Visual Question Answering – VizWiz izazov

Tamara Ranković
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad, Srbija
tamara.rankovic@uns.ac.rs

Eva Janković
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad, Srbija
eva.jankovic@uns.ac.rs

***Apstrakt*—Visual Question Answering kao složen problem zahteva rešavanje zadataka iz polja kompjuterske vizije i obrade prirodnog jezika. Kao takav, u pomenutim sferama AI zajednice izazvao je veliko interesovanje. Cilj sistema je pronalaženje odgovora na pitanje postavljeno na osnovu fotografije. Primene sistema variraju od automatskog opisa fotografija i pružanja prirodnije komunikacije sa računarskih sistemima (gde korisnik može pretraživati podatke kroz upite vezane za sadržaje fotografija), do pomoći slabovidim i slepim osobama u dnevnim aktivnostima. U ovom radu iznet je pristup rešavanju VQA problema u pogledu pomoći slepim osobama. Opisan je proces obrade odabranog skupa podataka, sveukupna arhitektura sistema i svaka od pojedinačnih komponenti. Opisane su upotrebljene arhitekture neuronskih mreža za određivanje regiona fotografija i izvlačenje informacija o sadržaju regiona, kao i arhitekture korišćenje za obradu ulaznih tekstualnih pitanja. Izneti su postignuti rezultati i ograničenja sistema u rukovanju tekstualnim sadržajem u okviru fotografija i rukovanju mutnim fotografijama. Predložena arhitektura pokazuje nivo rezultata koji se može ostvariti uz jednostavniju arhitekturu sistema i hardverska ograničenja.**

Keywords—VQA; kompjuterska vizija; NLP; duboke neuronske mreže;

I. UVOD

Razvoj različitih oblasti veštačke inteligencije konstantan je u proteklim godinama. Napredak na polju kompjuterske vizije doprineo je boljim performansama postojećih modela neuronskih mreža i razvoji potpuno novih arhitektura. Vodeće arhitekture za rešavanje raznih zadataka vezanih za slike dugo su konvolucione neuronske mreže (engl. *CNN*) [1]. Najčešći problemi koje kompjuterska vizija nastoji da reši su detekcija objekata, boja, teksta, kao i razna prebrojavanja. Sa druge strane, postoje veliki napreci i na polju obrade prirodnog jezika (engl. *NLP*), gde su takođe prisutne različite arhitekture neuronskih mreža. Najveću primenu su našle arhitekture bazirane na rekurentnim neuronskim mrežama [2] zbog svoje mogućnosti pamćenja vremenskog konteksta. Njihova uspešnost posledica je strukture govora i bitnosti konteksta na značenje reči. U prethodnim godinama vidimo sve veću primenu transformer modela i njihovu sve veću uspešnost u oba pomenuta polja veštačke inteligencije [3].

Kontinualni razvoj oba polja doveo je do potrebe za novim izazovima. Visual Question Answering (VQA) je problem koji podrazumeva odgovaranje na pitanja u kontekstu određene slike. Neophodno je da sistem može da prepozna šta je sadržaj slike, kao i da razume pitanje postavljeno prirodnim jezikom i na osnovu njih da odgovarajući smisleni odgovor. Kako je za svaki postupak obučavanja potreban skup podataka, razvoj VQA oblasti uslovljen je pojavljivanjem velikog broja skupova podataka. U okviru rada [4] dat je detaljan pregled pristupa rešavanja VQA problema, kao i skupova podataka koji se u te svrhe koriste. VQA nalazi veliku primenu u različitim sferama kao što su:

- interakcija čovek-računar, gde može pružiti ljudima bliži način komunikacije kroz pitanja i odgovore,
- pretraživanje korpusa slika, gde korisnik može zatražiti, na primer, sve slike na kojima pada kiša i na osnovu odgovora da li na određenoj slici pada kiša sistem bira da li da je prikaže,
- pomoć slabovidim i slepim osobama u svakodnevnim aktivnostima.

Glavna prepreka koja se sreće u polju poslednje navedene primene su bili sami skupovi podataka koji su pretežno bili sačinjeni od slika koje su po svojim karakteristikama (zamućenost, fokus, postojanje prepreka) dosta odstupale od fotografija koje bi ciljna grupa mogla da uslika. Kreiranjem WizViz skupa podataka [5], ova prepreka je umanjena. WizViz je specijalizovan skup podataka sastojan od fotografija i pitanja nastalih u svakodnevnom korišćenju aplikacije za pomoć slepim osobama. One bi uslikale pojam od interesa i snimile audio zapis pitanja. Skup prikupljen na ovakav način odgovara podacima koji bi se sretali u realnom svetu.

