Stefan Janković

Boosted SMRNN: on-line prepoznavanje govora

Upoređivane su različite arhikteture dvosmernih rekurentnih neuronskih mreža na problemu automatskog prepoznavanja govora nezavisno od govornika, na neograničenom rečniku, baziranog na hibridnom sistemu neuronska mreža / skriveni Markovljev model. BSMRNN arhiktetura je postigla iznenađujuće bolje rezultate u odnosu na ostale topologije. Uz pomoć algoritma AdaBoost.M1 su postignuti bolji rezultati sa dve mreže koje, za razliku od dvosmernih rekurentnih mreža, rade u realnom vremenu, bez dodatne cene u vidu kompjuterskog vremena. Našli smo da mali broj fonema srpskog jezika pogoduje ovakvom sistemu.

Uvod

Modeli za automatsko prepoznavanje govora

Problem automatskog prepoznavanja govora (automatic speech recognition, u daljem tekstu ASR) sastoji se u pisanju kompjuterskog programa koji je u stanju da na osnovu analize zvuka emituje istu sekvencu reči koju bi čula osoba koja sluša taj zvuk. Razvijena su dva osnovna modela za prepoznavanje govora: perceptivni i generativni.

Perceptivni model se sastoji iz dve odvojene komponente. Prva modeluje relaciju između kratkih zvučnih intervala i lingvističkih jedinica. Nezavisne verovatnoće da određen zvučni interval pripada određenoj lingvističkoj jedinici prosleđuje se do druge komponente, koja skup nezavisnih verovatnoća interpretira u skup smislenih reči. Proces u generativnom modelu se kreće u suprotnom smeru: polazi se od pretpostavljene reči, i modeluje se veza između pretpostavke i lingvističkih jedinica, i tek na kraju veza između lingvističkih jedinica i intervala zvuka.

Najčešća lingvistička jedinica koja se koristi unutar sistema za ASR jeste fonema, zbog njihovog relativno malog broja. Za neke specijalizovane aplikacije (npr. prepoznavanje izgovorenih cifara), osnovna lingvistička jedinica može biti i cela reč. Svaki kompletni sistem za automatsko prepoznavanje govora ima ograničen rečnik. Čak je i kod ljudi primećeno da nasumično izgovorene reči prepoznaju teže nego kada imaju informaciju o temi o kojoj se razgovara.

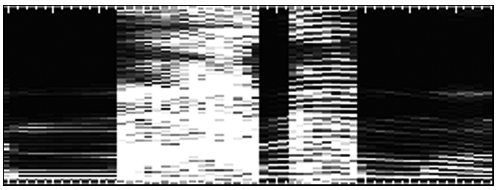
Stefan Janković (1987), Beograd, Obilićev venac 30/1, učenik 4. razreda Matematičke gimnazije u Beogradu

U ovom radu je korišćen perceptivni model sa fonemama kao osnovnim lingvističkim jedinicama. Svaki frejm je bio klasifikovan nezavisno od ostalih. Sistem je bio nezavisan od govornika i nije bilo rečnika.

Veštačke neuronske meže i automatsko prepoznavanje govora

Dok se za generativni model najčešće koristi skriveni Markovljev model (hidden Markov model, u daljem tekstu HMM) (Rabiner 1989), koji je u stanju da objedini oba dela generativnog procesa, za perceptivni model se koristi hibrid veštačke neuronske mreže i HMM-a (Morgan *et al.* 1994).

Glavni problem tokom klasifikacije fonema po frejmu je što se tokom govora za različite foneme emituje isti zvuk. Ljudi isti zvuk čuju drugačije u zavisnosti od prethodnog i budućeg zvuka, tj. subjektivni doživljaj zvuka je kontekstualno zavistan. Postoji više arhiktetura neuronskih mreža koje su u stanju da se nose sa kontekstualnom zavisnošću. U članku na http://www.idsia.ch/~juergen/ se pokazuje da LSTM arhiktetura postiže bolje rezultate od tradicionalne rekurentne mreže, kao i da dvosmerne rekurentne mreže nadmašuju jednosmerne rekurentne mreže.



