行政院及所屬各機關出國報告
（出國類別：開會）

參加「2024年國際AI機器學習(ML)大會(ICML 2024)」出國報告

服務機關：財政部財政資訊中心

姓名職稱：廖婉淑組長

鄭仕謙助理程式設計師

派赴國家：奧地利／維也納

出國期間：113年7月24日至113年7月30日

報告日期：113年10月21日

摘要

國際機器學習會議（ICML）中，探討了關於 LLMs 的幾個主題演講，並深入分析了各演講中所提出的挑戰和解決方案。其中，長文本基礎模型的研討會探討了模型評估、計算效率和數據需求等方面的挑戰。在高效長序列生成的演算法與硬體協同設計的演講中重點分析了鍵值緩存瓶頸問題，並提出了靜態壓縮和動態壓縮兩種技術，最後介紹透過 GPU 與 CPU 的協同設計來提升效率。狀態空間模型的權衡分析的演講則探討了狀態空間模型和變換器模型的優缺點，並提出了 Mamba 模型，結合兩者的優勢。

數據導向機器學習研究的研討會中，則強調了數據集的重要性，探討了數據整理、數據集的著作權以及以數據為中心的視角等問題，並介绍了一些常用的模型和方法。整體而言，這場會議深入淺出地介紹了 LLMs 領域的進展，並分析了相關的技術挑戰和解決方案，為與會者提供了對該領域理解的思考方向。

目次

壹、會議介紹及參與目的

一、會議介紹

人工智慧（Artificial Intelligence, AI）的發展歷史可追溯1950年，最早的先驅之一艾倫·圖靈(Alan Turing)，他提出了著名的「圖靈測試」，是一種用以評估機器能否模擬人類智能的方法。接著1956年在美國新罕布什爾州的達特茅斯學院召開了一次關鍵性的會議，這也被廣泛認為是人工智慧的正式起點。在這次會議上，與會者提出了許多基礎性的AI概念，探討了如何讓機器具備學習、推理和解決問題的能力。他們的願景是創建能夠模擬人類思考過程的機器，這一理念奠定了後來AI研究的基礎。

但在接下來的幾十年，因為技術限制和資源不足，AI的發展陷入低潮。直到1980年代，隨著計算能力的提升和神經網絡的興起，AI再次獲得關注。

進入21世紀，機器學習和深度學習的進步使AI實現了飛躍性的發展。隨著各科技大廠，如Google、Amazon等科技巨頭投入大量資源，推動了自然語言處理和計算機視覺等應用發展。AI技術逐漸被應用於各行各業，如醫療診斷、金融分析和自動駕駛等。目前，AI的發展正朝向更加智能化和自主化的方向前進。

本次參與國際機器學習會議（International Conference on Machine Learning, ICML）是機器學習領域最具影響力的學術會議之一，首次舉辦於1980年。旨在促進學術界、產業界和研究者之間的交流，聚焦於機器學習及其應用的最新研究成果。

ICML會議每年吸引全球數千名研究者和專家參加，會議內容涵蓋多種主題，包含監督學習、無監督學習、強化學習、生成模型、深度學習等。除了技術報告，會議還設有多種形式的活動，如研討會、講座和海報展示，藉此促進參與者之間的互動。

除了學術研究，ICML 會議也積極促進產業界的參與，設立特別的論壇和展示區，讓公司和創新者分享他們的應用和技術進展。這種產學合作的模式使得會議成為技術交流和應用探索的熱點。

二、參與目的

近年來，隨著 OpenAI 推出 Chatgpt 的使用，各類生成式 AI 百花齊放，在本次的 ICML 會議中最熱門的焦點莫過於大型語言模型與生成式 AI 模型的發展及應用。因應創新科技趨勢之下，期望能導入相關的應用，輔助機關同仁們日常工作，提升效率。透過參與本次的會議，希望藉此機會多了解相關的應用，不僅可做為自身的參考，也同時透過會議中專家學者們提出的想法，促進自身的思考，提升對於該領域的敏感度，在運用與導入相關 AI 模型時能更好的評估其效益。

貳、會議內容摘述

本次會議為期一周，除了各項主題演講外，還有各主題投稿論文發表及不同主題的研討會供與會人員自由參與。因時程上的安排，本次會議僅參與最後兩天的研討會活動。

最後兩天的研討會場次繁多，多達 29 場，每場次也均安排整天的議程。考量討論主題能運用的範圍結合業務上的使用，挑選了兩場次較為完整參與的研討會進行描述，主要針對基礎模型及數據資料的相關概念，以下分別詳述。

一、長文本基礎模型(Long-Context Foundation Models)

此研討會旨在召集研究人員針對長文本模型開發及討論，目前主要遇到的挑戰有以下三點：

1. 計算上的效率：隨著文本長度的增加，計算上的複雜度會隨著呈現二次增加。
2. 可使用數據的匱乏：要訓練此類的模型，勢必需要有大量長文本且連續的數據供使用，但可用的數據集有限，此需求難以被滿足。
3. 評估的複雜性：評估長文本基礎模型的效果本身就非常複雜，因為收集、構建或驗證此類評估數據的成本很高，皆需人工作業。

(一)主題演講1：如何評估長文本語言模型

第一場主題演講邀請的是來自馬薩諸塞大學阿默斯特分校的計算機科學副教授 Mohit Iyyer，主要的研究為自然語言處理與機器學習。而講者這次要探討的主題是「如何評估長文本的語言模型」。

一開始先提到，隨著近來大型語言模型規模的增加，從2,000個標記(Token)提升至128,000個標記，甚至有些模型提升至1,000萬個標記，

這一切得益於注意力機制(Attention Mechanism)[1]及硬體的發展，使得這一進程能快速推進，但這也導致超出對長文本模型有效評估其性能的能力。

接者講者提出目前在評估長文本檔案時所遭遇的挑戰：

1.缺乏可靠的自動化指標：現有的自動評估指標在處理非常長的文檔時效果不佳，難以迅速評估正確性。

2.人工評估：雖然人工評估非常徹底，但對於非常大的文檔來說，需耗費高昂的成本，且同時須耗費大量的時間。

3.數據汙染：公開可用的資料基準通常包含模型在訓練過程中可能已經記住的數據，從而扭曲產生的結果。

為了解決這類的問題，目前常用的方式為 尋找隱藏樣本(Needle in a Haystack, NIAH)，這種做法是將一小段訊息插入至長文本中，例如：製作三明治的秘密原料為「維吉麥醬」(Vegemite)，然後再詢問模型製作三明治的秘密原料為何？一般來說，模型可以找到正確的答案。此種方式雖然對於快速評估有效，但其主要測試的是模型的檢索能力，而非更深層次的理解或總結技能。

接著講者提到兩種他們團隊進行的專案(Fables項目及 NoCha項目)，用以驗證長文本模型的總結的能力，測試標的為長篇科幻小說：

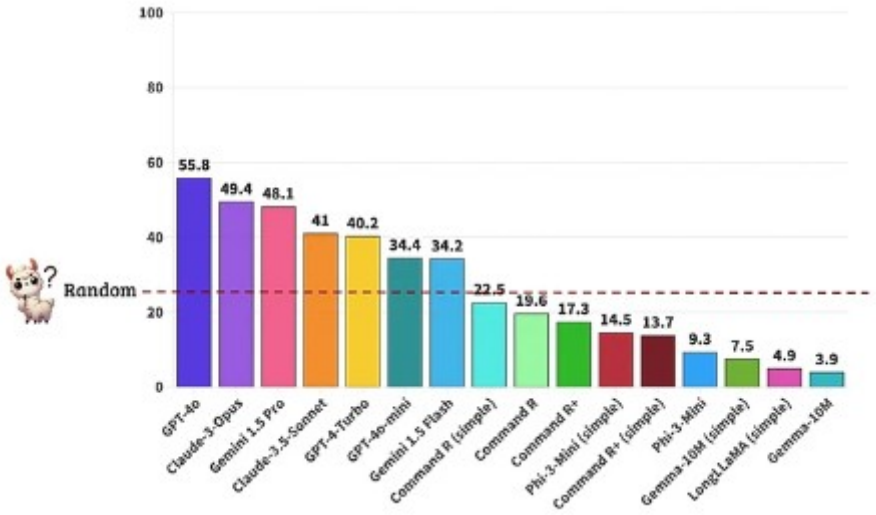
1. Fables 項目：此項目專注於評估各種模型生成的摘要的事實準確性。會先讓各模型做出書裡的總結，再讓其產生多項書中描述的事實，接著讓註釋者去比對書中的資料。儘管總結了最近各種模型呈現的結果，但人工評估者準確看到錯誤還是相當困難，且耗時耗力。下圖左邊框出各模型的準確度，右邊呈現人工評估能確認的準確度。

