113 年

3 月號

AI TRENDS

GOOGLE 在 APS 2024 展示量子計算的重大進展

Google

ELON MUSK 發布全新開源 LLM

xAI Grok

NVIDIA 推出 BLACKWELL，迎來運算新革命

NVIDIA

人工智慧技術月報

Artificial Intelligence Technology Monthly Report

目錄

精選文章

- Google 在 APS 2024 展示量子計算的重大進展 9
- 伊隆·馬斯克的科技企業向公眾揭示Grok的力量 10
- NVIDIA 以 Blackwell 計算平台在生成式 AI 上掀起革命 11

模型技術

- Alida 透過 Amazon Bedrock 的 AI 加強顧客反饋分析 14
- Microsoft Research 發布 Orca-Math：在專為數學設計的小型語言模型方面的重大進展 15
- 關於在人工智慧領域不斷創新的挑戰與OpenAI的使命演進 16
- 介紹Croissant：改變機器學習數據組織遊戲規則的創新 17
- 揭開 AI 在塑造未來中的角色：來自 Microsoft Research 論壇的洞見 18
- Microsoft本週研究焦點：重塑數位及現實世界互動的突破性技術 20
- Google Research 揭開語言模型社會學習的新紀元 21
- Microsoft Research 的突破：增強大型語言模型以更好地理解表格數據 22
- 利用 Amazon Bedrock 與 AWS Step Functions 革新圖像編輯 23
- Google 推出 Derm Foundation 與 Path Foundation，為醫學影像突破開創新紀元 24

- Hugging Face 開展開創性的開源機器人項目 26
- 以 Google 的Chain of Table革新表格推理 27
- 突破性AI技術改變食道癌檢測 28
- Google Research 揭示教導 AI 理解圖表的突破 30
- 革新語言模型：介紹Cappy 31
- Anthropic 發布 Claude 3 Haiku：AI 速度和成本效益上的重大突破 32
- 以 Amazon Bedrock 的生成式 AI 革新程式碼審查 33
- 開闢新視野：Amazon SageMaker JumpStart 中的 Gemma 模型 34
- Google 研究引入 HEAL：向 AI 中的健康平等邁進 35
- 用RAFT革命性地推進語言模型：在語言模型上的一大步 36
- OpenAI 從阿聯酋獲得財務支持以自行開發晶片 37
- 利用 AWS 上的聯邦學習轉變醫療保健 38
- 使用 Google 的 MELON 技術，革命性地改變了 3D 重建 39
- 加速未來：Microsoft Research 揭露 Garnet，一款下一代開源快取儲存系統 40
- 革新頭像互動：Microsoft Research 的洞見 41
- 使用NVIDIA與Amazon SageMaker革命性地推進語言模型部署 42
- 利用 Amazon SageMaker JumpStart 上精調的 Code Llama 模型提升您的編碼技巧 43
- NVIDIA 在路線優化上創下 23 項世界紀錄，樹立新標準 44
- 顛覆道路：NVIDIA 在 2024 年 GTC 推動未來 45
- NVIDIA 為未來資料中心揭開前沿方法 46
- NVIDIA推出新存儲驗證計劃，簡化企業AI部署 47
- NVIDIA及其合作夥伴將量子計算推向新時代 48
- NVIDIA Maxine：革命化視訊會議體驗 49

- 解鎖未來使用者介面的鑰匙：Google 的 ScreenAI 51
- 透過 Microsoft Research 的 AI 輔助方法革新雲端服務監控 52
- NVIDIA的Blackwell架構：生成式AI的巨大飛躍 53
- 用生成式AI革新工業作業：深入剖析AWS的最新創新 54
- Amazon Bedrock 釋放生成式語言模型自我一致性提示的力量 55
- NVIDIA 在創新與工程卓越方面居領導地位 56
- 內容創作的未來：NVIDIA 最新的技術奇蹟 57
- 利用 Google 的 AI 技術全球性地革新洪水預報 58
- Microsoft Research：值得關注的創新技術 59
- 在AI的虛擬領域中：NVIDIA推出新工具及應用程式加速AI開發 61
- 利用深度學習革新分子科學：Microsoft的M-OFDFT模型 62
- NVIDIA以LATTE3D模型革新3D生成技術 63
- AI新紀元的黎明：Transformer模型革命 64

資訊安全

- 用 AWS Nitro Enclaves 革新 AI 交互中的數據隱私保護 66
- 透過 Amazon S3 存取點為 SageMaker 筆記本實現安全跨帳戶數據共享 68

應用

- Amazon SageMaker 為生物醫學應用革命性地微調蛋白質語言模型 70
- VistaPrint透過AI驅動的個性化推薦來革新小型企業的行銷 71
- 革新線上互動：AWS 推出以 AI 驅動的聊天審查工具 72
- OpenAI 與 Le Monde 及 Prisa Media 合作，利用 ChatGPT 改變新聞體驗 73
- 革新粉絲互動：PGA TOUR 的 AI 虛擬助理 74
- 聯邦學習：醫療數據共享與診斷的前進一大步 75
- 顛覆道路：NVIDIA 與夥伴共同塑造車載 AI 的未來 76
- 利用Google的SCIN資料集革命性地推進皮膚科研究 78
- 利用NVIDIA的AIOps生態系統轉型企業IT運營 79
- 創新新創企業透過尖端AI對抗氣候變遷 80
- NVIDIA的合作夥伴網絡獎突顯人工智慧創新 81
- 利用 Google 的 AI 革命性地改變肺癌篩查 82
- 革命性的癌症偵測：引領方向的AI工具 83
- 連結夢想與現實：NVIDIA 對人工智慧未來的展望 84
- 利用 Contentful 和 Amazon Bedrock 革命化您的內容創作 85

服務

- OpenAI以董事會變革和治理增強迎接新時代 87
- OpenAI 歡迎新的遠見者加入其董事會 88
- OpenAI宣布重大董事會改組及新治理結構 89
- 利用 AWS 解鎖生成式 AI 的力量：簡化指南 90
- AWS 推出的 AI 驅動訂單處理代理人：顛覆客戶服務的革命 92

01 精選文章

Google 在 APS 2024 展示量子計算的重大進展

量子計算 APS 2024 Google Quantum AI 量子錯誤糾正 量子模擬 Qualtran XPRIZE/Google Quantum AI 比賽

2024-03-04

在 2024 年美國物理學會 (American Physical Society) 3 月會議於 Minneapolis 舉行的激動人心更新中，Google 的 Quantum AI 團隊走到了前台，展示了他們在推進量子計算技術方面的開創性努力。這次會議為物理學家和科技愛好者提供了一個理想的交流平台，Google 在此展示了其創新的進展。

Google 在會議上的廣泛參與包括超過 50 場演講，涵蓋了從量子錯誤糾正到量子模擬動態等多個主題。值得關注的是，關於「Crumble」的演示，這是一個開創性的互動工具，旨在視覺化展示量子錯誤糾正 (QEC) 電路，讓複雜的量子概念更加易於理解。此外，Google 公開的開源庫「Qualtran」，承諾將革命性地估算容錯算法的資源需求，標誌著量子計算研究向前邁進了重要一步。

另一個引人注目的時刻是宣布了 500 萬美元的 XPRIZE/Google Quantum AI 比賽，旨在加速量子技術的應用。這一舉措突顯了 Google 不僅致力於推進量子計算，也致力於培育該領域的創新者社群的承諾。

隨著量子計算繼續突破新地界，Google 在 APS 2024 的積極參與和貢獻展示了他們在探索量子前沿方面的領導力和奉獻。從促進複雜計算到揭開量子領域的神秘面紗，Google 的 Quantum AI 團隊走在前列，為一個量子計算改變我們跨行業解決問題方法的未來鋪平了道路。

請繼續關注 Google 和量子計算如何繼續演進，承諾帶來技術進步及更多領域的激動人心的可能性。

閱讀更多

伊隆·馬斯克的科技企業向公眾揭示Grok的力量

伊隆·馬斯克 AI Grok 開源 語言模型 Apache 2.0許可證 X Premium+ ChatGPT 言論自由 GitHub

2024-03-18

對全球的AI愛好者和開發者來說，伊隆·馬斯克的xAI採取了重大步驟，將其先進的語言模型Grok作為開源軟體公之於眾，這一決定令人振奮。擁有驚人的3140億參數，Grok以其深入的語言理解能力脫穎而出，為AI領域設定了新的標桿。

這一開創性的決定允許任何有網路接入的人下載、修改，甚至在用戶友好的Apache 2.0許可證下分發Grok。這一決策凸顯了馬斯克加速AI發展和促進該領域創新的承諾。

Grok不僅僅是另一個AI模型；它是一個專家混合模型，意味著它結合了眾多專業AI模型的智慧，以提供更準確、更細膩的回應。直到現在，Grok是馬斯克社交網路X的獨家功能，僅對X Premium+服務的訂閱者開放。通過將Grok開源，xAI使得對尖端AI技術的訪問民主化，使創意人士得以在任何地方實驗和創新。

發布包括模型的權重和網絡架構，為開發者提供了堅實的基礎。然而，值得注意的是，原始訓練數據和實時數據流訪問，這些可能給專有版本帶來優勢的內容並不包括在內。

以道格拉斯·亞當斯的《銀河系漫遊指南》系列命名的Grok，被設計為一個更開放、也許更具幽默感的對手，與OpenAI的ChatGPT等其他模型形成對比。這一舉動也反映了馬斯克對言論自由的立場，以及他對AI中審查和「覺醒」意識形態的批評。

雖然開源社區對這一舉措表示讚賞，但它也引發了關於賦予對強大AI工具的不受限制訪問的潛在風險和倫理考量的討論。

對於那些對Grok的可能性感到好奇的人來說，它現在已在GitHub上可供探索和創新。

這一來自xAI的開源姿態不僅承諾豐富AI開發景觀，也邀請了更廣泛的對話，關於在人工智能時代創新與責任之間的平衡。

閱讀更多

NVIDIA 以 Blackwell 計算平台在生成式 AI 上掀起革命

NVIDIA Blackwell 生成式AI 計算平台 AI模型 實時生成 電腦晶片 Tensor Core GPU Grace CPU Omniverse Cloud API NIM 微服務

2024-03-18

在一項突破性的公告中，NVIDIA 的 CEO Jensen Huang 向世界介紹了 Blackwell，這是一個即將為生成式人工智能（AI）領域帶來巨大推動的計算平台。Blackwell 處於 NVIDIA 雄心勃勃的願景核心，旨在處理萬億參數 AI 模型的巨大需求，開啟了實時生成式 AI 能力的新時代，這種能力有望從醫療保健到機器人技術等行業進行轉型。

Blackwell 的力量：計算效率的飛躍

想像一下，一種先進的電腦晶片，它能顯著增強 AI 的潛力，讓最複雜的任務看起來變得輕而易舉。那就是 Blackwell。在性能上遠超其前輩，它在訓練能力上帶來了 2.5 倍的提升，並在推理任務上實現了驚人的 5 倍改進。但魔法並未止步。Blackwell 平台是先進技術的交響曲，包括一個將 NVIDIA 的 B200 Tensor Core GPU 與 Grace CPU 配對的超級晶片，為 AI 操作創建了一個強大的動力。

機會的宇宙：NVIDIA 對 AI 驅動未來的願景

NVIDIA 不僅僅是關於構建更快的晶片。它是關於創建 AI 能夠茁壯成長的生態系統。Omniverse Cloud API 和 NIM 微服務的揭幕是 NVIDIA 致力於將 AI 帶入日常應用更近一步的證明。這些工具將重新定義我們與 AI 的互動方式，使其更加易於訪問並融入我們的數字生活。

生成式 AI 的黎明：轉型行業

NVIDIA 的 Blackwell 平台不僅僅是一項技術奇觀；它是各個行業創新的催化劑。從設計更智能的醫療保健解決方案到創造更高效的製造過程，Blackwell 的增強計算能力將加速生成式 AI 革命。隨著技術巨頭的認可和全球雲服務提供商的採用，Blackwell 有望成為 NVIDIA 歷來最成功的推出。

用 NIM 革新軟件創建

在軟件開發領域，NVIDIA 引入的 NIM 微服務代表了一種範式轉變。通過組建 AI 模型團隊，公司現在可以以前所未有的效率和創造力處理複雜任務。這種對軟件創建的新方法承諾解鎖新穎的應用和服務，進一步擴大 AI 對我們生活的影響。

結論：NVIDIA 在 AI 創新前沿

NVIDIA 在 GTC 大會上的最新公告凸顯了其在塑造生成式 AI 未來方面的領導角色。有了 Blackwell 平台，NVIDIA 不僅僅是在計算性能上設置了新的標準；它正在為一個 AI 的變革潛力得以完全實現的世界鋪平道路。當我們站在這個新時代的門檻上，NVIDIA 的願景讓我們一窺生成式 AI 將以我們僅能開始想像的方式塑造我們世界的未來。

閱讀更多

02 模型技術

Alida 透過 Amazon Bedrock 的 AI 加強顧客反饋分析

Alida Amazon Bedrock 顧客回饋分析 自然語言處理 NLP 生成式 AI 情感分析 顧客經驗
2024-03-04

Alida 透過 Amazon Bedrock 的 AI 強化顧客回饋分析

Alida 與 Amazon Bedrock 合作，透過創新的方式重新定義了對顧客回饋的理解，將開放式調查回應的分析精準度提高了 4 至 6 倍。這一重大的技術進步使 Alida 能夠比以往任何時候都更快地為品牌提供關於顧客經驗和需求的更豐富洞察。

傳統的自然語言處理（NLP）方法經常僅僅觸及表面，錯過了顧客表達的複雜、細膩情感。識別到這一差距，Alida 轉向 Amazon Bedrock 和 Anthropic 的 Claude Instant 模型。通過這樣做，他們成功地深入顧客回饋的核心，以前所未有的準確度識別情感和主題，這是傳統 NLP 解決方案無法達到的。

Amazon Bedrock 簡化了使用生成式 AI 的過程，通過單一介面提供一系列高效能的基礎模型。這意味著 Alida 能夠快速推進其創新服務的開發，而無需面對通常與機器學習部署相關的複雜性。

這項新服務的一個關鍵特點是它使用的「提示鏈接」策略，這一策略將分析分解成較小、邏輯性的步驟，允許在每個階段進行更細緻的檢查和微調。這種細緻的方法與傳統 NLP 的廣泛筆觸形成鮮明對比，提供的見解不僅相關，而且深刻響應顧客反饋的質性細節。

Alida 與 Amazon Bedrock 的合作不僅僅是技術升級，它是通往顧客洞察新時代的大門，這個時代的理解深度與數據的規模相匹配。對於那些尋求從顧客反饋中獲得真正、可行見解的公司來說，Alida 的新服務代表了一個巨大的進步。

閱讀更多

Microsoft Research 發布 Orca-Math：在專為數學設計的小型語言模型方面的重大進展

Orca-Math Microsoft Research 小型語言模型 數學問題解決 教育應用 人工智能 訓練計劃 AutoGen 數據合成 迭代學習

2024-03-05

在增強教育應用中人工智能方面，Microsoft Research 已經引入 Orca-Math，這是一個旨在解決小學數學問題的模型，具有前所未有的準確性。Orca-Math 在其前身 Orca 和 Orca 2 所奠定的基礎上建立，這兩個模型專注於提煉訓練信號和小型語言模型（SLMs）的方法，Orca-Math 通過展示模型專業化的力量而脫穎而出。

Orca-Math 是一個簡化的模型，擁有 70 億個參數，已從 Mistral 7B 模型中進行了微調，以在 GSM8K pass@1 上達到驚人的 86.81%。這一表現不僅超過了包括一般和專門數學模型在內的大得多的模型，還從基礎 Mistral-7B 模型在同一基準上的 37.83% 大幅跳躍。這一成就展示了 Orca-Math 在解決複雜的小學數學問題方面的卓越能力，這在歷史上一直是 SLMs 難以克服的挑戰。

Orca-Math 的最先進性能背後的秘密在於其訓練計劃，其中包括使用多代理系統（AutoGen）生成的 200,000 個數學問題的高質量合成數據。這種創新方法不僅允許更快和更具成本效益的訓練，還引入了一個迭代學習過程。在這裡，模型被訓練，允許練習解決問題，然後根據反饋進行改進，模仿真實世界的學習情境。

這項研究闡明了小型模型在專門應用中的潛力。通過使用緊湊的數據集並且不依賴外部輔助或組合就達到如此高的準確度，Orca-Math 代表了在教育設置中語言模型的效率和能力方面的重大進步。

Microsoft Research 正在公開數據集和訓練程序的詳細報告，鼓勵在該領域進一步探索和改進。這項倡議不僅突顯了 SLMs 在專門角色中的實力，還為持續學習和模型增強開辟了新途徑，為未來 AI 驅動的教育工具設立了新的基準。

請繼續關注這項突破性技術及其對教育中 AI 未來的影響的更多更新。

閱讀更多

關於在人工智慧領域不斷創新的挑戰與OpenAI的使命演進

人工智慧 OpenAI Elon Musk 通用智能 GPT-4 技術創新 資金籌集

2024-03-05

在一段顯著的旅程中，凸顯了在人工智慧（AI）領域推動創新所面臨的複雜性和挑戰，OpenAI 分享了其使命及發展過程中的洞見——特別是與科技願景家Elon Musk的關係。起初，Musk 提議的 10 億美元資金承諾，為OpenAI 設定了一個大膽的步伐，旨在將其定位為在開發先進的通用智能（AGI）技術競賽中的強大競爭者。然而，隨著對AGI 的追求不斷增加資源需求，夥伴關係發生了轉變。實現AGI 將需要數十億規模的資金這一認識，促使了將OpenAI 轉型為盈利實體的討論。這一轉變對於籌集所需資源至關重要，但也引入了與Musk 在願景上的分歧。

Musk 從OpenAI 的離開及其後決定通過Tesla 獨立追求AGI 開發的決定，凸顯了科技世界中雄心、原則與實際之間錯綜復雜的關係。OpenAI 對其使命的承諾導致了像GPT-4 這樣的技術創造，使AI 工具變得更加易於獲取且對社會有益。從增強像阿爾巴尼亞這樣的國家的歐盟加盟過程到簡化手術同意書，OpenAI 的貢獻涵蓋了各個領域，展示了AI 改善日常生活的強大能力。

這一展開的故事不僅突顯了在AI 領域開創先河的挑戰，也強調了在追求可能造福人類的技術突破時，願景一致性、資金和策略的重要性。

閱讀更多

介紹Croissant：改變機器學習數據組織遊戲規則的創新

Croissant 機器學習 數據組織 元數據格式 schema.org 資料集準備 負責任的AI 數據生命周期管理 ML安全性

2024-03-06

介紹Croissant：機器學習數據組織的遊戲規則改變者

在一項激動人心的開發中，Google Research Blog揭露了「Croissant」，一種突破性的元數據格式，旨在簡化機器學習(ML)資料集準備的方式。這項創新旨在解決ML領域中一個持續存在的挑戰：理解和組織各種格式和結構的數據集這一繁瑣任務。

傳統上，ML從業者必須穿越數據表示的迷宮，從文本和結構化數據到圖像、音頻和視頻。每個數據集都有其獨特的排列，使準備過程耗時且經常阻礙進展。Croissant作為效率的領航者，為描述和組織準備就緒的ML數據集提供了一種標準化方法，認識到這一瓶頸。

Croissant建立在schema.org的基礎上，這是一種在網絡上廣泛使用的結構化數據標準，通過整合為ML量身定做的層面，如數據資源、組織和默認ML語義，來加以增強。這不僅促進了更容易的數據集發現和工具開發，還簡化了ML框架的數據準備步驟。

