

Prevođenje teksta sa engleskog jezika na nemački jezik

Seminarski rad
Matematički fakultet,
Univerzitet u Beogradu

Dijana Zulfikarić, Stefan Pantić

February 18, 2019

Sadržaj

1	Opis problema	1
2	Zašto rešenja korišćenjem rekurentnih neuronskih mreža nisu dovoljno dobra?	1
2.1	Kako funkcionišu rekurentne neuronske mreže?	1
2.2	Šta mogu biti problemi u ovakvom pristupu?	2
3	Attention is all you need	2
3.1	Šta je osnovna ideja ovakvog pristupa?	2
3.2	Pristup koji je korišćen u rešavanju ovog problema	3
4	Pravljenje modela	4
4.1	Dobijanje podataka	4
4.2	Implementacija modela	4
5	Analiza treninga i testa	4
5.1	Predikcije	5
6	Primena modela na rešavanje drugačijeg problema	6
6.1	Prikaz rezultata i analiza rešenja	6
6.1.1	Grafički prikaz treninga	6
6.1.2	Predikcije	8
7	Zaključak	9

1 Opis problema

Problem kojim smo se bavili u ovom radu jeste generisanje prevoda engleskih rečenica na nemački jezik. Kao jedan od problema kojem je bilo pristupa na različite načine, postojalo je više algoritama kojima smo se mogli baviti u ovom slučaju. Za nas je najinteresantniji pristup bio pristup korišćen u radu [Attention is all you need](#).

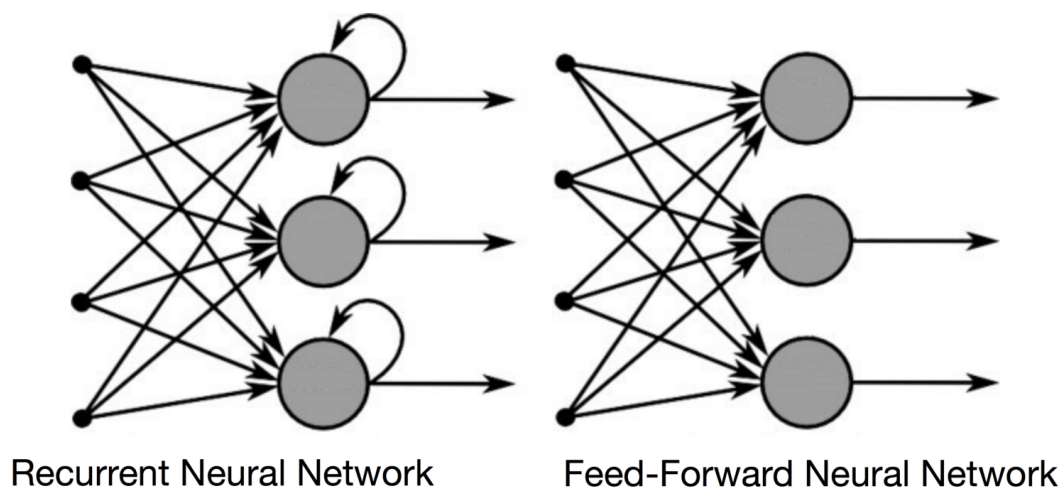
2 Zašto rešenja korišćenjem rekurentnih neuronskih mreža nisu dovoljno dobra?

Jedan od najkorišćenijih pristupa prilikom rešavanja ovakvih problema jeste korišćenje *rekurentnih neuronskih mreža*. Ovaj pristup, bez obzira na svoje mogućnosti, ima i određenih nedostataka koji će biti navedeni u narednom delu.

2.1 Kako funkcionišu rekurentne neuronske mreže?

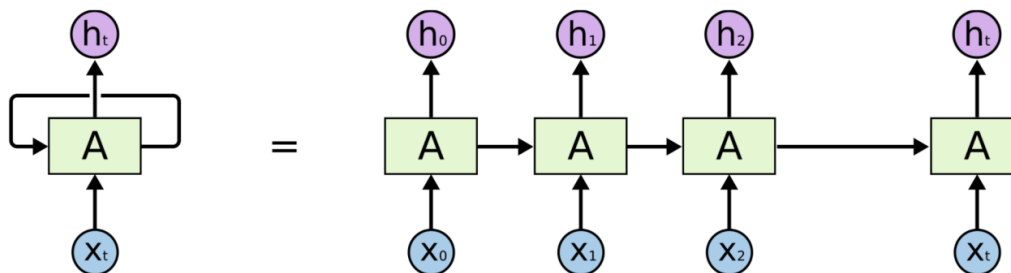
Rekurentne neuronske mreže su trenutno industrijski standard za obradu sekvencijalnih podataka iz razloga što predstavljaju prvi algoritam koji pamti pređašnja stanja i uzima ih u obzir za generisanje novih izlaza. Bez obzira na to što sam algoritam nije toliko nov, svoj pun potencijal mogao je iskazati tek u nekoliko prethodnih godina zbog značajnog pojeftinjenja hardverskih komponenti. Ideja rekurentnih neuronskih mreža jeste da pamte značajne informacije koje su dobile, što omogućava povećanje preciznosti prilikom predikcija narednog izlaza. Upravo zbog toga su preferirani algoritam koji se koristi za obradu sekvencijalnih podataka kao što su vremenske serije, obrada govora, teksta, zvuka itd.