1 注意力機制(Attention Mechanism)是一種能夠根據重要性動態加權信息的技術，從而幫助模型專注於關鍵部分以提升性能。

Model	Faithful	Unfaithful	Partial support	Can't verify
GPT-3.5-TURBO	72.07	10.52	13.01	4.41
MIXTRAL	70.04	10.46	16.72	2.78
GPT-4	78.55	4.54	15.53	1.38
GPT-4-TURBO	78.16	7.62	11.41	2.82
CLAUDE-3-OPUS	90.66	2.03	7.06	0.26

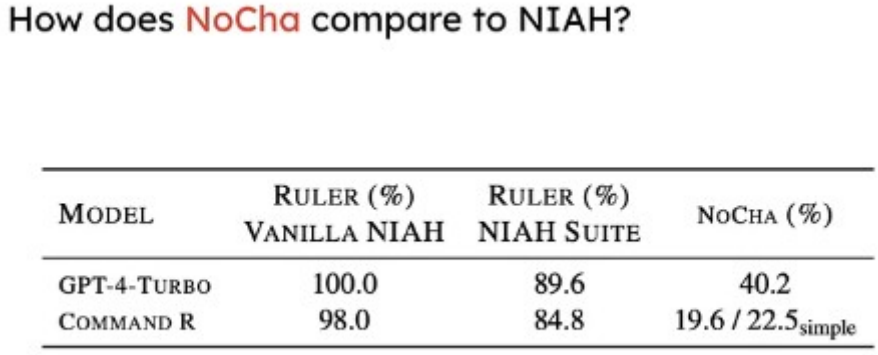
2. NoCha (A Novel Challenge for long-text LLMs)項目：此項目請註釋者寫出多組書中正確與錯誤的描述，並測試各模型在閱讀小說後是否能正確答出答案。為了驗證這些描述並非難以理解的句子，各個註釋者也分別回答了各項問題，正確率達到97%，藉此也表示只要有讀完該書，這些描述基本上都是可以理解的內容。

但經過測試，各種常見模型回答問題的正確率，得到不盡理想的結果，如下圖所示，突顯了當前模型的不足，有些模型的表現甚至低於隨機準確率。



接著，在各模型對於長文本的理解力上，考量 Fables 方法因無法確認其結果準確性，會議中講者以 NoCha 專案的方法對比 NIAH 對於長

文本理解力的表現，如下圖所示，相較於 NoCha 驗證各模型理解力的結果， NIAH 良好的能力並不是有效的推理或是描述與事實的相關內容，而僅是證明其擅長的為簡單的檢索查詢任務，在碰上文本的理解議題上，還是有相當大的挑戰。



MODEL	RULER (%) VANILLA NIAH	RULER (%) NIAH SUITE	NOCHA (%)
GPT-4-TURBO	100.0	89.6	40.2
COMMAND R	98.0	84.8	$19.6 / 22.5_{simple}$

最後講者提到，未來長文本模型的評估還是需要有更穩健的自動化方法來取代人工評估，減少成本的消耗。

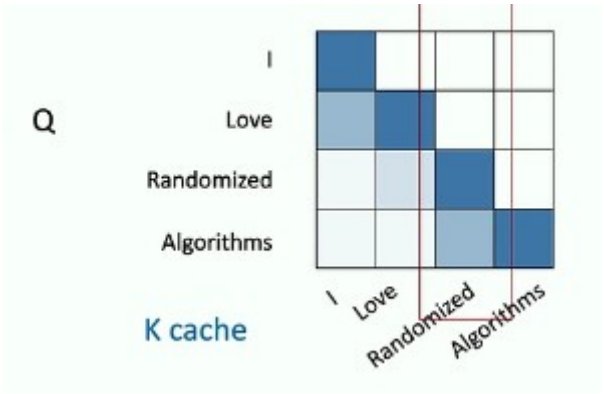
(二)主題演講2：高效長序列生成的演算法與硬體協同設計

第二場主題演講邀請的是來自卡內基梅隆大學擔任電機與計算機工程助理教授 Beidi Chen，其研究領域為演算法與硬體協同設計，藉以加速優化大規模機器學習系統。

講者一開始引言講述到，最近各種大語言模型規模持續增加，為了要部署這些模型，所需要的花費必定會增加，各項基礎設備也需要跟著升級。講者的目標是有效管理內存和處理需求，以容納這些大型語言模型，而不產生過高的成本，因此要設計一款更高效的演算法及系統，以支持長文本的大型語言模型推理。

遇到的第一個挑戰鍵值緩存瓶頸(KV Cache Bottleneck)[2]，隨著大型語言模型規模逐漸擴大，所需要的 GPU 內存也會跟著增加，往往會超出原有的基礎設施容量，而對於長文本模型的長序列內容又更消耗內存的使用。

而在之前的研究，在硬體支援有限的清況下，勢必要做到鍵值內存移除(KV cache eviction)[3]，因此必須找出需要移除的鍵值為何？講者提出選擇對鍵值評分找到常用的標記，評分高的標記作為保留的對象，這能夠保持效能同時減少計算量。但遇到另一個問題，大部分的鍵值評分過度集中於最起始的部分，造成有注意力沉沒(Attention sink)的狀況發生。



為了處理此一問題，利用串流式大型語言模型(StreamingLLM)[4]的技術，允許模型逐步接收輸入，再依序推理，藉此達到標記擴張至全文的效果，但這又產生另一個問題，這會使模型忘記中間的內容，只記得最初與部分關鍵值。

2 鍵值緩存瓶頸(KV Cache Bottleneck)指的是隨著輸入序列長度增加，鍵值存取因記憶體限制和計算負擔過重，導致模型效率下降的問題。

3 鍵值內存移除(KV cache eviction)是指為了減少內存使用和計算負擔，主動移除不再需要的鍵值對，以維持高效運行的過程。

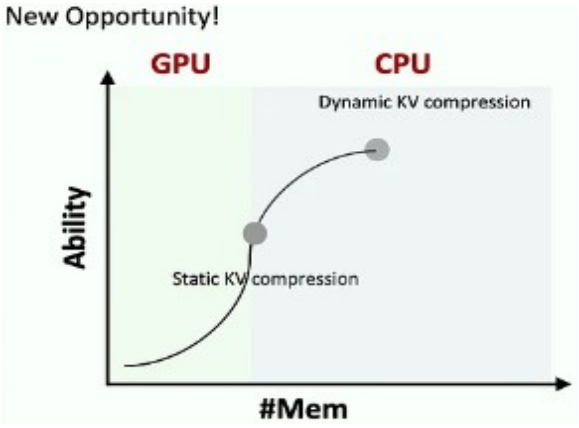
4 串流式大型語言模型（Streaming LLM）是一種逐步接收輸入數據並生成輸出，而非一次性處理整個輸入的模型，特別適合用於即時應用，如聊天機器人和語音助手。

講者提出靜態壓縮的技術，透過將被移除的鍵值標記壓縮其內容信息，藉此減少內存的使用，但是一旦移除的鍵值信息，就無法恢復，這也導致推論時常出錯。

而相對靜態壓縮，動態壓縮技術不再捨棄鍵值信息，而是動態選擇使用的鍵值對象，搭配推測解碼(Speculative Decoding)的技術[5]，利用小模型推測大模型使用的標記鍵值，減少計算量同時加速整個過程，而且可做到無損推論，與靜態壓縮相比，鍵值緩存瓶頸從大模型的問題轉移成需要紀錄所有鍵值的問題上。

