OPTIMIZACIJA MODELA DUBOKOG UČENJA PRIMJENOM KVANTIZACIJE U REAL-TIME SISTEMIMA AUTONOMNIH VOZILA / OPTIMIZATION OF DEEP LEARNING MODELS USING QUANTIZATION IN REAL-TIME SYSTEMS FOR AUTONOMOUS VEHICLES

Rudolf Petrušić[1] MA, doc.dr. Nešad Krnjić[2],
[1]Internacionalni univerzitet Travnik, Aleja konzula – Meljanac bb, 72270 Travnik, BiH,
[2]Sveučilište „Vitez“, ul. „Školska“ br. 7, 72270 Travnik, BiH,
e-mail: rudolf.petrusic@iu-travnik.com, prof.nesad@gmail.com

Pregledni članak
https://www.doi.org/10.58952/zr20261501371
UDK / UDC 004.8:629.33

Sažetak
Razvoj autonomnih vozila zavisi od sposobnosti ugrađenih sistema da u realnom vremenu obrađuju velike količine senzorskih podataka prikupljenih putem kamera, LiDAR-a i radara. Za zadatke poput detekcije objekata, prepoznavanja saobraćajnih znakova i identifikacije traka koriste se modeli dubokog učenja, čija visoka složenost predstavlja izazov za implementaciju na platformama sa ograničenim resursima. Kvantizacija modela predstavlja efikasan pristup optimizaciji, kojim se smanjuje numerička preciznost težina i aktivacija neuronskih mreža. Zamjenom 32-bitnih reprezentacija nižim preciznostima (16-bitnim ili 8-bitnim) smanjuju se memorijski zahtjevi i ubrzava izvođenje modela, što je ključno za real-time sisteme. Pored kvantizacije, primjenjuju se i tehnike poput pruning-a i knowledge distillation-a radi smanjenja kompleksnosti uz očuvanje performansi. Ove metode mogu se kombinovati kako bi se postigao optimalan balans između tačnosti, brzine i hardverskih ograničenja. Rad analizira ulogu kvantizovanih modela u autonomnim vozilima, njihove prednosti i ograničenja, s posebnim fokusom na implementaciju na GPU, TPU i AI akceleratorima te uticaj na performanse sistema.
Ključne riječi: *autonomna vozila, kvantizacija modela, optimizacija neuronskih mreža, real-time sistemi, duboko učenje*
JEL klasifiakcija: O33, C45, L62

Abstract
The development of autonomous vehicles depends on the ability of embedded systems to process large volumes of sensor data in real time, collected through cameras, LiDAR, and radar. Deep learning models are used for tasks such as object detection, traffic sign recognition, and lane detection, but their complexity poses challenges for deployment on resource-constrained platforms. Model quantization is an effective optimization technique that reduces the numerical precision of neural network weights and activations. By replacing 32-bit representations with lower precision formats (16-bit or 8-bit), memory usage is reduced and inference speed is improved, which is essential for real-time systems. In addition, techniques such as pruning and knowledge distillation are used to reduce model complexity while maintaining performance. These methods can be combined to achieve an optimal balance between accuracy, speed, and hardware constraints. This paper analyzes the role of quantized models in autonomous vehicles, highlighting their advantages and limitations, with a focus on implementation on GPUs, TPUs, and AI accelerators and their impact on system performance.
Keywords: *autonomous vehicles, model quantization, neural network optimization, real-time systems, deep learning*
JEL classification: O33, C45, L62

UVOD

Malo koja primjena vještačke inteligencije postavlja tako oštre i istovremene zahtjeve kao autonomna vožnja. Vozilo mora, dok se kreće, neprekidno odgovarati na pitanja koja čovjek rješava intuitivno: šta se nalazi ispred, koliko je daleko, kreće li se i kuda, smije li se proći. Iza svakog od tih odgovora stoji lanac modela dubokog učenja koji obrađuje tok podataka sa više senzora i mora to učiniti dovoljno brzo da odluka stigne na vrijeme.[112] Upravo ta dvostruka obaveza — biti i tačan i pravovremen — čini problem zanimljivim, ali i nezahvalnim. Posljednju deceniju obilježio je nagli rast kapaciteta percepcijskih modela. Duboke konvolucijske mreže, a u novije vrijeme i transformeri, donijele su skok u tačnosti, ali po cijeni desetina ili stotina miliona parametara i milijardi operacija po okviru.[113] U podatkovnom centru takvi modeli rade bez napora; u vozilu, gdje su računarska snaga, potrošnja energije i odvođenje toplote strogo ograničeni, oni se naprosto ne uklapaju u zadati budžet. Iz tog raskoraka između onoga što istraživački modeli mogu i onoga što ugrađeni hardver dopušta izrasla je čitava disciplina optimizacije neuronskih mreža. Među raspoloživim tehnikama kvantizacija zauzima posebno mjesto. Za razliku od pristupa koji mijenjaju arhitekturu, ona zadržava model takav kakav jeste, a mijenja samo način na koji se brojevi predstavljaju i računaju — i upravo zato se relativno lako uklapa u postojeće razvojne tokove.[114] Cilj ovog rada nije iscrpan katalog metoda, nego pregled koji kvantizaciju smješta u realan automobilski kontekst: pokazuje zašto je potrebna, koliko donosi, gdje su joj granice i kako se odnosi prema srodnim tehnikama. Nakon uvoda, poglavlje 2 opisuje percepcijski lanac autonomnog vozila; poglavlje 3 razmatra računarsku platformu i sigurnosne zahtjeve; poglavlje 4 detaljno obrađuje kvantizaciju; poglavlje 5 je posvećeno komplementarnim tehnikama; poglavlje 6 hardveru; poglavlje 7 uticaju na percepciju i odlučivanje; a poglavlja 8 i 9 donose diskusiju i zaključak.