Za razliku od običnih neuronskih mreža sa propagacijom unapred kod kojih se informacija prenosi samo u jednom pravcu (od ulaznog sloja, preko skrivenih sloja do izlaznog sloja) tako da informacija nikad ne dolazi do istog čvora dva puta, kod rekurentnih neuronskih mreža informacija se ciklično vraća u čvorove. Zbog ovoga obične mreže ne mogu da se "sete" nijedne informacije iz prošlosti, osim onih koje su dobijene prilikom treninga. Na sledećoj slici može se videti poređenje strukture ove dve vrste neuronskih mreža.



Obične rekurentne neuronske mreže imaju kratkotrajno pamćenje, dok u kombinaciji sa LSTM-ovima dobijaju i dugotrajno pamćenje. Još jedan slikovit primer rada ovih mreža u poređenju sa običnim mrežama sa propagacijom unapred može biti generisanje reči "neuron". Naime, u trenutku kada mreža sa propagacijom unapred dođe do karaktera "u", ona je već zaboravila pređašnje karaktere "n" i "e", što čini zadatak predviđanja narednog karaktera gotovo nemogućim. Rekurentna neuronska mreža zahvaljujući svojoj unutrašnjoj memoriji ove informacije zadržava, pa je i rešavanje ovakvog problema znatno lakši.

Ovakva arhitektura neuronske mreže poseduje dva ulaza - prvi predstavlja sadašnjost, a drugi prošlost. Rekurentne neuronske mreže računaju i ažuriraju težine na oba ulaza korišćenjem gradijentnog spusta ili neke modifikacije gradijentnog spusta. Na slici ispod grafički je razmotan sloj rekurentne mreže. \mathbf{X} predstavlja ulaznu sekvencu, a \mathbf{H} predstavlja izlaznu sekvencu. Možemo videti da je izlaz iz prvog čvora takođe jedan od ulaza u sledeći čvor.



2.2 Šta mogu biti problemi u ovakvom pristupu?

Problemi prilikom korišćenja rekurentnih neuronskih mreža mogu biti nestajući ili eksplodirajući gradijenti, koji se rešavaju uvođenjem LSTM-a. LSTM (*Long Short Term Memory*) predstavlja poboljšanje rekurentnih mreža uvođenjem dužeg "pamćenja". Sama LSTM ćelija sadrži tri tzv. *kapije* - ulazna, izlazna i zaboravljanje. Ulazna kapija odlučuje da li je novi ulaz potreban za dalju propagaciju, izlazna kapija odlučuje da li je potrebno da izlaz iz trenutnog čvora utiče na izlaz iz trenutnog vremenskog trenutka, i kapija zaboravljanja odlučuje koliko je prethodnih informacija relevantno za dalji tok učenja u trenutnom vremenskom trenutku.

Jasno je da uvođenjem LSTM-a možemo rešiti neke od problema koji nastaju zbog same arhitekture rekurentnih neuronskih mreža, ali i da to znatno usložnjava i usporava učenje. Takođe, mogu se javiti problemi ukoliko su sekvence koje se koriste za učenje predugačke, jer se ipak gube informacije zbog nestajućih gradijenata, zbog čega se predugačke zavisnosti ipak ne mogu naučiti.