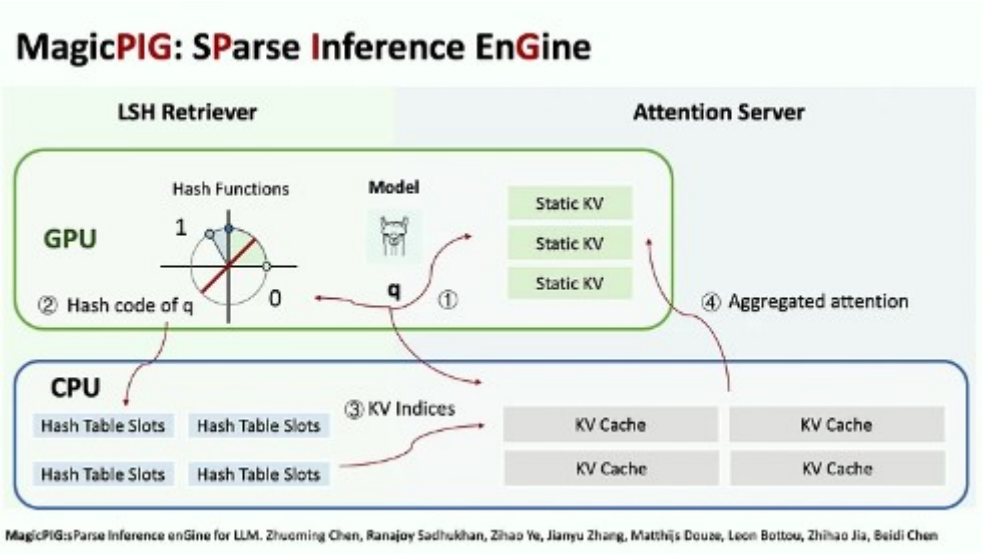
5 推測解碼（Speculative Decoding）是一種用於加速文本生成的技術，它在生成過程中預測接下來的字詞，並在得到部分輸入後進行多條可能的解碼，然後根據預設標準選擇最合適的輸出。

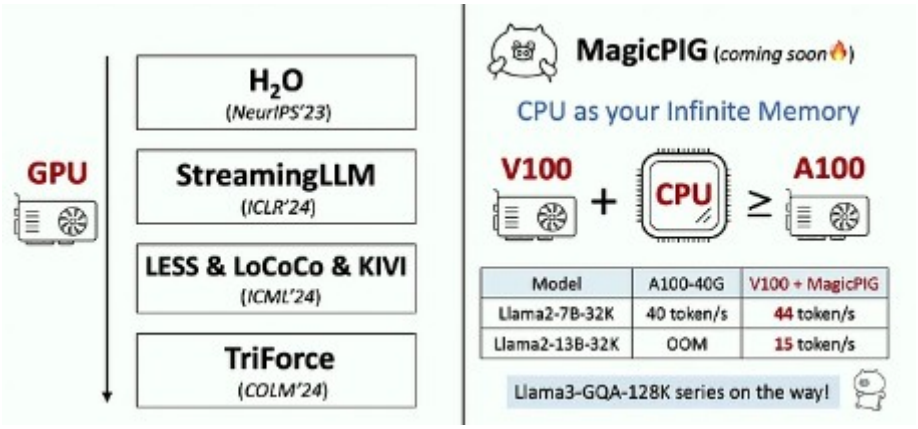
從上述內容可得到一個結論，靜態壓縮下可以節省內存的使用，但會捨去推論的能力；動態壓縮則是保持推論的能力，但需要大量的儲存空間。



因為時間因素，講者接著直接帶出了本次研究的結論，在MagicPIG項目中，為了解決以上的難題，透過以 GPU 與 CPU 的協同設計，將靜態的鍵值存放在 GPU 上，其他動態所需的鍵值存放在 CPU 上，同時以 CPU 做計算，等同於將 CPU 當成 GPU 的無限儲存記憶體。而執行的過程為：每當查詢(q)，在 GPU 計算 雜湊函數(Hash function)[6] 找到雜湊值(hash code)[7] 傳給 CPU，同時處理靜態鍵值相關的查詢部分，再由CPU找到動態鍵值傳回GPU以完成整個推論的流程。

6 雜湊函數(Hash function)是一種將任意大小的輸入數據轉換為固定長度雜湊值的數學函數。
7 雜湊值(hash code)是由雜湊函數生成的固定長度字符串或數字，代表輸入數據的唯一標識，通常用於數據完整性檢查和快速查找。





最後對於未來的發展方向，講者提到除了 CPU 與 GPU 的協同設計也可以從內存的架構、查詢的演算法及持續發展硬體設備的儲存體著手。

(三)主題演講3：狀態空間模型的權衡分析

第三場主題演講邀請的是來自卡內基梅隆大學的機器學習助理教授 Albert Gu，其致力於研究讓人工智慧能具備類似記憶的能力，同時其也是 Cartesia AI 創新公司的聯合創辦人。

講者一開始先大致帶過這場演講所要描述的內容，講到自動回歸模型(Autoregressive Modeling)是一種透過原本依序輸入的內容去預測接下來組成的機器學習模型。接者講者定義何為序列模型(Sequence model)，此種模型對於要處理的輸入資料數據是有順序或是有時間排序的，例如循環神經網絡(Recurrent Neural Networks, RNNs)[8]。

接著帶出此演講的重要概念：自動回歸的狀態(Autoregressive State)。這個「狀態」指的是序列模型中，時間點之間所記憶保留下的內容，舉例來說，當我們進行語言生成時，每個標記的產生是一個接著一個，而在每個標記之間，模型會紀錄下一些信息資訊，這些資訊是用來產生下一個標記的必要產物，這類的模型也可視為狀態空間模型(State space model, SSM)[9]。

RNN 就是典型的例子，一次一個單位的處理所輸入的序列資料，狀態空間模型可看做為一種RNN模型，但與過去RNN不同的地方在於隱藏狀態層的維度會大於輸入層的維度。使用RNN的好處在於狀態大小都是固定的，當序列資料輸入時，處理時間可以保持一致，也能確保輸出可以在常數時間推理完成，但反面的來說，當資料量大時，更多的訊息被壓縮至固定大小的狀態內，對於訊息量大的資料會有應用上的限制，舉例來說，語言的描述，如果經過壓縮，剩下的意義可能就有所不同，而變換器(Transformer)[10]就顯得重要。

注意力(Attention)模型，這種模型會在每一步的處理時比對所有資料的互動關係，這也就表示在模型推理的時間內需要處理很大量的資料比對工作，講者將此種模型視為對資訊無壓縮的模型，例如變換器就是

8 循環神經網絡（RNNs）是一種能夠處理序列數據的深度學習模型，通過隱藏狀態將先前的信息保留在計算中，這使得它們能夠捕捉到時間序列中的長期依賴關係。

9 狀態空間模型（State Space Model, SSM）是一種分析時間序列數據的數學框架，通過狀態變量和觀測方程捕捉內部狀態與可觀測輸出之間的關係。

10 變換器(Transformer)核心就是注意力機制。它讓模型能夠將輸入序列中的每個元素與其他所有元素進行比較，計算出每個元素對其他元素的注意力分數，然後根據注意力分數對輸入序列進行加權求和。

一種注意力模型。相較於 SSM ，雖然可以更好的處理訊息量大、相依關係更遠的資料，但訓練所耗費的時間也呈平方成長。

根據上述的內容，為了處理如「語言」這類的資料，講者提出三個如何讓 SSM 更有效率不可或缺的核心元素：

1.狀態大小：自動回歸的狀態可以儲存多少的訊息？為了處理更多狀態，SSM 的隱藏狀態層使用「狀態擴展」的方式，一維的輸入資料，進入到狀態隱藏層用多維去儲存更多的訊息資訊，至於儲存的維度要多大，則視使用者所認為需要儲存的資訊量而定，例如線性注意力(Linear Attention)模型[11]。

2.狀態更新：狀態能有多強的轉譯表達能力？考量狀態容量是有限的，勢必要權衡什麼訊息是必須放入狀態內的，在 Mamba模型[12]中採用選擇性(Selectivity)[13]的方式進行狀態的更新處理，而其邏輯採用的是輸入依賴轉換矩陣(Input-dependent transition matrix)[14]，讓模型自行去依輸入的內容判斷什麼該記住、哪些該忽略。

3.計算效率：為了有更大的狀態儲存訊息，同時能更好的處理狀態更新的函式，需要更好的利用 GPU 去進行關聯式掃描提升效率。

對於近期的語言模型，雖然還是有其他面向需要考慮，但以上的三個元素格外重要，Mamba 模型即為結合以上優點的模型，Mamba2 最近也推出，相較於 Mamba 雖然犧牲了一些狀態的表達能力，但擴大了狀態的