1. PERCEPCIJSKI LANAC AUTONOMNOG VOZILA

1.1. SENZORI I FUZIJA

Nijedan senzor sam za sebe nije dovoljan. Kamera daje gustu informaciju o boji i teksturi, ali slabo procjenjuje udaljenost i osjetljiva je na osvjetljenje; LiDAR pruža precizan trodimenzionalni oblak tačaka, ali je rjeđi i skuplji; radar pouzdano mjeri brzinu i radi po lošem vremenu, ali je niske rezolucije. Zbog toga se u praksi pribjegava fuziji senzora, u kojoj se njihovi komplementarni nedostaci međusobno poništavaju.[115] Fuzija, međutim, znači i da kroz sistem teče više paralelnih tokova podataka, pa se računsko opterećenje sabira — što dodatno zaoštrava potrebu za efikasnim modelima.

1.2. ZADACI PERCEPCIJE

Percepcija nije jedan zadatak nego njihov splet. Detekcija objekata locira i klasifikuje druge učesnike u saobraćaju; semantička i instancna segmentacija dodjeljuju značenje svakom pikselu i razdvajaju pojedinačne objekte; detekcija voznih traka i prepoznavanje saobraćajnih znakova omogućavaju poštovanje propisa; procjena dubine i praćenje (engl. tracking) održavaju geometriju scene i identitet objekata kroz vrijeme, što je preduslov za predviđanje kretanja.[116]

---

[112]S. Grigorescu, B. Trasnea, T. Cocias i G. Macesanu, „A Survey of Deep Learning Techniques for Autonomous Driving“, Journal of Field Robotics, sv. 37, br. 3, 2020, str. 362–386.

[113]I. Goodfellow, Y. Bengio i A. Courville, „Deep Learning“, MIT Press, Cambridge, 2016, pogl. 1 i 9.

[114]A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney i K. Keutzer, „A Survey of Quantization Methods for Efficient Neural Network Inference“, arXiv:2103.13630, 2021.

[115]J. Janai, F. Güney, A. Behl i A. Geiger, „Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art“, Foundations and Trends in Computer Graphics and Vision, sv. 12, 2020.

[116]A. Geiger, P. Lenz i R. Urtasun, „Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite“, Proc. IEEE CVPR, 2012, str. 3354–3361.

Svi ovi zadaci izvršavaju se istovremeno i kontinuirano, najčešće nad nekoliko kamera i jednim ili više LiDAR-a, tako da ukupno opterećenje daleko premašuje ono što sugeriše posmatranje pojedinačnog modela.

1.3. ARHITEKTURE: OD KONVOLUCIJA DO TRANSFORMERA

Konvolucijske neuronske mreže dugo su bile okosnica percepcije zahvaljujući efikasnom korištenju prostorne lokalnosti. Rezidualne veze, uvedene u ResNet-u, omogućile su pouzdano treniranje vrlo dubokih mreža i ostale su sastavni dio gotovo svih kasnijih arhitektura.[117] Za detekciju u realnom vremenu prelomni su bili jednoprolazni detektori tipa YOLO, koji lokalizaciju i klasifikaciju objedinjuju u jednom prolazu i tako uklanjaju usko grlo višefaznih pristupa.[118] Za ugrađene uslove razvijene su namjenski štedljive arhitekture — MobileNet sa separabilnim konvolucijama po dubini i EfficientNet sa uravnoteženim skaliranjem dubine, širine i rezolucije.[119]

Posljednjih godina transformeri, izvorno osmišljeni za jezik, ušli su i u vid: Vision Transformer i detektor DETR pokazali su da se globalne zavisnosti u sceni mogu modelirati bez ručno projektovanih komponenti.[120] Cijena je, očekivano, veća — pažnja (engl. attention) skalira kvadratno s brojem ulaznih elemenata, pa transformerski modeli percepcije po pravilu traže više memorije i računa od uporedivih konvolucijskih mreža. Time se, pomalo paradoksalno, potreba za kompresijom samo pojačava upravo tamo gdje su modeli najtačniji.

2. RAČUNARSKA PLATFORMA U VOZILU I SIGURNOSNI ZAHTJEVI

Da bi se razumjela uloga optimizacije, treba poći od ograničenja koja vozilo nameće. Prvo je latencija. Percepcijski lanac mora isporučiti rezultat unutar budžeta jednog okvira, jer tek nakon percepcije slijede predviđanje, planiranje i upravljanje; svaka milisekunda potrošena na percepciju oduzeta je ostatku sistema.[121] Drugo je memorija — ne samo kapacitet za pohranu parametara, nego i propusni opseg za njihovo dohvaćanje, koji je na rubnim uređajima često stvarno usko grlo, izraženije od same računske snage.[122] Treće je energija i toplota: vozilo ne može neograničeno hladiti procesore niti trošiti snagu koja umanjuje domet, što je posebno osjetljivo kod električnih vozila.

