# DUBOKO UČENJE I UMJETNE NEURALNE MREŽE

U ovom poglavlju će se generalno upoznati s pojmovima koji su bitni za razumijevanje ovog diplomskog rada. Prvo će se objasniti što je to duboko učenje, što je motivacija dubokog učenja i gdje se ono danas primjenjuje. Nakon toga će se opisati što su tradicionalne ili standardne umjetne neuralne mreže kako bi se stvorila podloga za upoznavanje s konvolucijskim mrežama koje su glavna tema ovog rada. Potom će se dati kratki, općeniti pregled rada i strukture neuralnih mreža kako bi se opisalo kako točno podaci prolaze kroz mrežu. Detaljan način rada će se opisati u jednom od sljedećih poglavlja. Ovo poglavlje će nakratko usporediti te dvije vrste mreža i dati razlog zašto su se konvolucijske mreže uopće razvile i zašto se koriste.

## ŠTO JE DUBOKO UČENJE

Duboko učenje (eng. *deep learning*) je područje strojnog učenja koje 'uči' računala ono što ljudi rade po svojoj prirodi: uče na primjerima. Ono je jedna od ključnih tehnologija iza autonomnih vozila. Čini ih sposobnima prepoznati znak stop ili razlikovati pješaka od uličnog svjetla.

On uči računalni model uči ispravno klasificirati podatke iz vizualnih, tekstualnih ili audio podataka te nekada ti modeli imaju bolju izvedbu od ljudi. Ti modeli su trenirani korištenjem velikog skupa označenih podataka što je značajka nadziranog učenja (eng. *supervised learning*).

Duboko učenje se prvi put spominje 1980.-ih, ali samo u teoretskom smislu. Dva su razloga zašto je to tako:

1. Duboko učenje zahtijeva velike količine označenih podataka. Na primjer, razvoj autonomnog vozila zahtijeva milijune slika i tisuće sati videa.
2. Duboko učenje zahtijeva znatnu računalnu moć. Grafičke kartice visokih performansi s mogućnošću paralelnog izvođenja pokazale su se efikasnima u dubokom učenju. Kada se one udruže s računarstvom u oblaku (eng. *cloud computing*), znatno se smanjuje vrijeme potrebno za treniranje dubokih neuralnih mreža. Govori se o prvotnom vremenu od nekoliko tjedana do konačnog vremena od nekoliko sati ili manje.

Širok je spektar industrija u kojima se primjenjuje duboko učenje. Neke od tih industrija su:

1. industrija autonomnih vozila
2. zračna industrija
3. medicinska industrija
4. industrijska automatizacija
5. marketinška industrija
6. robotika [2]

## UVOD U UMJETNE NEURALNE MREŽE

Modeli dubokog učenja nazivaju se umjetne neuralne mreže (eng. *artificial neural networks*). Umjetna neuralna mreža je računalni model inspiriran strukturom i načinom rada biološke neuralne mreže (kao što je ljudski mozak). One se sastoje od velikog broja međusobno povezanih jednostavnih jedinica koje se zovu neuroni, čvorovi ili perceptroni (eng. *neurons, nodes, perceptrons*). [1][2]

Svaki neuron može donijeti jednostavnu odluku i proslijediti ju neuronima s kojima je povezan preko veze (eng. *connection*). Neuroni su organizirani u međusobno povezane slojeve (eng. *layers*). Više povezanih slojeva čini umjetnu neuralnu mrežu. Samo neuroni koji pripadaju različitim slojevima su međusobno povezani vezama.[1][5]

Neuralna mreža na tako može oponašati skoro bilo koju funkciju i tako praktički odgovoriti na bilo koje pitanje nakon što se istrenira na dovoljnom broju uzoraka za treniranje.[5]

