UBRZANJE ALGORITAMA VEŠTAČKE INTELIGENCIJE SA PRIMENOM U PROSTORNOM PLANIRANJU KORIŠĆENJEM HUAWEI ASCEND 310 ARHITEKTURE

Aleksandar Peulić[1], Dušica Jovanović[1], Sanja Stojković[1]

Apstrakt: Kada je reč o veštačkoj inteligenciji (VI) i prostornom planiranju, Huawei Ascend 310 procesor može biti korišćen za ubrzavanje različitih aspekata procesa prostornog planiranja. Korišćenjem VI algoritama, moguće je identifikovati obrasce, prepoznati objekte, izvršiti klasifikaciju terena i analizirati prostorne karakteristike.Veštačka inteligencija može se koristiti za predviđanje kretanja i modeliranje različitih elemenata u prostoru. Na primer, može se primeniti za predviđanje saobraćajnog toka, kretanja ljudi ili identifikaciju potencijalnih lokacija za izgradnju infrastrukture. Ascend 310 procesor pruža mogućnost ubrzane obrade ovih algoritama, čime se omogućava brže i efikasnije planiranje. U prostornom planiranju, VI može se koristiti za optimizaciju korišćenja resursa poput vode, energije ili zemljišta. Huawei Ascend 310 procesor može ubrzati analizu i optimizaciju resursa, što omogućava bolje upravljanje i racionalno korišćenje prostornih resursa. VI može biti korisna i za simulaciju scenarija i vizualizaciju planiranih prostornih intervencija. Korišćenjem Ascend 310 procesora za ubrzavanje obrade simulacija i vizualizacija, može se dobiti realističan prikaz predloženih planova, što pomaže donosiocima odluka da bolje razumeju i procene implikacije planiranih intervencija. Huawei Ascend 310 je specijalizovani procesor (ASIC) za ubrzanje veštačke inteligencije koji je razvijen od strane kompanije Huawei. ASCEND je arhitektura za ubrzavanje VI koja koristi nisku snagu i visoku propusnost.

Ključne reči: Veštačka inteligencija, Huawei arhitektura, semantička segmenatacija, detekcija objekata

ACCELERATION OF ARTIFICIAL INTELLIGENCE ALGORITHMS WITH APPLICATION IN SPATIAL PLANNING USING HUAWEI ASCEND 310 ARCHITECTURE

Abstract: When we talk about artificial intelligence (AI) and spatial planning, the Huawei Ascend 310 chip can be used to accelerate various aspects of the spatial planning process. By using AI algorithms, it is possible to identify patterns, recognize objects, perform terrain classification, and analyse spatial features, can predict the flow of traffic or the movement of people, or identify potential locations for infrastructure construction. The Ascend 310 chip provides the ability to accelerate the processing of these algorithms, enabling more efficient planning. In spatial planning, AI can be used to optimize the use of resources such as water, energy, or land. The Huawei Ascend 310 chip can accelerate the analysis and optimization of resources, enabling better management and rational use of spatial resources. AI can also be useful for simulating scenarios and visualizing planned spatial interventions. By

[1] Univerzitet u Beogradu – Geografski fakultet, Studentski trg 3/III, Beograd, aleksandar.peulic@gef.bg.ac.rs; dusica.jovanovic@gef.bg.ac.rs; sanja.stojkovic@gef.bg.ac.rs

using the Ascend 310 chip to accelerate the processing of simulations and visualizations, a realistic representation of proposed plans can be achieved, helping decision makers better understand and evaluate the impact of planned interventions. The Huawei Ascend 310 is a dedicated artificial intelligence accelerator chip (ASIC) developed by Huawei. ASCEND is an AI acceleration architecture that operates with low power consumption and high bandwidth.

