EXPERTARTIKEL

TVÅ REFERENSKONSTRUKTIONER SOM VISAR PÅ SPÄNNVIDDEN I SENSAI.

Närvarodetektering med hjälp av en CMOS-sensor som har en upplösning på 64×64×3 och ansluts via en VGG8-buss. Systemet har en uppdateringsfrekvens på fem bilder per sekund och förbrukar 7 mW när det körs på en iCE40 UltraPlus.

För att räkna människor används en CMOS-sensor med en upplösning på 128×128×3 som ansluts över en VGG8-buss. Uppdateringsfrekvensen är 30 bilder per sekund och systemet drar 850 mW med en ECP5-FPGA som har 85 000 logikceller.

Kör AI i ändnoderna

Av Deepak Boppana, Lattice Semiconductor

Deepak Boppana har jobbat på Lattice Semiconductor sedan 2012. Datorseende, hårdvarucybersäkerhet, artificiell intelligens och maskininlärning – det är några av hans arbetsområden idag. Tillsammans med chefer för produktgrupper och marknadssegment söker han med ljus och lykta efter nya tillväxtmöjligheter för Lattice.

Ett sätt att snabbt få ut mer beräkningskraft till ändnoderna utan att förlita sig på molnet är att använda programmerbar logik, FPGA:er. Parallelliteten hos dessa kan exempelvis användas för att snabba upp neurala nät.

Genom att använda FPGA:er som är optimerade för låg effektförbrukning och har litet fotavtryck kan konstruktörerna uppfylla de strikta krav på effekt och storlek som finns hos dagens produkter för både konsumenter och industrin. Två exempel är Lattice FPGA-familjer iCE40 UltraPlus och ECP5. De drar mellan 1 mW och 1 W och upptar inte mer än 5,5 till 100 kvadratmillimeter. Genom att kombinera ultralåg effekt, hög prestanda och noggrannhet med stöd för äldre kommunikationsgränssnitt har dessa FPGA:er tillräcklig flexibilitet för att möta alla behov.

För att underlätta utvecklingsarbetet har Lattice tagit fram SensAI, industrins första teknikstack som ger konstruktörerna alla de verktyg de behöver för att utveckla produkter med låg effektförbrukning och hög prestanda. Det handlar om tillämpningar för smarta hem, smarta fabriker, smarta städer och smarta bilar. SensAi lanserades 2018 och är en kombinerad hårdvaru- och mjukvarulösning avsedd för AI-inferenser i ändnoderna där produkterna ständigt är aktiva och därför måste ha låg effektförbrukning.

Vad finns då i detta ekosystem? De modulära hårdvaruplattformarna från Lattice, som iCE40 UPduino 2.0 med skölden HM01B0 och det ECP5-baserade utvecklingspaketet, ger en stabil grund för utveckling av applikationer. UPduino kan användas för AI-lösningar som bara konsumerar några få milliwatt medan EVDK stödjer tillämpningar som kräver mer effekt, men som normalt ligger under 1W.

Det går enkelt att instansiera mjukvara i form av IP-block i en FPGA för att snabba upp utvecklingsarbetet. Det finns kompakta IP-block för CNN (Convolutional Neural Networks) som gör det möjligt att implementera tillämpningar baserade på deep learning i FPGA:an iCE40 UltraPlus.

SensAI har också ett fullt parametriserbart acceleratorblock för CNN-nät som kan implementeras i Lattice FPGA-familj ECP5. Detta IP-block har stöd för data med valbart antal bitar vilket i sin tur gör att konstruktörerna kan bestämma om de vill ha högre noggrannhet eller lägre effektförbrukning.

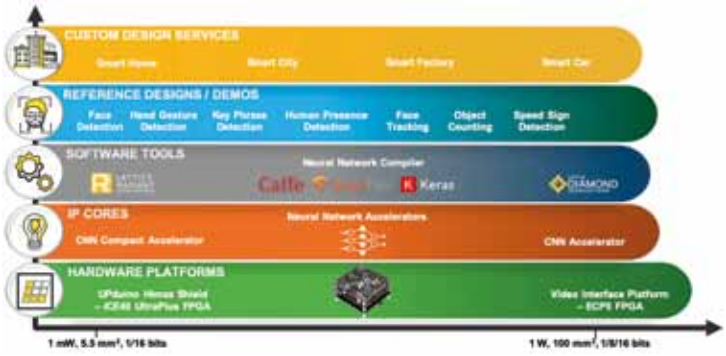
SENSAI GÖR DET MÖJLIGT för konstruktörerna att utforska olika möjligheter och se hur olika val påverkar resultatet via en lättanvänd verktygskedja. Det går att träna neurala nät med standardiserade lösningar som Caffe, TensorFlow och Keras. Utvecklingsmiljön har också en kompilator för neurala nät som mappar de tränade näten på en heltalsrepresentation och har stöd för olika kvantifieringsnivåer i vikterna. Konstruktören kan använda kompilatorn för att analysera, simulera och kompilera olika nät som ska implementeras på Lattice IP-kärnor utan att för den sakens skull ha kunskap om RTL. Konstruktörerna kan sedan använda traditionella FPGA-verktyg som Radiant och Diamond från Lattice för att implementera den kompletta konstruktionen.

För att snabba upp konstruktionsfasen har SensAI ett växande antal referenskonstruktioner och exempel. Listan innehåller bland annat ansiktsigenkänning, handgester, nyckelfraser i tal, närvarodetektion, ansiktsföljning, räkning av objekt och avläsning av hastighetsskyltar. Slutligen så behöver designteamen ofta en unik kompetens för att slutföra arbetet. För att möta det behovet har Lattice byggt relationer med ett antal konsultföretag runt om i världen som kan stödja kunderna lokalt om de själva saknar den nödvändiga kunskapen om artificiell intelligens och maskininlärning.

