Jazykom riadená vizuálna pozornosť - konekcionistický model

Igor Farkaš

Katedra aplikovanej informatiky / Centrum pre kognitívnu vedu
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského v Bratislave

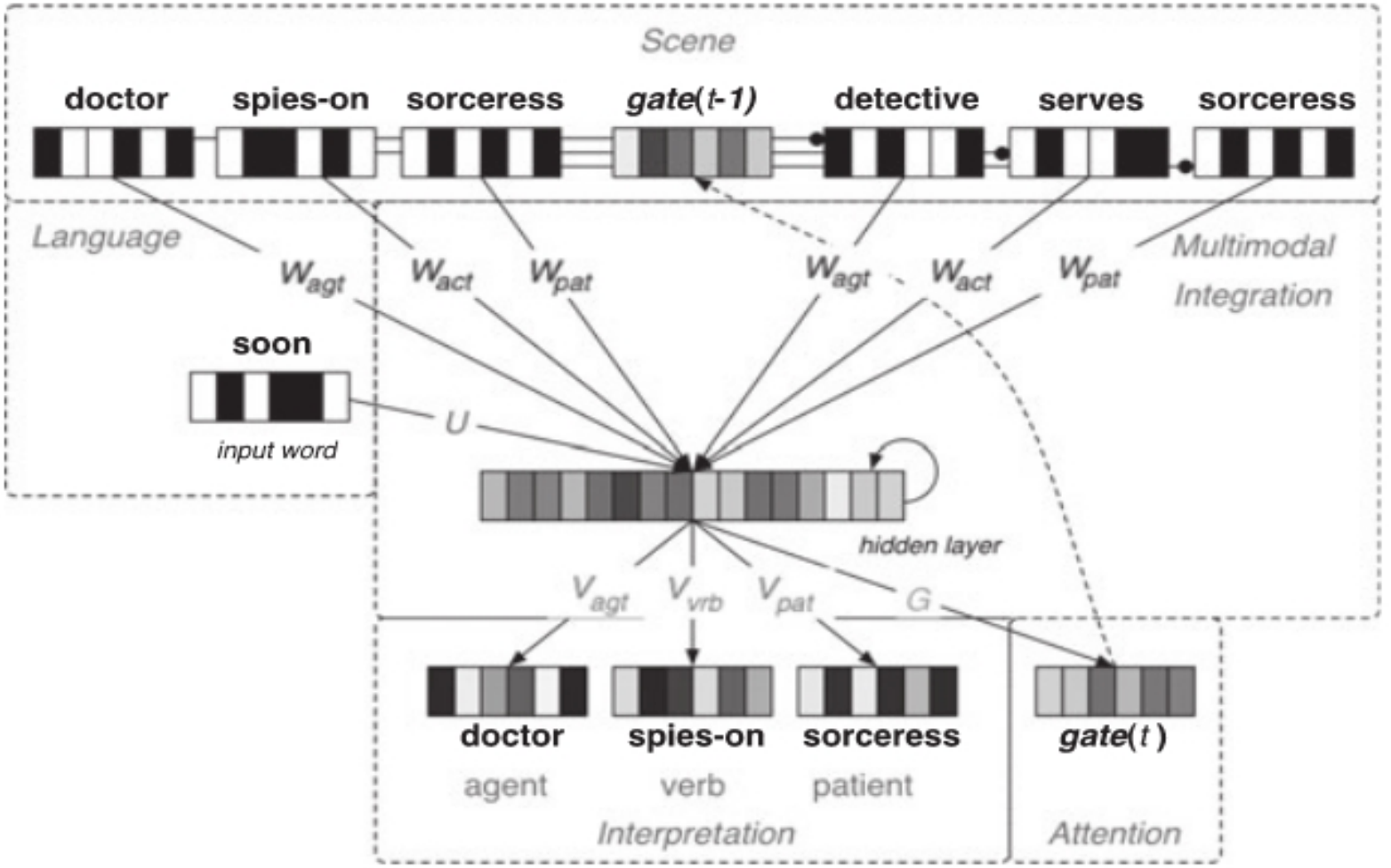
Seminár CNC, 9.11.2011

Výskum jazyka a kognície

- Tradičný výskum jazyka ako separovanej kognitívnej schopnosti
  - Syntaktické aspekty, sémantika – výpočtové modely
- Jazyk a ostatná kognícia sú však úzko prepojené (ukotvená kognícia)
- Paradigma vizuálneho sveta – využíva skutočnosť, že poslucháč má prirodzený sklon pozerať sa na relevantné elementy vizuálnej scény, ktoré sa v reči spomínajú alebo sa dajú očakávať (meranie pohybu očí – eye-tracking).
- Hovorený jazyk môže usmerňovať pozornosť v relevantnej vizuálnej scéne a informácia na scéne môže okamžite ovplyvniť proces porozumenia (Tanenhaus et al., 1995).

