

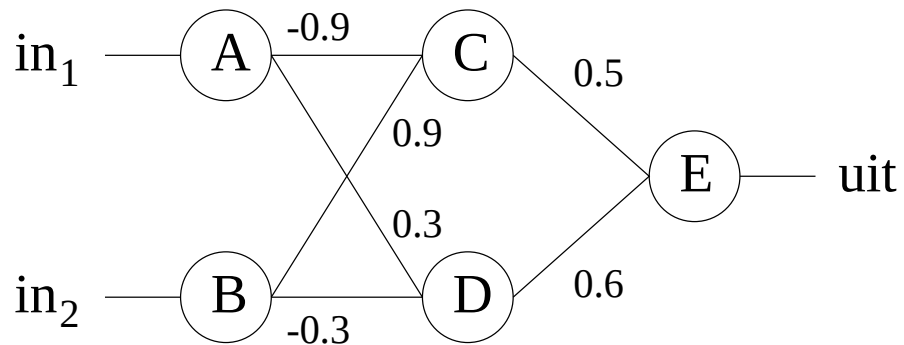
Neurale Netwerken voor Taaltechnologie

Erik F. Tjong Kim Sang
Institutionen för Lingvistik
Uppsala Universiteit
Zweden
erik.tjong@ling.uu.se

Neurale netwerken

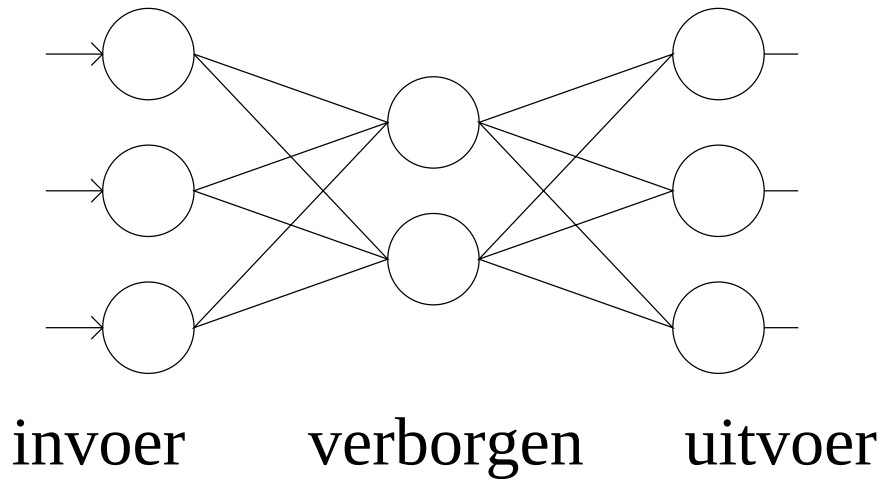
- Rekenmodellen bestaande uit cellen en verbindingen tussen de cellen.
- De cellen kunnen eenvoudige berekeningen uitvoeren en sturen communicatie-signalen over de verbindingen.
- Verbindingen kunnen signalen versterken of verzwakken.
- Neurale netwerken kunnen taken leren door hun verbindingsgewichten aan te passen.
- Zij zijn gebaseerd op een eenvoudig model van het menselijk brein.

Voorbeeld feedforward netwerk



in_1	in_2	uit_A	uit_B	uit_C	uit_D	uit
0	0	0	0	0	0	0
0	1	0	1	1	0	1
1	0	1	0	0	1	1
1	1	1	1	0	0	0

Leren in feedforward netwerken



Error backpropagation: de afwijking van de gemeten uitvoerwaarde van de uitvoercellen wordt gebruikt om de gewichten van de verbindingen aan te passen.

Formules

$$\Delta w_{c_1 c_2, l} = \eta * \delta_{c_2, l} * uit_{c_1, (l-1)}$$

$$\delta_{c, l} = (doel_{c, l} - uit_{c, l}) * (1 - uit_{c, l}) * uit_{c, l}$$

$$\Delta w_{c_1 c_2, l}(t) = \eta * \delta_{c_2, l} * uit_{c_1, (l-1)}(t) + \alpha * (\Delta w_{c_1 c_2, l}(t-1))$$

$$\delta_{c_1, l} = \left(\sum_{c_2} \delta_{c_2, (l+1)} * w_{c_1 c_2, (l+1)} \right) * (1 - uit_{c_1, l}) * uit_{c_1, l}$$

Neurale netwerken verwerken getallen.

Kunnen ze ook letters, woorden en zinnen verwerken?

