

Lecture 1: Introduction to Deep Learning and Its Foundation

Tao LIN

February 16, 2023



1 Introduction to Deep Learning

- From ANNs to Deep Learning
- Current Applications and Success

2 Review: Linear Algebra

- Notation
- Vectors
- Matrices

3 Review: Probability Theory

- Elements of probability
- Random variables
- Two random variables

Table of Contents

1 Introduction to Deep Learning

- From ANNs to Deep Learning
- Current Applications and Success

2 Review: Linear Algebra

3 Review: Probability Theory

Reference

- François Fleuret. Deep Learning Course. <https://fleuret.org/dlc/>

Many applications require the automatic extraction of “refined” information from raw signal (e.g. image recognition, automatic speech processing, natural language processing, robotic control, geometry reconstruction).



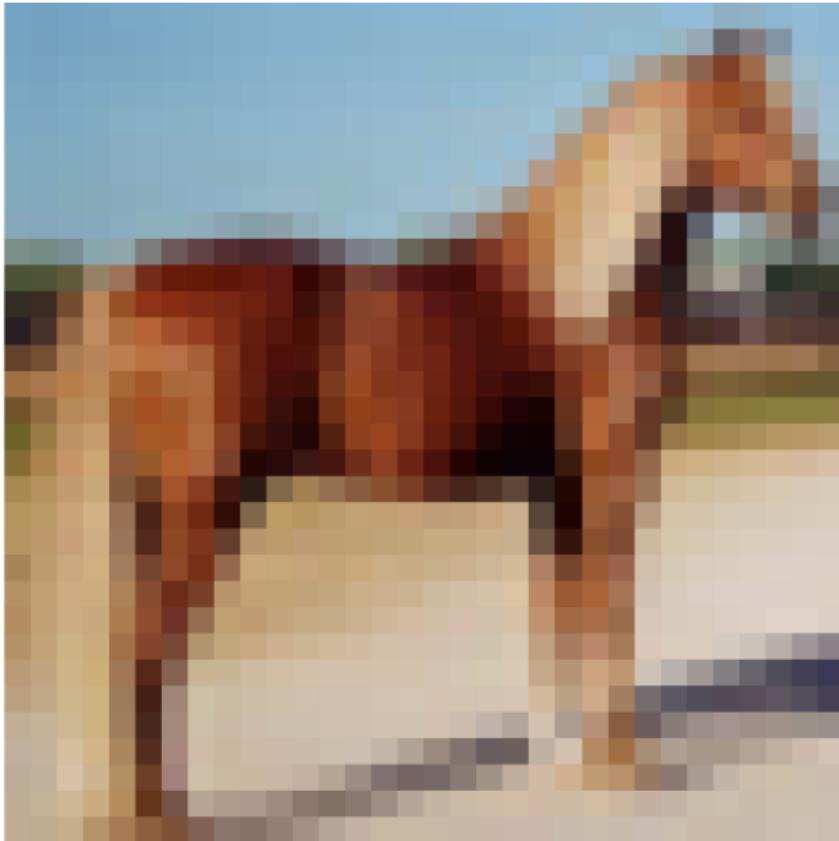
(ImageNet)

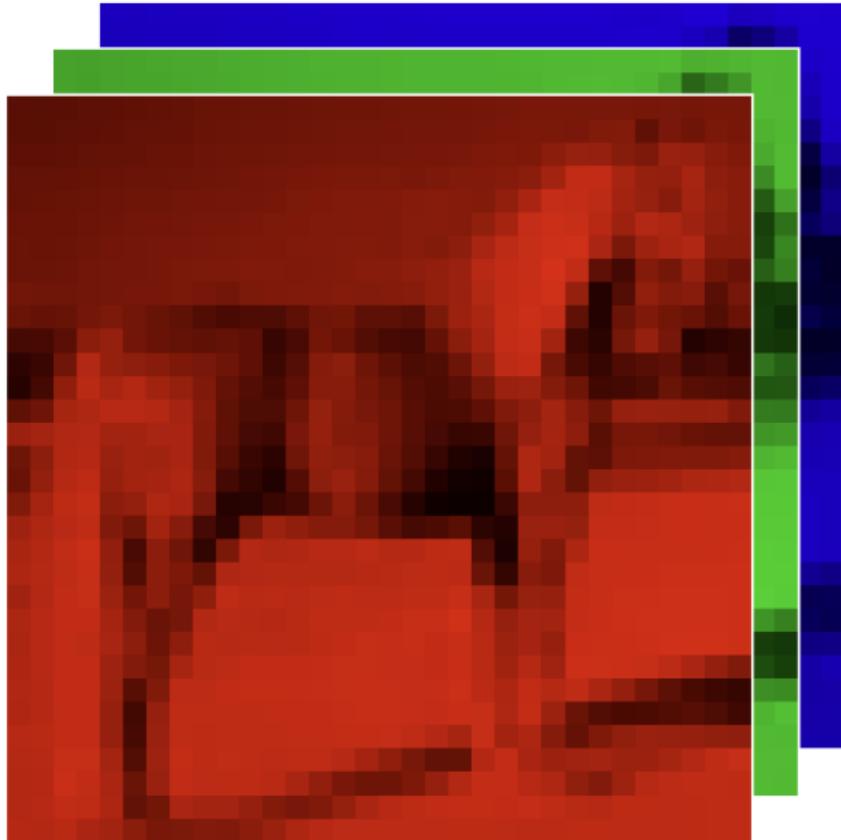
Our brain is so good at interpreting visual information that the “semantic gap” is hard to assess intuitively.

Our brain is so good at interpreting visual information that the “semantic gap” is hard to assess intuitively.



This: is a horse





```
>>> from torchvision.datasets import CIFAR10
>>> cifar = CIFAR10('./data/cifar10/', train=True, download=True)
Files already downloaded and verified
>>> x = torch.from_numpy(cifar.train_data)[43].transpose(2, 0).transpose(1, 2)
>>> x[:, :4, :8]
tensor([[[ 99,  98, 100, 103, 105, 107, 108, 110],
         [100, 100, 102, 105, 107, 109, 110, 112],
         [104, 104, 106, 109, 111, 112, 114, 116],
         [109, 109, 111, 113, 116, 117, 118, 120]],

        [[166, 165, 167, 169, 171, 172, 173, 175],
         [166, 164, 167, 169, 169, 171, 172, 174],
         [169, 167, 170, 171, 171, 173, 174, 176],
         [170, 169, 172, 173, 175, 176, 177, 178]],

        [[198, 196, 199, 200, 200, 202, 203, 204],
         [195, 194, 197, 197, 197, 199, 200, 201],
         [197, 195, 198, 198, 198, 199, 201, 202],
         [197, 196, 199, 198, 198, 199, 200, 201]]], dtype=torch.uint8)
```

Extracting semantic automatically requires models of extreme complexity, which cannot be designed by hand.

Techniques used in practice consist of

- ① defining a parametric model, and
- ② optimizing its parameters by “making it work” on training data.

Extracting semantic automatically requires models of extreme complexity, which cannot be designed by hand.

Techniques used in practice consist of

- ① defining a parametric model, and
- ② optimizing its parameters by “making it work” on training data.

This is similar to biological systems for which the model (e.g. brain structure) is DNA-encoded, and parameters (e.g. synaptic weights) are tuned through experiences.

Extracting semantic automatically requires models of extreme complexity, which cannot be designed by hand.

Techniques used in practice consist of

- ① defining a parametric model, and
- ② optimizing its parameters by “making it work” on training data.

This is similar to biological systems for which the model (e.g. brain structure) is DNA-encoded, and parameters (e.g. synaptic weights) are tuned through experiences.

Deep learning encompasses software technologies to scale-up to billions of model parameters and as many training examples.

There are strong connections between standard statistical modeling and machine learning.

There are strong connections between standard statistical modeling and machine learning.

Classical ML methods combine a “learnable” model from statistics (e.g. “linear regression”) with prior knowledge in pre-processing.

There are strong connections between standard statistical modeling and machine learning.

Classical ML methods combine a “learnable” model from statistics (e.g. “linear regression”) with prior knowledge in pre-processing.

“Artificial neural networks” pre-dated these approaches, and do not follow that dichotomy. They consist of “deep” stacks of parametrized processing.

Table of Contents

1 Introduction to Deep Learning

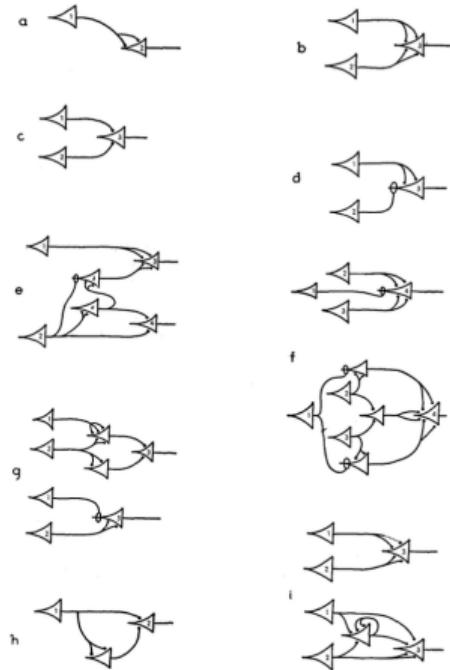
- From ANNs to Deep Learning
- Current Applications and Success

2 Review: Linear Algebra

- Notation
- Vectors
- Matrices

3 Review: Probability Theory

- Elements of probability
- Random variables
- Two random variables



Networks of “Threshold Logic Unit”

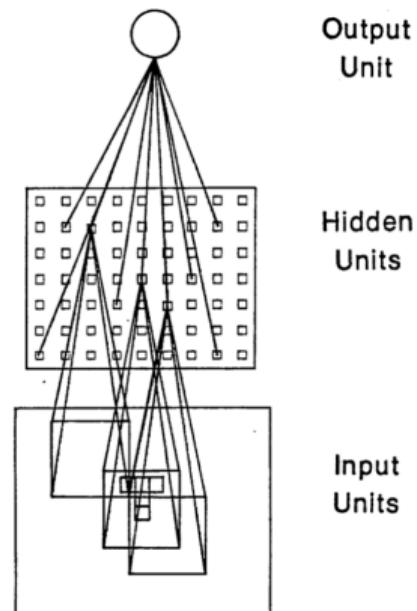
(McCulloch and Pitts, 1943)

- 1949 - Donald Hebb proposes the Hebbian Learning principle.
- 1951 - Marvin Minsky creates the first ANN (Hebbian learning, 40 neurons).

