Új eredmények a mély neuronhálós magyar nyelvű beszédfelismerésben*

Grósz Tamás[1], Kovács György[2], Tóth László[2]

[1] Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2., groszt@inf.u-szeged.hu
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos krt. 103., {gykovacs, tothl}@inf.u-szeged.hu

Kivonat 2006-os megjelenésük óta egyre nagyobb népszerűségnek örvendenek az akusztikus modellezésben az ún. mély neuronhálók. A hagyományos neuronhálókkal ellentétben a mély hálók sok rejtett réteget tartalmaznak, emiatt a hagyományos módszerekkel tanítva őket nem lehet igazán jó eredményeket elérni. Cikkünkben röviden bemutatunk négy új tanítási módszert a mély neuronhálókhoz, majd a mély neuronhálókra épülő akusztikus modelleket beszédfelismerési kísérletekben értékeljük ki. A különböző módszerekkel elért eredményeket összevetjük a korábban publikált eredményeinkkel.

Kulcsszavak: mély neuronhálók, akusztikus modellezés, beszédfelismerés

1. Bevezetés

A neuronhálós beszédfelismerési technika a reneszánszát éli, köszönhetően a mély neuronhálók feltalálásának. Tavalyi cikkünkben [1] mi is bemutattuk a módszer alapötletét, és az első mély neuronhálós felismerési eredményeinket magyar nyelvű adatbázisokon. A technológia iránti érdeklődés azóta sem csökkent, példának okáért az MIT „Tech Review's" listája a mély neuronhálókat beválogatta a 2013-as év 10 legfontosabb technológiai áttörést jelentő módszere közé. Mindeközben sorra jelennek meg az újfajta mély hálózati struktúrákat vagy tanítási módszereket publikáló cikkek. Jelen anyagunkban néhány olyan új ötletet mutatunk be, amelyekkel az eredeti tanítási algoritmus eredményei még tovább javíthatók, majd a módszereket magyar nyelvű beszédfelismerési adatbázisokon értékeljük ki.

A mély neuronhálók hatékony betanításához az eredeti szerzők az ún. DBN előtanítási módszert javasolták [2], ami egy elég komplex és műveletigényes algoritmus. Egy jóval egyszerűbb alternatívaként vetették fel nemrég az ún. discriminative pre-training („diszkriminatív előtanítás") algoritmust [3]. Ezen módszer

* Jelen kutatást a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett.

esetén az előtanítás felügyelt módon történik: kezdetben egy hagyományos (egy rejtett réteges) hálóból indulunk ki, néhány iteráción keresztül tanítjuk, ezután egy új rejtett réteget illesztünk be a kimeneti réteg alá és a kimeneti réteget újrainicializáljuk. Az így kapott neuronhálót újra tanítjuk néhány iteráción keresztül, az új rétegek hozzáadását pedig addig ismételjük, amíg a rejtett rétegek száma el nem éri a kívánt mennyiséget. A módszer előnye, hogy a tanítás során – mindkét fázisban – csak a backpropagation algoritmust kell használunk.

Egy másik mostanában javasolt, előtanítást nem igénylő módszer az ún. rectified („egyenirányított") neuronok használata. Ezek nevüket onnan kapták, hogy egyenletükben a szokásos szigmoid aktivációs függvény le van cserélve egy olyan komponensre, amelynek működése egy egyenirányító áramkörre hasonlít (matematikailag a $max(0, x)$ függvényt valósítja meg). A rectifier neuronokra épülő mély neuronhálók használatát eredetileg képfeldolgozásban vetették fel, csoportunk az elsők között próbálta ki őket beszédfelismerésben [4]. Eredményeink egybevágnak a más kutatók által velünk párhuzamosan publikált eredményekkel: úgy tűnik, hogy az egyenirányított mély neuronhálók hasonló vagy kicsit jobb felismerési pontosságot tudnak elérni, mint hagyományos társaik, viszont a betanításuk jóval egyszerűbb és gyorsabb [5,6].

