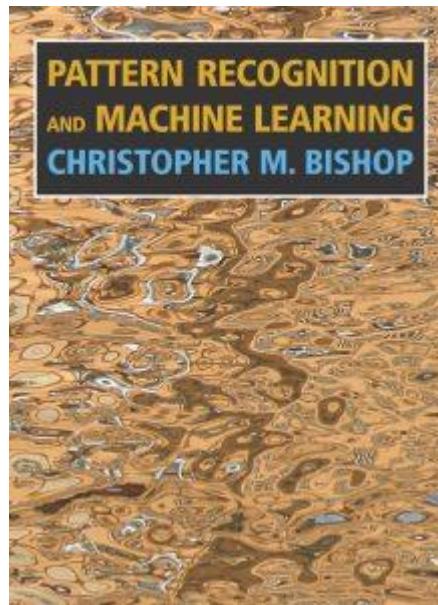
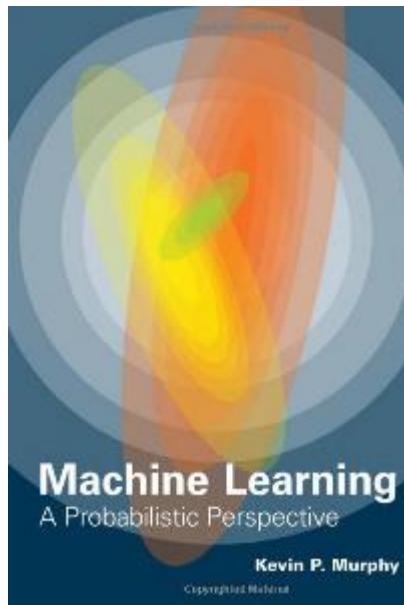
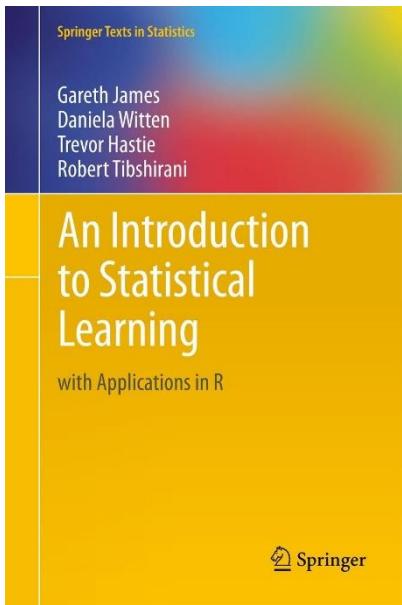
Mihalis A. Nicolaou

Nov 2023

Overview

- 1. Introduction**
- 2. Recent Developments**
- 3. Intro to ML and DL**
- 4. Learning Representations from High-dim Data**
- 5. Deep Learning**
- 6. M/T methods in Deep Learning**

Reference Books

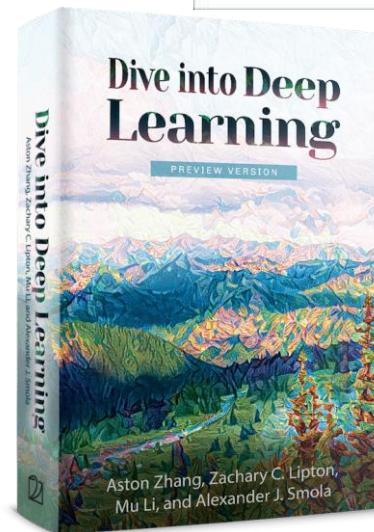


Neural Networks and Deep Learning

Neural Networks and Deep Learning is a free online book.

The Matrix Cookbook
[<http://matrixcookbook.com>]

<http://www.matrixcalculus.org/>



Advice for new PhD students

- Many talks and notes written on the web on advice to new PhD students from experienced academics
- Search for yourselves – e.g., depending on field etc.
 - More or less same advice holds across fields
- Critically think for yourselves – not every advice is up-to-date or necessarily correct for your case

How to get your PhD?

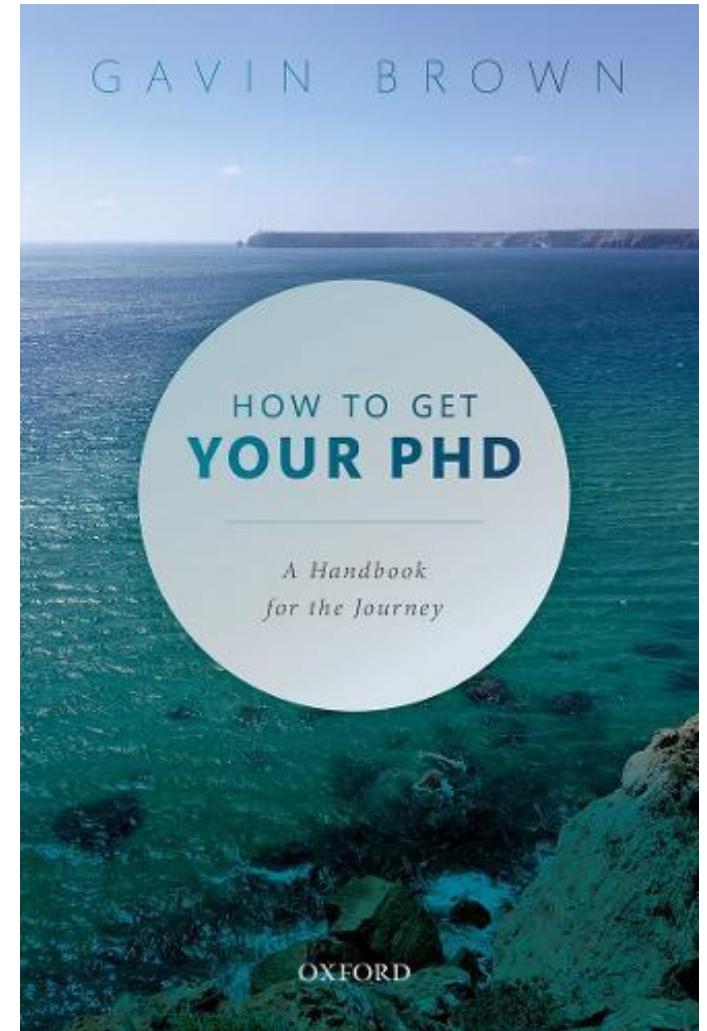
Free chapters more than enough!

Part 1: How to get your PhD

- [Preface](#) FREE
- [Chapter 1: What is a PhD?](#)
- [Chapter 2: Things you need to know](#)
- [Chapter 3: Things that will happen](#) FREE
- [Chapter 4: How to find your new idea](#)
- [Chapter 5: How to work with a supervisor](#)
- [Chapter 6: Work on yourself](#)
- [Chapter 7: How to read](#) FREE
- [Chapter 8: How to write](#)
- [Chapter 9: How to give presentations](#)
- [Chapter 10: What happens in a viva?](#)
- [Chapter 11: Your post-PhD life](#)
- [Chapter 12: Skills you will have gained](#) FREE

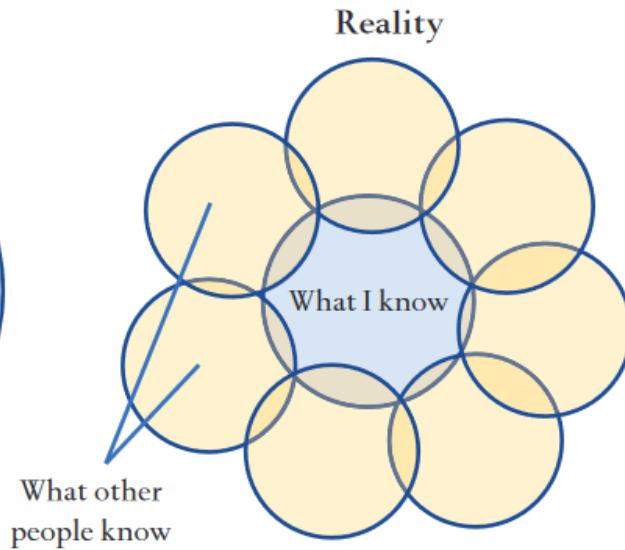
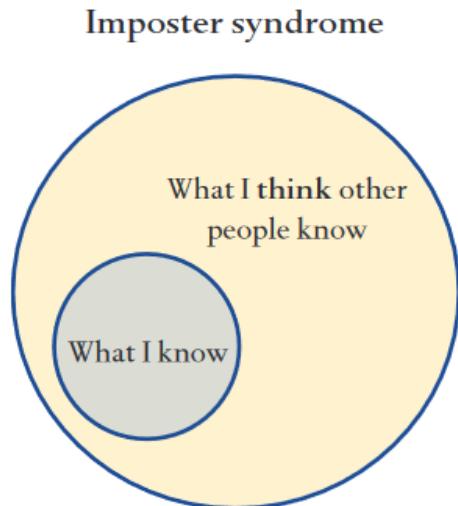
Part 2: Top Tips from Leading Minds

- [I. Choosing a career in science](#)
 - By Professor Dame Nancy Rothwell, DBE, FRS (University of Manchester)
- [II. Becoming a reflective professional](#) FREE



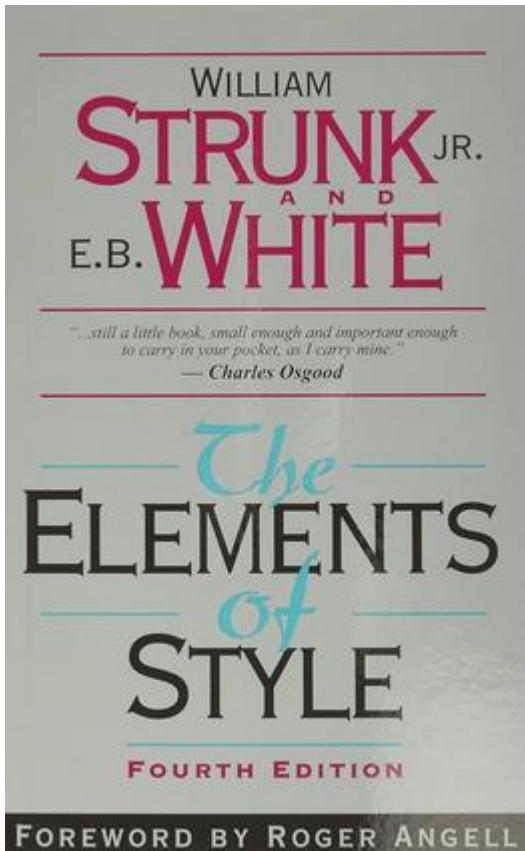
<https://profgavinbrown.github.io/howtogetyourphd/>

How to get your PhD?



Writing

- [Ten Simple Rules for Mathematical Writing](#)
 - https://www.mit.edu/~dimitrib/Ten_Rules.html



"You don't need to be a grammar nerd to enjoy this one....Who knew grammar could be so much fun?" —*Newsweek*

The #1 New York Times Bestseller

Eats, Shoots & Leaves



The Zero Tolerance Approach to Punctuation

! LYNNE TRUSS

With a Foreword by Frank McCourt,
author of *Angela's Ashes*

Recent Developments.

Developments (past decade+)



ARTIFICIAL INTELLIGENCE
AND LIFE IN 2030

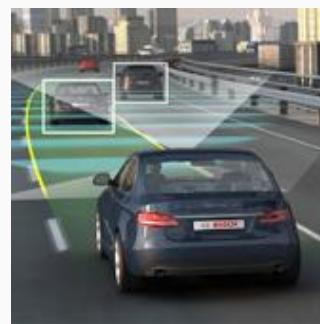
FT Financial Times
[AI breakthrough ChatGPT raises alarm over student cheating](#)



A robot wrote this entire article. Are you scared yet, human?
GPT-3

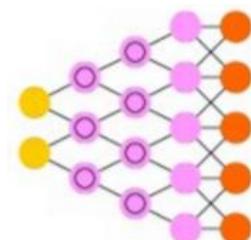


2014



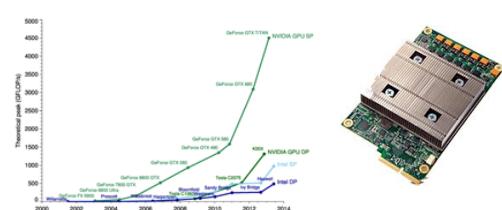
2018

↑
Foundations



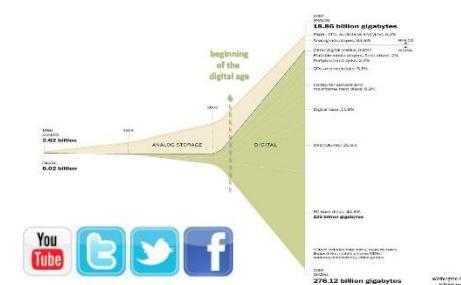
Algorithms

↑
Compute Availability



Parallelizable, Scalable

↑
Data Availability



Large datasets, annotations

M. A. Nicolaou

M I N D
 A QUARTERLY REVIEW
 OF
 PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND
 INTELLIGENCE

By A. M. TURING

1. The Imitation Game.

I PROPOSE to consider the question, ‘Can machines think ?’ This should begin with definitions of the meaning of the terms ‘machine’ and ‘think’. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words ‘machine’ and ‘think’ are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, ‘Can machines think ?’ is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the ‘imitation game’. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either ‘X is A and Y is B’ or ‘X is B and Y is A’. The interrogator is allowed to put questions to A and B thus :

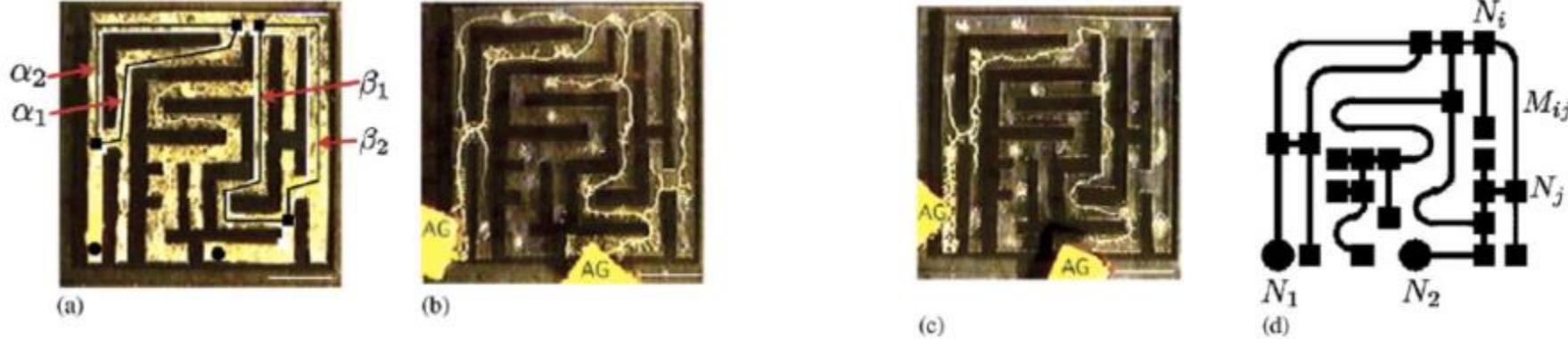
C: Will X please tell me the length of his or her hair ?
 Now suppose X is actually A, then A must answer. It is A’s

‘Can machines think ?’

“I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of **machines thinking without** expecting to be **contradicted**.”

“I’m confident that they think”
 - Geoffrey Hinton, 2023

Intelligence, working definition



- How can we define intelligence (e.g., in terms of energy)

Slime mold Physarum

- Single-celled amoebae
- No brain, no nervous system
- Yet, they somehow manage to be ‘intelligent’:
 - Can collectively find the shortest path between two pieces of food

Why ML?

- There exists knowledge (e.g., tacid) that we *can not* describe with if-then rules or analytical functions
- More generally, in many cases we simply do not know the function that maps a problem state to a solution
- Even if we do, we can potentially do it faster (e.g., simulations)
- This is also largely due to the vast variability and noise manifesting in real-world data/problems (e.g., in faces: illumination, pose, expression)



POSED, ILLUMINATION VARIATION



OCCLUSION



UNCONTROLLED SETTINGS

Deformations, Invariances, Modes of variation

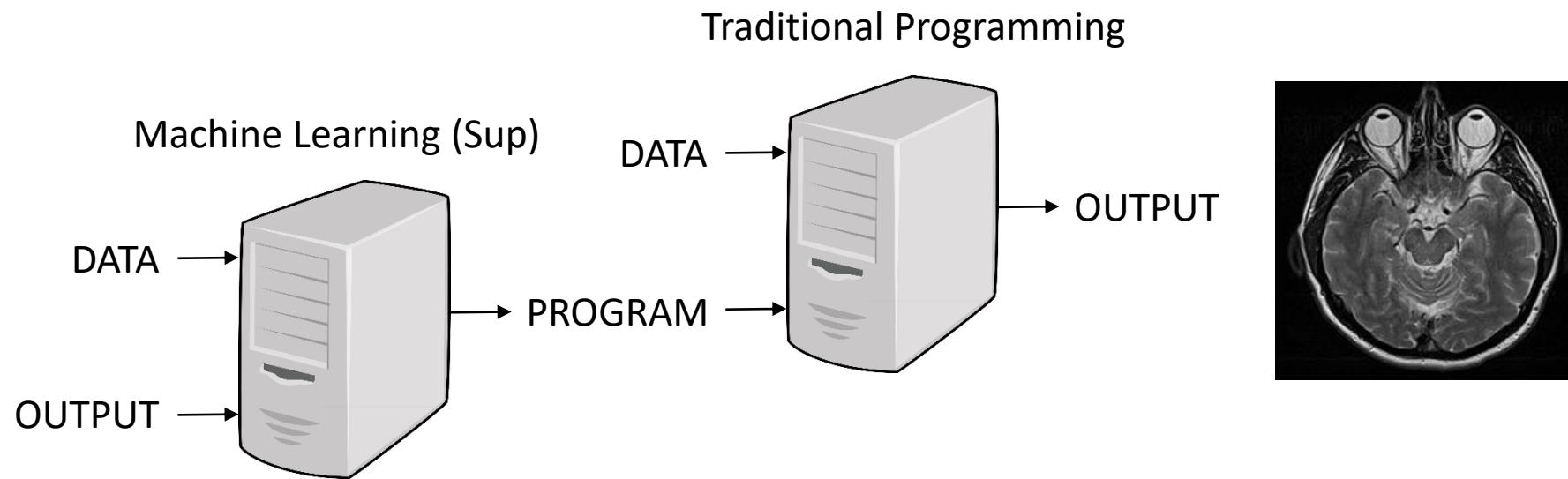
Problem: model all possible deformations/variations that an object might have while retaining its identity



After how much deformation does a giraffe stop being a giraffe?