3 Attention is all you need

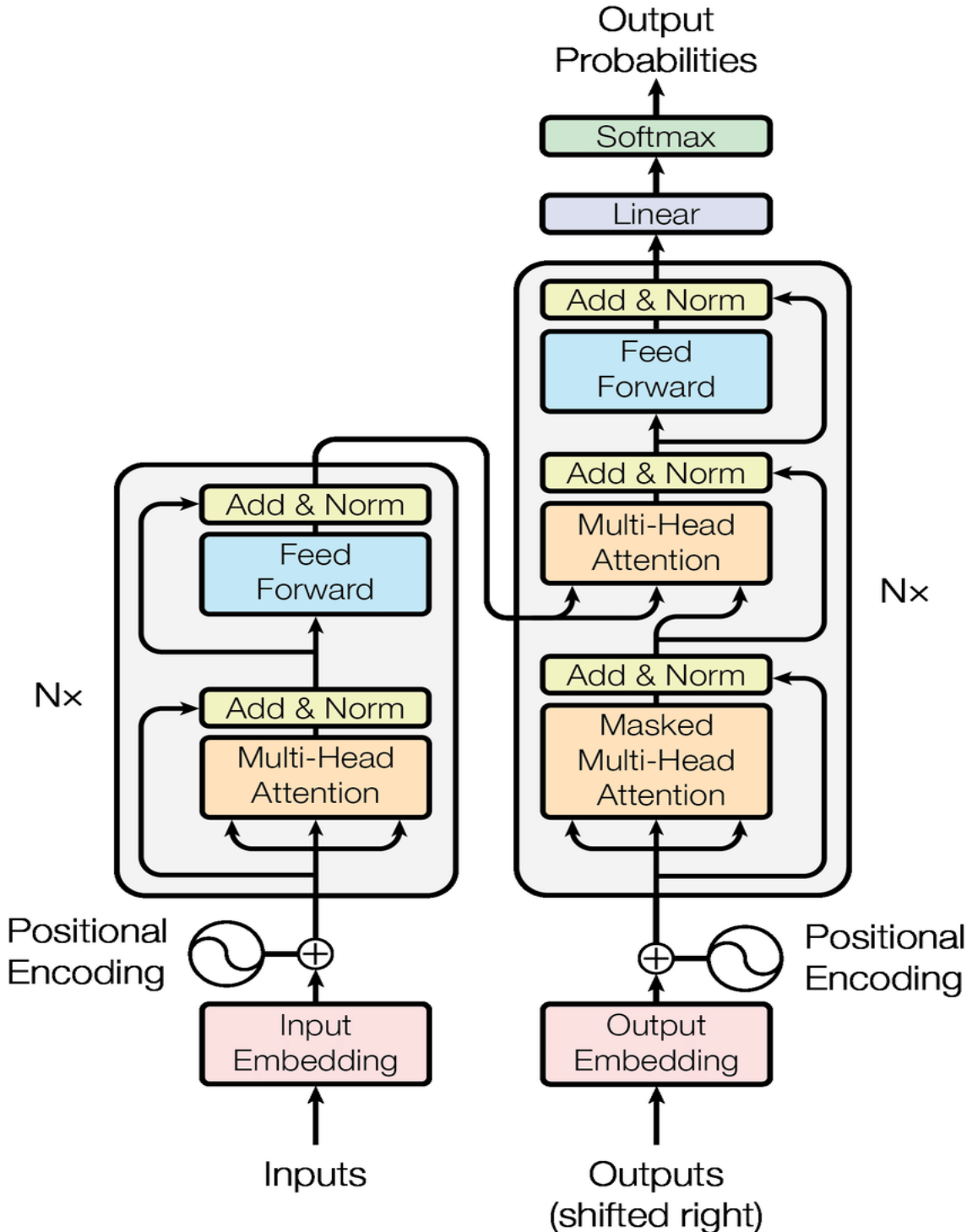
U narednom delu pozabavićemo se opisom nove arhitekture koja se može koristiti prilikom analize sekvencijalnih podataka bez korišćenja rekurentnih neuronskih mreža.

3.1 Šta je osnovna ideja ovakvog pristupa?

Umesto korišćenja standardnog RNN pristupa, arhitektura *Transformer* koristi *Attention* mehanizme kako bi procesirao ulazne sekvence i učio semantička mapiranje između izvorne rečenice i njenog prevoda. Ovakva arhitektura može se posmatrati kao dvoslojna arhitektura u kojoj prvi sloj predstavlja **enkoder** a drugi **dekoder**. Oba sloja vrše isto procesiranje ulaza, imaju *embedding* sloj koji je praćen pozicionim enkodiranjem. Zbog toga što se ne koristi klasičan RNN pristup koji prirodno enkodira sekvencijalne zavisnosti, potrebno je da ih ručno uvedemo. Ovo je urađeno korišćenjem sinusoida različitih frekvencija koje uvode instinktivnu vremensku komponentu podacima, ali i dopuštaju da se podatak procesira u celosti u svakom trenutku.