Key words: Artificial intelligence, Huawei architecture, semantic segmentation, object detection

UVOD

Prema arhitekturi, procesore koji se koriste za procesiranje algoritama veštačke inteligencije(VI) možemo podeliti na četiri osnovna tipa, Centralna procesorska jedinica (CPU), Grafička procesorska jedinica (GPU), Integrisano kolo za specifične aplikacije (application specific integrated circuit ASIC) i Field programmable gate array (FPGA). CPU je integrisano kolo, koje predstavlja jezgro računarskog sistema i kontrolnu jedinicu računara. Njegova uloga je izvršavanje i obrada instrukcija. Grafička procesorska jedinica (GPU) je grafički procesor. To je mikroprocesor koji obrađuje slike na parsonalnim računarima, radnim stanicama, konzolama za igru i mobilnim uređajima, kao što su tablet računari i pametni telefoni. Integrisano kolo za specifičnu aplikaciju (ASIC) je integrisano kolo dizajnirano za određenu namenu. (FPGA) je dizajniran da se njegova struktura hardvera može fleksibilno konfigurisati i menjati u realnom vremenu na osnovu zahteva projektanata sistema. U odnosu na primenu, VI procesori mogu biti podeljeni na one koji se koriste za trening i one koji se koriste za izvršavanje i donošenje odluka. U fazi treninga, složeni model neuronske mreže treba da bude obučen kroz veliki broj unosa podataka ili metoda. Proces treninga zahteva veliku količinu podataka o obuci i složenu duboku strukturu neuronske mreže. Ovo je zahtevan process koji zahteva ultra-visoke performanse uključujući računarsku snagu, preciznost i skalabilnost procesora. Nvidia GPU klaster i Google TPU se obično koriste za ove namene. Izvršavanje i donošenje odluka se izvodi korišćenjem obučenih modela i novih podataka. Na primer, uređaj za video nadzor koristi model duboke neuronske mreže u pozadini da prepozna snimljeno lice. Iako je proces izvršavanja i donošenja odluka u pogledu procesne moći hardvera manje zahtevan od procesa treniranja, sastoji se od velikog broja matričnih operacija. GPU, FPGA i ASIC se koriste u procesu izvršavanja i donošenja odluka. Huawei Ascend VI procesore karakteriše jedinica za obradu neuronske mreže (Neural-network processing unit NPU) koja koristi skup instrukcija dubokog učenja za obradu velikog broja neurona i sinapsi simuliranih na nivou logičkog kola, na način da se jedna instrukcija koristi za obradu grupe neurona. Jedna od četiri glavne arhitekture Ascend VI procesora je VI računarski modul, koji se sastoji od VI jezgra (Da Vinči arhitektura) i VI CPU-a. DaVinči arhitektura je razvijena da poboljša računarsku snagu za obradu algoritama VI i služi kao jezgro Ascend VI računarskog modula i VI procesora. Glavne komponente DaVinčijeve arhitekture su računarska jedinica i sistem za skladištenje. Kontrolna jedinica obezbeđuje upravljanje instrukcijama za ceo računarski proces. Ascend 310 koristi napredni proces proizvodnje procesora za postizanje visoke integracije i performansi. Procesor je dizajniran da podrži veliki broj jezgara, sa velikom gustinom tranzistora i efikasnim korišćenjem prostora na procesoru. NPU (Neural Processing Unit) je glavni element arhitekture Ascend 310, dizajniran za brzu i efikasnu obradu VI zadataka. Ascend 310 koristi Huavei-ovu Da Vinči arhitekturu, koja pruža dodatne optimizacije za VI obradu. Da Vinči arhitektura uključuje posebne blokove za obradu podataka i optimizovane putanje podataka, kako bi se postigao visok nivo performansi i efikasnosti. Ascend 310 koristi paralelizaciju na različitim nivoima da bi postigao brzu i efikasnu obradu. Ovo uključuje paralelizaciju na nivou jezgra, kao i paralelizaciju na nivou čitavog procesora. Paralelizacija omogućava istovremeno izvršavanje više zadataka i optimizuje vremenske potrebe obrade VI. Ascend 310 podržava heterogenu obradu, što znači da može da koristi CPU i NPU paralelno da ubrza VI zadatke. Ovo omogućava bolje korišćenje resursa i povećava ukupnu efikasnost

