Bahnbrechende Effizienz in der Bereitstellung von NLP-Modellen

Da die Modelle für Natural Language Processing (NLP) immer größer werden, nimmt die GPU-Leistung und -Funktionalität exponentiell ab, sodass Unternehmen in einer Vielzahl von Branchen eine qualitativ hochwertigere Sprachverarbeitung benötigen, jedoch zunehmend durch die heutigen Lösungen eingeschränkt werden.

Moderne industrielle NLP-Modelle folgen während ihrer gesamten Lebensdauer einem Rhythmus. Sie beginnen mit einem einmaligen, aufgabenunabhängigen Vortraining und durchlaufen dann ein aufgabenspezifisches Training mit sich schnellen ändernden Benutzerdaten. Diese regelmäßig aktualisierten Modelle werden schließlich bereitgestellt, um massive Online-Inferenzanforderungen von Anwendungen zu bedienen.

Ein aktueller aktiver Forschungstrend ist der Einsatz modernster NLP-Modelle wie BERT für die Online-Inferenz. Da die Modelle von Jahr zu Jahr größer werden, wird immer häufiger darüber diskutiert, wie diese Modelle in Echtzeit-Pipelines implementiert werden können. Um eine Implementierung in der Praxis zu ermöglichen, wurden verschiedene Techniken entwickelt, um große Modelle auf kompakte Varianten zu reduzieren. In Anwendungen wie digitalen Assistenten und Suchmaschinen sind diese kompakten Modelle der Schlüssel, um Modelle mit geringer Latenz und hoher Genauigkeit zu erzielen, die den Service-Level-Anforderungen gerecht werden.

SambaNova Systems bietet eine Lösung zur Untersuchung und Bereitstellung dieser kompakten Modelle – von einer einzelnen SambaNova Systems Reconfigurable Dataflow Unit (RDU) bis zu mehreren SambaNova DataScale-Systemen – und bietet so gegenüber herkömmlichen Accelerators beispiellose Vorteile für eine hochpräzise Online-Inferenz mit geringer Latenz.

BEWÄHRTE LEISTUNGSSTÄRKE DER DATENFLUSSAUSFÜHRUNG AUF RDU

Die Latenz von kompakten Modellen auf GPUs wird durch den Kernel-basierten Ausführungsmodus grundsätzlich begrenzt. Für Online-Inferenz mit Batchgröße 1 kann der Overhead von Kontextwechsel und Zugriff auf Off-Chip-Speicher für Betriebs-Kernels in der herkömmlichen Architektur die Latenz dominieren. Die SambaNova-RDU baut auf der SambaNova Systems Reconfigurable Dataflow Architecture (RDA) auf, um diese Barriere zu beseitigen. Insbesondere bei einem kürzlich vorgeschlagenen kompaktem BERT-Modell, TinyBERT, kann die RDU im Vergleich zur V100-GPU für MNLI, einer beliebten Textklassifizierungs-Task, um 5,8-mal bessere Latenzzeiten erreichen.

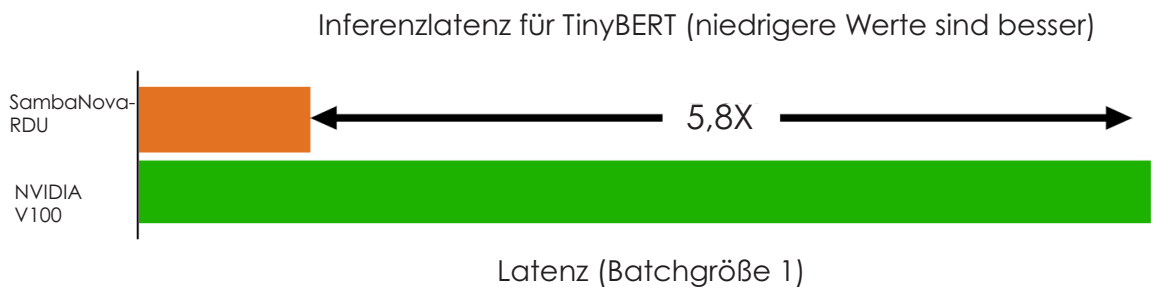

ABB. 1: LATENZVERGLEICH FÜR ONLINE-INFERENZ

In Anwendungen wie digitalen Assistenten oder Suchmaschinen sind die Eingabedaten NLP-Token mit kurzer Sequenzlänge, z. B. Abfragen des Smartphone-Assistenten wie „Wie ist das Wetter in Berlin?“ Bei diesen Szenarien hat eine reduzierte Sequenzlänge in der Regel eine zu vernachlässigende Auswirkung auf die Genauigkeit von kompakten Modellen. Das ist ein weiteres Merkmal, das eng mit dem Latenzvorteil von RDUs verbunden ist. Während die Latenz von GPUs bei kompakten Modellen mit reduzierter Sequenzlänge gesättigt wird, verbessert sich die Latenz von RDUs bei reduzierter Sequenzlänge.