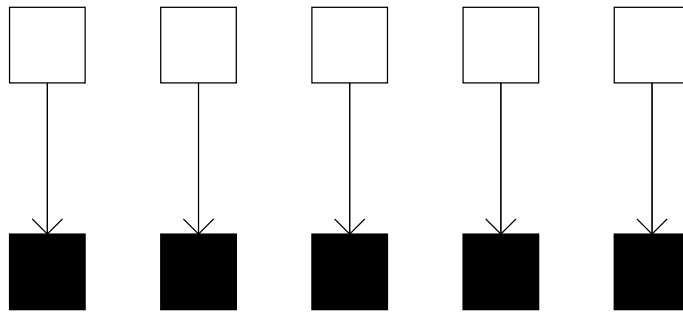
Lokale representatie

Representeer taalsymbolen als bitreeksen:

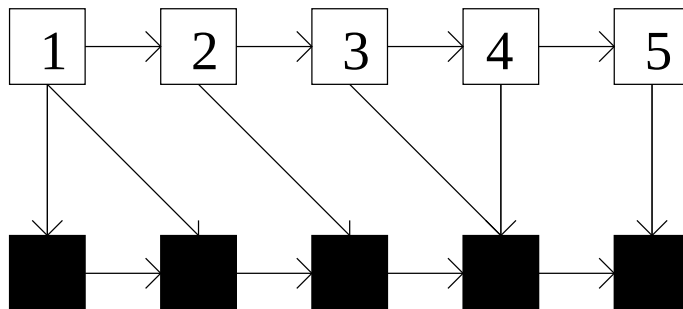
1	0	0	0	0		a
0	1	0	0	0		b
0	0	1	0	0		c
0	0	0	1	0		d
0	0	0	0	1		e

Ieder bit komt overeen met een netwerkkcel.

Feedforward netwerken zijn goed in het leren van verbanden tussen individuele symbolen:



Natuurlijke taal is een geordende reeks symbolen:

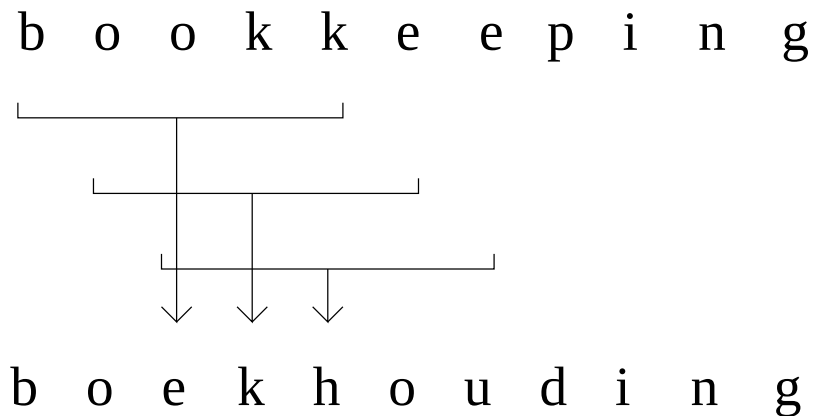


Voor standaard feedforward netwerken is het moeilijk om alle interne taalrelaties te representeren.

Mogelijke oplossing

Laat het feedforward netwerk symbolen verwerken in een context.

Voorbeeld

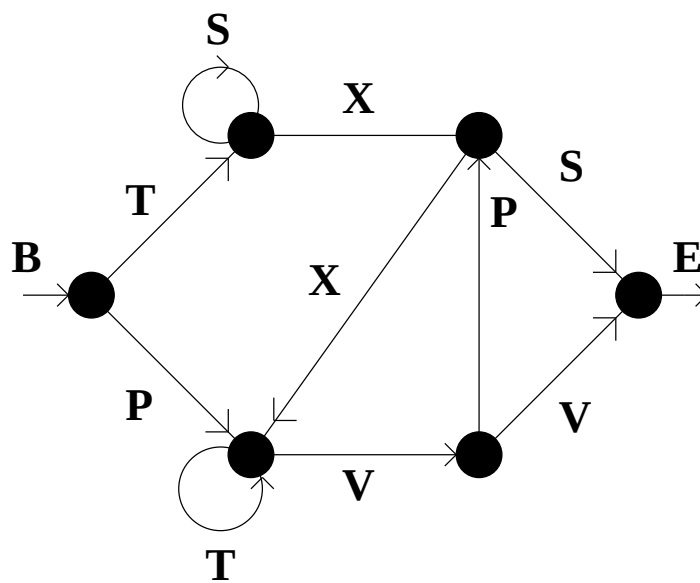


Maar wat is de beste contextgrootte?

Cleeremans 1993

Toepassing van een neuraal netwerk voor het bepalen van de grammaticaliteit van reeksen gebouwd met een formele grammatica.