- 1949 - Donald Hebb proposes the Hebbian Learning principle.
- 1951 - Marvin Minsky creates the first ANN (Hebbian learning, 40 neurons).
- 1958 - Frank Rosenblatt creates a perceptron to classify 20×20 images.

- 1949 - Donald Hebb proposes the Hebbian Learning principle.
- 1951 - Marvin Minsky creates the first ANN (Hebbian learning, 40 neurons).
- 1958 - Frank Rosenblatt creates a perceptron to classify 20×20 images.
- 1982 - Paul Werbos proposes back-propagation for ANNs.

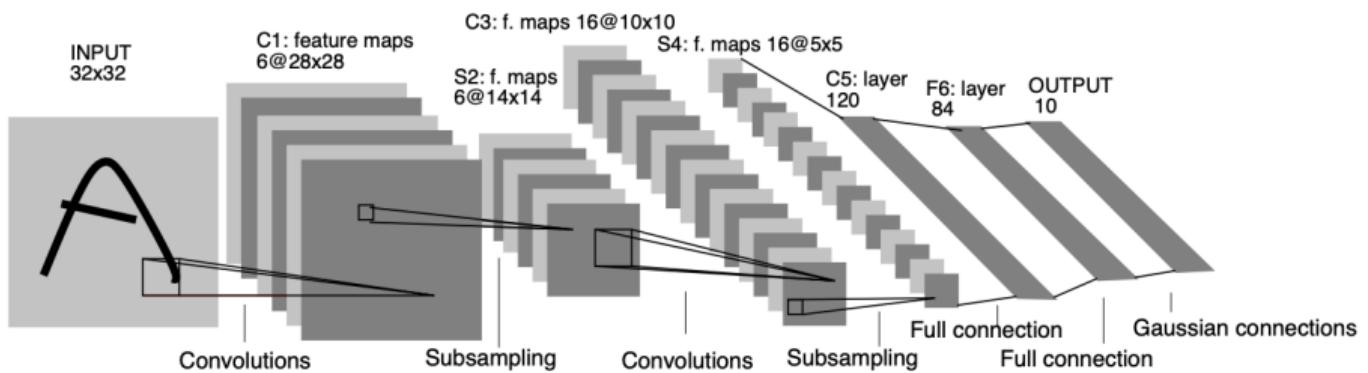
Network for the T-C problem



Trained with back-prop.

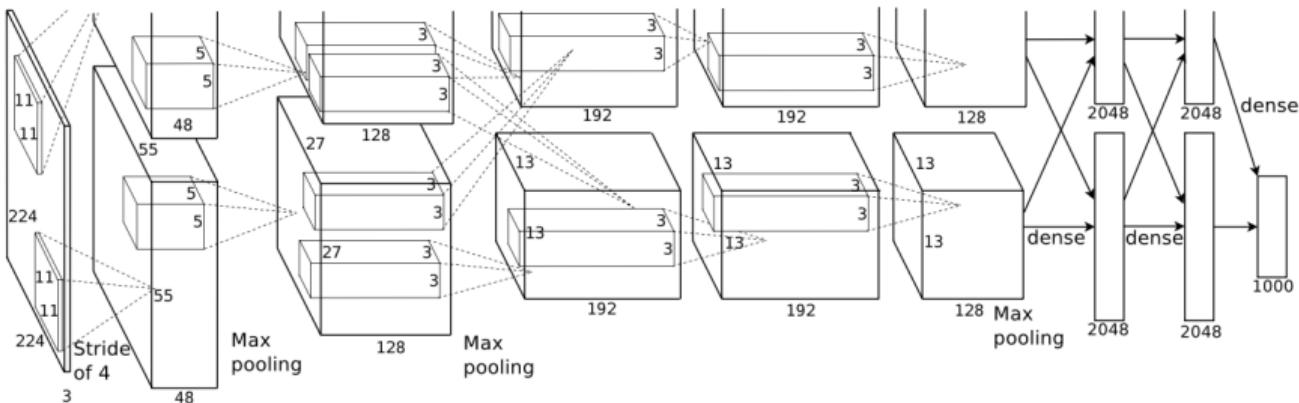
(Rumelhart et al., 1988)

LeNet-5



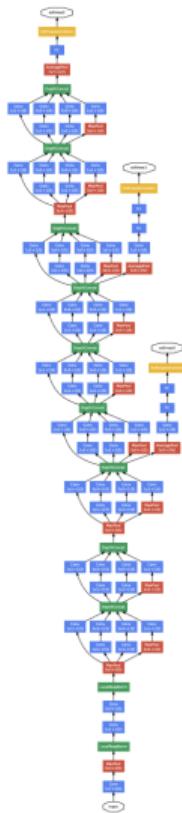
(leCun et al., 1998)

AlexNet

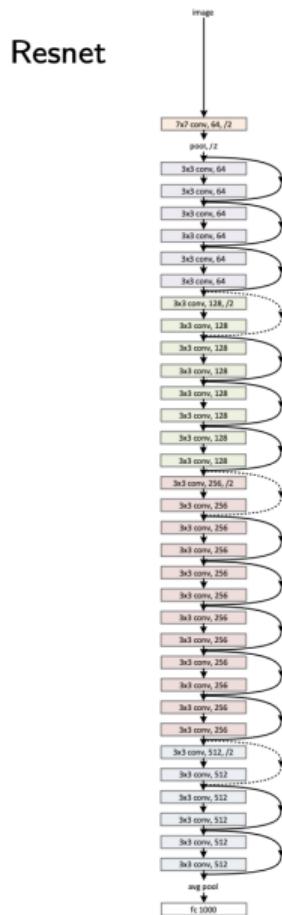


(Krizhevsky et al., 2012)

GoogLeNet



(Szegedy et al., 2015)



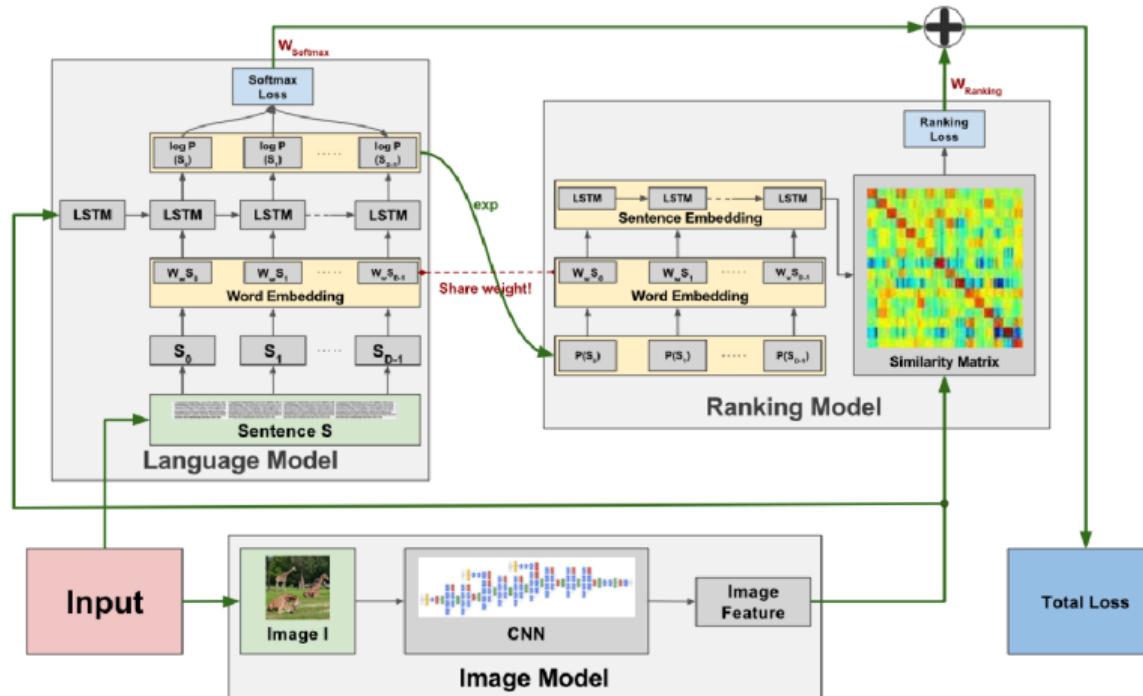
(He et al., 2015)

Deep learning is built on a natural generalization of a neural network: **a graph of tensor operators**, taking advantage of

- the chain rule (aka “back-propagation”),
- stochastic gradient decent,
- convolutions,
- parallel operations on GPUs.

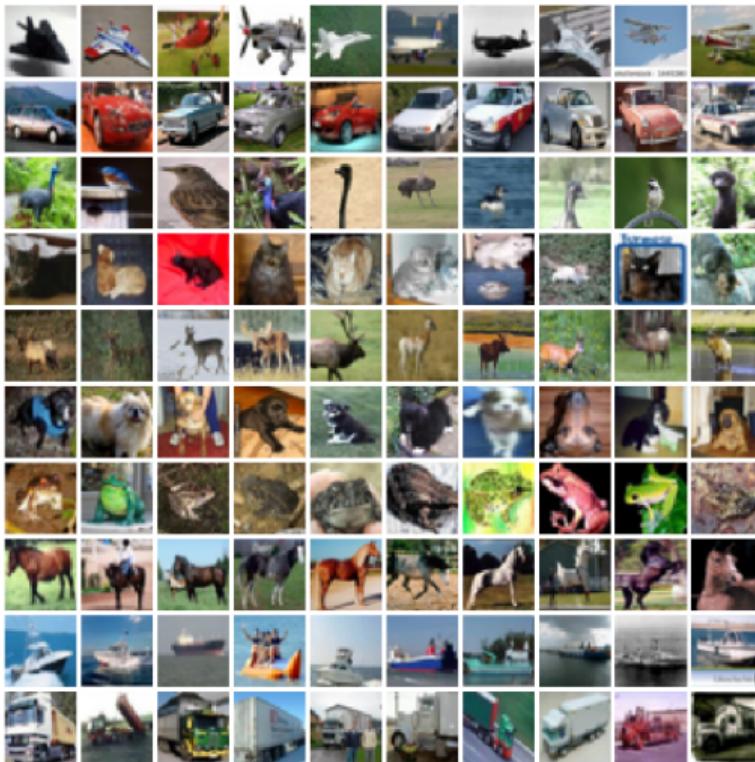
This does not differ much from networks from the 90s.

