Poenostavljen model globoke nevronske mreže za razpoznavanje govora

Robert Rozman

Fakulteta za računalništvo in informatiko
Univerza v Ljubljani
Tržaška 25, 1001 Ljubljana, Slovenija
E-pošta: robert.rozman@fri.uni-lj.si

Simplified Deep Neural Network Model for Speech Recognition

Current Speech Recognition Systems (SRSs) are rather complex and static structures trained on large databases. As an alternative, compact SRSs with limited vocabulary that are based on phonemes and use neural network as an acoustic model emerged. Their structure matches better with recent developments in the fields of distributed and parallel processing.

Recently, deep neural networks (DNN) appeared as a promising concept in speech recognition and showed the best results on the TIMIT phoneme recognition task. But their structure and training are quite complex. Therefore, we have created simplified DNN model for speech recognition that includes a classical parameterization process for MFCC features on lower part and common 3 layer classification structure on upper part of the network; network structure is now much simpler and training significantly faster. Now also parameterization part of the network can take part in the training process; consequently also the parameterization layers can be optimized. Practical tests confirm our expectations and show lower classification errors when additional parameterization layers are included in the optimization.

1 Uvod

Sistemi za razpoznavanje govora (v nadaljevanju SRG) so že zelo dolgo predmet intenzivnih raziskav. Razpoznavanje govornih signalov spada med splošno znane probleme, ki ves čas spremljajo razvoj širšega področja digitalnega procesiranja signalov. V današnjem času je potreba po uspešni govorni komunikaciji med človekom in računalnikom še posebej izrazita. V dobi intenzivnih govornih komunikacij si težko predstavljamo vsakdanje življenje brez uspešnega avtomatskega razpoznavanja govora.

Prav zaradi svoje daljše zgodovine je razpoznavanje govornih signalov zelo zanimivo področje; razvoj je namreč potekal dokaj specifično. Ves čas so nanj vplivale omejitve trenutno dostopnih tehnologij. Marsikatera ideja je zaradi tega ostala neuresničena, saj takratna tehnologija ni omogočala njene uspešne realizacije. Po drugi strani pa so zaradi želje po čim hitrejšem napredku odpadale morda tudi obetavne ideje, ki že v začetni fazi niso pokazale zadostnega napredka. Zato na tem področju dokaj pogosto srečamo pojavitve izboljšanih idej oziroma konceptov iz bližnje in včasih tudi bolj oddaljene preteklosti.

Nevronske mreže so tipičen predstavnik koncepta, ki je v razpoznavanju govora prisoten že daljše obdobje. Kljub temu večinoma ni zasedal vodilne vloge – po dokaj uspešnem začetku so na področju razpoznavanja govora prevladali statistični modeli, t.i. Prikriti Modeli Markova (angl. HMM). Ti so očitno prinašali nekemu obdobju bolj primerno razmerje med vgrajenim znanjem in možnostjo učenja iz govornih zbirk posnetkov ter nenazadnje tudi zelo dobre rezultate.

V novejšem času smo ponovno priča bolj intenzivnemu dogajanju na tem področju. Zmogljivost sistemov za digitalno procesiranje signalov narašča, vendar predvsem v razvijalcem manj prijazni smeri paralelizacije in porazdeljenega procesiranja, kar daje prednost temu trendu bolj primernim konceptom.

Nevronske mreže seveda imajo tudi svoje slabosti. Predvsem je na področju razpoznavanja govora problematična omejena možnost interpretacije in uporabe znanja, ki se zbira v mreži med postopkom učenja. Tudi časovno modeliranje, ki je pri govoru sicer zelo pomembno, ni najbolj blizu mrežnim konceptom. Prav tako se slednji med delovanjem nekoliko težje prilagajajo trenutnim razmeram. Seveda večina teh slabosti tekom razvoja postopoma izginja, zato mreže počasi, a zanesljivo spet prihajajo v ospredje tudi na tem področju.

Koncept t.i. globokih nevronskih mrež (angl. »Deep Neural Network« - DNN) je trenutno najbolj aktualen in tudi na področju razpoznavanja govora kaže dokaj obetavne rezultate [1]. Seveda je potrebno pojasniti, da gre za nekoliko enostavnejši problem razpoznavanja elementarnih govornih enot – fonemov. Zato je težko predvideti, ali se bo koncept uspešno razširil tudi v sisteme s kompleksnejšimi slovarji in jezikovnimi modeli.

