Edge AI - finale_Layout 1 14-07-26 2:30 PM Page 3

Edge AI u praksi

AI projekti za Raspberry Pi sa AI HAT+ modulom



Günter Spanner

Agencija Eho
www.infoelektronika.net

Edge AI - finale_Layout 1 14-07-26 2:30 PM Page 4

Sva prava zadržana. Nijedan deo ove knjige ne sme biti reprodukovan u bilo kom materijalnom obliku, uključujući fotokopiranje ili slučajno ili nenamerno smeštanje na bilo koji elektronski medijum sa ili uz pomoć bilo kog elektronskog sredstva, bez pismenog odobrenja nosioca autorskih prava osim u skladu sa odredbama zakona o autorskim pravima, dizajnu i patentima iz 1988. godine ili pod uslovima izdatim od Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London, England W1P 9HE. Prijave za pismene dozvole radi štampanja bilo kog dela ove publikacije upućuje se izdavaču ove knjige.

Izjava: Autor i izdavač su uložili najveće napore da bi se obezbedila tačnost informacija sadržanih u ovoj knjizi. Autor i izdavač ne mogu da pretpostave neprijatnosti i ovom izjavom isključuju bilo kakvu odgovornost za bilo koju stranku koja bi imala gubitke ili štetu uzrokovanu greškama ili propustima u ovoj knjizi, bez obzira da li su greške ili propusti nastali usled nemara, nezgode ili bilo kog drugog razloga.

ISBN 978-86-80134-628

Edge AI u praksi
Naslov originala: Edge AI Made Practical

Autor: Günter Spanner

Prevod: Vladislav Stefanović
Recenzent: Vladimir Savić

Izdaje i štampa: Agencija Eho, Niš
e-mail: redakcija@infoelektronika.net

Tiraž: 200
Godina izdanja: 2026

CIP - Каталогизација у публикацији Народна библиотека Србије, Београд

004.383/.384(035)
004.8:004.383.3(035)

ШПАНЕР, Гинтер, 1965-
Edge AI u praksi : AI projekti za Raspberry Pi sa AI HAT+ modulom / Günter

Spanner ; [prevod Vladislav Stefanović]. - Niš : Agencija Eho, 2026 (Niš : Agencija

Eho). - 204 str. : ilustr. ; 25 cm

Upozorenja **9**

Preuzimanje programa **9**

Poglavlje 1 • Uvod **10**

1.1. Zašto je AI u malim sistemima relevantan? 10
1.2. Tehničke prednosti Raspberry Pi AI HAT+ 13

Poglavlje 2 • Osnove veštačke inteligencije i mašinskog učenja **16**

2.1. Uvod u koncepte AI i ML 16
2.2. Tipovi mašinskog učenja: nadgledano, nenadgledano i učenje potkrepljivanjem 17
2.3. Osnove neuronskih mreža i dubokog učenja 18
2.4. Edge AI naspram Cloud-AI 21

Poglavlje 3 • Etička razmatranja i buduće perspektive **25**

3.1. Etička i društvena pitanja 25
3.2. Perspektive 25

Poglavlje 4 • Raspberry Pi AI HAT+ **27**

4.1. Tehničke specifikacije Raspberry Pi AI HAT+ 28
4.2. Oblasti primene 28

Poglavlje 5 • Puštanje Raspberry Pi računara u rad **31**

5.1. Potrebne komponente 31
5.2. Instalacija Raspberry Pi OS 32
5.3. Osnovno podešavanje i konfiguracija 33
5.4. Napajanje za Raspberry Pi 35
5.5. Aktivni hladnjak 36

Poglavlje 6 • Vreme za praktičan rad - fizičko računarstvo **39**

6.1. Početak rada sa Thonny okruženjem - korisnički pristupačan Python IDE 39
6.2. Instalacija Thonny okruženja 40
6.3. Pisanje i pokretanje prvog programa 41
6.4. Kod za fizičko računarstvo u Pythonu 42
6.6. Semafori i signalne kule 45
6.8. Upravljanje paketima u Thonnyju 57
6.9. Biti ili ne biti: Superkorisnik 58

Poglavlje 7 • Montaža hardvera Raspberry Pi AI HAT+ 61
7.1. GPIO stacking header konektor.....61
7.2. Kompletna montaža.....63
Poglavlje 8 • Instalacija softvera 66
8.1. Omogućavanje PCIe Gen3.....66
8.2. Instalacija Hailo paketa 68
8.3. Provera instalacije softvera.....68
Poglavlje 9 • Raspberry Pi uči da vidi: kamere 72
9.1. Pregled Raspberry Pi kamera.....72
9.2. Kablovi za Raspberry Pi kamere – šta treba da znate 73
9.3. Originalne Pi kamere.....74
9.4. USB web-kamere.....77
9.5. Korišćenje kamera.....78
9.6. Više kamera.....81
Poglavlje 10 • Prvi ulazak u Edge AI: osnovni tokovi obrade podataka 83
10.1. Detekcija osoba 84
10.2. Procena poze 87
10.3. Primer segmentacije instanci 89
10.4. Procena dubine 91
Poglavlje 11 • YOLO detekcija objekata i osoba 93
11.1. YOLO u radu 93
11.2. Pogled ispod haube.....95
11.3. Callback metoda.....98
11.4. Lako se prepadaš? Pametno upozorenje za detekciju osobe! 98
11.5. Žurka - prebrojte goste!.....102
11.6. Brojač objekata za automobile, šolje ili krave... 106
11.7. Koji objekti se prepoznaju?.....113
11.8. Proširenja i vežbe.....116
11.9. Lociranje objekata ili ljudi.....117
Poglavlje 12 • Hvatanje položaja tela 124
12.1. Upravljanje gestovima.....124
12.2. Računanje uglova.....129
12.3. Proširenja i vežbe.....133

Poglavlje 13 • Koristite servomotore za upravljanje fizičkim svetom.....134
13.1. Dobijanje uglova tela.....135
13.2. Upravljanje robotskom rukom.....136
13.3. Proširenja i vežbe.....144
Poglavlje 14 • Efikasna detekcija pomoću rpicam-apps.....145
14.1. Brza alarmna detekcija upada.....145
14.2. Pametni detektor objekata.....149
14.3. WLAN alarm za upad.....151
14.4. Proširenja i vežbe.....156
Poglavlje 15 • Profesionalno igranje: Catch A Star!.....158
15.1. Upravljanje igrom.....158
15.2. Položaj šake pomoću procene poze.....166
15.3. Proširenja i vežbe.....169
Poglavlje 16 • Razumevanje čipa Hailo-8L.....170
16.1. Arhitektura i specifikacije čipa Hailo-8L.....170
16.2. Performanse: 13 ili 26 TOPS.....172
16.3. Poređenje sa drugim AI akceleratorima.....173
16.4. Potrošnja energije i efikasnost.....174
Poglavlje 17 • Budućnost Edge AI-ja.....176
17.1. Trendovi i tehnologije u razvoju.....176
17.2. Potencijalni uticaji i primene.....178
Poglavlje 18 • Python za AI – Praktični vodič.....180
18.1. Osnova: Komentari u Python-u za AI projekte.....180
18.2. Komentarisanje u Thonny IDE.....181
18.3. Izlaz sa print() – Komunikacija rezultata na Raspberry Pi.....182
18.4. Struktura putem uvlačenja i blokova.....182
18.5. Promenljive i konstante.....184
18.6. Brojevi i tipovi promenljivih.....185
18.7. Konvertovanje tipova brojeva.....187
18.8. Operatori.....187
18.9. Nizovi i liste.....189
18.11. Kontrolisani završetak programa.....193
18.12. Upravljanje pinovima i portovima za digitalne ulaze i izlaze.....194

Poglavlje 19 • Hardvereski kompendijum.....196
19.1. Eksperimentalne pločice (Breadboards) - Alat za prototipiranje AI inženjera.....196
19.3. Kratkospojnici (Wire Jumpers).....198
19.3. Otpornici.....199
19.4. Svetleće diode (LED).....200
19.5. Kondenzatori.....201
19.6. Elektrolitski kondenzatori.....202
Poglavlje 20 • Otklanjanje grešaka.....203
Poglavlje 21 • Izvori i literatura.....204

Upozorenja

- Kola i ploče opisani u ovoj knjizi mogu se napajati isključivo proverenim, dvostruko izolovanim sigurnosnim izvorima napajanja. Kvar izolacije u jednostavnom izvoru napajanja mogao bi da dovede do pojave životno opasnih napona na neizolovanim komponentama.
- LED diode velike snage mogu da izazovu oštećenje vida. Nikada ne gledajte direktno u LED diodu!
- Autor i izdavač ne prihvataju odgovornost za bilo kakvu štetu nastalu izradom opisanih projekata.
- Elektronska kola mogu da emituju elektromagnetne smetnje. Pošto izdavač i autor nemaju uticaj na tehničku realizaciju korisnika, korisnik je isključivo odgovoran za poštovanje važećih granica emisije.

