AI Trends

113 年 2 月號 人工智慧技術月報

01 模型技術

革命性預測：Google Research以TimesFM實現的突破

Google Research TimesFM 預測模型 時間序列 零售 金融 製造業 醫療保健 自然科學 預訓練語料 Google Cloud Vertex AI 機器學習

2024-02-02

革命性預測：Google Research以TimesFM實現的突破

在一項劃時代的發展中，Google Research推出了TimesFM，這是一種專為時間序列預測量身定做的全新解碼器模型。這種創新模型旨在解釋和預測各個領域的模式，包括零售、金融、製造業、醫療保健和自然科學。準確的預測的重要性不言而喻，因為它可以顯著提升零售業的庫存管理並增加收入，以及其他諸多好處。

TimesFM的特點在於它能夠在不需要額外訓練的情況下，對新的未見過的時間序列數據提供可靠的預測。這一特性是一大突破，讓使用者能夠立即獲得洞見並為特定應用調整預測，例如需求規劃。

TimesFM以其尺寸和效率區別於眾，儘管其參數只有2億個——相比我們已經習慣的大型語言模型(LLMs)要少，但其在多樣化數據集和領域的零槍性能幾乎與專門為這些任務訓練的最新模型匹敵。這一表現凸顯了該模型的堅固性和其革新多個領域預測的潛力。

該模型基於驚人的1000億實際時間點的預訓練語料運作，包括來自Google Trends和Wikipedia Pageviews的數據。這一龐大的數據集使TimesFM能夠把握複雜的模式和趨勢，顯著提高其預測準確性。TimesFM的架構利用了變壓器層和一種獨特的方法來處理時間點作為標記，這使得它能夠對各種預測範圍進行高效和準確的預測。

今年晚些時候，Google計劃將TimesFM整合到Google Cloud Vertex AI中，為外部客戶提供訪問這一強大預測工具的途徑。這一整合將使企業和研究人員能夠利用一種已經證明其在零槍設置中超越傳統和一些基於深度學習的預測方法的模型，來預測未來事件，擁有前所未有的準確性。

TimesFM承諾提供接近最新技術水準的預測結果，最初設置最少且無需針對特定數據集訓練，代表時間序列預測技術的一個重大進步。這不僅僅是機器學習向前邁進的一步——它是向著預測全球各行各業的不可預知變為可觸及現實的一大飛躍。

閱讀更多

Google對抗AI模型中偏見的新策略：早期讀出和特徵遺忘

AI 偏見 早期讀出 特徵遺忘 機器學習 深度學習 公平性 強韌性

2024-02-02

Google透過早期讀出和特徵遺忘來對抗AI模型中的偏見新方法

在針對機器學習中偏見進行鬥爭的重大進展中，Google研究人員引入了創新策略來解決AI模型中由於虛假特徵和簡單性偏見所帶來的挑戰。通過他們最新的研究，他們提議使用早期讀出和特徵遺忘作為解決這些普遍問題的有效方案。

機器學習模型經常因包含統計偏見的有限數據而產生偏差。認識到這一點，Google的團隊發現深度學習網絡由於所謂的簡單性偏見而傾向於放大這些偏見。

為了對抗這一點，研究人員建議採用早期讀出——從網絡的初始層面做出的預測——這可以揭示對虛假特徵的依賴。這種方法有助於識別模型由於這些偏見而錯誤地過於自信的情況，為在模型訓練中糾正這些不準確性提供了一條途徑。

此外，特徵遺忘技術直接針對簡單性偏見。通過識別然後「遺忘」早期學習的簡單特徵，鼓勵網絡發現更複雜和更準確的預測因子。這種方法，被稱為特徵篩選，交替進行發現問題特徵和抹除它們，讓模型專注於更相關的數據點。

這些方法在提高模型公平性和準確性方面顯示出了有希望的結果，橫跨各種數據集，展現了在機器學習公平性和強韌性方面的重大進展。Google對負責任的AI實踐的承諾是這項研究的基礎，凸顯了他們持續努力使AI對所有用戶更加公平和有效的努力。

閱讀更多

HSR.health 利用 Amazon SageMaker 打造創新的疾病監測技術

Amazon SageMaker 疾病監測技術 人畜共患病 地理空間能力 機器學習 衛星影像 風險指數

2024-02-05

HSR.health 利用 Amazon SageMaker 開創性的疾病監測技術

在健康風險分析領域中，HSR.health 利用 Amazon SageMaker 先進的地理空間能力，開發了一項創新的人畜共患病溢出風險指數。這一尖端工具將革新健康科學家追蹤和減少人畜共患病傳播方式——人畜共患病是指可以從動物傳播到人類的疾病。

人畜共患病，一個持續的全球健康威脅，一直是許多疫情爆發的核心，包括 COVID-19 大流行。COVID-19 的迅速全球擴散凸顯了更有效疾病監測系統的迫切需求。為了解決這個需求，HSR.health 的指數利用衛星影像、機器學習（ML）以及大量的地理、社會和環境數據的結合來預測疾病溢出的高風險區域。

通過整合超過 20 個已知影響人野生動物互動的不同因素，該指數提供了一個人畜共患病溢出風險的全面評估。這些因素包括通過衛星可見的環境變化，如森林砍伐，以及像人口密度和社會經濟地位等社會決定因素。利用 SageMaker 的地理空間能力允許有效處理這些複雜的數據，使創建詳細的風險地圖和預測模型成為可能。

這種方法不僅有助於早期識別潛在的疾病熱點，還增強了疾病預防和控制的資源戰略配置。對健康當局來說，這意味著能夠迅速有效地行動，從源頭控制疫情，潛在地挽救無數生命。

HSR.health 在疾病監測中利用地理空間 ML 的創新使用，體現了科技在應對全球健康挑戰中的力量。通過整合多元數據集和尖端 ML 技術，人畜共患病溢出風險指數作為持續對抗大流行的創新典範。

閱讀更多

探索資料未來：TensorFlow的圖神經網絡（GNN）

TensorFlow 圖神經網路 GNN 異構圖 tfgnn.GraphTensor Keras API 機器學習 數據分析
2024-02-06

探索資料未來：TensorFlow的圖神經網路（GNN）

在一個充滿複雜關係的世界裡，無論是在社交網絡、知識圖譜的錯綜複雜網絡，還是交通系統內的迷宮般連接中，理解這些網絡超越了僅僅觀察個別元素。認識到這一點，Google Research推出了TensorFlow圖神經網絡（TF-GNN）1.0，這是在如此規模上模擬這些相互連接的數據的一次革命性飛躍。

圖神經網絡（GNNs）已經革新了我們解釋和預測網絡內行為的方式，不僅考慮實體，還有它們之間豐富的關係網。無論是預測一個分子的性質、學術文章的主題關注，還是產品被一起購買的可能性，GNNs已證明是從複雜的關係性數據中提取意義的有力工具。

TF-GNN的特別之處在於其處理「異構圖」的能力——不同種類的對象和關係共存的網絡。這更準確地反映了真實世界，其中實體及其連接差異巨大。在TensorFlow內部，這些圖被表示為`tfgnn.GraphTensor`，一種封裝了圖的結構及其特徵的複雜數據類型，使其成為開發者和研究人員的強大盟友。

TF-GNN庫簡化了GNN模型構建的艱鉅任務，提供從高級Keras API支持創建自定義模型到動態子圖抽樣技術的一切，這些技術允許在龐大的數據集上進行高效訓練。無論是通過檢查其鄰居來理解節點的隱藏狀態，還是為復雜數據集訓練這些模型的編排，TF-GNN為初學者和專家都提供了在圖神經網絡領域內輕鬆導航所需的工具。

隨著我們站在機器學習和數據分析新時代的門檻上，TF-GNN 1.0邀請開發者、研究人員和技術愛好者探索圖數據的未開發潛力。通過其創新的方法來呈現和分析複雜網絡，TF-GNN不僅僅是一個工具，而是理解塑造我們世界的錯綜複雜關係的門戶。

閱讀更多

Microsoft揭曉AI控制器介面，革命性地推進生成式AI與安全性

Microsoft AI控制器介面 生成式AI 安全性 大型語言模型 輕量級虛擬機 敏感領域 客製化管理

2024-02-07

Microsoft研究院推出AI控制器介面，革命性地推動生成式AI與安全性\n\n在人工智能領域邁出重要一步的Microsoft研究院引介了AI控制器介面（AICI），結合了生成式AI與一種先進的輕量級虛擬機（VM）。這項創新技術預計將轉變大型語言模型（LLMs）的互動方式，確保與特定格式如JSON的高度準確性，同時在敏感領域加強保密性和安全性。\n\nAICI設計了獨特的\"指令即程式\"介面，允許用戶生成的代碼與雲端的LLM輸出生成無縫整合。它支援現有的安全框架和快速實驗，並能啟用特定於應用程式的功能。這系統提供對生成式AI基礎架構的細緻控制，無論是本地還是基於雲端的LLM處理，都能進行客製化管理。\n\n在這個介面的核心是一個輕量級的AI控制器VM，為開發者和研究人員提供了高效與LLMs工作的靈活性。這種設置不僅允許對輸出進行微調，還可以同時處理多個請求而不影響性能。此外，AI控制器程式可以為特定的應用或任務開發，在CPU上與GPU上的模型處理并行運行。\n\nAICI透過有效的受限解碼和資訊流約束展現其實力，確保文本創建遵循特定格式，並讓用戶控制提示和數據對輸出的影響。這項技術在使LLMs更可靠、安全和多樣化的應用方面標誌著一個重要的進步。\n\n展望未來，Microsoft對AICI的願景包括在各種LLM基礎架構中進行廣泛整合，並開發一套標準的AI控制器，承諾開啟一個控制更嚴密、效率更高、安全性更強的AI新時代。

閱讀更多

探索人工智慧的未來：來自 Microsoft 研究論壇的洞見

Microsoft AI 人工智慧 賦能 AI 解碼人類邏輯 經濟化 AI 從互動中學習 語言模型 數據分析 技術創新

2024-02-07

探索 AI 的未來：來自 Microsoft 研究論壇的洞見

在最近一次 Microsoft 研究論壇的開幕會上，科技愛好者和專家深入探討了人工智慧（AI）不斷演進的風景，分享了對於未來的承諾和面臨的挑戰的見解。該活動標誌著一系列對話和合作的開始，旨在促進 AI 研究領域領先思想的開放對話和協作。

Microsoft 研究院以開創性技術聞名，引領了幾項變革性 AI 倡議的討論，關注這些進步如何能成為我們日常生活和工作環境的盟友。主要亮點包括：- 賦能 AI：開發提供現實世界協助的系統，將 AI 從數位領域轉移到實體效用。- 解碼人類邏輯：解鎖人類推理的組件，增強 AI 的理解和互動能力。- 經濟化 AI：旨在通過降低大小和成本，使 AI 技術更易於取得，從而提升性能和可用性。- 從互動中學習：使 AI 能夠透過與用戶的直接互動進化，超越僅僅回答問題的角色。

在一個「閃電回合」中，研究人員重點介紹了正在進行的努力，以提煉預訓練的大型語言模型，解碼基礎模型，並在分子科學和決策支持創新方面領先。這些努力凸顯了 Microsoft 對於不僅推進 AI，而且確保其負責任地開發和部署的承諾。

此外，論壇還重點介紹了兩個引人注目的研究項目：- 增強語言模型推理：通過一種創新方法，稱為 LAyer-SElective Rank reduction（LASER），Microsoft 研究人員展示了一種通過在訓練後精緻化它們的權重矩陣，顯著提升語言模型性能的方法。這一突破承諾將提高 AI 在理解和生成人類語言的能力。- 優化數據分析：針對處理大量高基數數據集的挑戰，揭示了一種新的 cache-conscious 算法。該方法旨在簡化對於商業智能至關重要的 top-k 聚合查詢，實現顯著的效率改進，為更快、更準確的數據洞察鋪平了道路。

該論壇不僅展示了 Microsoft 研究院在 AI 方面的進展，而且還慶祝了六名其研究人員被認定為 2023 年 ACM Fellows，表彰他們對該領域的貢獻。

