Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/22BE20Cincurak

GENERISANJE TEKSTUALNOG OPISA SLIKE POMOĆU MAŠINSKOG UČENJA

IMAGE CAPTIONING USING MACHINE LEARNING

Ivan Činčurak, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Automatsko opisivanje slike je postalo atraktivna tema u poslednjih nekoliko godina. Postoji velika potreba mašinskog opisivanja situacija u automobilskoj industriji. Google Image pretraga bi takođe mogla biti poboljšana. Takođe, moguće bi bilo unaprediti nadzorne kamere uklanjanjem potrebe za postojanje osobe koja bi konstantno morala da nadgleda kamere i čeka da se određena situacija desi, umesto da pogleda samo kada je opis slike na videu približan nekom unapred definisanom skupu tekstova. U ovom radu isprobana su tri načina za automatsko generisanje opisa slike. Prvi je primenom enkoder-dekoder arhitekture sa mehanizmom pažnje, drugi je bez ovog mehanizma, dok je treći upotrebom rekurentnih neuronskih mreža. Rešenje je evaluirano metrikama BLeU, ROUGE i Doc2Vec. Modeli su trenirani i testirani na MSCOCO skupu podataka. Dodatno, model je testiran podacima scrape-ovanim sa Google Images pretrage.*

Ključne reči: *opisivanje slika, metrike za sličnost teksta, enkoder-dekoder arhitektura, mehanizam pažnje*

Abstract – *Image captioning has become an attractive topic in the last few years. There is a substantial need for machine description of situations in the automotive industry. Google Image search could be improved. Also, it would be possible to enhance surveillance cameras by eliminating the need for a person to constantly monitor the cameras and wait for a certain situation to occur instead of only looking when the description of the image in the video is close to some predefined set of texts. In this paper, three methods for automatic image description were tested. The first is by applying the encoder-decoder architecture with the attention mechanism, the second is without this mechanism, and the third is by using recurrent neural networks. The solution was evaluated with BLeU, ROUGE, and Doc2Vec metrics. The models were trained and tested on the MSCOCO dataset. Additionally, the model was tested with scraped data from a Google Image search.*

Keywords: *image captioning, text similarity metrics, encoder-decoder architecture, attention mechanism*

1. UVOD

Razvoj tehnologije svakodnevno donosi nove izazove. Svedoci smo neverovatnog razvoja automobilske industrije koja želi da što je više moguće olakša vožnju svojim korisnicima. Jedna od stvari koja pomaže u tome je automatsko opisivanje okoline u kojoj se nalazi auto. Ovo bi moglo biti od izuzetne važnosti za vozača koji zbog smanjene vidljivosti ne može tačno da proceni šta se nalazi ispred njega na putu. Isto tako bi automobil uz pomoć dobrog mehanizma za opisivanje slika mogao da ga upozori na opasnost ukoliko se nešto neočekivano desi ispred automobila.

U ovom radu isprobane su metode predložene u radovima koji će biti predstavljeni u nastavku i koji su rešavali problem automatskog generisanja opisa slike. Isprobana je enkoder-dekoder arhitektura sa i bez mehanizma pažnje [2, 3] i rekurentna neuronska mreža čiji ulaz predstavljaju informacije dobijene upotrebom konvolutivnih neuronskih mreža koje su prethodno trenirane.

U radu je eksperimentisano sa varijacijom ovih metoda. Varijacije uključuju korišćenje raznih modela koji su predstavljali enkoder. Za dekoder su korišćene LSTM i GRU arhitekture. Pored arhitektura, eksperimentisano je sa različitim načinima pretprocesiranja teksta.

Za potrebe treniranja je korišćen MSCOCO skup podataka [7]. Za svaku sliku korišćeno je pet različitih opisa. Radi poređenja sa referentnim radovima, modeli su testirani nad MSCOCO skupom podataka i pokazalo se da su rezultati dosta slični, u proseku pet procenta razlike.

Dodatno, model je testiran podacima *scrape*-ovanim sa *Google Images* pretrage. Kao što je očekivano, model je dao lošije rezultate nad ovim skupom nego nad MSCOCO. Ovakvo ponašanje je normalno jer se prilikom evaluacije nad MSCOCO skupom podataka za svaku sliku koristilo pet različitih opisa za istu sliku što je modelu davalo veći dijapazon mogućih načina za opisivanje slike.

