Opis slika konvolucionim i rekurentnim neuralnim mrežama

Nikola Zeljković, Tamaš Tarjan
Departman za računarstvo i informatiku
Fakultet tehničkih nauka, Univerzitet u Novom Sadu
Novi Sad, Srbija

***Apstrakt*—Opisivanje slika je veoma koristan ali isto tako i veoma zahtevan proces. Zadatak ovog procesa je da za konkretnu sliku izgeneriše deskriptivnu smislenu rečenicu koja će semantički ispravno opisati zadatu sliku. Neke od najtežih prepreka u ovom procesu su dvosmislenost i veliki broj raznovrsnih slika. U prvom delu ovog rada dato je objašnjenje na koji način je moguće napraviti model za opisivanje slika upotrebom LSTM (eng. *Long short-term memory*) jedinica, koje su dale značajne rezultate u poslednjih par godina. LSTM predstavlja jednu od mogućih gradivnih jedinica rekurentnih mreža koja omogućava čuvanje vrednosti ili stanja u kratkim odnosno dugim vremenskim intervalima. Glavna prednost LSTM jedinica je ta što rešavaju problem smanjivanja nagiba funkcije korisnosti koja ocenjuje produktivnost date neuralne mreže (eng. *vanishing gradient problem*), međutim topologija ovakvih mreža je komplikovana. Takođe, predstavićemo značaj Faster R-CNN modela prilikom rešavanja ovog problema. U drugom delu rada rada ćemo predstaviti jednostavniju mrežu koja koristi konvolucione slojeve. Za treniranje i testiranje ovih modela koristićemo Microsoft COCO skup slika.**

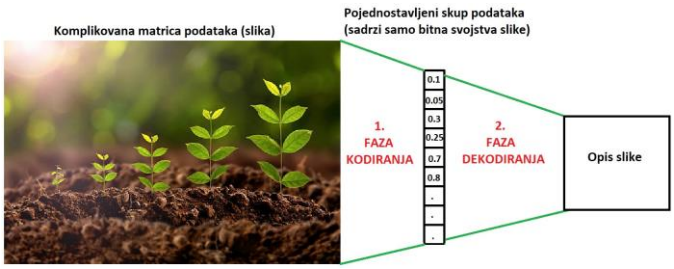
Ključne reči—neuralne mreže, prepoznavanje slika, prepoznavanje objekata, duboke konvolucione mreže, rekurentne mreže

I. Uvod

Ideja mašinske inteligencije potiče iz antikviteta i često se pojavljuje u starim fikcionalnim radovima. Alan Turingova teorija o računarstvu predstavlja ideju da jednostavan proces koji se zasniva na binarnim brojevima može da simulira bilo kakavu vrstu matematičke dedukcije. Ideja da digitalni računari mogu da simuliraju bilo kakav proces formalnog rezonovanja je takođe poznata kao *Church-Turing thesis*. S obzirom na ubrzan razvoj računarskih komponenti, principi mašinskog učenja su postali sve zastupljeniji u upotrebi pa je samim tim i proces opisivanja slika postao sve interesantniji za proučavanje.

Poslednjih nekoliko godina, duboke konvolucione mreže su postigle značajan uspeh u oblasti procesiranja slika, kao što su klasifikacija slika [1,2,3] i detekcija objekata [4,5,6]. Procesiranje slika takođe predstavlja osnovu prilikom mnogo kompleksnijih operacija kao što su generisanje priča [7] i generisanje kratkog opisa multimedijalnih sadržaja [8].

Bitan faktor prilikom procesiranja slika je naučiti mašinu da ima duboko razumevanje slike i scenarija koji se nalazi na slici a ne samo razumevanje pojedinačnih objekata na slici. U ovom radu predstavićemo rešenje za opisivanje slika. Prilikom opisivanja slika potrebno je odrediti koji objekti se nalaze na slici a nakon toga odrediti u kojim se oni relacijama nalaze kako bismo uspeli da stvorimo pravo razumevanje slike i uspešno opišemo sliku rečima koje sačinjavaju semantički smislenu rečenicu. U našem rešenju prikazanom na slici 1, data slika se prvo kodira u manji vektor podataka, ovim procesom zadržavamo samo bitne informacije sa date slike. Drugi deo pristupa je dekodiranje dobijenog vektora u semantički ispravan opis date slike. Srž ovog problema predstavlja kako izvršiti kodiranje početne slike, i kako iz dobijenog vektora generisati opis date slike. U prvom koraku koristi se model za detekciju objekata na slici u kombinaciji sa dubokom konvolucionom mrežom koja je zadužena za prepoznavanje odnosa detektovanih objekata na slici. Sve ove informacije se reprezentuju nizom vektora koji se koriste u fazi dekodovanja koja će izgenerisati semantički korektnu rečenicu.