Gebruikte grammatica: Rebergrammatica:

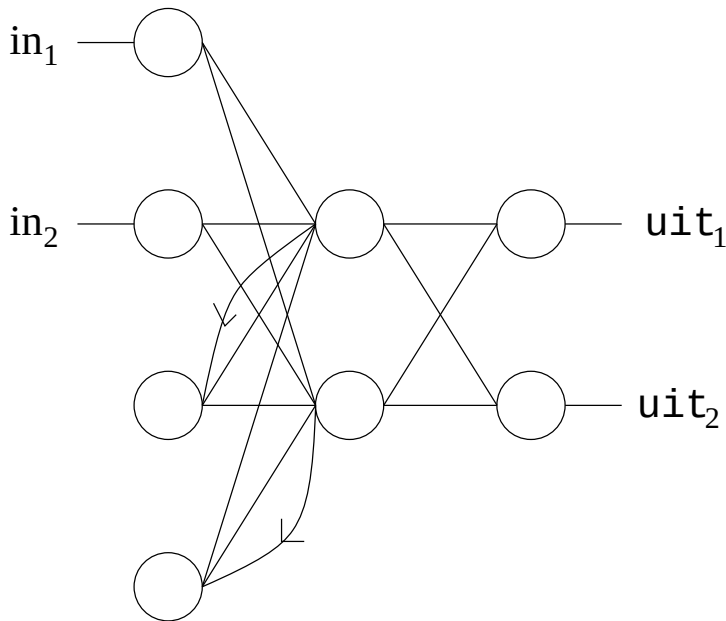


Resultaatpercentages

- Goedgekeurde correcte reeksen: 100%
- Afgekeurde incorrecte reeksen: 100%

Gebruikte neurale netwerk: SRN

Simple Recurrent Network (SRN)



Een feedforward netwerk met verbindingen van de verborgen cellen naar extra contextcellen.

Deze contextcellen vormen het geheugen van het netwerk.

Taakdefinitie

Het ontwikkelen van een neurale netwerk dat kan bepalen of een letterreeks een Nederlands woord is of niet.

Mogelijke toekomstige toepassingen

- Detectie van fouten in OCR-text.
- Detectie van spelfouten.

Extra beperkingen

1. Beperk data tot éénsyllabische woorden.
2. Gebruik alleen positieve leerdata.

Werkmethode

- Letters worden gerepresenteerd als een reeks van 29 bits (a-z, ', start_woord en einde_woord).
- Letters worden één voor één aan het netwerk aangeboden.
- Het netwerk moet proberen de volgende letter te voorspellen.
- Zowel woorden als letters ontvangen een score.
- Het netwerk krijgt alleen correcte woorden te zien tijdens het leerproces.
- Het netwerk wordt getest met andere correcte woorden en met incorrecte woorden.
- Alleen woorden met een score hoger dan een bepaalde drempelscore worden goedgekeurd.

Bepaling letterscores

1. De uitvoerwaarde van de corresponderende uitvoercel.
2. De uitvoerwaarde van de corresponderende uitvoercel.
3. De geometrische afstand tussen de werkelijke uitvoerletter en de gewenste uitvoerletter:

$$score_c = 1 - \sqrt{\frac{1}{29} \sum_{i=1}^{29} (doel_i(s) - werkelijk_i(s))^2}$$

Bepaling woordscores

1. Minimumletterscore.
2. Product van alle letterscores.
3. Product van alle letterscores.

Bepaling drempelscore voor woorden

De drempelscore voor woorden is de laagste score die het netwerk heeft toegekend aan de leerdata.

Voorbeeld

De respectieve uitvoercellen voor *boek* genereren de waarden: 1,00, 0,08, 0,16, 0,21, 0,06 en 0,32.

Maat	drempel	score	resultaat
1	$1 \cdot 10^{-3}$	$6 \cdot 10^{-2}$	geaccepteerd
2	$1 \cdot 10^{-7}$	$5 \cdot 10^{-5}$	geaccepteerd
3	$3 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	geaccepteerd

Lengtecorrectie

Langere woorden krijgen gemiddeld lagere scores dan kortere woorden voor maat 2 en maat 3.

Oplossing: vermenigvuldig woordscore een aantal malen met een correctiefactor afhankelijk van de woordlengte.

Netwerkinstellingen

versnelling $\eta = 0,01$

momentum $\alpha = 0,05$

aantal invoercellen = 29

aantal uitvoercellen = 29

aantal verborgen cellen = 4, 10 en 21

netwerkttest iedere 50 ronden

stabiliteitscriterium = totale leerfout verschilt

minder dan 1% van vorige leerfout

Ieder experiment is vijf keer uitgevoerd en het netwerk met de laagste leerfout is getest met de testdata.