Egy harmadik nemrégen feltalált módszer a neuronháló backpropagation tanítási algoritmusát módosítja. Az ún. dropout („kiejtéses") tanulás lényege, hogy a neuronháló tanítása során minden egyes tanítópélda bevitelekor véletlenszerűen kinullázzuk („kiejtjük") a hálót alkotó neuronok kimenetének valahány (általában 10-50) százalékát [7]. Ennek az a hatása, hogy az azonos rétegbe eső neuronok kevésbé tudnak egymásra hagyatkozni, így a probléma önálló megoldására vannak kényszerítve. Ennek köszönhetően lényegesen csökken a túltanulás veszélye. A módszert eredetileg javasló cikkben kiugró, 10-20 százaléknyi relatív hibacsökkenéseket értek el képi alakfelismerési és beszédfelismerési feladatokon.

A javasolt módszerek hatékonyságát először az angol nyelvű TIMIT adatbázison szemléltetjük, mivel ezen számtalan összehasonlító eredmény áll rendelkezésre. Ezután két magyar nyelvű adatbázissal kísérletezünk. Az egyik egy híradós adatbázis, amelynek méretét tavaly óta jelentősen sikerült megnövelnünk. A másik pedig a „Szindbád történetei" című hangoskönyv hangzóanyaga, amelyen szintén publikáltunk már eredményeket korábban.

2. Mély neuronhálók

A hagyományos neuronhálók és a mély hálók között az alapvető különbség, hogy utóbbiak több (általában 3-nál több) rejtett réteggel rendelkeznek. Ezen mély struktúrájú neuronhálók használatát igazolják a legújabb matematikai érvek és empirikus kísérletek, melyek szerint adott neuronszám mellett a több rejtett réteg hatékonyabb reprezentációt tesz lehetővé. Ez indokolja tehát a sok, relatíve kisebb rejtett réteg alkalmazását egyetlen, rengeteg neuront tartalmazó réteg helyett.

A sok rejtett réteges mély neuronhálók tanítása során több olyan probléma is fellép, amelyek a hagyományos egy rejtett réteges hálók esetén nem vagy alig

megfigyelhetőek, és ezen problémák miatt a betanításuk rendkívül nehéz. A hagyományos neuronhálók tanítására általában az ún. backpropagation algoritmust szokás használni, ami tulajdonképpen a legegyszerűbb, gradiensalapú optimalizálási algoritmus neuronhálókhoz igazított változata. Több rejtett réteg esetén azonban ez az algoritmus nem hatékony. Ennek egyik oka, hogy egyre mélyebbre hatolva a gradiensek egyre kisebbek, egyre inkább „eltűnnek" (ún. „vanishing gradient" effektus), ezért az alsóbb rétegek nem fognak kellőképp tanulni [8]. Egy másik ok az ún. „explaining away" hatás, amely megnehezíti annak megtanulását, hogy melyik rejtett neuronnak mely jelenségekre kellene reagálnia [2]. Ezen problémák kiküszöbölésére találták ki az alább bemutatásra kerülő módszereket.

2.1. DBN előtanítás

A mély neuronhálók legelső tanítási módszerét 2006-ban publikálták [2], lényegében ez volt az a módszer, amely elindította a mély neuronhálók kutatását. A módszer lényege, hogy a tanítás két lépésben történik: egy felügyelet nélküli előtanítást egy felügyelt finomhangolási lépés követ. A felügyelt tanításhoz használhatjuk a backpropagation algoritmust, az előtanításhoz azonban egy új módszer szükséges: a DBN előtanítás.

A DBN előtanítással egy ún. „mély belief" hálót (Deep Belief Network, DBN) tudunk tanítani, amely rétegei korlátos Boltzmann-gépek (RBM). A korlátos Boltzmann-gépek a hagyományosaktól annyiban térnek el, hogy a neuronjaik egy páros gráfot kell hogy formázzanak. A két réteg közül a látható rétegen keresztül adhatjuk meg a bemenetet, a rejtett réteg feladata pedig az, hogy az inputnak egy jó reprezentációját tanulja meg.

Az RBM-ek tanításához a kontrasztív divergencia algoritmust (CD) használhatjuk, amely egy energiafüggvény alapú módszer. Egy RBM a következő energiát rendeli egy látható (v) és a rejtett réteg (h) állapotvektor-konfigurációhoz:

$$E(v,h;\Theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}v_ih_j - \sum_{i=1}^{V} b_iv_i - \sum_{j=1}^{H} a_jh_j. \tag{1}$$

Az egy lépéses kontrasztív divergencia algoritmus (CD-1) esetén a következő update szabályt alkalmazzuk a látható-rejtett súlyokra:

$$\Delta w_{ij} \propto \langle v_ih_j\rangle_{input} - \langle v_ih_j\rangle_1, \tag{2}$$

ahol $\langle . \rangle_1$ a látható és a rejtett rétegek Gibbs-mintavételezővel egy lépésben történő mintavételezése utáni kovarianciája.

