Poređenje arhitektura neuronskih mreža za prepoznavanje govora

Mašinsko učenje - Projekat

Matei Jon Stanču, 1137/2015

2018-09-26

Sadržaj

1 Uvod

U ovom radu predstavljene su osnove sistema za prepoznavanje govora zasnovan na konvolutivnoj neuronskoj mreži, koji može da prepoznaje deset različitih reči. Dok su u praksi sistemi za prepoznavanje govora mnogo složeniji, cilj ovog rada je samo da pokaže osnovne koncepte i tehnike na kojima sistemi za prepoznavanje govora počivaju. Kreirane su, obučavane i testirane različite arhitekture konvolutivnih neuronskih mreža.

2 Opis problema

Kako je zvuk jednodimenzioni signal kroz vreme pri čemu prepoznavanje izgovorene reči dosta zavisi od konteksta, prirodno rešenje ovom problemu bi predstavljala rekurentna neuronska mreža. Međutim, zvuk može biti interpretiran i kao dvodimenzioni signal ako se uzme u obzir njegov spektrogram, tj. frekvencijski spektar u nekom vremenskom intervalu određene širine, pa se na osnovu takve reprezentacije zvuka mogu koristiti konvolutivne neuronske mreže identično kao u domenu prepoznavanja objekata na slikama.

Ideja je definisati vremenski prozor u kom mogu da stanu izgovorene reči, i zatim transformisati zvučni signal iz tog prozora u njemu odgovarajući spektrogram. To se radi računanjem frekvencijskog spektra svakih nekoliko milisekundi signala, koji predstavlja vektor jačina frekvencija prisutnih u tom intervalu od nekoliko milisekundi, nakon čega se svi ti vektori koji definišu ceo prozor poređaju hronološki i time se dobija matrica. Ova matrica može da se interpretira kao jednokanalna slika koja predstavlja spektrogram i predstavljena je slikom 1.

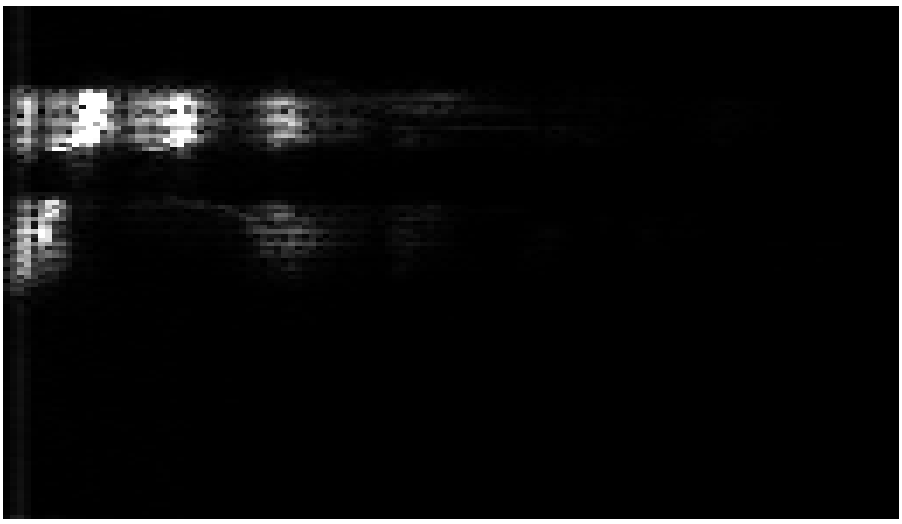

Figure 1: Spektrogram reči "Happy".

