

小試身手



BiomedicalData.c
sv

您能否找出這筆些數據異常的部分

	A	B	C	D	E	F	G	H	I	
1	ID	Gender	Age	Height	Weight	BMI	Smoking	Alcohol	ExerciseHabit	
2		1 Female	42	158	78	31.2	No	No	Occasional	
3		2 Male	45	163	98	36.9	No	Yes	Regular	
4		3 Female	45	169	67	23.5	No	Yes	None	
5		4 F	44	171	50	17.1	No	Yes	Regular	
6		5 Female	48	158	97	38.9	Yes	No	Regular	
7		6 Female	49	189	60	16.8	No	Yes	Occasional	
8		7 Male	7	183	61	18.2	Yes	No	Occasional	
9		8 Male	21	191	69	18.9	Yes	No	Occasional	
10		9 Female	25	174	60	19.8	No		Occasional	
11		10 Female	6	161	81	31.2	No		Occasional	
12		11 Male		300	200		No	No	None	
13		12 Female	28	162	87	32.7	Yes	Yes	Regular	

課程主題一



長庚大學資工系



AI與數據分析的基本概念

上課資料 & 課前問卷：

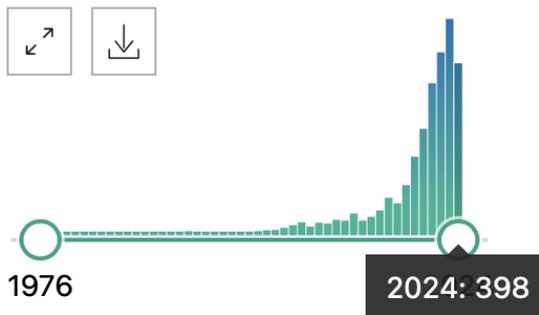


<https://forms.gle/LoapE8Kr1y3MuSkA6>

長庚大學 資工系 李季青

chichinglee@cgu.edu.tw





Rapid prediction of electron-ionization mass spectrometry using neural networks

JN Wei, D Belanger, RP Adams, D Sculley - ACS central science, 2019 - ACS Publications
 ... be easily retrained to predict mass spectra for other ionization ... mass spectra for small molecules from the NIST 2017 Mass ... similar to previously reported machine learning models, but ...
 ☆ 儲存 引 用 被引用 131 次 相關文章 全部共 16 個版本

Combining desorption electrospray ionization mass spectrometry imaging and machine learning for molecular recognition of myocardial infarction

K Margulis, Z Zhou, Q Fang, RE Sievers... - Analytical ..., 2018 - ACS Publications
 ... (mass spectral peak ... machine learning technique to categorize mass spectra obtained from the sweat of human volunteers. (30) In this present study, we extended this machine learning ...
 ☆ 儲存 引 用 被引用 28 次 相關文章 全部共 11 個版本

Fully automated unconstrained analysis of high-resolution mass spectrometry data with machine learning

DA Boiko, KS Kozlov, JV Burykina... - Journal of the ..., 2022 - ACS Publications
 ... machine learning-enabled deisotoping and untargeted spectra analysis (MEDUSA)—a framework for mass spectra processing, which includes a novel approach for spectra deisotoping...
 ☆ 儲存 引 用 被引用 31 次 相關文章 全部共 4 個版本

Unknown metabolite identification using machine learning collision cross-section prediction and tandem mass spectrometry

CK Asef, MA Rainey, BM Garcia, GJ Gouveia... - Analytical ..., 2023 - ACS Publications
 ...) measurements and machine learning CCS predictions to ... to poor matching of the machine learning training sets, limited ... Liquid chromatography mass spectrometry (LC-MS) remains ...
 ☆ 儲存 引 用 被引用 19 次 相關文章 全部共 4 個版本

Personal information from latent fingerprints using desorption electrospray ionization mass spectrometry and machine learning

Z Zhou, RN Zare - Analytical chemistry, 2017 - ACS Publications
 ... selection using the GDBT machine learning model, the ... identified by tandem mass spectrometry. This information provides ... the mass spectrometry combined with machine learning ...
 ☆ 儲存 引 用 被引用 104 次 相關文章 全部共 13 個版本

Toward Machine Learning-Driven Mass Spectrometric Identification of Trichothecenes in the Absence of Standard Reference Materials

BP Mayer, ML Dreyer, MC Prieto Conaway... - Analytical ..., 2023 - ACS Publications
 ... machine learning (ML) techniques toward identification of discriminative fragment ions from mass spectrometric ... a series of gradient-boosted machine learners are then used to develop ...
 ☆ 儲存 引 用 相關文章 全部共 4 個版本

Single-cell classification using mass spectrometry through interpretable machine learning

YR Xie, DC Castro, SE Bell, SS Rubakhin... - Analytical ..., 2020 - ACS Publications
 ... of resolving hundreds of molecules in each mass spectrum. We developed a machine learning workflow to classify single cells according to their mass spectra based on cell groups of ...
 ☆ 儲存 引 用 被引用 57 次 相關文章 全部共 10 個版本

New Approach Combining Molecular Fingerprints and Machine Learning to Estimate Relative Ionization Efficiency in Electrospray Ionization

AW Mayhew, DO Topping, JF Hamilton - ACS omega, 2020 - ACS Publications
 ... Electrospray ionization (ESI) is widely used as an ionization source for the analysis of complex mixtures by mass spectrometry. However, different compounds ionize more or less ...
 ☆ 儲存 引 用 被引用 17 次 相關文章 全部共 8 個版本

Utilization of machine learning for the differentiation of positional NPS isomers with direct analysis in real time mass spectrometry

JL Bonetti, S Samanipour, AC van Asten - Analytical Chemistry, 2022 - ACS Publications
 ... success in tackling this challenge, a machine learning technique, such as the Random ... by effectively utilizing minor differences in the mass spectra to assign the correct isomeric form. ...
 ☆ 儲存 引 用 被引用 13 次 相關文章 全部共 7 個版本

Sort by:



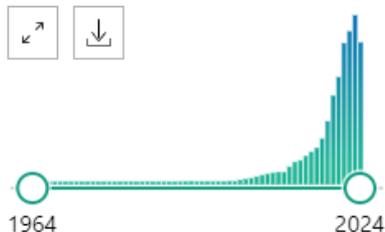


MY NCBI FILTERS

17,540 results

 << < Page of 1,754 > >>

RESULTS BY YEAR


 [Machine learning and statistical methods for clustering single-cell RNA-sequencing data.](#)

1

Cite Petegrosso R, Li Z, Kuang R.

Brief Bioinform. 2020 Jul 15;21(4):1209-1223. doi: 10.1093/bib/bbz063.

 Share PMID: 31243426 [Review](#).

This article reviews the **machine learning** and statistical methods for clustering scRNA-seq transcriptomes developed in the past few years. ...Several software packages developed to support the cluster **analysis** of scRNA-seq data are also reviewed and experimen ...

Sort by:



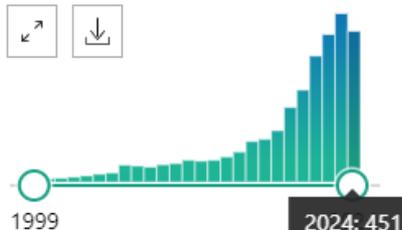


MY NCBI FILTERS

2,985 results

 << < Page of 299 > >>

RESULTS BY YEAR


 [Using machine learning approaches for multi-omics data analysis: A review.](#)

1

Reel PS, Reel S, Pearson E, Trucco E, Jefferson E.

Cite Biotechnol Adv. 2021 Jul-Aug;49:107739. doi: 10.1016/j.biotechadv.2021.107739. Epub 2021 Mar 29.

 PMID: 33794304 [Review](#).

Share **Machine learning** methods offer novel techniques to integrate and analyse the various omics data enabling the discovery of new biomarkers. ...It provides insight and recommendations for interdisciplinary professionals who envisage employing **machine learning** ...

The history of AI

1940s-1950s

Foundations of AI

In the 1940s, the first artificial neurons were conceptualised. The 1950s introduced us to the Turing Test and the term "Artificial Intelligence".



1960s-1970s

Early Development

The 60s and 70s brought the birth of ELIZA, simulating human conversation, and Dendral, the first expert system, showcasing the early potentials of AI.



1980s

AI Winter & Expert Systems

The 80s faced reduced AI funding but saw the inaugural National Conference on AI. The backpropagation concept rejuvenated neural networks.



1990s

Revival & Emergence of ML

The 90s witnessed IBM's Deep Blue defeating chess champion Garry Kasparov and the inception of the LOOM project, laying the foundations for GenAI.



2000s

The Genesis of Generative AI

Geoffrey Hinton propelled deep learning into the limelight, steering AI toward relentless growth and innovation.



2010s

Rise of AI

In 2011, IBM Watson won "Jeopardy!", highlighting AI's language skills. The 2010s marked major AI milestones, including pioneering work in image recognition and the birth of GANs in 2014, followed by OpenAI's founding in 2015.



2020s

GenAI Reaches New Horizons

At the start of this decade, we've seen significant strides in GenAI notably with OpenAI's GPT-3 and DALL-E. 2023 welcomed advanced tools like ChatGPT-4 and Google's Bard, alongside Microsoft's Bing AI, enhancing accessibility and reliability of information.

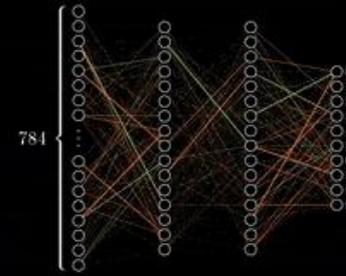


2020s

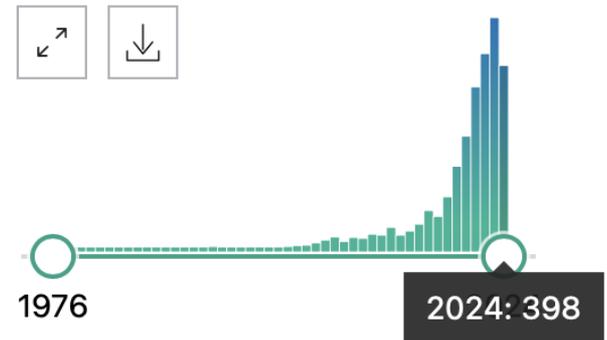
Neural Network in Action (<https://gfycat.com/ko/gifs/search/deep+learning>)

The History of AI: A Timeline from 1940 to 2023 + Infographic

Training in progress...



訓練需要大量計算
網路越複雜，計算量越高



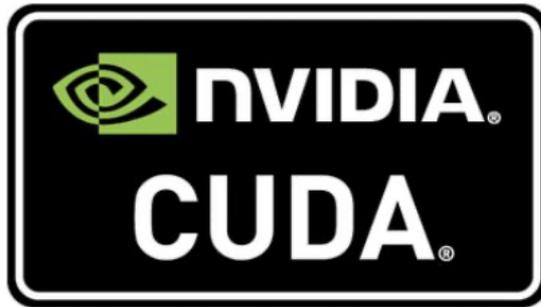
1999

NVIDIA 首開先河，推出 GPU

NVIDIA 發明創造了繪圖處理器，並導入產線，重新改造了整個產業。GeForce 256 是當時推出的全球首款 GPU，NVIDIA 對此一名詞的定義為「整合變換、光亮、三角形設定/修剪的單晶片處理器，並能將引擎轉換為每秒至少處理 1000 萬個多邊形。」當時最新的 GPU 每秒可處理超過 20 億個多邊形。