Koncept globokih mrež smo tudi sami preizkušali še pred njegovim bolj množičnim pojavom v znanstveni literaturi [2]. Nadaljevanje tega dela je predstavljeno v tem članku. V naslednjem poglavju je najprej predstavljen koncept enostavnejših razpoznavalnikov v povezavi z globokimi nevronskimi mrežami. Nato je opisan poenostavljen model globoke mreže, ki v celoto združi sicer ločena postopka parametrizacije govornega signala in klasifikacijo v osnovne govorne enote. Sledijo še opis in rezultati praktičnega preizkusa na konkretni govorni zbirki.

2 Enostavni sistemi za razpoznavanje govora – prednost ali omejitev ?

Na področju razpoznavanja govora je seveda končni cilj razvoja sistem, ki bo razpoznaval govor vseh govorcev s praktično neomejenim slovarjem in poljubno vsebino pogovora. Praksa pa nam vedno ponudi realnejšo sliko, kjer se moramo pogosteje zadovoljiti z različnimi omejitvami oziroma poenostavitvami. Vendar tak kompromis ne pomeni nujno samo nazadovanja – v nekaterih primerih se pokaže prav nasprotno. Zato mislimo, da imajo sistemi z omejenimi slovarji ali/in so namenjeni razpoznavanju enostavnejših govornih enot (v nadaljevanju krajše kar »enostavni sistemi«) pomembno vlogo na tem področju.

Že zdaj so omenjeni sistemi dokaj pomembni v primerih, ko so viri za razvoj razpoznavalnikov omejeni; pri tem ne mislimo samo na osnovne atribute sistemov, kot je npr. število govorcev, besed v slovarju ampak tudi na manjše jezikovne skupnosti. Te imajo bistveno bolj omejene vire za razvoj sodobnih jezikovnih tehnologij. V tem se skriva tudi paradoks, ker so ravno manjše jezikovne skupnosti s tega vidika bolj ogrožene od večjih, kjer razvoj poteka precej hitreje in producira boljše rešitve.

2.1 Prednosti enostavnih SRG

Enostavni sistemi imajo kar nekaj potencialnih prednosti [3]. Med njimi je pomembna tudi možnost tesnejše povezave z manjšo skupino uporabnikov, katerim se lahko sistem bolj učinkovito prilagodi. Posledično lahko njihov govor bolj uspešno razpoznava, v luči najnovejših trendov pa se lahko z uporabnikom poveže tudi na drugih ravneh – npr. lahko razpoznava razpoloženje, se nauči navad, čustev, potreb in uporabniku nudi pomoč v precej širšem smislu od same govorne komunikacije.

Poleg tega so enostavni sistemi za razpoznavanje govora bolj primerni za implementacijo v sodobnih sistemih, ki temeljijo na paradigmah vzporednega, porazdeljenega procesiranja, vseprisotnega računalništva in povezanega sveta (angl. »Internet of Things«).

2.2 Uporaba mrež v enostavnih sistemih

V preteklosti so nevronske mreže kazale boljše rezultate v enostavnih sistemih – v tistem trenutku, ko se je razvoj SRG obrnil v smeri zapletenejših sistemov, so dobile manj pomembno vlogo.

Potrebno je poudariti, da pridejo potencialne prednosti mrežnega koncepta najbolj do izraza prav v enostavnih SRG, kjer trenutno v obliki globokih mrež kaže zelo obetavne rezultate [1]. Sicer je v tem primeru proces učenja zelo zahteven, vendar so doseženi rezultati razpoznave fonemov med najboljšimi na govorni zbirki TIMIT [4]; ta je postala neke vrste neuradni standard za medsebojno primerjavo uspešnosti SRG.

2.3 Globoke mreže v razpoznavanju govora

Bistvo ideje globokih nevronskih mrež ni samo večje število skritih plasti, ampak tudi ustrezen način učenja, ki se lahko od sistema do sistema precej razlikuje. Vedno pa pri učenju rešujemo optimizacijski problem določitve parametrov nevronske mreže za čim boljšo prilagoditev učnim primerom (nadzorovano učenje). Klasični algoritem vzvratnega razširjanja napake (angl. »Back Propagation«) je enostaven in splošno znan, vendar brez izboljšav za tovrstno strukturo ni najbolj primeren. Zato trenutno obstaja kar nekaj različnih pristopov; vsi pa imajo nekaj skupnih značilnosti:

- na uspešnost učenja oziroma mreže zelo vpliva tudi začetno stanje; pri globokih mrežah se to zdi še bolj pomembno, zato se zelo raziskujejo možni postopki predhodne priprave na učenje (angl. »Pre-Training«),
- k učenju večjega števila notranjih plasti se običajno pristopi v hierarhičnem smislu – najprej se posamezne plasti nenadzorovano učijo tako, da se pri učenju nove uporabijo do sedaj naučene plasti,
- učenje se po navadi zaključi z nadzorovano fino prilagoditvijo vseh plasti učnim primerom – pogosto se to naredi kar z že omenjenim vzvratnim razširjanjem napake.

