Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Nik Prinčič

Vizualno sledenje na vgrajenih napravah

diplomsko delo

visokošolski strokovni študijski program
prve stopnje
računalništvo in informatika

Mentor: doc. dr. Luka Čehovin Zajc

Ljubljana, 2023

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Kandidat: Nik Prinčič
Naslov: Vizualno sledenje na vgrajenih napravah
Vrsta naloge: Diplomska naloga na visokošolskem programu prve stopnje Računalništvo in informatika
Mentor: doc. dr. Luka Čehovin Zajc

Opis:
Besedilo teme diplomskega dela študent prepiše iz študijskega informacijskega sistema, kamor ga je vnesel mentor. V nekaj stavkih bo opisal, kaj pričakuje od kandidatovega diplomskega dela. Kaj so cilji, kakšne metode naj uporabi, morda bo zapisal tudi ključno literaturo.

Title: Visual tracking on embedded devices

Description:
opis diplome v angleščini

Na tem mestu zapišite, komu se zahvaljujete za pomoč pri izdelavi diplomske naloge oziroma pri vašem študiju nasploh. Pazite, da ne boste koga pozabili. Utegnil vam bo zameriti. Temu se da izogniti tako, da celotno zahvalo izpustite.

Svoji dragi Alenčici.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
AI	Artificial inteligence	Umetna inteligenca
ANN	Artificial neural network	Umetna nevronska mreža
BNN	Biological neural network	Biološka nevronska mreža
DNN	Deep Neural Network	Globoka nevronska mreža
CNN	Convolutional neural network	Konvolucijska nevronska mreža
FPS	Frames per second	Sličice na sekundo
OAK	OpenCV AI Kit	OpenCV AI komplet
OAK	OpenCV AI Kit	OpenCV AI komplet
BBOX	Bounding box	Omejitveni okvir
IoT	Internet of things	Internet stvari
SOT	Single object tracking	Sledenje posameznega objekta
MOT	Multiple object tracking	Sledenje več objektom
VOT	Visual object tracking	Vizualno sledenje
VPU	Visual processing unit	Vizualna procesna enota
ROI	Region of interest	Območje interesa
USB	Universal serial bus	Univerzalno serijsko vodilo
PoE	Power over ethernet	Napajanje preko etherneta

Povzetek

Naslov: Vizualno sledenje na vgrajenih napravah

Avtor: Nik Prinčič

V okviru diplomskega dela je bilo implementirano in ovrednoteno delovanje vizualnega sledilnika na vgrajeni napravi Luxonis OAK-1. Izbran je bil sledilnik STARK, spada v družino sledilnikov, ki jih sestavljajo globoke nevronske mreže. Bolj specifično sledilnik uporablja arhitekturo transformer, ki je trenutno uporabljena v vseh najboljših vizualnih sledilnikih. Sledilnik je bilo potrebno rahlo predelati ter ga prevesti v OpenVINO format, ki omogoča uporabno na vgrajeni napravi. Poleg tega je bilo potrebno zasnovati cevovod, po katerem se podatki na napravi pretakajo. Z vsem naštetim smo dosegli, da lahko vgrajena naprava izvaja vse potrebne funkcije popolnoma avtonomno, vse kar potrebuje od gostiteljskega sistema (npr. osebni računalnik) je začetni omejitveni okvir tarče *(ang. bounding box)*, gostiteljskemu sistemu pa vrača vse naslednje omejitvene okvirje tarče. S tem smo dosegli to, da so performance sledenja neodvisne od gostiteljskega sistema.

Ključne besede: računalniški vid na vgrajenih napravah, DepthAI, vizualni sledilnik.

Abstract

Title: Visual tracking on embedded devices

Author: Nik Prinčič

This sample document presents an approach to typesetting your BSc thesis using LaTeX. A proper abstract should contain around 100 words which makes this one way too short.

Keywords: embedded computer vision, DepthAI, visual tracker.

Poglavje 1

Uvod

Računalniški vid je področje, ki se v zadnjih letih zelo hitro razvija. Z napredkov v razvoju avtonomnih sistemov, kot so avtonomni avtomobili, droni, roboti in še mnogi drugi, in vse večjem številu IoT naprav opremljenih s kamero, se vedno bolj pojavlja želja po uporabi modernih pristopov na majhnih, manj zmogljivih napravah. Področje računalniškega vida zavzema kar nekaj sklopov, v tem delu smo se osredotočili na vizualno sledenje, natančneje sledenju posameznega objekta *(ang. single object tracking, VOT)*.

Vizualno sledenje je področje, ki se ukvarja z iskanjem in sledenjem objektov v videu. V tem delu smo se osredotočili na SOT, kjer je cilj slediti samo enemu objektu. Sledilniku je naprej potrebno podati omejevalni okvir tarče *(ang. bounding box, BBOX)*, kateri želimo slediti, nato pa sledilnik na podlagi prostorske, pri najbolj modernih pa celo časovno-prostorske informacije sledi želeni tarči in nam za vsako naslednjo sličico vrne pripadajoč BBOX. V preteklosti so bil najbolj popularni tako imenovani klasični algoritmi (npr. KCF [6], MOSSE [2], ...), sedaj pa prevladujejo sledilniki, ki temeljijo na nevronskih mrežah. V zadnjih nekaj letih je upravičeno vedno bolj popularna arhitektura transformer [13], ki je bila primarno razvita in uporabljena za razumevanje in generacijo teksta *(ang. natural language processing, NLP)*, v zadnjih letih pa je bila adaptirana na vizualne sledilnike, kjer dosega odlične rezultate.

Da bi lahko zagotovili, dobre performance, pri manjši porabi energije, so se začele razvijati namenske procesorske enote VPU *(ang. visual processing unit)*, ki so optimizirane za izvajanje nevronskih mrež in pospešeno izvajanje operacij na slikovnimi tokovi. V to družino procesorskih enot spada tudi čip, Intel Movidius Myriad X, ki je v osrčju uporabljen naprave v tem diplomskem delu.