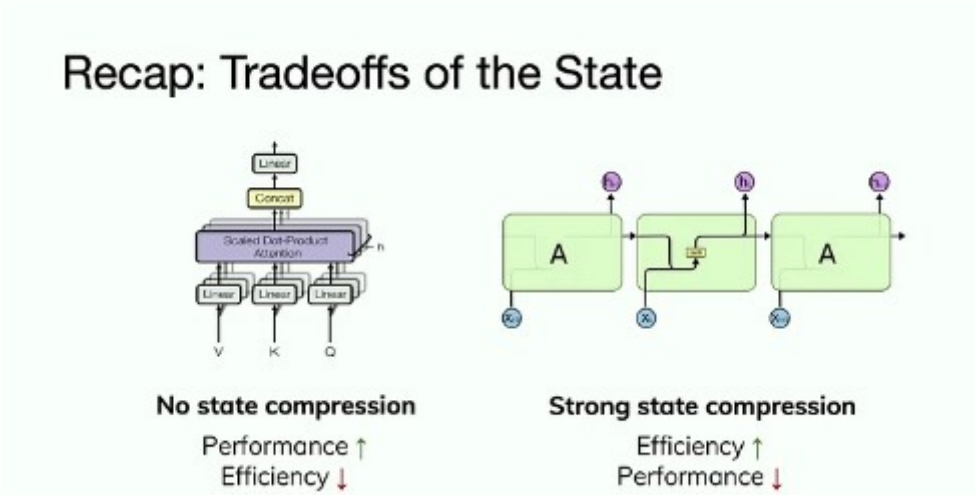
11 線性注意力（Linear Attention）是一種改進的注意力機制，通過將計算複雜度從二次方降到線性，從而能夠處理更長的序列，通常用於大型語言模型和其他序列任務中，以減少內存使用和計算時間。

12 Mamba模型是作者Albert Gu 提出的一種新架構，旨在克服傳統變換器模型在處理長序列數據時所面臨的效率和準確性問題。是一種選擇性狀態空間模型（Selective State Space Model）。

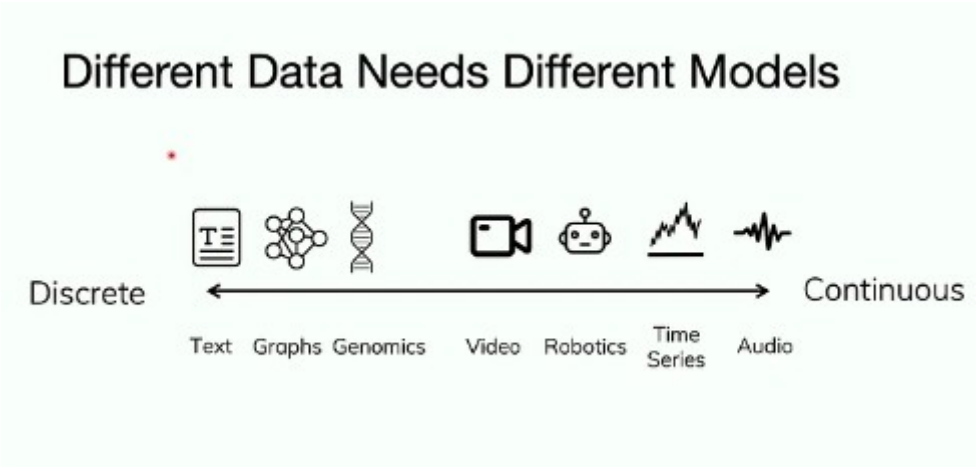
13 選擇性(Selectivity)指的是面對大量數據或複雜任務時，能夠準確地從眾多選項中挑選出最相關或最可能結果的能力。

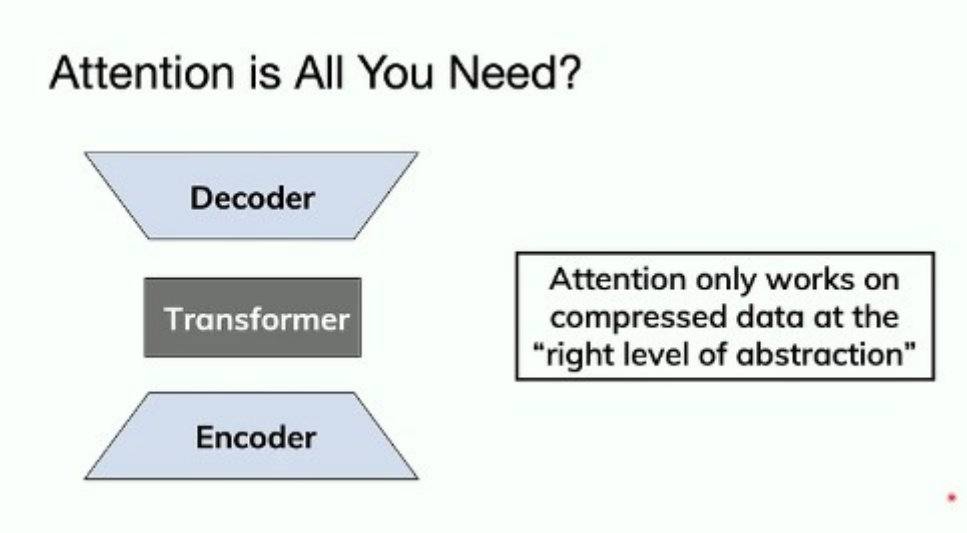
14 輸入依賴轉換矩陣(Input-dependent transition matrix)是一個矩陣，它的元素值會根據系統的輸入而改變，從而描述系統在不同輸入下的狀態轉換關係。

大小，且更有效率的在GPU上進行運算。但不管哪一個模型，都還是受限於有限的狀態儲存，這也是為何需要討論變換器，並與 SSM 進行比較。



接著講者提到，不同類型的數據（如語言與圖像）需要不同的模型。變換器在處理分詞的離散數據時表現出色，但在連續或非常高維的數據上可能表現不理想，在這裡講者也建議了變換器模型的使用情境，可以運用於經壓縮後的狀態。因此在某些情境下，混合模型可能表現更好。





最後講者總結提到混和模型結合 SSM 與變換器的優勢，利用變換器的機制強化 SSM 的表現，可產生更高效的模型，總體來說還是依照所需進行的任務選定使用的模型。講者結尾時將其與人類的智慧進行比較，指出模糊的記憶和明確數據存儲在模型中都是有價值的，未來可朝這方面去持續研究。

(四)本場研討會得獎論文

本場研討會對於這次投稿的論文頒發了最好論文獎項，而此篇論文的作者在本次的研討會也有進行其口頭報告，以下為此篇論文概述：

Improved Algorithms for Kernel Matrix-Vector Multiplication

核矩陣-向量乘法的改進演算法

1.研究動機

本文研究了核矩陣-向量乘積的計算效率，特別是聚焦於高斯核矩陣(Gaussian kernel matrix)[15]，這在像變換器這樣的機器學習應用中

15 高斯核矩陣(Gaussian kernel matrix)將數據點之間的相似性轉換成一個矩陣，矩陣中的每個元素代表著兩個數據點之間的相似度，相似度越高，對應的矩陣元素值就越大。

至關重要。它強調了與核矩陣-向量乘法的直接算法相關的二次時間複雜性挑戰，突顯出對更快演算法的需求。本文的重點在於開發快速演算法來計算涉及非對稱高斯核矩陣的矩陣-向量乘積。

2.關鍵假設

本文開發的演算法依賴於一個關鍵的建模假設：核矩陣K 中各條目的總和會隨n線性增長，而不是呈現最壞情況的二次增長。通過這次的實驗評估來驗證這一假設，也證明其適用於各種機器學習背景中遇到的高斯核矩陣，包括在大型語言模型（LLMs）中的快速注意力計算。

本文講者提出的演算法專注於有效估計來自核矩陣中重鍵和輕鍵的貢獻，利用建模假設來降低計算複雜性。本文中概述了一個預處理步驟，以處理輸入向量中的大條目，並採用抽樣策略來估計來自不太顯著條目的貢獻。

3.實現技術

本文採用了幾種創新的方法，以實現快速的核矩陣-向量乘法算法。首先定義基於一組 n 個鍵 k1, k2, ..., kn 和查詢 q1, q2, ..., qn 的高斯核矩陣 K。目標是計算給定向量 x 的乘積 Kx 。

以下是關鍵技術的概述：

(1)局部敏感雜湊(Locality-Sensitive Hashing, LSH)[16]

利用LSH 函數高效處理高維近似最近鄰搜索問題，其核心思想是：相似的資料點被映射到相同對應位置的機率比不相似的資料點高。這一技術有助於識別每個查詢的「重要」鍵值，即對核矩陣-向量乘積貢獻

16 局部敏感雜湊（Locality-Sensitive Hashing, LSH）是一種用於在大規模數據集中快速查找近似最近鄰的概率性算法。

顯著的鍵值。如果核函數 k(qi, kj) 是大的，則認為一個鍵值是重要的，這有助於縮小每個查詢所需的計算。

(2)鍵值分類(Key Classification)

根據鍵對核矩陣-向量乘積的貢獻將鍵值分類為「重鍵」和「輕鍵」。對於每個查詢，首先識別重鍵，然後準確計算其貢獻。

(3)輸入向量的前處理(Pre-processing of Input Vector)