- **Inductive Bias** of algorithm: e.g., CNN's are translation equivariant
- **From Data:** Extract invariances from data; capacity of model, large-scale dataset
 - Annotations and labels: expensive, error-prone, not scalable
- **Language:** Language can offer ‘supervision’ for ‘free’

Machine Learning



Arthur Samuel (1959)

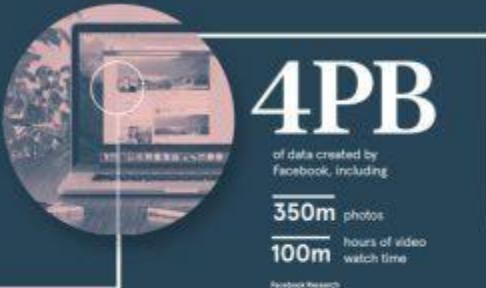
“... gives computers the ability to learn without being explicitly programmed”

Data: From few to exabytes

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devices - are hard to comprehend, particularly when looked at in the context of one day

 500m
tweets are sent every day

 4PB
of data created by Facebook, including
350m photos
100m hours of video watch time

DEMYSTIFYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data.

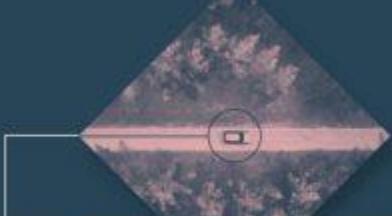
Unit	Value	Size
B	bit	1/8 of a byte
B	byte	1 byte
KB	kilobyte	1,000 bytes
MB	megabyte	1,000 ² bytes
GB	gigabyte	1,000 ³ bytes
TB	terabyte	1,000 ⁴ bytes
PB	petabyte	1,000 ⁵ bytes
EB	exabyte	1,000 ⁶ bytes
ZB	zettabyte	1,000 ⁷ bytes
YB	yottabyte	1,000 ⁸ bytes

*In lowercase 'B' is used as an abbreviation for bytes, while an upper case 'B' represents bytes.

 294bn
billion emails are sent

 3.9bn
people use emails

 3.9bn
people use emails

 4TB
of data produced by a connected car

ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB

2013

44ZB

2020

 65bn
messages sent over WhatsApp and two billion minutes of voice and video calls made

 Searches made a day

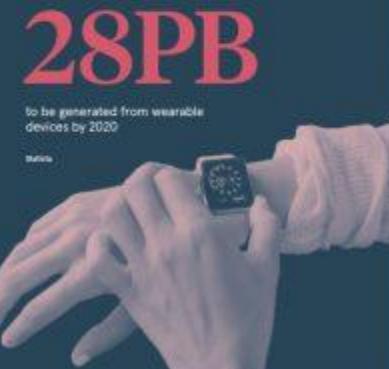
 Searches made a day from Google

 463EB

of data will be created every day by 2025

 95m
photos and videos are shared on Instagram

Instagram Business

 28PB
to be generated from wearable devices by 2020

RAconteur

Large-scale Data, Large-scale Models

Annotated Data: ImageNet (2009)



14M images / 20K categories



2012: Alexnet ~ 85%

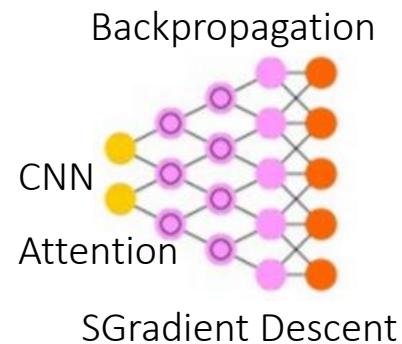
Vision Transformer (ViT)



EfficientNET

CLIP

Deep (hierarchical) Learning: Image Recognition



Parallelizable



> 60M parameters

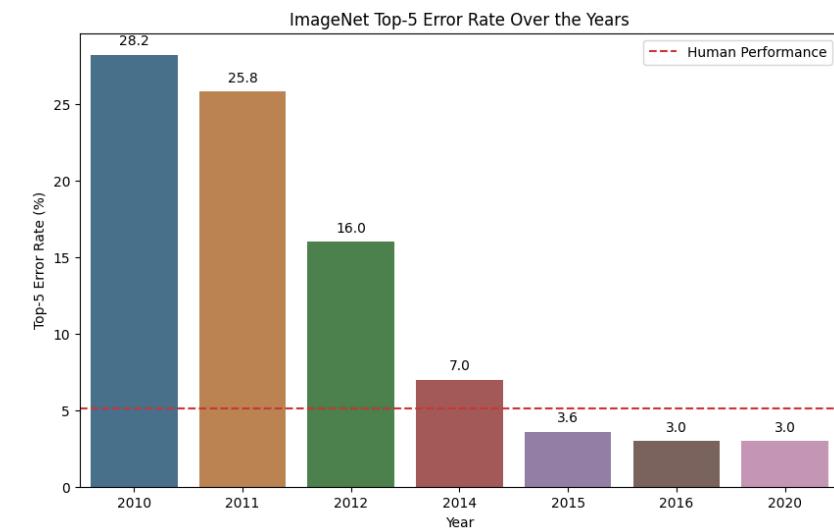


Now: Classification > 97%

+

=

Scalable



(B) Millions of parameters (DL) + (B) millions of data: *paradigm shift* in AI, ML and CV [and lately even trillions]

Text-to-image

Stable Diffusion



Hipster Llama wearing a hat, studio lighting, award winning photography.

Imagen



a teddy bear on a skateboard
in times square

DALL-E 2



A group of teddy bears in suit in a corporate office celebrating the birthday of their friend. There is a pizza cake on the desk.

M Can you write python code that solves my machine learning assignment: perform gradient descent for logistic regression

[View other drafts](#) ▾



✨ Sure, here is a Python code that solves the machine learning assignment of performing gradient descent for logistic regression:

Python

```
import numpy as np

def sigmoid(z):
    """Computes the sigmoid function of z."""
    return 1.0 / (1.0 + np.exp(-z))

def lrCostFunction(y, h):
    """Computes the logistic regression cost function for the given predictions h and true labels y."""
    m = len(y)
    return -np.sum(y * np.log(h) + (1 - y) * np.log(1 - h)) / m

def lrGradient(X, y, theta, alpha):
    """Computes the gradient of the logistic regression cost function with respect to the model parameters theta."""
    m = len(y)
    h = sigmoid(np.dot(X, theta))
    error = h - y
    grad = np.dot(X.T, error) / m
    return grad

def gradientDescent(X, y, theta, alpha, num_iters):
    """Performs gradient descent for logistic regression."""
    J_history = []
    for i in range(num_iters):
        grad = lrGradient(X, y, theta, alpha)
        theta -= alpha * grad
        J_history.append(lrCostFunction(y, sigmoid(np.dot(X, theta))))
```



The Bandwagon

CLAUDE E. SHANNON

INFORMATION theory has, in the last few years, become something of a scientific bandwagon. Starting as a technical tool for the communication engineer, it has received an extraordinary amount of publicity in the popular as well as the scientific press. In part, this has been due to connections with such fashionable fields as computing machines, cybernetics, and automation; and in part, to the novelty of its subject matter. As a consequence, it has perhaps been ballooned to an importance beyond its actual accomplishments. Our fellow scientists in many different fields, attracted by the fanfare and by the new avenues opened to scientific analysis, are using these ideas in their own problems. Applications are being made to biology, psychology, linguistics, fundamental physics, economics, the theory of organization, and many others. In short, information theory is currently partaking of a somewhat heady draught of general popularity.

subject are aimed in a very specific direction, a direction that is not necessarily relevant to such fields as psychology, economics, and other social sciences. Indeed, the hard core of information theory is, essentially, a branch of mathematics, a strictly deductive system. A thorough understanding of the mathematical foundation and its communication application is surely a prerequisite to other applications. I personally believe that many of the concepts of information theory will prove useful in these other fields—and, indeed, some results are already quite promising—but the establishing of such applications is not a trivial matter of translating words to a new domain, but rather the slow tedious process of hypothesis and experimental verification. If, for example, the human being acts in some situations like an ideal decoder, this is an experimental and not a mathematical fact, and as such must be tested under a wide variety of experimental situations.

Although this wave of popularity is certainly pleasant and exciting for those of us working in the field, it carries at the same time an element of danger. While we feel that information theory is indeed a valuable tool in providing fundamental insights into the nature of communication problems and will continue to grow in importance, it is certainly no panacea for the communication engineer or, *a fortiori*, for anyone else. Seldom do more than a few of nature's secrets give way at one time. It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized that the use of a few exciting words like *information*, *entropy*, *redundancy*, do not solve all our problems.



THE USEFULNESS OF USELESS KNOWLEDGE

BY ABRAHAM FLEXNER

Is it not a curious fact that in a world steeped in irrational hatreds which threaten civilization itself, men and women—old and young—detach themselves wholly or partly from the angry current of daily life to devote themselves to the cultivation of beauty, to the extension of knowledge, to the cure of disease, to the amelioration of suffering, just as though fanatics were not simultaneously engaged in spreading pain, ugliness, and suffering? The world has always been a

mental problems. I have no quarrel with this tendency. The world in which we live is the only world about which our senses can testify. Unless it is made a better world, a fairer world, millions will continue to go to their graves silent, saddened, and embittered. I have myself spent many years pleading that our schools should become more acutely aware of the world in which their pupils and students are destined to pass their lives. Now I sometimes won-

Intro to ML and DL.

Traditional Types of ML

Supervised Learning.

- **Data (x,y)**
- **Goal: Learn mapping x->y**
- **Examples:** Classification, regression, image recognition, ...

Unsupervised Learning.

- **Data (x)**
- **Goal: Learn some underlying structure from x**
- **Examples:** Clustering, Density estimation, dimensionality reduction...
- Discover some underlying structure of data

Supervised Learning Setup

We are given a dataset D :

$$D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} \subseteq \mathbb{R}^p \times \mathcal{C}$$

- \mathbb{R}^p is the p-dimensional feature space
- \mathbf{x}_i is the input vector of the i-th sample
- y_i is the label of the i-th sample
- C is the label space

The samples are drawn from an **unknown** distribution $\mathcal{P}(X, Y)$

- **Goal:** learn a function h such that for a new pair $(\mathbf{x}, y) \sim \mathcal{P}$, we have that $h(\mathbf{x}) \approx y$?
- **Q:** How would we guess y if we knew the true distribution $\mathcal{P}(X, Y)$?

Training and testing error, generalization

- After selecting a hypothesis class, we want to find a **hypothesis** function that works well on our training data (D_{TR})

$$h^*(\cdot) = \operatorname{argmin}_{h(\cdot) \in \mathcal{H}} \frac{1}{|D_{\text{TR}}|} \sum_{(\mathbf{x}, y) \in D_{\text{TR}}} \ell(\mathbf{x}, y | h(\cdot)),$$

- We then evaluate our function on the **unseen test data** (D_{TE}),

$$\epsilon_{\text{TE}} = \frac{1}{|D_{\text{TE}}|} \sum_{(\mathbf{x}, y) \in D_{\text{TE}}} \ell(\mathbf{x}, y | h^*(\cdot)).$$

- Given many samples drawn i.i.d. from the same distribution P , the test loss above is an unbiased estimator of the true **generalization loss**

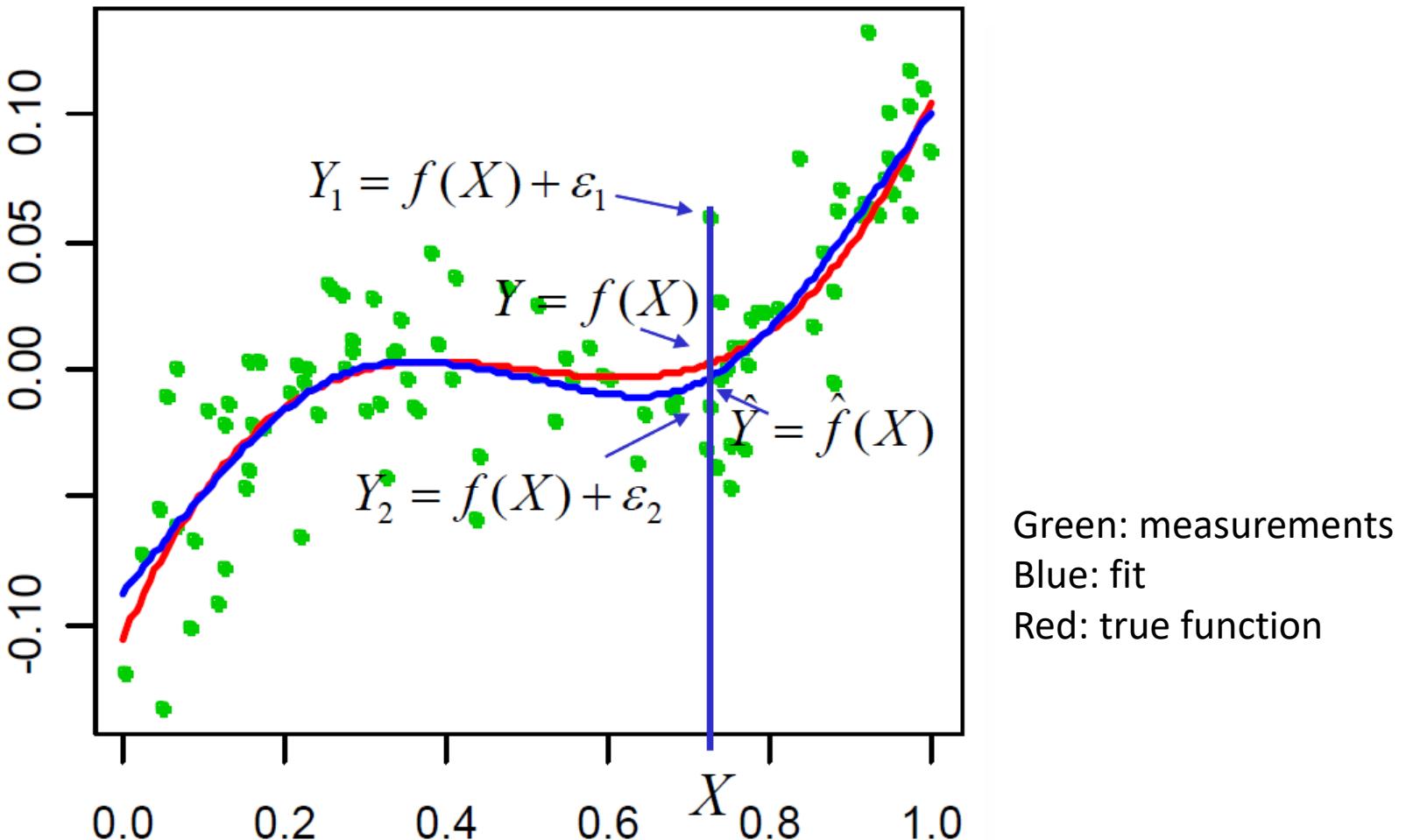
$$\epsilon = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{P}} [\ell(\mathbf{x}, y | h^*(\cdot))].$$

- When $|D_{\text{TE}}| \rightarrow \infty$, then $\epsilon_{\text{TE}} \rightarrow \epsilon$
 - ▶ Weak law of large numbers – empirical average of data drawn from a distribution converges to its expected value

Reducible and Irreducible error

$$E[(Y - \hat{Y})^2] = E[(f(X) + \epsilon - \hat{f}(X))^2] = [f(X) - \hat{f}(X)]^2 + \text{Var}(\epsilon)$$

Reducible Irreducible



Classification – Bayes optimal classifier

- Now, assume we already know $P(y|x)$ (which is almost never the case)
 - ▶ What would the optimal prediction be for a given x ?
- The Bayes optimal classifier predicts the most likely label using the conditional distribution

$$y^* = h_{\text{opt}}(\mathbf{x}) = \operatorname{argmax}_y P(y|\mathbf{x})$$

- Can the Bayes optimal classifier be wrong?
- Yes – when the sample label is not the most likely

$$\epsilon_{BayesOpt} = 1 - P(h_{\text{opt}}(\mathbf{x})|\mathbf{x}) = 1 - P(y^*|\mathbf{x}) \text{ (error rate)}$$

- E.g., when $P(+1|x)=0.8$ and $P(-1|x)=0.2$, the error would be 0.2

Supervised Learning is not the (only) way we learn

“When we’re learning to **see, nobody’s** telling us what the right answers are -we just **look**. Every so often, your mother says “**that’s a dog**”, but that’s very little information.

You’d be lucky if you got a **few bits** of information - even one bit per second - that way. The brain’s visual system has **10^{14}** neural connections. And you only live for **10^9** seconds.

