Visual Question Answering (VQA)

Natalija Krsmanović
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
krsmanovic.natalija99@gmail.com

Srđan Topić
Fakultet tahničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
topicsrdjan99@gmail.com

***Sažetak*— Kako vještačka inteligencija napreduje, jedna od najuzbudljivijih oblasti jeste vizuelno odgovaranje na pitanja, tj. odgovaranje na pitanja koja su vezana za određenu sliku. Ova oblast objedinjuje dvije cjeline. Prva je razumjevanje prirodnog jezika u tekstu, koja može omogućiti čovjeku da prirodniju i intuitivniju interakciju sa mašinom, čak i u tolikoj mjeri da je skoro neprimjetna razlika u razgovoru sa čovjekom ili mašinom. Druga cjelina jeste prepoznavanje sadržaja slika, koje obuhvata određivanje određenih karakteristika na slikama, na osnovu kojih se može odrediti šta se tačno nalazi na slici. Ideja ovog projekta je stvaranje modela, koji spaja navedene oblasti na način koji omogućuje postavljanje pitanja, koja se odnose na sadržaj slike, gdje bi model probao dati tačan odgovor. Korićeni su modeli dubokih neuronskih mreža. Za ekstrakciju karakteristika sa slika upotrebljeni su VGG16 i ResNet50, koji su naknadno modifikovani, dok su za ekstrakciju važnih informacija iz pitanja upotrebljeni LSTM i BERT. U okviru rada detaljno su objašnjene izabrane arhitekture i kako se one međusobno kombinuju. Evaluacija modela je mjerena putem tačnosti i upoređena sa rezultatima iz sličnih radova.**

I. Uvod

Vještačka inteligencija (AI) je oblast koja se značajno razvija i primijenjuje posljednjih nekoliko godina u svakodnovnom životu i mnogim domenima. Nakon što su pronađena rješenja za specifične probleme, porasla je tendencija za razvojem AI kao multidisciplinarne oblasti, gdje bi se određeni algoritmi i rješenja koristili nad problemima koji su kombinacija više njih. Jedan od takvih problema jeste *Visual Question Answering* (VQA). VQA je oblast koja spaja primjenu algoritama za razumijevanje teksta (NLP), kao i algoritama za izvlačenje informacija sa slika, pri čemu je cilj da se poboljša mašinsko razumijevanje na način da mašina uspije da poveže sliku i pitanje u jednu cjelinu i na osnovu toga pruži adekvatan odgovor. Iako je ljudima intuitivno da na osnovu slike odgovore na osnovna pitanja koja zahtijevaju prebrojavanje ili prepoznavanje određene boje, predmeta, radnje itd. , mašini je potrebno jako detaljno i duboko razumijevanje kako teksta tako i slike da bi dala smislen odgovor. Motivacija za izradu ovog rada leži u potrebi da se mašinama omogući bolje razumijevanje pitanja koja nisu unaprijed definisana u vezi sa vizuelnim sadržajem, radi praktične primjene. Ovakav sistem bi na primjer pomogao osobama sa oštećenim vidom da postavljaju pitanja u vezi sa određenim slikama radi boljeg razumijevanja sredine i lakšeg snalaženja. Takođe imao bi veliku primjenu u medicini za analizu i uspostavljanje nalaza i terapije na osnovu rentgenskih slika. Pored navedenog VQA se može primjenjivati u obrazovne svrhe, industriji, marketingu i mnogim drugim oblastima.

S obzirom da dati problem zahtjeva upotrebu više algoritama, postoji već mnogo predloženih rješenja i pristupa. U okviru ovog rada cilj je da se isproba i uporedi više arhitektura modela i njihovih kombinacija kako bi se stekla šira slika o mogućim problemima kao i daljem proširenju. Najčešći pristup jeste upotreba dubokih neuronskih mreža koje se obučavaju nad slikama i tekstualnim sadržajem radi uočavanja bitnih karakteristika. Nad ovim pristupom zasnivaće se naše rješenje.

