Univerzita Komenského
Fakulta Matematiky, Fyziky a Informatiky
Ústav informatiky

Miloš Černák

Učenie nesusedných závislostí pomocou rekurentných neurónových sietí

Diplomová práca

Diplomový školiteľ: Ing. Igor Farkaš, PhD.

Bratislava 2005

Týmto prehlasujem, že som diplomovú prácu vypracoval samostatne s odbornou pomocou školiteľa.

Bratislava, apríl 2005 Miloš Černák

Poďakovanie:

Moje veľké poďakovanie patrí môjmu školiteľovi Ing. Igorovi Farkašovi, PhD. za cenné rady a konzultácie.

Abstrakt

ČERNÁK,Miloš: *Učenie nesusedných závislostí pomocou rekurentných neurónových sietí*, Diplomová práca, Univerzita Komenského, Fakulta matematiky, fyziky a informatiky, Katedra informatiky, diplomový vedúci: Ing. Igor Farkaš PhD., Bratislava, 2005, 44 strán

Medzi základné problémy, ktoré ľudia musia pri učení sa jazyka zvládnuť, patrí segmentácia slov a generalizácia. Predpokladá sa, že na segmentáciu je potrebný štatistický mechanizmus, nie je však jednotný názor, či postačuje aj na generalizáciu, alebo je potrebný ešte ďalší, algebraický mechanizmus. Budeme skúmať úspešnosť riešenia spomenutých dvoch problémov modelmi rekurentných neurónových sietí (Elmanova rekurentná sieť, Jaegerova sieť s echo stavmi), pokúsime sa zistiť, či je na toto postačujúci štatistický mechanizmus. Porovnáme úspešnosť modelov sietí v našich experimentoch s empirickými výsledkami podobných experimentov, uskutočnených na ľuďoch. Budeme skúmať vplyv variabilnosti na úspešnosť a rýchlosť učenia nesusedných závislostí a zistíme, či sú získané výsledky v súlade s hypotézou variabilnosti.

Kľúčové slová: nesusedné závislosti, rekurentná neurónová sieť, Elmanova sieť, sieť s echo stavmi, hypotéza variabilnosti, segmentácia, generalizácia

Obsah

1 Úvod **7**

1.1 Motivácia 7

1.2 Obsahy jednotlivých kapitol 9

2 Implicitné učenie **11**

2.1 Susedné závislosti 11

2.2 Nesusedné závislosti 13

3 Umelé neurónové siete **15**

3.1 Formálny neurón 15

3.2 Dopredná neurónová sieť 16

3.3 Rekurentná neurónová sieť 17

3.3.1 Elmanova architektúra rekurentnej siete 18

3.3.2 Neurónová sieť s echo stavmi 21

4 Špecifikácia použitých modelov a dát **23**

4.1 Vstupy 24

4.2 Terminológia 25

5 Segmentácia slov **27**

5.1 Experiment 1: Echo state network 28

5.2 Experiment 2: Elmanova sieť 29

6 Generalizácia **31**

6.1 Experiment 3: Echo state network 31

6.2 Experiment 4: Elmanova sieť 32

6.3 Experiment 5: Elmanova sieť 33

7 Záver **39**

Kapitola 1

Úvod

1.1 Motivácia

Predmetom dlhodobého záujmu je spôsob a mechanizmy, pomocou ktorých si človek osvojuje materinský jazyk. Mnohé empirické štúdie dokazujú, že tento proces vyžaduje nielen vystavenie učiaceho sa patričnému jazykovému prostrediu, ale aj nejaké jeho vrodené dispozície pre spracovanie a učenie sa časovo usporiadaných vzoriek istým spôsobom (viď [1, 19, 23]). Presné fungovanie tohto procesu však nie je známe, ani to, čo zabezpečuje jeho rýchlosť a úspešnosť. Zdá sa, že osvojovanie si materinského jazyka u detí je inštinktívne a nevyžaduje žiadnu zvláštnu námahu, napriek tomu, že prebieha v prostredí bohatom na zmyslové vstupy. Ďalej môžeme predpokladať, že hlavným cieľom dieťaťa je porozumieť tomu, čo sa hovorí, teda samotné učenie sa jazyka je skôr iba akýmsi sprievodným javom.

V procese učenia získavame rôzne druhy znalostí, špecifické pre daný jazyk.

Ilustrované na slovenčine: vieme napríklad, že slovo *jesť* sa častejšie viaže so slovom *chlieb* ako so slovom *suseda*, alebo že po písmenách *st* niekedy

nasleduje *ĺ*, ale písmeno *x* nikdy.

Okrem tohto máme aj abstraktnejšie znalosti: prídavné meno *zelená* spájame s podstatným menom ženského rodu, väčšinu prídavných mien môžeme vystupňovať z 2.stupňa na tretí pomocou predpony *naj-* (*väčší* → *najväčší*)

Oba tieto druhy znalostí sú v istom zmysle štatisticky spoľahlivým odrazom skutočného stavu. Zdá sa však, že schopnosť nadobudnúť ich a reprezentovať sa opiera o rôzne mechanizmy. Prvý z nich sa viaže na vzťahy medzi konkrétnymi prvkami. Druhý možno nazvať *algebraický*, pretože sa sústreďuje na vzťahy medzi premennými, snaží sa vytvárať schémy, v ktorých možno nahradiť každú časť nejakým konkrétnym prvkom.

Existencia oboch mechanizmov bola testovaná pri viacerých príležitostiach. Saffran, Aslin a Newport [21] zistili, že už 8-mesačné deti prejavovali citlivosť na štatistické informácie obsiahnuté v hovorených vetách umelého jazyka. Deti boli v prvej fáze pokusu vystavené 2-minútovej nahrávke, nepretržitému reťazcu slabík (napr. *tibudopabikudaropigolatupabikutibudogolatudaropidaropitibudopabikugolatu*). Počas tejto prípravej fázy boli isté slabiky vždy nasledované jednoznačnou sekvenciou iných (napr. po *ti* vždy nasledovalo *budo*), na druhej strane existovali aj slabiky s viacerými možnosťami pokračovania (napr. po *pi* nasledovalo s pravepodobnosťou 1/3 *tibu*, *gola* alebo *daro*).

V druhej fáze pokusu testovali, ako dlho sa deti sústredili na blikajúce svetlá. Za týmito svetlami boli ukryté reproduktory, cez ktoré boli prehrávané testovacie slová (napr. *pabiku* a *pigola*). Zistili, že deti sa dlhšie sústredili pri prezentovaní slov typu *pabiku* ako typu *pigola*. Toto preukázalo, že deti počas prípravej fázy nadobudli informáciu o tom, ako často konkrétnu slabiku nasledovali iné.

Pokusy týkajúce sa druhého mechanizmu uskutočnili Marcus, Vijayan,

Bandi Rao a Vishton [12]. Ich výsledky ukazujú, že 7-mesačné deti dokážu zovšeobecňovať jednoduché gramatické pravidlá, ktoré sa naučia počas 2-minútovej prípravej fázy. Jeden z nich zahŕňa dve skupiny detí. Prvá si privykala na gramatiku typu ABA (vety boli napr. *ga ti ga* alebo *li na li*), druhá na gramatiku ABB (vety ako *ga ti ti*, či *li na na*). Po 2-minútách boli deti z oboch skupín testované vetami ako *wo fe fe* a *wo fe wo*, ktoré pozostávali z úplne nových slov. Pre obe skupiny detí bola polovica testovacích viet konzistentá s gramatikou, ktorej boli vystavené, a druhá polovica nie. Očakávalo sa, že deti budú schopné abstrahovať gramatiku viet, ktoré počuli v prípravnej fáze, a teda sa budú zaoberať dlhšie vetami, ktoré nie sú konzistentné s ich naučenou gramatikou (dieťa zo skupiny ABA pri počúvaní *wo fe fe*). Tieto očakávania boli správne, čo naznačuje, že schopnosť zovšeobecňovať vzniká aj bez výslovných inštrukcií, dlho pred tým, než sú deti schopné plynule narábať s jazykom. Podobné výsledky dosiahli Gomez a Gerken [4]. Ďalšie uskutočnené experimenty vylúčili možné špekulácie o vplyve povahy gramatiky na predchádzajúci výsledok. Deti rozlišovali medzi gramatikami AAB (*la la ti*) a ABB (*la ti ti*) - tu sa schopnosť rozlišovať nedá redukovať na detekciu za sebou sa opakujúcich prvkov, na rozdiel od [12].

V našej práci sa pokúsime preskúmať fungovanie predpokladaných mechanizmov učenia u ľudí na modeloch rekurentých neurónových sietí.