V okviru te diplomske naloge smo se osredotočili na sledilnik STARK [14], ter vgrajeno napravo Luxonis OAK-1. Cilj naloge je bil sledilnik prilagoditi uporabi na tej napravi, ga prevesti v potreben format, ter ga umestiti v cevovod. Cevovod je bilo potrebno tudi smiselno zasnovati, da v model pridejo pravilno oblikovani podatki.

Diplomsko delo je razdeljeno v 5 delov. V poglavju 2 bomo prestavili pregled področja. V poglavju 3 bomo opisali metodologijo, to zavzema hiter opis nevronskih mrež na splošno, arhitekture CNN in transformer ter predstavitev izbranega sledilnika. V poglavju 4 bomo opisali podrobnosti implementacije, kar vključuje opis uporabljene vgrajene naprave, prilagoditev modela ter njegovo pretvorbo v pravilen format, opis implementacije cevovoda in opis 4 različnih načinov delovanja, ki smo jih pripravili. V poglavju 5 bomo prestavili rezultate evalvacije, osredotočili se bomo na primerjavo med sledilnikom, ki je bil pognan samostojno na vgrajeni napravi proti tistemu, ki je pognan na osebnem računalniku. Predstavili bomo tudi rezultate primerjave med dvema načinoma delovanja, robni *ang. edge mode* proti gostiteljskemu *ang. host mode* načinu delovanja. V poglavju 6 je zaključek, ki povzema ugotovitve, ter predstavi možne izboljšave za nadaljnji razvoj.

Poglavje 2

Pregled področja

V tem poglavju bomo pregledali dosedanje delo na področju vizualnega sledenja na splošno in na vgrajenih napravah. Kot merilo uspešnosti sledilnika bom uporabili rezultate izziva VOT *(ang. VOT challenge)* [7], ki je eden najbolj uveljavljenih testov na področju vizualnega sledenja. VOT izziv vsako leto priredi evalvacijo novih sledilnikov. Sledilnike so v zadnji izvedbi, VOT2022 [8], testirali na v sedmih različnih kategorijah. Glede na rezultate izziva v zadnjih nekaj letih, lahko opazimo porast v popularnosti in uspešnosti sledilnikov, ki uporabljajo arhitekturo transformer. Iz rezultatov VOT2022, objavljenih v [8], lahko razberemo da 9 od najboljših 10 sledilnikov uporablja arhitekturo transformer, opazimo lahko tudi, da je kar 47% vseh testiranih sledilnikov uporabilo to arhitekturo.

Na področju vizualnega sledenja na vgrajenih napravah, je bilo objavljenih že nekaj del, a nobeno od njih ni uporabilo sledilnika na osnovi arhitekture transformer. V članku *Evaluation of Visual Tracking Algorithms for Embedded Devices*[9] so primerjali performance med 5 klasičnimi algoritmi, teste so izvedli na napravi Raspberry Pi 3 B v1.2 in ugotovili, da z uporabo sledilnika KCF dobijo najboljše razmerje med hitrostjo in natančnostjo. V članku *Real-Time Multiple Object Visual Tracking for Embedded GPU Systems*[4], so predstavili MOT *(ang. multiple object tracking)* na napravi Nvidia Jetson TX2, kjer so uporabili več stopenjsko arhitekturo, detektor (YOLOv3 [12])

in sledilnik (KCF [6]), in dobili dobre rezultate.

Poglavje 3

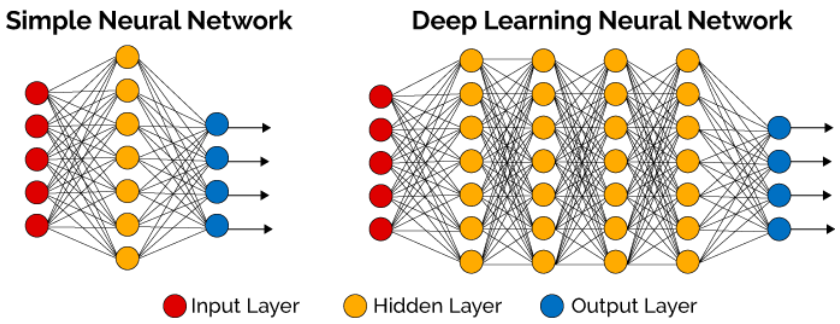
Metodologija

Ker je bil cilj dela, preizkusiti možnost uporabe enega izmed najsodobnejših vizualnih sledilnikov na vgrajeni napravi, ki temelji na nevronskih mrežah, bomo v tem poglavju predstavili delovanje nevronskih mrež, arhitekture CNN in transformer in opisali izbrani sledilnik.

3.1 Umetne nevronske mreže

Umetne nevronske mreže *(ang. Artificial Neural Network, ANN)* so vrsta modelov strojnega učenja, ki posnemajo delovanje bioloških nevronskih mrež *(ang. Biological Neural Network, BNN)*. Sestavljene so iz množice umetnih nevronov, ki so med seboj povezani z uteženimi povezavami in združeni v sloje. Sloje delimo na tri vrste - vhodni sloj *(ang. input layer)*, skrite sloje *(ang. hidden layers)* in izhodni sloj *(ang. output layer)*. Mreže z več kot enim skritim slojem imenujemo globoke nevronske mreže *(ang. Deep Neural Network, DNN)* ostale pa smatramo kot plitve nevronske mreže. Na sliki 3.1 je prikazana primerjava zgradbe med plitvo nevronsko mrežo (levo) in globoko nevronsko mrežo (desno).