Preuzimanje programa

Programi iz ove knjige mogu se preuzeti sa:

www.elektor.com ili **infoelektronika.net**

Ako program nije identičan onome koji je opisan u knjizi, treba koristiti verziju iz preuzimanja, jer je ona ažurnija.

Poglavlje 1 • Uvod

Ugrađivanje veštačke inteligencije u male uređaje - često nazvano „Edge AI" - brzo menja način na koji koristimo svakodnevne elektronske uređaje. Smeštanjem „mozga" veštačke inteligencije direktno u kamere, telefone, nosive uređaje i drugu kompaktnu elektroniku dobijaju se moćne funkcije bez slanja podataka udaljenim serverima.

Edge AI označava primenu algoritama veštačke inteligencije (AI) lokalno na uređajima, odnosno „na ivici" mreže, umesto na centralizovanim serverima u Cloud AI.

Zamislite sigurnosnu kameru koja trenutno prepoznaje posetioca na vratima, razlikuje prijatelje i članove porodice, a ignoriše kućne ljubimce. Nema kružnog kašnjenja do Cloud AI, pa upozorenje stiže odmah kada se neko približi. Zamislite pametni telefon koji prevodi govor u hodu bez slanja zvuka na internet, omogućavajući privatnu komunikaciju u inostranstvu. Zamislite nosivi zdravstveni uređaj koji u realnom vremenu uočava nepravilan srčani ritam i podstiče korisnika da reaguje pre nego što se oglasi alarm - bez čekanja da centralni server obradi podatke.

To više nisu futurističke zamisli; to su praktične prednosti izvršavanja AI obrade direktno na samom uređaju. Lokalna obrada podataka znači da se odluke donose u milisekundama, čime se poboljšava odziv u životno važnim i vremenski kritičnim scenarijima. Pošto sirovi podaci nikada ne napuštaju uređaj, lični video-snimci, glasovni zapisi i zdravstveni parametri ostaju privatni i bezbedni. Na mestima gde je povezivanje nestabilno ili skupo - na seoskim imanjima, gradilištima ili čak u tunelu metroa - Edge AI uređaji nastavljaju da rade bez prekida.

Energetski efikasni AI čipovi projektovani za rad na periferiji mreže (edge computing) mogu da obavljaju bilione operacija u sekundi uz potrošnju od svega nekoliko vati, čime se produžava autonomija baterijski napajanih uređaja. Konačno, postavljanje hiljada malih, samostalnih jedinica može biti skalabilnije i otpornije od oslanjanja na nekoliko centralizovanih servera u Cloud AI; ukoliko jedan uređaj otkaže, ostali nastavljaju nesmetan rad.

Ugrađivanjem AI funkcija u samu strukturu uređaja stvara se svet trenutnih uvida, veće privatnosti, nižih troškova i stvarne mogućnosti rada van mreže - što menja način na koji živimo, radimo i koristimo tehnologiju.

1.1. Zašto je AI u malim sistemima relevantan?

Prvo treba razmotriti brzinu. Kada se podaci sa senzora, fotografija ili kratak audiozapis šalju udaljenom serveru na analizu, uvodi se kašnjenje. Čak i kašnjenje od nekoliko stotina milisekundi može napraviti značajnu razliku. U situacijama kao što je izbegavanje prepreka kod autonomne igračke, svaki milisekund je presudan. Kada se inferencija, odnosno korak donošenja odluke AI modela, izvršava direktno na ure-

đaju, eliminiše se komunikacija napred-nazad. Kao rezultat, odgovori se dobijaju gotovo trenutno, pa kamere, dronovi ili roboti mogu da deluju bez oklevanja.

Privatnost se takođe znatno povećava. Kada se sirovi podaci - kao što su glas, snimci lica ili kućni video-zapisi - prenose preko interneta, otvaraju se pitanja privatnosti. Postavlja se pitanje ko može da vidi podatke i gde se oni čuvaju. Kod Edge AI pristupa osetljive informacije ostaju na uređaju. Obradu lokalno obavlja AI model, a šalju se samo minimalni, anonimizovani rezultati: kratka poruka od nekoliko bitova umesto celog video-zapisa veličine više megabajta ili čak gigabajta. Time se smanjuju komplikacije u vezi sa privatnošću i olakšava usklađivanje sa propisima kao što je GDPR (Opšta uredba o zaštiti podataka) u Evropi ili strogi propisi u zdravstvu, koji zahtevaju rigorozno postupanje sa ličnim podacima.

Upotreba interneta i troškovi predstavljaju ključna razmatranja. U mnogim regionima povezivanje je sporo ili skupo. Mobilne ili satelitske veze često se naplaćuju prema količini prenetih podataka. Ako bi se video visoke rezolucije neprekidno prenosio u Cloud AI, nastali bi značajni troškovi. Međutim, ako Edge AI kamere lokalno filtriraju video-tokove, otpremaju se samo neophodni snimci, na primer kada se detektuje životinja ili kada se pojavi problem na mašini. To drastično smanjuje prenos podataka i održava troškove niskim.

Pouzdanost je još jedna prednost. Uređaji koji zavise od cloud servisa rade onoliko dobro koliko to mreža dozvoljava. Ako se izgubi Wi-Fi ili mobilni signal, pametne funkcije mogu prestati da rade. Kod AI obrade na uređaju sistemi nastavljaju da funkcionišu i van mreže. Na primer, dronovi koji mapiraju udaljene šume mogu nastaviti da identifikuju vrste drveća, a nosivi zdravstveni monitori mogu i dalje izdavati upozorenja o abnormalnom srčanom ritmu, čak i tokom događaja kao što su maratoni u tunelima. Mogućnost rada van mreže obezbeđuje pouzdanost na bilo kojoj lokaciji.

Trajanje baterije i energetska efikasnost su od presudne važnosti, naročito kod prenosivih ili udaljenih uređaja. Izvršavanje složenih AI modela na procesorima opšte namene ili stalno slanje podataka putem mobilnih modema može brzo isprazniti bateriju. Specijalizovani AI čipovi - kao što je Hailo-8L - projektovani su da izvršavaju bilione operacija u sekundi po svakom utrošenom vatu. U praktičnom smislu, značajna AI obrada obavlja se uz minimalnu potrošnju energije, omogućavajući baterijski napajanim kamerama ili senzorima rad tokom dužih perioda - danima ili nedeljama umesto satima.

Skalabilnost i ekonomičnost su prirodno povezane. Kada se hiljade pametnih senzora postavljaju u urbanom okruženju - za nadzor saobraćaja, praćenje kvaliteta vazduha ili javnu bezbednost - opremanje svakog čvora Edge AI sposobnošću uklanja potrebu za velikom cloud serverskom infrastrukturom i povezanim troškovima računanja. Obrada se izvršava nezavisno na svakom uređaju. Ovaj distribuirani model

se efikasno skalira: uređaji se mogu dodavati po potrebi, bez naglog rasta troškova Cloud AI.

Neočekivana prednost ogleda se u otpornosti sistema. Centralizovane cloud platforme mogu imati prekide rada, održavanje ili sajber-napade, što može onesposobiti pametne funkcije. Autonomni uređaji, međutim, održavaju funkcionalnost nezavisno. Na primer, kućne sigurnosne kamere nastavljaju da detektuju kretanje čak i tokom ažuriranja ili prekida rada Cloud AI proizvođača.

Iz razvojne perspektive, Edge AI otvara mogućnosti za inovacije. AI modeli mogu se prilagoditi radu u okviru ograničenja kao što su mala memorija, skromna procesorska snaga i ograničena dostupna energija. Alati kao što je Hailo Dataflow Compiler pomažu u pretvaranju široko korišćenih formata neuronskih mreža u optimizovan kôd pogodan za izvršavanje na edge čipovima. To omogućava brzu iteraciju i brzo prototipovanje novih primena - kao što su senzori za praćenje biljaka koji detektuju suvo zemljište ili robotske ruke upravljane gestovima.

Naravno, postoje i izazovi pri primeni AI na malim uređajima. Modeli se moraju smanjiti tehnikama kao što su kvantizacija (manje bitova po broju) ili orezivanje mreže, odnosno pruning (uklanjanje retko korišćenih veza). Mora se pažljivo uravnotežiti tačnost, brzina i potrošnja energije. Srećom, ekosistem alata i podrške zajednice brzo napreduje, pa je integracija složenih modela u kompaktne sisteme sve jednostavnija.

