Ordered Neurons: Integrating Tree Structures Into Recurrent Neural Networks

Ким Алёна

НИУ ВШЭ

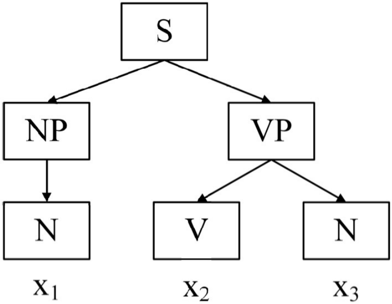
10 октября, 2019

План

- строение языка, структурирование
- что такое ordered neurons
- $cumax()$
- LSTM vs. ON-LSTM
- эксперименты
  - language modeling
  - unsupervised constituency parsing
  - targeted syntactic evaluation
- результаты

Естественные языки

- не строго последовательны
- имеют древоподобную структуру
- можно выделить так называемые составляющие **"constituents"**



Естественные языки

Интеграция древовидной структуры:

- иерархическое представление с увеличивающимся уровнем абстракции
- композиционные эффекты языка
- долгосрочные зависимости

Подходы

Supervised syntactic parser:

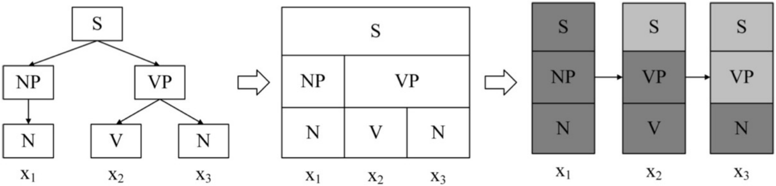
- мало разметки
- синтаксические правила в некоторых областях нарушаются (tweets)
- язык меняется, а вместе с ним правила

RNN - работают хорошо, но есть некоторые проблемы:

- не прослеживают долгосрочные зависимости
- способность к обобщению
- учитывание отрицания

LSTM потенциально могут кодировать в скрытых состояниях древовидную структуру языка

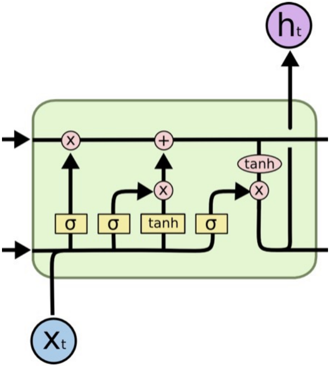
Ordered Neurons



(a) Constituency tree (b) Block view (c) ON-LSTM cell states

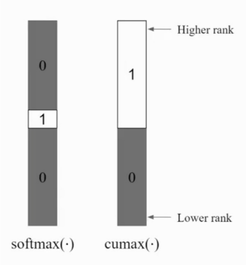
- заканчивается бóльшая составляющая, значит заканчиваются все мéньшие составляющие
- обновляется high-ranking neuron, значит все lower ranking neurons должны обновиться

Long Short Term Memory



$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$\hat{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$h_t = o_t \circ \tanh(c_t)$$

$cumax()$



$$cumax(\cdot) = cumsum(softmax(\cdot))$$

$$g = (0, .., 0, 1, .., 1)$$

d - позиция первой единицы в g

$$p(d) = softmax(\cdot)$$

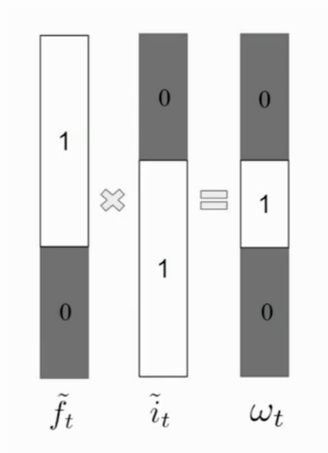
$$(d \leq k) = (d = 0) \vee (d = 1) \vee \ldots \vee (d = k)$$