U radovi [6][7] autori predlažu arhitekture za rešavanje VQA problema bazirane na izvlačenju vektora atributa slika, upotrebom CNN-a, i teksta upotrebom long short-term memory (LSTM) mreže i njihovom kombinovanju. U okviru oba rada, kao najbitniji segment arhitekture izdvaja se *attention* sloj baziran na pitanju na osnovu koga sistem zna na koji segment slike treba da se fokusira. Na osnovu radova, uviđena je prednost koju može doneti attention sloj i isti je uključen u arhitekturu sistema.

Rad [8] predlaže arhitekturu baziranu na kombinovanju vektora atributa dobijenih na osnovu teksta pitanja i ulaznih slika. Kako bi dobili vektor atributa za sliku, autori prvo koriste CNN kako bi izdvojili regione slike i zatim kreiraju vektor atributa za svaki od regiona. Tako dobijeni vektori se matrično

množe sa vektorima atributa ulaznog pitanja, u cilju uviđanja veza između pitanja i svakog od regiona; koliko je region značajan za postavljeno pitanje. Na osnovu rada, u arhitekturu sistema uključena je i komponenta za izdvajanje regiona ulaznih fotografija.

U radu [9] detaljno je predstavljena arhitektura bazirana na kombinovanju vektora atributa pitanja dobijenog upotrebom word embedding-a i Gated Reccurent Unit (GRU) neuronske mreže i vektora atributa slike dobijenih za svaki region slike upotrebom ResNet konvolucione neuronske mreže. Dodatno, u okviru arhitekture uključen je i attetion sloj baziran na pitanjima. Poslednji segment arhitekture čini linearni klasifikator koji povezuje kreirani vektor atributa za odgovorima. Skup podataka nad kojim je arhitektura testirana je WizViz skup koji je odabran i za ovaj rad. Predložena arhitektura postiže dobre rezultate i predstavlja bazu za arhitekturu predstavljenu u okviru ovog rada.

Transformer modeli trenutno pokazuju najbolje rezultate pri rešavanju različitih problema na polju NLP-a. U radu [10] predstavljen je BERT transformer, njegova arhitektura, postupak treniranja i primeri moguće primene. Autori prikazuju kako se pretrenirani BERT uz dodatno podešavanje parametara može upotrebiti kao ekstraktor atributa zadatog ulaza i kako se taj rezultat može dalje koristiti za različite zadatke. Sem ovakvog pristupa, u radu se ističe i mogućnost izvlačenja osobina iz pretreniranog BERT-a bez daljeg obučavanja na konkretnom zadatku. Benefit navedene alternative jeste da se ona može obaviti u fazi pretprocesiranja čime se postiže znatno veća efikasnost.

U okviru rada [11] istaknuti su dostupni skupovi podataka, algoritmi i problemi s kojima se algoritmi sreću prilikom rešavanja VQA problema. Od velikog značaja je i uporedni pregled evaluacionih metrika sa prednostima i manama svake od njih. Najčešći pristupi evaluaciji algoritama su preciznost (engl. *accurecy*), WUPS [12], metrika zasnovana na konsenzusu, i ručna evaluacija. Dodatno je istaknuto da i dalje ne postoji univerzalna metrika za istraživani problem i da odabir iste dosta zavisi od samog skupa podataka (upotreba određenih metrika može dovesti do neopravdano visokih rezultata), kao i od načina dolaska do odgovora (generisanje ili klasifikacija). Na osnovu datog pregleda i smernica pruženih u okviru rada koji definiše odabran skup podataka, metrike koje su korišćene su Accurecy i VQA Accurecy metrika koja spada u metrike bazirane na konsenzusu.

U nastavku rada dat je detaljan prikaz skupa podataka, isprobanih pristupa rešavanju problema, konačne arhitekture sistema i ostvarenih rezultata. II poglavlje sadrži detaljan opis skupa podataka i procesa obrade koji je nad njim odrađen. U III poglavlju nalazi se pregled korišćenih metodologija uz poseban osvrt na svaku konačnu komponentu arhitekture posebno. IV poglavlje sadrži pregled rezultata postignutih upotrebom opisane arhitekture, kao i diskusiju vezanu za iste. U poslednjem poglavlju je dat zaključak koji sumarizuje rad i sadrži predloge mogućih unapređenja rešenja.

