Generisanje opisa slike

Jelena Vlajkov, Ilija Brdar
Univerzitet u Novom Sadu
Fakultet tehničkih nauka
Novi Sad, Srbija
{jelena.vlajkov, brdar.ilija}@uns.ac.rs

***Apstrakt*—Zbog sve veće prisutnosti vizuelnog sadržaja u mnogim sferama savremenog života, pred računare se nameće zadatak opisa tih sadržaja kako bi se automatizovale razne aktivnosti. Primena dubokog mašinskog učenja na ovaj problem beleži najbolje rezultate. U ovom radu primenjene su dve enkoder-dekoder arhitekture dubokog učenja. Prvi eksperiment sproveden je koristeći InceptionV3 zajedno sa LSTM, a drugi koristeći EfficientNet sa transformer modelom. Za evaluaciju oba modela korišćena je BLEU sličnost, a za trening i samu evaluaciju korišćen je Flickr8k skup podataka. Dobijeni rezultati svedoče da obe arhitekture daju zadovoljavajuće rezultate, tj. generišu gramatički i semantički smislene rečenice. Arhitektura koja bolje rešava ovaj problem jeste EfficientNet u kombinaciji sa transformer modelom.**

Ključne reči—opis slike, transformeri, konvolutivne mreže, rekurentne mreže, enkoder, dekoder

I. Uvod

Vizuelni sadržaj je postao neizostavan deo svakodnevice svakog čoveka. Slike su sastavni deo dokumenata, interneta, televizijskog programa i ostalih medija sa kojima se čovek često susreće. Čovek ima urođenu sposobnost da protumači i opiše sliku koju vidi. Međutim, zbog sve veće prisutnosti takvog sadržaja, ovaj zadatak se stavlja pred računare. Od računara se očekuje da obezbedi generisanje opisa vizuelnog sadržaja koji bi se kasnije mogao upotrebiti za optimizaciju i ubrzavanje raznih procesa.

Jedna od primena ove tehnike jeste dobavljanje slika na osnovu njihovog sadržaja (eng. *Content-Based Image Retrieval*). Ova primena može drastično ubrzati pretragu lokalne baze slika ili čak pretragu celog interneta [1]. Ubrzanje je primetno u poređenju sa pređašnjim sistemima dobavljanja na osnovu sadržaja, a naročito u poređenju sa manuelnim radom ljudskog eksperta. Društvene mreže nalaze novu primenu ove tehnike u generisanju opisa korisničke objave [1]. Opisi mnogu da uključe lokacije i dešavanja koji su prikazani na slici. Postoje i humanije primene kao što je pomoć ljudima sa oštećenim vidom u razumevanju digitalnog vizuelnog sadržaja.

Zadatak generisanja opisa slike (eng. *Image captioning*) pored prepoznavanja objekata koji se nalaze na slici obuhavata i prepoznavanje odnosa između tih objekata i krajnju formulaciju opisa izraženu prirodnim jezikom. Zbog toga se ovaj zadatak ne smatra lakim i uvek postoji mesto za poboljšanje u odnosu na postojeće tehnike [1].

Tehnika koja trenutno daje najbolje rezultate jeste upotreba enkoder-dekoder arhitekture [1]. U ulozi enkodera najčešće se nalazi konvolutivna neuronska mreža (CNN) koja je zadužena za izvlačenje vektora karakteristika (eng. *feature vector*) iz slike. Dekoder najčešće predstavlja rekurentna neuronska mreža (RNN) čiji je zadatak da na osnovu vektora karakteristika slike generiše njen tekstualni opis.

Skorašnji radovi kao enkodere koriste brze R-CNN mreže, a dekodere koriste kompleksnije transformer modele i postižu *state-of-the-art* performanse. Ključ uspešnosti transformer modela leži u *attention* mehanizmu koji određenim svojstvima slike daje veći značaj [2].