Habár az RBM energiafüggvénye rendkívül jól működik bináris neuronok esetén, beszédfelismerésben azonban valós bemeneteink vannak, ennek kezelésére szükséges az energiafüggvény (1) módosítása. A valós bemenetekkel rendelkező RBM-et Gaussian-Bernoulli korlátos Boltzmann-gépnek (GRBM) nevezzük, energiafüggvénye:

$$E(v,h|\Theta) = \sum_{i=1}^{V}\frac{(v_i-b_i)^2}{2} - \sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}v_ih_j - \sum_{j=1}^{H} a_jh_j. \tag{3}$$

Ezen új energiafüggvény esetén a CD-1 algoritmusban csupán a Gibbs-mintavételezés módját kell módosítani, a súlyok frissítése pedig továbbra is (2) szerint történik.

A DBN előtanítás során a hálót rétegpáronként tanítjuk. Az első lépésben az inputot és a legelső rejtett réteget egy GRBM-nek tekintve a CD-1 algoritmussal tanítjuk. A továbbiakban a következő RBM-nek a látható rétege az előzőleg tanított RBM rejtett rétege lesz, az új rejtett rétege pedig a következő rejtett réteg a hálóban. Az így inicializált hálók felügyelt tanulással finomhangolva lényegesen jobb eredményeket tudnak elérni, mint az előtanítás nélkül tanítottak.

2.2. Diszkriminatív előtanítás

A diszkriminatív előtanítást (Discriminative pre-training, DPT) a DBN előtanítás alternatívájaként javasolták [3]. Ahogy az elnevezésből sejthető, ez a módszer is két fázisból áll, a különbség, hogy az előtanítást is felügyelt tanítással, a backpropagation algoritmussal valósítjuk meg. Az algoritmus kezdetben egy hagyományos egy rejtett réteges neuronhálóból indul ki, amit néhány iteráción keresztül tanítunk. A következő lépésben egy új rejtett réteget illesztünk be a kimenet és a legfelső rejtett réteg közé, a kimeneti réteg súlyait újrainicializáljuk, majd az egész hálót tanítjuk néhány iteráción keresztül. Mindezt addig ismételjük, amíg a rejtett rétegek száma a kívánt mennyiséget el nem éri. A módszer előnye, hogy nem igényel külön tanítási algoritmust.

A tanítás során felmerül egy fontos kérdés, mégpedig, hogy az előtanítás során meddig tanítunk. Az eredeti cikk [3] szerint az eredmények romlanak, ha minden előtanítási lépésben a teljes konvergenciáig tanítunk. Javasolt csak néhány iterációnyit tanítani - a szerzők 1 iterációnyit javasolnak - mi a 4 iterációnyi előtanítást találtuk a legeredményesebbnek, azonban megemlítjük, hogy ha a tanító adatbázis mérete megnő, akkor az 1 iterációnyi előtanítás is elegendőnek tűnik.

2.3. Rectifier neuronhálók

Tekintve, hogy az előző két előtanításos módszernek rendkívül nagy az időigénye, sok kutató olyan módszereket próbált kidolgozni, amelyek nem igényelnek előtanítást. Az egyik ilyen javaslat nem a tanítóalgoritmust módosítja, hanem a hálót felépítő neuronokat. Az ún. rectified („egyenirányított”) neuronok nevüket onnan kapták, hogy a szokásos szigmoid aktivációs függvény le van cserélve egy olyan komponensre, amelynek működése egy egyenirányító áramkörre hasonlít (matematikailag a $max(0, x)$ függvényt valósítja meg). A rectifier neuronokra épülő mély neuronhálók használatát eredetileg képfeldolgozásban javasolták, csoportunk az elsők között próbálta ki őket beszédfelismerésben [4]. Eredményeink egybevágnak a más kutatók által velünk párhuzamosan publikált eredményekkel [5,6]: úgy tűnik, hogy az egyenirányított mély neuronhálók előtanítás nélkül is hasonló vagy kicsit jobb felismerési pontosságot tudnak elérni, mint hagyományos társaik előtanítással.