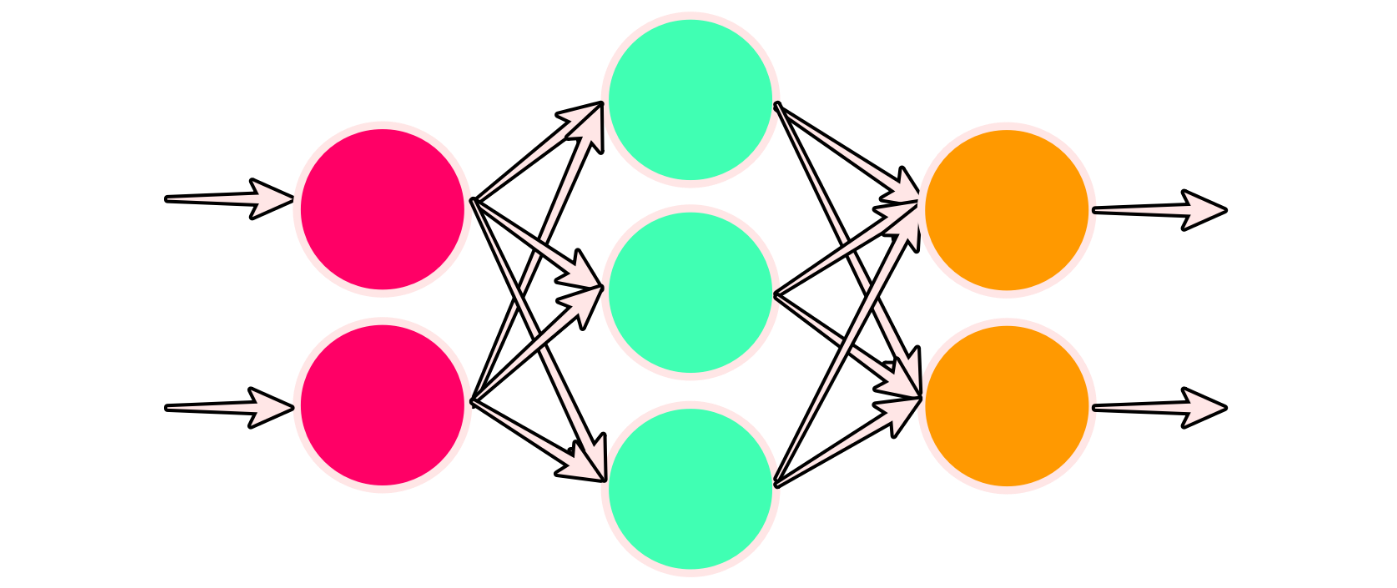
Umjetna neuralna mreža se u literaturi može naći i pod nazivima:

1. mreža (eng. *net*)
2. neuralna mreža (eng. *neural net*)
3. model

### SLOJEVI U NEURALNOJ MREŽI

'Plitka' neuralna mreža, kakva je prikazana na slici, sadrži samo tri sloja:

1. ulazni sloj (eng. *input layer*) koji prima neovisne varijable ili ulaze u model
2. jedan skriveni sloj (eng. *hidden layer*) koji se nalazi između ulaznog i izlaznog sloja
3. izlazni sloj (eng. *output layer*) koji generira predviđanja modela.[5]