Croissant的發布得到了主要ML數據集存儲庫（如Kaggle, Hugging Face和OpenML）和流行ML框架（包括TensorFlow、PyTorch和JAX）的支持。這種廣泛的采用強調了Croissant大大減少數據開發負擔的潛力，特別是對資源有限的研究和初創公司。

此外，Croissant致力於支持負責任的AI (RAI)，通過引入涵蓋關鍵RAI考慮因素的詞彙擴展，如數據生命周期管理和ML安全性。

有了Croissant，找到、精煉和訓練正確的數據集變成了一個更加流暢的過程，為ML研究和開發的進步鋪平了道路。Google鼓勵社區加入完善和擴展Croissant的能力，承諾一個數據挑戰在通往AI創新之旅中成為過去的未來。

閱讀更多

揭開 AI 在塑造未來中的角色：來自 Microsoft Research 論壇的洞見

人工智慧 Microsoft Research 醫療保健 科學發現 材料發現 生成模型 AutoGen 元認知 語言模型 癌症病理學 社會影響

2024-03-06

在最近一集的 Microsoft Research 論壇中，焦點放在人工智慧 (AI) 如何革命性地改變醫療保健、自然科學以及其廣泛的社會影響。這次聚會是一個思想的大熔爐，探索 AI 技術的變革力量。

用 AI 革新科學發現 Chris Bishop，領導著 Microsoft Research AI4Science，分享了使用 AI 模擬和預測自然現象的突破。Bishop 強調 AI 在科學發現中的至高無上的角色，突顯其潛力徹底改變我們對世界的理解。

AI 在醫療保健中：資料利用的新時代 由 Tristan Naumann 和其他領袖主導的討論，深入探討了 AI，特別是生成式 AI 和大型語言模型，如何開啟醫療保健數據的全部潛力。這項進步使得分析更加有結構且高效，為創新的醫療保健解決方案鋪平了道路。

材料發現與 AI 的影響 Tian Xie 和同僚展示了 AI 對傳統材料發現過程的顛覆性影響。通過整合機器學習，該領域正在見證加速進步，增強了我們在材料設計方面的創新能力。

擁抱生成模型以實現科學進步 Rianne van den Berg 介紹了擴散模型和基於分數的生成模型這一有前景的領域。這些模型以其在高解析度圖像中的成功而聞名，現在正被量身定制用於科學發現，標誌著研究方法論的重大飛躍。

AutoGen：塑造 AI 應用的未來 Chi Wang 介紹了 AutoGen，一個為下一代 AI 應用設計的多代理框架。這個創新框架已經展示出卓越的性能，展示了 AI 在有效處理複雜任務方面的潛力。

元認知與生成式 AI：增強可用性 Lev Tankelevitch 對元認知的見解揭示了使用生成式 AI 系統的認知需求。這一視角對於設計增強人類工作流程和決策過程的 AI 至關重要。

未來之路：模塊化語言模型 Alessandro Sordoni 分享了創建一個多功能的專家語言模型庫的進展。這種方法旨在提高 AI 系統對各種任務的適應性，預示著 AI 應用中定制化和效率的新時代。

GigaPath：用 AI 革命化癌症病理學 Naoto Usuyama 介紹了 GigaPath，一項利用大型視覺變壓器分析千萬像素病理圖像的先鋒計劃。這個項目強調了 AI 在轉變癌症護理和研究方面的潛力，利用實際病人數據獲得突破性見解。

導航數字公共領域：AI 的社會影響 Madeleine Daepp 和 Vanessa Gathecha 強調了生成式 AI 呈現的挑戰與機會，特別是在全球選舉和假訊息的背景下。他們的工作強調需要通過協作方式負責任地利用 AI 的潛力。

Microsoft Research 論壇第二集展示了 AI 在各個領域承諾的光明未來。從醫療保健和材料科學到社會的基本組織，AI 的角色變得越來越關鍵，為探索和創新開啟新的前沿。

閱讀更多

Microsoft本週研究焦點：重塑數位及現實世界互動的突破性技術

生成式萬花筒網絡 深度學習 多模態數據 Holoportation 虛擬會議 醫療保健 機器學習 Text Diffusion with Reinforced Conditioning 文本生成 對話AI

2024-03-06

在這週的Microsoft研究焦點中，我們深入探討了一些真正突破性的技術，這些技術承諾將重塑我們與數位及現實世界環境互動的方式。

首先，我們探索了生成式萬花筒網絡的領域。這些網絡代表了神經網絡（深度學習背後的大腦）解讀數據方式的一大飛躍。通過一個有趣的「過度概括」現象，這些網絡可以將多個輸入映射到單一輸出上，顯著增強了數據處理能力。想像一下轉動萬花筒，看著不同的圖案匯聚成一個穩定、可識別的圖像。這就是生成式萬花筒網絡背後的原理，這些網絡在轉變我們處理多模態數據（包括圖像和聲音）的方式方面展現了顯著的潛力。

在電信領域，Microsoft引入了Holoportation™，這是一項創新技術，承諾將革命性地改變虛擬會議。這項技術允許更具沉浸感的體驗，使遠程互動感覺就像每個人都在同一個房間裡。它在醫療保健中特別有影響力，醫生和患者可以以以前不可能的方式進行互動，克服地理障礙，增強護理的提供。

最後，我們關注了機器學習中一項名為Text Diffusion with Reinforced Conditioning (TREC)的新進展。文本生成一直是一個具有挑戰性的領域，因為語言的複雜性。TREC通過一個獨特的模型解決了這個問題，該模型反覆精煉文本生成，承諾更加連貫和高質量的輸出。這一發展可能顯著改善從自動內容創建到更細膩複雜的對話AI的應用。

這些來自Microsoft的技術不僅在各自的領域推動了可能性的界限，而且還為人機互動變得更加直觀和有意義的未來提供了一瞥。

閱讀更多

Google Research 揭開語言模型社會學習的新紀元

社會學習 大型語言模型 隱私保護 合成範例 秘密分享者 教學方法

2024-03-07

在一項開創性的探索中，Google Research 引入了大型語言模型（LLMs）之間的社會學習概念，標誌著朝向更有效率及更私密的機器學習方式邁出了重要的一步。受到人類社會學習的啟發，即知識透過觀察和指導進行轉移，Google 的團隊開發了一個框架，讓 LLMs 能夠使用自然語言相互教學和學習，而不共享敏感資訊。

想像 LLMs 如同在虛擬教室中的學生和老師。一個在特定任務（如垃圾郵件檢測或解決數學問題）上表現出色的老師 LLM，可以通過提供指導或範例的方式，引導學生 LLM 提升其技能，這種方式仿照了人類的教學方法。這個過程，被稱為社會學習，對於提升模型的表現而不冒露出私人資料的風險充滿了希望。

為了解決隱私顧慮，Google 團隊設計了一種方法，其中教師模型產生「合成範例」 - 新的、教育性的範例，與原始資料足夠不同以保護隱私，但又足夠相似以便於學習。這種創新方法確保了敏感資訊的保密性，同時仍對 LLMs 的集體智慧做出貢獻。

該團隊的研究還引入了「合成指導」的概念，教師為任務制定口頭指示，推動了 LLMs 如何從有限的資料中更有效地學習的界限。他們的發現表明，這種方法可以達到與傳統少數樣本學習技術相當的準確度。

此外，研究還通過「秘密分享者」方法，仔細處理了潛在的隱私風險，確保社會學習過程最小化資料記憶和洩漏。這種細膩的方法再次肯定了 Google 對隱私的承諾，同時推進了 LLMs 的能力。

隨著 Google Research 展望未來的發展，社會學習在提升機器智能方面的潛在應用範圍廣泛，從改善垃圾郵件過濾器到推進教育工具。這項倡議為更具合作性、效率性和安全性的人工智能領域的進步鋪平了道路。

承認團隊的集體努力，這項研究不僅凸顯了 LLMs 模仿和超越人類學習方法的能力，還為注重隱私的機器學習設定了新的標準。

閱讀更多

Microsoft Research 的突破：增強大型語言模型以更好地理解表格數據

Microsoft Research 大型語言模型 表格數據 結構化數據 LLMs 結構理解能力 SUC HTML CSV 自我增強提示

2024-03-07

Microsoft Research 的突破：增強大型語言模型以更好地理解表格數據

在一個數據為王的數位時代，表格如同皇冠上的寶石，以結構化、易於消化的格式組織著大量信息。認識到表格所扮演的關鍵角色，Microsoft Research 已踏上了一段激動人心的旅程，旨在彌合大型語言模型 (LLMs) 與對表格數據理解之間的差距。

在著名的第17屆 ACM 國際網絡搜索與數據挖掘會議 (WSDM 2024) 上發表，Microsoft 的研究論文名為「表格遇上 LLM：大型語言模型能理解結構化表格數據嗎？一項基準和實證研究」，深入探討了 LLMs 與結構化數據之間的互動。這項開創性研究引入了一個新的基準，稱為結構理解能力 (SUC)，旨在評估和提升 LLMs 解讀和利用表格格式（從 HTML 到 CSV）的能力。

通過細緻的實驗，Microsoft Research 團隊發現了引人注目的見解。令人驚訝的是，當測試 LLMs 時，HTML 格式比分隔符分隔的格式（如 CSV）表現得更好，說明了表格理解的複雜性。這項研究還探討了各種提示技巧，包括自我增強提示，這是一種模型不可知的方法，利用 LLMs 固有的知識來更好地掌握結構化數據。

這項開創性研究不僅揭示了 LLMs 在理解表格方面的當前局限，還為未來的探索勾畫了一條路線。團隊建議，結構信息的整合和可能使用外部工具，可以顯著推進 LLMs 的能力，為遠遠超出傳統基於文本任務的應用開啟新門。

隨著我們站在這一技術進化的臨界點，Microsoft Research 的發現有望革新我們與結構化數據的互動和利用方式，使其對機器和人類都更加可訪問和有意義。

閱讀更多

利用 Amazon Bedrock 與 AWS Step Functions 革新圖像編輯

Amazon Bedrock **AWS Step Functions** 圖像編輯 自動化 **Amazon Titan Image Generator G1** 生成式 **AI** 圖像處理技術

2024-03-07

利用 Amazon Bedrock 和 AWS Step Functions 革新圖像編輯

在創新的一大步中，Amazon 推出了一項解決方案，大幅簡化了更改圖像背景的過程，使創意廣告、電子商務和媒體等行業的相關工作變得輕而易舉。手動編輯，一個以前耗時的任務，現在得益於 Amazon Bedrock 和 AWS Step Functions 而變得自動化且可擴展。

這場革命的核心在於 Amazon Bedrock 的一部分 Amazon Titan Image Generator G1 模型。該模型利用一種稱為外繪製的方法來自動更改任何圖像的背景。當與 AWS Step Functions 結合使用時，這些技術創建了一個無縫的工作流程，自動化地跨多個圖像更改背景的過程，效率高且輕鬆。

這個解決方案的特點是其能夠快速處理大批量的圖像，將本可花費數小時的任務轉化為只需幾分鐘即可完成的任務。用戶僅需通過網絡應用程序上傳圖像，通過文字提示指定他們想要的背景，然後讓系統完成剩下的工作。過程包括檢測圖像標籤，更新數據庫中的圖像細節，以及調用 Amazon Titan Image Generator G1 模型來生成具有新背景的圖像。

這種自動化不僅提高了生產力，還通過允許快速嘗試不同的背景，開啟了新的創意可能性。此外，該系統設計用於並行處理任務，確保無論圖像量多大都能高效處理。

對於有興趣實施這一尖端解決方案的人來說，GitHub 上提供了詳細的操作指南。這種方法不僅代表了圖像處理技術的重大進步，也展示了生成式 AI 如何應用於實際的日常任務，重新定義了數位領域中創意和效率的極限。

閱讀更多

Google 推出 Derm Foundation 與 Path Foundation，為醫學影像突破開創新紀元

Google Derm Foundation Path Foundation 機器學習 深度學習 醫學影像 皮膚病學 病理學 診斷模型 數字向量 嵌入

2024-03-08

Google 推出 Derm Foundation 和 Path Foundation，為醫學影像突破鋪路

為了解決全球醫學影像專業知識的短缺，Google 開發了兩項革命性工具，Derm Foundation 和 Path Foundation，旨在透過機器學習（ML）改變皮膚病學和病理學的格局。這些工具在使用 ML 解讀醫學影像的準確度與效率上，標誌著一個重要的飛躍。

醫學影像是診斷和治療眾多疾病的關鍵工具，往往因高品質數據、ML 專業知識和計算能力的稀缺而面臨瓶頸。Google 的 Derm Foundation 和 Path Foundation 採用深度學習（DL）將圖像轉換為數字向量或「嵌入」。這些嵌入捕捉醫學影像的基本特徵，大幅減少了訓練有效模型所需的資源。

想像一位經驗豐富的吉他手，憑藉豐富的技能和理解，可以憑耳朵學習新歌。同樣地，這些嵌入工具能迅速識別醫學影像中的模式，使診斷模型的快速開發成為可能。這種方法類似於將複雜的交響樂轉換為簡單、可識別的旋律，使先進的診斷更容易獲取且效率更高。

Path Foundation 特別擅長處理病理學影像，例如用於癌症診斷的苯胺藍和伊紅染色幻燈片影像。它通過從這些影像的獨特複雜性（包括它們的巨大大小和組織結構的細微差別）中學習，超越傳統模型。這產生的嵌入可以顯著提高跨多種組織類型和疾病的診斷模型的表現。

另一方面，Derm Foundation 在皮膚病學領域大放異彩。它利用針對皮膚病學的特定預訓練來快速理解和分類皮膚病況。即使是在數據集較小的情況下，研究人員也可以利用這些嵌入來構建用於皮膚病診斷的強大模型，展示了這一工具革命性地改變皮膚病學研究和治療的潛力。

通過向研究社群提供這些工具，Google 旨在促進新診斷模型的開發，提高皮膚病學和病理學醫學影像診斷的質量和效率。這些技術承諾解鎖眾多應用，從提高診斷準確性到發現疾病的新生物標記。

當我們站在醫學影像這個新時代的門檻上時，Google 的 Derm Foundation 和 Path Foundation 設定了全球研究人員和臨床醫生的賦能，為更快、更精確的診斷鋪路，最終，實現更好的患者結果。

閱讀更多

Hugging Face 開展開創性的開源機器人項目

開源 機器人 人工智慧 AI Hugging Face 技術民主化 全球科技社群 創新

2024-03-08

在一次大膽的舉動中，將軟體與可觸及的創新緊密結合，Hugging Face 已揭露其進軍機器人領域的計畫，這是一個開源項目，承諾將重塑機器人和人工智慧（AI）的未來。這項計畫由才華橫溢的 Remi Cadene 領銜，他之前是 Tesla 智囊團的一員，對於一家主要以其對機器學習軟體領域的貢獻而聞名的公司來說，這是一次重大的轉向。

這家位於巴黎的企業不僅僅是關於創造機器人；它是關於機器人的民主化。Hugging Face 的宗旨是設計、建造並維護可接入的機器人系統，這些系統無縫集成 AI 技術，Hugging Face 的使命是讓每個人都能使用先進的機器人。該項目熱衷於使用低成本材料和現成的零組件，讓創新更容易超越研究實驗室的限制，進入真實世界。

這個項目的特別之處在於其對開源原則的承諾。在一個科技經常被專有秘密所包圍的時代，Hugging Face 把牌放在桌上，邀請來自全球的工程師、開發者和愛好者共同參與、學習和創新。這種協作精神可能導致機器人技術的快速進步，由全球科技社群的集體智慧所推動。

該項目目前正在招募狂熱的機器人工程師，準備探索具體化 AI 和機器人學內的未知領域。這項努力不僅凸顯了 Hugging Face 對開源項目的承諾，也預示著一個 AI 和機器人在我們日常生活中更加緊密結合的未來。

通過這個雄心勃勃的項目，Hugging Face 不僅僅是在創造機器人；它正在激勵新一代技術人員夢想大膽，建造更大。創新的潛力是無限的，對從醫療保健到製造業等行業的影響可能是深遠的。敬請關注，隨著這個開源機器人項目的展開，為 AI 和機器人的激動人心的交匯處開闢新的航道。

閱讀更多

以 Google 的Chain of Table革新表格推理

Google Chain of Table 數據分析 自然語言處理 大型語言模型 LLMs PaLM 2 GPT 3.5

2024-03-11

以 Google 的Chain of Table革新表格推理在數據分析和自然語言處理的領域中，表格是組織和解釋複雜資訊的骨幹。面對大型語言模型（LLMs）在理解表格的結構性質上所遇到的挑戰，Google Research 提出了一種突破性的方法：Chain of Table。這個創新框架旨在通過訓練 LLMs 逐步精煉表格，使其更貼近特定的問題或問題，從而增強表格理解。通過將表格分解為更簡單、可管理的段落，LLM 能夠深入了解每一部分，揭示更深層次的理解和分析。這種方法之所以脫穎而出，是因為它在該領域創造了顯著的進步，樹立了新的標杆。鏈式表格的核心操作是引導 LLMs 通過一系列操作，逐步轉換原始表格。每一項操作都基於上下文學習選擇，反映了一個動態規劃過程，這個過程模仿了人類解決問題的方式。這導致了一系列中間表格，每一個都代表了推理過程的一個階段，最終形成了一個清晰回答所提問題的最終表格。Chain of Table的獨特之處在於其適應性和效率，即使是在處理大型表格時，也能通過專注於一小部分行進行操作轉換。這不僅提高了性能，而且通過中間結果的可見性揭示了 LLM 的推理途徑。Google 使用像 PaLM 2 和 GPT 3.5 這樣的模型在各種基準測試中進行實驗，顯示出Chain of Table在提供更準確的答案和對抗複雜問題及大型表格的堅韌性方面的優越能力。這種方法不僅提高了 LLMs 的推理能力，也為利用表格結構進行數據分析開辟了新的途徑，承諾在基於表格的推理和自然語言處理方面取得飛躍。

閱讀更多

突破性AI技術改變食道癌檢測

AI 食道癌檢測 突破性技術 Microsoft Research Cyted EndoSign 早期檢測 篩檢裝置 開放源碼 Azure Machine Learning

2024-03-11

在醫療領域的一大進展中，Microsoft Research 與 Cyted 合作，引進了一種創新方法，顯著改善食道癌的早期檢測，這是一種因晚期診斷和低存活率而臭名昭著的疾病。這種新方法利用人工智慧（AI）來增強對巴雷特食道（BE）的篩檢過程，這是一種顯著增加發展成食道癌風險的病狀。

Microsoft Research 和 Cyted 開發的AI模型已顯示與病理學家採用的傳統手工工作流程的診斷準確性匹配，並有潛力將他們的工作量減少驚人的63%。這一突破可能大大提高識別高風險患者的效率，為早期干預提供了一線希望。

介紹 EndoSign®：一種革命性篩檢裝置

這一進步的核心在於 EndoSign®，一種由 Cyted 首創的新型膠囊海綿裝置。這種最小侵入性工具旨在收集食道細胞進行分析，與傳統的內窺鏡活檢相比，簡化了程序，降低了成本，並減少了不適。

然而，挑戰在於分析收集到的細胞是否為BE，這一過程需要審查大量的組織病理學幻燈片。進入AI模型，這些模型已經被訓練以高效率掃描這些幻燈片，尋找BE的指標，準確度驚人，有潛力繞過更昂貴且耗時的染色方法。

