Uporaba globokih konvolucijskih nevronskih mrež na surovem besedilu

Žiga Pušnik

Motivacija

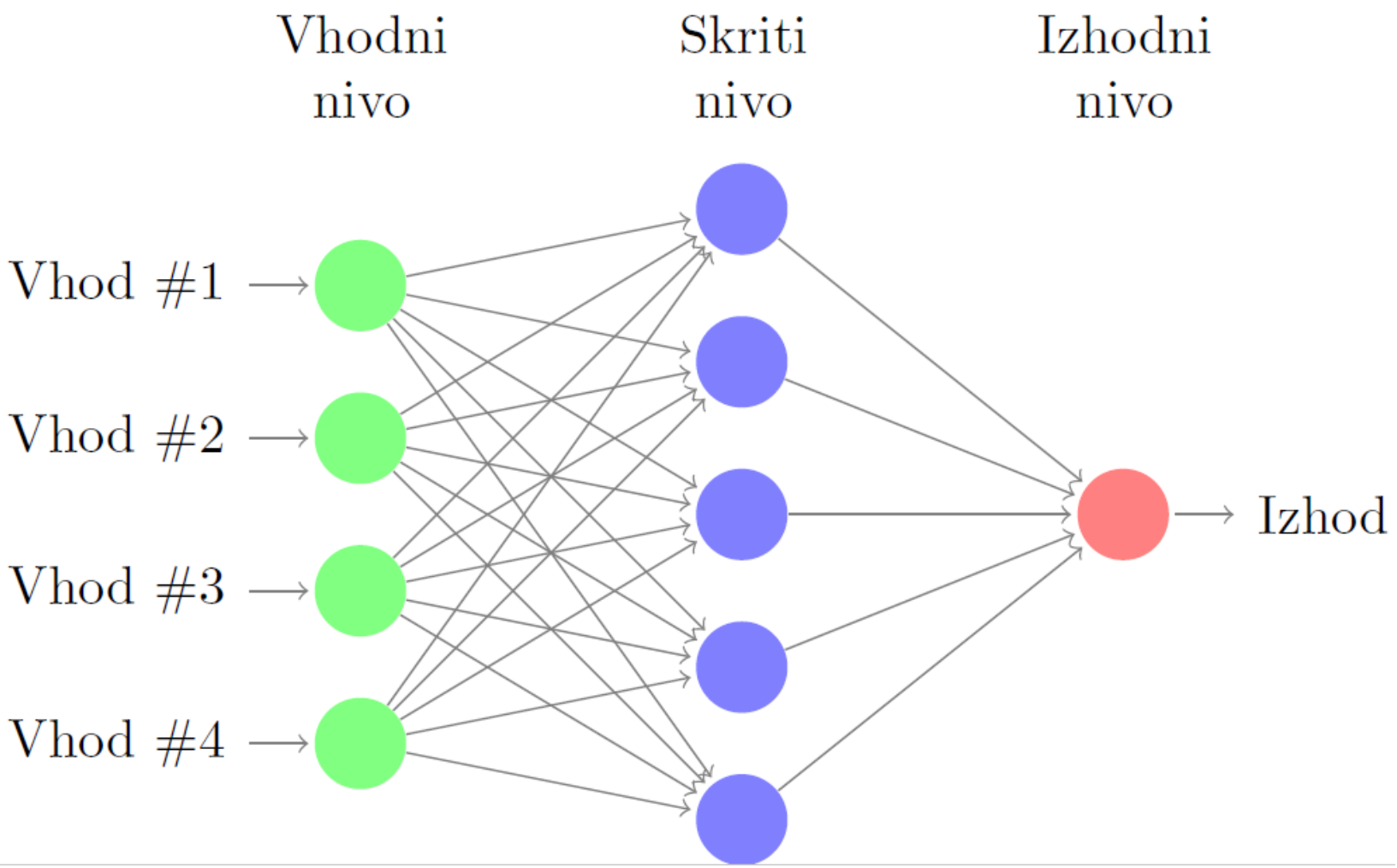
- Ali lahko računalnik razume besedilo s konvolucijskimi nevronskimi mrežami?
- *Text understanding from scratch*
  - *(98% – 99%) klasifikacijska točnost*
- klasifikacija člankov