Prilikom testiranje modela nad *scrape*-ovanim slikama postojao je samo jedan mogući opis, te je logično da je model davao lošije rezultate. Prilikom testiranja su korišćene BLeU metrika, kao i ROUGE i Doc2Vec uz kosinusnu sličnost vektora.

2. PREGLED STANJA U OBLASTI

U radu [1] opisano je rešenje problema automatskog generisanja opisa slike primenom mehanizma pažnje koji je ubačen u enkoder-dekoder arhitekturu. Enkoder prestavlja CNN koja sliku reprezentuje vektorom. Ovaj vektor predstavlja ulaz u dekoder koji predstavlja RNN. Prilikom treniranja korišćen je vokabular veličine 10000 reči. Pri evaluaciji rešenja [1] korišćen je MSCOCO,

---

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio prof. dr Aleksandar Kovačević.

Flickr30k i Flickr8k skup podataka. Evaluacija rezultata je vršena upotrebom BLeU i Meteor metrika. Rezultati koji su dobijeni su 67 i 20,30 nad Flickr8k, 66,9 i 18,49 nad Flick30k i 71,8 i 23,9 nad MSCOCO skupom podataka. Pored generisanja opisa slike, u ovom radu analizirano je koje tačno piksele model razmatrao prilikom formiranja opisa slike. Ovakva analiza je moguća zbog upotrebe mehanizma pažnje.

U radu [2] je prikazano još četiri rešenja problema automatskog generisanja opisa slike zasnovano na enkoder-dekoder arhitekturi sa i bez mehanizma pažnje, primenom *multimodal learning*-a kao i kompozicionih arhitektura. Autori su koristili skup podataka kao što je u prvom radu [1]. Ukupan skup podataka je podeljen tako da je 5000 slika korišćeno za testiranje i validaciju iz MSCOCO skupa, dok je po 1000 slika uzeto iz Flickr skupova. Korišćene metrike su: BLeU, Meteor, ROUGE-L i CIDEr. Zanimljivo je napomenuti da su rezultati dobijeni enkoder-dekoder arhitekturom sa mehanizmom pažnje identični kao u prvom radu [1]. Najbolje rešenje dalo je poslednje od četiri navedena rešenja. Dobijenu su sledeći rezultati: 0,72, 0,248, 0,53, 0,95 gde brojevi odgovaraju redom navedenim metrikama.

U radu [3] je rešavan problem automatskog generisanja slike bez mehanizma pažnje. Cilj je bio napraviti LSTM model koji prima sliku i niz reči i vrši predikciju reči koja ima najveću verovatnoću da se pojavi sledeća u rečenici. Za evaluaciju su korišćene BLeU, Meteor, CIDEr, SBU i Pascal VOC. Rezultati dobijeni u ovom radu su dosta slični onima dobijenim u prethodna dva [1, 2].

Rad [4] predstavlja rešavanje ovog problema upotrebom GAN mreže. Arhitektura se sastojala iz dva dela, generatora i evaluatora. Prvi od njih je služio kako bi se napravio opis slike, dok je drugi deo tu kako bi odredio koliko dobro generisana rečenica opisuje ulaznu sliku. Prilikom treniranja korišćeni su MSCOCO i Flickr30k skupovi podataka. Za evaluaciju su korišćene sledeće metrike: BLeU-3, BLeU-4, Meteor. ROUGE-L, CIDEr i SPICE kao i dve dodatne metrike E-GAN i E-NGAN koje su detaljno opisane u ovom radu. Dobijeni rezultati su redom: 0,39, 0,19, 0,24, 0,52, 1, 0,2, 0,52, 0,62.

U radu [5] pokušano je rešavanje problema upotrebom samo konvolutivnih neuronskih mreža bez dekodera. Kao ulaz model je primao feature iz slike upotrebom pretrenirane VGG16 mreže i vektor dobijen od rečenice. Model za pretvaranje reči u vektore je treniran od početka. Skup podataka koji je korišćen je MSCOCO a metrike koje su korišćene su BLeU, Meteor, ROUGE-L, CIDEr i SPICE dok su rezultati redom: 0,713, 0,247, 0,525, 0,912, 0172. Može se primetiti da su autori uspeli ovim pristupom da dobiju rezultate kao i u prethodnim radovima.

Radovi [1, 2, 3] predstavljaju osnovu za ovaj rad i daju uvid u to kakvim rezultatima treba da težimo ukoliko bismo želeli da automatizujemo proces opisivanja slike.