$$p(g_k = 1) = p(d \leq k) = \sum_{i \leq k} p(d = i)$$

$$p(d \leq k) = cumsum(softmax(\cdot)) = \mathbb{E}[g_k]$$

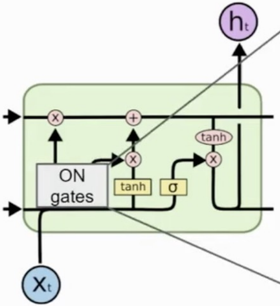
$$\hat{g} = \mathbb{E}[g]$$

Structure Gating Mechanism



- Master Forget Gate
  $\widetilde{f}_t = cumax(W_{\widetilde{f}} x_t + U_{\widetilde{f}} h_{t-1} + b_{\widetilde{f}})$
- Master Input Gate
  $\widetilde{i}_t = 1 - cumax(W_{\widetilde{i}} x_t + U_{\widetilde{i}} h_{t-1} + b_{\widetilde{i}})$
- Overlap
  $\omega_t = \widetilde{f}_t \circ \widetilde{i}_t$

Structure Gating Mechanism



$$
\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
\tilde{f}_t &= \mathrm{cumax}(W_{\tilde{f}} x_t + U_{\tilde{f}} h_{t-1} + b_{\tilde{f}}) \\
\tilde{i}_t &= 1 - \mathrm{cumax}(W_{\tilde{i}} x_t + U_{\tilde{i}} h_{t-1} + b_{\tilde{i}}) \\
\omega_t &= \tilde{f}_t \circ \tilde{i}_t \\
\hat{f}_t &= f_t \circ \omega_t + (\tilde{f}_t - \omega_t) \\
\hat{i}_t &= i_t \circ \omega_t + (\tilde{i}_t - \omega_t)
\end{aligned}
$$

Experiments. Language Modeling

Model	**Parameters**	**Validation**	**Test**
Zaremba et al. (2014) - LSTM (large)	66M	82.2	78.4
Gal & Ghahramani (2016) - Variational LSTM (large, MC)	66M	–	73.4
Kim et al. (2016) - CharCNN	19M	–	78.9
Merity et al. (2016) - Pointer Sentinel-LSTM	21M	72.4	70.9
Grave et al. (2016) - LSTM	–	–	82.3
Grave et al. (2016) - LSTM + continuous cache pointer	–	–	72.1
Inan et al. (2016) - Variational LSTM (tied) + augmented loss	51M	71.1	68.5
Zilly et al. (2016) - Variational RHN (tied)	23M	67.9	65.4
Zoph & Le (2016) - NAS Cell (tied)	54M	–	62.4
Shen et al. (2017) - PRPN-LM	–	–	62.0
Melis et al. (2017) - 4-layer skip connection LSTM (tied)	24M	60.9	58.3
Merity et al. (2017) - AWD-LSTM - 3-layer LSTM (tied)	24M	60.0	57.3
ON-LSTM - 3-layer (tied)	25M	$\mathbf{58.29 \pm 0.10}$	$\mathbf{56.17 \pm 0.12}$
Yang et al. (2017) - AWD-LSTM-MoS*	22M	56.5	54.4