Slika 1 Ilustracija prve i druge faze procesa. U prvoj fazi, slika se svodi na kraći niz podataka iz kojeg se u drugoj fazi generiše opis.

U poglavlju II predstavićemo bitne istorijske događaje koje su omogućili ubrzan razvoj veštačke inteligencije. U poglavlju III ćemo detaljno opisati jedan model koji rešava problem opisivanja slika. U poglavlju IV ćemo opisati detalje treniranja predstavljenog modela, nakon toga u poglavlju V ćemo predstaviti alternativni model koji koristi konvolucione mreže umesto rekurentnih mreža prilikom dekodiranja. U poglavlju VI opisaćemo bitne faktore prilikom testiranja neuralnih mreža. Na kraju ćemo u poglavlju VII opisati ograničenja ovih pristupa.

II. Istorija veštačke inteligencije

Veštačka inteligencija (AI, eng. *Artificial intelligence*) je inteligencija koju pokazuju mašine, u suprotnosti sa prirodnom inteligencijom koju pokazuju ljudi i druge životinje. U informatici i računarstvu istraživanje veštačke inteligencije definisano je kao izučavanje „inteligentnih

agenata“ koji predstavljaju bilo koji uređaj koji je svestan svog okruženja i sposoban da preduzme akcije koje će maksimizovati njegovu šansu da ostvari određene ciljeve. Kolokvijalno, termin veštačka inteligencija primenjuje se kada mašina oponaša „kognitivne“ funkcije koje ljude asociraju na ljudski um poput učenja i rešavanja problema.

Obim veštačke inteligencije je često sporan. Mašine postaju sve sposobnije i time se zadaci za koje se smatralo da iziskuju „inteligenciju“ izbacuju iz definicije veštačke inteligencije. Ovaj fenomen naziva se *AI* efekat i dovodi do toga da definicija veštačke inteligencije postaje „Veštačka inteligencija je ono što još uvek nije urađeno.“ Na primer, optičko prepoznavanje karaktera je često isključeno iz obima veštačke inteligencije zbog toga što je postalo rutinska tehnologija. Sposobnosti koje se klasifikuju kao veštačka inteligencija od 2017. godine su uspešno razumevanje ljudskog govora, takmičenje na najvišem nivou strateških igara (kao što su šah i *Go*), autonomni automobili, *intelligent routing in content delivery network* [16, 17], i vojne simulacije.

Veštačka inteligencija osnovana je kao akademska disciplina 1956. godine i do danas je prolazila kroz mnoge periode optimizma praćene razočaranjima i gubitkom finansijske podrške (poznatim kao „*AI winter*“), nakon čeka bi usledili novi pristupi i uspesi kao i vraćanje finansijske podrške. U većem delu svoje istorije istraživnje veštačke inteligencije je bilo podeljeno u polja koja često nisu međusobno sarađivala. Ova polja bazirana su na tehničkim razlikama, kao što su različiti ciljevi (na primer robotika ili mašinsko učenje), korišćenje različitih instrumenata (logika ili neuralne mreže) ili duboke filozofske razlike. Polja su takođe bazirana na socijalnim faktorima kao što su različite institucije ili rad različitih naučnika.

Tradicionalni problemi (ciljevi) istraživanja veštačke inteligencije uključuju rasuđivanje, reprezentovanje znanja, planiranje, učenje, procesuiranje prirodnog jezika, percepciju i mogućnost pomeranja ili manipulisanja objektima. Opšta inteligencija je među dugoročnim ciljevima ovog istraživanja. Pristupi u istraživanju uključuju statističke metode, računarsku inteligenciju, i tradicionalnu simboličnu veštačku inteligenciju. Mnoge alatke se koriste u istraživanju veštačke inteligencije, uključujući verzije pretraga i matematičke optimizacije, neuralne mreže i metode bazirane na statistici, verovatnoći i ekonomiji. Polje veštačke inteligencije potpomognuto je kompjuterskim naukama, matematikom, psihologijom, lingvistikom, filozofijom i drugim naukama.