1.2 Obsahy jednotlivých kapitol

V **druhej kapitole** sa budeme venovať implicitnému učeniu u ľudí. Spomenieme bližšie niektoré experimenty, týkajúce sa schopnosti ľudí učiť sa susedné a nesusedné závislosti, a predpokladané mechanizmy umožnujúce toto učenie. V **tretej kapitole** popíšeme umelé neurónové siete, ich archi-

tektúru a spôsob učenia. V **štvrtej kapitole** uvedieme modifikácie použitých modelov oproti štandardným a spoločné parametre architektúr v našich experimentoch. Popíšeme sady trénovacích vstupov a kódovanie ich prvkov a zavedieme tiež niekoľko pojmov, ktoré budeme využívať pri popise experimentov. Obsahom **piatej kapitoly** budú dva experimenty, týkajúce sa segmentácie slov, ich výsledky porovnáme s empirickými výsledkami [22]. Ďalšie tri experimenty popíšeme v **šiestej kapitole**. Ich cieľom bude porovnať úspešnosť generalizácie u rekurentných neurónových sietí a u ľudí, referenčným pokusom bude [16]. Poslednú kapitolu tvorí záver našej práce.

Kapitola 2

Implicitné učenie

2.1 Susedné závislosti

Pri osvojovaní si jazyka nemajú dospelí výhodu nad deťmi, na rozdiel od väčšiny ostatných aspektov kognitívneho vývinu. Počiatočné fázy osvojovania si jazyka sú charakteristické nezávislosťou na veku. V neskorších fázach tí, ktorí začali ako deti, prekonávajú tých, ktorí začali ako dospelí (napr. [8], [10], [13], [24])

Experiment, ktorý uskutočnili Saffran, Newport, Aslin, Turnick a Barrueco [22], mal za cieľ zistiť, či aspoň počiatočné fázy osvojovania si jazyka spĺňajú nasledovnú charakteristiku: sú uskutočniteľné *implicitne*, sú nezávislé na veku a využívajú štatistické výpočty. Konkrétna činnosť, ktorou sa zaoberali, bola segmentácia slov. Skôr, než môžu deti začať získavať informácie o syntaxi, musia byť schopné rozpoznať slová ich jazyka. Tento problém je o to ťažší, že hovorené vety sú väčšinou súvislým prúdom zvukov, bez jednoznačných prestávok alebo iných zvukových značiek, ktoré by určovali hranice slov. Dospelí čelia podobnému problému, ak narazia na nové slovo včlenené do súvislého rečového prejavu. Na rozdiel od detí, ktoré ešte nemajú vybu-

dovanú slovnú zásobu, môžu použiť už známe slová ako oddeľovače, a takto určiť, kde začína a končí nové slovo. Napriek zjavnej obtiažnosti segmentácie slov dokazujú experimenty, že túto úlohu zvládajú deti 8-mesačné deti, teda oveľa skôr, ako u nich nastupuje schopnosť tvoriť slová [9].

Jedným riešením by mohlo byť, že deti sú schopné využiť štatistické náznaky hraníc slov. Zvuky, ktoré sa vyskytujú v jazykovej vzorke v rámci slova, navzájom korelujú viac, ako páry zvukov na rozhraní medzi slovami. Jednoduchý príklad, majme sekvenciu slov *veľkékreslo*. Za *veľ* nasleduje s nejakou pravdepodobnosťou *ké*. Za *ké* však nasleduje *kre* iba v prípade, že slovo na *kre* začínajúce nasleduje iné slovo, končiace na *ké*. Teda prechodové pravdepodobnosti z jedného zvuku na ďalší budú vo všeobecnosti najvyššie pri dvoch za sebou nasledujúcich slabikách v rámci slova.

Späť k spomínanému experimentu - pokusným osobám nebolo povedané, že nahrávka, ktorú počúvajú obsahuje nejaký jazyk, iba to, že v pozadí budú počuť nejakú zvukovú nahrávku, ktorá môže ovplyvniť ich umelecké schopnosti. Ich úlohou bola nakresliť nejaké obrázky pomocou počítačového programu. Prvý pokus bol na vzorke 12 dospelých a 13 detí. Počúvali 21-minútovú nahrávku pozostávajúcu z 300 výskytov každého zo 6-tich trojslabičných nezmyselných slov (napr. *bupada*, *patubi*) nahovorených v náhodnom poradí pomocou rečového syntetizátora. Tento prúd zvukov neobsahoval žiadne pauzy ani iné akustické náznaky hraníc slov. Jedinými náznakmi boli prechodové pravdepodobnosti, vyššie vnútri slov (0.31-1.0) ako na hranici medzi slovami (0.1-0.2). Ďalej bolo za účelom testovania vytvorených 6 novotvarov, zo slabík pôvodného jazyka použitých v poradí, v akom sa zatiaľ nevyskytli v nahrávke. Počas testovania bolo vytvorených 36 párov pôvodné slovo - novotvar, testovaná osoba mala pre každý pár rozhodnúť,ktorý zvuk počula počas kreslenia. Priemerné skóre u dospelých bolo 21.2 z 36 (58.6%),

u detí 21.3 (59.2%). Druhý pokus na vzorke 12 dospelých a 11 detí bol rozložený do dvoch dní. V prvý deň si vypočuli 21-minútovú nahrávku, na druhý deň si ju opäť vypočuli a boli následne testovaní rovnakým spôsobom, ako pri prvom pokuse. Priemerné skóre u dospelých bolo 26.3 (73.1%), u detí 24.6 (68.3%). Zdvojnásobenie dĺžky expozície malo veľký vplyv na úspešnosť. Medzi deťmi a dospelými nebol žiaden výrazný rozdiel. Taktiež sa nepotvrdil výrazný vzťah medzi vekom a dĺžkou expozície.

2.2 Nesusedné závislosti

Napriek tomu, že deti ako aj dospelí sú schopní sledovať prechodové pravdepodobnosti medzi susednými slabikami [21], pri nesusedných slabikách (teda aspoň v prípade, že spolu so vstupom nie sú prezentované nejaké pomocné náznaky, napr. pauzy oddeľujúce jednotlivé slová), to už nie je také jednoduché a dôkazy nie sú jednoznačné ([14],[17],[18]). Dôležitú úlohu pri tom, ako ľahko je závislosť medzi prvkami zistiteľná, hrá variabilita výplne, t. j. prvkov medzi závislými prvkami [3]. Učenie sa zlepšuje s rastúcou variabilitou výplne. Keď je množina prvkov zúčastňujúcich sa na závislosti malá relatívne k množine ostatných, tvoriacich výplň, vystupujú nesusedné závislosti ako nemenná štruktúra oproti premenlivejšej výplni na pozadí. Tento efekt pôsobí aj v prípade, že výplň, zdieľaná viacerými nesusednými prvkami, nemá žiadnu variabilitu. Je to možno preto, že výplň sa stáva nemennou vzhľadom na premenlivé závislosti [15]. Príkladmi v prirodzenom jazyku sú rôzne vzťahy (na väčšiu vzdialenosť), napr. zhoda čísla medzi podstatným menom a slovesom, oddelené tou istou výplňou (*knihy na poličke sú zaprášené* versus *kniha na poličke je zaprášená*). Súhrnne budeme nazývať účinky nulovej a veľkej variability nazývať *hypotéza variabilnosti (variability hypothesis).*

Všetky experimenty zlyhali v snahe ukázať generalizáciu na základe štatistických informácií, pokiaľ neboli poskytnuté na vstupe nejaké dodatočné náznaky, ako napr. pauzy medzi jednotlivými slovami. Na základe toho Peña a kol.[18] argumentovali, že generalizácia vyžaduje výpočtový mechanizmus založený na pravidlách, kdežto segmentácia slov spolieha na štatistické výpočty na nižšej úrovni. Onnis a kol. [16] vysvetľujú neúspechy experimentov pomocou hypotézy variabilnosti. Pri spomenutých experimentoch bola variabilita výplne malá, je to teda konzistentné s hypotézou variabilnosti, podľa ktorej je takéto učenie ťažké. Ich výsledky ukazujú, že dospelí sú schopní sledovať susedné aj nesusedné závislosti a miera úspešnosti je ovplyvnená variabilitou. Toto podporuje hypotézu, že na segmentáciu slov i generalizáciu postačuje jediný mechanizmus, ktorý je založený na štatistických informáciách.

Kapitola 3

Umelé neurónové siete

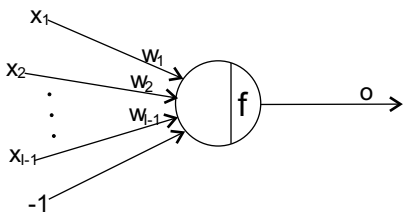
3.1 Formálny neurón

Formálny neurón (alebo **perceptrón**, viď obrázok 3.1), nech ma I vstupných synáps ($I \in \mathbb{N}$), prijíma vstupné signály

$$\mathbf{x} = (x_1, x_2, \ldots, x_I), \quad x_1, x_2, \ldots, x_I \in \mathbb{R}$$

cez synapsy váhované váhovým vektorom

$$\mathbf{w} = (w_1, w_2, \ldots, w_I), \quad w_1, w_2, \ldots, w_I \in \mathbb{R}$$



Obrázok 3.1: Perceptrón s aktivačnou funkciou f.