Ne treba zanemariti ni uticaj na životnu sredinu. Veliki data centri troše ogromne količine energije i zahtevaju značajno hlađenje. Distribuiranjem inferencije na brojne efikasne uređaje može se smanjiti ukupan ugljenični otisak AI aplikacija. Kada se obrada obavlja lokalno umesto centralizovano u cloud farmama zasnovanim na GPU procesorima, ukupna potrošnja energije često opada. U kontekstu globalnih ciljeva održivosti, Edge AI nudi zeleniju alternativu.

Zaključno, ugrađivanje AI direktno u male uređaje omogućava prednosti u pogledu:

- brzine,
- privatnosti,
- troškova,
- pouzdanosti,
- efikasnosti,
- skalabilnosti.

Podaci ostaju lokalni, upotreba propusnog opsega se smanjuje na minimum, a funkcionalnost se održava čak i u udaljenim ili zahtevnim okruženjima. Energetski efikasni AI čipovi omogućavaju uređajima da obavljaju složene zadatke bez brzog pražnjenja baterija, dok distribuirana inteligencija povećava otpornost sistema. Kako se alati i metode budu poboljšavali, pojavljivaće se nove kreativne primene - u pametnim domovima, industrijskom nadzoru, nosivim zdravstvenim uređajima i au-

tonomnim mašinama. Ova evolucija predstavlja više od tehnološkog trenda - ona čini osnovu budućnosti ispunjene bezbednim, održivim i odzivnim inteligentnim sistemima.

Konačno, Edge AI se lako i povoljno skalira. Umesto izgradnje jednog ogromnog data centra, organizacije mogu distribuirati stotine ili hiljade malih jedinica, od kojih svaka obavlja sopstvenu lokalnu obradu. Dodavanje novih uređaja je jednostavno i ne dovodi do naglog rasta cloud-compute troškova. Pošto je inteligencija decentralizovana, problem u jednoj jedinici - kao što je hardverski kvar ili ažuriranje softvera - ne obara ceo sistem.

Kako specijalizovani čipovi, softverski alati i tehnike optimizacije modela nastavljaju da se poboljšavaju, Edge AI će se širiti u svim oblastima života. Pametni domovi će učiti naše rutine, fabrike će se samostalno korigovati u realnom vremenu, lekari će imati trenutnu dijagnostičku podršku pored pacijenta, a sistemi za praćenje životne sredine štitiće divlje životinje i poljoprivredno zemljište bez ljudske intervencije. Premeštanjem AI obrade na ivicu dobijamo brže reakcije, jaču privatnost, niže troškove, veću pouzdanost i bolju energetsku efikasnost - zbog čega uređaji ne postaju samo povezani, već zaista inteligentni.

1.2. Tehničke prednosti Raspberry Pi AI HAT+

Raspberry Pi AI HAT+ predstavlja značajan iskorak u primeni veštačke inteligencije na krajnjim uređajima, odnosno na ivici mreže. Komplet kombinuje specijalizovani AI procesor kompanije Hailo sa razvojnim alatima, omogućavajući brzo prototipovanje i efikasno izvršavanje modela dubokog učenja na ugrađenim uređajima. Tehničke prednosti Raspberry Pi AI HAT+ obuhvataju energetsku efikasnost, skalabilnu arhitekturu i podršku za industrijski standardne AI okvire. Raspberry Pi AI HAT+ koristan je u primenama od detekcije objekata u realnom vremenu do inteligentne analize senzorskih podataka bez oslanjanja na Cloud AI. Neke od glavnih prednosti su:

- **Visoke računske performanse:** Do 26 TOPS (tera operacija u sekundi), čime nadmašuje mnoge CPU/GPU jedinice u jednopločnim računarima
- **Energetska efikasnost:** Troši samo 2,5-3 W, idealno za baterijski napajane sisteme
- **Podrška za okvire:** Kompatibilnost sa TensorFlow, PyTorch, ONNX i OpenCV
- **Jednostavna GStreamer integracija:** Gotovi i optimizovani tokovi za obradu video-signala u AI aplikacijama

Sa AI HAT+ dodatkom Raspberry Pi može da obavlja video-analitiku u realnom vremenu, što bi bilo jednostavno nemoguće na njegovom CPU-u, pa čak i GPU-u. Tipični slučajevi primene obuhvataju:

- Brza detekcija objekata i lica: Za nadzorne kamere ili sigurnosne sisteme
- Optička kontrola kvaliteta: Uočavanje nedostataka na proizvodnoj liniji
- Prepoznavanje teksta visokih performansi (OCR - Optical Character Recognition): Automatizacija obrade dokumenata ili očitavanja registarskih tablica

Zahvaljujući GStreamer podršci, AI HAT+ se može direktno uključiti u žive tokove kamere radi detekcije i praćenja objekata u hodu.

Jedna od najvećih prepreka kod dronova i robota jeste obrada senzorskih podataka u realnom vremenu. Raspberry Pi AI HAT+ to omogućava i obuhvata:

- **Navigacija i izbegavanje prepreka:** Autonomna vozila i roboti mogu brže da analiziraju okruženje
- **Upravljanje gestovima:** Prepoznavanje pokreta ruke radi upravljanja robotima ili dronovima
- **Praćenje objekata:** Omogućava dronovima da prate pokretne mete ili mapiraju teren

Mala potrošnja energije znači da je AI HAT+ pogodan za mobilne, baterijski napajane platforme kao što su roboti i dronovi. U kućnoj automatizaciji AI HAT+ omogućava:

- **Lokalno glasovno upravljanje i prepoznavanje:** Nije potrebna veza sa Cloud AI
- **Detekcija anomalija:** Uočavanje sumnjivih pokreta ili neuobičajenih zvukova
- **Automatsko upravljanje osvetljenjem i klimom:** Podešavanje parametara na osnovu detektovanog prisustva

Edge computing sa AI HAT+ dodatkom obezbeđuje da osetljivi podaci nikada ne moraju da napuste kućnu mrežu.

Na kraju, fabrike i industrijski pogoni imaju korist od AI obrade u realnom vremenu na proizvodnoj liniji:

- **Prediktivno održavanje:** Analiza senzorskih podataka radi predviđanja kvarova opreme
- **Automatska inspekcija:** Detekcija grešaka na proizvodnim sklopovima
- **Prepoznavanje registarskih tablica:** Naplatni sistemi / upravljanje parkingom

Performanse AI HAT+ dodatka čine ga pogodnim čak i za zahtevne industrijske primene u realnom vremenu.

Zašto jednostavno ne koristiti GPU? GPU procesori su uobičajeni AI akceleratori, ali na ivici mreže donose izazove:

- **Veća potrošnja energije:** GPU procesori često zahtevaju 10-100 W, dok AI HAT+ troši samo 2,5-3 W
- **Velike dimenzije:** Suviše su glomazni za kompaktne jednopločne računare kao što je Raspberry Pi
- **Veća cena:** Uređaji kao što su NVIDIA Jetson moduli ili spoljašnji GPU sistemi obično su znatno skuplji od AI HAT+ dodatka

Pored toga, namenski NPU (Neural Processing Unit), kao što je Hailo-8L, optimizovan je za AI opterećenja bez dodatnog opterećenja koje nosi GPU opšte namene.

Danas programeri mogu da grade moćne AI aplikacije, koje su ranije zahtevale velike GPU procesore ili cloud servere, na uređajima malog formata. Ako istražujete moderne AI tehnologije ili tražite efikasna edge rešenja, Raspberry Pi AI HAT+ svakako zaslužuje razmatranje.

Poglavlje 2 • Osnove veštačke inteligencije i mašinskog učenja

Mašinsko učenje (ML) je grana veštačke inteligencije koja omogućava računarima da uče obrasce iz podataka i donose predviđanja ili odluke bez eksplicitnog programiranja za svaki pojedinačni zadatak.

Poslednjih godina veštačka inteligencija i mašinsko učenje prešli su put od uskih istraživačkih oblasti do centralnih stubova moderne tehnologije. Od glasovnih asistenata i sistema preporuka do medicinske dijagnostike i autonomnih vozila, ove tehnologije preoblikuju industrije i svakodnevni život. Ovo poglavlje pruža osnovno razumevanje AI i ML - šta su, kako rade i zašto su važni. Počinjemo istraživanjem istorijskog razvoja AI, uz razlikovanje simboličkog rezonovanja i učenja zasnovanog na podacima. Zatim razmatramo osnovne principe mašinskog učenja, uključujući nadgledano, nenadgledano i učenje potkrepljivanjem, i uvodimo ključne pojmove kao što su modeli, algoritmi, obučavanje i evaluacija.