轉化臨床工作流程

通過將這些AI模型整合到臨床工作流程中，篩檢食道癌的過程可能會大大簡化。一種提議的工作流程可能會減少對病理學家審查的需求，僅限於AI模型識別潛在BE的案例，從而為高優先級案例節省寶貴的時間和資源。

開放源碼創新，影響更廣

為了加快癌症檢測的進展，這項技術已經開放源碼，允許全球的研究人員和機構適應並改進它。這一舉措托管在Azure Machine Learning上，旨在促進創新和協作，不僅對抗食道癌，也可能對抗其他癌症。

癌症檢測的一大飛躍

Microsoft Research 和 Cyted 之間的合作代表了食道癌早期檢測的重大進步。通過利用AI的力量和創新醫療裝置，我們正步入一個新時代，早期干預可以顯著改善患者結果，為受這種毀滅性疾病影響的成千上萬人提供希望。

閱讀更多

Google Research 揭示教導 AI 理解圖表的突破

人工智能 圖表理解 大型語言模型 Google Research GraphQA

2024-03-12

Google Research 的科學家 Bahare Fatemi 和 Bryan Perozzi 在一項令人振奮的進展中，介紹了一種新方法，以增強大型語言模型（LLMs）理解和處理圖表的能力。傳統上，LLMs 擅長解釋和生成基於文本的內容，但在理解圖表的複雜互連結構方面則遇到困難，這種結構在我們的數位世界中無所不在，從互聯網的網站網絡到社交網絡。

圖表，它們將各種實體（節點）及其連接（邊）之間的關係映射出來，對於組織複雜的數據集合至關重要。然而，它們的複雜性對人工智能而言是一項重大挑戰。Google 最新的研究「像圖表一樣交談：為大型語言模型編碼圖表」旨在彌合這一差距。研究人員設計了一種方法，將圖表數據的複雜網絡轉換成 LLMs 不僅能理解，還能有效利用來解決基於圖表的問題的語言。

這項研究的基石是一個名為 GraphQA 的新基準，旨在評估 LLMs 處理圖表特定任務的能力。這一基準測試了 LLMs 對於基本的與圖表相關的問題，從識別節點之間的連接到檢測圖表內的循環。這些任務雖然基本，但對於更深入的圖表分析至關重要，並且具有廣泛的應用，從優化網絡結構到提升搜索引擎的性能。

令人驚訝的是，研究揭示了三個關鍵因素影響 LLMs 在圖表任務上的表現。首先，用於將圖表信息翻譯成文本的方法顯著影響了 LLMs 理解和處理數據的能力。其次，與圖表相關的問題的性質也起作用。對於 LLMs 而言，某些問題本質上更具挑戰性。最後，圖表本身的結構或「形狀」可以影響性能，某些配置對於 LLMs 來說更難以導航。

這項開創性的研究不僅為更加精密的能夠理解圖表中所代表的複雜關係的人工智能鋪平了道路，還推出了 GraphQA，這是一個無疑將進一步推動該領域創新的工具。通過細緻的編碼和巧妙的問題構建，Google 的研究讓我們更接近於真正能「像圖表一樣交談」的人工智能，為數據分析和更廣泛的領域解鎖新的潛能。

閱讀更多

革新語言模型：介紹Cappy

自然語言處理 Cappy 大型多任務語言模型 評分器 RoBERTa 效率 適應性

2024-03-14

革新語言模型：介紹Cappy

在自然語言處理（NLP）領域裡，Google的團隊推出了Cappy，這是一項重大的進展，它是一款緊湊卻強大的工具，旨在增強大型多任務語言模型（LLMs）的能力。與需要大量計算資源的傳統模型不同，Cappy只需一小部分的參數就能運作，使它成為效率和性能的遊戲規則改變者。

Cappy的亮點在於，它能夠評估LLMs根據給定指令生成的回應品質。它從0到1分配一個分數，反映了回應的準確度。這個創新的評分器能夠與現有的模型無縫協作，包括那些被商業機密所包圍的模型，通過一個簡單卻有效的評分系統。它是分類任務的多才多藝的盟友，也是生成挑戰的強大增強者，所有這些都不需要與調整大模型相關的龐大計算需求。

Cappy成功的秘訣在於其基於強大的RoBERTa模型之上的連續預訓練，並輔以包含問答、情感分析和摘要任務的多樣化數據集。這種方法使Cappy能夠對LLMs的輸出提供精確的調整，顯著提高它們在眾多任務中的效果。

Cappy的引入標誌著NLP領域的一個轉折點，簡化了專業任務適應過程，無需龐大的記憶體要求或直接訪問LLM參數的後勤噩夢。它能夠與其他適應策略協同工作，進一步放大其實用性，為語言模型效率和適應性的新時代奠定了基礎。

展望未來，Cappy的潛在應用是無限的，承諾在語言理解和生成方面解鎖新的性能水平。這一突破不僅僅是一步，而是在追求更智能、更易於訪問和更高效的語言模型的征途上的一次巨大飛躍。

閱讀更多

Anthropic 發布 Claude 3 Haiku：AI 速度和成本效益上的重大突破

Anthropic **Claude 3 Haiku** 人工智慧 企業級任務 處理速度 安全性 **Claude API** **Amazon Bedrock** **Google Cloud Vertex AI**

2024-03-14

在最近的一項突破中，Anthropic 推出了 Claude 3 Haiku，為人工智慧（AI）領域的速度和成本效益設立了新標準。Haiku 專為企業級任務設計，擁有無與倫比的能力，使其成為同類中最快的 AI 模型。這一創新有望革命性地改變企業處理大型數據集的方式，提供快速的分析和輸出，且成本只是通常情況的一小部分。

Claude 3 Haiku 在處理速度上表現出色，能夠每秒處理 21,000 個令牌（約 30 頁），用於不超過 32,000 令牌的提示。這種快速處理能力使企業能夠同時執行多種任務——從客戶支持到吸引人的聊天體驗——使運營效率達到前所未有的水平。

此外，Haiku 在各種基準上的卓越視覺能力和強大性能，包括推理、數學和編程，凸顯了其在滿足多樣化企業需求上的多功能性。值得注意的是，其輸入到輸出令牌比率為 1:5，確保即使是複雜的提示也能有效處理，使其成為分析大量文件（如季度報告、合同或法律案件）的理想解決方案，且成本減半。

安全性是 Haiku 設計的另一個基石。Anthropic 實施了一套全面的安全框架，包括持續監控、安全編碼實踐、強大的數據加密和嚴格的測試，以降低風險並保護企業數據。

Claude 3 Haiku 通過 Claude API 和 Claude.ai 提供給 Claude Pro 訂閱者，並將在 Amazon Bedrock 上標示其存在，並即將在 Google Cloud Vertex AI 上推出。這種可訪問性進一步放大了其在各個部門的潛在影響，承諾在前所未有的速度和負擔性上帶來 AI 驅動的效率和安全的新時代。

通過提供一個不僅更快，而且比同類產品更具成本效益的解決方案，Anthropic 的 Claude 3 Haiku 準備好改變企業運營，體現人工智慧技術向前的重大飛躍。

閱讀更多

以 Amazon Bedrock 的生成式 AI 革新程式碼審查

Amazon Bedrock **生成式 AI** **程式碼審查** **AWS** **自動化** **程式碼品質** **安全性**

2024-03-14

以 Amazon Bedrock 的生成式 AI 革新程式碼審查

在快速變化的軟體開發世界中，確保新程式碼的品質、安全性和功能性是落在專案經理肩上的重要任務。然而，這個過程充滿挑戰，從需要深厚的技術知識到在緊迫的期限內需要審查的變更量大。傳統上，這意味著大量的手動努力和一個緩慢、笨重的審批過程。

進入 Amazon Bedrock，一個在程式碼審查和批准領域的遊戲改變者。利用生成式人工智慧（AI）的力量，Amazon Bedrock 提供了一個不僅加快程式碼審查過程，還提高其效率和準確性的解決方案。這項技術與 AWS 的強大部署工具套件無縫整合，自動化分析和總結程式碼變更。這是一個端到端的工作流程，將 AI 的最佳實踐與 AWS 服務的靈活性和力量結合在一起。

這是它的運作方式：

1. 開發人員推送程式碼變更，觸發一個自動化部署流程。
2. AWS CodeBuild 處理這些變更，執行必要的掃描和測試。
3. 變更由 Amazon Bedrock 的 AI 分析，產生一個簡潔的摘要。
4. 管理者被通知並提供所有需要的資訊來做出知情的批准決定。

使 Amazon Bedrock 脫穎而出的是其使用的基礎模型（FMs），使得選擇最適合您特定需求的 AI 模型成為可能。這些模型可以用您的數據進行客製化，確保在程式碼審查過程中的相關性和準確性。最棒的是，這一切都無需您管理任何基礎設施。

通過採用 Amazon Bedrock，經理人現在可以滿懷信心地應對程式碼審查的挑戰，得到生成式 AI 的見解和速度的支持。結果？更快、更安全、品質更高的軟體部署。

擁抱程式碼審查的未來，體驗 Amazon Bedrock 和生成式 AI 的效率。

閱讀更多

開闢新視野：Amazon SageMaker JumpStart 中的 Gemma 模型

Amazon SageMaker JumpStart **Gemma 模型** **機器學習** **語言理解** **推理能力** **預先訓練的模型** **負責任生成式 AI 工具箱**

2024-03-14

開闢新視野：Amazon SageMaker JumpStart 中的 Gemma 模型

在不斷演進的機器學習領域中，Amazon SageMaker JumpStart 引入 Gemma 模型，成為眾所矚目的焦點。這些受 Google 的 Gemini 啟發而設計的模型，旨在將語言理解和推理能力提升到前所未有的水平。現在，用戶可以訪問這些預先訓練的模型，包括 Gemma 2B 和 Gemma 7B 兩個版本，它們擁有高達 70 億的參數，並在驚人的 6 兆代幣文本上進行訓練。這一進步不僅承諾在編碼、常識推理等多個領域提升性能，還通過新發布的負責任生成式 AI 工具箱，為更安全的 AI 應用鋪平道路。

Amazon SageMaker JumpStart 作為這些創新的大門，提供了與其生態系統的無縫整合。無論您是在機器學習領域試水，還是為專業任務微調模型，JumpStart 都簡化了旅程。憑藉隨時可用的基礎模型，SageMaker 減輕了從頭開始訓練的沉重成本和複雜性，讓用戶能夠為他們特定的需求微調這些巨型模型。

但真正讓 Gemma 脫穎而出的是什麼？其輕量級框架和最先進的能力確保它不僅僅是眾多模型中的一滴水。Gemma 模型在理解和生成類人文本方面表現出色，為 AI 能夠以更自然、直觀的方式互動和協助開闢了新的前沿。

在 SageMaker 的全面環境下部署和微調 Gemma 模型輕而易舉。無論您是部署一個模型以前所未有的準確性回答問題，還是為對話數據集進行微調，SageMaker JumpStart 都提供了實現 Gemma 全潛力所需的工具和指導。

當我們邁入機器學習的新時代，Amazon SageMaker JumpStart 中的 Gemma 模型不僅象徵著技術進步，也代表著致力於使尖端 AI 更加可及和負責任的承諾。機器學習的未來充滿希望，正是這些工具將引領我們實現其全部潛力。

閱讀更多

Google 研究引入 HEAL：向 AI 中的健康平等邁進

HEAL 健康平等 人工智慧 機器學習 Google 研究 健康不平等 公平性能 負責任 AI
2024-03-15

在健康差距極為明顯的時代，Google 研究部門已著手進行一項開創性的項目——機器學習性能的健康平等評估（HEAL）。這個創新框架旨在確保機器學習（ML）和人工智慧（AI）技術的進步不會忽視那些處於醫療可及性邊緣的人群，而是努力減少健康不平等。

HEAL 基於健康平等與平等之間的根本區別，強調 AI 技術需要提供對現有健康差異敏感的性能，以便每個人都有機會獲得最佳健康。HEAL 引入了一個四步驟過程，旨在評估基於 ML 的健康技術是否提供了公平的利益。這包括識別和量化健康差異、衡量工具在不同子群體中的性能，以及評估這些技術在改善受健康差異影響最大的人群成果方面的優先程度。

一項關於皮膚病學 AI 模型的案例研究說明了 HEAL 框架的實際應用。這個模型經過訓練，能夠從圖像和病人信息中識別皮膚病徵，並進行評估以確定其性能是否符合減少健康不平等的目標。研究發現，不同種族/民族和性別之間有希望實現公平性能的可能性，儘管它突出了在非癌症皮膚病條件的年齡相關差異方面需要改進的領域。

Google 的倡議強調了在 AI 技術的開發和評估中納入健康平等考慮因素的重要性。通過優先考慮與現有健康差異相關的性能，HEAL 框架旨在指導 AI 工具的開發，對減少健康結果的不平等產生積極影響。這種方法不僅展示了對負責任 AI 開發的承諾，還為所有社會成員受益的更具包容性的健康技術鋪平了道路。

在探索健康平等的複雜性時，Google 研究認識到 HEAL 背後的合作努力和在各種 AI 工具和使用案例中不斷完善和應用的持續需求。最終目標是培養對 AI 技術對健康平等影響的全面理解，超越指標，實現與受健康差異影響最大的人群需求相一致的共享目標。

閱讀更多

用RAFT革命性地推進語言模型：在語言模型上的一大步

RAFT 語言模型 LLM RAG Meta Llama 2 Azure AI Studio 領域適應性 自然語言處理 AI技術 生成性AI

2024-03-15

用RAFT革命性地推進語言模型：在語言模型上的一大步Microsoft與UC Berkeley的研究人員合作，引入了一種創新方法來增強大型語言模型（LLMs）的性能，名為RAFT（Retrieval Augmented Fine-tuning，檢索增強的細調）。這種突破性的方法結合了檢索增強生成（RAG）和細調的優勢，提供了一種更有效的方式來使LLMs適應特定領域，如銀行、法律和醫療行業。在傳統設置中，教導LLMs涉及使用特定領域知識進行細調或利用RAG，後者根據文檔的語義相似性來檢索文檔以回答查詢。然而，這兩種方法都有其限制。細調可能導致不準確性和幻覺，而RAG儘管提供了上下文，但在區分真正相關的文檔和分心項目方面存在困難。RAFT通過本質上允許LLM在被查詢之前「學習」相關文檔來引入一種新解決方案，類似於學生為開卷考試做準備。這種方法利用Meta Llama 2和Azure AI Studio，通過使它們更善於從提供的上下文中提取有用信息，顯著提高了LLM的性能。該方法從準備一個合成數據集開始，該數據集包括問題、一系列相關和不相關的文檔，以及一個生成的答案連同一個思考過程的解釋。然後使用這個數據集來細調Llama 2模型，增強其領域適應性並減少過度擬合的風險。RAFT不僅通過將LLM的語氣和風格與領域對齊，而且通過提煉從檢索到的上下文生成的答案的質量而脫穎而出。通過使LLMs更適用於像醫療保健或法律等特定領域，RAFT為特定領域AI開發設定了新的標準。Microsoft和UC Berkeley的合作示範了如何利用基礎模型和用戶友好的平台進行細調，以實現對先進自然語言處理能力的普及。這為各行各業的創新、定制化解決方案鋪平了道路，降低了企業和開發者希望利用生成性AI和LLMs的能力的門檻。本質上，RAFT不僅僅是AI技術的進步；它是朝著創建一個更包容和多元的AI生態系統邁出的一步，最新的突破對所有人都是可訪問的，促進了特定領域AI解決方案的新時代。

閱讀更多

OpenAI 從阿聯酋獲得財務支持以自行開發晶片

OpenAI 晶片開發 阿聯酋 MGX Nvidia 半導體技術 國際合作 AI 應用

2024-03-15

OpenAI，這個站在人工智能技術創新前沿的組織，正報導與阿布達比的國家支持實體 MGX 進行談判，以資助其開發自家半導體晶片的雄心勃勃項目。這個合作可能標誌著科技世界的一個轉捩點，可能減少 OpenAI 對 Nvidia 的依賴，Nvidia 目前領先於半導體產業。

OpenAI 創造這些晶片不僅僅是為了獲得技術獨立，更是為了推動 AI 能夠達成的極限。有了阿聯酋的財務支持，OpenAI 進軍晶片製造業顯示出公司控制 AI 技術全部堆棧，從硬件到軟件的一個新時代即將到來。

這一舉措發生之際，英國半導體部門也因該國參與歐盟的「晶片聯合承諾」而獲得提升。這一計劃承諾提供對一個龐大研究基金的加強訪問，凸顯了半導體技術在當今數位時代的戰略重要性。

MGX 與 OpenAI 的合作，以及對英國和歐盟半導體產業的更廣泛影響，凸顯了全球加強科技基礎設施的轉變。這些發展不僅承諾在 AI 能力上的進步，也突顯了在創新尋求中國際合作趨勢的增長。

隨著這些技術的發展，它們有可能轉變行業，從醫療保健到娛樂，通過使 AI 更加強大和易於使用。阿聯酋對 OpenAI 的支持可能是一個遊戲規則的改變者，為一個未來設定了舞台，其中 AI 解決方案更加定制、高效，並且重要的是，在那些揮舞它們的人的控制之下。

閱讀更多

利用 AWS 上的聯邦學習轉變醫療保健

聯邦學習 AWS 醫療保健 數據隱私 機器學習 Amazon Elastic Kubernetes Service Amazon SageMaker FedML

2024-03-15

在 AWS 上使用聯邦學習轉變醫療保健

在提升數據隱私的同時利用機器學習（ML）獲得更好的醫療保健成果方面，AWS 推出了一種使用 FedML、Amazon Elastic Kubernetes Service（EKS）和 Amazon SageMaker 的聯邦學習（FL）方法。這項創新將革命性地改變醫療數據的利用方式，解決了隱私和安全的首要關切。

聯邦學習是一種分散式訓練 ML 模型的方法，可大幅降低隱私風險。與傳統的 ML 不同，在傳統 ML 中數據被匯集和暴露，FL 確保敏感數據永遠不會離開其原始位置。相反，只有模型更新被分享到中央服務器進行聚合，而不會危及底層數據。這種方法在醫療保健領域尤為關鍵，因為患者數據既敏感又受到嚴格的規範。

在 AWS 中實施這種方法利用了 FedML，這是一個為 FL 設計的框架，它在不同的計算環境中提供靈活性，支持眾多的 FL 範式。這個開源解決方案使得 FL 應用的開發和部署變得無縫，確保了數據隱私，增強了合作而不暴露敏感信息。

一項激動人心的應用案例利用了來自多個組織的心臟病數據。通過使用 FL，模型可以在這些分散的數據上學習，提高預測心臟病的準確性，而無需分享患者的私人信息。這不僅保護了隱私，還利用了更廣泛的數據集獲得了更堅實的醫療保健洞察力。

此外，AWS 已將此解決方案與 Amazon SageMaker Experiments 集成，允許對 ML 實驗進行嚴謹的追踪和分析。這一集成確保了 ML 項目的透明度和可重現性，賦予研究人員以信心對其模型進行改進和演進。

AWS 上的這個聯邦學習解決方案為醫療保健和其他領域中更私密、更安全且更高效的數據使用鋪平了道路。通過使組織能夠在不危及數據安全的情況下合作開發 ML 模型，AWS 為負責任的 AI 開發設定了新標準。