Polje veštačke inteligencije bazirano je na tvrdnji da ljudska inteligencija može biti toliko precizno objašnjena da je moguće napraviti mašinu koja može da je simulira. Ova tvrdnja pokrenula je mnoge filozofske argumente o prirodi ljudsog uma, kao i etici kreiranja veštačkih bića koja će imati inteligenciju poput ljudske. Filozofski i etički problemi kreiranja takvog bića razmatrani su u mitovima, fikciji i filozofiji još od antičkih vremena. Neki ljudi smatraju da veštačka inteligencija predstavlja pretnju čovečanstvu ukoliko joj se dozvoli da se nesmetano razvija. Drugi smatraju da će veštačka inteligencija, u suprotnosti sa drugim aspektima razvoja tehnike, dovesti do masovne nezaposlenosti.

U 21. veku tehnike veštačke inteligencije doživele su oživljavanje koje prati paralelan razvoj mogućnosti računara, velike količine podataka i teorijskog razumevanja. Tehnike veštačke inteligencije postale su ključan deo tehnološke industrije tako što pomažu u rešavanju mnogih problema u računarskim naukama.

Tipična veštačka inteligencija je svesna svog okruženja i preduzima akcije koje će maksimizovati njenu šansu da uspešno postigne svoje ciljeve. Namenjen cilj veštačke inteligencije može biti jednostavan („1 ukoliko veštačka inteligencija pobedi u igri *Go*, u suprotnom 0“) ili kompleksan („Preduzmi akcije koje su matematički slične onima koje su dovele do nagrade u prošlosti“). Ciljevi mogu biti eksplicitno definisani, ili mogu biti indukovani. Ako je veštačka inteligencija programirana učenjem uslovljavanjem (eng. *reinforcement learing*) , ciljevi mogu biti implicitno indukovanim nagrađivanjem jedne vrste ponašanja, a kažnjavanjem druge. Alternativno, evolutivni sistem može indukovati ciljeve korišćenjem *fitness* funkcije za mutiranje i preferencijalnu replikaciju sistema koji ostvaruju visok rezultat. Ovo je slično tome kako su životinje evoluirale tako da imaju urođene ciljeve kao što je pronalaženje hrane, ili kako psi mogu biti uzgajani veštačkom selekcijom tako da imaju određene željene osobine. Nekim sistemima veštačke inteligencije, poput onih koji koriste algoritam najbližeg komšije (eng. *nearest-neighbor*) umesto rasuđivanja analogijom, nisu zadati ciljevi, osim do nivoa u kom su ciljevi implicitni u podacima kojima treniraju. Takvim sistemima cilj može biti da uspešno obave neki zadatak klasifikacije.

Veštačka inteligencija se česno zasniva na korišćenju algoritama. Algoritam predstavlja skup instrukcija koje računar može da izvrši. Kompleksiniji algoritmi se često grade od drugih jednostavnijih algoritama.

Mnogi od algoritama koje veštačka inteligencija koristi imaju sposobnost učenja iz podataka; oni mogu sami sebe da unaprede učenjem novih heuristika (strategija koje su radile dobro u prošlosti), ili mogu sami da pišu nove algoritme. Neki od ovih algoritama, uključujući Bajesove mreže (eng. *Bayesian network*), stabla odlučivanja, algoritam najbližeg komšije, bi teoretski mogli, ako bi im bili dostupni neograničeni podaci, vreme i memorija, da nauče da aproksimiraju bilo koju funkciju, uključujući i neku kombinaciju matematičkih funckija koji bi najbolje opisali ceo svet. Ovi algoritmi sposobni da uče bi stoga mogli, u teoriji, izvesti bilo koje moguće znanje razmatranjem svake moguće hipoteze i povezivanjem te hipoteze sa podacima. U praksi, gotovo nikad nije moguće razmatrati svaki ishod zbog toga što time vreme potrebno da se reši problem eksponencijalno raste (eng. *combinatorial explosion*). Veliki deo istraživanja veštačke inteligencije uključuje shvatanje kako identifikovati i izbeći razmatranje širokog spektra ishoda koji verovatno neće uroditi plodom. Na primer, kada gledamo mapu u potrazi za najkraćim putem od Denvera do Njujorka na istoku, u većini slučajeva nećemo gledati ni jedan put koji prolazi kroz San Francisko ili druge oblasti na zapadu, stoga, veštačka

inteligencija koja koristi algoritam za pronalaženje puta poput algoritma A* može da izbegne razmatranje svakog mogućeg puta koje bi dovelo do eksponencijalnog rasta vremena koje je potrabno za pronalaženje onog najbližeg.