Nech hodnota I-teho vstupu je vždy -1, váhu w_I budeme označovať ako prah excitácie perceptrónu (tiež ozn. ako θ). Výstup perceptrónu dostaneme

pomocou vzťahu

$$o = f(net) = f(\mathbf{w} \cdot \mathbf{x}) = f(\sum_{i=1}^{I} w_i x_i) = f(\sum_{i=1}^{I-1} w_i x_i - \theta)$$

kde f je **aktivačná funkcia** (premennou net budeme označovať sumu váhovaných vstupov).

Hlavným dôvodom pre použitie aktivačnej funkcie je, že chceme zabezpečiť, aby bol výstup neurónu nejakým spôsobom rozumne ohraničený. Požadujeme tiež, aby bola aktivačná funkcia monotónna. Často používanou aktivačnou funkciou je sigmoida (aj vstupno-výstupná charakteristika u biologického neurónu má sigmoidálny tvar), a to buď unipolárna alebo bipolárna. Definovaná je nasledovne:

unipolárna: $f(net) = \dfrac{1}{1 + e^{(-\lambda net)}}$

bipolárna: $f(net) = \dfrac{2}{1 + e^{(-\lambda net)}} - 1$

Parameter λ je tzv. strmosť sigmoidy.

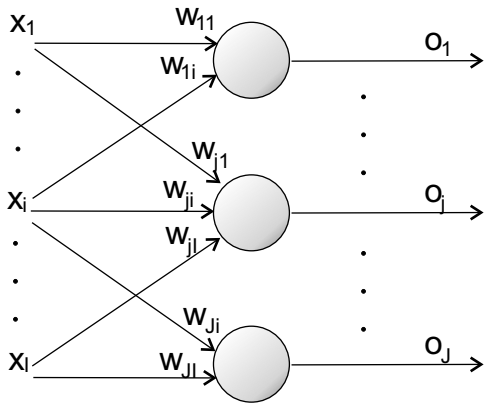
3.2 Dopredná neurónová sieť

Majme jednoduchú doprednú neurónovú sieť (viď obrázok 3.2) zloženú z $J \in \mathbb{N}$ neurónov napojených na tie isté vstupy $x_1, x_2, \ldots, x_I$.
Výstup j-teho neurónu je

$$o_j = f(net_j) = f(\sum_{i=1}^{I} w_{ji} x_i) \quad j = \overline{1, J}$$

kde $w_{j1}, w_{j2}, \ldots, w_{jI}$ sú vstupné váhy prislúchajúce j-temu neurónu. Výstupný vektor celej siete je potom

$$\mathbf{o} = (o_1, o_2, \ldots, o_J) = W_{J \times I} \cdot \mathbf{x}$$



Obrázok 3.2: Jednovrstvová dopredná neurónová sieť

kde W je matica váh. Zložením viacerých jednoduchých dopredných sietí (výstup nižšej vrstvy je vstupom pre vyššiu vrstvu) dostávame viacvrstvovú doprednú sieť. Najnižšou vrstvou je vstupná vrstva, nasleduje niekoľko skrytých vrstiev a najvyššou je výstupná vrstva. V praxi sa obyčajne používa 1 alebo 2 skryté vrstvy, neplatí totiž, že čím viac skrytých vrstiev, tým väčšia úspešnosť siete. Vektor výstupných aktivít neurónov skrytých vrstiev zvykneme označovať ako *stav siete*.

3.3 Rekurentná neurónová sieť

Pri spracúvaní postupností sa často stretávame s tým, že jednému vstupnému vektoru zodpovedá viacero možných výstupných vektorov, a to v závislosti od časového kontextu. Inak povedané, samotný výstup nie je závislý iba na vstupe, ale aj na nejakom počte predošlých vstupov (napríklad slová generované regulárnou gramatikou). Jednou z architektúr, ktoré je možné použiť pre takéto dáta, je neurónová sieť s oknom do minulosti (time delay neural network, TDNN)[11]. Je to isté rozšírenie viacvrstvovej doprednej siete, ok-

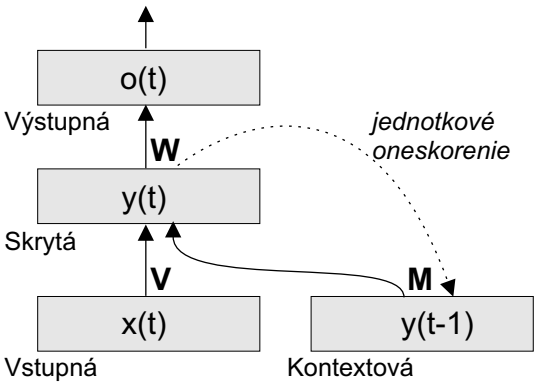
no do minulosti znamená, že sieť má okrem normálneho vstupu prístup aj k fixnému počtu predošlých vstupov. Je to architektonicky najjednoduchšie riešenie, jeho slabosťou je práve fixná dĺžka okna, takáto sieť nemusí byť schopná zachytiť časovú štruktúru v trénovacích dátach. Ak chceme byť schopní spracúvať postupnosti vyžadujúce okno do minulosti premenlivej dĺžky (odhadnúť správnu veľkosť okna len na základe trénovacej množiny nie je jednoduché ani v prípade, že TDNN je postačujúca), potrebujeme použiť inú architektúru. Rekurentná sieť je taktiež rozšírením doprednej siete, okrem vstupného vektora dostávajú neuróny na vstup aj stav siete v minulom kroku (t. j. výstupné aktivity všetkých neurónov). Toto je realizované pomocou rekurentných spojení (spojenia medzi výstupom neurónu a iným neurónom s jednotkovým oneskorením). Rekurentné spojenia majú váhu 1 a táto váha je nemenná.

3.3.1 Elmanova architektúra rekurentnej siete

Navrhnutá Elmanom v roku 1990 [2], býva označovaná aj ako Simple Recurrent Network (SRN), viď obrázok 3.3 . Kontextová vrstva má rovnaký počet neurónov, zo skrytej vrstvy do nej vedú rekurentné spojenia. Váha týchto spojení je nastavená na 1 a spojenia sú iba medzi príslúchajúcimi si neurónmi (i-ty na skrytej vrstve s i-tym na kontextovej). Kontextová vrstva takto obsahuje v každom kroku aktivácie skrytej vrstvy z predošlého kroku.

Na trénovanie RNS možno použiť rôzne metódy, napr. spätné šírenie chyby (error backpropagation)[20] alebo metódu známu ako rekurentné učenie v reálnom čase (real time recurrent learning, RTRL)[25]. Druhá zo spomenutých metód je výpočtovo náročnejšia, ale silnejšia, preto ju budeme používať.

Prejdime teraz k jej bližšiemu popisu. Uvažujme o sieti s I neurónmi vo vstupnej vrstve, J neurónmi v skrytej i kontextovej a K neurónmi vo



Obrázok 3.3: Elmanova sieť

výstupnej vrstve. Majme v čase t vstup siete $\mathbf{x}^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_I^{(t)})$, výstup siete $\mathbf{o}^{(t)} = (o_1^{(t)}, o_2^{(t)}, \ldots, o_K^{(t)})$ a očakávaný výstup vstup siete $\mathbf{d}^{(t)} = (d_1^{(t)}, d_2^{(t)}, \ldots, d_K^{(t)})$. Váhy spojení medzi vstupnou a skrytou vrstvou budeme označovať V, medzi kontextovou a skrytou ako M a váhy medzi skrytou a výstupnou vrstvou ako W.

