黃仁勛、小扎巔峰對談:萬字長文揭秘 Meta 的 AI 未來

之三

(接上期)因此,所有的硬件和系統最終都會被優化,以便非常好地運行這項技術,這將使所有人受益,同時它也能很好地兼容我們正在構建的系統。而這一點,我認爲只是展示了(開源)這種方式最終會變得非常有效的一個例子。

扎克伯格:所以,我認爲開源策略作爲商業策略將會是一個好的選擇。我想人們可能還沒完全意識到我們有多么熱愛它。

黃仁勛: 我們圍繞它建立了一個生態系統。我們創造了這個東西。

扎克伯格: 是的, 我發現這一點了。你們團隊的表現一直很出色。每次我們推出新産品, 你們是首批發佈並優化使其運作的團隊。我很感激這一點。

黃仁勛: 我還能說什么呢?我們確實有不少優秀的工程師。

扎克伯格:(笑)而且,你們也總是迅速抓住這些機會。

黃仁勛:所以,我雖是長者,但行動敏捷。這就是 CEO 必須做的。

我認識到了一件重要的事情。我認爲 Llama 確實非常重要。我們圍繞它構建了一個名爲 AI 工廠(AI Foundry)的概念,以便幫助每個人構建 AI。很多人,他們有構建 AI 的願望。對他們來說,擁有 AI 非常重要,因爲一旦他們將其融入數據飛輪,這就是他們公司機構知識被編碼並嵌入 AI 的方式。但他們承擔不起讓那個 AI 飛輪,數據飛輪在其他地方(通過購買服務)轉起來的成本。因此開源允許他們這樣做。但他們並不眞正知道如何將這一切轉化爲 AI。所以我們創建了這個名爲 AI Foundry 的東西。我們提供工具,提供專業知識,Llama 的技術,我們有能力幫助他們將這一切轉化爲 AI 服務。然後當我們完成這一切後,他們接手,他們擁有,其輸出就是我們所說的 NIMM。這個 NIMM,這個神經微型 NVIDIA 推理微服務,他們只需下載,帶走並在任何他們喜歡的地方運行, 包括本地部署。我們擁有一個完整的合作夥伴生態系統,從能夠運行 NIMMs 的 OEM 到像 Accenture 這樣的 GSIs, 我們培訓並與他們合作創建基于 Llama 的 NIMMs 和管道。現在我們正在幫助全球各地的企業實現這一目標。我的意思是,這確實是一件非常令人興奮的事情。這一切實際上都是由 Llama 的開源引發的。

AI 産業的未來:模型不會一家獨大,從小到大的模型都有場景

扎克伯格: 是的,我認爲,幫助人們從大型模型中提煉出自己的模型的能力,將成爲一個眞正有價值的新事物。就像我們在産品方面討論的那樣,至少我不認爲會有一個每個人都會去跟它交流的核心 AI 智能體。在同一水平上,我也不認爲必然會有一個模型是每個人都會使用的。

黃仁勛: 我們有一個芯片 AI, 芯片設計 AI。我們有一個軟件編碼 AI。我們的軟件編碼 AI 理解 USD, 因爲我們用 USD 爲 Omniverse 編寫代碼。我們有一個理解 Verilog 的軟件 AI, 我們的 Verilog。我們擁有能夠理解我們的缺陷數據庫的軟件 AI, 並且知道如何幫助我們分類缺陷並將其發送給正確的工程師。

這些 AI 中的每一個都是基于 Llama 進行微調的。我們會對它們進行微調,並設置防護措施。你知道, 如果我們有一個用于芯片設計的 AI,我們並不希望詢問它關於政治、宗教之類的問題。所以我們會對它進行防護限制。因此,我認爲每家公司基本上都會爲它們擁有的每一個功能,配備專門爲此構建的 AI。他們需要幫助來實現這一點。

扎克伯格:是的,我認爲未來的一大問題是,人們將多大程度上使用更大、更複雜的模型,而不是僅僅針對他們的需求訓練自己的模型。至少我可以肯定,未來會有各種各樣、數量龐大的不同模型涌現。