Ostali radovi [4, 5] nisu implementirani u ovom radu jer je akcenat bio na istraživanju mehanizma pažnje u enkoder-dekoder arhitekturama i koliko on zapravo unapređuje rešavanje problema kojim se bavi ovaj rad.

3. METODOLOGIJA

U ovom poglavlju opisani su implementacioni detalji rešenja za automatsko opisivanje slika. Na početku poglavlja je opisan skup podataka, kao i prikupljanje podataka za test skup. Potom sledi deo koji se odnosi na obradu podataka. Poglavlje se završava navođenjem arhitekture rešenja.

3.1. Skup podataka

Za rešavanje problema, prilikom treniranja korišćen je MSCOCO skup podataka. Sastoji se od 83000 slika za treniranje i 41000 za validaciju. Ukupna količina podataka je 19GB. Svaka slika ima 5 različitih opisa koji joj odgovaraju. Za dodatno testiranje modela je korišćen ručno napravljen skup podataka. U svrhu pravljenja skupa za testiranje koristila se tehnika *scrape*-ovanja. Sa sajta http://www.google.com/imghp su skidane slike po sledećem principu. Kod je uzimao rečenicu iz unapred definisanog skupa rečenica od 100 primeraka i išao na navedeni sajt. Nakon učitavanja rezultata skidana je prva slika koju je *Google* izbacio i to je skladišteno u direktorijum gde je naziv svake slike odgovarao rečenici koja je uneta u cilju pretrage na navedeni sajt.

3.2. Obrada podataka

U radu su isprobana dva različita pristupa obradi podataka pre treniranja. Prvi pristup podrazumeva uklanjanje znakova interpunkcije. Druga obrada je bila nešto složenija. Pored osnovne obrade korišćene u prvom pristupu, dodato je uklanjanje stop reči i svođenje reči na njihov koren.

Obrada podataka je vršena i nad slikama čiji se opis generisao. Svaka slika se najpre smanjivala na dimenzije 299×299, a zatim se vršila normalizacija slike tako da svaki piksel ima vrednost od -1 do 1.

3.3. Arhitektura rešenja

U cilju generisanja opisa slike korišćena su tri različita pristupa. Prvi pristup koji je isproban je korišćenje enkoder-dekoder arhitekture podržane mehanizmom pažnje. Prvi deo predstavlja CNN koja kao ulaz prima sliku, dok kao izlaz daje *feature* koji je dobijen na osnovu te slike. Drugi deo predstavlja RNN koja prima izlaz iz prethodnog sloja, odnosno vektor i daje na izlazu rečenicu koja opisuje sliku. Za CNN korišćeni su pretrenirani modeli kao što su: *InceptionV3*, *EfficientNet*, *Retina* i *VGG*. Ovi modeli su morali biti donekle izmenjeni u odnosu na originalne verzije. Potrebno je bilo skinuti poslednji sloj koji kao ulaz prima vektor, a kao izlaz vrši kvalifikaciju. Mreže nije bilo potrebno dodatno trenirati.

Kada su u pitanju RNN, u našem slučaju je korišćen model koji ima više ulaza i jedan izlaz (ulaz je sekvenca reči , a izlaz jedna reč). Za ovaj pristup je korišćen GRU i LSTM model. Rezultati ova dva modela su bili skoro identični, pa će zbog toga u daljem radu biti prikazani samo rezultati GRU modela koji se pokazao bolje.

U drugom pristupu pravljena je arhitektura koja ima CNN i RNN samo bez mehanizma pažnje. Cilj ovog pristupa je bila provera koliko stvarno ovaj mehanizam pomaže modelu u opisivanju slika.

Treći pristup je opisivanje slike upotrebom LSTM-a. Slika se kao i u prethodnim primerima pušta kroz CNN, koja nije deo celog modela. Ona nam služi samo kako bi dobili vektore od slika za potrebe treniranja. Ono što je zanimljivo na ulaz LSTM modela pored ovog vektora, dovodi se niz reči. Ako zamislimo početak predikcije, model će na ulaz dobiti vektor slike i reč koja predstavlja početak rečenice. U našem slučaju je to bila reč START.

Model vrši predikciju i kao izlaz daje reč koja je najverovatnija da se nađe nakon prosleđene reči. Sada se u naredni ulaz stavlja ponovo vektor slike ali se sada se stavlja i reč START i ona koja je dobijena od strane modela. Postupak se ponavlja dokle god ne dođemo do oznake koja predstavlja kraj rečenice, END, ili dok model ne izbaci broj reči koji smo odredili da je maksimalan mogući broj reči u opisu.