Najpogosteje se za učenje nevronskih mrež uporablja algoritem vzvratnega razširjanja *(ang. backpropagation)*. Algoritem se izvaja v več iteracijah, pri katerih se s pomočjo kriterijske funkcije izračuna napaka, ki je



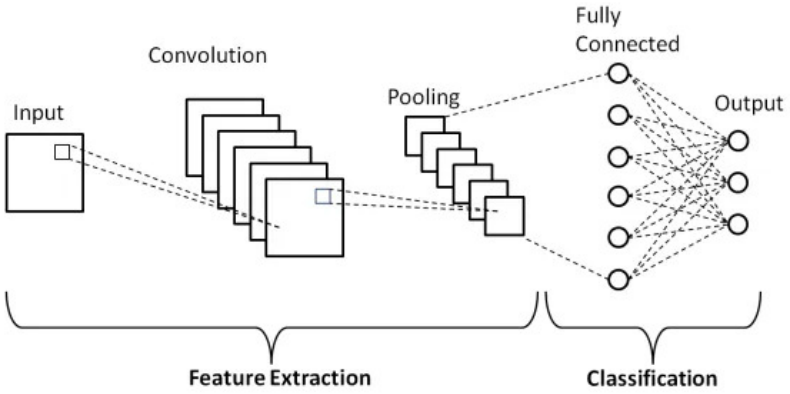
Slika 3.1: Primerjava zgradbe NN in DNN [11].

razlika med želenim izhodom in dejanskim izhodom. Napako uporabimo za izračun gradienta, ki nam pove, kako moramo spremeniti vrednost uteži, da se napaka zmanjša. Kriterijsko funkcijo morem pravilno izbrati glede na vrsto problema, ki ga želimo rešiti. Za reševanje regresijskih problemov se najpogosteje uporablja srednjo kvadratno napako *(ang. Mean Squared Error, MSE)*, za klasifikacijske probleme pa najpogosteje kategorično križno entropijo *(ang. Categorical Cross-Entropy)* ali pa binarno križno entropijo *(ang. Binary Cross-Entropy)*. Za iskanje optimalnih vrednosti uteži se uporabljajo različni optimizacijski algoritmi, ki na podlagi gradientnega spusta *(ang. gradient descent)* iščejo minimum kriterijske funkcije. Med njimi sta najbolj znan stohastični gradientni spust *(ang. Stochastic Gradient Descent, SGD)* in Adam *(ang. Adaptive Moment Estimation)*.

3.2 Konvolucijske nevronske mreže

Konvolucijske nevronske mreže *(ang. Convolutional Neural Network)*, so vrsta umetnih nevronski mrež, ki se pogosto uporablja v nalogah računalniškega vida, kot so prepoznavanje objektov, klasifikacija slik, detekcija obrazov in drugo. CNN modeli so tipično sestavljeni iz več ponovitev konvolucijskih slojev *(ang. convolutional layers)* in združevalnih slojev *(ang. pooling layers)*. Za zadnjim združevalnim slojem najpogosteje sledi nekaj polno-povezanih slojev. Poenostavljena arhitektura konvolucijske nevronske mreže je prika-

zana na sliki 3.2.



Slika 3.2: Poenostavljena arhitektura konvolucijkse nevronske mreže [1].

Namen konvolucijskih slojev je pridobivanje značilk *(ang. feature extraction)*, namen združevalnih slojev je zmanjševanje dimenzionalnosti podatkov, polno-povezani sloji pa so namenjeni preslikovanju pridobljenih značilk v končni izhod.

Konvolucijski sloji delujejo na principu matematične operacije konvolucije, ki je definirana na dveh funkcijah. Rezultat konvolucije nam pove, kako oblika ene spremeni obliko druge. Definirana je z enačbo:

$$(f * g)(t) = \int_{\tau=-\inf}^{\inf} f(\tau)g(t-\tau)d\tau \tag{3.1}$$

Kjer je f predstavlja vhodno funkcijo, g pa jedro konvolucije. Ker se pri obdelavi digitalnih podatkov pogosto uporablja diskretna konvolucija, z enačbo 3.1 spremenjena v:

$$(f * g)[t] = \sum_{k=-\inf}^{\inf} f[k]g[n-k] \tag{3.2}$$

3.3 Arhitektura transformer

Transformer je arhitektura nevronske mreže, ki temelji na mehanizmu pozornosti, ki je bil predstavljen leta 2017 v članku *Attention is all you need*

[13]. Primarno je bila arhitektur razvita za naloge procesiranja naravnega jezika *(Natural language processing, NLP)*, sejda pa postaja popularna tudi v drugih domenah strojnega učenja, med drugim tudi v računalniškem vidu. Osnovna arhitektura, kji je bila predstavljena v [13], vključuje kodirni *(ang. encoder)* in dekodirni modul *(ang. decoder)*. Kodirnik je sestavljen iz dveh podslojev, dekodirnik pa iz treh.

3.3.1 Mehanizem pozornosti

Mehanizem pozornosti deluje na podlagi ključev *(ang. key, K)*, poizvedb *(ang. query, Q)* in vrednosti *(ang. value, V)*. Mehanizem iz matrike ključev in matrike poizvedb izračuna matriko pozornosti. Z matričnim množenjem matrike pozornosti in matrike vrednosti dobimo linearno kombinacijo vrednosti, ki predstavljajo izhod. Razlikujemo med samo-pozornostjo in med-pozornostjo. O samo-pozornosti govorim, ko vse tri vhodne parametre dobimo iz iste množice podatkov, pri med-pozornosti pa poizvedbe pridobimo iz ene množice podatkov, ključe in vrednosti pa iz druge.

Vhodne vektorje $x_i, x_{i+1}, ..., x_n$ združimo v matriko $X_{n \times d_m}$, kjer d_m predstavlja dimenzionalnost modela. Naučene parametre pa predstavljajo matrike $W^Q_{d_m \times d_m}$, $W^K_{d_m \times d_m}$ in $W^V_{n \times d_m}$. Iz navedenih matrik lahko izračunamo

$$\begin{aligned} Q &= XW^Q, \\ K &= XW^K, \\ V &= XW^V, \end{aligned} \tag{3.3}$$

Matrika pozornosti je definirana z enačbo:

$$A(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_m}})V \tag{3.4}$$

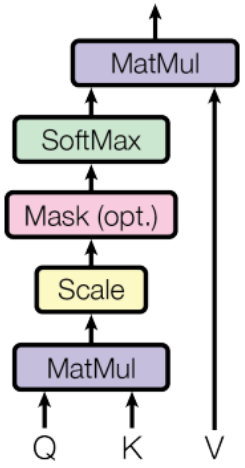
kjer funkcija *softmax* spremeni vektor n realnih števil v vektor verjetnostne porazdelitve.

