Language translation model using neural machine translation with seq2seq architecture

Projekat iz Računarske inteligencije

Jelena Zarić

Septembar 2023

Sadržaj

1 Uvod

Neuronsko mašinsko prevođenje (NMT) predstavlja pristup učenju od početka do kraja za automatizovani prevod. Njegova snaga dolazi iz činjenice da direktno uči mapiranje iz ulaznog teksta u povezani izlazni tekst. Dokazano je da je efikasniji od tradicionalnog prevođenja zasnovanog na frazama, koje zahteva mnogo više truda u dizajniranju modela. S druge strane, NMT modeli su skupi za obuku, posebno na velikim skupovima podataka za prevod. Takođe su značajno sporiji tokom inferencije zbog velikog broja korišćenih parametara. Druge ograničavajuće faktore predstavljaju njegova otpornost pri prevođenju retkih reči i poteškoće u prevođenju svih delova ulazne rečenice. Kako bi se prevazišli ovi problemi, već postoje neka rešenja, kao što je korišćenje mehanizma pažnje za kopiranje retkih reči.

2 Opis problema

Izrada modela za prevođenje sa engleskog na francuski jezik koristeći neuronsko mašinsko prevođenje s arhitekturom seq2seq.

3 Skup podataka

Koristili smo fra-eng skup podataka koji sadrži oko 170 000 linija, pri čemu u svakoj liniji se nalazi najpre rečenica na engleskom, a zatim njen prevod na francuski jezik.

Za potrebe projekta nismo iskoristili svih 170 000, već smo uzeli manji podskup od 20 000 linija, radi bržeg treninga.

```
Go.     Va !
Hi.     Salut !
Run!    Cours !
Run!    Courez !
Who?    Qui ?
Wow!    Ça alors !
Fire!   Au feu !
Help!   À l'aide !
Jump.   Saute.
Stop!   Ça suffit !
Stop!   Stop !
Stop!   Arrête-toi !
Wait!   Attends !
Wait!   Attendez !
Go on.  Poursuis.
Go on.  Continuez.
Go on.  Poursuivez.
Hello!  Bonjour !
Hello!  Salut !
I see.  Je comprends.
I try.  J'essaye.
```

Slika 1: Primer elementa dataseta

4 Pretprocesiranje ulaznih i izlaznih podataka

Ulazni podaci nisu pretprocesirani, samo smo podelili ulazne linije na dva dela, jedan za ulazne recenice (na engleskom) i jedan za ciljne rečenice (na francuskom). Koristimo tabulator kao početni znak sekvence za ciljane elemente, i novi red kao znak završetka sekvence.

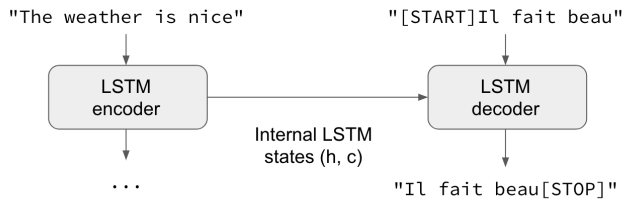
Rečenice teksta mogu biti različite dužine. Međutim, LSTM algoritam očekuje ulazne instance iste dužine. Stoga, konvertujemo naše ulazne i izlazne rečenice u vektore fiksne dužine.

Proširujemo ulazne rečenice do dužine najduže ulazne rečenice tako što dodajemo nule na početak, dok se reči zadržavaju na kraju jer je izlaz enkodera zasnovan na rečima koje se pojavljuju na kraju rečenice.

Proširujemo izlazne rečenice do dužine najduže izlazne rečenice tako što dodajemo nule na kraju, u slučaju dekodera, jer obrada počinje od početka rečenice.