Konkretno za obradu slika i ekstrakciju bitnih osobina (*Image Features Extraction*) koristiće se modeli koji u osnovi koriste konvolutivne neuronske mreže (CNN). Oni predstavljaju nadogradnju odnosno primjenu dubokih mreža sa velikim brojem slojeva koji su u mogućnosti da uoče više detalja na slikama. Za treniranje ovakve mreže potrebni su ogromni resursi za smještanje podataka i izračunavanja kao i vrijeme, pa se iz tog razloga koriste već pretrenirani modeli. Ovi modeli su većinom pretrenirani nad skupom podataka ImageNet[1], koji sadrži preko million labeliranih slika sa 1000 kategorija. Konstruisan je u svrhu primjene u oblasti istraživanja na temu detekcije objekata. Korišćenje pretreniranih modela se može unaprijediti za specifičan domen primjenom *fine-tuning*-a. Ovaj process se izvodi tako da se prednji slojevi duboke mreže zamrznu, ali se slojevi poslije toga ponovo treniraju nad novim skupom podataka da bi se uočile karakteristike koje su specifične samo nad datim skupom, a nisu bile uočene u okviru ImageNet slika. Na sličan način se konstruišu i modeli za procesiranje teksta. U radu su predstavljena dva načia. Prvi podrazumijeva korišćenje rekurentnih neuronskih mreža (RNN), a drugi korišćenje transformera. U slučaju transformera, iz istih razloga kao kod obrade slike korišćen je pretrenirani model.

Kako bi se došlo do odgovora nakon analize slike i pitanja, potrebno je te rezultate sjediniti i dodatno analizirati. Za ove potrebe predložena su dva pristupa, gdje se jedan zasniva na konkatenaciji prethodno navedenih izlaza, a drugi na množenju vektora. U oba slučaja rezultat se prosljeđuje potpuno povezanoj mreži (MLP). U poglavlju 2 biće dat pregled postojećih rješenja sa detaljima primjenjenih modela I njihovim rezultatima.

II. RELEVANTNA LITERATURA

Prilikom traženja rješenja na ovu temu, u obzir su uzeti radovi koji koriste arhitekturu povezivanja modela kao što je opisano u uvodnom dijelu. Među pronađenim rješenjima izdvajamo "VQA: Visual Question Answering"[2]. U ovom radu je iskorišćen skup podataka VQA. Dati skup će biti korišćen u našem radu takođe I zato će detaljno biti opisan u narednom poglavlju. U navedenom radu analize su vršene nad cijelim skupom podataka, što podrazumijeva obradu realnih i apstraktnih slika, kao i sve vrste dostupnih *open-ended* i *multiple-choice* pitanja i tipova odgvora. Kako bi imali osnovu za poređenje performansi modela, naveli su više opcija *baseline* modela. Neki od njih su izbor random odgovora od najčešćih 1000, postavljanje svakog odgovora na "yes" jer je to najčešći odgovor i drugi.

Za ekstrakciju karakteristika sa slika u navedenoj literaturi upotrebljen je VGGNet[3] na dva načina, pri čemu se kao izlaz uzimao vektor iz posljednjeg skrivenog sloja dužine 4096, gdje se nad njim dodatno vršila L2 normalizacija u drugom slučaju.

Kada je u pitanju analiza teksta odnosno pitanja, predložene su tri opcije: Bag of Words (BoW)[4], LSTM i deeper LSTM. Za primjenu Bag of Words izvučeno je po 10 najčešćih riječi za prve tri pozicije u pitanjima. To je rezultovalo vektorima duzine 1030 za svako pitanje. Prvi model za LSTM se sastoji samo od jednog skrivenog sloja, pri čemu se rezultujući vektor dobija konkatenacijom 512-dimenzionih vektora iz stanja zadnje ćelije i skrivenog stanja. Drugi model LSTM je proširen dodavanjem još jednog skrivenog sloja i potpuno povezanog sloja na kraju, kako bi se na izlazu opet našao vektor dužine 1024.