A rectifier függvény két alapvető dologban tér el a szigmoid függvénytől: az első, hogy az aktivációs érték növekedésével a neuronok nem „telítődnek”, ennek köszönhetően nem jelentkezik az eltűnő gradiens effektus. A rectifier neuronok esetén emiatt egy másik probléma jelentkezhet, mégpedig hogy a gradiens értékek „felrobbannak” (ún. „exploding gradient” effektus), azaz egyre nagyobb értékeket vesznek fel [8]. A probléma kiküszöbölése céljából a neuronok súlyait a tanítás során időről időre normalizálni szokták, mi a kettes norma szerint normalizáltunk. A másik fontos különbség, hogy negatív aktivációs értékekre 0 lesz a neuronok kimenete, aminek következtében a rejtett rétegeken belül csak a neuronoknak egy része lesz aktív adott input esetén. Ez utóbbi tulajdonságról azt is gondolhatnánk, hogy megnehezíti a tanulást, hiszen megakadályozza a gradiens visszaterjesztését, azonban a kísérleti eredmények ezt nem támasztják alá. A kísérletek azt igazolták, hogy az inaktív neuronok nem okoznak problémát mindaddig, amíg a gradiens valamilyen úton visszaterjeszthető.

Összefoglalva: a rectifier hálók nagy előnye, hogy nem igényelnek előtanítást, és a hagyományos backpropagation algoritmussal gyorsan taníthatók.

2.4. Dropout módszer

Az ún. dropout („kiejtéses”) tanulás lényege, hogy a neuronháló tanítása során minden egyes tanítópélda bevitelekor véletlenszerűen kinullázzuk („kiejtjük”) a hálót alkotó neuronok kimenetének valahány (általában 10-50) százalékát [7]. Ennek az a hatása, hogy az azonos rétegbe eső neuronok kevésbé tudnak egymásra hagyatkozni, így a probléma önálló megoldására vannak kényszerítve. Ennek köszönhetően lényegesen csökken a túltanulás veszélye. A módszert eredetileg javasló cikkben kiugró, 10-20 százaléknyi relatív hibacsökkenéseket értek el képi alakfelismerési és beszédfelismerési feladatokon.

A dropout technika előnye, hogy roppant egyszerűen implementálható, és elvileg minden esetben kombinálható a backpropagation algoritmussal. Az eredeti cikkben előtanított szigmoid hálók finomhangolása során alkalmazták, de azóta többen megmutatták, hogy rectifier neuronhálók tanításával kombinálva is remekül működnek [5]. További javulás érhető el az eredményekben, ha a tanítás során minden inputvektort többször (2-3-szor) is felhasználunk egy iteráción belül, különböző neuronkieséssel. Ugyan ez némileg javít az eredményeken, de az algoritmus futásidejét sokszorosára növeli, ezért mi csak egyszer használtunk fel minden inputvektort egy tanítási iterációban.

3. Kísérleti eredmények

A továbbiakban kísérleti úton vizsgáljuk meg, hogy a mély neuronhálók különböző tanítási módszerekkel milyen pontosságú beszédfelismerést tesznek lehetővé. Az akusztikus modellek készítése az ún. hibrid HMM/ANN sémát követi [9], azaz a neuronhálók feladata az akusztikus vektorok alapján megbecsülni a rejtett Markov-modell állapotainak valószínűségét, majd ezek alapján a teljes

megfigyeléssorozathoz a rejtett Markov-modell a megszokott módon rendel valószínűségeket. Mivel a neuronhálóknak állapotvalószínűségeket kell visszaadniuk, ezért minden esetben első lépésben egy rejtett Markov-modellt tanítottunk be a HTK programcsomag használatával [10], majd ezt kényszerített illesztés üzemmódban futtatva kaptunk állapotcímkéket minden egyes spektrális vektorhoz. Ezeket a címkéket kellett a neuronhálónak megtanulnia, amihez inputként az aktuális akusztikus megfigyelést, plusz annak 7-7 szomszédját kapta meg.

A modellek kiértékelését háromféle adatbázison végeztük el. Mindhárom esetben azonos volt az előfeldolgozás: e célra a jól bevált mel-kepsztrális együtthatókat (MFCC) használtuk, egész pontosan 13 együtthatót (a nulladikat is beleértve) és az első-második deriváltjaikat. A híradós adatbázis esetében szószintű nyelvi modellt is használtunk, a többi adatbázis esetén pusztán egy beszédhang bigram támogatta a beszédhang szintű felismerést.