- predprocesiranje besedila pri obdelavi naravnega jezika

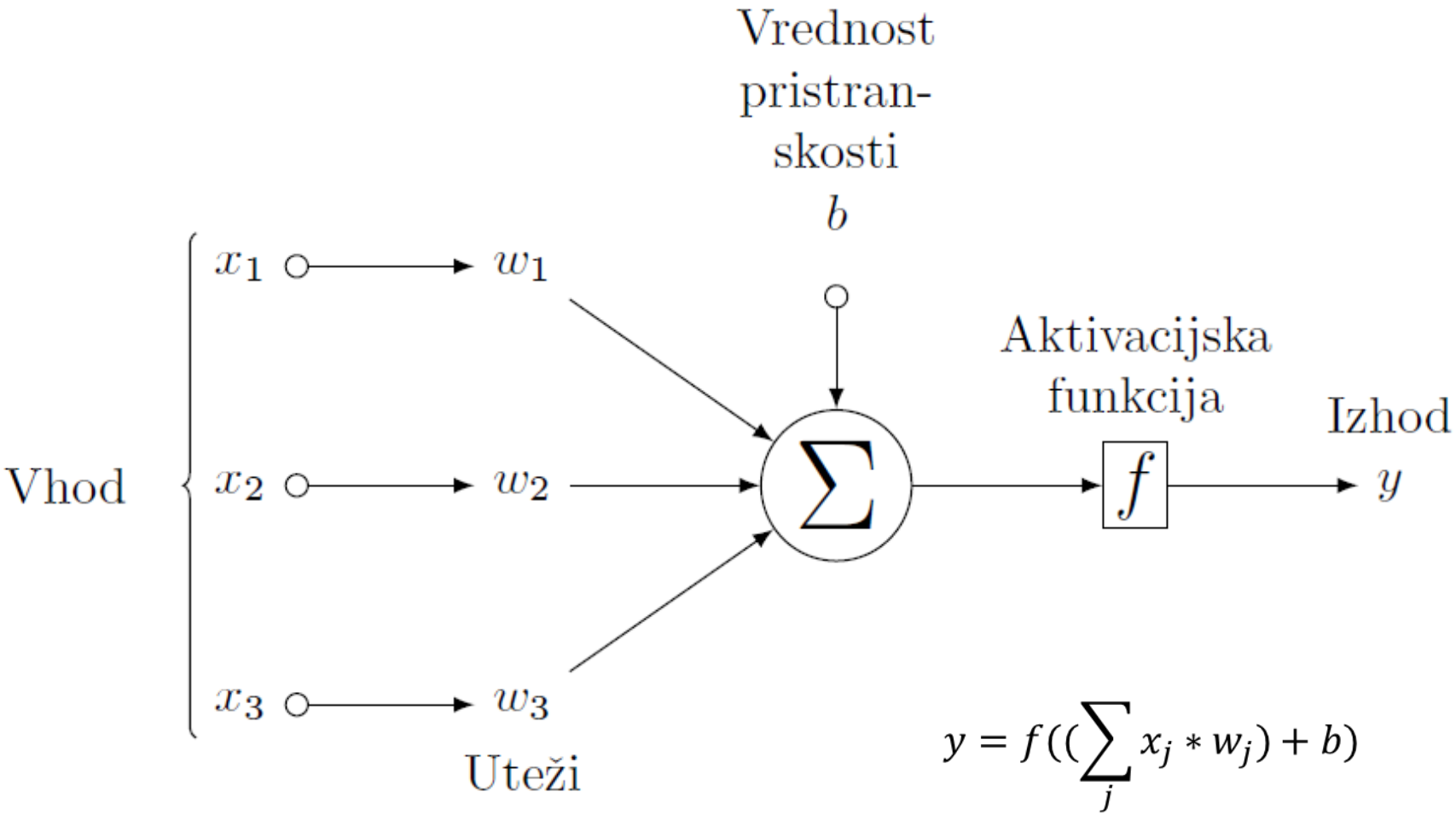
Nevronske mreže

- vzporednica z možgani
- aproksimacija poljubne funkcije (teoretično)
- globje arhitekture za kompleksnejše funkcije
- težave pri globokih arhitekturah