Iz 3.3 in 3.4 lahko izračunamo končno izhodno vrednost mehanizma

$$S = AV, \tag{3.5}$$

na sliki 3.3 je potek izračuna prikazan v obliki diagrama.

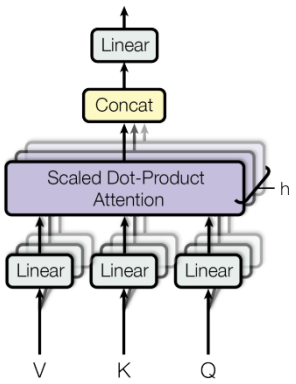
Ko se mehanizem pozornosti uporabi v dekodirnem modulu je potrebno določen del podatkov skriti. Bolj natančno, modelu je treba preprečiti, z podatki, ki jih še ni napovedal.



Slika 3.3: Diagram izračuna pozornosti [13].

3.3.2 Več-glava pozornost

Da bi model lahko zajel več lastnosti vhodnih podatkov je potrebno mehanizem pozornosti nadgraditi. Več-glava pozornost vzporedno izračuna več ločenih pozornosti (glav) in jih združi v končni rezultat. Vsaka glava se osredotoči na eno lastnost podatkov. Vsaka glava i ima svoje matrike naučenih uteži, končni rezultat pa se pridobi s strnitvijo posameznih matrik pozornosti. Na sliki 3.4 je prikazan diagram poteka več-glave pozornosti.



Slika 3.4: Diagram izračuna več-glave pozornosti [13].

3.3.3 Vhodna vdelava in pozicijsko kodiranje

Preden vhodni podatki prispejo do kodirnega bloka jih je potrebno najprej pravilno predelati. Za ta namen se uporablja vdelava vhodnega niza v d_m dimenzionalni latentni prostor *(ang. embedding space)*. Vhodna vdelava *(ang. input embedding)* vhodne besede pretvori v vektorje, s tem model pridobi informacijo o pomenu posamezne besede. Pozicijsko kodiranje *(ang. positional encoding)* pa modelu poda informacijo o vrstnem redu besed v nizu. Da bi se izognili velikim vrednostim v pozicijskem kodiranju, se za kodiranje uporabljata sledeči funkciji 3.6

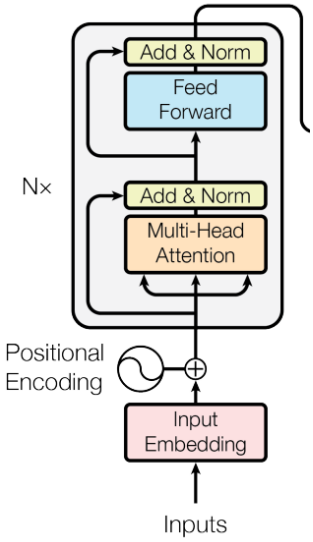
$$\begin{aligned} PE(pos, 2i) &= sin(pos/10000^{2i/d_m}) \\ PE(pos, 2i+1) &= cos(pos/10000^{2i/d_m}) \end{aligned} \tag{3.6}$$

v enačbi 3.6 predstavlja *pos* pozicijo besede v nizu, i pa dimenzijo kodiranja, kar pomeni da ima vsaka dimenzija pozicijskega kodiranja pripadajočo sinusoidno vrednost.

Informacijo o pomenu in poziciji posamezne besede združimo tako, da matriki seštejemo.

3.3.4 Kodirni modul

Kodirni modul ali kodirnik je sestavljen iz N identičnih slojev, ki so sestavljeni iz dveh pod-slojev. Prvi pod sloj je več glava pozornost *(ang. multi-headed attention)* in polno povezane usmerjene nevronske mreže *(ang. Feedforward neural network, FFN)*.



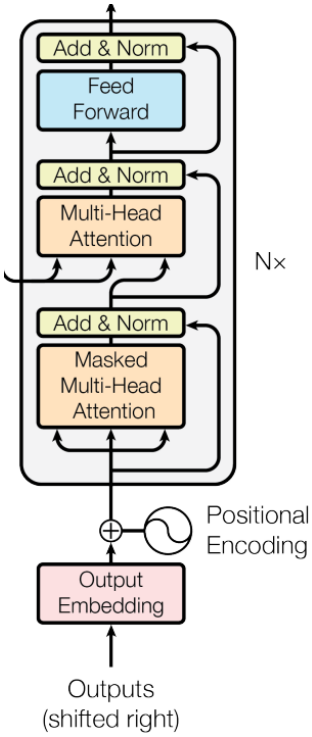
Slika 3.5: Diagram kodrinega modula [13].

3.3.5 Dekodirni modul

Dekodirni modul ali dekodirnik je sestavljen iz N identičnih slojev, ki so sestavljeni iz treh pod-slojev. Prvi pod sloj je več glava pozornost z možnostjo maskiranja vhodnih podatkov. Drugi pod sloj je več glava pozornost, tretji sloj pa predstavlja polno povezana nevronska mreža.

3.4 Model STARK

Model STARK [14] spada v družino SOT *(ang. single object tracking)* sledilnikov. Primarno je sestavljen iz dveh arhitektur, konvolucijskih nevronskih



Slika 3.6: Diagram dekodirnega modula [13].

mrež in arhitekture transformer, ki je bila prilagojena za vizualne sledilnike. Navdih za njegov nastanek je bil predhodni model za detekcijo DETER [3]. Ena od novosti, ki so jo uvedli v tem modelu je uporaba časovne in prostorske komponente *STARK_ST*. Prostorska komponenta vsebuje informacijo o izgledu objekta, kateremu sledi. Časovna komponenta pa nosi informacijo o spremembi pozicije objekta skozi čas. Arhitektura, ki so jo predlagali vsebuje tri ključne elemente: kodirni modul, dekodirni modul in napovedovalno glavo *(ang. predictio head)*. Model kot vhod prejme trenutno sliko, začetno matrico *(ang. template)* in dinamično matrico, ki se skozi čas dinamično posodablja. Z uporabo dinamična matrice, ki se skozi čas posodablja, lahko model zajame prostorsko in časovno informacijo o objektu, ki mu sledimo.