Učením sa snažíme minimalizovať chybovú funkciu E určenú vzťahom

$$E^{(t)} = \frac{1}{2}\sum_{k=1}^{K}(d_k^{(t)} - o_k^{(t)})^2$$

Váhy synáps upravujeme v každom kroku, pomocou metódy negatívneho gradientu.

$$\Delta w_{kj}^{(t)} = -\alpha\frac{\partial E^{(t)}}{\partial w_{kj}} \qquad \Delta v_{ji}^{(t)} = -\alpha\frac{\partial E^{(t)}}{\partial v_{ji}} \qquad \Delta m_{jl}^{(t)} = -\alpha\frac{\partial E^{(t)}}{\partial w_{jl}} \tag{3.1}$$

Výstup siete $\mathbf{o}^{(t)}$ je daný

$$\begin{aligned} o_k^{(t)} &= f(net_k^{(t)}) = f(\sum_{j=1}^{J} w_{kj} y_j^{(t)}) \text{ kde} \\ y_j^{(t)} &= f(net_j^{(t)}) = f(\sum_{i=1}^{I} v_{ji} x_i^{(t)} + \sum_{l=1}^{J} m_{jl} y_l^{(t-1)}) \end{aligned}$$

Úpravou vzťahov 3.1 dostávame:

pre zmenu váh medzi skrytou a výstupnou vrstvou platí

$$\Delta w_{kj}^{(t)} = \alpha \delta_k^{(t)} y_j^{(t)} = \alpha (d_k^{(t)} - o_k^{(t)}) f_k^{'}(net_k^{(t)}) y_j^{(t)} \tag{3.2}$$

pre zmenu váh medzi vstupnou vrstvou a sktytou

$$\Delta v_{ji}^{(t)} = \alpha \sum_{k=1}^{K} \left[\delta_k^{(t)} \sum_{h=1}^{J} w_{kh} \frac{\partial y_h^{(t)}}{\partial v_{ji}} \right]$$

kde

$$\frac{\partial y_h^{(t)}}{\partial v_{ji}} = f^{'}(net_h^{(t)}) \left[x_i^{(t)} \delta_{jh}^{Kron.} + \sum_{l=1}^{J} m_{hl} \frac{\partial y_l^{(t-1)}}{\partial v_{ji}} \right]$$

a podobne medzi kontextovou a skrytou vrstvou

$$\Delta m_{ji}^{(t)} = \alpha \sum_{k=1}^{K} \left[\delta_k^{(t)} \sum_{h=1}^{J} w_{kh} \frac{\partial y_h^{(t)}}{\partial m_{ji}} \right]$$

kde

$$\frac{\partial y_h^{(t)}}{\partial m_{ji}} = f^{'}(net_h^{(t)}) \left[y_i^{(t-1)} \delta_{jh}^{Kron.} + \sum_{l=1}^{J} m_{hl} \frac{\partial y_l^{(t-1)}}{\partial m_{ji}} \right]$$

Parameter α je rýchlosť učenia a $\delta_{jh}^{Kron.} = \begin{cases} 1 & ak\ j = h \\ 0 & inak \end{cases}$ (Kroneckerova delta).

Pri našich experimentoch použijeme modifikovanú verziu SRN siete, zmenený bude spôsob výpočtu aktivácií neurónov na výstupnej vrstve. Namiesto štandardnej sigmoidy bude výstup $\mathbf{o}^{(t)} = o_1^{(t)}, o_2^{(t)}, \ldots, o_K^{(t)}$ počítaný pomocou **softmax**-u

$$o_k^{(t)} = \frac{e^{(net_k)}}{\sum_{l=1}^{K} e^{(net_l)}}$$

Pri softmax-e sa taktiež mení chybová funkcia, používa sa tzv. **vzájomná entropia** (cross entropy)

$$E^{(t)} = -\sum_{k=1}^{K} d_k^{(t)} ln(o_k^{(t)})$$

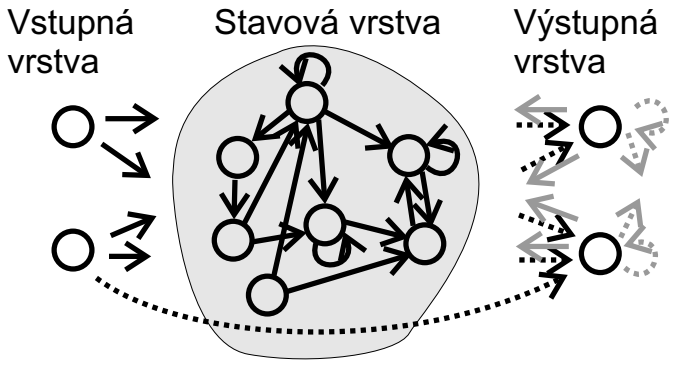
Vďaka tomu, že

$$\frac{\partial E^{(t)}}{\partial net_k^{(t)}} = o_k^{(t)} - d_k^{(t)}$$

sa praktický výpočet úprav váh zmení len málo, vo vzťahu (3.2) vypadne člen $f_k'(net_k^{(t)})$.

3.3.2 Neurónová sieť s echo stavmi

Vo svojich prácach[6][7] uviedol Herbert Jaeger novú architektúru nazvanú neurónová sieť s echo stavmi (Echo state network, ESN). Za istých podmienok je stav rekurentnej siete funkciou histórie vstupov, teda môže byť chápaný ako ozvena (echo) tejto histórie, odtiaľ pravdepodobne pochádza pomenovanie tejto architektúry. Základná architektúra siete je zobrazená na obrázku 3.4 . Najjednoduchšia verzia ESN je takmer zhodná s SRN. Oproti nej má



Obrázok 3.4: Neurónová sieť s echo stavmi. Čiarkovane sú zobrazené spojenia, ktorých váhy sa pri trénovaní modifikujú. Šípky čiernej farby predstavujú spojenia základnej verzie, sivé šípky predstavujú možné spojenia.

však navyše aj spojenia zo vstupnej na výstupnú vrstvu. Zložitejšie verzie môžu obsahovať tzv. *feedback* spojenia, teda oneskorené rekurentné spojenia z výstupnej vrstvy do stavovej, prípadne *laterálne*, teda oneskorené rekurent-

né spojenia medzi neurónmi výstupnej vrstvy. Podstatnou zmenou u ESN je, že pri trénovaní sa upravujú iba váhy výstupnej vrstvy. Toto značne zjednodušuje proces trénovania, vyžaduje si to však dodatočné obmedzenia danej siete. Jaeger preto zaviedol pojem *echo state vlastnosť*. Sieť má echo state vlastnosť, akk je výsledný stav rovnaký limitne vzhľadom na dĺžku vstupnej postupnosti bez ohľadu na počiatočný stav.

Uvedieme ešte postačujúcu podmienku pre *echo state vlastnosť* (podľa [7])

Veta 3.3.1. *Nech* $\mathbf{W}$ *je matica váh oneskorených rekurentných spojení zo stavovej vrstvy do stavovej a nech* $T(\mathbf{x}^{(t)}, \mathbf{u}) = \mathbf{x}^{(t+1)}$ *označuje aktualizáciu stavu siete zo stavu* $\mathbf{x}^{(t)}$ *spracovaním vstupu* u *na stav* $\mathbf{x}^{(t+1)}$ *a nech* Λ *je najväčšia hodnota v matici* $\mathbf{W}$*. Ak platí* $\Lambda < 1$*, potom* $d(T(\mathbf{x}, \mathbf{u}), T(\mathbf{x}', \mathbf{u})) < \Lambda \cdot d(\mathbf{x}, \mathbf{x}')$ *pre ľubovoľný vstup* $\mathbf{u}$ *a ľubovoľné stavy* $\mathbf{x}, \mathbf{x}' \in [-1, 1]^N$*, a teda sieť má echo state vlastnosť.*

Skontrolovanie platnosti tejto podmienky je nenáročné, budeme sa o ňu opierať pri nastavovaní parametrov ESN v našich experimentoch.

Vďaka jednoduchosti trénovania a možnosti učit sieť on-line nachádza ESN uplatnenie v robotike. V článku [5] je ESN využitá na hypotetizovanie faktov o situácii (napr. *je_v_mietnosti*), v ktorej sa robot momentálne nachádza. Berie pri tom do úvahy minulé vstupy a pomáha filtrovať významné senzorické vstupy od momentálne menej postatných. Bežne bývajú pravidlá pre platnosť nejakého faktu zadávané ručne, v rámci modelovania domény, v ktorej robot operuje. [5] demonštruje, že podobné pravidlá môžu byť aj naučené.

Kapitola 4

Špecifikácia použitých modelov a dát

SRN - počet neurónov na vstupnej a výstupnej vrstve bol pre S1 7, pre S2 10 a pre S3 35, počet neurónov skrytej vrstvy bol vždy 10. Ako aktivačnú funkciu sme používali unipolárnu sigmoidu (so strmosťou $\lambda = 2$) s výnimkou neurónov výstupnej vrstvy. Inicializácia váh - vstupné váhy boli náhodné z intervalu $\langle -0.5, 0.5 \rangle$, váhy spojení vedúce z kontextovej vrstvy a výstupné váhy boli náhodné z intervalu $\langle -1, 1 \rangle$. Rýchlosť učenia α bola 0.01.