För att möta de snabbt ökande kraven på prestanda för AI i ändnoderna släppte Lattice under 2019 en större prestandaförbättring i SensAI, liksom förbättringar av designflödet. Den uppdaterade stacken ger nu tio gånger mer prestanda än den ursprungliga. Förbättringen beror på flera saker inklusive snabbare minnesaccess, en förbättrad CNN-kärna, åttabitarskvantifiering, sammanslagna lager och en dubbel DSP-kärna.

Åttabitarsdata halverar inte bara antalet minnesaccesser utan gör det också möjligt

Lattice SensAI är en komplett lösning med hård- och mjukvara



att använda bilder med högre upplösning vilket ger noggrannare resultat.

FÖR ATT YTTERLIGARE LYFTA PRESTANDA har Lattice optimerat beräkningarna i convolution-skiktet i SensAI:s neurala nät och därmed minskat den totala beräkningstiden. Företaget har dubblerat antalet tillåtna beräkningsmotorer i kretsarna vilket ytterligare minskar beräkningstiden med uppskattningsvis 50 procent.

Givet att Lattice har ökat prestanda utan att effektförbrukningen ökat har konstruktörerna nu möjlighet att byta till en krets med färre logikgrindar i ECP5-familjen. Ett exempel på förbättringarna är närvarodetektering med hjälp av en CMOS-sensor som har en upplösning på 64x64x3 och ansluts via en VGG8-buss. Systemet har en uppdateringsfrekvens på fem bilder per sekund och förbrukar 7 mW när det körs på en iCE40 UltraPlus.

Ett annat exempel handlar om att räkna människor och nyttjar även det en CMOS-sensor med en upplösning på 128×128×3 över en VGG8-buss. Uppdateringsfrekvensen är 30 bilder per sekund och systemet drar 850 mW med en ECP5-FPGA som har 85 000 logikceller.

SensAI stödjer dessutom nya neuronnätsmodeller, verktyg för maskininlärning och snabbare designcykler. Nya referenskonstruktioner som går att modifiera underlättar utvecklingsarbetet av populära tillämpningar som närvarodetektering samtidigt som en växande skara samarbetspartners erbjuder sina tjänster till dem som så önskar. Blockdiagrammet i figur 4, visar vilka komponenter som Lattice nu erbjuder inklusive träningsmodeller, dataset för träning, träningsskript, uppdaterade IP-block med neurala nät och en uppdaterad kompilator för neurala nät.

En del av arbetet med att förbättra användarupplevelsen handlar om att bredda antalet verktyg för maskininlärning. Den första versionen av SensAI stödde Caffe och Tensorflow medan den nya versionen även fungerar med open sourceverktyget Keras som är skrivet i Python och designat för att köras ovanpå Tensorflow, Microsoft Cognition Toolkit och Theano. Det ger utvecklarna en möjlighet att snabbt testa idéer på djupa neurala nät. Keras gör det också möjligt att snabbt ta fram en prototyp.

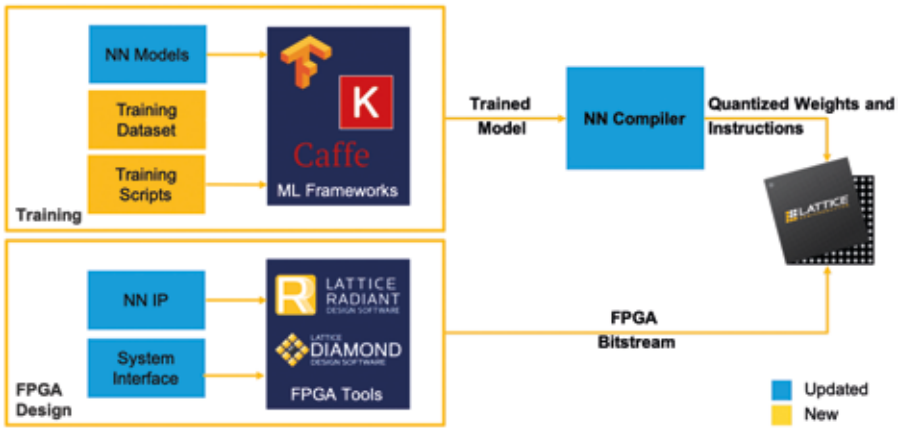
Det var ursprungligen tänkt som ett gränssnitt – inte ett fristående verktyg för maskininlärning – och ger en hög abstraktionsnivå som snabbar upp utvecklingsarbetet av modeller baserade på deep learning.

För att ytterligare underlätta arbetet har Lattice uppdaterat kompilatorn i SensAI så att den automatiskt väljer det antal bitar som ger det noggrannaste resultatet när den konverterar modellen till en körbar fil. SensAI har också en hårdvarudebugger som gör att användaren kan läsa och skriva till varje lager i nätet.

Efter simuleringen av mjukvaran vill utvecklarna veta hur bra deras neurala nät kommer att fungera på den riktiga hårdvaran. Verktyget gör att det bara tar minuter att få svaret.

Slutsats

De närmaste åren kommer att vara avgörande för utvecklingen av smarta produkter som används långt ut i näten och som ständigt är påslagna. I takt med att tillämpningarna blir allt mer komplexa kommer utvecklarna att behöva verktyg som ger bättre prestanda till lägre effektförbrukning. ■



Designflödet i SensAI inkluderar programvara för maskinlärning, dataset för träning och skript plus den IP för neurala nät som behövs för design och träning av AI i ändnoder.