Prednost tega sledilnika je v tem, da ne potrebuje kompleksne predobdelave *(ang. preprocessing)* vhodnih podatkov in naknadne obdelave *(ang.*

postprocessing) izhoda. Ob inicializaciji prejem kot vhod sliko in omejitveni okvir tarče. Na podlagi teh dveh vhodnih podatkov se najprej iz celotne vhodne slike izreže iskalno območje *(ang. search area)*, ki se uporabi za izračun matrice. Ob vsaki naslednji sličici, pa so vhodni podatki iskalno območje izrezano iz vhodne slike, začetna matrica in dinamična matrica, sledilnik pa vrne izračunani omejitveni okvir.

3.4.1 Arhitektura modela stark

V tem delu smo zaradi manjše procesorske in prostorske zahtevnosti uporabili, različico modela STARK, ki uporablja samo prostorsko komponento *STARK_S*. V nadaljevanju bomo predstavili delovanje in arhitekturo modela.

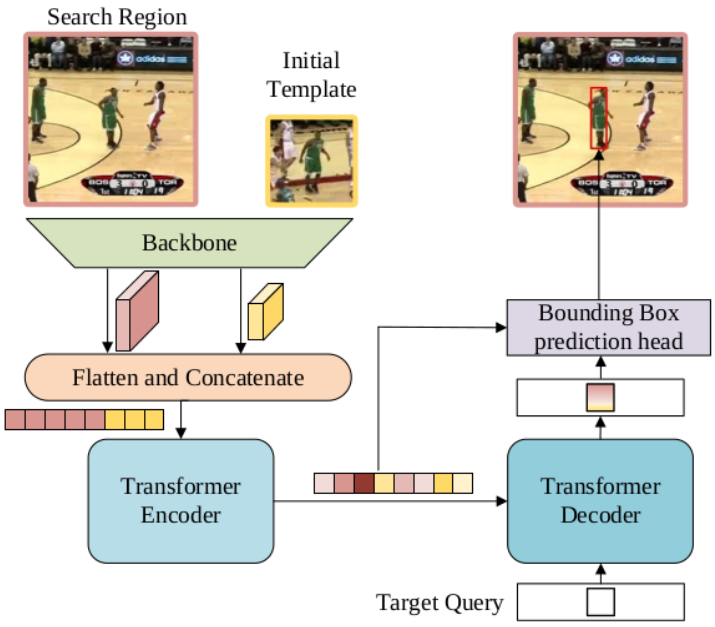
Arhitekturo modela lahko razdelimo na tri glavne dele: *(1)* konvolucijska hrbtenica *(ang. convolutional backbone)*, *(2)* kodirni-dekodirni transformer in predikcijsko glavo za omejitveni okvir *(ang. bounding box prediction head)*. V nadaljevanju bomo predstavili vsakega od teh delov.

Konvolucijska hrbtenica

Konvolucijsko hrbtenico sestavlja model ResNet [5], kateremu so odstranili zadnjo sekcijo in polno povezane sloje. Kot vhod prejme hrbtenica začetno matrico $z \in \mathbb{R}^{3\times H_z\times W_z}$ in iskalno območje $x \in \mathbb{R}^{3\times H_x\times W_x}$. Po prehodu čez hrbtenico dobimo dve matrici značilk $f_z \in \mathbb{R}^{C\times \frac{H_z}{s}\times \frac{W_z}{s}}$ in $f_x \in \mathbb{R}^{C\times \frac{H_x}{s}\times \frac{W_x}{s}}$.

3.4.2 Kodirnik

Na matriki značilk, ki jo izračuna hrbtenica, je najprej potrebno izvesti predobdelavo. Predobdelava vključuje sloj z ozkim grlom *(ang. bottleneck layer)*, ki matrikam zmanjša prostorsko dimenzionalnost iz C na d. Nato je matriki potrebno sploščiti in združiti vzdolž prostorske dimenzije. Rezultat prejšnjih dveh operaciji proizvede sekvenco značilk dolžine $\frac{H_z}{s}\frac{W_z}{s} + \frac{H_x}{s}\frac{W_x}{s}$ in dimenzionalnosti d. Novo pridobljeni sekvenci značilk prištejemo vrednosti pozicijskega kodiranja vhoda. Ta sekvenco značilk je vhod za kodirni mo-



Slika 3.7: Diagram arhitekture modela STARK_S [13].

dul. Kodirni modul je sestavljen iz N identičnih slojev. Sloji so sestavljeni iz več-glave samo-pozornosti in FFN. Kodirni modul zajame odvisnosti med vsemi značilkami v vhodni sekvenci in s tem omogoča modelu, da se nauči o diskirminativnih značilkah, katere se uporabijo za lokalizacijo objekta.

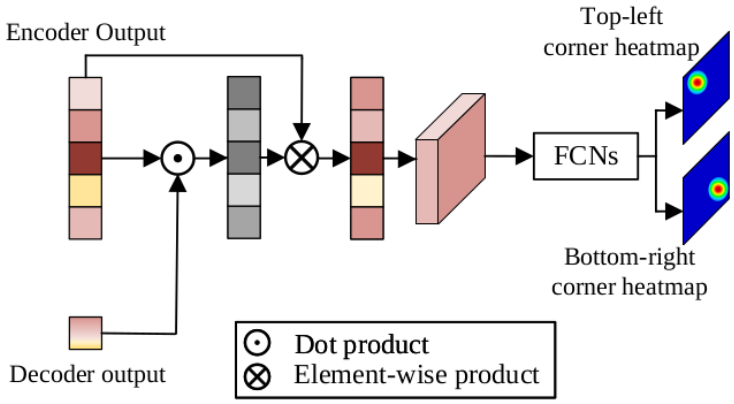
3.4.3 Dekodirnik

Dekodirnik kot vhod prejem sekvenco značilk, ki jo je izračunal kodirnik, in eno poizvedbo. Podobno kot kodirnik je tudi dekodirnik sestavljen iz M identičnih slojev. Vsak sloj je sestavljen iz samo-pozornosti, med-pozornosti kodirnik-dekodirnik in FFN. V sloju med-pozornosti lahko ciljna poizvedba deluje nad vsemi pozicijam na vhodni matrici in iskalnem območju, kar omogoča učenje robustnih reprezentacij predikcij končnega omejevalnega okvirja.