Connectionist Model of Situated Language Comprehension

CIAnet



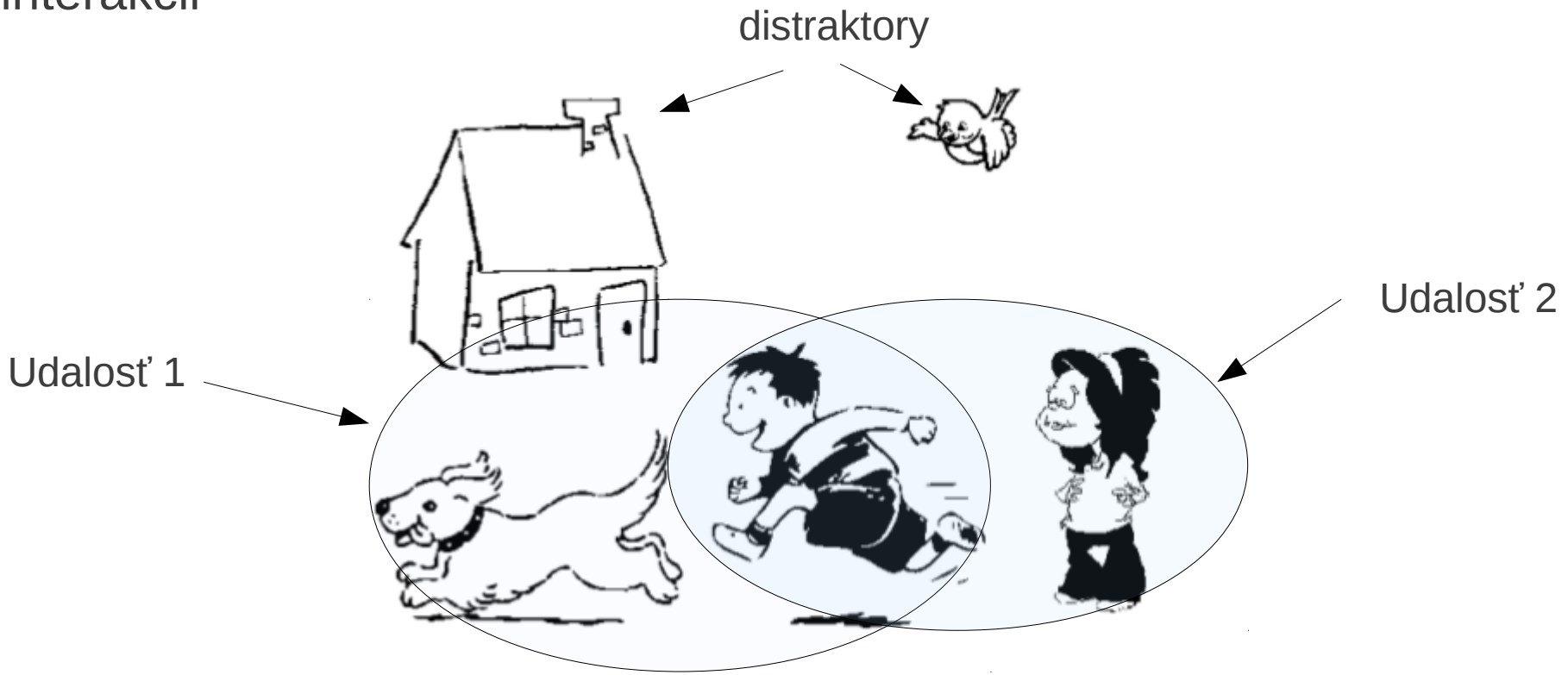
(Mayberry, Crocker, Knoeferle, 2009)

Príklad vizuálneho stimulu a opisu udalosti

RPIT No-Conflict (NC) Conditions		
(3)	Stereo	*Den Piloten verzaubert gleich der Zauberer.* The pilot$_{acc}$ enchants soon the wizard$_{nom}$. ''The wizard will soon enchant the pilot.''
(4)	Scene	*Den Piloten verköstigt gleich der Detektiv.* The pilot$_{acc}$ serves soon the detective$_{nom}$. ''The detective will soon serve the pilot.''

