Aducerea Inteligenței Artificiale la margine

Autor: **Cliff Ortmeyer**
Director Global Marketing
Tehnic şi Dezvoltare de Soluţii
Farnell AN AVNET COMPANY

În puţin mai mult de un deceniu, de când cercetătorii au descoperit noi tehnici de îmbunătăţire a eficienţei şi eficacităţii sale, deep learning (învăţarea profundă) a devenit o tehnologie practică, pe care se sprijină acum un număr de aplicaţii ce necesită inteligenţă artificială (IA).

Multe dintre aceste aplicaţii sunt găzduite în cloud pe servere puternice, după cum sarcinile implică uneori procesarea unor surse bogate în date, precum imagini, video şi audio. Aceste servere fac apel adesea la performanţele adiţionale ale unui hardware de accelerare, care pleacă de la unităţi de procesare grafică până la dispozitive particularizate. Acest lucru devine important pentru procesele intensive, în timpul cărora o reţea neurală este instruită cu date noi.

În mod tipic, procesul de deducţie, care utilizează o reţea de instruire, pentru evaluarea noilor date, utilizează mult mai puţine calcule intensive decât în cazul instruirii. Există, de asemenea, sarcini de lucru care implică surse ce necesită calcul mai puţin intens, precum citirea unor senzori de la dispozitive IoT, în care atât instruirea cât şi deducţia pot fi realizate pe hardware de performanţă mai joasă. Ca rezultat, proiectanţii de sisteme descoperă că aceste sarcini de lucru nu au nevoie neapărat să fie localizate în cloud odată ce modelul IA a fost instruit, deşi multe servicii curente o fac, din motive de afacere. În schimb, modelul instruit poate fi transferat pe o maşină locală pentru procesare mai aproape de sursa de date.

AVANTAJELE IA LA MARGINEA REŢELEI

Există un număr de motive pentru a aduce modelele IA mai aproape de marginea reţelei. Un factor major îl reprezintă intimitatea şi acceptarea de către utilizator. De exemplu, utilizatorii care folosesc dispozitive precum difuzoare inteligente, devin îngrijoraţi în legătură cu faptul de a avea conversaţiile lor private înregistrate şi încărcate în cloud în cazul în care aceste servicii nu pot fi suportate local.

În aplicaţiile de control industrial, care încep acum să utilizeze IA pentru monitorizarea condiţiilor echipamentului sau pentru optimizarea proceselor, există îngrijorări asupra confidenţialităţii datelor de producţie, care vor solicita în mod similar ca necesitatea de transfer în cloud să fie cât mai mică posibil. Pentru numeroase aplicaţii industriale, există probleme de siguranţă în funcţionare şi de viteză de conectare la cloud.

Numeroase sisteme, fie situate la nivel de hală, fie într-o locaţie cu control de la distanţă, nu dispun de conexiuni în bandă largă care sunt necesare pentru a suporta deciziile bazate pe cloud.

Sistemele de control sunt, de asemenea, afectate în mod divers de o latenţă mare a comunicaţiilor. Dacă modelele IA sunt utilizate în sisteme de control în buclă închisă, orice întârziere în preluarea unei actualizări din cloud va conduce la imprecizii şi la generarea de instabilitate. Unele sisteme pot utiliza o mixtură de cloud şi procesare locală.

Camerele de supraveghere de exemplu, vor păstra lăţimea de bandă a reţelei prin identificarea locală a ameninţărilor imediate, dar apoi apelează la cloud pentru a realiza procesări suplimentare pentru situaţii ce nu sunt acoperite de modelele locale.

Durata de viaţă a bateriei este, de asemenea, crescută semnificativ atunci când IA se află la margine, deoarece mai puţine date sunt trimise în reţea, reducând atât costurile cu reţeaua cât şi cele legate de cloud.