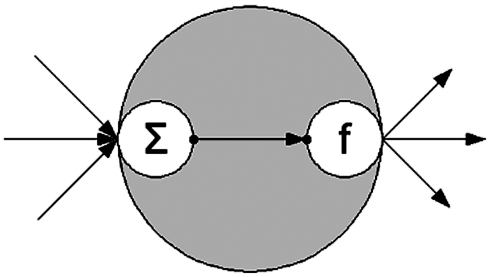
Slika 1. Dva osvetljena dela pretstavljaju spektograme foneme 'A' u različitim kontekstima

Figure 1. Two lighter parts are spectrograms of phoneme 'A' in different contexts

U ovom radu su upoređene tri različite arhikteture dvosmernih rekurentnih neuronskih mreža: LSTM (Long-Short term memory), koja je u stanju da modeluje zavisnosti između udaljenih vremenskih intervala, SMRNN (Segmented-Memory Recurrent Neural Network) (Jinmiao *et al.* 2004), koja je pogodna za modelovanje zavisnosti između srednje udaljenih vremenskih intervala, i tradicionalne RNN (Recurrent Neural Network), koja je u stanju da modeluje zavisnost isključivo između bliskih vremenskih intervala (Hochreiter 1998). Iako ASR nije on-line problem, vrlo je poželjno da sistem radi u realnom vremenu. Glavni nedostatak dvosmernih rekurentnih mreža jeste što su u stanju da klasifikuju foneme tek na kraju sekvence. Prevazilaženje ovog problema je u ovom radu pokušano preko metoda za kreiranje komiteta.

Koncepti veštačkih neuronskih mreža

Neuron – svaka veštačka neuronska mreža se sastoji od neurona, najčešće organizovanih u slojeve. Neuron sumira ulaze, i daje izlaz koji zavisi od aktivacione funkcije, najčešće sigmoidalne (slika 2). Izlaz se



Slika 2. Tipični veštački neuron

Figure 2. Typical artificial neuron

dalje prosleđuje kroz veze do drugih neurona. Veze imaju težine, tako da se izlaz neurona množi sa težinom veze. Svaka mreža se sastoji od ulaznih, izlaznih i, najčešće, od skrivenih neurona.

Višeslojni perceptron (Multi Layer Perceptron – MLP) – jeste veštačka neuronska mreža koja se sastoji iz slojeva neurona i veza koje strogo pokazuju od slojeva bližih ulaznom sloju ka slojevima bližim izlaznom sloju. Modelovanje kontekstualnih zavisnosti unutar MLP-a se postiže povećanim ulaznim prozorom, tj. mreža prima informacije iz više vremenskih koraka odjednom (Zurada 1992).

Rekurentna neuronska mreža (RNN) – za modelovanje kontekstualnih zavisnosti između udaljenih vremenskih koraka, MLP nije pogodan iz razloga što zahteva veliki broj veza. Zato se koristi RNN, koja sadrži veze koje pokazuju od neurona iz prethodnih ka neuronima iz budućih koraka, ili suprotno. Za modelovanje kontekstualnih zavisnosti i iz prošlosti i iz budućnosti koristi se odloženo ciljanje ili dvosmerne rekurentne neuronske mreže. Pri odloženom ciljanju, mreža pokušava da klasifikuje zvučni signal koji joj je prosleđen nekoliko koraka ranije. Ovakav pristup ograničava mrežu da koristi ograničenu količinu informacija iz budućnosti. Dvosmerne mreže su ograničene samo svojim sposobnostima, ali ne rade u realnom vremenu.

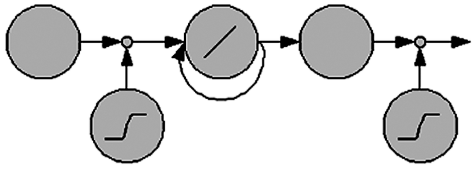
Treniranje neuronskih mreža – u ovom radu je korišćen trening opadanja greške duž gradijenta (gradient descent), pri čemu su parcijalni izvodi izračunati algoritmom propagacije greške unazad kroz vreme (Back-Propagation Through Time – BPTT). Trening se sastoji od izračunavanja parcijalnih izvoda funkcije greške po promenljivoj (u slučaju neuronskih mreža – veze), i promene promenljivih srazmerno njihovim izvodima i nekoj konstanti (učeći koeficient) (www.idsia.ch). Ciklus se ponavlja sve dok se ne dobiju zadovoljavajući rezultati.