黃仁勛: 我們使用的是最大型的那些模型。而這樣做的原因在于,我們的工程師時間極其寶貴。因此,我們現在正針對性能優化 405B 版本的 Llama 3.1。如你所知,無論 GPU 多大,405B 都無法完全適配。這就是爲什么 NVLink 的性能如此關鍵。我們採用了這種技術, 通過一個名爲 NVLink 的非阻塞交換機, 將每塊 GPU 連接起來。

例如,在一個 HGX 中,就有兩個這樣的交換機。我們使得所有這些 GPU 能夠協同工作,運行 405B 時性能極爲出色。我們這樣做的原因是,工程師的時間對我們來說極其寶貴。你知道,我們希望使用儘可能最佳的模型。即便這樣做在成本上只節省了幾分錢,誰又會在乎呢?因爲我們希望確保向他們展示的是最優質的結果。

扎克伯格:是的,我的意思是,我認爲 405B 的成本大約是 GPT-4 模型的一半。所以,從這個層面來說,它已經相當不錯了。但我認爲人們正在設備上使用或需要更小型的模型,他們會將其精簡。因此, 這就像是 AI 運行的一整套不同的服務。

黃仁勛: 讓我們假設一下, 我們正在用于芯片設計的 AI 每小時可能只需 10 美元成本。你知道,如果你持續使用它,並且將那個 AI 共享給衆多工程師,那每個工程師可能都有一個成本不高的 AI 陪伴着他們。這個 AI 的成本其實並不高。而我們支付給工程師的薪酬卻很高。

因此,對我們來說,每小時幾美元就能大幅提陞那些尚未接入 AI 的人的能力。立刻行動,接入一個 AI 吧。我們想說的就是這些。

那么,讓我們談談下一波趨勢。你們所做的工作中,我特別喜歡的一點是,計算機視覺,我們內部大量使用的一個模型是 Segment Everything。你知道, 我們現在正在視頻上訓練 AI 模型,以便我們能更好地理解世界模型。我們的應用場景是機器人技術和工業數字化, 將這些 AI 模型接入 Omniverse, 以便我們能更好地建模和表達物理世界。讓機器人在這些 Omniverse 世界中更好地運作。因此你的應用,Ray-Ban Meta 眼鏡,你將 AI 引入虛擬世界的願景眞的很有趣。給我們講講吧。

扎克伯格:好的,嗯,這里面有很多值得探討的內容。你所談論的那個 Segment Everything 模型, 我們實際上在 SIGGRAPH 上展示了它的下一個版本,Segment Everything 2。現在它已經能正常運行了,速度更快了。它現在也能處理視頻了。我覺得這些實際上是我考艾島牧場的牛。順便說一下,它們被稱爲馬克的小牛,美味的小牛。

黃仁勛:美味的馬克的小牛。所以馬克下次來我家,我們得一起做費城奶酪牛排。你直接帶牛過來。

扎克伯格: 然後你做奶酪,我就當副廚。這牛眞的非常好吃。

黃仁勛: 這是副廚的評價。

扎克伯格: 好吧,聽着。然後在晚上結束時,你就像," 嘿,你吃得夠多了,對吧?" 然後我就說:" 不知道,我還能再吃一個。" 你就像:" 眞的嗎?"

黃仁勛: 我肯定像是在說:" 對, 我們還要再做一些。我們還要再做一些。你吃飽了嗎?" 通常你的客人會說:" 哦,是的,我很好。"

扎克伯格: “再給我做一個芝士牛排,老黃。”

黃仁勛: 所以,爲了讓你知道他有多強迫症,我轉過身去,我在準備芝士牛排。然後我說:" 馬克,切一下西紅柿。" 遞給馬克一把刀。

扎克伯格: 對,我是個精確的切割者。

黃仁勛: 然後他就切西紅柿。每一個都切得精確到毫米。但眞正有趣的是,我以爲所有的西紅柿都會被切成片,然後像一副撲克牌那樣疊起來。當我轉過身時,他說他需要另一個盤子。而他這樣做的原因是,他切的每一片西紅柿,彼此之間都不能有接觸。一旦他把一片西紅柿與其他西紅柿分開,它們就不應該再碰到一起。

扎克伯格: 是啊,你看,夥計,如果你希望它們接觸,你應該提前告訴我。你需要...我只是一個副廚,好嗎?