Náš model – Prezentácia vizuálnej scény

- Každá scéna = niekoľko udalostí (môžu zdieľať konštituenta) + distraktory
- Každá udalosť = agent v činnosti resp. s objektom/paciensom v interakcii

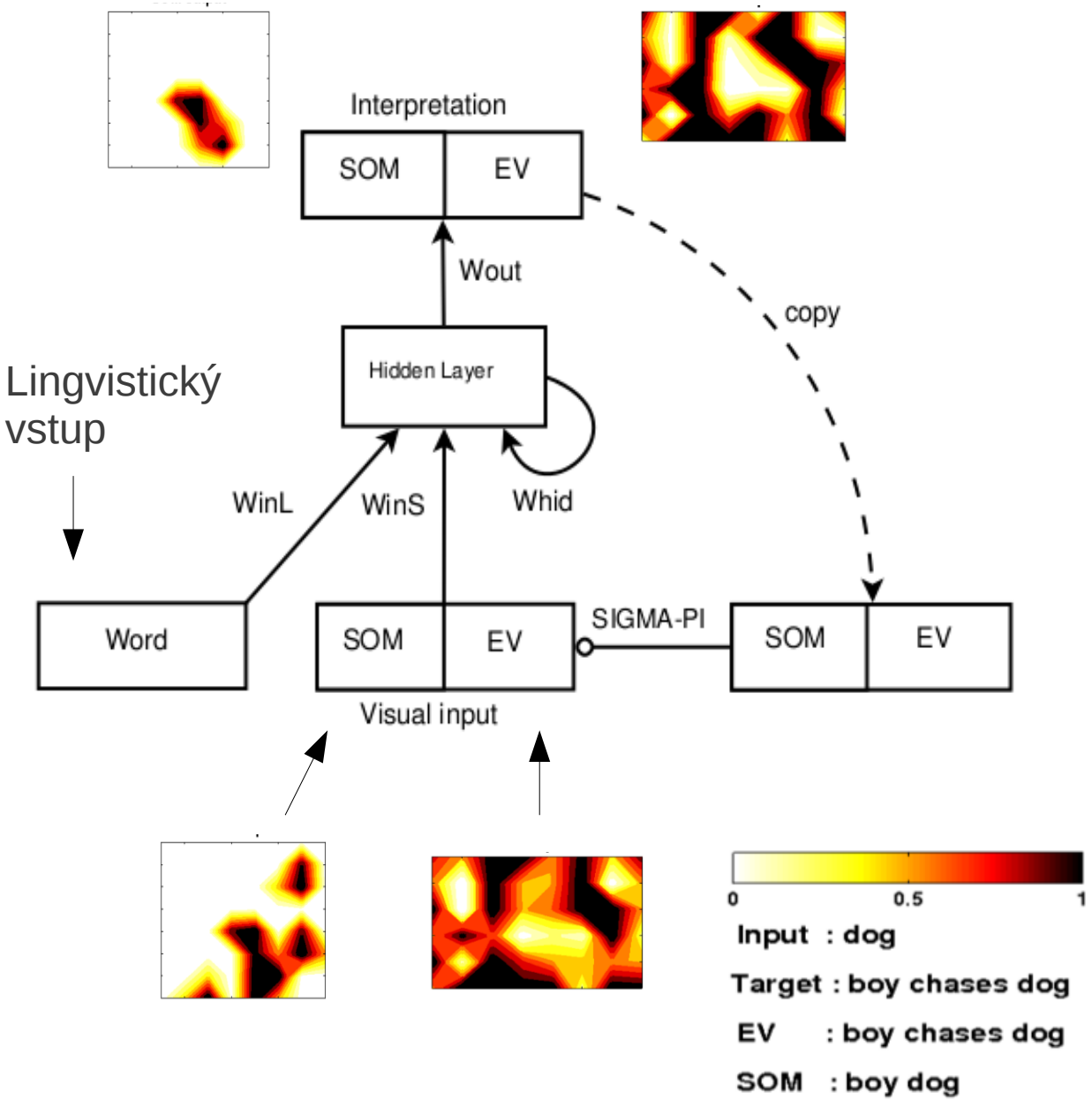


Dve udalosti: (lingvistický opis)

“Boy chases dog.”