2007

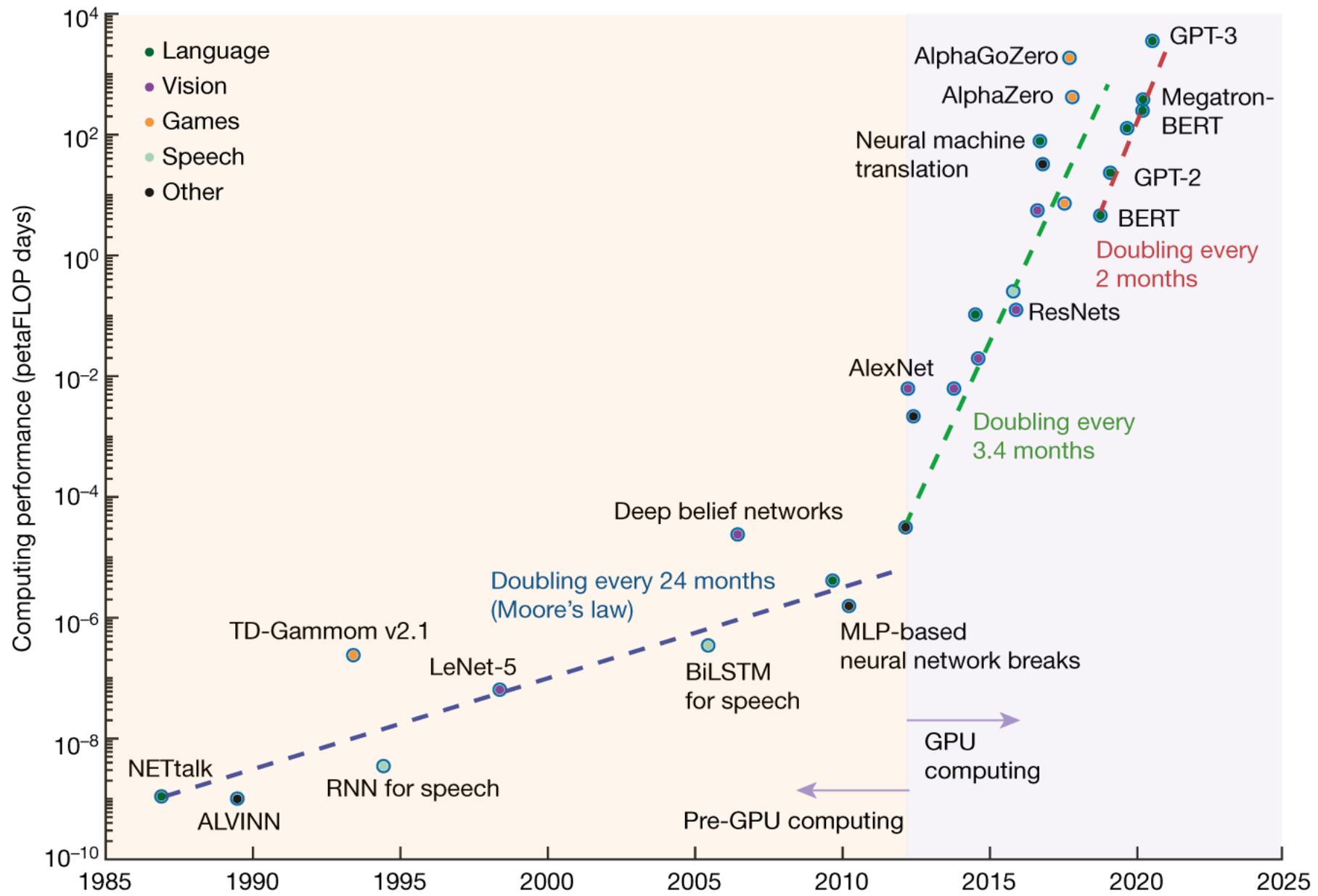


CUDA 允許軟體開發者和程式設計師使用 NVIDIA 的圖形處理單元 (GPU) 來進行非常複雜的運算任務。

CUDA 讓普通的圖形卡不僅能處理電玩和圖形渲染，還能處理科學計算和工程問題，例如天氣預測、分子動力學模擬等。

Compute Unified Device Architecture (CUDA) , 17 years ago

a Computing power demands



Part 1

AI 的基礎概念

AI 基礎概念

1. Encoding (讓電腦看得懂)

2. Training

3. AI 跑起來

從閉著眼睛分辨水果開始

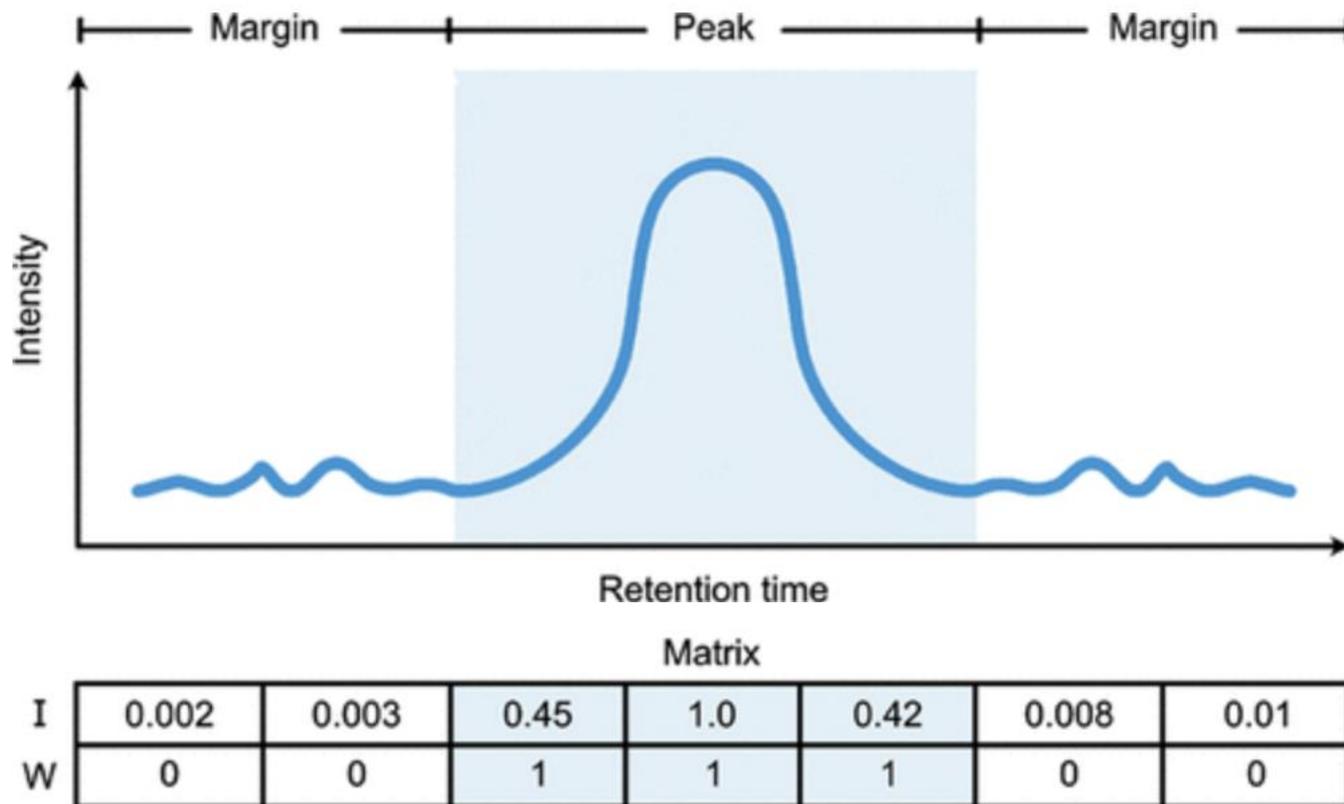


1. Encoding : 讓電腦看得懂



	氣味	脆度	多汁	刺刺	皮光滑	重量
Apple	1	1	0	0	1	0
Pineapple	1	0	1	1	0	1
Passion Fruit	0	0	1	0	0	0
Durian	1	0	0	1	0	1

1. Encoding : 讓電腦看得懂



2. Training



企鵝狗

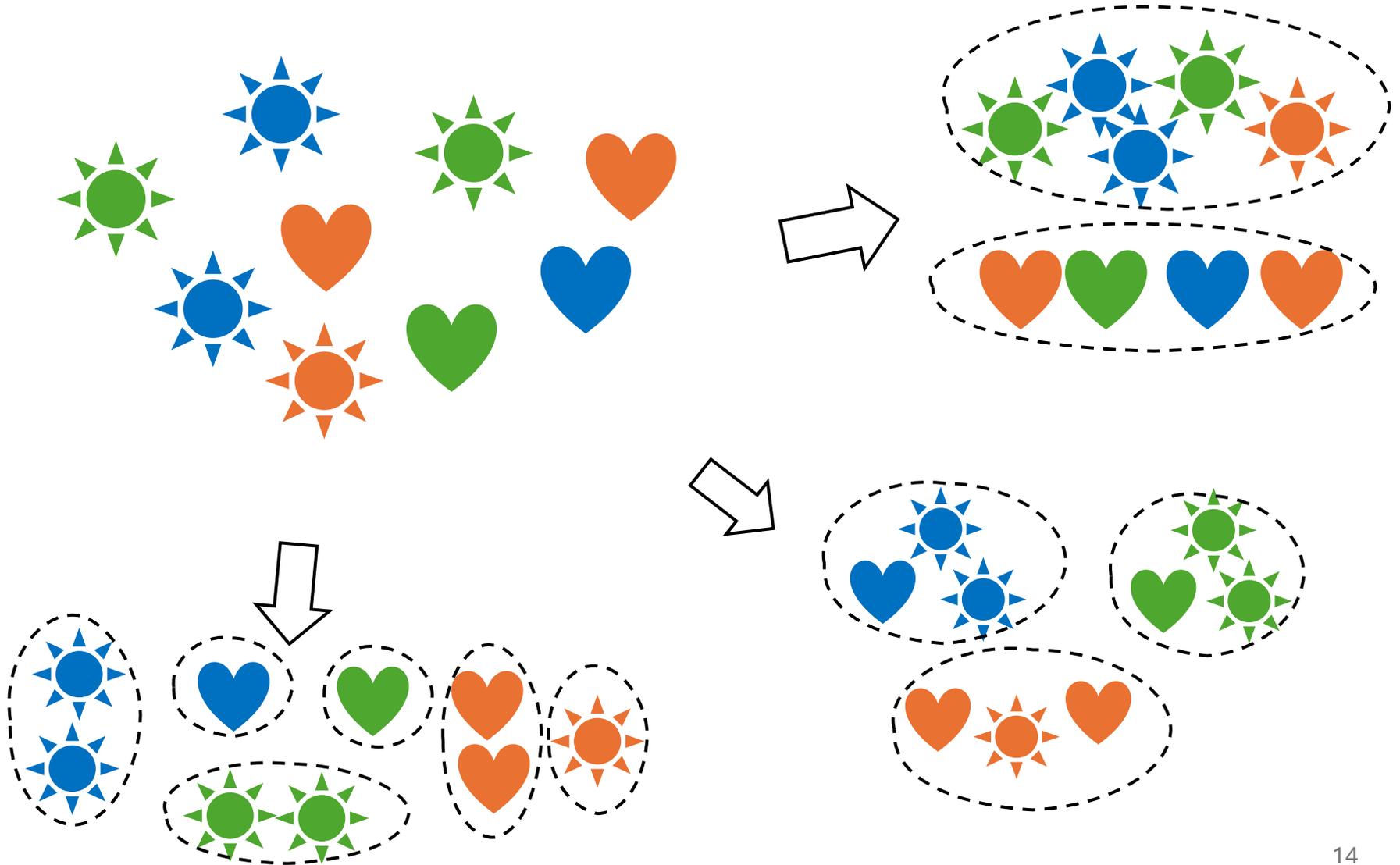


Supervise learning
監督式學習

模型在已標記的數據集上進行訓練。這些數據集包括輸入數據和相應的正確輸出（標籤）

Unsupervised learning 非監督式學習

不依賴於預先標記好的訓練數據。
非監督式學習使用未標記的數據來從中找出
內在的模式和結構。





氣味	脆度	多汁	刺刺	皮光滑	沉甸甸
1	1	0	0	1	0
1	0	1	1	0	1
0	0	1	0	0	0
1	0	0	1	0	1

加起來

3

4

1

3

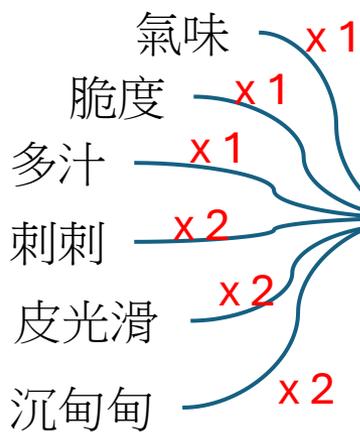


=





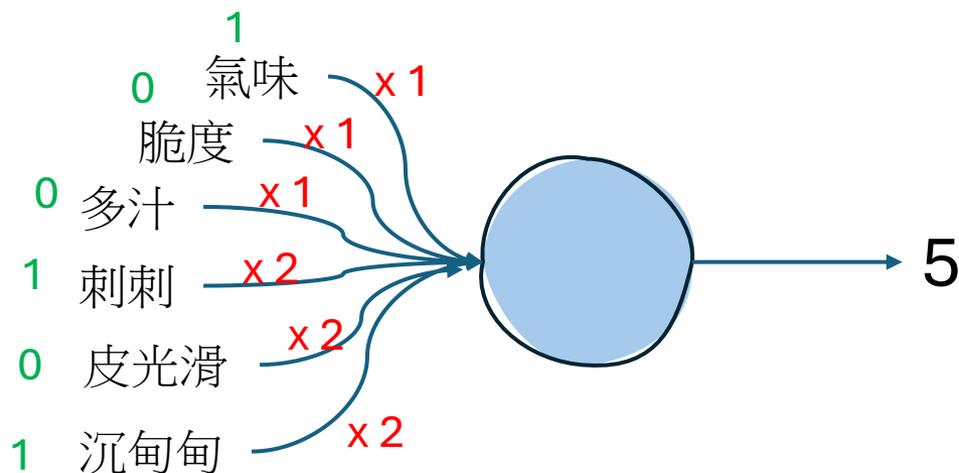
Smell 氣味	Firmness 脆度	Juiciness 多汁	Spikiness 刺刺	Smoothness of the skin 皮光滑	Weight 重量	加起來
1 x 1	1 x 1	0 x 1	0 x 2	1 x 3	0 x 3	5
1 x 1	0 x 1	1 x 1	1 x 2	0 x 3	1 x 3	7
0 x 1	0 x 1	1 x 1	0 x 2	0 x 3	0 x 3	1
1 x 1	0 x 1	0 x 1	1 x 2	0 x 3	1 x 3	6



权重
Weighting

5

Loss Function
計算真實答案
跟預測值的差異



$$y = \text{氣味} \times 1 + \text{脆度} \times 1 + \text{多汁} \times 1 + \text{刺刺} \times 2 + \text{皮光滑} \times 2 + \text{沉甸甸} \times 2$$

一開始，權重為隨機亂數

(Loss function)