Wie Abb. 2 zeigt, erreicht das Modell TinyBERT bei der MNLI-Benchmark-Task, die wir als Hilfe heranziehen, über Sequenzlängen von 64 bis 256 Spitzenwerte bei der Modellgenauigkeit. Abb. 3 zeigt, dass die GPU über die Sequenzlängen hinweg dieselbe Latenz aufweist. Allerdings erhöht sich die Beschleunigung der RDU gegenüber der GPU bei einer reduzierten Sequenzlänge von 64 auf das 8,7-Fache.

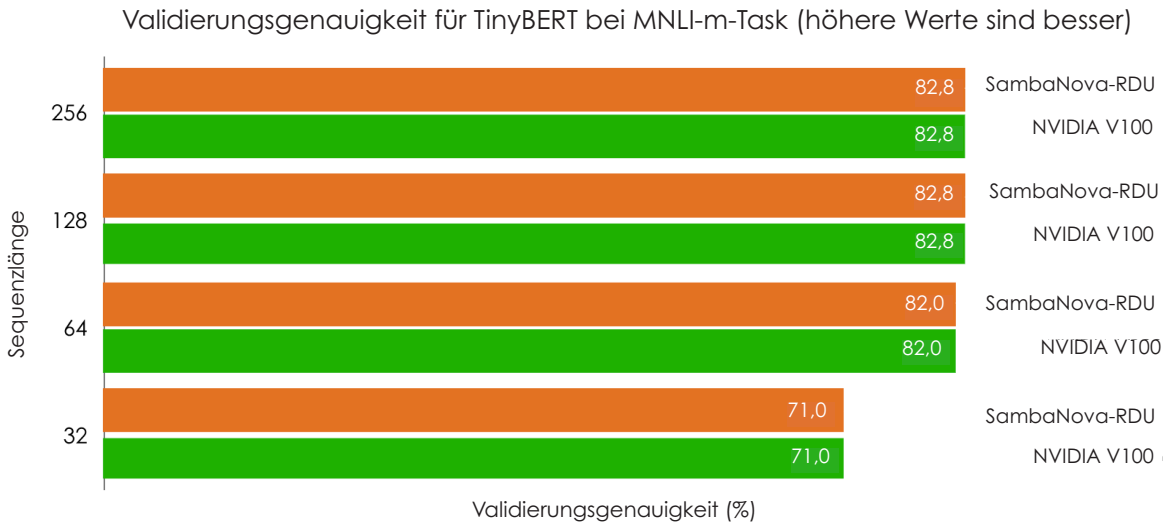

ABB. 2: RDU- UND GPU-MODELLGENAUIGKEIT FÜR UNTERSCHIEDLICHE SEQUENZLÄNGEN

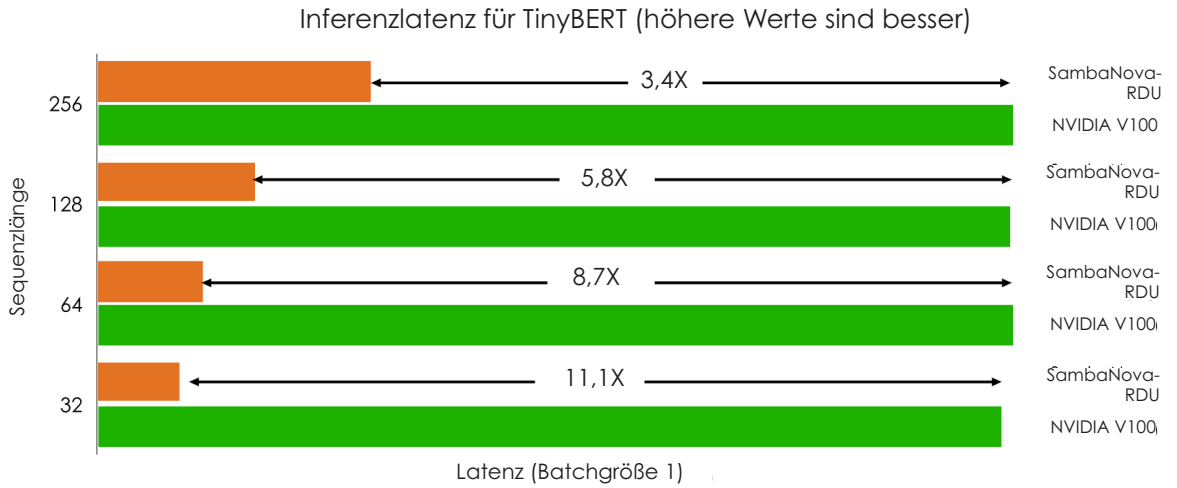

ABB. 3: BALKENDIAGRAMM DER RDU- UND GPU-LATENZ FÜR UNTERSCHIEDLICHE SEQUENZLÄNGEN

HÖHERE GENAUIGKEIT MIT SAMBANOVA SYSTEMS DATASCALE

Der auf Datenflüsse ausgelegte Chip von SambaNova bietet beispiellose Funktionen für Online-Inferenz mit geringer Latenz bei kompakten Modellen. Unter Ausnutzung dieser Funktionen haben die Forschungslabore von SambaNova außerdem gezeigt, dass sich mit dem SambaNova DataScale-Komplettsystem (8 Sockets) eine Genauigkeit mit Spitzenwerten erzielen lässt und es bei kompakten NLP-Modellen gleichzeitig eine Inferenz mit geringer Latenz ermöglicht.