把輸入的向量x 做前處理，讓其集中於其重要的條目。極大的條目會被明確計算其對Kx的貢獻，而極小的值則被四捨五入為零，以減少誤差。這一步確保剩餘的 x 值可管理並在可控範圍內。

2.隨機抽樣

對於輕鍵，實施了一個隨機抽樣程序。每個輕鍵以1/n的概率均勻子抽樣，讓算法能夠有效地估計所有輕鍵對核矩陣-向量乘積的貢獻。這種方法有助於減少計算中使用的估計器(estimator)[17] 的變異數(variance)[18]。

3.變異數的減少

分析輕鍵的估計器的變異數，並確定為達到所需精度所需的重複次數。這是通過快速的高斯核密度估計原理實現的，有助於優化抽樣過程。

4.結論

該論文的結論是，所提出的算法顯著提高了核矩陣-向量乘法的效率，同時解決了機器學習應用中的實踐挑戰。但其也承認了使用上的限制，雖然實證驗證支持這個假設，但實踐中的矩陣多樣性意味著沒有單

17 估計器(estimator)是一種用來估計未知參數的工具。根據我們已有的數據，來推測出我們不知道的資訊。

18 變異數(variance)是一個用來描述一組數據分散程度的統計量。簡單來說，它告訴我們一組數據中的每個數值離平均值有多遠。變異數越大，表示數據分散得越開，反之則表示數據聚集得越緊密。

一的假設可以完美地模擬所有場景，這也為這項研究的貢獻提供了平衡的觀點。

二、資料導向機器學習研究：基礎模型的數據集 (Data-centric Machine Learning Research (DMLR): Datasets for Foundation Models)

這場研討會強調資料集的重要性以及資料的質量，在這次的會議中會把重點放在基礎模型的數據集議題。

這場研討會希望帶給大家的重點：

1. 數據很「重要」
2. 數據的「質量」很「重要」
3. 數據集的創新會帶領機器學習模型的進步

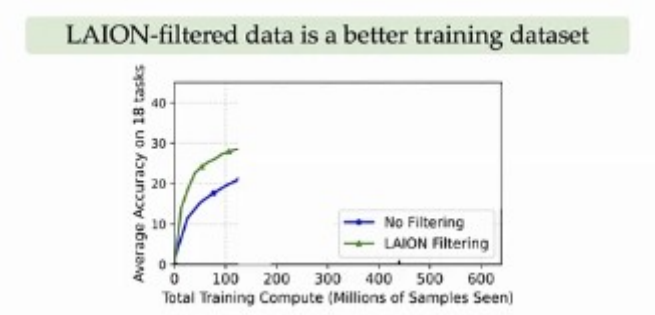
(一)主題演講1：質量與數量：數據整理(Data Curation)的困境

這場的講者來自卡內基梅隆大學計算機科學系的助理教授 Aditi Raghunathan，他致力於打造可靠的機器學習系統，確保它們在現實世界中的運行是可行的。

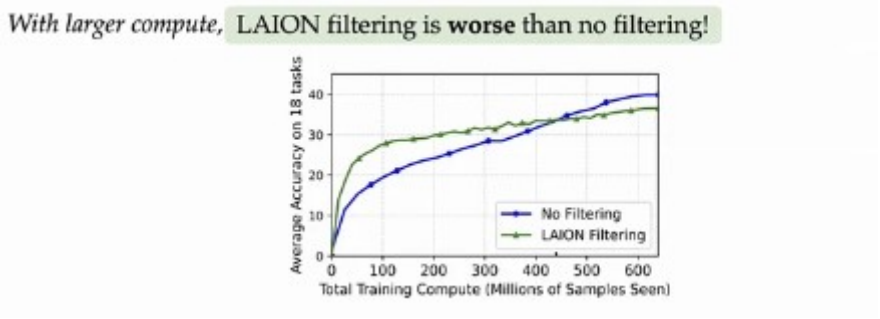
首先講者提到我們所熟知的基礎模型大多是透過大量的網路資料進行訓練，而大多的網路資料皆為異質的，需要花很大努力對其進行數據整理，而這其中說不定會有些資料對於模型的訓練來說是關鍵的元素，因此在提取網路上的異質資料時要非常小心，藉此取得高品質的資料數據。

對於異質的網路資料，講者提到會對其進行適當的分級，數據整理的方式則是利用過濾器挑選出良好的訓練數據集，但這個過濾器該如何做？首先需要先定義何謂良好的訓練數據集，

這裡講者舉了一個 CLIP[19] 的訓練模型在使用了 LAION[20] 過濾的數據，同時也使用未過濾的數據進行比對，講者特別強調使用的過濾器不是這裡要討論的重點，經過比對可以發現經過過濾的數據所訓練出來的模型正確率高於未經過濾的數據。



可是這結果是否就是正確的？當提高了算力時，未過濾的數據所訓練出來的模型正確率會提高，甚至超過經過濾的數據集。這裡講者給了一個小結，「數據整理無法忽視計算資源的影響」，從這裡獲得一個收穫，有多少計算資源將決定數據整理的方式。



接著討論到質量與數量的權衡，當計算資源較低時，模型訓練的週期(epoch)[21] 較少，這時候使用較高品質的數據集，會有較好的結果；但當計算資源提高時，模型訓練的週期增加，用同樣的高品質數據集變得

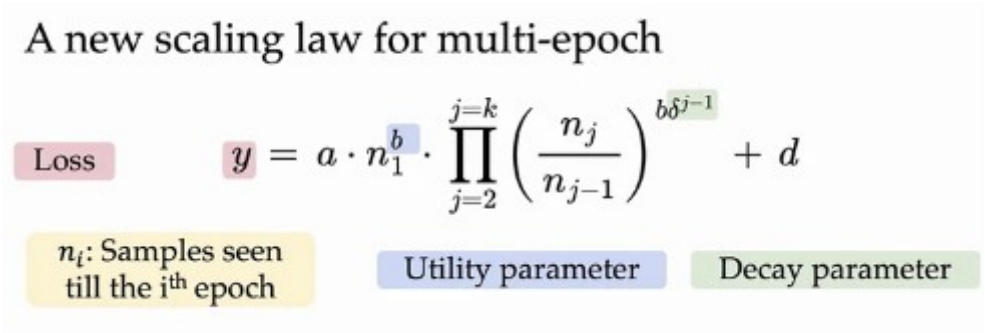
19 CLIP（Contrastive Language-Image Pretraining）是一種由OpenAI開發的多模態學習模型，能夠將文字和影像連結起來。它透過大量的文字、影像配對數據進行訓練，學習到文字和影像之間的關聯性。

20 LAION (Large-scale Artificial Intelligence Open Network) 是一個非營利組織，致力於提供開源的人工智慧模型和數據集，以促進機器學習研究。

21 週期(epoch)指的是整個訓練數據集被完整地迭代(iteration)神經網絡一次。舉例，假設你有1000張圖片的訓練集，每次訓練只取100張（batch size=100），那麼完成一個週期需要10次迭代（1000/100=10）。

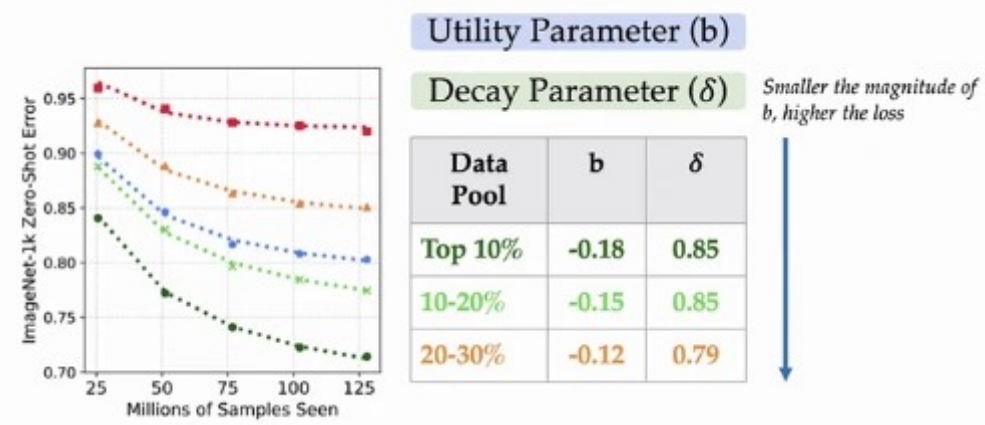
不再那麼有效，這時加上一些其他資料，可能品質較低也沒關係，反而可以提升模型的訓練成效，這時候數量的重要性就超越質量了。