黃仁勛: 這就是爲什么他需要一個不帶偏見的 AI。

扎克伯格: 是的。(笑)

黃仁勛: 這眞的很酷。好的,所以它在識別牛的足迹...它在追踪牛的足迹。

扎克伯格: 用這個可以製作很多有趣的特效。而且因爲它會被廣泛開放,行業內還會有更多嚴肅的應用。所以我的意思是,科學家用這些東西來研究珊瑚礁、自然棲息地以及地形的演變等等。但我的意思是,它能夠在視頻中實現這一點,告訴它你想追踪的內容,你就能擁有 B-roll 鏡頭並能夠與之互動。這眞是非常酷的研究。

黃仁勛: 我舉個例子, 告訴你們我們使用它的場景。比如,你有一個倉庫,里面裝滿了攝像頭。倉庫的 AI 正在監控着所有發生的事情。假設一堆箱子倒塌了, 或者有人在地面上灑了水,或者即將發生任何意外,AI 識別到這一情況,生成文字信息,發送給相關人員,救援就會在路上。這就是使用它的一個方式。不是記錄所有事情,如果有事故發生, 不是記錄每一納秒的視頻,然後再回去檢索那個時刻, 它只記錄重要的內容,因爲它知道自己在看什么。因此,擁有一個視頻理解模型,一個視頻語言模型對於所有這些有趣的應用來說,確實非常強大。那么,你們接下來還會研究些什么呢?Ray,跟我談談...

計算平台的未來:別 XR 了,就 AI+ 潮眼鏡就能賣 10 億個

扎克伯格:是的,還有所有智能眼鏡。我認爲,當我們考慮下一代計算平台時,我們會將其分解爲混合現實(XR)、頭戴設備和智能眼鏡。而智能眼鏡,我認爲人們更容易接受並佩戴它們,因爲現在幾乎每個戴眼鏡的人最終都會陞級爲智能眼鏡。這在全球有超過十億人。所以這將是一個相當大的市場。

VR/MR 頭戴設備,我認爲有些人覺得它們在遊戲或其他用途上很有趣,有些人則還不這么認爲。我的觀點是,它們都將存在于這個世界。

我認爲智能眼鏡將類似于手機,是常駐型計算平台的下一個版本。而混合現實頭顯則更像你的工作站或遊戲主機,當你坐下來進行更沉浸式的體驗,並需要更多計算資源時。眼鏡(大小)只是非常小的形式因素。因爲算力將帶來很多限制,就像你不能在手機上進行(和計算機)同樣級別的計算一樣。

黃仁勛: 它恰好出現在所有這些生成式 AI 突破發生的時候。

扎克伯格: 是的,所以我們基本上,對於智能眼鏡,我們一直從兩個不同的方向來解決這個問題。一方面,我們一直在構建我們認爲的那種理想全息增強現實(AR)眼鏡所需的技術,並且我們正在進行所有定制硅芯片的工作,所有定制顯示堆棧的工作,就像爲了實現這一目標所需做的所有事情。而且它們是眼鏡,對吧?不是頭戴式設備。不像 VR/MR 頭顯。它們看起來就像普通眼鏡。但它們與你現在戴的眼鏡相比,還有相當大的差距。我是說你的眼鏡非常薄。

但即便我們製造的雷朋眼鏡,目前還無法將實現全息 AR 所需的所有技術完全融入其中。我們正在接近,我認爲未來幾年我們會越來越接近。它仍然會相當昂貴,但我們還是會把它推出成一個産品。

我們考慮的另一個角度是,先從外觀好看的智能眼鏡開始。通過與全球頂尖的眼鏡製造商 EssilorLuxottica 合作, 他們基本上涵蓋了你所熟知的所有大牌,比如雷朋、奧克利、奧利弗·皮普爾斯,以及其他少數幾個品牌。這些幾乎都屬於 EssilorLuxottica 旗下。

黃仁勛:(就像)NVIDIA 眼鏡 (那么大牌)。(笑)

扎克伯格: 我想,你知道,他們大概也會喜歡這種比喻。我是說,誰會不喜歡呢?在當下,誰會不想要這樣的眼鏡呢?