U ovom radu korišćena je popularna enkoder-dekoder arhitektura, a eksperimenti su sprovođeni upotrebom različitih modela. Sprovedena su dva eksperimenta gde su korišćeni:

1. *Inception Net* u ulozi enkodera i LSTM u ulozi dekodera
2. *Efficient Net* u ulozi enkodera i transformer model u ulozi dekodera

Korišćeni skup podataka je javno dostupan pod nazivom Flickr8k [3]. Sadrži osam hiljada slika različitih pojava i scena, kako bi se razvio diverzitet situacija koje treba opisati. Svaka slika anotirana je sa pet različitih opisa.

Za evaluaciju performansi modela korišćena je BLEU mera sličnosti. Za svaku sliku iz testnog skupa podataka računata je sličnosti generisanog opisa sa opisima iz skupa podataka. Prosek izračunatih mera sličnosti je uzet za meru performansi modela.

Dobijeni rezultati eksperimenata svedoče da obe arhitekture generišu smislene i gramatički ispravne opise slika. Bolje rezultate ostvaruje arhitektura sačinjena od EfficientNet i transformer modela (57% BLEU mera). Ona je nadmašila arhitekturu InceptionV3 + LSTM za 4% BLEU mere. Zaključak koji se nameće na osnovu dobijenih rezultata jeste da se obe isprobane arhitekture zaista mogu koristiti u rešavanju problema generisanja opisa slike, ali da svakako ima prostora za njihovo poboljšanje.

Ostatak ovog rada je struktuiran na sledeći način: Sekcija II sadrži pregled prethodno objavljene literature na ovu temu, kao i teoretske osnove korišćenih modela. Sekcija III sadrži opis metodologije sprovođenja eksperimenata. Sekcija IV izlaže dobijene razultate, dok sekcija V sadrži zaključke izvedene iz pomenutih rezultata.

II. Prethnodni radovi

U ovoj sekciji biće izloženi prethodni radove i poznate metodologije koje se bave problemom generisanja opisa slika.

A. Metode generisanja teksta

Hossain i ostali [1] dele metode davanja opisa slici u tri grupe: opisivanje na osnovu šablona (eng. *Template-based image captioning*), na osnovu prethodnih opisa (eng. *Retrival-based image captioning*) i generisanje novih opisa (eng. *Novel caption generation*).

Metode zasnovane na šablonima podrazumevaju postojanje fiksiranih šablona opisa sa praznim mestima koje je potrebno popuniti. Zadatak modela je da na slikama detektuje različite objekte, atribute i njihove veze, a zatim da popuni prazna mesta u šablonu adekvatnim rečima. Ova metoda može da izgeneriše gramatički ispravne rečenice, ali svaka ima predefinisanu dužinu, što smanjuje kvalitet opisa.

Metode zanovane na prethodnim opisima novoj slici dodeljuju opis iz skupa već postojećih opisa. Podrazumeva se postojanje slika sa već dodeljenim opisima. Kada je potrebno pronaći opis za novu sliku, pronalaze se vizuelno najsličnije slike sa svojim opisima. Rezultujući opis se bira iz ovog skupa (eng. *caption pool*). Ova metoda generiše gramatički korektne opise, ali su oni često vrlo generički i semantički siromašni.

Metode generisanja novih opisa slike koriste duboko duboko učenje za analiziranje vizuelnog sadržaja, a zatim i generisanja novih opisa. Za analizu slike se često koristi konvolutivna neuronska mreža, a za generisanje novog sadržaja jezički model (eng. *language model*). Na ovaj način dobijaju se semantički bogati opisi. Zbog toga će ova metoda biti u fokusu ovog rada.

Novije metode generisanja opisa obuhvataju upotrebu učenja uslovljavanjem (eng. *Reinforcement Learning*) kao i upotrebu GAN modela. Iako ovi modeli mogu proizvesti semantički bogate i gramatički ispravne opise, i dalje imaju velike limitacije na čijem se prevazilaženju aktivno radi.