閱讀更多

利用 Amazon 的 AI 革新抵押貸款詐騙偵測

Amazon Web Services 機器學習 抵押貸款詐騙偵測 Amazon Fraud Detector 自動化 數據分析

2024-02-07

利用 Amazon 的人工智慧革新抵押貸款詐騙偵測

在創新的大躍進中，Amazon Web Services (AWS) 推出了一項先進的解決方案，用於偵測抵押文件詐騙，利用其強大的機器學習能力。這項突破性技術，在最近的 AWS 機器學習博客文章中詳細介紹，旨在顯著簡化審核流程，同時提升詐騙偵測的準確性。

透過使用 Amazon Fraud Detector，一個完全管理的服務，這種新方法自動化了在抵押文件中識別欺詐活動。該服務允許創建自定義的詐騙偵測模型，這些模型從歷史數據中學習，使得能夠配置符合特定業務需求的決策邏輯。Amazon Fraud Detector 簡化了風險決策工作流的協調，承諾提高詐騙預測的準確度以及審核過程的效率。

該解決方案的核心在於其處理和分析大量歷史數據的能力，這些數據儲存在 Amazon Simple Storage Service (Amazon S3) 中，以訓練一個專門設計來識別文件篡改和詐騙跡象的機器學習模型。一旦訓練完成，這個模型就可以準確預測新抵押申請中詐騙的可能性，使用業務定義的規則和複雜的算法相結合。

此外，該系統設計為自適應，具有基於新數據微調其檢測方法的能力，確保其隨著詐騙技術的演進而保持有效。部署此模型涉及一系列直接的步驟，從數據上傳和模型訓練到性能審查和部署，最終創建一個能夠解釋模型分數並根據預定義的規則觸發適當行動的檢測器。

這項技術不僅承諾使抵押申請過程對金融機構更安全，而且更高效，有潛力減少手動審查申請以發現詐騙所需的時間和成本。它證明了機器學習在革新傳統流程和增強安全措施方面的力量。

通過利用 Amazon Fraud Detector 的能力，公司現在可以以前所未有的精確度和自動化程度來處理抵押詐騙偵測，展示了向更智能、更安全的金融交易邁出的重要一步。

閱讀更多

NVIDIA在世界政府峰會倡導全球採用主權AI

主權AI NVIDIA 數據中心 GPU 雲計算 自主系統 經濟發展 數位自主權 大型語言模型 綠色數據中心

2024-02-12

在杜拜的世界政府峰會上，NVIDIA的執行長Jensen Huang發表了引人注目的討論，強調每個國家都需要發展和控制自己的人工智能（AI）基礎設施的迫切需求，他將這一概念稱為「主權AI」。Huang的願景與阿聯酋人工智能部長His Excellency Omar Al Olama共享，強調AI的潛力在於能夠透過擁有數據和智慧產出，封裝並保留每個國家獨特的文化、歷史和智識遺產。

根據Huang的說法，主權AI不僅是關於技術；它關乎將一個國家的精髓嵌入數位領域，使國家能夠確保其數位自主權並促進針對其特定需求量身打造的創新。這種方法在AI在經濟發展中的角色日益重要之際尤為關鍵，預測顯示到2030年，由於AI的進步，中東經濟將增加3200億美元。

為了強調各國採用主權AI的可行性，Huang指出，開發必要的基礎設施既不昂貴也不複雜。這項努力的核心是將一個國家的語言和文化數據編碼到大型語言模型中，透過NVIDIA的GPU的進步使這項任務變得可行。根據Huang的說法，這些GPU通過提供一個普遍平台來民主化AI，該平台支持從雲計算到自主系統的廣泛創新。

在一個有趣的轉折中，Huang反對長期以來的建議，即掌握計算機科學對於在資訊時代中蓬勃發展至關重要。相反，他認為，計算的未來涉及創建「會說人話」的技術，從而使編程對每個人來說都是可接近的，而不僅僅是那些擁有技術專業知識的人。

進一步鞏固阿聯酋成為全球IT樞紐和倡導主權AI的承諾，Moro Hub與NVIDIA合作，宣布建立一個綠色數據中心的計劃。這項舉措是將AI整合到各個領域的更廣泛運動的一部分，承諾帶來突破性的創新，這些創新超越了傳統資訊技術的界限。

隨著全球各國在AI對其未來的影響上掙扎，NVIDIA對主權AI的推動提供了一個藍圖，用於利用這一變革性技術，同時保留每個國家的獨特身份和自主權。

閱讀更多

NVIDIA 推動 AI 與視覺化新紀元，發布 RTX 2000 Ada 世代 GPU

NVIDIA RTX 2000 Ada GPU 視覺化 AI 加速 光線追踪 Tensor 核心 CUDA 核心 DLSS 3 AV1 編碼器

2024-02-12

NVIDIA 開創 AI 與視覺化的新紀元，推出 RTX 2000 Ada 世代 GPU

在 AI 加速設計與視覺化領域裡，NVIDIA 推出了其 RTX 2000 Ada 世代 GPU，為現代專業人士量身打造的性能與多功能性強大工具。這款尖端 GPU 承諾將大幅提升各種應用程序的速度——從複雜的3D環境到工業設計，使前一代的能力成為過去。

RTX 2000 Ada 之所以脫穎而出，是因為它與其前身 RTX A2000 12GB 相比，在專業工作流程中提供了高達1.5倍的性能提升。這種提升不僅僅是量的飛躍，更是質的飛躍，使各領域的專業人士能夠達到前所未有的效率和創造力水平。建築師、工程師、產品設計師和內容創作者現在可以享受加速的視覺化工作流程、快速的產品設計迭代和無縫的高解析度視頻編輯，這一切都得益於 GPU 的強大16GB記憶體和 NVIDIA RTX 技術的力量。

但真正讓 RTX 2000 Ada 脫穎而出的是它基於最新的 NVIDIA Ada Lovelace GPU 架構。這包括第三代 RT 核心，可提供高達1.7倍更快的光線追踪速度，第四代 Tensor 核心，能夠提高 AI 處理量達1.8倍，以及提升 FP32 處理量1.5倍的 CUDA 核心——同時保持相同的功耗效率。此外，該卡還引入了 DLSS 3 用於 AI 驅動的圖形增強和一個 AV1 編碼器，承諾相比 H.264 提供 40% 的效率提升，為廣播和內容創作者開啟新視野。

NVIDIA 對於提升專業工作流程的承諾進一步體現在支持 RTX 2000 Ada 的 RTX 企業驅動程序的新功能上。這些包括 Video TrueHDR，為網絡內容提供擴展的色彩和亮度，改進的視頻質量增強功能，以及將工作負載從 CPU 移至 GPU 以加快任務完成的工具。

通過全球分銷商以及即將由 Dell Technologies、HP 和 Lenovo 等領先技術提供商提供的 NVIDIA RTX 2000 Ada 世代 GPU，正準備重新定義 AI 加速設計和視覺化的風景，承諾專業人士能夠以更少的努力和更大的創造力實現更多。

閱讀更多

Microsoft 發布 GraphRAG：在敏感數據的大型語言模型發現中的一大進步

Microsoft GraphRAG LLM 知識圖譜 數據分析 機器學習 私人敘事數據 問答系統
2024-02-13

Microsoft 推出 GraphRAG：在敏感數據的 LLM 探索中邁出的一大步

在 Microsoft Research 的一項開創性開發中，一項被稱為 GraphRAG 的新技術即將革新大型語言模型（LLMs）分析和理解私人敘事數據的方式。這項創新旨在解決擴展 LLMs 能力至未經訓練數據的挑戰，為數據調查開闢新途徑，如在未探索數據集中的主題識別和語義概念分析。

GraphRAG 的突出之處在於利用 LLM 生成的知識圖譜進行文檔分析，顯著提升問答性能。與使用向量相似度進行信息搜索的傳統檢索增強生成（RAG）系統不同，GraphRAG 引入了一種新的方法。它採用由 LLM 創建的知識圖譜，在複雜信息中導航並在查詢時增強提示。這種方法在從分散的信息片段中合成洞見和全面理解大型文檔或集合方面顯示出顯著的改進。

例如，在探索來自新聞文章的暴力事件信息（VIINA）數據集時，GraphRAG 展示了其提供深入且相互關聯答案的能力，這是基線 RAG 系統無法匹配的。這在需要深入理解數據集主題和敘事的查詢中尤為明顯。

此外，GraphRAG 通過提供證據來源，引入了一層可靠性。GraphRAG 生成的每個回應都以數據集為基礎，並為斷言提供引用來源，使用者可以輕鬆地將 LLM 生成的結果與原始材料進行核對。

展望未來，Microsoft Research 計畫將 GraphRAG 應用於各個領域，同時專注於提升指標和評估框架。這項技術不僅承諾將提升 LLMs 處理私有數據集的效率，還將為從社交媒體分析到科學研究的領域鋪平新的研究和應用可能性之路。

閱讀更多

NVIDIA 以 Chat with RTX 革命性地改變了個人電腦使用體驗：您專屬的客製化聊天機器人

NVIDIA Chat with RTX GPT 聊天機器人 RTX GPU TensorRT-LLM 生成式 AI 隱私 安全 Windows PC

2024-02-13

NVIDIA 以 Chat with RTX 革新個人電腦：您自己的客製化聊天機器人

NVIDIA 發布了 Chat with RTX，一個免費下載的技術展示，標誌著個人電腦領域的一大飛躍。這項創新使任何擁有 NVIDIA RTX GPU 的用戶都能在他們的 Windows PC 上創建並與自己的個性化 GPT 聊天機器人互動。

Chat with RTX 利用最先進的技術，包括檢索增強生成 (Retrieval-Augmented Generation, RAG)、NVIDIA TensorRT-LLM 軟體，以及 NVIDIA RTX 加速，將生成式 AI 能力直接帶到您的本地 PC。這意味著更快的回應和更個性化的互動，因為聊天機器人可以訪問並從您自己的檔案和內容中學習。

想像一下，當您問您的 PC「我在拉斯維加斯時被推薦的餐廳是哪家？」並立即獲得帶有完整上下文的回答。Chat with RTX 可以掃描本地檔案 — 不論是文字文件、PDF，還是您提供的 YouTube 視頻 URL，並根據您自己的數據給出精確的資訊或建議。

在 Windows RTX PC 上本地運行的 Chat with RTX 確保您的數據保持私密和安全，處理敏感資訊而無需在線分享或依賴雲端服務。使用此功能需要 NVIDIA GeForce RTX 30 系列 GPU（或更高）且至少有 8GB 的 VRAM，以及 Windows 10 或 11 和最新的 NVIDIA GPU 驅動程式。

對於開發者來說，Chat with RTX 的潛力是巨大的。它是基於 TensorRT-LLM RAG 開發者參考項目構建的，為在 RTX GPU 上開發和部署客製化、加速的 RAG 基礎應用程序開啟了新的可能性。

NVIDIA 的 Chat with RTX 不僅僅是一個聊天機器人，而是一扇通往個人電腦未來的窗口，在這裡您的 PC 以深度個性化的方式理解並與您互動，使日常任務更快、更簡單、更直觀。

閱讀更多

Google 應對資料隨時間變化挑戰的新方法

Google 人工智慧 AI模型 概念漂移 CLEAR基準測試 Instance-Conditional Timescales of Decay for Non-Stationary Learning 訓練資料 物體分類

2024-02-14

Google 對抗資料隨時間變化挑戰的新策略

在一個唯一不變的是變化的世界裡，Google Research 開創了一種改進人工智慧（AI）模型隨時間學習不斷演變資料的方式的創新方法。這項創新特別關鍵，因為傳統的AI模型在訓練資料的特徵改變時，常常難以保持相關性—這種現象被稱為概念漂移。

傳統上，AI模型是基於歷史資料進行訓練，假設這些資料會持續代表未來的情境。然而，隨著世界的變化，這些訓練資料的相關性也在變化。Google的研究凸顯了這個假設的不足，透過CLEAR基準測試展示了在十年的時間裡，圖像中物體的外觀如何顯著演變，對物體分類模型構成挑戰。