“Girl looks-at boy.”

Porozumenie vetám vo vizuálnom kontexte



Predpoklady:

- SOM – lokalistické reprezentácie objektov
- EV – distribuované reprezentácie udalostí
- Pozornostný mechanizmus zhora-nadol = výsledok učenia
- 2-3 slovné vety typu SV(0)

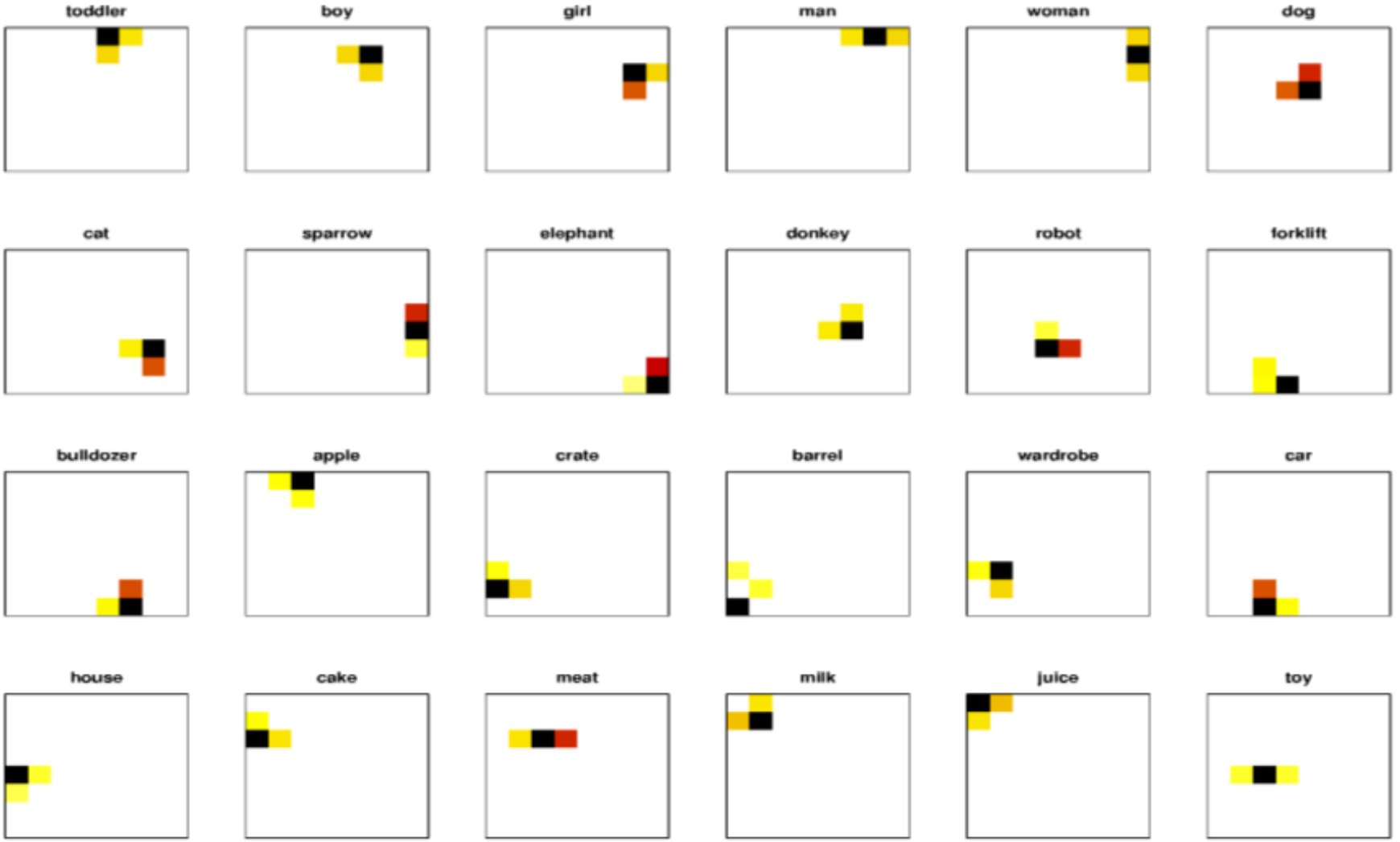
(Švantner, Farkaš, Crocker, CogSci 2011)