Bistvo ideje globokih mrež je pravzaprav samodejno (nenadzorovano) učenje hierarhične predstavitve vhodnih podatkov – od enostavnejših podrobnosti, ki se na višjih plasteh sestavljajo v kompleksnejše objekte. Tovrstno učenje je precej zahtevno in je potrebna tudi razmeroma obsežna zbirka učnih primerov. Zato smo zasnovali poenostavljen model globoke mreže. Glede na to, da imamo na področju razpoznavanja govora dolgo časa razmeroma nespremenjen način izračuna predstavitve govornega signala – t.i. parametrizacijo, smo se odločili, da ga uporabimo in z nekaj poenostavitvami združimo z običajnim klasifikacijskim delom v enotno večplastno mrežo, opisano v nadaljevanju.

3 Enostavna zasnova globoke mreže za razpoznavanje govora

Klasični postopek parametrizacije lahko z nekaj prilagoditvami realiziramo v obliki večih plasti nevronov; v našem modelu bodo to spodnje plasti, ki jih bomo označili kot parametrizacijski del globoke mreže.

Izhodna plast spodnjega dela je hkrati vhodna plast zgornjega dela, ki skupaj z dvema plastema (skrita in izhodna) tvori klasifikacijski del globoke mreže. Običajno srečamo v SRG le klasifikacijski del v nevronski mreži, medtem, ko se parametrizacija izvaja ločeno.

Globoko mrežo in vse potrebne postopke smo implementirali s programskim paketom »NICO Toolkit« [5], ki omogoča gradnjo tudi nestandardnih arhitektur in razmeroma enostavno dodajanje lastnih tipov nevronov.

3.1 Parametrizacijski del globoke mreže

Celoten klasični postopek parametrizacije značilk MFCC smo modelirali s pomočjo globoke (večplastne) mreže. Vse osnovne korake tega procesa smo implementirali v naslednjih plasteh:

- plast »okno«:
  dodali smo nevrone z linearno aktivacijsko funkcijo (v nadaljevanju jih bomo krajše imenovali kar kot linearne) in utežmi, ki ustrezajo Hammingovemu oknu,
- plast »DFT«:
  dodali smo linearne nevrone z utežmi, ki ustrezajo izračunu DFT; pri tem smo ločili veji za realno in imaginarno komponento,
- plast »množica filtrov« oziroma »FB«:
  dodali smo linearne nevrone z utežmi, ki ustrezajo po melodični lestvici (angl. »Mel Scale«) razporejeni množici filtrov, ki so običajni del vseh postopkov parametrizacije,
- plast »MFCC«:
  dodali smo linearne nevrone z utežmi, ki ustrezajo diskretni kosinusni transformaciji; ta stopnja je sicer najbolj potrebna zaradi dekorelacije pri uporabi modelov HMM, v našem primeru pa je bolj pomembna zaradi glajenja spektra iz predhodne plasti.

Za opisane plasti smo v programski paket dodali nove ukaze, ki jih vstavijo v mrežo in ustrezno povežejo, da kot celota ustrezajo klasičnemu postopku izračuna značilk MFCC.

Nekatere podrobnosti standardnega izračuna teh značilk smo morali izpustiti ali poenostaviti, da smo jih lahko predstavili z globoko mrežo. Tako smo izpustili poudarjanje (angl. »Preemphasis«), nadomestitev prve značilke MFCC z energijo okvirja, izračun odvodov značilk (t.i. delta značilke) in še nekaj drugih manj pomembnih podrobnosti.

3.2 Zakaj parametrizacija?

V primerjavi z obstoječimi SRG je največja razlika našega modela v tem, da se izračun značilk MFCC opravi na parametrizacijskih plasteh iste mreže in ne kot ločen postopek, ki običajno ni predmet optimizacije oziroma prilagoditve obstoječim podatkov. S tem, ko smo ga vključili v mrežo, lahko podobno kot klasifikacijski del učimo in spreminjamo vrednosti tudi v parametrizacijskem delu. V bistvu smo s tem v mrežo vgradili znanje v obliki implementacije klasičnega izračuna značilk MFCC, ki ga sedaj lahko po potrebi tudi prilagodimo in optimiziramo enakovredno z ostalimi plastmi.