3.4.4 Glava za predikcijo omejevalnega okvirja

Model uporablja novo zasnovani sistem za predikcijo omejevalnega okvirja, ki deluje na predikcij verjetnostne distribucije robov okvirja. Glava kot vhod prejme sekvenco značilk iskalnega območja, ki jih je izračunal kodirni modul in izhodno vdelavo *(ang. output embedding)*, ki jo je izračunal dekodirnik in med njimi izračuna podobnosti. Podobnosti pomnoži z sekvenco značilk iskalnega območja, s tem se poveča pomembnost pomembnih regij. Rezultat te operacije je nova sekvenca značilk, ki jo je potrebno preoblikovati v matriko $f \in \mathbb{R}^{d \times \frac{H_s}{s} \times \frac{W_s}{s}}$. Novo pridobljena matrika je posredovana v polno povezano konvolucijsko mrežo *(ang. fully connected convolutional network)*, ki iz matrike značilk izračuna dve verjetnostni matriki, $P_{tl}(x, y)$ in $P_{br}(x, y)$, ki predstavljata levi zgornji in desni spodnji kot omejevalnega okvirja. Končne koordinate omejevalnega okvirja $(\hat{x_{tl}}, \hat{y_{tl}})$ in $(\hat{x_{br}}, \hat{y_{br}})$. Na sliki 3.8 je z diagramom predstavljen opisan potek predikcije omejevalnega okvirja, zadnji korak računanja kordinat iz verjetnostnih matrik pa je opisan v formuli 3.7.

$$
\begin{aligned}
(\hat{x_{tl}}, \hat{y_{tl}}) &= (\sum_{y=0}^{H} \sum_{x=0}^{W} x \cdot P_{tl}(x, y), (\sum_{y=0}^{H}) \sum_{x=0}^{W} y \cdot P_{tl}(x, y)), \\
(\hat{x_{br}}, \hat{y_{br}}) &= (\sum_{y=0}^{H} \sum_{x=0}^{W} x \cdot P_{br}(x, y), (\sum_{y=0}^{H}) \sum_{x=0}^{W} y \cdot P_{br}(x, y)).
\end{aligned}
\tag{3.7}
$$



Slika 3.8: Diagram poteka izračuna omejitvenega okvirja [13].

Poglavje 4

Implementacija

V tem poglavju bomo najprej predstavili vgrajeno napravo Luxonis OAK-1, ogrodji OpenVINO in DepthAI, na kratko predstavili vsa ostala uporabljena orodja in tehnologije ter na koncu podrobno opisali postopek implementacije.

4.1 Luxonis OAK-1

Luxonis je Ameriško podjetje, ki se ukvarja z razvojem naprav za uporabo na področju prostorske umetne inteligence *(ang. spatial AI)* in računalniškega vida. Od ostalih jih razlikuje odptokodnost vseh njihovih naprav in ogrodja *(ang. framework)* DepthAI. Ponujajo več različnih različic naprav, vse pa je skupno to, da imajo integriran čip Intel Movidus MyriadX, ki ponuja relativno visoke performance, pri tem pa zavzame malo prostora in porabi malo energije. Njihove izdelke lahko razdelimo na grobo razdelimo glede na 2 karakteristiki. Glede na zmožnost zajemanja slike (mono ali stereo) in glede na način napajanja in komunikacije (USB ali Ethernet in PoE). V tem delu smo uporabili napravo OAK-1, ki omogoča zajem mono slike, napaja in komunicira pa preko USB.

V osrčju naprave je modul RVC2 *(Robotic Vision Core 2)*. Modul ponuja 4 TOPS procesorske poči, od katerih he je 1.4 TOPS rezerviranih za izvajanje nevronskih mrež. Podpira pa tudi hardversko kodiranje slikov-

Slika 4.1: Slika prikazuje uporabljeno napravo Luxonis OAK-1 [10].

nih tokov (H.264, H.265, MJPEG), pospešeno izvajanje pogostih operacij v računalniškem vidu (skaliranje, rezanje, zaznavanje robov, itd.). V osrčju modula je Intelov sistem na čipu *(ang. system on chip, SoC)* Movidus Myriad X vključuje Intelov NCE *(Neural compute engine)*, 16 vektorskih procesorskih enot SHAVE, 20 hardverskih pospeševalnih enot poimenovanih *Enhanced Vision Accelerators*, ter 2.5 MB vgrajenega hitrega homogenega spomina.

4.2 DepthAI

DepthAI je hkrati programsko ogrodje *(ang. framework)* in tudi ekosistem odprtokodne programske in hardverske opreme, ki ga razvija podjetje Luxonis. Ogrodje je na voljo v dveh izvedbah, ogrodje za programski jezik Python in izvedba za programski jezik C++. Ogrodje nam olajša uporabo naprav, saj ponuja programski vmesnik *(ang. Application Programming Interface, API)*, s katerim lahko odstopamo do resursov naprave. Princip delovanja stoji na cevovodni arhitekturi. Cevovod je sestavljen iz med-seboj povezanih vozlišč, ki se izvajajo na napravi. V ogrodju imamo na razpolago več različnih tipov vozlišč, spodaj je navedenih nekaj najbolj uporabljenih:

- vozlišče za manipuliranje slike *ImageManip*, ki nam omogoča enostavno manipulacijo s slike (skaliranje, izrezovanje, pretvorba formatov, itd.),
- vozlišče za konfiguriranje in interakcijo z kamero *ColorCamera*,

- vozlišče za pretok podatkov preko USB povezave v semri iz naprave na gostiteljski sistem *XLinkOut*,
- vozlišče za pretok podatkov preko USB povezave v semri iz gostiteljskega sistema v napravo *XLinkIn*,
- vozlišče za izvajanje pomeri narejenih skript na naparvi *Script*. Skripte morajo biti napisane v programskem jeziku Python,
- vozlišče za uporabo pomeri narejenih nevronskih mrež na napravi *NeuralNetwork*. Nevronske mreže morajo bit prevedene v pravilen format,

4.3 OpenVINO

OpenVINO je komplet odprtokodnih orodji, ki nam omogočajo optimizacijo in prevajanje modelov v format, ki je primeren za delovanje na najrazličnejših napravah, med drugimi tudi VPU Intel Movidus MyriadX. Vsak model je potrebno najprej pravilno prilagoditi, da ustreza zahtevam in omejitvam ciljnega sistema.