Sistem čine tri komponente, prvu komponentu predstavlja izdvajač atributa koji računa 40 dimenzione atribute predstavljene na Mel skali koji se računaju svakih 25 milisekundi sa

klizecim prozorom od 10 milisekundi, koji se zatim prosleđuju komponenti sa propagacijom unapred. Komponenta sa propagacijom unapred se sastoji od tri skrivena sloja sa po 128 skrivenih jedinica i jednim slojem mekog maksimuma. Svaki skriveni sloj koristi ReLU kao aktivacionu funkciju. Izlaz iz mekog maksimum sloja sadrži po jedan izlaz za svaku ključnu reč koja se prepoznaje i dodatan izlaz koji označava sve prozore koji ne pripadaju nijednoj ključnoj reči. Težine mreže su obučavane tako da optimizuju unakrsnu entropiju funkciju greske koristeći distribuirani asinhroni gradijentni spust [6]. Na kraju u okviru treće komponente dobijaju se ocene za pojedinačne prozore pri čemu se one kombinuju kako bi se dobila finalna ocena koja odgovara nekoj od klučnih reči. Detaljnije u [7].

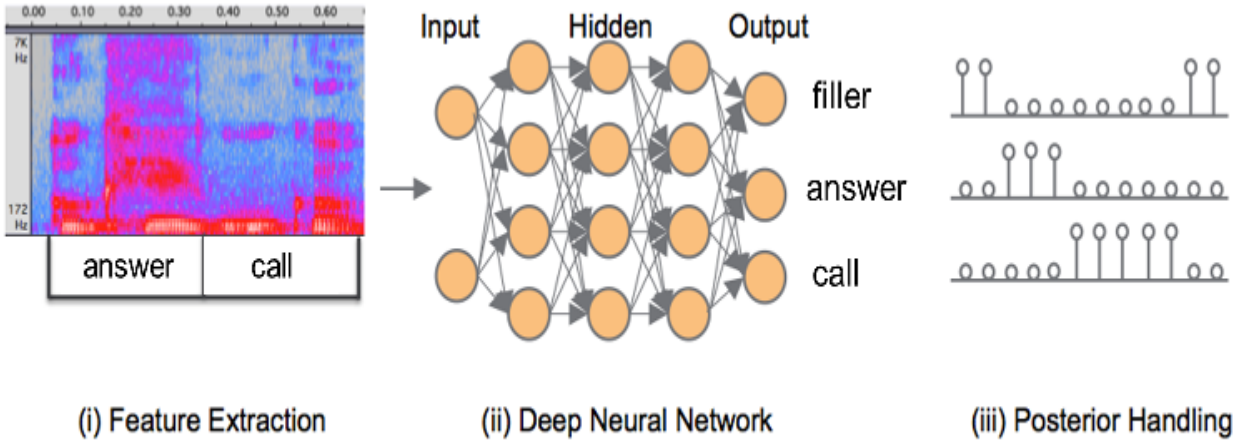

Figure 2: Zadatak detekcije ključnih reči.

3 Skup podataka

Skup podataka koji je korišćen za obučavanje modela je preuzet sa Tensorflow-ove stranice i sadrži 30 različitih reči izgovorenih od strane više ljudi, preko 105000 zvučnih zapisa u 16-bitnom PCM kodiranom WAVE formatu uzorkovanih sa podrazumevanom frekvencijom od 16000 Hz.

3.1 Priprema podataka

Prilikom obučavanja modela omogućena je promena parametara kao što je frekvencija uzorkovanja ili trajanje pojedinačnog zvučnog zapisa. Pre obučavanja modela datoteke su razvrstane po direktorijumima koji nose ime kao klase kojima pripadaju, s tim što su sve klase koje se ne razmatraju u konkretnoj klasifikaciji smeštene u jedan zajednički direktorijum koji predstavlja klasu nepoznatih reči. Ovo je neophodno kako bi mreža mogla da nauči da prepoznaje zvuke

koji nisu relevantni. Takođe je moguće podestiti procenat nepoznatih reči u ukupnom broju reči, pri čemu je podrazumevana vrednost 10% .