II. Opis skupa podataka

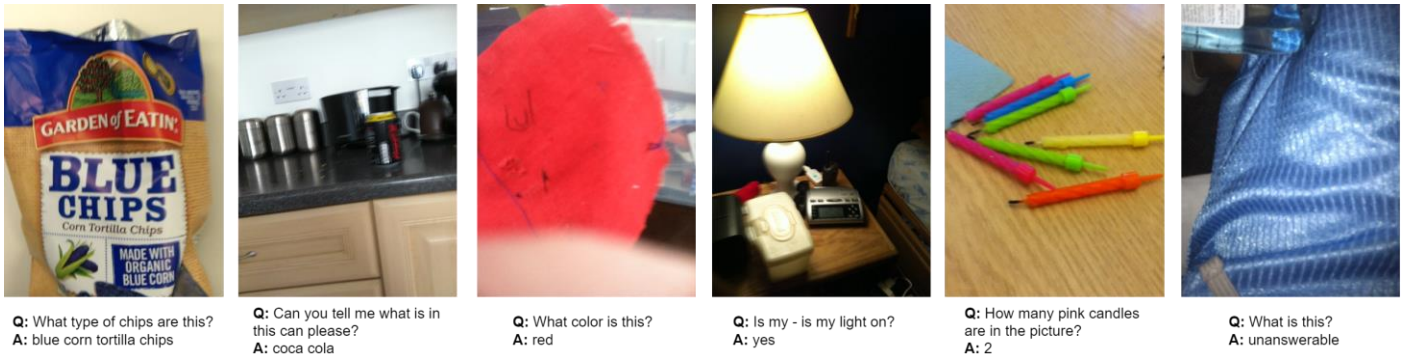
Skup podataka odabran za rešavanje opisanog problema objavljen je u okviru VizWiz izazova [5] i javno je dostupan [13]. U okviru izazova postoje različiti zadaci koji pokrivaju širok spektar oblasti kompjuterske vizije. Neki od postojećih zadataka su Image Captioning, Image Quality Issues, Visual Privacy, Reasons for Answer Differences, kao i tri posebna zadatka vezana za VQA. Za svaki od zadataka postoji poseban skup podataka. Primarni skup podataka upotrebljen u okviru ovog rada je Visual Question Answering skup [14].

A. Struktura skupa podataka

Odabrani skup podataka sačinjen je od fotografija snimljenih od strane slepih osoba, kao i pitanja vezanih za fotografije koja su kreirana na osnovu audio zapisa govora. Opisan način prikupljanja podataka pruža realističan kvalitet i tip slika, kao i pitanja, koji se mogu očekivati prilikom realne upotrebe sistema za pomoć slepim osobama. Shodno tome, postoje fotografije na osnovu kojih se ne može dati odgovor na pitanja koja su postavljena, kao i fotografije slabijeg kvaliteta (mutne fotografije i fotografije van fokusa). Za svaku fotografiju i pitanje vezano je i 10 odgovora na dato pitanje prikupljenih od strane nezavisne grupe pojedinaca (engl. *crowdsourcing*). Svaki odgovor takođe ima i naznaku da li je anotator siguran u dati odgovor (*answer_confidence – yes/no*). Dodatno, za svaku opisanu strukturu vezana je informacija o tipu pitanja i mogućnosti odgovora na isto (*answerable – yes/no*). Na slici 1 izdvojene su fotografije sa pitanjima i najčešćim odgovorom koje pripadaju različitim tipovima pitanja.

Skup podataka podeljen je na trening, validacioni i test skup. Fotografije i anotacije (pitanja i odgovori) su odvojeni zbog zauzeća memorije samih fotografija, i mogu se preuzeti pojedinačno. Vezu između anotacije i fotografije predstavlja ime fotografije. Anotacije su date u vidu tri JSON datoteke, vezane za svaki deo skupa, koje sadrže liste objekata koji okupljaju ime fotografije, pitanje, listu odgovora, tip pitanja i da li je na pitanje moguće dati odgovor.

Trening skup podataka sadrži 20 523 para fotografija i pitanja, kao i 205 230 odgovora na njih. Validacioni skup sadrži 4319 parova fotografija i pitanja, kao i 43 190 odgovora. Test skup sadrži 8000 parova fotografija i pitanja, bez odgovarajućih odgovora.



Slika 1 Izdvojene fotografije sa pitanjima i odgovorima iz WizViz skupa

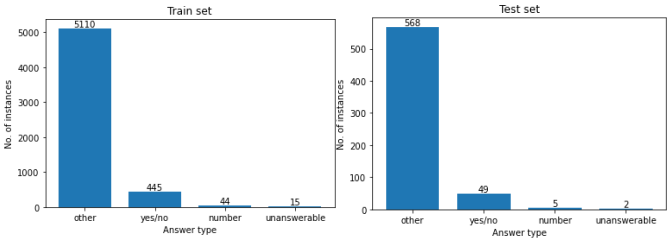
B. *Pretprocesiranje skupa podataka*

Daljim istraživanjem podataka primećen je veliki procenat fotografija koje zahtevaju Optical Character Recognition (OCR) tehnike kako bi se došlo do odgovora na pitanje. Neka od pitanja ovakvog tipa su:

- „What is the captcha on this screenshot?“
- „Hi, what can you see in this page that I have scanned with my phone camera?“
- „What book is this?“