為了解決這個問題，Google的團隊，由Nishant Jain和Pradeep Shenoy領導，開發了一種創新的方法，這種方法基於訓練資料的年齡和對當前條件的相關性動態調整每份訓練資料的重要性。這個方法，名為「Instance-Conditional Timescales of Decay for Non-Stationary Learning」，涉及一個輔助模型，該模型為訓練實例分配重要性分數，考慮到它們的內容和收集它們的時間長短。這種技術允許優化AI模型在未來資料上的表現，標誌著相比於傳統方法的重大進步。

他們的方法結合了兩個世界的最佳特點：它保留了過去資料的寶貴資訊，同時優先考慮最近的資料以保持模型更新。這種方法在物體分類任務中展示了高達15%的準確性提升，跨越了大約3900萬張照片的廣泛數據集，這些照片橫跨十年，展示了這項技術顯著增強AI模型隨時間保持相關性和堅韌性的潛力。

Google的創新策略不僅提供了解決概念漂移挑戰的解決方案，還為AI研究開辟了新途徑，強調了在不斷變化的世界中適應性學習的重要性。

閱讀更多

NVIDIA 在語音翻譯技術領域取得新突破

NVIDIA 語音翻譯 人工智能 LIMMITS 挑戰賽 多語言 P-Flow NVIDIA Riva
2024-02-14

NVIDIA 在語音翻譯技術上取得新進展

在人工智能領域取得顯著的飛躍，NVIDIA 團隊贏得了備受矚目的 2024 年 LIMMITS 挑戰賽，推出了一個突破性的語音翻譯模型。這項創新的 AI 現在能夠使用僅僅三秒鐘的語音片段，將一個人的聲音翻譯成七種語言中的任何一種。這項技術不僅捕捉到了講話者聲音的精髓，還能準確反映出孟加拉語、恰蒂斯加爾語、印地語、卡納達語、馬拉地語和泰盧固語的細微差別和口音 - 為超過十億的母語使用者賦予了聲音。

NVIDIA 的模型因提供個性化的文字轉語音翻譯而脫穎而出，超越了現有服務，在準確模仿目標語言的口音或講話者的聲音細節方面往往表現不佳。通過細緻的開發，團隊為結果中的自然度設定了新標準，使我們比以往任何時候都更接近真實的語音介面。

這項技術為個性化的多語言體驗鋪平了道路，範圍橫跨廣播、電信、教育、電子商務和線上遊戲等多個平台。這種創建多語言內容的能力，既輕鬆又準確，承諾在全球範圍內打破語言障礙。

被稱為 P-Flow 的這款模型，不久將被整合進 NVIDIA Riva，這是 NVIDIA AI 企業軟件平台的一部分。這項整合將使用戶能夠跨平台部署這項尖端技術，無論是在數據中心、個人系統還是雲服務上，都讓多語言語音翻譯變得前所未有的容易。

NVIDIA 的成就不僅僅是在比賽中獲勝；它是向加強全球通信邁出的重要一步，預示著一個語言障礙僅僅是過去時的未來。隨著這項技術的發展和開始重塑我們的世界，請繼續關注更多更新。

閱讀更多

革新製造業：Amazon SageMaker Canvas 推出無需編碼的異常檢測功能

Amazon SageMaker Canvas 無需編碼 異常檢測 製造業 機器學習 實時監控

2024-02-15

革新製造業：Amazon SageMaker Canvas 推出無需編碼的異常檢測功能

在製造業領域裡，Amazon 推出了其 SageMaker Canvas 的更新版，使得業內專業人士比以往更加簡單地檢測機器數據中的異常。這一創新的發展縮短了領域專家與機器學習技術之間的距離，賦予了具備深厚行業知識的人士創建和部署複雜模型的能力，而無需編寫代碼。

Amazon SageMaker Canvas 利用了一個無代碼介面，使用者可以輕鬆設計出能夠預測、分類或檢測操作數據中異常的模型。這對於製造業尤為重要，在這個行業中，及早檢測到不規則情況可能意味著在常規維護檢查和昂貴的系統故障之間的差異。

過程始於領域專家挑選關鍵數據特徵並訓練模型以識別正常運作模式。當異常出現時——比如，馬達溫度的突然升高——SageMaker Canvas 的模型將標記這些偏差，為工程師提供早期警告。

部署直接且適合實時應用，允許持續監控並對潛在問題立即做出反應。此外，對於那些旨在將模型無縫整合到現有系統的企業來說，SageMaker Canvas 模型可以直接作為實時端點共享或部署，由 AWS 的強大雲基礎設施所支持。

這一創新不僅使機器學習技術的使用民主化，還簡化了異常檢測過程，使其對業內專業人士來說更加可訪問和可行。憑藉 Amazon SageMaker Canvas，公司可以利用機器學習的力量提高運營效率，並在潛在問題影響生產力之前預先解決。

本質上，Amazon 對 SageMaker Canvas 的最新更新，將改變製造業在維護和運營效率方面的處理方式，提供一種智能的、可訪問的、高效的異常檢測解決方案。

閱讀更多

NVIDIA 發布 Eos：AI 超級計算未來的一瞥

NVIDIA Eos AI 超級計算機 DGX H100 Quantum-2 InfiniBand 生成式 AI 加速計算
2024-02-15

NVIDIA 發布 Eos：窺探 AI 超級計算的未來

在最近一次劃時代的揭示中，NVIDIA 揭開了其最先進的超級計算機 Eos 的神秘面紗，承諾將重新定義人工智慧（AI）的格局。Eos 作為 NVIDIA 非凡創新的證明，不僅是全球前 10 大超級計算機之一，而且旨在引領 AI 革命。

Eos 的核心是 NVIDIA DGX H100 系統，配備了 576 單位，總計提供驚人的 18.4 exaflops 的 FP8 AI 性能。這個強大的系統旨在處理最需求苛刻的 AI 任務，從訓練廣泛的語言模型到模擬複雜的量子現象。擁有 4,608 個 H100 GPU 的 Eos，不僅僅是一台超級計算機，更是 AI 領域可能性的燈塔。

Eos 配備 NVIDIA Quantum-2 InfiniBand 網絡技術，這一尖端技術支援高達 400Gb/s 的驚人數據傳輸速度。這種能力對於快速移動大型數據集至關重要，這是訓練複雜 AI 模型的基本要求。

Eos 的特點是其針對需要超低延遲和高吞吐量連接的 AI 工作負載的設計，遍及大型計算集群。這對於希望擴大其 AI 創新的企業來說，是一個理想的解決方案，提供了先進的加速計算、網絡和軟體解決方案的組合，包括 NVIDIA Base Command 和 NVIDIA AI Enterprise。

隨著世界更多地依靠生成式 AI —— 從醫療保健到自主技術的各個領域進行轉型 —— 對一個「AI 工廠」的需求變得明顯。Eos 滿足了這一需求，提供了一個始終可用、可擴展的引擎，以加速 AI 模型的開發。其為 AI 優化的架構，為全球開發者和企業提供了關鍵資源，承諾加速走向 AI 融合應用的旅程。

NVIDIA 的 Eos 不僅展示了其技術在大規模運作時的潛力，還鞏固了 NVIDIA 推動 AI 技術和基礎設施界限的承諾。憑藉 Eos，NVIDIA 不僅僅是開啟了每一天的大門，如其名字所示，還開啟了 AI 創新的未來。

閱讀更多

OpenAI 推出 Sora：影片生成的未來

OpenAI Sora 影片生成 Transformer 語言理解 數字模擬

2024-02-15

OpenAI 推出 Sora：影片生成的未來

在透過科技模擬實體世界方面，OpenAI 進行了一次重大的飛躍，推出了革命性的影片生成模型 Sora。Sora 脫穎而出的特點在於其能夠生成高保真度、長達一整分鐘的影片，涵蓋了廣泛的時長、解析度和長寬比。這一創新模型基於將影片和圖像轉化為統一表徵的基礎之上運作，允許進行大規模的生成模型訓練。

Sora 採用了 Transformer 架構，利用視頻和圖像潛碼的時空片段，使其能夠熟練地處理廣泛的視覺數據類型。這個模型一個顯著的特點是其擴散機制，能夠細緻地精煉輸入數據來產生清晰而詳細的影片輸出，隨著訓練計算量的增加，顯著提高樣本質量。

但這不僅僅是關於生成隨機影片。Sora 擁有獨特的能力，理解語言，將文字提示轉化為生動、高質量的影片，準確遵循用戶指令。這個模型超越了文字到影片的轉換；它可以在時間上延伸影片，甚至根據用戶提示編輯影片，展現了其靈活性和在從內容創建到數字模擬等廣泛應用領域的潛力。

此外，Sora 的能力包括生成具有驚人細節的各種尺寸的圖像，以及其對模擬物理和數字世界的各個方面的潛力，包括 3D 移動和長期連貫性，凸顯了影片生成技術的光明前景。

雖然 Sora 展現出優異的能力，但它仍在成長的道路上，持續發展以克服諸如準確模擬物理或維持物體一致性等限制。儘管如此，OpenAI 在 Sora 上取得的進展照亮了向創建我們世界的通用目的模擬器邁進的令人振奮的旅程，橋接了數字模擬與現實之間的差距。

閱讀更多

利用 Amazon SageMaker 的動態端點轉型 ML 模型管理

Amazon SageMaker **ML 模型管理** **動態端點** **計算資源調整** **DJLServing** **成本節省** **效率提升**

2024-02-19

利用 Amazon SageMaker 的動態端點轉型 ML 模型管理

Amazon 為處理機器學習（ML）模型在變動和不可預測的流量需求下所推出的一項開創性更新。Amazon SageMaker 多模型端點（MMEs）現在擁有一項創新功能，能夠根據實際流量動態調整計算能力，為開發者和企業帶來顯著的成本節省和效率提升。

之前，SageMaker MMEs 靜態分配計算資源，導致在變動的流量負載下產生效率低下。這種新方法利用 DJLServing，一個高性能模型服務器，根據每個模型的流量模式動態調整計算資源。這意味著無論是少數模型接收到請求的激增，還是流量均勻分布，SageMaker 都能有效管理資源，無需閒置計算能力的浪費。

這項能力對於經營多個 ML 模型的企業特別有利，使他們能夠在單一端點上部署數千個模型，無需靜態資源分配。現在，不常用的模型可以根據需要動態加載，確保後續請求的較快響應時間。

實際上，這種動態調整允許更不依賴於流量模式的方法。接收到突然激增的請求的模型可以自動分配更多計算資源，確保不需要手動干預即可保持一致的性能。對於企業來說，這不僅代表了運營成本的降低，也為他們的客戶提供了更可靠和高效的服務。

DJLServing 的整合進一步增強了 SageMaker MMEs 的靈活性和可擴展性。它支持多個 ML 框架，並提供自動縮放和動態批處理等功能，使管理多樣化的 ML 模型組合比以往更加容易。

隨著公司繼續在其運營中利用機器學習，能夠有效地管理和根據現實世界需求擴展 ML 模型的能力變得至關重要。Amazon SageMaker 的最新創新滿足了這一需求，提供了一種強大的解決方案，簡化了 ML 模型的部署和擴展，確保企業能夠在快速發展的數位景觀中保持適應性。

閱讀更多

Amazon Titan 在AI驅動影像創建和搜尋上開創新紀元

Amazon Titan Image Generator Titan Multimodal Embeddings AI驅動 影像創建 影像搜尋 內容創作者 安全措施 水印 負責任的使用

2024-02-19

Amazon Titan 開啟AI驅動影像創建和搜尋的新領域

Amazon 已經推出兩項革命性技術，Titan Image Generator 和 Titan Multimodal Embeddings，致力於轉變我們創建和搜尋影像的方式。Titan Image Generator 對內容創作者來說是一項遊戲規則的改變者，它能夠從簡單的英文文字提示創建高品質、逼真的影像。想像一下，只需幾個字就能產生廣告或娛樂用的工作室品質影像 - 這就是 Titan Image Generator 的力量。這個先進的AI理解包含多個物體的複雜指令，提供如提示迭代、自動背景編輯以及生成多種場景變化等功能。此外，它允許創作者融入他們自己的數據以製作個性化、符合品牌的影像，確保每個輸出都是獨特和量身定制的。