Skup podataka takođe sadrži poseban direktorijum sa datotekama koje sadrže buku proizvedenu od strane različitih svakodnevnih aktivnosti. Ove zapisi se mogu koristiti za mešanje sa zapisima koji predstavljaju ključne reči koje je potrebno prepoznati kako bi se postigao efekat pozadinske buke i time napravilo da snimci budu realističniji. Potrebno je bilo napraviti i direktorijum sa snimcima koji predstavljaju tišinu, pri čemu se u realnim uslovima nikada ne očekuje potpuna tišina, uvek će postojati neki šum ili neki zvuk koji je dosta utišan pa je zbog toga bilo potrebno iskoristiti snimke iz direktorijuma pozadinske buke. Parametar koji određuje procenat tišitne u svim snimcima takođe može da se podesi i njegova podrazumevana vrednost je 10% .

3.2 Pretprocesiranje

Pošto je ljudsko uho osetljivije na određene frekvencije u odnosu na ostale, u praksi često se vrši dodatno pretprocesiranje kako bi se spektrogram dalje transformisao u skup MFC koeficijenata (eng. Mel-Frequency Cepstral coefficients). Ova reprezentacija je i dalje matrica pa takođe može da se interpretira kao slika i nakon ovog pretprocesiranja može da se prosledi konvolutivnoj neuronskoj mreži. Postupak transformacije frekvencijskog spektra u niz MFC koeficijenata čine sledeći koraci [2] [3]:

- Transformacija polaznog signala Furijeovom transformacijom u njemu odgovarajući frekvencijski spektar.
- Preslikavanje frekvencijskog spektra u interval Mel skale.
- Primena logaritamske funkcije na transformisanom spektru.
- Transformacija log-Mel skaliranih koeficijenata diskretnom kosinusnom transformacijom kao da je signal.

Nakon primene diskretne kosinusne transformacije, amplitude rezultujućeg spektra predstavljaju MFC koeficijente polaznog signala što predstavlja reprezentaciju koja je odovarajuća za prosledjivanje neuronskoj mreži.

4 Korišćeni modeli

Konvolutivne mreže predstavljaju bolje rešenje u odnosu na klasične duboke neuronske mreže iz više razloga. Prvo, jer klasične neuronske mreže zanemaruju strukturu ulaza, u smislu da ulaz može biti predstavljen bilo kojim redosledom a da pri tom mreža ne proizvodi značajnu razliku u performansama. Pošto su spektrogrami reprezentacije govora koje imaju jake korelacije

u vremenu i frekvenciji, modelovanje tih korelacija konvolutivnom neuronskom mrežom koja sadrži lokalno deljene težine se pokazalo kao dosta dobro rešenje.

Druga prednost konvolutivnih neuronskih mreža u odnosu na klasične bi bila to da klasične neuronske mreže nisu eksplicitno dizajnirane da modeluju različite varianse u različitim signalima kao što je obično prisutno u govoru zbog različitih stilova govora, tj. različiti stilovi proizvode translacije vrednosti u frekvencijskom domenu. Zapravo klasične duboke neuronske mreže dovoljne veličine bi mogle da uhvate ovaj efekat ali bi pored ogromne veličine, mreža zahtevala i jako mnogo primeraka u skupu za obučavanje, dok bi konvolutivna neuronska mreža to mogla da uradi sa neuporedivo manjim brojem parametara tako što uprosečava izlaze iz skrivenih jedinica u različitim lokalnim vremenskim i frekvencijskim regionima.

U opštem slučaju CNN arhitektura koja rešava zadatak detekcije ključnih reči prikazana je na slici 3. Dat nam je ulazni signal $V \in R^{t \times f}$, gde su t i f vremenska i frekvencijska dimenzija ulaza. Matrica težina $W \in R^{(m \times r) \times n}$ je konvolvirana sa celim ulazom V, pri čemu matrica W pokriva manji deo vremensko-frekvencijskog prostora iz ulaza veličine $m \times r$, gde je $m \leq t$ i $r \leq f$. Deljenje težina u ovoj mreži omogućava da se dobro modeluju lokalne zavisnosti u vremensko-frekvencijskom prostoru u ulaznom signalu. Matrica težina ima n skrivenih jedinica, odnosno mapa atributa. Filter može da se pomera za pozitivnu vrednost s u vremenskoj dimenziji i p u frekvencijskoj dimenziji, tako da je ceo konvolucioni proces proizvodi n mapa atributa dimenzije $\frac{(t-m+1)}{s} \times \frac{(f-r+1)}{v}$.