閱讀更多

使用 Google 的 MELON 技術，革命性地改變了 3D 重建

Google MELON 3D 重建 計算機視覺 神經輻射場 NeRF 電子商務 自動駕駛

2024-03-18

用 Google 的 MELON 技術革新 3D 重建

在計算機視覺領域，Google 的 MELON 技術正改變著從圖像中重建 3D 物體的遊戲規則，即使在相機的位置或「姿態」未知的情況下。這項創新解決了該領域一個長期存在的挑戰：不需要預先確定的相機角度，就能從少量二維圖像中準確構建三維模型。

MELON 的突出之處在於不需要對相機位置的初步猜測、複雜的訓練過程，或是預標記的數據，簡化了將其整合到當前技術如神經輻射場（NeRF）中的過程。僅需 4-6 張圖像，MELON 就能在重建 3D 物體方面達到驚人的準確度，展示了其在電子商務、自動駕駛等領域的潛力。

秘密武器？一種動態的、輕量級神經網絡，能直接從圖像中推導出相機姿態，以及一種巧妙的「模數損失」方法，考慮到物體的對稱性，讓系統能識別出最合理的視點。這些突破使 MELON 能以前所未有的精確度，導航著物體重建的棘手水域。

值得注意的是，MELON 也展示了對噪聲的抵抗力，即使是從高度扭曲的圖像中，也能成功重建物體。這種穩健性暗示了 MELON 更廣泛的適用性及其在現實世界情景中的應用前景，其中條件遠非理想。

隨著 Google 持續改進 MELON 以適用於實際應用，這項技術預示著 3D 成像的新時代，使複雜的計算機視覺世界變得更加觸手可及，也更加令人興奮。

閱讀更多

加速未來：Microsoft Research 揭露 Garnet，一款下一代開源快取儲存系統

Microsoft Research Garnet 快取儲存系統 開源 .NET 技術 群集模式 數據儲存 API 高效率

2024-03-18

在近期的一大突破中，Microsoft Research 介紹了 Garnet，一個創新的快取儲存系統，旨在透過使應用程式和服務更快速且高效來加速它們。這個開源的奇蹟，現已可供下載，正在設定新的數據儲存與存取技術標準。

Garnet 之所以脫穎而出，是因為它解決了互動式網頁應用程式和服務中對於快速數據處理的日益增加的需求。通過提高速度並降低運營成本，它承諾在傳統快取儲存系統上實現顯著的性能改進。

Garnet 究竟有何獨特之處？首先，它建立在最新的 .NET 技術上，使其跨平台、高度可擴展且具有未來證明性。這意味著各平台的開發者現在可以享受增強的可擴展性、吞吐量，以及重要的是，在關鍵性能百分位數上更低的延遲。

此外，Garnet 引入了群集模式，具有分片、複製和動態關鍵字遷移等功能，擴大了其適用性和效率。其 API 支援從簡單的字串操作到複雜的分析任務的廣泛操作，同時啟用 C# 中的自訂擴展。

Microsoft 對於開源創新的承諾透過 Garnet 彰顯無遺，邀請開發者來貢獻、擴展並進一步優化這個系統。它在多個 Microsoft 情境中的部署已經展示了其重新定義應用程式性能和成本效率的潛力。

本質上，Garnet 不僅僅是一個新的快取儲存系統；它是通往網頁和應用服務未來的大門，承諾一個速度、效率和可擴展性不僅僅是理想，而是現實的世界。隨著 Garnet 持續發展，它設定成為開發者工具箱中不可或缺的工具，推動我們數位世界可能性的界限。

閱讀更多

革新頭像互動：Microsoft Research 的洞見

數位通訊 虛擬互動 動作捕捉 頭像 生態效度 社會文化影響 同理心 數字頭像

2024-03-18

在數位通訊不斷進化的領域中，Microsoft Research 正在開創使我們的虛擬互動更加真實和包容的方式。他們最新的研究，在著名的 ANIVAE 2024 上發表，深入探討了以頭像為基礎的通訊細節，提供了關於上下文、文化和角色如何影響我們的數字體驗的開創性見解。

頭像——我們在視頻會議、增強現實（AR）和虛擬現實（VR）平台中的數字對應物，正在變得越來越複雜。然而，它們面臨著運動噪音——抖動和扭曲——這可能會減損真實感。Microsoft 的研究人員正在探索創新的動作捕捉技術，以最小化這些干擾，目標是通過創建能夠準確反映真實世界文化和能力多樣性的頭像，來提升用戶體驗。

該研究通過強調生態效度，或實驗與現實生活情境的相似程度，在評價頭像噪音方面引入了新視角。這種方法揭示了參與者對於頭像的反應基於互動上下文顯著不同，強調了設計真實生活中的頭像的重要性。

通過在涉及建築師-客戶討論家庭裝修的動態場景中動畫化頭像，研究人員觀察到全上下文情境導致對噪音的更大容忍度，凸顯了社會上下文的關鍵作用。此外，文化觀察揭示了隱性偏見，例如性別角色影響噪音感知，指出實驗中高生態效度的必要性，以捕捉社會文化影響對頭像感知的影響。

此外，該研究還涉及了同理心的角色，發現在豐富上下文情境中，同理心得分較低的個體對噪音更加苛刻。這一見解暗示了個人特質與數字通訊偏好之間的複雜相互作用。

Microsoft Research 在這一領域的努力不僅僅是關於精煉技術，還關於培養更有意義和包容的數字互動。通過在頭像研究中優先考慮真實情境，他們旨在橋接數字與物理通訊之間的差距，確保我們的虛擬代表像我們一樣細膩和多元。

這項研究標誌著提高數字頭像的真實感和包容性的重要一步，承諾將革新我們在數字領域中的連接方式。

閱讀更多

使用NVIDIA與Amazon SageMaker革命性地推進語言模型部署

NVIDIA Amazon SageMaker 語言模型 部署 NIM微服務 推論 自然語言處理 TensorRT Triton推理伺服器 定製容器 成本管理 性能優化

2024-03-18

使用NVIDIA與Amazon SageMaker革命性地推進語言模型部署

在一次破天荒的合作中，NVIDIA的NIM微服務已與Amazon SageMaker整合，為部署和優化大型語言模型(LLMs)設定了新標準。這次合作為在幾分鐘內部署最先進的LLMs鋪平了道路，將原本需要數日完成的任務轉變為迅速、高效的過程。

NIM作為NVIDIA AI企業軟體平台的一大亮點，提供了推論微服務，賦予應用程式先進的自然語言處理和理解能力。無論是開發複雜的聊天機器人、摘要大量文件，還是打造其他以NLP為驅動的應用程式，NIM微服務都能簡化這一過程。透過使用技術如NVIDIA TensorRT、NVIDIA TensorRT-LLM和NVIDIA Triton推理伺服器，這些服務確保您的LLMs在SageMaker內的NVIDIA加速實例上達到巔峰性能。

NIM的精髓在於其提供針對流行模型的預優化引擎，促進無縫模型部署。對於那些努力定制的人，NIM提供了製作定製容器的工具，以加強您獨特的應用需求。

此外，NIM引入了先進的託管技術，如即時批處理。該技術通過與進行中的請求同時處理新請求，顯著提高了計算利用率，確保了高效率的部署環境。

將NIM與SageMaker整合不僅優化了性能，而且還細致地管理了成本。它利用SageMaker的規模調整、部署策略和工作負載評估能力，同時通過Amazon CloudWatch提供全面的可觀察性。

展望未來，NIM旨在通過支援參數高效微調(PEFT)定制方法和擴展後端支持，進一步革命化LLM部署。這一舉措不僅強調了對性能和成本效率的承諾，同時也簡化了LLM部署，使尖端自然語言處理比以往任何時候都更加易於取得。

與NVIDIA NIM微服務和Amazon SageMaker一同擁抱LLM部署的未來，並體驗這次合作帶來的無與倫比的好處。

閱讀更多

利用 Amazon SageMaker JumpStart 上精調的 Code Llama 模型提升您的編碼技巧

Amazon SageMaker JumpStart **Code Llama** 編碼技巧 模型精調 機器學習 程式設計 **Meta** **PyTorch FSDP** **PEFT/LoRA** **Int8 量化** **Python SDK** 代碼生成 除錯

2024-03-18

利用精調的 Code Llama 模型在 Amazon SageMaker JumpStart 中提升您的編碼技巧

在不斷進化的程式設計領域中，Amazon Web Services (AWS) 為程式設計師和機器學習從業者帶來一項激動人心的發展：使用 Amazon SageMaker JumpStart 精調 Code Llama 模型的能力。Code Llama 是 Meta 的心血之作，專為編碼任務而設計，範圍從代碼生成到除錯，橫跨多種流行的程式語言，如 Python、Java 等。

精調這些預訓練模型可以提高準確性和解釋性，提供針對性的編碼解決方案。透過在 Amazon SageMaker Studio UI 中幾次點擊或使用 SageMaker Python SDK，用戶可以輕鬆地精調和部署 Code Llama 模型。此過程利用了如 PyTorch FSDP、PEFT/LoRA 和 Int8 量化等先進技術進行高效的模型優化。

Code Llama 的應用不僅僅局限於代碼生成；它還可以解釋自然語言提示來寫函數、完成代碼和協助除錯。精調進一步提煉了這一能力，確保模型適應特定的編碼領域和任務，通過在如 HumanEval 和 MBPP 數據集等基準中的改進表現得以證明。

在 SageMaker JumpStart 中開始是簡單的，且通過 SageMaker Studio UI 精調無需編碼。對於那些偏好編碼的人，SageMaker Python SDK 為自然語言處理任務的模型精調提供了靈活性，增強了它們處理未見任務的能力，使用零樣本提示。

這一進步不僅只是關於改善代碼生成，而是關於創造一種更直觀、更準確和更高效的編碼體驗。無論您是經驗豐富的開發者還是編碼新手，在 Amazon SageMaker JumpStart 上精調 Code Llama 模型為創新開啟了新門戶，使編碼變得更加容易和愉快。深入 AWS 的增強編碼世界，探索您今天能夠用精調的 Code Llama 模型實現的成就！

閱讀更多

NVIDIA 在路線優化上創下 23 項世界紀錄，樹立新標準

NVIDIA cuOpt 路線優化 物流管理 AI Omniverse GTC 2024 物流和製造業 真實時間動態規劃

2024-03-18

在物流和運營效率方面，NVIDIA 以確立 23 項新的世界紀錄，為這個領域帶來令人振奮的飛躍。這項成就凸顯了 NVIDIA 的 cuOpt 路線優化引擎在不同產業中簡化運營的無與倫比能力，從與川崎重工合作提升鐵路安全，到與 SyncTwin 攜手優化製造業。

在 GTC 2024 會議期間，NVIDIA 的 CEO Jensen Huang 公布 cuOpt 正式向公眾開放，標志著物流管理新時代的來臨。這款開創性的引擎不僅僅是尋找從 A 點到 B 點的最短路徑，它大膽地重新想像了貨物、服務和資訊在全球網絡中的流動方式，承諾將革命性地改變配送、服務呼叫乃至倉庫和工廠的內部運作。

NVIDIA 的 cuOpt 之所以與眾不同，是因為它融入了 NVIDIA AI Enterprise 軟體平台，允許實時動態重新規劃路線、工廠優化，甚至進行機械人模擬。當與 NVIDIA Omniverse 配對時，公司可以創建超逼真的 3D 模擬，用於規劃和優化複雜的自動化環境。這強大的組合進一步得到 NVIDIA Metropolis for Factories 的增強，確保在精確度和可靠性至關重要的營運中更安全、更高效。

cuOpt 引擎的實力不僅僅是理論上的。它已經在最具挑戰性的基準測試中經受考驗，速度比傳統基於 CPU 的解決方案快達 100 倍。這項技術奇蹟正在幫助像川崎重工這樣的公司，通過先進的 AI 驅動系統自動化並提升鐵路軌道檢查的安全性和效率，每年潛在節省數百萬。

同樣地，汽車座椅巨頭 SyncTwin 正在利用 cuOpt 和 Omniverse，協調其廣闊供應鏈中原材料的流動。這不僅使生產流程更為簡化，還為運營效率和可持續性開辟了新的視野。

有了 NVIDIA cuOpt，物流和製造業的未來不僅僅是優化，而是轉型。GTC 2024 的與會者獲得了這個未來的前排席位，全球運作的複雜性遇到了更智能、更快速、更可適應的解決方案，全部多虧了 NVIDIA 的遠見技術。

閱讀更多

顛覆道路：NVIDIA 在 2024 年 GTC 推動未來

NVIDIA GTC AI DRIVE Thor 自動駕駛 車載娛樂系統 GPU 電動車 自主交通
2024-03-18

本週，聖荷西成為了汽車創新的震中，全球人工智慧會議（GTC）2024 展示了汽車行業 AI 技術的最新進展。這次會議集結了世界領先的汽車製造商和技術專家，揭幕了有望改變我們駕駛體驗的新型號與技術。

活動的一大亮點是 NVIDIA 執行長 Jensen Huang 宣布推出 NVIDIA DRIVE Thor 平台。這項尖端技術結合了先進的駕駛輔助和車載娛樂系統，由專為轉換器和生成式人工智慧任務而設計的新 NVIDIA Blackwell GPU 架構提供動力。這項宣布讓業界領袖電光石火間振奮，像是 BYD、Hyper 和 XPENG 等頂尖電動車（EV）製造商已計劃將 DRIVE Thor 納入其下一代車輛中。

此外，NVIDIA 不僅加速個人交通的發展，還在通過與 Nuro、Plus、Waabi、WeRide 和 DeepRoute.ai 的合作關係，革命化卡車運輸、機器人計程車和貨物配送。這些合作旨在利用 DRIVE Thor 創建更智能、更安全的自動解決方案。

GTC 展覽廳內，展示了由 NVIDIA 驅動的車輛，包括 Lucid Air 和 Mercedes-Benz Concept CLA Class，展現了 NVIDIA 技術在提升車輛安全性、效率和用戶體驗方面的無縫整合。

除了汽車進步之外，NVIDIA 還推出了 Omniverse Cloud APIs。這些 API 將通過高保真感測器模擬加速車輛的自主性，對於開發和驗證自動駕駛車輛（AVs）至關重要。這一發展凸顯了 NVIDIA 致力於引領智能、AI 驅動車輛的發展，邀請開發者和軟件供應商加入這場變革之旅。

隨著 2024 年 GTC 的繼續進行，AI 在重新塑造我們對交通和移動性的看法方面的潛力比以往任何時候都更加生動。隨著 NVIDIA 的 DRIVE Thor 領銜前進，前方的道路看起來更安全、更智能、更互聯。請繼續關注來自汽車創新前線的更多更新。

閱讀更多

NVIDIA 為未來資料中心揭開前沿方法

NVIDIA 資料中心 數位雙胞胎 Omniverse 液體冷卻系統 NVIDIA Grace CPU NVIDIA Blackwell GPU CAD 數據中心設計 效率 可擴展性 能源效率

2024-03-18

在 2024 年 3 月 18 日這個令人興奮的日子裡，NVIDIA 介紹了一個旨在打造下一代資料中心的革命性數位藍圖。這個前瞻性的計畫，得以透過支持 Omniverse 數位雙胞胎的概念，以及與行業巨頭如 Ansys、Cadence、PATCH MANAGER、Schneider Electric、Vertiv 等的合作夥伴關係得以實現。此舉為建立高效能 AI 基礎設施奠定了基礎，有望重新塑造資料中心的構想與運作方式。

設計和管理現代資料中心需要在性能、能源效率和可擴展性之間取得微妙的平衡。NVIDIA 的方法包括組建一支頂尖工程師團隊來解決計算和網絡設計、電腦輔助設計 (CAD) 建模，以及機械、電氣和熱設計的複雜性。

NVIDIA 宣布的核心是發布一個基於 NVIDIA GB200 NVL72 液體冷卻系統的大型集群。這個強大的設置擁有 18 個 NVIDIA Grace CPU 和 36 個 NVIDIA Blackwell GPU，分佈在兩個機架上，所有這些都通過先進的第四代 NVIDIA NVLink 交換機相互連接。

使這個項目與眾不同的是，利用 NVIDIA Omniverse 平台創建了一個完全運作資料中心的數位雙胞胎。這個數位雙胞胎使工程師能夠以前所未有的準確性和效率設計、模擬和優化新的資料中心。通過利用 Omniverse，NVIDIA 及其合作夥伴創建了一個共生環境，CAD 數據集在完全的物理準確性和照片實景中融合，為更加流暢和高效的資料中心開發過程鋪平了道路。

這個項目的一項關鍵創新是技術公司 Kinetic Vision 使用 NavVis VLX 可穿戴激光掃描儀。這個掃描器產生高度精確的點雲數據和全景照片，當與 Prevu3D 軟體結合使用時，可以轉換成 3D 網格。這個網格形成了一個物理精確的設施模型的基礎，在此之內模擬了新的數位資料中心。

此外，該項目展示了協作和整合的力量，工具如 PATCH MANAGER 和 NVIDIA Air 在設計和優化集群和網絡基礎設施的物理布局中發揮了關鍵作用。共同努力導致了更有效和優化的設計，大大減少了啟動資料中心所需的時間。

NVIDIA 的這項舉措不僅凸顯了在設計和優化資料中心時數位雙胞胎的潛力，還為行業在效率、準確性和創新方面設定了新的標準。這是對資料中心技術未來的一瞥，數字和實體世界融合創造了更可持續和高效的基礎設施。

敬請關注 NVIDIA 在重塑資料中心和 AI 基礎設施景觀中的進一步更新。

閱讀更多

NVIDIA推出新存儲驗證計劃，簡化企業AI部署

NVIDIA 存儲驗證計劃 AI部署 NVIDIA OVX 存儲系統 人工智慧 數據處理 DDN Dell PowerScale NetApp Pure Storage WEKA L40S GPU NVIDIA AI Enterprise

2024-03-18

NVIDIA為企業推出新的存儲驗證計劃，簡化AI部署

NVIDIA為了促進企業更順暢、更快速地整合人工智慧，推出了一項旨在優化其NVIDIA OVX計算系統存儲方案的新計劃。這項創新舉措承諾將革命化企業部署人工智慧的方式，解決建置AI基礎設施的日益增加的複雜性與時間挑戰。

這項創新的核心在於NVIDIA存儲夥伴驗證計劃，旨在簡化選擇與NVIDIA OVX伺服器相容的高效能存儲系統。這些伺服器是複雜AI和圖形密集型任務（從聊天機器人到搜索工具）背後的強大動力，這些任務需要龐大的數據處理能力。高速、高效的存儲對於最大化這些系統的性能至關重要。

參與這一開創性計劃的行業領導者包括DDN、Dell PowerScale、NetApp、Pure Storage和WEKA，他們全部已成功完成NVIDIA嚴格的OVX存儲驗證。這一過程涉及全面測試，以確保存儲解決方案滿足多樣化企業AI工作負載的高要求，著重於存儲性能和輸入/輸出擴展等因素。

這個計劃不僅僅是一個驗證印章，它提供了一個結構化且靈活的框架，允許企業自定義他們的存儲和系統配置。這種適應性確保了由NVIDIA L40S GPU驅動並配備NVIDIA AI Enterprise軟體的NVIDIA OVX伺服器的先進計算能力，可以有效地整合到現有的資料中心環境中。

憑藉NVIDIA認證的OVX伺服器，企業現在可以更有信心和效率地開始他們的AI之旅，依靠該計劃的指導來確保他們的基礎設施為生成式AI應用優化。這項計劃不僅強調了NVIDIA致力於促進各個行業採納AI的承諾，也標誌著朝著簡化與AI部署相關的技術複雜性邁出的重要一步。