3.2 Pristup koji je korišćen u rešavanju ovog problema

Arhitektura **enkodera** sastoji se od proizvoljnog broja *Multi-Head Scaled Dot-Product Attention* (*MHDPA*) blokova koji uzimaju poziciono-ekodirane podatke i iz njih izvlače *key*, *query*, *value* trojku, enkodira relacije korišćenjem skalarnog proizvoda i proizvodi izlaznu sekvencu. Arhitektura **dekodera** sastoji se od proizvoljnog broja *Masked MHDPA* i *MHDPA* blokova i vrši slične operacije enkoderu. Razlika je u tome što se ključ i vrednost enkodera prosleđuju dekoderu, dok je *query* koji ulazi u dekoder izlaz iz *Masked MHDPA* bloka. Izlaz dekodera propagira se kroz jedan linearni sloj koji ima *softmax* aktivacionu funkciju i proizvodi izlaznu sekvencu. Na narednoj slici predstavljena je arhitektura modela **Transformer**:



4 Pravljenje modela

4.1 Dobijanje podataka

Podaci predstavljaju fajlove u kojima se nalaze rečenice na engleskom jeziku i korespondirajuće rečenice na nemačkom jeziku. Mogu se preuzeti pokretanjem sledeće komande:

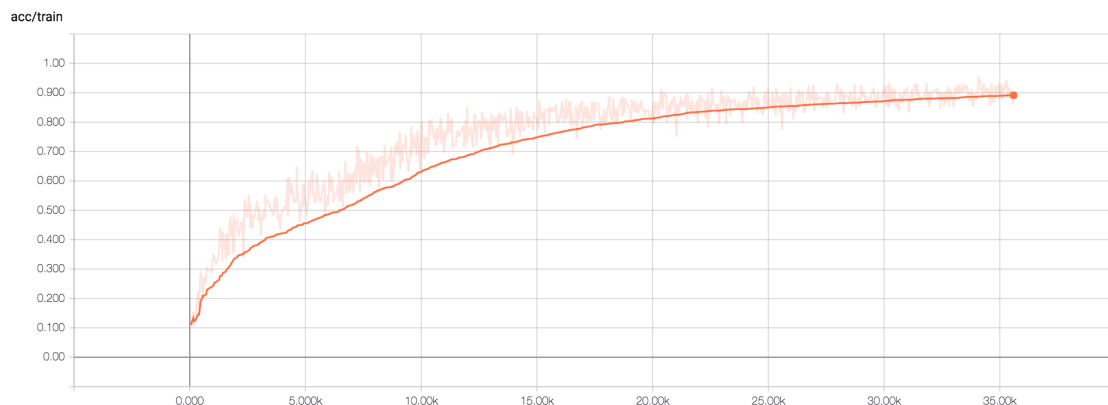
```
1 wget -qO- --show-progress https://wit3.fbk.eu/archive/2016-01//texts/de/en/de-en.tgz
2 tar xz; mv de-en data
```

4.2 Implementacija modela

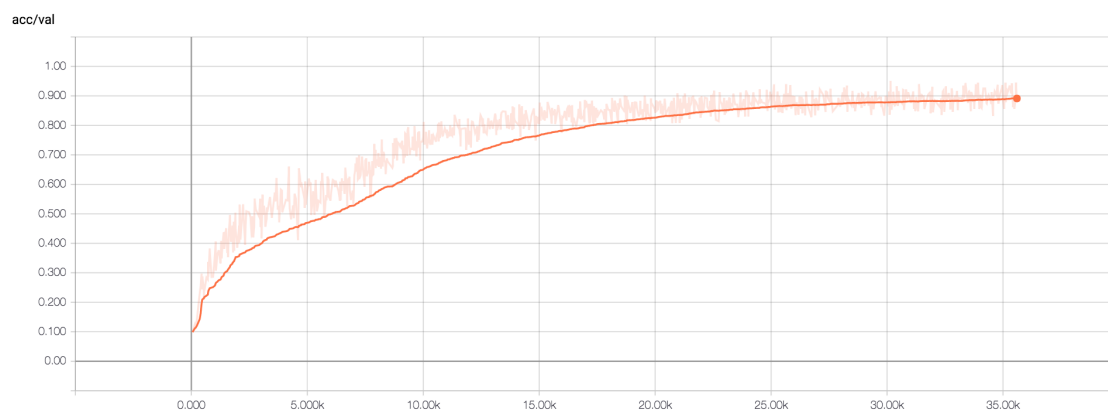
Prilikom rada osnovna biblioteka koja je bila korišćena je biblioteka *tensorflow*, dok su ostale pomoćne biblioteke navedene u datoteci [requirements.txt](#). Ostali implementacioni detalji i kodovi mogu se pogledati na [GitHub repozitorijumu](#).