obrade algoritama veštačke inteligencije. Ascend 310 ima integrisanu memoriju velikog kapaciteta koja je optimizovana za radna opterećenja algoritama veštačke inteligencije. Takođe podržava različite interfejse za povezivanje, kao što su PCIE (Peripheral Component Interconnect Express) i DDR (Double Data Rate) za brz prenos podataka. Ascend 310 je dizajniran sa naglaskom na nisku potrošnju energije. To ga čini pogodnim za ugradnju u različite uređaje sa ograničenom potrošnjom energije, kao što su pametni telefoni, IoT (Internet of Things) uređaji i druge platforme koje podržavaju veštačku inteligenciju, [1]. VI u prostornom planiranju, kao i u mnogim drugim sferama nauke i tehnike je mnogostruka, samim tim i potreba za efikasnim hardverom za obradu zahtevnih algoritama je velika. U ovom radu je prikazana efikasna primena Huawei Ascend 310 procesora u implementaciji semantičke segmentacije i You Only Look Once (YOLO) kod prepoznavanja i analize protoka i učesnika saobraćaja sa ciljem da se doprinese poboljšanju planiranja saobraćajnica.

MATERIJALI I METODE

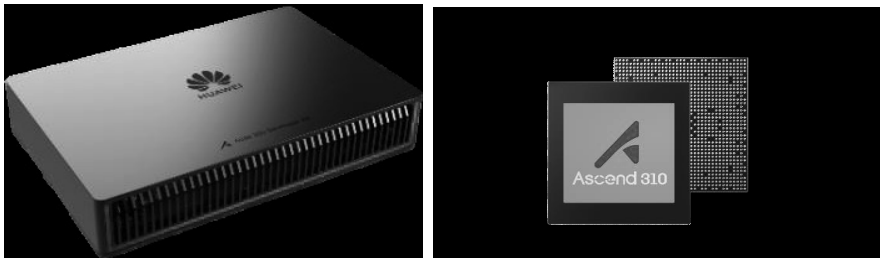
Semantička segmentacija je algoritam računarskog vida koji podrazumeva podelu slike na više regiona ili segmenata i dodeljivanje svakog piksela na slici određenoj klasi ili kategoriji. Cilj je da se svaki piksel označi odgovarajućim semantičkim značenjem, kao što je identifikacija objekata, regiona ili granica unutar slike. Za razliku od klasifikacije slika, gde je zadatak da se dodeli jedna oznaka celoj slici, semantička segmentacija obezbeđuje detaljnije razumevanje slike i dodeljivanjem oznaka svakom pikselu. Pristup koji ide do nivoa piksela omogućava detaljnije razumevanje scene, omogućavajući naprednije aplikacije kao što su detekcija objekata, segmentacija slike, autonomnu vožnju, analizu medicinske slike i još mnogo toga. Semantička segmentacija se obično izvršava korišćenjem tehnika dubokog učenja, posebno konvolucionih neuronskih mreža(CNN). CNN mogu efikasno naučiti prostorne hijerarhije i uhvatiti lokalni kontekst, čineći ih pogodnim za zadatke klasifikacije na nivou piksela kao što je semantička segmentacija. Popularne arhitekture za semantičku segmentaciju uključuju U-Net, Fully Convolutional Networks (FCN), DeepLab i Mask R-CNN. Proces treninga za semantičku segmentaciju uključuje obezbeđivanje obeleženih podataka za trening, gde je svaki piksel na slici ručno označen svojom odgovarajućom klasom. Mreža uči iz ovih označenih podataka tako što optimizuje svoje parametre koristeći tehnike kao što su propagacija unazad i stohastički gradijentni pad. Jednom treniran, model može da predvidi oznake klasa za nevidljive slike i generiše mape segmentacije na nivou piksela. Sve u svemu, semantička segmentacija je moćna tehnika u kompjuterskom vidu koja omogućava detaljno i precizno razumevanje sadržaja slike, otvarajući put za širok spektar primena u prostornom planiranju. Semantička segmentacija se može matematički objasniti kao problem klasifikacije na nivou piksela. Cilj je dodeliti semantičku oznaku svakom pikselu na slici. Matematički, možemo predstaviti sliku kao 2D ili 3D niz, gde svaki element predstavlja intenzitet piksela ili vrednosti boje. Izlaz segmentacije možemo predstaviti kao 2D niz, gde svaki element odgovara predviđenoj oznaci klase za odgovarajući piksel [2,3,4,5].