Rezultati za slike i pitanja za par BoW I VGGNet su konkatenirani i prosljeđeni potpuno povezanoj mreži sa dva skrivena sloja. U slučaju kombinacije LSTM I VGG nad njihovim vektorima se primjenjuje množenje pojedinačnih elemenata i to dalje prosljeđuje MLP. Svaki sloj MLP se sastoji od 1000 neurona, a *dropout* je postavljen na 0,5. Aktivaciona fukcija između na kraju ovih slojeva je *tanh*, a kao zadnji sloj postavljen je *softmax* kako bi se na izlazu nalazilo K kategorija odnosno mogućih odgvora. Za potrebe testiranja modela nad *open-closed* dogovorima biran je onaj koji od K mogućih ima najveću vjerovatnoću. Za *multiple-choice* odgovore konačan je onaj koji ima najveću aktivaciju među ponuđenim odgovorima.

Za evaluaciju modela korišćena je tačnost. Najbolji rezultat dao je model koji kombinuje LSTM sa dva sloja i VGG sa L2 normalizacijom. Tačnost tog modela za *open-ended* zadatak je 57,75%, a za *multiple-choice* 62,70%.

Na osnovu rezultata ovog rada, mi ćemo u našem iskoristiti ideju o dvije vrste kombinovanja rezultujućih vektora, kao i dodavanje potpuno povezanog sloja na VGG I dodavanje još jednog skrovenog sloja u LSTM.

U drugom radu "Visual Question Answering" [5] pomenuti su i drugi skupovi podataka nad kojima se mogu trenirati modeli za ovaj problem. Kao i u prošlom radu upotrebljen je VQA v2.0 skup podataka, ali od svih pitanja odabrana su samo *multiple-choice* i problem je predstavljen kao klasifikacioni sa K najčešćih odgovora.

Za izdvajanje osobina sa slike primijenjen je VGG19 zajedno sa potpuno povezanim slojevima. Svaka slika je bila enkodovana putem vektora dužine 1000, kao rezultat *softmax* funkcije nad posljednjim povezanim slojem.

Međuzavisnosti i riječi u pitanju enkodovane su putem pretreniranog Google modela Infer Sent [6] koji je treniran nad *Stanford Natural Language Interface,* koji je labeliran ručno. Dati model se sastoji od bidirekcione LSTM. Svaka riječ prije ulaza u LSTM reprezentovana je putem GloVe[7] vektora. Rezultat ovog modela je vektor dužine 4096 za svako pitanje. Sledeći korak je izveden konkatenacijom enkodovane slike i pitanja čime se dobija 5096-dimenzioni vektor. Pored ovog pristupa iskorišteni su i dodatni povezani slojevi kako bi oba rezultujuća vektora bila dužine 1000 da bi se mogli množiti.

Eksperimenti i klasifikacija su isprobani sa tradicionalnim mašinskim učenjem korišćenjem Support Vector Machine(SVM) algoritma, kao i sa dubokim mašinskim učenjem- MLP. U ovom slučaju MLP se sastojao od tri sloja sa po 1024 neurona, pri čemu je *dropout* postavljen na 0,5. Prilikom treniranja modela skup podataka je podijeljen na odgvore iz grupe "yes/no" i "other". Kao i u prvom navedenom radu, rezultati su najbolji za opciju sa množenjem vektora . Za "yes/no" tačnost je 72,76%, a za "other" sa 100 kategorija 57,54%.