Kako je odlučeno da OCR izlazi iz opsega projekta, neophodno je bilo odrediti koji sve parovi fotografija-pitanje zahtevaju OCR kao veštinu kompjuterske vizije. Postupak određivanja neophodnih veština koje sistem treba da poseduje kako bi odgovorio na pitanje je sam po sebi kompleksan i zahteva posebnu arhitekturu i posebno anotiran skup podataka. Kako bi opisan problem bio prevaziđen, upotrebljen je dodatni skup podataka iz VizWiz izazova. Vision Skills for VQA [15] je skup podataka koji se koristi za određivanje potrebne veštine iz polja kompjuterske vizije neophodne za odgovaranje na postavljeno pitanje vezano za fotografiju. Skup ima sličnu strukturu kao i primarni skup podataka na kom je baziran rad. Podeljen je na trening, validacioni i test skup i sadrži redom 14 259, 2248 i 5719 parova fotografija i odgovora. Anotacije potrebnih veština su redom: Text Recognition, Color Recognition, Object Recognition i Counting i svakom paru fotografija-pitanje pridružena je vrednost od 0 do 5 za svaku od pomenutih osobina. Javno dostupne anotacije postoje samo za trening skup [16].

Spajanjem trening i validacionog skupa početnog VQA skupa podataka sa anotiranim (trening) delom skupa Visual Skills for VQA, izbacili smo sve fotografije koje su zahtevale nivo Text Recognition veštine iznad 1. Nakon filtriranja početnog skupa podataka, novodobijeni skup podeljen je na trening i test skup u odnosu 90:10. Prilikom podele posebna pažnja posvećena je očuvanju odnosa tipova pitanja u okviru podskupova. Na slici 2 prikazani su udeli svakog od tipova pitanja u okviru trening i test skupa. Na ovaj način kreirani je trening skup sa 5614 i test skup sa 624 para fotografija i pitanja, sa odgovarajućim brojem odgovora.



Slika 2 Zastupljenost tipova pitanja u okviru trening i test skupa

III. METODOLOGIJA

Kako bi model uspešno rešio VQA problem potrebno je da poseduje mogućnost razumevanja postavljenog pitanja i slike na koju se ono odnosi. Uzimajući u obzir da se pitanje vrlo često odnosi samo na određeni deo, a ne na celu sliku, korisno je pronaći potencijalne regione od interesa [8] i formirati pogodnu reprezentaciju osobina svakog od njih. Nakon određivanja feature vektora regiona slike i samog pitanja, upotrebom sloja pažnje (engl. *attention*) određuje se relevantnost regiona za zadato pitanje i svakom od njih pridružuje se određena težina. Sledeći korak jeste formiranje jedinstvene reprezentacije čitave slike jednim vektorom, na osnovu težinskih vrednosti za regione, sa kojim će se zatim kombinovati feature vektor pitanja, čime se ulaz svodi na jedan vektor koji nosi informacije o oba ulaza i vezama među njima. Navedeni vektor predstavlja ulaz u multi-label klasifikator, čiji je broj mogućih odgovora jednak veličini rečnika formiranog na osnovu odgovora dostupnih u trening skupu. U daljem tekstu biće detaljno opisan svaki od delova predložene arhitekture, čiji je prikaz dat na slici 3, i biće data objašnjenja za donete odluke.

A. *Rukovanje fotografijama*

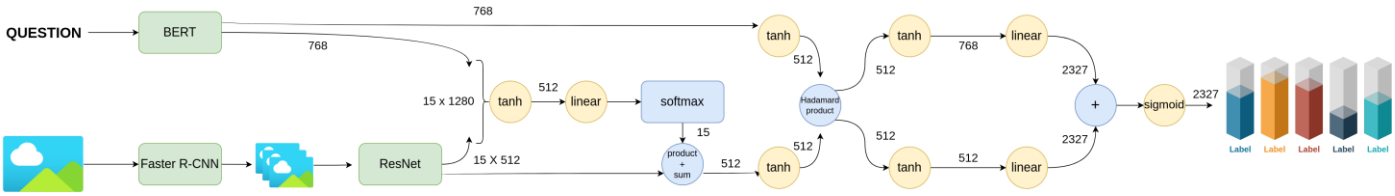
Početni korak transformacije jeste detektovanje regiona od interesa, koji predstavljaju objekte na toj fotografiji. Za taj zadatak upotrebljena je Faster R-CNN mreža za predlaganje regiona [21], koja kao kičmu (engl. *backbone*) koristi ResNet CNN [22] sa 50 slojeva. Čitava mreža pretrenirana je na MIcrosoft COCO skupu podataka [23]. Izlaz iz modela predstavlja niz regiona određenih koordinatama i svakom je pridružena vrednost između 0 i 1 koja predstavlja sigurnost predložene detekcije. Prag sigurnosti iznad kog su regioni prihvaćeni i poslati na dalju obradu je 0.5. Broj regiona varirao je od 0 do 100. Kako je samo 0.89% fotografija imalo iznad 15 regiona, maksimalan broj je ograničen na tu vrednost radi efikasnosti, a preostali regioni su odbačeni. Problem su predstavljale slike na kojima nisu detektovani objekti od interesa, a njih je bilo 28% u trening skupu. Kako takve fotografije nisu proizvodile nikakav ulaz u fazi ekstrakcije osobina, pridružen im je jedan region koji predstavlja čitavu fotografiju. Nakon pridruživanja, prosečan broj regiona po slici iznosi 2.52.