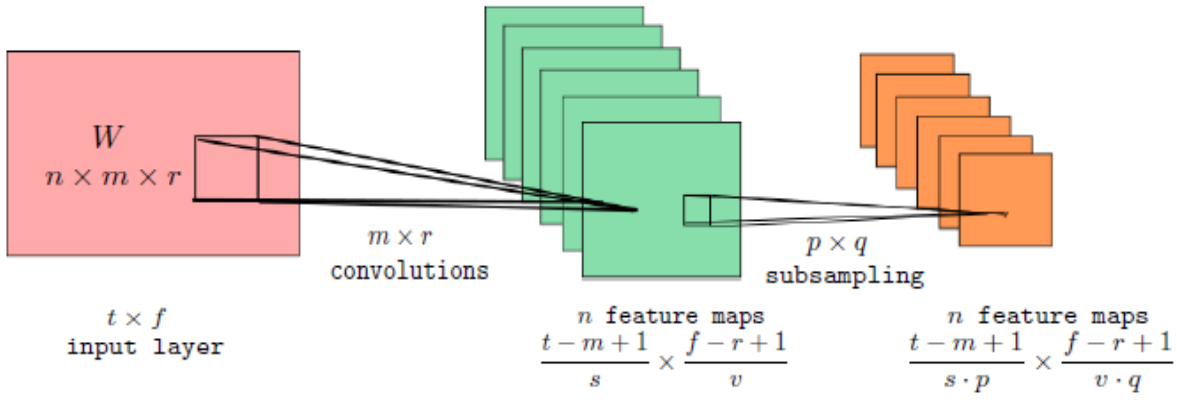

Figure 3: Zadatak detekcije ključnih reči.

Nakon primene konvolucije sledi primena sloja koji vrši agregaciju izdvajanjem maksimuma kako bi se uklonila variabilnost u vremensko-frekvencijskom prostoru koja se javlja zbog različitih stilova u govoru ili distorzija u kanalima. Agregacija je veličine $p \times q$ i predstavlja nepreklapajuću agregaciju nakon čega vremensko-frekvencijsa dimenzija je $\frac{(t-m+1)}{s \cdot p} \times \frac{(f-r+1)}{v \cdot q}$. U nastavku su razmatrane dve arhitekture konvolutivnih mreža, jedna fokusirana na pouzdanost klasifikacije, druga na brzinu klasifikacije.

4.1 Standardna arhitektura

Standardna konvolutivna neuronska mreža koja je puno puta korišćena, testirana i pokazivala veoma dobre rezultate [8] [9] sadrži dva konvolutivna sloja. Ako je ulaz u mrežu $t \times f = 98 \times 40$, tada će prvi sloj imati 64 filtera veličine $m \times r = 20 \times 8$, pri čemu se konvolucija vrši pomeranjem filtera za $s = 1$ po vremenskoj osi i za $v = 1$ po frekvencijskoj osi. Zatim se vrši nepreklapajuća agregacija veličine $q = 3$ izdvajanjem maksimuma po frekvencijama. Sledeći konvolutivni sloj ima filter veličine $m \times r = 10 \times 4$ dok je veličina agregacije $q = 1$, odnosno agregacija ne postoji. Ova arhitektura sadrži još jedan sloj poravnavajna pre nego što se podaci dalje proslede potpuno povezanom sloju koji dalje prosleđuje sloju koji vrši meki maksimum nad predviđanjima. Na slici 4 prikazan je broj parametara po slojevima standardne arhitekture.

type	m	r	n	p	q	Par.	Mul.
conv	20	8	64	1	3	10.2K	4.4M
conv	10	4	64	1	1	164.8K	5.2M
lin	-	-	32	-	-	65.5K	65.5K
dnn	-	-	128	-	-	4.1K	4.1K
softmax	-	-	4	-	-	0.5K	0.5K
Total	-	-	-	-	-	244.2K	9.7M

Figure 4: Parametri standardne arhitekture.