So it’s no use learning one bit per second. You need more like **10^5** bits per second. And there’s only one place you can get that much information: from the **input** itself. “

Self-supervised Learning

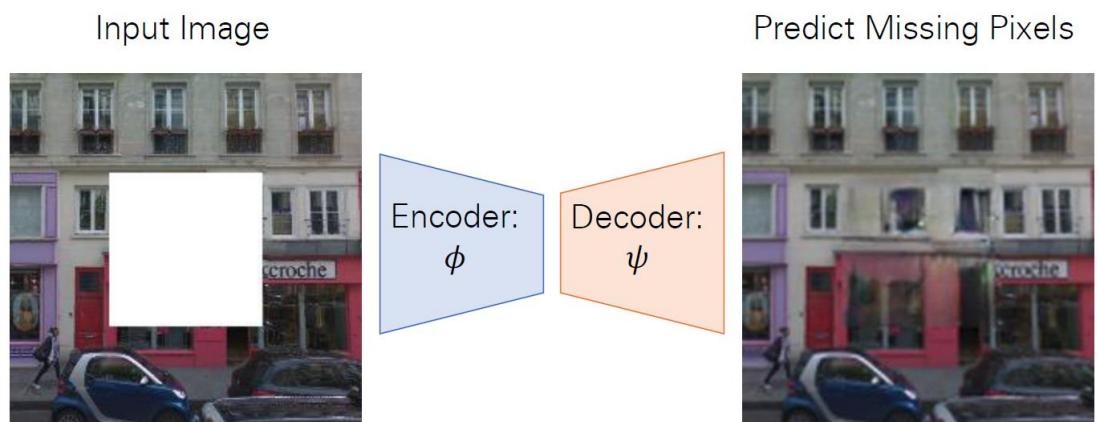
Methods that learn from ‘raw’ data (e.g., predict naturally-occurring signal in raw data rather than labels)

- E.g., Boston Dynamics robots predict physical world shortly in the future to adapt body inertia (and this is what experienced athletes do)

Intuition: Force the model to capture representations that are ‘semantically meaningful’

Example:

- Predict any part of the input from other parts
 - (e.g., future, past, present)
- Occlude parts of data, predict them



Discriminative, Generative, Foundation models

Discriminative Models

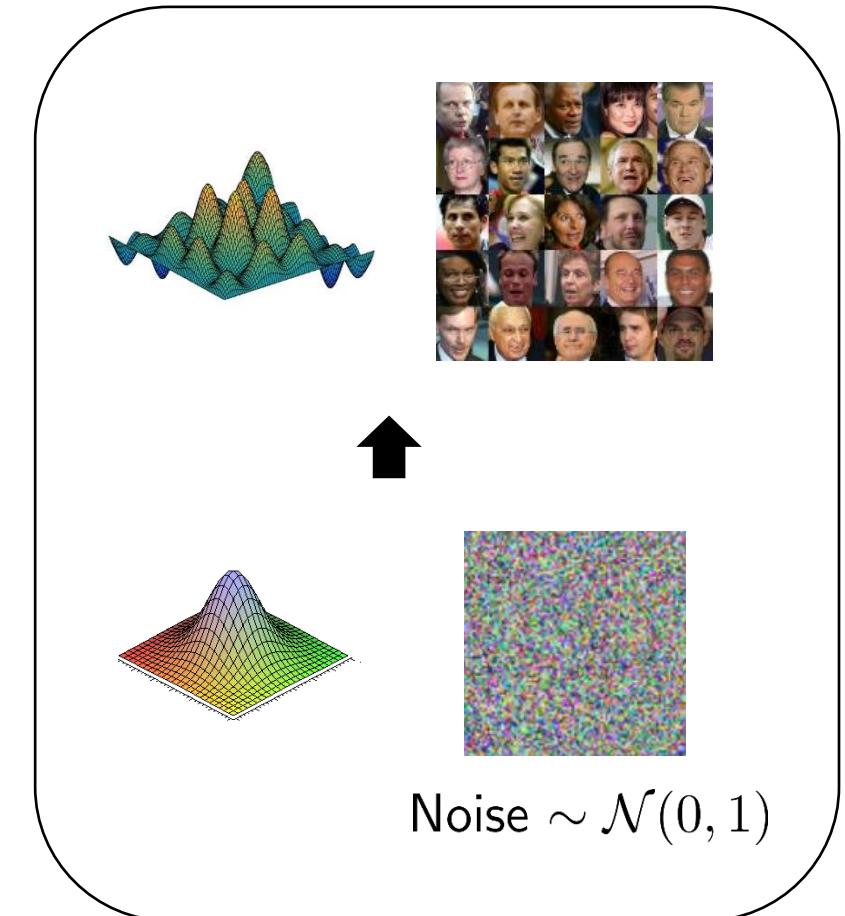
- Learn conditional probability $P(X|Y)$
- Usually supervised
- Good classifiers with less data

Generative Models

- Learn joint probability $P(X,Y)$
- Need more data but can synthesize data

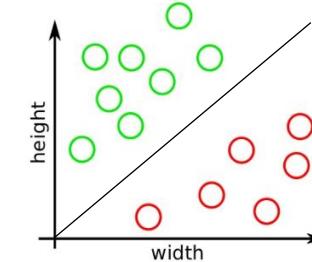
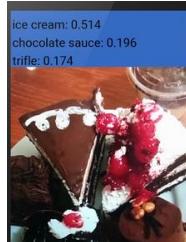
Foundation models: trained on very large datasets, adapted to downstream tasks (GPT, LLaMA, Stable Diffusion, DALL-E)

Architectures can be similar (modularity)

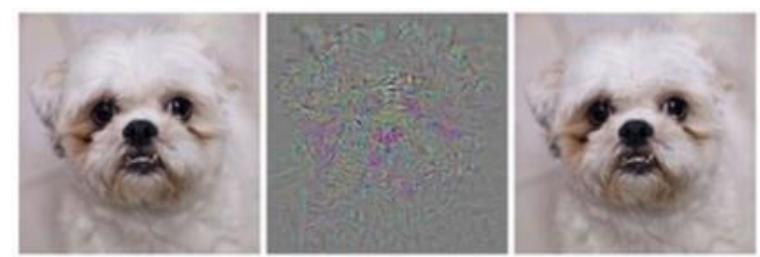
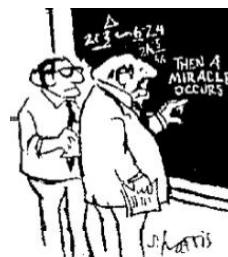


Still, many challenges and limitations

- Geometry-aware learning
- Reliability, Accuracy, Robustness
- Interpretability, Explainability, Uncertainty Quantification
- Learning invariances, deformations
- Geometry-aware learning
- Bias and fairness
- Efficiency
- ...

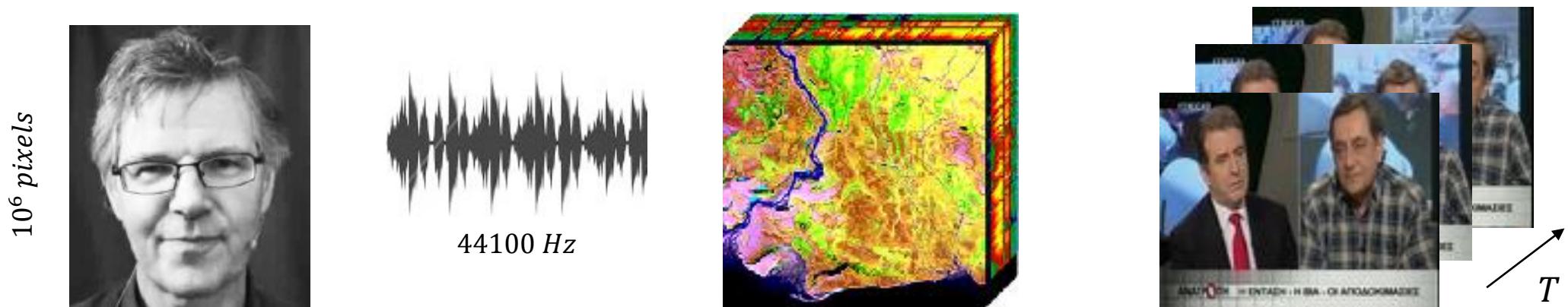


JAMES RIVELLI	ROBERT CANNON
Prior Offenses	Prior Offense 1 petty theft
Subsequent Offenses	Subsequent Offenses None
LOW RISK	MEDIUM RISK



Learning from High-Dim Data

Most interesting signals are high-dimensional



high intra-class variability

Learning from High-Dimensional Data

- Given a dataset with samples $\{\mathbf{x}_i, y_i\}$, estimate the function $f(\mathbf{x}_i) = y_i$ in such way that we can perform predictions on an unseen sample?
(regression, classification, ...)

Question

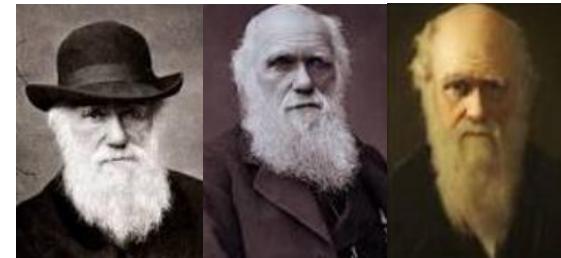
- Can we approximate function f by using data and local interpolation?

2.7×10^4



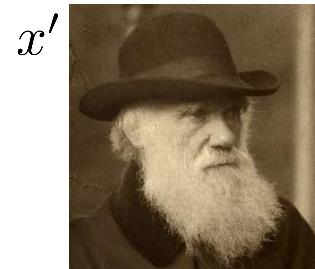
$$y_i = 1$$

7.5×10^4



$$y_i = 0$$

(regularity)?

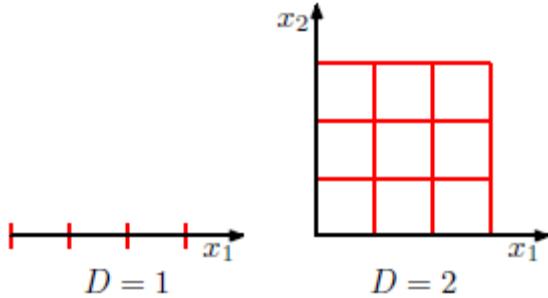


$$y' = ?$$

$$y' = f(\mathbf{x}_i^*), i^* = \operatorname{argmin}_i \|\mathbf{x}_i - \mathbf{x}'\|$$

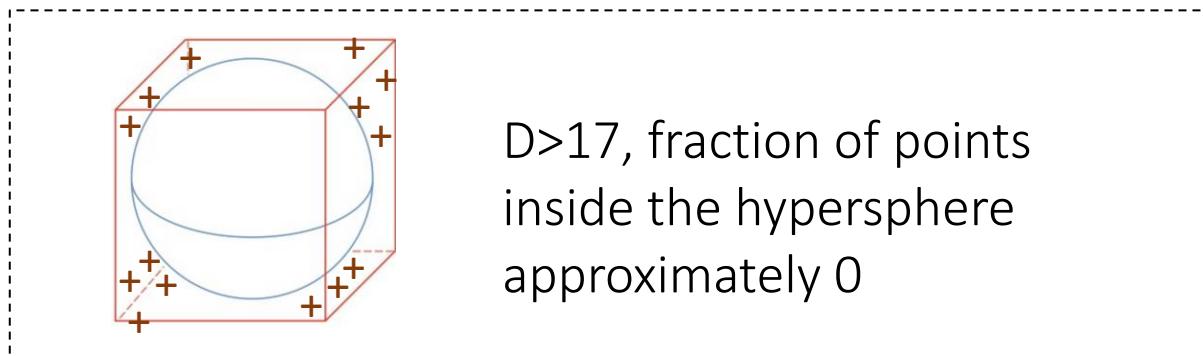
Learning from High-Dimensional Data

Curse of Dimensionality

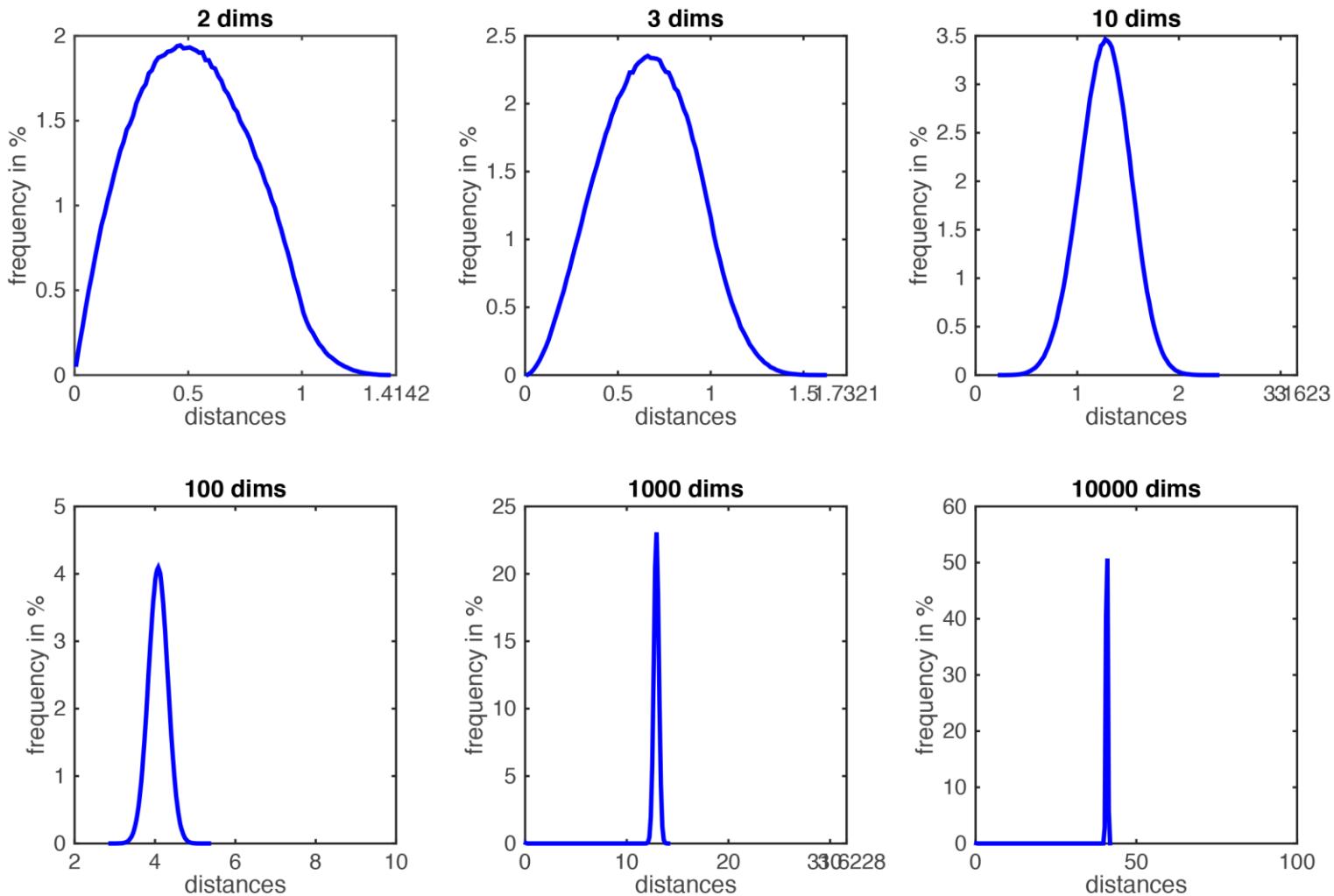


- 10 samples for space $[0,1]$ with $\varepsilon=0.1$
- 10^2 samples for space $[0,1]^2$
- 10^D samples for space $[0,1]^D$
- $D=100?$ $D=10^6?$ (atoms universe $\approx 10^{80}$)

- We need an exponential amount of data wrt. dimensionality!
- There are no close neighbours in high dimensions (breaking locality)
- The distance between samples is always large

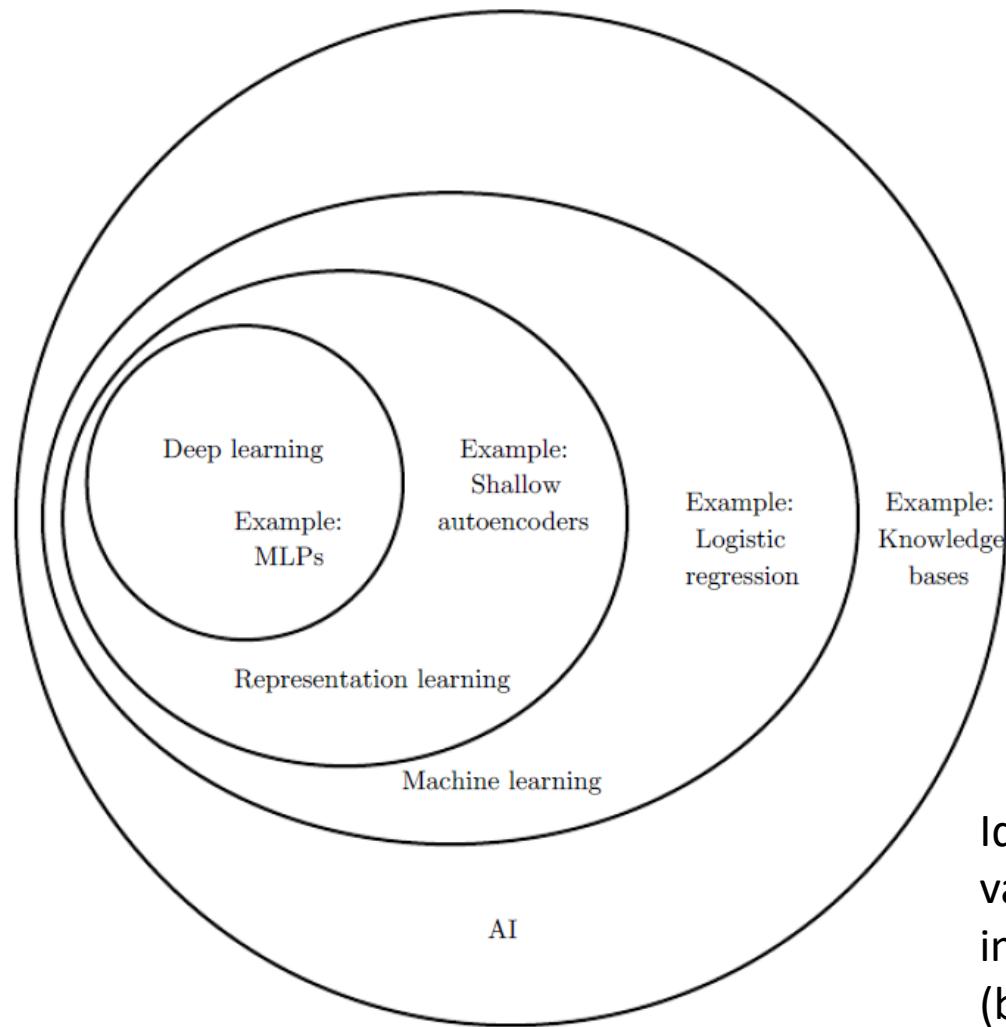


Curse of dimensionality



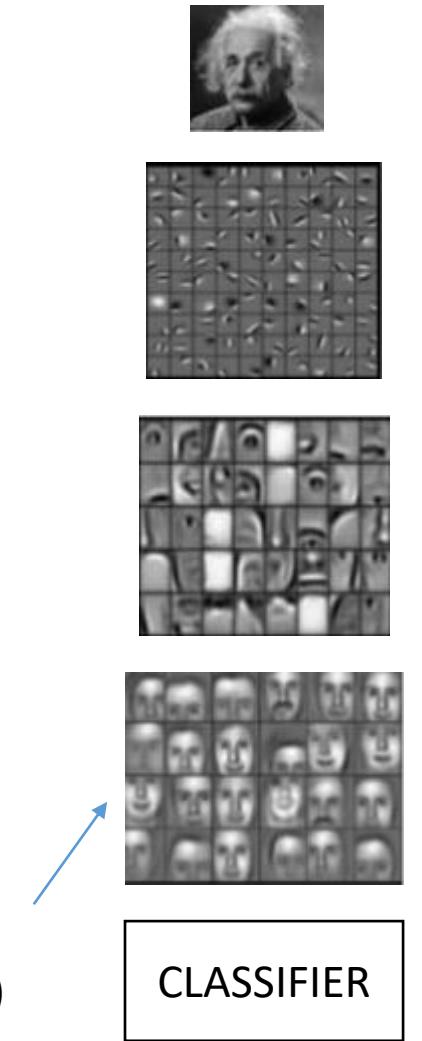
Distributions of pairwise distances between randomly distributed points with a growing number of dimensions