為了找出最適當的權重，我們不斷的把水果丟進去，跟標準答案比較偏差多遠

朝著影響答案比較大的那些輸入值為方向去調整權重

透過 偏微分 這個工具，來幫忙找出哪些輸入影響比較大

$$y = 5x$$

$$y' = 5$$

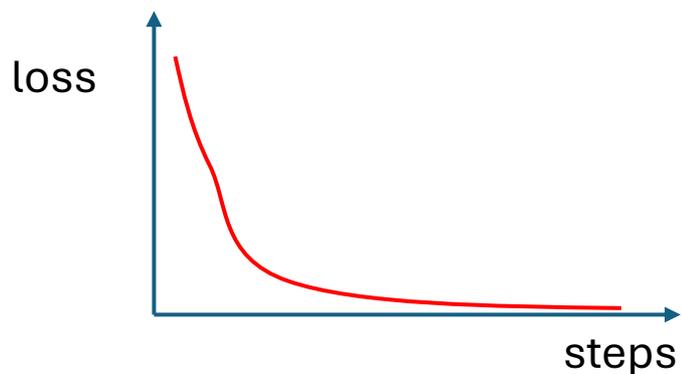
$$y = 10x$$

$$y' = 10$$

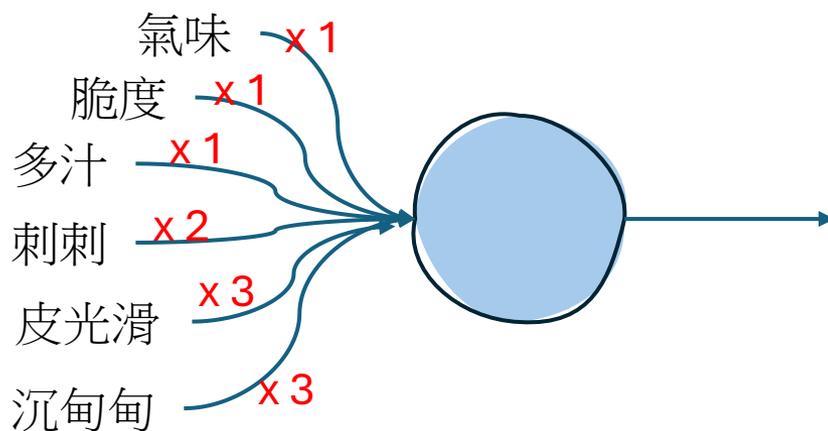
$$y = 5x^3$$

$$y' = 15x^2$$

隨著水果越丟愈多，權重不斷被改變，越來越接近標準答案



Loss 下降到一定程度後，把所有的權重記下來，模型訓練完成



3. AI 跑起來

正確答案：

5



1
1
0
0
1
0

7



1
0
1
1
0
1

1

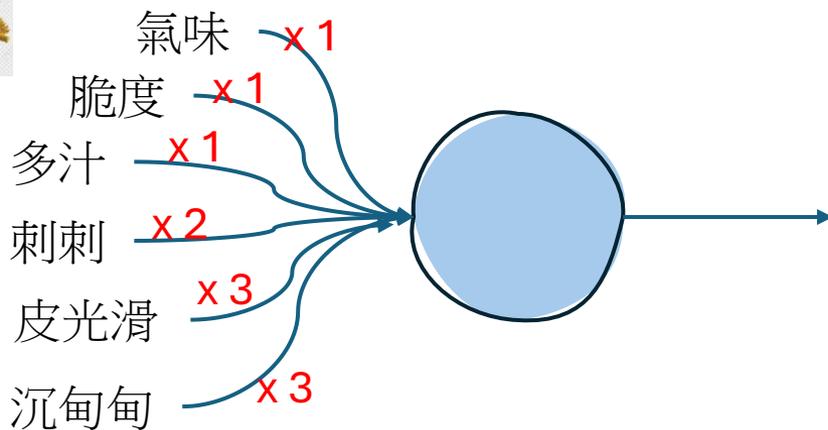


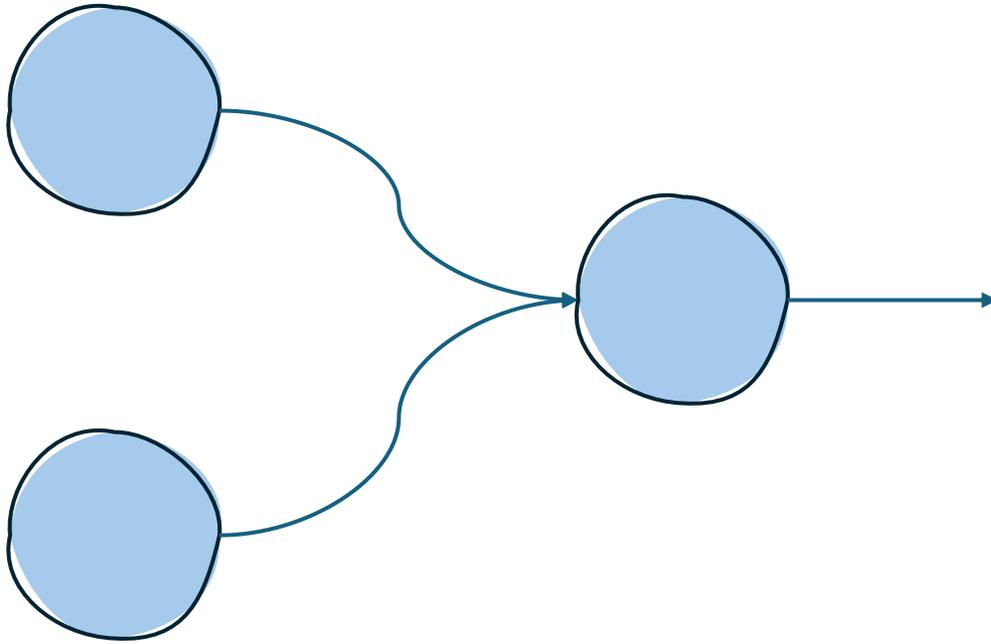
0
0
1
0
0
0

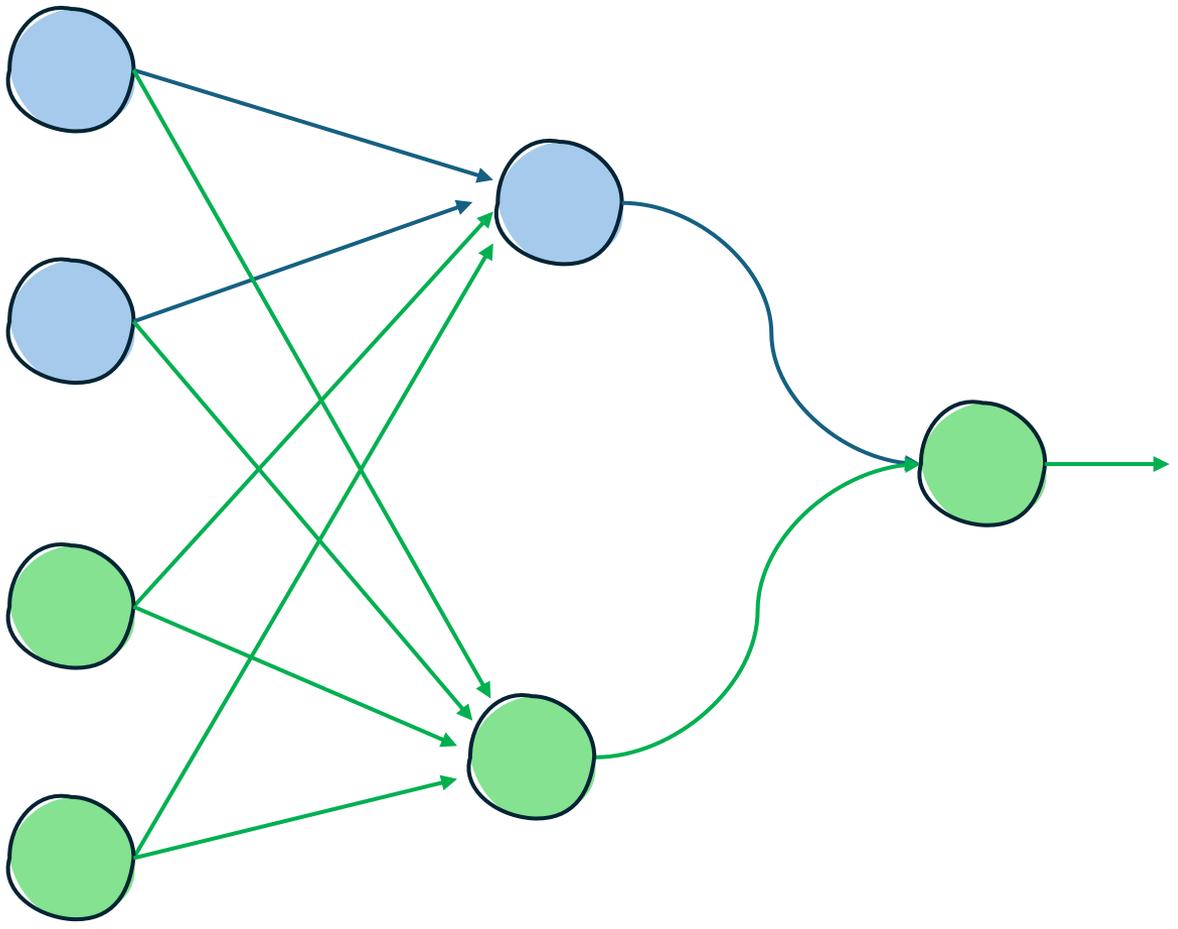
6



1
0
0
1
0
1

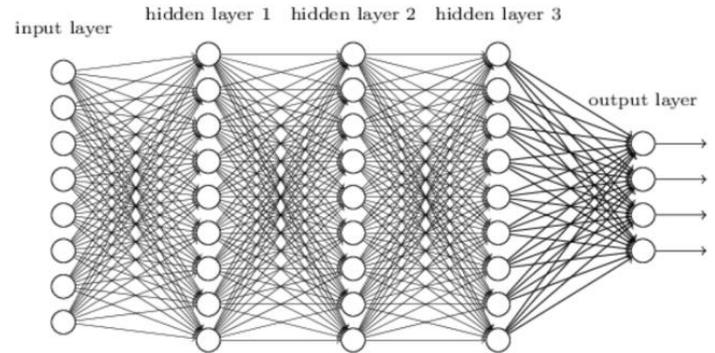
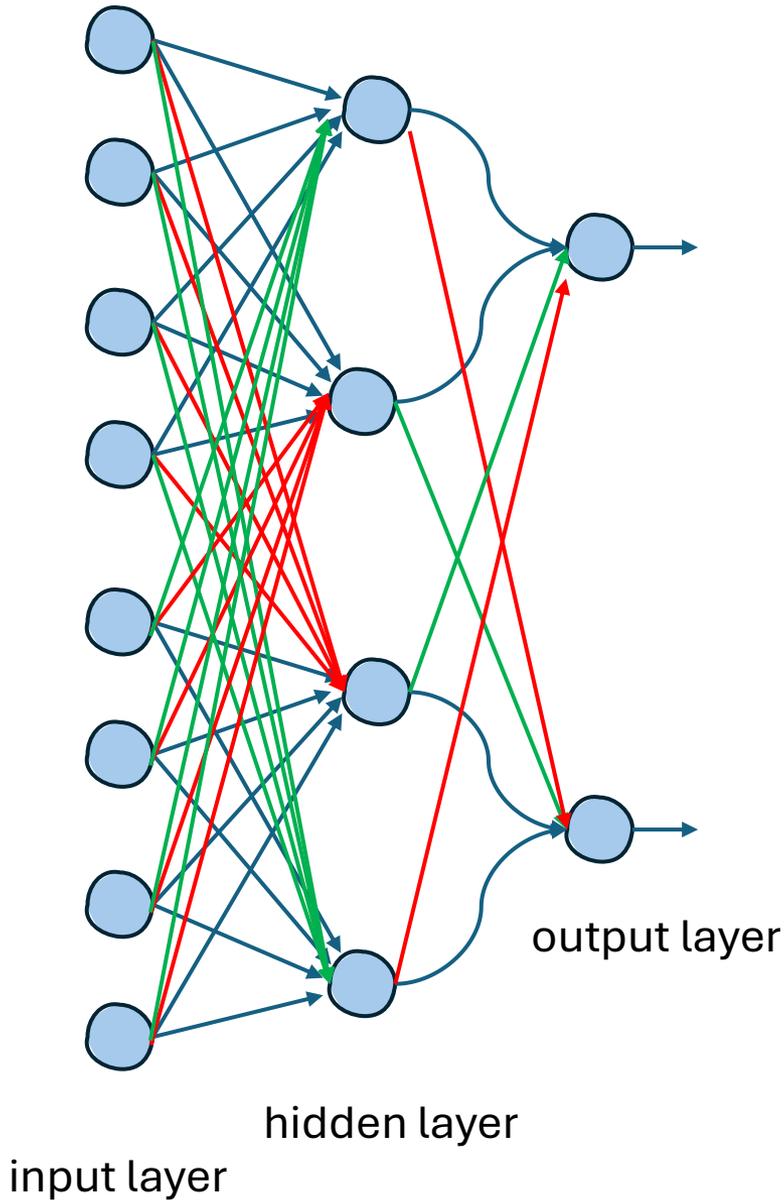




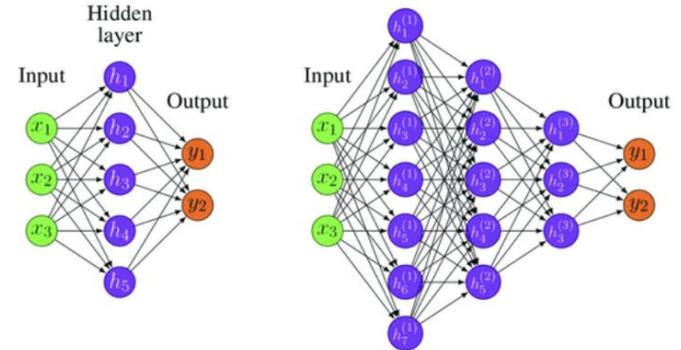


$32 + 8 = 40$ 個參數的 ANN

(Artificial Neuron Network)



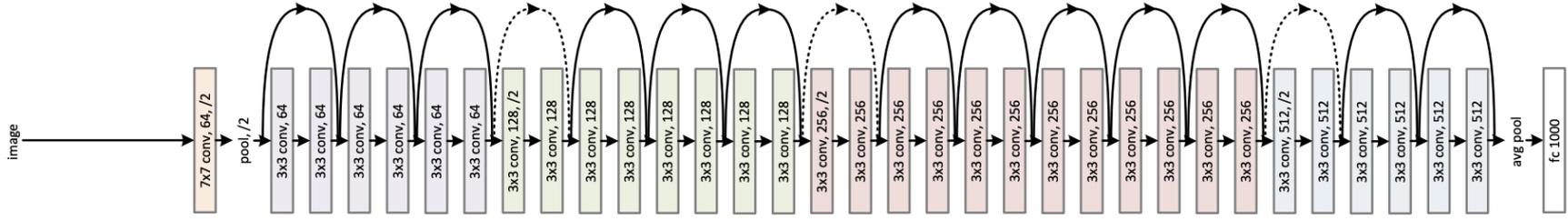
Network Architecture



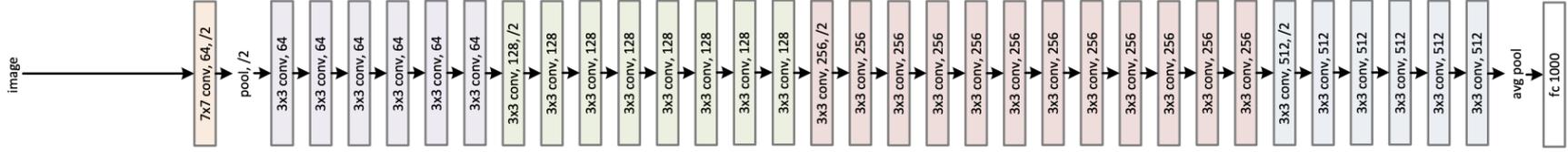
Frontiers in Artificial Intelligence 3:4
February 20203:4. DOI:10.3389/frai.2020.00004

ResNet-152

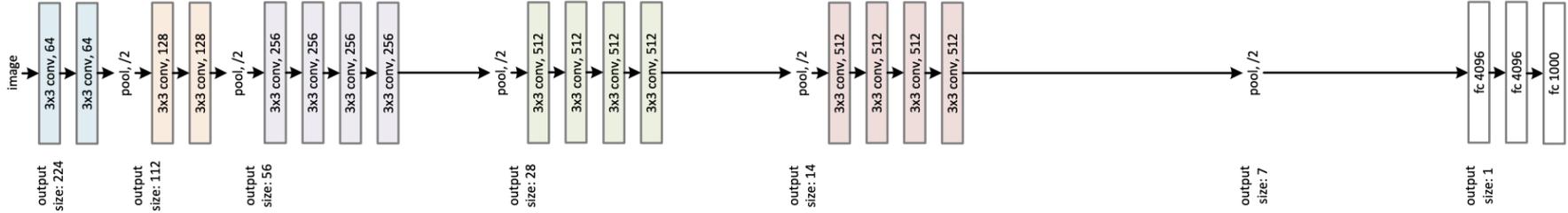
34-layer residual



34-layer plain



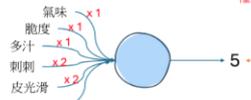
VGG-19



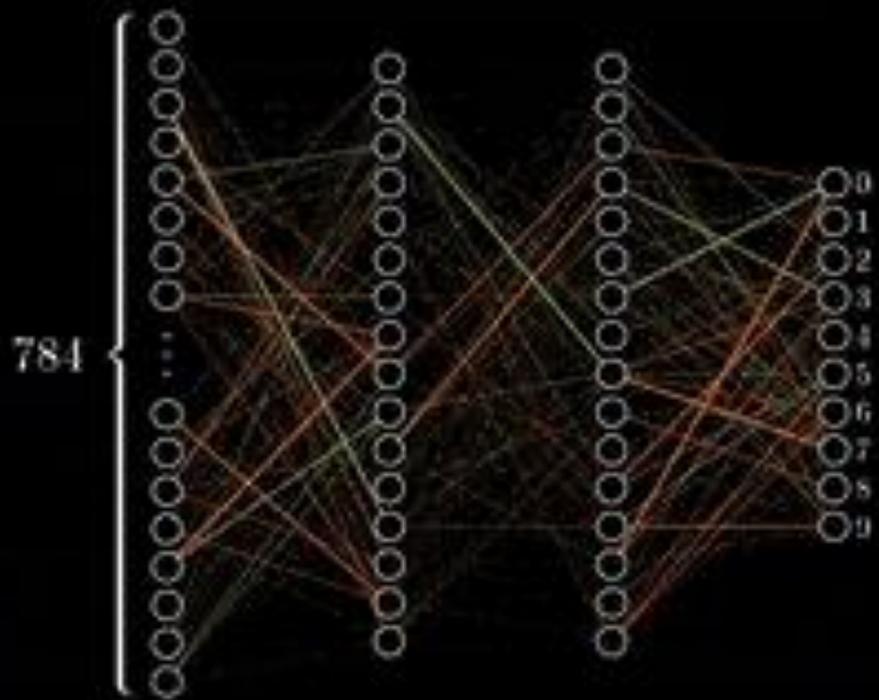
Training in progress...

	氣味	脆度	多汁	刺刺	皮光滑	沉甸甸	加起來
	1	1	0	0x2	1x3	0x3	5
	1	0	1	1x2	0x3	1x3	7
	0	0	1	0x2	0x3	0x3	1
	1	0	0	1x2	0x3	1x3	6

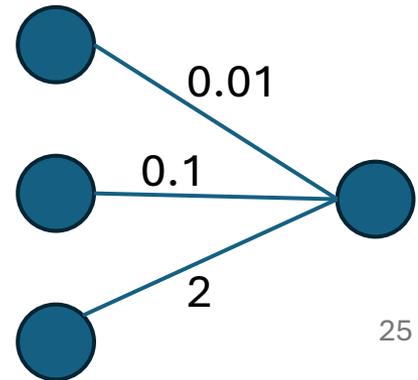
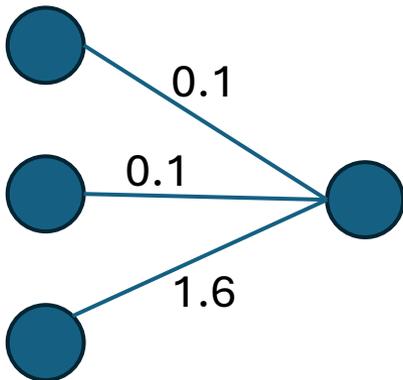
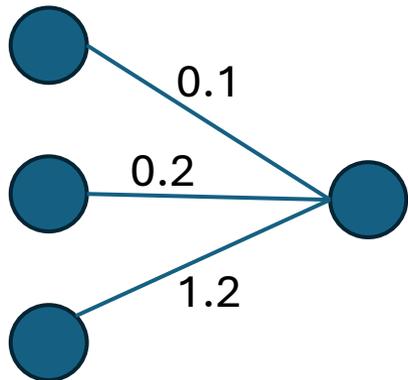
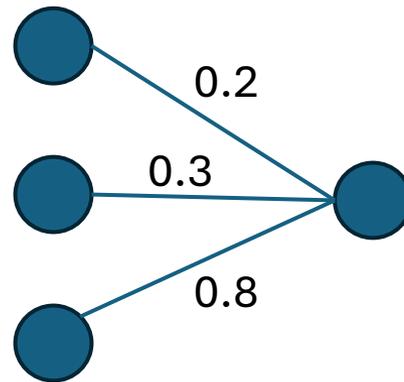
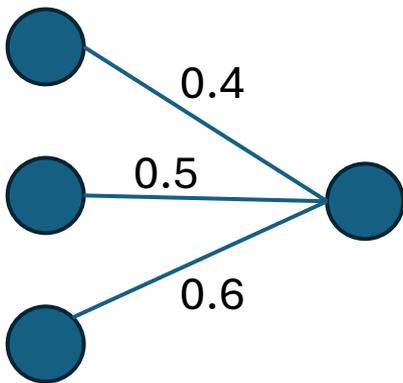
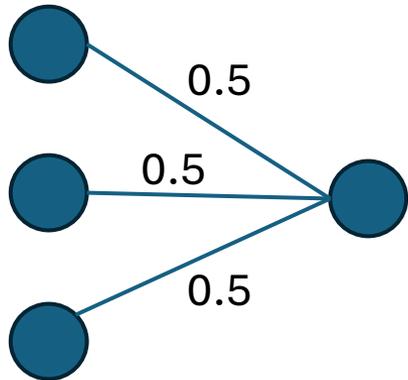
權重



