ELEKTRONIK PRAXIS
Fachwissen für Elektronik Professionals

FPGAs beschleunigen Inferenz in smarten Edge-Produkten

03.12.19 | Autor / Redakteur: Hussein Osman * / Michael Eckstein

Intelligenz am Edge: Endgeräte wie Überwachungskameras können mithilfe von Embedded-KI Inferenzberechnungen übernehmen.(Bild: gemeinfrei / Pixabay)

Firmen zum Thema

Lattice GmbH

Syslogic GmbH

ET System electronic GmbH

FlowCAD EDA-Software Vertriebs GmbH

Intelligente IoT-Produkte fordern Entwickler heraus: Sie benötigen immer mehr Rechenleistung, müssen kompakt sein und dürfen nur wenig Strom verbrauchen. Eine mögliche Lösung sind Field Programmable Gate Arrays.

Für geringen Stromverbrauch und kompakte Abmessungen optimierte FPGAs können dort zum Einsatz kommen, wo hohe Energieeffizienz gefordert und wenig Platz verfügbar ist. Zum Entwickeln von Edge-Produkten mit KI-Unterstützung hat Lattice den sensAI-Technologie-Stack zusammengestellt. Grundlage bilden modulare Hardwareplattformen wie iCE40 UPduino 2.0 mit HM01B0 Shield und das ECP5-basierte Embedded Vision Development Kit (EVDK).

Die FPGAs lassen sich mit Soft-IP programmieren, etwa für das Entwickeln Neuronaler Netze. Passend dazu umfasst sensAI IP für einen kompakten CNN-Beschleuniger (Convolutional Neural Network). Für die ECP5-FPGAs ist ein vollständig parametrierbarer CNN-Beschleuniger-IP-Core verfügbar. Diese IPs unterstützen variable Quantisierung und ermöglichen es Entwicklern, Datengenauigkeit und Stromverbrauch für ihre Applikation abzustimmen.

Die erweiterte Unterstützung für Machine-Learning-Frameworks soll Entwicklern einen durchgängigen und benutzerfreundlichen Design-Flow bieten. Neben Caffe und TensorFlow unterstützt sensAI auch Keras, ein in Python geschriebenes neuronales Open-Source-Netzwerk, das auf TensorFlow, Microsoft Cognition Toolkit oder Theano läuft. Keras soll Ingenieuren helfen, mit geringem Aufwand mit mehrschichtigen Neuronalen Netzen (Deep Neural Networks) experimentieren zu können.

Die Software ist eine benutzerfreundliche, modulare und erweiterbare Umgebung für schnelles Prototyping. Keras war ursprünglich nicht als ein autonomes Machine-Learning-Framework konzipiert, sondern als eine Schnittstelle. Sie bietet Entwicklern ein hohes Abstraktionsniveau, das die Entwicklung von Deep-Learning-Modellen beschleunigen hilft.

Darüber hinaus hat Lattice seinen sensAI Neural Network Compiler dahingehend erweitert, dass das Tool jetzt bei der Konvertierung eines Machine-Learning-Modells in die Firmware-Datei automatisch die genauesten Fraction-Bits wählt. Der aktualisierte sensAI-Stack umfasst auch ein Hardware-Debugging-Tool, das es dem Anwender ermöglichen soll, lesend oder schreibend auf sämtliche Netzwerkschichten zuzugreifen.

Nach der Softwaresimulation werden Entwickler wissen wollen, wie sich ihr Netzwerk auf echter Hardware verhält. Mithilfe dieses Tools erhalten Entwickler innerhalb von Minuten Einblick in die reale Hardware. So lassen sich Edge-Lösungen mit einem Stromverbrauch zwischen 1 mW und 1 W auf kompakten Plattformen mit einer Grundfläche zwischen 5,5 und 100 mm^2 entwickeln.

Vier Anwendungsbeispiele für FPGA-Beschleuniger

Lattice sieht vier mögliche Anwendungsszenarien, in denen die Beschleunigung durch den neuen sensAI-Stack zum Tragen kommt. Im ersten Szenario verwenden Entwickler den Stack zur Realisierung von Lösungen, die im Standalone-Modus laufen sollen. Diese Systemarchitektur auf der Basis von Lattice iCE40 UltraPlus oder ECP5 FPGAs ermöglicht Entwicklern die sichere Implementierung von permanent betriebsbereiten, integrierten Lösungen mit kurzen Latenzzeiten, bei denen FPGA-Ressourcen zur Systemsteuerung

verwendet werden können. Ein typisches Beispiel für ein solches Szenario ist ein Stand-alone-Sensor für die Anwesenheitserkennung oder Zählung von Personen.