4. EKSPERIMENTI I REZULTATI

Prilikom evaluacije su korišćene sledeće metrike: BLeU, ROUGE-1, ROGUE-2, ROUGE-L, Doc2Vec + kosinusna sličnost.

4.1. BLeU metrika

U ovom radu je korišćen sledeći način računanja sličnosti uz pomoć BLeU metrike. Računate su metrike gde je n pripadalo skupu brojeva 1, 2, 3, 4 i zatim je uzimana srednja vrednost tih metrika. Ovaj način računanja preuzet je iz rada [4]. Naravno, n odgovara dužini sekvenci reči (n-gram).

Tabela 1. *Rezultati korišćenjem BLeU metrike*

Mehanizam pažnje	Obrada podataka	Model	Skor
DA	Osnovno	Inception + GRU	**0,44**
DA	Osnovno	EffNet + GRU	0,36
DA	Napredno	Inception + GRU	0,39
DA	Napredno	EffNet + GRU	0,3
NE	Osnovno	Inception + GRU	0,27
NE	Osnovno	EffNet + GRU	0,25
NE	Osnovno	Retina + GRU	0,26
NE	Osnovno	VGG + LSTM	0,4
NE	Osnovno	Inception + LSTM	0,4

Tabela 1. sastoji se od četiri kolone. Prva kolona govori o tome da li je u eksperimentu korišćen mehanizam pažnje ili ne. Druga kolona odgovara obradi podataka. Pod osnovnom obradom se smatra uklanjanje znakova interpunkcije dok se naprednom smatra uklanjanje stop reči i svođenje reči na njihov koren.

Treća kolona služi za opisivanje modela, odnosno kaže koji se model koristio za enkoder a koji za dekoder.

Četvrta predstavlja BLeU skor. Bitno je napomenuti da poslednja dva reda u tabeli odgovaraju eksperimentima gde pored slike u model ulazi i sekvenca reči.

Ono što se može primetiti je da napredna obrada podataka samo pogoršala rezultate. Još jedna zanimljiva činjenica koja se može shvatiti na osnovu ovih rezultata je da se model bez mehanizma pažnje najgore ponaša. Naravno, kao što je i očekivano, kombinacija Inception + RNN + mehanizam pažnje se pokazala kao najbolja kombinacija i postignu su rezultati slični kao u relevantnom radu.

4.2. Doc2Vec + kosinusna sličnost

Nedostatak BLeU i ROUGE metrika je da ne mogu verodostojno izračunati sličnost između rečenica ako su reči drugačije ili sinonimi ili ako je redosled reči drugačiji i time promenjen smisao rečenicu. Pristup predstavljen u radu [6] je najbliži realnoj proceni kvaliteta.

Rezultati dobijeni nad MSCOCO skupu je 84% što govori da je nad tim slikama model dosta uspešno radio i davao dosta dobre opise.

Tabela 2. *Rezultati nastali primenom Doc2Vec metrike*

Mehanizam pažnje	Obrada podataka	Model	Skor
DA	Osnovno	Inception + GRU	**0,84**
DA	Osnovno	EffNet + GRU	0,82
DA	Napredno	Inception + GRU	0,78
DA	Napredno	EffNet + GRU	0,72
NE	Osnovno	Inception + GRU	0,75
NE	Osnovno	EffNet + GRU	0,68
NE	Osnovno	Retina + GRU	0,65
NE	Osnovno	VGG + LSTM	0,83
NE	Osnovno	Inception + LSTM	0,8

U tabeli 2. su prikazani rezultati dobijeni istim eksperimentom kao u prethodnoj tabeli, ali primenom Doc2Vec metrike. Ono što je i očekivano je da je ova metrika dala mnogo bolje rezultate nego prethodne dve.

4.3. ROUGE metrika

ROUGE je još jedna poznata metrika za proveru kvaliteta mašinski proizvedenih rečenica [5]. Postoje razne varijacije ove metrike od kojih su u ovom radu korišćene ROUGE-1, ROUGE-2 i ROUGE-L. ROUGE-N, gde u našem slučaju N predstavlja jedna i dva, predstavlja preklapanje n-grama između prave rečenice i one koja je proizvod modela. ROUGE-L predstavlja izračunavanje najduže sekvence reči koja se nalazi u obe rečenice. Ono po čemu se ova metrika razlikuje od BLeU je to što je bazirana na *recall*-u a ne na *precision*-u.