下一個目標是要評估何時要以質量為重，何時又需要改以數量為重？而高品質的數據隨著模型訓練的週期越多，同樣的數據能帶來的效用就會降低，講者提出了結合不同品質的數據集來訓練模型，這裡要先找到一套縮放法則(Scaling laws)[22]來做為過濾器的標準，藉此找到適合的數據集組合方式。



透過以上縮放公式計算數據集的損失函數(Loss function)[23]，Utility parameter 指的是一開始訓練資料能帶來的效用，Decay parameter 指的是隨著同樣的數據訓練次數增加會造成效用遞減的速度。

把公式套用至不同品質的資料集上得到以下結果：

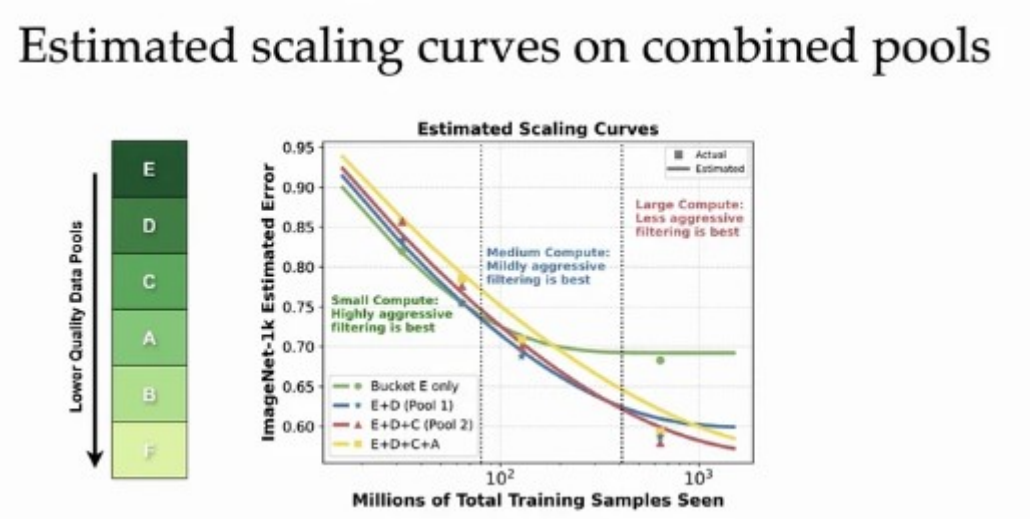


22 縮放法則(Scaling laws)描述模型的性能隨著模型大小、數據量和計算資源的增加而變化的規律。

23 損失函數(Loss function)是一個用來衡量模型預測結果與真實標籤之間差異的函數。計算出的損失值越小，表示模型的預測結果越接近真實值。

當品質低的數據集(如上圖20%-30%的數據集)，他的損失函數計算出的結果較高，遞減的較緩慢，這也表示高品質的數據集能有較好的正確性。

而對於混合數據集的結果呈現如下：

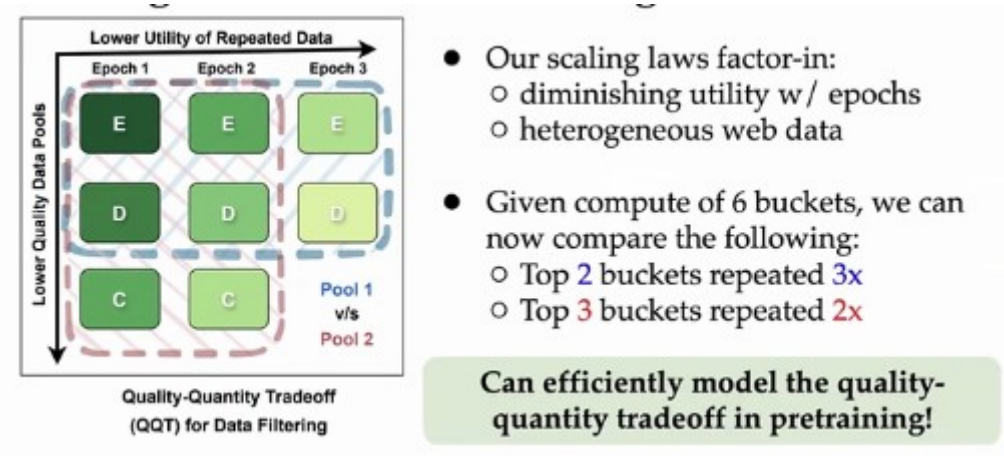


根據不同的算力可以得到以下結果(上圖各資料集的資料量是一樣的，只差在過濾的資料品質，顏色越深表示品質高，能提供較好的效果)：

1.算力小：經過過濾品質高的數據集有較小的損失值，可以有較好的訓練結果。

2.算力中等：數據集平均過濾程度中等的表現較佳。

3.算力大：數據集平均過濾程度小結果表現較好。



當要訓練的資料量一致時，可得到上圖的結果：

1.數據集1(Pool 1)：因訓練次數多，即使資料訓練的效果遞減，也可達到訓練目的。

2.數據集2(Pool 2)：訓練次數少，需要較高品質的資料來達到預期的效果。

根據以上研究，講者做出以下結論：

1.數據整理不能忽略算力。

2.講者與他的團隊所提出的縮放法則可以算出不同訓練次數下的資料效用遞減情形。

(二)主題演講2：數據集創作者應該知道的著作權

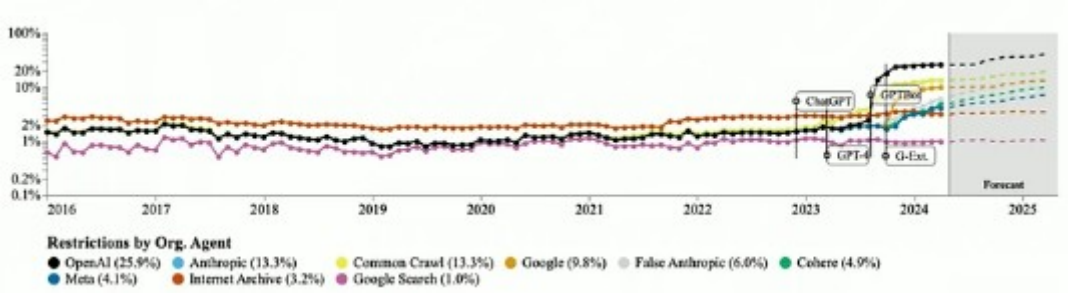
本場演講的講者Stella Biderman本身是EleutherAI的執行董事，也被視為開源大型語言模型運動的領軍人物。她的工作著重在於使整個語言建模流程透明公開，特別是一些常被忽視的組成部分，如訓練數據、訓練庫和評估協議。在她參與的項目中強調了文檔化和公開發布訓練數據的實際和道德必要性。

講者先提出疑問開頭，為什麼我們要在乎數據的著作權？對於數據集來說，著作權是其重要的元數據(Metadata)[24]，在理想情況下，每個數據集中的數據點都應有版權信息的記錄，也因為公開的數據集可能未標示清楚或是使用者未注意，產生侵權行為。而侵權行為可能發生在開發過程的不同階段。這包括獲取訓練數據、處理和操作該數據以及生成的AI 模型。

另一個問題，人們對於自己的數據被使用感到不開心，因此最好的做法就是對資料加上層層限制。這也導致近年來網路資料的授權使用及

24 元數據(Metadata)是描述數據的數據，提供關於這個資料的各種資訊。

相關限制改變，對於AI使用者來說越來越嚴格，特別是針對自動爬取資料的程式，從原本1%~2%的資料會阻擋機器搜尋，到現在平均20%上下的資料會被阻擋。



對於講者來說，因為其致力於開放資料領域的推行，所有他所產出的資料等等都會申請著作權並授權他人使用，但並非每個人都有義務去授權他的資料或是所開發的模型。在缺乏明確的相關法規之下，各資料授權時間點不同，可能因為覺得這數據集有未來前景，就申請著作權加上使用限制等，或是當數據集的所有者發現使用其數據集的AI模型訓練結果不錯時，他就可能較對其提起訴訟，這些狀況都為開發者創造了不確定性和模糊性，這種模糊性可能導致無意的版權侵權和隨後的法律糾紛。