Slika : prikaz plitke, potpuno povezane umjetne neuralne mreže

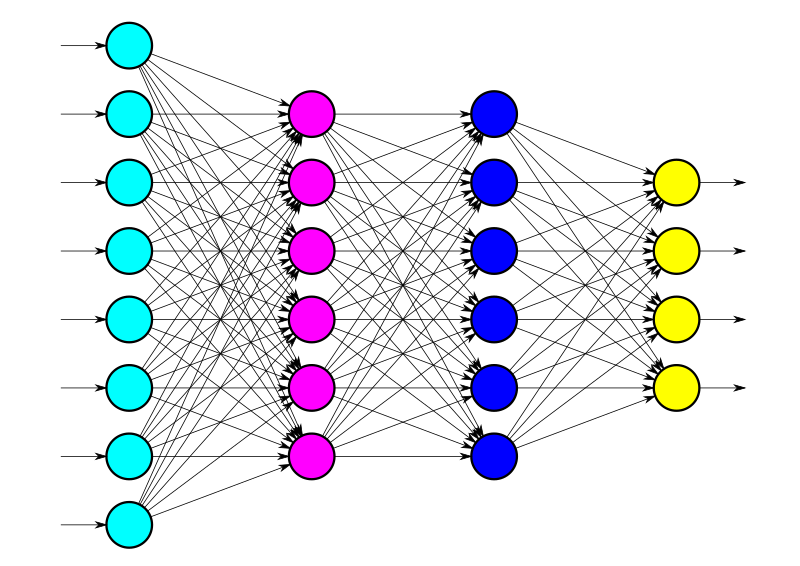
Različiti slojevi izvode različite vrste transformacija na vlastitim ulazima korištenjem aktivacijskih funkcija (eng. *activation function*). Podaci se prosljeđuju mreži preko ulaznog sloja, obično u obliku nekog višedimenzionalnog vektora. Broj čvorova u ulaznom sloju jednak je broju dimenzija ulaznog podatka. Kada bi se kao primjer uzela gore prikazana slika, broj dimenzija ulaznog podatka bi bio broj 2. Ulazni sloj te podatke, takvi kakvi jesu prosljeđuje skrivenim slojevima.[1][3]

Skriveni slojevi donose određene odluke na temelju podataka koje su dobili na svom ulazu od prethodnog sloja i na temelju aktivacijske funkcije koju kroiste. Te odluke prosljeđuju sljedećem sloju. Proces se ponavlja sve dok se ne dosegne izlazni sloj. Broj čvorova u skrivenom sloju je proizvoljan.[3]

Izlazni sloj izvodi konačnu operaciju kako bi prikazao predviđanja modela. Broj čvorova u izlaznom sloju jedan je broju klasa predviđanja (eng. *prediction classes*). Ako se kao primjer uzme gornja slika, broj klasa predviđanja, odnosno izlaznih čvorova, bi bio broj 2.[3]

Duboka neuralna mreža ima sličnu strukturu. Razlika je u tome što duboka neuralna mreža sadrži 2 ili više 'skrivenih' slojeva koji obrađuju ulazne podatke.[5]

Primjer takve mreže prikazan je na slici.



Slika : prikaz strukture potpuno povezane duboke umjetne neuralne mreže

Iako se i plitke neuralne mreže mogu nositi sa složenim problemima, duboke neuralne mreže daju ispravnija predviđanja i ta ispravnost raste kako im se dodaje više skrivenih slojeva. Kod tradicionalnih dubokih umjetnih neuralnih mreža najčešće je optimalno imati do 9 ili 10 skrivenih slojeva. Pokazalo se da dodavanjem više skrivenih slojeva od toga broja, preciznost mreže počinje opadati. Danas, duboke mreže imaju najčešće između 3 i 10 skrivenih slojeva.[5]

Obe slike prikazuju 'potpuno povezanu neuralnu mrežu' (eng. *fully-connected neural network*) s takozvanim 'gustim slojevima' (eng. *dense layers*). Taj naziv se odnosi na to da je neuron u određenom sloju povezan sa svakim neuronom iz susjednog sloja.

Postoje različite vrste slojeva u umjetnim neuralnim mrežama. Neke od vrsta su:

1. gusti (ili potpuno povezani) slojevi
2. konvolucijski slojevi (eng. *convolutional layers*)
3. slojevi udruživanja (eng. *pooling layers*)
4. slojevi povrata (eng. *recurent layers*)
5. slojevi za normalizaciju (eng. *normalization layers*)

Razlog, zbog kojeg postoji više vrsta skrivenih slojeva, je taj što različiti slojevi izvode različite transformacije na svojim ulaznim podacima te su stoga neki slojevi prikladniji za pojedini zadatak u odnosu na druge slojeve. [1]

### VEZE IZMEĐU SLOJEVA

Kao što je spomenuto u jednom od prethodnih poglavlja, neuroni, koji se nalaze u različitim slojevima, međusobno su povezani preko veza. Neuroni, koji pripadaju istom sloju, ne mogu biti međusobno povezani.