ESN - počty neurónov v jednotlivých vrstvách, aktivačné funkcie a rýchlosť učenia boli rovnaké ako u SRN. Vstupné váhy boli -0.1 a 0.1, obe hodnoty s pravdepodobnosťou 0.5. Výstupné váhy boli náhodné z intervalu $\langle -1, 1 \rangle$. Váhy medzi kontextovou a výstupnou vrstvou nadobúdali hodnoty -0.47, 0.47 a 0 s pravdepodobnosťou 0.1, 0.1 a 0.8. Nastavenie týchto váh je pre tento model veľmi dôležité, keďže ich hodnoty sa počas trénovania už nemenia. Globálne škálovanie matice váh $1 \rightarrow 0.47$ zabezpečuje splnenie podmienky 3.3.1, a tým *echo state vlastnosť* siete. Veľká pravdepodobnosť, že spojenie bude mať váhu 0, zasa zlepšuje dynamiku siete (graf spojení, ak vynecháme

tie s nulovou váhou, bude riedky).

4.1 Vstupy

Budeme používať 3 sady vstupných údajov (S1, S2, S3), každá sada obsahuje 3-slabičné slová z umelého jazyka. Tieto slová sú tvaru AXB, kde A jednoznačne určuje B. Budeme používať 3 rámce A_B v každom jazyku a budú spoločné pre všetky 3 sady: *ba_ te, gu_ do* a *pi_ ra*. Jednotlivé sady vstupov sa od seba líšia počtom slabík, ktoré sú použité ako X.

$$
\begin{aligned}
X1 &= \{di\} \\
X2 &= \{di, ku, to, pa\} \\
X3 &= \{be, bi, bo, bu, by, ta, ti, to, tu, ty, ga, ge, gi, go, gy, \\
&\quad da, de, di, dy, pa, pe, po, pu, py, re, ri, ro, ru, ry\}
\end{aligned}
$$

|X1|=1, |X2|=4 a |X4|=29. Treba zdôrazniť, že slabiky, ktoré sú v úlohe X sa nevyskytujú v A ani v B, taktiež množiny A a B sú disjunktné.

Kódovanie vstupu

Vstupná postupnosť je zadávaná po slabikách, t. j. v každom takte je na vstupe zakódovaná jedna slabika. Slabiky sú *kódované lokalisticky* (tzv. one-hot kódovanie), to znamená, že pre každú slabiku má len jeden prvok vstupného vektora hodnotu 1, ostatné majú hodnotu 0. Tento spôsob kódovania určuje architektúru siete, presnejšie počet neurónov na vstupnej vrstve (taktiež na výstupnej vrstve, keďže aj výstup budeme kódovať *lokalistickým kódovaním*).

4.2 Terminológia

Zavedieme teraz niekoľko pojmov, ktoré budeme používať pri popise experimentov, budeme ich ilustrovať na vstupnej sade S1.

$$S1 = \{badite, gudido, pidira\}$$

Vstup siete pri trénovaní tvorí postupnosť slabík z $A \cup B \cup X$, ktorá vznikne zreťazením slov príslušnej vstupnej sady v náhodnom poradí. Príklad časti možnej vstupnej postupnosti: *gudidobaditepidiragudidopidirapidirabadite* Pod pojmom **part-word** (ozn. PW) budeme rozumieť slovo, ktoré vznikne spojením tretej slabiky jedného slova sady a prvej a druhej nasledujúceho slova, alebo spojením druhej a tretej slabiky jedného slova a prvej nasledujúceho.

gudido$\overline{$*badi*$\underline{$**te**$}$**pidi**$\underline{$*ra*$\overline{$*gudido*$\underline{$*pi*$}$**dira**$}$**pi**$\overline{$*dira*$\underline{$*badite*$}$$}$$}$$}$

(PW sú zvýraznené, nadčiarknuté a podčiarknuté sú pôvodné slová)

PW zastrešujú ostatné možnosti delenia vstupnej postupnosti slabík na slová pevnej, vopred známej dĺžky (v našom prípade dĺžky 3). PW budeme teda využívať pri testoch experimentov zaoberajúcich sa segmentáciou slov. Ich prostredníctvom môžeme zistiť, nakoľko testovaná sieť preferuje delenie vstupnej postupnosti zodpovedajúce slovám z príslušnej sady pred iným delením.

Rule-word (ozn. RW) bude označovať slovo, ktoré má tvar $A_iA_jB_i$ alebo $A_iB_jB_i$, kde $A_i\_B_i$ a $A_j\_B_j$ sú rámce vstupnej sady. Príkladmi RW pre S1 sú *bagute* alebo *pidora*.

RW predstavujú nové slová, zložené zo slabík použitých vo vstupnej postupnosti. Tieto slabiky sú v novom slove usporiadané v poradí, v akom sa v pôvodnej vstupnej postupnosti nevyskytli. Prechodové pravdepodobnosti

$P(A_j|A_i)$ a $P(B_i|A_j)$ sú nulové ($P(Y|X)$ označuje pravdepodobnosť, že slabika Y nasleduje bezprostredne za slabikou X), jediná väzba je medzi 1. a 3. slabikou (A_i a B_i). RW je slovo vytvorené podľa abstraktného pravidla, budeme pomocou neho testovať schopnosť siete zovšeobecňovať.

Kapitola 5

Segmentácia slov

Pri nasledujúcich experimentoch otestujeme schopnosť rekurentných sietí riešiť problém segmentácie slov. Cieľom je zrealizovať pokus, podobný [22] (ktorý bol uskutočnený na ľuďoch), pomocou rekurentných neurónových sietí.

Vyhodnocovanie experimentov [22] (príprava a priebeh experimentu sú podrobnejšie popísané v časti 2.1) prebiehalo tak, že pokusnej osobe bola prezentovaná dvojica slov, jedno z nich bolo „slovo“ (t. j. slovo zo vstupnej sady), to druhé bolo „neslovo“. „Neslovo“ je zložené z 3 slabík vstupnej postupnosti v poradí, v akom sa vo vstupnej postupnosti nevyskytovali (pre S1 (viď 4.1) sú to napr. *badodi* či *tedora*). Pokusná osoba mala následne určiť, ktoré zo slov sa podobá tomu, čo počula počas trénovania. Pri použití neurónovej siete narážame na jeden základný problém. Náš prístup musí zahŕňať nejaké kritérium, pomocou ktorého sieť rozhodne, ktoré zo slov sa jej viac „páči“. Architektúra našej siete neumožňuje zadať na vstup siete 2 slová naraz a nejak priamo získať výsledok, navyše musíme jednotlivé slová zadávať postupne po slabikách v priebehu viacerých taktov.

Spôsob vyhodnocovania bude teda nasledovný:

- vyberieme rovnaký konštantný počet (štandardne budeme používať

vzorku veľkosti $N = 20$) „slov“ a „neslov“

- pre každé slovo zistíme jeho ohodnotenie
- vyberieme N slov s najvyšším ohodnotením a zistíme, aká časť z nich patrí do prvej množiny („slová“), a koľko do druhej („neslová“)

Ohodnotenie slova získame tak, že:

- zresetujeme kontextovú vrstvu siete (t. j. nastavíme neuróny príslušnej vrstvy na 0), to spôsobí, že sieť bude izolovaná od predchádzajúcich vstupov
- postupne zadávame na vstup prvých $M - 1$ slabík (kde M je počet slabík slova) v $M-1$ taktoch a zaznamenáme hodnoty $u_1, u_2, \ldots, u_{M-1}$ (výstup neurónu zodpovedajúceho nasledujúcej slabike v každom takte)
- ohodnotenie slova u definujeme ako $u = \sum_{m=1}^{M-1} u_m$

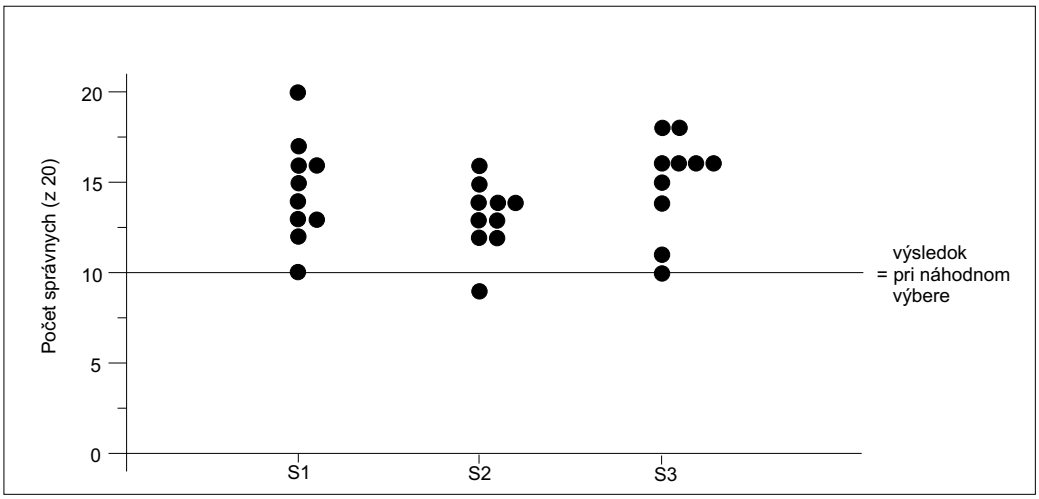
5.1 Experiment 1: Echo state network

Pre každú zo vstupných sád bolo natrénovaných 10 ESN sietí.

vstup	rozsah trénovania
S1	100×200
S2	100×100
S3	100×50

Rozsah trénovania $a \times b$ znamená, že trénovanie prebiehalo v b epochách, a v každej epoche pozostávala vstupná postupnosť z $a \cdot (\#slov\ v\ sade)$ slov (vyberaných náhodne zo „slov“. Rozsah trénovania bol určený experimentálne tak, aby po danom počte epoch bol počet chybne predikovaných slov sady

(bez dodatočného resetovania kontextu) nulový pre väčšinu trénovaných sietí (v niekoľkých prípadoch sieť toto kritérium nespĺňala, bola preto ignorovaná a namiesto nej bola natrénovaná nová sieť).