Eksponencionalno opadanje greške

Sepp Hochreiter (1998) je pokazao da pri gradient descent treningu, bez obzira koja se funkcija greške i koji se algoritam za računanje izvoda koristi, RNN nije u stanju da modeluje kontekstualne zavisnosti većeg vre-

menskog intervala. Posmatrano iz ugla algoritma BPTT, osnovna ideja je u sledećem: greška koja se propagira od neurona *j* do neurona *i* srazmerna je prvom izvodu aktivacione funkcije neurona *i*, i vrednosti veze između tih neurona. Kako je najveća vrednost prvog izvoda sigmoidalne funkcije kodomena (0,1) jednaka 0.25, to, ako je absolutna vrednost veze između neurona manja od 4.0, propagirana greška eksponencijalno opada. Posledice su da signal greške propagiran iz vremenski udaljenog koraka ne utiče značajno na promenu veza. Ukoliko je absolutna vrednost veze izmedju neurona veća od 4.0, očekuje se da signal greške eksponencijalno raste, što dovodi do velikih oscilacija vrednosti promenljivih tokom treninga.

Visoke vrednosti veza mogu dovesti do loše generalizacije, te nisu pogodne. Veliki broj neurona ne utiče značajno na smanjivanje opadanja signala greške, jer se signali primljeni iz različitih neurona međusobno potiru ukoliko su neki veći, a neki manji od nule. Veliki broj promenljivih parametara takođe dovodi do loše generalizacije. Hochreiter na osnovu teorijske analize pokazuje da RNN sa sigmoidalnim neuronima može da modeluje kontekstualnu zavisnost do 10 vremenskih koraka. Baldi (et al. 2000) navodi da je empirijski primećeno da na problemu predikcije sekundarne strukture proteina RNN može da modeluje kontekstualnu zavisnost do 15 vremenskih koraka.



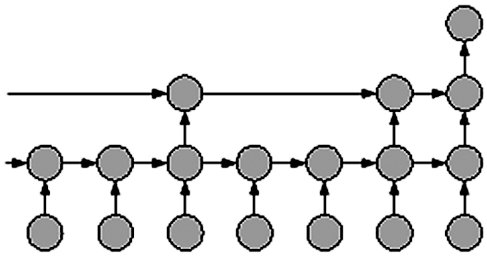
Slika 3. Osnovni LSTM blok

Figure 3. Basic LSTM block

Za modelovanje kontekstualnih zavisnosti većih vremenski intervala dobre rezultate su pokazale dve arhikteture. Pri jednoj, neuroni se organizuju u LSTM blokove (long-short term memory, slika 3). Ovi blokovi se sastoje od linearnih neurona koji imaju rekurentnu vezu konstantne vrednosti 1.0 koja pokazuje samo na njih same. Ovo omogućava da greška koja se propagira unazad kroz tu vezu ne menja vrednost, osim pri spoljašnjem uticaju. Linearni neuroni su okruženi ulaznim neuronima i ulaznim kapijama sa jedne, i izlaznim neuronima i izlaznim kapijama sa druge strane. Ulazne i izlazne kapije su neuroni sa aktivacionom funkcijom kodomena (0,1). Izlazi iz ulaznog neurona i kapije se množe i prosleđuju do linearnog neurona. Izlaz iz linearnog neurona se prosleđuje do izlaznog neurona i aktivacija izlaznog neurona se množi sa aktivaciom izlazne kapije. Kasnije su LSTM blokovi proširени sa tzv. kapijama za zaboravljanje, kao i direktnim vezama od linearnih neurona do kapija – pogledati http://www.idsia.ch/~juergen/. U ovom radu su korišćene proširene

verzije LSTM blokova. Eksperimenti (Hochreiter 1998) su pokazali da su LSTM neuronske mreže u stanju da modeluju kontekstualne zavisnosti između ulaza udaljenih i do preko 1000 vremenskih koraka.

Druga arhiktetura, koja je dala dobre rezultate na problemu predikcije sekundarne strukture proteina (Jinmiao *et al.* 2004), je rekurentna mreža sa segmentiranom memorijom (segmented-memory recurrent neural network – SMRNN, slika 4). Ova topologija je nastala na osnovu analize ljudskog pamćenja. Ljudi imaju običaj da duge vremenske intervale dele u segmente i da pamte osobine segmenata, ne svakog pojedinačnog trenutka. SMRNN se sastoji od dva rekurentna sloja. Jedan sloj, sloj na nivou simbola, sadrži rekurentne veze između dva uzastopna vremenska koraka. Drugi sloj operiše na nivou segmenata. Na početku svakog segmenta, neuroni iz ovog sloja prosleđuju informacije do neurona u narednih d koraka. Ako je obična rekurentna mreža u stanju da modeluje kontekstualnu zavisnost do 10 vremenskih koraka, onda je SMRNN u stanju da modeluje vremensku zavisnost do manje od 100 vremenskih koraka.