Umesto da se usredsredi isključivo na tehničku implementaciju, ovo poglavlje naglašava konceptualnu jasnoću, sa ciljem da čitaocima pruži čvrstu osnovu za dalje učenje i praktičnu primenu. Bez obzira na to da li ste student, programer ili jednostavno radoznali u vezi sa inteligentnim sistemima, razumevanje ovih osnova prvi je korak ka uključivanju u svet AI koji se brzo razvija.

2.1. Uvod u koncepte AI i ML

AI i ML danas spadaju među najtransformativnije oblasti računarstva i tehnologije podataka. Zajedno omogućavaju mašinama da obavljaju zadatke koji obično zahtevaju ljudsku inteligenciju - kao što su prepoznavanje obrazaca, razumevanje jezika, donošenje odluka i prilagođavanje novim informacijama. Kako ove tehnologije sve više pokreću aplikacije u zdravstvu, finansijama, obrazovanju, transportu i drugim oblastima, solidno razumevanje njihovih osnova postalo je neophodno i za studente i za profesionalce.

Veštačka inteligencija označava širi cilj stvaranja sistema koji mogu da simuliraju inteligenciju nalik ljudskoj. Ona obuhvata niz podoblasti, uključujući obradu prirodnog jezika, robotiku, računarski vid i ekspertne sisteme. AI sistemi projektuju se tako da opažaju svoje okruženje, razmatraju složene probleme, uče iz iskustva i preduzimaju radnje koje povećavaju verovatnoću ostvarivanja određenog cilja.

U okviru AI, mašinsko učenje se izdvojilo kao ključni pristup. Umesto da budu eksplicitno programirani za svaki zadatak, ML sistemi uče obrasce i odnose iz podataka. Mašinsko učenje se usredsređuje na razvoj algoritama koji omogućavaju računarima da uče i donose predviđanja ili odluke na osnovu ulaznih podataka. Ovaj prelaz sa logike zasnovane na pravilima na učenje zasnovano na podacima omogućio je izgradnju fleksibilnijih, skalabilnijih i prilagodljivijih sistema.

Ovo poglavlje razmatra osnovne principe AI i ML, uključujući:

- **Istorijski kontekst i razvoj** veštačke inteligencije
- **Tipovi AI:** uska naspram opšte AI, kao i simbolički naspram konekcionističkog pristupa
- **Kategorije mašinskog učenja:** nadgledano, nenadgledano i učenje potkrepljivanjem
- **Ključne pojmovi:** skupove podataka, karakteristike, modele, obučavanje, testiranje i metrike evaluacije
- **Algoritme i modeli:** uključujući stabla odlučivanja, linearnu regresiju, neuronske mreže i tehnike klasterovanja
- **Izazovi i ograničenja:** kao što su pristrasnost, preprilagođavanje, interpretabilnost i zavisnost od podataka
- **Etičke i društvene aspekti:** primene AI u realnim okruženjima

Na kraju ovog poglavlja čitaoci će imati konceptualni alat za razumevanje toga kako se AI i ML sistemi grade, kako funkcionišu i gde se najefikasnije primenjuju. Ovo znanje postavlja osnovu za dublje proučavanje, praktično eksperimentisanje i informisanu diskusiju o razvoju i upotrebi inteligentnih tehnologija.

2.2. Tipovi mašinskog učenja: nadgledano, nenadgledano i učenje potkrepljivanjem

Mašinsko učenje se obično deli na tri glavna tipa: nadgledano učenje, nenadgledano učenje i učenje potkrepljivanjem. Svaki tip je definisan načinom na koji algoritam uči iz podataka i vrstom povratne informacije koju dobija tokom obučavanja.

Nadgledano učenje (Supervised learning) je najčešći i najintuitivniji oblik mašinskog učenja. U ovom pristupu algoritam se obučava na označenom skupu podataka, gde je svaki ulaz povezan sa poznatim izlazom. Cilj nadgledanog učenja je da se nauči preslikavanje sa ulaza na izlaze, tako da model može tačno da predvidi izlaz za nove, ranije neviđene podatke. Ova kategorija se dalje deli na zadatke klasifikacije i regresije. Klasifikacija se koristi kada je izlaz kategorijski, na primer pri određivanju da li je e-poruka neželjena pošta ili nije. Regresija se, s druge strane, bavi kontinuiranim izlazima, kao što je predviđanje cena kuća ili globalnih temperatura.

Uobičajeni algoritmi koji se koriste u nadgledanom učenju obuhvataju linearnu regresiju, logističku regresiju, stabla odlučivanja, mašine potpornih vektora i neuronske mreže. Iako nadgledano učenje može biti veoma efikasno, ono obično zahteva velike količine označenih podataka, čije pribavljanje može biti skupo i vremenski zahtevno.

Nasuprot tome, **nenadgledano učenje** (unsupervised learning) radi sa neoznačenim podacima. Algoritam ne dobija eksplicitna uputstva o tome šta treba predvideti, već ima zadatak da identifikuje skrivene strukture ili obrasce u podacima. Najčešće tehnike nenadgledanog učenja su klasterovanje i redukcija dimenzionalnosti. Algoritmi

klasterovanja grupišu slične tačke podataka na osnovu njihovih karakteristika, zbog čega su korisni za segmentaciju korisnika ili detekciju anomalija.

Tehnike redukcije dimenzionalnosti, kao što je analiza glavnih komponenti (PCA), pomažu u smanjenju složenosti podataka uz očuvanje bitnih informacija, pa se često koriste za vizuelizaciju podataka ili pretprocesiranje. Pošto nema oznaka, ocenjivanje performansi nenadgledanih modela može biti zahtevno, a rezultati nisu uvek jednostavni za tumačenje.

Učenje potkrepljivanjem (Reinforcement learning) je suštinski drugačiji pristup, inspirisan bihevioralnom psihologijom. U ovoj paradigmi agent interaguje sa okruženjem i uči da donosi odluke primajući povratne informacije u obliku nagrada ili kazni. Cilj je naučiti politiku - strategiju izbora akcija - koja maksimizuje kumulativne nagrade tokom vremena. Učenje potkrepljivanjem posebno je korisno u dinamičnim, složenim okruženjima u kojima akcije imaju dugoročne posledice. Široko se koristi u aplikacijama kao što su igranje igara (npr. AlphaGo), robotika, autonomna vozila i prilagodljivo upravljanje resursima. Neki poznati algoritmi učenja potkrepljivanjem uključuju Q-learning, duboke Q-mreže i metode gradijenta politike. Obučavanje modela učenja potkrepljivanjem može biti računarski intenzivno i često zahteva velike količine interakcije sa okruženjem da bi se efikasno učilo. Osim toga, izazovi kao što su balansiranje istraživanja (isprobavanje novih akcija) i eksploatacije (biranje najbolje poznate akcije) čine ovaj okvir učenja složenijim.

Ukratko, nadgledano učenje je najbolje u scenarijima u kojima su dostupni označeni podaci i potrebna su precizna predviđanja; nenadgledano učenje je pogodno za otkrivanje obrazaca u neoznačenim podacima; a učenje potkrepljivanjem je idealno za učenje optimalnih strategija putem interakcije i povratnih informacija. Razumevanje razlika i odgovarajućih primena ovih tipova učenja osnovno je za uspešnu primenu tehnika mašinskog učenja u različitim domenima.

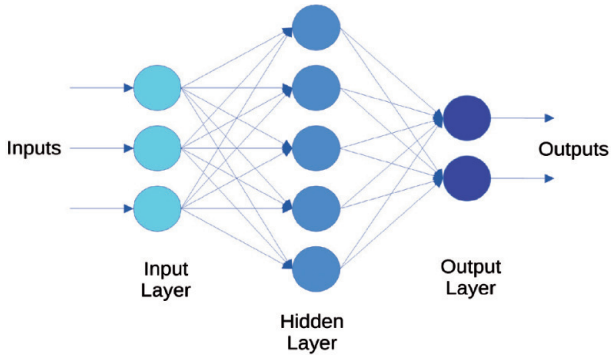
2.3. Osnove neuronskih mreža i dubokog učenja

Neuronske mreže su osnovni koncept moderne veštačke inteligencije, naročito u oblasti dubokog učenja, koje je postiglo izuzetne uspehe u oblastima kao što su prepoznavanje slika, obrada prirodnog jezika, sinteza govora i autonomni sistemi. Inspirisane strukturom i funkcionisanjem ljudskog mozga, neuronske mreže su računarski modeli sastavljeni od slojeva međusobno povezanih čvorova, nazvanih neuroni ili jedinice, koji obrađuju podatke da bi naučili obrasce i reprezentacije.