This generalization allows to design complex networks of operators dealing with images, sound, text, sequences, etc. and to train them end-to-end.



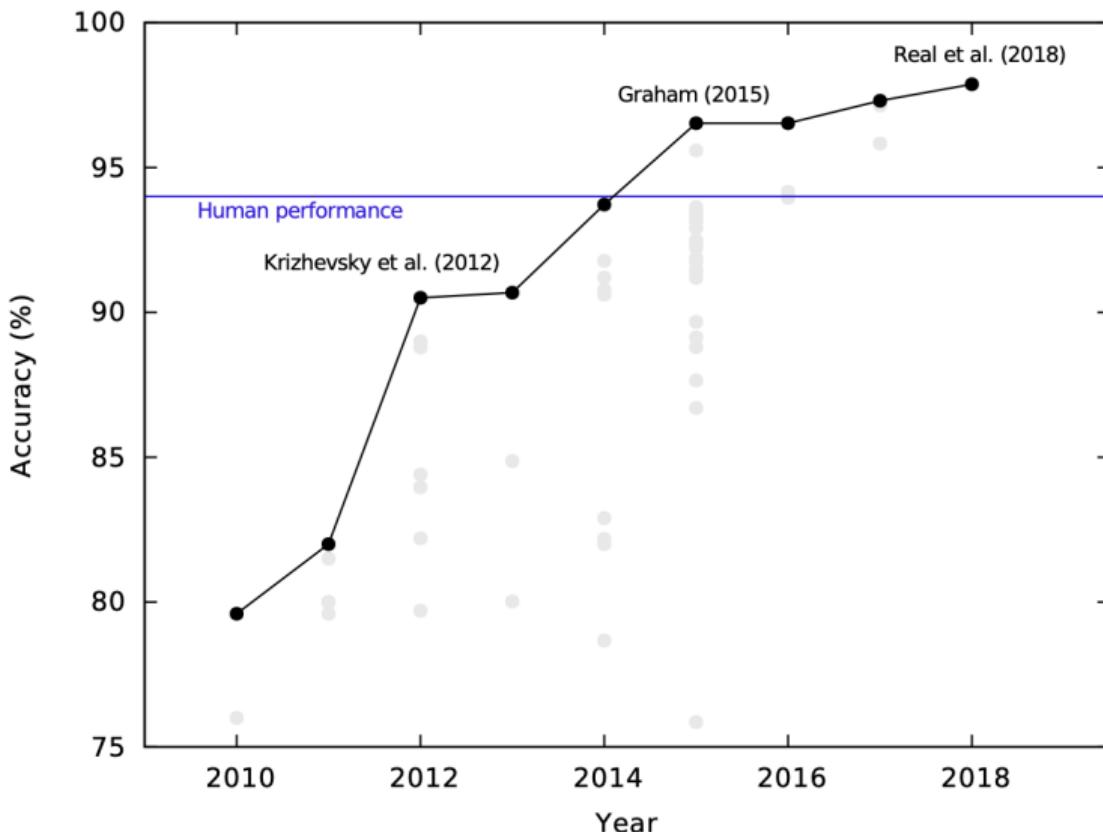
(Yeung et al., 2015)

CIFAR10



32 × 32 color images, 50k train samples, 10k test samples.

Performance on CIFAR10



ImageNet Large Scale Visual Recognition Challenge.

1000 categories, > 1M images

Hatchet
A small ax with a short handle used with one hand (usually to chop wood)

849 pictures

Numbers in brackets: (the number of synsets in the subtree).

- ImageNet 2011 Fall Release (32): plant, flora, plant life (4486) geological formation, formation natural object (1112) sport, athletics (176) artifact, artefact (10504)
 - instrumentality, instrumenta
 - device (2760)
 - implement (726)
 - tool (347)
 - abrader, abradant
 - bender (0)
 - clincher (0)
 - comb (1)
 - cutting implement,
 - bit (12)
 - blade (2)
 - cutter, cutlery, c
 - bolt cutter (0)
 - cigar cutter (0)
 - die (0)
 - edge tool (92)
 - adz
 - ax, axe (1)
 - broada

Treemap Visualization **Images of the Synset** **Downloads**

(<http://image-net.org/challenges/LSVRC/2014/browse-synsets>)

method	top-1 err.	top-5 err.
VGG [41] (ILSVRC'14)	-	8.43 [†]
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	19.38	4.49

Table 4. Error rates (%) of **single-model** results on the ImageNet validation set (except [†] reported on the test set).

method	top-5 err. (test)
VGG [41] (ILSVRC'14)	7.32
GoogLeNet [44] (ILSVRC'14)	6.66
VGG [41] (v5)	6.8
PReLU-net [13]	4.94
BN-inception [16]	4.82
ResNet (ILSVRC'15)	3.57

Table 5. Error rates (%) of **ensembles**. The top-5 error is on the test set of ImageNet and reported by the test server.

(He et al., 2015)

Table of Contents

1 Introduction to Deep Learning

- From ANNs to Deep Learning
- Current Applications and Success

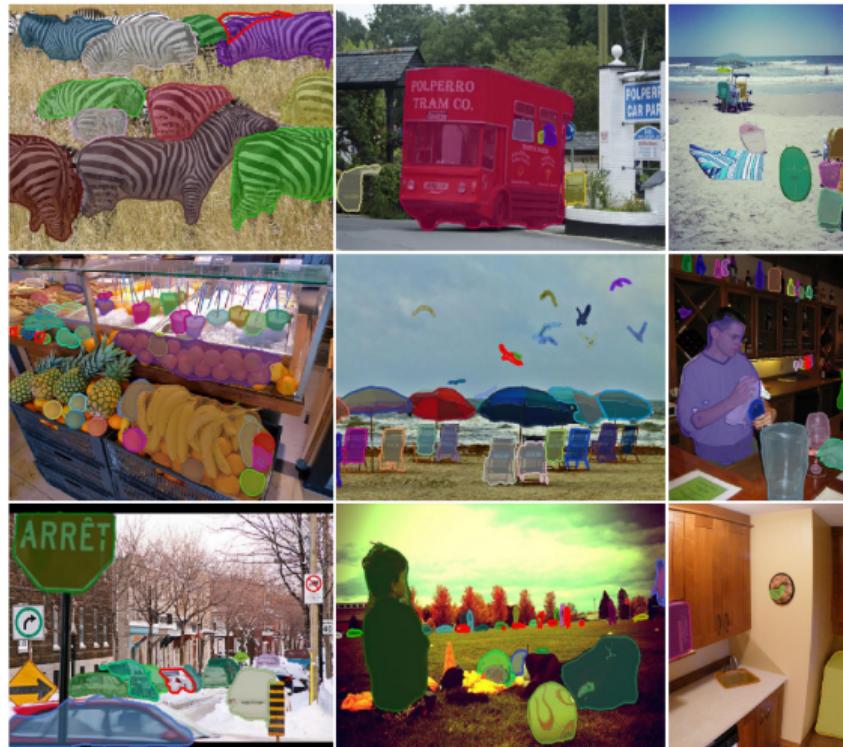
2 Review: Linear Algebra

- Notation
- Vectors
- Matrices

3 Review: Probability Theory

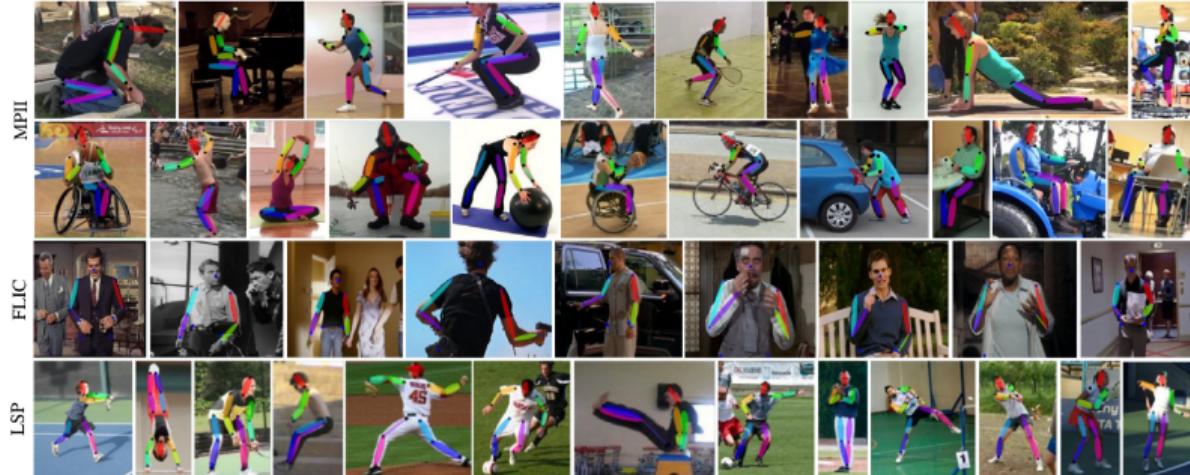
- Elements of probability
- Random variables
- Two random variables

Object detection and segmentation



(Pinheiro et al., 2016)

Human pose estimation



(Wei et al., 2016)

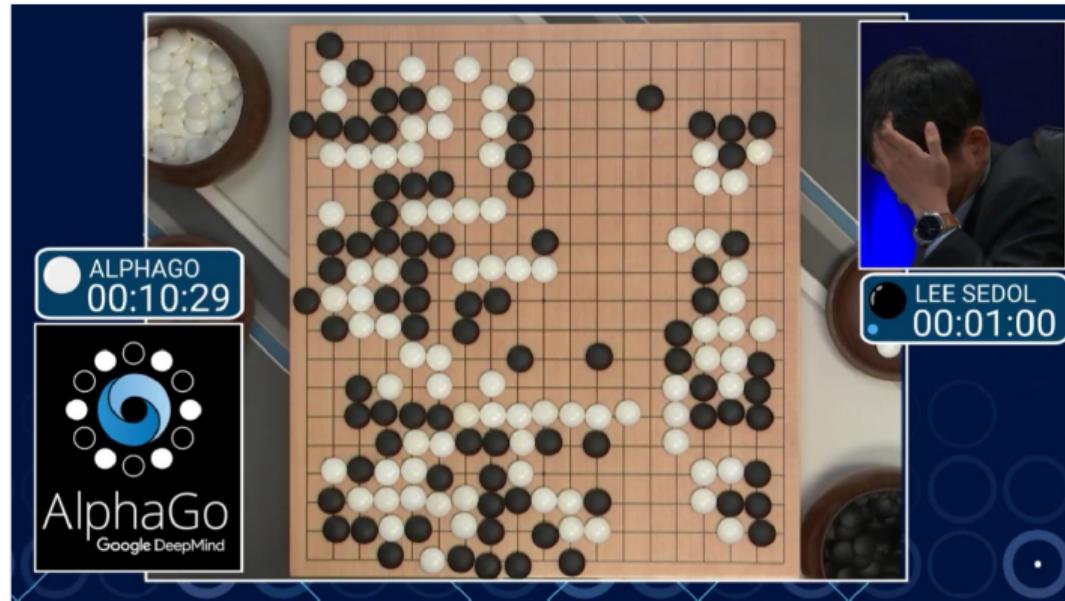
Reinforcement learning



Self-trained, plays 49 games at human level.