APLICAŢIA IA ÎN MENTENANŢA PREDICTIVĂ

Mentenanţa predictivă este la ora actuală cel mai cunoscut caz de utilizare pentru aplicaţii de industrie conectată. Ea dovedeşte o bună rentabilitate a investiţiei, datorită abilităţii de a reduce numărul de inspecţii la faţa locului. Durata de oprire a maşinii poate fi, de asemenea, redusă prin identificarea cu precizie a perioadei de timp de funcţionare rămase pentru o componentă, făcând posibilă exploatarea duratei de viaţă operaţionale, fără risc de defectări pe durata operării. Predicţia duratei de viaţă pe baza de date operaţionale ajută la programarea optimă a mentenanţei şi identificarea corectă a cerinţei de piese de schimb. Clienţii care utilizează mentenanţa predictivă văd deja un câştig în randament de 20 până la 25%.

Cu mentenanţă predictivă, senzorii unei maşini unelte pot detecta schimbări în temperatură cuplate cu o creştere a zgomotului sau vibraţiilor, determinate cu o combinaţie de microfoane şi accelerometre, ca indicativ al unei probleme potenţiale. Cu tehnicile algoritmice tradiţionale, pot fi extrem de dificil de corelat citirile senzorilor în timp real cu posibilele probleme. Modelele bazate pe IA pot interpreta serii de date în funcţie de timp împreună cu intrări de timp real pentru a determina cu precizie starea de sănătate a sistemului. Asemenea modele pot beneficia de avantajul unor tehnici precum fuziune de senzori pentru a determina când este necesară mentenanţă sau nu. Cerinţele de complexitate şi latenţă a modelului vor determina fezabilitatea rulării modelului local.

DE LA CLOUD LA GATEWAY SAU MARGINE: ABORDĂRI ÎN PROCESAREA LOCALĂ

Există două abordări în procesarea locală. Prima este de a utiliza puterea de procesare disponibilă pe dispozitiv. Fezabilitatea acestui lucru depinde de complexitatea modelului IA şi de puterea de procesare disponibilă. Pentru nodurile senzoriale de joasă putere, nu va fi posibilă rularea de modele de învăţare profundă, fie şi numai pentru deducţie, totuşi, de vreme ce algoritmii de învăţare automată pot lua multe forme, poate să nu fie necesară rularea complet local a unei sarcini atât de grele.

A doua opţiune este de a procesa prin trecerea datelor complet sau parţial pe alt dispozitiv. De exemplu, dispozitivul poate rula un model de IA simplificat care realizează o analiză iniţială a datelor. În cazul unui sistem de supraveghere care monitorizează activitatea, acest model poate realiza sarcina de determinare simplă dacă un semnal este zgomot de fond, precum vântul, sau se datorează unui intrus ce trece prin apropiere sau zgomotului brusc al spargerii unei ferestre. Funcţiile mai complexe pot fi transferate către un gateway care concentrează date de la multiple camere şi alte dispozitive de securitate. Dispozitivele pot oferi ceva preprocesare pentru a contribui la eficientizarea modelului de gateway.

Hardware-ul puternic precum dispozitivele cu logică programabilă Xilinx oferite de platforma Ultra96 de la Avnet pot realiza deducţii pe baza unor modele complexe de învăţare profundă şi a altor algoritmi complecşi de învăţare automată. Gateway-ul local poate chiar actualiza modelul

prin colectarea datelor şi realizarea de pachete de-date de instruire atunci când sarcina de lucru de deducţie este uşoară. Alternativ, gateway-ul poate colecta datele care sunt importante de reţinut şi să le trimită pe serverul cloud cu o frecvenţă zilnică, săptămânală sau lunară.

IA la nivelul dispozitivului oferă latenţa de comunicaţie cea mai mică. Cu toate acestea, vitezele de procesare mai mari ale gateway-ului local pot depăşi performanţele scăzute ale procesorului dispozitivului şi să ofere cei mai buni parametri de latenţă şi viteză de transfer de date. Transferul în cloud se va baza pe factori cum ar fi lăţimea de bandă a conexiunii Internet, precum şi distanţa până la servere: viteza finită a luminii impune o limită asupra a cât de joasă poate fi latenţa dacă un server cloud este utilizat pentru sarcini de deducţie.