Na te zahtjeve odgovaraju namjenski automobilski sistemi-na-čipu, koji na jednom silicijumu objedinjuju grafički procesor, dubinske akceleratore i procesor opće namjene. Platforme poput NVIDIA DRIVE Orin nude reda nekoliko stotina cjelobrojnih TOPS unutar energetskog okvira prihvatljivog za vozilo, dok rješenja kao Mobileye EyeQ idu putem krajnje energetske štedljivosti.[123]

[117]K. He, X. Zhang, S. Ren i J. Sun, „Deep Residual Learning for Image Recognition“, Proc. IEEE CVPR, 2016, str. 770–778.

[118]J. Redmon, S. Divvala, R. Girshick i A. Farhadi, „You Only Look Once: Unified, Real-Time Object Detection“, Proc. IEEE CVPR, 2016, str. 779–788.

[119]A. G. Howard i dr., „MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications“, arXiv:1704.04861, Google, 2017.

[120]N. Carion i dr., „End-to-End Object Detection with Transformers“, Proc. ECCV, 2020.

[121]Pojam „realno vrijeme“ (engl. real-time) ovdje ne znači „što brže“, nego obradu unutar tvrdo zadanog vremenskog budžeta po okviru; za percepciju u vožnji uobičajeni cilj je 30 i više okvira u sekundi, tj. latencija ispod ~33 ms po okviru.

[122]V. Sze, Y.-H. Chen, T.-J. Yang i J. S. Emer, „Efficient Processing of Deep Neural Networks“, Synthesis Lectures on Computer Architecture, Morgan & Claypool, 2020.

[123]NVIDIA Corporation, „NVIDIA TensorRT Developer Guide“ i „NVIDIA DRIVE Orin SoC“ tehnička dokumentacija, 2024.

Zajedničko im je da svoju nominalnu propusnost ostvaruju prvenstveno u sniženim preciznostima — što kvantizaciju iz poželjne pretvara u praktično neizbježnu.[124]
Najzad, automobilska percepcija je sigurnosno kritična i podliježe funkcionalnoj sigurnosti prema standardu ISO 26262. Funkcijama čiji otkaz može dovesti do teških posljedica dodjeljuje se najstroži nivo integriteta, ASIL D, što pred svaku izmjenu modela — pa i optimizaciju — postavlja zahtjev dokazive, a ne tek prosječne pouzdanosti.[125] Drugim riječima, ubrzanje koje donosi kvantizacija ima smisla samo ako se pri tome može jamčiti da tačnost ne pada ispod sigurnosnog praga, i to upravo u najtežim, a ne u prosječnim uslovima.[126]
Vrijedi se na trenutak zadržati na samom formatu brojeva. Modeli se treniraju u 32-bitnoj aritmetici s pokretnim zarezom (FP32), koja nudi širok dinamički raspon, ali je rasipna: svaka operacija množenja i akumulacije u FP32 troši osjetno više energije i silicijumske površine od ekvivalentne cjelobrojne operacije.[127] Ključni uvid, na kojem počiva cijela ideja kvantizacije, jeste da preciznost potrebna za stabilno treniranje daleko premašuje preciznost dovoljnu za pouzdano zaključivanje. Taj višak je prostor koji optimizacija nastoji iskoristiti.

3. KVANTIZACIJA MODELA DUBOKOG UČENJA

3.1. OSNOVNI PRINCIP I MATEMATIČKA FORMULACIJA

Kvantizacija preslikava vrijednosti iz visoko-preciznog skupa u manji, diskretni skup nivoa. U najčešćem, uniformnom afinom obliku, realna vrijednost r predstavlja se cijelim brojem q pomoću faktora skale s i nulte tačke z, prema $q = \text{round}(r / s) + z$, dok se rekonstrukcija (dekvantizacija) vrši kao $r \approx s \cdot (q - z)$.[128] Faktor skale određuje korak kvantizacije i izvodi se iz opsega vrijednosti tenzora, a nulta tačka dopušta asimetrično preslikavanje opsega koji nije centriran oko nule. Kod simetrične varijante nulta tačka je nula, čime se pojednostavljuje cjelobrojno množenje u hardveru — sitnica koja na nivou miliona operacija po okviru postaje važna.
Prelazak s FP32 na 8-bitne cijele brojeve (INT8) sam po sebi četvorostruko smanjuje veličinu modela i otvara put efikasnim cjelobrojnim jedinicama za množenje i akumulaciju, što na odgovarajućem hardveru donosi višestruko ubrzanje uz znatno manju potrošnju po operaciji.[129]
Glavna poteškoća nije u samoj ideji, nego u izboru opsega odsijecanja (engl. clipping range): preuzak opseg gubi izražajne ekstreme, a preširok rasipa dragocjene nivoe na rijetke vrijednosti i povećava grešku zaokruživanja. Pronalaženje te ravnoteže je, u praksi, najveći dio posla.[130]

3.2. GRANULARNOST I RASPODJELA NIVOA

Kvantizacija se može primijeniti s različitom granularnošću — po cijelom tenzoru, po kanalu ili po grupi. Iskustvo pokazuje da kvantizacija po kanalu, u kojoj svaki izlazni kanal konvolucije ima vlastiti faktor skale, gotovo redovno znatno smanjuje gubitak tačnosti, jer se prilagođava različitim raspodjelama težina pojedinih filtera, a trošak je zanemariv.