5 Analiza treninga i testa

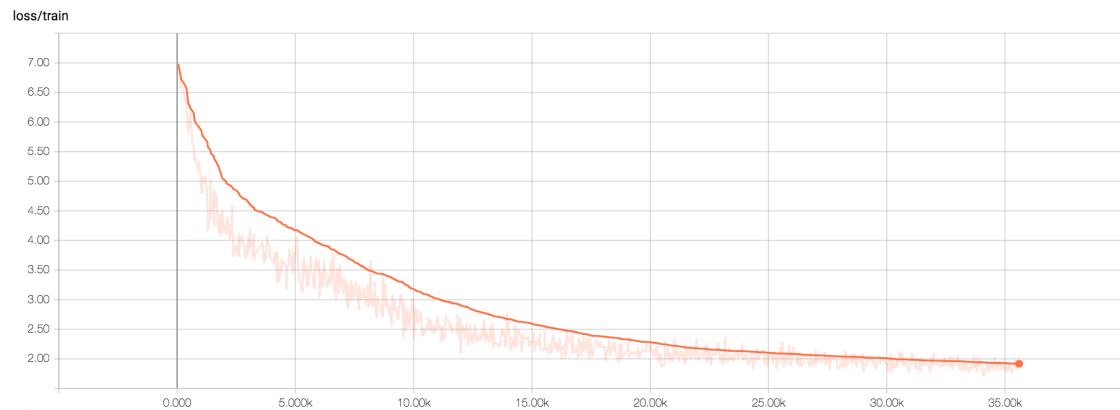
Prilikom procesa učenja pratili smo preciznost i funkciju gubitka koja je bila optimizovana koršćenjem tehnike gradijentnog spusta (kao funkcija greške korišćena je *categorical cross-entropy*). Slede slike grafika dobijene prilikom učenja na trening skupu i skupu za validaciju sa korišćenjem alata *tensorboard*:



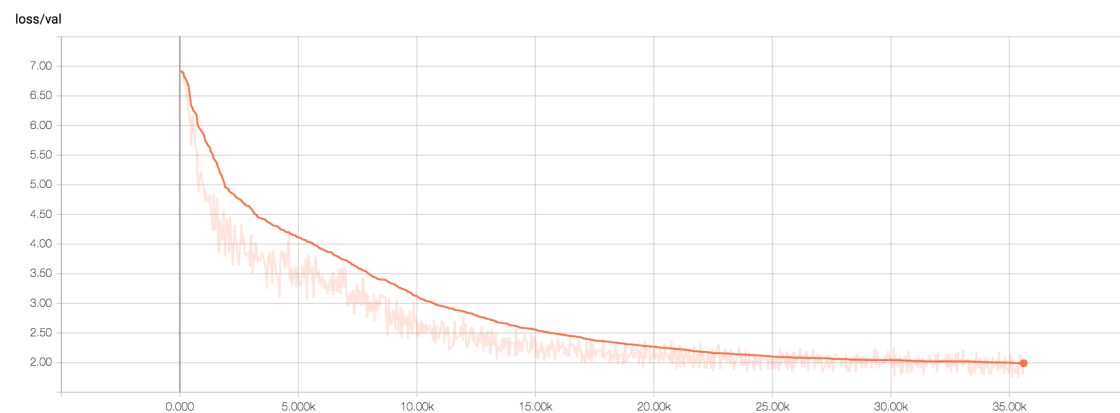
Preciznost na trening skupu.



Preciznost na skupu za validaciju.



Funkcija gubitka na trening skupu.



Funkcija gubitka na skupu za validaciju.

5.1 Predikcije

U sledećem delu prikazaćemo nekoliko nasumično odabranih primera test skupa kao i predikciju modela za iste.

input: And it wasn't social intelligence
 true: Uns es war nicht soziale Intelligenz
 prediction: Könnte es war nicht tief Werbung

input: This is a mountain
 true: Das ist ein Berg
 prediction: Das ist ein Medikament

input: <UNK> wear funny white <UNK>
 true: <UNK> tragen <UNK> weiße <UNK>
 prediction: <UNK> folgen <UNK> Hey <UNK>

input: They pretty much do nothing
 true: Sie machen eigentlich gar nichts
 prediction: Sie machen eigentlich gar nichts

input: He's a <UNK> just like <UNK> inside the whale
 true: Er ist ein <UNK> so wie <UNK> im <UNK>
 prediction: Er ist ein <UNK> so wie <UNK> im <UNK>

input: Great memories are learned
true: Gute <UNK> sind <UNK>
prediction: John <UNK> sind <UNK>

input: Because she's <UNK> everybody knows it
true: Weil sie <UNK> ist und jeder weiß das
prediction: Weil sie <UNK> ist und jeder weiß das

input: I don't know what I'm supposed to do
true: Ich weiß nicht was ich machen soll
prediction: Ich weiß nicht was ich machen soll