Slika 4.
Dinamika SMRNN mreže za interval $d = 3$

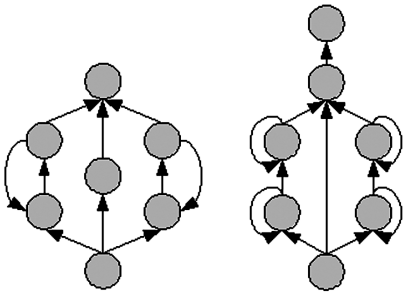
Figure 4.
Dynamics of SMRNN for interval $d = 3$

U ovom radu se prvi put primenjuje SMRNN arhiktetura na problem klasifikacije fonema. Motiv za njihovu primenu je što, iako su dve arhikteture u stanju da ne koriste obavezno informacije iz dalekih vremenskih intervala, problem klasifikacije fonema po frejmu ne predstavlja problem veoma dugih vremenskih zavisnosti. Promenom intervala d se omogućava eksplicitna kontrola nad maksimalnim protokom signala greške kroz mrežu.

Komiteti veštačkih neuronskih mreža

Više neuronskih mreža se mogu organizovati u komitet. Ukoliko su greške koje mreže (ili bilo koji drugi sistem za klasifikaciju) daju statistički nezavisne, komitet omogućava da se ukupna greška smanji (Ranawana 2006).

Malo je verovatno da će greške biti statistički nezavisne prostim treniranjem više mreža. Postoji više načina za rešavanje ovog problema. U ovom radu je korišćeno nekoliko algoritama, ali je algoritam AdaBoost.M1



Slika 5. Levo: BRNN za predikciju sekundarne strukture proteina (Baldi *et al.* 2000)

Desno: BSMRNN za predikciju sekundarne strukture proteina (Jinmiao *et al.* 2004)

Figure 5. Left: BRNN for protein secondary structure prediction

Right: BSMRNN for protein secondary structure prediction

dao najbolje rezultate. Pri ovom algoritmu, mreže se treniraju sekvencijalno. Pri treniranju prve mreže, distribucija primera (težina greške) je uniformna. Za svaku sledeću mrežu distribucija se menja tako da se netačno klasifikovanim primerima iz prethodnog ciklusa dodeljuju veća distribucija. Pogledati (Freund *et al.* 1996).

Rezultati i diskusija

Eksperimentalna postavka

Govorna baza. Govorna baza koja je korišćena u svim eksperimentima je S70W120s120, preuzeta od AlfaNum tima. Baza je snimljena 1983. godine na gramofonskim pločama, ali je kasnije presnimljena u digitalni format. Format digitalnog zapisa je mono PCM, 16 bita po odmerku, 22050 odmeraka po sekundi. Deo baze koja je korišćena se sastoji od 114 govornika, koji su ukupno izgovorili 6632 rečenice. Ukupno je bilo izdvojeno 4.944 (oko 75%) rečenica za trening skup, 340 (oko 5%) za potvrđujući skup, i 1.348 (oko 20%) za test skup. Između skupova nije bilo istih govornika. Rečenice su pretprocesirane uz pomoć HTK toolkit-a (http://htk.eng.cam.ac.uk), u 12 standardnih MFC (Mel-Frequency Cepstral) koeficienata iz 26 kanala, logaritam energije i prve izvode ovih koeficienta, sa normalizovanim energijama, što je ukupno niz od 26 koeficienta po frejmu. Ulazni prozor je bio 25 ms, i uzastopni prozori su se preklapali 15 ms. Tokom preliminarnih eksperimenata primećeno je da se foneme "dž", "đ", "č" i "ć" teško raspoznaju putem neuronskih mreža, te su foneme "dž" i "đ", kao i "č" i "ć" posmatrane kao jedna. Običaj u sadašnjim sistemima za prepoznavanje govora je da se spajaju različite foneme sa sličnim akustičnim osobinama, ako ne utiču na promene rezultata pri određenom rečniku. Zajedno sa tišinom ukupno je bilo 29 različitih fonema.