Minden módszer esetében 128-as batch-eken tanítottunk, a momentumot 0.9-re állítottuk és backpropagation algoritmus esetén a korai leállást használtuk, a betanított mély hálók minden rejtett rétege 1024 neuronból állt.

A DBN előtanítás esetén a paraméterezés annyiban változott a tavalyi cikkünkben közölthöz képest, hogy lényegesen kevesebb epochon keresztül futtattuk a kontrasztív divergencia algoritmust az egyes RBM-ekre: 5 epoch a GRBM esetén és 3 a többi esetén a tavalyi 50-20 helyett. Tapasztalataink szerint ez volt az az iterációszám, amely során a rekonstrukciós hiba lényegesen csökkent, az ezt követő epochokban a súlyok is már csak minimálisan változtak. Az epochszám jelentős csökkentésével a tanításhoz szükséges idő is számottevően csökkent.

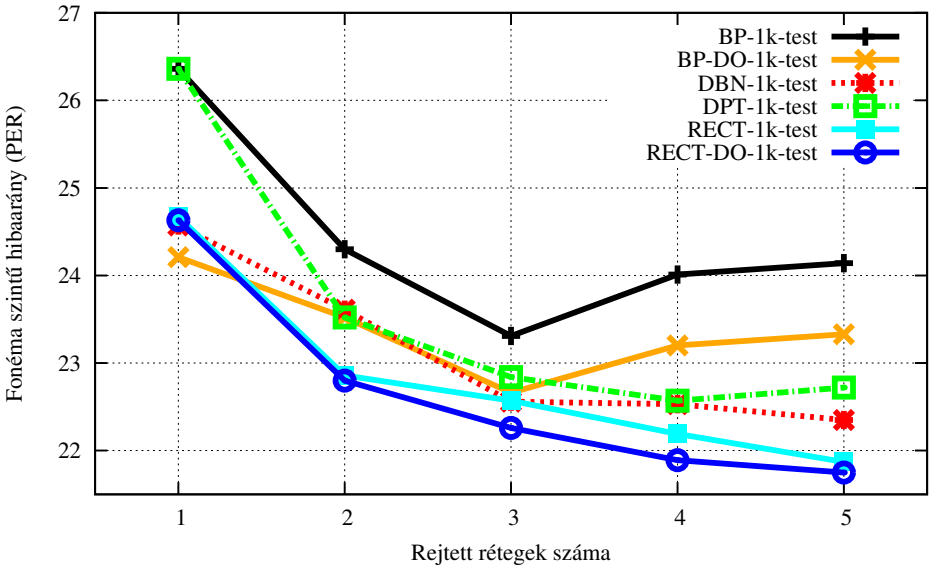
A diszkriminatív előtanítás esetén minden új rejtett réteg hozzáadása után 4 iteráción keresztül előtanítottunk 0.01-es fix tanulási rátával.

A rectifier neuronhálók estében a tanulási ráta 0.001 volt, illetve minden iteráció végén a súlyokat normalizáltuk, hogy az egy neuronhoz tartozó súlyok 2-es normája 1 legyen.

A dropout módszer esetén szigmoid hálókra a 10%-os neuronkiesési valószínűséget találtuk a legjobbnak, rectifier hálók estén pedig a 20%-ot. A tanítási iterációk végén a [7]-ben javasolt módon a súlyokat csökkentjük 10, illetve 20%-kal (a neuronkiesési valószínűséggel), hogy kompenzáljuk az a tényt, hogy tesztelés során a neuronok nem „esnek ki” véletlenszerűen.

3.1. TIMIT

A TIMIT adatbázis a legismertebb angol nyelvű beszédadatbázis [11]. Habár mai szemmel nézve már egyértelműen kicsinek számít, a nagy előnye, hogy rengeteg eredményt közöltek rajta, továbbá a mérete miatt viszonylag gyorsan lehet kísérletezni vele, ezért továbbra is népszerű, főleg ha újszerű modellek első kiértékeléséről van szó. Esetünkben azért esett rá a választás, mert több mély neuronhálós módszer eredményeit is a TIMIT-en közölték, így kézenfekvőnek tűnt a használata az implementációnk helyességének igazolására. A TIMIT adatbázis felosztására és címkézésére a tavalyi cikkünkben [1] ismertetett (és amúgy sztenderdnek számító) módszert használtuk. A továbbiakban csak monofón eredményeket közlünk.