6 Primena modela na rešavanje drugačijeg problema

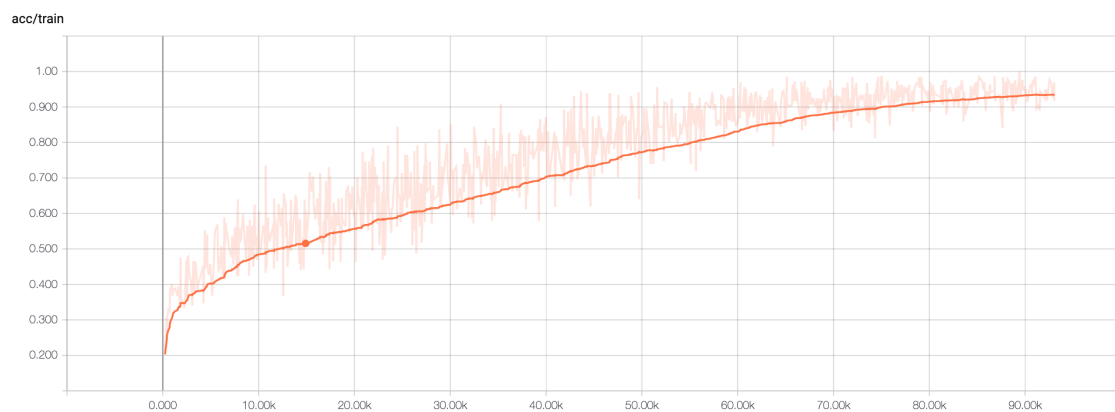
Nakon korišćenja modela u rešavanju problema prevođenja, rešili smo da isprobamo dobijeni model na drugačiji tip problema. Odlučili smo da pokušamo da sumirujemo recenzije hrane sa Amazona. Skup podataka može se naći na sajtu [Kaggle](#). Sam model nije menjan, dok je od hiperparametara promenjena dužina ulaznih i izlaznih sekvenci zbog značajnog uvećanja dužine ulazne i izlazne sekvence u odnosu na jednu rečenicu koja je bila u slučaju prevođenja. Takođe, sam skup podataka je znatno veći od predašnjeg, pa je radi dodatne optimizacije korišćena biblioteka *Ray* za omogućavanje asinhronog izvršavanja. Model je treniran približno 96h, u poređenju sa prethodnim treningom koji je trajao oko 32h.

6.1 Prikaz rezultata i analiza rešenja

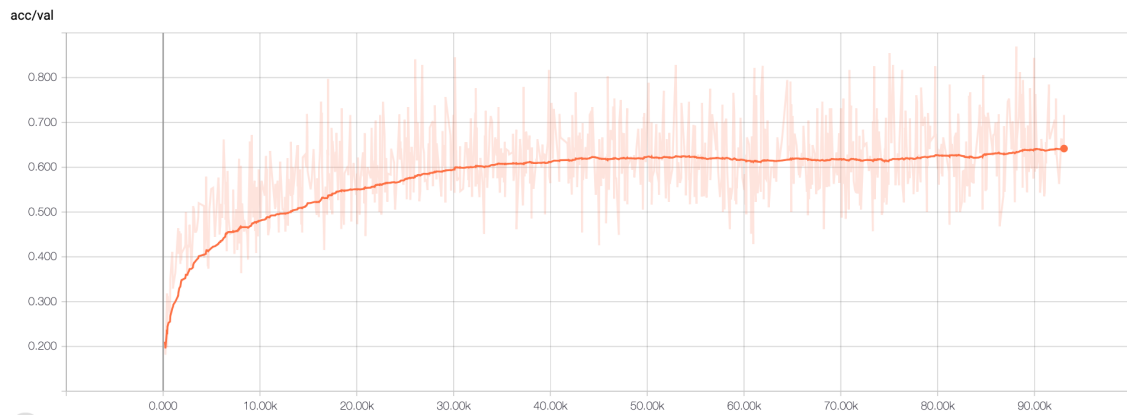
U narednim segmentima bavićemo se analizom treninga i rezultata izvršavanja prikazanih na nekoliko primera.