Rezultati dobijeni nad MSCOCO skupom su 0,7, 0,35 i 0,5. Poslednji broj se odnosi na metriku koja je korišćena i u nekim navedenim radovim [2, 4, 5]. Rezultat koji smo dobili je samo za manje od par procenata lošiji nego rezultati iz tih radova.

U tabeli 3. prikazani su rezultati dobijeni istim eksperimentom kao u prethodne dve tabele, ali primenom tri prethodno navedenih metrika.

Tabela 3. *Rezultati nastali primenom ROUGE metrike*

Mehanizam pažnje	Obrada podataka	Model	ROUGE-1	ROUGE-2	ROUGE-l
DA	Osnovno	Inception + GRU	**0,62**	**0,26**	**0,58**
DA	Osnovno	EffNet + GRU	0,57	0,25	**0,58**
DA	Napredno	Inception + GRU	0,58	0,22	0,54
DA	Napredno	EffNet + GRU	0,51	0,2	0,51
NE	Osnovno	Inception + GRU	0,53	0,2	0,53
NE	Osnovno	EffNet + GRU	0,49	0,15	0,49
NE	Osnovno	Retina + GRU	0,5	0,18	0,48
NE	Osnovno	VGG + LSTM	0,55	0,23	0,55
NE	Osnovno	Inception + LSTM	0,51	0,2	0,51

Prva stvar koja se zaključuje iz tabele 3. je ta da je međusobno poređenje modela paralelno donekle poređenju BLeU metrikom. Vidimo da se i ovde može doći do istog zaključka koji se model najbolje a koji nagore ponašao. Zanimljiva stvar koja se može primetiti je da ROUGE-2 metrika davala dosta niske rezultate, što znači da se sekvence reči dužine dva iz tačnih rečenica nisu pojavljivale mnogo u kandidatskim rečenicama. Na osnovu toga se može zaključiti da bi se za ROUGE-3, 4 i svaki sledeći dobila još manja tačnost.

5. ZAKLJUČAK

U ovom radu su opisana tri načina rešavanja *Image captioning* problema. Ispostavilo se da je pristup sa enkoder-dekoder arhitekturom podržanom mehanizmom pažnje dao najbolje rezultate. Prikazani su različiti načini pripreme podataka za treniranje modela kao i različiti modeli koji su se trenirali. Iz svega navedenog u radu je jasno da je ovaj problem rešiv i da je moguće istrenirati modele koji će davati opise koji stvarno odgovaraju slikama. Možemo primetiti da je u ovom radu dostignuta tačnost od 84% iako model nije treniran na više od 20 epoha, što daje osnovu za dalje istraživanje.

U cilju popravljanja rezultata moguće se fokusirati na neku određenu oblast odnosno kategoriju slika. Recimo mogao bi da se odvoji takav skup podataka koji bi sadržao samo slike ljudi i njihovih aktivnosti. Ovakav pristup bi sigurno doveo do značajnog povećanja tačnosti modela, ali bi isto tako izgubio na svojoj generičnosti. Naravno, pored svega navedenog moguće je pokušati sa drugačijim arhitekturama koje mogu rešiti ovaj problem.

Kao dalje razvijanje moguće je ugraditi ovaj model u softver koji prati dešavanja koja se dešavaju na nekom snimku tako što bi vršio opisivanje *frame*-a na svakih nekoliko sekundi. Jasno je da su mogućnosti neograničene.

6. LITERATURA

[1] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. PMLR, 2015.

[2] Bai, S., & An, S. (2018). A survey on automatic image caption generation. Neurocomputing, 311, 291-304

[3] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2016). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence, 39(4), 652-663.

[4] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

[5] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318)

[6] Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. Information Sciences, 307, 39-52

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra-manan, P. Doll ar, and C. L. Zitnick. Microsoft coco: Com-mon objects in context. InEuropean Conference on Computer Vision, pages 740–755. Springer, 2014.

Kratka biografija:

Ivan Činčurak rođen je 8. januara 1998. godine u Novom Sadu, Srbija. Fakultet tehničkih nauka upisao je 2016. godine na studijskom programu Računarstvo i automatika. Diplomski rad odbranio je 2020. godine. Master akademske studije upisao je iste godine na usmerenju Inteligentni sistemi. Kontakt: ivan.cincurak98@gmail.com