Svaka veza između dva neurona ima pridruženu težinu (eng. *weight*) koja je neka brojčana vrijednost. Svaka težina predstavlja snagu veze između dva čvora. Kada mreža na ulazu dobije neku ulaznu vrijednost (eng. *input value*), ona se dodjeljuje čvoru u ulaznom sloju. Nadalje, kada se ta ulazna vrijednost prosljeđuje određenim čvorovima u sljedećem, ovdje skrivenom, sloju, bit će pomnožena s težinom koja je pridružena određenoj vezi. Ta nova vrijednost, koja se prosljeđuje neuronu u sljedećem sloju, čini izlaznu vrijednost prethodnog neurona (eng. *output value*).[1]

Čvor, u ovdje skrivenom, sloju, na svoj ulaz prima više ulaznih vrijednosti iz više čvorova, u ovom primjeru ulaznog sloja. Konačnu će ulaznu vrijednost u određeni čvor činiti ponderirana suma (eng. *weighted sum*) koju čine umnošci određenih vrijednosti čvorova i njima pridruženih veza.

Tako, izlazna vrijednost iz danog čvora može biti izražena na sljedeći način:

Težine veza su inicijalizirane nasumičnim brojčanim vrijednostima i čine jedan od parametara koje mreža treba naučiti tijekom procesa učenja kako bi što ispravnije preslikavala dani ulaz u ispravnu klasu predviđanja.

(možda slika i primjer?)

### PROSLJEĐIVANJE PREMA NAPRIJED

Nastavljajući primjer iz prethodnog poglavlja, jednom kada se dobije izlazna vrijednost za dani čvor, taj dobiveni izlaz se pojedinačno množi svakom pojedinačnom težinom veze koja povezuje dani čvor sa čvorovima koji se nalaze u sljedećem sloju. Te nove vrijednosti se prenose kao ulazne vrijednosti prikladnim čvorovima.

Taj se opisani proces ponavlja sve dok se ne dostigne izlazni sloj s klasama predviđanja. Ako se za primjer uzme jedna od prije prikazanih slika, može se primijetiti da se proces ponavlja od ulaznog sloja prema izlaznom sloju, odnosno s lijeva na desno. Tako se cijeli proces prolaska podataka kroz umjetnu neuralnu mrežu naziva 'prosljeđivanje prema naprijed'. (eng. *forward pass*).

## KONVOLUCIJSKE NEURALNE MREŽE

Jedno od najvećih ograničenja tradicionalnih umjetnih mreža jest da se počinju mučiti pod računalnom složenošću koju zahtijeva obrada slika. Uobičajeni skupovi podataka za strojno učenje, kao što je MNIST skup podataka koji sadrži slike ručno pisanih znamenki dimenzija 28x28, prikladni su za većinu tipova umjetnih mreža. Neuron bi u prvom skrivenom sloju standardne umjetne mreže trebao imati prihvatljiv broj od 784 veze (28x28x1).[3]

Međutim, kada bi tradicionalna neuralna mreža trebala obraditi RGB sliku dimenzija 64x64, neuron u prvom skrivenom sloju bi imao 12,288 veza. Tradicionalna mreža u tom slučaju ne bi raspolagala dovoljnom računalnom moći i bila bi sklona *overfitting*-u.[3]

Drugim riječima, struktura potpuno povezane neuralne mreže pokazala se neefikasnom tamo gdje je potrebno analizirati slike (ili podatke) velikih dimenzija. Konvolucijska neuralna mreža koristi strukture kod kojih neuroni u jednom sloju ne komuniciraju sa svim neuronima u sljedećem sloju. Umjesto toga, svaki skup neurona analizira malo područje ili značajku slike. [4]