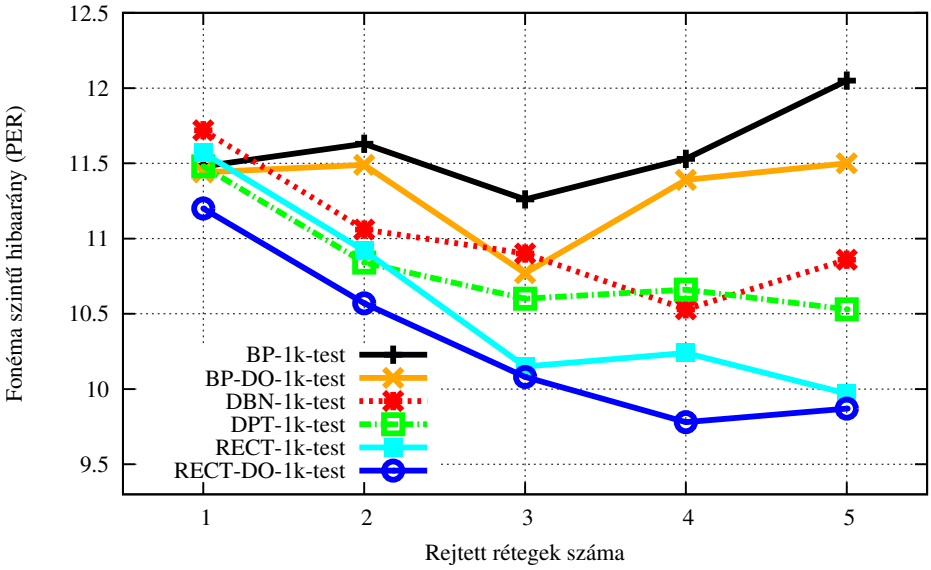
1. ábra. A különböző módszerek eredményei[4]a TIMIT core teszt halmazon a rejtett rétegek számának függvényében

A TIMIT adatbázison elért beszédhang szintű eredményeket láthatjuk a 1. ábrán. Jól látható, hogy a hagyományos backpropagation tanítóalgoritmusnál mindegyik ismertetett módszer jobban teljesített, ezen felül az is megfigyelhető, hogy a két előtanításos módszer nagyjából azonos eredményeket ért el. A legjobb eredményeket a rectifier hálókkal tudtuk kihozni: 21.75%, ami nagyjából 3%-os relatív javulás az előtanításos módszerekhez képest, illetve 7%-os relatív javulás a legjobb hagyományos (előtanítás nélküli) módszerhez képest. A korábbi cikkünkben közölt legjobb monofón eredményünkhöz (22.8%) képest a legjobb módszerrel több mint 1%-os javulást sikerült elérnünk, azonos módszerrel pedig 22.35%-ot, ami igazolja, hogy célszerű kevesebb előtanítást alkalmazni.

Az 1. ábrán megfigyelhető a dropout módszer hatékonysága is: míg szigmoid hálók esetén átlagosan 1%-os javulást hozott, ami 4%-os relatív javulásnak felel meg, addig a rectifier hálók esetén lényegesen kisebb a javulás. Ez utóbbinak az oka abban keresendő, hogy megfigyeléseink szerint a rectifier hálók neuronjainak átlagosan 70%-a inaktív tanítás során, ezt a dropout módszerrel kb. 75%-ra tudtuk növelni, ami nem hozott jelentős javulást az eredményekben.

Megvizsgáltuk továbbá, hogy a legjobban teljesítő mély neuronhálónkkal megegyező paraméterszámú hagyományos, egy rejtett réteges háló milyen eredményeket tud elérni. Az így kapott 23.5% lényegesen rosszabb mint a mély struk-

[4] Jelmagyarázat: **BP**: backpropagation, **BP-DO**: backpropagation+dropout, **DBN**: DBN előtanítás, **DPT**: diszkriminatív előtanítás, **RECT**: rectifier háló, **RECT-DO**: rectifier háló+dropout



2. ábra. A különböző módszerek eredményei a hangoskönyv adatbázis teszthalmazán a rejtett rétegek számának függvényében