Deep Learning is a kind of representation learning



Deep Learning (Goodfellow, Bengio, Courville)

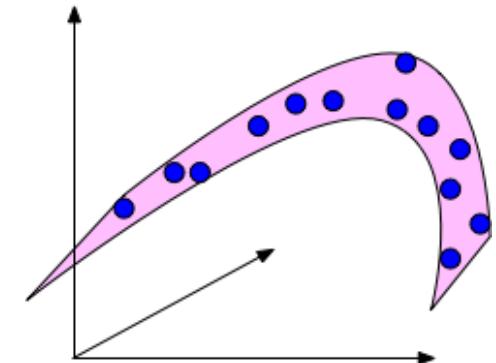
Ideally, kills variation that is invariant to task (build invariances)



Discovering (Latent) Representations

- Instead of working directly on high-dimensional data space X , introduce **latent representation** z_i (usually compressed/dimensionality reduction)

$$\begin{array}{ccccc} \mathbf{x}_i & \xrightarrow{g(\mathbf{x}_i)} & \mathbf{z}_i & \xrightarrow{\hat{g}(\mathbf{x}_i)} & \hat{\mathbf{x}}_i \\ \mathbf{x}_i \in \mathbb{R}^D \rightarrow \mathbf{z} \in \mathbb{R}^d, d \ll D, \text{s.t.} \dots & & & & \end{array}$$



$$\min ||\mathbf{X} - \hat{g}(g(\mathbf{X}))|| \quad \text{← Reconstruction loss}$$

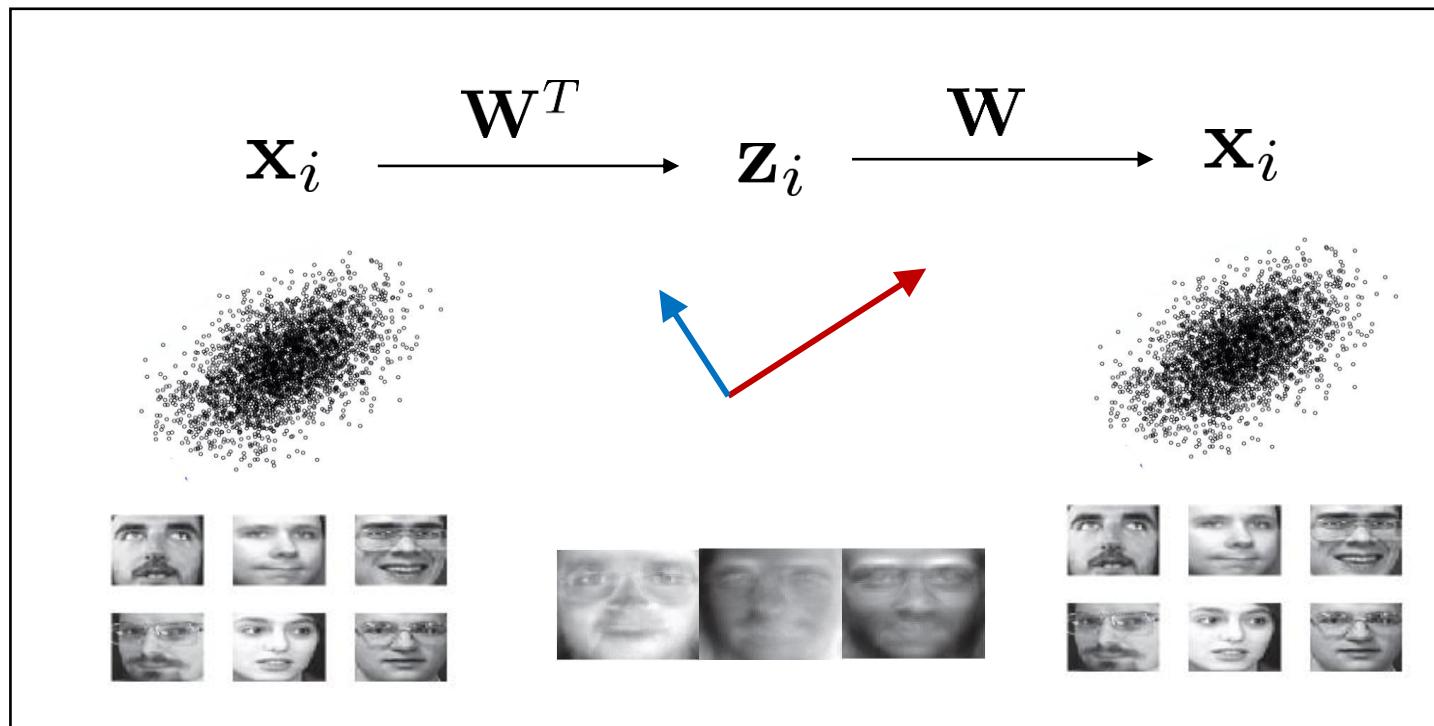
- Eliminate nuisance factors, capture useful signal properties
- Reduced dimensionality of z_i ensures non-trivial solution, but this can also be enforced otherwise
- (e.g., images with human faces are high-dimensional. However, a low-dimensional representation well captures the variation)

Discovering Latent Representations

Principal Component Analysis (PCA)

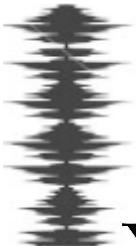
- Dataset $\{\mathbf{x}_i, y_i\}$. Instead of learning $f(\mathbf{x}_i) = y_i$ directly, introduce a latent representation \mathbf{z}_i that **maximizes variance** in the latent projected space.

$$\mathbf{x}_i \in \mathbb{R}^D \rightarrow \mathbf{z} \in \mathbb{R}^d, d \ll D, \text{s.t.} \dots$$



Discovering Shared Representations

Canonical Correlation Analysis



\mathbf{X}_1

\mathbf{X}_2

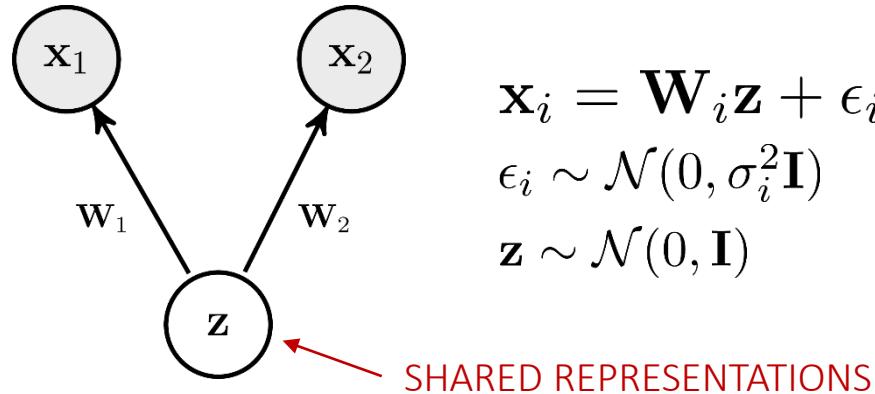
In many problems, data may come from multiple views or domains (e.g., audio-visual, multi-sensory, infrared, RGB...)

Probabilistic Interpretation

Deterministic

$$\begin{aligned} \max & (\mathbf{W}_1^T \Sigma_{12} \mathbf{W}_2) = \\ \min & \|\mathbf{W}_1^T \mathbf{X}_1 - \mathbf{W}_2^T \mathbf{X}_2\|_F^2 \\ \text{s.t. } & \mathbf{W}_i^T \Sigma_{ii} \mathbf{W}_i = \mathbf{I} \end{aligned}$$

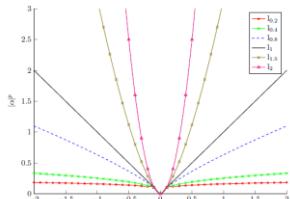
Probabilistic View – Generative Model



Several DL variants

Discovering Latent Representations

From Principal Component Analysis (PCA) to autoencoders (AE)

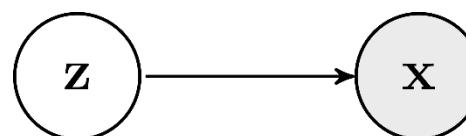
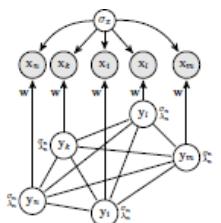


$$\mathbf{x} \xrightarrow{\mathbf{W}^T} \mathbf{z} \xrightarrow{\mathbf{W}} \tilde{\mathbf{x}}$$

$$\begin{aligned} \max (\mathbf{W}^T \Sigma \mathbf{W}) = \\ \min \|\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\| \\ \text{s.t. } \mathbf{W}^T \mathbf{W} = \mathbf{I} \end{aligned}$$

PCA

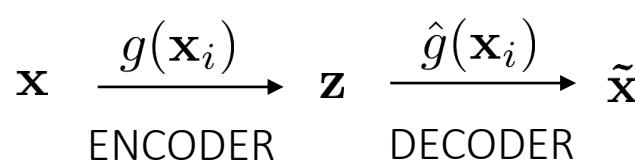
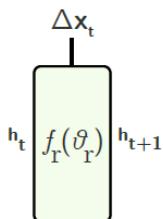
Optimization



$$\begin{aligned} \mathbf{x}|z \sim \mathcal{N}(\mathbf{W}z + \mu, \sigma^2 \mathbf{I}) \\ z \sim \mathcal{N}(0, \mathbf{I}), \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \end{aligned}$$

PPCA

Probabilistic View



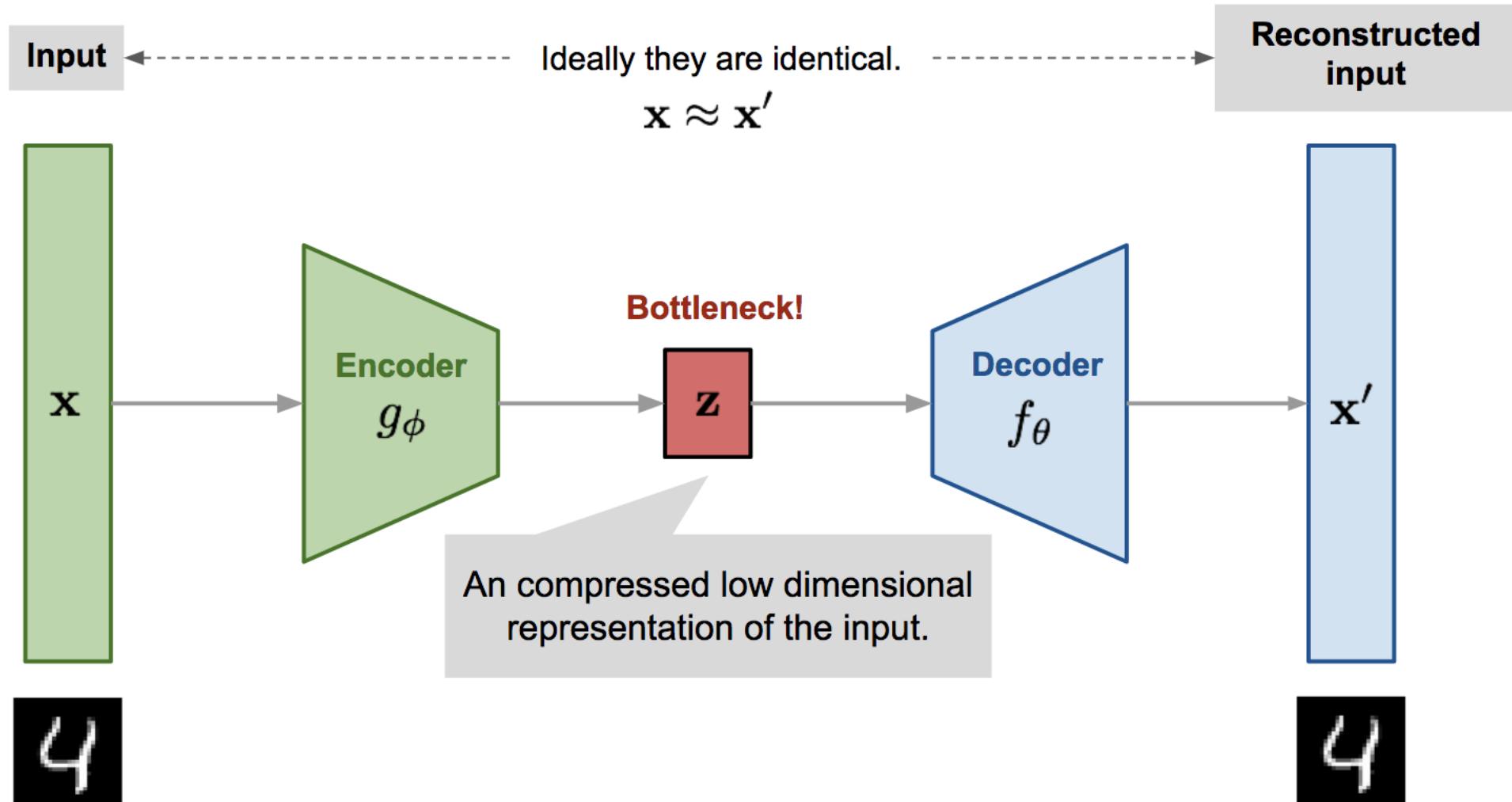
Deep Learning

$$\begin{aligned} \min \|\mathbf{X} - g(\hat{g}(\mathbf{X}))\| \\ \text{where } g(\mathbf{x}_i), \hat{g}(\mathbf{x}_i) \text{ are NN} \\ \text{(and can be CNNs / Deep)} \end{aligned}$$

AE

Properties: interpretability, scalability, robustness, generalization, capacity/DoF

Autoencoder

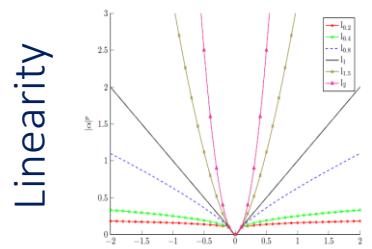


Discovering Latent Representations

- Instead of working directly on high-dimensional data space X , introduce **latent representation** z_i (usually compressed)

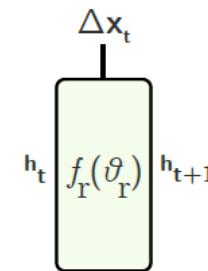
$$\mathbf{x}_i \xrightarrow{g(\mathbf{x}_i)} \mathbf{z}_i \xrightarrow{\hat{g}(\mathbf{x}_i)} \mathbf{x}_i$$
$$\mathbf{x}_i \in \mathbb{R}^D \rightarrow \mathbf{z} \in \mathbb{R}^d, d \ll D, \text{s.t.} \dots$$

Principal Component Analysis (PCA, 1910)



$$\mathbf{x} \xrightarrow{\mathbf{W}^T} \mathbf{z} \xrightarrow{\mathbf{W}} \tilde{\mathbf{x}}$$
$$\min \|\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\|$$
$$\text{s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I}$$

Deep Learning



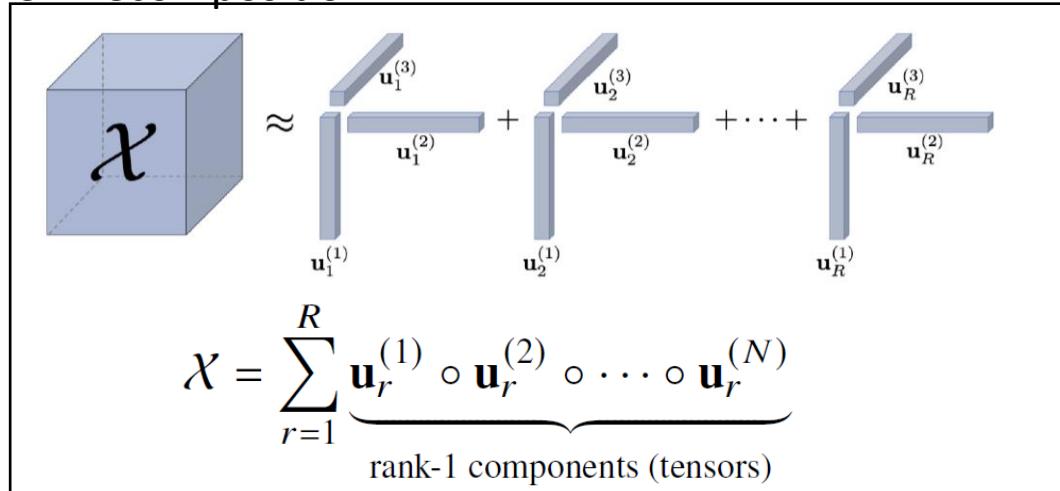
Autoencoder

$$\mathbf{x} \xrightarrow{\text{ENCODER}} \mathbf{z} \xrightarrow{\text{DECODER}} \tilde{\mathbf{x}}$$
$$g(\mathbf{x}_i) \xrightarrow{\hat{g}(\mathbf{x}_i)} \tilde{\mathbf{x}}$$
$$\min \|\mathbf{X} - \hat{g}(g(\mathbf{X}))\|$$