另一方面，Titan Multimodal Embeddings 模型旨在重新定義我們如何搜尋和推薦內容。通過理解文字和影像，它能將它們轉換成語義向量，使搜尋結果更加精確。不論你是將產品描述與照片結合用於線上商店，或是尋找精確的推薦，這個模型簡化了過程，提供快速且精確的結果。

這兩項技術不僅在其各自領域提供了前所未有的便利和效率，而且還配備了內建的安全措施，如在AI生成的影像上加入看不見的水印，促進負責任的使用並防止錯誤資訊的傳播。這些創新標誌著朝向更易於使用、高效和安全的影像生成和搜尋的一大步，為創意和內容發現開闢新途徑。Amazon 致力於推動AI應用的界限，使這些複雜的技術可供全世界的開發者和創作者使用，這一點是顯而易見的。

閱讀更多

革命化內容本地化：ZOO Digital憑藉AI技術的飛躍

ZOO Digital AI技術 本地化 WhisperX Amazon SageMaker OpenAI Faster Whisper Wav2Vec2 pyannote

2024-02-20

革命化內容本地化：ZOO Digital憑藉AI技術的飛躍

在內容比以往更快跨越國界的時代，將電視節目和電影本地化以匹配不同語言、地區和文化的挑戰變得日益複雜。作為提供端到端本地化和媒體服務的領導者，ZOO Digital正在開創一種利用突破性AI技術簡化此過程的解決方案。

傳統上，本地化一集30分鐘的節目可能需要長達3小時，因為在配音前需要進行的講話者分離（即確定何時由誰發言）涉及手動工作。ZOO Digital的雄心壯志是什麼？將這個時間縮短到30分鐘以下。

踏入WhisperX模型，這是自動講話者分離領域創新的一大亮點。部署在Amazon SageMaker上，這個模型證明了AI在轉變媒體內容工作流程中的力量。WhisperX建立在OpenAI的Whisper之上，不僅轉錄音訊，還能識別不同的講話者，顯著減少內容本地化涉及的時間和勞動。

但是，是什麼讓WhisperX脫穎而出？它結合了多項AI技術，包括用於快速轉錄的Faster Whisper、用於精確時間戳對齊的Wav2Vec2，以及用於準確識別講話者的pyannote模型。每個組件都在確保配音和字幕工作能夠滿足全球內容消費的快節奏需求中發揮至關重要的作用。

ZOO Digital與AWS Prototyping的合作不僅展示了WhisperX在現實場景中的有效性，還突顯了節省成本和時間的潛力。通過自動化分離過程，ZOO Digital可以利用AI支持並增強其廣泛的自由職業者網絡的技能，確保以前所未有的速度提供高質量的本地化。

這項技術進步不僅是ZOO Digital的勝利，也是全球內容創作者和消費者的勝利。隨著我們繼續擁抱數位時代，像WhisperX這樣的AI技術正在為我們跨文化分享和體驗故事設定新的標準。

這項舉措展示了機器學習和AI加速內容本地化的轉型潛力，隨著技術持續發展，這是一個令人興奮的觀察領域。

閱讀更多

Microsoft Research 揭示尖端創新

Microsoft Research 雲端運算 CaaSPER 自動擴展算法 CPU 資源 攝影機定位技術 增強實境 虛擬實境 機器人技術 神經網絡 企業系統可用性量表 ESUS

2024-02-21

Microsoft Research 發表前沿創新技術

在優化雲端運算的最新進展中，Microsoft Research 推出了一種名為 CaaSPER 的創新垂直自動擴展算法。這一工具旨在動態調整雲中單體應用（如有狀態的資料庫）的 CPU 資源，以確保最佳利用並最小化資源浪費。這種反應與預測策略的獨特結合，為節省成本和提升性能提供了重大進展，展示了在高效雲資源管理方面的一大步。

Microsoft Research 的另一項顯著成就是在攝影機定位技術方面的進步，這對於增強實境、虛擬實境和機器人技術至關重要。通過開發更準確的場景地標檢測方法，研究人員顯著降低了當前定位技術的儲存和速度限制。這種新方法利用緊湊型神經網絡，提供了與傳統方法相同的準確度，但效率更高，標誌著沉浸式技術應用的一個重大突破。

此外，Microsoft Research 通過引入企業系統可用性量表（ESUS）改進了企業應用的可用性評估。這一新問卷簡化了可用性測量過程並與企業環境的當前需求保持一致，提供了對用戶滿意度更準確的反映。ESUS 代表了對於高科技產品和服務用戶互動理解的演進。

來自 Microsoft Research 的這些創新不僅展示了在雲計算、攝影機定位和可用性評估方面的重大進步，而且還承諾將引入跨越各種技術應用的新功能和效率。

閱讀更多

Google利用VideoPrism在視頻理解方面取得新突破

Google VideoPrism 影片理解 基礎視覺編碼器 算法 影片分析 技術成就 應用範圍

2024-02-22

Google憑藉VideoPrism在影片理解領域取得新進展

在影片充斥我們數位世界的時代，Google最新的創新，VideoPrism，提供了一種開創性的影片理解方法。這個基礎視覺編碼器利用超過618百萬影片剪輯的龐大資料集，結合高品質的影片-文字配對與機器生成的字幕來訓練其精密的算法。

設計用於掌握從分類和字幕生成到複雜問題回答等一系列影片理解任務——VideoPrism代表了一個重大的飛躍。它脫穎而出的地方在於有效地消化影片所提供的動態豐富畫面，不僅捕捉外觀，還有複雜的運動和內在關係。

VideoPrism真正革命性的地方在於其適應性和應用的廣度。使用單一未經更改的模型，它在各種基準測試中達到了無與倫比的表現，決定性地超越了現有方法。它不僅僅是識別影片中的內容；VideoPrism理解本質、運動和上下文，為影片分析設定了新的標準。

超越其令人印象深刻的技術成就，VideoPrism的潛在應用範圍廣泛。從增強如行為學和生態學等領域的科學研究到轉變教育工具和醫療診斷，其影響可能是深遠的。Google承諾在這一領域進一步負責任的研究，預示著更多令人興奮的發展在前方。

VideoPrism不僅是技術上的勝利；它是影片分析未來的一扇窗口，承諾解鎖對以影片形式捕捉的世界更深入的理解。

閱讀更多

革命性的3D設計：Rhino 3D 與 OpenUSD 釋放新可能性

3D設計 Rhino 3D OpenUSD 計算機輔助設計 3D建模 NVIDIA Omniverse 數位化

2024-02-22

革命性的3D設計：Rhino 3D與OpenUSD釋放新可能性

在3D建模和設計的領域中，一項突破性的更新已經到來，承諾將改變設計師和創造者將他們的想像概念實現的方式。Rhino 3D，領先的計算機輔助設計（CAD）軟體，已經引入了對OpenUSD（通用場景描述）的支持，這一舉措將革命性地改變3D建模的格局。

OpenUSD作為一個強大的框架，使得跨各種應用程序的無縫協作和效率成為可能。這一整合不僅提升了3D建模體驗，還為更加動態和互聯的工作流程鋪平了道路。無論是用於建築設計、珠寶創造，還是海洋模型製作，Rhino 3D的最新更新都使專業人士能夠以前所未有的輕鬆和靈活性實現他們的願景。

這次更新的一個亮點功能是能夠直接從Rhino中導出OpenUSD文件，這大大簡化了設計過程。對於像Tanja Langgner這樣的藝術家和設計師來說，這意味著可以在應用程序之間無縫轉移資產，優化可視化，並以更高的精確度將創意想法實現。同樣，像來自New Jersey Institute of Technology的Mathew Schwartz那樣的教育工作者和研究人員，現在可以利用OpenUSD和NVIDIA Omniverse平台將設計、分析和可視化融為一體的工作流程，開啟設計研究和教育的新視野。

Rhino 8的發布帶來了一系列旨在提升3D建模體驗的增強功能。值得注意的改進包括高級建模功能，如PushPull直接編輯，引入ShrinkWrap功能以創建水密網格，以及對細分曲面的更好控制。此外，用戶可以期待一個更加流暢的繪圖和插圖過程，具有提升精確度的增強功能和更好的紋理協調，確保他們的設計細節被以最高的忠實度捕捉。

展望未來，Rhino計劃進一步擴大其導出能力，使OpenUSD文件在各種3D環境中更加易於訪問。這一持續的發展突顯了對培育一個更加整合和高效的3D設計生態系統的承諾。

對於那些對OpenUSD的潛力以及最新的3D建模進步感到好奇的人來說，即將到來的NVIDIA GTC會議是一個不容錯過的活動。在這裡，與會者可以深入了解OpenUSD的世界，從專家那裡學習，並發現生成性AI啟用的3D管道是如何塑造工業數位化的未來。

當我們站在3D設計的新時代門檻上時，Rhino 3D對OpenUSD的擁抱標誌著一個重要的里程碑，承諾解鎖新的創意可能性，並重新定義3D建模和開發中的可能邊界。

閱讀更多

02 資訊安全

NVIDIA 與 NIST 攜手以推動 AI 安全標準

NVIDIA NIST AI 安全 NeMo 欄杆 AISIC 負責任的 AI 生態系統 AI 風險管理 模型透明度
2024-02-08

NVIDIA 與 NIST 攜手推進 AI 安全標準

在確保人工智能（AI）的安全性和可信度方面，NVIDIA 與國家標準與技術研究院（NIST）合作，成立了人工智能安全研究所聯盟（AISIC）。這一合作標誌著在開發和部署安全 AI 技術方面的一個關鍵步驟。

該聯盟旨在建立一套工具、方法論和標準框架，指導創建可被信任以安全和負責任地運作的 AI 系統。NVIDIA 將其在 AI 領域的廣泛專業知識帶入聯盟，利用其與政府、學術界和產業界合作的經驗，促進 AI 技術的安全應用。

NVIDIA 的貢獻核心是開發 NeMo 欄杆，一個開源軟體計畫，確保 AI 生成的回應是準確的、相關的且安全的。這一點，加上 NVIDIA 近期對拜登政府自願性 AI 安全承諾的支持，以及對國家科學基金會 AI 研究計畫的 3000 萬美元投資，凸顯了該公司致力於培育一個負責任的 AI 生態系統的承諾。

AISIC 專注於促進知識交流和在可信賴 AI 方面的研究進展。擁有來自各個領域的 200 多個關鍵利益相關者，包括領先的 AI 創造者、研究人員和民間社會組織，聯盟有望在 AI 安全性和治理方面推動重大進步。

NVIDIA 的參與不僅將提升聯盟的技術能力，還將促進 AI 風險管理框架和模型透明度實踐的發展。通過這一個項目，NVIDIA 和 AISIC 正在為負責任的 AI 技術進化設置新的標準，確保它們最符合社會的最佳利益。

請持續關注 NVIDIA 和 AISIC 如何塑造值得信賴 AI 的未來更多更新。

閱讀更多

在數位時代保護隱私：Google Research推出DP-Auditorium

DP-Auditorium Google Research 差分隱私 隱私保護 開源庫 數據安全

2024-02-13

在數位隱私日益受到關注的時代，Google Research推出了一款新工具，DP-Auditorium，承諾將革新我們審核數據安全的方式。想像一個你的資訊被一層無法穿透的隱私保護所包圍的世界；DP-Auditorium是創造這個現實的一步。這個開源庫旨在測試差分隱私（DP）機制的有效性 - 這些過程將你的數據匿名化，使其無法追踪到你。

差分隱私就像是在合唱中混合個別聲音一樣，確保沒有單一聲音脫穎而出 - 這是保護跨行業和政府服務（如美國人口普查）使用的大型數據集中的個人信息的關鍵技術。然而，準確設置這些隱私機制既困難又容易出錯。這就是DP-Auditorium發光發熱的地方。它提供了一種「黑箱」方法，這意味著它可以評估一個隱私機制而不需要了解其內部運作，使用創新的測試算法來發現任何隱私泄露。