Vidi se da u konvolutivnim slojevima postoji jako mnogo množenja što je vremenski dosta zahtevno. To može da se reši izostavljanjem jednog konvolutivnog sloja.

4.2 Arhitektura za brzu klasifikaciju

.....

5 Obučavanje i evaluacija modela

Standardna neuronska mreža je obučavana u 18000 koraka pri čemu prvih 15000 koraka je parametar brzine ulčenja postavljen na 0.001 dok je preostalih 3000 brzina učenja postavljena na 0.0001. Mreža je obučavana da klasifikuje 10 različitih ključnih reči na engleskom jeziku, to su: "yes", "no", "up", "down", "left", "right", "on", "off", "stop", "go". Dodatno mreža je sposobna da razlikuje tišine i nepoznate reči od prethodno navedenih.

Obučavanje standardne mreže je trajalo 5 dana i 1 sat, pri čemu već u prvim koracima mreže je počela dobro da razlikuje tišine od zvuka. Nakon svih 18000 koraka mreža je ostvarila preciznost klasifikacije od 88.5% što je daleko ispod najboljih sistema za prepoznavanje govora što trenutno postoje koji postižu 98%. Finalna matrica konfuzije je prikazana na slici 5 i može se videti da nakon obučavanja standardna mreža najbolje razlikuje tišine od nepoznatih i poznatih reči, dok je najmanju preciznost postizala u razlikovanju poznatih od nepoznatih reči.

```
INFO:tensorflow:Confusion Matrix:
 [[369   0   0   0   0   0   1   0   0   0   1   0]
 [  4 258   2   5  14  15  13  19  16   1   5  19]
 [  2   5 370   8   1   2   6   2   0   0   0   1]
 [  3   9   5 351   5   4   8   2   0   1   3  15]
 [  3   5   0   0 322   1   1   0   3   4  10   1]
 [  4   4   2  20   1 331   3   0   0   0   4   8]
 [  3   4  13   2   5   0 316   6   0   0   2   1]
 [  2   6   1   0   2   0   4 345   3   0   0   0]
 [  4  14   0   0   6   1   0   0 327  11   0   0]
 [  4   3   0   0  25   0   2   2   8 325   3   1]
 [  5   6   0   0  11   0   0   1   0   3 322   2]
 [  3  16   0  30   3  17   0   1   0   1   5 296]]
INFO:tensorflow:Step 18000: Validation accuracy = 88.5% (N=4445)
```

Figure 5: Finalna matrica konfuzije.

Izmerene su preciznosti na skupu za obučavanje u svakom koraku i na skupu za validaciju na svakih 400 koraka, pri čemu su ocene na oba skupa bile slične tokom celog obučavanja što znači da mreža nije pokazivala znake potprilagođavanja ili preprilagođavanja. Na slikama 6 i 7 prikazane su preciznosti i vrednosti funkcije greske unakrsne entropije na svkom koraku obučavanja.