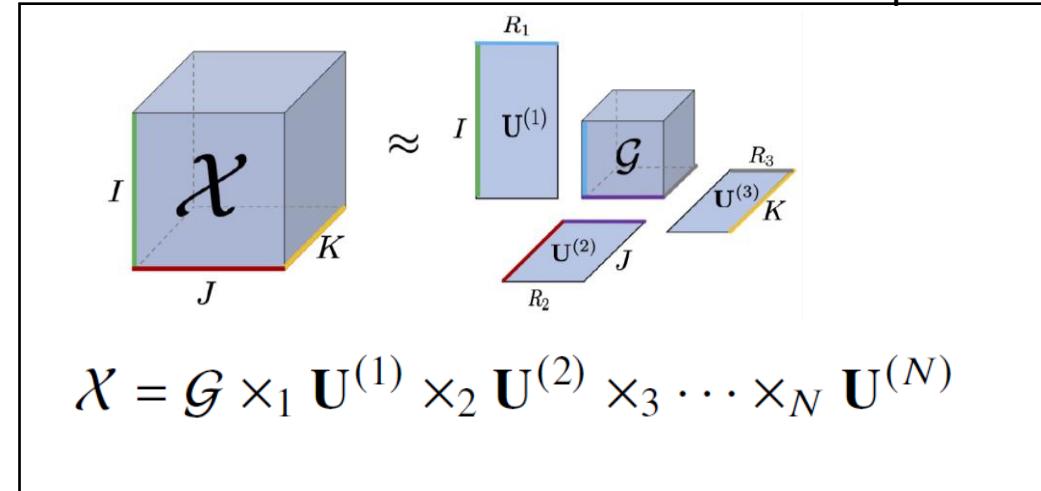
Tensor (Matrix) Methods & Deep Learning

- Multidimensional arrays;
 - ▶ *generalize* the concept of a matrix to multiple dimensions.
- Naturally **represent** the structure of multi-dimensional data and mappings
 - ▶ E.g. RGB images, brain scans, longitudinal data, hyperspectral data
 - ▶ also, intermediate activations and latent representations of deep networks, convolutional kernels, attention mechanisms

CP Decomposition



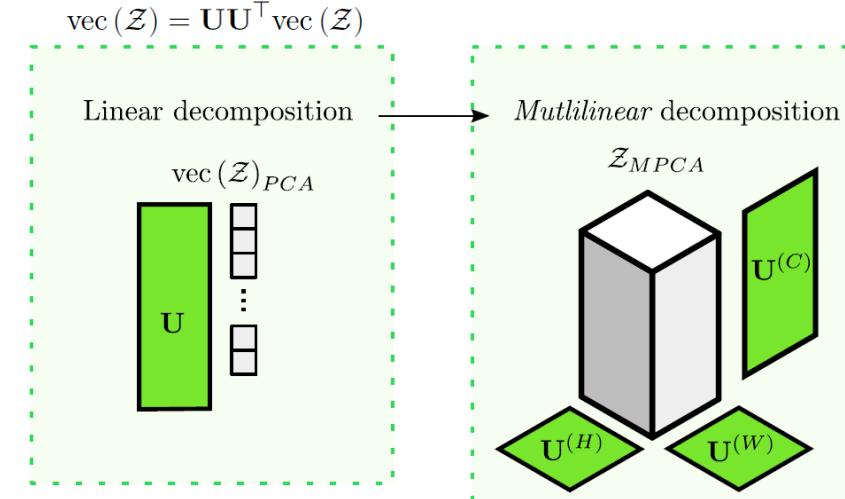
Tucker Decomposition



Tensor Methods & Deep Learning

github.com/tensorly/Proceedings_IEEE_companion_notebooks

- Multidimensional arrays;
 - ▶ *generalize* the concept of a matrix to multiple dimensions.
- Naturally **represent** the structure of multi-dimensional data and mappings
 - ▶ E.g. RGB images, brain scans, longitudinal data, hyperspectral data
 - ▶ also, intermediate activations and latent representations of deep networks, convolutional kernels, attention mechanisms
- Favourable properties:
 - ▶ Structure-preserving mappings
 - ▶ High-order (multiplicative) interactions
 - ▶ Robust and regularized models
 - ▶ Model Interpretability
 - ▶ Efficient & Lightweight networks



Deep Learning

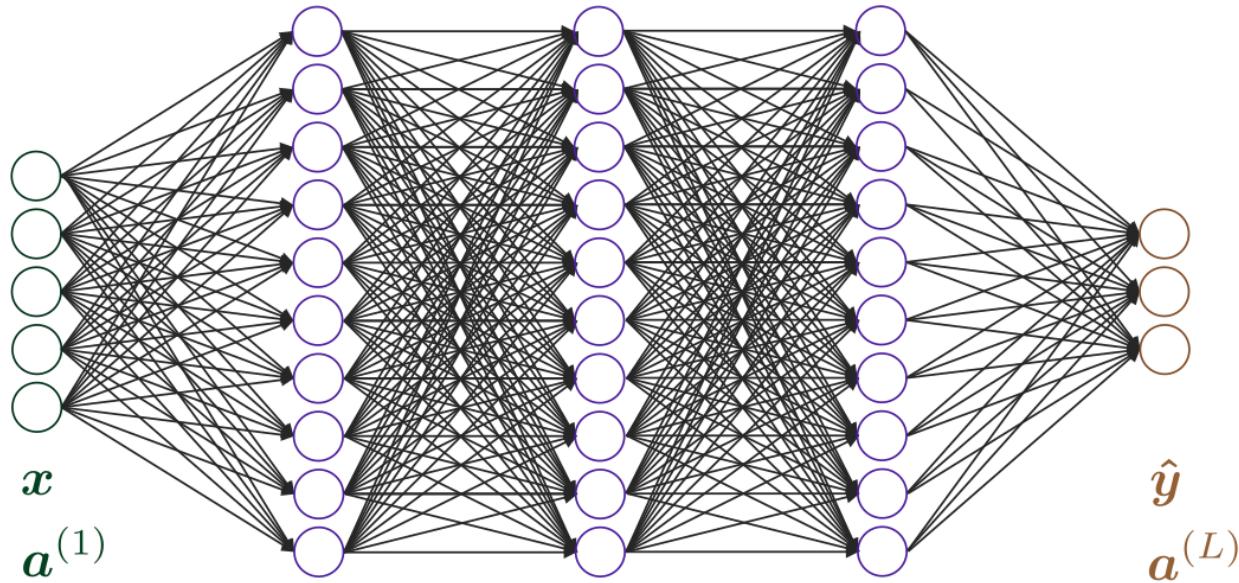
Why does DL work on natural signals?

- **Audio** signals can be represented as 1D vectors over time
- **Images** are 2D signals – however each point can be a vector
 - ▶ E.g., for RGB images, the size of this vector is 3
- **Language** can be represented by 1-hot vectors
 - ▶ Each vector is the size of a vocabulary

Some properties of natural signals

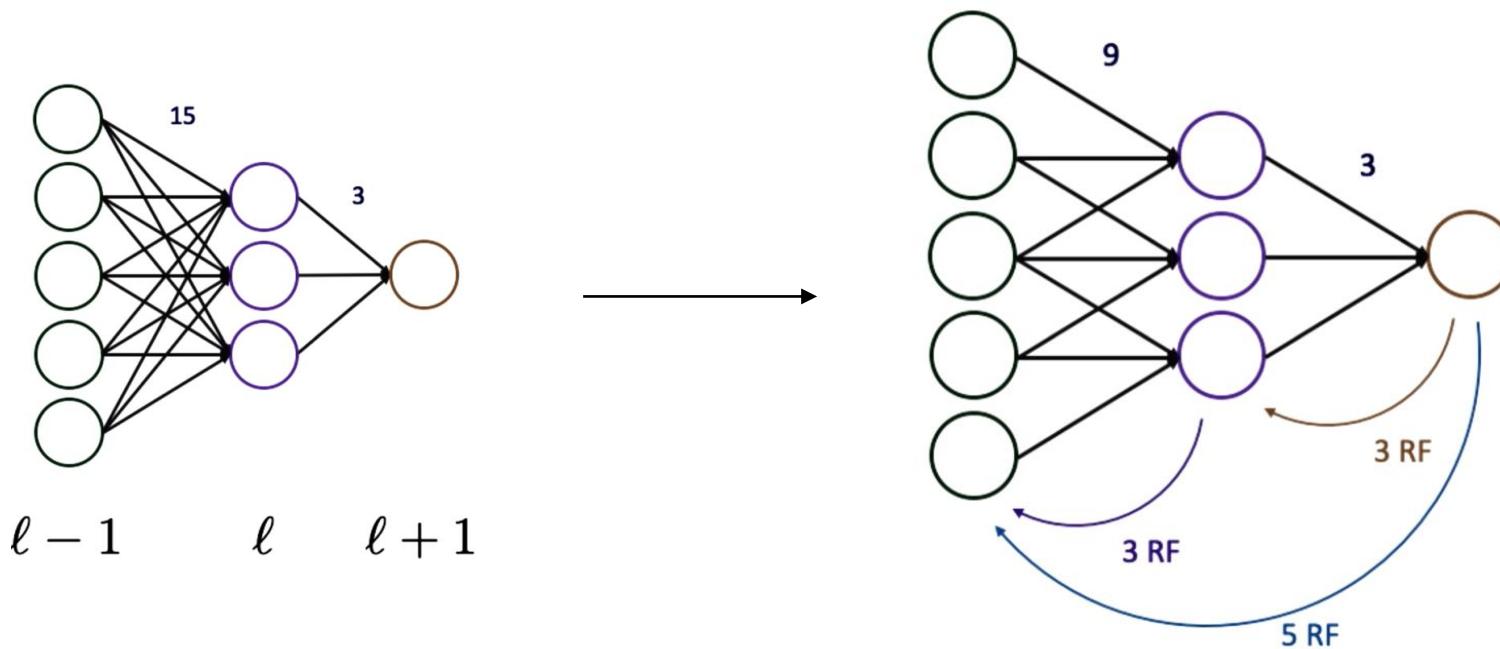
- **Stationarity:** Patterns repeated throughout a signal
 - ▶ Audio: repeating patterns (e.g., peaks) across time
 - ▶ Images: similar visual patterns across the spatial dimensions
- **Locality:** Neighbours are more related (correlated) than other points
 - ▶ Audio: if a peak is observed, then more likely that neighbouring points will also have high values
 - ▶ Images: If a pixel at a position has a specific colour, one would expect that neighbouring pixels are more likely to have the same colour
- **Compositionality:** Signals may be composed of parts/sub-parts
 - ▶ E.g., words form sentences, and sentences documents

Locality \Rightarrow Sparsity



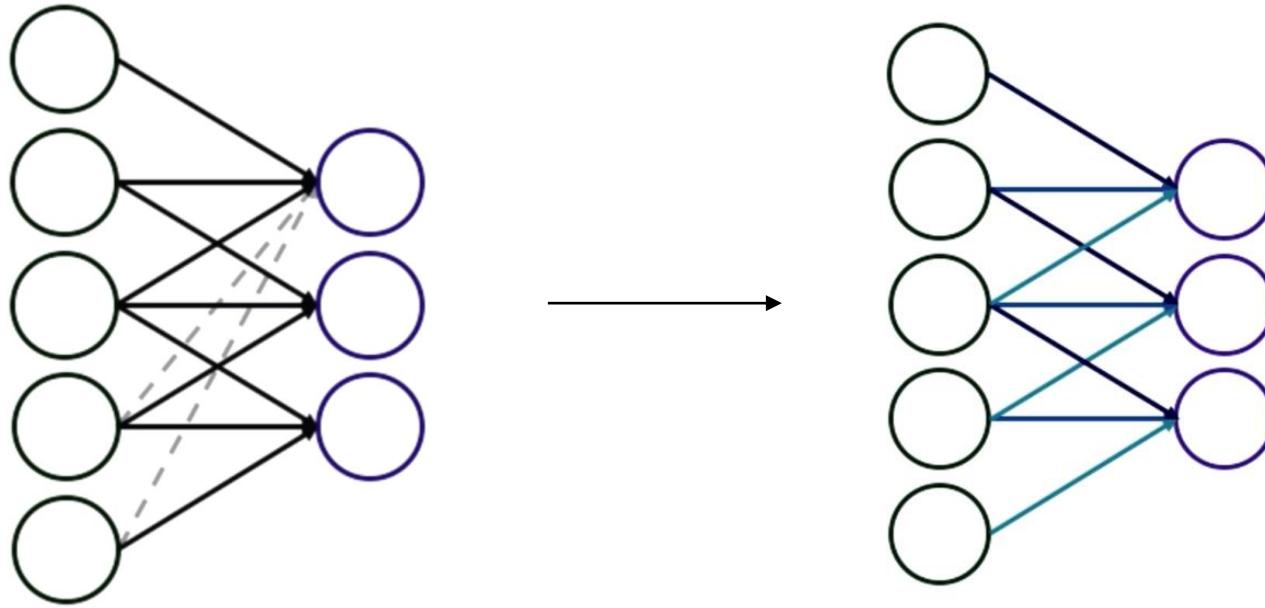
- Imposing sparsity in NN means dropping connections to far away neurons instead of having a fully connected network

Locality \Rightarrow Sparsity



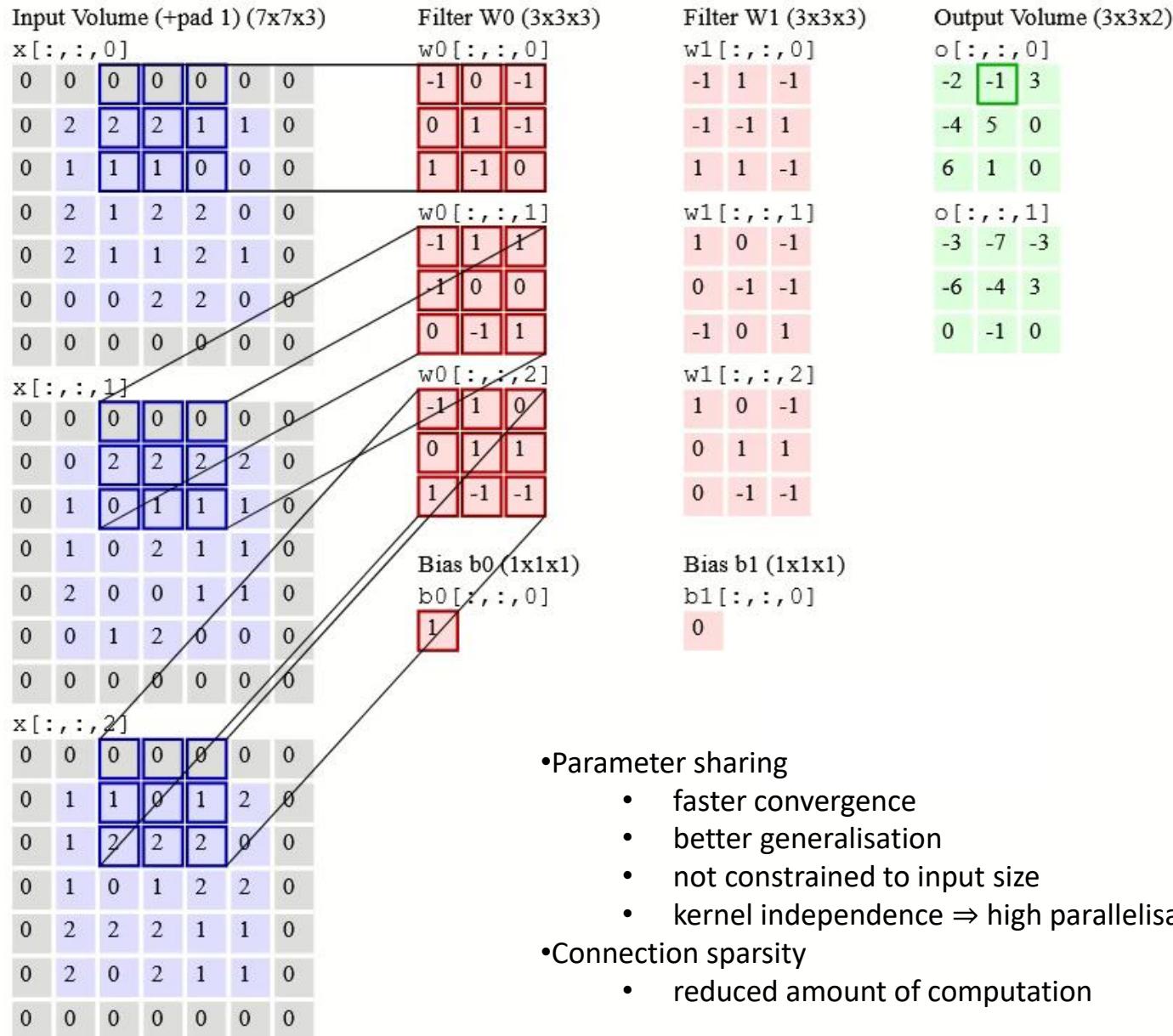
- Hidden layer neurons do not span whole input
 - ▶ However the overall architecture is able to
- **Receptive Field (RF):** Number of neurons in a previous layer that are visible or taken into account.