Nevronske mreže

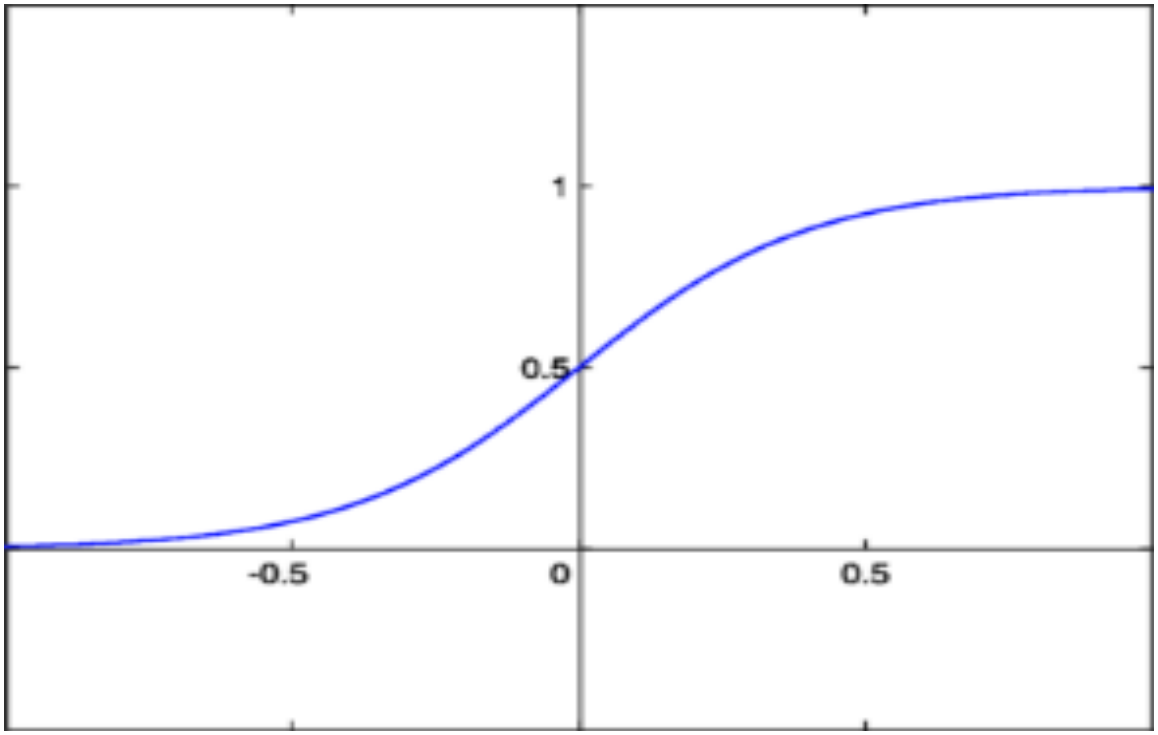


Shema nevrona



$$y = f((\sum_j x_j * w_j) + b)$$

Sigmoida



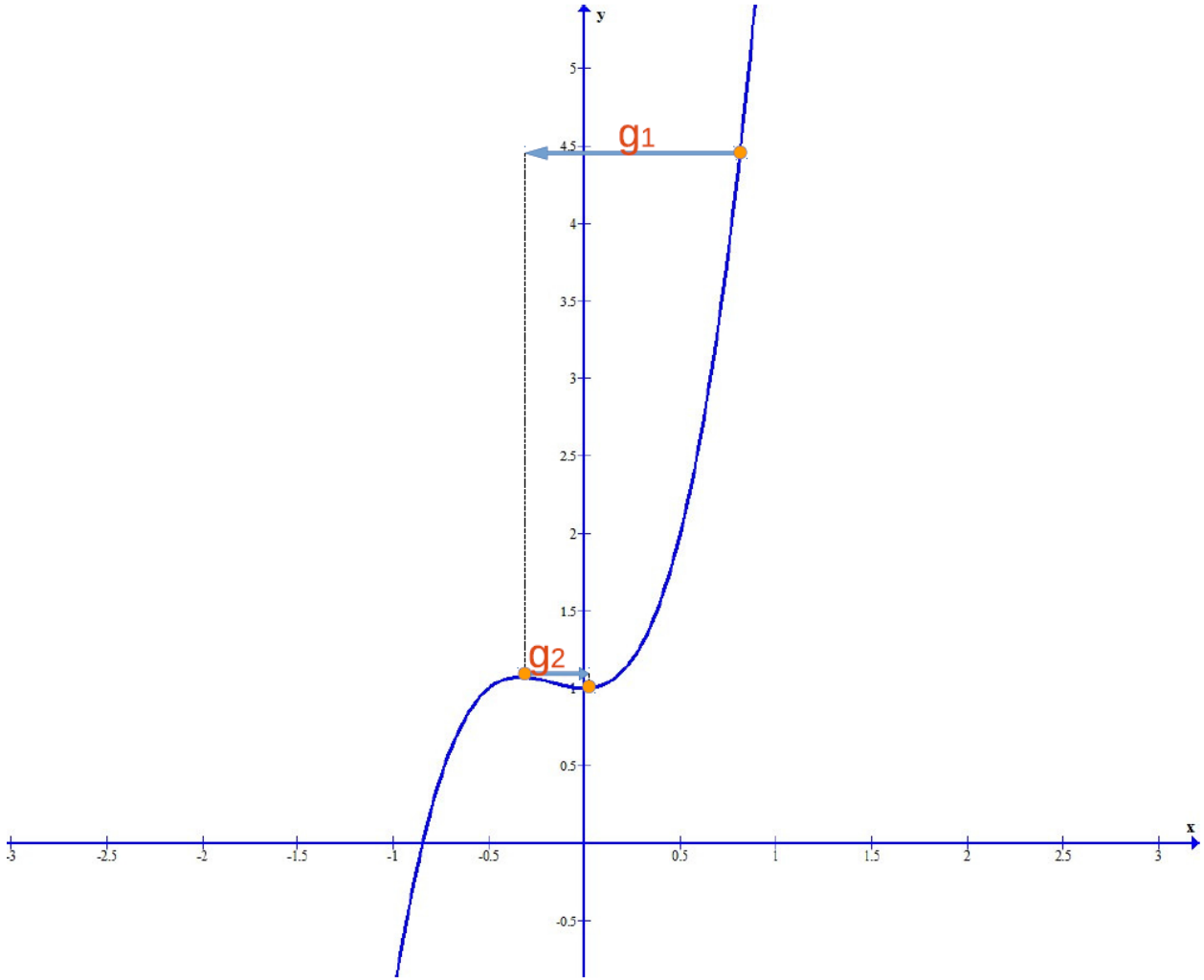
$$S(t) = \frac{1}{1 + e^{-t}}$$

Učenje nevronske mreže

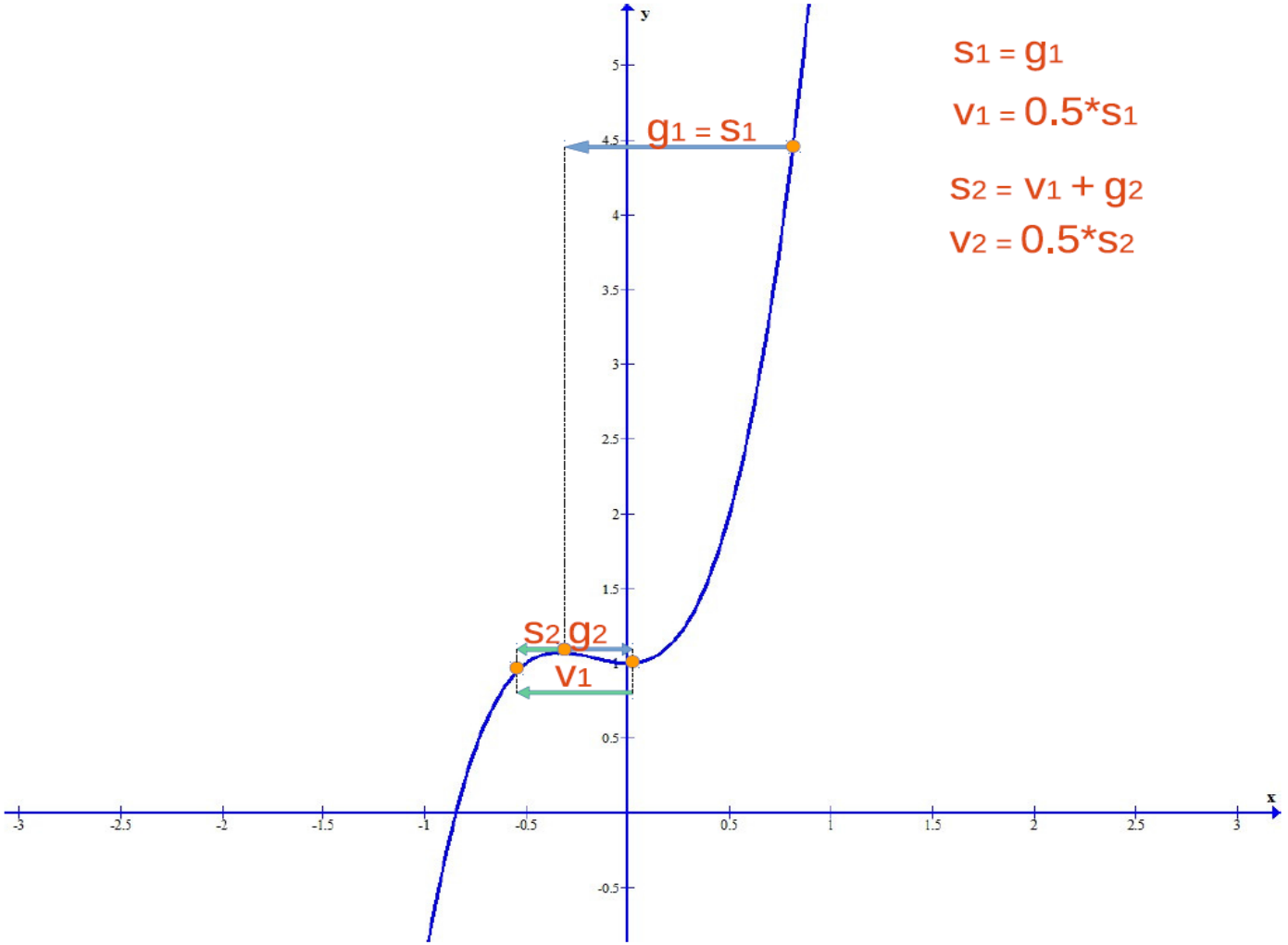
- gradientni spust (minimiziramo funkcijo napake)
  - $\Delta w_{ij} = \alpha \Delta_i x_j$ $\Delta_i = (y_i - S(in_i)) * S'(in_i)$
  - $\Delta b_i = \alpha \Delta_i$
  - α ... stopnja učenja

- vzratno širjenje napake
  - $\Delta_j = S'(in_i) * \sum_i w_{ij} \Delta_i$

Gradientni spust



Moment



$s_1 = g_1$

$v_1 = 0.5 * s_1$

$s_2 = v_1 + g_2$

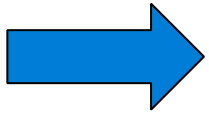
$v_2 = 0.5 * s_2$

Konvolucija

- matematična operacija
- drseče okno (filter)
- vsota slikovnih elementov (pikslov) pomnoženih z utežmi filtra