Kada je svakoj od ulaznih fotografija pridružen niz regiona dužine između 1 i 15, bilo je potrebno odrediti feature vektor svakog od njih. Na ResNet mrežu sa 18 slojeva pretreniranu na ImageNet skupu podataka [24] poslat je svaki izdvojeni region svih fotografija. Regioni su pre prosledjivanja svedeni na fotografije dimenzija 224x224 isecanjem i skaliranjem. Preuzet je izlaz enkodera mreže koji predstavlja vektor dužine 512.

Ulazne fotografije ne pripadaju nekom specifičnom i uskom domenu, već mogu biti raznovrsne prirode i sadržaja. Stoga je doneta odluka da se opisane transformacije obave u fazi pretprocesiranja, kako specijalizacija korišćenih modela njihovim dodatnim treniranjem ne bi donela benefite, dok je na ovaj način faza treniranja učinjena efikasnijom.

B. *Rukovanje pitanjima*

Formiranje feature vektora pitanja u tekstualnom formatu izvršeno je pomoću BERT transformer modela sa 12 slojeva enkodera, pretreniranog na BooksCorpus [25] skupu podataka i engleskoj Wikipedia enciklopediji [26]. Priprema ulaza za BERT podrazumeva WordPiece tokenizaciju [27] koja ulazni tekst deli na reči ili delove reči koje su deo rečnika. Prvi token uvek predstavlja specijalni [CLS] token nakon čega slede



Slika 3 Arhitektura predloženog sistema za rešavanje VQA problema

tokeni ulaznog teksta. Pitanje predstavlja jednu rečenicu i stoga se ne koristi [SEP] token koji ih razdvaja. Maksimalan broj ulaznih tokena koji BERT podržava je 512, ali sva pitanja su znatno kraća od te granice tako da se nije javila potreba za njihovim skraćivanjem i uklanjanjem reči. Za feature vektor čitavog pitanja očitana je vrednost poslednjeg skrivenog stanja koje odgovara [CLS] tokenu, kako bi se preuzela agregacija semantike čitavog pitanja [10] u jedan vektor dužine 768. Iz istih razloga kao i pri rukovanju fotografijama, ova transformacija obavljena je u fazi pretprocesiranja.

C. *Rukovanje odgovorima*

Svaki par fotografija-pitanje za sebe ima vezano 10 odgovora koji su sakupljeni postupkom *crowdsourcing*-a. Kako su odgovori unošeni u slobodnoj formi, postoje slučajevi gde su različite osobe na drugi način unele isti ili sličan odgovor. Kako bismo smanjili korpus odgovora i izbacili slične odgovore koji ne doprinose semantici, primenjeni su sledeći koraci pretprocesiranja:

- pretvaranje Unicode karaktera u ASCII,
- izbacivanje belina (razmaka, tabova i novih redova),
- izbacivanje specijalnih karaktera,
- svođenje svih odgovora na mala slova,
- svođenje odgovora na numerička pitanja (*answer_type* = *number*) na samo brojeve.

Zbog velike dimenzionalnosti multilabel problema gde svaki par fotografija-pitanje možemo vezati sa 1 do N odgovora (gde je N ukupan broj odgovora), odabrani su samo odgovori za koje su anotatori naveli da su sigurni u njih (*answer_confidence* = *yes*). Takođe, uzimajući u obzir namenu sistema, uzeti su samo odgovori oko kojih su se barem dva anotatora složila. Pitanja koja nemaju nijedan ovakav odgovor se labeliraju kao *unanswerable*. Na opisani način, kreiran je korpus od 2327 različitih odgovora na osnovu trening skupa. Odgovori prisutni u test skupu su pretprocesirani na isti način i zatim je proveravano da li se nalaze u već kreiranom skupu odgovora. Očekivano, određeni broj (68) odgovora na pitanja u okviru test skupa nije se našao u korpusu odgovora, i u tim slučajevima, sve labele su postavljene na 0.

Problem odgovora koji se sastoje od više reči nije prevaziđen na predložen način. Dodatno je isprobana upotreba POS tagging [13] mehanizma u pokušaju određivanja glavne reči u okviru odgovora. Ideja je bazirana na odbacivanju dodatnih prideva/priloga i izdvajanja glavne imenice/glagola, ali zbog strukture odgovora, uspešnost nije bila visoka i ideja je odbačena.