B. *Generisanje opisa dubokim učenjem*

Za generisanje novih opisa slike upotrebom dubokog učenja najčešće se koriste dva pristupa: standardna enkoder-dekoder arhitektura i kompozitna arhitektura.

U enkoder-dekoder arhitekturi, ulogu enkodera preuzima konvolutivna neuronska mreža koja ima zadatak ekstrakcije *feature* vektora iz slike. U *feature* vektoru treba da budu sadržana najvažnija svojstva slike: objekti, atributi i veze. U ulozi dekodera nalazi se LSTM koji prima *feature* vektor slike i generiše tekstualni opis. Kako se *feature* vektor slike dovodi na ulaz LSTM mreže samo u prvom koraku, kod generisanja dugačkih opisa moguće je da dođe do problema nestajućeg gradijenta (eng. *vanishing gradient*). Neki pokušaji rešavanja ovog problema obuhvataju upotrebu gLSTM mreža (eng. *guided* LSTM). Neka jednostavnija rešenja obuhvataju kontrolisanje dužine opisa, kako bi se preventirao problem dugačkih opisa. Slika 1 prikazuje opisanu enkoder-dekoder arhitekturu.

Slika 1 *Grafički prikaz enkoder-dekoder arhitekture*

Kompozitne arhitekture obuhvataju CNN za ekstrakciju *feature* vektora iz slike, jezičke modele za generisanje više opisa na osnovu *feature* vektora i modele sličnosti (eng. *similarity model*) koji selektuju najkvalitetnije generisane opise. Ovaj metod beleži značajna poboljšanja u generisanju opisa u odnosu na prethodne arhitekture.

Opisana enkoder-dekoder arhitektura i njene validacije nisu u mogućnosti da se fokusiraju na prostorne aspekte slike koji su važni za generisanje opisa. Nasuprot tome, oni generišu opis posmatrajući sliku kao celinu. Modeli zanovani na pažnji (eng. *Attention-based model*) prevazilaze ovaj problem tako što se fokusiraju na različite delove slike za vreme generisanja opisa. Tipična arhitektura ovih modela sastoji se od CNN-a u ulozi ekstraktora *feature* vektora i jezičkih modela za generisanje opisa. U svakom koraku faze generisanja teksta (eng. *time step*) fokusiraju se različiti regioni slike i opisi se ažuriraju. Nakon poslednjeg koraka jezičkog modela, dobija se finalni opis slike.

Kako je upotreba transformer modela sve dominantnija u oblasti procesiranja prirodnog jezika, postoji tendencija da se upotrebljavaju i u drugim sferama. Tako u [2] autori predstavljaju arhitekturu koja se sastoji od RCNN mreže za ekstrakciju *feature* vektora iz slike i transformer enkoder i dekoder blokova koji su zaduženi za generisanje opisa. Kako ova arhitektura postiže *state-of-the-art* performanse, ona je usvojena i u ovom radu.

C. *Evaluacione metrike*

Postoji veliki broj evaluacionih metrika koje se koriste za merenje kvaliteta generisanog opisa. U nastavku su navedene najpopularnije.

BLEU (eng. *Bilingual evaluation understudy*) [4] je jedna od prvih metrika koja se koristila za merenje kvaliteta mašinski prevedenog teksta i bliska je ljudskom rasuđivanju. Individualni segmenti generisanog teksta se porede sa referentnim tekstom i za svaki segment se računa mera sličnosti. Ukupan kvalitet generisanog teksta se dobija tako što se mere za svaki segment teksta uproseče. Mane BLEU metrike su te što ne obraća pažnju na sintaksnu korektnost teksta i radi dobro ukoliko je generisani tekst kratak.

METEOR (eng. *Metric for Evaluation of Translation with Explicit ORdering*) [5] poredi sličnost segmenata generisane rečenice sa segmentima referentne rečenice. Uzima u obzir i sinonime.