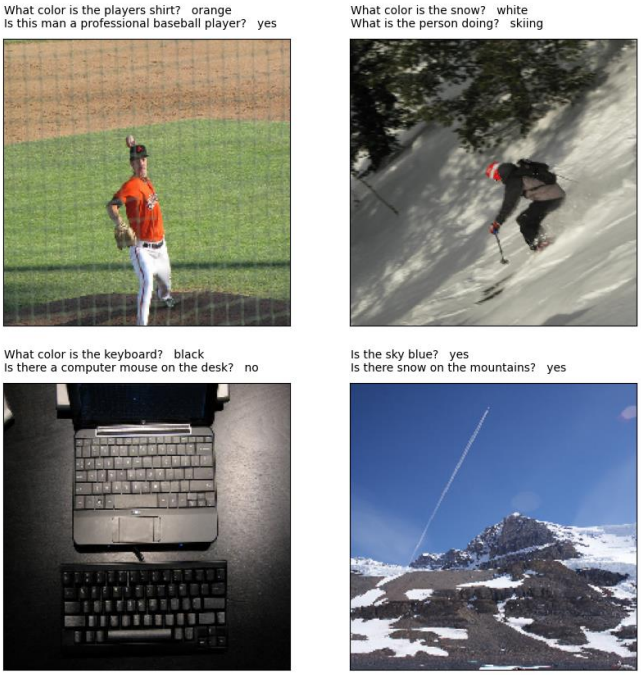
III. ANALIZA SKUPA PODATAKA

Po uzoru na relevantnu literaturu skup podataka koji je posmatran jeste VQA v2.0[8]. Dati skup je podijeljen na tri foldera. Prvi "Images" sadrži slike podijeljene na trening, validacioni i test skup sa po 82 783, 40 504 i 81 434 slika redom. Drugi folder "Questions" sadrži pitanja zajedno sa njihovim id vrijednostima, kao i id-em slike na koju se odnose. Trening skup sadrži 443 757 pitanja, validacioni 214 354, a test 447 793. U datim skupovima postoje pitanja koja su ponovljena, odnosno ista pitanja su povezana sa više različitih slika. Treći folder pod imenom "Annotations" sadrži podatke o odgovorima. Konkretno sadrži id pitanja, id slike, najcesci odgovor, tip pitanja i tip odgovora. Trening skup sadrži 4 437 570 odgovora, a validacioni 2 143 540.

Navedene brojke se odnose samo na realne slike. Apstraktne slike nisu uzete u obzir. Kako bi se lakše izvršila analiza, dati uzorci po folderima su spojeni u jedan konačan skup podataka. Na slici 1 je prikazan primjer pitanja i odgovora koji su u vezi sa nekom slikom.

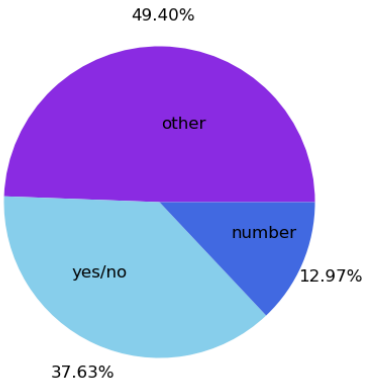
A. Odgovori

S obzirom das mo se odlučili da problem predstavimo kao klasifikacioni za određeni broj najčešćih odgovora potrebno ih je dodatno analizirati. U okviru i trening i test skupa postoji podjela na tri vrste odgovora. Prva su „yes/no" odgovori, potom „number" i „other". Na slici 2 je prikazana raspodjela vrsta odgovora nad cijelim trening skupom.



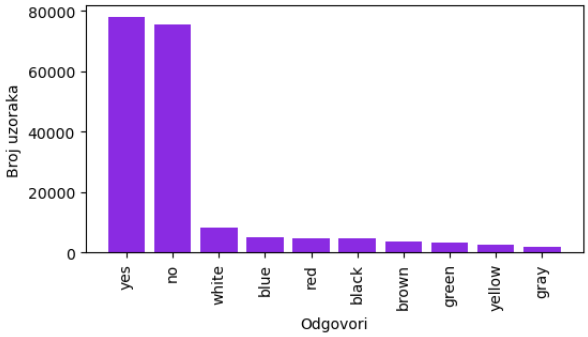
Slika 1. Primjer uzoraka sa slikama i pridruženim pitanjima i odgovorima

Vidimo da najveći udio uzoraka tj. čak polovina ima odgovore koji su tipa "other". Odmah nakon toga skoro 40% pitanja ima odgovor sa da ili ne. Iz ovoga se može zaključiti da najviše ima odgovora sa da ili ne s obzirom da su samo te dvije opcije moguće iz ove grupe, dok postoji veliki broj različitih odgovora iz grupe "other" sa manjim brojem pojavljivanja. U okviru te grupe prvobitno je očekivano da postoje odgovori sa više riječi, ali međutim prisutna su samo tri odgovora koja sadrže dvije ili tri riječi. Najmanje odgovora se odnosi na brojeve, koji se odnose najviše na prebrojavanje određenih objekata na slici, ili prepoznavanja koji broj je napisan na slici. Pošto dati problem čitanja infomracija sa slike podrazumijeva upotrebu dodatnih OCR [9] alata, iz daljnjeg razmatranja uklonićemo sve uzorke čiji odgovor je tipa "number".