Vstupy modelu A-SRN

- Jazykový vstup $\mathbf{l}_{in}$: jednoduché vety typu *subjekt-sloveso-objekt* tvorené pomocou 40 slov
- Vizuálny vstup $\mathbf{s}_{in}$:
  - Predpoklad: reprezentácia scény má dve úrovne
  - Objektový vstup $\mathbf{c}_{in}$: 24 možných vizuálnych objektov je reprezentovaných pomocou SOM
  - Vstup kódujúci udalosti $\mathbf{e}_{in}$: možné udalosti sú reprezentované pomocou auto-asociatívnej siete

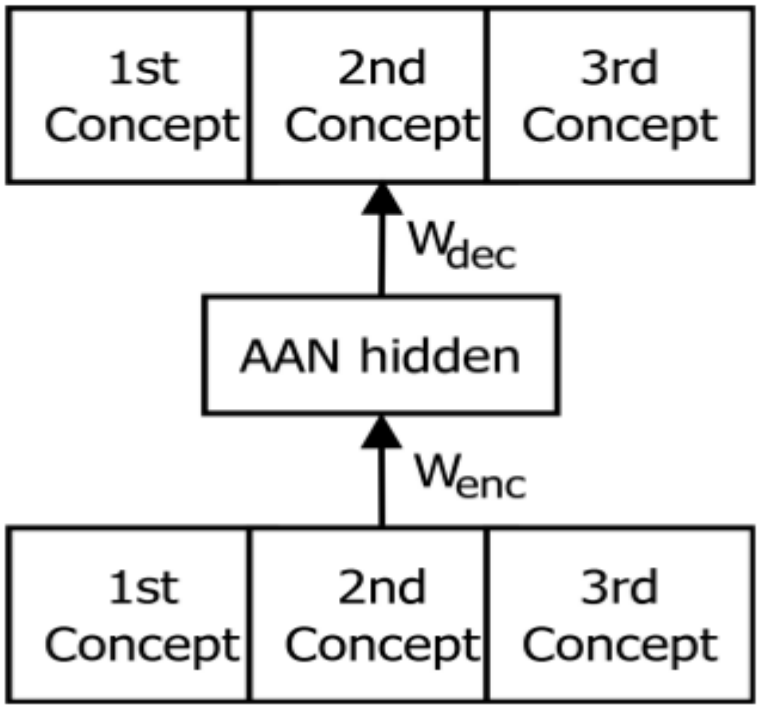
Reprezentácia objektov (pomocou SOM)

$$\mathbf{c}_{\text{in}}^{\text{all}} = \mathbf{c}_{\text{in}}^{(1)} + \ldots + \mathbf{c}_{\text{in}}^{(m)} + \mathbf{c}_{\text{D}}^{(1)} \ldots + \mathbf{c}_{\text{D}}^{(n)},$$



Reprezentácia akcií (pomocou autoasociátora)

$$\mathbf{e}_{\text{in}}^{\text{all}} = \mathbf{e}_{\text{in}}^{(1)} + \ldots + \mathbf{e}_{\text{in}}^{(k)}$$



Action features:

	anim	cnct	motn	trns	efrt	tmp	ego	flow
Walk	1	0	1	0	0	0	0	0
Run	1	0	1	0	1	0	0	0
Sit	1	0	0	0	0	0	0	0
Meditate	0	0	0	0	0	0	1	0
Lift	1	1	1	1	1	1	0	0
Push	1	1	1	1	1	0	0	0
Pull	1	1	1	1	1	0	1	0
Touch	1	1	0	1	0	1	0	0
Hold	1	1	0	1	1	0	0	0
Point-at	0	0	0	1	0	0	0	0
Look-at	1	0	0	1	0	0	0	0
Greet	0	0	0	1	0	1	0	0
Hit	0	1	1	1	1	1	0	0
Chase	1	0	1	1	1	0	0	0
Eat	1	0	0	0	0	1	1	0
Drink	1	0	0	0	0	1	1	1