ROUGE (eng. *Recall-Oriented Understudy for Gisting Evaluation*) [6] poredi sekvence reči, parove reči i n-grame generisanog teksta i referentnog teksta. Postoje različite varijacije ove metrike za rad sa specifičnim vrstama teksta.

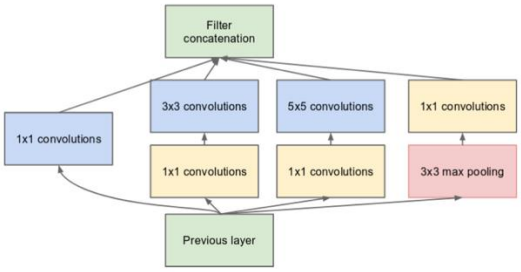
III. METODOLOGIJA

U ovom segmentu rada biće opisani eksperimenti koji su izvedeni u sklopu ovog rada. Za podelu skupova korišćena je podela od 80/20, 80% slika za treniranje i 20% slika za testiranje.

Prvi od eksperimenata jeste enkoder-dekoder arhitektura preko InceptionV3 mreže u ulozi enkodera i LSTM u ulozi dekodera, dok je drugi eksperiment izveden preko EfficientNet neuronske mreže u ulozi enkodera, i transformer arhitekture u ulozi dekodera.

A. *InceptionV3 i LSTM arhitektura*

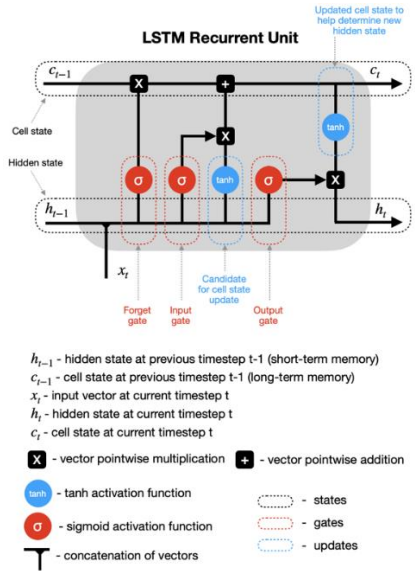
Inception je konvolutivna neuronska mreža nastala sa ciljem da se smanje resursi potrebni za obučavanje neuronskih mreža. Sastavljena je od 9 modula kao sa Slike 2, linearno poređanih. Ima 22 slojeva, 27 uključujući i *pooling* slojeve. InceptionV3 je nastala kako bi se mane originalne Inception arhitekture prevazišle, npr. smanjili potrebni računarski resursi za treniranje faktorizacijom filtera i ubacivanje regularizacione komponente za sprečavanje *over fitting*-a.



(b) Inception module with dimension reductions

Slika 2 *Arhitektura Inception modula*

LSTM (*Long Short-Term Memory*) je rekurentna neuronska mreža. Ono što je razlikuje od obične rekurentne neuronske mreže je to što u jednoj rekurentnoj jedinici umesto kombinacije prethodnog stanja i trenutnog stanja koje se šalje na aktivacionu funkciju, LSTM, kako se može videti na Slici 3, ubacuje razne kapije na kojima se odlučuje koje stanje će se „zapamtiti“ ili „zaboraviti“. Samim tim, arhitektura LSTM neuronske mreže je kompleksnija i više računski zahtevna od obične rekurentne.



Slika 3 *Arhitektura LSTM rekurentne jedinice*

Feature vektori generisani InceptionV3 neuronskom mrežom su vektori dužine 2048. Neuronska mreža kao ulaz prima slike dimenzije 299x299, stoga je prvo neophodno redefinisati dimenzije slika u skupovima.

Kreiran je rečnik svih jedinstvenih reči, kako bi postojao način predstavljanja reči preko indeksa. Ulaz u LSTM predstavlja kombinaciju podataka, *feature* vektori slike zajedno sa svim kombinacijama sekvenci reči, dok je izlaz modela naredna reč. Problem nestajućih gradijenata rešen je ograničavanjem dužine opisa slike, uzimajući najduži opis iz trening skupa. Zbog ograničenih resursa, model je treniran na 20 epoha.