隨著這些經過驗證的存儲解決方案變得可用，NVIDIA繼續與系統廠商密切合作，以支持AI無縫地融入業務運營，為全球企業承諾一個創新和效率的新時代。

閱讀更多

NVIDIA及其合作夥伴將量子計算推向新時代

NVIDIA 量子計算 人工智慧 CUDA-Q 量子機器學習 教育 醫療保健 化學 金融詐騙 數字支付

2024-03-18

在邁向計算未來的重大轉變中，NVIDIA及其合作夥伴在量子計算領域取得了重大進展。從研究到教育的各項倡議不僅拓展了我們對量子世界的理解，也為其在各行各業的實際應用鋪平了道路。

這些進步的核心是人工智慧與量子計算的整合，以開啟新的可能性。加拿大和美國的研究人員已開發出一個簡化量子模擬的大型語言模型，允許深入探索分子。這項創新由Alan Aspuru-Guzik教授領導，利用NVIDIA的CUDA-Q，一種混合編程模型，並運行在NVIDIA的H100 GPU超級電腦Eos上。這一突破預示著將量子演算法與機器學習融合的新方法，其潛在應用包括醫療保健、化學等領域。

在金融領域，匯豐銀行正在用量子機器學習來對抗詐騙。該銀行已成功利用NVIDIA GPU模擬了165個量子比特，這一成就顯著超過了傳統基準。這種量子機器學習演算法有望在數字支付中革新詐騙檢測，展示了量子計算在處理複雜數據方面的可擴展性和力量。

教育也獲得了量子加持。NVIDIA正與近兩打大學合作，為未來的計算機科學家準備量子時代。通過圍繞CUDA-Q設計課程，該倡議旨在彌合傳統計算與量子系統之間的差距，確保一支準備充分的勞動力隊伍，準備迎接量子計算帶來的挑戰和機會。

混合量子-古典計算的生態系統進一步擴展，隨著日本的ABCI-Q和丹麥的NVIDIA DGX SuperPOD等新系統的部署。這些專門用於量子計算研究的系統，凸顯了全球致力於推進這項尖端技術的承諾。

此外，與初創公司的合作以及與CUDA-Q的整合正在推動量子計算的能力。從模擬更好的電池材料到開發適用於網絡防禦的量子啟用解決方案，這些合作突顯了量子計算的多樣化應用。

隨著NVIDIA繼續推動量子計算的極限，其影響預計將貫穿多個部門，不僅有望增強我們的計算能力，也將解鎖解決世界上一些最複雜問題的方案。

為了深入了解這些量子飛躍，NVIDIA邀請愛好者和專業人士在NVIDIA GTC進一步探索這些主題，這是一場全球AI會議，將更加聚焦量子計算的未來。

閱讀更多

NVIDIA Maxine：革命化視訊會議體驗

NVIDIA Maxine 人工智慧 視訊會議 AI驅動特性 音訊品質 擴增實境 視線校正 語音字型模型 虛擬體育娛樂

2024-03-18

在NVIDIA的一次突破性更新中，視訊會議、呼叫中心和串流應用領域將被徹底轉型。介紹Maxine開發者平台 - NVIDIA的最新創新，旨在重新定義價值100億美元的視訊會議產業。

Maxine利用人工智慧來賦予開發者力量，授予他們整合AI驅動特性的工具，這些特性承諾提供個性化、引人入勝且高效的視訊會議體驗。這個平台透過NVIDIA AI Enterprise軟體取得，為提高視訊和音訊品質開啟了可能性的大門，並將擴增實境效果引入日常視訊通話中。

透過先進AI特性提升參與度

Maxine脫穎而出，提供一系列旨在提高用戶參與度的特性。從噪聲減少、視訊去噪和視訊升級到突破性的功能，如視線校正、實時人像和甚至實時3D轉換，Maxine正在為視訊會議品質和個人連接設定新標準。

一個值得注意的特色，眼神接觸模型，利用視線重定向來促進更沉浸式的會議體驗。此外，語音字型模型在適應目標聲音的同時保留講話者的語言細節，確保通訊的清晰度和個性化。

令人興奮的發展在地平線上

NVIDIA不會就此止步。該平台承諾未來將增強，包括Speech Live Portrait和Studio Voice，設計來確保參與者始終展現和聲音最佳狀態。此外，Maxine 3D和視頻重燈模型的早期訪問也在計畫中，承諾帶來更吸引人的視訊會議功能。

真實世界應用和全球影響

超越技術進步，Maxine的AI綠幕特性已經開始在虛擬體育娛樂世界引起波瀾。與阿森納足球俱樂部的合作展示了Maxine的潛力，使用AI為俱樂部的全球粉絲基礎創造沉浸式虛擬體驗。

視訊會議的新時代

像Gemelo、Pexip、Spectacle和VideoRequest這樣的公司是最早利用Maxine的新特性的公司之一，將標準通訊工具轉變為專業工作室。這標誌著在創建個性化視訊消息和提高用戶參與度方面向前邁出了重要的一步。

隨著NVIDIA Maxine現已在NVIDIA AI Enterprise平台上可用，視訊會議的未來是光明的，承諾提供無與倫比的個性化、效率和參與度水平。

敬請期待，我們將繼續關注NVIDIA在用AI驅動創新重塑視訊會議格局的旅程。

閱讀更多

解鎖未來使用者介面的鑰匙：Google 的 ScreenAI

Google ScreenAI 視覺語言模型 UI PaLI pix2struct

2024-03-19

解鎖未來使用者介面的鑰匙：Google 的 ScreenAI

在科技與創意交匯的領域中，Google 以其最新創新：ScreenAI，邁出了巨大的一步。想像一下，一個電腦不僅能看到，還能理解並與使用者介面（UIs）和資訊圖表中的視覺元素無縫互動的世界。這就是 ScreenAI 的承諾，一個尖端的視覺語言模型，旨在導航圖表、圖解、表格等複雜圖景。

ScreenAI 脫穎而出的特點是掌握 UIs 的視覺和文字內容，識別按鈕、圖示、文本框等元素，並了解它們的用途和信息。其區別在於其處理這些元素的上下文能力，識別它們在螢幕上的關係和功能。這是通過一系列創新技術實現的，包括 PaLI 架構用於整合圖像和文本數據，以及 pix2struct 的靈活貼片策略，適應各種圖像長寬比。

ScreenAI 的核心是在一個廣泛的數據集上訓練，該數據集包含來自多種設備的螢幕截圖，使用先進技術標記 UI 元素及其描述。這個模型不僅僅是關於理解圖像；它是關於橋接視覺信息和語言之間的差距，使其能夠生成問題解答、UI 導航和內容摘要任務。

Google 的 ScreenAI 代表了使技術更易於使用者訪問和直觀的重要里程碑。通過使機器能夠理解人們每天使用的圖形介面，ScreenAI 為我們的數字世界與使用者之間的更自然和高效互動鋪平了道路。

閱讀更多

透過 Microsoft Research 的 AI 輔助方法革新雲端服務監控

Microsoft Research AI 輔助 雲端服務監控 深度學習 效率 可靠性

2024-03-19

在複雜的雲端服務世界中，有效監控這些服務的挑戰正變得日益重要。Microsoft Research 已經向著解決這項挑戰邁出了一大步，引入了一種 AI 輔助的雲端服務監控方法。這種創新方法旨在提高監控的可靠性和效率，確保服務能夠順暢且高效地運行。

傳統的雲端服務監控設置方法，大量依賴於服務管理員的專業知識和試錯嘗試。這經常導致採取反應式的問題解決方法，只有在問題出現後才進行調整。Microsoft Research 的解決方案則提出了一種主動策略，利用人工智慧來預測和緩解潛在問題，以免它們成為問題。

這種方法的核心是一個深度學習框架，根據服務的獨特屬性推薦特定的監控器。這個模型利用資源和服務等級目標（SLO）類別的結構化本體論來進行精確推薦。例如，它可以高精度推薦服務的 API 或 CPU 使用情況的監控器，確保有效監控服務的所有關鍵方面。

這種 AI 輔助監控的一個最大益處是它能夠減少錯誤檢測和不必要的警報數量，這些可能會消耗資源並妨礙服務性能。通過分析 Microsoft 一年的生產事件，研究團隊識別出錯誤檢測的共同原因，並利用這些見解來改進他們的監控。

展望未來，Microsoft Research 計劃進一步開發這項技術，引入監控評分卡。這些評分卡將評估監控的性能，提供可行的見解以進一步提升監控策略。這代表著朝著更智能、高效和可靠的雲端服務監控邁出的關鍵一步，承諾將改變我們確保基於雲的應用和服務的健康和性能的方式。

承認 Microsoft Research 的杰出團隊的集體努力，這一雲端服務監控的突破性進展，體現了人工智慧在解決複雜技術挑戰上的力量。隨著我們進入更多依賴雲的計算環境，這些創新對於維持用戶期望和依賴的可靠性和效率至關重要。

閱讀更多

NVIDIA的Blackwell架構：生成式AI的巨大飛躍

NVIDIA Blackwell架構 生成式AI 加速運算 GPU Transformer Engine NVLink Grace CPU 雲服務

2024-03-19

NVIDIA，一家在加速運算領域的領導者，最近宣布了其新的Blackwell GPU架構的重大突破。這項創新將革命性地影響生成式AI領域，為處理複雜AI模型提供更具成本效益和節能的解決方案。

以著名數學家David Harold Blackwell命名的Blackwell架構引入了六項突破性技術。這些進展包括世界上最強大的晶片，擁有2080億個電晶體，以及第二代Transformer Engine，使計算能力和模型大小翻倍。它還擁有第五代NVLink，用於快速的多GPU通信，以及專門用於可靠性、安全性和數據解壓縮的先進引擎。

這個架構的核心是NVIDIA GB200 Grace Blackwell Superchip，它透過高達900GB/s的NVLink互連將兩個B200 Tensor Core GPU與一個Grace CPU整合在一起。通過結合多個GB200 Superchips，系統可以大規模擴展，達到驚人的1.4 exaflops的AI性能。

包括Amazon Web Services和Google Cloud在內的主要雲服務提供商，以及Dell Technologies和Meta等公司，已經計劃提供搭載Blackwell的產品。這種廣泛的支持凸顯了這個平台激發各行各業新進展的潛力，從工程和晶片設計到科學計算等領域。

有了Blackwell，NVIDIA不僅僅是推出了一個新的GPU架構；它正在為下一代AI能力鋪路。這可能會是將實時萬億參數AI應用從概念轉變為現實的引擎，標誌著向更先進和高效的生成式AI解決方案邁進的重要里程碑。

閱讀更多

用生成式AI革新工業作業：深入剖析AWS的最新創新

生成式AI AWS 工業運作 數據分析 機器學習 自然語言查詢 PandasAI Amazon Bedrock

2024-03-19

用生成式AI革新工業作業：深入剖析AWS的最新創新

在快速發展的製造業世界裡，人工智能（AI）與機器學習（ML）的結合不僅是一種趨勢；它正在革新工業運作的方式。從提升生產力到簡化複雜的數據分析，AI和ML正站在工業轉型的最前線。然而，這場創新之旅並非沒有挑戰。由傳感器和機械生成的大量非結構化數據構成了一個重大障礙。為每個獨特的工業場景開發定制ML模型既耗時又耗資源，經常成為AI解決方案廣泛應用的障礙。

進入生成式AI及其解鎖新潛能的驚人能力。通過利用大型預訓練基礎模型，如Claude，生成式AI通過從簡單文本提示生成多樣化內容改變了遊戲規則。這種稱為零次提示的方法，大大減少了數據科學家手動開發模型的需求，使AI在整個行業中更為普及，包括對於較小的製造商。結果如何？提升了生產力，主動異常檢測，優化庫存，並做出了明智的決策。

然而，旅程並未在此止步。傳統基礎模型在處理複雜工業數據時有其局限性，特別是受限於它們能處理的數據大小。為了克服這一點，AWS推出了對自然語言查詢（NLQs）的代碼生成的創新使用。像PandasAI這樣的Python庫擴展了生成式AI的能力，使得與複雜工業數據集進行更精確的互動成為可能。

真正的魔法在於引入了針對NLQs的多次提示。這種技術通過為模型提供多個所需輸出的示例，提高了代碼生成的準確性，顯著改善了模型在複雜查詢上的表現。這個方法利用了嵌入式表示，或文本的數值表示，提供了豐富的上下文提示，指導AI生成更準確的回應，用於如時間序列數據分析和高級數學計算等任務。

AWS並未止步於數據分析。Amazon Bedrock及其知識庫的引入簡化了生成式AI應用的開發，通過檢索增強生成（RAG）方法，實現了更準確的異常檢測和設備診斷。這一生態系統以各種方式支持工業運營，從維護團隊評估資產健康到促進異常的根本原因分析。

本質上，AWS的創新解決方案為工業運營設定了新的標準，使生成式AI的力量變得可訪問、可擴展且更有效。通過橋接複雜數據與可操作見解之間的差距，AWS不僅解決了當今的挑戰，還為製造業更高效、更有生產力和更多AI驅動的未來鋪平了道路。

閱讀更多

Amazon Bedrock 釋放生成式語言模型自我一致性提示的力量

Amazon Bedrock 生成式語言模型 自我一致性提示 機器學習 自然語言處理 AI 應用 隱私保護 道德 AI

2024-03-19

Amazon Bedrock 釋放生成式語言模型自我一致性提示的力量

在機器學習和自然語言處理 (NLP) 的領域中，Amazon 通過其 Amazon Bedrock 服務向前邁出了一大步。這個創新平台現在引入了一種稱為自我一致性提示的方法，顯著提升了生成式語言模型的性能——特別是在處理複雜推理任務時。

生成式語言模型是聊天機器人、虛擬助手和自動內容創建技術的核心。然而，這些模型在需要分析性推理或多步驟邏輯的任務上往往掙扎。Amazon Bedrock 的自我一致性提示方法通過生成多個潛在答案並選擇最一致的一個來提供解決方案。這種方法與依賴單一、確定性反應的傳統方法形成對比，後者經常導致準確性和創造力的限制。

在其核心，Amazon Bedrock 作為通往來自領先 AI 公司的大量高性能基礎模型的橋樑，通過單一 API 可訪問。這為開發人員提供了一個多功能的工具集，用於構建以安全性、隱私保護和道德 AI 實踐為重點的 AI 驅動應用。

自我一致性提示背後的方法論涉及三個步驟：通過提示引出推理、利用隨機解碼生成多樣化的推理路徑，以及聚合這些來確定最一致的答案。這項技術在算術推理和多項選擇題答題中展現了顯著的性能提升。

例如，在應用於算術問題解決時，自我一致性提示的模型在抽樣 30 個推理路徑時，準確率從 51.7% 提升到 68%。這展示了該方法顯著增強 AI 模型解決問題能力的潛力。

此外，Amazon Bedrock 的靈活性使得可以在不同的模型和任務上實施這種方法。在 AWS 認證考試題目上使用 AI21 Labs Jurassic-2 Mid 模型進行的測試中，自我一致性提示的模型超越了使用傳統方法的對手。

這一突破展示了 Amazon 對推進人工智能領域的承諾，為開發人員提供強大的工具，以創建更智能、響應更快、能力更強的 AI 應用。無論是用於教育、商業還是娛樂，這項技術的潛在應用範圍廣泛而多樣，承諾未來的 AI 能夠更有效地理解和與周圍的世界互動。

閱讀更多

NVIDIA 在創新與工程卓越方面居領導地位

NVIDIA 人工智能 AI CUDA NVIDIA Omniverse GPU 生成性 AI 高性能圖形處理單元
2024-03-19

在一項顯著的成就中，NVIDIA 被《快公司》雜誌評為全球最具創新性的公司。這一殊榮是對 NVIDIA 在技術領域，特別是人工智能（AI）領域的開創性貢獻的肯定。這一榮譽緊隨 NVIDIA 創始人兼 CEO Jensen Huang 被引入美國國家工程學院之後，凸顯了 NVIDIA 工作對工程領域的重大影響。

《快公司》的選擇受到 NVIDIA 在利用 AI 於各個不同行業（包括汽車、醫療保健、遊戲和零售）中的影響力所影響。NVIDIA 創新實力的一個關鍵亮點是其 CUDA 計算平台和 NVIDIA Omniverse 平台，這已經革命性地改變了與像 Nissan Z 運動型轎車這樣的物理對象的數位副本的互動方式。

NVIDIA 的貢獻不僅限於提供尖端的計算硬體；該公司積極參與塑造 AI 的未來，取得了他人在 AI 領域建立之上的重大進展。這種創新領導地位使 NVIDIA 與以往的領導者如 OpenAI 和製藥創新者 Moderna 和 Pfizer-BioNTech 身處同一聯盟。

《快公司》的排名基於創新、影響力、時效性和相關性等標準，表彰那些推出了突破性產品並促進了積極社會影響的組織。自1999年發明 GPU 以來，NVIDIA 一直處於現代 AI 時代的前沿，現在正推動加速計算和生成性 AI 的轉變，這正在以前所未有的速度轉變全球行業。

上個月，Jensen Huang 因其對高性能圖形處理單元的貢獻以及推動 AI 革命而被選入國家工程學院，進一步鞏固了 NVIDIA 作為技術創新領導者的地位。像 Microsoft 的 Satya Nadella 和斯坦福大學的 John Hennessy 這樣的傑出人物讚賞了 Huang 的遠見卓識領導力和對計算及各個領域的變革性貢獻。

NVIDIA 從增強 3D 圖形到主導 AI 革命的旅程強調了其持久的創新遺產以及其在塑造一個由智能計算驅動的未來中的角色。

閱讀更多

內容創作的未來：NVIDIA 最新的技術奇蹟

NVIDIA 內容創作 生成式 AI RTX 技術 數位藝術 Adobe Substance 3D OBS DLSS 3.5 Ray Reconstruction RTX Remix Omniverse Audio2Face iClone 動畫

2024-03-19

對於全球的創作者來說，NVIDIA 展示了一系列工具和更新，這些都承諾將革命性地改變內容創作過程，這無疑是一件令人興奮的事情。這些創新的核心在於生成式 AI 和 RTX 技術的進步，在全球 AI 會議 GTC 中展示。

對於那些涉獵數位藝術的人來說，Adobe 的 Substance 3D Stager 和 Sampler 通過 Adobe Firefly 設定接收重大升級。想像一下，僅從一個描述中創造出錯綜複雜的背景，或是無縫整合 3D 模型至這些環境中，帶來前所未有的真實感。這些工具不僅僅是關於創造；它們關於用精確和便捷將您最狂野的想像帶入生活。

視頻內容創作者也沒有被遺忘。OBS 30.1 beta 介紹了向 YouTube 的高動態範圍串流，感謝即時消息協議 —— 一項創新，確保您的觀眾在不妥協的情況下獲得最佳的視覺體驗。

但也許最吸引人的更新是將 DLSS 3.5 與 Ray Reconstruction 引入 RTX Remix Open Beta。這項功能對於模組製作人來說是一場遊戲規則的改變，允許顯著提升圖形保真度。更新後的 Steam 上的 Portal With RTX 作為這些能力的見證，展示了前所未有的增強光線追蹤圖像。

NVIDIA Omniverse Audio2Face for iClone 8 將面部動畫提升到新的高度，使創作者能夠僅從音頻輸入就產生逼真的表情和唇同步動畫。這個工具對於動畫師來說是一個福音，簡化了給角色賦予生命的過程。