---

[124] TOPS (engl. Tera Operations Per Second) je mjera teorijske računske propusnosti akceleratora pri cjelobrojnim operacijama, najčešće INT8; stvarno ostvarena propusnost po pravilu je niža od deklarisane.

[125] ISO 26262:2018, „Road vehicles — Functional safety“, International Organization for Standardization, Ženeva, 2018.

[126] ASIL (engl. Automotive Safety Integrity Level) je klasifikacija rizika prema standardu ISO 26262; najstrožijem nivou ASIL D podliježu funkcije čiji otkaz može izazvati teške posljedice.

[127] V. Sze, Y.-H. Chen, T.-J. Yang i J. S. Emer, „Efficient Processing of Deep Neural Networks“, Synthesis Lectures on Computer Architecture, Morgan & Claypool, 2020; operacija množenja i akumulacije u pokretnom zarezu troši višestruko više energije i silicijumske površine od ekvivalentne cjelobrojne.

[128] B. Jacob i dr., „Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference“, Proc. IEEE CVPR, 2018, str. 2704–2713.

[129] H. Wu, P. Judd, X. Zhang, M. Isaev i P. Micikevicius, „Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation“, arXiv:2004.09602, NVIDIA, 2020.

[130] M. Nagel i dr., „A White Paper on Neural Network Quantization“, arXiv:2106.08295, Qualcomm AI Research, 2021.

Pored uniformne, postoji i neuniformna kvantizacija (logaritamska ili zasnovana na klasterovanju), koja nivoe zgušnjava tamo gdje su vrijednosti učestalije. Ona zna biti tačnija, ali se teže preslikava na cjelobrojnu aritmetiku, pa je u automobilskim cjevovodima rjeđa nego u istraživačkim radovima.

3.3. KVANTIZACIJA NAKON TRENIRANJA (PTQ) I KVANTIZACIJSKI SVJESNO TRENIRANJE (QAT)

Dva su osnovna pristupa, i izbor između njih je tipičan inženjerski kompromis. Kvantizacija nakon treniranja (PTQ) primjenjuje se na već istreniran model i traži tek mali reprezentativni skup za kalibraciju opsega aktivacija; brza je, ne zahtijeva ponovno treniranje i često je sasvim dovoljna za INT8.[131] Kada PTQ ne zadovolji — što se dešava kod osjetljivih arhitektura ili nižih preciznosti — pribjegava se kvantizacijski svjesnom treniranju (QAT), koje simulaciju kvantizacije ugrađuje u sam proces učenja, pa mreža uči težine otporne na grešku zaokruživanja. Gradijenti se kroz nediferencijabilnu operaciju zaokruživanja propagiraju tehnikom prolaznog procjenitelja (engl. straight-through estimator).[132] QAT po pravilu vraća tačnost na razinu blisku FP32 modelu i kod INT8, ali po cijeni dodatnog treniranja i složenijeg toka — trošak koji se za sigurnosno kritične funkcije gotovo uvijek isplati.[133]

3.4. STATIČKA I DINAMIČKA KVANTIZACIJA, MJEŠOVITA PRECIZNOST I EKSTREMNI REŽIMI

Prema obradi aktivacija razlikuju se statička kvantizacija, kod koje se opsezi unaprijed određuju kalibracijom, i dinamička, kod koje se računaju u trenutku izvođenja. Za real-time percepciju statička je gotovo uvijek poželjnija, jer izbjegava dodatni trošak tokom inference.[134] Mješovita preciznost ide korak dalje: osjetljive slojeve zadržava u višoj preciznosti (FP16), a robusne kvantizuje agresivnije, i u praksi nudi najbolji odnos tačnosti i brzine.[135] Na drugom kraju spektra su 4-bitne, ternarne i binarne mreže; binarni pristupi poput BinaryConnect i XNOR-Net zamjenjuju množenja logičkim operacijama i donose drastične uštede, ali uz pad tačnosti koji ih, barem zasad, čini teško prihvatljivim za sigurnosno kritičnu percepciju u vozilu.[136] Sumarno, INT8 — naročito u sprezi s granularnošću po kanalu i QAT-om — danas predstavlja praktičan standard za inferencu, jer pogađa povoljan odnos između očuvane tačnosti i dobiti u brzini i memoriji.[137]

---

[131]R. Krishnamoorthi, „Quantizing Deep Convolutional Networks for Efficient Inference: A Whitepaper“, arXiv:1806.08342, Google, 2018.

[132]Pravoprolazni procjenitelj (engl. straight-through estimator) propušta gradijent kroz nediferencijabilnu operaciju zaokruživanja kao da je riječ o identiteti, i tako omogućava treniranje uprkos stepenastoj prirodi kvantizacije.

[133]B. Jacob i dr., „Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference“, Proc. IEEE CVPR, 2018, str. 2704–2713; pokazano je da kvantizacijski svjesno treniranje vraća tačnost INT8 modela na razinu blisku FP32.

[134]M. Nagel i dr., „A White Paper on Neural Network Quantization“, arXiv:2106.08295, Qualcomm AI Research, 2021; statička kvantizacija unaprijed određuje opsege aktivacija kalibracijom i izbjegava trošak njihovog računanja u vrijeme izvođenja.