D. *Attention sloj vođen pitanjem*

Izdvojenim regionima fotografije, odnosno njihovim feature vektorima, dodeljuje se težina u zavisnosti od toga koliku važnost imaju za postavljeno pitanje, a u tu svrhu koristi se attention mehanizam. Ovaj sloj kao ulaz preuzima vektor osobina fotografije dimenzija 512x15 i vektor osobina pitanja dimenzije 1x768. Za fotografije sa manje od 15 regiona nedostajući feature vektori popunjeni su nulama, što ne utiče na rezultujući vektor ovog sloja zbog prirode operacije koje se sprovode. Svaki vektor regiona (sem onih popunjenih samo nulama, kojima se dalje konkatenira samo vektorom nula) konkatenira se sa vektorom pitanja i zatim se šalje kao ulaz nelineranom sloju veličine 512 sa tangens hiperbolik aktivacionom funkcijom, čiji se izlaz dalje prosleđuje linearnom sloju koji proizvodi skalarnu vrednost. Nakon što se za svaku kombinaciju regiona i pitanja preuzme izlaz linearnog sloja, primenjuje se softmax funkcija čiji izlaz su težine kojima se množe vektori odgovarajućih regiona, a zatim sabiraju kako bi se dobio jedan vektor koji reprezentuje čitavu fotografiju.

E. *Klasifikator*

Vektor pitanja i težinski vektor slike, koji je izlaz attention sloja, prolaze kroz nelinearne slojeve sa tangens hiperbolik aktivacionom funkcijom koji daju izlaz dužine 512. Takva dva izlaza spajaju se operacijom Hadamardovog proizvoda [28]. Nakon toga mreža se deli na dve grane gde svaka poseduje nelinearan sloj praćen lienarnim slojem, čije se dimenzije mogu videti na slici 3. Izlaz iz oba linearna sloja je veličine formiranog rečnika ponuđenih odgovora, odnosno 2327. Ta dva vektora sabiraju se i zatim prolaze kroz poslednji, ovog puta sigmoid, sloj. Odabir sigmoid aktivatione funkcije potiče od prirode problema koji predstavlja multi-label klasifikaciju. Poslednji sloj koristi i dropout mehanizam [29], sa stopom 0,67. Motivacija za obučavanje samo trećine poslednjeg sloja mreže pri svakom prolasku potiče iz činjenice da bi delovi mreže nakon dovoljno epoha mogli početi da se specijalizuju za sliku i pitanje koji pripadaju određenom tipu odgovora koji spada pod answerable pitanja (3 tipa – yes/no, number, other). Klasifikator je formiran tako da rešava multi-label klasifikaciju iz razloga što su neki od anotatora pri davanju odgovora davali odgovore koji su sinonimni ili pitanje prosto poseduje više različitih validnih odgovora, a poželjno je da mreža razume koji odgovori su slični i da ne isključuje sve sem jednog odgovora.

F. *Treniranje*

Mreža se trenira AdaDelta algoritmom zbog svojih osobina da ne zahteva podešavanje stope učenja kroz vreme, a za početnu stopu odabrana je vrednost 1, kao što je predstavljeno u radu [30] u kom je i sam algoritam predstavljen. Za funkciju

gubitka odabrana je Binary Cross-Entropy, pogodna za multi-label klasifikacione probleme. Broj primeraka iz trening skupa (engl. *batch size*) pri svakom prolasku je 30.

Kako bi se sprečio overfitting mreže, iz trening skupa izdvojen je deo podataka (10%) za formiranje validacionog skupa, koji je poslužio kao signal za ranije zaustavljanje treniranja ukoliko kategorička tačnost (engl. *categorical accuracy*) kroz više od 50 epoha nije rasla. Od svih stanja modela tokom treniranja, sačuvano je samo ono u trenutku kada je kategorička tačnost nad validacionim skupom bila najviša.

IV. Rezultati

Performanse modela izmerene su pomoću VQA metrike koju VizWiz izazov propisuje, a koja originalno potiče iz VQA izazova [31]. Svako pitanje iz test skupa poseduje deset odgovora koje su anotatori naveli. Kako se mera sličnosti predikcije i tačnih odgovora bazira na tome koliko anotatora je dalo određeni odgovor, odnosno prihvata se samo jedna predikcija, a problem smo postavili kao multi-label klasifikaciju, odabran je odgovor koji je dao vrednost najbližu jedinici. Formalni zapis metrike je sledeći:

$$\text{accuracy} = \min\left(\frac{\#\text{ humans that provided that answer}}{3}, 1\right)$$

Ostvareni rezultati nad test skupom prikazani su u tabeli 1. Date su performanse koje model postiže za svaki tip odgovora pojedinačno, kao i ukupan rezultat koji je ostvaren nad celim test skupom. Broj pitanja u test skupu prema tipovima odgovora može se pronaći na slici 2. Mreža se najbolje pokazala na pitanjima koja zahtevaju da/ne odgovor i ostvarila je rezultat 63.95, dok su performanse najslabije za odgovore koji predstavljaju brojeve i iznosi 6.67. Tip unanswerable ima rezultat 50, dok je za poslednji tip (ostalo), koji obuhvata sve odgovore koji ne pripadaju prva tri tipa, model pokazao rezultat 33.39. Ukupan rezultat koji je model ostvario nad čitavim test skupom iznosi 35.63.