Pri prilaghajanju moremo biti pozorni na tipe slohjev ki so uporabljeni v modelu, saj vsi sloji niso podprti na vseh napravah. Upoštevati je tudi treba, model ne more več pomniti dinamičnega stanja, predstavljamo si lahko, da model ni več kos programske opreme, temveč samo zaporedje matematičnih operacij, ki prejme vhod in vrne izhod.

Po prilagoditvi modela sledi korak optimizacije. Pri tem koraku se s pomočjo orodja *Model optimizer*, izboljša računska in prostorska poraba modela. V tem koraku se podat tudi ciljni podatkovni tip uteži in ostalih fiksnih parametru model, imena izhodnih in vhodnih podatkov ter dimenzionalnost vhodnih podatkov. Pri podajanju dimenzionalnosti vhdonih podatkov moramo upoštevati, da nekatere ciljne naprave ne podpiranjo dinamičnih velikosti.

Po optimizacijskem koraku sledi še zandji korak, prevajanje. Model je potrebno prevesti v pravilen format. Pri prevajanju moramo podati tip končne

napave, podatkovni tip vhodnih podatkov in število vektorskih procesorjev SHAVE, ki jih bo model uporabil. Rezultat prevajanja je binarna datoteka v formatu *.blob*.

4.4 Uporabljene tehnologije

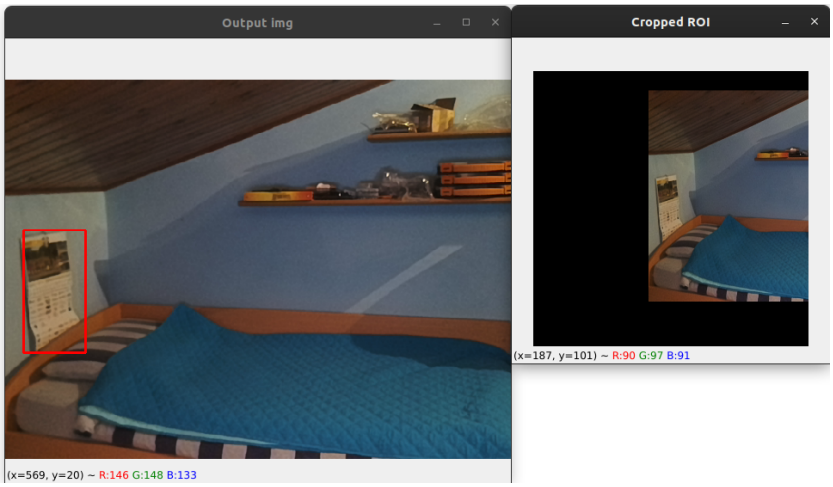
Uporabljen je bil že naučen model STARK, bolj specifično STARK-Lightning. Gre za različico sledilnika, ki uporablja samo prostorsko informacijo, poleg tega pa se različica Lightning razlikuje še po tem, da porabi precej manj procesorske moči in prostora, seveda pa je posledica tega slabša natančnost. Model so [14] implementirali v ogrodju PyTorch, z uporabo programskega jezika Python, v katerem smo tudi mi nadaljevali z razvojem. Za lažji razvoj sta se uporabili dve konetejnerizirani okolji Docker. Eno okolje je bilo namenjeno razvoju modela v programskem jeziku Python, drugo okolje pa je bilo uporabljeno za namestitev in uporabo kompleta orodji OpenVINO. prednost uporabe kontejneriziranjih okolji je prenosljivost in reproduciranje rezultatov.

4.5 Prilagoditev modela

Model pri delovanju potrebuje vhdono iskalno območje, ki je izrezano iz vhdone slike. Iskalno območje je fiksne dimenzije, če gre za inicializacijski korak je to dimenzija $128 * 128$, pri vseh nadljnih korakih pa je dimenzija iskalnega območja $320 * 320$. Izrezovanje iskalnega območja deluje tako, da iz vhdone slike izreže območje ki ga omejuje omejevalni okvir pomnožen z faktorjem iskanja, ki je v primeru inicializacije 2, v vseh naslednjih korakih pa 5. Pri tem je potrebno upoštevati, da je omejevalno območje pri robu slike, v tem primeru se sliki dodajo dodatni robovi *(ang. padding)* z vrednostjo 0. Da dodani robovi ne bi uplivali na končni rezultat jih je potrebno maskirati, zato se pole iskalnega območja pripravi tudi maska. Maska ima vrednost 1, kjer je bil dodan rob, na ostalih mestih pa ima vrednost 0. V začetni imple-

mentaciji je bilo izdelovanje maske vmeščeno v predprocesiranje, ker pa si na vgrajeni napravi ne smemo privoščiti preveč obsežnega predprocesiranja smo se odločili da izdelovanje maske vključimo v sam model. Na začetek modela smo dodali dodaten modul, ki izračuna masko. Naj bo $X_{1\times3\times H\times W}$ vhodna iskalna regija, maks pa se izračuna po naslednjih korakih:

1. izračunamo povprečje po 2. dimneziji matrike. Rezultat je matrika $M_{1\times1\times H\times W}$,
2. matriko preoblikujemo v $M_{1\times1\times H\times W}$,
3. da izločimo morebitne napake, ki bi jih lahko ta pristop povzročil (predpostavimo, da obstaja neničelna verjetnost, da bo kamera proizvedla vrednost piksla, katerag povreča vrednost bo 0), izvedemo še operacijo maksimalnega združevanja *(ang. max pooling)* z velikostjo okna 3×3.