U osnovi neuronske mreže nalazi se veštački neuron, poznat i kao perceptron, koji oponaša sposobnost biološkog neurona da se aktivira kao odgovor na stimuluse. Svaki neuron prima ulaze, često numeričke vrednosti, obrađuje ih kroz ponderisanu sumu, dodaje bias član i prosleđuje rezultat kroz nelinearnu aktivacionu funkciju kao što su sigmoid, tanh ili ReLU (Rectified Linear Unit). Ova aktivacija određuje da li se i koliko snažno neuron „aktivira”, doprinoseći ukupnom procesu donošenja odluka u mreži.

Neuronske mreže se obično organizuju u slojeve:

1. **Ulazni sloj:** Prima sirove ulazne podatke, kao što su vrednosti piksela slike ili numeričke karakteristike iz skupa podataka.
2. **Skriveni slojevi:** Jedan ili više međuslojeva u kojima se obavljaju proračuni. Svaki neuron u skrivenom sloju povezan je sa neuronima u susednim slojevima, a ti slojevi uče da iz ulaza izdvajaju sve apstraktnije reprezentacije.
3. **Izlazni sloj:** Daje konačan rezultat mreže, kao što je oznaka klasifikacije ili numeričko predviđanje.



Slika 2.1: Slojevi neuronske mreže

Kada neuronska mreža sadrži mnogo skrivenih slojeva - često više od tri - naziva se duboka neuronska mreža, a proces učenja naziva se duboko učenje. Ove duboke arhitekture omogućavaju modelima da uče složene hijerarhijske karakteristike, što ih čini moćnim za zadatke kao što su prepoznavanje lica, prevođenje jezika ili generisanje teksta.

Obučavanje neuronske mreže uključuje proces nazvan propagacija greške unazad (backpropagation), koji izračunava gradijent funkcije gubitka, odnosno mere koliko je predviđanje modela udaljeno od stvarne oznake, u odnosu na svaku težinu u mreži. Korišćenjem optimizacionog algoritma kao što je stohastički gradijentni spust (SGD) ili Adam, mreža iterativno ažurira svoje težine kako bi minimizovala funkciju gubitka, čime se poboljšava tačnost.

Razvijeno je nekoliko važnih arhitektura neuronskih mreža za rad sa različitim tipovima podataka:

- **Potpuno povezane mreže (Dense Networks)** su osnovne arhitekture u kojima je svaki neuron povezan sa svakim neuronom u prethodnom i narednom sloju. Pogodne su za strukturirane podatke ili jednostavne klasifikacione probleme.
- **Konvolucione neuronske mreže (CNN)** posebno su projektovane za obradu slikovnih podataka. One koriste konvolucione slojeve za prepoznavanje pro-

stornih obrazaca u slici, počev od jednostavnih elemenata kao što su konture, teksture i osnovni oblici, pa sve do složenijih struktura i objekata.

- **Rekurentne neuronske mreže (RNN)** i njihove napredne varijante kao što su **LSTM (Long Short-Term Memory)** i **GRU (Gated Recurrent Units)** prilagođene su sekvencijalnim podacima, kao što su vremenske serije ili prirodni jezik, jer mogu održavati unutrašnju memoriju prethodnih ulaza.

- **Transformeri**, novija inovacija, uvela je revoluciju u obradi prirodnog jezika (NLP) omogućavanjem paralelne obrade sekvencijalnih podataka umesto oslanjanja na sekvencijalne korake kao kod RNN mreža. Uvedeni u uticajnom radu „Attention is All You Need" (Vaswani i dr., 2017), transformeri koriste mehanizam nazvan **samopažnja** (self-attention), koji modelu omogućava da odmeri značaj različitih reči u sekvenci jednih u odnosu na druge. Ova arhitektura stoji u osnovi velikih modela kao što je GPT, kao i drugih osnovnih modela, i postala je kamen temeljac savremenih primena dubokog učenja u razumevanju i generisanju jezika.

Još jedan važan aspekt dubokog učenja jeste učenje karakteristika. Za razliku od tradicionalnih modela mašinskog učenja, koji se često oslanjaju na ručno definisane karakteristike, duboke neuronske mreže automatski uče karakteristike iz sirovih podataka. Kako informacije prolaze kroz uzastopne slojeve, mreža postepeno izdvaja apstrakcije višeg nivoa - na primer, kod obrade slika rani slojevi mogu detektovati ivice i teksture, dok dublji slojevi identifikuju oblike, objekte i kontekst.

Obučavanje dubokih neuronskih mreža nosi nekoliko izazova. Čest problem je preprilagođavanje (overfitting), kada model dobro radi na podacima za obučavanje, ali loše na ranije neviđenim podacima. Za suzbijanje toga obično se koriste tehnike kao što su dropout, odnosno nasumično deaktiviranje neurona tokom obučavanja, regularizacija i rano zaustavljanje. Pored toga, obučavanje dubokih modela zahteva značajne računarske resurse i često velike označene skupove podataka, što je dovelo do razvoja transfernog učenja. To je metoda u kojoj se prethodno obučen model fino podešava za konkretan zadatak pomoću manjeg skupa podataka.

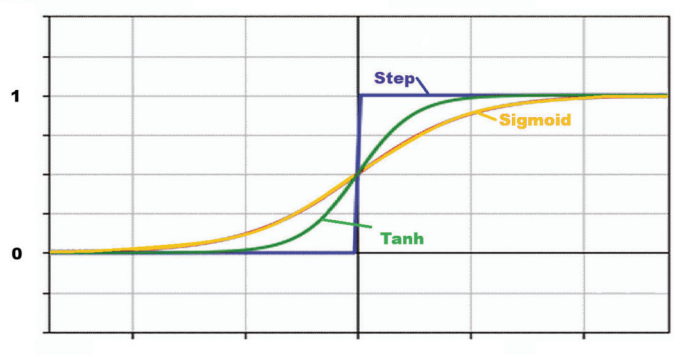
Još jedna ključna komponenta u tokovima rada dubokog učenja jeste upotreba aktivacionih funkcija. Ove funkcije uvode nelinearnost u mrežu, omogućavajući joj da modeluje složene odnose u podacima. Uobičajene aktivacione funkcije obuhvataju:

- **ReLU (Rectified Linear Unit):** Najčešće se koristi u skrivenim slojevima zbog jednostavnosti i efikasnosti.

- **Sigmoid:** Često se koristi u izlaznim slojevima kod problema binarne klasifikacije.

- **Softmax:** Pretvara izlaze u verovatnoće za zadatke višeklasne klasifikacije.

- **Step:** Daje diskretne izlaze, na primer 0 ili 1, u zavisnosti od toga da li ulaz prelazi prag.
- **Tanh (hiperbolički tangens):** Daje vrednosti između -1 i 1, obezbeđuje aktivacije centrirane oko nule i često se koristio u skrivenim slojevima pre nego što je ReLU postao popularan.



Slika 2.2: Aktivacione funkcije

Pored arhitektura i tehnika obučavanja, okviri za duboko učenje kao što su TensorFlow, PyTorch i Keras značajno su ubrzali razvoj obezbeđivanjem alata za izgradnju, obučavanje i implementaciju modela. Ove biblioteke apstrahuju mnoge niskonivojske detalje i nude unapred pripremljene module, čineći duboko učenje dostupnijim istraživačima, inženjerima i hobistima.

Ukratko, neuronske mreže čine računarsku okosnicu dubokog učenja. Slaganjem slojeva, pažljivim podešavanjem težina i upotrebom moćnih algoritama za obučavanje, ovi modeli mogu naučiti da rešavaju veoma složene zadatke u različitim domenima. Kako dostupnost podataka i računarska snaga nastavljaju da rastu, neuronske mreže i duboko učenje ostaće centralni za napredak veštačke inteligencije i inteligentnih sistema.

2.4. Edge AI naspram Cloud-AI

Kako se veštačka inteligencija sve više integriše u uređaje, aplikacije i servise, osnovna infrastruktura koja pokreće AI radna opterećenja postala je od ključnog značaja. Tradicionalno, AI modeli su razvijani, obučavani i implementirani u Cloud AI, odnosno u snažnim udaljenim data centrima povezanim putem interneta. Međutim, sa eksplozivnim rastom povezanih uređaja, senzora i aplikacija u realnom vremenu došlo je do značajnog pomeranja ka obradi AI zadataka na ivici mreže - blizu mesta gde se podaci generišu. Ova paradigma poznata je kao Edge AI.

Ovo poglavlje daje sveobuhvatno poređenje između Edge AI i Cloud AI pristupa, razmatrajući njihove arhitekture, slučajeve primene, kompromise i načine na koje se dopunjuju u savremenim AI sistemima. Razumevanje ovih paradigmi neophodno je za

izgradnju efikasnih, skalabilnih i odzivnih inteligentnih sistema, naročito u domenima kao što su IoT, robotika, autonomna vozila i pametni gradovi.