[135]A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney i K. Keutzer, „A Survey of Quantization Methods for Efficient Neural Network Inference“, arXiv:2103.13630, 2021; mješovita preciznost slojevima različite osjetljivosti dodjeljuje različite bit-širine.

[136]M. Courbariaux, Y. Bengio i J.-P. David, „BinaryConnect: Training Deep Neural Networks with Binary Weights During Propagations“, Proc. NeurIPS, 2015.

[137]H. Wu, P. Judd, X. Zhang, M. Isaev i P. Micikevicius, „Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation“, arXiv:2004.09602, NVIDIA, 2020; INT8 s kvantizacijom po kanalu i QAT-om preporučuje se kao podrazumijevani izbor za inferencu.

Preciznosti ispod 8 bita donose dalje uštede, ali zahtijevaju naprednije metode kompenzacije greške i, što je presudno, hardver koji ih izvorno podržava — bez toga ostaju akademska vježba.[138]

4. KOMPLEMENTARNE TEHNIKE OPTIMIZACIJE MODELA

Kvantizacija nije jedini, niti uvijek dovoljan, alat. Da bi se njeno mjesto ispravno ocijenilo, treba je posmatrati uz tehnike s kojima dijeli cilj — smanjenje modela uz očuvanje performansi.

4.1. PRUNING (REDUKCIJA TEŽINA)

Pruning uklanja parametre koji malo doprinose izlazu. Nestrukturirana varijanta postavlja pojedinačne težine na nulu i daje rijetke matrice koje teorijski smanjuju broj operacija, ali za stvarno ubrzanje traže poseban hardver ili formate. Strukturirana varijanta uklanja cijele filtere, kanale ili glave pažnje, pa se model smanjuje i ubrzava i na standardnom hardveru — zbog čega je u automobilskoj praksi često poželjnija.[139] U oba slučaja pruning se izvodi iterativno, uz dotreniravanje nakon svakog koraka kako bi se nadoknadio gubitak tačnosti.

4.2. DESTILACIJA ZNANJA

Destilacija znanja prenosi „znanje“ s velikog, tačnog modela (učitelja) na manji model (učenika): učenik uči ne samo tvrde oznake, nego i meke vjerovatnoće učitelja, koje nose bogatiju informaciju o sličnosti klasa.[140] Tako kompaktan model dostiže tačnost veću od one koju bi postigao učeći isključivo na izvornim oznakama, što ga čini pogodnim za ugrađenu percepciju. U dobro postavljenom cjevovodu destilacija i kvantizacija djeluju saglasno, ne suprotstavljeno.

4.3. FAKTORIZACIJA NISKOG RANGA I PRETRAGA ARHITEKTURA

Faktorizacija niskog ranga aproksimira velike matrice težina proizvodom manjih i tako smanjuje broj parametara i operacija. Komplementarno tome, automatska pretraga arhitektura (engl. Neural Architecture Search) i pažljivo ručno projektovane efikasne mreže već na nivou dizajna smanjuju složenost, pa se kvantizacija i pruning primjenjuju nad ionako kompaktnim modelima — što je, po pravilu, lakše i zahvalnije.[141]

4.4. KOMBINOVANJE TEHNIKA

Najvažnije je da ove metode nisu konkurentske. Pionirski rad o dubokoj kompresiji pokazao je da kombinacija pruninga, kvantizacije i entropijskog kodiranja može smanjiti veličinu modela za red veličine uz zanemariv gubitak tačnosti.[142]

U automobilskoj praksi ustalio se cjevovod čiji koraci slijede tu logiku: efikasna bazna arhitektura, zatim destilacija znanja, potom strukturirani pruning i, na kraju, kvantizacijski svjesno treniranje. Tek tako složene, tehnike daju zbir veći od svojih dijelova.

---

[138]M. Rastegari, V. Ordonez, J. Redmon i A. Farhadi, „XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks“, Proc. ECCV, 2016.

[139]S. Han, J. Pool, J. Tran i W. J. Dally, „Learning Both Weights and Connections for Efficient Neural Networks“, Proc. NeurIPS, 2015.

[140]G. Hinton, O. Vinyals i J. Dean, „Distilling the Knowledge in a Neural Network“, arXiv:1503.02531, NIPS Deep Learning Workshop, 2015.

[141]M. Tan i Q. Le, „EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks“, Proc. ICML, 2019, str. 6105–6114; uravnoteženim skaliranjem dubine, širine i rezolucije postiže se visoka tačnost uz bitno manje parametara i operacija.

[142]S. Han, H. Mao i W. J. Dally, „Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding“, Proc. ICLR, 2016.

5. HARDVERSKE PLATFORME I AKCELERATORI

Vrijedi naglasiti ono što se u teorijskim raspravama lako previdi: korist od kvantizacije postoji samo ako je hardver izvorno podržava. Bez cjelobrojnih jedinica i odgovarajućeg programskog lanca, INT8 model umije se izvršavati jednako sporo kao FP32 — ušteda ostane na papiru.[143]

5.1. GRAFIČKI PROCESORI I AUTOMOBILSKI SoC

Grafički procesori nude masovni paralelizam, a savremeni modeli sadrže namjenske tenzorske jedinice koje ubrzavaju matrične operacije pri sniženim preciznostima (FP16, INT8, pa i INT4). U vozilu su, međutim, ključne heterogene SoC platforme male potrošnje koje na jednom čipu spajaju GPU, dubinske akceleratore i procesor opće namjene; one omogućavaju inferencu reda desetina do nekoliko stotina cjelobrojnih TOPS unutar energetskog okvira primjerenog vozilu.[144] Razvojne ploče iz iste porodice (npr. NVIDIA Jetson) zato su postale uobičajen poligon za prototipiranje percepcijskih modela prije prelaska na automobilski razred.