(Mnih et al., 2015)

Strategy games



March 2016, 4-1 against a 9-dan professional without handicap.

(Silver et al., 2016)

Translation

"The reason Boeing are doing this is to cram more seats in to make their plane more competitive with our products," said Kevin Keniston, head of passenger comfort at Europe's Airbus.

- "La raison pour laquelle Boeing fait cela est de créer plus de sièges pour rendre son avion plus compétitif avec nos produits", a déclaré Kevin Keniston, chef du confort des passagers chez Airbus.

When asked about this, an official of the American administration replied:
"The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington."

- Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu:
"Les États-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington"

(Wu et al., 2016)

Auto-captioning

A person riding a motorcycle on a dirt road.



Two dogs play in the grass.



A group of young people playing a game of frisbee.



Two hockey players are fighting over the puck.



A herd of elephants walking across a dry grass field.



A close up of a cat laying on a couch.



Question answering

I: Jane went to the hallway.
I: Mary walked to the bathroom.
I: Sandra went to the garden.
I: Daniel went back to the garden.
I: Sandra took the milk there.
Q: Where is the milk?
A: garden

I: It started boring, but then it got interesting.
Q: What's the sentiment?
A: positive

(Kumar et al., 2015)

Image generation



(Brock et al., 2018)

ChatGPT



Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →



Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests



Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

Why does it work now?

The success of deep learning is multi-factorial:

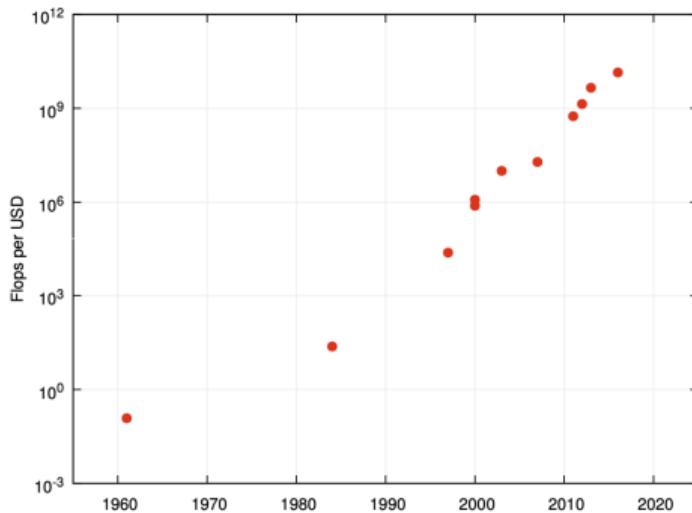
- Five decades of research in machine learning,
- CPUs/GPUs/storage developed for other purposes,
- lots of data from “the internet”,
- tools and culture of collaborative and reproducible science,
- resources and efforts from large corporations.

Five decades of research in ML provided

- a taxonomy of ML concepts (classification, generative models, clustering, kernels, linear embeddings, etc.),
- a sound statistical formalization (Bayesian estimation, PAC),
- a clear picture of fundamental issues (bias/variance dilemma, VC dimension, generalization bounds, etc.),
- a good understanding of optimization issues,
- efficient large-scale algorithms.

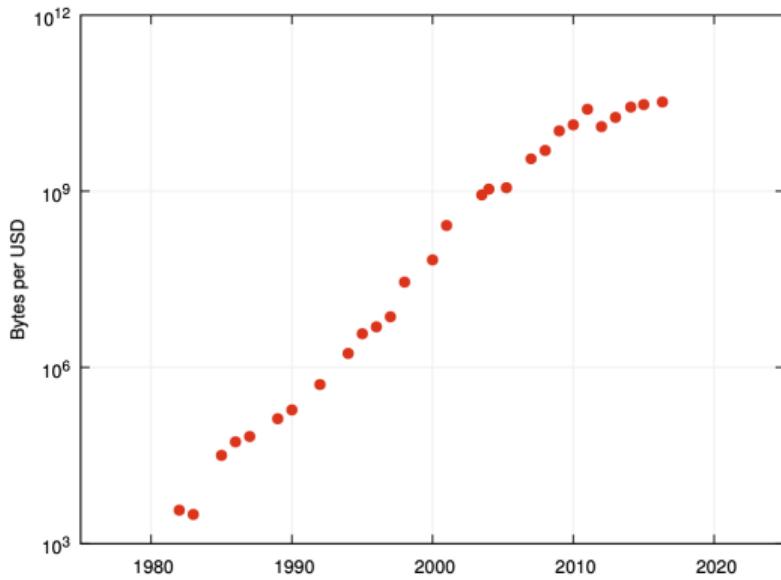
From a practical perspective, deep learning

- lessens the need for a deep mathematical grasp,
- makes the design of large learning architectures a system/software development task,
- allows to leverage modern hardware (clusters of GPUs),
- does not plateau when using more data,
- makes large trained networks a commodity.



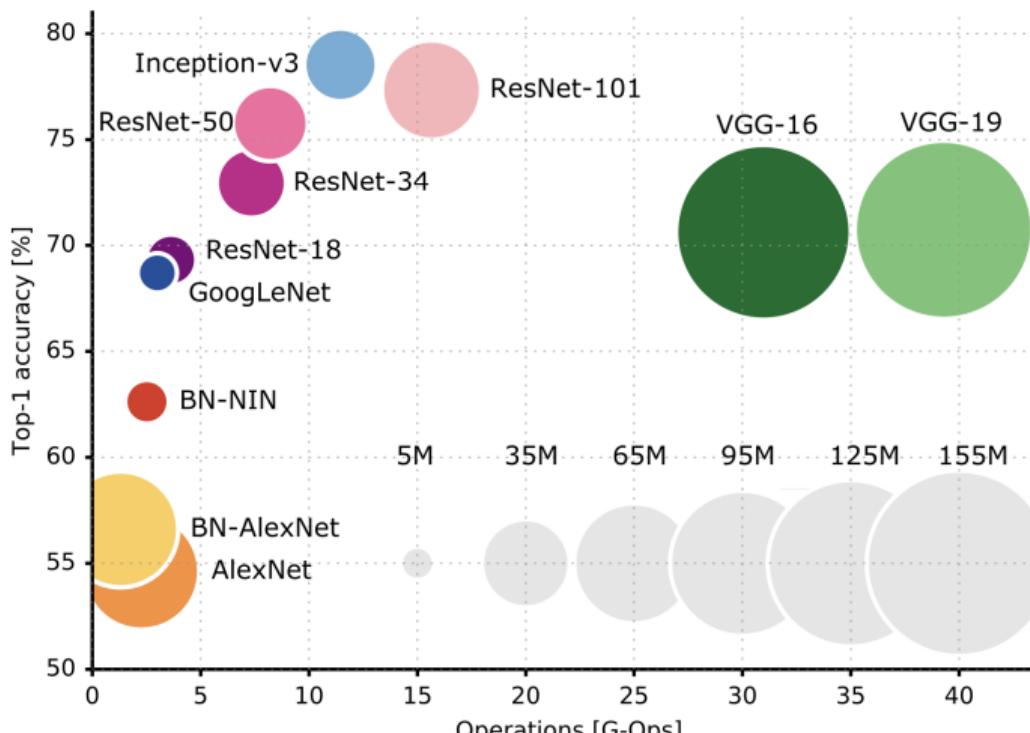
(Wikipedia “FLOPS”)

	TFlops (10^{12})	Price	GFlops per \$
Intel i7-6700K	0.2	\$344	0.6
AMD Radeon R-7 240	0.5	\$55	9.1
NVIDIA GTX 750 Ti	1.3	\$105	12.3
AMD RX 480	5.2	\$239	21.6
NVIDIA GTX 1080	8.9	\$699	12.7



(John C. McCallum)

The typical cost of a 4Tb hard disk is \$120 (Dec 2016).



(Canziani et al., 2016)

Data-set	Year	Nb. images	Resolution	Nb. classes
MNIST	1998	6.0×10^4	28×28	10
NORB	2004	4.8×10^4	96×96	5
Caltech 101	2003	9.1×10^3	$\simeq 300 \times 200$	101
Caltech 256	2007	3.0×10^4	$\simeq 640 \times 480$	256
LFW	2007	1.3×10^4	250×250	—
CIFAR10	2009	6.0×10^4	32×32	10
PASCAL VOC	2012	2.1×10^4	$\simeq 500 \times 400$	20
MS-COCO	2015	2.0×10^5	$\simeq 640 \times 480$	91
ImageNet	2016	14.2×10^6	$\simeq 500 \times 400$	21,841
Cityscape	2016	25×10^3	$2,000 \times 1000$	30

Table of Contents

1 Introduction to Deep Learning

2 Review: Linear Algebra

- Notation
- Vectors
- Matrices

3 Review: Probability Theory

Reference

- Zico Kolter and Chuong Do, Linear Algebra Review and Reference.
<http://cs229.stanford.edu/section/cs229-linalg.pdf>, 2012
- Volkan Cevher, Mathematics of Data: From Theory to Computation.
<https://edu.epfl.ch/coursebook/en/mathematics-of-data-from-theory-to-computation-EE-556>, 2015

Table of Contents

1 Introduction to Deep Learning

- From ANNs to Deep Learning
- Current Applications and Success

2 Review: Linear Algebra

- Notation
- Vectors
- Matrices

3 Review: Probability Theory

- Elements of probability
- Random variables
- Two random variables

Notation

- **Scalar** are denoted by lowercase letters (e.g., k).
- **Vectors** by lowercase boldface letters (e.g., \mathbf{x}).
- **Matrices** by uppercase boldface letters (e.g., \mathbf{A}).
- **Component** of a vector \mathbf{x} , matrix \mathbf{A} are x_i , $a_{i,j}$, and $A_{i,j,k}$, respectively.
- **Sets** by uppercase calligraphic letters (e.g., \mathcal{S}).