講者接著解釋著作權的相關內容。著作權指的是對於智慧的產物所獨佔的所有控制權，而其規範又因為不同目的、不同地區而有所不同。但主要的目的如下：

1.「複製權」其最主要的核心，讓人可以產出複製品。

2.「衍生品的製作」，提供人們產製與原作相關的其他作品。舉例：超級英雄漫畫授權電影公司製作超級英雄電影。

3.「複製品的傳遞」，讓人們可以使用複製品的權利。例如：作者的書提供給出版社大量印刷並販售。

4.「公開使用」，例如表演或是電影播放等等，泛指在公開場合下使大眾能接觸體驗到的方式，現今來說，放在網路上也是規範之內。

5.「授權」，授權使用以上任一目的。對於人們來說，預設立場是無法使用以上任一種使用目的，除非擁有者授權使用才行。

而對於著作權講者提到相關的使用，而最常發生的是被用來當作起訴的依據。相較於較負面的使用，對於著作權還是有正面的影響：

1.決定誰可以使用資料

2.決定資料可以怎麼被使用

但也有些較為模糊的許可授權規範，如需要合乎道德的使用或是不可觸犯法規等，這些在資料的使用上常常不容易被界定。

接者講者提到，「文本與資料探勘例外」，指的是允許研究人員和組織使用自動化技術分析大量文本或數據，強調不會侵犯著作權。這一例外在學術界被認為是很重要的，只要不商用都沒關係。但講者提醒，這都是學界人士所認為的，實際上還是要看使用狀況決定，依據使用條款的不同，能用的範圍也不同，有些條款也會有誤導的狀況發生，使用前還是諮詢專業的律師較為妥當。

大部分的資料集來自不同來源的組合，而在著作權中針對這類的資料也有具體的規範。在美國的相關法規中有提到「『編輯作品』是指通過收集和組裝已有材料或數據形成的作品，這些材料或數據經過選擇、協調或排列，使得整體作品構成原創的著作權作品。」，表示組合不同資料集所，經不同排列會是過濾挑選產生的資料集，其著作權會是屬於蒐集資料並彙編成新資料集的人。但這裡講者提到，在機器學習領域不該這樣定義，即使資料集授權給大眾使用，但個別資料本身的授權不一

定是合規的，資料的授權使用還是要回歸到基礎數據，也就是所使用的個別資料的授權使用上，這條規定常常被資料集的創作者錯誤的引用。

總結以上內容，講者講述了幾個認為針對授權的方面我們可以做的更好的部分：

1.著作權及授權許可的相關資訊是非常重要的元數據(Metadata)。

2.對於創作者發布的資料，有義務將相關的授權資訊整理並分享告知。

3.對於數據使用者要使用或散布的資料，要勤奮的搜尋其授權資訊。

4.保存好著作權的元數據是法律義務。舉例來說，在美國著作權法中，移除資料上的著作權元數據會被視為是違法的行為。

最後是講者提醒的重點：

1.請多與相關法律專業或是了解政策的人員請益。

2.數據的使用者有義務針對要使用或散布的資料去查詢數據的著作權及授權資訊。

3.針對資料使用越來越嚴謹，現在會是關鍵的時期去弄清楚如何取的資料授權及使用權利。

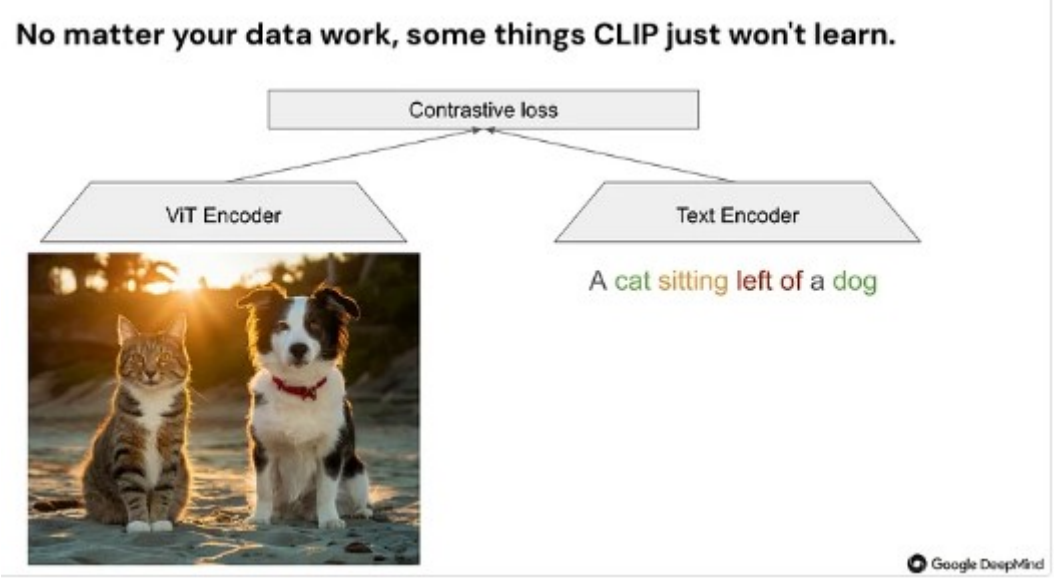
4.讓大家更注重授權可以使機器學習領域的發展變得更好！

(三)主題演講3：大型語言模型的想像：以數據為中心的視角

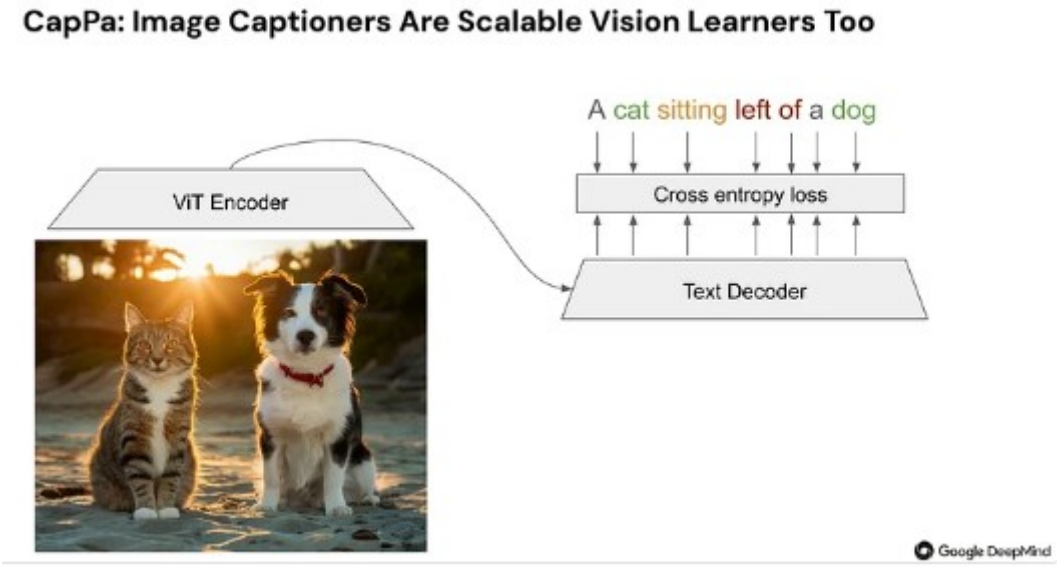
本場的講者Lucas Beyer來自Google DeepMind團隊，目前從事視覺和語言模型的研究。

講者開頭用「語言」比喻做API，視其為串接人們想像的工具。講者比較了過去所使用的圖像數據集以及現在常看到的圖像數據集。過去的數據集多為監督式資料，且分類的用詞較為簡單且單一，如汽車、飛機、鳥等等。但如今的數據集多為來自網路上的資料，文字的描述變得

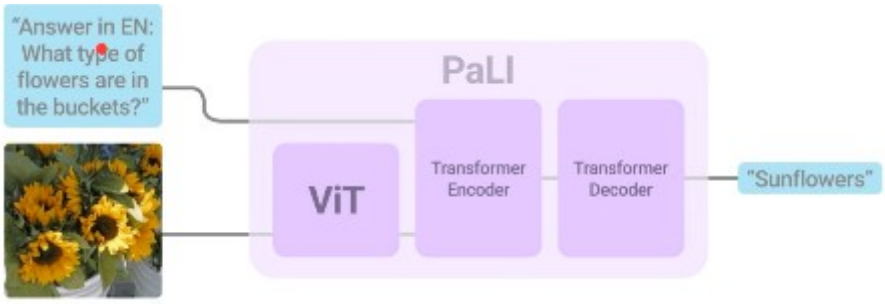
較為詳細，如法蘭克福機場的天際線 -2017等。在這裡講者推薦使用 CLIP 模型，它可以依據使用者輸入的文字進行圖片分類，效果不差。而這也表示，這個領域已從純粹手動設計特徵轉向數據驅動的方法，讓電腦自動產出。