這些更新強調了 NVIDIA 對於提升創意工作流程的承諾。隨著這些工具在內容創作領域中變得更加嵌入，我們即將見證到令人屏息的數位藝術、沉浸式視頻內容和模糊虛擬與現實界限的動畫的激增。歡迎來到由 NVIDIA 驅動的內容創作的未來。

閱讀更多

利用 Google 的 AI 技術全球性地革新洪水預報

Google AI 洪水預報 機器學習 災害管理 氣候變化
2024-03-20

利用 Google 的 AI 技術全球性地革新洪水預報

在創新的大躍進中，Google 正在利用人工智慧 (AI) 的力量，革新全球的洪水預報。洪水被認為是最普遍的自然災害，每年造成約 500 億美元的經濟損失，並對全球近 15 億人構成重大風險。Google 的這項計劃根基於提高全球對這類災難的抵抗力的承諾，旨在將準確及時的洪水預報擴展到服務不足的地區，每年可能挽救數千人的生命。

在這項開創性努力的核心是先進機器學習模型的開發，這些模型已經顯示出在缺乏傳統數據源的地區預測極端洪水的驚人能力。Google 在著名期刊 Nature 發表的研究概述了一種 AI 驅動的方法，將洪水預報的可靠性從幾乎零提高到提前五天，對非洲和亞洲的地區尤其有益。

Google 的 Flood Hub 是這一技術進步的見證，為 80 多個國家提供實時河流預報，最長可達七天。這個由機器學習驅動的工具，綜合了大量數據以對全球任何河流位置進行預測，克服了經濟不利地區數據稀缺的限制。

Google 的河流預報模型背後的方法論涵蓋了長短期記憶 (LSTM) 網絡和靜態流域屬性的複雜組合，使系統能夠區分不同的水文行為。這種方法不僅超越了傳統的洪水預報系統，還為早期警報機制打開了新的途徑，成為全球脆弱社區的希望之光。

隨著 Google 繼續擴大其洪水預報能力並與國際夥伴合作，這一倡議代表著我們集體對抗氣候變化和自然災害的關鍵一步。通過利用 AI 和機器學習，Google 不僅推進科學研究，還為數百萬面臨洪水風險的人提供了救生索，展示了技術在促進更具韌性的世界方面的轉型潛力。

閱讀更多

Microsoft Research：值得關注的創新技術

Microsoft Research 增強大型語言模型 個性化生產力工具 CoT-Influx GPT-4 LongRoPE Robin 生成式AI

2024-03-20

本週，Microsoft Research 充滿了可能重新定義我們與技術互動方式的突破性成果。從增強大型語言模型（LLM）推理能力、個性化生產力工具到擴展LLM上下文窗口，這裡有一些創新快照，為未來的發展奠定了基礎。

用CoT-Influx提升LLM推理能力

想像一個可以更有效地通過簡單過濾雜訊來解決複雜數學問題的工具。Microsoft 的 CoT-Influx 正是通過採用一種名為增強型上下文修剪的新方法來改善LLM的數學推理能力。通過為提示選擇和精煉思考鏈（CoT）示例，CoT-Influx 在LLM性能上取得了顯著改進，甚至在不需要任何微調的情況下超越了競爭對手！

定製化生產力觸手可及

在我們不斷追求效率的過程中，Microsoft 通過一個由GPT-4驅動的個性化生產力代理人邁出了一大步。這個代理人源於用戶調查和遙測數據，提供了可以適應個人工作風格和偏好的定製化協助，承諾以更直觀有效的方式處理日常任務。

用LongRoPE打破界限

是否因為LLM的限制而在處理文本數量上遇到瓶頸？LongRoPE在此打破這一障礙，將LLM的上下文窗口擴展到超過200萬個標記。這一創新意味著更深入、更全面的數據分析，無需昂貴的微調成本或性能折衷，徹底改變我們處理大量文本數據的方式。

Robin：對話式調試助手

調試常常感覺像在大海撈針。Robin是一個設計用來讓調試變得輕鬆的對話式AI助手。通過與開發人員進行合作對話，Robin不僅能識別錯誤，還能測試假設並提出修正建議，顯著減少解決時間並提高生產力。

導航生成式AI的諷刺

雖然生成式AI擁有極大的潛力來提高生產力，但它並非沒有缺點。Microsoft最新的研究闡明了「自動化的諷刺」，提供了對於GenAI系統有時可能會阻礙勝於幫助的洞見。通過正面應對這些挑戰，這項研究為更加用戶友好和高效的AI工具鋪平了道路。

請繼續關注Microsoft Research，了解這些創新技術如何很快轉變我們的數位景觀。

閱讀更多

在AI的虛擬領域中：NVIDIA推出新工具及應用程式加速AI開發

NVIDIA AI開發 生成式AI RTX AI平台 大型語言模型 TensorRT-LLM NVIDIA AI Workbench NVIDIA NIM微服務

2024-03-20

在AI的虛擬領域中：NVIDIA推出新工具和應用程式，以加速AI開發

在技術進化的核心，NVIDIA在GTC會議上的最新展示設定了PC和工作站上AI開發的新基準。揭示了一系列開發者工具和應用程式，代表了使生成式AI更加容易獲取和高效的重大進步。以RTX AI平台為先鋒，NVIDIA不僅在塑造未來；它正在革命化它。

其中一個亮點是「與RTX聊天」，一項技術奇蹟，將大型語言模型(LLMs)與用戶數據無縫整合，提供個性化AI體驗。這個工具是更廣泛的TensorRT-LLM生態系統的一部分，這個生態系統已經顯著成長，為超過500個AI應用程式承諾一個加速的未來。

但創新並未止步。NVIDIA AI Workbench和NVIDIA NIM微服務已經成為改變遊戲規則的因素，簡化了AI應用程式開發和部署過程。這些工具旨在使AI變得更加靈活，大大減少從概念到執行的時間。

生成式AI也隨著新的整合和更新得到了重大推動。從在遊戲中啟用即時、AI驅動的互動，到用生成式AI功能增強生產力應用程式，NVIDIA處於轉變數位景觀前沿。最近的開發者競賽獲獎者展示了這些技術的巨大潛力，應用範圍從AI輔助的電子郵件撰寫到理解人類語言的命令行介面。

NVIDIA對加速AI開發的承諾在其持續努力將TensorRT-LLM與流行的開發框架整合中顯而易見。這不僅承諾了增強的性能，也為開發者和愛好者提供了更大的可訪問性。

當我們站在AI新時代的邊緣時，NVIDIA在GTC會議上的公告作為開發者、研究人員和創造者的燈塔。揭示的工具和技術不僅僅是進步；它們是未來的建築塊，一個AI更加融合、直觀和有影響力的未來。

閱讀更多

利用深度學習革新分子科學：Microsoft的M-OFDFT模型

深度學習 分子科學 **M-OFDFT** 藥物發現 材料科學 計算效率 人工智能

2024-03-21

在一項突破性的發展中，Microsoft Research引入了一種新的方法，利用人工智能解決分子科學中最複雜的問題之一。該團隊由資深研究員劉昌（Chang Liu）領導，在分子性質計算領域取得了顯著進展，他們推出了M-OFDFT模型。

M-OFDFT，即分子軌域自由密度泛函理論（Molecular Orbital-Free Density Functional Theory），利用深度學習大幅提高識別分子性質的準確性和效率。這項創新可能對藥物發現和材料科學等行業產生深遠的影響，在這些行業中，了解和預測分子性質至關重要。

M-OFDFT解決的核心問題是計算分子的電子結構這一長期挑戰。傳統方法雖然在某種程度上有效，但往往涉及準確性和計算成本之間的折衷。然而，M-OFDFT利用深度學習顯著提高了這些計算的準確性，而不會帶來以往方法所伴隨的高昂計算成本。

這項技術在現有的無軌域密度泛函理論（OFDFT）的基礎上進行了構建，通過深度學習提高了其準確性，特別是在OFDFT表現不佳的分子系統上。結果是一種達到與傳統密度泛函理論（DFT）相同準確性水平的方法，但計算成本僅為一小部分。

實際上，M-OFDFT展示了驚人的結果。例如，在涉及一個含有700多個原子的蛋白質分子的測試中，M-OFDFT的速度幾乎是傳統DFT方法的30倍。這種速度和效率意味著，M-OFDFT可能會顯著加快分子性質計算和分子動態模擬的過程，特別是對於大分子而言。這種加速對於提高我們開發新藥和發現新材料的能力至關重要。

此外，M-OFDFT凸顯了人工智能在解鎖科學研究新機會方面的潛力。通過展示AI技術如何提高分子性質計算的準確性和效率，劉昌及其團隊的工作為最終打破分子科學中準確性和計算效率之間長期存在的折衷開闢了新的方法。

本質上，M-OFDFT不僅僅是分子科學的一步前進；它是向著一個未來的飛躍，在這個未來中，比以往任何時候都更容易地理解和操控世界的微觀構建塊。這可能會革新行業和研究領域，使得用高準確性和效率模擬分子世界的曾經不可能的任務成為可能。

閱讀更多

NVIDIA以LATTE3D模型革新3D生成技術

NVIDIA LATTE3D 3D生成技術 AI GPU 虛擬環境 視頻遊戲開發 廣告 設計 機器人訓練

2024-03-21

NVIDIA以LATTE3D模型革新3D生成技術

對於創作者和開發者來說，NVIDIA推出的LATTE3D模型是一個激動人心的前進步伐 - 這是一種生成式AI，能夠在不到一秒的時間內將文字轉換成詳細的3D形狀。這項由NVIDIA的AI研究團隊開發的突破性技術，承諾將徹底改變虛擬環境的創建方式，提供適用於視頻遊戲開發、廣告、設計和機器人訓練的應用。

就在一年前，生成相同質量的3D視覺效果還需要一小時。現在，由Sanja Fidler領導的NVIDIA團隊大大加快了這一過程，使幾乎實時的文字到3D創建成為了實際可能。LATTE3D的速度關鍵在於它能夠在單個GPU上運行，演示中展示了NVIDIA RTX A6000。

LATTE3D的靈活性是其另一強項。它不僅能夠從一個簡單的文字提示生成多個3D選項，還允許快速優化所選對象以獲得更高質量。此外，其潛力遠不僅僅限於其最初在動物和日常物體上的訓練。通過在不同的數據集上重新訓練模型，LATTE3D可以幫助設計多種環境，從茂盛的花園到完全家具齊全的家，從而促進廣泛的創意和開發過程。

這一創新建立在NVIDIA的A100 Tensor Core GPU上，並結合了由ChatGPT生成的多樣化文字提示，以增強其對用戶輸入的響應能力。憑藉這些進步，NVIDIA繼續推動AI和計算機視覺的界限，提供有望改變創意工作流程和跨行業應用開發的工具。

閱讀更多

AI新紀元的黎明：Transformer模型革命

Transformer模型 **生成式AI** **深度學習** **適應性計算** **NVIDIA** **GPU技術大會** **Attention Is All You Need**

2024-03-21

在最近一次具有歷史意義的GPU技術大會(GTC)上，NVIDIA的CEO Jensen Huang與《Attention Is All You Need》這篇具有開創性的研究論文背後的先驅者們坐下來討論。這場標誌性的對話匯聚了七位原始作者，他們的工作顯著改變了人工智慧的發展進程，催生了我們今天所知的生成式AI時代。

他們的發明，即Transformer模型，最初旨在解決當時盛行的遞歸神經網絡(RNNs)在處理語言時的缺點。然而，它超越了其最初的目的，不僅革新了語言處理領域，還改變了深度學習領域。這個模型的效率，類似於從蒸汽機到內燃機的飛躍，已成為推動當今生成式AI能力向前發展的基石。

作者們最初是Google的一個團隊，此後已經在AI產業中分散開來，每個人都在他們的事業中不斷推動AI的界限。他們一開始的共同願景很明確：創造出能夠理解和轉換任何數據類型的多功能AI模型，從文本和圖像到音頻等等。他們在這方面的成功是顯而易見的，他們的原始論文收到了超過100,000次的引用，他們的想法繼續影響著該領域的新發展。

展望未來，這些AI願景家們將“適應性計算”視為下一個前沿——AI模型將根據任務的複雜性智能地分配計算資源，優化效率和能源使用。他們預期會超越Transformer模型，開啟AI性能的新時代。

會議在一個感人的時刻結束，當Huang向研究人員們贈送了一個象徵性的禮物，以表彰他們對AI世界的巨大貢獻——一塊NVIDIA DGX-1 AI超級電腦的簽名蓋板，象徵著他們在改變世界中的作用。

這次對話不僅慶祝了AI歷史上的一個關鍵時刻，也為未來十年的創新定下了基調，承諾在人工智慧的世界帶來難以想象的新能力。

閱讀更多

03 資訊安全

用 AWS Nitro Enclaves 革新 AI 交互中的數據隱私保護

AWS Nitro Enclaves 數據隱私 大型語言模型 加密 隔離

2024-03-12

用 AWS Nitro Enclaves 革新 AI 交互中的數據隱私保護

在人工智慧領域的數據隱私保護上邁出了一大步，AWS 推出了一種創新的方式，以在大型語言模型（LLM）交互過程中保護敏感信息。這種創新解決方案利用了 AWS Nitro Enclaves，為處理個人和健康相關信息提供了一個避風港，無需擔心資料暴露的風險。

大型語言模型是我們日常遇到的許多智能功能背後的支撐，從回答我們查詢的聊天機器人到翻譯語言甚至生成代碼的系統。然而，這些強大的工具在處理敏感數據時，存在隱私泄露的風險。AWS Nitro Enclaves 通過創建一個加強的環境，將這些交互與潛在的漏洞隔離開來，正面解決了這一問題。

它是如何運作的？

想象一個場景，您需要一個聊天機器人來處理包含個人或健康信息的查詢。通常，這些敏感數據有被意外暴露的風險。然而，使用 AWS Nitro Enclaves，數據和交互被鎖在一個安全的堡壘中，這是雲中的一種虛擬堡壘。這個堡壘如此安全，以至於即使是擁有雲環境管理訪問權限的人也無法窺視其內容。

在這個堡壘內，數據保持加密，只在這個安全空間內被解密和處理。整個交易過程不會將信息暴露給更廣泛的雲環境或任何人眼。這個過程不僅創新，對於維護數字交互中的信任也是至關重要的，尤其是在像醫療和金融這樣隱私至關重要的領域。

幕後的技術天才

AWS Nitro Enclaves 通過幾個關鍵特性確保了這些操作的安全：

- 隔離： 通過利用 Nitro Hypervisor，Nitro Enclaves 將其 CPU 和內存從雲環境的其他部分隔離開來，確保敏感任務在一個與潛在威脅分開的隔間中處理。
- 無外部訪問： Enclaves 被設計為無外部訪問、持久存儲或互動功能，意味著在其中處理的數據保持機密且防篡改。
- 加密證明： 這一特性驗證了堡壘的完整性和身份，確保只有經授權的操作才能在其中執行。

賦能安全創新

通過將 AWS Nitro Enclaves 整合到他們的運營中，開發者和企業現在可以設計出在不妥協隱私的情況下自信處理敏感數據的應用程序。這標誌著邁向更安全、值得信賴的 AI 驅動交互的重要一步，為處理私人信息的行業開啟了創新的新門戶。

本質上，AWS Nitro Enclaves 不僅僅是一項技術成就；它是對數字時代數據隱私的一個承諾，確保隨著我們的交互變得更加智能，它們仍然保持機密和安全。

閱讀更多

透過 Amazon S3 存取點為 SageMaker 筆記本實現安全跨帳戶數據共享

Amazon S3 SageMaker 跨帳戶數據共享 IAM VPC 數據安全

2024-03-13

透過 Amazon S3 存取點為 SageMaker 筆記本實現安全的跨帳戶數據共享

在數據為王的世界裡，確保數據訪問的安全與效率至關重要，特別是對於從事尖端機器學習和人工智慧項目的數據科學家來說。Amazon S3 存取點是一項重要的進步，它極大地簡化了跨不同帳戶管理和保護數據訪問的方式，特別是對於共享數據集的應用程式而言。

想像一下，來自一個帳戶（我們稱之為帳戶 A）的數據科學家需要訪問另一個帳戶（帳戶 B）中存儲的數據集，以用於他們的項目。這些數據集可能用於從欺詐檢測算法到優化交易策略等任何事物。隨著使用案例的數量和複雜性增加，使用桶策略管理訪問權限的傳統方法迅速變得繁瑣且不足夠。這時，Amazon S3 存取點就派上用場了。

這些存取點作為獨特的主機名稱，為訪問共享數據的應用程式提供安全且特定的訪問權限。它們支援使用 AWS 的內部網絡和虛擬私人雲（VPCs）在同一區域內的高速數據傳輸，確保數據在必要時安全且高效地跨帳戶移動。此外，限制訪問到 VPCs 的能力意味著數據可以在私有網絡內有效地被防火牆保護。

設置此結構涉及配置數據擁有帳戶（帳戶 B，帶有 S3 桶中的數據集）和數據消費帳戶（使用 SageMaker 筆記本的帳戶 A）。通過定義特定的 IAM 權限和策略，以及為 SageMaker 筆記本創建專用存取點，確立了無縫且安全的跨帳戶數據訪問。

這種方法不僅解決了傳統桶策略所施加的大小限制，而且還允許按規模管理訪問權限，確保數據科學家可以在不妥協合規性或安全要求的情況下安全訪問他們需要的數據。

這項技術的潛在應用範圍廣泛，涵蓋了在數據安全、合規和高效訪問管理至關重要的行業和使用案例。它代表了公司管理其數據格局的方式，特別是在多帳戶環境中，向前邁出了重要一步。

對於有興趣設置跨帳戶訪問的人，AWS 提供了詳細的操作指南，確保數據科學家和 IT 專業人員可以順利且安全地實施這一解決方案。

總之，Amazon S3 存取點為跨帳戶管理數據訪問提供了一種強大且可擴展的解決方案，成為需要訪問共享數據集的數據科學家在不妥協安全或效率的情況下工作的寶貴工具。

閱讀更多

04 應用

Amazon SageMaker 為生物醫學應用革命性地微調蛋白質語言模型

Amazon SageMaker **蛋白質語言模型** **生物醫學** **機器學習** **藥物發現** **雲計算** **ESM-2** **Gradient Accumulation** **Gradient Checkpointing** **Low-Rank Adaptation**

2024-03-06

Amazon SageMaker 為生物醫學應用革命性地微調蛋白質語言模型

在機器學習與生命科學領域的一項顯著進展中，Amazon Web Services 展示了一種利用多功能的 Amazon SageMaker 來精煉 ESM-2 蛋白質語言模型的高效方法。這個模型，類似於語言模型理解文本的方式，解析蛋白質的複雜語言——這些生物分子對於各種身體功能至關重要，且是開發新藥的核心。

蛋白質由氨基酸組成，現在可以通過這個模型更準確地分析，預測它們的行為和相互作用。通過利用 Amazon SageMaker 的力量，研究者現在可以微調這個模型來執行特定任務，如預測蛋白質的細胞位置，這是理解其功能和作為藥物目標潛力的關鍵步驟。

這一突破代表了向前的一大步，使用了 Gradient Accumulation、Gradient Checkpointing 和 Low-Rank Adaptation 等尖端技術來提高訓練效率。這些方法不僅加速了訓練過程，還大幅降低了計算需求，使研究更加容易取得且成本效益高。

這種精煉模型的應用可能會革命性地改變我們對藥物發現和開發的方法，提供一種更快、更便宜、更準確的預測蛋白質行為的方法。這是機器學習和雲計算正在成為醫學進步不可或缺工具的一個閃亮示例。