Loss Function: 計算真實答案跟預測值的差異

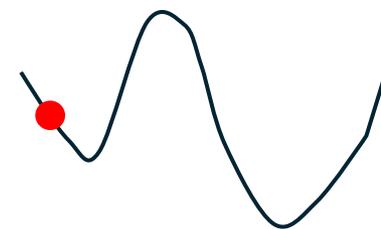
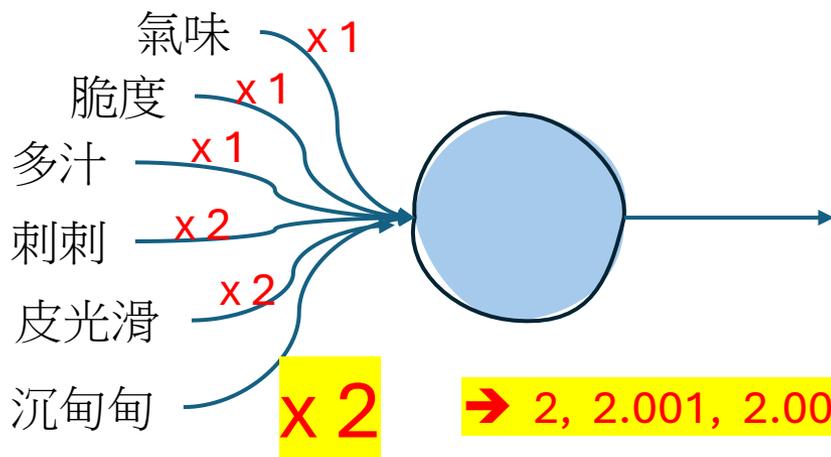


反向傳播算法 (Backpropagation Algorithm) 、 Loss Function 、 偏微分 → 調整權重





反向傳播算法 (Backpropagation Algorithm) 、 Loss function 、 偏微分 → 調整權重的技巧



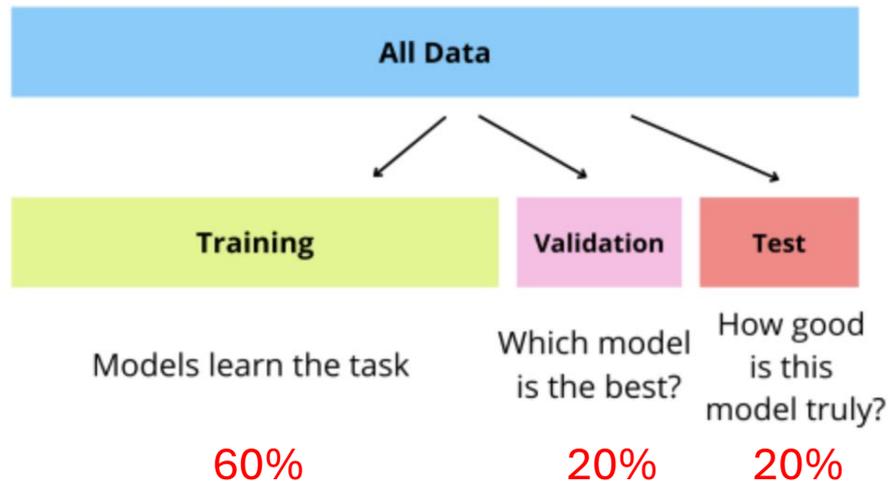
→ 2, 2.001, 2.002, 2.003..... 計算慢，容易以為最佳解

→ 2, 5, 10, 15,..... 計算快，但容易跳過最佳解

Learning Rate

把數據分拆成三群

- Training, Validation and Testing subsets



防止過擬合、評估泛化能力

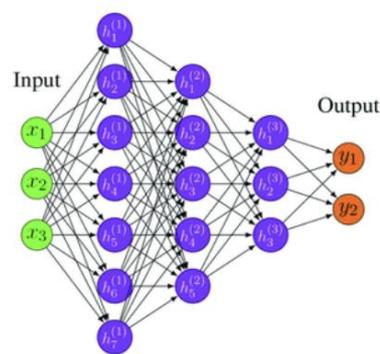
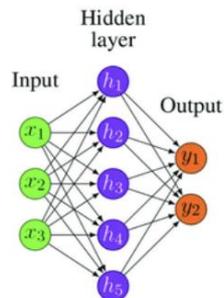
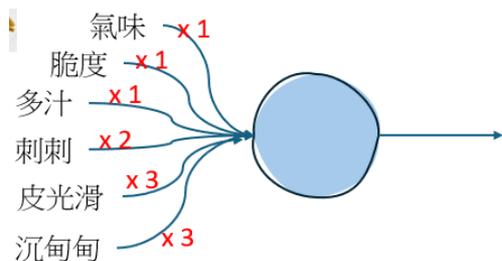
◆ 模型對未見過的新數據進行準確預測的能力

→ 模型在訓練數據上的表現非常好，但在測試數據或新數據上表現差

超參數 VS 參數

超參數 (Hyperparameters) 是指在訓練開始前需要設置的參數，這些參數不會隨著數據的學習而自動調整。

參數 (Parameters) 是在學習過程中自動生成/自動調整的內部值，不是由人工直接設定。



超參數

學習率 (Learning Rate) :

決定每次更新模型參數時步伐的大小。例如，學習率設為0.01表示每次更新時，權重的調整幅度為梯度值的1%。

迭代次數 (Epochs) :

指模型通過完整訓練數據集的次數。例如，設置迭代次數為10，表示訓練過程中模型會完整地掃過訓練數據集10次。

批次大小 (Batch Size) :

在每次迭代中，模型使用的訓練樣本數量。例如，批次大小設為32，表示每次更新模型時使用32個樣本來計算梯度和更新參數。

隱藏層的神經元數量 (Number of Units in Hidden Layers) :

每一個隱藏層的神經元數量。例如，一個兩層神經網絡的隱藏層單元數設置為64和32，表示第一層隱藏層有64個神經元，第二層有32個神經元。

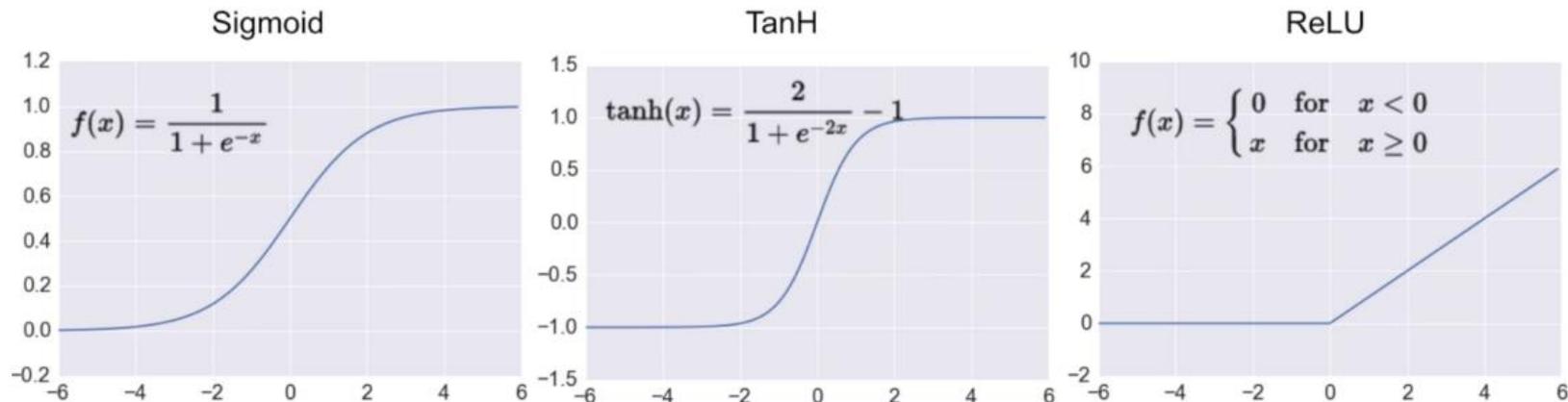
激活函數 (Activation Function) :

決定神經元輸出的形式。例如，選擇ReLU作為激活函數表示當輸入為正時，輸出為該正值，否則輸出為0。

優化算法 (Optimizer) :

決定如何更新模型的權重。例如，使用Adam優化算法可以自動調整學習率，通常能在較少的迭代次數內達到收斂。

激活函數 (Activation Function)



Sigmoid函數

取值範圍(0, 1)

雙曲正切函數 (TANH)

(hyperbolic tangent)

取值範圍(-1, 1)

線性整流函數 (ReLU)

小於0為0，輸出等於輸入

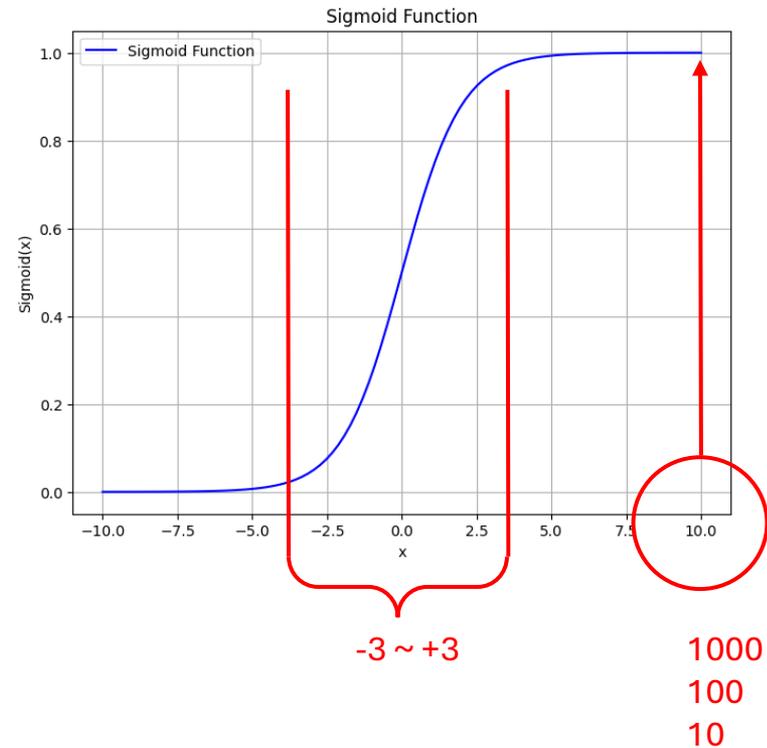
10 → sigmoid $1/(1+e^{10}) \rightarrow 0.00004539786$

10 → tanh $(2/(1+e^{-20}))-1 \rightarrow -1.12283715876$

10 → ReLU → 10

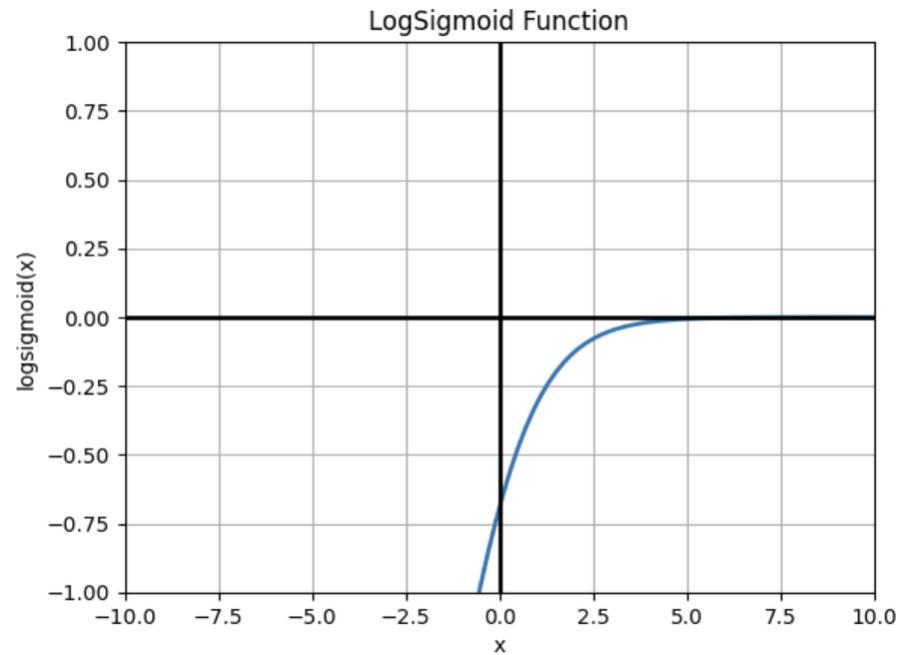
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

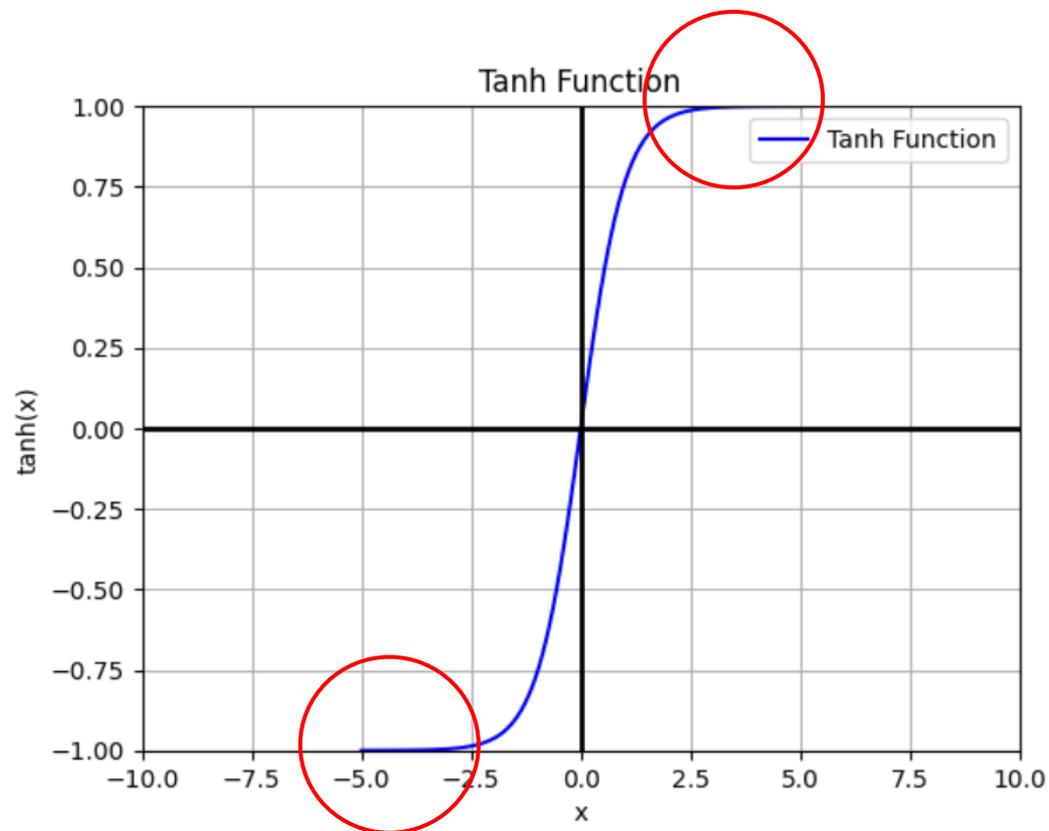


LogSigmoid

$$f(x) = \log\left(\frac{1}{1+e^{-x}}\right)$$



Tanh (hyperbolic tangent)



超參數調校

- 訓練機器學習或深度學習模型時，調整模型的超參數以提高模型性能的過程。
 - 網格搜索 (Grid Search)
 - 隨機搜索 (Random Search)
 - 遺傳演算法 (Genetic Algorithm)
 -

<https://playground.tensorflow.org/>

Regularization : 防止過擬合的技術。在模型中添加額外的約束或信息,減少模型的複雜度,從而提高模型的泛化能力。



Epoch: 000,000 | Learning rate: 0.03 | Activation: Tanh | Regularization: None | Regularization rate: 0 | Problem type: Classification

DATA

Which dataset do you want to use?



