Sistem za generisanje opisa slika

Mentor: Milica Škipina Student: Vulin Svetozar SW-57-2018

Softversko Inženjerstvo i Informacione Tehnologije, FTN, Novi Sad

Uvod

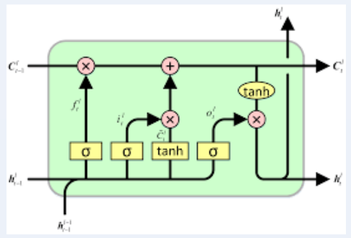
Sa razvojem neuronskih mreža, proširuje se i njihova primena na rešavanje različitih problema. Konvolucione neuronske mreže (CNN) su se izdvojile kao jedan od najboljih algoritama u rešavanju problema kompjuterske analize i čitanje slika, a rekurentne neuronske mreže (RNN) su se pokazale dobro u analizi i obradi teksta.

two dogs are playing in the grass

Motivacija i cilj

Koristeći CNN i posebnu vrstu RNN-a – LSTM, sistem za generisanje slika je u mogućnosti da za priloženu sliku uoči i prepozna karakteristike na istoj i da na osnovu toga izgeneriše njen opis. To se postiže tako što već gotova CNN mreža Xception, izdvajaja iz slike bitne karakteristike, i zatim se tim karakteristikama pridodaju opisi slike obrađenih uz LSTM mrežu i taj model se trenira da na osnovu dobijenih karakteristika predvidi tekst.

Xception



Podaci i obrada

Za potrebe ovog projekta, korišćen je Flickr8k dataset. Ovaj dataset se sastoji od 8000 slika, koje uz sebe imaju po 5 opisa. Izgled dataseta se može videti na primeru ispod:

```
1022975728_75515238d8.jpg#0     A black dog running in the surf .
1022975728_75515238d8.jpg#1     A black lab with tags frolicks in the water .
1022975728_75515238d8.jpg#2     A dog splashes in the water
1022975728_75515238d8.jpg#3     The black dog runs through the water .
1022975728_75515238d8.jpg#4     This is a black dog splashing in the water .
102351840_323e3de834.jpg#0      A man drilling a hole in the ice .
102351840_323e3de834.jpg#1      A man is drilling through the frozen ice of a pond .
102351840_323e3de834.jpg#2      A person in the snow drilling a hole in the ice .
102351840_323e3de834.jpg#3      A person standing on a frozen lake .
102351840_323e3de834.jpg#4      Two men are ice fishing .
1024138940_f1fefbdce1.jpg#0     Two different breeds of brown and white dogs play on the beach
1024138940_f1fefbdce1.jpg#1     Two dogs are making a turn on a soft sand beach .
1024138940_f1fefbdce1.jpg#2     Two dogs playing in the sand at the beach .
1024138940_f1fefbdce1.jpg#3     Two dogs playing together on a beach .
1024138940_f1fefbdce1.jpg#4     Two large tan dogs play along a sandy beach .
1024551765_5f8ead62d5.jpg#0     A man uses ice picks and crampons to scale ice .
```

Pre treniranja potrebno je obraditi podatke. Obrada podataka se deli u dve faze:

1. generisanje karakteristika slika
2. obrada teksta

Prva faza predstavlja generisanje karakteristika sa slika. Prilikom ove faze je korišćena *transfer learning* tehnika. Ovom tehnikom se podrazumeva korišćenje gotove i istrenirane neuronske mreže radi dobijanja određenih podataka koji će se dalje koristiti u treniranju sopstvenog modela. Korišćen je Keras-ov Xception, CNN model, koji se inače koristi za klasifikaciju slika. Skidanjem poslednjeg sloja modela je postignuto to da se izgenerišu samo bitne karakteristike iz slika, bez klasifikacije šta je na slici. Nakon pripremanja slike i puštanja kroz mrežu, karakteristike se dobijaju kao vektor od 2048 elemenata i čuvaju se u fajlu.