1	4	0
0	1	-1
3	-2	0

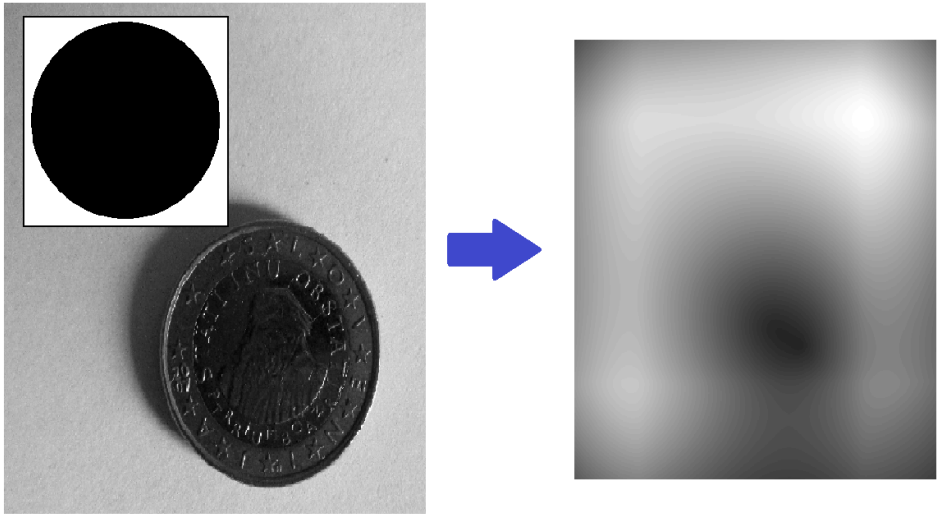
1	-1
0	2



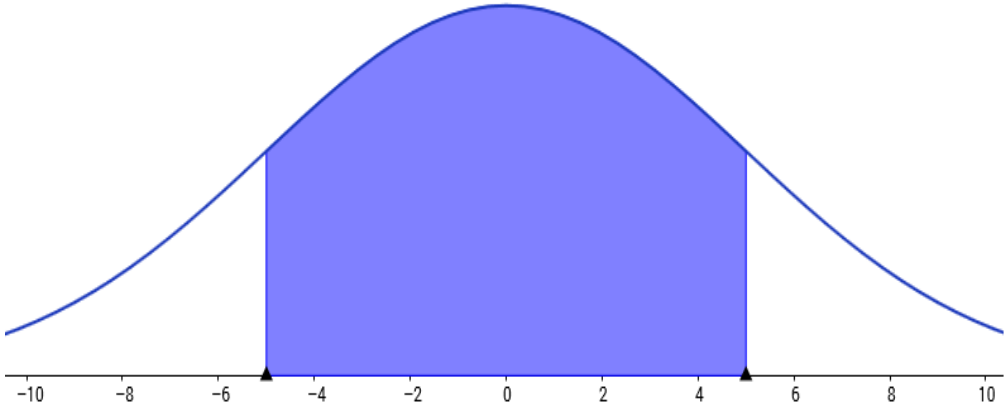
-1	2
-5	2

Konvolucija

- detekcija objektov
- detektiranje robov
- transformacija slik



Konvolucija



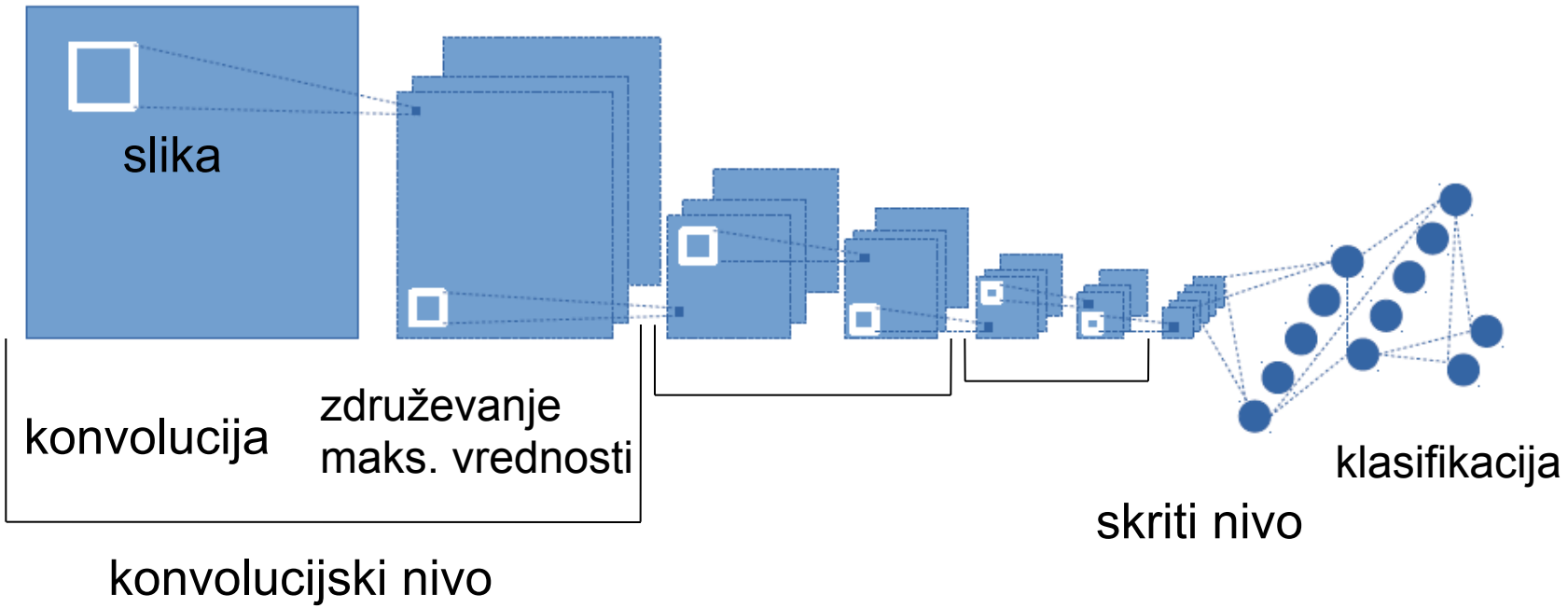
Konvolucija

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Konvolucijske nevronske mreže

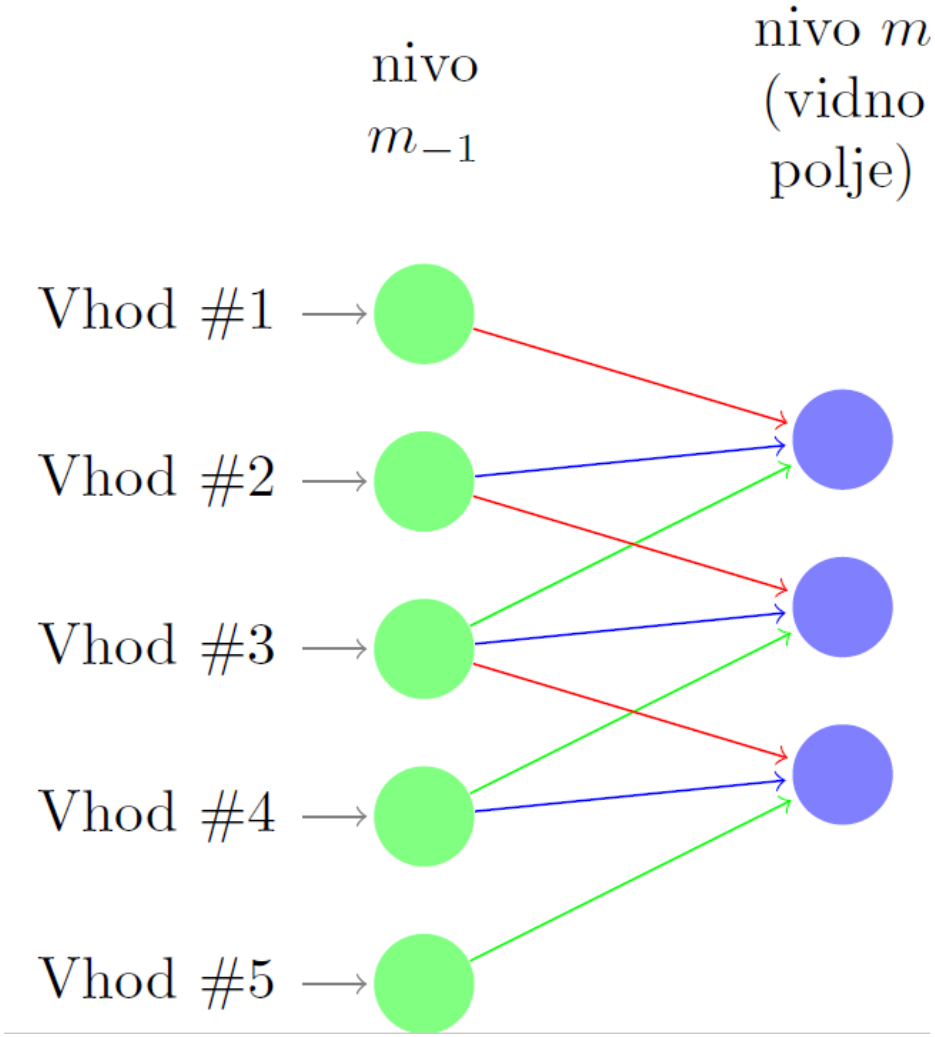
- namenjene vizualnemu zaznavanju
- nadgradnja preprostih nevronskih mrež
- globoke arhitekture
- veliko konvolucijskih nivojev
- prepoznavanje obrazov, ročno pisanih številk, razlikovanje med spoloma, ...