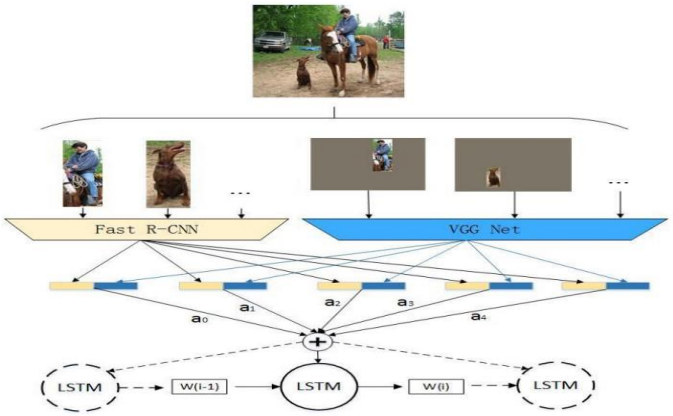
Najraniji (i najlakši za razumevanje) pristup veštačkoj inteligenciji je simbolizam (poput formalne logike): „Ako inače zdrava odrasla osoba ima temperaturu, onda ona možda ima grip“. Drugi, malo generalniji pristup je Bajesova logika [21]: „Ako trenutni pacijent ima temperaturu, prilagodite verovatnoću da on ima grip na određeni način“. Treći važan pristup, ekstremno popularan u rutinskim primenama veštačke inteligencije, je analogni, poput SVM-a (od eng. *Support vector machine*) i najbliži komšija [22]: „Posle pregleda iz zapisa o poznatim prethodnim pacijentima čija su temperatura, simptomi, godine i drugi faktori uglavnom isti kao kod trenutnog pacijenta vidi se da je X% tih pacijenata imalo grip.“ Četvrti pristup je teži za intuitivno razumevanje, ali je inspirisan time kako mozak funkcioniše: neuralne mreže koriste veštačke „neurone“ koji mogu da nauče tako što porede sami sebe sa željenim ishodom i menjaju snagu veza između svojih internih neurona da učvrste veze koje se čine korisnim [18]. Ova četiri glavna pristupa mogu se preplitati jedni s drugima, kao i sa evolutivnim sistemima; na primer, neuralne mreže mogu naučiti da prave zaključke, generalizuju, i prave analogije. Neki sistemi implicitno ili eksplicitno koriste više ovih pristupa, pored drugih algoritama veštačke inteligencije i onih koji to nisu. Najbolji pristup se često razlikuje u zavisnosti od problema [18, 19, 20].

III. Naš pristup

U ovom delu detaljno ćemo opisati topologiju svih mreža koje se koriste u modelu prikazanom u ovom radu. Ulazni podatak u model je slika S, dok je izlaz deskriptivna rečenica R koja sadrži određeni broj kodovanih reči: $S = \{w_1, w_2, \ldots, w_k\}$.

Deo modela koji je zadužen za kodiranje slike se sastoji od modela koji detektuje objekte i od duboke konvolutivne mreže koja za date objekte generiše vektore koje možemo posmatrati kao prostorne relacije datog objekta. Primer ovakve relacije može biti relacija između hrane i tanjira u rečenici “Na slici se nalazi vegetarijanska hrana u tanjiru.” Ove informacije su predstavljene numeričkim vektorima. Model za kodiranje generiše određeni broj anotacija, koje su predstavljene M-dimenzionalnim vektorima. Svaki od tih vektora sadrži po jednu kodiranu reprezentaciju objekata kao i lokaciju određenog objekta. Vektor lokacije sadrži sve informacije potrebne za određivanje odnosa datog objekta i nekog drugog objekta koji se nalazi na slici.

Mreža koja je zadužena za dekodiranje i generisanje semantički ispravnog opisa date kodirane slike se sastoji od duboke rekuretne neuralne mreže kao na slici 2.



Slika 2 Pregled opisanog modela. Deo za kodiranje prvo generiše informacije o pronađenim objektima (levi deo vektora) kao i informacije o njihovim prostornim lokacijama (desni deo vektora)

A. *Kodiranje slike*

Ovaj model je smišljen ugledajući se na proces kojim ljudi identifikuju sadržaj slika. Intuitivno je prvo identifikovati objekte koji se nalaze na slici i na osnovu njihovih lokacija odrediti međusobni odnos između datih objekata.