Optimale resultaat per maat

maat	verborgen cellen	Goedgekeurde correcte reeksen	Afgekeurde incorrecte reeksen
1	10	100%	1,0%
2	4	99,8%	5,5%
3	4	100%	8,3%

- goede resultaten voor correcte reeksen
- slechte resultaten voor incorrecte reeksen
- netwerk accepteert bijna alles
- werkmethode dezelfde als die van Cleere-mans et.al.
- aanpassingen leermethode hielpen niet

Wat is de verklaring voor dit resultaat?

De uitvoerwaarde van een uitvoercel van een SRN geeft aan wat de kans is dat het corresponderende symbool de volgende in de reeks is

In de Rebergrammatica zijn deze kansen 0 of 0,5.

In de éénsyllabische Nederlandse data kan deze kans bijna elke waarde tussen 0 en 0,8 (veel waarden kleiner dan 0,1).

De netwerkuitvoer bevat ruis: uitvoerwaarden kunnen tot 0,2 afwijken van de werkelijke kans.

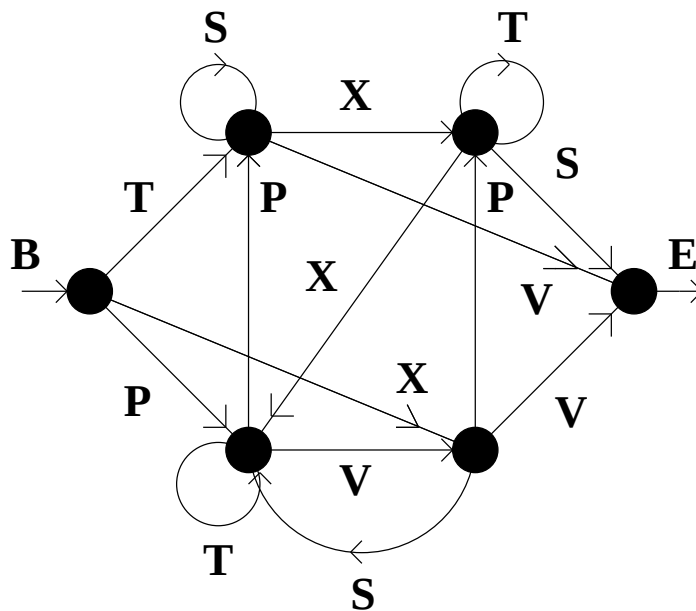
Het netwerk heeft moeite om het verschil te herkennen tussen kleine kansen en nul-kansen.

Hypothese

Als in de Rebergrammatica meer dan twee symbolen een reeks zouden kunnen voortzetten dan zou het netwerk meer problemen hebben met het leren van de grammatica.

Test

Laat een SRN zinnen leren die zijn geproduceerd door een Rebergrammatica met respectievelijk twee, drie en vier voortzettingssymbolen.



Optimale resultaat per aantal voortzettings- symbolen

aantal voortzettings- symbolen	maat	Goedgekeurde correcte reeksen	Afgekeurde incorrecte reeksen
2	1	100%	100%
3	1	100%	92,2%
4	1	100%	82,2%

Conclusies

1. Een SRN kan de structuur van een eenvoudige grammatica leren van positive voorbeelden.
2. Het netwerk heeft problemen met het leren van complexere grammatica's.
3. Een toepassing van een SRN in een taaltechnologieproduct lijkt op dit moment nog ver weg.

Slotwoord

Stoianov & Nerbonne 1998:

Een SRN kan op de taak "Leren van monosyllabische Nederlandse woorden" acceptatiepercentages van meer dan 99% correcte reeksen halen terwijl het meer dan 99% van de incorrecte reeksen afkeurt.

Experimentkenmerken:

- frequente woorden in tekstcorpora komen vaker voor in de leerdata
- netwerk gebruikt negatieve data voor het bepalen van de optimale drempelscore
- negatieve reeksen die "dichtbij" een positieve reeks liggen, zijn buiten beschouwing gelaten.

Adres

Postadres: Institutionen för Lingvistik
Box 513
751 20 Uppsala
Sweden

telefoon: +46 18 18 11 13

fax: +46 18 18 14 16

E-mail: erik.tjong@ling.uu.se

WWW: <http://stp.ling.uu.se/~erikt/>

Literatuur

- Axel Cleeremans, "Mechanisms of Implicit Learning", The MIT Press, 1993. ISBN 0-262-03205-8.
- Erik F. Tjong Kim Sang, The Limitations of Modeling Finite State Grammars with Simple Recurrent Networks. In Toine Andernach, Mark Mollen Anton Nijholt editors, "CLIN V, Papers from the Fifth CLIN Meeting", Taaluitgeverij, Enschede, The Netherlands, 1995. Ook beschikbaar op <http://stp.ling.uu.se/~erikt/papers/clin95.ps>

Overhead sheets

<http://stp.ling.uu.se/~erikt/misc/antwerp98.ps>