Table of Contents

1 Introduction to Deep Learning

- From ANNs to Deep Learning
- Current Applications and Success

2 Review: Linear Algebra

- Notation
- **Vectors**
- Matrices

3 Review: Probability Theory

- Elements of probability
- Random variables
- Two random variables

Vector spaces

Note 1

We focus on the **field of real numbers** (\mathbb{R}) but most of the results can be **generalized** to the **field of complex numbers** (\mathbb{C}).

A vector space or *linear space* (over the field \mathbb{R}) consists of

- (a) a **set** of vectors \mathcal{V}
- (b) an **addition** operation: $\mathcal{V} \times \mathcal{V} \rightarrow \mathcal{V}$
- (c) a **scalar multiplication** operation: $\mathbb{R} \times \mathcal{V} \rightarrow \mathcal{V}$
- (d) a **distinguished element** $\mathbf{0} \in \mathcal{V}$

and satisfies the following properties:

- 1 $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}, \quad \mathbf{x}, \mathbf{y} \in \mathcal{V}$
- 2 $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z}), \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$
- 3 $\mathbf{0} + \mathbf{x} = \mathbf{x}, \forall \mathbf{x} \in \mathcal{V}$
- 4 $\forall \mathbf{x} \in \mathcal{V} \quad \exists (-\mathbf{x}) \in \mathcal{V} \text{ such that } \mathbf{x} + (-\mathbf{x}) = \mathbf{0}$
- 5 $(\alpha\beta)\mathbf{x} = \alpha(\beta\mathbf{x}), \quad \alpha, \beta \in \mathbb{R} \quad \forall \mathbf{x} \in \mathcal{V}$
- 6 $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}, \quad \forall \alpha \in \mathbb{R} \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{V}$
- 7 $1\mathbf{x} = \mathbf{x}, \forall \mathbf{x} \in \mathcal{V}$

commutative under addition

associative under addition

$\mathbf{0}$ being additive identity

$-\mathbf{x}$ being additive inverse

associative under scalar multiplication

distributive

1 being multiplicative identity

Vector spaces contd.

Example 2 (Vector space)

- 1 $\mathcal{V}_1 = \{\mathbf{0}\}$ for $\mathbf{0} \in \mathbb{R}^p$
- 2 $\mathcal{V}_2 = \mathbb{R}^p$
- 3 $\mathcal{V}_3 = \sum_{i=1}^k \alpha_i \mathbf{x}_i$ for $\alpha_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^p$

Definition 3 (Subspace)

A **subspace** is a vector space that is a *subset* of another vector space.

Example 4 (Subspace)

\mathcal{V}_1 , \mathcal{V}_2 , and \mathcal{V}_3 in the example above are subspaces of \mathbb{R}^p .

Vector spaces contd.

Definition 5 (Span)

The **span** of a set of vectors, $\{x_1, \dots, x_k\}$, is the set of all possible **linear combinations** of these vectors; i.e.,

$$\text{span}\{x_1, \dots, x_k\} = \{\alpha_1 x_1 + \dots + \alpha_k x_k \mid \alpha_1, \dots, \alpha_k \in \mathbb{R}\}. \quad (1)$$

Definition 6 (Linear independence)

A set of vectors, $\{x_1, \dots, x_k\}$, is **linearly independent** if

$$\alpha_1 x_1 + \dots + \alpha_k x_k = \mathbf{0} \Rightarrow \alpha_1 = \alpha_2 = \dots = \alpha_k = 0. \quad (2)$$

Vector spaces contd.

Definition 7 (Basis)

The **basis** of a vector space, \mathcal{V} , is a set of vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ that satisfy (a) $\mathcal{V} = \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$, (b) $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ are linearly independent.

Definition 8 (Dimension)

The **dimension** of a vector space, \mathcal{V} , (denoted $\dim(\mathcal{V})$) is the number of vectors in the basis of \mathcal{V} .

Vector Norms

Definition 9 (Vector norm)

A norm of a vector in \mathbb{R}^p is a function $\|\cdot\| : \mathbb{R}^p \rightarrow \mathbb{R}$ such that for all vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ and scalar $\lambda \in \mathbb{R}$

- (a) $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in \mathbb{R}^p$ non-negativity
- (b) $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$ definitiveness
- (c) $\|\lambda\mathbf{x}\| = |\lambda| \|\mathbf{x}\|$ Homogeneity
- (d) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ triangle inequality

- There is a family of ℓ_q -norms parameterized by $q \in [1, \infty]$;
- For $\mathbf{x} \in \mathbb{R}^p$, the ℓ_q -norm is defined as $\|\mathbf{x}\|_q := (\sum_{i=1}^p |x_i|^q)^{1/q}$.

Example 10

1 ℓ_2 -norm: $\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^p x_i^2}$

2 ℓ_1 -norm: $\|\mathbf{x}\|_1 := \sum_{i=1}^p |x_i|$

3 ℓ_∞ -norm: $\|\mathbf{x}\|_\infty := \max_{i=1, \dots, p} |x_i|$

Vector norms contd.

Definition 11 (Quasi-norm)

A **quasi-norm** satisfies all the norm properties except (d) triangle inequality, which is replaced by $\|\mathbf{x} + \mathbf{y}\| \leq c (\|\mathbf{x}\| + \|\mathbf{y}\|)$ for a constant $c \geq 1$.

Definition 12 (Semin(pseudo)-norm)

A **semi(pseudo)-norm** satisfies all the norm properties except (b) definitiveness.

Definition 13 (ℓ_0 -“norm”)

$$\|\mathbf{x}\|_0 = \lim_{q \rightarrow 0} \|\mathbf{x}\|_q^q = |\{i : x_i \neq 0\}|$$

The ℓ_0 -norm counts the non-zero components of \mathbf{x} . It is **not** a norm—it does not satisfy the property (c) \Rightarrow it is also neither a **quasi-** nor a **semi-norm**.

Definition 14 (Norm balls)

Radius r in ℓ_q -norm: $\mathcal{B}_q(r) = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_q \leq r\}$.

Inner products

Definition 15 (Inner product)

The **inner product** of any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ (denoted by $\langle \cdot, \cdot \rangle$) is defined as
 $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^p x_i y_i$.

The inner product satisfies the following properties:

- ① $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ symmetry
- ② $\langle (\alpha \mathbf{x} + \beta \mathbf{y}), \mathbf{z} \rangle = \langle \alpha \mathbf{x}, \mathbf{z} \rangle + \langle \beta \mathbf{y}, \mathbf{z} \rangle, \forall \alpha, \beta \in \mathbb{R}, \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^p$ linearity
- ③ $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0, \forall \mathbf{x} \in \mathbb{R}^p$ positive definiteness

Important relations involving the inner product:

- Hölder's inequality: $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_q \|\mathbf{y}\|_r$, where $r > 1$ and $\frac{1}{q} + \frac{1}{r} = 1$.
- Cauchy-Schwarz is a special case of Hölder's inequality ($q = r = 2$).

Vector norms contd.

Definition 16 (Inner product space)

An **inner product space** is a **vector space** endowed with an **inner product**.

Definition 17 (Dual norm)

Let $\|\cdot\|$ be a norm in \mathbb{R}^p , the the **dual norm** denoted by $\|\cdot\|^*$ is defined:

$$\|\mathbf{x}\|^* = \sup_{\|\mathbf{y}\| \leq 1} \mathbf{x}^\top \mathbf{y}, \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^p \quad (3)$$

- The **dual** of the *dual norm* is the **original (primal) norm**, i.e., $\|\mathbf{x}\|^{**} = \|\mathbf{x}\|$.
- Hölder's inequality $\Rightarrow \|\cdot\|_q$ is a **dual norm** of $\|\cdot\|_r$ when $\frac{1}{q} + \frac{1}{r} = 1$.

Example 18

- 1 $\|\cdot\|_2$ is **dual** of $\|\cdot\|_2$ (i.e., $\|\cdot\|_2$ is **self-dual**): $\sup\{\mathbf{z}^\top \mathbf{x} \mid \|\mathbf{x}\|_2 \leq 1\} = \|\mathbf{z}\|_2$.
- 2 $\|\cdot\|_1$ is **dual** of $\|\cdot\|_\infty$ (and vice versa): $\sup\{\mathbf{z}^\top \mathbf{x} \mid \|\mathbf{x}\|_\infty \leq 1\} = \|\mathbf{z}\|_1$.

Metrics

- A **metric** on a set is a function that satisfies the minimal properties of a distance.

Definition 19 (Metric)

Let \mathcal{X} be a set, then a function $d(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a metric if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$:

- ① $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all \mathbf{x} and \mathbf{y} *(non-negativity)*
- ② $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$ *(definiteness)*
- ③ $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ *(symmetry)*
- ④ $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$ *(triangle inequality)*

- A **pseudo-metric** satisfies (a), (c), and (d) but not necessarily (b)
- A **metric space** (\mathcal{X}, d) is a set \mathcal{X} with a metric d defined on \mathcal{X}
- **Norm** induce **metrics** while **pseudo-norms** induce **pseudo-metrics**

Table of Contents

1 Introduction to Deep Learning

- From ANNs to Deep Learning
- Current Applications and Success

2 Review: Linear Algebra

- Notation
- Vectors
- Matrices

3 Review: Probability Theory

- Elements of probability
- Random variables
- Two random variables

Basic matrix definitions

Definition 20 (The identity matrix)

The **identity matrix**, denoted $\mathbf{I} \in \mathbb{R}^{n \times n}$, is a square matrix with ones on the diagonal and zeros everywhere else. That is,

$$I_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (4)$$

- For all $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{AI} = \mathbf{A} = \mathbf{IA}$ (the dimensions of \mathbf{I} are inferred from context)

Definition 21 (Diagonal matrix)

A **diagonal matrix** is a matrix where all non-diagonal elements are 0. This is typically denoted $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$ with

$$D_{i,j} = \begin{cases} d_i & i = j \\ 0 & i \neq j \end{cases} \quad (5)$$

- Clearly, $\mathbf{I} = \text{diag}(1, \dots, 1)$.

Basic matrix definitions contd.

Definition 22 (Transpose)

The **transpose** of a matrix results from “flipping” the rows and columns. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, its transpose, written $\mathbf{A}^\top \in \mathbb{R}^{n \times m}$, is the $n \times m$ matrix whose entries are given by

$$(\mathbf{A}^\top)_{i,j} = A_{j,i}. \quad (6)$$

- $(\mathbf{A}^\top)^\top = \mathbf{A}$
- $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$
- $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$

Basic matrix definitions contd.