Opisan model ima nekaj prednosti in tudi nekaj slabosti. Med prednosti vsekakor sodi njegova manjša računska zahtevnost. Z vključitvijo parametrizacije smo vnaprej določili vrednosti nekaterih plasti in s tem bistveno zmanjšali njihovo razsežnost in skrajšali postopek učenja; faza nenadzorovanega učenja sedaj ni več potrebna.

Seveda pa smo po drugi strani s tem omejili sposobnost prilagajanja mreže učnim primerom; zaradi tega uspešnost mreže ne bo več primerljiva s kompleksnejšo zasnovo, vendar bo računsko precej manj zahtevna. Kljub omejitvam pa nam lahko tako poenostavljen model verjetno ponudi kar nekaj koristnih izkušenj in spoznanj, predvsem v zvezi s postopkom parametrizacije. Med najpomembnejšimi je vprašanje ali je postopek parametrizacije, kot ga že dolga leta poznamo, res optimalen in kje bi ga morda lahko izboljšali? Postopek učenja nevronske mreže namreč poskuša zmanjševati klasifikacijsko napako in s tem v našem modelu izboljšuje tudi sam parametrizacijski postopek.

3.3 Učenje globoke mreže

Za preliminarni preizkus smo uporabili kar enostaven splošno znan postopek vzvratnega razširjanja napake. Ker nas med drugim zanima tudi vpliv posameznih parametrizacijskih plasti na uspešnost razpoznavanja, smo postopke učenja izvajali v več zaporednih korakih:

- začetno stanje mreže:
  - parametrizacijske plasti se postavijo na vrednosti, ki ustrezajo klasičnemu postopku parametrizacije,
  - klasifikacijske plasti pa se postavijo na manjše naključne vrednosti,
- osnovno učenje; izvede se več iteracij učenja klasifikacijskih plasti (parametrizacijski se ne spreminjajo),
- dodatna sprostitev plasti in ponovitev učenja; spreminjajo se tako klasifikacijske kot tudi izbrane plasti parametrizacijskega dela.

Takojšnja sprostitev vseh plasti ob uporabi vzvratnega razširjanja napake ne bi bila smiselna, ker bi se parametrizacijski del prehitro spremenil. Zato smo mrežo učili postopoma po opisanih korakih. Pri tem nas je predvsem zanimalo, ali dodatno sproščanje posameznih parametrizacijskih plasti pripomore k bolj uspešni razpoznavi.

4 Praktični preizkus modela

Praktični preizkus smo izvedli na testnem sistemu za klasifikacijo osnovnih glasov v 5 glavnih razredov. Ta preizkus je pripravljen po vzoru splošno znanega testa z imenom »5 broad classes«, ki je dodan znanemu programskemu paketu HTK [6] in je splošno priznan in uveljavljen. Mi smo ga izbrali zato, kjer je uspešnost sistemov pri tej klasifikaciji nekoliko višja od sicer bolj običajne fonemske in so zato razlike v uspešnosti nekoliko bolj izrazite in posledično omogočajo lažjo analizo.

Še pred samim preizkusom smo najprej določili vse potrebne podrobnosti postopka parametrizacije. V našem primeru smo govorni signal najprej razdelili na 16ms dolge okvirje; pri frekvenci vzorčenja 8000Hz to pomeni 128 vzorcev digitalnega signala v enem okvirju. Okvirji so si sledili na 10ms, kar pomeni 100 okvirjev v eni sekundi. Mreža je torej imela 128 vhodov za vse vzorce posameznega okvirja. V drugi plasti so se te vrednosti pomnožile s Hammingovim oknom (plast »okno«). Na naslednji plasti se je opravil izračun amplitudnega odziva s pomočjo DFT (plast »DFT«). Sledi še razdelitev spektra na 24 širših »kritičnih« pasov s pomočjo množice trikotnih filtrov (plast »FB«). Na

tako dobljenem spektru uporabimo še diskretno kosinusno transformacijo (DCT) in dobimo končni vektor 13 značilk MFCC. Te predstavljajo vhodne podatke oziroma prvo (vhodno) plast klasifikacijskega dela mreže; nad njo se nahaja še klasična skrita plast s 400 nevroni in na koncu še izhodna plast.

Množice učnih, validacijskih in testnih primerov smo pripravili iz govorne zbirke ŠTEVKE [7]. Pri tem smo posnetke le razdelili na posamezne okvirje po 128 vzorcev – ves ostali izračun značilk MFCC se namreč opravi v sami mreži. Vseh 5 razredov je bilo približno enako zastopanih v vseh množicah.