Trening. Tokom svih eksperimenata ulazni sloj se sastojao od 26 linearnih neurona, po jedan za svaki koeficient. Izlazni sloj se sastojao od 29 neurona sa SoftMax aktivacionom funkcijom, po jedan za svaku

fonemu. Korišćena je Cross-Entropy objektna funkcija. Svi neuroni skrivenih i izlaznog sloja su imali pomeraj (bias). Sve veze su bile inicijalizovane unutar granica (–0.1, 0.1). Sve mreže su bile trenirane na trening skupu i testirane na potvrđujućem skupu. Težine veza na mreži su postavljene na one koji su pokazivali najbolje rezultate na potvrđujućem skupu, i mreža je testirana na test skupu. Za rezultate je uziman procenat tačnih klasifikacija koje je mreža vršila. Posle završetka treninga vrednosti unutar mreža su bile postavljene na one koje su pokazivale najmanju grešku na potvrđujućem skupu, i bile testirane na test skupu. Sve mreže su trenirane preko postepenog opadanja greške. Iako se za SMRNN originalno koristi proširen oblik RTRL algoritma (Real-Time Recurrent Learning), da bi smo učinili uslove mreža jednakim, parcijalni izvodi su računati preko BPTT algoritma, koristeći momenat 0.9 i različite učeće koeficiente. Veze su osvežavane nakon svake rečenice. Sve mreže su imale oko 75000 veza.

Poređenje različitih topologija neuronskih mreža

U toku ovog eksperimenta, testirane su dve različite arhikteture za modelovanje kontekstualnih zavisnosti dugih vremenskih intervala. Takođe, testirana je i tradicionalna BRNN sa sigmoidalnim neuronima. Testirane su sledeće arhikteture (tabela 1):

Tabela 1. Poređenje različitih topologija.

Topologija	Rezultati na test skupu	Rezultati na trening skupu	Broj ciklusa tokom treninga	idsia*
BRNN	67.7%	71.4%	52	69.0%
BLSTM	71.7%	75.2%	75	69.8%
BSMRNN	75.6%	79.5%	86	–

* http://www.idsia.ch/~juergen

1. BRNN, potpuno povezana, sa sigmoidalnim neuronima unutar po jednog skrivenog sloja, za kontekst iz budućnosti i iz prošlosti, sa učećim koeficientom $\alpha = 10^{-5}$.

2. BLSTM, potpuno povezana, sa po jednim skrivenim slojem sa LSTM blokovima za kontekst iz budućnosti i iz prošlosti, sa učećim koeficientom $\alpha = 10^{-6}$.

3. BSMRNN sa po 117 neurona u slojevima nivoa simbola, i sa po 88 neurona u slojevima nivoa segmenata sa intervalom 4. Svi neuroni su imali sigmoidalnu aktivacionu funkciju kodomena (0,1). Koeficient učenja je bio $\alpha = 8 \cdot 10^{-6}$.

Različiti koeficienti su korišćeni da bi mreže imale iste uslove, tj. da mreže koje brže konvergiraju ne bi preskočile minimum. Interval u sloju nivoa segmenata BSMRNN arhikteture je 4. Pri ovom intervalu, očekuje se da je mreža u stanju da modeluje kontekstualne zavisnosti od oko 50 vremenskih koraka pre i posle trenutno posmatranog koraka, tj. oko 1 sekunde emitovanog zvuka.

U tabeli 1 su dati rezultati poređenja različitih topologija. BRNN nije davala bolje rezultate do 75-og, dok BLSTM i BSMRNN nisu postigle bolje rezultate do 100-og trening ciklusa. BLSTM se posle 75-og ciklusa naglo specijalizovala, tj. razlika između trening i test skupa naglo je počela da raste. Poslednja kolona pokazuje rezultate na test skupu TIMIT baze govora na engleskom jeziku.

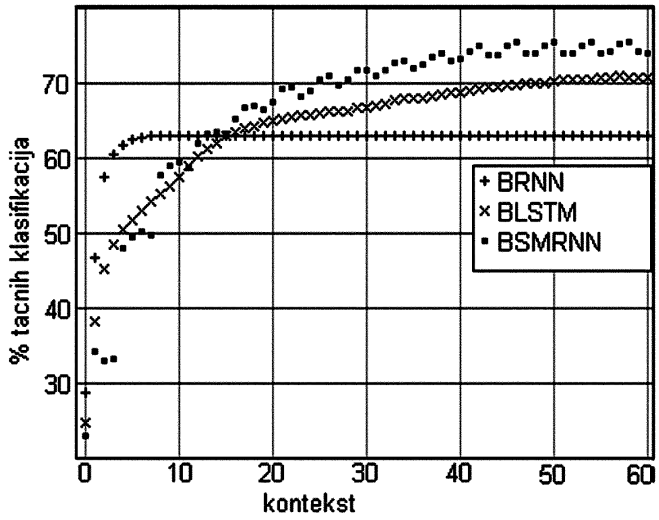
BSMRNN arhiktetura je prvi put primenjena na ovom problemu. Izgleda da njena mogućnost da se skoncentriše na određenu količinu informacija predstavlja značajnu prednost u odnosu na ostale arhikteture. Tokom preliminarnih eksperimenata, nijednom drugom kompleksnijom arhikteturom (npr. sa više skrivenih slojeva, rekurentnim vezama između različitih slojeva) nisu postignuti bolji rezultati.