Bild 1: Lattice sensAI ist ein vollständiger Hard- und Software-Lösungsstapel für die Entwicklung von Edge-AI-Anwendungen.(Bild: Lattice)

Möglich ist auch die Implementierung zweier unterschiedlicher Arten von Pre-Processing-Lösungen. Die erste analysiert Sensordaten vorab mithilfe eines stromsparenden iCE40-UltraPlus-FPGAs und minimiert dadurch die Kosten der Datenübertragung zu einem SoC oder in die Cloud. In einer smarten Türklingel beispielsweise ließen sich so die eingehenden Sensordaten dahingehend analysieren, ob ein Mensch vor der Tür steht oder lediglich eine Katze vorbeiläuft. Nur im ersten Fall würde das SoC aufgeweckt oder eine Verbindung zur Cloud hergestellt und eine genauere Analyse initiiert. Dadurch sinken die Datenübertragungskosten und der Stromverbrauch. Dieser Ansatz reduziert das vom System zu analysierende Datenvolumen und die dafür aufzuwendende Energie erheblich – wichtig für ständig verfügbare Edge-Anwendungen.

In einer weiteren Pre-Processing-Anwendung beschleunigt ein ECP5-FPGA ein NN. Immer öfter möchten Unternehmen eine bewährte MCU-basierte Legacy-Lösung um KI-Fähigkeiten erweitern, ohne Komponenten ersetzen oder ihr Design grundlegend überarbeiten zu müssen. In einigen Fällen ist die MCU relativ leistungsschwach. Ein typisches Beispiel könnte eine smarte Industrie- oder Heimanwendung sein, in der Bilddaten vor der Durchführung der Analyse nach bestimmten Kriterien gefiltert werden müssen. Hier könnten Entwickler entweder eine weitere MCU hinzufügen, wobei sie eine zeitaufwendige Revalidierung ihres Designs in Kauf nehmen müssten. Oder sie fügen zwischen MCU und Rechenzentrum einen Beschleuniger ein, der die Daten nachbereitet und das zur Cloud zu sendende Datenvolumen minimiert. Dieser Ansatz ist besonders attraktiv für Entwickler von IoT-Geräten, die mit KI-Funktionalität ausgestattet werden sollen.

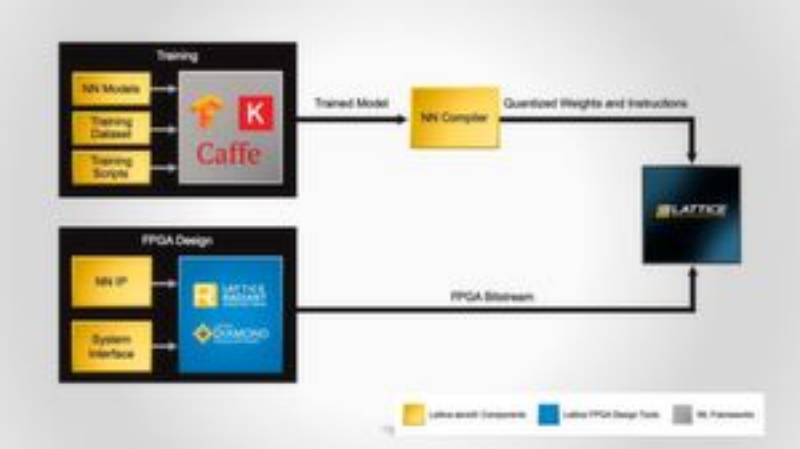

Bild 2: Der sensAI-Design-Flow umfasst branchenführende Machine-Learning-Frameworks, Trainingsdaten und -skripte sowie Neural-Network-IP, die für Design und Training von Edge-AI-Geräten erforderlich sind.(Bild: Lattice)

sensAI-Beschleuniger können auch Post-Processing-Funktionen übernehmen. Anstatt Daten zur Verarbeitung in die Cloud zu senden, kann ein ECP5-FPGA am Netzwerkrand als Beschleuniger eingesetzt werden. So lassen sich beispielsweise in Überwachungskameras oder smarten Türklingeln das Hauptsystem-SoC durch ein ECP ergänzen. Dieses kann in den Bilddaten, die vom Image Signal Processor (ISP) im SoC bereinigt wurden, Muster erkennen. Durch die Mustererkennung am Netzwerkrand erspart man sich die Übertragung sensibler Daten in die Cloud. Das spart Bandbreite und ermöglicht einen besseren Datenschutz.