Slika 4.2: Primer situacije kjer je izrezanemu iskalnemu območju dodan rob.

Za enostavnejšo umestitev modela v cevovod, je model razdeljen na 2 dela. Prvi model, od sedaj naprej ga bomo poimenovali samo **backbone**, zajema samo hrbtenico, sloj ozkega grla in pozicijsko kodiranje. Ta model se uporabi ob inicializaciji. Drugi model, od sedaj naorej ga bomo poimenovali

complete, pa ostaja nespremenjen in se uporablja pri vsakem nadaljnem koraku.

4.6 Prevajanje modela

Model je najprej potrebno iz ogrodja PyTorch izvoziti v format *ONNX*. *ONNX* je odprtokodni format za shranjevanje modelov, ki ga je ustvarila Microsoft. Za izvoz lahko uporabi funckionalnost, ki je vgrajena v PyTorch. Pri izvotu moramo podati, primer vhodnih podatkob, ter poimenovati - labelirati vhodne argumente. V tabeli 4.1 so podani uporabljeni argumenti.

labela	**opis**	**dimenzije**	**podatkovni tip**
Model: backbone			
img	iskalno območje	$1 \times 3 \times 128 \times 128$	float16
Model: complete			
img_x	iskalno območje	$1 \times 3 \times 320 \times 320$	float16
feat_z	sekveca značilk matrice	$64 \times 1 \times 128$	float16
mask_z	maska matrice	1×64	bool
pos_z	pozicijsko kodiranje matrice	$64 \times 1 \times 128$	float16

Tabela 4.1: Tabela prikazuje uporabljene argumnte pri izvozu modelov v format *ONNX*.

Ko je model uspešno izvožen v format *ONNX*, ga lahko z orodjem, iz paketa OpenVINO, *Model optimizer*

Članki v revijah

[4] Mauro Fernández-Sanjurjo, Manuel Mucientes in Víctor Manuel Brea. "Real-Time Multiple Object Visual Tracking for Embedded GPU Systems". V: *IEEE Internet of Things Journal* 8.11 (2021), str. 9177–9188. DOI: 10.1109/JIOT.2021.3056239.

[6] João F Henriques in sod. "High-speed tracking with kernelized correlation filters". V: *IEEE transactions on pattern analysis and machine intelligence* 37.3 (2014), str. 583–596.

[7] Matej Kristan in sod. "A Novel Performance Evaluation Methodology for Single-Target Trackers". V: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.11 (nov. 2016), str. 2137–2155. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2516982.

[12] Joseph Redmon in Ali Farhadi. "Yolov3: An incremental improvement". V: *arXiv preprint arXiv:1804.02767* (2018).

https://www.overleaf.com/project/609ce2055f917cb2f776732e

Članki v zbornikih

[2] David S. Bolme in sod. "Visual object tracking using adaptive correlation filters". V: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, str. 2544–2550. DOI: 10.1109/CVPR.2010.5539960.

[3] Nicolas Carion in sod. "End-to-end object detection with transformers". V: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer. 2020, str. 213–229.

[5] Kaiming He in sod. "Deep residual learning for image recognition". V: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, str. 770–778.

[9] Ville Lehtola in sod. "Evaluation of Visual Tracking Algorithms for Embedded Devices". V: *Image Analysis*. Ur. Puneet Sharma in Filippo Maria Bianchi. Cham: Springer International Publishing, 2017, str. 88–97. ISBN: 978-3-319-59126-1.

[13] Ashish Vaswani in sod. "Attention is All You Need". V: 2017. URL: https://arxiv.org/pdf/1706.03762.pdf.

[14] Bin Yan in sod. "Learning spatio-temporal transformer for visual tracking". V: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, str. 10448–10457.

Celotna literatura

[1] Sai Balaji. *Binary Image classifier CNN using TensorFlow*. URL: https://medium.com/techiepedia/binary-image-classifier-cnn-using-tensorflow-a3f5d6746697 (pridobljeno 2023).

[2] David S. Bolme in sod. "Visual object tracking using adaptive correlation filters". V: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, str. 2544–2550. DOI: 10.1109/CVPR.2010.5539960.

[3] Nicolas Carion in sod. "End-to-end object detection with transformers". V: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer. 2020, str. 213–229.

[4] Mauro Fernández-Sanjurjo, Manuel Mucientes in Víctor Manuel Brea. "Real-Time Multiple Object Visual Tracking for Embedded GPU Systems". V: *IEEE Internet of Things Journal* 8.11 (2021), str. 9177–9188. DOI: 10.1109/JIOT.2021.3056239.

[5] Kaiming He in sod. "Deep residual learning for image recognition". V: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, str. 770–778.

[6] João F Henriques in sod. "High-speed tracking with kernelized correlation filters". V: *IEEE transactions on pattern analysis and machine intelligence* 37.3 (2014), str. 583–596.

[7] Matej Kristan in sod. "A Novel Performance Evaluation Methodology for Single-Target Trackers". V: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.11 (nov. 2016), str. 2137–2155. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2516982.

[8] Matej Kristan in sod. *The Tenth Visual Object Tracking VOT2022 Challenge Results*. 2022.

[9] Ville Lehtola in sod. "Evaluation of Visual Tracking Algorithms for Embedded Devices". V: *Image Analysis*. Ur. Puneet Sharma in Filippo Maria Bianchi. Cham: Springer International Publishing, 2017, str. 88–97. ISBN: 978-3-319-59126-1.

[10] *Luxonis*. URL: https://www.luxonis.com/ (pridobljeno 2023).

[11] Nisarg Patel. *What Is Deep Learning?* URL: https://medium.com/@nnpatel4583/what-is-deep-learning-4daa22ceea4e (pridobljeno 2023).

[12] Joseph Redmon in Ali Farhadi. "Yolov3: An incremental improvement". V: *arXiv preprint arXiv:1804.02767* (2018).

[13] Ashish Vaswani in sod. "Attention is All You Need". V: 2017. URL: https://arxiv.org/pdf/1706.03762.pdf.

[14] Bin Yan in sod. "Learning spatio-temporal transformer for visual tracking". V: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, str. 10448–10457.