DP-Auditorium的核心在於其兩個主要組件：屬性測試器和數據集查找器。屬性測試器就像偵探，檢查樣本以發現任何隱私洩露，而數據集查找器則是偵察兵，確定這些洩露可能發生的條件。它們共同確保旨在保護你的隱私的機制正確地執行其職責。

Google Research的測試已經展示了DP-Auditorium發現隱私漏洞的能力，這些漏洞可能被其他方法忽略，證明了它在保護我們數位生活安全中的價值。通過讓DP-Auditorium向所有人開放，Google Research邀請全球的開發者和研究人員加強隱私保護技術的防禦。這一舉措不僅提高了差分隱私機制的可靠性，也標誌著邁向更安全、更私密數位時代的重要一步。

閱讀更多

OpenAI為ChatGPT增添記憶功能與用戶新控制選項

OpenAI ChatGPT 記憶功能 使用者控制 隱私 安全 更新

2024-02-13

在一次突破性的更新中，OpenAI為ChatGPT引入了一項新功能，使其能夠記住您對話中的細節。這一發展將徹底改變用戶與ChatGPT的互動方式，使重複互動變得更加連貫且相關，無需重復提供信息。不論是您對會議筆記格式的偏好、有關您咖啡店的細節，還是個人趣聞，ChatGPT現在都能保留這些信息，以便根據您的需求量身定制未來的對話。

OpenAI強調對這一新記憶功能的使用者控制。您可以命令ChatGPT記住或忘記特定細節，查看它記住了哪些內容，甚至可以通過設置完全關閉記憶功能。這確保了您的互動始終在您的控制之下，提供自訂化設置以增強您的ChatGPT體驗。

這一更新最初將向一群選定的ChatGPT免費用戶和Plus用戶推出，計劃在不久的將來進行更廣泛的發布。

此外，此次更新引入了臨時對話功能，用於無記憶的對話，以及自定義指令，以實現更個性化的互動。OpenAI也十分關注隱私和安全，實施措施以防止ChatGPT保留敏感信息，除非明確指示。

對於商業和企業用戶來說，這一記憶功能意味著ChatGPT的使用更加高效，使其能夠隨著時間的推移學習並適應您特定的工作流程和偏好。

最後，個別的GPT，如Books GPT或Artful Greeting Card GPT，也將配備自己的記憶功能，進一步個性化不同應用中的用戶體驗。

隨著這些更新，OpenAI繼續推動AI所能提供的邊界，使ChatGPT成為個人和專業使用中更加寶貴的工具。

閱讀更多

OpenAI 對抗國家附屬網路威脅中的人工智慧濫用

OpenAI Microsoft Threat Intelligence 網絡威脅 人工智慧濫用 數位安全 負責任使用 AI 安全增強措施 AI 生態系統

2024-02-14

OpenAI 對抗國家附屬的網路威脅行動中的人工智慧濫用

在一項開創性的舉措中，OpenAI 與 Microsoft Threat Intelligence 合作，採取重大步驟，以阻止國家附屬的網絡威脅行動者濫用人工智慧（AI）。這項行動導致了來自五個團體的活動受到干擾：Charcoal Typhoon 和 Salmon Typhoon（中國附屬）、Crimson Sandstorm（伊朗附屬）、Emerald Sleet（北韓附屬）、以及 Forest Blizzard（俄羅斯附屬）。

這些行動者試圖利用 OpenAI 的先進服務，進行各種惡意目的，包括網絡安全漏洞、資訊釣魚和旨在危害數位安全的腳本任務。這些團體使用 AI 的方式多種多樣，從翻譯技術文件和除錯代碼，到研究網絡安全工具中的漏洞，以及撰寫釣魚內容。

OpenAI 的主動態度涉及使用他們的 AI 模型以及 Microsoft 的洞察來檢測和干擾這些威脅。這種夥伴關係強調了負責任使用 AI 的承諾，確保其好處不被濫用。OpenAI 的方法包括持續監控、與產業夥伴交換資訊，以及迭代的安全增強措施，以領先於不斷進化的網絡威脅。

通過揭露這些活動並實施嚴格措施，OpenAI 旨在促進一個安全且透明的 AI 生態系統。這項努力不僅保護了數位基礎設施，也確保 AI 繼續作為一股正面力量，改善生活而不成為惡意行動者的工具。

這項倡議標誌著促進負責任 AI 使用的關鍵一步，凸顯了在保護數位環境免受複雜威脅的重要性，以及警覺和合作的重要性。

閱讀更多

03 人物故事

Ivan Tashev：微軟研究院的音訊信號處理大師

Ivan Tashev 音訊信號處理 微軟研究院 Kinect Microsoft Teams 神經網絡 語音增強
2024-02-01

在微軟研究院播客系列「What's Your Story」的一集中，Johannes Gehrke 揭開了Ivan Tashev 的旅程故事，一位對音訊信號處理充滿深厚熱情的合夥軟體架構師。這一部分深入探討了Tashev 對工程學早期的著迷以及在數學奧林匹克競賽中取得關鍵勝利如何引領他走上創新與探索的道路。

在Tashev 工作的核心是他對於微軟創新產品如 Kinect、Teams 和 HoloLens 音訊組件的貢獻。他從保加利亞好奇的孩子成長為音訊處理的開拓者，強調好奇心和追求卓越的力量。

Tashev 在 Kinect 音訊系統開發中的專業知識尤其閃亮，他解決了多通道聲學回音消除的看似不可能的挑戰——一項曾被認為不可能的壯舉。通過引入包含旋律信號的校準過程，Tashev 使 Kinect 成為第一個能夠在其揚聲器的噪音中從多達4 1/2 米的距離識別人類語音的裝置。

他的旅程並未止步。Tashev 在語音增強技術的演進中扮演了關鍵角色，從傳統的統計信號處理方法轉變為利用神經網絡。這一轉變標誌著音訊處理能力的顯著飛躍，並在2020年於 Microsoft Teams 中部署了一種先進的基於神經網絡的語音增強算法。這個算法不僅改善了音訊質量，也為即時通訊設定了新的標準。

Ivan Tashev 的故事證明了奉獻和創新的變革力量。他在微軟研究院的工作不僅推動了音訊信號處理領域的發展，也豐富了我們與技術的日常互動，使數位通訊更加自然和沉浸。

閱讀更多

揭開DevOps的神秘面紗：Nicole Forsgren從排球到軟體工程的旅程

DevOps 軟體工程 技術創業 Google Microsoft Research AI 開發者體驗

2024-02-15

在Microsoft Research的播客系列「What's Your Story」中一集鼓舞人心的節目裡，Johannes Gehrke深入探討了開發者體驗和DevOps領域的翹楚Nicole Forsgren的生活和職業生涯。Forsgren的職業軌跡非常了不起，從IBM的軟體工程師到學者，最終變成一位技術創業家，其啟動的公司吸引了Google的注意。

在愛達荷州的一個小鎮長大，那裡的職業抱負往往是有限的，Forsgren的故事始於一個意想不到的轉變，從心理學到計算機資訊系統，由於一場突如其來的家庭危機和一次關於技術領域有利可圖的機遇的偶然談話驅動。基於一位學生的建議，這一躍進決定了她進入技術和軟體工程的道路。

她在IBM擔任軟體工程師期間的任職以及她的重要貢獻，包括一項旨在增強對抗冷啟動攻擊的安全專利，展示了她解決複雜技術挑戰的能力。Forsgren的學術之旅進一步豐富了她的專業知識，導致了DevOps的關鍵研究——一個今天與高效合作的軟體開發實踐同義的術語。

Forsgren的創業之舉，DevOps Research and Assessment (DORA)，體現了她對軟體工程創新方法的典範，最終以被Google收購告終。這一舉動不僅凸顯了她作為研究人員和企業家的能力，也突出了她在塑造現代軟體開發實踐中的影響力。

目前，在Microsoft Research，Forsgren繼續突破開發者體驗和生產力的界限，探索AI如何革新軟體工程。她的旅程證明了擁抱變化、挑戰常規以及在研究、技術和創業交匯處蘊藏的無窮可能性的力量。

Nicole Forsgren的故事是任何進入技術領域的人的燈塔，證明了擁有好奇心、韌性和開放轉變的態度，可以深刻影響技術格局。

閱讀更多

04 應用

Accenture透過AWS生成式AI革命化藥物文件撰寫

生成式AI 藥品文件撰寫 AWS Accenture 監管文件 加密技術

2024-02-06

Accenture利用AWS生成式AI革新藥品文件撰寫

在一項突破性的發展中，Accenture利用了AWS的生成式AI服務，創建了一個最先進的解決方案，用於撰寫監管文件，簡化了將新藥推向市場的繁瑣過程。這個巧妙的系統主要旨在解決生成共同技術文件（CTD）的復雜性，這些文件對於像美國食品和藥物管理局（FDA）這樣的監管機構批准藥物至關重要。

創建CTD可能非常耗時，對於一家大型藥品公司來說，每年可能需要高達100,000小時的工作。這些文件包含超過100份詳細報告，對藥物研究和測試階段至關重要。Accenture的AI驅動解決方案承諾通過自動生成CTD大幅減少這項勞動密集型工作。利用Amazon SageMaker JumpStart和其他AWS AI服務，該系統有效地從測試報告中提取關鍵數據，以正確格式生產CTD。

這個解決方案的特點不僅僅是繁瑣任務的自動化，還包括其對高標準的控制、安全性和可審計性的堅持，融入了AWS Well-Architected原則。這確保了敏感數據得到了最嚴謹的處理，採用加密技術提升了安全性。

通過採用AWS生成式AI，Accenture不僅希望在製藥領域內提升效率，也希望加快創新治療的批准。這在行業內標誌著一個重大的飛躍，有可能將突破性治療更快地交到患者手中。

在初步測試中，Accenture的解決方案已展示了在撰寫CTD所需時間上減少了驚人的60-65%。這反映了藥物提交過程的顯著優化，承諾更快的響應時間和系統性能的持續改進。Accenture和AWS之間的這次合作示範了生成式AI如何改變受監管行業的風景，為效率和創新設定了新的標杆。

閱讀更多

利用 QnABot 與 ServiceNow 整合優化 IT 支援

QnABot ServiceNow IT 支援 AWS Amazon Lex AWS Lambda 生成式 AI IT 服務管理 ITSM 自動化客戶支援

2024-02-06

利用 QnABot 與 ServiceNow 整合優化 IT 支援

在當今快節奏的數位世界中，透過電話報告 IT 問題已成為過去式。AWS 推出了一個尖端解決方案 QnABot，徹底改變組織處理 IT 支援的方式。這個開源工具，建基於 AWS 的強大服務如 Amazon Lex 和 AWS Lambda，從 5.4 版本開始，現在擁有生成式 AI 能力。它旨在快速解決常見的 IT 問題，提升員工體驗，並讓 IT 代理能夠集中精力處理更複雜的問題。

透過將 QnABot 與 ServiceNow 整合，後者作為 IT 服務管理（ITSM）的領導者，被 Gartner Magic Quadrant 2023 認定，組織現在可以簡化其事件管理。這意味著用戶可以簡單地與 QnABot 聊天來分類 IT 服務問題。如果聊天機器人確定需要人工介入，它可以立即通過直接從用戶那裡收集細節來在 ServiceNow 中開啟一個事件票證。

這個解決方案不僅加快了 IT 服務事件的解決過程，也豐富了整個支援流程。QnABot 擁有多通道、多語言能力，可以無縫地整合到 IT 服務台工作流程中，確保員工的 IT 問題能夠得到迅速且高效的解決。

對於那些希望通過創新 AI 技術提升其 IT 支援系統的人來說，將 AWS 上的 QnABot 與 ServiceNow 整合提供了一個全面的解決方案。它代表了在自動化客戶支援方面向前邁出的重要一步，釋放了寶貴的 IT 資源，並改善了用戶滿意度。