Definition 23 (The Trace)

The **trace** of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, denoted $\text{trace}(\mathbf{A})$, is the sum of diagonal elements in the matrix:

$$\text{trace}(\mathbf{A}) = \sum_{i=1}^n A_{i,i} \quad (7)$$

- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\text{trace}(\mathbf{A}) = \text{trace}(\mathbf{A}^\top)$.
- For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$, $\text{trace}(\mathbf{A} + \mathbf{B}) = \text{trace}(\mathbf{A}) + \text{trace}(\mathbf{B})$.
- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\alpha \in \mathbb{R}$, $\text{trace}(\alpha\mathbf{A}) = \alpha\text{trace}(\mathbf{A})$.
- For \mathbf{A}, \mathbf{B} such that \mathbf{AB} is square, $\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$.
- For $\mathbf{A}, \mathbf{B}, \mathbf{C}$ such that \mathbf{ABC} is square, $\text{trace}(\mathbf{ABC}) = \text{trace}(\mathbf{BCA}) = \text{trace}(\mathbf{CAB})$

Basic matrix definitions contd.

Definition 24 (Recall the definition of Span)

The **span** of a set of vectors, $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$, is the set of all possible **linear combinations** of these vectors; i.e.,

$$\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_k\} = \{\alpha_1 + \dots + \alpha_k \mathbf{x}_k \mid \alpha_1, \dots, \alpha_k \in \mathbb{R}\}. \quad (8)$$

- If $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ is a set of k linearly independent vectors of $\mathbf{x}_i \in \mathbb{R}^n$, then $\text{span}(\{\mathbf{x}_1, \dots, \mathbf{x}_k\}) = \mathbb{R}^n$.
⇒ Any vector $v \in \mathbb{R}^n$ can be written as a linear combination of \mathbf{x}_1 through \mathbf{x}_k .
- The **projection** of a vector $y \in \mathbb{R}^n$ onto the span of $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$, is the vector $\mathbf{v} \in \text{span}(\{\mathbf{x}_1, \dots, \mathbf{x}_k\})$ such that \mathbf{v} is as close as possible to \mathbf{y} :

$$\text{Proj}(y; \{\mathbf{x}_1, \dots, \mathbf{x}_k\}) = \arg \min_{\mathbf{v} \in \text{span}(\{\mathbf{x}_1, \dots, \mathbf{x}_k\})} \|y - \mathbf{v}\|_2 \quad (9)$$

Basic matrix definitions contd.

Definition 25 (Range of a matrix)

The **range** of a matrix, $\mathbf{A} \in \mathbb{R}^{n \times p}$, (denoted by $\text{range}(\mathbf{A})$) is defined as

$$\text{range}(\mathbf{A}) = \{\mathbf{Ax} \mid \mathbf{x} \in \mathbb{R}^p\} \subseteq \mathbb{R}^n \quad (10)$$

- $\text{range}(\mathbf{A})$ is the **span** of the columns (or the **column space**) of \mathbf{A} .
- The projection of a vector $\mathbf{y} \in \mathbb{R}^n$ onto the range of \mathbf{A} (let \mathbf{A} be full rank and that $p < n$) is given by:

$$\text{Proj}(\mathbf{y}; \mathbf{A}) = \arg \min_{\mathbf{v} \in \text{range}(\mathbf{A})} \|\mathbf{v} - \mathbf{y}\|_2 = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} \quad (11)$$

⇒ equivalent to the least-square objective.

Basic matrix definitions contd.

Definition 26 (Nullspace of a matrix)

The **nullspace** of a matrix, $\mathbf{A} \in \mathbb{R}^{n \times p}$, (denoted by $\text{null}(\mathbf{A})$) is defined as

$$\text{null}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{Ax} = \mathbf{0}\} \quad (12)$$

- $\text{null}(\mathbf{A})$ is the set of vectors mapped to **zero** by \mathbf{A} .
- $\text{null}(\mathbf{A})$ is the set of vectors **orthogonal** to the *rows* of \mathbf{A} .
- The vectors in $\text{range}(\mathbf{A})$ are of size n , while the vectors in the $\text{null}(\mathbf{A})$ are of size p .
 \Rightarrow vectors in $\text{range}(\mathbf{A}^\top)$ and $\text{null}(\mathbf{A})$ are both in \mathbb{R}^p .
- $\text{range}(\mathbf{A}^\top)$ and $\text{null}(\mathbf{A})$ are disjoint subsets that together span the entire space of \mathbb{R}^p .
Sets of this type are called **orthogonal complements** $\Rightarrow \text{range}(\mathbf{A}^\top) = \text{null}(\mathbf{A})^\perp$

Basic matrix definitions contd.

Definition 27 (Rank of a matrix)

The **rank** of a matrix, $\mathbf{A} \in \mathbb{R}^{n \times p}$, (denoted by $\text{rank}(\mathbf{A})$) is defined as

$$\text{rank}(\mathbf{A}) = \dim(\text{range}(\mathbf{A})) \quad (13)$$

- $\text{rank}(\mathbf{A})$ is the maximum number of **independent** columns (or rows) of \mathbf{A} ,
 $\Rightarrow \text{rank}(\mathbf{A}) \leq \min(n, p)$. If $\text{rank}(\mathbf{A}) = \min(n, p)$, then \mathbf{A} is said to be **full rank**.
- $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^\top)$; and $\text{rank}(\mathbf{A}) + \dim(\text{null}(\mathbf{A})) = n$.
- For $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times p}$, $\text{rank}(\mathbf{AB}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$
- For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, $\text{rank}(\mathbf{A} + \mathbf{B}) \leq \text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B})$

Basic matrix definitions contd.

Definition 28 (The inverse)

The **inverse** of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is denoted \mathbf{A}^{-1} , and is the unique matrix such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1} \quad (14)$$

- Not all matrices have inverses. Non-square matrices do not have inverses by definition.
- \mathbf{A} is **invertible** or **non-singular** if \mathbf{A}^{-1} exists, and **non-invertible** or **singular** otherwise.
- In order for a square matrix \mathbf{A} to have an inverse \mathbf{A}^{-1} , then \mathbf{A} must be full rank.
- The following are properties of the inverse (all assume that $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ are non-singular):
 - $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
 - $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
 - $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$

Basic matrix definitions contd.

Definition 29 (Orthogonal matrices)

Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are **orthogonal** if $\mathbf{x}^\top \mathbf{y} = 0$. A vector $\mathbf{x} \in \mathbb{R}^n$ is **normalized** if $\|\mathbf{x}\|_2 = 1$. A square matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ is **orthogonal** if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being **orthonormal**):

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I} = \mathbf{U} \mathbf{U}^\top \quad (15)$$

- The inverse of an orthogonal matrix is its transpose.
- Operating on a vector with an orthogonal matrix will not change its Euclidean norm, i.e.,

$$\|\mathbf{Ux}\|_2 = \|\mathbf{x}\|_2 \quad (16)$$

for any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{U} \in \mathbb{R}^{n \times n}$ orthogonal.

Definition 30 (Symmetric matrix)

A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **symmetric** if $\mathbf{A} = \mathbf{A}^\top$. It is **anti-symmetric** if $\mathbf{A} = -\mathbf{A}^\top$.

Basic matrix definitions contd.

Definition 31 (Positive semi-definite & positive definite matrices)

A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **positive semi-definite** (denoted $\mathbf{A} \succeq 0$) if $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$ for all $\mathbf{x} \neq 0$; while it is **positive definite** (denoted $\mathbf{A} \succ 0$) if $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$.

- $\mathbf{A} \succeq 0$ iff all its eigenvalues are non-negative, i.e., $\lambda_{\min}(\mathbf{A}) \geq 0$.
- Similarly, $\mathbf{A} \succ 0$ iff all its eigenvalues are positive, i.e., $\lambda_{\min}(\mathbf{A}) > 0$.
- \mathbf{A} is negative semi-definite if $-\mathbf{A} \succeq 0$; while \mathbf{A} is negative definite if $-\mathbf{A} \succ 0$.
- Semi-definite ordering of two symmetric matrices, \mathbf{A} and \mathbf{B} : $\mathbf{A} \succeq \mathbf{B}$ if $\mathbf{A} - \mathbf{B} \succeq 0$.

Example 32 (Matrix inequalities)

- 1 If $\mathbf{A} \succeq 0$ and $\mathbf{B} \succeq 0$, then $\mathbf{A} + \mathbf{B} \succeq 0$
- 2 If $\mathbf{A} \succeq \mathbf{B}$ and $\mathbf{C} \succeq \mathbf{D}$, then $\mathbf{A} + \mathbf{C} \succeq \mathbf{B} + \mathbf{D}$.
- 3 If $\mathbf{B} \preceq 0$, then $\mathbf{A} + \mathbf{B} \preceq \mathbf{A}$
- 4 If $\mathbf{A} \succeq 0$ and $\alpha \geq 0$, then $\alpha \mathbf{A} \succeq 0$
- 5 If $\mathbf{A} \succ 0$, then $\mathbf{A}^2 \succ 0$
- 6 If $\mathbf{A} \succ 0$, then $\mathbf{A}^{-1} \succ 0$

Matrix decompositions

Definition 33 (Eigenvalues & Eigenvectors)

The vector \mathbf{x} is an **eigenvector** of a *square* matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ if $\mathbf{Ax} = \lambda\mathbf{x}$, where $\lambda \in \mathbb{R}$ is called an **eigenvalue** of \mathbf{A} .

Lemma 34 (Properties of eigenvalues and eigenvectors for a symmetric \mathbf{A})

- *The trace of \mathbf{A} is equal to the sum of its eigenvalues, $\text{trace}(\mathbf{A}) = \sum_{i=1}^n \lambda_i$*
- *The determinant of \mathbf{A} is equal to the product of its eigenvalues, $\det(\mathbf{A}) = \prod_{i=1}^n \lambda_i$*
- *The rank of \mathbf{A} is equal to the number of non-zero eigenvalues of \mathbf{A}*
- *If \mathbf{A} is non-singular, then $1/\lambda_i$ is an eigenvalue of \mathbf{A}^{-1} with associated eigenvector \mathbf{x}_i , i.e., $\mathbf{A}^{-1}\mathbf{x}_i = (1/\lambda_i)\mathbf{x}_i$*
- *The eigenvalues of a diagonal matrix $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$ are just the diagonal entries d_1, \dots, d_n*

Matrix decompositions contd.