5.2. TENZORSKE PROCESORSKE JEDINICE (TPU)

Tenzorske procesorske jedinice su namjenski akceleratori (ASIC) građeni oko sistoličkog niza za množenje matrica i optimizovani za cjelobrojnu aritmetiku niske preciznosti. Mjerenja pokazuju da pri inferenci postižu osjetno veću propusnost i energetsku efikasnost po operaciji od savremenih procesora i grafičkih kartica iste generacije.[145] Njihova arhitektura izravno koristi prednosti kvantizovanih modela, jer je cjelobrojno matrično množenje samo jezgro dizajna — pristup koji je danas duboko uticao i na automobilske akceleratore.

5.3. FPGA I NAMJENSKI AI AKCELERATORI

FPGA uređaji nude rekonfigurabilnost i mogućnost prilagođavanja širine aritmetike potrebama konkretnog modela, što je pogodno za istraživanje proizvoljnih bit-širina, ali traži specijalizovano znanje. Namjenski dubinski akceleratori (engl. Deep Learning Accelerator) integrisani u automobilske SoC-ove, naprotiv, nude fiksne, energetski vrlo efikasne cjevovode za uobičajene slojeve neuronskih mreža, i upravo se na njih oslanja serijska proizvodnja.[146]

5.4. PROGRAMSKI LANAC

Hardver je tek polovina priče; drugu polovinu nosi programski lanac. Okviri za optimizaciju i izvođenje, poput NVIDIA TensorRT-a, automatizuju kalibraciju za INT8, spajanje slojeva (engl. layer fusion), izbor optimalnih jezgara i raspoređivanje na akceleratore, i tek oni teorijsku prednost kvantizacije pretvaraju u stvarno ubrzanje na ciljnoj platformi.[147] Razmjenski formati poput ONNX-a pri tome olakšavaju prenos modela iz okruženja za treniranje u okruženje za izvođenje u vozilu, čime se skraćuje put od istraživanja do serije.

---

[143]B. Jacob i dr., „Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference“, Proc. IEEE CVPR, 2018, str. 2704–2713; dobit od INT8 ostvaruje se samo na hardveru s izvornim cjelobrojnim jedinicama za množenje i akumulaciju.

[144]V. Sze, Y.-H. Chen, T.-J. Yang i J. S. Emer, „Efficient Processing of Deep Neural Networks“, Synthesis Lectures on Computer Architecture, Morgan & Claypool, 2020; heterogene platforme koje spajaju GPU, namjenske akceleratore i procesor opće namjene daju najveću efikasnost po vatu.

[145]N. P. Jouppi i dr., „In-Datacenter Performance Analysis of a Tensor Processing Unit“, Proc. ISCA, 2017, str. 1–12.

[146]NVIDIA Corporation, „NVIDIA Deep Learning Accelerator (NVDLA)“ i „NVIDIA DRIVE“ tehnička dokumentacija, 2024; automobilski SoC-ovi integrišu fiksne, energetski efikasne cjevovode za uobičajene slojeve neuronskih mreža.

[147]NVIDIA Corporation, „NVIDIA TensorRT Developer Guide“, 2024; alat automatizuje INT8 kalibraciju, spajanje slojeva (engl. layer fusion) i izbor optimalnih jezgara za ciljni akcelerator.

5.5. UPOREDNI PREGLED TEHNIKA OPTIMIZACIJE

Tabela 1 sažima glavne tehnike kompresije prema mehanizmu, tipičnoj dobiti i ograničenjima, posmatrano iz ugla real-time percepcije u vozilu.

Tabela 1. Uporedni pregled tehnika optimizacije modela u kontekstu automobilske percepcije.

Tehnika	Mehanizam	Tipična dobit	Ograničenja
Kvantizacija (INT8)	Niža preciznost težina i aktivacija	≈4× manja memorija, znatno ubrzanje i ušteda energije	Mogući pad tačnosti; traži hardversku podršku
Pruning	Uklanjanje težina/kanala	Manje parametara i operacija	Nestrukturirani traži poseban hardver; nužno dotreniravanje
Destilacija znanja	Prenos znanja učitelj→učenik	Mali model uz veću tačnost	Traži kvalitetan model učitelja i dodatno treniranje
Efikasne arhitekture / NAS	Dizajn kompaktnih modela	Manja složenost od samog početka	Pretraga arhitekture računski skupa

6. UTICAJ NA PERCEPCIJU, ODLUČIVANJE I SIGURNOST

Optimizaciju modela u vozilu nije moguće odvojiti od sigurnosti, i tu se inženjerski pogled mora podrediti onome šta je dopušteno. Kvantizacija uvodi grešku koja se očituje kao blagi pad metrika — srednje prosječne preciznosti (mAP) kod detekcije ili srednjeg presjeka nad unijom (mIoU) kod segmentacije.[148] Uz INT8 i QAT taj pad je najčešće mali, nerijetko unutar jednog procentnog poena, ali prosjek ovdje vara: u sigurnosno kritičnom kontekstu presudni su rubni slučajevi — slabo osvijetljen pješak, udaljen ili djelimično zaklonjen objekt, kiša ili magla — gdje i mala degradacija može biti odlučujuća.[149]

Zato se kvantizovani modeli ne smiju ocjenjivati samo na prosječnim metrikama, nego se moraju validirati na reprezentativnim i namjerno teškim skupovima podataka, te u uslovnim podjelama (engl. scenario-based testing) koje izoluju upravo te rubne situacije.[150] S druge strane, dobitak nije samo brzina radi brzine: kraće vrijeme percepcije ostavlja više vremena modulima za predviđanje i planiranje i tako skraćuje ukupno vrijeme reakcije vozila, što je samo po sebi sigurnosna prednost.

