Styrkretsar nära kanten

EXPERTARTIKEL

Djup maskininlärning i edge-tillämpningar kommer utföras av styrkretsar

Av Ali Osman Ors, NXP Semiconductors

Ali Osman Ors har 20 år i halvledarindustrin bakom sig. Maskininlärning och datorseende är hans specialiteter och han har tidvis ansvarat för NXP:s strategier inom det området, bland annat för platformar för adas och autonoma fordon. Idag ansvarar han för NXP:s strategier inom edge-processorer. Ali Osman Ors rekryterades från CogniVue där han var forsknings- och utvecklingschef för dess systemkretsar inom AI och bild. Han läste till ingenjör i Ottawa, Kanada.

Intresset för att göra maskininlärning direkt på video- och stillbildskameror med ML-uppgraderade mikrostyrkretsar har vuxit i takt med att IoT-beräkningar flyttar allt närmare ”kanten”, the edge.

Utvecklingen av approcessorer och neuronnätsacceleratorer i mikrostyrkretsar går snabbt och med jämna mellanrum dyker det upp nya smarta, effektivare lösningar. Trenden är att försöka konsolidera allt mer AI-funktionalitet, som artificiella neuronnät och applikationsprocessorer i mikrostyrkretsarna, utan att dramatiskt öka strömförbrukningen eller storleken.

Dagens maskininlärningsmodeller tränas på kraftfulla cpuer och gpuer och implementeras sedan på mikrostyrkretsar. Inferensmotorer som TensorFlow Lite används för att krympa modellerna till en storlek anpassad till mikrokontrollerns resurser. Det är enkelt att skala modellerna för att möta högre ML-krav.

Snart torde det bli möjligt att inte bara göra inferenser utan även träning på mikrostyrkretsarna, vilket kan göra dem till en ännu mer formidabel konkurrent till stora, dyra beräkningslösningar.

För bara några år sedan trodde man allmänt att maskininlärning (ML) och djup maskininlärning (DL, deep learning, djupinlärning) var något som krävde kraftfull hårdvara. I edge-system hänvisades maskininlärning och inferensdragning till gateways, edgeservrar och datacenter.

Det var en fullt rimlig hypotes för sin tid. Trenden att distribuera beräkningar mellan moln och edge hade knappt startat. Men idag är scenen ny. Efter tunga forskningsinsatser inom såväl akademi som industri är hypotesen falsifierad.

Maskininlärning kräver inte längre processorer med prestanda som räknas i biljoner operationer per sekund (teraops, Tops). Allt oftare kan maskininlärning i edgetilllämpningar utföras av mikrostyrkretsar som utrustats med integrerade acceleratorkärnor för ändamålet.

Det är ett uppdrag som de klarar utmärkt. Och de utför det dessutom till låg kostnad och med extremt låg strömförbrukning. De ansluter till molnet bara när det är absolut nödvändigt.

Det känns alltmer som om mikrostyrkretsar med ML-acceleratorer är nästa steg i den utveckling som inneburit att sensorer som mikrofoner, kameror och miljöövervakningssensorer för varje generation utrustats med allt mer beräkningskraft. Det här är för övrigt den typ av sensorer som kommer att hantera de stora datavolymerna när IoT ska realisera sin potential.

Steg för steg mot kanten

”Edge” definieras typiskt som de yttersta noderna i ett IoT-nät. Konkret kan det syfta på gateways och edgeservrar. Men det är egentligen inte där kanten går, utan den går ända framme vid sensorerna vid användaren. Det logiska är att placera så mycket beräkningskraft så nära användaren som möjligt. Och det är ett uppdrag som mikrostyrkretsar är som gjorda för.

Man skulle kanske kunna argumentera för att edgeberäkningarna borde hanteras av kortdatorer som har fantastisk prestanda numera – i kluster blir de rena rama superdatorn. Men de är fortfarande för stora och dyra för att kunna distribueras i hundratal eller tusental vilket är vad som krävs i storskaliga IoT-tillämpningar.

