

FROM ENTROPY TO COMPRESSION

2025.6.3

王媛婷



AGENDA

- Entropy
 - Shannon's source coding theorem
 - Python example
 - Compression
- AI-based Compression
 - Case example
- Conclusions
- Reference



ENTROPY

熱力學 Thermodynamics

- TENET
- Rudolf Clausius 於1850年提出
- 熱力學第二定律：在一個孤立系統中，熵永遠不會自行減少，只會維持不變或增加，且不可逆
 - 熱水會變涼，冷水不會自己變熱
- 熵增定律原理：系統的不確定性、混亂或能量無法利用的程度在增加
 - 整齊的房間不維持整理，則會隨時間變亂

→ 熵是用來描述一個系統的無序程度或能量分散的程度



ENTROPY

資訊理論 Information Theory

- Claude Shannon → 通訊系統中的資料具有不確定性與冗餘性，而這種「不確定性」可以被量化
- **Shannon定義的熵 (entropy) :**

$$H(X) = - \sum_i p_i \log_2 p_i$$

資訊量越高 → 越難預測 → 要越多位元來表示

- Shannon's coding theorem

$$\frac{\text{entropy}}{\ln k} \leq \text{mean}(L) \leq \frac{\text{entropy}}{\ln k} + 1$$

↑
平均編碼長度

- ! • Huffman Coding : total coding length

$$\left\lceil \frac{N \cdot \text{entropy}}{\ln k} \right\rceil \leq b \leq \left\lfloor \frac{N \cdot \text{entropy}}{\ln k} + N \right\rfloor$$

↑
總編碼長度

$$b = \text{mean}(L) \cdot N$$



SHANNON'S CODING THEOREM

Shannon編碼定理實作

請輸入你最近兩週的早餐紀錄(共14天,以英文逗號分隔):

蛋餅, 吐司, 吐司, 貝果, 漢堡, 貝果, 蛋餅, 貝果, 貝果, 蛋餅, 蛋餅, 貝果, 貝果, 蛋餅

breakfast summary

早餐吃什么	機率	編碼長度 (bit)	Shannon 編碼
貝果	0.429	2	00
蛋餅	0.357	2	01
吐司	0.143	3	110
漢堡	0.071	4	1110

report:

早餐總數: 14

早餐種類: 4

熵值: 1.727 bits

理論mean(L)範圍: 2.492 ~ 3.492 bits

實際mean(L): 2.286 bits

Possibility:

$$P(\text{貝果}) = 6/14 = 0.429$$

$$P(\text{蛋餅}) = 5/14 = 0.357$$

$$P(\text{吐司}) = 2/14 = 0.143$$

$$P(\text{漢堡}) = 1/14 = 0.071$$

Coding length:

$$\text{貝果} : \text{ceil}(-\log_2 0.427) \cong 2$$

$$\text{蛋餅} : \text{ceil}(-\log_2 0.357) \cong 2$$

$$\text{吐司} : \text{ceil}(-\log_2 0.143) \cong 3$$

$$\text{漢堡} : \text{ceil}(-\log_2 0.071) \cong 4$$



SHANNON'S CODING THEOREM

Shannon編碼定理實作

請輸入你最近兩週的早餐紀錄(共14天,以英文逗號分隔):

蛋餅,吐司,吐司,貝果,漢堡,貝果,蛋餅,貝果,貝果,蛋餅,蛋餅,貝果,貝果,蛋餅

breakfast summary

早餐吃什么	機率	編碼長度 (bit)	Shannon 編碼
貝果	0.429	2	00
蛋餅	0.357	2	01
吐司	0.143	3	110
漢堡	0.071	4	1110

report:

早餐總數: 14

早餐種類: 4

熵值: 1.727 bits

理論mean(L)範圍: 2.492 ~ 3.492 bits

實際mean(L): 2.286 bits

累積機率:

$$C(\text{貝果}) = 0$$

$$C(\text{蛋餅}) = 0.429 \rightarrow P(\text{貝果})$$

$$C(\text{吐司}) = 0.786 \rightarrow P(\text{蛋餅}) + P(\text{貝果})$$

$$C(\text{漢堡}) = 0.857 \rightarrow P(\text{吐司}) + P(\text{蛋餅}) + P(\text{貝果})$$

Shannon (coding length)

貝果 : 00

蛋餅 : 01

吐司 : 110

漢堡 : 1110

「機率」決定每個符號的編碼長度 → 越常出現的符號編碼越短



COMPRESSION

Shannon's coding theorem is based on possibility

MP3

- 聲音範圍經過小波/頻率分析轉換
- 按照聲音成分「出現機率」分配編碼長度
- 使用 Huffman 編碼 進行最後一階的資料壓縮

→ 搭配人耳不易察覺的頻率範圍，對不同頻段進行壓縮，也就是對熵較低的部分可用較少位元表示

JPEG

- 對圖像進行 DCT 轉換 + 量化
- 分析所剩值的出現機率
- Huffman coding/arithmetic coding

→ 將影像轉為頻譜後，對較常出現的係數使用更短的編碼，例如使用 Huffman coding，其編碼長度也與 Shannon 提出的理論有密切關聯

Shannon 編碼理論中量化了我們要測量的那一個量的不確定性，而熵則定義了壓縮的極限



AI-BASED COMPRESSION

為什麼需要AI-based Compression?

傳統壓縮技術限制

- 依賴人工設計的固定轉換與統計模型 (如 DCT, Huffman, JPEG)
- 無法理解資料語意
- 難以針對特定內容做最佳化壓縮

技巧



理解

固定算法



可學習模型

AI壓縮技術

圖像

→ 了解場景內容，自動分配壓縮比

聲音

→ 自適應語速與語者特性

影片

→ 預測畫面內容，只傳輸必要變化

模型

→ 降低LLM訓練的儲存與運算成本



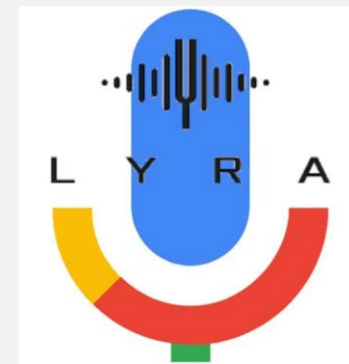
→ 不只是壓縮資料，而是學會「理解資料內容」，只保留最關鍵訊息



AI-BASED COMPRESSION

AI 壓縮技術

- 減少記憶體需求
- 降低通訊成本
- 加快推論速度與部署效率



Case I: Google Lyra

壓縮種類: 聲音

- 優勢：
- 壓縮比極高：可將聲音壓到**每秒只需 3kb 資料** → 一般聲音約需 16kbps
 - 使用神經網路 (CNN + 自編碼器) 學習聲音的特徵空間 → 只傳遞最有意義的資訊
 - Lyra 就透過模型去萃取**關鍵語意特徵**，有效地去除冗餘
 - Lyra 不使用固定編碼表