Druga faza je obrada teksta. Svaka slika dolazi sa 5 tekstualnih opisa koje je potrebno obraditi da bi se povećala preciznost predikcije. Obrada teksta podrazumeva:

1. prebacivanje teksta u mala slova
2. brisanje svih znakova interpunkcije ili specijalnih karaktera
3. brisanje prekratkih reči (a ili ‘s)
4. brisanje brojeva

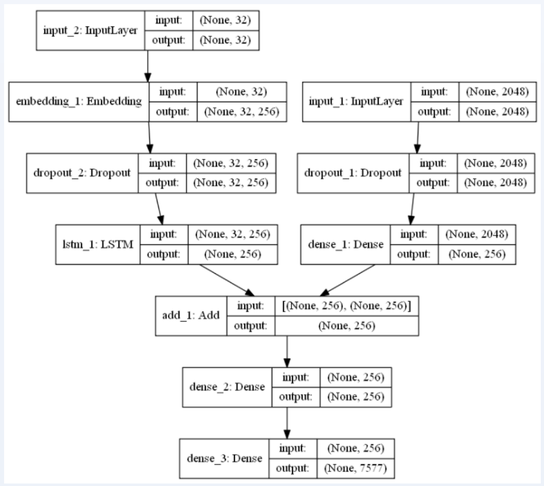
Nakon što je tekst obrađen, potrebno je pripremiti tekst za predikciju tako što će na početak svakog opisa biti dodato <start> i <end> na kraj. Ovo će pomoći pri predikciji, jer će služiti kao početna i krajnja tačka u generisanju teksta. Obrađen tekst se čuva u fajlu.

Treniranje mreže

Arhitektura mreže se sastoji iz dva ulaza, koji imaju vektore od po 2048 i 32 elementa. Drugi ulaz ima 32 elementa jer je to najveća dužina opisa iz dataset-a. Ulaz sa vektorom od 2048 elemenata služi za obradu karakteristika slika.

Sa druge strane, ulaz sa 32 elementa prima vektor reči iz opisa koje su pomoću Keras-ovog Tokenizer-a namapirane na brojne vrednosti. Svaka reč iz rečnika se predstavlja jednim brojem. Zatim se na taj ulazni sloj vezuje Embedding sloj koji koristi tehniku *word embedding.* To je tehnika koja uz dobijenu reč pravi vektorski prostor kojem pridodaje reči koje se često koriste u kontekstu sa posmatranom reči. Nakon ove obrade, dalje se koristi LSTM mreža koja vrši procesiranje teksta.

Ove dve strane se zatim kombinuju i propuštaju kroz još jedan sloj koji kreira izlaz od onoliko elemenata koliko reči postoji u vokabularu. Arhitektura je data na slici ispod.



Za treniranje mreže je izdvojen skup od 6000 slika. Zbog velikog broja slika i opisa, za treniranje bi bila potrebna velika kompjuterska moć, zato se prilikom treniranja koristi generator koji generiše na osnovu reči i slike sledeću reč. Zatim se te 2 reči i slika ponovo pošalju na predikciju, da bi se dobila 3. reč. Ovo se ponavlja sve dok se ne izgeneriše opis ili ne pređe maksimalna dužina dozvoljenog opisa. (U ovom slučaju 32)

Rezultati treniranja

Keras-ova evaluacija loss funkcije na kraju treniranja 10. epohe je prikazala vrednost oko 2.7. Nakon empirijskog testiranja, zaključak je bio da je došlo do *overfittinga.* Glavni pokazatelj ove teorije je bio da mreža na mnogo različitih slika daje sličan ili isti opis.

Tehnike koje su primenjene da bi se rešio problem overfittinga su:

- L2 regularizacija
- Promena dropout-a
- Promena optimizatora i stepena učenja
- Izmena arhitekture mreže

Rezultati nakon primenjenih tehnika se nisu znatno promenili, pa je zaključak da je možda problem u veličini skupa podataka što nije bilo moguće testirati zbog hardverskih ograničenja.

two people are sitting on bluff overlooking the ocean

two dogs are playing in the grass

Reference

- Chollet, François. "Xception: Deep learning with depthwise separable convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- https://colah.github.io/posts/2015-08-Understanding-LSTMs/