Uspostavlja se, dakle, kompromis između tačnosti i pravovremenosti koji se ne može razriješiti na nivou jednog modela, nego na nivou cijelog sistema percepcije i odlučivanja, pa i njegove redundanse. Razuman inženjerski princip, koji slijedi i logiku standarda ISO 26262, jeste birati najvišu kvantizaciju koja u najgorem, a ne prosječnom slučaju zadržava tačnost iznad sigurnosnog praga — uz svjesno ostavljenu rezervu.[151]

[148]M. Cordts i dr., „The Cityscapes Dataset for Semantic Urban Scene Understanding“, Proc. IEEE CVPR, 2016.
[149]H. Caesar i dr., „nuScenes: A Multimodal Dataset for Autonomous Driving“, Proc. IEEE/CVF CVPR, 2020, str. 11621–11631.
[150]P. Sun i dr., „Scalability in Perception for Autonomous Driving: Waymo Open Dataset“, Proc. IEEE/CVF CVPR, 2020, str. 2446–2454; validacija na velikim i raznolikim skupovima nužna je da bi se otkrila degradacija u rubnim uslovima.
[151]ISO 26262:2018, „Road vehicles — Functional safety“, International Organization for Standardization, Ženeva, 2018; za funkcije nivoa ASIL D zahtijeva se dokaziva pouzdanost u najnepovoljnijim, a ne prosječnim uslovima.

7. DISKUSIJA: PREDNOSTI I OGRANIČENJA

Glavna prednost kvantizacije je povoljan odnos uloženog i dobijenog. U poređenju s pruningom i destilacijom, ona donosi velike i, što je važno, predvidljive uštede memorije i energije uz razmjerno jednostavnu primjenu — posebno u obliku PTQ-a, koji ne traži ponovno treniranje.[152] Za razliku od nestrukturiranog pruninga, INT8 kvantizacija izravno se preslikava na široko dostupne cjelobrojne jedinice, pa je ubrzanje stvarno, a ne tek teorijsko — što u serijskoj proizvodnji vrijedi više od svake elegancije na papiru.[153]

Ograničenja, naravno, postoje i ne treba ih umanjivati. Agresivna kvantizacija ispod 8 bita traži napredne metode i izvornu hardversku podršku, a osjetljivost slojeva i arhitektura na grešku nije ujednačena — transformerski modeli percepcije, primjera radi, znaju biti znatno osjetljiviji od konvolucijskih.[154] Stoga kvantizaciju treba shvatiti ne kao samostalno rješenje, nego kao dio strategije: u sprezi s efikasnim arhitekturama, destilacijom i strukturiranim pruningom, uz mješovitu preciznost za najosjetljivije dijelove mreže.[155] Konačan izbor uvijek je uslovljen ciljnom platformom, sigurnosnim zahtjevima i raspoloživim razvojnim vremenom, i tu nema univerzalnog recepta.

ZAKLJUČAK

Kvantizacija se s pravom izdvaja kao jedna od najpraktičnijih tehnika optimizacije modela dubokog učenja za real-time sisteme autonomnih vozila. Smanjenjem numeričke preciznosti težina i aktivacija postižu se opipljive uštede memorije i energije te ubrzanje izvođenja, što je presudno za ugrađene platforme s ograničenim resursima. INT8 kvantizacija, naročito uz granularnost po kanalu i kvantizacijski svjesno treniranje, danas je praktičan standard koji zadržava tačnost blisku izvornom FP32 modelu.[156]

Ipak, kvantizacija nije samodovoljna. Njen puni potencijal ostvaruje se u sprezi s komplementarnim metodama — pruningom, destilacijom i efikasnim arhitekturama — i uz hardver koji izvorno podržava cjelobrojnu aritmetiku niske preciznosti. Za sigurnosno kritičnu primjenu nezaobilazna je pažljiva validacija kvantizovanih modela u rubnim uslovima, jer prosjek nije mjerodavan tamo gdje su posljedice teške. Budući pravci uključuju robusniju kvantizaciju ispod 8 bita, kvantizaciju transformerskih modela percepcije i tješnju integraciju optimizacije modela s hardverskim ko-dizajnom — kombinaciju od koje se opravdano očekuje da dodatno suzi jaz između istraživačkih modela i njihove pouzdane primjene u vozilu.[157]

---

[152]R. Krishnamoorthi, „Quantizing Deep Convolutional Networks for Efficient Inference: A Whitepaper“, arXiv:1806.08342, Google, 2018; kvantizacija nakon treniranja (PTQ) ne zahtijeva ponovno treniranje i traži tek mali skup za kalibraciju.