扎克伯格: 但我們一直在與他們合作開發雷朋系列,現在已經是第二代了。

我們的目標是,好吧,讓我們將外形限制在看起來非常棒的範圍內。然後在這個框架內儘可能多地融入技術,儘管我們知道在技術上還無法達到我們理想中的完美整合,但最終,它們將會是外觀出色的眼鏡。到目前爲止,我們配備了攝像頭傳感器,因此您可以拍照和録像,您實際上可以實時直播到 Instagram, 您可以在 WhatsApp 上進行視頻通話並實時傳輸給對方, 您知道,您所看到的畫面。它配備了麥克風和揚聲器,我是說,那個揚聲器實際上非常出色。它就像,開放式耳道設計, 因此許多人覺得它比耳塞更舒適。您可以聽音樂,就像擁有一段私密的體驗,這相當不錯。人們很喜歡這一點,您可以在上面接聽電話。

但我們發現, 這些傳感器組合正是與 AI 對話所需的。所以這可以說是個意外的發現。如果五年前你問我, 我們會先實現全息 AR 還是 AI? 我可能會說,大概是全息 AR 吧。對吧,我的意思是,這看起來就像是所有虛擬現實和混合現實技術上的進步,以及構建新的顯示技術棧。我們正朝着這個方向不斷取得進展。然後,大型語言模型(LLMs)取得了突破,結果是我們現在擁有了高質量的人工智能,並且在全息增強現實(AR)出現之前,它的改進速度非常快。

所以這是一種我未曾預料到的轉變。幸運的是,我們處于有利位置,因爲我們一直在研究這些不同的産品,但我認爲最終你會得到一系列不同價格、不同技術水平的潛在眼鏡産品。所以,基于我們現在看到的 Ray-Ban Meta 的情況,價格在 300 美元左右的無顯示屏 AI 眼鏡將成爲一個非常熱門的産品,最終可能會有數千萬甚至數億人使用。屆時, 你將與高度互動的 AI 進行對話。

黃仁勛: 它會有你剛纔展示的視覺語言理解技術。實時翻譯功能。你可以用一種語言與我交談。我聽到的是另一種語言。

扎克伯格: 當然,有顯示屏也會很棒,但這會增加眼鏡的重量,使其價格更高。因此,我認爲很多人會想要那種全息顯示效果,但也有很多人希望最終能擁有類似超薄眼鏡的産品。

黃仁勛 對於工業應用和某些工作場合,我們確實需要(虛擬現實)這樣的技術。

扎克伯格: 對於消費品也是如此。我在疫情期間思考過這個問題, 當時大家都短暫地遠程工作了一段時間。就像你整天都在 Zoom 上,這還算過得去。雖然現在我們有這些工具已經挺棒了, 但未來我們離實現虛擬會議並不遙遠。到那時我並不實際在場, 你看到的只是我的全息影像, 但感覺就像我們眞的在那里, 身體上也在場。我們可以一起工作,共同協作。但我認爲這在與 AI 合作時尤爲重要。

黃仁勛: 我可以接受一個不需要我時刻佩戴的設備。

扎克伯格: 哦,是的,但我認爲我們終將達到那個實際應用的階段。我的意思是,在眼鏡中,有細框和粗框,還有各種風格, 但我認爲我們離擁有全息眼鏡的形式還有一段時間, 不過我認爲在不久的將來, 擁有一副時尚且稍顯粗框的眼鏡並不遙遠。

黃仁勛: 這些(小型智能)太陽鏡現在越來越貼合人臉了。我能看出來。

扎克伯格: 是的,你知道嗎,這是一個非常有用的風格。無論你信不信,但我正努力成爲一名潮男,以便在眼鏡上市前影響這種風格。(笑)

扎克伯格: 我能看出你在嘗試。你的潮男之路進行得怎么樣了?(笑)