Stationarity \Rightarrow parameter sharing



- **Parameter sharing:** use a small set of parameters multiple times across the architecture
- Above network: reduces free parameters from 9 to 3
 - ▶ Same amount of data, less parameters to train

Sparsity + Param. Sharing \Rightarrow Convolutional Kernels



$$Y(i, j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(i + m, j + n) \cdot K(m, n)$$

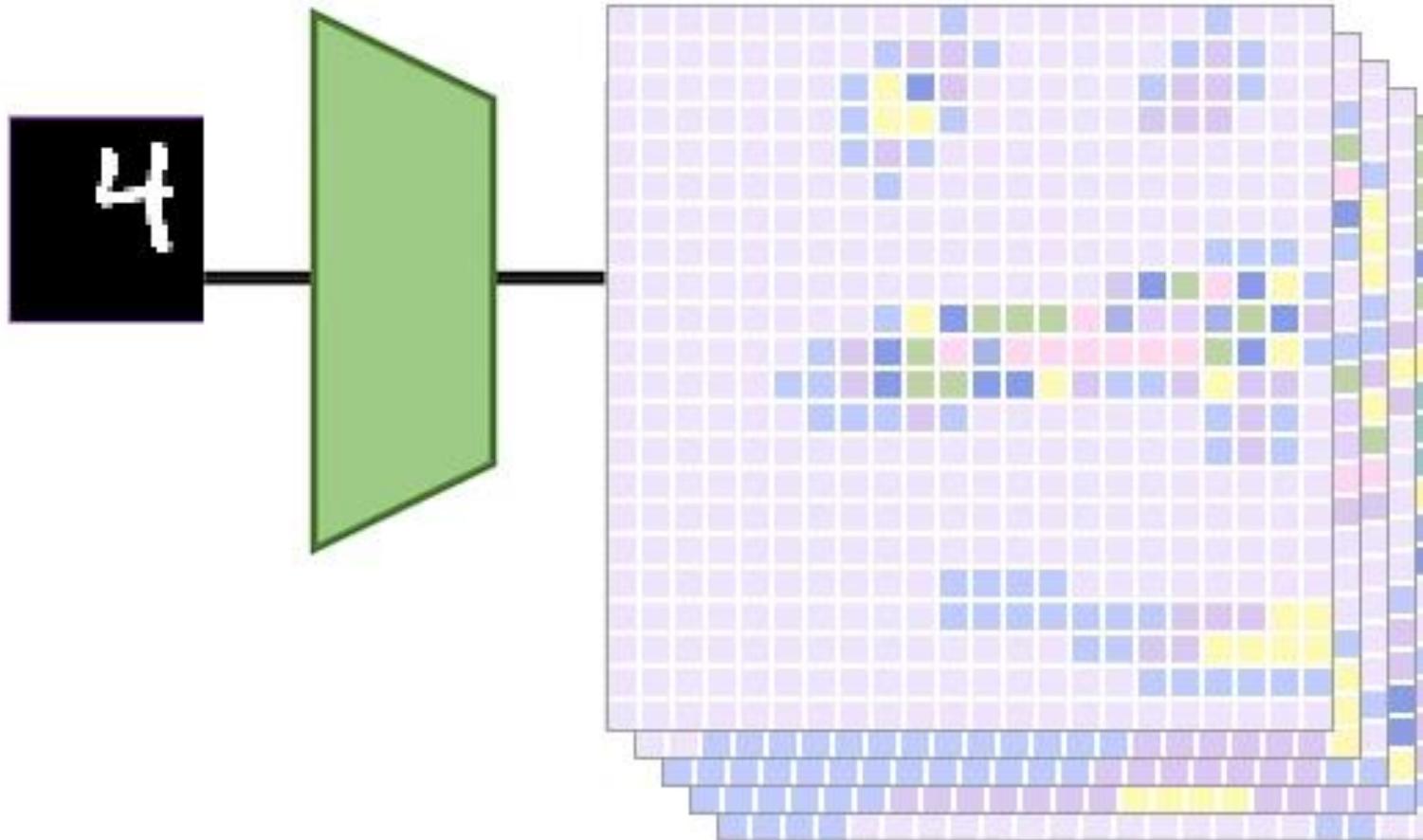
- Parameter sharing

- faster convergence
- better generalisation
- not constrained to input size
- kernel independence \Rightarrow high parallelisation

- Connection sparsity

- reduced amount of computation

Translation Equivariance

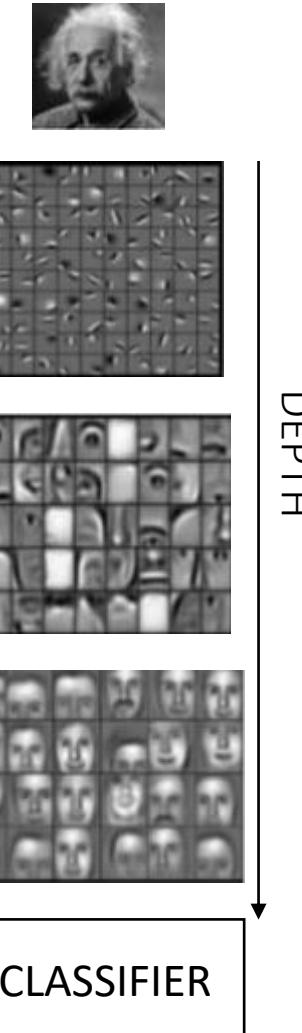
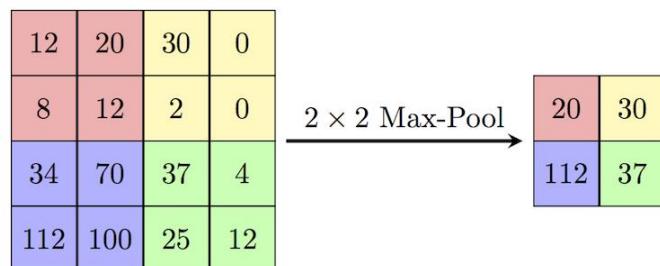


Compositionality

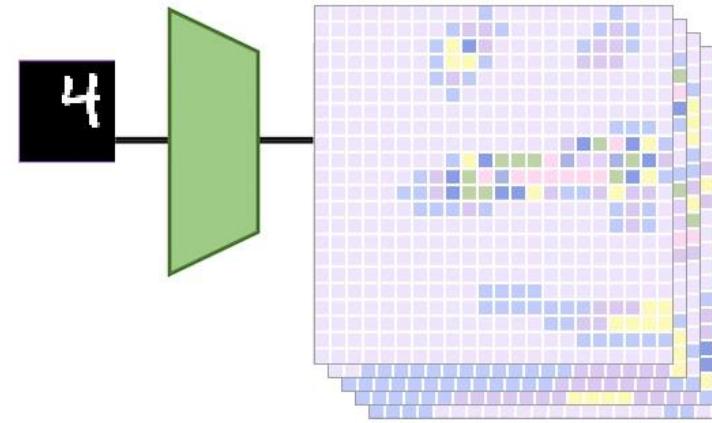
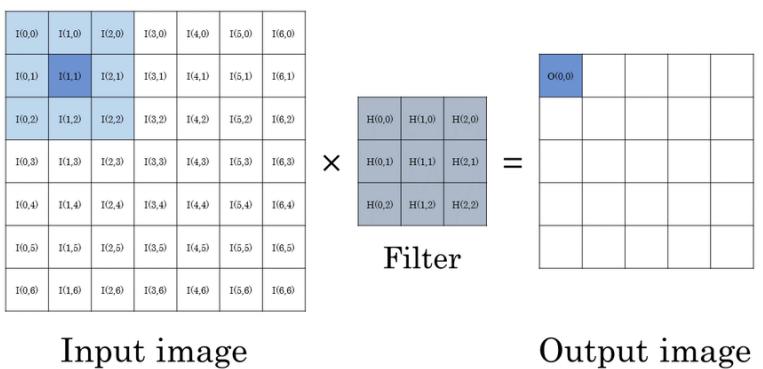
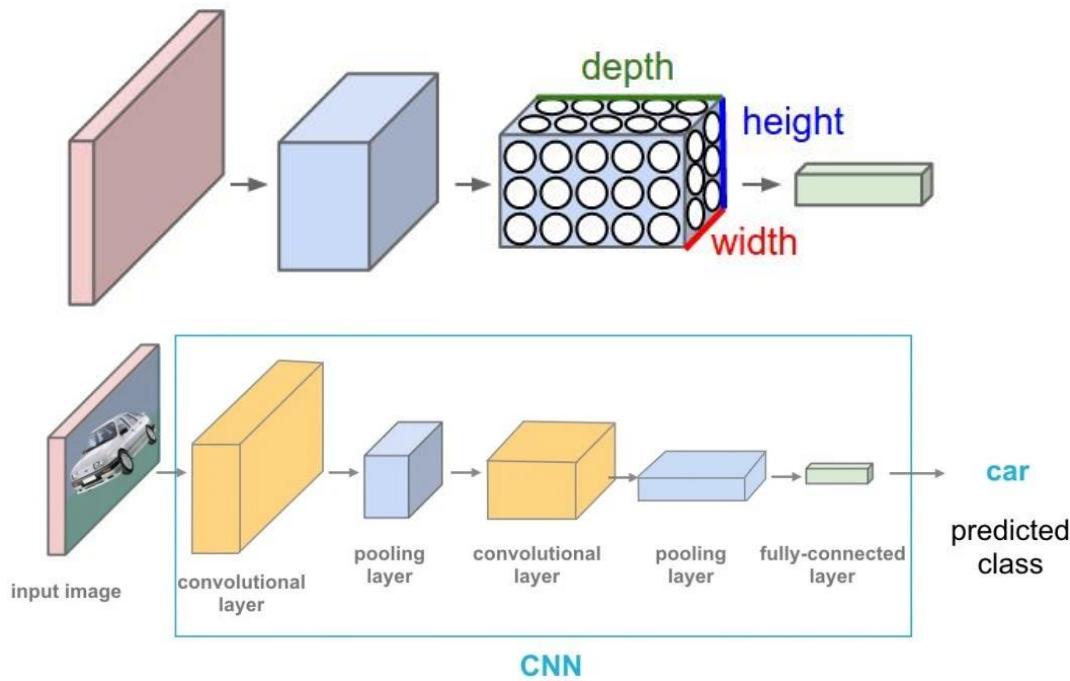
- Deep (convolutional or not) networks are inherently modular

- We are learning representations in a hierarchical fashion

- Multiple layers include:
 - ▶ Convolutional kernels
 - ▶ Non-linear functions (e.g., $\text{ReLU}(z)=\max(0,z)$)
 - ▶ Pooling layers
 - ▶ Normalization layers
 - ▶ Residual connections



Summary: Convolutional Networks



Some favourable properties (of conv kernels) related to vision and natural signals:

- Translation Equivariance
- Parameter sharing (fast, parallelizable, generalization)
- Connection sparsity (reduced amount of regularization), Locality

Vision Transformer (ViT, 2021)

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

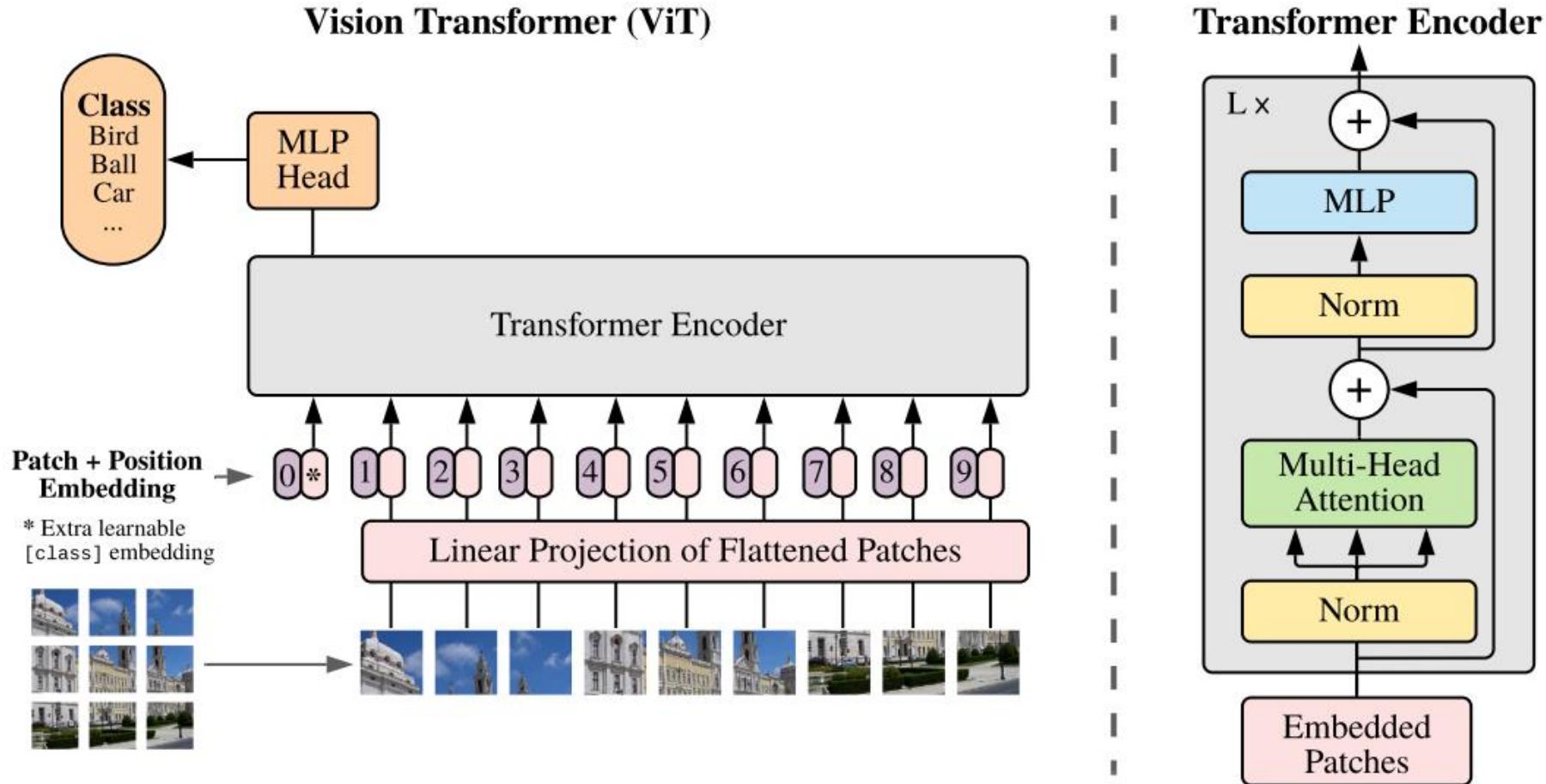
**Alexey Dosovitskiy^{*,†}, Lucas Beyer^{*}, Alexander Kolesnikov^{*}, Dirk Weissenborn^{*},
Xiaohua Zhai^{*}, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby^{*,†}**

^{*}equal technical contribution, [†]equal advising

Google Research, Brain Team

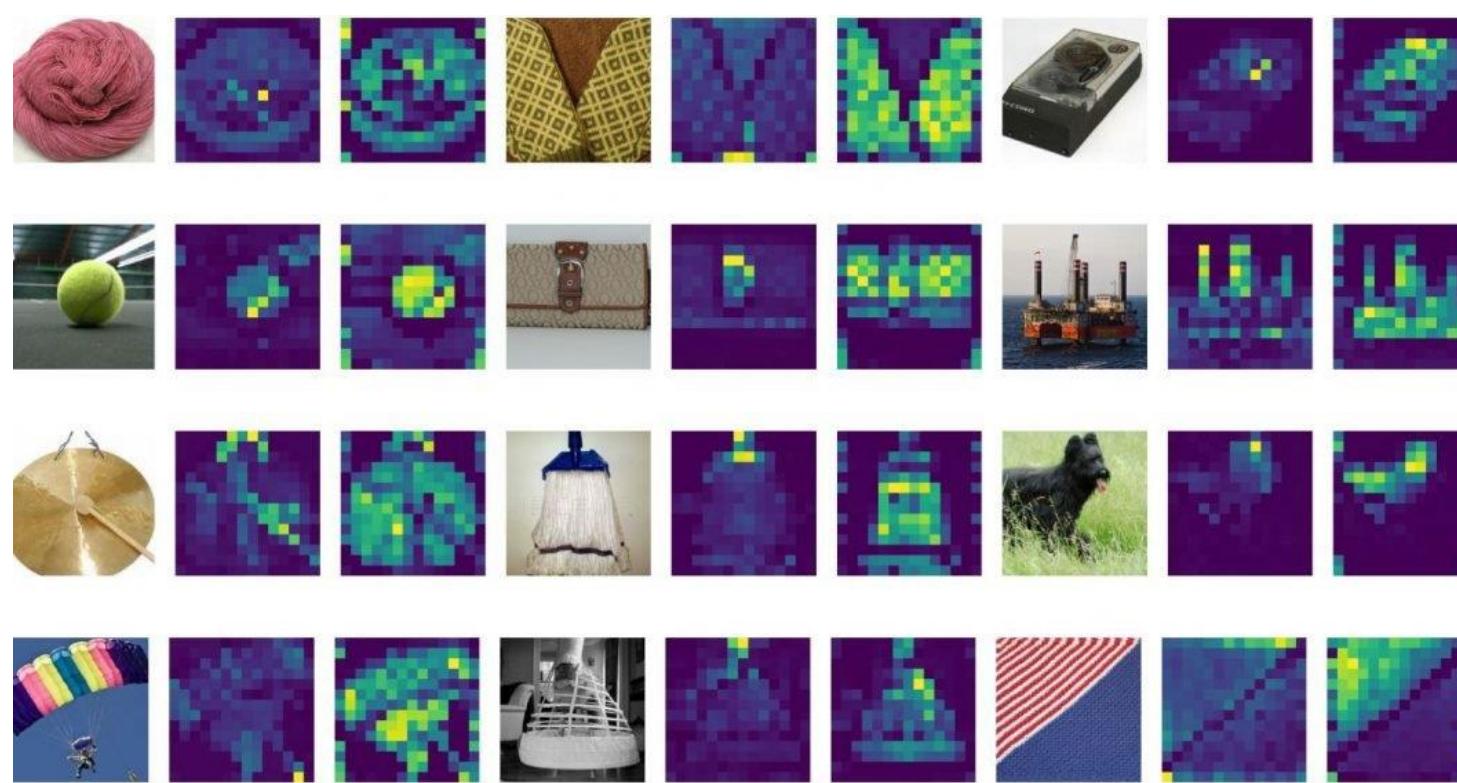
{adosovitskiy, neilhoulsby}@google.com

Vision Transformer

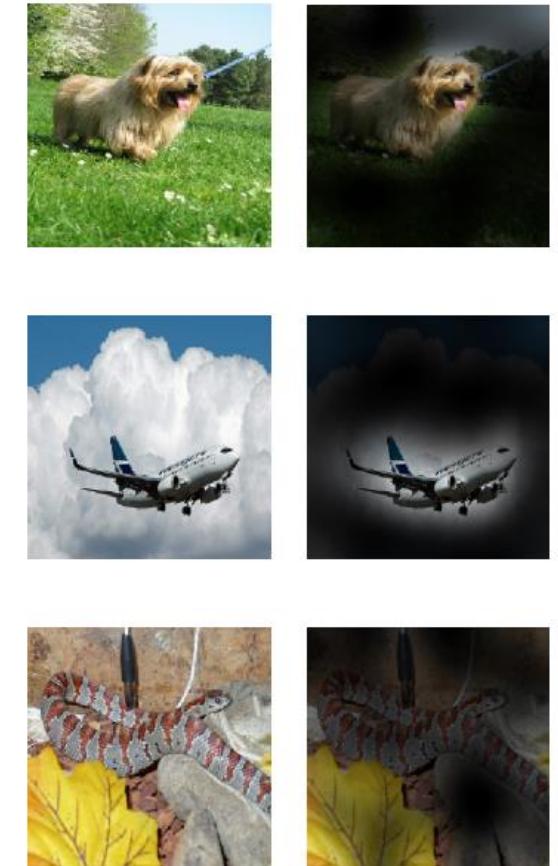


Can also use CNNs to extract patch representations

Vision Transformer



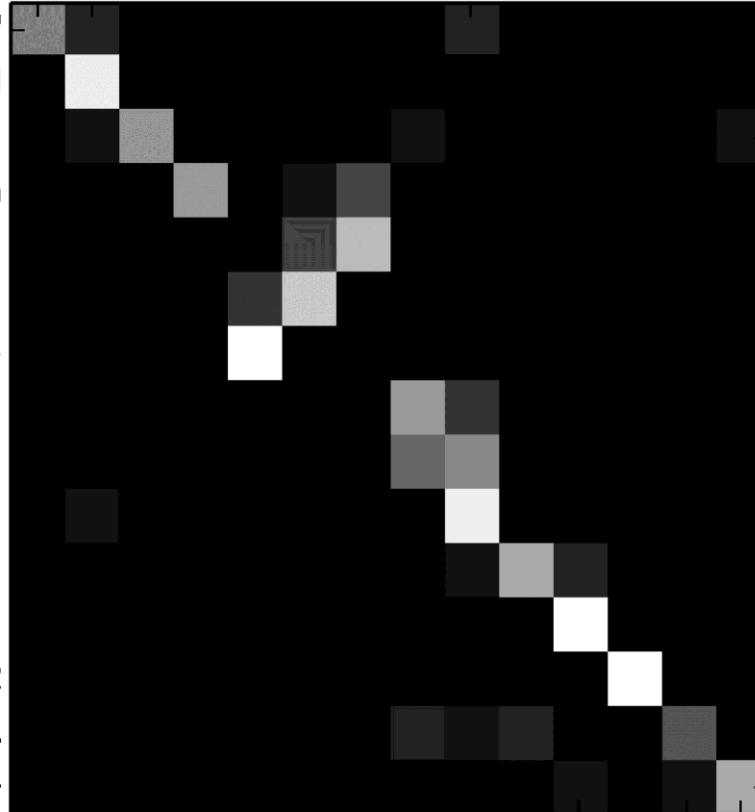
Input Attention



- Global representation from shallow layers, local info also important
- ViT considered to retain more spatial information than ResNet
- High-quality latent representations with large-amounts of data