YOLO je popularan algoritam za detekciju objekata u realnom vremenu koji istovremeno vrši lokalizaciju i klasifikaciju objekata. Privukao je široku pažnju zbog svoje brzine i tačnosti. Dok je originalni YOLO algoritam evoluirao u nekoliko verzija (npr. YOLO v2, YOLO v3, YOLO v4, YOLO v6, YOLO v6, YOLO v7), svi oni dele sličnu matematičku pozadinu. Ulazna slika je podeljena na $\boldsymbol{S \times S}$ mrežu ćelija. Svaka ćelija je odgovorna za predviđanje objekata čiji centri spadaju u tu ćeliju. Svaka ćelija mreže predviđa $\boldsymbol{B}$ graničnih regiona, zajedno sa njihovim rezultatima pouzdanosti. Granični region je predstavljen koordinatama centra (x, y), širine (w) i visine (h). Ocena pouzdanosti ukazuje na verovatnoću da region sadrži objekat i tačnost predviđanja objeka. Svaki granični okvir ima pridruženu ocenu tačnosti, koja predstavlja verovatnoću da sadrži objekat. Ocena tačnosti se izračunava korišćenjem funkcije aktiviranja logističke regresije, koja ograničava vrednost između 0 i 1. YOLO vrši višeklasnu klasifikaciju za svaki granični okvir i predviđa verovatnoće $\boldsymbol{C}$

klase za svaku ćeliju, gde je ***C*** ukupan broj klasa u skupu podataka. Verovatnoće klasa se izračunavaju korišćenjem ***softmax*** aktivacione funkcije, obezbeđujući da su predviđene verovatnoće suma od 1 do 5. YOLO algoritam koristi specifičnu funkciju gubitka za obuku mreže. Nakon što mreža napravi predviđanja, primenjuje se korak naknadne obrade koji se zove Non-Maximum Suppression (NMS) da bi se filtrirali suvišni i preklapajući granični okviri. NMS eliminiše granične regione sa niskim rezultatom pouzdanosti zadržavajući samo najpouzdanije regione koji se ne preklapaju. Kombinovanjem ovih matematičkih komponenti, YOLO postiže efikasnu detekciju objekata tako što obrađuje celu sliku u jednom prolazu kroz mrežu. To eliminiše potrebu za koracima naknadne obrade. YOLO algoritam uspostavlja ravnotežu između brzine i tačnosti, što ga čini veoma pogodnim za aplikacije za detekciju objekata u realnom vremenu, [6,7,8,9,10].

REZULTATI I DISKUSIJA

U ovom radu implementirani su algoritmi semantičke segmentacije i YOLO na Huawei Atlas 200 razvojnom sistemu koji pokreće Acend 310 procesor (Slika 1).



Slika 1. Huawei Atlas 200 razvojni sistem i Ascend 310
(Official Textbooks for Huawei ICT Academy, Springer (eBook))