黃仁勛: 還早。但我覺得,如果未來業務的一大塊是打造人們佩戴的時尚眼鏡,那我可能應該開始多關注這方面了。

扎克伯格: 沒錯。完全同意。我們得讓那個每天穿同樣衣服的我退休了。(笑)但這就是眼鏡的特別之處。我覺得它不同于手錶或手機,人們眞的不希望都長得一樣。

黃仁勛: 對,所以我認爲,你知道的,這是一個平台, 我覺得它會傾向于我們之前討論的主題,成爲一個開放的生態系統,因爲人們對於形式多樣性和風格的需求將會非常大。並不是每個人都會想要那種你知道的,其他人設計的類似的那種眼鏡,我不認爲這對這次來說會行得通。

扎克伯格: 是的,我認爲你說得對。

結束語:軟件 3.0 時代已經到來

黃仁勛: 馬克,我們正處在一個整個計算堆棧正在被重新發明的時代,這眞是不可思議。我們對軟件的思考方式,你知道安德烈稱之爲軟件 1.0 和軟件 2.0, 而現在我們基本上處于軟件 3.0 的時代了。從通用計算到生成式神經網絡處理方式的轉變,我們能夠開發的計算能力和應用在過去的想象中是不可思議的。這種生成式 AI 技術,我記不清還有哪項技術能以如此快的速度影響消費者、企業、行業和科學界,並且能夠跨越從氣候技術到生物技術再到物理科學等所有不同科學領域。在我們遇到的每一個領域,生成式 AI 都正處于這一根本性轉變的核心。

而且, 除了你提到的生成式 AI 將在社會中産生深遠影響之外, 我非常興奮的一件事是,有人之前問我是否會有一個 "Jensen AI",那正是你所說的創造性 AI,我們可以構建自己的 AI,並加載所有我寫過的內容,通過我回答問題的方式進行微調,希望隨着時間的積累,它能成爲一個眞正出色的助手和夥伴,爲那些只想提問或交流想法的人服務。這個版本的 Jensen AI 不會評判你,你不必擔心被評判, 因此你可以隨時與它互動。我認爲這些都是非常了不起的事情。我們經常寫很多東西,而僅僅給它三四個主題,讓它以我的聲音寫出我想要的基本主題, 並以此爲起點,這是多么不可思議。現在我們可以做的事情實在太多了,與你合作眞的很棒。

我知道建立一家公司並不容易,你將你的公司從桌面轉向移動、VR 再到 AI,所有這些設備,這眞的非常非常不尋常。我們自己也多次轉型,我知道這有多難。多年來,我們倆都經歷了很多挫折,但這就是成爲先驅和創新者所需要的。看着你眞的很棒,同樣,我不確定這是否算轉型,如果你一直在做之前的事情,但你也增加了新的內容,這意味着還有更多的章節等待我們。我認爲對你來說也是如此,看着你們的旅程很有趣。

我的意思是,你們經歷了一個時期,當時每個人都認爲一切都會轉向這些設備,計算會變得非常便宜,但你們一直堅持不懈,就像實際上你們會需要這些能夠並行處理的大型系統一樣。

扎克伯格: 是的, 我們曾經一段時間內製造越來越小的設備,但後來我們讓計算機變得時尚了一段時間,然後變得不那么時尚,甚至超級不時尚,但現在又酷起來了。

黃仁勛: 我們製造圖形芯片 GPU,現在你部署 GPU。Zuck 在他的數據中心里有成千上萬的 H100。我想你們即將達到 600,000 塊 GPU 的機組。

扎克伯格: 我們是優質客戶。這就是如何在大會上獲得 Jensen 問答環節的秘訣。(笑)

當他們說 " 你知道嗎, 幾周後我們在 SIGGRAPH 有個活動 ",我就想," 是啊,我覺得那天我沒什么安排,不在丹佛,聽起來很有趣,我那天下午也沒什么安排 "。

黃仁勛:所以)你就這么出現了,但關鍵是,你們構建的這些系統,它們是巨大的系統,非常難以協調,非常難以運行,你知道你比大多數人晚進入 GPU 領域,但你運營的規模比任何人都大,看着你所做的一切眞是令人難以置信,恭喜你所做的一切,你現在眞是個潮流引領者。**(完)**