Konvolucijske nevronske mreže



Pri združevanju maksimalnih vrednosti posredujemo samo največjo vrednost.

Konvolucijske nevronske mreže



Klasifikacija člankov

- članki pridobljeni z Dbpedie
- 10 razredov:
  - athlete, animal, musical work, village, artist film, infrastructure, building, company, natural place
- 30.000 učnih primerov
- 10.000 testnih primerov

Obdelava člankov

- 150 zaporednih znakov iz vsakega članka
- lematizacija in krnjenje,
  - (NLTK korpusi, Lancaster stemmer)
- golo besedilo v ASCII kodiranju (bitno polje)
- vektorizacija z 1 do 45 kodiranjem, znaki
- a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, '.', '!', '?', ',', ':', '-', '"', '(', ')'

Obdelava člankov (primer)

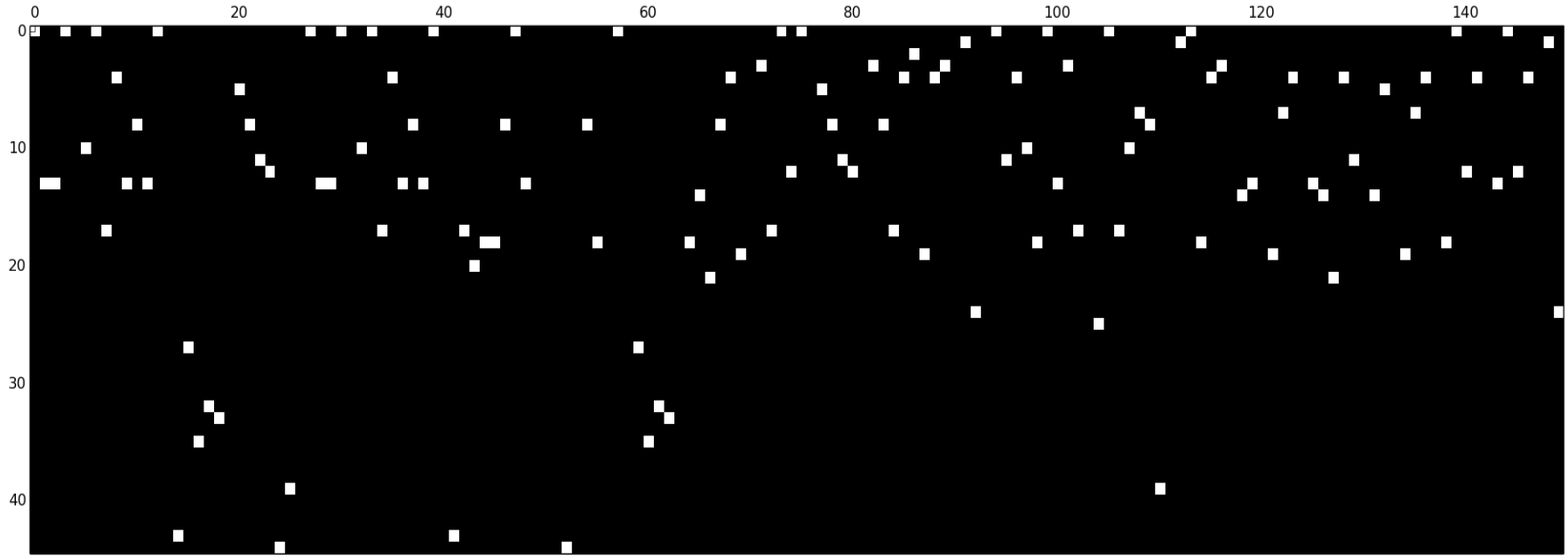
- Anna_Karenina_(1967_film), Anna Karenina (Russian: Анна Каренина) is a 1967 Soviet drama film directed by Aleksandr Zarkhi, based on the novel of the same name by Leo Tolstoy. It was listed to compete at the 1968 Cannes Film Festival, but the festival was cancelled due to the events of May 1968 in France.

