# Peta generacija računara - šta se očekuje? Neuralne procesorske jedinice

Mladen Puzić, 18/2020

Prethodne godine definisane su naglim rastom popularnosti veštačke inteligencije. Skoro svaka oblast naših života je ili već promenjena njenim razvojem, ili se to očekuje u skorijoj budućnosti. Dostupnost ogromne količine podataka neophodnih za treniranje sistema mašinskog učenja i napredak u algoritmima doveli su do svetskih poznatih modela koji se svakodnevno koriste.

Ogromne investicije privatnog sektora, uključujući vodeće tehnološke kompanije poput Gugla, Amazona i Majkrosofta, dodatno su podstakle popularizaciju veštačke inteligencije. Ovo ulaganje dovodi do razvoja naprednih AI platformi, alata i usloga koje su postale pristupačne široj javnosti. Ovo dovodi do sve veće integracije veštačke inteligencije u svakodnevni život ljudi, kroz uključivanje u svaki aspekat njihove produktivnosti.

Ovakav rast je, pored napretka softvera i teorije podataka, zahtevao i značajan napredak na strani hardvera.

Neuralne procesorske jedinice (n*eural processing units - NPUs*), drugačije poznati i kao procesori dubokog učenja (*deep learning processors*) ili VI akcelaratori (*AI accelerators*) predstavljaju ključan napredak računarskih sistema za upotrebe u domenu mašinskog učenja. U pitanju su specijalizovani čipovi dizajnirani za ubrzanje operacija nad podacima potrebnim za svrhe veštačke inteligencije. Njihovom primenom i razvojem promeniće se način na koji koristimo računare.

Glavna korist neuralnih procesorskih jedinica jeste njihova mogućnost brzog izvršavanja složenih matematičkih operacija, čime se ubrzavaju performanse i smanjuje potrošnja energije. Posebno su efikasne obrade matrica i vektora, što ih čini idealnim za obradu podataka u neuralnim mrežama. Pored toga, sjajno su optimizovani za paralelan rad sa podacima, što omogućava rad u sistemima koji zahtevaju obradu u realnom vremenu, poput autonomnih vozila.

Potrošnja energije se takođe ne može zanemariti kao velika prednost. Za razliku od tradicionalnih procesora, koji mogu da se koriste u mnoštvu različitih situacija, specijalizacija procesora dubokog učenja podrazumeva manju potrošnju energije. Ovo doprinosi tome da se oni mogu koristiti u uređajima sa ograničenom baterijom ili napajanjem - mobilnim uređajima poput telefona i laptopova, kao i IoT (*Internet of Things*) sistemima.

Kako su nastajale specijalizovane procesorske jedinice (poznati i kao *koprocesori)* poput grafičkih i zvučnih kartica, a uz sve veću popularnost veštačke inteligencije, bilo je pitanje vremena kada će nastati procesorska jedinica specijalizovana za mašinsko učenje.

Prvi pokušaji ovoga nastali su još krajem osamdesetih godina prošlog veka, kada je Intel predstavio svoj *ETANN 80170NX* analogni čip za implementaciju neuralnih mreža. Ipak, zbog slabih performansi tadašnjih procesora, bilo je prerano za ovakvu tehnologiju.

Veću popularnost NPU-ovi su dostigli tek početkom 2010-tih, uz porast popularnosti veštačke inteligencije. *DianNao* procesor (*Električni mozak* na kineskom), specijalizovan za rad sa dubokim neuralnim mrežama, na površini od samo 3.02mm2 i snagom od 485mW, uspeo je da izvrši 452 Giga operacije u sekundi.

Mnoge kompanije su nakon toga krenule da prave svoje NPU-ove, bilo za komercijalne svrhe, bilo za svrhe sopstvenog razvoja. Neki od poznatijih su *Qualcomm Hexagon DSP, Samsung Neural Processing Solution, Apple Neural Engine, Huawei Da Vinci, Google Tensor Processing Unit.*

Većina komercijalnih NPU-ova dolazi integrisano uz centralnu procesorsku jedinicu, dok se industrijski uglavnom pojavljuju kao odvojena komponenta na matičnoj ploči.

Neuralne procesorske jedinice su prvi put dodate u mobilni telefon počevši od *Qualcomm Snapdragon 820* procesora, u 2015. godini. Pored NPU-ova, za zahtevnije operacije često se koriste i grafičke procesorske jedinice (GPU) (uglavnom za treniranje modela), dok je primena NPU-ova češće za inferenciju.

Jedna od ključnih primena NPU-ova jesu autonomni sistemi i robotika. Pošto su vrlo sposobni za brzu obradu podataka, omogućavaju autonomnim vozilima (automobilima, letelicama...), robotima i sl. da prepoznaju prepreke, planiraju putanje, reaguju na okolinu i donose odluke u realnom vremenu. Ovo sa sobom donosi primenu u mnoštvu industrija poput industrijskih robota, vojnih robota, autonomnim letelicama, dijagnoziranju u medicini, video pozivima, medicinskih robota.

Pored toga, koristan je i u drugim oblastima, poput prepoznavanja glasa, obradi prirodnih jezika, prevođenju između jezika, prepoznavanju lica.

Neuralne procesorske jedinice su prava revolucija u računarstvu - njihovo brzo i energetski efikasno izvršavanje operacija mašinskog učenja već dovodi do napretka u mnoštvu industrija i naučnih polja. Iako toga često nismo ni svesni, svako od nas ih već svakodnevno koristi kako bi optimizovali svoj dan i povećali produktivnost.

Uz napredak algoritama, optimizacija i hardvera možemo očekivati da se njihov uticaj širi i na nove oblasti i uređaje i da još puno čujemo o doprinosima koje procesori duboko učenja donose nama.