Aktivácia a trénovanie

Aktivácia:

$$\begin{aligned} \mathbf{a}_{\text{hid}}(t) &= \sigma(\mathbf{W}_{\text{inL}}.\mathbf{l}_{\text{in}}(t) + \mathbf{W}_{\text{inS}}.\mathbf{s}'_{\text{in}}(t) + \mathbf{W}_{\text{hid}}.\mathbf{a}_{\text{hid}}(t-1)) \\ \mathbf{a}_{\text{out}}(t) &= [\mathbf{c}_{\text{out}}(t), \mathbf{e}_{\text{out}}(t)] = \sigma(\mathbf{W}_{\text{out}}.\mathbf{a}_{\text{hid}}(t)) \end{aligned}$$

kde $\mathbf{s}'_{\text{in}}(t)$ vyjadruje:

$$\mathbf{s}'_{\text{in}}(t) = \begin{cases} \mathbf{s}_{\text{in}}(t) & \text{pre ESN a SRN} \\ \mathbf{s}_{\text{in}}(t). * \mathbf{a}_{\text{out}}(t-1) & \text{pre A-SRN} \\ \gamma\ \mathbf{s}_{\text{in}}(t) + (1-\gamma)\ \mathbf{s}_{\text{in}}(t). * \mathbf{a}_{\text{out}}(t-1) & \text{pre A-SRN}^{+} \\ \mathbf{s}_{\text{in}}(t). * \sigma(\mathbf{W}_{\text{bck}}.\mathbf{a}_{\text{hid}}(t-1)) & \text{pre A-SRN}_{\text{bck}} \end{cases}$$

SIGMA-PI spojenie (ozn. '.*'): násobenie vektorov po zložkách

Trénovanie:

- SRN, A-SRN: back-propagation through time
- ESN: výpočet váh cez pseudoinverziu

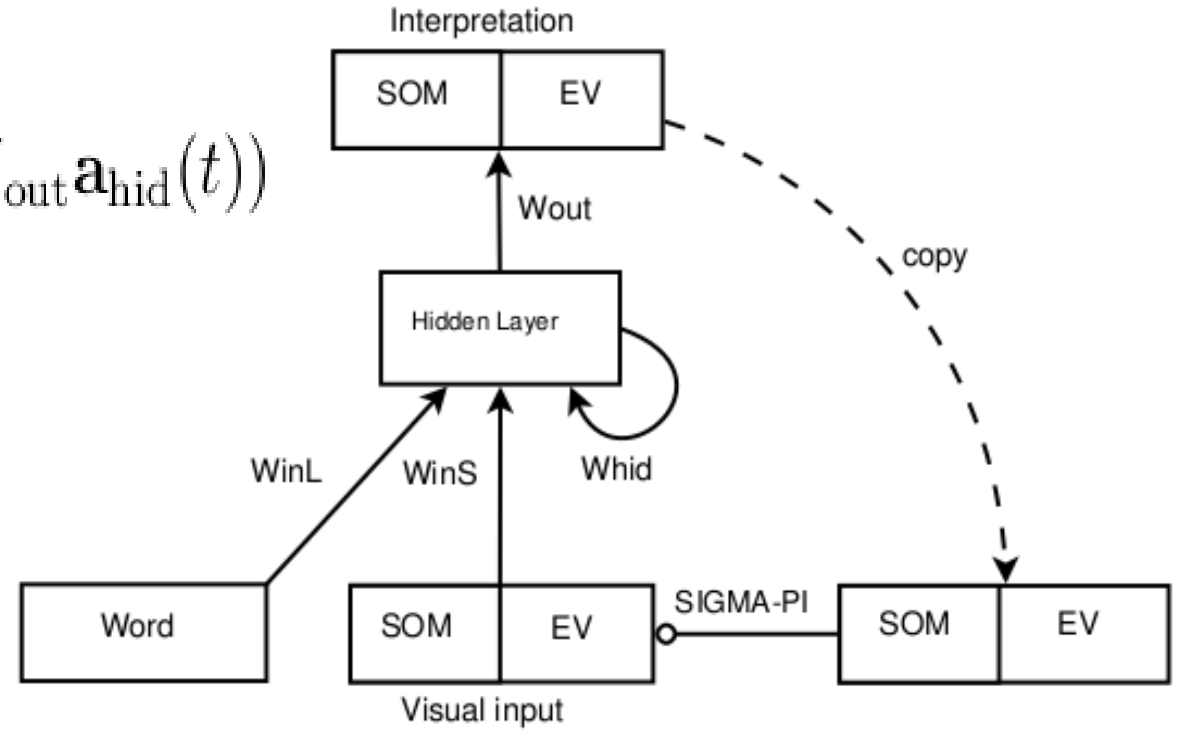
$$\mathbf{W}_{\text{out}} = \mathbf{A}_{\text{tgt}}\mathbf{A}_{\text{hid}}^{\top}(\mathbf{A}_{\text{hid}}\mathbf{A}_{\text{hid}}^{\top} + \alpha^2 I)^{-1}$$