Ratio of training to test data: 50%

Noise: 0

Batch size: 10

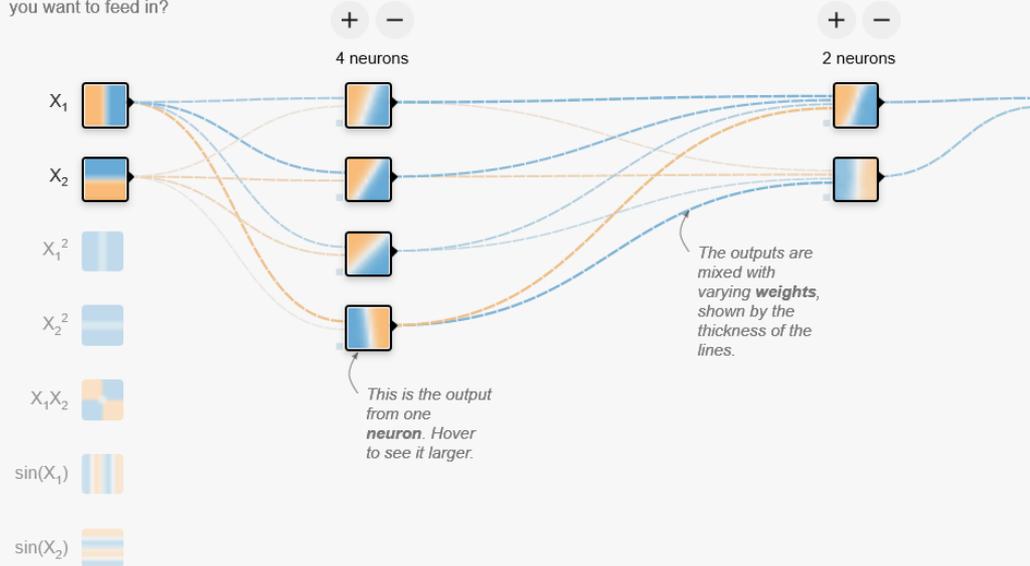
REGENERATE

FEATURES

Which properties do you want to feed in?

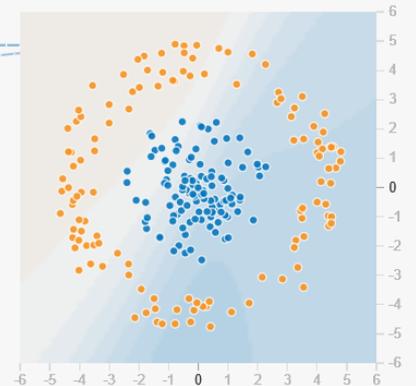
- X_1
- X_2
- X_1^2
- X_2^2
- $X_1 X_2$
- $\sin(X_1)$
- $\sin(X_2)$

2 HIDDEN LAYERS



OUTPUT

Test loss 0.508
Training loss 0.491



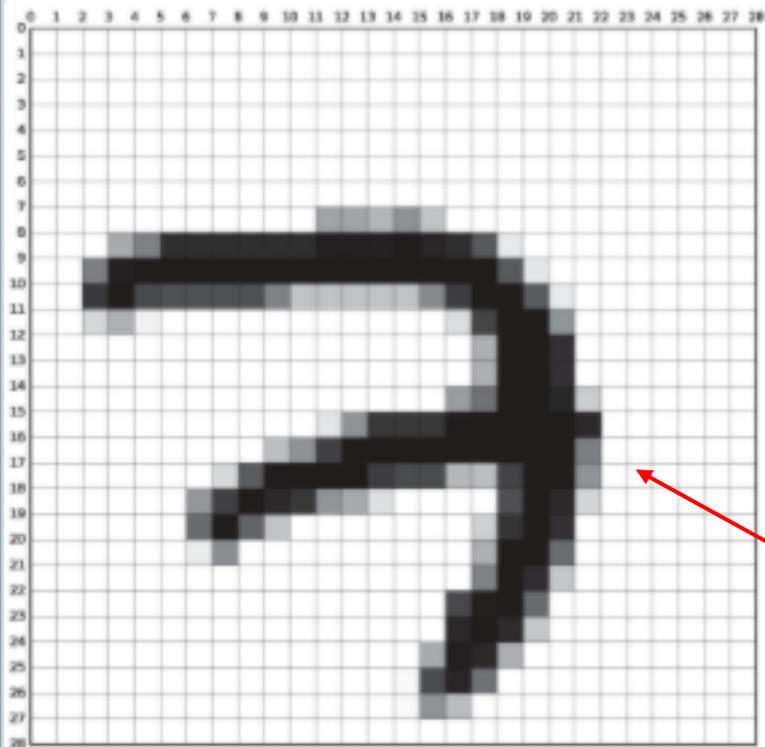
Colors shows data, neuron and weight values.

Show test data Discretize output

Part 2

卷積神經網路

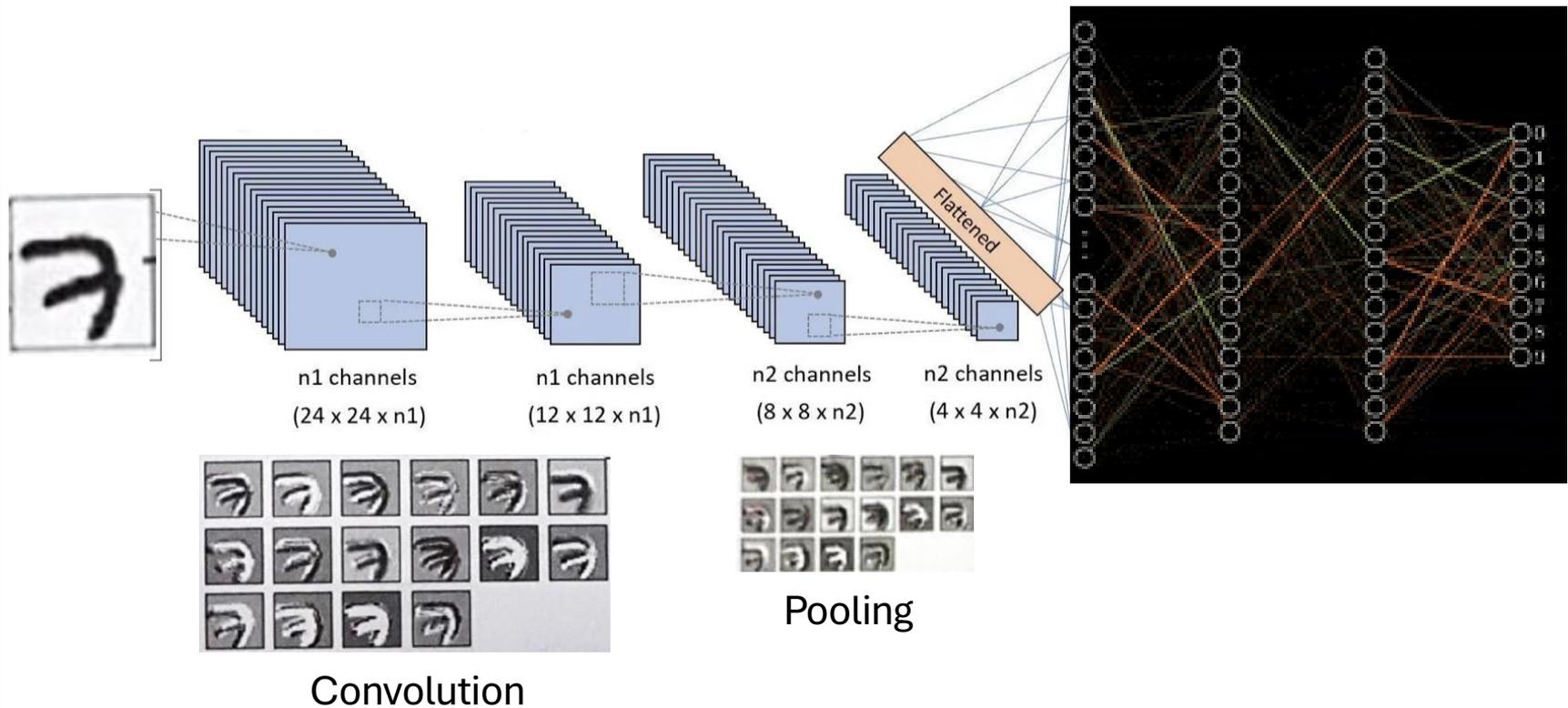
28格

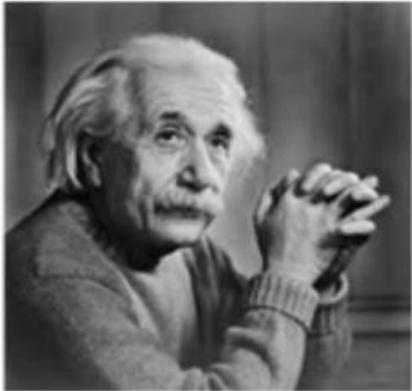


0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9

CNN (Convolutional neural network)

卷積神經網路





-1	0	1
-1	0	1
-1	0	1

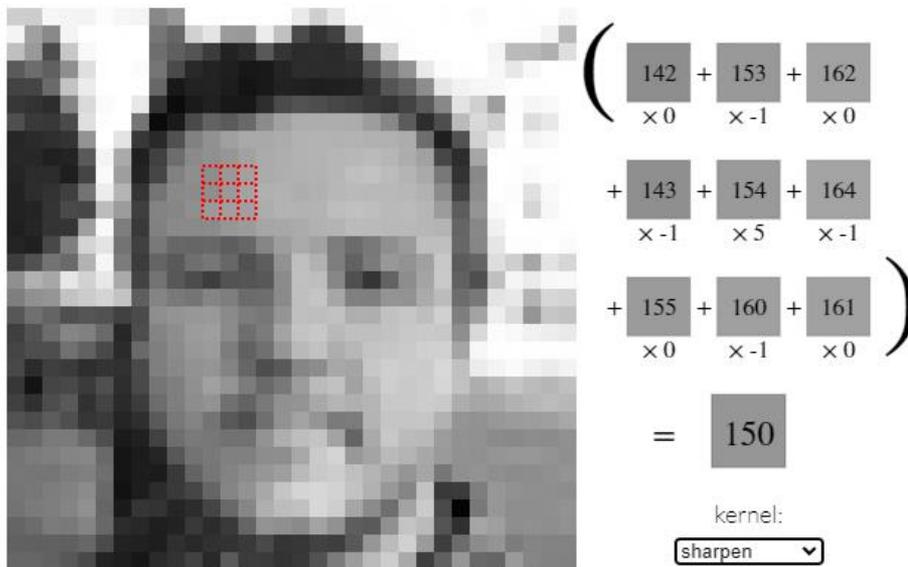


-1	-1	-1
0	0	0
1	1	1



Image Kernels

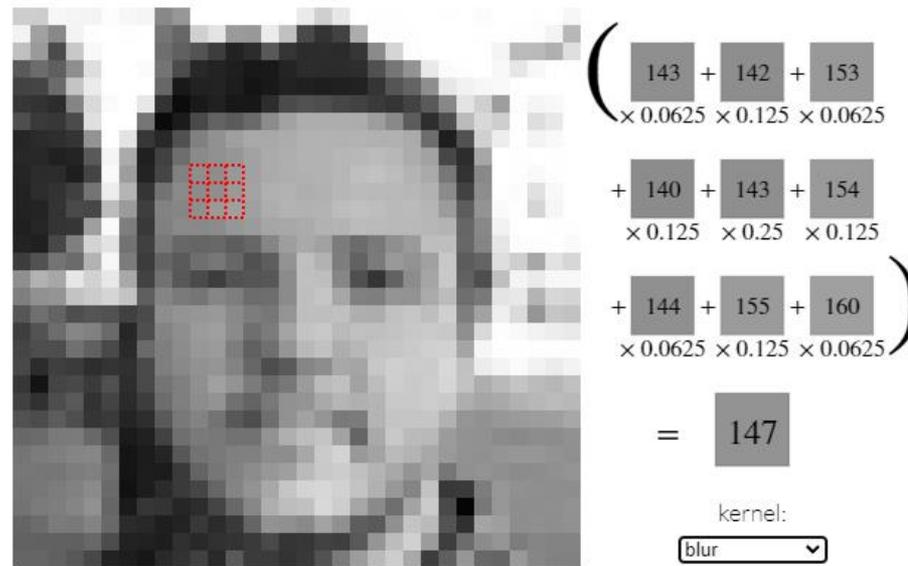
- **邊緣檢測卷積核 (Edge Detection Kernels)**
 - Sobel核: 檢測圖像中的水平或垂直邊緣。
 - Sobel X核會強調圖像中的水平邊緣，Sobel Y核則會強調垂直邊緣。
 - 這些邊緣檢測器常用在圖像邊緣提取和特徵檢測。
- **模糊卷積核 (Blurring Kernels)**
 - 平均模糊 (Box Blur) : 每個像素的值是周圍像素的平均值，去除圖像中的噪聲或平滑圖像。
 - 高斯模糊 (Gaussian Blur) : 減少噪聲和細節。
- **銳化卷積核 (Sharpening Kernels) :**
 - 強化圖像的細節、強調圖像中的特徵。
- **拉普拉斯卷積核 (Laplacian Kernels) :**
 - 強調圖像中的快速變化部分，特別是在檢測細節和邊緣



