18

Excel@FIT 2018

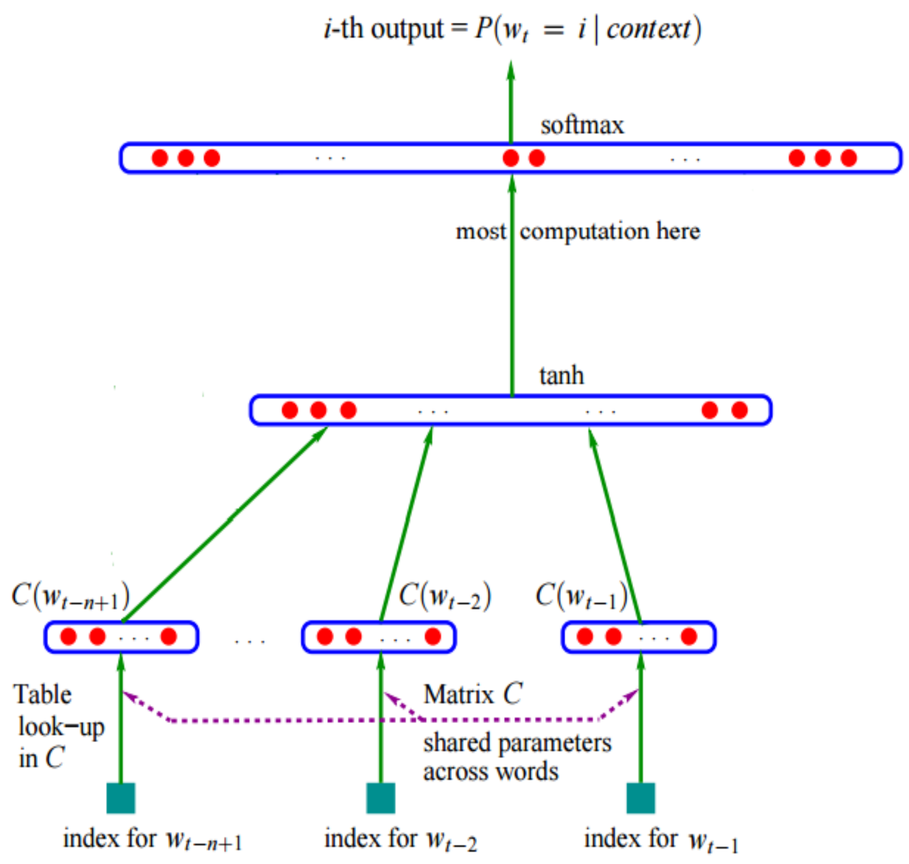
Akcelerácia neurónovej siete pre jazykové modelovanie

Jazykové modelovanie, neurónová sieť?

Jazykový model je funkcia v tvare:

$$f(w_N, w_{N-1}, ..., w_1) = P(w_N | w_1^{N-1})$$

Tento model vytvoríme pomocou neurónovej siete. Štruktúra nášho modelu je:

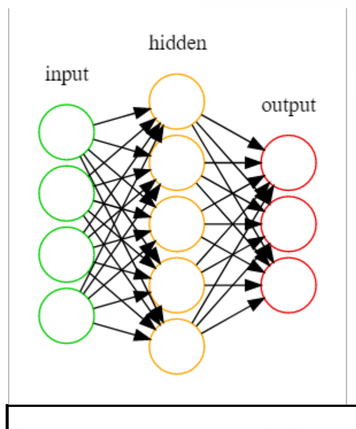


$$f(w_t, w_{t-1}, ..., w_{t-n+1}) = g(w_t, \mathbf{C}(w_{t-n+1}), ..., \mathbf{C}(w_{t-1}))$$

Výpočet neurónovej siete:

$$\mathbf{h} = \tanh(\mathbf{d} + \mathbf{Hc})$$

$$\mathbf{y} = \mathbf{b} + \mathbf{Uh}$$



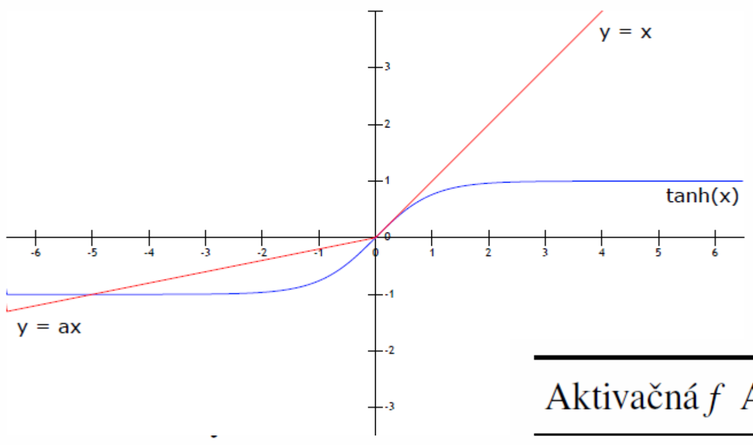
Model	Min	Max	Avg
tanh 50	4,72 ms	4,88 ms	4,75 ms
tanh 150	5,55 ms	5,81 ms	5,60 ms
tanh 250	6,27 ms	6,54 ms	6,30 ms

Function	CPU time
Embedding	145.200us
narrow	4.750us
narrow	2.451us
cat	27.543us
view	3.093us
t	4.036us
expand	3.338us
addmm	77.838us
tanh	147.768us
t	2.813us
expand	2.613us
addmm	1530.979us
log_softmax	4002.025us

Akcelerácia

Zmena aktivačnej funkcie

$$\text{PReLU}(x) = max(0,x) + a \cdot min(0,x)$$



Model	Min	Max	Avg
tanh()	5,55 ms	5,81 ms	5,60 ms
PReLU()	5,37 ms	5,41 ms	5,39 ms

Aktivačná f	Avg
tanh()	146,0 μs
PReLU()	55,7 μs

Predpočítanie matíc pre výpočet skrytej vrstvy

$$\mathbf{h} = \tanh(\sum_{j=0}^{n-1} \mathbf{M}_j[w_j])$$

Model	Čas na získanie skrytej vrstvy
baseline	254,8 μs
projekcia	100,2 μs

Model	Min	Max	Avg
baseline	5,55 ms	5,81 ms	5,60 ms
projekcia	5,37 ms	5,39 ms	5,38 ms

Ukladanie histórie

Model	Min	Max	Avg
baseline	0,68 ms	1,04 ms	0,72 ms
cache 150	0,54 ms	1,61 ms	1,01 ms
cache 300	0,54 ms	1,73 ms	1,04 ms
cache 500	0,55 ms	1,96 ms	1,17 ms

Model f	Úspešnosť
cache 150	11,3 %
cache 300	15,2 %
cache 500	18,2 %

Dominik Labaš