

Revolutionäre NPU-Technologie: 60% schnellere KI bei 44% weniger Energie!

Koreanische Forscher entwickeln energieeffiziente NPU-Technologie, die AI-Leistung um 60% steigert und Energiekosten senkt.



NACHRICHTEN AG

Tokio, Japan - Was geht in der Welt der künstlichen Intelligenz? An einem spannenden Punkt der Technologiewende stehen die Forscher am Korea Advanced Institute of Science and Technology (KAIST). Sie haben eine neuartige, energieeffiziente NPU-Technologie entwickelt, die die Leistung bei der Verarbeitung von generativen AI-Modellen um beeindruckende 60% steigert, während der Energieverbrauch um 44% gesenkt wird. [cloudcomputing-news.net] berichtet von den Fortschritten, die im Rahmen dieser Forschung unter der Leitung von Professor Jongseok Park, in Zusammenarbeit mit HyperAccel Inc., erzielt wurden.

Aktuelle große AI-Modelle wie OpenAIs ChatGPT-4 und Googles

Gemini 2.5 sind wahre Rechenmonster, die eine hohe Speicherkapazität und Bandbreite benötigen. Thus ist es kein Wunder, dass Unternehmen wie Microsoft und Google Hunderttausende von NVIDIA GPUs einkaufen, um diesen Anforderungen gerecht zu werden. Doch das könnte sich ändern, denn die neue NPU-Technologie zielt darauf ab, bestehende Engpässe in der AI-Infrastruktur zu lösen und hält dabei den Energieverbrauch in Schach.

Fortschritt durch spezialisierte Hardware

Im Kern dieser Innovation steht die Optimierung des Speichermanagements durch den Einsatz einer spezialisierte Hardwarearchitektur. Minsu Kim und Dr. Seongmin Hong von HyperAccel Inc. haben gemeinsam an der Forschung mitgearbeitet, deren Ergebnisse auf dem International Symposium on Computer Architecture (ISCA 2025) in Tokio präsentiert wurden. Die Forschungsarbeit trägt den Titel Oaken: Fast and Efficient LLM Serving with Online-Offline Hybrid KV Cache Quantization und fokussiert sich stark auf die KV-Cache-Quantisierung, die in generativen AI-Systemen einen enormen Teil des Speicherverbrauchs ausmacht. Mit dieser Technologie kann die gleiche Leistung mit weniger NPU-Geräten erreicht werden, was nicht nur Kosten senkt, sondern auch die Umwelt schont.

Durch die Implementierung eines dreigleisigen Quantisierungsalgorithmus können die Ingenieure den Genauigkeitsverlust während der Inferenz minimal halten. Dazu gehören kämpferisch passende Strategien wie schwellenbasierte Online-Offline-Hybridquantisierung, Gruppenverschiebungsquantisierung und eine fusionierte dichte sowie spärliche Kodierung. Diese Ansätze sind dafür ausgelegt, mit der begrenzten Bandbreite und Kapazität der aktuellen Systeme optimal umzugehen und damit die Langlebigkeit der KI-Infrastruktur zu fördern.

Ein Schritt in die nachhaltige Zukunft

Energieeffizienz in der KI lässt sich jedoch nicht mit einer einzelnen Lösung bewerkstelligen. Insgesamt stellt die Forschung einen bedeutenden Fortschritt in Richtung nachhaltiger AI-Infrastruktur dar, dessen Auswirkungen sich erst mit der Skalierung und Implementierung in kommerziellen Umgebungen richtig entfalten werden. Der CO2-Fußabdruck von AI-Cloud-Diensten könnte durch die neuen Entwicklungen erheblich gesenkt werden. Hierbei spielt auch die aktuelle Forschung zu weiteren Techniken eine Rolle, die auf die Reduktion des Energieverbrauchs abzielen, wie beispielsweise die Modellbeschneidung, Quantisierung oder effiziente Architekturen. Wissenschaftler und Unternehmen sind gefordert, diese Ziele in Einklang mit steigenden Anforderungen an die Leistung zu bringen.

Doch was ist der Schlüssel für die Zukunft der KI-Energieeffizienz? Laut [focalx.ai] sind nicht nur innovative Hardware und spezialisierte Algorithmen gefragt, sondern auch eine verstärkte Zusammenarbeit zwischen Technikern und Unternehmen, um nachhaltige Lösungen zu schaffen. Die Herausforderungen sind vielfältig: von der Balance zwischen Leistung und Effizienz bis hin zu möglichen Hardwareeinschränkungen.

Insgesamt stehen wir vor einer spannenden Entwicklung, die uns zeigen könnte, wie leistungsstarke, energieeffiziente Infrastrukturen in der KI umgesetzt werden können. Es bleibt abzuwarten, welche Impulse diese neuartige NPU-Technologie in der Branche setzen wird und wie sie dazu beitragen kann, die ökologischen Fußabdrücke der Unternehmen zu verringern.

Behalten wir also die Entwicklungen im Auge – die Zukunft klingt vielversprechend!

Details	
Ort	Tokio, Japan

Details	
Quellen	• www.cloudcomputing-news.net • itc.ua • focalx.ai

Besuchen Sie uns auf: wom87.de