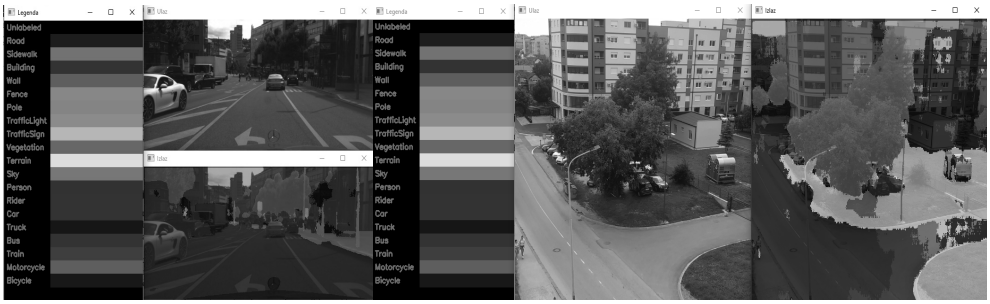
SEMANTIČKA SEGMENTCIJA

Da bi smo pokrenuli semantičku segmentaciju na Huawei Ascend 310 procesoru, potrebno je da se izvrše sledeći koraci:

- Obuka modela semantičke segmentacije na zasebnoj mašini koristeći alogoirtam dubokog učenja kao što je TensorFlow ili PyTorch. Ovo uključuje prikupljanje označenog skupa podataka, definisanje odgovarajuće mrežne arhitekture (npr. U-Net, DeepLab) i obuku modela da nauči klasifikacije objekata na nivou piksela.
- Kada se model obuči, treba ga optimizovati za zaključivanje na Huawei Ascend 310 procesoru. Ovo može uključivati tehnike kao što su kvantizacija modela, kompresija da bi se smanjila veličina modela i poboljšale performanse zaključivanja na procesoru.
- Konverzija obučenog i optimizovanog modela u format kompatibilan sa Huawei Ascend 310 procesorom. Ovo obično uključuje pretvaranje modela u ONNX (Open Neural Network Exchange) format, koji je široko podržan format za modele dubokog učenja.
- Za postavljanje konvertovanog modela na Huawei Ascend 310 procesor radi zaključivanja, koristiti se paket za razvoj softvera Huawei Ascend (SDK) i biblioteke za povezivanje sa procesorom i obavljanje zadatka semantičke segmentacije na ulaznim slikama ili video zapisom.

- Nakon obavljanja semantičke segmentacije na Ascend 310, moguče je primeniti tehnike naknadne obrade da bi se precizirali rezultati segmentacije ili izdvojile značajne informacije. Zatim se vrši vizualizacija rezultata segmentacije da bi se interpretirali i analizirali izlazi.

Za potrebe ovog rada, kao ulazni podaci koriste se slike raskrsnica u Beogradu i Kragujevcu. Koristi se gotov, već treniran model za prepozanvanje objekata na slici, a algoritam je realizvan u Python 3.7 instaliranom na Atlas 200 razvojnoj platformi. Na Slici 2 prikazan je rezultat semantičke segmentacije na Huawei Ascend 310, prikazana je legenda, ulazna slika i rezultat segmentacije. Algoritam se izvršava veomo efikasno, a kao rezultat su detektovani, segmentirani objekti na raskrsnici po klasama navedenim u legendi, koje je moguće kasnije analizirati u cilju unpređenja i planiranja toka saobraćaja.



Slika 2. Rezultat semantičke segmentacije na Huawei Ascend 310

YOLO

Huawei Ascend 310 procesor se takođe može koristiti za pokretanje YOLO algoritma za detekciju objekata. YOLO je poznat po svojim mogućnostima detekcije objekata u realnom vremenu, a računarska moć Ascend 310 procesora može poboljšati performanse YOLO zaključivanja, izvršavanjem sledečih koraka;

- Obuka YOLO modela na zasebnoj mašini koristeći okvir dubokog učenja kao što je TensorFlov ili PyTorch. Ovo uključuje prikupljanje označenog skupa podataka, definisanje YOLO arhitekture (npr. YOLO v3, YOLO v4) i obuku modela da detektuje i klasifikuje objekte na slikama ili video zapisima.
- Kada se YOLO model obuči, potrebno je da se optimizuje za zaključivanje na Huawei Ascend 310 procesoru. Ovo može uključivati tehnike poput kvantizacije modela ili kompresije da bi se smanjila veličina modela i poboljšale performanse zaključivanja na procesoru.
- Konverzija obučenog i optimizovanog YOLO modela u format kompatibilan sa Huawei Ascend 310 procesoru.
- Postavljanje konvertovanog YOLO modela na Huawei Ascend 310 procesoru za detekciju objekata u realnom vremenu. Treba napomenuti da optimizacija YOLO modela za performanse na Huawei Ascend 310 procesoru može zahtevati eksperimentisanje i fino podešavanje da bi se postigli najbolji rezultati. Pored toga, implementacija YOLO -a u realnom vremenu na Ascend 310 može uključiti analizu protoka podataka, iskorišćenja memorije i ukupne efikasnosti sistema kako bi se postigla brza detekcija objekata.

