Automatsko prepoznavanje govora

Seminarski rad u okviru kursa
Metodologija stručnog i naučnog rada
Matematički fakultet

Vladimir Vuksanović, Aleksa Kojadinović, Lazar Čeliković
kontakt email prvog, drugog, trećeg autora

15. novembar 2021.

Sažetak

U ovom tekstu je ukratko prikazana osnovna forma seminarskog rada. Obratite pažnju da je pored ove .pdf datoteke, u prilogu i odgovarajuća .tex datoteka, kao i .bib datoteka korišćena za generisanje literature. Na prvoj strani seminarskog rada su naslov, apstrakt i sadržaj, i to sve mora da stane na prvu stranu! Kako bi Vaš seminarski zadovoljio standarde i očekivanja, koristite uputstva i materijale sa predavanja na temu pisanja seminarskih radova. Ovo je samo šablon koji se odnosi na fizički izgled seminarskog rada (šablon koji *morate* da koristite!) kao i par tehničkih pomoćnih uputstava. Pročitajte tekst pažljivo jer on sadrži i važne informacije vezane za zahteve obima i karakteristika seminarskog rada.

Ključne reči: prepoznavanje govora

Sadržaj

1 Uvod **2**

2 Izazovi **2**

3 Statisticki model **3**
3.1 Procesiranje zvucnog signala 3
3.2 Akusticki model 4
3.3 Model izgovora 4
3.4 Jezicki model 4
3.5 Dekodiranje 4

4 End-to-end model **4**
4.1 CTC model 4
4.2 modeli zasnovani na paznji 6

5 Metrike za evaluaciju **6**

6 Zaključak **6**

Literatura **6**

A Pregled skupova podataka **6**

1 Uvod

Govor je za ljude najintuitivniji i prirodniji nacin komunikacije. Zbog toga je od samog nastanka kompjutera, nastala i ideja da koristimo isti nacin komunikacije da interagujemo sa njima. To bi znatno smanjilo potrebno predznanje za koriscenje kompjutera i ucnilio ga pristupacnijim vecem broju ljudi. Najveca prepreka ovim sistemima do skoro je bio kako sa velikom tacnosti prepoznati sta je korisnik rekao. Taj postupak se naziva automatsko prepoznavanje govora.

Automatsko prepoznavanje govora (eng. *Automatic Speech Recognition, ASR*) je proces pretvaranja zvučnog signala govora u sekvencu reči pomoću kompjutera. Neke od najznacajnijih primena ovih sistema su: pametni licni asistenti (Google Assistant[1], Apple Siri[2],...), transkripcija snimaka, pretrazivanje audio sadrzaja i pristupacnost.

Iako su istrazivanja na ovu temu pocela jos sredinom dvadesetog veka, popularnost je pocela da dobija tek u poslednjoj deceniji kada je uvodjenje dubokih neuronskih mreza drasticno povecalo performanse ovih sistema. Ta razlika je bila dovoljna da ucini ove sisteme prakticno primenljivim umesto nezgodnim za upotrebu zbog velikog broja gresaka. Jedan od najznacajnijih postignuca je ostvareno 2016. godine je kompanija Majkrosoft napravila sistem koji je ostvario iste rezultate kao ljudski ekperti na Switchboard skupu podataka [1]. Za glavne uzroke ovog naglog poboljsanja se smatraju [2]:

1. Sakupljanje velike kolicine tanskribovanih skupova podataka
2. Nagli porast u performansama grafickih procesorskih jedinica (GPU)
3. Poboljsanje algoritama za ucenje i arhitektura modela

U nastavku cemo prvo navesti neke izazove koje treba da resimo da bi smo napravili dobar sistem za prepoznavanje govora, zatim cemo opisati nacin rada dva najpopularnija modela: statisticki i end-to-end i na kraju cemo predstaviti nacin za njihovu evaluaciju.

2 Izazovi

Prepoznavanje govora je veoma tezak zadatak zato sto je potrebno da radi podjednako dobro u razlicitim uslovima. Ideal kojem se tezi kod ovih sistema je nezavisnost od govornika (speaker independence).

Neki od najvecih izazova su:

- Mala kolicina podataka za trening - Za ostvarivanje dobrih rezultata, potrebno je sakupiti vise stotina sati labeliranih zvicnih snimaka koji treba da sadrze vise govornika. Ovako velike skupove je tesko naci, i jos teze verifikovati. Zbog toga se istrazuju alternativni nacini za treniranje kao st su samo-treniranje [3], iterativno treniranje [4] ili treniranje koristeci kompjuterski generisan glas [google?].
- Stil govora
- Varijacije u brzini govora, jacini i frekvenciji glasa, starosti
- Varijacije u dijalektu i akcentima
- Pozadinska buka
- Velicina recnika