Najprej smo vzpostavili začetno stanje mreže. Parametrizacijski del smo nastavili tako, da je izvrševal že opisan postopek izračuna MFCC. Klasifikacijske plasti smo nastavili na manjše naključne vrednosti in smo v zaporedju sprožili najprej osnovno učenje mreže v 120 iteracijah ob sproščenem klasifikacijskem delu (»klasifikator«). Nato smo na tako naučeni mreži sprožili še dodatne postopke učenja z dodatno sproščenimi izbranimi parametrizacijskimi plastmi. Za vsako smo dobili nov sistem, katerega uspešnost smo preizkusili v praksi. V tabeli 1 smo jih označili kar z imeni plasti, ki so bile dodatno sproščene. Za boljšo primerjavo smo na nekoliko drugačen način naučili še dva sistema :

- »parametrizacija«:
  sprostili smo vse parametrizacijske plasti in ponovili učenje,
- »kontrolni test«:
  ponovili smo učenje brez sprostitve katerekoli parametrizacijske plasti; ta preizkus služi ovrednotenju učinka dodatnega učenja na nespremenjeni mreži.

V vseh primerih dodatnega učenja smo postopek izvedli v manjšem številu iteracij (40) in z nekoliko manjšim učnim koeficientom. Namerno ga nismo preveč zmanjšali zato, da bi našli tudi kakšno sosednjo, bolj izrazito točko optimalnosti in se izognili preprostemu globljemu sestopu v lokalni optimum. Ta situacija je verjetno glavna slabost obstoječega postopka vzvratnega razširjanja napake – po začetnem učenju se verjetno nahaja blizu lokalnega optimuma in se kljub sprostitvi parametrizacijskih plasti ne bo preveč oddaljil od obstoječe točke. Ta problem bo vsekakor predmet naših nadaljnjih raziskav.

Rezultati preizkusa so prikazani v tabeli 1. Vidimo, da ima sproščanje parametrizacijskih plasti pozitivni vpliv na uspešnost razpoznavanja. Kontrolni test pa hkrati pokaže, da samo ponovitev učenja brez sprostitve dodatnih plasti ni enako uspešen. To pomeni, da se mreža s ponovnim učenjem lahko še bolje prilagodi učnim primerom le s sproščanjem dodatnih plasti.

Tabela 1. Primerjava uspešnosti klasifikacije v 5 osnovnih razredov za različne postopke učenja globoke mreže.

Oznaka sistema	Srednja kvadratna napaka na učni množici	Učna množica	Testna množica
»klasifikator«	1.800	57,7	57,6
»okno«	1.676	62,8	62,2
»FB«	1.627	64,8	64,2
»FB« in »MFCC«	1.541	67,6	66,4
»parametrizacija«	1.544	67,2	66,1
»kontrolni test«	1.661	63,6	62,9

5 Zaključek

Opisani preizkus pomeni samo uvod v delo na tem področju. Ideja enotne globoke mreže je zelo privlačna, seveda pa potrebuje nadaljnje raziskave in izboljšave. Obravnavan model globoke nevronske mreže za razpoznavanje govora namreč bistveno poenostavi njegovo učenje in uporabo. Seveda pa se mora najprej dokazati s svojo uspešnostjo, ki pa ni edini in izključni cilj tovrstnih raziskav. Zelo veliko se lahko namreč naučimo tudi o postopku parametrizacije, ki se še danes bolj ali manj nespremenjen uporablja praktično v vseh SRG. Potencialna možnost njegove optimizacije bi imela zelo močan vpliv na nadaljnji razvoj SRG. Prikazani rezultati so v tem pogledu zelo obetavni in bomo z raziskavami na tem področju vsekakor nadaljevali.

Literatura

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE Signal Processing Magazine, **29**, November 2012 (in press).

[2] ROZMAN, Robert, ŠTRANCAR, Andrej, KODEK, Dušan. Povečevanje robustnosti sistemov za razpoznavanje govora in optimizacija procesa parametrizacije. V: ZAJC, Baldomir (ur.). *Zbornik desete Elektrotehniške in računalniške konference ERK 2001,* zv. B, str. 257–260.

[3] Robert Rozman, "Enostavnejša zasnova sistema za razpoznavanje govora", *Elektrotehniški vestnik*, letn. 80, št. 4, str. 171-176, 2013.

[4] Garofolo, John, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[5] The NICO Toolkit, http://nico.nikkostrom.com/.

[6] The Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk/.

[7] ROZMAN, Robert. *Nesimetrične okenske funkcije v sistemih za razpoznavanje govora : doktorska disertacija*. Ljubljana: [R. Rozman], 2005. VI, 128 f., ilustr.