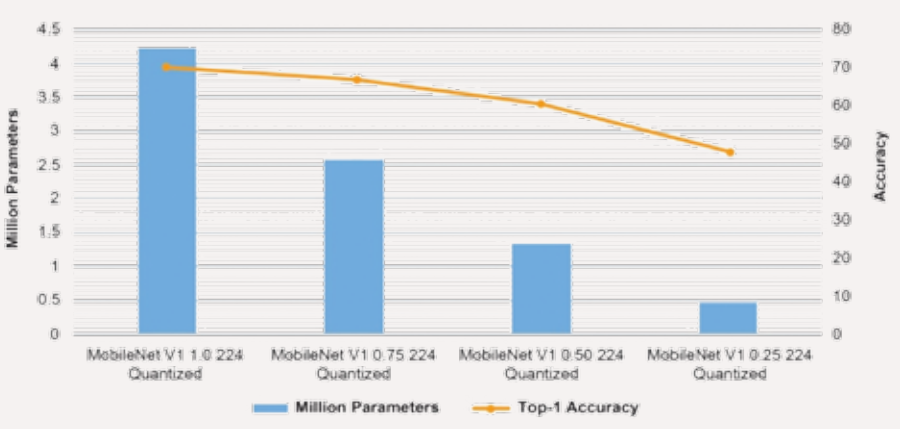
De kräver dessutom en extern likströmskälla vilket kan vara ogörligt. Förbrukningen i mikrostyrkretsar mäts å andra sidan i milliwatt. De kan drivas av några knappceller eller solceller.

Det är alltså inte så konstigt att maskininlärning i edgetillämpningar på mikrostyrkretsar har blivit ett hett område för nyutveckling. Det har till och med fått ett namn: TinyML.

Målet är att inferenser – och i sinom tid även träning – ska göras på strömsnåla, resursbegränsade komponenter, specifikt mikrostyrkretsar, snarare än på mer kraftfulla plattformar eller i molnet.

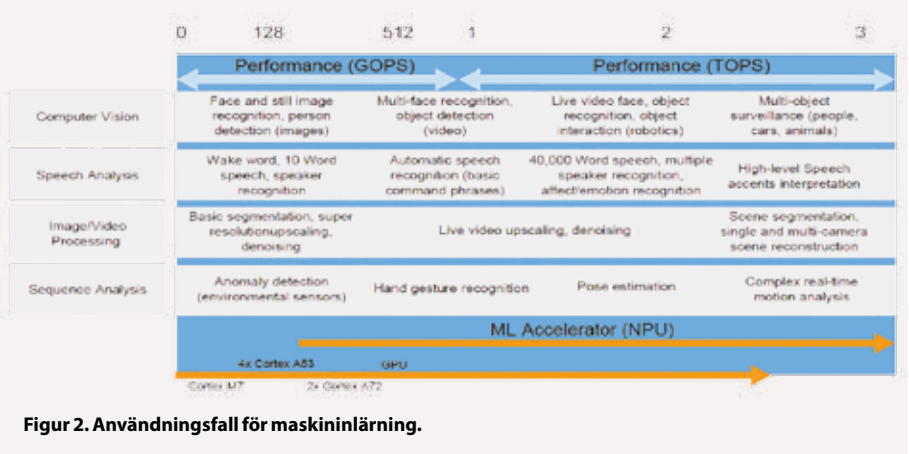
Detta kräver att neuronnätsmodellerna rättas i storlek efter sina begränsade beräknings-, lagrings- och bandbreddsresurser. Detta måste ske utan att funktionalitet och noggrannhet blir signifikant sämre.

Resurserna måste optimeras. Komponenterna tar in tillräckligt med sensordata för att kunna göra sina beräkningar, samtidigt som noggrannheten justeras ner. Även om



Figur 1. Ju mindre breddmultiplikator (α∈{1,0, 0,75, 0,5, 0,25}) desto färre parametrar och beräkningar i modellen MobileNet V1. Observera dock att steget från 1,0 till 0,75 påverkar noggrannheten bara lite grand samtidigt som antalet parametrar och beräkningar minskar dramatiskt.

vill gå på djupet



Figur 2. Användningsfall för maskininlärning.

data till slut skickas till molnet – eller möjligen först till en edge-gateway och sedan till molnet – behöver de inte skickas i lika stora volymer eftersom betydande delar av analysen redan har utförts.