如何作用：

- QnABot 與 ServiceNow 的整合允許實時創建事件票證。
- 使用者可以透過聊天與 QnABot 互動，描述他們的問題，並在 ServiceNow 中開啟一個票證，無需手動干預。
- 這種整合利用了 AWS 的強大雲服務和 ServiceNow 的先進 ITSM 能力。

這種在 IT 支援自動化方面的突破，正準備重新定義支援工作流程，為無縫、AI 驅動的客戶服務解決方案的未來展開了一瞥。

閱讀更多

Amazon SageMaker 利用 AI 革命性地推動藥物安全

Amazon SageMaker **藥物安全** **機器學習** **自然語言處理** **BioBERT** **數據保護** **HIPAA** **Falcon-7B** **Falcon-40B** **Adverse Drug Reaction Dataset**

2024-02-06

Amazon SageMaker 以 AI 革命性地推進藥物安全

為了加強藥物安全，Amazon 利用其先進的機器學習平台 Amazon SageMaker，自動化偵測藥物不良反應——這在製藥業是一大步。鑑於每天產生大量的健康數據，傳統的手動監控藥物安全事件既耗時又昂貴。Amazon 的機器學習方案提出了一種高效自動化的方法，利用在 Hugging Face 上可獲得的 Adverse Drug Reaction Dataset，一個著名的自然語言處理（NLP）任務平台。

這一開創性解決方案的核心是 BioBERT 模型，該模型已在 Pubmed 數據集的大量醫學文獻上精心預訓練，使其獨特地適用於理解醫學術語的複雜性。當在 SageMaker 上進行微調時，該模型在從社交媒體、電子郵件乃至手寫筆記等多樣化數據源中分類藥物不良事件方面展示出了卓越的能力。

Amazon 的方法不僅滿足了對更快、更可靠的藥物安全監測的迫切需求，而且通過納入基本的加密措施，確保符合最嚴格的數據保護標準，如 HIPAA。

通過整合最先進的 AI 模型，如 Falcon-7B 和 Falcon-40B，Amazon 進一步豐富了數據集，用合成生成的事件樣本增強了模型檢測不良事件的能力，準確性顯著提高。結果說明了一切：經過合成數據豐富的微調 BioBERT 模型，在識別藥物不良反應方面提供了前所未有的性能，展示了 AI 在大幅改善製藥領域中病人安全和結果方面的潛力。

這在自動化藥品監測方面的飛躍表明 Amazon 致力於利用最新的機器學習技術來解決複雜的醫療挑戰。隨著這項技術的不斷發展，藥物安全監測的未來看起來充滿希望，AI 領航保障公共健康。

閱讀更多

透過 Amazon Bedrock 的 AI 代理革新保險理賠

Amazon Bedrock **AI 代理** 保險理賠 **Generative AI** 數據完整性 隱私 **負責任的 AI 使用**
2024-02-08

透過 Amazon Bedrock 的 AI 代理革新保險理賠

在保險領域內，Amazon Web Services (AWS) 透過 Amazon Bedrock 平台，引入了 Generative AI 代理和知識庫的創新應用，這項尖端技術旨在顯著提升運營效率、增強客戶服務並簡化管理保險理賠的決策過程。

這項技術奇蹟的核心是利用專門的 AI 代理，能夠自動化保險理賠的整個生命周期。從啟動新的理賠到發送待辦文件的提醒以及收集必要的證據，這些代理配備了處理傳統上需要大量人工介入的各種任務的能力。代理透過理解自然語言輸入工作，使它們在處理請求時極為友好且高效。

Amazon Bedrock 利用基礎模型 (FMs)，提供了一座通往 Amazon 及其他領先 AI 公司提供的強大 AI 工具的橋樑。這些 AI 代理，當與知識庫相結合時，提供了一個能夠根據輸入的特定數據和指令執行任務的協調系統。這個系統不僅能夠自動化常規任務，還能夠參與複雜的、多步驟的工作流程，將任務分解成更小、更易管理的步驟，並確保其高效執行。

這項技術為保險公司提供了改善理賠處理速度和準確性的重大機會，從而提升客戶滿意度。其潛在應用不僅限於簡單的任務自動化，還能使公司透過更好的決策支持系統擴展其業務和改善知識管理。

此外，這個解決方案在設計時考慮到了安全性、隱私以及負責任的 AI 使用，確保企業在堅持數據完整性和道德 AI 使用的最高標準時能夠利用這項技術。

總之，通過 Amazon Bedrock 為保險理賠管理引入 Generative AI 代理和知識庫，標誌著 AI 技術在保險行業應用中的一大飛躍。這項創新不僅承諾將革新保險理賠的處理方式，還為該行業的運營效率和客戶服務設定了新的標準。

閱讀更多

革新機器學習實驗：Booking.com與Amazon SageMaker的突破性合作

Amazon SageMaker 機器學習 模型調整 模型訓練 超參數調整 模型解釋性 Booking.com

2024-02-12

革新機器學習實驗：Booking.com與Amazon SageMaker的突破性合作

在一次開創性的合作中，Booking.com利用了Amazon SageMaker的力量，大幅改善了其機器學習(ML)實驗框架。這一舉措不僅優化了他們的搜索和推薦算法，也顯著提升了全球數百萬線上旅遊體驗。

邁向現代化的一大步

面對ML模型訓練和實驗資源長時間等待的挑戰，Booking.com的排名團隊決定現代化其ML基礎設施。通過整合Amazon SageMaker，他們取得了顯著的成果：

- 大幅減少等待時間： 透過在SageMaker上使用即時實例，團隊將等待時間縮短了十倍，使模型迭代和洞察生成更快。
- 先進ML功能觸手可及： 通過SageMaker Automatic Model Tuning和SageMaker Clarify的添加，使團隊能以前所未有的精準度和洞察力精煉他們的模型。
- 開發效率： 現代化努力將反饋迴圈從幾分鐘縮短到即時，大幅加速了ML模型的開發週期。

賦予ML科學家可定制的管道

這一轉型的關鍵方面是創建一個友好的環境，允許ML科學家輕鬆測試假設，而無需深入複雜的編碼。通過在SageMaker中設置簡化的管道，包括一個全面的配置檔案（`config.ini`），科學家可以輕鬆自定義他們的ML模型建構過程。

模型建構和調整的創新

與Amazon SageMaker的旅程引入了幾個創新步驟，包括：

- 自動模型調整(AMT)： 通過AMT，排名團隊現在可以通過運行數百個並行訓練作業，有效地確定最佳超參數，顯著加速模型開發過程。
- 使用SageMaker Clarify進行模型解釋性： 這一功能幫助團隊理解不同輸入特徵如何影響他們的ML模型，促進了改善模型性能的知情決策。

令人印象深刻的業務成果

遷移到Amazon SageMaker帶來了相當多的好處：

- 增加模型訓練頻率： 每月模型訓練作業增加了五倍，與管線運行失敗率的顯著下降相結合。
- 優化訓練時間： 由於在SageMaker上有效的GPU訓練，模型訓練時間減少了80%。
- 增強的ML能力： 引入如超參數調整和模型解釋性等先進功能，顯著推進了Booking.com的ML實驗能力。

結論

Booking.com以Amazon SageMaker現代化其ML實驗框架的舉措，證明了基於雲的解決方案在推動創新和效率方面的強大力量。這次合作不僅解決了現有的挑戰，也為線上旅遊服務領域的運營卓越設定了新標準。

閱讀更多

革新零售業：BigBasket搭載AI的結帳大躍進

BigBasket AI Amazon SageMaker 零售業 結帳流程 計算機視覺 卷積神經網絡
2024-02-13

革新零售業：BigBasket搭載AI的結帳大躍進

在零售業內一項令人矚目的創新中，印度領先的線上食品和雜貨商店BigBasket，已經利用Amazon SageMaker的力量，顯著改善其實體店面的結帳流程。這一前進的飛躍，不僅承諾為超過500個城市的1000萬顧客帶來更佳的購物體驗，還為零售效率和客戶服務設定了新的標準。

AI帶來的尖端結帳體驗

手動結帳和易丟失的重量標籤的日子一去不復返了。BigBasket的AI啟用系統利用先進的計算機視覺技術，迅速識別快速消費品(FMCG)，將結帳體驗轉變為一個無縫、高效的過程。通過使用經Amazon SageMaker細緻訓練的複雜計算機視覺模型，BigBasket已將模型訓練時間減少了顯著的50%，並實現了20%的成本節省。

技術勝利

將Amazon SageMaker整合進BigBasket的運營體系，帶來了實質性的增強。利用帶有ResNet152的卷積神經網絡(CNN)進行圖像分類，該系統現在能夠熟練處理超過12,000件商品的龐大目錄，每月以600件的速度迅速增加新商品。這個AI驅動的結帳不僅加快了流程，還顯著減少了人為錯誤，確保了準確愉快的購物體驗。

光明的未來在前方

這個技術奇蹟，已運行全面生產超過六個月，證明了BigBasket對創新和客戶滿意度的承諾。隨著AI自助結帳系統的推出，BigBasket正在為零售業的新時代鋪路，承諾帶來更順暢、更高效的購物旅程。與AWS的成功合作以及使用Amazon SageMaker，標誌著利用技術更好地服務客戶和重新定義零售格局的里程碑。

隨著我們向前邁進，BigBasket與AWS和SageMaker的旅程，成為了一個閃耀的範例，展示了如何使用技術解決現實世界的挑戰，使我們的日常生活變得更加輕鬆和高效。

閱讀更多

利用Amazon生成式AI聊天機器人革命化旅遊規劃

生成式AI 聊天機器人 旅遊規劃 Amazon Amazon Redshift Amazon Bedrock 客戶體驗 個人化 數據安全

2024-02-14

利用Amazon的生成式AI聊天機器人革新旅遊規劃

在一項突破性的舉措中，Amazon利用生成式AI聊天機器人引入了一種新的製定個人化旅行計劃的方式，透過Amazon Redshift和Amazon Bedrock的力量。這項創新有望改變我們對旅遊規劃的方式，提供前所未有地符合個人偏好和興趣的定制行程。

生成式AI在各個行業中引起了轟動，Amazon正處於前沿，利用其能力顯著提升客戶體驗。這項技術使得開發能夠理解和解釋人類語言、生成恰當回應，甚至自動化複雜文件處理過程的智能應用成為可能。這一AI應用的飛躍承諾在研究加速、洞察發現和生產力提升方面達到前所未有的規模。

Amazon Bedrock是一項全面管理的服務，簡化了構建可擴展生成式AI應用的過程。它提供了來自領先AI實體的印象深刻的高性能基礎模型選擇，可通過單一API訪問。這項服務不僅促進了這些模型與您的數據的定制，而且還確保了符合安全和隱私標準，為希望保持領先的企業提供了寶貴的工具。

魔法在於提示工程——一種設計用戶輸入以指導AI生成所需輸出的技術。這種方法通過微調提示以基於用戶數據產生相關和準確的資訊，從而創建更高效和有效的生成式AI應用，用戶數據存儲在諸如Amazon Redshift之類的數據庫中。

Amazon最新的創新展示了這項技術在旅遊和款待業中的實際應用。通過打造一個個人化的旅行行程規劃器，Amazon展示了顯著豐富客戶體驗的潛力。該系統利用用戶數據，包括愛好、興越和旅行預訂細節，創建一份根據每位用戶偏好量身定制的獨特旅行指南。這不僅增強了用戶體驗，而且還展示了個人化旅遊規劃未來的一瞥。

本質上，Amazon進軍生成式AI聊天機器人旅遊規劃不僅僅是技術進步——它是向著一個技術理解並滿足我們個別偏好的未來邁進的一步，使每一次旅程都獨特屬於我們自己。

閱讀更多

為未來加速：NVIDIA 在汽車產業中推動數位轉型

NVIDIA 汽車產業 數位轉型 自動駕駛 數位雙胞胎 模擬 NVIDIA Omniverse

2024-02-14

為未來加速：NVIDIA 在汽車產業中推動數位轉型

在軟體與人工智慧引導創新的時代中，NVIDIA 正在汽車產業中燃起一場重大的轉變。通過將物理與數位結合，NVIDIA 及其合作夥伴正在引領車輛開發和製造的新階段。這場數位進化不僅僅是升級引擎蓋下的技術；它正在徹底改革車輛的構思、設計與測試方式。