通過利用 Amazon SageMaker 的能力，研究者可以突破蛋白質分析的界限，開啟了理解生命基石和開發治療各種疾病治療方法的新途徑。

閱讀更多

VistaPrint透過AI驅動的個性化推薦來革新小型企業的行銷

個性化推薦 小型企業行銷 Amazon Personalize 購物體驗 數據倉庫 客戶數據平台 Twilio Segment

2024-03-11

VistaPrint透過AI驅動的個性化推薦革新小型企業行銷

在小型企業行銷領域裡，VistaPrint透過利用Amazon Personalize的強大能力，革新了其產品推薦的交付方式，這是向前邁出的一大步。這種創新的方法不僅通過提供及時且相關的建議來增強購物體驗，而且顯著提高了VistaPrint的轉換率。

傳統上，作為尋求高品質行銷產品的小型企業的燈塔，VistaPrint面臨其本地部署產品推薦系統的挑戰，該系統難以快速擴展並適應變化的需求。轉向使用Amazon Personalize的雲原生系統不僅解決了這些問題，而且還將總體擁有成本降低了30%，同時將轉換率提高了令人印象深刻的10%。

VistaPrint成功的秘訣在於各種技術的無縫整合。通過聚合和轉換現代數據倉庫中的歷史數據並使用Amazon Personalize，VistaPrint現在能夠提供超個性化的產品推薦。這個系統通過分析過去的互動和類似項目的興趣，智能預測客戶需求，無論是在其網站上的實時還是通過針對性的電子郵件營銷活動。

此外，VistaPrint創新地使用像Twilio Segment這樣的客戶數據平台(CDP)，允許從多個接觸點收集全面的客戶數據。這些數據對於創建客戶的360度視圖至關重要，使得推薦更加精確和個性化。

VistaPrint從一個內部開發的本地部署解決方案到一個尖端的雲原生個性化推薦系統的轉變，體現了利用正確技術可以顯著增強客戶體驗和業務成果的方式。通過Amazon Personalize的力量和策略性數據整合，VistaPrint繼續設定個性化行銷的新標準，賦能全球的小型企業實現其全部潛力。

閱讀更多

革新線上互動：AWS 推出以 AI 驅動的聊天審查工具

AWS AI 聊天審查 線上互動 Amazon Transcribe Amazon Comprehend Amazon Bedrock Amazon OpenSearch Service 大型語言模型 生成式AI

2024-03-13

革新線上互動：AWS 推出以 AI 驅動的聊天審查工具

在一項突破性的舉措中，AWS 引入了一種新穎的方法來審查線上遊戲和社交社群中的語音和文字聊天。這個新方案利用了一系列 AWS 服務，包括 Amazon Transcribe、Amazon Comprehend、Amazon Bedrock 和 Amazon OpenSearch Service，來解決線上互動中日益增長的仇恨言論、網絡欺凌、騷擾和詐騙等問題。

這項創新的精髓在於它使用先進的 AI 和大型語言模型（LLMs）來自動化審查過程。傳統上，監控線上聊天以尋找有害內容一直是一項勞動密集型的任務，且高度可能出現錯誤和不一致。AWS 的最新工具旨在簡化這個過程，使其更有效率和有效。

這個解決方案採用了雙管齊下的方法，既適用於音頻也適用於文字聊天。對於音頻，過程從 Amazon Transcribe 將口語文字化開始，同時評估內容的有害程度。如果對話的有害程度超過了預定的閾值，系統就會啟動 Amazon Bedrock 的 LLMs 進一步根據特定政策分析內容。

文字聊天也進行了類似的審查，以 Amazon Comprehend 為先鋒在有害程度檢測上發揮作用。然後，被標記為潛在有害的訊息將由 LLMs 進行最終判定。

AWS 的審查工具之所以脫穎而出，是因為它們的靈活性。它們不僅支持即插即用的部署，還允許進行重大自訂化以與多樣的平台政策保持一致。這種簡單性、速度、成本效益和適應性的平衡使 AWS 的新解決方案成為確保更安全線上社群的一項變革工具。

通過整合生成式 AI，這些審查工作流程承諾能夠動態增強用戶安全性，適應線上通信不斷演變的景況。這項舉措代表了利用 AI 潛能促進更加包容和尊重的數位空間的關鍵步驟。

為了探索 AWS 審查工具的全部功能並獲得實際操作經驗，鼓勵使用者訪問 GitHub 上可用的範例代碼。這項技術進步凸顯了 AWS 利用 AI 創建更安全、更積極的線上環境的承諾。

閱讀更多

OpenAI 與 Le Monde 及 Prisa Media 合作，利用 ChatGPT 改變新聞體驗

OpenAI ChatGPT 新聞體驗 合作夥伴關係 Le Monde Prisa Media AI 數位新聞

2024-03-13

對全球新聞愛好者來說，一個令人興奮的發展是 OpenAI 宣布與 Le Monde 和 Prisa Media 建立戰略合作夥伴關係。此次合作旨在透過將如 El País 和 El Huffpost 等知名出版社的法文和西班牙文內容直接整合進 ChatGPT，徹底革新我們消費新聞的方式。想像一下，透過 ChatGPT 的創新 AI 技術，獲得最新、權威的新聞討論就在你指尖之間。

OpenAI 的首席運營官 Brad Lightcap 強調了這些合作夥伴關係的目標：利用 AI 來支持新聞業在我們社會中的關鍵角色，使全球用戶的新聞更富互動性和洞察力。不久，ChatGPT 的用戶將有機會接觸到精煉的新聞摘要，並附有完整文章的鏈接，以獲得深入閱讀體驗。

慶祝其成立 80 周年的 Le Monde 和西班牙語媒體巨頭 Prisa Media 相信，這種參與不僅將擴大他們的影響範圍，還將堅持他們提供經過驗證且平衡的新聞故事的承諾。這一進步被視為技術與人類專業知識的融合，旨在通過提供包括音頻和視頻在內的創新數位格式新聞，豐富讀者體驗。

此項舉措是 OpenAI 更廣泛願景的一部分，旨在開發賦能各行各業並解決複雜挑戰的 AI 工具，其中新聞業是主要焦點。通過這樣的合作夥伴關係，OpenAI 繼續支持新聞傳播的演進，確保 AI 用戶能夠訪問可信的資訊，同時維護新聞的完整性。

隨著我們開啟數位新聞的新章節，AI 與人類新聞業的協同作用承諾將世界帶得更靠近您，敬請關注。

閱讀更多

革新粉絲互動：PGA TOUR 的 AI 虛擬助理

AI 虛擬助理 PGA TOUR 生成式AI 檢索增強生成 Amazon Athena Amazon Kendra 大型語言模型 Anthropic Claude v2 Claude Instant 粉絲互動

2024-03-14

革新粉絲互動：PGA TOUR 的 AI 虛擬助理

在粉絲互動方面邁出驚人的一步，PGA TOUR 與 AWS 合作，推出了一款開創性的虛擬助理，旨在轉變粉絲與該運動互動的方式。這款由生成式 AI 驅動的助理被設計為能即時提供關於賽事、運動員等更多詳細回答，運用最新的 AI 技術。

這項創新的核心是一種稱為檢索增強生成 (Retrieval-Augmented-Generation, RAG) 的技術，結合了巧妙的文本到SQL方法。這使得助理能夠以自然語言理解問題，提取相關數據，並生成有用的答案。系統利用 Amazon Athena 執行 SQL 查詢和 Amazon Kendra 進行文檔的索引和搜索，確保快速而精確的信息檢索。

使這款助理與眾不同的是，它能深入 PGA TOUR 豐富的歷史數據，其中一些數據可以追溯到 1800 年代，並為粉絲提供互動式、對話般的體驗。無論是查詢 Tony Finau 的最長開球距離還是探索運動員的統計數據，助理都使用結構化和非結構化數據來提供引人入勝的內容。

開發過程是一個縝密的過程，涉及從 Amazon Bedrock 選擇最合適的基礎模型，並與不同的大型語言模型 (LLMs) 進行廣泛測試，以確保最高質量的回應。團隊最終選擇了 Anthropic 的 Claude v2 和 Claude Instant，並注意到它們在生成回應方面的卓越表現。

展望未來，PGA TOUR 旨在擴大助理的功能，使其成為一個個性化的、全渠道的體驗，可通過網絡、移動和語音界面訪問。這一舉措代表著利用 AI 提升粉絲互動的重要一步，使體育數據變得比以往更加易於訪問和引人入勝。

這個項目不僅體現了 AI 在轉變粉絲體驗方面的潛力，也為其他行業樹立了範例。通過這種創新的方法，PGA TOUR 確實處於利用技術豐富體育領域中粉絲體驗的前沿。

閱讀更多

聯邦學習：醫療數據共享與診斷的前進一大步

聯邦學習 醫療數據 心臟中風診斷 機器學習 數據隱私 醫學研究 AWS SageMaker 雲原生

2024-03-15

聯邦學習：醫療數據共享與診斷的前進一大步

在不斷進化的醫療領域中，及時診斷心臟中風始終是一個重大挑戰。在美國，每年有超過795,000人遭受中風，迅速、準確的診斷對於病人的結果有著重大的影響。傳統上，CT和MRI在診斷中風方面發揮了關鍵作用，但這一過程可能會相當耗時，尤其是在擁擠的急診部門。

引入聯邦學習（FL），這是一種突破性的機器學習（ML）方法，承諾將革新醫療數據的使用方式，使得更快的診斷和更明智的決策成為可能。這種分散式ML方法允許跨不同組織協同訓練ML模型，而不需要數據離開其原始位置的安全。通過保留敏感的病人數據在組織的防火牆內，並僅共享ML模型及其元數據，FL直面了隱私和數據安全的重大關切。

這種創新方法不僅加速了診斷過程，還通過利用之前在單個醫療機構內孤立的豐富多樣的數據集，提高了醫學研究的質量。通過FL，醫院、研究所和其他關注健康的組織可以協同改進模型的準確性和可靠性，同時保持最高標準的數據隱私和安全。

FL的潛力不僅限於心臟中風診斷。其在醫學成像、病人相似性學習和預測模型方面的應用，正在為醫學研究和病人護理的新時代鋪平道路。在像AWS SageMaker這樣的平台支持下，實施FL變得更加容易，提供了一種雲原生的方法，簡化了採用過程，而不需要與在地解決方案相關的重計算需求。

隨著我們走向一個更加互聯和以數據驅動的醫療領域，聯邦學習作為創新的一束光芒脫穎而出，承諾將數據孤島之間的鴻溝橋接起來，釋放醫療數據的全部潛力。這不僅僅是技術上的一步前進，更是通過協作和數據共享的力量，拯救生命和改變醫療的一大飛躍。

閱讀更多

顛覆道路：NVIDIA 與夥伴共同塑造車載 AI 的未來

車載 AI NVIDIA 生成式 AI 自動駕駛 AI 助理 安全駕駛 智慧伴侶

2024-03-18

革命性的道路：NVIDIA 與夥伴共同推動車載 AI 的未來

想像您的車不僅僅是一輛車，而是一位懂您需求、提供協助並以個人化的方式確保您安全的智慧伴侶。在人工智慧領域中的巨擘 NVIDIA，正透過其尖端的生成式 AI 技術，引領我們邁向這個未來。NVIDIA 與 Cerence、Geely 及 NIO 等行業領袖合作，轉變車載體驗，使其更安全、更舒適且驚人地直觀。

在這場創新的核心是 NVIDIA 的雲到邊緣技術，為從智慧 AI 助理到先進的駕駛員和乘客監控系統等一切提供動力。這些技術不僅僅是未來概念，它們正在 NVIDIA GTC 活動中展出，凸顯車輛如何演變為路上的智慧伴侶。

打造會對話的汽車

其中一項突出的創新是汽車助理的開發。這些數位副駕駛能夠進行自然對話，感謝 NVIDIA 的 Avatar Cloud Engine (ACE) 和大型語言模型 (LLMs)，它們能夠提供即時協助和個性化互動。想像一下，向您的汽車詢問最近的咖啡店或交通更新，並得到像人類一樣的回應！

眼見為實

但這不僅僅關於對話。NVIDIA 也在增強汽車感知世界的方式。透過 AI 強化的周邊視覺化，車輛現在能夠理解並詮釋 360 度相機視角，通過從每個角度識別潛在危險，讓駕駛更安全。

用生成式 AI 推動未來

NVIDIA 的幾個夥伴正在 GTC 上展示他們的最新進展。從 Cerence 那為下一代車載計算平台提供動力的專為汽車設計的 LLMs，到使用 NVIDIA DRIVE Thor 的 Waabi 自動駕駛貨車解決方案，生成式 AI 在汽車領域的潛力全面展現。其他創新包括 Li Auto 的 Mind GPT 用於場景理解，以及 NIO 的 NOMI GPT 通過由 NVIDIA AI 堆疊提供動力的高效計算平台提供大量功能體驗。

通過先進 AI 為車輛賦能

此外，MediaTek 的新型晶片系統將提供由 AI 驅動的車艙內體驗，確保車輛不僅理解並回應語音指令，還能在多個顯示器上提供豐富內容並確保駕駛員警覺。

隨著 NVIDIA 及其夥伴繼續推動這些 AI 驅動的創新，智慧型、自主型車輛的夢想正在成為現實。前方的道路令人興奮，很明顯，AI 不僅改變了我們的駕駛方式，還重新定義了我們與車輛的關係。敬請關注，看看交通運輸如何繼續演進，使我們的旅程更安全、更智能、更愉快。

閱讀更多

利用Google的SCIN資料集革命性地推進皮膚科研究

SCIN資料集 皮膚科研究 Google Research 多樣性 AI工具 隱私

2024-03-19

用Google的SCIN資料集革新皮膚科研究

在皮膚科領域裡，Google Research最近推出了皮膚狀況圖片網絡（SCIN），這是一個旨在豐富和多元化皮膚病學研究領域的全面性資料集。這項創新資源因其收錄了超過10,000張來自美國各地志願者提供的反映各種皮膚、指甲和頭髮狀況的圖片而脫穎而出。SCIN的特色在於其對常見但被低估的狀況，如過敏、感染和炎症，尤其是初期階段的表現，有所代表。

SCIN資料集的一個關鍵特點是其對多樣性的擁抱。它展示了不同膚色和身體部位的條件，解決了現有資料集常常忽略的膚色範圍的關鍵缺口，特別是那些比目前主流記錄還要深的膚色。通過包括自我報告的人口統計信息和皮膚類型，SCIN旨在使未來皮膚科領域的AI工具更加包容和有效。

創建SCIN時採用了一種新穎的群眾外包方法，通過網絡搜索廣告接觸到個人，並賦予他們直接為醫療研究做出貢獻的能力。這種方法不僅產生了高質量的資料集，還確保了廣泛代表了在健康系統中鮮見的病況。

Google Research通過重視貢獻者隱私和免費提供資料集，希望促進皮膚科研究、教育以及AI工具的發展。SCIN證明了合作式、包容性研究方法創建更具代表性的資料集的潛力，為醫學研究的進步以及跨不同人群更好的病患護理鋪平了道路。

閱讀更多

利用NVIDIA的AIOps生態系統轉型企業IT運營

AIOps NVIDIA IT運營 網絡安全 人工智能 自動化 雲服務

2024-03-19

利用NVIDIA的AIOps生態系統轉型企業IT運營

在不斷演進的企業IT運營領域中，變化的速度和複雜性往往超過了IT團隊的處理能力，NVIDIA正在引領一種突破性的方法。其AIOps合作夥伴生態系統正在革新企業處理IT運營和網絡安全的方式，利用人工智能的強大能力。這種AI和IT運營的融合，被稱為AIOps，自動化日常任務並強化安全措施，使對廣泛挑戰的響應更加迅速和精確，從簡單的技術故障到關鍵的安全威脅。

在NVIDIA策略的核心是其AI Enterprise軟件套件，一個加速計算的強大工具，賦予廣泛的合作夥伴將先進的AI驅動解決方案整合到他們的IT框架中。這些工具 - 包括NVIDIA NIM、Morpheus和NeMo - 滿足各種需求，從基於AI的網絡安全到生成AI聊天機器人、增強搜索功能等等。

特別令人興奮的是，從NVIDIA的合作夥伴中湧現出來的AIOps解決方案範圍，每一個都利用這些技術突破IT運營的界限。例如，Dynatrace Davis超模態AI整合了數種AI技術，為IT和商業運營提供更精確和可行的見解。同樣地，Elastic的Elasticsearch關聯引擎和New Relic的AI助手框架正在重新定義IT團隊如何監控、診斷和解決運營問題。

此外，像AWS、Google Cloud和Microsoft Azure這樣的雲服務巨頭正在利用NVIDIA的AI能力提供大量服務，通過自動化和優化增強IT運營，利用雲的龐大資源實現更大的效率和可擴展性。

NVIDIA的AIOps生態系統對企業的影響深遠。通過自動化日常任務和加強網絡安全，公司不僅可以簡化他們的運營，還可以加強自身以抵禦不斷增長的網絡威脅範圍。隨著數字景觀的持續演進，NVIDIA的創新正在為企業IT運營卓越設定新的標準，展現了一個IT運營更加安全、高效和前瞻性的未來景象。

閱讀更多

創新新創企業透過尖端AI對抗氣候變遷

AI 氣候變遷 NVIDIA Earth-2 預測天氣 洪水風險管理 海洋數據 高性能計算 永續運算
2024-03-19

創新新創企業透過尖端AI對抗氣候變遷

在氣候變遷對人類構成重大挑戰的世界裡，科技創新成為希望的燈塔。NVIDIA作為AI與永續運算的領導者，站在這場戰鬥的最前線，提供像是Earth-2這樣的平台，賦予新創企業革命性的能力，來預測、管理、並適應氣候變遷。

1. Tomorrow.io: 精準預測天氣

想像一下，如果能準確預測極端天氣條件，允許更好的準備和回應。總部位於波士頓的新創公司Tomorrow.io正在將這個想法變為現實。利用NVIDIA的Earth-2平台，Tomorrow.io結合先進的AI和機器學習技術及來自衛星和感測器的全球數據集，提供精準的天氣預報。他們的工作旨在簡化天氣預測的複雜性，讓每個人都能做出知情的決策。

2. ClimaSens: 洪水風險管理的新紀元

從墨爾本到紐約，ClimaSens正在改變洪水風險分析的遊戲規則。透過將歷史和即時氣候數據與未來天氣預測結合先進AI模型，ClimaSens的FloodSens模型提供潛在洪水的詳細評估。這種創新方法，由Earth-2和FourCastNet模型支持，代表著保護社區和促進永續未來的重要一步。

3. North.io: 揭開我們海洋的秘密

我們廣闊未知的海洋領域擁有全球食品安全和可再生能源的鑰匙。總部位於德國基爾的North.io專注於使用來自自主水下車輛(AUVs)的數據來繪製海底地圖。他們的TrueOcean平台，得到NVIDIA的Earth-2 APIs支持，不僅有助於收集和分析海洋數據，也有助於利用AI驅動的天氣預測來規劃和管理AUV操作。這項倡議顯著降低了探索危險的離岸環境中的人類風險。

這些作為NVIDIA Inception計劃的一部分的新創企業，展示了AI和高性能計算在對抗氣候變遷中的潛力。他們的工作不僅推進了我們對環境挑戰的理解和回應，也為更具韌性和永續的未來鋪平了道路。