6.1.1 Grafički prikaz treninga

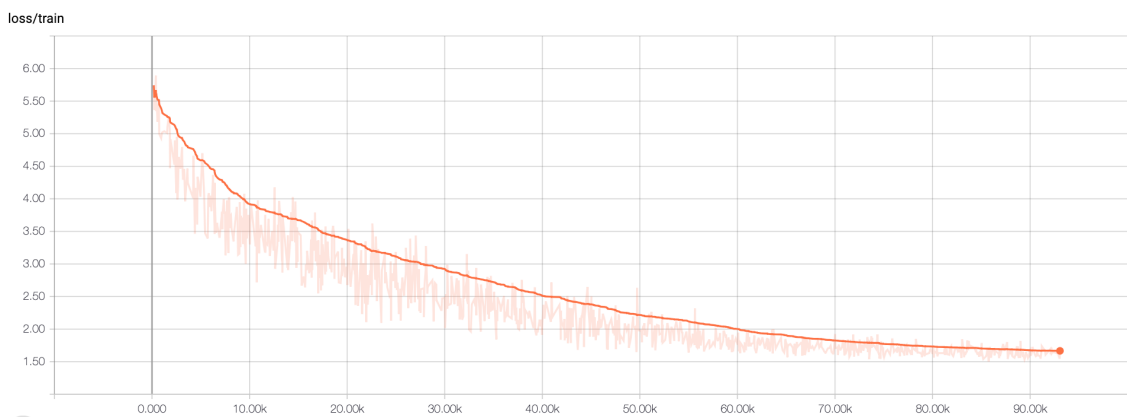
Na narednim slikama možemo videti preciznost i funkciju gubitka koja je optimizovana na isti način kao u prethodnom slučaju. Slede slike grafika dobijene prilikom učenja na trening skupu i skupu za validaciju sa korišćenjem alata *tensorboard*:



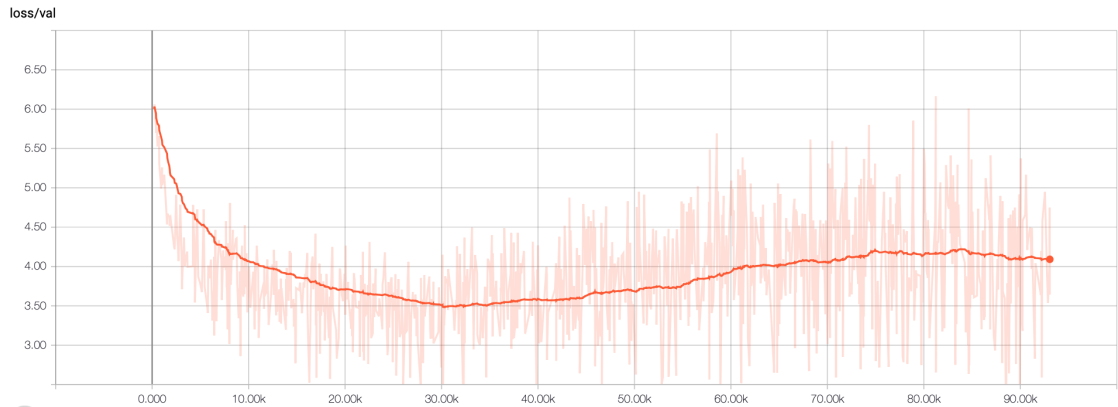
Preciznost na trening skupu.



Preciznost na skupu za validaciju.



Funkcija gubitka na trening skupu.



Funkcija gubitka na skupu za validaciju.

Praćenjem greške na trening skupu i skupu za validaciju, možemo videti da je model počeo da se prilagođava trening skupu nakon približno 48h. Takođe, bitno je napomenuti da je maksimalna dostignuta preciznost oko 80%, što je za 15% slabiji rezultat nego prilikom pokretanja modela pri rešavanju prvog problema, te stoga dolazimo do zaključka da se model mnogo bolje snalazi sa kraćim sekvencama, dok je za duže sekvence nailazio na veće probleme prilikom učenja pravilnosti.