Upoređujući rezultate sa onim na http://www.idsia.ch/~juergen/, nije jasno da li srpski jezik omogućuje lakše raspoznavanje u odnosu na druge jezike usled malog broja fonema. Primetimo da su se tamo koristile veće mreže (sa oko 100.000 veza), i TIMIT baza sadrži manju količinu šuma (S70W120s120 baza je prvobitno snimljena na gramofonskim pločama), ali i da su u ovom radu tokom predprocesiranja emitovane energije bile normalizovane, da bi se omogućilo lakše upoređivanje različitih govornika. Takođe, specifičnosti samih baza i druge razlike tokom predprocesiranja otežava poređenje. U najgorem slučaju, deluje da se na srpskom jeziku mogu postići makar uporedivi rezultati.

U sledećem testu mreže su imale za zadatak da klasifikuju foneme sa parcijalnim kontekstualnim informacijama. Test je sproveden na onim frejmovima unutar test skupa koji su okruženi sa barem 75 frejmova sa obe strane. Mreže su u početku pokušale da klasifikuju izolovane primere, zatim im je dozvoljeno da koriste kontekstualne informacije 1 frejm pre i posle, itd. Rezultati su prikazani na slici 6.

BRNN nije koristila informacije udaljenije od 9 vremenskih koraka. BSMRNN nije davala drugačije rezultate pri dozvoljenom kontekstu udaljenijem od 45 koraka, dok je BLSTM modelovala kontekstualne informacije udaljene do oko 60 koraka. Pri potpunim kontekstualnim informacijama, BLSTM je pokazivala nešto lošije rezultate. Oscilacije u rezultatima BSMRNN potiču od arhikteture (oscilacije imaju period 4, kao i interval nivoa segmenata BSMRNN mreže).

Rezultati ukazuju da ni BSMRNN ni BLSTM arhikteture nisu u mogućnosti da u potpunosti eksploatišu vremenski bliske kontekstualne infor-



Slika 6.
Procenat tačnih klasifikacija u zavisnosti od odgovarajućeg teksta pre i posle posmatranog koraka. Rezultati za BSMRNN su konvergirali približno za 45 koraka, za BLSTM za oko 60

Figure 6.
Percentage of correct classification depending on suitable text before and after observed iteration. Results for BSMRNN were stabilized after approximately 45 iterations, for BLSTM after 60 iterations.

macije. Pri BSMRNN arhikteturi, ovaj problem bi mogao da se reši dodavanjem direktnih veza između sloja nivoa simbola i izlaznog sloja. Za BLSTM arhikteturu, u (Schmidhuber *et al.* 2005) se napominje da kompleksnije arhikteture nisu pokazale bolje rezultate.

Rezultati na komitetu neuronskih mreža

Komitet se sastojao od dve SMRNN mreže, obe su posmatrale rečenice od početka ka kraju, bez odloženog ciljanja. Obe mreže su imale isti broj neurona kao dva dela BSMRNN iz eksperimenta 2. Dodato je 29 veza, (od 'bias'-a do izlaznih neurona), da bi dve mreže bile u potpunosti nezavisne, što nije značajno. Ovakva postavka omogućava da sistem za ASR radi u realnom vremenu, pri istoj ceni u vidu kompjuterskog vremena kao i dvosmerna mreža.

U radu Hernández-Espinosa *et al.* (2003) je uporedjeno više algoritama za kreiranje komiteta. Prateći njihove rezultate, u toku ovog eksperimenta su primenjeni algoritmi CELS (Liu *et al.* 1999), AdaBoost.M1 i AdaBoost.M2 (Freund 1996). Kombinovanje rezultata koje daju mreže unutar komiteta se može vršiti na više načina. Pri svim algoritmima najbolji rezultati su postignuti koristeći kombinovanje rezultata nalaženjem geometrijske sredine. Kombinovani rezultati su dalje normalizovani, tako da suma svih izlaza koje daje komitet bude jednaka 1. Pogledati dodatak A za sve rezultate.