[1] https://assistant.google.com/
[2] https://www.apple.com/siri/

3 Statisticki model

Dugo vremena statisticki pristup je bio dominantan za sisteme za prepoznavanje govora. Cilj ovih sistema je da pronadju najverovatniju transkripciju za zadati signal. Formalno, neka je $\hat{W}$ optimalan niz reci za transkripciju nekog zvucnog signala X. Cilj je optimizovati formulu [5]:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|X)$$

primenom Bajesove formule to mozemo da zapisemo kao:

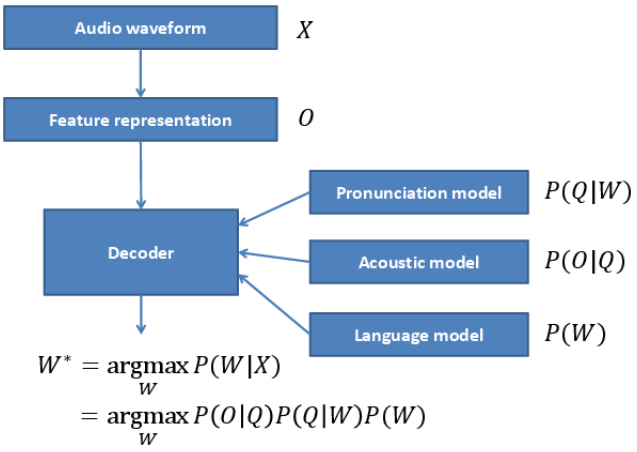
$$\hat{W} = \underset{W}{\operatorname{argmax}} \frac{P(X|W)P(X)}{P(W)}$$

a kako je $P(X)$ konstantno za konkretan ulaz, mozemo da ga eliminisemo:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(X|W)P(W)$$

Prvi deo, $P(X|W)$, racuna akusticki model, a drugi, $P(W)$, racuna jezicki model. Ideja je da odvojeno optimizujemo obe prethodne velicine i ocekujemo da ce se tako maksimizirati ukupna verovatnoca. U praksi, $P(X|W)$ je tesko precizno odrediti na nivou reci zbog velikih varijacija u izgovoru, pa se cesce odredjuje na nivou glasova ili fonema. Zbog toga se dodatno uvodi i model izgovora (eng.) koji kombinuje te manje jedinice u reci.

Na slici 1 je prikazana cela struktura statistickog modela od zvucnog signala do transkripcije: U nastavku ce biti opisana svaka od pomenutih



Slika 1: Statisticki model

komponenti, njena uloga i nacin rada.

3.1 Procesiranje zvucnog signala

Sirovi zvucni signal je veoma nepogodan za koriscenje zato sto sadrzi veliku kolicinu nepotrebnih informacija i šuma. Zbog toga se pre prosledjivanja akustickom modelu, signal prvo obradjuje tako da ostanu samo kljucne karakteristike i smanji sum i velicina reprezentacije.

Signal delimo na kratke segmente koje zovemo okviri (eng. frame). Svaki od njih je fiksne duzine (obicno 10-30 milisekundi) sa kratkim preklapanjem susednim okvirima radi smanjenja naglih promena prilikom prelaska iz jednog u drugi. Pretpostavka je da je u svakom od okviru glas konstantan, to jest da se glasovi mogu menjati samo prelaskom iz jednog okvira u drugi. Na svaki od tih novodobijenih delova se zatim primenjuje neka vrsta spektralne analize kojom se izdvajaju samo njegove najbitnije karakteristike. Tacna reprezentacija koja se koristi varira u zavisnosti od modela, ali jedna od najpopularnijih je MFCC [?].

3.2 Akusticki model

Akusticki model je zaduzen da pretvori sekvencu okvira u sekvencu fonema. Za modelovanje svake foneme se koriste skriveni Markovljevi modeli (eng. *Hidden Markov Models*) [6]. Ti modeli se zatim nadovezuju da grade slova, reci...

3.3 Model izgovora

U slucaju da zelimo da prepoznajemo samo malu kolicinu reci, prethodni model bi bio sasvim dovoljan pod uslovom da imamo dobar trening skup za te reci. Problem nastaje ako zelimo da prepoznamo vecu kolicinu reci, tada cemo za neku rec imati mozda samo par primera u celom skupu. To nije dovoljno da se istrenira model.

$$P(X|W) = P(X|Q) \sum_{Q} P(Q|W)$$

$P(Q|W)$ zovemo model izgovora (eng. *Pronunciation Model*). On mapira niz fonema u odgovarajucu rec ili niz slova koja se tako izgovara.

3.4 Jezicki model

Odredjuje najverovatniju kombinaciju reci

3.5 Dekodiranje

U praksi je recnik prilicno veliki pa nije prakticno racunati verovatnocu za svaku kombinaciju reci u recenici. Umesto toga se koristi beam search algoritam.