Slika 2. Udio vrsta odgovora u trening skupu

Na slici 3 su prikazani najčešći odgovori, na osnovu kojih može da se zaključi da će vjerovatno model više težiti ka tome da kao rezultat predvidi odgovor sa da ili ne, i nad tim pitanjima će se bolje obučiti u odnosu na pitanja koja zahtjevaju drugačiji odgovor.



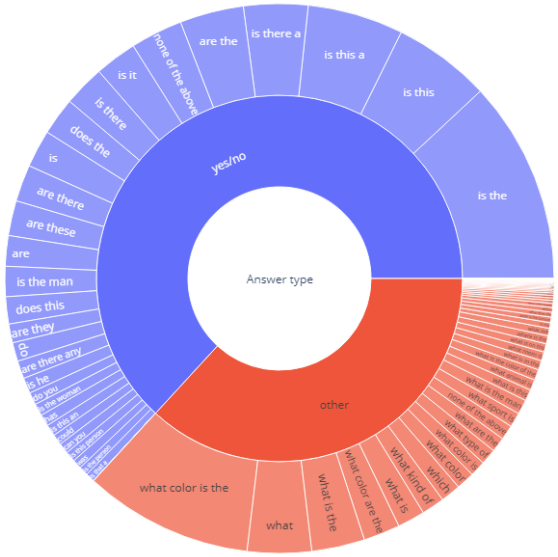
Slika 3. Top 10 najčešćih odgovora

B. *Pitanja*

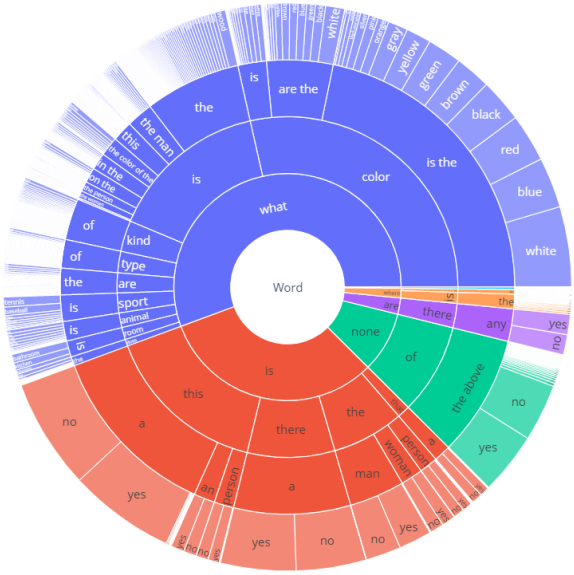
Kao što je prethodno spomenuto, u okviru skupa podataka postoji polje koje označava tip pitanja. U okviru i trening i test skupa zabilježeno je 65 različitih tipova pitanja. Na slici 4 je prikazana distribucija tipova pitanja u odnosu na vrstu odgovora. Može se primjetiti da pitanja koja započinju sa riječi "what" imaju odgovor koji pripada grupi "other". Za da/ne pitanja mnogo je veći raspon i raznolikost u početnim riječima pitanja.

Kako bi se uočile dodatne veze između odgovora i riječi u pitanjima, izdvojene su prve tri riječi iz svakog pitanja. Njihova distribucija prikazana je na slici 5. Zadnji prsten datog grafika predstavlja odgovore na ta pitanja. Može se uočiti da je najviše pitanja koja započinju sa riječju "what". Na pitanja koja započinju sa "is" ili "none" odgovori su većinski iz grupe da/ne.

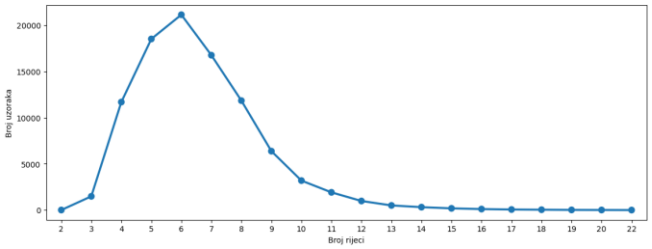
Na slici 6 je prikazan broj riječi koji je zastupljen po svakom pitanju. Može se primijetiti da najveći broj pitanja sadrži oko 6 riječi, dok za više od 10 riječi postoji jako mali broj uzoraka.