Obrázok 5.1: Výsledky experimentu 1 pre jednotlivé ESN siete

Pri testovaní úspešnosti sme nepoužívali „slová“ a „neslová“ ako v [22], ale „slová“ a *PW*. Nami zvolený test bol obtiažnejší, očakávateľná bola teda nižšia miera úspešnosti.

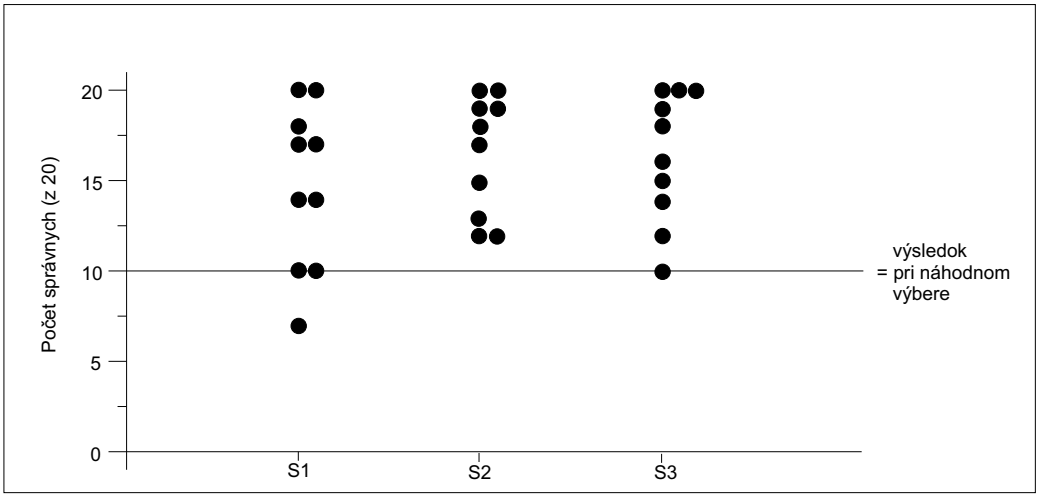
Výsledky nášho experimentu sú zobrazené na obrázku 5.1 . Priemerný výsledok pre S1 bol 14.6 z 20-tich (73%), pre S2 13.2 (65%) a pre S3 15.0 (75%). Pre porovnanie spomenieme výsledky z [22], dospelí dosiahli úspešnosť 26.3 z 36 (73.1%) a deti 24.6 (68.3%). Môžeme konštatovať, že problém segmentácie slov je ESN sieť schopná úspešne riešiť.

5.2 Experiment 2: Elmanova sieť

Pre každú zo vstupných sád bolo natrénovaných 10 SRN sietí.

vstup	rozsah trénovania
S1	20×50
S2	20×50
S3	20×50

Testovanie prebiehalo rovnako ako v predchádzajúcej časti (5.1)



Obrázok 5.2: Výsledky experimentu 2 pre jednotlivé SRN siete

Výsledky experimentu sú zobrazené na obrázku 5.2 . Priemerný výsledok pre S1 bol 14.7 z 20-tich (73.5%), pre S2 16.5 (82.5%) a pre S3 16.4 (82%). Ako vidíme, aj SRN sieť zvláda segmentovať slová, a to úspešnejšie ako ESN aj napriek menšiemu rozsahu trénovania.

Kapitola 6

Generalizácia

Pod problémom generalizácie (viď 2.2) v našom prípad budeme rozumieť rozpoznávanie *RW*, resp. ich preferovanie pred *PW*.

6.1 Experiment 3: Echo state network

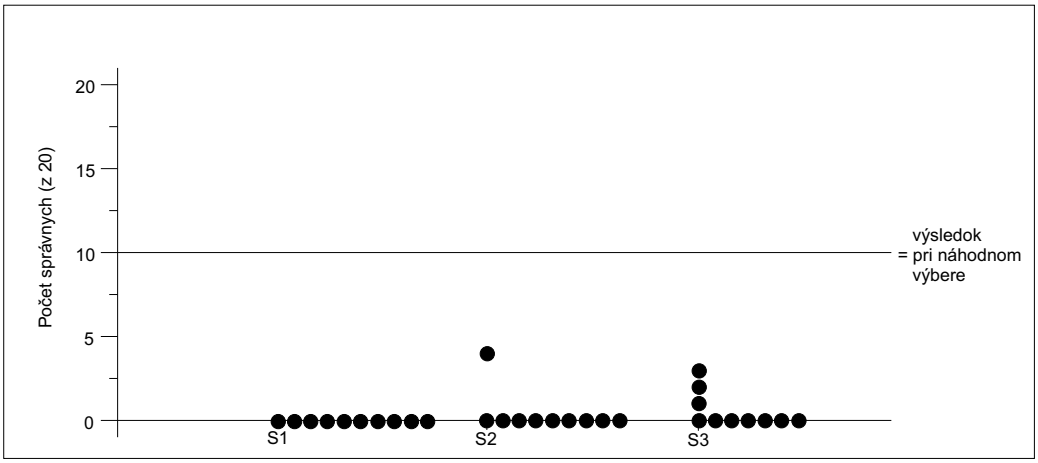
Pre každú zo vstupných sád bolo natrénovaných 10 ESN sietí.

vstup	rozsah trénovania
S1	100×200
S2	100×100
S3	100×50

Testovanie prebiehalo podobne ako v časti (5.1), porovnávali sme však ohodnotenia RW a PW.

Výsledky experimentu sú zobrazené na obrázku 6.1 . Priemerný výsledok pre S1 bol 0.0 z 20-tich (0%), pre S2 0.4 (0.02%) a pre S3 0.6 (0.03%).

Ako vidíme, úspešnosť pri tomto experimente je extrémne nízka napriek veľkému rozsahu trénovania. Sieť po naučení úspešne predpovedá nasledu-



Obrázok 6.1: Výsledky experimentu 3 pre jednotlivé siete

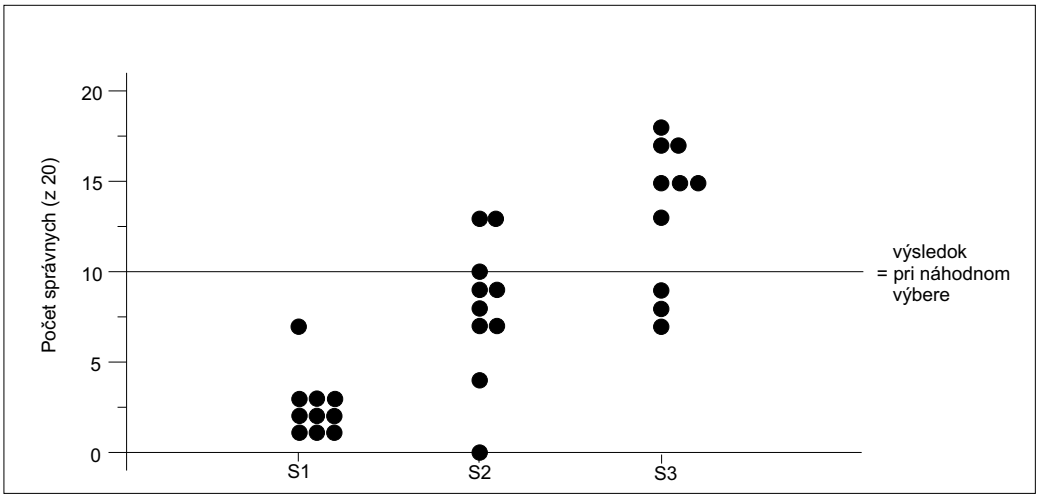
júcu slabiku v učiacej sekvencii, natrénovanie tohto modelu však nemá badateľný vplyv na tendenciu zovšeobecňovať. Sieť jednoznačne preferuje PW pred RW. Vysvetlenie tohto javu môže ležať v spôsobe učenia tejto siete. Na podchytenie a využitie pravidla (AXB) pri predikcii nestačí iba trénovanie váh výstupnej vrstvy. Detailnejšie sa preto týmto modelom zaoberať nebudeme.