Ett populärt exempel på TinyML är kamerabaserad objektidentifiering. Det begränsade lagringsutrymmet skulle normalt kräva att alla de högupplösta originalbilderna komprimeras. En kamera med egen inbyggd analyskapacitet kan istället som alternativ söka upp objekt av intresse i scenen och bara leverera bilder av dem, istället för den kompletta scenen. Det resulterar i färre bilder som därmed kan bibehålla sin höga upplösning. Vanligen tänker man sig att funktionen utförs av en kraftfullare beräkningsnod, men med TinyML går det att klara sig med mikrostyrkretsar.

Liten men ett lejon

Trots att TinyML är ett relativt nytt paradigm har det redan gett överraskande goda resultat för inferens (även på ganska klena styrkretsar) och träning (på mer kraftfulla styrkretsar) med minimal förlust i precision. På sistone har det presenterats fungerande exempel på röst- och ansiktsigenkänning, röststyrning och naturlig språkbehandling och till och med på att köra flera komplexa algoritmer för datorseende parallellt.

Med andra ord har en mikrokontroller med en prislapp under två dollar – på en Arm Cortex M7 på 500 MHz med mellan 28 och 128 kbyte minne – numera prestanda nog för att addera intelligens till en sensor.

Mikrostyrkretsar på denna pris- och prestandanivå har inte bara Ethernet, USB, SPI och säkerhetsfunktioner som AES-128 utan även externt minne av flera slag, SPDIF och I2C. Vidare har de olika sensorer, Bluetooth och Wi-Fi eller åtminstone gränssnitt till dem.

För ett lite högre pris kan du addera en Arm Cortex-M7 på 1 GHz, en Cortex-M4 på 400 MHz, 2 Mbyte RAM och grafikacceleration. Strömförbrukningen är vanligen inte mer än några milliampere från en 3,3 VDC-matning.

Några ord om TOPS

Konsumenter är inte ensamma om att förenkla allt till en enda parameter när de bedömer prestanda. Konstruktörer gör det hela tiden. Marknadsavdelningar älskar det eftersom det gör det enkelt att sätta en rubrik som skiljer tydligt mellan olika produkter i sortimentet.

Det är i alla fall så det känns. Det klassiska exemplet är att cpu:er år efter år bara värderades efter sin klockfrekvens. Lyckligtvis för både konstruktörer och konsumenter är detta inte längre fallet. Att bedöma en cpu efter ett enda parameter är ungefär som att betygsätta en bil från motorns maxvarvtal. Det är inte helt meningslöst, men har väldigt lite att göra med hur kraftfull motorn är eller hur bra bilen presterar, eftersom det bestäms av många faktorer tillsammans.

Olyckligtvis har neuronnätsacceleratorer gått i samma fälla. Mikroprocessorer och mikrostyrkretsar för maskininlärning döms återigen efter ett enda värde som är på miljarder eller biljoner operationer per sekund, GOPS och TOPS. Det är en enkel siffra att lägga på minnet. Men i praktiken är det bara ett ensamt mätvärde som även om det är precist och korrekt bara är artificiellt – som en teoretisk mätning i labb snarare än under verklig drift.

Till exempel tar Tops-måttet inte hänsyn till minnesbandbredd, cpu-overhead och för- och efterbearbetning. När dessa och andra faktorer beaktas – genom att till exempel mäta prestanda på systemnivå på ett specifikt kort i verklig drift – kan den gott bara vara 50 eller 60 procent av databladets Tops-värde.

Siffran visar nämligen bara produkten av antalet beräkningselement och klockfrekvensen. De avslöjar inte hur ofta kretsen kan leverera data. Om data alltid fanns omedelbart plats, energiförbrukning inte var ett problem, minneskapaciteten var oändlig och algoritmen passade hårdvaran som hand i handske – då skulle Tops vara mer meningsfullt. Men i den verkliga världen finns inga sådana ideala miljöer.

Och mätvärdet är ännu mer missvisande för ML-acceleratorer i mikrostyrkretsar. Små kretsar av det här slaget har typiskt värden mellan 1 och 3 TOPS – för det är ofta inte mer inferenskapacitet än så som behövs i ML-applikationer. Kretsarna är byggda kring Arm Cortexprocessorer som är speciellt utformade för strömsnål ML. Om man väger in mikrokontrollerns stöd för heltals- och flyttalsoperationer och många andra funktioner, är det uppenbart att Tops, eller godtyckligt annat ensamt mätvärde, är inte är tillräckligt för att bedömda prestanda vare sig för en ensam krets eller ett system. ■