The
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
. <end>

L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
. <end>



Vision Transformer (ViT) vs CNN

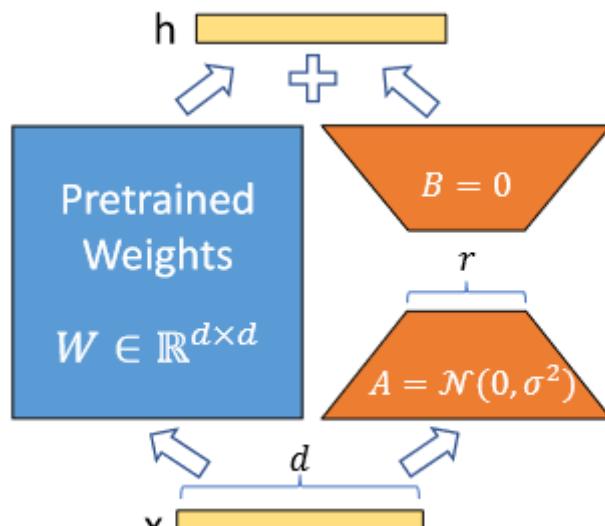
- The inductive biases of CNNs favour visual and other natural data (e.g., locality sensitive, translation equiv)
 - ▶ Learn faster, but less representational power (can get ‘stuck’)
 - ▶ Not as good in capturing global relationships, really good in local ones
- Attention-based architectures represent larger family of functions than convolutions
 - ▶ Need more data (need to learn these biases from data)
 - ▶ Better able to learn dependencies
 - ▶ Can outperform CNNs, and be more robust (occlusion, distribution shift, adversarial attacks)

Rule of thumb: ViT trained > 10-14M data can outperform CNN-based architectures.
Otherwise use ResNet/EfficientNet

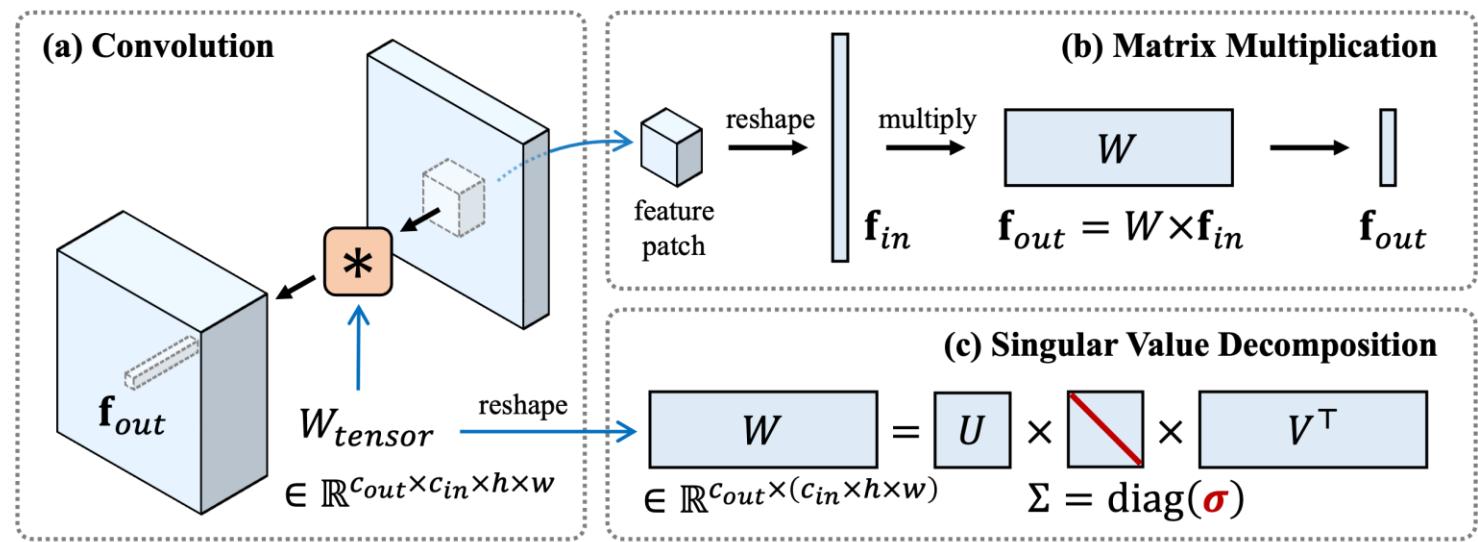
Matrix/Tensor Factorizations in Deep Learning

Parameter-efficient Learning

- Most parameter-efficient methods (e.g., LoRA, SVDiff) are based on low-rank decompositions (factorizations), e.g., SVD.

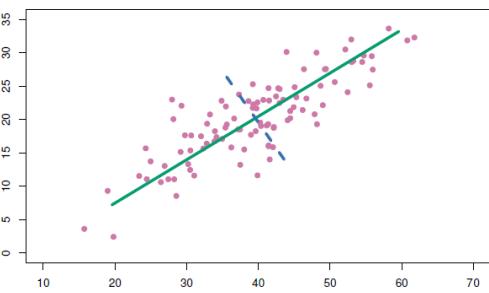


LoRA

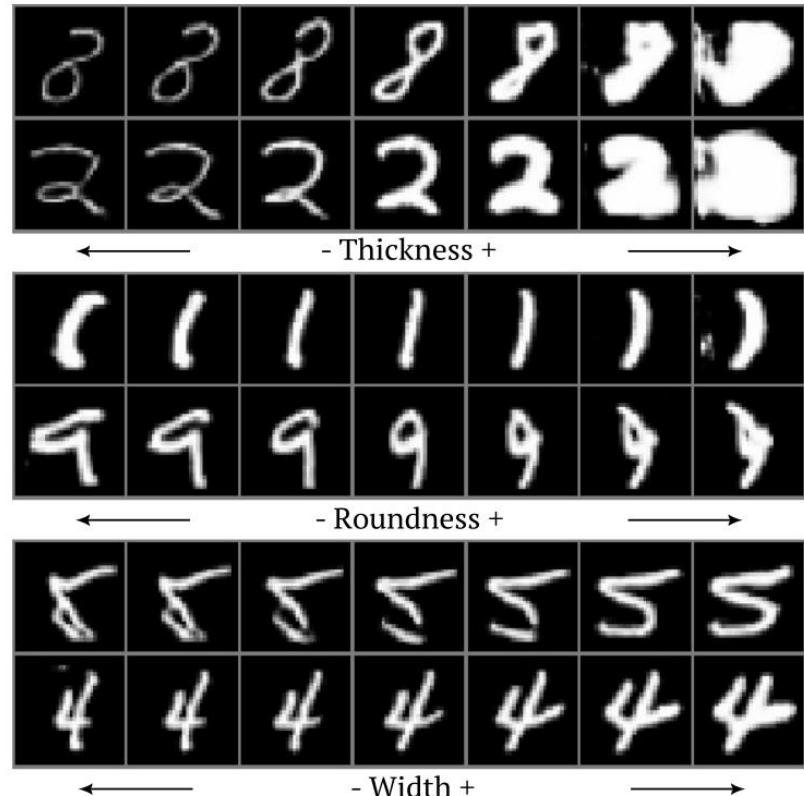


SVDiff

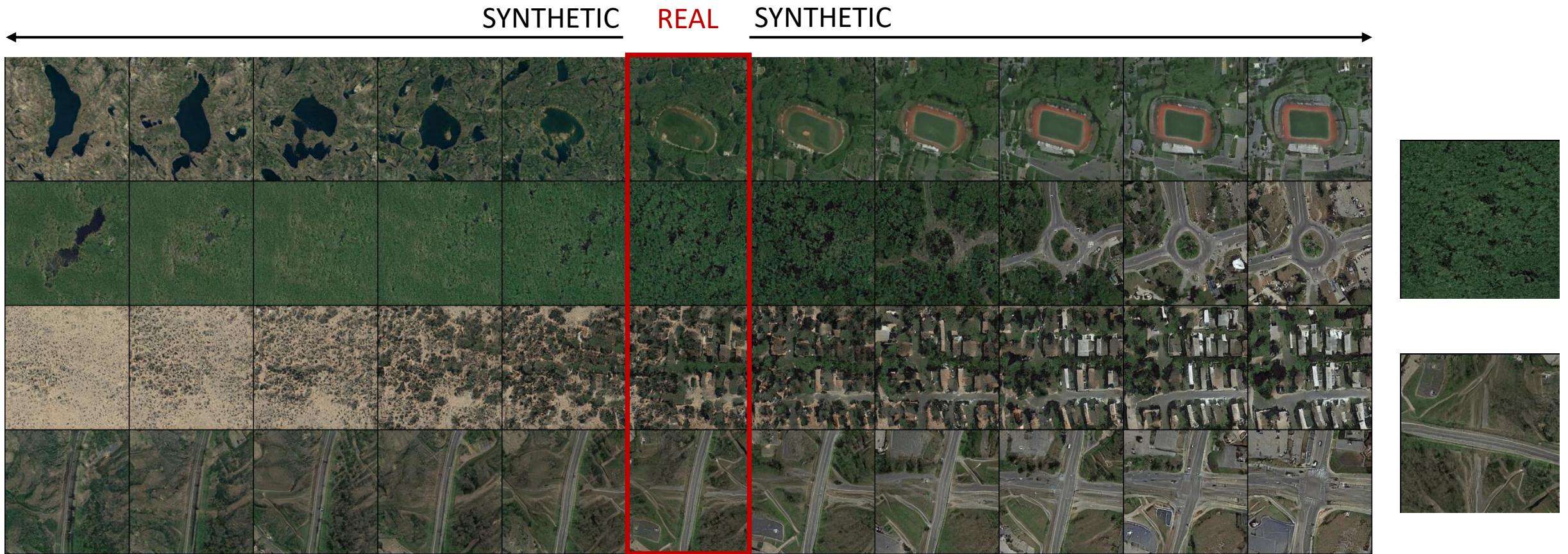
Interpreting the Latent Space of Generative Models



- Can we interpret the latent space (e.g., *activations* or *weights*) of large generative models by some form of appropriate factorization?
 - ▶ Yes- variance preserving models works well (e.g., PCA: discover principal directions corresponding to modes of largest variation) – fully unsupervised



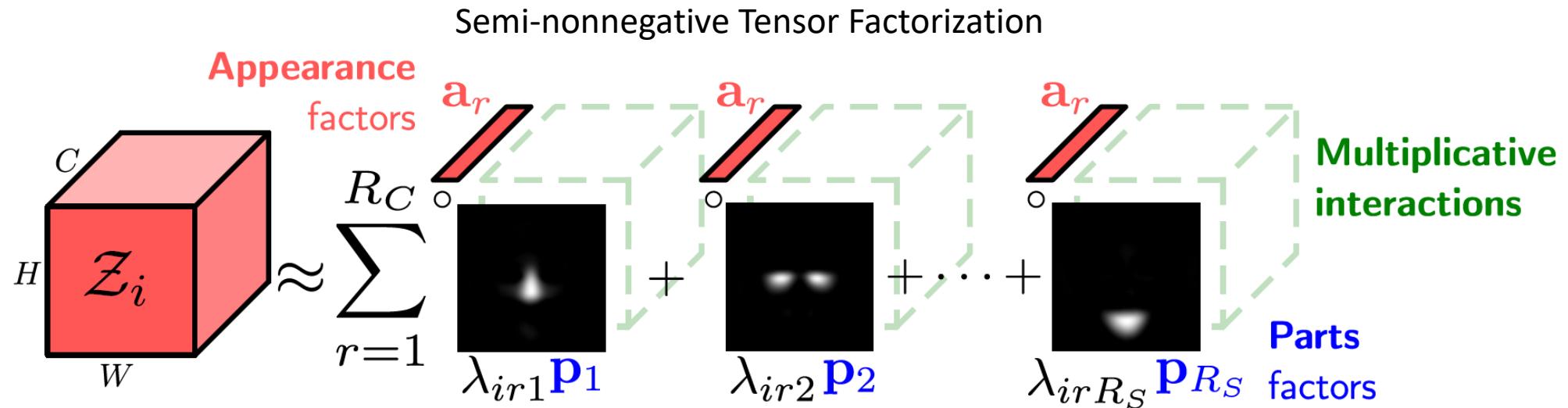
Interpreting Generative Models for Satellite Imagery



- Starting from the image in the centre, we can guide the network to produce images that increase (decrease) **urbanization** in a satellite image
- Improved results across datasets on multiple configurations, even on **unseen** datasets
- *Complements traditional baseline augmentations*

Unsupervised Learning of Parts and Appearances in the Feature Maps of GANs (PandA)

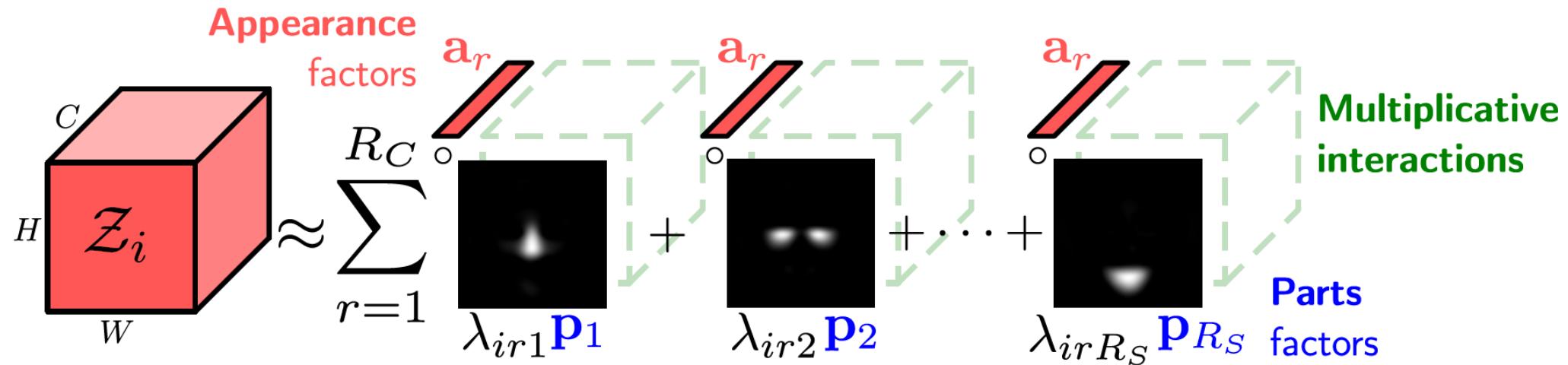
- Can we enable *pixel-level localized* control of pre-trained generative networks (architecture agnostic – convolutional generators)
- Can we discover **factors** that represent the **parts** and **appearances** emerging from the activations of a generative network trained on a given dataset?
 - ▶ Exploit intermediate tensor structure: spatial dimensions/ information, channels: appearance



Parts operate as semantic masks, obtained in unsupervised manner: **Localize modes of variation**

Unsupervised Learning of Parts and Appearances in the Feature Maps of GANs (PandA)

- Proposed Semi-nonnegative Tensor Factorization:



- Optimization problem (block-coordinate descent)