6.2 Experiment 4: Elmanova sieť

Pre každú zo vstupných sád bolo natrénovaných 10 ESN sietí.

vstup	rozsah trénovania
S1	20×50
S2	20×50
S3	20×50

Testovanie prebiehalo rovnako ako v predchádzajúcej časti (6.1)



Obrázok 6.2: Výsledky experimentu 4 pre jednotlivé siete

Výsledky experimentu sú zobrazené na obrázku 6.2 . Priemerný výsledok pre S1 bol 2.5 z 20-tich (12.5%), pre S2 8.0 (40%) a pre S3 13.4 (67%). Nami získané výsledky sa vo veľkej miere zhodujú s výsledkami experimentu 2 v [16]. Spomínaný experiment sa týkal generalizácie, kde vstupná sekvencia bola súvislá (teda hranice slov neboli označené pauzami). Pokus robili pri nízkej (|X|=3) a vysokej(|X|=24) variabilite (v našom pokuse nízka (|X|=4) a vysoká (|X|=29)). Výsledné úspešnosti preferencie RW pred PW v ich experimente boli 41.9% pre nízku variabilitu (čo zodpovedá S2 a 40%) a 67% pre vysokú variabilitu (zodpovedá S3 a 67%). V ich experimente nie je uvedená úspešnosť pre nulovú variabilitu, takže túto hodnotu nemôžeme porovnať.

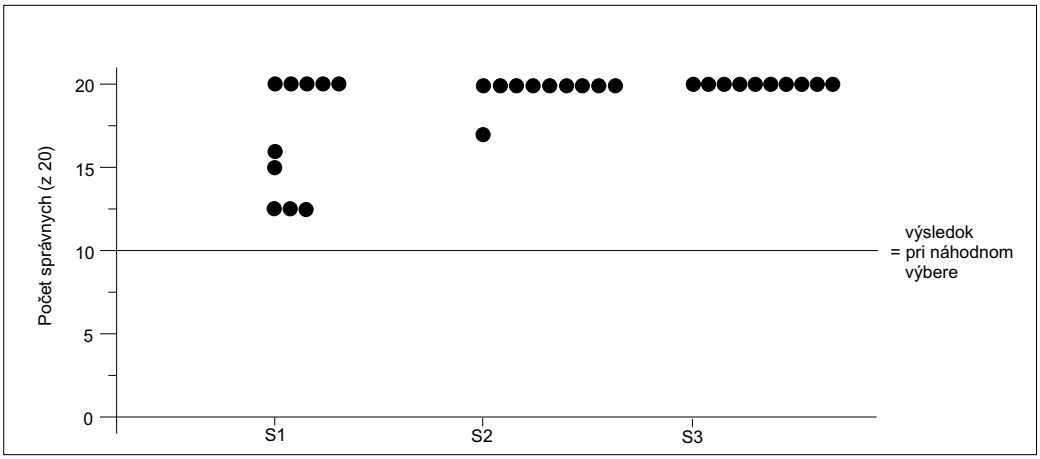
6.3 Experiment 5: Elmanova sieť

Pozrime sa teraz, ako ovplyvní úspešnosť učenia pridanie *páuz* do vstupnej postupnosti. Tieto *pauzy* budu realizované zresetovaním kontextovej vrstvy

v takte, v ktorom sieť dostáva na vstup zakódovanú prvú slabiku slova. Pre každú zo vstupných sád bolo natrénovaných 10 ESN sietí.

vstup	rozsah trénovania
S1	20×50
S2	20×50
S3	20×50

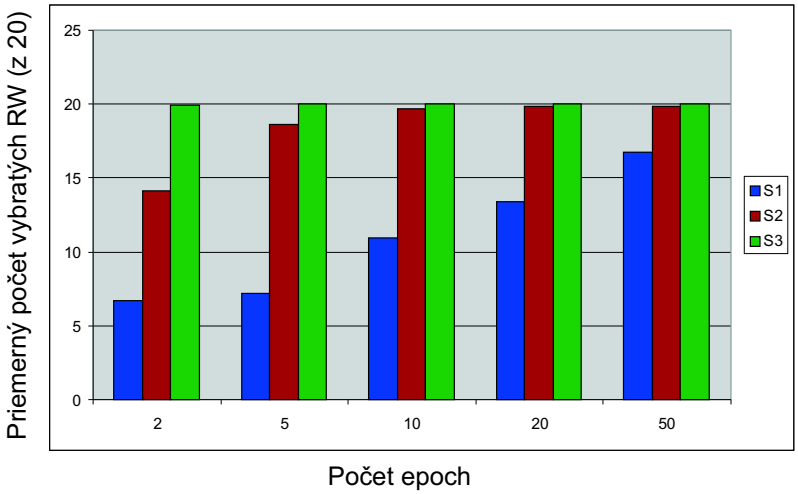
Testovanie prebiehalo rovnako ako v časti (6.1)



Obrázok 6.3: Výsledky experimentu 5 pre jednotlivé siete

Výsledky experimentu sú zobrazené na obrázku 6.3 . Priemerný výsledok pre S1 bol 16.7 z 20-tich (83.5%), pre S2 19.7 (98.5%) a pre S3 20.0 (100%). Pozorujeme, že úspešnosť preferovania RW pred PW sa pri nezmenenom rozsahu trénovania rapídne zvýšila. Keďže úspešnosť je veľmi veľká pre všetky sady (teda aj pre všetky testované hladiny variability), skúsime sa ešte pozrieť, ako sa vyvíjala úspešnosť počas trénovania, resp. zistiť, či má variabilita viditeľný efekt na rýchlosť učenia. Podľa hypotézy variabilnosti môžeme očakávať, že rýchlosť učenia bude pri nulovej a veľkej variabilite

väčšia ako pri malej variabilite.



Obrázok 6.4: Vývoj priemerných výsledkov jednotlivých skupín sietí v čase

Ako vidíme na obrázku 6.4 , rýchosť učenia skutočne závisí od variability, čím väčšia variabilita, tým väčšia je rýchlosť učenia. Toto tvrdenie obstojí aj v prípade, ak neporovnávame rýchlosť učenia na základe počtu epoch, ale počtu slabík, ktoré sieť počas učenia spracovala. Ďalej však môžeme skonštatovať neúspech v prípade S1. Podľa hypotézy variabilnosti sme očakávali, že učenie S1 (nulová variabilita) bude úspešnejšie ako S2 (malá variabilita). Tento negatívny výsledok žiaľ nevieme presne zdôvodniť.

Pozrime sa teraz na náš experiment z hľadiska prechodových postupností. U RW tvaru $A_iA_jB_i$ sú prechodové pravdepodobnosti $P(A_j|A_i)$ a $P(B_i|A_j)$ nulové, kedže v tomto poradí sa dané slabiky vo vstupnej postupnosti nikdy nevyskytli. Ohodnotenie slova pri testovaní bude teda závisieť takmer výnimočne od úspešnosti predikcie poslednej slabiky. U PW je ohodnotenie súčtom rovnomernejšie rozdelených čísel. Uvažujme o PW tvaru $B_jA_iX_i$, prechodová pravdepodobnosť $p(A_i|B_j)$ je 0.33 a je rovnaká pre všetky sady, $P(X_i|A_i)$ je však nepriamo úmerná veľkosti množiny X príslušnej vstupnej

sady. Je teda rozumné predpokladať, že ohodnotenie PW bude s rastúcou variabilitou klesať a sieť sa posunie viac k preferovaniu RW. Tento predpoklad môžeme skonfrontovať s experimentálnymi výsledkami. Prvky výstupného vektora dávajú v súčte 1, hodnotu výstupu neurónu môžeme chápať ako pravdepodobnosť, ktorú sieť pridelila javu, že nasledujúcou slabikou bude tá, ktorú spomínaný neurón reprezentuje (pri lokalistickom kódovaní).

Zaujímavý efekt môžeme pozorovať pri použití vstupnej sady S1 (tá má množinu X={*„di“*}). Aktivity výstupných neurónov naznačujú, že pri predikcii druhej slabiky slova sieť takmer ignoruje vstup a stabilne predikuje *„di“*. U RW je spomínaným vstupom nejaké A_i, u PW buď X_j alebo B_j. Toto správanie môžeme chápať ako prejav schopnosti siete generalizovať. Tým, že predikuje stále rovnakú druhú slabiku (ktorá je z X), dáva najavo, že nezáleží na tom, aká je druhá slabika (v kontexte RW a PW nie je druhá slabika nikdy z X). Ak sa pozrieme na výstupy pri použití S2 (|X|=4), zistíme, že natrénovaná sieť dáva pri predikcii druhej slabiky na 4 výstupných neurónoch hodnoty okolo 0.25 (boli to práve tie neuróny, ktoré reprezentovali prvky z X), ostatné mali zanedbateľné výstupy. Opäť treba zdôrazniť, prvá vstupná slabika tieto hodnoty takmer neovplyvnila. U S3 (|X|=29) sme dostali analogické výsledky, významných výstupov bolo pri predikcii druhej slabiky 29 a mali hodnoty v rozmedzí od 0.3 do 0.4.