Za potrebe ovog rada koristi se YOLO coco model, algoritam je realizovan u programskom jeziku Python 3.7 na Huawei Ascend 310 procesoru. Na slici 3 prikazan je rezultat detekcije objekata primenom YOLO algoritma detekcije na Ascend 310.

Slika 3. Rezultat YOLO detekcije na Huawei Ascend 310

ZAKLJUČAK

Ascend 310 koristi posebnu arhitekturu koja je optimizovana za obradu VI zadataka. Ova arhitektura omogućava efikasno izvršavanje alogotirama veštačke inteligencije, kao što su matrične operacije i procesranje neuronske mreže. Ovaj processor koristi naprednu poluprovodničku tehnologiju za postizanje visoke efikasnosti, što znači da je napravljen sa niskom potrošnjom energije kao prioritetom, što ga čini energetski efikasnim u poređenju sa drugim procesorima koji se koriste za VI obradu. Efikasno upravljanje ovim resursima omogućava bolje performanse i smanjuje potrošnju energije, tako da je Ascenf 310 primenljiv za različite VI zadatke, uključujući prepoznavanje slika, koje svoju prmenu mogu naći u prostornom planiranju. Njegova efikasnost je posledica posebnih optimizacija hardvera i softvera koji su prilagođeni ovim specifičnim zadacima. Sve ove karakteristike zajedno čine Ascend 310 efikasnim VI procesorom za obradu, koji može da obezbedi brze performanse uz minimalnu potrošnju energije, tako da sa svojom softverskom podrškom koju obezbedjuje Huawei, predstavlja efikasan i vrlo upotrebljiv alat za primenu u različitim aspektima prostornog planiranja.

ZAHVALNICA

Ovaj rad je podržan od strane Huawei EBG Tim, Beograd, Srbija

LITERATURA

Huawei Technologies (2023). Artificial Intelligence Technology. Official Textbooks for Huawei ICT Academy, Springer (eBook). doi: 10.1007/978-981-19-2879-6.

Long, J., Shelhamer, E. & Darrel, T. (2014). Fully Convolutional Networks for Semantic Segmentation, Cornell University. Preuzeto sa https://arxiv.org/abs/1411.4038.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2016). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, Cornell University. Preuzeto sa https://arxiv.org/abs/1606.00915

He, K., Gkioxari, G., Dollar, P. & Girshick, R. (2017). Mask R-CNN, Cornell University. Preuzeto sa https://arxiv.org/abs/1703.06870

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F. & Hartwig, A. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, Cornell University. Preuzeto sa https://arxiv.org/abs/1802.02611

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection, Cornell University. Preuzeto sa https://arxiv.org/abs/1506.02640

Redmon, J. & Farhadi, A. (2015). YOLO9000: Better, Faster, Stronger, Cornell University. Preuzeto sa https://arxiv.org/abs/1612.08242

Redmon, J. & Farhadi, A. (2018). YOLOv3: An Incremental Improvement, Cornell University. Preuzeto sa https://arxiv.org/abs/1804.02767

Bochkovskiy, A., Wang, C. Y. & Hong-Yuan, M. L. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. Cornell University. Preuzeto sa https://arxiv.org/abs/2004.10934

Redmon, J. & Farhadi, A. (2016). YOLOv2: Improved Real-Time Object Detection. Cornell University. Preuzeto sa https://arxiv.org/abs/1612.08242