U proteklih nekoliko godina naučna oblast koja se bavi prepoznavanjem objekata u multimedijalnom sadržaju je značajno napredovala. Ideje koje su omogućile brojna napredovanja su metode koje služe za predlaganje mogućih regiona određenih slika, tako da verovatnoća da dati regioni sadrže objekat bude maksimalna (npr. [13]). Ovaj model koristi *Faster* R-CNN (*Region-based Convolutional Neural Networks*) [6]. *Faster* R-CNN je model koji se sastoji od dve komponente. Prva komponenta *Faster* R-CNN-a je duboka čista konvoluciona mreža za generisanje regiona koji verovatno sadrže objekat. Druga komponenta je R-CNN detektor koji za date regione proverava da li oni zaista sadrže objekat. Ceo sistem predstavlja jedinstvenu mrežu za detekciju objekata. Konvoluciona mreža koja generiše regione govori R-CNN-u gde da traži objekte.

Da bi generisali regione, autori rada [6] pomeraju malu neuralnu mrežu kroz dobijenu mapu svojstava, koja predstavlja izlaz poslednjeg sloja prve komponente *Faster* R-CNN mreže i sadrži niz brojeva koji je mreža izgenerisala. Ovaj pristup smanjuje složenost mape svojstava i kao rezultat daje vektore manjih dužina. Ovi vektori predstavljaju svojstva date slike. Ta svojstva se dalje šalju u dva odvojena potpuno povezana sloja (eng. *fully-connected layer*). Prvi sloj je *box regression layer* koji generiše regione, a drugi sloj je *box classification layer* koji ocenjuje verovatnoću da region sadrži objekat.

Nakon treniranja, ovaj model prima sliku kao ulaz, i daje listu mogućih regiona i listu verovatnoća da dati regioni sadrže objekat. Ove liste se sortiraju tako da verovatnoće budu u opadajućem redosledu, što znači da će regioni sa najvećom verovatnoćom biti prvi u listi. Svaki od datih regiona se prosledi u potpuno povezan sloj i sa tim se mapira na vektor svojstava. Tačnije, za svaku ulaznu sliku se detektuje N objekata i svaki od tih objekata se predstavlja kao D-dimenzionalni vektor.

B. *Lokalizacija objekata*

Ovaj deo modela služi za izvođenje kodiranih svojstava koje predstavljaju prostorne informacije datog objekta. Sa ovim svojstvima možemo odrediti relaciju između pronađenih objekata na slici. Prilikom detekcije objekata, znamo da je za svaku ulaznu sliku izlaz određeni broj četvorougaonih regiona sa verovatnoćom da dati region sadrži objekat. Za svaki objekat na slici pravimo novu sliku tako da sačuvamo prvobitne vrednosti datog regiona, a ostalim delovima slike postavljamo prosečnu vrednost skupa podataka za treniranje kao na slici 2. Ovim pristupom se dobija nova slika, koja sadrži samo jedan region sa jednim objektom. Ako smo detektovali N objekata, onda ćemo ovim pristupom dobiti N novih slika. Svaka nova slika se prosleđuje kao ulaz u VGG mrežu [2] koja generiše vektorsku reprezentaciju lokacije datog objekta pomoću jedne duboke konvolucione mreže. Ovim dobijamo još N vektora koji predstavljaju lokacije pronađenih objekata.

Nakon opisanog procesa rezultujući vektori se sastoje od dva dela. Prvi deo je vektor svojstava koji opisuju dati objekat dok je drugi deo vektor svojstava koji opisuju lokaciju datog objekta. Glavna ideja prikazanog pristupa je da se skup informacija date slike redukuje na kraći vektor bitnih svojstava i time pojednostavi proces generisanja opisa.