Tabela 1 VQA metrika po tipovima odgovora, prikazana u procentima

	Tipovi odgovora				***Ukupno***
	Yes/No	***Numbers***	***Other***	***Unanswerable***	
VQA metrika	63.95	6.67	33.39	50	**35.63**

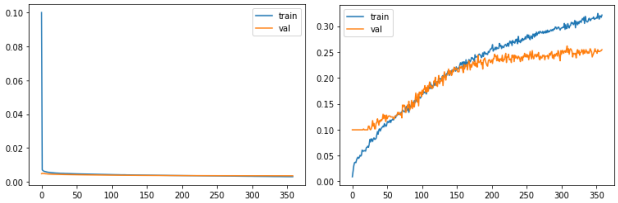
A. *Diskusija*

Rezultati koje model pokazuje variraju u zavisnosti od tipa odgovora, što je objašnjivo time da odgovori različitih tipova zahtevaju drugačije sposobnosti prepoznavanja i razumevanja slike i pitanja, od kojih su neke kompleksnije od drugih. Dok su pitanja čiji su odgovori da/ne tipa uglavnom jednostavnija i zahtevaju poznavanje manje detalja slike. Pitanja vezana za brojeve mogu obuhvatati neke ili sve od tehnika detekcije objekata, klasifikacije tih objekata, prepoznavanje teksta i slično.

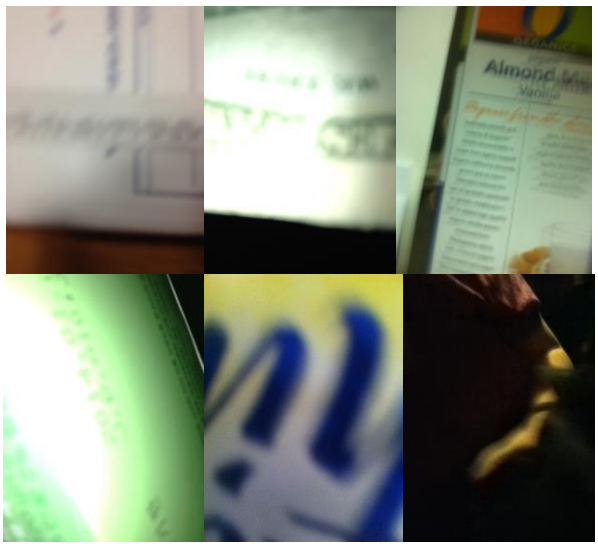
Problem koji se javio kod nekih složenijih pitanja i slika jeste nedovoljan broj njihovih primera; pitanja vezanih za brojeve u test skupu ima svega 5. Kako bi se model mogao dodatno istrenirati, a zatim i evaluirati na većem broju primeraka, jedno od daljih poboljšanja moglo bi uključiti proširenje skupova podataka. Iako neusklađenost u kompleksnosti pojedinačnih pitanja može dovesti do pojave da mreža preuči određene vrste pitanja, a bude nedovoljno istrenirana za druge, slika 4, na kojoj su prikazani gubitak i kategorička tačnost kroz epohe za vreme treniranja, pokazuju da mreža nije počela da uči napamet, jer gubitak i dalje opada na validacionom skupu dok tačnost raste. Stoga, povećanjem trening skupa stvara se prostor za dodatno učenje i poboljšanje performansi.

Dodatna prepreka prilikom obučavanja bile su i brojne fotografije koje poseduju loš kontrast i osvetljenje ili su mutne, što je prikazano na slici 5. Pokazatelj slabog kvaliteta je i veliki procenat (28%) fotografija na kom mreža za detekciju regiona nije uspela da pronađe nijedan region. Iz takvih fotografija je teško generisati feature vektore koji nose značajne informacije za dalji tok zaključivanja. Dodatna poboljšanja kvaliteta fotografija u fazi pretprocesiranja potencijalno bi mogla učiniti neke od fotografija pogodnijima za jasniju detekciju objekata, samim tim i za formiranje semantički bogatijih feature vektora.