數位藍圖：加速設計與開發

想像將新車上市的時間縮短數年的情景。這是當公司擁抱數位化時所面臨的現實。通過創建數位雙胞胎 —— 物理車輛的虛擬複製品 —— 工程師能夠模擬組成汽車的30,000個零件，簡化開發流程，並大幅減少實體原型的需求。這不僅節省時間，也大幅降低成本，使得更多創新與實驗成為可能。

模擬：推動更安全自動駕駛車輛的動力

自動駕駛之路鋪滿了數據和模擬。借助NVIDIA的技術，公司能夠創建複雜的虛擬環境來測試和完善將來引導自駕車的人工智慧系統。這種方法為確保這些系統能夠處理現實世界駕駛的不可預測性提供了一種安全、成本效益高的方式，從突如其來的天氣變化到意料之外的道路障礙。

NVIDIA Omniverse：汽車設計的新維度

在這場數位轉型的核心是NVIDIA Omniverse平台，這是一個合作空間，全球團隊可以聚集在一起，實時設計和測試新的車輛概念。這種數位與物理領域的融合不僅僅改變了製造車輛的方式；它正在重新塑造汽車產業的整體價值主張。適應這種數位-物理混合的公司不僅能夠生存下來，而且將蓬勃發展，以前所未有的速度和規模推出創新。

汽車產業正處於一個十字路口，而NVIDIA正在引導它走向一個每個車輛設計、開發和部署的方面都由數位化驅動的未來。這一旅程剛剛開始，前方的可能性無限。

密切關注NVIDIA如何繼續重新定義交通運輸。加入這場旅程，發現NVIDIA GTC中汽車產業的未來，數位與實體世界的融合將成為舞台的中心。

閱讀更多

超級計算與安全：俄亥俄超級計算中心在 NASCAR 未來的角色

超級計算 安全 OSC NASCAR Open OnDemand 模擬 教育 工業創新 醫療突破
2024-02-14

超級計算與安全：俄亥俄超級計算中心在 NASCAR 未來的角色

在 NASCAR 競速世界這個快節奏的領域中，速度與安全至關重要，科技進步永遠在地平線上。而俄亥俄超級計算中心 (OSC) 作為創新的燈塔，正利用超級計算的力量，重新定義賽車未來及更廣泛的領域。

在最近一集 NVIDIA AI Podcast 中，主持人 Noah Kravitz 與 OSC 的戰略計劃總監 Alan Chalker 聊到了超級計算的轉型影響。焦點放在 OSC 的開創性 Open OnDemand 計劃上，這是一個提供簡易訪問先進計算資源的網絡平台。這項倡議不僅僅適用於學術界，還擴展到包括 NASCAR 這樣高能量領域的廣泛行業中。

對於 NASCAR，OSC 正在將虛擬變成現實，模擬賽車設計以確保它們滿足安全和效率的不斷進化的需求。這種合作只是超級計算革命化各個行業潛力的一瞥，提供了無與倫比的精準度和遠見。

但 OSC 及其 Open OnDemand 計劃的未來又會如何呢？前景看起來很有希望，持續的努力是為了讓超級計算對更廣泛的應用去謎化和民主化。無論是出於教育目的、工業創新，還是醫療突破，OSC 都在前沿，推動計算的未來朝著更安全、更智能、更可持續的方向前進。

敬請關注，看看俄亥俄超級計算中心如何一次模擬一次加速進步。

閱讀更多

電信產業擁抱生成式 AI 以獲得競爭優勢，NVIDIA 調查發現

生成式人工智慧 電信產業 NVIDIA 5G 客戶體驗 數據保護

2024-02-15

電信產業擁抱生成式人工智慧以獲得競爭優勢，NVIDIA 調查發現

最新的 NVIDIA 調查顯示，在電信部門中有一個引人注目的趨勢 —— 生成式人工智慧正在迅速成為創新和競爭優勢的基石。超過 400 名來自全球的行業專業人士參與了這項調查，這強調了生成式人工智慧從默默無聞到成為重新定義電信格局中心角色的流星般崛起。

在部署 5G 網絡五年後，電信公司現在正在利用生成式人工智慧和大型語言模型來革新他們的運營。這個轉變不僅僅是為了保持領先；這是關於利用人工智慧來提升客戶體驗，簡化運營，並最終提高收入並降低成本。

有驚人的 43% 受訪者已經在投資生成式人工智慧，且有相當數量的人計畫在來年擴大他們的人工智慧使用案例。這種熱情在執行主管的展望中有所反映，超過半數的人認為人工智慧的採納對於獲得競爭優勢至關重要。此外，有 56% 的業界利益相關者相信人工智慧在他們公司未來成功中扮演著關鍵角色，這比前一年的情緒有顯著增長。

專注於客戶體驗是顯而易見的，接近一半的受訪者表明這是他們利用人工智慧的主要目標。從改善客戶服務到提升網絡運營和行銷，生成式人工智慧走在轉變電信產業解決各種商業挑戰方法前沿。

人工智慧的投資正在增長，有一個顯著的轉向更大預算專門用於人工智慧基礎設施。這項投資不僅僅關於內部能力，還涉及與合作夥伴合作共同開發人工智慧解決方案，尊重行業的嚴格數據保護標準。

NVIDIA 的調查凸顯了電信產業邁向人工智慧整合旅程中的一個關鍵時刻。隨著公司繼續探索人工智慧的潛力，對客戶體驗、收入增長和成本削減的重視預示著一個充滿希望的未來，生成式人工智慧在塑造該行業演進中扮演著中心角色。

閱讀更多

顛覆客服：引進Amazon Bedrock的情境聊天機器人

Amazon Bedrock 聊天機器人 情境聊天 檢索增強生成 大型語言模型 客戶支持

2024-02-19

顛覆客服：引進Amazon Bedrock的情境聊天機器人

在今日繁忙的數位化市場中，客服已找到新夥伴：聊天機器人。Amazon透過其Amazon Bedrock的知識庫，將這項創新提升了一個層次，將聊天機器人從簡單的問答機器人轉變為複雜的數位代理人。這些升級的聊天機器人能夠提供個性化、情境化的回應給客戶詢問，這要感謝它們與內部知識庫的深入整合。

這種在聊天機器人智能上的飛躍背後的祕密是檢索增強生成（Retrieval Augmented Generation, RAG）架構，一個將大型語言模型（Large Language Models, LLMs）的預測能力與知識庫的事實基礎結合起來的系統。當用戶與聊天機器人互動時，他們的詢問不只是被孤立地回答。相反地，它是透過從廣泛的數據庫中拉取相關資訊而增強的，確保回應既相關又基於現實世界的知識。

對於企業來說，這意味著聊天機器人現在可以根據過去的購買提供推薦、通過訪問特定的客戶紀錄來提供細緻的支持，甚至可以調整它們的語言以匹配用戶的技術專業程度。這種個性化的水平不僅令人印象深刻；它對於客戶參與來說是一場遊戲改變者。

設定這樣一個先進的系統可能聽起來令人望而卻步，但Amazon Bedrock簡化了這個過程。它是一個無服務器解決方案，處理從管理向量資料庫中的文本嵌入到提供檢索和生成文本的API的所有繁重工作。企業現在可以專注於他們最擅長的事情，將聊天機器人開發的複雜性留給Amazon Bedrock。

這種對會話式AI的突破性方法不僅使得24/7客戶支持成為現實，還顯著提升了這些互動的質量。有了Amazon Bedrock的知識庫，聊天機器人正在進化為虛擬顧問，以前所未有的精緻度提供洞察和協助。

閱讀更多

NVIDIA 透過最新 Studio 驅動程式與應用程式 Beta 版本革新創意工作流程

NVIDIA Studio 驅動程式 創意工作流程 beta 版本 AI 驅動的 Studio 應用程式 「增強語音」工具 RTX 加速 「與 RTX 對話」 大型語言模型

2024-02-22

NVIDIA 透過最新的 Studio 驅動程式和應用程式 Beta 版本，為創意工作流程帶來革命性進步

在一次對創意人士和玩家都具有改變遊戲規則的更新中，NVIDIA 推出了其 2 月份的 Studio 驅動程式，以及全新 NVIDIA 應用程式的 beta 版本。這一飛躍向前的更新，承諾將在各種應用程式中簡化和增強創意過程，使其成為任何創作者工具箱中的必備品。

Studio 驅動程式，經過精心設計以優化創意軟體的性能，已經與應用程式開發者進行了廣泛的測試以確保無懈可擊的兼容性和性能提升。這個最新的驅動程式專注於增強功能、自動化過程和加速工作流程，為創意軟體的優化設定了新的標準。

與此同時，NVIDIA 應用程式的 beta 版本旨在為 NVIDIA GPU 擁有者統一和現代化的使用者體驗，提供一站式的最新驅動程式、AI 驅動的 Studio 應用程式等。這個新應用程式簡化了更新過程，幫助發現並安裝必要的 NVIDIA 軟體，並為 GPU 控制中心進行了改進，為整合和友好的用戶界面。

在眾多亮點功能中，Adobe Premiere Pro 的 AI 驅動的「增強語音」工具現已普遍可用，並且從 NVIDIA RTX 加速中獲益匪淺。這個革命性的工具大幅減少了不需要的噪聲，使得對話片段聽起來仿佛是專業錄製的 - 在特定 NVIDIA GPU 上快達 75%。

此外，NVIDIA 介紹了「與 RTX 對話」，一款技術展示應用程式，允許與大型語言模型進行個性化互動，為 GeForce RTX 擁有者承諾快速且安全的結果。

強調這些創新的現實世界影響，電影製作人 James Matthews 分享了他使用 NVIDIA Studio 驅動的工作流程創作短片「Dive」的經驗。Matthews 從概念化到最終產品的旅程凸顯了 NVIDIA 技術的轉型效果，使創作者能夠完全沉浸在他們的工藝中，並實現無縫的創意流程。

這種尖端技術與實際應用的結合，體現了 NVIDIA 致力於賦予創意人士力量的承諾，提供的工具不僅增強了他們當前的項目，也激發了未來創新的靈感。

敬請期待，見證 NVIDIA 如何一次更新一次重新定義創意與技術的界限。

閱讀更多

05 服務

Amazon SageMaker Canvas 引進新 AI 模型與即時互動功能

Amazon SageMaker Canvas | AI 模型 | 即時互動 | Llama 2 | Mistral | 生成式 AI | 機器學習 | 無編碼

2024-02-05

Amazon SageMaker Canvas 引進新的 AI 模型與即時互動功能

在一項令人振奮的發展中，Amazon SageMaker Canvas 擴大了其功能，整合了尖端的生成式 AI 模型，包括 Meta 的 Llama 2 與 Mistral.AI 的 Mistral，並推出了一項用於串流回應的新功能。這項升級旨在賦予用戶能力，使他們能夠毫不費力地構建和部署機器學習模型，無需任何編碼知識。

有什麼新變化？

- Llama 2 模型整合：在生成式 AI 領域中領先，Llama 2 因其執行各種任務的能力而聞名，從參與連貫的對話到生成新內容和提取資訊。憑藉由 Amazon Bedrock 驅動的 130 億和 700 億參數版本，以及透過 Amazon SageMaker JumpStart 的 70 億參數版本，Llama 2 提供了卓越的擴展性和多功能性。
- Mistral 模型新增：由法國 AI 新創公司 Mistral.AI 開發，Mistral 7B 模型擁有 73 億參數，採用群組查詢注意力（GQA）技術以更快的推論速度。這項創新使 Mistral 模型能夠高效地提供高性能，使其成為生成式 AI 應用的一個引人注目的選項。
- 串流回應：為了提升用戶體驗，SageMaker Canvas 現在提供了串流回應功能，允許即時互動和更自然、更流暢的聊天機器人應用對話流程。這項功能有望顯著改善性能和用戶滿意度。