6.1.2 Predikcije

U sledećem delu prikazaćemo nekoliko nasumično odabranih primera test skupa kao i predikciju modela za iste.

```
input:      <UNK> <UNK> <UNK> <UNK> br <UNK> dog's favorite treat is <UNK> chicken
<UNK> I did a <UNK> taste test with treats this one the chicken <UNK> and a regular
<UNK> biscuit expecting her to take the <UNK> first but then see how fast she came
back for the Newman's Own Dog treat To my actual surprise she took the Newman's Own
first This was the first time I ever offered it to her I usually give her a few
treats in the morning and she smells them all and <UNK> her favorite <UNK> br I
will definitely be buying more <UNK> out the <UNK> I asked her to give the treat
back to me she's a good dog and choose again just to make sure She chose the
Newman's own again and <UNK> it She did come back for the chicken <UNK> when she
was done with the Newman's Own She left the Milk <UNK> where it <UNK> br In
addition to her liking it I liked that the ingredient list was so simple without
<UNK> ingredients ie it was Barley Lamb <UNK> If I like the ingredient list and she
loves the treat that's all I need to <UNK> br Very very happy with this product
true:      <UNK> that my dog chose this over her favorite treat
predicted: <UNK> fed a dog chose this over her favorite food
```

```
input:      The <UNK> Microwave Pork rinds that I received was <UNK> I ordered bags
and the bags only popped of the bag All rinds popped but for the amount of the
money that I paid for them I could have gone to the store and bought a bag already
popped and gotten a full bag I feel that it should have popped at least half of the
bag I understand that everyone is <UNK> their <UNK> but I also think that the <UNK>
should get something for their money I really enjoy this product but I probably
will not order any more for this reason
true:      <UNK> Microwave Pork rinds review
predicted: <UNK> Kettle normal rinds sodas
```

```
input:      I use goat milk for everything I use it to replace <UNK> milk in all my
recipes and for coffee and it works well for me The price is great Highly recommend
it
true:      Love <UNK> Milk
```

predicted: Love <UNK> Milk

input: I have purchased several herbs from <UNK> which have been of good quality <UNK> leaf is always one of the most difficult to purchase due to the likely chance of <UNK> more <UNK> than leaf This <UNK> leaf purchase was of average to good quality I did have about <UNK> to leaf cut in the package so I was <UNK> with the selection When I dry my own <UNK> leaf I usually do about the same or a little less <UNK> to leaf ratio for cooking use <UNK> of the <UNK> leaf are just as <UNK> as the leaf and can be used to cook with <UNK>
true: <UNK> <UNK>
predicted: <UNK> <UNK>

input: These cookies are every bit as good as their gluten counter part <UNK> to make Delicious and much cheaper than purchasing the already made <UNK> packaged gluten free cookies
true: Great for the whole family
predicted: Great for the whole family

input: First of all these things taste pretty bad If you can handle the taste while they melt on your tongue maybe it wouldn't be AS bad But I really didn't notice much change in the taste of foods <UNK> tasted sweet but that's all I noticed I considered trying two at a time but they just tasted so bad to me that I have never used them again
true: <UNK>
predicted: <UNK>

input: We've been drinking Yuban since and find this the most convenient way to get it The price is great and it <UNK> comes every month If we are having guests we can <UNK> up the order Fantastic
true: Yuban <UNK>
predicted: Cinnamon <UNK>

input: I am a <UNK> <UNK> coffee guy I enjoy <UNK> about coffee with the <UNK> of my local Starbucks I also have my own <UNK> <UNK> French press espresso maker and drip coffee <UNK> cup and cup So where does <UNK> fit into my coffee life I typically use a <UNK> packet when I want a quick cup of coffee and don't want to take the time or effort to brew a pot cup or cup coffee <UNK> <UNK> Fast and <UNK> Taste great for an instant coffee it is not quite as good as fresh brewed but <UNK> better than other instant coffeebr <UNK> It is expensive So watch the price on Amazon it will vary <UNK> over time I tell my wife we are not <UNK> the full price people <UNK> the item in your <UNK> and then if <UNK> have time check your <UNK> daily for a while I do this for all Amazon <UNK> since the price does really <UNK> for most items For <UNK> I have found that watching it can save me or more on Amazon the price of this item locally never <UNK> <UNK> It is not the same as good brewed coffee Don't expect this to be the same as fresh brewed coffee all else <UNK> <UNK> filtered water etc
true: Great fast way to get good coffee
predicted: Good alternative and to are good food

7 Zaključak

Analizirajući dobijene rezultate došli smo do sledećih zaključaka:

- Model je dobro razumeo povezanost između podataka pri rešavanju prvog problema, što se može zaključiti iz gorenavedenih slika grafika dobijenih prilikom izvršavanja na trening i validacionom skupu.
- Rezultati dobijeni prilikom testa su takođe zadovoljavajući.

- Međutim, kada smo isti model probali da treniramo na podacima sa znatno dužim sekven-
cama, rezultati su bili i do 15% gori.
- Istu arhitekturu modela možemo koristiti na različitim tipovima problema, uz dodatne opti-
mizacije koje su potrebne. Model se dovoljno dobro snalazi u različitim problemima, i iako
rezultati drugog treninga nisu dobri kao prvi, mogu se koristiti za generisanje dovoljno dobrih
sumarizacija teksta.
- Bez obzira na drugačiju arhitekturu i izostavljanje rekurentnih neuronskih mreža prilikom
rešavanja ovog tipa problema, treniranje i pokretanje ovog modela je još uvek izuzetno za-
htevno kako vremenski, tako i memorijski.