Unsupervised Constituency Parsing

Model	**Training Data**	**Training Object**	**Vocab Size**	**Parsing F1** WSJ10 $\mu\ (\sigma)$	WSJ10 max	WSJ $\mu\ (\sigma)$	WSJ max	**Depth WSJ**	**Accuracy on WSJ by Tag** ADJP	NP	PP	INTJ
PRPN-UP	AllNLI Train	LM	76k	66.3 (0.8)	68.5	38.3 (0.5)	39.8	5.8	28.7	65.5	32.7	*0.0*
PRPN-LM	AllNLI Train	LM	76k	52.4 (4.9)	58.1	35.0 (5.4)	42.8	6.1	37.8	59.7	**61.5**	**100.0**
PRPN-UP	WSJ Train	LM	15.8k	62.2 (3.9)	70.3	26.0 (2.3)	32.8	5.8	24.8	54.4	17.8	*0.0*
PRPN-LM	WSJ Train	LM	10k	70.5 (0.4)	71.3	37.4 (0.3)	38.1	5.9	26.2	**63.9**	24.4	*0.0*
ON-LSTM 1st-layer	WSJ Train	LM	10k	35.2 (4.1)	42.8	20.0 (2.8)	24.0	5.6	38.1	23.8	18.3	**100.0**
ON-LSTM 2nd-layer	WSJ Train	LM	10k	65.1 (1.7)	66.8	47.7 (1.5)	**49.4**	5.6	**46.2**	61.4	55.4	*0.0*
ON-LSTM 3rd-layer	WSJ Train	LM	10k	54.0 (3.9)	57.6	36.6 (3.3)	40.4	5.3	44.8	57.5	47.2	*0.0*
300D ST-Gumbel	AllNLI Train	NLI	–	–	–	19.0 (1.0)	20.1	–	*15.6*	*18.8*	*9.9*	59.4
w/o Leaf GRU	AllNLI Train	NLI	–	–	–	22.8 (1.6)	25.0	–	18.9	24.1	*14.2*	51.8
300D RL-SPINN	AllNLI Train	NLI	–	–	–	*13.2 (0.0)*	*13.2*	–	*1.7*	*10.8*	*4.6*	50.6
w/o Leaf GRU	AllNLI Train	NLI	–	–	–	*13.1 (0.1)*	*13.2*	–	*1.6*	*10.9*	*4.6*	50.0
CCM	WSJ10 Full	–	–	–	71.9	–	–	–	–	–	–	–
DMV+CCM	WSJ10 Full	–	–	–	77.6	–	–	–	–	–	–	–
UML-DOP	WSJ10 Full	–	–	–	**82.9**	–	–	–	–	–	–	–
Random Trees	–	–	–	31.7 (0.3)	32.2	18.4 (0.1)	18.6	5.3	17.4	22.3	16.0	40.4
Balanced Trees	–	–	–	43.4 (0.0)	43.4	24.5 (0.0)	24.5	4.6	22.1	*20.2*	*9.3*	55.9
Left Branching	–	–	–	*19.6 (0.0)*	*19.6*	*9.0 (0.0)*	*9.0*	12.4	–	–	–	–
Right Branching	–	–	–	56.6 (0.0)	56.6	39.8 (0.0)	39.8	12.4	–	–	–	–

Unsupervised Constituency Parsing

	ON-LSTM	LSTM
Short-Term Dependency		
SUBJECT-VERB AGREEMENT:		
Simple	0.99	**1.00**
In a sentential complement	0.95	**0.98**
Short VP coordination	0.89	**0.92**
In an object relative clause	0.84	**0.88**
In an object relative (no *that*)	0.78	**0.81**
REFLEXIVE ANAPHORA:		
Simple	**0.89**	0.82
In a sentential complement	**0.86**	0.80
NEGATIVE POLARITY ITEMS:		
Simple (grammatical vs. intrusive)	0.18	**1.00**
Simple (intrusive vs. ungrammatical)	**0.50**	0.01
Simple (grammatical vs. ungrammatical)	0.07	**0.63**
Long-Term Dependency		
SUBJECT-VERB AGREEMENT:		
Long VP coordination	**0.74**	**0.74**
Across a prepositional phrase	0.67	**0.68**
Across a subject relative clause	**0.66**	0.60
Across an object relative clause	**0.57**	0.52
Across an object relative (no *that*)	**0.54**	0.51
REFLEXIVE ANAPHORA:		
Across a relative clause	0.57	**0.58**
NEGATIVE POLARITY ITEMS:		
Across a relative clause (grammatical vs. intrusive)	0.59	**0.95**
Across a relative clause (intrusive vs. ungrammatical)	**0.20**	0.00
Across a relative clause (grammatical vs. ungrammatical)	**0.11**	0.04

References

[1]. https://arxiv.org/pdf/1810.09536.pdf
[2]. https://colah.github.io/posts/2015-08-Understanding-LSTMs/