Eine Studie des Forschungslabors von SambaNova Systems zeigt, dass die Genauigkeit von TinyBERT mit einer Mehrheitswahl (Majority-Voting-Prinzip) über mehrere Modellinstanzen erheblich gesteigert werden kann (Abb. 4). Das SambaNova DataScale-System ist perfekt für die effiziente Nutzung dieser Genauigkeitssteigerungen konzipiert. SambaNova demonstriert, dass mehrere TinyBERT-Modelle auf allen acht Sockets des SambaNova DataScale-Systems bereitgestellt werden können. Wie Abb. 5 zeigt, steigt beim Ensembling von TinyBERT-Modellen die Klassifizierungsgenauigkeit um 0,4 % (gegenüber zu vernachlässigenden Latenzeinbußen) im Vergleich zu einem einzigen TinyBERT-Modell mit einer RDU.

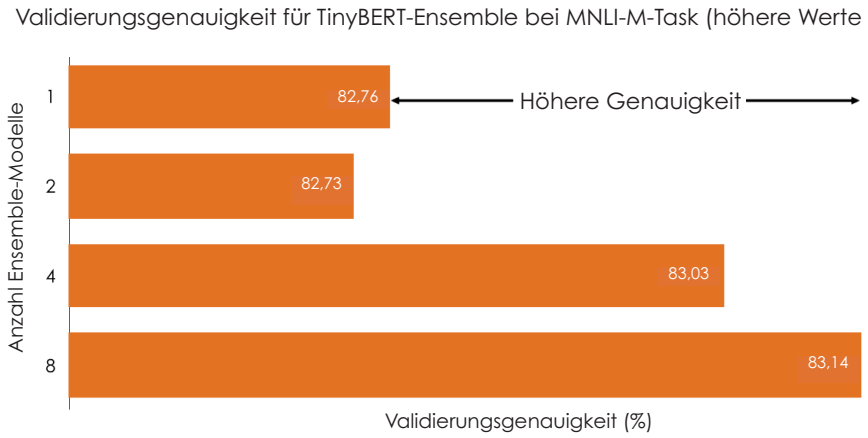

ABB. 4: MODELLGENAUIGKEIT MIT UNTERSCHIEDLICHER ANZAHL VON EXPERTEN FÜR DAS ENSEMBLE

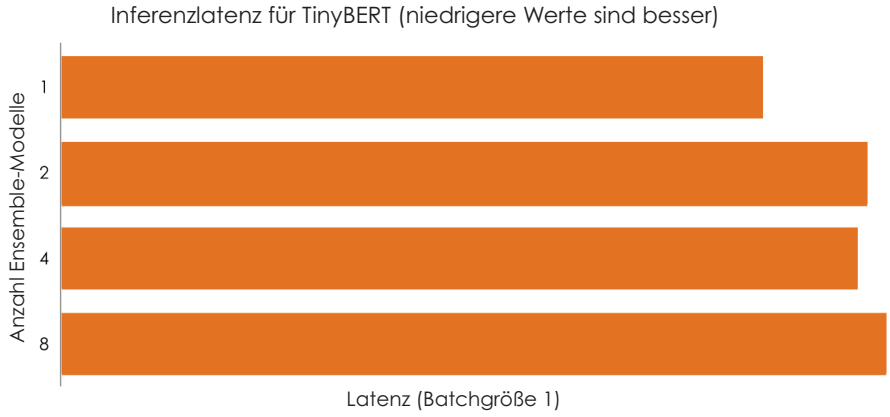

ABB. 5: VERGLEICH DER LATENZ FÜR EINZELNEN TINYBERT MIT EINER RDU GEGENÜBER 8 EXPERTEN AUF 8-SOCKET-SYSTEMEN

Das kompakte BERT-Modell ist nur ein wichtiger Fall, bei dem SambaNova Systems DataScale eine maßgeschneiderte Lösung für hochgenaue Online-Inferenz mit geringer Latenz bietet.