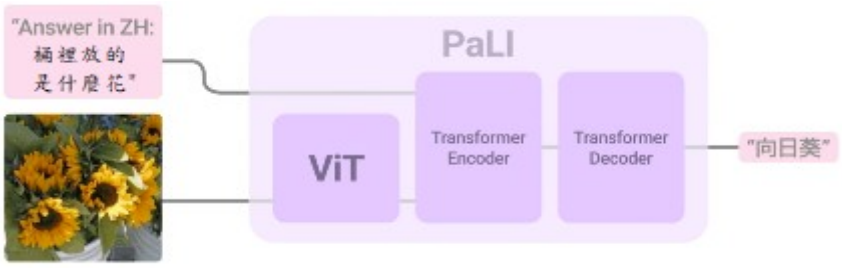
但使用 CLIP 還是有其限制，對於相對位置的的描述它是無法理解的，為了解決這類問題，講者提出使用 Captioner 的方式處理，其概念是讓圖片自行轉文字描述，再與輸入的文字做交叉比對，藉此得到關聯性，而這種方式進行訓練也確實提高其準確性。



接者講者提到，現階段的模型訓練多採用多階段的訓練(Multiple stages of training)，相較於原本的大語言模型，先預訓練一個模型，使用於不同場景可能是，零樣本(Zero shot)[25] 使用或是微調(fine tuning)[26]。講者介紹另一個模型 PaLI，這是一個多語言視覺模型，它的運作模式是把文字問題的輸入以及圖片的輸入，透過Google的圖像分類模型(ViT)將轉換的資訊與輸入的文字作用，產出文字的輸出。模型的架構如下圖：



25 零樣本(Zero shot)是一種機器學習方法，在沒有見過特定任務或類別的情況下，能夠進行預測或分類。
26 微調(fine tuning)指的是在一個已經預訓練好的模型上，利用新的、特定領域的數據進行額外的訓練，以適應新的任務或提升模型在特定任務上的表現。



輸入問題時告知要產生的語言，配合圖片的解析，PaLI 模型會產出相對應的結果，例如用英文和中文詢問圖片中的花是什麼，對於模型來說需要先知道「花」與圖片的關係，找到後判斷其分類，再用指定的語言產出結果。

但如果只有文字的輸出感覺不太夠？講者提出可以使用一些常見的切割或是標記模型，在圖片中框選出相對應文字產出的圖像內容，這可以讓產出內容更豐富。但 PaLi 模型是屬於私人模型，未開放使用。這裡講者介紹了他們另一個開放模型：PaliGemma 。這款模型是有授權供大眾下載，於各種領域使用，但不可使用於違法等不良用途，與前場演講相呼應。與 PaLI 架構上的不同在於 PaliGemma 僅使用語言圖片編碼器與語言解碼器，少了語言編碼器這層架構。

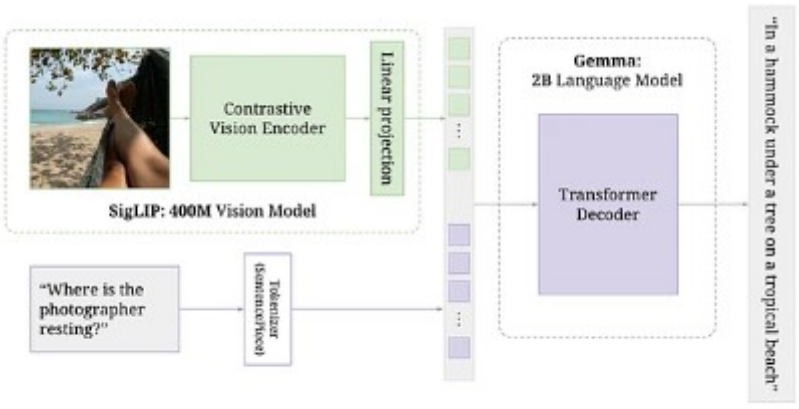


Figure 1 | PaliGemma's architecture: a SigLIP image encoder feeds into a Gemma decoder LM.

接著提到 PaliGemma 的訓練方式。在預訓練裡，採用的是先前提到的多階段的訓練方式，

stage0：單模態訓練(Unimodel pretraining)：圖片編碼器單獨訓練、語言解碼器也獨自訓練，始可使用開放的圖片模型或是語言模型。如果使用開放模型，在這個階段幾乎可以算是零成本的開始，只需要下載即可。

stage1：多模態訓練(Multimodel pretraining)：使用混和的資料在上述架構下訓練，此預訓練耗時較長，且所使用的混合資料需要挑選。

stage2：解析度提升(Resolution increase)：此階段使用短的預訓練去訓練高解析度的圖像。

stage3：轉換(Transfer)：將前述訓練完的基礎模型轉換為特定任務的模型。講者強調這比模型微調所需時間更少，更方便。

最後講者提到透過強化學習(Reinforcement Learning, RL)[27]去做到真正想做的，畢竟監督式的微調學習效果常常無法得到我們所想要的結果。其概念是在我們預訓練完的模型上再加上RL用於模型的微調，以逐漸符合使用的需求。

講者在最後結論提出"Things are converging!"，可以理解為隨著發展很多研究成果勢必要整合彙整起來，以解決未來的問題。

參、心得及建議

本次的 ICML 會議深刻感受到人工智慧與機器學習領域的蓬勃發展。從來自於世界各地的與會人們熱情的討論，到各面向主題的深入探討，都充分展現著 AI 浪潮帶給這世界的衝擊與影響。會議中關於長文本基礎模型、數據導向機器學習研究以及著作權議題的探討，更讓我們對未來 AI 發展趨勢和挑戰有了更清晰的認識。參與會議的心得整理如下。

27 強化學習(Reinforcement Learning, RL)是一種通過獎懲機制訓練 AI 模型在特定環境中做出最佳決策的機器學習方法。

一、長文本基礎模型的挑戰與進展

在本場研討會中了解到對於長文本基礎模型所面臨的挑戰：

計算效率：文本長度增加導致計算複雜度呈二次增長。

數據匱乏：缺乏足夠的長文本數據用於模型訓練。

評估複雜性：評估長文本模型效果困難且成本高昂。

(一)現有評估方法的局限性

目前自動評估指標對於模型評估效果不佳，對於人工評估又太耗費成本，簡單來說評估方法尚待人們的努力。換個角度來看，講者所提出的長文本評估的問題，也可視作模型的使用限制，最主要的限制還是在於模型的理解力方面。

(二)演算法與硬體協同設計

由講者提出透過硬體和演算法協同設計可以解決大型語言模型推理過程中遇到的鍵值緩存瓶頸問題。透過 CPU 和 GPU 的協同工作，有效地提升了長文本模型的推理效率。從中發現的使用限制在於大型語言模型的規模與硬體基礎設備上，如何搭配以取的最好的效果，是未來需要持續研究的課題。

(三)狀態空間模型

講者提出混和模型結合 SSM 與變換器的優勢，利用變換器的機制強化 SSM 的表現，可產生更高效的模型。針對所需使用的任務，挑選適合的模型。

透過這場研討會可以了解到長文本模型最大的痛點還是在於如何讓模型理解內容並進行推論，講者們的研究都著重於在有限的資源下讓模型效果最佳化。

二、數據導向機器學習的重點

(一)數據質量的重要性

數據的質量對於模型訓練效果至關重要。在計算資源有限的情況下，高質量數據集能帶來更好的訓練效果；而當計算資源充足時，混合不同質量數據集反而能提升模型性能。簡單來說，數據整理策略需要考慮到可用的計算資源。

(二)著作權問題

在著作權的問題上，講者強調數據集創作者應重視著作權問題，並在發布數據集時提供清晰的授權資訊。這對於保障數據安全和促進 AI 領域的健康發展至關重要。

(三)大型語言模型的應用與發展趨勢

大型語言模型在圖像理解方面具有巨大潛力。例如 PaliGemma 模型是一個開源的多語言視覺模型，可以用於各種不同的任務，有多模態的產出。而強化學習可以用於模型微調，以提升模型在特定任務上的表現。

在這場研討會最大的收穫莫過於數據著作權的認識，這也是未來在使用或是訓練模型時必然會遇到且越來越受關注的議題。

在這次的 ICML 會議中，可以發現 AI 領域變化相當快速，而講者 Lucas Beyer 所說的 "Things are converging" 可以視為是最能代表這場會議的一句話，如何在資源有限的清況下，透過組合現有模型達到所預期的效果，這將會是未來要不斷面對的課題。而在這場會議中所學習到的知識與概念，可做為未來推動應用 AI 模型的參考。

肆、參考資料

The Forty-First International Conference on Machine Learning
https://icml.cc/Conferences/2024