V nasledujúcej tabuľke sú uvedené výstupy neurónov prislúchajúce uvedeným slabikám pri predikovaní 2. a 3. slabiky slova sieťou učenou na S1:

	A_1	B_1	A_2	B_2	A_3	B_3	X_1
didopi:							
2.sl.	0.01	0.01	0.01	0.02	0.01	0.01	0.99
3.sl.	0.38	0.10	0.14	0.06	0.25	0.10	0.02
ragudi:							
2.sl.	0.01	0.01	0.01	0.01	0.01	0.01	0.99
3.sl.	0.06	0.16	0.03	0.32	0.04	0.39	0.04

Príklady výstupov pre S3:

	A_1	B_1	A_2	B_2	A_3	B_3	X_1	...	X_{29}
gyragu:									
2.sl.	0.01	0.01	0.01	0.01	0.01	0.04	0.05	...	0.03
3.sl.	0.07	0.73	0.07	0.01	0.13	0.01	0.01	...	0.01
dopibe:									
2.sl.	0.02	0.01	0.01	0.01	0.02	0.01	0.03	...	0.03
3.sl.	0.01	0.10	0.01	0.03	0.01	0.67	0.01	...	0.01

Pozrime sa ešte na výstupy pri predikcii tretej slabiky. Všetky slová uvedené v tabuľkách sú PW, prvé v každej z nich je tvaru XBA, druhé tvaru BAX. Pre S1, ak testovacie slovo začína slabikami XB, sieť sa snaží udržať postupnosť ...AXBAXB..., najvyššie hodnoty výstupov majú neuróny zodpovedajúce slabikám z A. Ak testovacie slovo začína slabikami BA, sieť v niektorých prípadoch predikuje slabiku z X, v iných z B. Ak však zvýšime variabilitu X (príklady pre S3), pozorujeme, že tvar PW prestáva hrať úlohu pri predikovaní tretej slabiky. Tentokrát patrí najvyššia hodnota výstupného vektora vždy neurónu reprezentujúcemu slabiku z B. Taktiež sa výrazne zý-

šil odstup medzi najvyššou a druhou najvyššou hodnotou, tento rozdiel býva viac ako 0.5. Sieť teda pri vyššej variabilite nepreferuje iba množinu slabík (v tomto prípade B), ale vyberá z nej už aj konkrétny prvok.

Kapitola 7

Záver

V tejto práci sme reflektovali predpokladané mechanizmy učenia sa prirodzeného jazyka na rekurentných neurónových sieťach. Architektúry, ktoré sme použili pri testovaní, boli Elmanova SRN (*Simple recurrent network*) a Jaegerova ESN (*Echo state network*). V prvom experimente sme testovali schopnosť ESN riešiť problém segmentácie slov. Zistili sme, že dosahuje úspešnosť približne porovnateľnú s ľudskými subjektami v podobnom experimente. V druhom experimente sme skúmali ten istý problém, použili sme však SRN sieť. Dosiahnutá úspešnosť siete bola vyššia ako u ľudí. V experimetoch 3 až 5 sme sa zaoberali problémom generalizácie a vplyvom variability na dosahovanú úspešnosť. V experimente 3 sme použili nesegmentovanú vstupnú postupnosť a ESN sieť. Ukázalo sa že, sieť mala takmer nulovú úspešnosť, túto skutočnosť vysvetľujeme nedostatočnou silou trénovacieho postupu pre tento typ problému. Experiment 4 sledoval ten istý postup ako experiment 3, pri použití architektúry SRN. Dosiahnuté výsledky boli porovnateľné s výsledkami ľudských subjektov a boli v súlade s hypotézou variabilnosti. Výsledky experimentu 4 taktiež potvrdili, že na zvládnutie problému generalizácie nie je nutný osobitný mechanizmus, ale problém generalizácie a

segmentácie slov je riešiteľný iba za pomoci štatistických informácii extrahovateľných zo vstupu. V experimente 5 sme skúmali, aký vplyv má na úspešnosť generalizácie informácia o hraniciach slov vo vstupne postupnosti. Z dôvodu nízkej úspešnosti ESN v experimente 3 sme tento efekt skúmali iba na SRN sieti. Dodatočná informácia mala výrazný vplyv na rýchlosť učenia, pri zachovaní rozsahu trénovania z experimentu 4 sa úspešnosť pohybovala blízko 100%. Preto sme uskutočnili dodatočný experiment, kde hlavným parametrom bol rozsah učenia. Potvrdil sa pozitívny vplyv variability na úspešnosť. Negatívnu časť výsledkov exp. 5 (t. j. menšiu úspešnosť pri nulovej ako pri malej variabilite) sme nedokázali dostatočne zdôvodniť. Detailne sme analyzovali predikciu jednotlivých slabík slova a popísali vplyv variability na ňu.

Literatúra

[1] N Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, 1965.

[2] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[3] R. Gómez. Variability and detection of invariant structure. *Psychological Science*, 13:431–436, 2002.

[4] R.L. Gómez and L.-A. Gerken. Artificial grammar learning by 1 year-olds leads to specific and abstract knowledge. *Cognition*, 70(1):109–135, 1999.

[5] J. Hertzberg, H. Jaeger, and F. Schönherr. Learning to ground fact symbols in behavior-based robots. In *ECAI 2002. Proceedings of the 15th European Conference on Artificial Intelligence*, volume 77 of *Frontiers in Artificial Intelligence and Applications*, pages 708–712, Amsterdam, 2002. IOS Press. Lyon, France, July, 21–26.

[6] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Technical Report GMD Report 148, German National Research Center for Information Technology, 2001.

[7] H. Jaeger. Short term memory in echo state networks. Technical Report GMD Report 152, German National Research Center for Information Technology, 2001.

[8] J. Johnson and E. Newport. Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive Psychology*, 21:60–99, 1989.

[9] P.W. Jusczyk and R.N Aslin. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29:1–23, 1995.

[10] S.D. Krashen, M.H. Long, and R.C. Scarcella. Age, rate, and eventual attainment in second language acquisition. In S.D. Krashen, R.C. Scarcella, and M.H. Long, editors, *Child-adult differences in second language acquisition*, pages 161–172. Rowley,MA, Newbury House, 1982.

[11] K.J. Lang, A.H. Waibel, and G.E. Hinton. A time-delay neural network architecture for isolated word recognition. *Neural Netw.*, 3(1):23–43, 1990.

[12] G.F. Marcus, S. Vijayan, S. Bandi Rao, and P.M. Vishton. Rule learning in 7-month-old infants. *Science*, 283:77–80, 1999.

[13] E.L. Newport. Maturational constraints on language learning. *Cognitive Science*, 14:11–28, 1990.

[14] E.L. Newport and R.N. Aslin. Learning at a distance i. statistical learning of nonadjacent dependencies. *Cognitive Psychology*, 48:127–162, 2004.

[15] L. Onnis, M. Christiansen, N. Chater, and R. Gómez. Reduction of uncertainty in human sequential learning: Evidence from artificial gram-

mar learning. *Proceedings of the 25th Annual Conference of the Cognitive Science Society.*, pages 887–891, 2003.

[16] L. Onnis, P. Monaghan, M.H. Christiansen, and N. Chater. Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. *Proceedings of the 26th Annual Conference of the Cognitive Science Society.*, pages 1047–1052, 2004.

[17] L. Onnis, P. Monaghan, K. Richmond, and N. Chater. Phonology impacts segmentation in speech processing. *Journal of Memory and Language*, 2005.

[18] M. Pena, L. L. Bonatti, M. Nespor, and J. Mehler. Signal-driven computations in speech processing. *Science*, 298:604–607, 2002.

[19] S Pinker. *The language instinct*. Morrow, New York, 1994.

[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[21] J. Saffran, R. Aslin, and E. Newport. Statistical learning by 8-month old infants. *Science*, 274:1926–1928, 1996.

[22] J.R. Saffran, E.L. Newport, R.N. Aslin, R.A. Tunick, and S. Barrueco. Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, 8:101–105, 1997.

[23] M.S. Seidenberg. Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275:1599–1603, 1997.

[24] G.R. Slavoff and J.S. Johnson. The effects of age on the rate of learning a second language. *Studies in Second Language Acquisition*, 17:1–16, 1995.

[25] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1:270–280, 1989.