Obdelava člankov (primer)

- anna_karenina_ (1967_film) , ann karenin (russ :) is a 1967 soviet dram film direct by aleksandr zarkh , bas on the novel of the sam nam by leo tolstoy . it was list to compet at the 1968 can film fest , but the fest was cancel due to the ev of may 1968 in frant .

Obdelava člankov

- podobnost z Brajevo pisavo



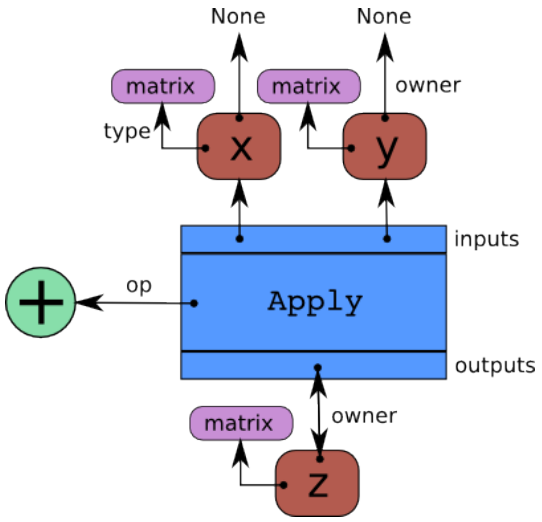
Obdelava člankov

- primer: 1 - 7 kodiranje, "c, e, i, j, n, o, s"

s	o	n	c	e		s	i	j	e
0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	1
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0
0	0	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0

Implementacija

- programski jezik python
- knjižnjica Theano
  - optimizacija simboličnih izrazov
  - simbolično odvajanje

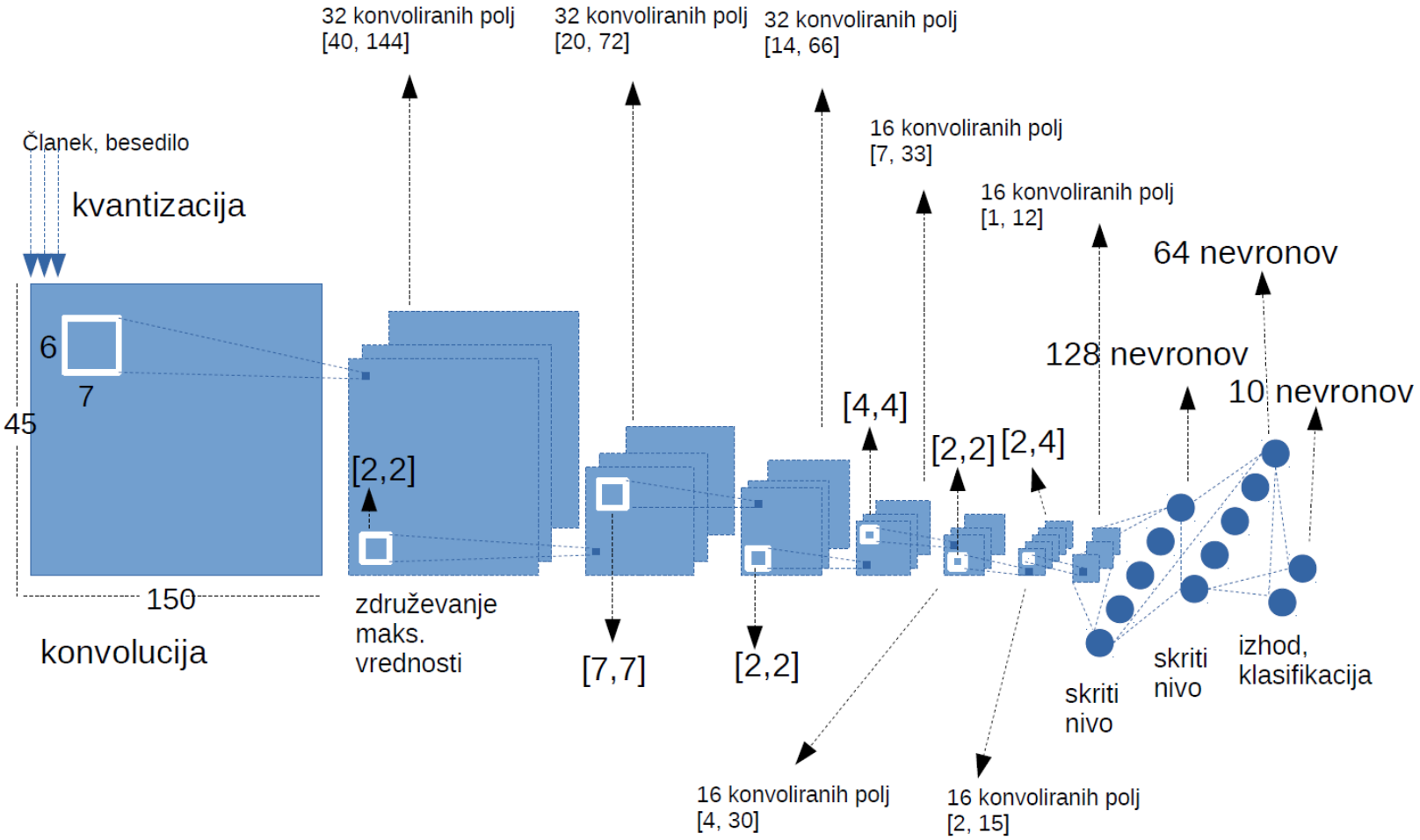


http://deeplearning.net/software/theano/tutorial/symbolic_graphs.html

Razvoj ideje in implementacija

- Kako globoko arhitekturo potrebujemo?
- Kakšne so primerne dimenzije filtrov?
- model za kvantizirane podatke in golo besedilo
  - 4 konvolucijski nivoji, 2 skrita nivoja, izhodni nivo
  - filtri za golo besedilo so enodimenzionalni