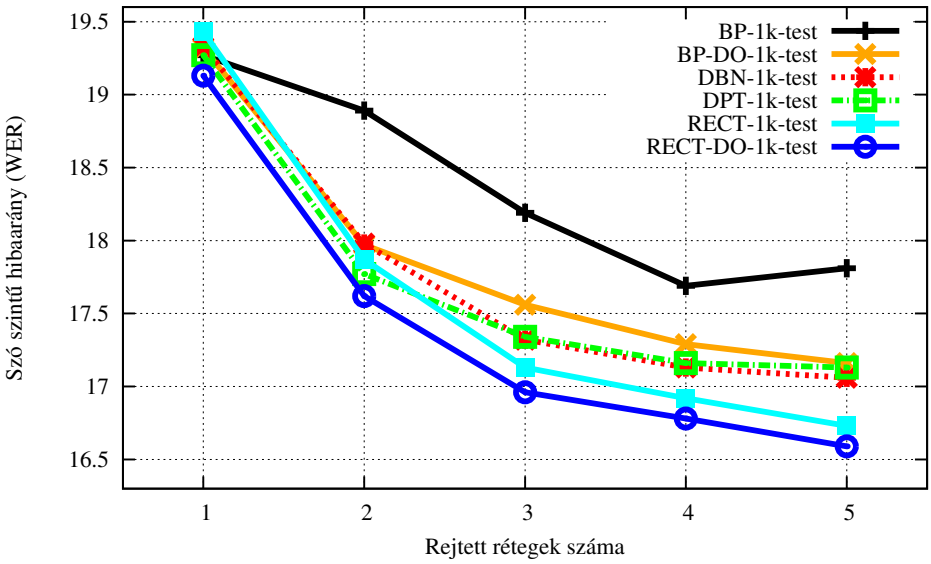
túrával elérhető eredmények, ami igazolja, hogy célszerű azonos paraméterszám esetén a mély struktúrájú hálót választani.

3.2. Hangoskönyv

A hangoskönyv adatbázisunk megegyezik a tavaly használttal. A 2. ábrán a különböző rétegszámmal elért eredményeket láthatjuk.

Megfigyelhető, hogy a különböző tanítási módszerek eredményei már jobban eltérnek, mint a korábbi adatbázison, viszont továbbra is megállapíthatjuk, hogy ha 2 vagy annál több rejtett réteget használunk, akkor a hagyományos módszer mindig a legrosszabb. A TIMIT-en elért eredményekhez hasonlóan itt is a rectifier hálók teljesítettek a legjobban, a legjobb eredményt (9.78%-ot) 4 rejtett réteggel dropout módszerrel tanítva értük el, ez több mint 13%-os hibacsökkenést jelent.

Az adatbázis sajátosságai miatt az figyelhető meg, hogy a hagyományosan tanított hálók esetén a rétegszám növelésével nem tudunk jelentős javulást elérni. A dropout módszer szigmoid hálók esetén 3 rejtett réteggel teljesített a legjobban – nagyjából 0.5%-kal jobb eredményt ért el –, rectifier háló esetén pedig 4 rejtett réteg esetén javított jelentősebben. A hagyományosan (azaz csak backpropagation algoritmussal) tanított, illetve a backpropagation+dropout módszerrel tanított szigmoid hálók kivételével mindegyik módszer esetén 4 vagy 5 rejtett réteggel értük el a legjobb eredményt. A tavalyi legjobb monofón eredményhez (10.62%) képest idén jelentős javulást tudtunk elérni (9.78%).



3. ábra. A különböző módszerek eredményei a híradós adatbázis teszthalmazán a rejtett rétegek számának függvényében

3.3. Híradós adatbázis

A híradós adatbázis, amely méretét tavaly óta sikerült jelentősen megnövelnünk, nagyjából 28 órányi hanganyagot tartalmaz. Az adatbázis felosztása: 22 órányi anyag a betanítási rész, 2 órányi a fejlesztési halmaz és a maradék 4 órányi hanganyag pedig a tesztelő blokk.

A híradós adatbázison szószintű felismerést tudtunk végezni, az ehhez szükséges nyelvi modellt az origo (www.origo.hu) hírportál szövegei alapján készítettük. Az így előálló korpusz nagyjából 50 millió szavas, mivel a magyar nyelv agglutináló (toldalékoló) nyelv. A korpusz lecsökkentése érdekében csak azokat a szavakat használtuk, amelyek legalább kétszer előfordultak a híranyagban, így 486982 szó maradt. A szavak kiejtését a Magyar Kiejtési Szótárból [12] vettük. A trigram nyelvi modellünket a HTK [10] nyelvi modellező eszközei segítségével hoztuk létre.