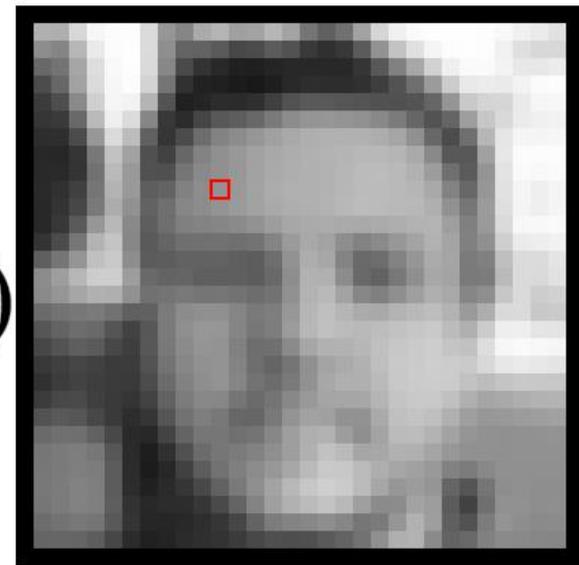
input image



output image

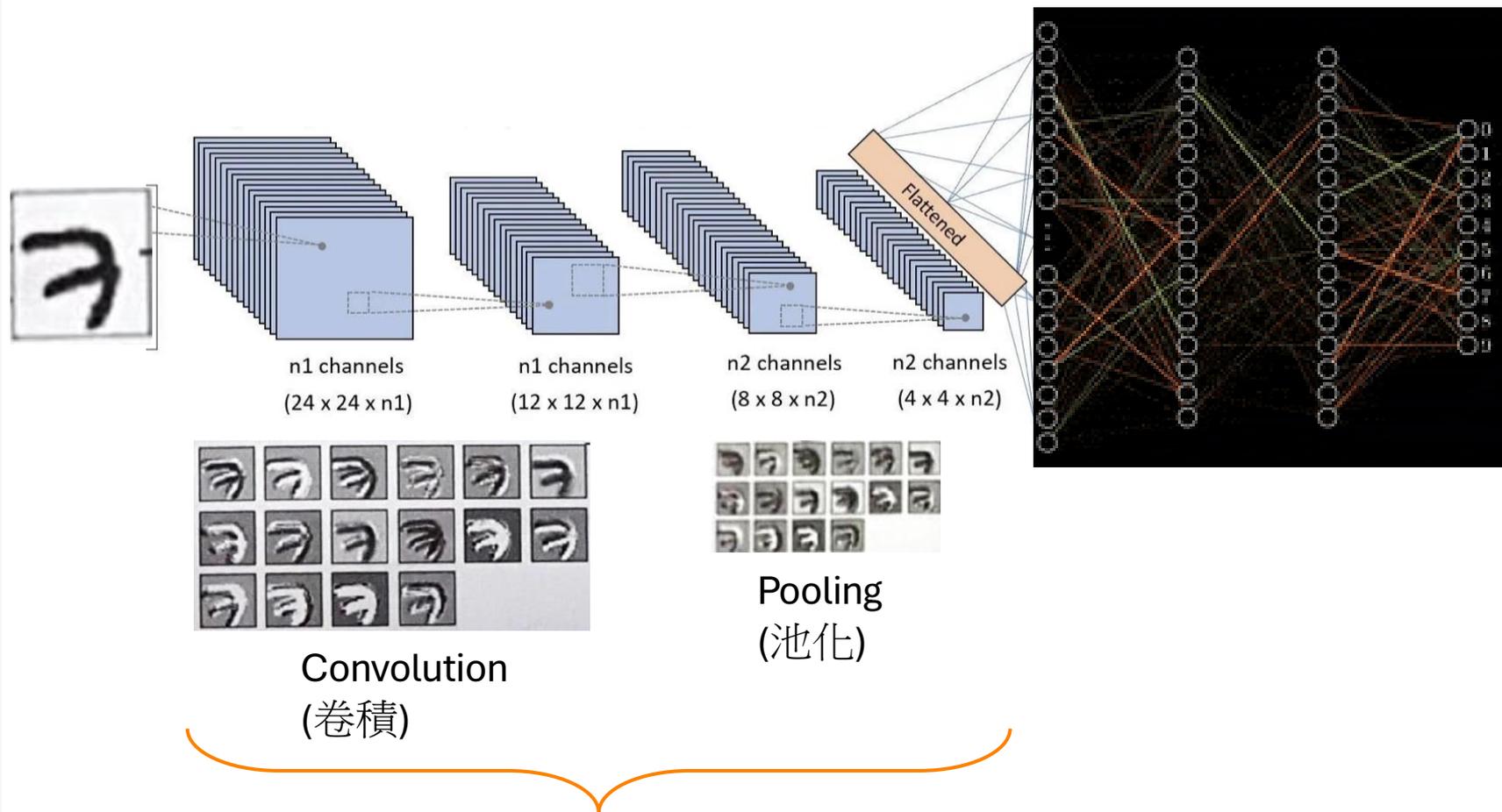


input image



output image

CNN (Convolutional neural network) 卷積神經網路



這邊也會作權重更新

流程



28x28

程式碼



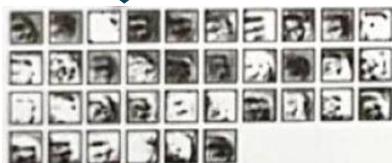
第一次卷積
(28x28 16張)

請幫我
把圖片套上16種濾鏡



縮減取樣
(14x14 16張)

請幫我
把圖片縮小，相鄰二個顏色取深的



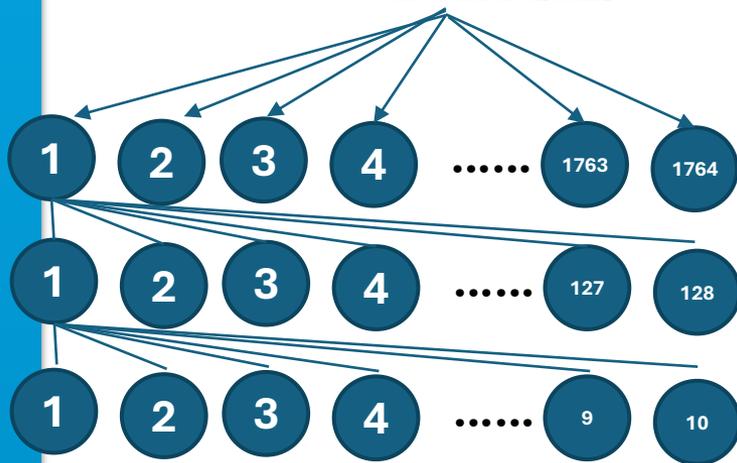
第二次卷積
(14x14 36張)

請幫我
把圖片套上36種濾鏡



縮減取樣
(7x7 36張)

請幫我
把圖片縮小，相鄰二個顏色取深的



平坦層 (reshape成一維)
(7x7x36 = 1764個input)

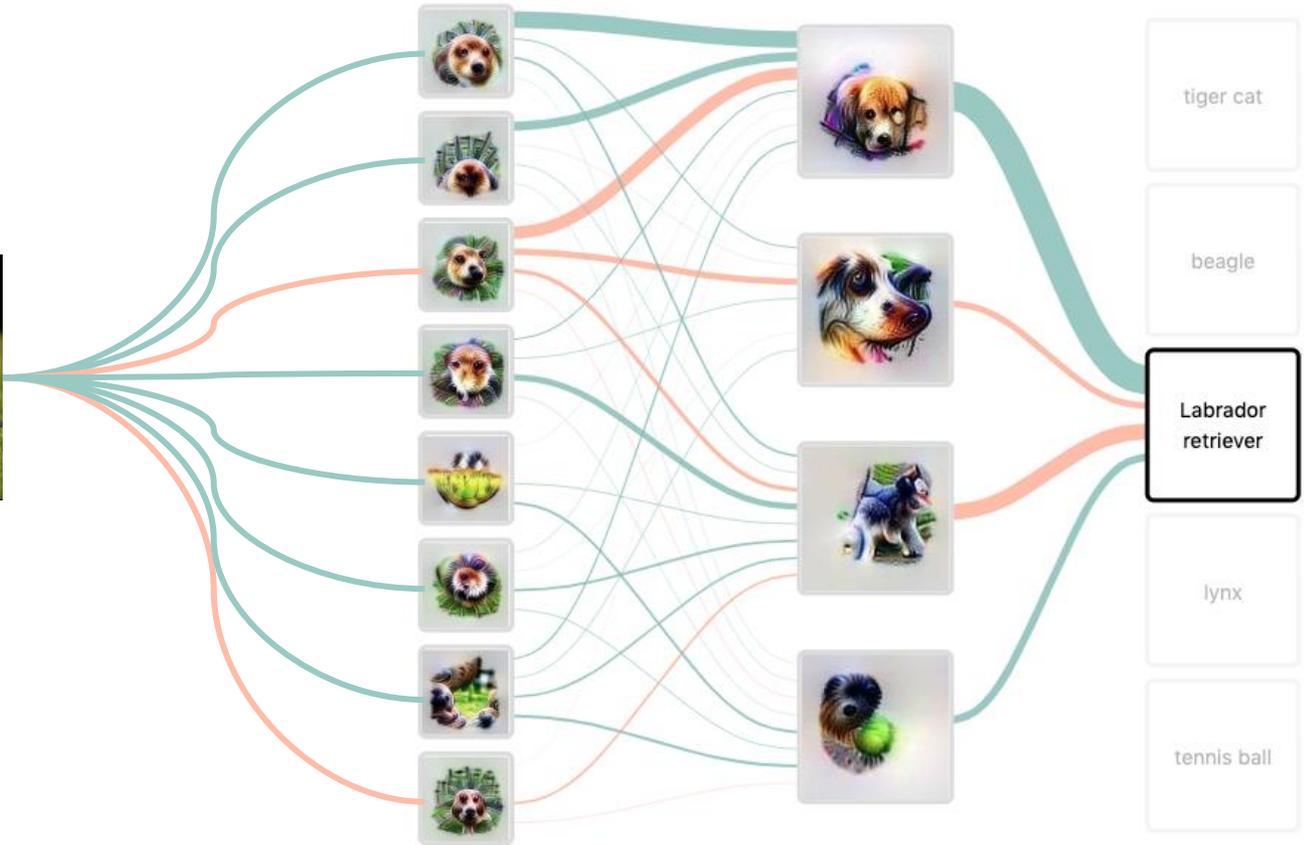
請幫我
把像素攤開

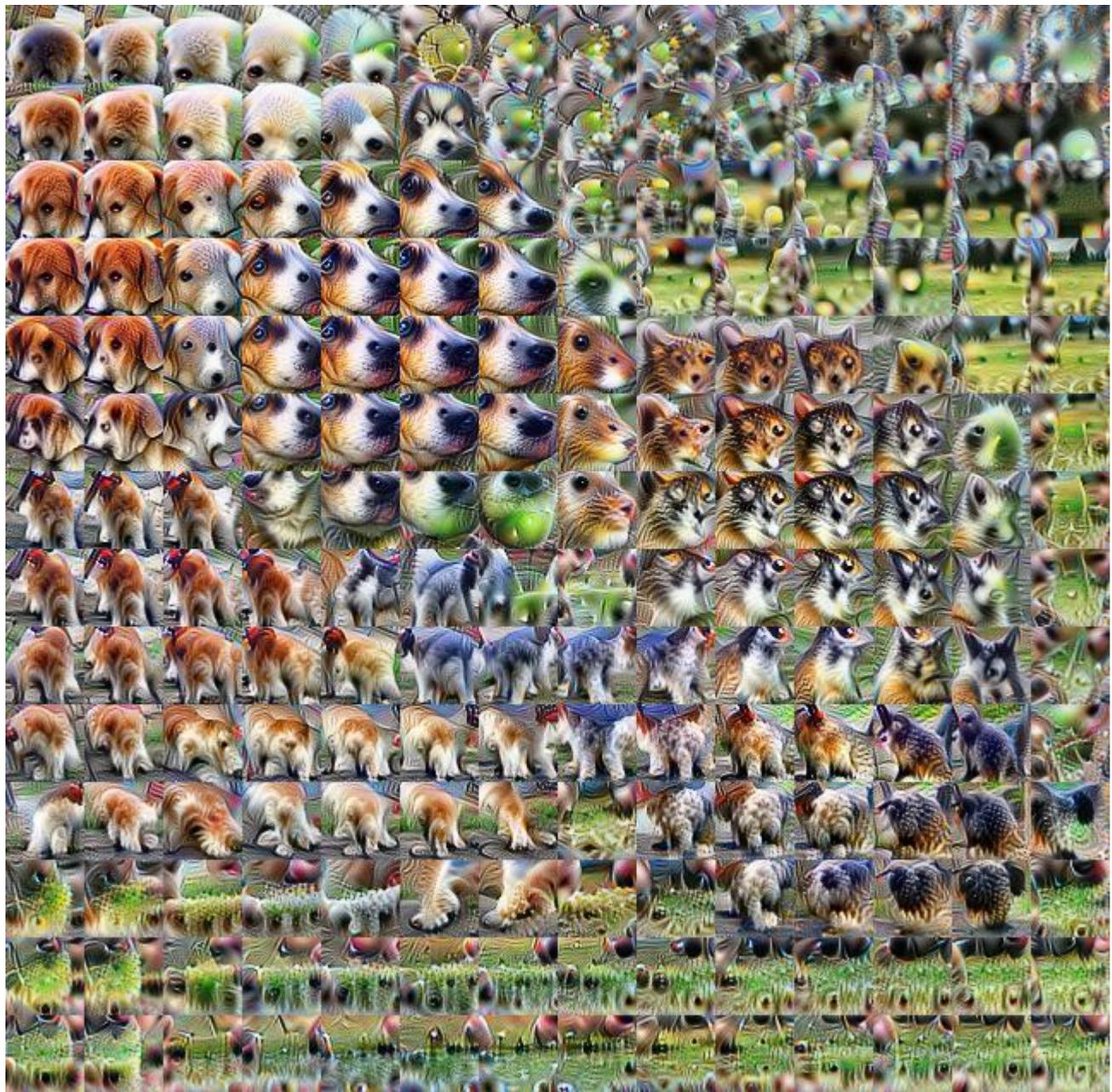
隱藏層128個

丟進去128個神經元計算

輸出層10個

輸出分群10群





圖片來源 (經修改) : https://distill.pub/2018/building_blocks/
作者 : Chris Olah 等人

Part 3

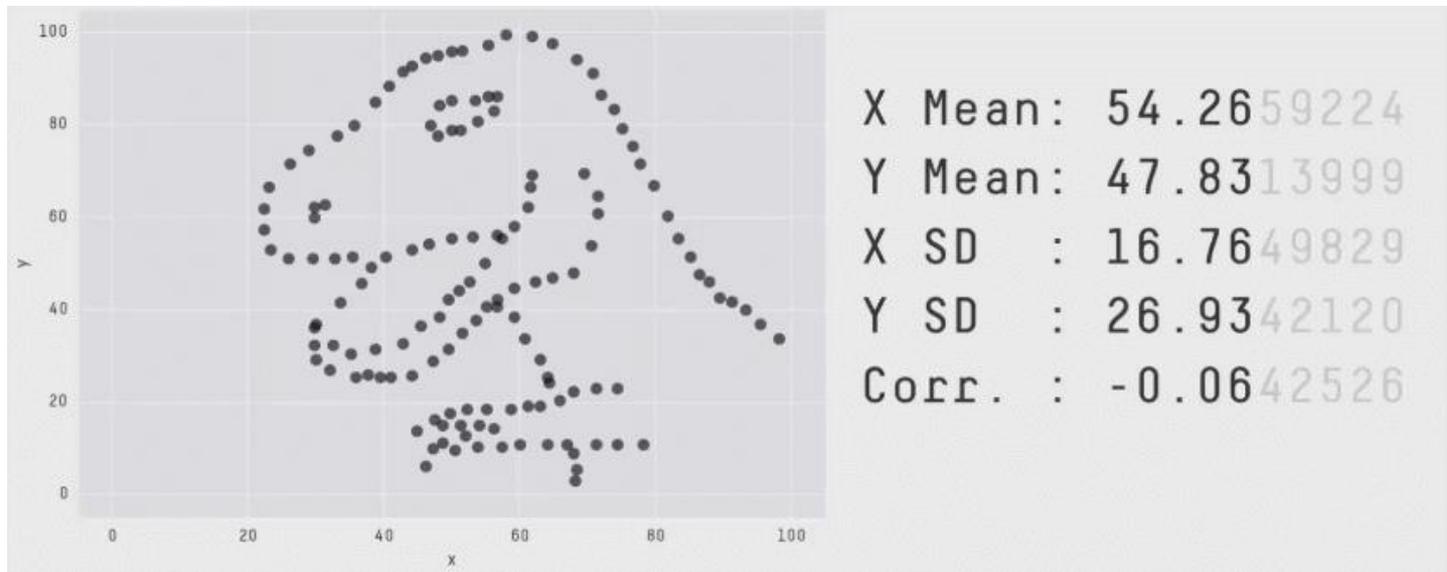
Exploration Data Analysis (EDA)

探索式資料分析

Exploration Data Analysis (EDA)

探索式資料分析

是一種初步分析數據的過程，用來理解數據的基本結構、發現模式、檢測異常、測試假設，並通過可視化的繪圖、視覺化等快速簡易的方式，從各種面向先了解資料的狀況，以利後續分析。