OPŢIUNI DE TEHNOLOGII DE ÎNVĂŢARE

Dezvoltatorul are de ales ce tip de model să folosească pentru o aplicaţie specifică. Modelele timpurii de IA apelau la utilizarea unor tehnici precum structuri arborescente de decizie sau sisteme expert care solicitau un efort mare de la specialiştii din domeniu pentru a relaţiona combinaţii de intrări, la cauze şi rezultate probabile.

Învăţarea automată a îmbunătăţit arborele decizional prin realizarea de tehnici precum crearea unor "păduri" aleatorii. O astfel de abordare construieşte arbori decizionali multipli de la pregătirea de date şi aplicarea lor în paralel pentru calculul unei medii care reprezintă rezultatul cel mai probabil bazat pe datele de instruire. O "pădure" aleatoare este un exemplu de sistem de învăţare supervizat; el corelează datele furnizate de dezvoltatorii sistemului cu intrările de la care modelul de învăţare automată recunoaşte relaţiile.

Învăţarea profundă este un alt exemplu de tehnologie de învăţare supervizată, deoarece se bazează pe datele de instruire, care sunt etichetate în prealabil. De exemplu, într-un sistem de clasificare după imagini, modulul îşi preia datele din imagine şi le aplică unui strat de neuroni simulat. Ieşirea după primul strat de neuroni este trecută succesiv prin mai multe straturi. Unele dintre acestea combină ieşirile de la neuroni multipli din stratul precedent pentru a produce o singură valoare care este trecută către următorul strat. În acest mod, reţelele neurale profunde realizează o reducere a dimensiunii: o treaptă vitală atunci când se convertesc date complexe multi-dimensionale precum datele ca imagine sau radio, ce pot fi utilizate pentru a oferi o clasificare finală.

Reţeaua neurală este capabilă de a realiza clasificări deoarece ea poate învăţa modul prin care diferite aranjamente de coeficienţi din stratul neural răspund diferitelor imagini pentru a oferi ieşiri care se potrivesc cu etichetele de instruire.

Instruirea unui model de învăţare automată nu trebuie să se bazeze numai pe date etichetate. Algoritmii de învăţare automată nesupervizată, precum gruparea, pot găsi tipare în date fără ajutor adiţional. Un asemenea proces poate fi foarte folositor în sistemele de control industrial unde sunt utilizaţi senzori multipli sau atunci când este important comportamentul în timp al intrărilor. În cazul monitorizării condiţiilor pentru o maşină unealtă, magnitudinea vibraţiei poate să nu indice o problemă, ci poate fi pur şi simplu consecinţa procesului. Totuşi, un model (tipar) al modificării datelor seriale temporale asociate cu schimbarea rapidă în temperatură poate fi indicativul unei probleme care necesită mentenanţă. Asemenea diferenţe pot fi evidenţiate de date prin separarea în grupuri în care sunt uşor de diferenţiat atunci când datele sursă, dacă sunt utilizate direct, nu prezintă un model clar.

Un sistem de învăţare automată nesupervizată poate găsi modele în combinaţiile de citire a senzorilor care pot fi utilizate pentru a reduce cantitatea de date care este apoi trimisă către algoritmul supervizat, care este pregătit pe maşina ce operează sub diferite tipuri de stres şi diferite stări ale reparaţiei. După cum gruparea reduce cantitatea de date necesar a fi transferate, un model de proiectare util pentru implementare în sistemele IoT este de a realiza această activitate pe nodul dispozitivului. Modelul derivat din învăţarea supervizată poate rula pe gateway sau chiar în cloud, în funcţie de echilibrul între latenţă, performanţe şi cost.

ADUCEREA INTELIGENŢEI ARTIFICIALE LA MARGINE

Un factor important de luat în considerare cu învăţarea automată ca tehnologie, este acela că implementarea sa necesită adesea o cunoaştere în profunzime şi experienţă cu scopul de obţine cât mai mult din diferitele forme de IA disponibile. Octonion a dezvoltat o soluţie software la această problemă care este construită pe dispozitivul SmartEdge Agile de la Avnet. Soluţia permite inginerilor să dezvolte soluţii IA pentru sisteme IoT care implică procesare semnificativă la margine fără implicarea unui ajutor scump sau fără atingerea unei experienţe în profunzime.