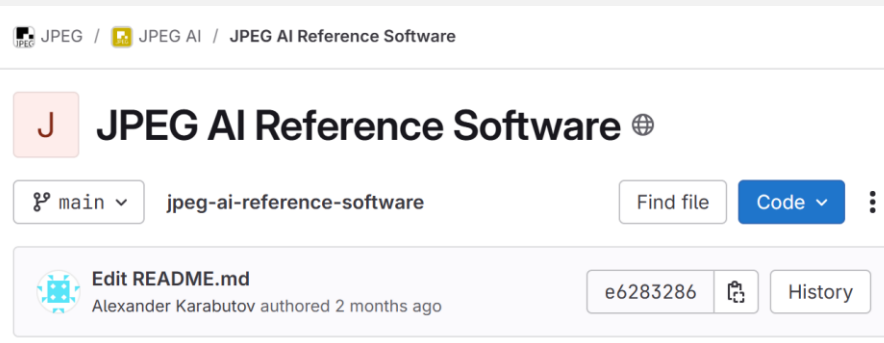
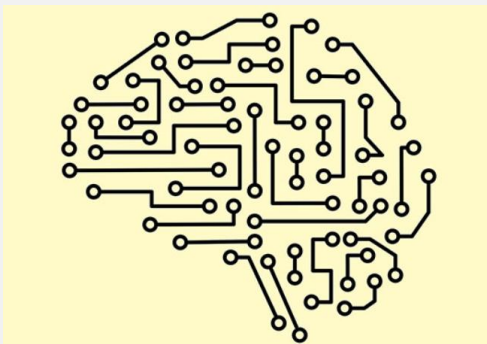
Lyra 用模型學會提取哪部分訊號是有價值，並知道該怎麼編碼。



AI-BASED COMPRESSION

AI壓縮技術

- 減少記憶體需求
- 降低通訊成本
- 加快推論速度與部署效率



Case2: JPEG AI

壓縮種類: 圖像

- 優勢：
- 捨棄了傳統JPEG使用的DCT，改用深度學習模型來學習影像的壓縮方式
 - 使用**Autoencoder**架構將圖片轉換成潛在特徵
 - 壓縮過後的潛在特徵會被再量化與編碼，解碼器再從這些壓縮特徵中還原出圖像
 - 讓模型自動學習圖像中哪些特徵最有資訊價值

JPEG AI 利用 AI 自動學習圖像，不僅壓得更有效率，還能保留更高的畫面品質，失真更少



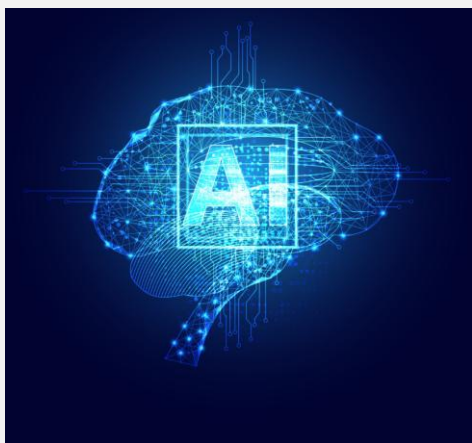
CONCLUSIONS

- **Shannon 編碼建立壓縮理論基礎**

→ 根據機率調整符號編碼長度，越常見的符號編碼越短。

- **傳統壓縮技術應用Shannon編碼理論**

→ 像 JPEG、MP3 等壓縮技術，都先分析資料分布，再進行透過變長編碼



- **AI 壓縮技術突破熵極限限制**

→ 利用模型學習資料結構與語意，不再只靠統計特性進行壓縮。

- **AI壓縮技術正支撐 AI 發展需求**

→ 透過理解資料的語意與結構，實現超越傳統壓縮極限的效率

→大幅降低訓練成本 / 節省記憶體、降低運算成本



REFERENCE

- Entropy <https://en.wikipedia.org/wiki/Entropy>
- 《天能》裡那些有點難的物理學：一個「逆熵」的世界為何不合理？
<https://pansci.asia/archives/190653>
- Shannon's source coding theorem
https://en.wikipedia.org/wiki/Shannon%27s_source_coding_theorem
- http://boson4.phys.tku.edu.tw/thermodynamics/Ch-15_Information_theory.html
- Google lyra <https://www.ithome.com.tw/news/153412>
- JPEG AI <https://jpeg.org/jpegai/>
- JPEG AI github <https://gitlab.com/wgl/jpeg-ai/jpeg-ai-reference-software>
- Compression: ADSP-8 P275
- Entropy and coding length: ADSP-9 P325-327、P338
- ChatGPT