Trénovacie dáta

- Objekty: životné, neživotné,... 24 ks.
- Akcie: pohyb, manipulácia, konzumácia, sociálne... 16 ks
- Lexikon = 40 slov, one-to-one mapping.
- Korpus: 10000 párov udalosť-veta, 7000 na trénovanie
- Príklady: Toddler looks-at crate, Woman walks,...

A-SRN: s extra explicitnou spätnou väzbou

$$\mathbf{a}_{\text{out}}(t) = F(\mathbf{W}_{\text{out}}\mathbf{a}_{\text{hid}}(t))$$



$$\mathbf{a}_{\text{hid}}(t) = \sigma(\mathbf{W}_{\text{inL}}.\mathbf{l}_{\text{in}}(t) + \mathbf{W}_{\text{inS}}.\mathbf{s}'_{\text{in}}(t) + \mathbf{W}_{\text{hid}}.\mathbf{a}_{\text{hid}}(t-1))$$

$$\mathbf{s}'_{\text{in}}(t) = \mathbf{s}_{\text{in}}(t).*\mathbf{a}_{\text{out}}(t-1)$$

sigma-pi spojenie

Výsledky (1) – presnosť na konci viet

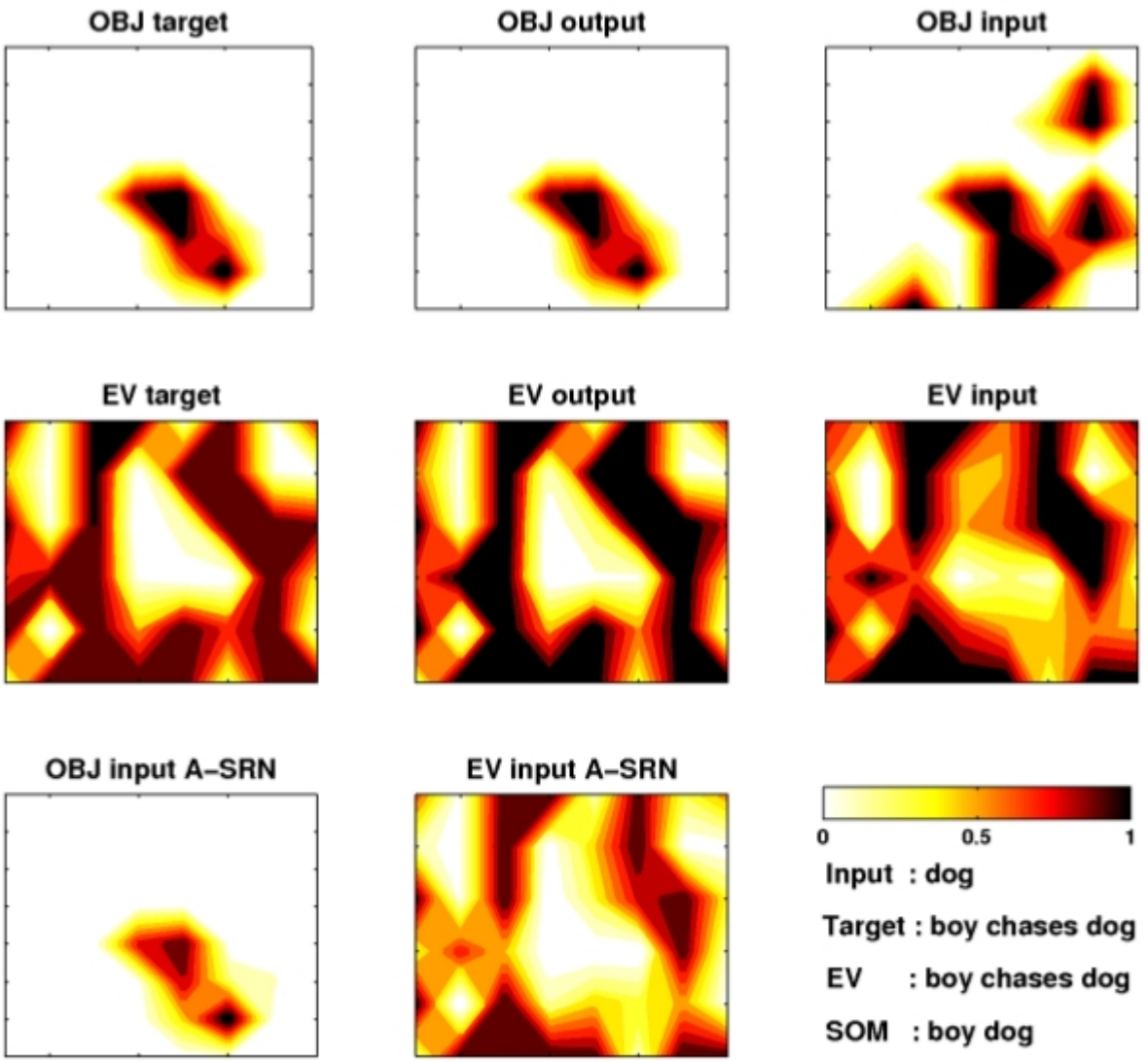
- **cos :** kosínus medzi cieľovým a skutočným výstupom siete
- **EV :** presnosť pri dekódovaní udalosti na konci vety. Sieť je úspešná iba pri správnom dekódovaní trojice agent-akcia-objekt
- **OBJ :** predikcia agenta a objektu získaná z aktivácie objektovej časti výstupu (SOM)

Target matching at sentence end

Model	SOM	EV
SRN	0.986	0.985
A-SRN	0.949	0.899
A-SRN+	0.976	0.949

γ=0.3

Príklad aktivácií natrénovanej A-SRN



Výsledky (2) – obmedzenie viz. vstupu

Motivácia: v snahe vylepšiť presnosť modelu, (a) nútiť model viac sa spoliehať na lingvistický vstup, (b) simulovať “len počúvanie.”