Edge AI označava primenu AI modela na edge uređajima kao što su pametni telefoni, mikrokontroleri, kamere, dronovi i ugrađeni sistemi. Ovi uređaji lokalno obrađuju podatke, često koristeći specijalizovani hardver kao što su NPU (Neural Processing Unit), GPU, FPGA ili namenski AI čipovi.

Ključne karakteristike
Inferencija na uređaju: AI proračuni se odvijaju na uređaju ili u njegovoj blizini, tamo gde se podaci generišu.

- Mala latencija: Donošenje odluka u realnom ili gotovo realnom vremenu.
- Mogućnost rada van mreže: Funkcioniše bez neprekidnog pristupa internetu.
- Ograničeni resursi: Ograničena procesorska snaga, memorija i energija.

Primeri
- Glasovni asistenti na pametnim telefonima
- AI sigurnosne kamere koje detektuju pokret ili lica
- Industrijski senzori koji u realnom vremenu prate stanje opreme

Cloud AI označava upotrebu centralizovanih data centara za obučavanje, čuvanje i posluživanje AI modela. Oslanja se na računarsku infrastrukturu visokih performansi i ogromne skupove podataka za razvoj i implementaciju složenih AI servisa. Korisnici Cloud AI servisima obično pristupaju preko API-ja ili platformi kao što su Google Cloud AI, AWS SageMaker ili Microsoft Azure AI.

Ključne karakteristike
- Centralizovana obrada: Teška računanja se prebacuju na cloud servere.
- Skalabilnost: Lako se skalira radi podrške velikim skupovima podataka i složenim modelima.
- Obučavanje i implementacija modela: Idealno za razvoj i upravljanje AI rešenjima velikog obima.
- Neprekidna povezanost: Zahteva pouzdanu internet vezu za inferenciju.

Primeri
- Cloud servisi za obradu prirodnog jezika
- AI pretraga i sistemi preporuka
- Udaljeno obučavanje i implementacija velikih modela dubokog učenja
- Arhitekturne razlike između ova dva pristupa prikazane su u sledećoj tabeli:

Aspekt	Edge AI	Cloud AI
Lokacija računanja	Na uređaju ili lokalnom gateway-u	Udaljeni cloud serveri
Latencija	Mala (realno vreme)	Veća (zavisna od mreže)
Povezivanje	Radi van mreže ili sa povremenim internetom	Zahteva stalnu internet vezu
Potrošnja energije	Optimizovano za rad male potrošnje	Velika potrošnja energije u data centrima
Privatnost podataka	Visoka, jer podaci ostaju lokalno	Podaci se šalju preko mreža (pitanja privatnosti)
Složenost modela	Ograničena zbog hardverskih ograničenja	Veoma visoka, koristi GPU klastere

Tabela 2.1: Arhitekturne razlike između Edge AI i Cloud AI

U oba slučaja postoje prednosti i ograničenja:

Edge AI: prednosti

- **Performanse u realnom vremenu:** Ključno za vremenski osetljive primene kao što su autonomna vožnja i robotika.
- **Smanjena upotreba propusnog opsega:** Manje podataka mora da se prenosi u Cloud AI.
- **Poboljšana privatnost i bezbednost:** Osetljivi podaci ostaju na uređaju.
- **Funkcionalnost van mreže:** Idealno za udaljena okruženja ili nestabilne mrežne uslove.

Edge AI: ograničenja

- **Ograničeni resursi:** Ograničenja procesorske snage, memorije i skladištenja.
- **Pojednostavljivanje modela:** Zahteva kompresiju, pruning (orezivanje nepotrebnih delova mreže) ili kvantizaciju.
- **Teška ažuriranja:** Ažuriranja firmware i modela mogu biti složena.

Cloud AI: prednosti

- **Skalabilno računanje:** Podržava obučavanje i inferenciju (izvršavanje obučenih modela nad novim podacima, u velikom obimu).
- **Centralizovana ažuriranja:** Lakše održavanje i unapređivanje modela.
- **Agregacija podataka:** Može koristiti ogromne skupove podataka radi poboljšanja tačnosti.

Cloud AI: ograničenja

- **Latencija i pouzdanost:** Zavisni od brzine i stabilnosti mreže.
- **Rizici po privatnost:** Osetljivi podaci moraju da se prenose spoljnim serverima.
- **Troškovi propusnog opsega:** Visoke stope prenosa podataka mogu biti skupe.

Sledeći pregled prikazuje najvažnije slučajeve primene i aplikacije:

Domen	Edge AI aplikacije	Cloud AI aplikacije
Zdravstvo	Nosivi zdravstveni monitori, prenosiva dijagnostika	Analiza medicinskih slika, uvidi na nivou populacije
Proizvodnja	Prediktivno održavanje, vizuelna inspekcija	Optimizacija proizvodnje, analitika lanca snabdevanja
Maloprodaja	Pametne police, autonomna naplata	Analitika kupaca, sistemi preporuka
Transport	Autonomna vozila, upravljanje semaforima	Upravljanje flotom, predviđanje saobraćaja
Poljoprivreda	Praćenje useva na terenu, detekcija štetočina	Analiza vremena, predviđanje prinosa velikog obima
Pametni gradovi	Nadzorne kamere, senzori kvaliteta vazduha	Gradska analitika, planiranje infrastrukture

Tabela 2.2: Najvažniji slučajevi upotrebe i aplikacije za Edge AI i Cloud AI

Savremeni AI sistemi sve češće usvajaju hibridni pristup, kombinujući prednosti i edge-a i Cloud AI. Na primer:

- **Federativno učenje: (Federated learning)** Decentralizovana tehnika učenja u kojoj se modeli obučavaju lokalno na edge uređajima, a centralnom serveru se šalju samo ažuriranja modela, ne i sirovi podaci.
- **Podeljeno računanje (Split computing):** Zadatak inferencije deli se između edge uređaja i Cloud AI kako bi se uravnotežile latencija i tačnost.

Ovi pristupi nude obećavajući put ka izgradnji skalabilnih, privatnih i efikasnih AI sistema.

Očekuje se da će se razlika između Edge AI i Cloud AI pristupa dodatno smanjivati kako:

- lokalni hardver postaje snažniji
- tehnike optimizacije AI modela napreduju
- 5G i edge-cloud infrastruktura smanjuju latenciju i omogućavaju neprimetne prelaze
- decentralizovani okviri i okviri za AI koji čuvaju privatnost, kao što su federativno učenje i diferencijalna privatnost, dobijaju na značaju

Pored toga, inicijative zelenog AI promovišu energetski efikasne AI modele, čineći edge implementaciju ne samo praktičnom već i ekološki odgovornom.

Edge AI i Cloud AI se međusobno ne isključuju; oni imaju komplementarne uloge u AI ekosistemu. Cloud AI obezbeđuje tešku obradu potrebnu za obučavanje i implementacije velikog obima, dok Edge AI donosi brzinu, privatnost i lokalnu inteligenciju na nivo uređaja. Izbor između njih - ili projektovanje hibridnog rešenja - zavisi od zahteva aplikacije, kao što su latencija, osetljivost podataka, računarski zahtevi i dostupnost infrastrukture.

Kako AI nastavlja da se širi u svaki aspekt našeg života, razumevanje ove dve paradigme ključno je za programere, sistemske arhitekte i donosioce odluka koji žele da grade inteligentne, odzivne i odgovorne AI sisteme.

Poglavlje 3 • Etička razmatranja i buduće perspektive

Kako veštačka inteligencija i mašinsko učenje nastavljaju da napreduju, njihov uticaj na društvo postaje sve dublji. Pored tehničkih performansi, neophodno je razmotriti etičke posledice ovih tehnologija, uključujući pravičnost, transparentnost, odgovornost, privatnost i moguće društvene posledice. Ovo poglavlje razmatra te etičke izazove i istražuje kako se AI može razvijati i primenjivati na odgovoran način. Pored toga, ono se osvrće na nove trendove, buduće mogućnosti i mere potrebne da bi AI pozitivno doprineo društvu uz što manje rizike.