請繼續關注更多氣候創新於NVIDIA GTC會議，探索最新的AI、高性能計算和永續運算在氣候研究方面的進展。

閱讀更多

NVIDIA的合作夥伴網絡獎突顯人工智慧創新

NVIDIA 人工智慧 合作夥伴網絡 加速計算 生成式人工智慧 金融服務 醫療保健

2024-03-19

在近期的創新與合作慶祝活動中，NVIDIA表彰了其美洲合作夥伴在推動各行各業的人工智慧驅動轉型中所作出的卓越貢獻。這種認可的核心是NVIDIA合作夥伴網絡（NPN）美洲年度合作夥伴獎，這些獎項聚焦於在整合人工智慧到他們運營中表現優異的組織，從而幫助跨部門的客戶在技術能力上邁進。

今年，NVIDIA引入了三個新的獎項類別，反映出人工智慧領域內日益增長的多樣性和專業化。這些類別旨在表彰那些不僅在人工智慧實施上表現優異，還在特定領域如金融服務展現出卓越奉獻的合作夥伴，或是在多個行業推進NVIDIA全面技術棧的發展。

在眾多傑出的得獎者中，Lambda因其利用NVIDIA加速計算和人工智慧企業技術的端到端人工智慧解決方案而受到讚譽。World Wide Technology因其透過NVIDIA的先進系統和解決方案促進人工智慧採用而脫穎而出。CDW因其在金融服務行業的影響力工作、Deloitte因其在生成式人工智慧的先驅使用、以及Mark III因其通過NVIDIA技術在醫療保健領域的進步而獲獎。

這些獎項凸顯了NVIDIA與其合作夥伴之間的互利關係，突顯了在人工智慧方面的協作和共享專業知識如何能夠帶來能夠轉變業務乃至於整個行業的開創性解決方案。透過在人工智慧上推動創新，這些合作夥伴關係不僅解決了複雜的商業挑戰，也為新的成長機會和效率鋪平了道路。

NVIDIA對其合作夥伴成就的認可，證明了在人工智慧領域中協作創新的力量。它展示了人工智慧技術的多樣化應用，從在金融服務中透過聊天機器人增強客戶體驗到推進研究和醫療保健解決方案，並凸顯了合作夥伴生態系統在實現技術飛躍中的關鍵角色。

閱讀更多

利用 Google 的 AI 革命性地改變肺癌篩查

AI 肺癌篩查 機器學習 Google Research 電腦輔助診斷系統 CT 影像 放射科醫師 Google Cloud 開源

2024-03-20

Google Research 在創新的大躍進中，引進了一項突破性的電腦輔助診斷系統，旨在改變肺癌篩查的方式。肺癌是全球癌症相關死亡的首要原因，尤其在診斷晚期時，呈現重大挑戰。儘管傳統的篩查方法有效，但一直受到偽陽性結果和大規模篩查處理效率低下的問題困擾。

Google 的解決方案利用機器學習（ML）來提高肺癌篩查的準確性和效率。通過採用先進的 ML 模型，該系統能夠自動檢測和分類計算機斷層掃描（CT）影像中潛在癌症的跡象。這些模型在識別可能的癌症方面，已展示出與專業放射科醫師相媲美的能力。

這項創新的核心是一個以用戶為中心的介面，旨在無縫整合進放射科醫師的工作流程。這個介面接收 CT 影像，並輸出一個跨四個類別的癌症懷疑評級，協助放射科醫師做出更加明智的決定。令人印象深刻的是，該系統已被證明能夠改善臨床醫生正確識別沒有可行動肺癌發現的病例的能力，可能減少不必要的後續程序，並減輕患者的焦慮。

部署在 Google Cloud 上的這項輔助肺癌篩查系統凸顯了 AI 在提升醫療保健交付中的力量。Google 的開放承諾顯而易見，他們已將處理 CT 影像的代碼開放給研究社群，鼓勵在這一關鍵領域進一步的進展。

這項倡議不僅僅是向提高肺癌篩查準確性邁進的一步；它是向著一個 AI 與醫療專業人員攜手合作拯救生命的未來邁出的一大步。透過夥伴關係和開源貢獻，Google 旨在使這項技術成為臨床設置中的主流，最終導致全球更可持續的肺癌篩查計劃。

閱讀更多

革命性的癌症偵測：引領方向的AI工具

AI工具 癌症偵測 早期偵測 個性化護理 臨床試驗

2024-03-21

在醫療保健領域的一項突破性發展中，一種名為Mia的新AI工具已成為早期癌症偵測的希望之光。這款創新工具展現出了驚人的能力，能夠識別出經驗豐富的醫生之前未曾注意到的癌症跡象。

在與英國NHS臨床醫生進行的試驗中，Mia經受了測試，分析了超過10,000名女性的乳房X光片。令人印象深刻的是，它不僅成功地檢測出所有參與者中的乳腺癌病例，還識別出了11個先前避開人類放射科醫生的額外病例。Mia在預測癌症存在方面達到了81.6%的令人驚異的準確率，在排除癌症方面的成功率為72.9%，展示了其革命性的癌症篩檢和診斷潛能。

乳腺癌是全球女性中最普遍的癌症，每年新增兩百萬病例。早期偵測和改進治療提高了存活率，但許多患者在治療後仍面臨嚴重的副作用。研究人員現在正在提高Mia的能力，以預測患者在治療後三年內經歷這些副作用的風險。這一進展可以使個性化護理計畫成為可能，為風險較高的人提供替代治療或額外支持。

Mia背後的團隊正準備通過一項名為Pre-Act的臨床試驗，進一步驗證這個AI風險預測模型，該試驗將涉及780名乳腺癌患者，持續兩年。最終目標是開發一個能夠對患者的預後和治療需求進行全面評估的AI系統，開啟個性化醫療保健的新時代。

這款AI工具的成功標誌著對抗癌症鬥爭中的一個重要里程碑，承諾了一個未來，在這個未來中，早期偵測和量身定制的治療計畫可以顯著改善患者的預後。

閱讀更多

連結夢想與現實：NVIDIA 對人工智慧未來的展望

人工智慧 NVIDIA 數位夢想 AI 互動 虛擬世界

2024-03-21

在不斷發展的人工智慧 (AI) 領域中，來自 Imbue 的 Kanjun Qiu 和 NVIDIA 的 Bryan Catanzaro 在 NVIDIA GTC 會議上分享了改變遊戲規則的觀點。他們深入探討了 AI 如何將單調乏味轉化為魔幻般的體驗，使我們的數位夢想成為現實。

想像一下，你對你的設備低語你的願望，它就像神燈精靈一樣，將它們實現。這種將 AI 從白日夢轉化為可行的數位行動的願景，標誌著向著一個想法與執行之間的差距大幅縮小的世界邁進的重大飛躍。然而，從我們今天的立場到這個未來情景的旅程充滿挑戰，主要在於讓 AI 無縫地理解和執行複雜的人類任務。

在他們的討論中，Qiu 和 Catanzaro 處理了圍繞 AI 的刻板印象，強調其自動化日常任務的潛力，從而豐富人類的創造力和生產力。他們突出了人類與 AI 互動中的當前障礙，指出對人類來說似乎簡單的事情，對 AI 系統來說往往是重大挑戰。

對話還談到了 AI 需要從基本的代碼生成器進化為成為連接人類與電腦的直覺介面的必要性。這種進化需要能夠與人類「理性」交談和與電腦「編碼」的 AI 系統。目標是什麼？使軟體創建像工業革命對製造業那樣，變得無所不在且具有變革性。

此外，Qiu 和 Catanzaro 探討了虛擬世界在個人計算的未來中的角色。他們設想 AI 不僅促進這些虛擬空間的創造，還將它們視為有效人類與 AI 互動的必要平台。這種整合預計將擴大人類能力，使科技成為我們生活中看不見卻賦能的力量。

這場在 NVIDIA GTC 的壁爐旁談話描繪了一個技術賦予每個人將其獨特願景實現的未來，打破夢想與行動之間的障礙。隨著 AI 的不斷演進，它承諾將改變我們與數位世界的互動方式，使我們最複雜的數位願望僅是一個簡單的指令之遙。

閱讀更多

利用 Contentful 和 Amazon Bedrock 革命化您的內容創作

Contentful **Amazon Bedrock** **AI 內容生成器** **內容管理平台** **生成式人工智慧** **隱私** **安全性** **語言模型**

2024-03-22

透過 Contentful 和 Amazon Bedrock 革新您的內容創造

在一次突破性的合作中，Contentful 和 Amazon 推出了一款創新工具，這款工具預計將轉變內容創造的世界。這款由 Amazon Bedrock 提供動力的 AI 內容生成器，現已在 Contentful 市場上提供，旨在簡化並提升創建、重寫、總結和翻譯內容的過程。

這款工具借助於通過 Amazon Bedrock 可訪問的生成式人工智慧 (AI) 模型的力量，讓內容創作者大幅減少發布高質量、一致性內容所需的時間。Contentful 是一個智能的內容管理平台，與 Amazon 合作，提供無縫且安全的內容生產體驗。

Amazon Bedrock 推出了一項全面管理的服務，允許立即訪問來自 AI21 Labs、Anthropic、Cohere、Meta、Stability AI 以及 Amazon 等行業領先大型語言模型的選擇。這項服務簡化了生成式 AI 應用的開發，同時確保了隱私和安全性。通過這次合作，Contentful 用戶可以直接在平台內使用這些模型，選擇最符合其語言風格、創意、速度和預算需求的模型。

其中一項顯著功能包括能夠自動重寫內容以適應不同渠道，如調整長度以適合較小的屏幕。這項功能可以在 Contentful 的介面內輕鬆訪問，促進內容適應的效率和創造性。

要開始使用，用戶可以免費註冊 Contentful 並從 Contentful 市場安裝 AI 內容生成器應用。安裝過程使用者友好，提供了定制和安全性選項，以符合組織政策。

Amazon Bedrock 與 Contentful 的整合為數位團隊帶來了遊戲規則的改變，開啟了內容生成、翻譯和定制的全新可能性，同時確保內容與品牌指南保持一致。這次合作不僅僅是關於提升內容質量；它關於革新我們的創造方式，使內容生成比以往任何時候都更加可訪問、高效和富有創造性。

閱讀更多

05 服務

OpenAI以董事會變革和治理增強迎接新時代

OpenAI 董事會變革 治理增強 Sam Altman Greg Brockman Sue Desmond-Hellmann Nicole Seligman Fidji Simo 審查 公司治理 利益衝突政策 使命與戰略委員會

2024-03-08

OpenAI以董事會變革和治理增強迎接新時代

在一項重大舉措中，OpenAI宣布在經過徹底審查後，Sam Altman和Greg Brockman將繼續擔任其領導角色。這一決定標誌著該組織的新篇章，因為它還迎來了三位新的董事會成員：Sue Desmond-Hellmann博士、Nicole Seligman和Fidji Simo。這些人憑藉他們在全球組織中的卓越職業生涯帶來豐富的經驗，承諾引導OpenAI的成長並確保其使命造福於人類。

由WilmerHale進行的審查對OpenAI的運營和領導動態進行了深入研究，檢查了超過30,000份文件並進行了眾多訪談。儘管去年11月發生了領導層變更的動盪，審查結論認為Altman和Brockman的領導對OpenAI的未來仍然至關重要。

這一新階段的關鍵是採納增強的治理結構，包括新的公司治理準則、加強的利益衝突政策，以及創建舉報熱線。這些改進措施，加上額外的董事會委員會，如使命與戰略委員會，旨在加強OpenAI對其核心使命的承諾。

隨著OpenAI向前邁進，董事會的擴大和治理增強表明了對於引導變革性技術以造福全球的堅定承諾。這一關鍵時刻強調了該組織在導航人工智能開發的複雜性同時確保其負責任進展的決心。

閱讀更多

OpenAI 歡迎新的遠見者加入其董事會

OpenAI 董事會 非營利 法律 人工普遍智能 全球法律和合規 消費技術
2024-03-08

OpenAI 歡迎新的遠見者加入其董事會

在向拓展其視野邁進的重要一步中，OpenAI 宣布了三位傑出專業人士加入其董事會，這標誌著其致力於擴大成長與創新的決定性步驟。新成員包括了 Dr. Sue Desmond-Hellmann，一位資深的非營利領袖以及比爾與梅琳達·蓋茨基金會的前CEO；Nicole Seligman，一位在 Sony Corporation 有著卓越紀錄的全球認可律師；以及 Fidji Simo，Instacart 的活躍CEO與董事長。

這些新增成員為 OpenAI 帶來了激動人心的新篇章，因為每位成員都在引領全球組織穿越複雜環境，包括科技、非營利部門與董事會治理方面擁有豐富的經驗。他們的集體專業知識預期將顯著貢獻於監督 OpenAI 的成長並確保其通過人工普遍智能為全人類的利益完成使命。

Dr. Desmond-Hellmann 的著名職業生涯涵蓋了在公共衛生和醫學方面的重大貢獻，Nicole Seligman 的法律專業知識使她能夠導航全球法律和合規事宜，而 Fidji Simo 在消費技術方面的廣泛經驗承諾將為 OpenAI 的前進之旅增添寶貴的見解。有了他們的加入，OpenAI 重申了其推動 AI 研究和應用邁向更大善的承諾。

閱讀更多

OpenAI宣布重大董事會改組及新治理結構

OpenAI 董事會改組 治理結構 Sam Altman 法律 透明度 責任 AI技術

2024-03-11

為了應對近期的挑戰，OpenAI宣布了一項重大的董事會改組計畫，以及新的治理結構。在經歷了一段動盪期，見證了Sam Altman短暫離開並重新回到CEO職位後，OpenAI通過增加三位新的獨立董事來加強其領導團隊。這三位新成員包括Bill and Melinda Gates Foundation的前CEO Sue Desmond-Hellmann；曾任Sony Corporation執行副總裁兼法務長的Nicole Seligman；以及Instacart的CEO兼董事長Fidji Simo。

OpenAI董事會主席Bret Taylor對新任命的董事表示熱烈歡迎，並強調他們預期對OpenAI發展人工普遍智能以造福人類的使命將作出貢獻。在這次重組期間，公司宣布了幾項治理改革措施。這些措施包括新一套公司治理指南、加強利益衝突政策、設立舉報熱線，以及成立額外的委員會，如一個旨在實施OpenAI核心使命的「使命與策略」小組。

面對因Altman短暫被解職引起的組織和公眾對治理的關注，這次董事會活化和治理改革應運而生。隨著對透明度和責任的重新承諾，OpenAI正準備以堅強的領導團隊和加強的治理結構，導航AI技術的未來。

閱讀更多

利用 AWS 解鎖生成式 AI 的力量：簡化指南

生成式 AI AWS Amazon Bedrock Amazon SageMaker 模型定制 提示工程 檢索增強生成
2024-03-14

利用 AWS 解鎖生成式 AI 的力量：簡化指南

生成式 AI 正在改變企業與數據互動的方式，為客戶體驗、生產力和流程優化帶來前所未有的創新機會。但是，要導航基礎模型 (FMs) 的複雜性以獲得這些好處可能是具有挑戰性的。這份指南將這個過程變得簡單，讓所有人都能輕鬆上手。

創新的基石：Amazon Bedrock 與 Amazon SageMaker

Amazon Web Services (AWS) 提供了豐富的資源，讓輕鬆建立生成式 AI 應用成為可能。首先，Amazon Bedrock 通過單一 API 簡化了與領先 AI 公司的高性能 FMs 的工作。這是一項完全管理的服務，讓您可以使用自己的數據私人定制模型，確保您的內容保持安全和符合 HIPAA 標準。這意味著您可以開始無憂無慮地且不必處理基礎設施管理的情況下，將生成式 AI 部署到您的應用中。

同時，Amazon SageMaker 充當一站式商店，用於創建、訓練和部署機器學習模型。有了 SageMaker JumpStart，探索和實施各種公共 FMs 變得輕而易舉，甚至讓非專家也能迅速深入生成式 AI 的世界。

建立生成式 AI 解決方案的方法

1. 提示工程 (Prompt Engineering)：這涉及到編寫提示來引導模型生成準確的回應。像是零樣本和少樣本提示這樣的技術，透過最少的例子調適模型至新任務，而思維鏈提示則將複雜的推理拆解成可管理的步驟。

2. 檢索增強生成 (Retrieval Augmented Generation, RAG)：RAG 通過納入新的或更新的信息來定制模型回應，顯著提升了在各種用例（如聊天機器人、寫作助手和摘要）中的表現。

3. 模型定制和微調：用您的數據定制 FMs，提升它們對您業務獨特背景的理解，確保生成的內容更能與您的受眾產生共鳴。Amazon Bedrock 和 SageMaker 促進了這一過程，允許更精確、特定領域的應用。

導航挑戰與最大化影響

採用生成式 AI 涉及克服數據隱私、安全性和確保高質量輸出等挑戰。然而，AWS 工具簡化了與現有系統的整合，使組織能夠在有效管理這些挑戰的同時利用生成式 AI。

開始您的生成式 AI 旅程

無論是增強客戶互動、簡化操作還是激發創新，生成式 AI 都提供了通往顯著業務價值的道路。AWS 提供工具和技術，使這一旅程盡可能平滑和有效，即使是技術知識有限的人也能解鎖生成式 AI 的潛力。

通過將複雜概念簡化為實際應用，AWS 正在使生成式 AI 的使用民主化，標誌著技術賦權在各行各業新時代的來臨。

閱讀更多

AWS 推出的 AI 驅動訂單處理代理人：顛覆客戶服務的革命

AWS AI 客戶服務 生成式 AI 大型語言模型 Amazon Lex Amazon Bedrock AWS Lambda 無服務器架構

2024-03-15

在快節奏的客戶服務世界中，企業持續尋找更有效率的方式來更好地服務客戶。AWS 認識到這一需求，推出了一項創新解決方案，該方案利用生成式 AI 的能力，特別是在像是自助咖啡店和快餐連鎖店這樣的環境中，轉變一對一客戶互動。

傳統上，這些場合主要依賴人類服務員，這種方法存在諸多挑戰，如規模擴張問題、錯誤的可能性，以及服務時間的限制。AWS 的創新方法利用生成式 AI 和大型語言模型 (LLMs) 開發出能夠迅速理解和處理自然語言請求的自動系統。

解決方案架構包括使用 Amazon Lex 建構能處理語音的訂單處理代理人，Amazon Bedrock 通過單一 API 存取一系列高效能的基礎模型，以及 AWS Lambda 進行無縫整合和操作。這三項 AWS 服務使企業能夠創建不僅能理解自然語言中的客戶請求，還能以驚人的效率作出回應和適應的代理人。

例如，當客戶下訂單時，系統可以確認菜單項目，必要時提供替代方案，並提供包括總成本在內的完整摘要——全部通過自然對話。這不僅通過使客戶體驗更加互動和個性化來增強客戶體驗，還簡化了訂單處理工作流程，顯著減少了手工錯誤和等待時間。

此外，這一解決方案建立在無服務器架構上，提供可擴展性和靈活性的同時，保持操作成本低。企業可以輕鬆自定義解決方案以適應各種使用案例，這得益於使用 Amazon S3 存儲提示模板和 DynamoDB 管理客戶數據。

通過將這種 AI 驅動的代理人整合到他們的服務渠道中，企業不僅可以確保更高水平的服務效率和準確性，還可以釋放人力資源以專注於更複雜的客戶需求。這代表客戶服務技術向前邁進了一大步，為未來該領域的創新提供了藍圖。

AWS 的新解決方案證明了生成式 AI 在重新定義客戶互動中的力量，對於希望在客戶服務的競爭性格局中保持領先的企業來說，這是一項令人興奮的發展。

閱讀更多