Slika 4. Distribucija tipova pitanja na osnovu tipova odgovora



Slika 5. Distribucija prve tri riječi iz pitanja zajedno sa konkretnim odgvoorima



Slika 6. Raspodjela uzoraka u odnosu na broj riječi unutar pitanja

IV. PRIMIJENJENI MODELI

U ovom poglavlju biće predstavljeni modeli koji su izabrani za izvlačenje karakteristika sa slike i iz pitanja i na koji način su kombinovani da bi se dobio konačan rezultat odnosno odgovor.

A. *Image Features Extraction*

U skupu podataka koji je razmatran svaka slika je različitih dimenzija. Kako bi se slike mogle proslijediti nekom od modela za modelovanje istih, potrebno ih je svesti na iste dimenzije. Na osnovu literature izabrano je da to bude 224 x 224 x 3.

Modeli koji su posmatrani:

1) *Pretrenirani VGG16*: zarad izvlačenja odnosno enkodovanja slika u arhitekturi ovog modela zadržani su samo konvolucioni slojevi. To znači da je parameter "include_top" postavljen na False, pa se time pri treniranju modela ne koristi dio sa potpuno povezanim slojevima, već se rezultat izvlači direktno iz posljednjeg konvolucionog.

2) *Pretrenirani ResNet50:* kao i kod VGG16 i u ovom slučaju izbačeni su posljednji potpuno povezani slojevi. Međutim pošto je izlaz iz posljednjeg konvolucionog sloja (broj_slika, 7, 7, 2048), dodat je još jedan sloj „GlobalAveragePooling2D" koji usrednjava vrijednosti i kao rezultat daje izlaz dimenzije (broj_slika, 2048).

Kako bi se izvučene karakteristike mogle dalje upotrebiti u kombinaciji sa rezultovanim vektorima pitanja, prebačene su da budu 2D. Odnosno (broj_slika, 25 088) u slučaju VGG16 I (broj_slika, 2048) u slučaju ResNet50.

B. *Question Features Extraction*

Za obradu teksta odnosno pitanja su predložena dva pristupa. Prvi se zasniva na rekurentnim mrežama, a drugi na transformerima.

Dati modeli su:

1) *LSTM1* :u prvom pristupu tokenizacija riječi je izvršena samo na osnovu riječi koje se nalaze u okviru pitanja u predstavljenom skupu podataka. Na osnovu tog riječnika za svako pitanje dodijeljena je lista brojeva, gdje je svaki za jednu riječ.

Pošto su sva pitanja različite dužine a LSTM zahtijeva sekvence fiksne dužine izvršen je *padding* nad tim sekvencama tako da se dopune nulama do dužine pitanja sa maksimalnim brojem riječi.Rezultat se prosljeđuje modelu koji sadrži jedan *Embedding* sloj i dva povezana LSTM sloja sa po 512 ćelija i parametrom *dropout* na 0,2. Rezulat je vektor (broj_pitanja, 512).

2) *LSTM2*: ovaj pristup koristi iste slojeve kao LSTM1 ali za tokenizaciju riječi unutar pitanja na koristi rječnik sagrađen nad riječima iz pitanja već reprenirani GloVe (glove-viki-gigaword-100)[10]. Na osnovu rezultata se kreira matrica

težina koja se proslijeđuje *Embedding* sloju, kome je parameter *trainable* podešen na False.
Rezultat je takođe 2D vektor (broj_pitanja, 512)

3) *BERT*: ovaj model se zasniva na primjeni transformera. Za tokenizaciju je upotrebljen pretrenirani BertTokenizer I to bert-base-uncased koji je treniram nad raznim tekstovima na internetu čiji se riječnik sastoji od riječi napisanim malim slovima. Nakon primjene ovog postupka svaka riječ je predstavljena vektorom dužine 768.