B. *EfficientNet i transformer arhitektura*

EfficientNet je konvolutivna neuronska mreza nastala sa ciljem postizanja boljih performansi, uz manji utrošak resursa. Naime, za postizanje veće tačnosti kod standardnih konvolutivnih mreža, obično se radi proširivanje (eng. *scale-up*) mreže u širinu i dubinu, uz zauzimanje više resursa. EfficientNet proračunava koeficijente za koje je potrebno izvršiti proširenje u tri dimenzije: širina mreže, dužina mreže i rezolucija slike. Izračunati koeficijenti predstavljaju optimalan odnos performansi modela i dostupnih resursa. Postoje reazličite vrste EfficientNet mreža u zavisnosti od izabranih koeficijenata. U ovom eksperimentu korišćena je pretrenirana EfficientNet-B0 mreža.

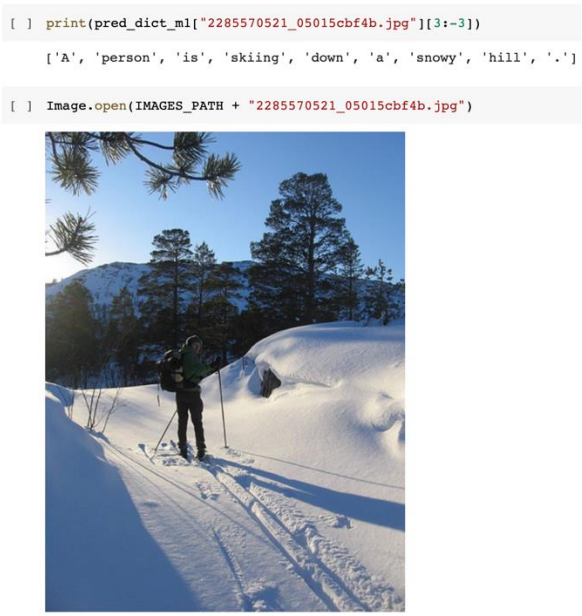
Transformer model sastoji se od enkoder i dekoder dela. Enkoder deo prihvata feature vektor iz konvolutivne mreže i generiše novu reprezentaciju. Transformer dekoder kao ulaze uzima novu reprezentaciju feature vektora iz enkodera i sekvencu reči koje predstavljaju opis zadate slike. Na taj način dekoder pokušava da nauči da generiše opis za svaku sledeću sliku. Transformer enkoder i dekoder su trenirani na 10 epoha zbog limitiranosti resursa.

Feature vektori generisani od strane EfficientNet mreže imaju dužinu 512. EfficientNet kao ulaz prima slike dimenzije 299x299, pa je neophodno prilagoditi slike zadatoj dužini. Neophodno je tokenizovati sve dostupne opise i napraviti jedinstven rečnik, gde bi svaka reč u opisu bila predstavljena svojim indeksom u rečniku. Problem nestajućeg gradijenta je rešen ograničavanjem dužine opisa na 25 reči.

IV. Rezultati

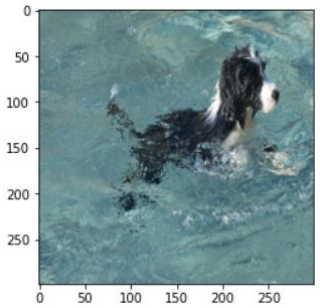
U ovom odeljku rada, biće prikazani rezultati izvršenih eksperimenata. Kao metrika za evaluaciju korišćena je BLEU metrika.

U sklopu prvog eksperimenata, posmatrana je BLEU metrika sa težinama (1, 0, 0, 0) označavajući da se posmatraju unigrami u generisanju sličnosti. LSTM putem *feature* vektora InceptionV3 mreže daje sličnost od skoro 57% sa originalnim opisima slike. Na Slici 4 se može videti primer generisanog opisa na osnovu slike.