5 Model mreže

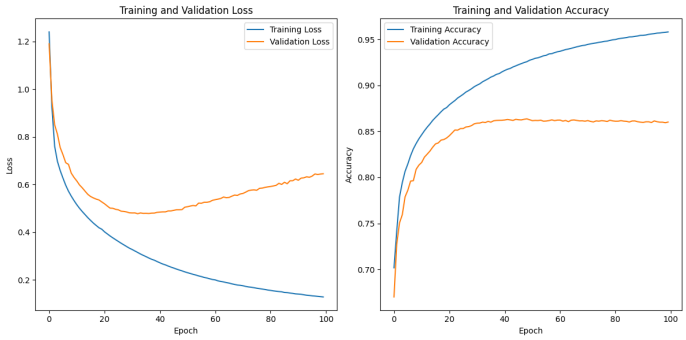
Ovaj model mreže je implementacija sekvencijskog prevodilačkog sistema koji koristi rekurentne neuronske mreže (LSTM) za prevođenje iz jednog jezika u drugi. Model se sastoji od dva glavna dela: enkodera i dekodera. Enkoder prima ulaznu sekvencu i obrađuje je putem LSTM sloja kako bi izvukao relevantne značajke. Nakon obrade, deo enkodera se odbacuje, a zadržavaju se unutarnja stanja. Dekoder zatim koristi ta unutarnja stanja kao početna stanja i generiše izlaznu sekvencu na ciljanom jeziku. Takođe, primenjuje se dropout za regularizaciju. Na kraju, model koristi sloj gustine sa softmax aktivacijom kako bi generisao konačan prevedeni tekst.



Slika 2: Prikaz arhitekture mreže

6 Testiranje i rezultati

Ovaj model za mašinsko prevođenje koristi optimizator 'rmsprop' kako bi prilagodio svoje težine kako bi minimizirao gubitak ('categorical_crossentropy'). Tokom procesa treniranja, koristi se paketna obrada u kojoj svaka grupa sadrži 64 primera (batch_size=64). Model se trenira tokom 100 epoha, što znači da će prolaziti kroz skup podataka 100 puta kako bi naučio bolje prevođenje. Takođe, kako bi se pratila progresija i izbegla prenaučenost, 20% podataka se koristi za validaciju (validation_split=0.2), što omogućava modelu da se ocenjuje na odvojenim podacima tokom treninga.



Slika 3: Prikaz rezultata treninga

Nakon treniga, kako testirati model ?
Ideja koju smo koristili zasniva se u 3 koraka:

- Enkodujte ulaz i dobijte početno stanje dekodera.
- Izvršite jedan korak dekodera s ovim početnim stanjem i početak sekvence tokenom kao ciljem. Izlaz će biti sledeći ciljni token.
- Ponavljajte sa trenutnim ciljnim tokenom i trenutnim stanjima sve dok ne dodjete do tokena za kraj ili je rečenica duža od maksimalne dozvoljene dužine.

Neki od rezultata koji su dobijeni:

- Input sentence: Call me. Decoded sentence: Appellez-moi !
- Input sentence: Come in. Decoded sentence: Venez !
- Input sentence: Be nice. Decoded sentence: Soyez gentilles !
- Input sentence: I'm 19. Decoded sentence: Je suis touchée.

7 Zaključak

Neuronsko mašinsko prevođenje predstavlja naprednu primenu obrade prirodnog jezika i uključuje veoma složenu arhitekturu. Ovu vrstu prevođenja možemo izvoditi putem arhitekture seq2seq, koja se zasniva na modelu enkoder-dekoder. Enkoder i dekoder su LSTM mreže. Enkoder kodira ulazne rečenice, dok dekoder dekodira ulaze i generiše odgovarajuće izlaze.

Arhitektura seq2seq je uspešna kada je reč o mapiranju odnosa ulaza na izlaz. Osnovna seq2seq arhitektura nije sposobna za hvatanje konteksta i jednostavno uči kako da mapira pojedinačne ulaze na pojedinačne izlaze.

Bolje je koristiti ugradnju reči (word embedding) jer duboki modeli za učenje rade s brojevima. Zbog toga moramo pretvoriti naše reči u njihove odgovarajuće numeričke vektorske reprezentacije, a ne samo u verziju niza celih brojeva. Na taj način model može naučiti povezanosti između reči.

Literatura

[1] François Chollet, *A ten-minute introduction to sequence-to-sequence learning in Keras*. Dostupno na: https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html.

[2] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, *Sequence to Sequence Learning with Neural Networks*. Dostupno na: https://arxiv.org/pdf/1409.3215.pdf.