Kako bi se dobio izlaz odnosno samo enkodovana sekvenca odnosno pitanje, preuyima se posljednji sloj I vrijednosti iz mprvog tokena koji sadrži bitne informacije tj. reprezentaciju pitanja. Rezultat je posljednično 2D vektor (broj_pitanja, 768)

C. *Kombinovanje enkodovane slike i pitanja*

Nakon što su ulazna slika I pitanje koje se odnosi na nju analizirani I enkodovani sa karakterističnim vektorima, da bi ih posmatrali kao jedna cjelina potrebno je da ih ukombinujemo na dva načina:

1) *Konkatenacija*: vektori su konkatenirani jedan na drugi bez prilagođavanja dimenzija. Može se uočiti da će postojati više različitih veličina konačnog vektora u zavisnoti koje modele od predloženih međusobno kombinujemo

2) *Element-wise množenje*: dati pristup podrazumijeva moženje vektora element po element, iz čega slijedi da I vektor slike i pitanja moraju biti iste dimenzije. Zbog ovog uslova određeni modeli će dodatno biti prošireni sa potpuno povezanim slojevima kako bi se smanjile dimenzionalnosti do poklapajućih.

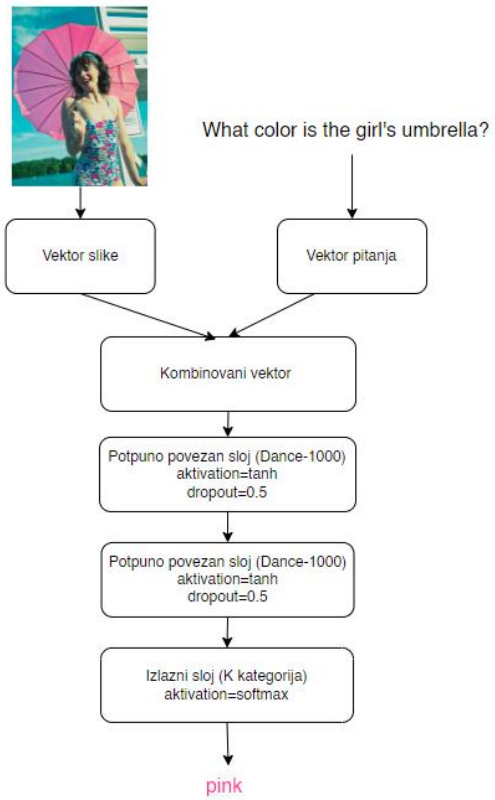
D. *Klasifikacija odgovora*

Posljednji korak jeste klasifikacija vektora iz poglavlja 4,3 na K izabranih klasa. Klase odgovora su birane tako da predstavljaju 100 najčešćih odgovora. Samim tim se ovaj problem svodi na izbor jedne od 100 klasa kao odgovor na zadatu sliku I pitanje.

S obzirom da u skupu podataka ipak postoji najviše odgovora sa da ili ne, moguće je da će i model težiti ka tim odgovorima i bolje se nad njima istranirati u odnosu na druge klase.

Za rješavanje problema klasifikacije izabran je Multi-Layer-Perceptron (MLP). Konačan model sadrži dva skrivena potpuno povezana sloja sa po 1000 neurona. Po uzoru na slične radove u oba sloja postavljen je *dropout* na 0.5, a aktivaciona funkcija je tanh. Posljednji sloj je *softmax* koji dodjejuje određenu vjerovatnoću svakoj od K klasa, pri čemu bi konačan odgovor bio onaj sa najvećom vjerovatnoćom.

Na slici 7 je prikazana arhitektura opisanog sistema sa svim komponentama.



Slika 7. Prikaz generalne arhitekture sistema za VQA

E. *Rezultati*

Svi predloženi modeli iz prethodnog poglavlja su trenirani samo nad 7000 uzoraka iz trening skupa zbog računarskih resursa. Rezultati svakog posebno modela su čuvani u posebne fajlove koji su nakndano učitani radi kombinovanja i klasifikacije.