[153]H. Wu, P. Judd, X. Zhang, M. Isaev i P. Micikevicius, „Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation“, arXiv:2004.09602, NVIDIA, 2020; za razliku od nestrukturiranog pruninga, INT8 se izravno preslikava na dostupne cjelobrojne jedinice, pa je ubrzanje mjerljivo.

[154]A. Gholami i dr., „A Survey of Quantization Methods for Efficient Neural Network Inference“, arXiv:2103.13630, 2021; transformerski modeli, zbog raspodjela aktivacija s izraženim ekstremima, po pravilu su osjetljiviji na kvantizaciju od konvolucijskih.

[155]S. Han, H. Mao i W. J. Dally, „Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding“, Proc. ICLR, 2016; najveće uštede postižu se kombinovanjem pruninga, kvantizacije i entropijskog kodiranja.

[156]R. Krishnamoorthi, „Quantizing Deep Convolutional Networks for Efficient Inference: A Whitepaper“, arXiv:1806.08342, Google, 2018; kvantizacija po kanalu uz QAT zadržava tačnost blisku FP32 modelu i pri INT8 preciznosti.

[157]M. Nagel i dr., „A White Paper on Neural Network Quantization“, arXiv:2106.08295, Qualcomm AI Research, 2021; kao pravci razvoja ističu se robusnija kvantizacija ispod 8 bita i tješnji ko-dizajn modela i hardvera.

LITERATURA

[1] Goodfellow, I., Bengio, Y. i Courville, A. (2016). Deep Learning. Cambridge, MA: MIT Press.
[2] Grigorescu, S., Trasnea, B., Cocias, T. i Macesanu, G. (2020). A Survey of Deep Learning Techniques for Autonomous Driving. Journal of Field Robotics, 37(3), 362–386.
[3] Janai, J., Güney, F., Behl, A. i Geiger, A. (2020). Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. Foundations and Trends in Computer Graphics and Vision, 12(1–3), 1–308.
[4] Caesar, H. i dr. (2020). nuScenes: A Multimodal Dataset for Autonomous Driving. Proc. IEEE/CVF CVPR, 11621–11631.
[5] Sun, P. i dr. (2020). Scalability in Perception for Autonomous Driving: Waymo Open Dataset. Proc. IEEE/CVF CVPR, 2446–2454.
[6] Redmon, J., Divvala, S., Girshick, R. i Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. Proc. IEEE CVPR, 779–788.
[7] Liu, W. i dr. (2016). SSD: Single Shot MultiBox Detector. Proc. ECCV, 21–37.
[8] Ren, S., He, K., Girshick, R. i Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Proc. NeurIPS.
[9] He, K., Zhang, X., Ren, S. i Sun, J. (2016). Deep Residual Learning for Image Recognition. Proc. IEEE CVPR, 770–778.
[10] Howard, A. G. i dr. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861.
[11] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. i Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proc. IEEE CVPR, 4510–4520.
[12] Tan, M. i Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proc. ICML, 6105–6114.
[13] Vaswani, A. i dr. (2017). Attention Is All You Need. Proc. NeurIPS.
[14] Dosovitskiy, A. i dr. (2021). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale (ViT). Proc. ICLR.
[15] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. i Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. Proc. ECCV.
[16] Jacob, B. i dr. (2018). Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. Proc. IEEE CVPR, 2704–2713.
[17] Krishnamoorthi, R. (2018). Quantizing Deep Convolutional Networks for Efficient Inference: A Whitepaper. arXiv:1806.08342.
[18] Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., van Baalen, M. i Blankevoort, T. (2021). A White Paper on Neural Network Quantization. arXiv:2106.08295.
[19] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W. i Keutzer, K. (2021). A Survey of Quantization Methods for Efficient Neural Network Inference. arXiv:2103.13630.
[20] Wu, H., Judd, P., Zhang, X., Isaev, M. i Micikevicius, P. (2020). Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation. arXiv:2004.09602.
[21] Courbariaux, M., Bengio, Y. i David, J.-P. (2015). BinaryConnect: Training Deep Neural Networks with Binary Weights During Propagations. Proc. NeurIPS.
[22] Rastegari, M., Ordonez, V., Redmon, J. i Farhadi, A. (2016). XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. Proc. ECCV.
[23] Han, S., Pool, J., Tran, J. i Dally, W. J. (2015). Learning Both Weights and Connections for Efficient Neural Networks. Proc. NeurIPS.
[24] Han, S., Mao, H. i Dally, W. J. (2016). Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. Proc. ICLR.
[25] Hinton, G., Vinyals, O. i Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

[26] Sze, V., Chen, Y.-H., Yang, T.-J. i Emer, J. S. (2020). Efficient Processing of Deep Neural Networks. Synthesis Lectures on Computer Architecture. Morgan & Claypool.
[27] Jouppi, N. P. i dr. (2017). In-Datacenter Performance Analysis of a Tensor Processing Unit. Proc. ISCA, 1–12.
[28] Cordts, M. i dr. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. Proc. IEEE CVPR.
[29] Geiger, A., Lenz, P. i Urtasun, R. (2012). Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. Proc. IEEE CVPR.
[30] ISO 26262:2018. Road vehicles — Functional safety. International Organization for Standardization, Ženeva.
[31] NVIDIA Corporation (2024). NVIDIA TensorRT Developer Guide & NVIDIA DRIVE Platform Documentation. Dostupno na: developer.nvidia.com.