Slika 4 *Primer generisanja opisa slike putem InceptionV3 i LSTM*

U sklopu drugog eksperimenata takođe je posmatrana BLEU metrika sa težinama (1, 0, 0, 0). Mere sličnosti opisa za svaku sliku iz testnog skupa su uprosečene i dobijena je mera sličnosti od skoro 61%. Na Slici 5 se može videti primer generisanog opisa na osnovu slike.




```
Predicted caption: a black
and white dog is swimming
in the pool
```

Slika 5 *Primer generisanja opisa slike putem EfficientNet i transformer modela*

Tabela 1 prikazuje sumarne rezultate eksperimenata. Arhitektura koja uključuje kombinaciju EfficientNet i transformer modela daje bolje rezultate u odnosu na arhitekturu koju čine InceptionV3 i LSTM (poboljšanje 4% BLEU mere).

Eksperiment	BLEU
InceptionV3 + LSTM	56,6%
EfficientNet + Transformer	60,8%

Tabela 1 *Tabelarni prikaz rezultata eksperimenata*

V. ZAKLJUČAK

Ovaj rad nastao je iz potrebe da se unaprede rešenja za digitalno konzumiranje sadržaja. Ovakvo rešenje bi se moglo koristiti za lakšu pretragu sadržaja u bazama podataka. Pored takvih primena, pomoglo bi i ljudima sa medicinskim smetnjama i oboljenjima koje ih sprečavaju da na lak način interpretiraju slike.

U ovom radu, izvedena su dva eksperimenta. U okviru prvog eksperimenta, izvlačili smo *feature* vektore uz pomoć InceptionV3 konvolutivne neuronske mreže, gde je takve vektore konzumirala LSTM rekurentna neuronska mreža u ulozi dekodera. U okviru drugog eksperimenta, *feature* vektori dobijeni su iz EfficientNet konvolutivne neuronske mreže, dok je takve vektore konzumirao transformer model.

Oba eksperimenta pokazala su zadovoljavajuće rezultate bez obzira na ograničene računarske resurse. Kao mera evaluacije korišćena je BLEU metrika. Prvi eksperiment dao je skoro 57% sličnosti generisanog opisa sa originalnim opisima, dok je drugi eksperiment dao bolje rezultate za čak 4%.

Daljim unapređivanjem tehnika za generisanje opisa slike se mogu performanse znatno povećati. Neki od načina poboljšanja performansi može biti proširenje skupova podataka, kao i raznolikost slika unutar skupova podataka. Takođe, mogu se isprobati neke druge tehnike, kao npr. korišćenje drugih neuronskih mreža i arhitektura. Performanse se mogu i poboljšati povećanjem računarskih resursa, gde bi i treniranje kompleksnijih arhitektura bilo moguće na veći broj epoha.

REFERENCES

[1] M. Z. Hossain, F. Sohel, M. F. Shiratuddin i H. Laga, „A Comprehensive Survey of Deep Learning for Image Captioning," *ACM Comput. Surv,* p. 36, 2018.

[2] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn i N. Pugeault, „Image Captioning through Image Transformer," *Computer Vision – ACCV,* 2020.

[3] M. Hodosh, P. Young i J. Hockenmaier, „Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research 47,* pp. 853-899, 2013.

[4] K. Papineni, S. Roukos, T. Ward i W.-J. Zhu, „BLEU: a method for automatic evaluation of," *Proceedings of the 40th annual meeting on association for computational linguistics,* p. 311–318, 2002.

[5] S. Banerjee i A. Lavie, „METEOR: An automatic metric for MT evaluation with improved correlation," *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine,* t. 29, p. 65–72, 2005.

[6] C.-Y. Lin, „Rouge: A package for automatic evaluation of summaries," *Text summarization branches out: Proceedings of the ACL-04 workshop,* t. 8, 2004.