Ezen adatbázis esetén környezetfüggő (trifón) modelleket használtunk, ennek eredményeképp az adatbázis mérete miatt 2348 állapot adódott, azaz ennyi osztályon tanítottuk a neuronhálókat.

A 3. ábrán láthatóak az elért szószintű eredmények különböző rejtett rétegszám mellett. Ezen adatbázis esetén is elmondható, hogy a hagyományos módszer adja a legrosszabb eredményt, továbbá az is megfigyelhető, hogy a tanító adatbázis megnövekedése miatt a különböző tanítási módszerek eredményei jóval kevésbé térnek el. Továbbra is a rectifier hálók adják a legjobb eredményt (16.6%), ez a hagyományos módszerrel elérhető legjobb eredményhez (17.7%) képest 6%-os relatív hibacsökkenés.

1. táblázat. Az 5 rejtett réteges háló különböző módszerekkel történő tanításához szükséges idők

Módszer	Előtanítási idő	Finomhangolási idő
Hagyományos	0 óra	4.5 óra
Dropout	0 óra	5.5 óra
DBN előtanítás	1 óra	4 óra
Diszkriminatív előtanítás	2.5 óra	3 óra
Rectifier háló	0 óra	4 óra
Rectifier háló + Dropout	0 óra	4.5 óra

A híradós adatbázishoz közölt korábbi legjobb eredményünkhöz (16.9%) [13] képest is sikerült javítanunk, pedig a rejtett rétegek neuronszáma 2048-ról 1024-re csökkent.

Végül megvizsgáltuk az egyes módszerek időigényét: a 1. táblázatban az 5 rejtett réteges mély hálók különböző módszerekkel történő betanításához szükséges időket láthatjuk a híradós adatbázisra, GeForce GTX 560 Ti grafikus kártyát használva. Megállapítható, hogy a rectifier hálók nem csak jobb eredményeket adnak, de a betanításukhoz is kevesebb idő szükséges, mint a többi módszer esetén.

4. Konklúzió

Cikkünkben bemutattuk a mély neuronhálókra épülő akusztikus modelleket, illetve a betanításukhoz legújabban javasolt algoritmusokat. A kísérleti eredmények egyértelműen igazolják, hogy az új algoritmusok jobb eredményeket tudnak adni, miközben egyszerűbbek és/vagy kisebb időigényűek, mint az eredeti DBN előtanításra alapuló megoldás. Az eredményeket és a tanítási időket figyelembe véve megállapíthatjuk, hogy a legjobb módszer – az itt ismertetettek közül – a rectifier hálók dropout módszerrel történő tanítása.

Hivatkozások

1. Grósz T., Tóth, L.: Mély neuronhálók az akusztikus modellezésben. In: Proc. MSZNY. (2013) 3–12
2. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation **18**(7) (2006) 1527–1554
3. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proc. ASRU. (2011) 24–29
4. Tóth, L.: Phone recognition with deep sparse rectifier neural networks. In: Proc. ICASSP. (2013) 6985–6989
5. Dahl, G.E., Sainath, T.N., Hinton, G.: Improving deep neural networks for lvcsr using rectified linear units and dropout. In: Proc. ICASSP. (2013) 8609–8613

6. Zeiler, M., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., Hinton, G.: On rectified linear units for speech processing. In: Proc. ICASSP. (2013) 3517–3521
7. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. In: CoRR. Volume abs/1207.0580. (2012)
8. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proc. AISTATS. (2010) 249–256
9. Bourlard, H., Morgan, N.: Connectionist speech recognition: a hybrid approach. Kluwer Academic (1994)
10. Young, S., et al.: The HTK book. Cambridge Univ. Engineering Department (2005)
11. Lamel L., Kassel R., S.S.: Speech database development: Design and analysis of the acoustic-phonetic corpus. In: DARPA Speech Recognition Workshop. (1986) 121–124
12. Abari, K., Olaszy, G., Zainkó, C., Kiss, G.: Hungarian pronunciation dictionary on Internet (in Hungarian). In: Proc. MSZNY. (2006) 223–230
13. Tóth, L., Grósz, T.: A comparison of deep neural network training methods for large vocabulary speech recognition. In: TSD. (2013) 36–43