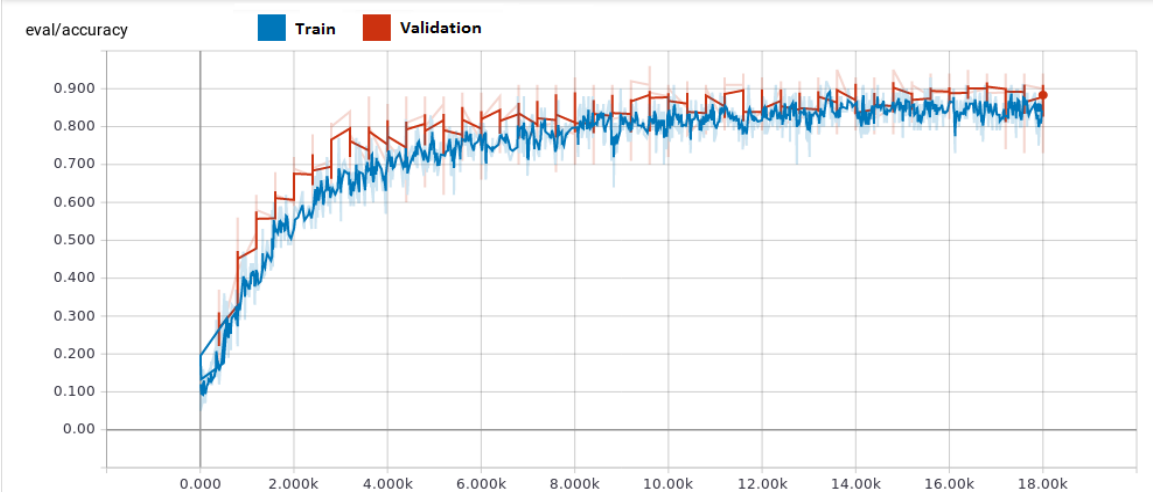

Figure 6: Preciznost klasifikacije na skupu za obučavanje i validaciju.

Literatura

[1] Mladen Nikolić, Anđelka Zečević. *Mašinsko učenje*. Matematički Fakultet, Univerzitet u Beogradu, 2018.

[2] Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian *HMM-Based Audio Keyword Generation*. Nanyang Technological University, Singapore, 2004

[3] Sahidullah Md., Saha, Goutam *Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition*. Indian Institute of Technology, Kharagpur, 2012

[4] Tara N. Sainath, Carolina Parada *Convolutional Neural Networks for Small-footprint Keyword Spotting*. Google Inc., New York, Interspeech 2015.

[5] Pete Warden, Google Brain. J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng. *Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition*. Mountain View, California, 2018.

[6] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng. *Large Scale Distributed Deep Networks*. NIPS, 2012.

[7] G. Chen, C. Parada, and G. Heigold. *Small-footprint Keyword Spotting using Deep Neural Networks*. ICASSP, 2014.

[8] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran *Deep Convolutional Neural Networks for LVCSR*. ICASSP, 2013.

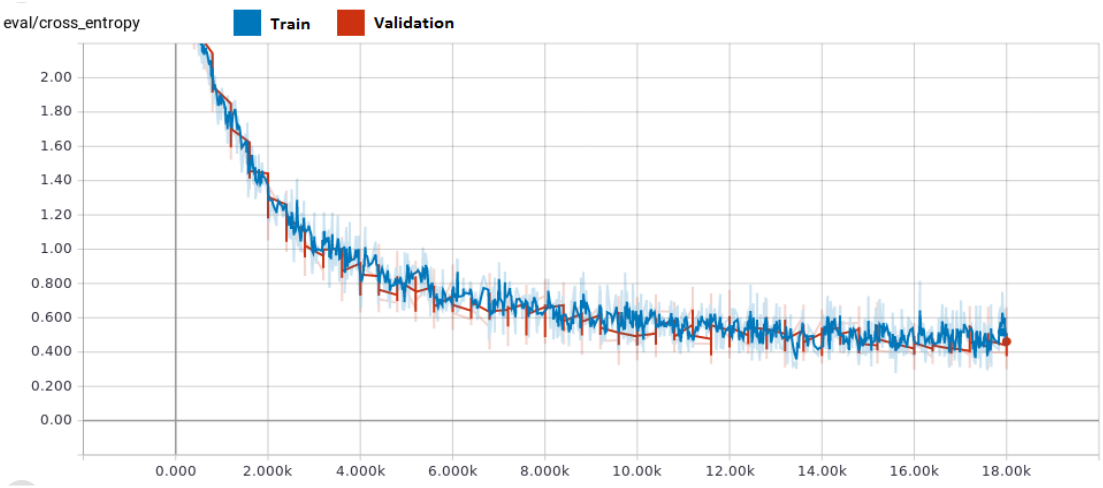

Figure 7: Ocena greške unakrsnom entropijom na skupu za obučavanje i validaciju.

[9] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak *Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks*. ICASSP, 2015.