Model


32 konvoliranih polj
[40, 144]
32 konvoliranih polj
[20, 72]
32 konvoliranih polj
[14, 66]
16 konvoliranih polj
[7, 33]
16 konvoliranih polj
[1, 12]
64 nevronov
128 nevronov
10 nevronov
Članek, besedilo
kvantizacija
6
7
45
150
konvolucija
[2,2]
združevanje
maks.
vrednosti
[7,7]
[2,2]
[4,4]
[2,2]
[2,4]
skriti
nivo
skriti
nivo
izhod,
klasifikacija
16 konvoliranih polj
[4, 30]
16 konvoliranih polj
[2, 15]

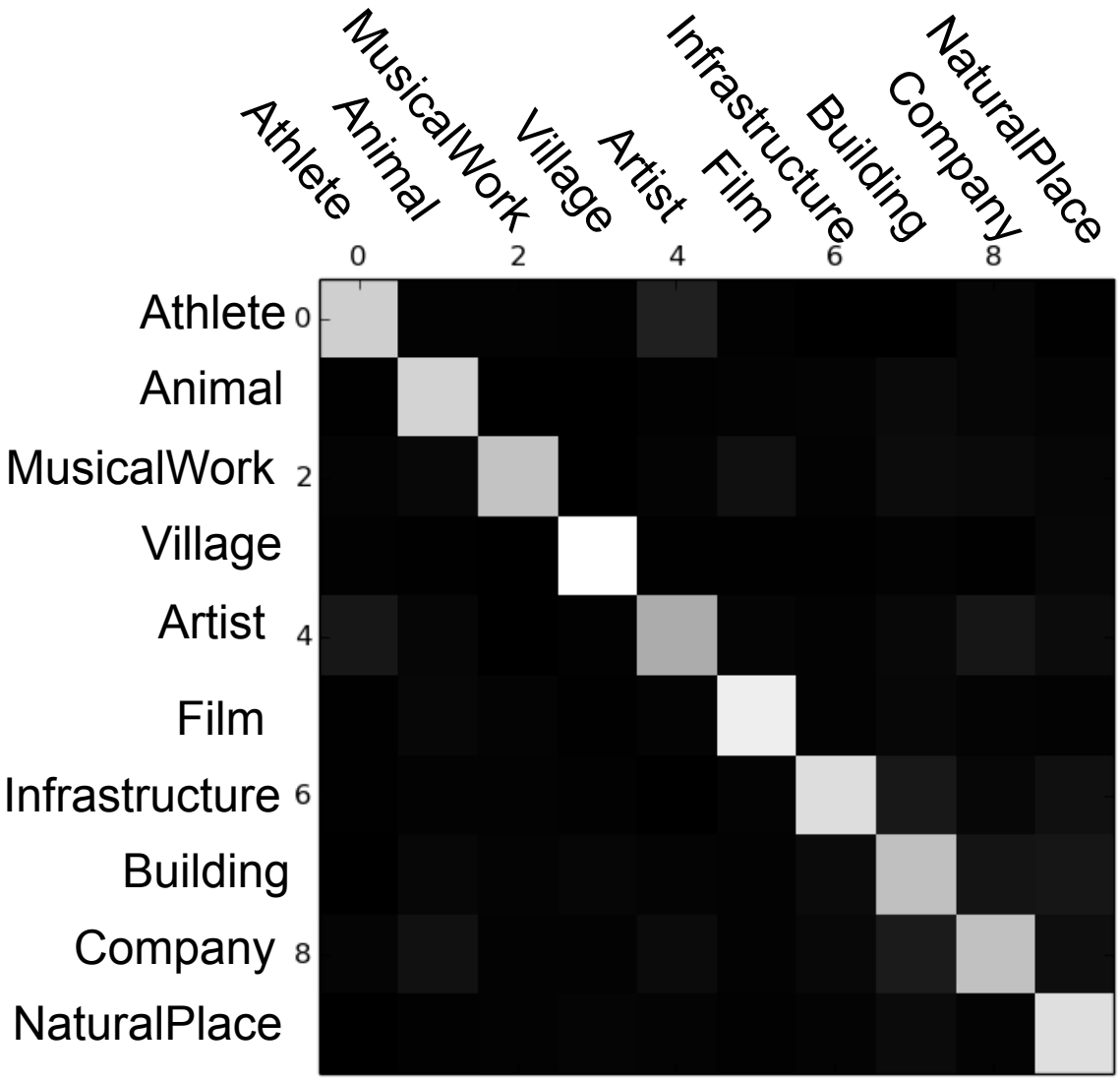
Rezultati - kategorizacija člankov

za stopnja učenja = 0.13 in stopnja momenta 0.04

klasifikacijska točnost (%)	krnjenje	lematizacija
bitno polje (ASCII koda)	20.79%	46.33%
vektorizacija	82.90%	80.90%

- dvodimenzionalna konvolucija prekaša enodimenzionalno

Matrika zmot



Zaključek

- velik potencial za jezikovne probleme
- primerne za malo obdelane podatke
- poiskati ustrezno parametrizacijo
- časovno potratne
- računsko zahtevne
- prilagoditve arhitekture
- več raziskav
- še globlje konvolucijske nevronske mreže

Zaključek

- klasifikacija člankov
- „sentiment analysis“
- učenje jezikovnih pravil (vejice)
- razčlenjevanje stavkov
- ...

Vprašanja?