3.1. Etička i društvena pitanja

Kako se sistemi veštačke inteligencije i mašinskog učenja sve više integrišu u svakodnevni život, oni otvaraju važna etička i društvena pitanja. Ove tehnologije utiču na donošenje odluka u oblastima kao što su zdravstvo, bezbednost, finansije i zapošljavanje - često sa značajnim posledicama u stvarnom svetu. Pitanja kao što su privatnost podataka, algoritamska pristrasnost, transparentnost, odgovornost i moguće potiskivanje ljudskog rada ukazuju na potrebu za odgovornim razvojem i primenom. Glavna pitanja su:

- **Privatnost:** Obrada velikih količina ličnih podataka zahteva usklađenost sa propisima kao što je GDPR u Evropi
- **Pristrasnost i diskriminacija:** ML modeli mogu da prenesu i pojačaju pristrasnosti prisutne u podacima, što dovodi do nepravednih ishoda, na primer kod alata za zapošljavanje koji pokazuju pristrasnost prema ženama
- **Odgovornost:** Utvrđivanje odgovornosti za odluke koje donosi AI predstavlja izazov, posebno u kritičnim oblastima kao što su zdravstvo ili pravo, gde se i dalje vode rasprave o regulaciji
- **Uticaj na radna mesta:** Automatizacija može promeniti radne uloge i zahtevati prilagođavanje radne snage, što je tema široko zastupljena u literaturi o etici AI
- **Transparentnost:** Problem „crne kutije”, gde je odluke AI teško objasniti, predstavlja zabrinutost, naročito u zdravstvenim ili pravnim primenama

Ova pitanja naglašavaju potrebu za odgovornim razvojem AI, uz ravnotežu između inovacija i dobrobiti društva.

3.2. Perspektive

Veštačka inteligencija i mašinsko učenje razvijaju se neviđenom brzinom, transformišući gotovo svaki aspekt društva, industrije i svakodnevni život. Gledano unapred, AI obećava podsticanje inovacija u zdravstvu, autonomnim sistemima, obrazovanju, finansijama, upravljanju životnom sredinom i mnogim drugim domenima. Međutim, ove mogućnosti prate i izazovi, uključujući etička razmatranja, bezbednosne rizike i društvene uticaje. Ovo poglavlje razmatra nove trendove, moguće

primene i društvene posledice AI, pružajući pogled unapred na to kako ove tehnologije mogu oblikovati našu budućnost.

Sledeća tabela prikazuje neke od najvažnijih razvojnih pravaca.

Domen	Buduća perspektiva	Mogući uticaj	Izazovi / rizici
Zdravstvo	Personalizovana medicina zasnovana na AI	Terapije prilagođene pacijentu, rano otkrivanje bolesti, bolji ishodi lečenja	Pitanja privatnosti, algoritamska pristrasnost, oslanjanje na tačne podatke
Autonomni sistemi	Samovozeći automobili i dronovi	Bezbedniji transport, automatizacija logistike, manje ljudskih grešaka	Bezbednost u složenim okruženjima, pitanja odgovornosti
Obrazovanje	Učenje i tutorstvo uz pomoć AI	Personalizovane putanje učenja, prilagodljiv sadržaj, veće angažovanje	Privatnost podataka, nejednak pristup, preterano oslanjanje na AI u nastavi
Finansije	Prediktivna analitika i otkrivanje prevara	Pametnije investicione strategije, povećana bezbednost	Pristrasnost modela, problemi transparentnosti, rizik od automatizovanih grešaka
Životna sredina	Modelovanje klime i optimizacija resursa	Bolja predviđanja, efikasnije korišćenje energije	Nenamerne ekološke posledice
Industrija i proizvodnja	Pametne fabrike i prediktivno održavanje	Veća efikasnost, manje zastoja	Sajberbezbednosne pretnje, gubitak radnih mesta
Saradnja čoveka i mašine	Donošenje odluka prošireno AI podrškom	Veća kreativnost i sposobnosti rešavanja problema	Preterano oslanjanje na AI, pitanja odgovornosti, mogući gubitak veština
Edge i IoT	AI na ivici mreže	Inferencija u realnom vremenu, manja latencija, bolja privatnost i autonomija	Hardverska ograničenja, bezbednosne ranjivosti, potrošnja energije

Tabela 3.1: Najvažniji razvojni pravci, mogući uticaji i izazovi primene

Poglavlje 4 • Raspberry Pi AI HAT+

Raspberry Pi je sam po sebi popularan jednokartični računar za ugradne projekte; međutim, ima prilično ograničenu računarsku snagu za AI. To predstavlja poseban izazov: AI modeli, naročito neuronske mreže, zahtevaju visoke računarske performanse. CPU i GPU računara Raspberry Pi 5 nisu dovoljni za mnoge primene.

Rešenje ovog problema pruža Raspberry Pi AI HAT+. On sadrži specijalizovanu jedinicu za obradu neuronskih mreža (NPU), koja značajno ubrzava proračune mašinskog učenja uz malu potrošnju energije.

Neural Processing Unit (NPU), odnosno jedinica za obradu neuronskih mreža, specijalizovani je hardverski akcelerator projektovan za efikasnu obradu zadataka veštačke inteligencije i mašinskog učenja, posebno operacija koje se koriste u neuronskim mrežama.

AI hardverje optimizovan za stvarne Edge AI primene kao što su:

- Analiza videa u realnom vremenu
- Računarski vid
- Obrada govora
- Detekcija i praćenje objekata
- Prepoznavanje lica i analiza emocija
- Detekcija anomalija i defekata (npr. u proizvodnji ili kontroli kvaliteta)
- Prediktivno održavanje pomoću senzorskih podataka
- Autonomna navigacija (roboti, dronovi i vozila)
- Pametni nadzor i bezbednosno praćenje
- Prepoznavanje gestova i poze
- Razumevanje prirodnog jezika (za ugradne asistente ili glasovne interfejse)
- Analitika podataka i donošenje odluka na ivici mreže
- Praćenje životne sredine (otkrivanje dima, požara, vremenskih obrazaca itd.)
- Pametna maloprodaja i analitika kupaca (brojanje ljudi, toplotne mape, praćenje ponašanja)
- Zdravstvena dijagnostika i monitoring (npr. detekcija vitalnih znakova iz videosnimaka)
- Upravljanje saobraćajem i mobilnošću (brojanje vozila, detekcija zagušenja)
- Poljoprivredna automatizacija (praćenje zdravlja useva, praćenje stoke)
- Optimizacija energije (pametne mreže i sistemi upravljanja zgradama)
- i još mnogo toga

Raspberry Pi AI HAT+ je Edge AI rešenje zasnovano na Hailo-8 NPU čipu. Posebno je projektovano za programere koji žele da pokreću snažne AI aplikacije na kompaktnim sistemima kao što je Raspberry Pi.

4.1. Tehničke specifikacije Raspberry Pi AI HAT+

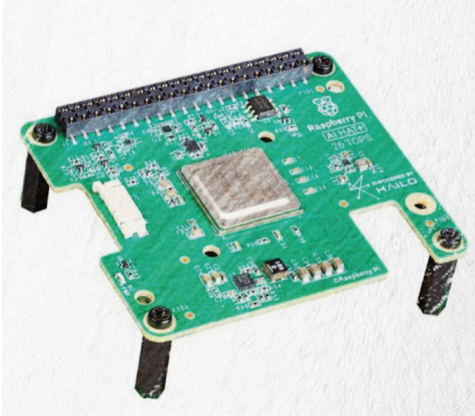
Raspberry Pi AI HAT+ ima sledeće karakteristike:

Hardver:

- Hailo-8 AI akcelerator (26 TOPS)
- Projektovan za SBC računare kao što su Raspberry Pi, NVIDIA Jetson i x86 PC računari
- Interfejsi: USB 3.0 / M.2 / PCIe
- Potrošnja energije: 2,5-3 W

Softver i razvojni alati:

- Hailo Dataflow Compiler: optimizuje AI modele za NPU
- Podržava TensorFlow, ONNX i PyTorch
- GStreamer integracija za video-AI tokove obrade
- Kompatibilno sa Raspberry Pi OS (Bookworm)



Slika 4.1: Raspberry Pi AI HAT+

4.2. Oblasti primene

Raspberry Pi AI HAT+ pogodan je za AI aplikacije na Raspberry Pi platformi, posebno za:

Računarski vid

- Prepoznavanje lica i detekcija objekata (npr. brojanje ljudi, sigurnosne kamere)
- OCR (optičko prepoznavanje znakova sa slika)
- Procena gestova i poze
- Analiza vozila, registarskih tablica i saobraćaja
- Kontrola kvaliteta i detekcija anomalija
- Segmentacija scene i praćenje okruženja

Edge AI i IoT

- Automatizacija pametne kuće pomoću AI
- Analiza senzora pomoću AI (temperatura, vibracije ili pokret)
- Prediktivno održavanje za IoT sisteme
- Upravljanje energijom i optimizacija
- Pametna poljoprivreda (praćenje useva i stoke)
- Senzorsko praćenje okruženja (kvalitet vazduha, dim ili detekcija poplave)