Koeficienti učenja su bili $\alpha = .2 \cdot 10^{-5}$. Rezultati u tabeli 2 su za algoritam AdaBoost.M1, uz pomoć koga su postignuti najbolji rezultati. Rezultati na trening skupu su za distribuciju koja je prema algoritmu dodeljena mrežama.

Tabela 2. Rezultati na komitetu neuronskih mreža

Mreža	Rezultat na test skupu	Rezultat na trening skupu	Broj trening epoha
SMRNN (1)	69.0%	73.3%	58
SMRNN (2)	68.0%	63.3%	40
Komitet (1) i (2)	75.8%	79.8%	/
BSMRNN	75.6%	79.5%	86

Kao što se iz tabele 2 vidi, SMRNN (1) nije postigla bolje rezultate do 75-og ciklusa, dok komitet nije postigao bolje rezultate do 65-og ciklusa. Zadnji red predstavlja rezultate iz eksperimenta 2.

Razlika u rezultatima između komiteta SMRNN i dvosmerne SMRNN nije značajna, ali je glavna prednost u tome što komitet radi u realnom vremenu.

Primetimo da bi se bolji rezultati komiteta mogli postići odloženim ciljanjem ili drugom funkcijom za raspodelu distribucije unutar algoritma AdaBoost. Takodje, bolji rezultati bi se mogli postići treniranjem više malih mreža.

Druga mreža unutar komiteta je davala bolje rezultate na test skupu nego na trening skupu. Za uniformnu distribuciju trening skupa, mreža je postizala 71.6% tačnih klasifikacija. Zaključuje se da su neki primeri izuzetno teški, i da bi neuniformna početna distribucija primera dovela do poboljšanja rezultata.

Rezultati na hibridu veštačka neuronska mreža/HMM

Iako neuronske mreže omogućuju da se svaki frejm klasifikuje po fonemi, sama rečenica se, na osnovu izlaza mreža, kreira uz pomoć HMM-a. Pri ovom testu je provereno da li su bolji rezultati na zadatku klasifikacije fonema po frejmu različitih mreža omogućili bolje rezultate na ukupnom prepoznavanju govora (tabela 3).

Tabela 3. Rezultati hibridnih sistema

Sistem	Greška
BLSTM / HMM	20.4%
BSMRNN / HMM	18.3%
Boosted SMRNN / HMM	17.9%
BLSTM / HMM (*)	33.8%

* http://www.idsia.ch/~juergen

Hibridni sistemi su se sastojali od HMM-a i različitih mreža. Trenirani su preko Viterbijevog kriterijuma, na trening skupu. Inicijalizacija parametara HMM-a je izvršena tako da parametri odgovaraju verovatnoći prelaska iz jednog stanja u drugo, unutar baze govora. Izlazne verovatnoće mreža su preko Bajesovog pravila transformisane u očekivanja za HMM. Greška je računata kao minimalni broj prepravki (ubacivanje, izbacivanje ili menjanje karaktera) potreban da bi se dobile tačne rečenice, relativan u odnosu na ukupan broj izgovorenih glasova.

Mreže koje su davale bolje rezultate pri klasifikaciji fonema su omogućile preciznije prepoznavanje govora. Primećuje se iznenađujuće mala greška koju daju sistemi. Upoređujući rezultate sa http://www.idsia.ch/~juergen/, izgleda da mali broj fonema unutar srpskog književnog jezika izuzetno pogoduje skrivenom Markovljevom modelu.

Zaključak

Preko tradicionalne BRNN arhikteture nije moguće u potpunosti modelovati kontekstualne zavisnosti između udaljenih frejmova unutar sistema za ASR baziranog na klasifikaciji fonema po frejmu. Sa druge strane, BLSTM arhiktetura je u stanju da modeluje zavisnosti vrlo udaljenih vremenskih intervala, ali loše modeluje zavisnosti vremenski bliskih frejmova. Našli smo da je BSMRNN arhiktetura naročito pogodna za ovaj problem, i da bi dalja poboljšanja mogla da se dobiju modifikacijom arhikteture. Uz pomoć algoritma AdaBoost.M1 postignuti su uporedivi rezultate sa dve SMRNN mreže, koje, za razliku od BSMRNN, rade u realnom vremenu. Moguće je da bi pri daljim izmenama komitet postizao bolje rezultate.