Soluţia rulează pe trei nivele: dispozitiv; gateway; cloud. Hardware-ul dispozitivului, Avnet SmartEdge Agile, este un nod senzorial autonom, de joasă putere.

Acest dispozitiv oferă acces la o varietate de tipuri de senzori care includ: accelerometru, giroscop, magnetometru, pe lângă cei pentru măsurarea presiunii, umidităţii şi proximităţii. Pentru înregistrarea intrărilor audio este disponibil un microfon. Hardware-ul de senzor inteligent comunică cu un gateway local ce utilizează BLE. Versiunile care suportă alte tehnologii de reţea orientate IoT, precum LoRaWAN, SigFox, celulară şi WiFi sunt în dezvoltare. Modulul gateway poate lua forma unui dispozitiv inteligent Android sau iOS; sau alternativ, software-ul de gateway poate rula pe un Raspberry Pi sau pe platforma Linux a dispozitivului care oferă acces la cloud. Nivelul de cloud poate fi dezvoltat pe AWS, Microsoft Azure sau pe soluţii server particularizate.

Coordonarea diferitelor straturi este oferită prin mediul software Brainium al Octonion, care oferă suport margine-la-cloud într-un mediu de tip zero-cod. Securitatea este cheia tuturor implementărilor IoT şi un element important al sistemului cu scopul de a proteja datele private utilizate pentru pregătirea modelelor, precum şi pe durata deducţiei. Sistemul dezvoltat pentru Brainium utilizează criptare hardware AES împreună cu stocare de chei criptografice bazate pe hardware rezistent la intruziuni pentru a proteja datele imediat ce au fost citite de microcontrolerul de pe dispozitivul Avnet SmartEdge Agile. Imaginile firmware utilizate de dispozitiv pentru a realiza funcţii IA sunt, de asemenea criptate, cu o validare prin semnătură digitală. Atunci când datele sunt trimise, canalele de comunicaţie necesită criptare TLS pentru a proteja mesajele de la a fi spionate de adversar, de la margine până în cloud.

Mediul software Brainium de la Octonion oferă o combinaţie de învăţare automată supervizată şi nesupervizată. De exemplu, el poate fi pregătit pentru anomalii cu privire la datele senzoriale brute provenite în timp. Suplimentar, el va grupa datele de la diferite scenarii pentru a identifica tipare de date comune. Acestea pot fi introduse într-o structură flexibilă de modele cu ajutorul software-ului AI Studio, astfel încât dezvoltatorii să poată alege modelul necesar pentru un caz de utilizare ţintă.

Pe durata procesului de învăţare, dispozitivul colectează eşantioane de date, le asigură şi le criptează înainte de a le conecta la AI Studio software ce rulează în cloud. Acest software învaţă de la eşantioane pentru a genera modele IA, pe care le transmite înapoi la dispozitivul de la margine pentru sarcina de deducţie. Orice dispozitiv din hardware-ul implementat de client poate recepţiona acest model IA şi poate opera într-un mod independent pentru a realiza monitorizare şi analiză. Mediul care rămâne este consistent de la prototipare la producţie pe scară largă. Pentru implementare, clienţii pot utiliza hardware-ul SmartEdge Agile de la Avnet, chiar de pe raft sau pot alege proiectul pentru utilizare în implementări proprii.

Prin valorificarea datelor de calcul inteligente la margine, proiectul Avnet SmartEdge Agile, pus în funcţiune de software-ul Brainium, minimizează volumul datelor necesare pentru a fi transmise de la fiecare dispozitiv aflat la marginea reţelei. Acolo unde se trece în cloud, în sistem trebuie implementată securitate ridicată. Rezultatul este un sistem care oferă toate uneltele necesare pentru dezvoltarea rapidă şi implementarea sistemelor activate de IA.

Farnell
ro.farnell.com

Farnell element14

Cea mai extinsă gamă de produse
de până acum de la mărci de renume

Peste 500.000 de conectori disponibili
Expedierea produselor în aceeași zi*

ro.farnell.com/connector-eguide

*Expedierea în aceeași zi disponibilă pentru produsele în stoc