Industrija i robotika

- Kontrola kvaliteta u proizvodnji uz pomoć AI
- Autonomni roboti sa AI navigacijom i izbegavanjem prepreka
- Praćenje industrijske bezbednosti i detekcija radnika
- Automatizacija skladišta i sortiranje objekata
- Inspekcija, mapiranje i dostava pomoću dronova

Obrada govora i zvuka

- Glasovno upravljanje i prepoznavanje govora
- Pretvaranje teksta u govor i sinteza glasa
- Identifikacija i verifikacija govornika
- Detekcija akustičnih događaja (alarmi, kvarovi mašina, zvuci divljih životinja)

Pametni nadzor i bezbednost

- Detekcija i praćenje uljeza
- Prepoznavanje lica za kontrolu pristupa
- Praćenje vozila i registarskih tablica
- Analiza perimetra i gužve
- Sistemi upozoravanja u realnom vremenu za bezbednost i sprovođenje zakona

Pametni gradovi i transport

- Praćenje i optimizacija toka saobraćaja
- Upravljanje parkiranjem pomoću AI detekcije
- Javna bezbednost i detekcija anomalija u masi ljudi
- Pametno osvetljenje i raspodela resursa prema nivou aktivnosti

Zdravstvo i blagostanje

- Praćenje pacijenata pomoću slike ili senzorskih podataka
- Detekcija pada i analiza kretanja
- Detekcija vitalnih znakova (srčani ritam, obrasci disanja)
- Dijagnostika i detekcija zdravstvenih anomalija uz pomoć AI

Maloprodaja i analitika kupaca

- Prepoznavanje proizvoda i praćenje popunjenosti polica
- Brojanje kupaca i praćenje ponašanja
- Upravljanje redovima i predviđanje vremena čekanja
- Ciljano oglašavanje kroz demografsku analizu

Analitika podataka i donošenje odluka

- Klasifikacija događaja i detekcija anomalija u realnom vremenu
- Edge AI inferencija bez oslanjanja na Cloud AI
- Federativno učenje i distribuirana AI na ivici mreže

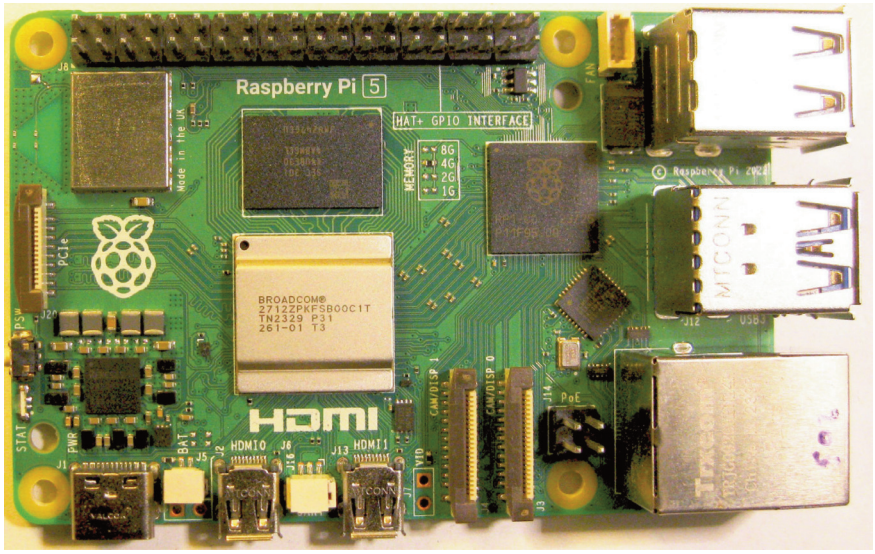
Obrazovanje i istraživanje

- Praktično učenje AI i ugradnih Sistema
- Izrada prototipova aplikacija za edge inferenciju
- Benčmark testiranje i merenje performansi AI modela

Naravno, u narednim poglavljima nije moguće detaljno obraditi sve primene; ipak, implementiran je niz zanimljivih primera koji ilustruju ključne tačke.

Poglavlje 5 • Puštanje Raspberry Pi računara u rad

Raspberry Pi 5 je najnovija verzija popularnog jednokartičnog računara, koja donosi bolje performanse i nove funkcije. Ovo poglavlje opisuje osnovno podešavanje Raspberry Pi 5, uključujući pripremu hardvera, instalaciju operativnog sistema i početnu konfiguraciju.



Slika 5.1: Raspberry Pi 5

5.1. Potrebne komponente

Za podešavanje Raspberry Pi 5 potrebne su sledeće komponente:

- Monitor sa HDMI ulazom
- USB tastatura i miš
- Mrežni kabl ili Wi-Fi pristup

Hardverski komplet

- Raspberry Pi 5
- Napajanje (USB-C konektor, najmanje 5 V / 3 A)
- MicroSD kartica (32 GB ili više)
- Raspberry Pi AI HAT+ (13 ili 26 TOPS)
- GPIO stacking header sa dugim pinovima
- Aktivni hladnjak
- Kamera sa kablom
- Mala probna ploča (breadboard)
- Crvena LED
- Žuta LED
- Zelena LED

- Plava LED
- Modul semafora
- 10 jumper („Dupont”) žica muško-žensko
- DHT22 senzorski modul
- Servomotor
- Micro-HDMI-na-HDMI kabl

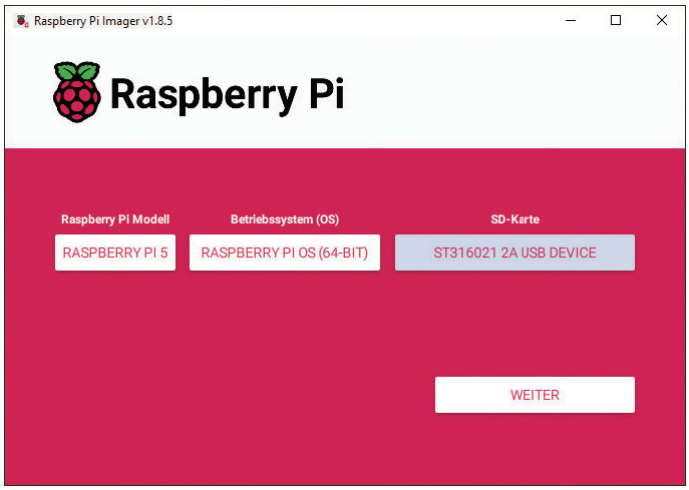
5.2. Instalacija Raspberry Pi OS

Instalacija Raspberry Pi OS je jednostavan postupak osmišljen tako da se uređaj brzo pripremi za rad. Zvanična i najčešće korišćena metoda koristi alat Raspberry Pi Imager, koji je dostupan za Windows, macOS i Linux.

Koraci za instalaciju operativnog sistema:

- **Preuzmite i instalirajte Raspberry Pi Imager:** Softver Imager može se preuzeti sa zvaničnog Raspberry Pi Veb-sajta. On pruža jednostavan interfejs za izbor i upis slike operativnog sistema na microSD karticu.
- **Pripremite microSD karticu:** Ubacite microSD karticu (preporučeno 16 GB ili veću) u računar pomoću čitača kartica. Imager će formatirati karticu i upisati sliku operativnog sistema na nju.
- **Izaberite operativni sistem:** Raspberry Pi Imager nudi više operativnih sistema, uključujući Raspberry Pi OS (ranije Raspbian) u različitim verzijama (sa desktop okruženjem, lite ili full). Izaberite odgovarajuću verziju prema potrebama.
- **Upišite sliku:** Nakon izbora operativnog sistema i ciljne microSD kartice, Imager će preuzeti (ako je potrebno) i upisati sliku na karticu. Ovaj proces obično traje nekoliko minuta.
- **Prvo pokretanje i konfiguracija:** Kada je kartica pripremljena, ubacuje se u Raspberry Pi. Pri prvom pokretanju Raspberry Pi OS pokreće čarobnjak za podešavanje jezika, tastature, Wi-Fi mreže i ažuriranje sistema.

Ovaj postupak obezbeđuje čistu i optimizovanu instalaciju najnovijeg Raspberry Pi OS, omogućavajući trenutnu upotrebu za različite projekte.



Slika 5.2: Raspberry Pi Imager



Slika 5.3: Raspberry Pi desktop posle prvog pokretanja

5.3. Osnovno podešavanje i konfiguracija

Nakon instalacije Raspberry Pi OS, sledeći korak je osnovno podešavanje i konfiguracija kako bi se uređaj pripremio za upotrebu. Pri prvom pokretanju Raspberry Pi će korisnika voditi kroz čarobnjak za podešavanje osnovnih postavki:

- **Jezik i region:** Izaberite željeni jezik, vremensku zonu i raspored tastature u skladu sa lokacijom korisnika.