Srpski jezik je za ovaj problem izuzetno pogodan iz dva razloga: pismo je fonetsko, tj. ne zahteva od sistema da modeluje zavisnost između fonetskih i pisanih jedinica, i mali broj fonema omogućava da se preko HMM-a postignu visoke performanse. Neke situacije su izuzetno teške i Cross-Entropy objektna funkcija nije dovoljna. Početna distribucija primera prema težini prepoznavanja bi omogućila bolje rezultate.

Zahvalnost. Želimo da se zahvalimo AlfaNum timu na ustupljenoj bazi govora.

Literatura

http://www.idsia.ch/~juergen/

Hernández-Espinosa C., Fernández-Redondo M., Ortiz-Gómez M. 2003. Ensemble Methods for Multilayer Feedforward Networks. Proc. of the European Symposium on Artificial Neural Networks Bruges (Belgium), 2003, d-side publi., ISBN 2-930307-03-X, pp. 261-266

Jiménez D. 1998. *Dynamically Weighted Ensemble Neural Networks for Classification*. San Antonio: The University of Texas Health Science Center

Freund Y., Schapire R. 1996. Experiments with a New Boosting Algorithm. Proc. of the Thirteenth Int. Conf. on Machine Learning, pp. 148-156.

Hochreiter S. 1998. Recurrent Neural Net Learning and Vanishing Gradient.

Zurada J. M. 1992. Introduction to Articifial Neural Systems. West Publishing Company

Jinmiao C., Chaudhari N. S. 2004. Improvement of Bidirectional Recurrent Neural Network for Learning Long-Term Dependencies. Icpr, pp. 593-596, 17th International Conference on Pattern Recognition (ICPR'04) - vol. 4

Rabiner L. R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. of the IEEE, vol. 77, no. 2, februar 1989

Liu Y., Yao X. 1998. A Cooperative Ensemble Learning System. Proc. of the World Congress on Computational Intelligence 1998, pp. 2202-2207

Morgan N., Bourlard H. 1994. Connectionist Speech Recognition – A Hybrid Approach. The Kluwer International Series in Engineering and Computer Science

Baldi P., Brunak S., Frasconi P., Pollastri G., Soda G. 2001. Bidirectional dynamics for protein secondary structure prediction. *Lecture Notes in Computer Science*, 1828: 80–104

Ranawana R. 2006. *Multi-Classifier Systems – Review and a Roadmap for Developers*. University of Oxford Computing Laboratory

---

Stefan Janković

Boosted SMRNN: On-Line Speech Racognition

Different architectures of two-way reccurent neural networks were compared in solving the problem of automated speech recognition (independent from a speaker), using unlimited vocabulary, based on hybrid system of neural networks/hidden Markov model. BSMRNN architecture gained supprisingly better results when comared to other topologies. Using AdaBost.M1, better results were accoplished for two networks, which, unlike two-way reccurent networks, work efficiently in real time. It is found that a small number of phonems in the Serbian language is suitable for this system.

Dodatak
Rezultati različitih metoda za kreiranje komiteta

Pri svim testovima su korišćene po dve SMRNN mreže. Korišćeni su algoritmi AdaBoost.M1, AdaBoost.M2 (Freund *et al.* 1996) i algoritam CELS (Liu 1998). Za kombinovanje rezultata korišćene su metode:

MAX – izlaz sa najvećom vrednošću iz svih neuronskih mreža je i izlaz koji daje komitet. Primetimo da pri ovoj metodi izlaz iz komiteta ne predstavlja verovatnoću pripadanja nekoj fonemi.

AVG – izlaz komiteta je jednak aritmetičkoj sredini izlaza mreža.

PROD – izlaz komiteta je jednak normalizovanoj geometrijskoj sredini izlaza mreža.

WAVG – u zavisnosti od ukupnog rezultata koje su davale mreže, za algoritme AdaBoost.M1 i AdaBoost.M2 izlazi iz mreža su imali težine prema formulama unutar algoritama.

DAVG – poverenje u mreže zavisi od izlaza koji daju. Pogledati (Jiménez 1998).

Za CELS su korišćeni koeficienti penala korelacije 0.1, 0.25 i 0.5. Najbolji rezultati su postignuti sa koeficientom 0.1, i ti rezultati su prikazani u sledećoj tabeli:

Tabela D1. Rezultati različitih metoda na test skupu

Metoda	Algoritam		
	AdaBoost.M1	AdaBoost.M2	CELS
MAX	74.0%	73.2%	73.3%
AVG	75.1%	74.0%	74.2%
PROD	**75.8%**	74.8%	74.9%
WAVG	75.1%	74.1%	–
DAVG	75.4%	74.3%	74.4%