Nakon što su sačuvane karakteristike iz pitanja i slika za sve modele, one su proslijeđene dalje MLP na klasifikaciju. Treniranje se vršilo nad 5000 uzoraka, od kojih je 20% bilo iskorišćeno za validaciju tačnosti kroz epohe. Testiranje je odrađeno nad 2000 uzoraka sa onim težinama koje su dale najveću tačnost tokom treninga.

Prve posmatrane kombinacije su između VGG16 i svih prethodno navedenih modela za ekstrakciju karakteristika iz pitanja (LSTM1,LSTM2,BERT). Kao što se može vidjeti u tabeli 1, eksperiment sa BERT modelom je lošiji u odnosu na LSTM. Kada se porede LSTM1 I LSTM2 tačnosti su iste nad test skupom ali pošto je loss bila manja za LSTM1 nad test skupom, odlučeno je da će naredni eksperimenti koristiti LSTM1 kao model za enkodovanje pitanja.

Može se primjetiti da kombinacija vektora slike koji se dobio putem ResNet modela daje bolje rezultate u odnosu na VGG. Zanimljivo je primijetiti i to da je BERT zajedno sa VGG imao najgore rezultate, dok je u konkatenaciji sa ResNet dao najbolji rezultat.

U posljednjem eksperimentu primijenjen je drugačiji pristup kombinovanju vektora odnosno koristi se množenje vektora. Iz tog razloga je iskorišćen VGG koji je dopunjen potpuno povezanim slojem kako bi konačan izlaz bio istih dimenzija kao LSTM. Iako je očekivano na osnovu rezultata iz literature da će ovaj pristup dati najbolje rezultate ipak je zauzeo drugo mjesto.

TABELA 1. TAČNOST ZA EKSPERIMENTE

Eksperimenti	Validacioni skup	Test skup
VGG16+LSTM1+Con	34.90	31.65
VGG16+LSTM2+Con	35.00	31.65
VGG16+BERT+Con	34.90	30.70
ResNet50+LSTM1+Con	34.90	32.55
ResNet50+BERT+Con	39.40	35.55
VGG16(MLP)+LSTM1+Mul	34.90	34.00

V. ZAKLJUČAK

Rezultati u ovom radu su dosta lošiji u odnosu na rezultate iz radova koji su navedeni u poglavlju 2. To se može objasniti time što je u pomenutim radovima za trening modela korišten skoro čitav skup dostupnih uzoraka. Na taj način modeli su uspjeli više da se prilagode podacima I uoče određene relevantne karakterisike. Takođe broj uzoraka doprinosi i činjenici da tipovi odgvoora nisu balansirani.

Ako se posmatra mjera tačnosti nad test skupom za izvedene eksperimente, primjetno je da joj upotreba ResNet modela doprinosi. Razlog tome može biti sama arhitektura modela koja je složenija i sa većim brojem slojeva od VGG16.

Kako bi se dobili bolji rezultati i proširio ovaj rad, predlaže se pokretanje najbolje pokazanih modela ali nad velikim skupom podataka, uz dodatno prilagođavanje parametara. Još jedan dodatan pristup bio bi podjela skupa na uzorke kojima je odgovor samo iz skupa “yes/no” i odvojeno iz skupa “other”.

Reference

[1] [Online] https://www.image-net.org/

[2] Aishwarya Agrawal , Jiasen Lu , Stanislaw Antol , Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh, “Visual Question Answering,”, arXiv:1505.00468v7 [cs.CL] 27 Oct 2016.

[3] Usman Muhamed,Weiquiang Wang, Sajid Ali, “Pre-trained VGGNet Architecture for Remote-Sensing Image Scene Classification”, 018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, August 20-24, 2018.

[4] [Online] https://www.mygreatlearning.com/blog/bag-of-words/

[5] Pankti Kansara, “Visual Question Answering”, San Jose State University, Spring 2018.

[6] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, “Supervised learning of universal sentence representations from natural language inference data,” arXiv preprint arXiv:1705.02364, 2017.M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[7] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[8] [Online] https://visualqa.org/

[9] [Online] https://www.simpleocr.com/tag/ocr-engine/

[10] [Online] https://huggingface.co/fse/glove-wiki-gigaword-100