4 End-to-end model

4.1 CTC model

Neka je ulazni zvuk uzorkovan u proizvoljnom broju jednakih vremenskih intervala, gde je svaki od njih predstavljen vektorom realnih brojeva duzine m. Neka je L konacna azbuka labela (oznaka). Cilj je napraviti preslikavanje h koje slika proizvoljan zvucni signal u niz labela:

$$h : (\mathbb{R}^m)^* \to L^* \tag{1}$$

U praksi fiksiramo broj vremenskih trenutaka na neku vrednost T. Iako je broj trenutaka fiksiran za zvucne signale, nizovi labela ne moraju biti iste duzine za svaku ulaznu instancu, stoga jednu trening instancu

predstavlja par $(\mathbf{x}, \mathbf{z})$ gde je z vektor labela duzine najvise T. Ako bismo test skup oznacili sa O, tada funkciju greske mozemo definisati na sledeci nacin:

$$LER(h, O) = \frac{1}{|O|} \sum_{(\mathbf{x},\mathbf{z})\in O} \frac{ED(h(\mathbf{x}), \mathbf{z})}{|\mathbf{z}|} \tag{2}$$

ED predstavlja edit distancu (minimalni broj izmena koji dovodi jedan niz karaktera do drugog, pri cemu dozvoljene izmene podrazumevaju brisanje, supstituciju i umetanje karaktera). Prethodna mera naziva se stopa greske labela (*eng. label error rate - LER*).

Formula 2 jeste prirodna ocena greske za probleme koji za cilj imaju minimizaciju greske prevodjenja.

Koristeci sva prethodna razmatranja konstruisemo rekurentnu neuronsku mrezu koja na ulazu ima mT ulaza, dok se na izlazu dobija T vektora dimenzija $L' = L \cup \{\epsilon\}$ pri cemu svaki predstavlja raspodelu verovatnoca oznaka za svaki trenutak prosirujuci azbuku blanko labelom ϵ.

Neformalno receno, prolaskom kroz izlazne softmax nizove dobijamo putanju $\pi \in (L')^T$ koja predstavlja jedan moguci odabir labela. Ako $y^t_{\pi_t}$ predstavlja softmax ???? vrednost t-tog trenutka oznake π_t tada verovatnocu odabira kompletne putanje dobijamo kao proizvod:

$$p(\pi|x) = \prod_{t=1}^{T} y^t_{\pi_t}$$

U praksi uzorkovanje zvuka se vrsi u veoma sitnim vremenskim intervalima (oko 10ms). Stoga je pojava blanko ili dupliciranih oznaka veoma cesta. Iz tog razloga uvodimo preslikavanje β cija je uloga preciscavanje nizova labela uklanjanjem blanko oznaka i susednih duplikata.

$$\beta : (L')^T \to L^U, U \leq T$$

Primetimo da za jednu preciscenu putanju l moze postojati vise mogucih izvornih putanja, pa je verovatnoca njenog odabira jednaka sumi po svim izvornim putanjama.

$$p(l|x) = \sum_{\pi\in\beta^{-1}(l)} p(\pi|x) \tag{3}$$

Imajuci u vidu sve prethodno navedeno, zadatak preslikavanja h je odabir najverovatnije preciscene putanje za dati ulaz.

$$h(x) = \operatorname*{argmax}_{l\in L^U} p(l|x) \tag{4}$$

Najjednostavniji algoritam jeste pohlepni odabir najbolje oznake za svaki vremenski trenutak ponaosob. Medjutim, ovakav pristup ne garantuje optimalnost. Naravno, postoje bolji algoritmi za resavanje datog problema. Isti nece biti obradjivani u ovom radu, ali se mogu naci u [].

4.2 modeli zasnovani na paznji

5 Metrike za evaluaciju

Standardna mera za procenu kvaliteta sistema za prepoznavanje govora je stopa pogresnih reci (eng. *Word Error Rate (WER)*).

$$WER = \frac{S + D + I}{N}$$

gde je S broj zamenjenih reci, D broj obrisanih reci, I broj umetnutih reci, i N ukupan broj reci u referentnoj recenici. Minimalna vrednost koju moze da dobije je 0, dok maksimalna vrednost moze da bude preko 1 (npr. veliki broj umetnutih reci). Stopa pogresnih reci se efikasno racuna dinamickim programiranjem, pomocu Vagner-Fiserovog algoritma (eng.)

Cilj sistema za prepoznavanje govora je da minimizuje ovu vrednost.

6 Zaključak

Literatura

[1] Microsoft, "Historic achievement: Microsoft researchers reach human parity in conversational speech recognition," 2016.

[2] A. Hannun, "The history of speech recognition to the year 2030," 2021.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[4] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," *Interspeech 2020*, Oct 2020.

[5] U. Kamath, J. Liu, and J. Whitaker, *Deep Learning for NLP and Speech Recognition*. Springer Publishing Company, Incorporated, 1st ed., 2019.

[6] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

A Pregled skupova podataka

TIMIT	cell2	cell3
SwitchBoard	cell5	cell6
LibriSpeech	cell8	cell9
CommonVoice	a	a