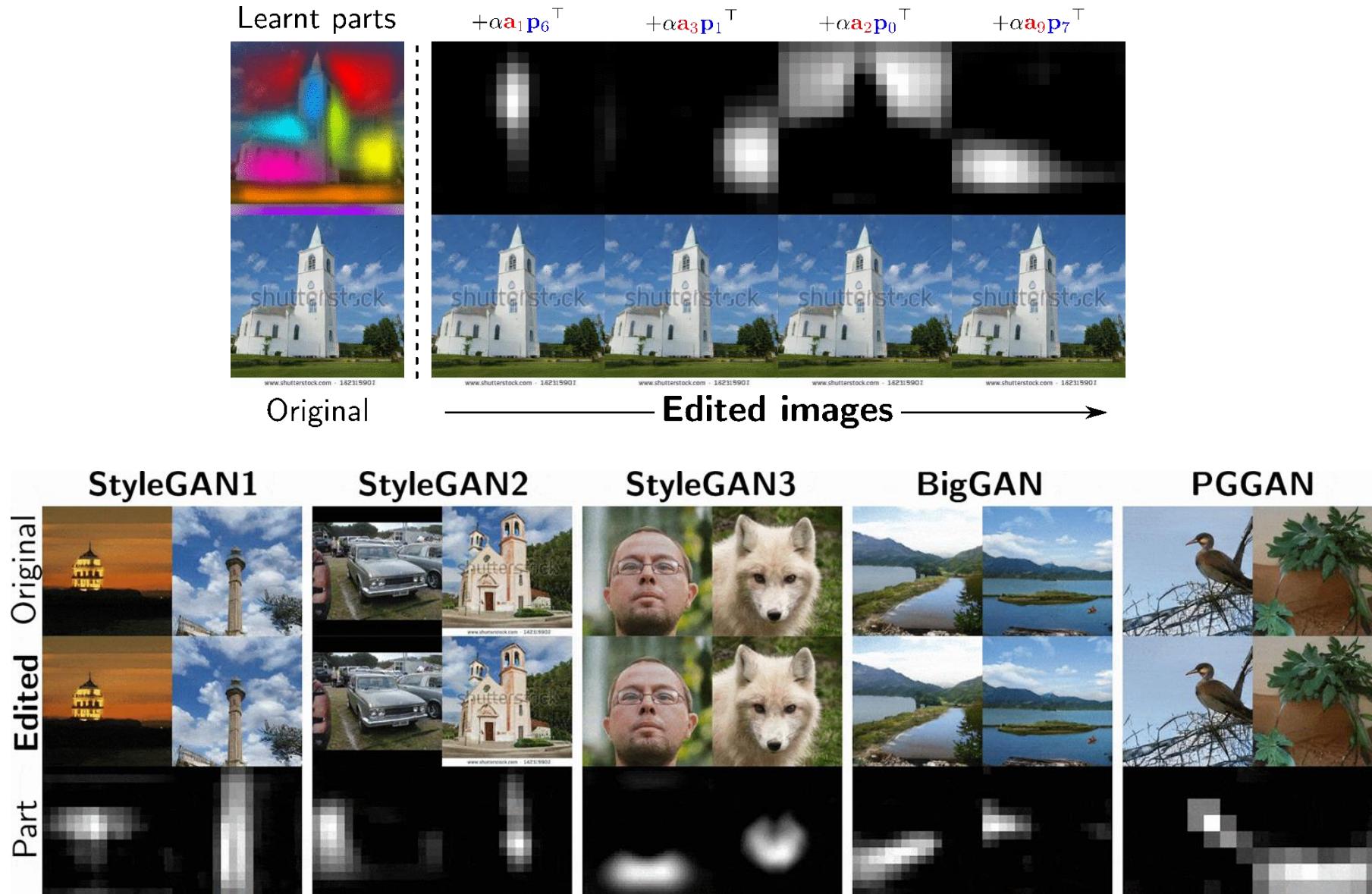
$$\min_{\mathbf{A}, \mathbf{P}} \sum_{i=1}^N \|\mathbf{Z}_i - \mathbf{A} (\mathbf{A}^\top \mathbf{Z}_i \mathbf{P}) \mathbf{P}^\top\|_F^2 \quad \text{s.t. } \mathbf{P} \geq \mathbf{0}$$

P: Orthogonality + non negativity leads to sparsity

A: mixed-sign, orthogonality

Facilitates editing specific parts:
 $\mathcal{X}'_i = G_{[l:]} (\mathbf{Z}_i + \alpha \mathbf{a}_j \hat{\mathbf{p}}_k^\top)$

Downstream tasks



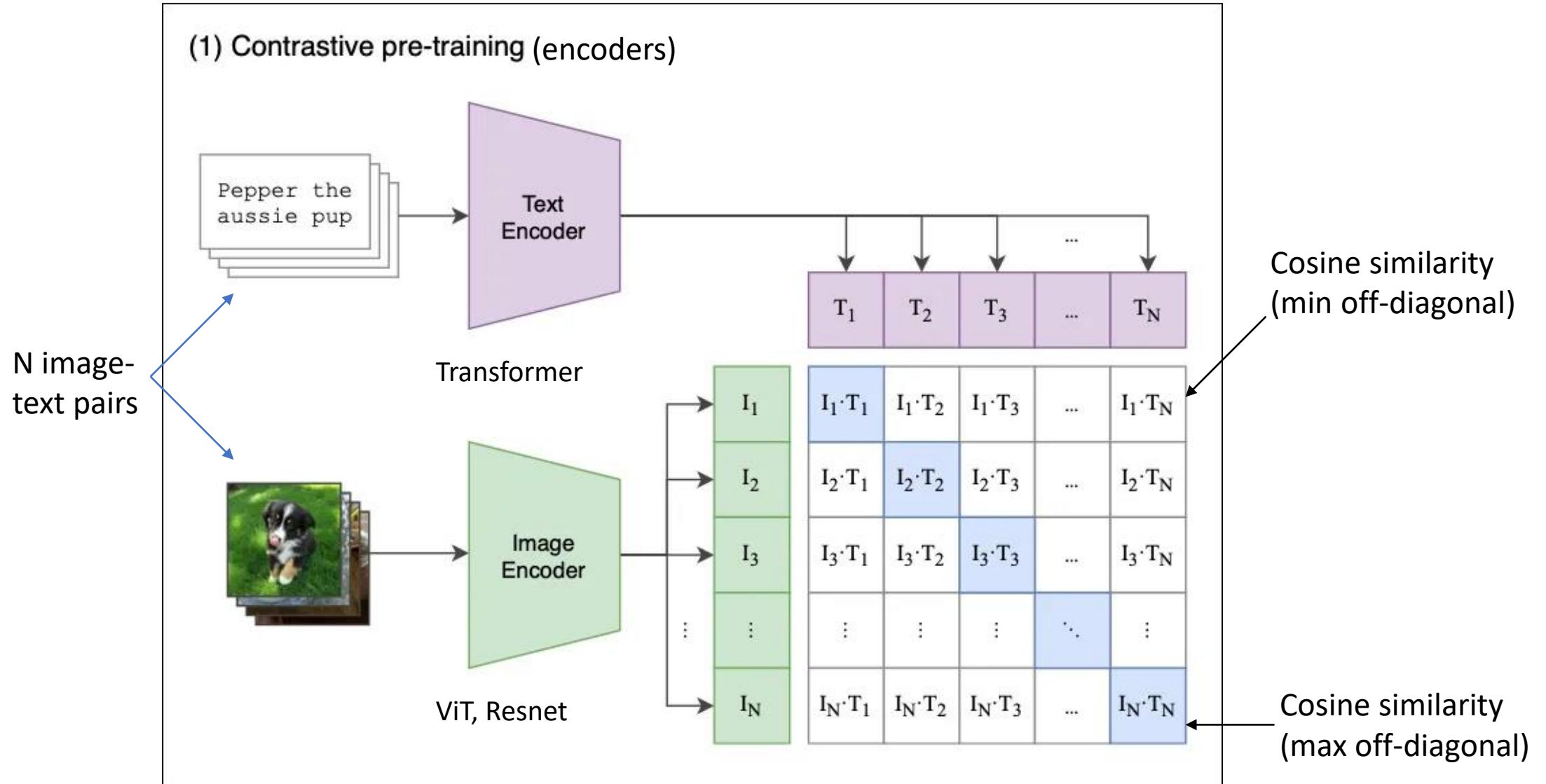
Joint Vision-Language Models: CLIP

Contrastive Language-Image Pre-training (CLIP); ICML 2021

- Connect benefits to NLP from scale to vision (multimodal)
- No predetermined object categories to limit flexibility
- No retraining for new concepts; ‘zero-shot’
 - Can predict most relevant text for image
- No need for manual annotations that are costly and tedious
- Use alt-text: trained on 400M text-image pairs
- 256 V100 GPUs >200K\$ cost on AWS

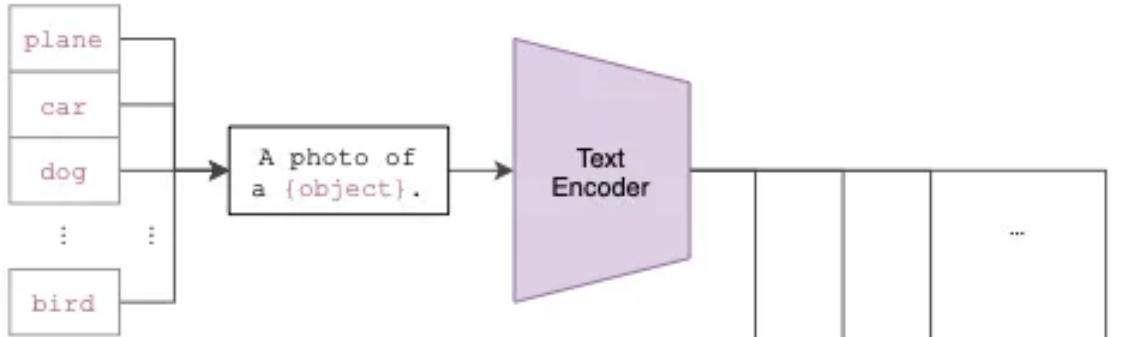
CLIP

symmetric cross-entropy loss (prediction of image from text, and text to image)

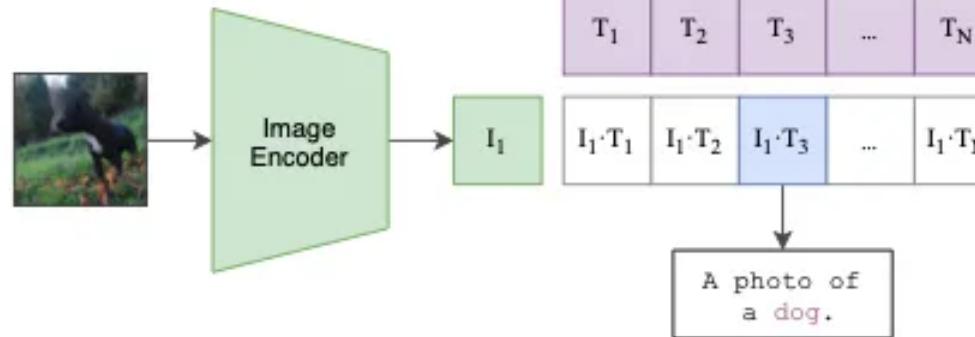


CLIP and Zero-shot Classification

(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



ChatGPT can write full code for this quite easily

	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		76.2	76.2	0%
ImageNetV2		64.3	70.1	+5.8%
ImageNet-R		37.7	88.9	+51.2%
ObjectNet		32.6	72.3	+39.7%
ImageNet Sketch		25.2	60.2	+35.0%
ImageNet-A		2.7	77.1	+74.4%

Robust to domain shifts, unseen images

PoS-Grounded Subspaces in Vision-Language Models

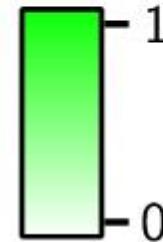
- There are often **multiple** important visual properties of an image:



CLIP

goldfish
golden
swimming

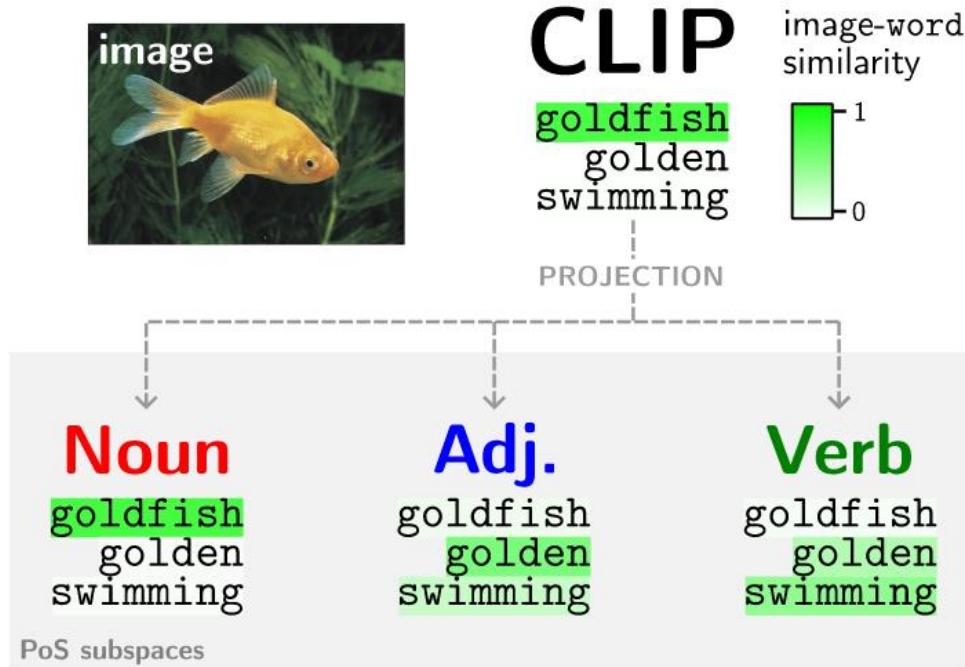
image-word
similarity



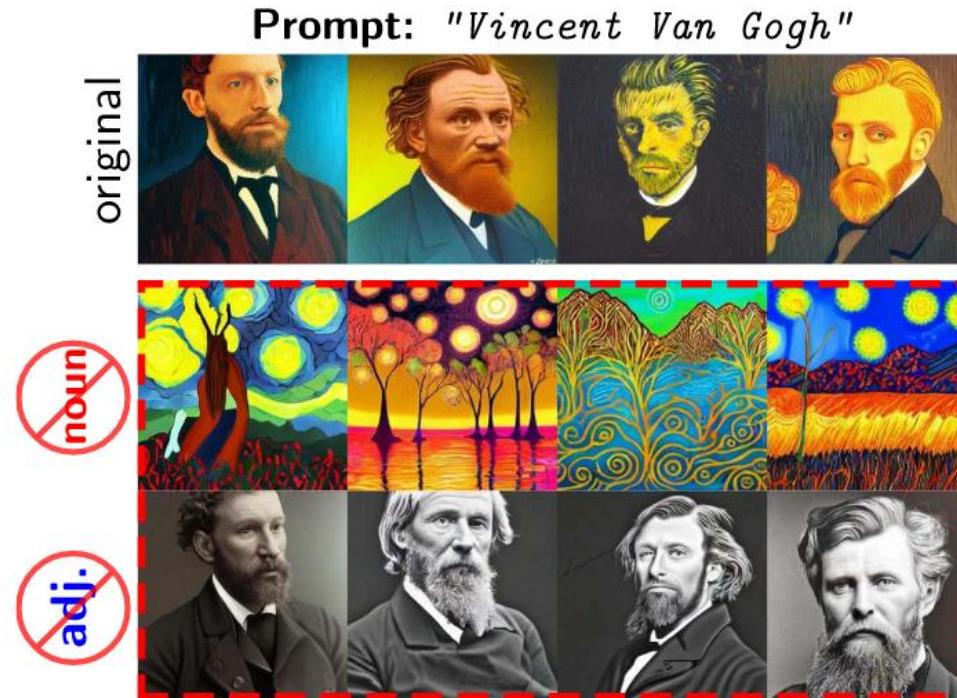
- In **CLIP** however, all visual modes of variation in an image / text are represented in a single joint vision-language embedding—in an **unpredictable** manner
 - i.e., no particular structure

$$\mathbf{W}_i = \arg \max_{\mathbf{W}_i^\top \mathbf{W}_i = \mathbf{I}_k} \left\{ (1 - \lambda) \|\mathbf{W}_i^\top \mathbf{X}_i\|_F^2 - \sum_{j \in \mathcal{C} \setminus \{i\}} \lambda \|\mathbf{W}_i^\top \mathbf{X}_j\|_F^2 \right\},$$

PoS-Grounded Subspaces in Vision-Language Models



(a) Image-word similarity with both CLIP's embedding and after projecting it onto the PoS subspaces.



(b) Text-to-image visualisation of the subspace disentanglement of phrases with multiple visual associations.

- We leverage the association between parts of speech and specific visual modes of variation to better separate the constituent visual components
- **Intuition:** **nouns** describe what something is, **adjectives** how it looks

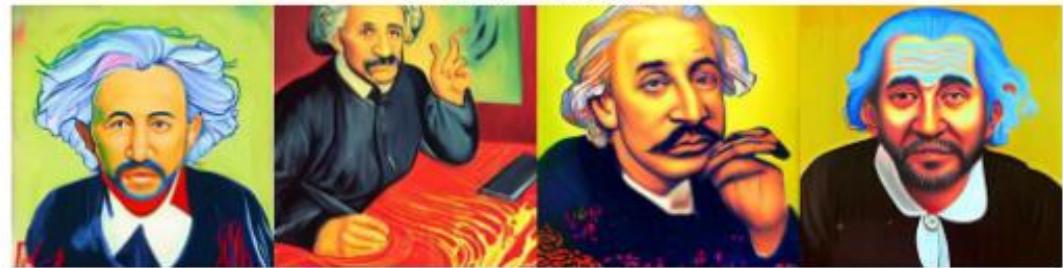
PoS-Grounded Subspaces: Style blocking

Prompt: "A painting of a mountain in the style of **Van Gogh**"

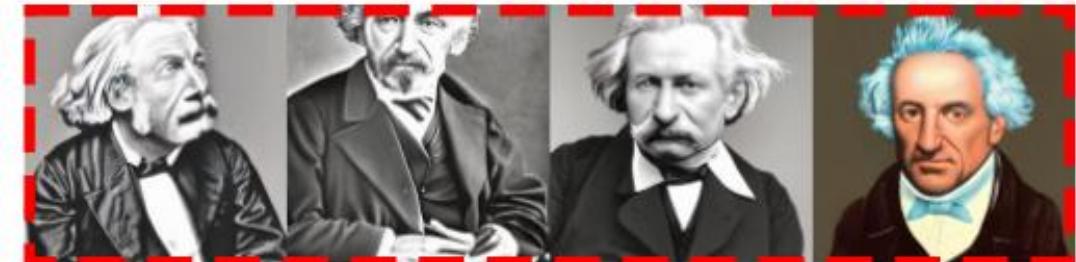
original



Prompt: "A **Gauguin** painting of Einstein"



adj.
no



- One can project onto the orth. complement of the adjective subspace to block the imitation of artists' styles in CLIP-based text-to-image models

PoS-Grounded Subspaces: Style blocking

Prompt: "A painting of a beach in the style of Qi Baishi"



original



(a) A custom subspace for ‘artistic style’.

Prompt: "A photo of a bloody rabbit carcass"



original



(b) A custom subspace for gory/bloody visual themes.

- Or learn additional subspaces corresponding to more specific visual appearances to remove entire visual themes from synthetic images (e.g. related to gore)