Pomohlo v prípade A-SRN, in 50% case

Restricting the situational input

Model	SOM-50	EV-50	SOM-0	EV-0
SRN	0.989	0.995	0.627	0.504
A-SRN	0.988	0.989	0.823	0.769
A-SRN+	0.990	0.992	0.688	0.671

Výsledky (3) – predikcia pred koncom vety

Testovali sme predikčnú schopnosť modelov (objekt=paciens) vzhľadom na:
a) presnosť cieľa (požadovaného výstupu)
b) to, či predikovaný objekt je konzistentný so svetom
c) to, predikovaný objekt je na aktuálnej scéne.

Presnosť dekódovania v SOM (paciens) = asi 50%.

Všetky 3 modely boli zhruba rovnako presné z pohľadu EV.

Patient prediction – all models - EV

	target	world	scene
At subject	0.5	0.8	0.8
At verb	0.65	0.95	0.85

Správanie modelu A-SRN

- Úspešná generalizácia modelov
- 100% zameranie pozornosti na relevantné konštituenty na konci vety
- Istá miera anticipovania (pred koncom vety)
- Mechanistické porozumenie pozornosti – dôležité v KV
- 4 fundamentálne procesy pozornosti: working memory, top-down sensitivity control, competitive selection, and automatic bottom-up filtering for salient stimuli (Knudsen, 2007).
- A-SRN ~ “top-down sensitivity control that regulates the strength of different signals that compete to access to working memory.”
- Dôležité vylepšenie modelu: samostatné 'what' a 'where' vizuálne subsystémy.

Fundamental components of attention

(Knudsen, 2007)