Hardverska ograničenja igrala su ulogu prilikom donošenja nekih odluka u procesu dizajniranja modela. Jedan od primera jeste odabir ResNet mreže sa 18 slojeva, umesto



Slika 4 Vrednosti gubitka i tačnosti kroz epoha u procesu treniranja



Slika 5 Primeri fotografija iz trening skupa koje su lošeg osvetljenja ili kvaliteta

kompleksnijih sa 50 ili 152 sloja, koje na izlazu enkodera daju duže i informacijama bogatije vektore, što može biti ključna odlika za pitanja koja zahtevaju visok nivo poznavanja detalja fotografije. Takođe, ukupan broj težina koje se mogu trenirati prilikom obučavanja ograničen je tako se hardverski može podržati.

V. ZAKLJUČAK

Sistemi za pomoć slabovidim i slepim osobama izuzetno je značajan za poboljšanje njihovog sveukupnog životnog kvaliteta. Upotreba sistema koji daje odgovor na pitanja koja postavljaju vezanih za svoju okolinu dodatno bi doprinela njihovoj samostalnosti u izvršavanju dnevnih aktivnosti. Razvoj oblasti kompjuterske vizije i obrade prirodnog teksta doprineo je razvoji sve kvalitetnijih rešenja za sisteme koji se baziraju na rešavanju opisanog, VQA, problema.

U okviru rada predložena je arhitektura sistema koja bi mogla biti primenjena za rešavanje ovog problema. Predložena arhitektura mogla bi se integrisati u veći sistem koji bi na osnovu Speech-To-Text komponenti dobijao ulazno pitanje, i na osnovu Text-To-Speech komponenti odgovor davao korisniku.

Načini na koje bi se prikazano rešenje moglo unaprediti svakako podrazumevaju razmatranje drugačijih arhitektura finalnog klasifikatora i njegovo dodatno usložnjavanje. Drugi pravac unapređenja bi bio izbacivanje klasifikacije i implementaciju generatora odgovora kao finalnog segmenta arhitekture, što je predloženo u okviru [18][19][20]. Na ovaj način bi bio prevaziđen problem nemogućnosti da sistem odgovori na pitanje ako se prethodno nije susreo sa sličnim odgovorom.

LITERATURA

[1] Rawat, Waseem, and Zenghui Wang. "Deep convolutional neural networks for image classification: A comprehensive review." Neural computation 29.9 (2017): 2352-2449.

[2] Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." Physica D: Nonlinear Phenomena 404 (2020): 132306.

[3] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[4] Wu, Qi, et al. "Visual question answering: A survey of methods and datasets." Computer Vision and Image Understanding 163 (2017): 21-40.

[5] Gurari, Danna, et al. "Vizwiz grand challenge: Answering visual questions from blind people." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[6] Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

[7] Xu, Huijuan, and Kate Saenko. "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering." European conference on computer vision. Springer, Cham, 2016.

[8] Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[9] Teney, Damien, et al. "Tips and tricks for visual question answering: Learnings from the 2017 challenge." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[10] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[11] Kafle, Kushal, and Christopher Kanan. "Visual question answering: Datasets, algorithms, and future challenges." Computer Vision and Image Understanding 163 (2017): 3-20.

[12] Wu, Z. "Palmer, M.: Verbs Semantics and Lexical Selection." Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics, Las Cruces, New Mexico. 1994.

[13] https://vizwiz.org/ - posećeno 2.7.2022.

[14] https://vizwiz.org/tasks-and-datasets/vqa/ - posećeno 2.7.2022.

[15] https://vizwiz.org/tasks-and-datasets/vision-skills/ - posećeno 2.7.2022.

[16] https://github.com/chiutaiyin/Vision-Skills/blob/master/csv/vizwiz_skill_typ_train.csv - posećeno 2.7.2022.

[17] Marquez, Lluis, Lluis Padro, and Horacio Rodriguez. "A machine learning approach to POS tagging." Machine Learning 39.1 (2000): 59-91.

[18] Malinowski, Mateusz, Marcus Rohrbach, and Mario Fritz. "Ask your neurons: A neural-based approach to answering questions about images." Proceedings of the IEEE international conference on computer vision. 2015.

[19] Jiang, Aiwen, et al. "Compositional memory for visual question answering." arXiv preprint arXiv:1511.05676 (2015).

[20] Gao, Haoyuan, et al. "Are you talking to a machine? dataset and methods for multilingual image question." Advances in neural information processing systems 28 (2015).

[21] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).

[22] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[23] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[24] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[25] Zhu, Yukun, et al. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." Proceedings of the IEEE international conference on computer vision. 2015.

[26] https://www.wikipedia.org/ - posećeno 2.7.2022.

[27] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).

[28] Horn, Roger A. "The hadamard product." Proc. Symp. Appl. Math. Vol. 40. 1990.

[29] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15.1 (2014): 1929-1958.

[30] Zeiler, Matthew D. "Adadelta: an adaptive learning rate method." arXiv preprint arXiv:1212.5701 (2012).

[31] https://visualqa.org/challenge.html - posećeno 2.7.2022.