如何開始

在 SageMaker Canvas 平台上可用，用戶可以輕鬆探索這些新模型和功能，只需導航至現成模型頁面並選擇所需選項。無論是用於構建聊天機器人、推薦系統，還是虛擬助理，Llama 2 和 Mistral 模型的整合，以及串流回應的引入，都將徹底改變我們與生成式 AI 互動的方式。

結論

此次對 Amazon SageMaker Canvas 的更新標誌著使生成式 AI 變得更加無障礙和互動的重要一步。通過整合像 Llama 2 和 Mistral 這樣的先進模型並引入串流回應，Amazon 為無需編碼專業知識就能開發創新 AI 應用鋪平了道路。

閱讀更多

顛覆姿態估計：Amazon SageMaker Ground Truth自定義標記工作流程

姿態估計 **Amazon SageMaker Ground Truth** 標記工作流程 計算機視覺 機器學習
2024-02-14

顛覆姿態估計：Amazon SageMaker Ground Truth自定義標記工作流程

在計算機視覺領域內，理解圖像或影片中物體的位置和方向，被稱為姿態估計，對於從機器人學到擴增實境的應用至關重要。然而，訓練準確的模型需要依靠擁有大量標有物體上每一興趣點精確座標的圖像數據庫——一個既勞動密集又容易出錯的過程。

這時，Amazon SageMaker Ground Truth引入了一種突破性的自定義標記工作流程，專為有效且準確地標註物件上的關鍵點——物體的骨骼點而設計。這個創新的工作流程不僅簡化了標記過程，還最小化了錯誤，確保訓練堅固的姿態估計模型的高品質數據。

這個自定義工作流程利用了一個使用者友好的網頁門戶，允許標記者直接在瀏覽器上的圖像上標註關鍵點。這個介面會根據預定義的骨骼結構自動連接這些點，提供即時的視覺反饋來捕捉並更正錯誤，如錯誤的點分配。結果？獲取進步姿態估計技術所需的高品質標記數據集的成本顯著降低。

此外，SageMaker Ground Truth促進了標記工作的管理和標記數據融入訓練過程。通過利用AWS Lambda函數進行標注前後的處理和將數據存儲在Amazon S3中，這個解決方案簡化了大規模、準確的姿態估計模型的開發。

這一突破代表了使複雜的姿態估計模型更加容易獲得和高效的一大步，為體育分析、安全性等領域的創新鋪平了道路。

閱讀更多

透過 Adobe 的 AI 創新和 NVIDIA 的 RTX 技術解鎖創意潛力

Adobe **NVIDIA RTX** **AI 創新** **數位創作者** **Enhance Speech** **GeForce RTX** **Photoshop** **Lightroom** **Omniverse OpenUSD**

2024-02-15

解鎖創意潛力：Adobe 的 AI 創新與 NVIDIA RTX 技術

在數位創作者的激動人心的進步中，Adobe 正在透過其最新的 AI 驅動工具套件，全部由 NVIDIA RTX 技術強力驅動，革新故事講述的藝術。在這場轉型的核心是 Adobe Firefly，Adobe 新增至其創意應用程式的成員，該程式托管於 NVIDIA 強大的雲端基礎設施上。

這些進步的關鍵是即將推出的 Adobe Premiere Pro 中的 Enhance Speech 工具。這個由 NVIDIA RTX 推動的突破性功能，承諾消除背景噪音並提升對話錄音的清晰度，賦予它們光鮮、專業的聲音品質。

在這波創新中受到矚目的是 Esteban Toro，Adobe 的資深社群關係經理，他巧妙地利用這些 AI 驅動的能力在 Adobe Photoshop 和 Lightroom 中。他的項目，電影式肖像系列，展示了充滿情感的照片和影像肖像，每個都有其獨特的故事 - 如 81 歲的韓國畫家 Kim Nam Soon，她在 65 歲時開始了她的藝術之旅。

Toro 的工作證明了 AI 在將原始素材轉化為引人入勝的敘事方面的力量。透過利用 Enhance Speech 工具，Toro 克服了在嘈雜環境中錄音的挑戰，確保了清晰的音質。此外，基於文本編輯工具的語音轉文字 AI 技術便於創建 18 種不同語言的字幕，簡化了編輯過程。填充詞檢測功能也發揮了關鍵作用，通過消除不必要的停頓和填充詞，確保了更乾淨、更準確的轉錄。

這些由 GeForce RTX 和 NVIDIA RTX 技術加速的 AI 驅動工具，不僅為 Toro 節省了大量時間，還使他能夠全面實現他的創意願景，不受技術限制。效率提升顯著 - 最終檔案導出速度是僅使用 CPU 處理的四倍快，多虧了 GPU 加速的 NVIDIA 視頻編碼器。

Adobe 的節省時間功能套件，包括 AI 驅動的自動重構工具和 Photoshop Lightroom 中的 RTX 加速的 Raw Details 和 Super Resolution 功能，突顯了創意與技術之間的協同作用。Adobe 與 NVIDIA 的合作為全球的創作者鋪平了道路，使他們能夠突破數位藝術的界限。

對於那些渴望深入了解 AI 驅動內容創作世界的人來說，NVIDIA 也在聚焦其 Omniverse OpenUSD 月。這一倡議慶祝 3D 世界創建、組合和合作的開放、可擴展生態系統，為創意探索提供了無限可能。

當我們站在數位創意的新時代門檻上時，像 Adobe 的 AI 驅動工具和 NVIDIA 的 RTX 技術這樣的技術不僅僅是工具，而是通往想象和故事講述未知領域的門戶。

閱讀更多

用 Code Llama 70B 在 Amazon SageMaker JumpStart 上革新您的開發工作流程

Code Llama **Amazon SageMaker JumpStart** 編碼效率 大型語言模型 軟件開發 模型部署 數據隱私

2024-02-16

用 Code Llama 70B 在 Amazon SageMaker JumpStart 上革新您的開發工作流程

您是一位尋求提升編碼效率的開發者嗎？別再尋找了！由 Meta 開發的開創性 Code Llama 70B 模型現已可透過 Amazon SageMaker JumpStart 使用，為編碼生產力帶來了非凡的飛躍。Code Llama 是一個大型語言模型（LLM），設計用於理解和生成來自編程及自然語言提示的代碼，是任何開發者工具箱中的亮點。

Code Llama 在各種程式語言中表現出色，包括 Python、C++、Java、PHP、C#、TypeScript 和 Bash。想像一下，減少編碼任務上花費的時間，簡化軟件開發工作流程，並輕鬆生產具有良好文件、高品質的代碼。有了 Code Llama，這一切都不僅僅是可能。

Code Llama 的突出之處在於其滿足廣泛程式設計需求的能力。它有三種變體：用於通用目的的基礎模型、專門用於 Python 專案的 Python 專門版本，以及針對理解自然語言指令定制的指令跟隨模型。無論您的專案複雜性或性質如何，Code Llama 都有適合您的變體。

使用 Amazon SageMaker Studio 的 JumpStart 部署 Code Llama 70B 模型非常簡單。只需幾次點擊，就能將這個強大的工具用於推理。此外，SageMaker 的基礎設施確保了您的數據隱私和安全，讓您在創新時安心。

像 Code Llama 這樣的基礎模型正在改變機器學習的格局，為文本摘要、數字藝術生成、語言翻譯，現在還有代碼生成，打開了新的大門。通過 SageMaker JumpStart 利用這些預先訓練的模型，您不僅僅是在採用尖端技術；您是在為當今快節奏的開發世界中的成功做好準備。

不要錯過用 Code Llama 70B 提升您的編碼專案的機會。不管您是在生成新代碼、完成現有代碼，還是將自然語言指令轉化為代碼，Code Llama 都是讓開發工作流程更高效、更愉快的解決方案。今天就在 Amazon SageMaker JumpStart 上探索 Code Llama 的潛力，革新您的開發流程！

閱讀更多

Google 的 Gemma 模型與 NVIDIA 的 AI 平台攜手合作

Google NVIDIA Gemma AI 平台 語言模型 GPU TensorRT-LLM 雲服務 Chat with RTX 數據私密性

2024-02-21

Google 的 Gemma 模型與 NVIDIA 的 AI 平台聯手

Google 與 NVIDIA 合作，為最新的語言處理技術 Gemma 注入強大動力。這次合作將 Gemma ——一個開創性的開放語言模型，推向 AI 發展的最前線，透過利用 NVIDIA 的 GPU 力量。Gemma 有兩個版本，分別擁有 20 億和 70 億參數，旨在設計成靈活且輕量的，使其能夠被部署在各種領域，以促進創新同時保持低成本。

對 Gemma 在所有 NVIDIA AI 平台上的優化，包括數據中心、雲環境和本地 RTX AI 電腦，標誌著 AI 可及性的一大飛躍。通過利用 NVIDIA 的 TensorRT-LLM —— 一個旨在優化大型語言模型推理的庫，Gemma 在 NVIDIA 的 GPU 上的性能得到顯著提升。這次合作確保開發者可以利用全球超過 1 億 NVIDIA RTX GPU，為高性能 AI 應用開闢了一片可能性的領域。

此外，Google 計畫通過整合 NVIDIA 的 H200 Tensor Core GPU 來增強其雲服務，進一步提升 Gemma 的能力。對於企業開發者來說，NVIDIA 提供了一套豐富的工具生態系統，如 NeMo 框架和 TensorRT-LLM，以便他們能夠無縫地調整和部署 Gemma 於生產應用中。

但這還不是全部。NVIDIA 正在將 Gemma 整合到其 Chat with RTX 技術中，為用戶提供一種創新的方式，直接從他們的 RTX 驅動 PC 與生成式 AI 進行交互。這不僅提供快速的結果，還確保了用戶數據的私密性，這在當今的數位時代是一個關鍵考慮。

請繼續關注更多有關 Gemma 與 NVIDIA 如何重塑 AI 交互，使其對開發者和用戶都更加可及和高效的更新。

閱讀更多

為何 GTC 2024 是年度必參加的科技盛會

GTC 2024 NVIDIA AI 科技會議 機器人學 創新 Silicon Valley
2024-02-22

為何 GTC 2024 是年度必參加的科技盛會

NVIDIA 的 GPU 技術會議 (GTC) 將於 2024 年 3 月 18 至 21 日，把 San Jose Convention Center 轉變為科技愛好者的夢想之地。這個活動不僅僅是一場會議；它是一些科技界最聰明頭腦的聚集，包括從 Adobe 到 Zoox 的頂尖公司。以下是你絕對不能錯過它的原因：

1. 創新的會議和展覽：擁有超過 900 場會議和 300 個展覽，GTC 2024 是最新 AI、計算等科技的中心。從 Amazon 到 Pixar，最亮眼的行業領袖將分享他們對科技未來的見解。
2. AI 的開拓者：體驗一次終身難忘的小組討論，特邀 Transformer 神經網絡架構的原始建築師，一個已經革命性改變 AI 的概念。由 NVIDIA 的 CEO Jensen Huang 主持，這個小組論壇承諾深入洞察 AI 的未來。
3. 獨家揭曉：成為第一批目睹下一代技術突破揭曉的人，繼去年宣布的「光速」計算突破之後。期待大型揭曉可能塑造科技未來。
4. 無與倫比的網絡：GTC 2024 提供與產業領袖和創新者連接的無與倫比機會。每一次對話都可能是開始的大事，從新角色到開創性項目。
5. 深入 AI 和機器人學：展覽不僅僅是展示；它們是互動式學習體驗。親手體驗最新的 AI 和機器人技術進步，保持在技術前沿。
6. 培養多樣性和創新：在 Women In Tech 早餐會開始您的一天，探索包括生成性 AI 藝術裝置在內的多元體驗。GTC 在一個獨特吸引人的環境中培養創造力。
7. 向願景家學習：直接與來自如 Disney Research 和 Google DeepMind 這樣的巨擘的思想領袖接觸。這些會議不僅僅是講座；它們是獲得可以推動您職業發展的實用知識的機會。

位於創新之心 Silicon Valley，GTC 2024 不僅僅是一場會議。它是一個協作平台，將與會者與尖端技術和想法連接。這是參與 AI 和科技轉型瞬間的機會。

閱讀更多