Definition 35 (Eigenvalue decomposition)

The **eigenvalue decomposition** of a **square** matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$, is given by:

$$\mathbf{A} = \mathbf{X}\Lambda\mathbf{X}^{-1} \tag{17}$$

- the columns of $\mathbf{X} \in \mathbb{R}^{n \times n}$, i.e. x_i , are **eigenvectors** of \mathbf{A}
- $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, where λ_i (also denoted $\lambda_i(\mathbf{A})$) are **eigenvalues** of \mathbf{A}
- A matrix that admits this decomposition is therefore called **diagonalizable** matrix

Definition 36 (Eigendecomposition of symmetric matrices)

If $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **symmetric**, the decomposition becomes $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top$, where $\mathbf{U} \in \mathbb{R}^{n \times n}$ is **unitary** (or **orthonormal**), i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and λ_i are real.

If we order $\lambda_1 \geq \lambda_2 \dots \geq \lambda_n$, $\lambda_i(\mathbf{A})$ becomes the i^{th} largest eigenvalue of \mathbf{A} :

- $\lambda_n(\mathbf{A}) = \lambda_{\min}(\mathbf{A})$ is the minimum eigenvalue of \mathbf{A}
- $\lambda_1(\mathbf{A}) = \lambda_{\max}(\mathbf{A})$ is the maximum eigenvalue of \mathbf{A}

Matrix decompositions contd

Definition 37 (Singular value decomposition)

The **singular value decomposition** (SVD) of a matrix, $\mathbf{A} \in \mathbb{R}^{n \times p}$, is given by:

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \quad (18)$$

- $\text{rank}(\mathbf{A}) = r \leq \min(n, p)$ and σ_i is the i^{th} **singular value** of \mathbf{A}
 - \mathbf{u}_i and \mathbf{v}_i are the i^{th} **left and right singular vectors** of \mathbf{A} respectively
 - $\mathbf{U} \in \mathbb{R}^{n \times r}$ and $\mathbf{V} \in \mathbb{R}^{p \times r}$ are unitary matrices (i.e., $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$)
 - $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$, where $\sigma_1 \geq \sigma_2 \dots \geq \sigma_r \geq 0$
-
- \mathbf{v}_i are **eigenvectors** of $\mathbf{A}^\top \mathbf{A}$; $\sigma_i = \sqrt{\lambda_i(\mathbf{A}^\top \mathbf{A})}$ (and $\lambda_i(\mathbf{A}^\top \mathbf{A}) = 0$ for $i > r$) since $\mathbf{A}^\top \mathbf{A} = (\mathbf{U}\Sigma\mathbf{V}^\top)^\top (\mathbf{U}\Sigma\mathbf{V}^\top) = (\mathbf{V}\Sigma^2\mathbf{V}^\top)$
 - \mathbf{u}_i are **eigenvectors** of $\mathbf{A}\mathbf{A}^\top$; $\sigma_i = \sqrt{\lambda_i(\mathbf{A}\mathbf{A}^\top)}$ (and $\lambda_i(\mathbf{A}\mathbf{A}^\top) = 0$ for $i > r$) since $\mathbf{A}\mathbf{A}^\top = (\mathbf{U}\Sigma\mathbf{V}^\top)(\mathbf{U}\Sigma\mathbf{V}^\top)^\top = (\mathbf{U}\Sigma^2\mathbf{U}^\top)$

Definition 38 (LU)

The **LU factorization** of a **non-singular square** (full-rank) matrix, $\mathbf{A} \in \mathbb{R}^{p \times p}$, is given by:

$$\mathbf{A} = \mathbf{PLU} \tag{19}$$

where **P** is a **permutation matrix**¹, **L** is **lower triangular** and **U** is **upper triangular**.

Definition 39 (QR)

The **QR factorization** of any matrix, $\mathbf{A} \in \mathbb{R}^{n \times p}$, is given by $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is an **orthonormal** matrix, i.e., $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$, and $\mathbf{R} \in \mathbb{R}^{n \times p}$ is **upper triangular**.

Definition 40 (Cholesky)

The **Cholesky factorization** of a **positive definite and symmetric** matrix, $\mathbf{A} \in \mathbb{R}^{p \times p}$, is given by $\mathbf{A} = \mathbf{LL}^\top$, where **L** is a **lower triangular** matrix with **positive** entries on the **diagonal**.

¹A matrix $\mathbf{P} \in \mathbb{R}^{p \times p}$ is **permutation** if it has only one 1 in each row and each column.

Matrix norms

Definition 41 (Matrix norm)

A norm of an $n \times p$ matrix is a map $\|\cdot\| : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}$ such that for all matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times p}$ and scalar $\lambda \in \mathbb{R}$

- 1 $\|\mathbf{A}\| \geq 0$ for all $\mathbf{A}^{n \times p}$ non-negativity
- 2 $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = 0$ definitiveness
- 3 $\|\lambda\mathbf{A}\| = |\lambda| \|\mathbf{A}\|$ homogeneity
- 4 $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ triangle inequality

Definition 42 (Matrix inner product)

Matrix inner product is defined as follows: $\langle \mathbf{A}, \mathbf{B} \rangle = \text{trace}(\mathbf{AB}^\top)$.

Definition 43 (Frobenius norm)

The **Frobenius norm** of $\mathbf{A} \in \mathbb{R}^{n \times p}$ is $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^p A_{i,j}^2} = \sqrt{\text{trace}(\mathbf{AA}^\top)}$.

Matrix norms contd.

Definition 44 (Operator norm)

The **operator norm** between ℓ_q and ℓ_r ($1 \leq q, r \leq \infty$) of a matrix \mathbf{A} is defined as

$$\|\mathbf{A}\|_{q \rightarrow r} = \sup_{\|\mathbf{x}\|_q \leq 1} \|\mathbf{Ax}\|_r . \quad (20)$$

Lemma 45 (Useful relation for operator norms)

The following **identity** holds

$$\|\mathbf{A}\|_{q \rightarrow r} := \max_{\|\mathbf{z}\|_r \leq 1, \|\mathbf{x}\|_q = 1} \langle \mathbf{z}, \mathbf{Ax} \rangle = \max_{\|\mathbf{z}\|_{q'} \leq 1, \|\mathbf{x}\|_{r'} = 1} \langle \mathbf{A}^\top \mathbf{z}, \mathbf{x} \rangle =: \|\mathbf{A}^\top\|_{q' \rightarrow r'} , \quad (21)$$

whenever $1/q + 1/q' = 1 = 1/r + 1/r'$.

Table of Contents

1 Introduction to Deep Learning

2 Review: Linear Algebra

3 Review: Probability Theory

- Elements of probability
- Random variables
- Two random variables

Reference

- Arian Maleki and Tom Do, Review of Probability Theory.
<http://cs229.stanford.edu/section/cs229-prob.pdf>, 2012

Table of Contents

1 Introduction to Deep Learning

- From ANNs to Deep Learning
- Current Applications and Success

2 Review: Linear Algebra

- Notation
- Vectors
- Matrices

3 Review: Probability Theory

- **Elements of probability**
- Random variables
- Two random variables

Definition 46 (Sample space Ω)

The sample space Ω of a random experiment is the set Ω of all the outcomes of that experiment.

Each outcome $\omega \in \Omega$ can be a complete description of the state of the real world at the end of the experiment.

Definition 47 (Set of events \mathcal{F})

A set whose elements $A \in \mathcal{F}$ are subsets of Ω , i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment.

\mathcal{F} should satisfy three properties

- ① $\emptyset \in \mathcal{F}$
- ② $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$
- ③ $A_1, \dots \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$

Definition 48 (Probability measure)

The probability measure is a function $P : \mathcal{F} \rightarrow \mathbb{R}$ that maps event \mathcal{F} onto the interval $[0, 1]$ and satisfies the following properties,

- $P(A) \geq 0$, for all $A \in \mathcal{F}$
- $P(\Omega) = 1$
- If A_1, \dots are disjoint / independent events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P(\cup_i A_i) = \sum_i P(A_i) \tag{22}$$

Properties:

- If $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A \cap B) \leq \min(P(A), P(B))$
- (Union Bound) $P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega \setminus A) = 1 - P(A)$
- (Law of Total Probability) If A_1, \dots, A_k are a set of disjoint events such that $\bigcup_{i=1}^k A_i = \Omega$, then $\sum_{i=1}^k P(A_i) = 1$.

Definition 49 (Conditional probability and independence)

Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as,

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}. \quad (23)$$

- $P(A|B)$ is the probability measure of the event A after observing the occurrence of B .
- Two events are called independent if and only if $P(A \cap B) = P(A)P(B)$.

Table of Contents

1 Introduction to Deep Learning

- From ANNs to Deep Learning
- Current Applications and Success

2 Review: Linear Algebra

- Notation
- Vectors
- Matrices

3 Review: Probability Theory

- Elements of probability
- **Random variables**
- Two random variables

Random Variables

Definition 50 (Random Variable)

A real-valued random variable, $X(\omega)$ or X , is a **function** $X : \Omega \rightarrow \mathbb{R}$, that associates a value to the outcome of a randomized experiment

Example 51

- Whether a coin flip was headed: a function from $\Omega = \{H, T\}$ to $\{0, 1\}$
- Number of heads in a sequence of n throws: function from $\Omega = \{H, T\}^n$ to $\{0, 1, \dots, n\}$.

Remark 52

- *Not every function is acceptable as a random variable.*
- *From a measure-theoretic perspective, random variables must be Borel-measurable functions².*

²Intuitively, this restriction ensures that given a random variable and its underlying outcome space, one can implicitly define the each of the events of the event space as being sets of outcomes $\omega \in \Omega$ for which $X(\omega)$ satisfies some properties.

Example 53 (Discrete random variable)

The probability of the set associated with a random variable X taking on some specific value k is

$$P(X = k) := P(\{\omega : X(\omega) = k\}). \quad (24)$$

Example 54 (Continuous random variable)

The probability that X takes on a value between two real constants a and b (where $a < b$) as

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\}). \quad (25)$$

Cumulative distribution functions

By using the function below we can calculate the probability of any event in \mathcal{F} .