Pristup korištenja konvolucijskih mreža se temelji na pretpostavci da model može ispravno funkcionirati na temelju lokalnog razumijevanja slike. Konvolucijska mreža koristi manje parametara u odnosu na potpuno povezane mreže tako što ponovno iskorištava parametre i po nekoliko puta. Dok potpuno povezana neuralna mreža generira težinu veze za svaki piksel na slici, konvolucijska neuralna mreža generira tek toliko težina kako bi skenirala malo područje na slici u danom trenutku.[4]

Ovaj pristup je koristan tijekom procesa treniranja – što mreža ima manje parametara, to će imati bolje izvođenje i brže će učiti.[4]

Konvolucijske neuralne mreže (eng. *convolutional neural networks)* su primarno bile korištene kako bi se riješili zadaci raspoznavanja uzoraka na slikama. One uče detektirati različite značajke ulaznih podataka korištenjem desetaka ili čak stotina skrivenih slojeva. Svaki sloj u konvolucijskoj mreži povećava kompleksnost značajki koje model uči. Na primjer, skriveni sloj na početku mreže uči detektirati rubove i svjetlinu, dok posljednji sloj uči detektirati značajke koje su specifične za objekt kojeg model treba moći znati prepoznati. [2][3][5]

Konvolucijske mreže su po strukturi analogne standardnim umjetnim neuralnim mrežama na način da se sastoje od neurona, veza i slojeva. I dalje će neuron primiti neki ulaz, izvesti određenu operaciju i dobivenu vrijednost proslijediti dalje u mrežu.[3]

Ulazi u konvolucijski sloj se još mogu nazivati ulaznim kanalima (eng. *input channels*), a izlazi se mogu nazivati izlaznim kanalima (eng. *output channels*). [1]

Međutim, skriveni slojevi konvolucijske neuralne mreže izvode operacije koje izmjenjuju ulazne podatke s namjerom da model nauči značajke koje su specifične za pojedine podatke. Tri najčešće korištena sloja su:

1. Konvolucijski sloj: obrađuje ulaznu sliku skupom konvolucijskih filtera. Svaki od filtera aktivira određene značajke na slici.
2. Aktivacijski sloj: aktivira značajke koje se prosljeđuju sljedećem sloju u mreži.
3. Sloj za udruživanje (eng. *pooling*): pojednostavljuje izlaz tako što izvodi ne-linearno uzorkovanje i tako reducira broj parametara koje mreža treba naučiti, ali u isto vrijeme čuva najbitnije informacije.

Ove operacije se ponavljaju kroz desetke ili stotine slojeva. Svaki sloj uči identificirati različite značajke.[5]

Konvolucijske neuralne mreže su popularne radi tri glavna faktora:

1. eliminiraju potrebu manualnog izvlačenja značajki – značajke se uče direktno od strane konvolucijske mreže,
2. daju visoku kvalitetu raspoznavanja objekata,
3. mogu se ponovno istrenirati za nove zadatke prepoznavanja i tako omogućuju razvoj nove mreže na temelju već postojeće.[5]

Konvolucijske mreže su osobito korisne za pronalaženje uzoraka u slikama s ciljem prepoznavanja objekata, lica i krajolika. Također su izrazito djelotvorne i kod klasificiranja ne-vizualnih podataka kao što su audio podaci, podaci s vremenskim serijama ili signalima. [4][5]

## LITERATURA ZA OVO POGLAVLJE

[1] <https://deeplizard.com>

[2] <https://ch.mathworks.com/discovery/deep-learning.html>

[3] <https://arxiv.org/pdf/1511.08458.pdf>

[4] <https://missinglink.ai/guides/convolutional-neural-networks/convolutional-neural-networks-image-classification/>

[5] <https://missinglink.ai/guides/neural-network-concepts/complete-guide-artificial-neural-networks/>

[6] <https://ch.mathworks.com/discovery/convolutional-neural-network-matlab.html>

\*pronać ili napravit koju sliku za CNN

\*više slika?