C. *Dekodiranje slike – generisanje opisa*

U ovom radu opisujemo metod dekodiranja koji se zasniva na LSTM mrežama sa mehanizmom pažnje (eng. *attention mechanism*). Mehanizam pažnje se prvi put koristio u neuralnim mrežama iz oblasti mašinskog prevođenja [12]. Koristeći principe ovog mehanizma, autori [13, 14, 11] su uveli korišćenje ovog procesa u oblast procesiranja slika, dok su autori [11] bili prvi koji su primenili ovaj pristup kod opisivanja slika. Ideja mehanizma pažnje je teorijski jednostavna. Zasniva se na tome da slike sadrže delove koji ne utiču na generisan semantički opis, zato u optimalnom slučaju mehanizam te delove slike potupuno ignoriše i obraća pažnju samo na bitne podregione. Vektore svojstava koji opisuju te podregione nazivamo anotacionim vektorima (eng. *annotation vectors*). Model prikazan u ovom radu prati ideju iz rada „*Neural image caption generation with visual attention*“[11] koja koristi dugotrajnu-kratkotrajnu memorijsku mrežu. Ova mreža u svakom koraku generiše po jednu reč, pri čemu je to generisanje uslovljeno kontekstnim vektorom $\boldsymbol{z_j}$ koji je generisan na osnovu svojstvenih vektora pronađenih objekata, prethodnim skrivenim stanjem $\boldsymbol{h_{j-1}}$ i predhodno generisanim rečima $\boldsymbol{w_{j-1}}$. Model se može opisati sledećim formulama:

$$In_j = \sigma(W_i E w_{j-1} + U_i h_{j-1} + Z_i z_j + b_i) \quad (1)$$

$$f_j = \sigma(W_f E w_{j-1} + U_f h_{j-1} + Z_f z_j + b_f) \quad (2)$$

$$c_j = f_j c_{j-1} + tanh(W_c E w_{j-1} + U_c h_{j-1} + Z_c z_j + b_c) \quad (3)$$

$$o_j = \sigma(W_o E w_{j-1} + U_o h_{j-1} + Z_o z_j + b_o) \quad (4)$$

$$h_j = o_j \, tanh(c_j) \quad (5)$$

$\boldsymbol{In_j}$ predstavlja stanje na ulazu dugotrajno-kratkotrajne memorijske mreže, odnosno stanje ulazne kapije (*input gate*); $\boldsymbol{f_j}$ predstavlja stanje u kom se briše memorija ćelije, odnosno stanje zaboravne kapije (forget gate); $\boldsymbol{c_j}$ predstavlja ćeliju, odnosno memoriju (*cell*); $\boldsymbol{o_j}$ je izlazna kapije (*output gate*) koja kontrološe šta će biti izlaz dugotrajno-kratkotrajne memorijske mreže; $\boldsymbol{h_j}$ predstavlja stanje skrivenog sloja mreže (eng. *hidden layer*). $\boldsymbol{W, U, Z}$ i $\boldsymbol{b}$ su “naučene” težinske i prednaponske matrice (eng. *weight and bias matrices*). $\boldsymbol{E}$ je matrica ugradnje (*embedding matrix*)[1] dok je $\boldsymbol{\sigma}$ sigmoid[2] aktivaciona funkcija, koja mapira rezultate množenja i sabiranja datih matrica u uniformnu raspodelu (-1, 1). Konekstni vektor $\boldsymbol{z_j}$ se generiše od anotacionih vektora koji predstavljaju svojstvene vektore od različitih objekata. Autori od [11] predstavljaju dva različita načina računanja konekstnog vektora $\boldsymbol{z_j}$. Ovaj model koristi „*soft*“ verziju računanja[3].

IV. Treniranje

U ovom delu ćemo analizirati primer treniranja predstavljenog modela. Prikupljanje dovoljnog broja podataka za treniranje neuralnih mreža je jedan od težih koraka u oblasti mašinskog učenja. Za model koji je sposoban da sa velikom preciznošću opiše veliki broj slika potrebno je sakupiti veliki broj raznovrsnih i unapred anotiranih slika. U Microsoftovom COCO[4] skupu podataka svaka slika je opisana već izgenerisanim vektorom svojstava $\{\boldsymbol{A_i}\}$ kao i izlaznim opisom slika, koji je predstavljen sekvencom reči $\{\boldsymbol{w_k}\}$. Parametri predstavljene mreže za kodiranje slike su unapred zadati. Zadatak procesa treniranja je pronalaženje težinskih parametara mreže, koja služi za dekodiranje kodirane slike. Funkcija koja ocenjuje rad dekodera je opisana jednačinom (6).

$$GREŠKA = -\sum_t log\,(p(w_j)) + \lambda \sum_i (1 - \sum_t a_{i,j})^2 \quad (6)$$

$\boldsymbol{w_j}$ je tekuća izlazna reč, dok $\boldsymbol{\lambda > 0}$ predstavlja faktor balansiranja između gubitka entropije i kazne na težinu pažnje. U ovom primeru treniranja za optimizaciju mreže se koristi stohastičko gradijentno spuštanje sa momentom 0.9. Funkcija $\boldsymbol{p}$ predstavlja verovatnoću tačnosti date reči $\boldsymbol{w_j}$. $\{\boldsymbol{a_{i,j}}\}$ predstavlja otežanu sumu vektora svojstva $\{\boldsymbol{Ai}\}$.