圖片來源：

<https://baubimedi.medium.com/速記ai課程-統計與資料分析-四-3cf14683b98f>

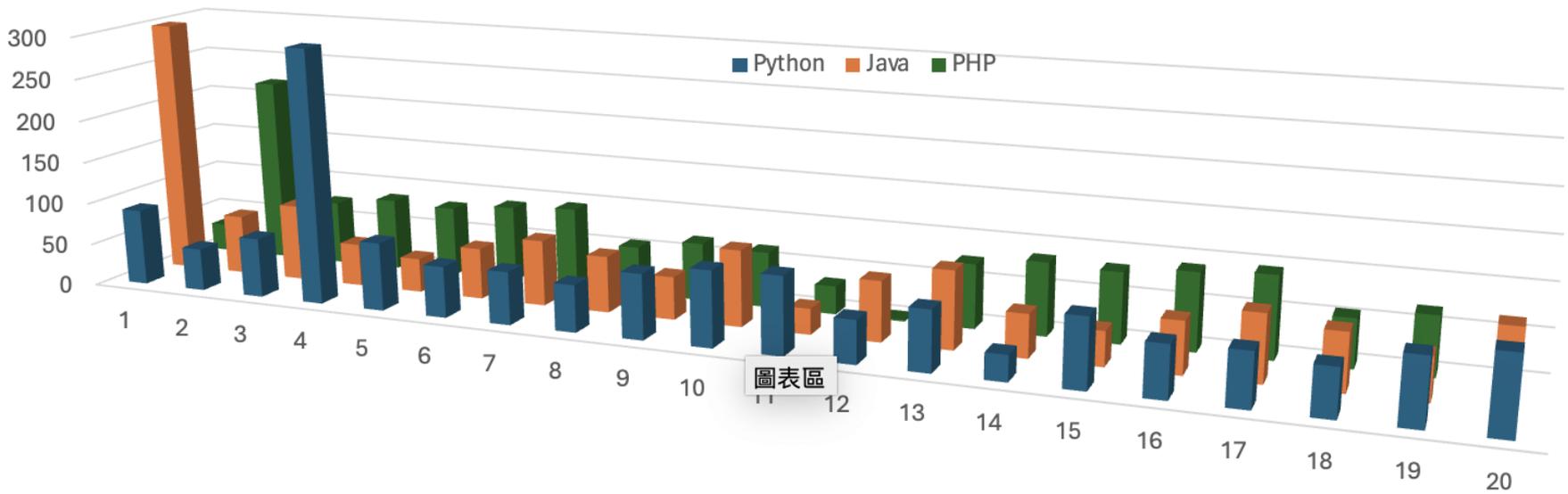
- **數據量 (Data volume):**
 - 影響模型的訓練效果和計算資源的需求。
- **目標特徵 (Target features):**
 - 目標特徵是指模型需要預測的結果變量，通常在監督學習中使用。這個變量可以是連續的數值（例如房價）或分類標籤（例如分類問題中的“是/否”）。舉例來說，在房價預測模型中，房價就是目標特徵；在垃圾郵件分類中，“垃圾郵件”或“非垃圾郵件”是目標特徵。
- **雜訊數據/ 異常值 (Noisy data/Outliers):**
 - 因觀測、紀錄錯誤而產生，例如一個人的年齡被記錄為**300歲**，這可能是錯誤的。
 - 異常值可能是正確的數據，但它們的數值明顯**偏離平均值**。
 - 對於常態分佈的數據集，異常值可能是距離**平均值3到6個標準差的數值**。
 - 當這些數值超過數據集的5%時，我們需要處理這些異常值。
- **缺失值 (Missing values):**
 - 數據集中某些樣本在某些特徵上的值缺失的情況。缺失值可能會影響模型的準確性，特別是在大量數據缺失時。
 - 處理缺失值的方法包括**刪除**含有缺失值的樣本、用**平均值**或**中位數**填補缺失值，或使用其他插補技術。
 - 在醫療數據集中，一些患者的血壓記錄可能缺失，這時需要對這些缺失值進行處理以確保模型的準確性。
- **定性特徵 (Qualitative features):**
 - 以文字、圖形、音頻或其他非數字格式表示的非數字數據。我們需要檢查數據集中是否包含定性特徵。如果存在定性特徵，我們需要使用數據編碼技術來處理它們。

Outliers processing

- The term "outliers" refers to data points in a sample that significantly deviate from the rest of the data points; outliers are also known as "anomalies."
- Having too many outliers can introduce bias to deep learning models.
- If necessary, we need to identify and analyze these outliers for processing.
- However, not all outliers require processing, as some outliers may represent meaningful values in practical applications.

```
import numpy as np
import pandas as pd
```

```
scores={
'Python':[90,50,70,300,80,60,62, 55, 76, 88, 90, 50, 70, 30, 80, 60, 62, 55, 76, 88],
'Java':[300, 70, 90, 50, 40, 60, 77, 66, 50, 89, 30, 70, 90, 50, 40, 60, 77, 66, 50, 89],
'PHP':[33, 220, 75, 85, 82, 90, 95, 56, 68, 65, 33, 2, 75, 85, 82, 90, 95, 56, 68, 65]
}
df=pd.DataFrame (scores)
print (df.shape)
```



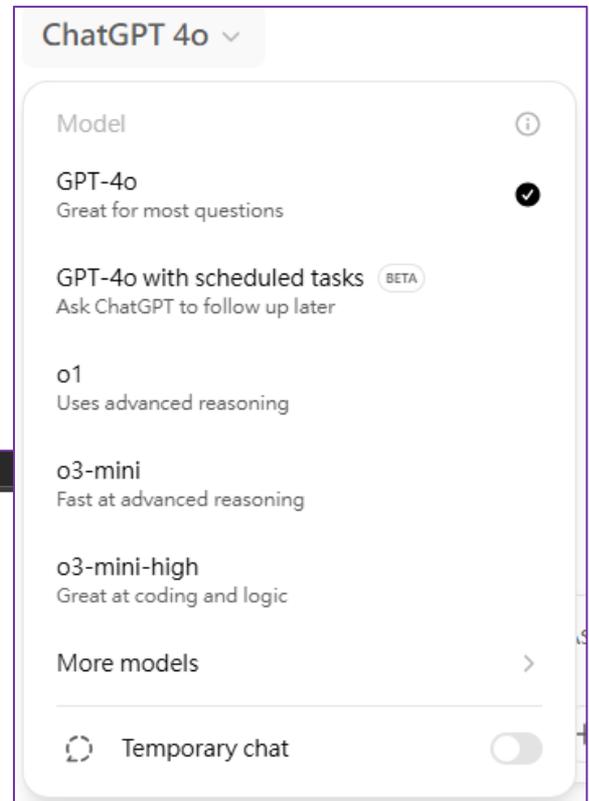
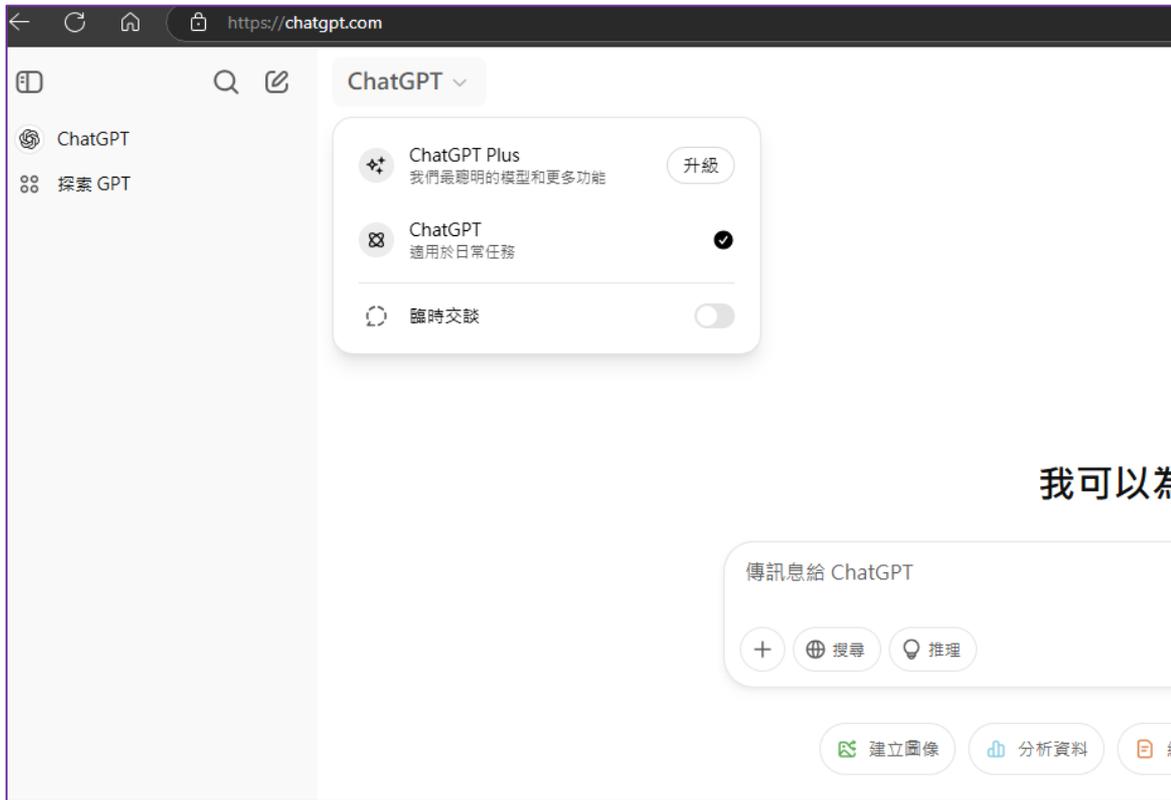
把 LLM 應用在 Exploration Data Analysis

ChatGPT <https://chatgpt.com/>

Claude.ai <https://claude.ai/>

Perplexity.ai <https://www.perplexity.ai/>

註冊帳號



Prompts (提示詞)

- 大型語言模型的 Prompts 指的是用來引導 AI 生成回應的輸入文本。透過不同的 Prompts，可以影響 AI 的輸出內容，確保回答符合期望的格式、風格或資訊需求。
- 大型語言模型可以幫助進行數據處理、視覺化、模型建立、結果解釋等。

- 指定明確的數據分析目標：
 - 我們擁有一組mRNA gene expression 數據，請幫我找出表現量異常的基因，並解釋為什麼。
- 要求 AI 逐步說明數據分析步驟：
 - 請幫我設計一個分析流程來分析mRNA gene expression 數據，包含數據清理、特徵工程、模型選擇與評估方法。
- 要求 AI 生成程式碼：
 - 請用 Excel VBA 撰寫一段程式碼，分析以找出潛在數據錯誤為目標這個mRNA gene expression 數據，並畫出對應的分佈圖。
- 要求 AI 提供執行方法指引：
 - 請逐步說明在 windows 11 的 Excel 上執行上述程式碼的步驟。

- 指定數據類型與格式：
 - 我有一個包含『基因Accession Id、RPKM、Fold Change、藥物作用後的IC50』的 CSV 檔，以逗號隔開每個欄位，請幫我寫 Python 程式來計算Fold Change與IC50的相關性。
- 請 AI 建議適合的統計檢定方法：
 - 比較兩組Gene set，請問應該用什麼統計檢定來判斷是否有顯著差異？
- AI 幫助報告撰寫：
 - 請幫我用 Markdown 格式整理數據分析報告，包括問題描述、數據來源、方法、結果與結論。

範例數據



DemoData.csv

	年齡				性別	身高
	A	B	C	D	E	F
1	Age	Feature1	Feature2	Feature3	Gender	Height
2	53	29	117	232	M	130
3	58	26	83	229	F	149
4	56	29	91	159	M	141
5	48	32	80	187	M	184
6	7	30	82	17500	F	164
7	57	26	93	209	M	176
8	46	32	120	154	M	137
9	52	31	114	19	F	217
10	47	35	102	153	M	146
11	55	28	97	219	F	208
12	51	26	107	221	F	198
13	59	30	111		F	207
14	43	33	93	210	F	175
15	53	26	112	202	M	187
16	60	33	90	235	F	170
17	53	29	83	229	M	186
18	41	28	95	213	F	201
19	60	3	97	199	M	161
20	45	32	107	225	F	157
21	45	20	02	186	M	124

Prompt: 找出資料表格內的缺值

好的，我會分析這個CSV檔案並找出其中的缺值。我會逐列檢查每個欄位，並報告任何缺失的數據。

以下是我發現的缺值:

1. Age 欄位:

- 第43行: 空值

2. Feature1 欄位:

- 第36行: 空值
- 第44行: 空值

3. Feature2 欄位:

- 沒有發現缺值

4. Feature3 欄位:

- 第11行: 空值
- 第43行: 空值
- 第83行: 空值

5. Gender 欄位:

- 第52行: 空值
- 第53行: 空值



DemoData.csv
試算表

找出資料表格內的缺值

我會先讀取你的檔案並檢查表格內是否有缺失值 (NaN) 。讓我來處理。

你的表格內有缺失值的欄位如下：

- Age : 1 個缺值
- Feature1 : 2 個缺值
- Feature2 : 2 個缺值
- Feature3 : 3 個缺值
- Gender : 3 個缺值
- Height : 4 個缺值



你想要如何處理這些缺失值呢？填補、刪除，還是進一步分析？ [-]

分析



python

```
import pandas as pd
```

永遠顯示詳細資料 複製

```
# 讀取上傳的 CSV 檔案
```

```
file_path = "/mnt/data/DemoData.csv"
```

```
df = pd.read_csv(file_path)
```

```
# 找出缺失值
```

```
missing_values = df.isnull().sum()
```

```
missing_values[missing_values > 0]
```

結果

```
Age      1
```

```
Feature1 2
```

```
Feature2 2
```

```
Feature3 3
```

```
Gender   3
```

```
Height   4
```

```
dtype: int64
```

Q: ChatGPT 或此類的AI
會不會把我的數據偷走
另做他用？

讓ChatGPT代寫程式，
在自己點電腦執行

Prompt: 寫一個VBA 程式，把缺值的儲存格標上黃色底色

Excel 2016 介面顯示，檔案名為 DemoData.csv，已儲存到此電腦。搜尋欄位於頂部右側。

功能區包含：檔案、常用、插入、繪圖、頁面配置、公式、資料、校閱、檢視、開發人員、說明。

常用功能區包含：剪貼簿、字型、對齊方式、數值、樣式。

通知欄顯示：可能發生資料遺失。如果將活頁簿儲存為逗號分隔 (.csv) 格式，會遺失部分功能。若要保留這些功能，請將活頁簿儲存為 Excel 檔案格式。 不再顯示 另存新檔...

工作表顯示：

	A	B	C	D	E	F	G	H	I	J	K	M
1	Age	Feature1	Feature2	Feature3	Gender	Height						
2	53	29	117	232	M	130						
3	58	26	83	229	F	149						
4	56	29	91	159	M	141						
5	48	32	80	187	M	184						

圖中有一紅色箭頭指向「另存新檔...」按鈕，以及一個黃色方塊標註「先另存成 .xlsm 格式」。

這是一個 VBA 程式，會尋找 Excel 表格中的缺失值（空白或 #N/A），並將其儲存格標記為黃色底色。

使用方式：

1. 打開 Excel 並按 **Alt + F11** 開啟 VBA 編輯器。
2. 插入模組：點擊 "插入" > "模組"。
3. 貼上以下 VBA 程式碼 並執行 `HighlightMissingValues`。

vba

複製 編輯

```
Sub HighlightMissingValues()
    Dim ws As Worksheet
```

Microsoft Visual Basic for Applications - DemoData.20250221.xlsm

檔案(F) 編輯(E) 檢視(V) 插入(I) 格式(O) 偵錯(D) 執行(R) 工具(T) 增益集(A) 視窗(W) 說明(H)

程序(P)...
自訂表單(U)
模組(M)
物件類別模組(C)
檔案(L)...