Definition 55 (CDF)

A **Cumulative Distribution Function (CDF)** is a function $F_X : \mathbb{R} \rightarrow [0, 1]$ which specifies a probability measure as,

$$F_X(x) \triangleq P(X \leq x) \tag{26}$$

Properties:

- $0 \leq F_X(x) \leq 1$
- $\lim_{x \rightarrow -\infty} F_X(x) = 0$
- $\lim_{x \rightarrow \infty} F_X(x) = 1$
- $x \leq y \Rightarrow F_X(x) \leq F_X(y)$

Probability mass functions

When a random variable X takes on a finite set of possible values (i.e., X is a discrete random variable).

Definition 56 (Probability mass functions)

A **Probability Mass Function (PMF)** is a function $p_X : \Omega \rightarrow \mathbb{R}$ such that

$$p_X(x) \triangleq P(X = x) \quad \text{for every } x \in \mathcal{X}. \quad (27)$$

In the case of discrete random variable, we use the notation \mathcal{X} for the set of possible values that the random variable X may assume.

Properties:

- $0 \leq p_X(x) \leq 1$ for every $x \in \mathcal{X}$
- $\sum_{x \in \mathcal{X}} p_X(x) = 1$
- $\sum_{x \in \mathcal{A}} p_X(x) = P(X \in \mathcal{A})$

Probability density functions

For some continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere.

Definition 57 (Probability density functions)

We define the **Probability Density Function** (PDF) as the derivative of the CDF, i.e.,

$$f_X(x) \triangleq \frac{dF_X(x)}{dx} \quad (28)$$

- For very small Δx , $P(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x$
- The value of PDF at any given point x is not the probability of that event, i.e.,

$$f_X(x) \neq P(X = x). \quad (29)$$

- Properties:

- The density is non-negative: i.e., $f_X(x) \geq 0$ for any x
- Probabilities integrate to 1: $\int_{-\infty}^{\infty} f_X(x) dx = 1$
- $\int_{x \in \mathcal{A}} f_X(x) dx = P(X \in \mathcal{A})$

Expectation

Definition 58 (Expectation of random variable)

Discrete random variable: Suppose that X is a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \rightarrow \mathbb{R}$ is an arbitrary function ($g(X)$ can be considered as a random variable). The expectation or expected value of $g(X)$ is defined as

$$\mathbb{E}[g(X)] \triangleq \sum_{x \in \text{Val}(X)} g(x)p_X(x). \quad (30)$$

Continuous random variable: If X is a continuous random variable with PDF $f_X(x)$, then the expected value of $g(X)$ is defined as

$$\mathbb{E}[g(X)] \triangleq \int_{-\infty}^{\infty} g(x)f_X(x)dx. \quad (31)$$

The expectation of $g(X)$ can be thought of as a “weighted average” of the values that $g(x)$ can be taken on for different values of x , where the weights are given by $p_X(x)$ or $f_X(x)$.

Properties:

- $\mathbb{E}[X]$ of a random variable itself is found by letting $g(x) = x$
- $\mathbb{E}[a] = a$ for any constant $a \in \mathbb{R}$
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$ for any constant $a \in \mathbb{R}$
- (Linearity of Expectation) $\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)]$
- For a discrete random variable X , $\mathbb{E}[1\{X = k\}] = P(X = k)$

Variance

Definition 59 (The variance of a random variable X)

The variance of a random variable X is defined as

$$\text{Var}(X) \triangleq \mathbb{E} [(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad (32)$$

Properties:

- $\text{Var}[a] = 0$ for any constant $a \in \mathbb{R}$
- $\text{Var}[af(X)] = a^2\text{Var}[f(X)]$ for any constant $a \in \mathbb{R}$

Some common random variables

Discrete random variables:

- $X \sim \text{Bernoulli}(q)$ (where $0 \leq q \leq 1$): one if a coin with heads probability p comes up heads, zero otherwise.

$$p(x) = \begin{cases} q & \text{if } x = 1 \\ 1 - q & \text{if } x = 0 \end{cases} \quad (33)$$

- $X \sim \text{Binomial}(n, q)$ (where $0 \leq q \leq 1$): the number of heads in n independent flips of a coin with heads probability q

$$p(x) = \binom{n}{x} q^x (1 - q)^{n-x} \quad (34)$$

- $X \sim \text{Geometric}(q)$ (where $q > 0$): the number of flips of a coin with heads probability q until the first heads.

$$p(x) = q(1 - q)^{x-1} \quad (35)$$

- $X \sim \text{Poisson}(\lambda)$ (where $\lambda > 0$): a probability distribution over the non-negative integers used for modeling the frequency of rare events:

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!} \quad (36)$$

Some common random variables

Continuous random variables:

- $X \sim \text{Uniform}(a, b)$ (where $a < b$): equal probability density to every value between a and b on the real line

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad (37)$$

- $X \sim \text{Exponential}(\lambda)$ (where $\lambda > 0$): decaying probability density over the non-negative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (38)$$

- $X \sim \text{Normal}(\mu, \sigma^2)$: also known as the Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad (39)$$

Table of Contents

1 Introduction to Deep Learning

- From ANNs to Deep Learning
- Current Applications and Success

2 Review: Linear Algebra

- Notation
- Vectors
- Matrices

3 Review: Probability Theory

- Elements of probability
- Random variables
- Two random variables

Joint and marginal CDFs

Definition 60 (Joint CDF of $F_{XY}(x, y)$)

The joint cumulative distribution function of X and Y is defined by

$$F_{XY}(x, y) = P(X \leq x, Y \leq y). \quad (40)$$

Definition 61 (Marginal CDF of $F_{XY}(x, y)$)

The joint CDF $F_{XY}(x, y)$ and the joint distribution functions $F_X(x)$ and $F_Y(y)$ of each variable separately are related by

$$F_X(x) = \lim_{y \rightarrow \infty} F_{XY}(x, y) \quad F_Y(y) = \lim_{x \rightarrow \infty} F_{XY}(x, y), \quad (41)$$

where $F_X(x)$ and $F_Y(y)$ are the **marginal cumulative distribution functions** of $F_{XY}(x, y)$.

- $0 \leq F_{XY}(x, y) \leq 1$
- $\lim_{x,y \rightarrow \infty} F_{XY}(x, y) = 1$
- $\lim_{x,y \rightarrow -\infty} F_{XY}(x, y) = 0$

Joint and marginal Probability Mass Functions (PMFs)

Definition 62 (Joint PMF)

If X and Y are discrete random variable, then the **joint PMF** $p_{XY} : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ is defined by

$$p_{XY}(x, y) = P(X = x, Y = y) \quad (42)$$

- $0 \leq P_{XY}(x, y) \leq 1$ for all x, y , and $\sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} P_{XY}(x, y) = 1$
- $p_X(x) = \sum_y p_{XY}(x, y)$

Definition 63 (Marginal PMF)

We refer to $p_X(x)$ as the **marginal PMF** of X .

Joint and marginal Probability Density Functions (PDFs)

Definition 64 (Joint PDF)

Let X and Y be two continuous random variables with joint distribution function F_{XY} . In the case that $F_{XY}(x, y)$ is everywhere differentiable in both x and y , then we can define the **joint PDF** as

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}, \quad (43)$$

Properties:

- $f_{XY}(x, y) \neq P(X = x, Y = y)$
- $\int \int_{x \in \mathcal{A}} f_{XY}(x, y) dx dy = P((X, Y) \in \mathcal{A})$
- The values of the PDF $f_{XY}(x, y)$ are always non-negative, but they may be greater than 1
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) = 1$

Definition 65 (Marginal PDF)

The marginal PDF of X is defined as $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$.

Conditional distributions

Definition 66 (Conditional distribution)

Discrete case: the conditional PMF of Y given X is simply

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}, \quad (44)$$

assuming that $p_X(x) \neq 0$.

Continuous case: the conditional PDF of Y given $X = x$ is simply

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}, \quad (45)$$

provided $f_X(x) \neq 0$.

Bayes's rule

Definition 67 (Bayes's rule)

In the case of **discrete random variable** of X and Y ,

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)} = \frac{p_{X|Y}(x|y)p_Y(y)}{\sum_{y' \in \text{val}(Y)} p_{X|Y}(x|y')p_Y(y')} \quad (46)$$

In the case of **continuous random variable** of X and Y ,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y')f_Y(y')dy'} \quad (47)$$

Constituents of $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$:

- $P(A)$, the **prior** probability, is the probability of A before B is observed.
- $P(A|B)$, the **posterior** probability, is the probability of A given B , i.e., after B is observed.
- $P(B|A)$ is the probability of observing B given A . As a function of A with B fixed, this is the **likelihood**.

Independence

Two random variables X and Y are **independent** if $F_{XY}(x, y) = F_X(x)F_Y(y)$ for all values of x and y .

- For discrete random variables, $p_{XY}(x, y) = p_X(x)p_Y(y)$ for all $x \in \text{Val}(X), y \in \text{Val}(Y)$.
- For discrete random variables, $p_{Y|X}(y|x) = p_Y(y)$ whenever $p_X(x) \neq 0$ for all $y \in \text{Val}(Y)$.
- For continuous random variables, $f_{XY}(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.
- For continuous random variables, $f_{Y|X}(x, y) = f_Y(y)$ whenever $f_X(x) \neq 0$ for all $y \in \mathbb{R}$.

Lemma 68

If X and Y are independent, then for any subsets $\mathcal{A}, \mathcal{B} \subseteq \mathbb{R}$, we have

$$P(X \in \mathcal{A}, Y \in \mathcal{B}) = P(X \in \mathcal{A})P(Y \in \mathcal{B}) \quad (48)$$

Remark 69

If X is independent of Y , then any function of X is independent of any function of Y .

Expectation and co-variance

Definition 70 (Expectation)

Discrete case: Suppose that we have two discrete random variables X, Y and $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a function of these two random variables. The expected value of $g(X, Y)$ is defined as:

$$\mathbb{E}[g(X, Y)] \triangleq \sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} g(x, y) p_{XY}(x, y) \quad (49)$$

Continuous case: For continuous random variables X, Y , the analogous expression is

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy. \quad (50)$$

Definition 71 (Co-variance)

The co-variance of two random variables X and Y is defined as

$$\text{Cov}[X, Y] \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \quad (51)$$

Properties:

- When $\text{Cov}[X, Y] = 0$, we say that X and Y are uncorrelated \Rightarrow it is not equivalent to state that X and Y are independent
- (Linearity of expectation) $\mathbb{E}[f(X, Y) + g(X, Y)] = \mathbb{E}[f(X, Y)] + \mathbb{E}[g(X, Y)]$
- If X and Y are independent, then $\text{Cov}[X, Y] = 0$
- If X and Y are independent, then $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$