Primer ovakvog treniranja naveden je u radu [23].

V. Konvolucioni pristup

LSTM mreže se smatraju standardom za opis slika zbog njihove mogućnosti da zapamte dugoročne zavisnosti. Međutim, kompleksno adresiranje i mehanizam prepisivanja u kombinaciji sa sekvencijalnim procesiranjem, kao i velikim memorijskim zahtevima predstavlja izazov u toku treniranja. Takođe, zbog njihove sekvencijalnosti, one su zahtevnije za inženjering kada se razmatraju novi zahtevi. Do skoro konvolucione mreže nisu mogle da se pozitivno porede sa LSTM mrežama kada je opis slika u pitanju. Za razliku od LSTM mreža, konvolucione mreže su *feed-forward* (veze između jedinica ne formiraju ciklus, zbog toga se razlikuju od

---

[1] Matrica ugradnje (*embedding matrix)* predstavlja linearno mapiranje prostora u kom vrednosti nisu povezane u prostor u kom svaka vrednost ima smislenu vezu.

[2] Sigmoid funkcija je matematička funkcija čiji grafik ima karakteristike slova „S“, odnosno sigmoid krive.

[3] „*Soft*“ verzija računanja predstavlja deterministično računanje koje je moguće jednostavno trenirati pomoću standardnih algoritama za proširenje unazad.

[4] Microsoft COCO dataset: cocodataset.org

rekurentnih mreža). Prvu konvolucionu mrežu koja se pozitivno poredi sa LSTM mrežama u oblasti opisa slika konstruisali su autori [15]. Njihov model sastoji se od tri sloja: prvi i poslednji sloj predstavljaju mapiranja između reči i vektora, dok središnji sloj, za razliku od rekurentnih mreža, koristi paralelne maskirane konvolucije. U narednim paragrafima opisaćemo model ovakve mreže, njegovo treniranje i rezultate.

A. *Zaključivanje*

U radu [11] korišćena je jednostavna *feed-forward* mreža, f_w, za modelovanje verovatnoće reči $p_{i,w}(y_i|I)$. Predikcija reči y_i oslanja se na prethodne reči $y_{<i}$ ili njihovu reprezentaciju:

$$p_{i,w}(y_i|y_{<i}, I) = f_w(y_i, y_{<i}, I) \quad (6)$$

Kako bi konvolucionim operacijama bilo onemogućeno da koriste informacije narednih tokena, korišćeni su maskirani konvolucioni slojevi koji mogu da vrše operacije samo nad podacima „iz prošlosti“. Sada zaključivanje može biti rađeno sekvencijalno, reč po reč. Stoga, zaključivanje započinje početnim tokenom i koristi *feed-forward* prolaz da generiše verovatnoću za reč y_1: $p_{1,w}(y_1|0,I)$. Nakon toga se uzorkuje $y_1 \sim p_{1,w}(y_1|0,I)$. Nakon uzorkovanja y_1 se vraća u *feed-forward* mrežu da bi se generisale sledeće reči y_2, i tako dalje, sve dok se ne predvidi poslednji token ili do dolaska do fiksne gornje granice od N koraka.

B. *Treniranje – prednosti ovog modela*

Zbog toga što nema rekurentnih veza sve posmatrane reči su dostupne u svakom vremenskom koraku *i*, i model može biti treniran u paraleli za sve reči.

C. *Rezultati*

Konvolucioni pristup ostvaruje jednako dobre rezultate kao pristupi bazirani na LSTM kada je u pitanju opis slika. Performanse se mogu unaprediti korišćenjem pretrage bima (eng. *beam search*), odnosno pretrage koja prvo širi čvor koji najviše obećava. Konvolucioni pristup takođe rezultuje većom entropijom u distribuciji izlaznih verovatnoća, daje bolju tačnost predikcije, i ima manje gradijenata koji nestaju. Izbegavanjem sekvencijalnog procesuiranja koje se koristi kod rekurentnih mreža omogućeno je da u sličnom vremenskom periodu bude obavljeno treniranje sa 1,5x više parametara.