VB
ThisWorkbook
工作表1 (DemoDat)

性 - 工作表1

工作表1 Worksheet

字母順序 性質分類

(Name)	工作表1
DisplayPageBreak	False
DisplayRightToLeft	False
EnableAutoFilter	False
EnableCalculation	True
EnableFormatCon	True
EnableOutlining	False
EnablePivotTable	False

DemoData.20250221.xlsm - Module1 (程式碼)

(一般)

```
Sub HighlightMissingValues()  
    Dim ws As Worksheet  
    Dim rng As Range  
    Dim cell As Range  
  
    ' 設定要檢查的工作表  
    Set ws = ActiveSheet  
    ' 設定要檢查的範圍 (可調整為你實際的資料範圍)  
    Set rng = ws.UsedRange  
  
    ' 遍歷範圍內的所有儲存格  
    For Each cell In rng  
        If IsEmpty(cell.Value) Or IsError(cell.Value) Then  
            ' 標記缺值為黃色底色  
            cell.Interior.Color = RGB(255, 255, 0)  
        End If  
    Next cell  
  
    ' 清理記憶體  
    Set ws = Nothing  
    Set rng = Nothing  
End Sub
```

這段 VBA 會：

- 檢查目前 Excel 工作表的所有使用範圍 (UsedRange)。
- 如果發現 空白儲存格或錯誤值 (如 #N/A)，則將其背景設為黃色 (RGB(255, 255, 0))。
- 你也可以手動指定檢查範圍，例如 Set rng = ws.Range("A1:D100")。

試試看吧！有需要修改的話，可以告訴我你的 Excel 資料範圍和結構。 😊

如何執行這個VBA程式？

步驟 3：執行 VBA 程式

1. 回到 Excel (但 VBA 視窗保持開啟) 。
2. 按下 `Alt + F8` ，會跳出「巨集」視窗。
3. 找到 `HighlightMissingValues` ，然後點擊 "執行" (Run) 。
4. Excel 會自動標記所有缺值 (空白或錯誤值) 為黃色底色。

額外提示

- 如果 VBA 巨集無法執行，請檢查：
 1. Excel 巨集安全性設定：
 - 到 `檔案` > `選項` > `信任中心` > `信任中心設定` > `巨集設定` ，選擇「啟用所有巨集」。
 2. 儲存檔案時使用 `.xlsm` 格式 (啟用巨集的 Excel 檔案) 。

篩選出有缺失值的欄位
 → 使用顏色篩選

	A	B	C	D	E	F
	Age	Feature1	Feature2	Feature3	Gender	Height
2	53	29	117	232	M	130
3	58	26	83	229	F	149

	A	B	C	D	E	F
1	Age	Feature1	Feature2	Feature3	Gender	Height
2	53	29	117	232	M	130
3	58	26	83	229	F	149
4	56	29	91	159	M	141
4	43	33	93	210	F	175
5	53	26	112	202	M	187
6	60	33	90	235	F	170
7	53	29	83	229	M	186
8	41	28	95	213	F	201
9	60	3	97	199	M	161
0	45	32	107	225	F	157
1	45	30	92	186	M	134
2	51	27	82	242	F	
3	47	28	109	188	M	210
4	43	34	83	192	F	164
5	50	33	105	188	M	147
6	47	25	83	176	M	131
7	57	29	108	202	F	167
8	59	35	80	173	F	189

顯示所有缺值(黃色)的資料列？

	A	B	C	D	E	F
1	Age	Feature1	Feature2	Feature3	Gender	Height
13	59	30	111		F	207
22	51	27	82	242	F	
36	59		116	203	M	209
44		30	105	222	M	199
47	49	30	119	243	M	
49	48	29	114		F	213
50	44		100	248	F	143
53	42	28	89	151		
55	56	30	99	151		141
85	53	35		233		170
92	59	29	87		F	168
97	300	34		217	N	157
98	47	25	100	239	M	

只顯示有黃色底色(缺失值)的資料列

上機練習

Prompt: 找出資料表格中的年齡 (Age)、身高(Height) 異常值

好的，我會分析年齡(Age)和身高(Height)欄位的數據，找出可能的異常值。我們通常將明顯偏離正常範圍的值視為異常值。讓我們逐一檢查這兩個欄位：

1. 年齡(Age)異常值：

人類的正常年齡範圍通常在0-120歲之間。在這個數據集中，我們發現了以下異常值：

- 第81行：160歲
- 第98行：300歲

這些年齡明顯超出了正常人類壽命範圍，很可能是數據輸入錯誤。

另外，還有一個值得注意的數據：

- 第5行：7歲（雖然不是嚴格意義上的異常值，但在這個數據集中顯得特別年輕，可能需要進一步驗證）

2. 身高(Height)異常值：

假設身高單位是厘米(cm)，正常人類身高範圍大約在50-250cm之間。在這個範圍內，我們沒有發現明顯的異常值。

然而，有些數值看起來偏低或偏高，可能需要進一步確認：

偏低的值：

- 第1行：130cm
- 第6行：130cm
- 第53行：130cm

偏高的值：

- 第4行：184cm

這些身高值雖然在可能的人類身高範圍內，但都處於極端情況，可能需要進一步驗證其準確性。

總結：

1. 年齡(Age)欄位有2個明顯的異常值：160歲和300歲。
2. 身高(Height)欄位沒有明顯超出人類可能範圍的異常值，但有一些極端值可能需要進一步確認。

建議：

1. 對於年齡的異常值，建議檢查原始數據源或與數據提供者確認，可能是輸入錯誤。
2. 對於身高的極端值，雖然在可能範圍內，但也建議進行驗證，確保數據的準確性。
3. 在進行後續分析時，可能需要考慮如何處理這些異常值或極端值，例如排除、替換或使用統計方法處理。

您需要我對這些異常值進行更深入的分析嗎？

Prompt: 請針對異常值進行更深入的 的分析 (1)

當然，我會對年齡(Age)和身高(Height)的異常值進行更深入的分析。我們將使用一些統計方法來識別可能的異常值，並探討這些異常值對整體數據的影響。

1. 年齡(Age)分析：

首先，讓我們計算一些基本統計數據：

平均值：52.7歲

中位數：52歲

標準差：24.9歲

第一四分位數 (Q1)：46歲

第三四分位數 (Q3)：56歲

四分位距 (IQR)：10歲

使用 $1.5 * IQR$ 法則來識別異常值：

下界： $Q1 - 1.5 * IQR = 31$ 歲

上界： $Q3 + 1.5 * IQR = 71$ 歲

基於這個標準，我們可以識別出以下異常值：

- 7歲 (第5行)
- 160歲 (第81行)
- 300歲 (第98行)

這些異常值嚴重影響了數據的分布和統計特徵。如果我們移除這些異常值：

新的平均值：51.1歲

新的標準差：5.7歲

這表明異常值大大增加了數據的離散程度。

Prompt: 請針對異常值進行更深入的分析 (2)

2. 身高(Height)分析：

基本統計數據：

平均值：177.1cm

中位數：180cm

標準差：26.5cm

Q1：159cm

Q3：198cm

IQR：39cm

使用 $1.5 * IQR$ 法則：

下界： $Q1 - 1.5 * IQR = 100.5\text{cm}$

上界： $Q3 + 1.5 * IQR = 256.5\text{cm}$

基於這個標準，我們沒有發現超出這個範圍的異常值。但是，我們可以注意到一些接近極限的值：

最低值：130cm (出現3次)

最高值：220cm

雖然這些值在統計上不被視為異常值，但它們仍然是極端值，可能值得進一步調查。

Prompt: 請針對異常值進行更深入的 分析 (3)

3. 異常值的影響和建議：

年齡(Age)：

- 160歲和300歲的數據點明顯是錯誤的，應該被移除或更正。
- 7歲的數據點雖然不常見，但可能是合法的。建議確認這個數據點的準確性。
- 移除這些異常值後，年齡數據的分布會更加集中，更有利於後續分析。

身高(Height)：

- 雖然沒有統計上的異常值，但130cm和220cm的身高在成年人中相當罕見。
- 建議檢查這些極端值的準確性，特別是如果它們與其他變量（如年齡、性別）不相符的情況。

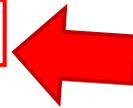
4. 後續步驟建議：

5. 數據清理：移除或更正明顯錯誤的年齡數據（160歲和300歲）。
6. 數據驗證：對於極端的身高值和7歲的年齡值，進行額外的驗證。
7. 考慮使用更穩健的統計方法：例如，使用中位數而不是平均數來表示中心趨勢，這樣可以減少異常值的影響。
8. 分組分析：考慮按性別或年齡段分組分析身高數據，這可能會揭示一些模式或進一步的異常值。
9. 可視化：創建箱形圖或散點圖來直觀地展示數據分布和異常值。

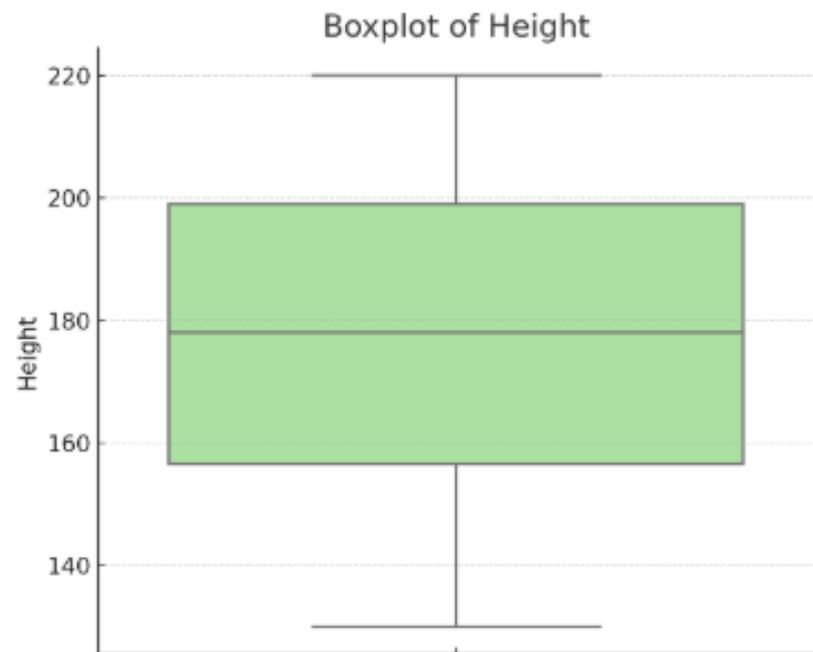
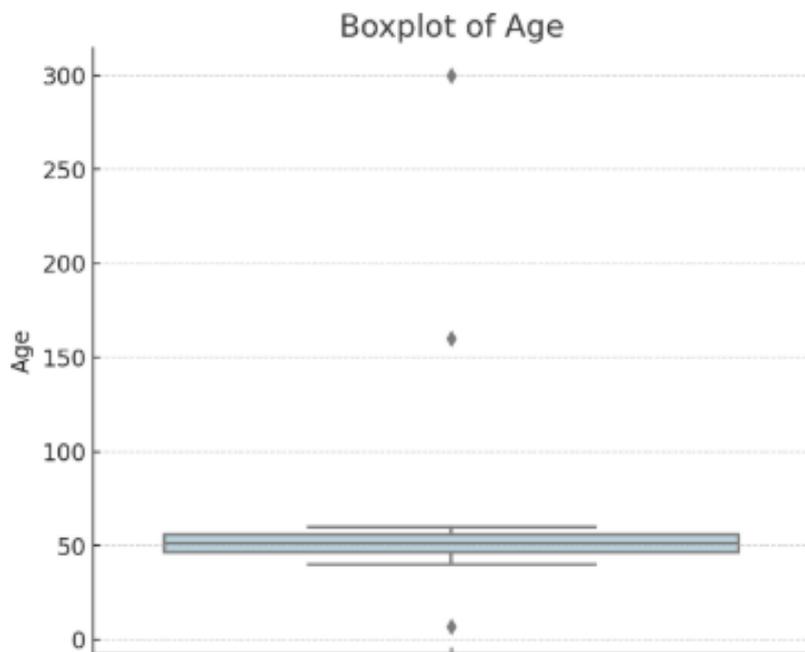


為了更深入分析 年齡 (Age) 和 身高 (Height) 的異常值，我將：

1. 統計異常值的數量 (各自有多少異常值)。
2. 計算異常值的平均數和標準差 (觀察異常數據的分佈情況)。
3. 繪製箱型圖 (Boxplot)，直觀顯示異常值的範圍和分佈。



現在來執行這些分析。



撰寫上回覆的箱型圖 (boxplot) 程式碼，用colab上面的python來實做，給我詳細步驟

1. Google Colab 操作步驟

(1) 進入 Google Colab

👉 打開瀏覽器，進入 [Google Colab](#)

(2) 上傳 CSV 檔案

- 點擊左側的 檔案 (Files)
- 按 上傳 (Upload) ，選擇你的 `DemoData.csv` 檔案

(3) 複製並執行以下 Python 代碼



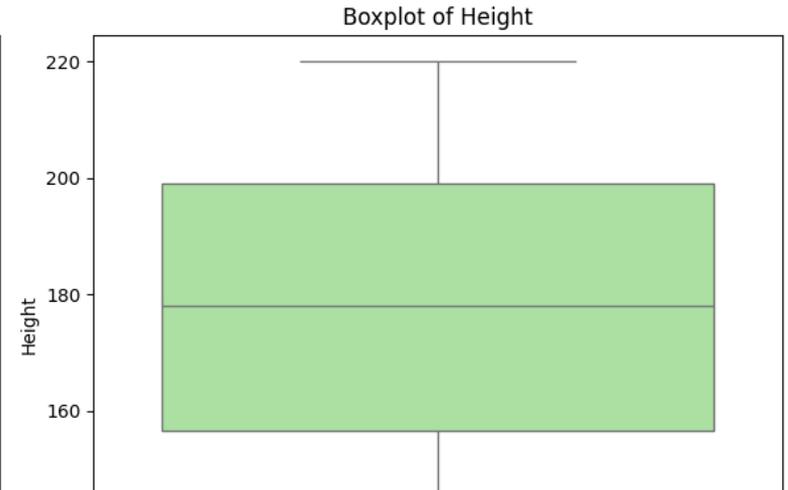
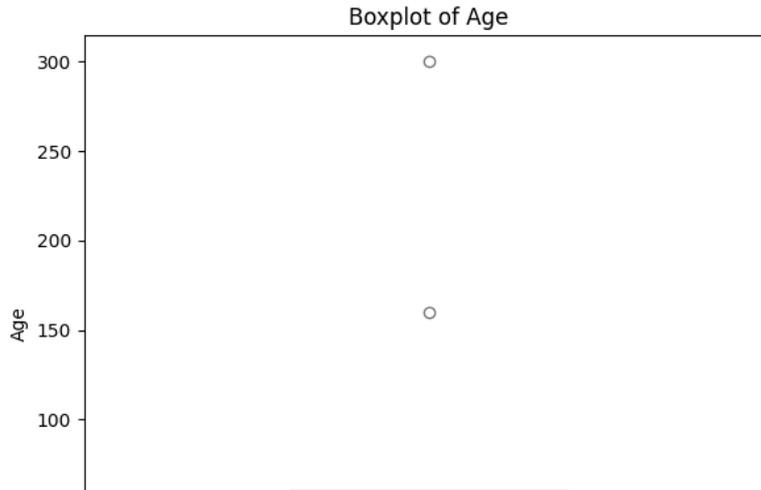
+ Code + Text

```
40 sns.boxplot(y=df["Height"], color="lightgreen")
41 plt.title("Boxplot of Height")
42 plt.ylabel("Height")
43
44 plt.tight_layout()
45 plt.show()
46
47 # ✓ 6. 顯示異常值統計結果
48 print("🔴 年齡 (Age) 異常值:")
49 print(age_outliers)
50
51 print("\n🔴 身高 (Height) 異常值:")
52 print(height_outliers)
53
```



選擇檔案 DemoData.csv

- DemoData.csv(text/csv) - 2066 bytes, last modified: 2024/8/26 - 100% done
- Saving DemoData.csv to DemoData.csv



<https://drive.google.com/>

This screenshot shows the Google Drive menu with the following items:

- 新資料夾 (Alt+C 再按 F)
- 檔案上傳 (Alt+C 再按 U)
- 資料夾上傳 (Alt+C 再按 I)
- Google 文件
- Google 試算表
- Google 簡報
- Google 表單
- 更多

The '更多' (More) option is expanded to show:

- Google 繪圖
- Google 我的地圖
- Google 協作平台
- Google Apps Script
- 連結更多應用程式 (highlighted with a red arrow)

This screenshot shows the Google Drive interface with the Google Workspace Marketplace overlay. The marketplace is titled "Google Workspace Marketplace" and features a search bar and filters. The main content area displays "Form Builder Plus" by Jivrus Technology, Inc. with a rating of 4.8 stars and 284,000 downloads. Below it, there is a section for "企業必備" (Essential for Business) with a link to learn more.

This screenshot shows the Google Drive menu with the '更多' (More) option expanded to show additional applications:

- Google 繪圖
- Google 我的地圖
- Google 協作平台
- Google Apps Script
- Google Colaboratory
- RealtimeBoard
- 連結更多應用程式

練習數據

利用 ChatGPT，



BiomedicalData.c
sv

找出這個數據的異常值、缺失值

繪製 Age 的 Boxplot、BMI 的 Bar chart

男生或女生哪一群體有比較多的運動習慣？

	A	B	C	D	E	F	G	H	I
1	ID	Gender	Age	Height	Weight	BMI	Smoking	Alcohol	ExerciseHabit
2		1 Female	42	158	78	31.2	No	No	Occasional
3		2 Male	45	163	98	36.9	No	Yes	Regular
4		3 Female	45	169	67	23.5	No	Yes	None
5		4 F	44	171	50	17.1	No	Yes	Regular
6		5 Female	48	158	97	38.9	Yes	No	Regular
7		6 Female	49	189	60	16.8	No	Yes	Occasional
8		7 Male	7	183	61	18.2	Yes	No	Occasional
9		8 Male	21	191	69	18.9	Yes	No	Occasional
10		9 Female	25	174	60	19.8	No		Occasional
11		10 Female	6	161	81	31.2	No		Occasional
12		11 Male		300	200		No	No	None
13		12 Female	22	162	67	25.7	Yes	Yes	Regular