VI. Testiranje

Najboji način testiranja neuralnih mreža je automatizovana provera dobijenih rezultata datog modela, sa ulaznim podacima za koje imamo unapred definisane očekivane izlaze. Možemo podeliti skup podataka na dva podskupa. Prvi podskup se koristi prilikom treniranja mreže, dok se drugi podskup koristi za testiranje. U praksi se ovi skupovi dele tako da 20% podataka pripada skupu za testiranje, a 80% podataka za treniranje. Mera performanse modela se dobija od funkcije ocenjivanja (eng. *fitness function*[5] ili *loss function*[6]). Funkcija ocenjivanja upoređuje dobijene rezultate sa očekivanim rezultatima.

U slučaju da dobijemo loše rezultate postoji nekoliko mogućih razloga zbog kojih mreža ne radi na očekivani način. Prvi i veoma čest razlog su loše obeleženi podaci. Postoji mogućnost da se u skupu podataka nalaze određeni bitni podaci sa loše definisanim izlazima. Ako ovo nije slučaj, sledeći objekat razmatranja je funkcija ocenjivanja. Ako je ova funkcija loše definisana, tada će neuralna mreža da optimizuje svoje parametre tako, da loše definisana funkcija ocenjivanja daje dobre ocene. Količina podataka je takođe bitan faktor prilikom treniranja mreže. U zavisnosti od topologije mreže i željenih rezultata, višak ili manjak podataka ima tendenciju da utiče na uspešnost modela. Postoji mogućnost da skupovi podataka za treniranje i testiranje sadrže podatke sa bitnim svojstvenim razlikama, što dovodi do loših rezultata prilikom testiranja. Takođe je moguće da topologija mreže nije dobro rešenje za dat problem.

VII. Zaključak

U ovom radu smo ižučavali različite pristupe za mašinsko opisivanje slika, treniranje neuralnih mreža i testiranje istih. Navedeni radovi i njihovi autori su pokazali da konvolucioni i rekurentni modeli imaju veoma dobru sklonost u oblasti opisivanja slika.

Konvolucioni pristup ima jednostavniju topologiju i daje slične rezultate kao pristup koji koristi rekurentne mreže. Iako su navedeni pristupi vrlo dobri za usko definisan problem opisivanja slika, treba da uzmemo u obzir činjenicu da neke slike sadrže implicitne informacije koje se ne mogu prepoznati bez kontekstnog domena. Jedna osoba ima puno iskustvenih informacija koje mogu biti korisne prilikom prepoznavanja zašto je jedna slika interesantna, tužna ili smešna. Postavlja se pitanje kakav mehanizam je potreban da neuralne mreže iskoriste konekstualni domen znanja i da daju boje rezultate.

VIII. Literatura

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[3] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[4] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(1): 142-158.

[5] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[6] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

[7] T. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. B. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling. CoRR, abs/1604.03968, 2016.

[8] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In NAACL HLT, 2015.

[5] *Fitness function* je funkcija koja određuje koliko je dato rešenje blizu očekivanom

[6] *Loss function* je funkcija koja određuje cenu procesa.

[9] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154-171.

[10] Jin J, Fu K, Cui R, et al. Aligning where to see and what to tell: image caption with regionbased attention and scene factorization[J]. arXiv preprint arXiv:1506.06272, 2015.

[11] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International Conference on Machine Learning. 2015: 2048-2057.

[12] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[13] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]//Advances in neural information processing systems. 2014: 2204-2212.

[14] Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention[J]. arXiv preprint arXiv:1412.7755, 2014.

[15] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2014.

[16] Stutzbach, Daniel et al. (2005). "The scalability of swarming peer-to-peer content delivery".

[17] Nygren., E.; Sitaraman R. K.; Sun, J. (2010). "The Akamai Network: A Platform for High-Performance Internet Applications"

[18] Rosenblatt, Frank (1957), The Perceptron--a perceiving and recognizing automaton. Report 85-460-1, Cornell Aeronautical Laboratory.

[19] Aizerman, M. A.; Braverman, E. M.; Rozonoer, L. I. (1964). "Theoretical foundations of the potential function method in pattern recognition learning". Automation and Remote Control. 25: 821–837.

[20] Bishop, Christopher M. (2006). Pattern Recognition and Machine Learning. Springer.

[21] Gelman, Andrew; Carlin, John B.; Stern, Hal S.; Dunson, David B.;Vehtari, Aki; Rubin, Donald B. (2013). Bayesian Data Analysis, Third Edition. Chapman and Hall/CRC. ISBN 978-1-4398-4095-5.

[22] Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks".

[23] Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang. Image Captioning with Object Detection and Localization