курс «Глубокое обучение»

Анализ текстов

Александр Дьяконов

30 марта 2020 года

План

Задачи с текстами, данные, понимания языка (Language Understanding)

Свёрточные модели для текста

Dynamic CNN, VD-CNN, Сравнение CNN vs RNN, C-LSTM

Модель seq2seq, обобщение

Механизм внимания, виды

Дифференцируемые структуры памяти

Neural Turing Machines, Pointer Network, Discrete Read/Write,
Memory Network «MemN2N», KV-MemNN

Задачи с текстами

текст → метка / метки

Определение темы / настроения / автора

Определение тональности

Разметка на части речи

текст → текст

Машинный перевод

Аннотирование

Чат-бот

текст, текст → текст

ответы на вопросы

справочная / экспертная система

… → текст

описание изображения

моделирование / генерация языка

текст → …

parse tree по предложению

Проблемы

текст – сигнал, к счастью, дискретный, с паузами между словами

один и тот же смысл передаётся по-разному

- haha
- hahahahahahahaha
- haaaahaaaa
- lol
- rotflmao
- lol!!!!!!!!!!!!!!
- wow that is big
- that is biiiiiig
- that. is. big.
- waaaaaaay big

Данные

маленькие: Penn Treebank, WikiText-2

long-term dependencies

LAMBADA (Paperno et al., 2016)

– предсказать последнее слово, используя, по крайней мере, 50 слов контекста

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernandez, R. Thelambada dataset: Word prediction requiring a broad discourse context // arXiv:1606.06031, 2016

Children's Book Test (Hill et al., 2015)

– угадать, какое слово пропущено – выбор из 10

Hill, F., Bordes, A., Chopra, S., and Weston, J. The goldilocks principle: Reading children's books with explicit memory representations // arXiv:1511.02301, 2015.

другие

The Winograd Schema challenge (Levesque et al., 2012, Trichelair et al. 2018)

«неоднозначности в тексте»

Trichelair, P., Emami, A., Cheung, J. C. K., Trischler, A., Suleman, K., and Diaz, F. On the evaluation of common-sense reasoning in natural language understanding // arXiv:1811.01778, 2018.

Conversation Question Answering dataset (CoQA) Reddy et al. (2018)

документы + диалоги о содержимом

Reddy, S., Chen, D., and Manning, C. D. Coqa: A conversational question answering challenge, arXiv:1808.07042, 2018

Данные

One Billion Word Benchmark (Chelba et al., 2013) (Al-Rfou et al., 2018)

Al-Rfou, R., Choe, D., Constant, N., Guo, M., and Jones, L.
Character-level language modeling with deeper self-attention.
arXiv preprint arXiv:1808.04444, 2018.

Summarization: CNN and Daily Mail dataset (Nallapati et al., 2016)

Translation: WMT-14 English-French

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. Abstractive text summarization using sequence-to-sequence rnns and beyond.
arXiv:1602.06023, 2016.

Question Answering: Natural Questions dataset (Kwiatkowski et al., 2019)

Kwiatkowski, T., Palomaki, J., Rhinehart, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin,J., et al.
Natural questions: a benchmark for question answering research. 2019.

Данные

GLUE (Wang et al., 2018)

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv:1804.07461, 2018.

decaNLP (McCann et al., 2018)

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. arXiv:1806.08730, 2018

IR-based QA

Stanford Question Answering Dataset (SQuAD) / SQuAD2.0

https://rajpurkar.github.io/SQuAD-explorer/

NewsQA

WikiQA

CuratedTREC

WebQuestions

WikiMovies

Russian: SberQUAD

IR-based QA

Dataset	Example	Article / Paragraph
SQuAD	**Q**: How many provinces did the Ottoman empire contain in the 17th century? **A**: 32	**Article**: Ottoman Empire **Paragraph**: ... At the beginning of the 17th century the empire contained 32 provinces and numerous vassal states. Some of these were later absorbed into the Ottoman Empire, while others were granted various types of autonomy during the course of centuries.
CuratedTREC	**Q**: What U.S. state's motto is "Live free or Die"? **A**: New Hampshire	**Article**: Live Free or Die **Paragraph**: "Live Free or Die" is the official motto of the U.S. state of New Hampshire, adopted by the state in 1945. It is possibly the best-known of all state mottos, partly because it conveys an assertive independence historically found in American political philosophy and partly because of its contrast to the milder sentiments found in other state mottos.
WebQuestions	**Q**: What part of the atom did Chadwick discover?† **A**: neutron	**Article**: Atom **Paragraph**: ... The atomic mass of these isotopes varied by integer amounts, called the whole number rule. The explanation for these different isotopes awaited the discovery of the neutron, an uncharged particle with a mass similar to the proton, by the physicist James Chadwick in 1932. ...
WikiMovies	**Q**: Who wrote the film Gigli? **A**: Martin Brest	**Article**: Gigli **Paragraph**: Gigli is a 2003 American romantic comedy film written and directed by Martin Brest and starring Ben Affleck, Jennifer Lopez, Justin Bartha, Al Pacino, Christopher Walken, and Lainie Kazan.

Table 1: Example training data from each QA dataset. In each case we show an associated paragraph where distant supervision (DS) correctly identified the answer within it, which is highlighted.

https://arxiv.org/pdf/1704.00051.pdf

SQuAD 1.0 → SQuAD 2.0

недостатки первой версии:
ответы на все вопросы есть в пределах параграфа
(во второй версии есть вариант «нет ответа»)

	SQuAD 1.1	SQuAD 2.0
Train		
Total examples	87,599	130,319
Negative examples	0	43,498
Total articles	442	442
Articles with negatives	0	285
Development		
Total examples	10,570	11,873
Negative examples	0	5,945
Total articles	48	35
Articles with negatives	0	35
Test		
Total examples	9,533	8,862
Negative examples	0	4,332
Total articles	46	28
Articles with negatives	0	28

Table 2: Dataset statistics of SQuAD 2.0, compared to the previous SQuAD 1.1.

https://arxiv.org/pdf/1806.03822.pdf

Данные

RACE

five classes of questions: word matching, paraphrasing, single-sentence reasoning, multi-sentence reasoning, insucient or ambiguous questions.

http://www.cs.cmu.edu/~glai1/data/race/

MS Marco

http://www.msmarco.org

CommonsenseQA

https://www.tau-nlp.org/commonsenseqa

SWAG

https://rowanzellers.com/swag/

HellaSWAG

https://rowanzellers.com/hellaswag/

Данные

CNN / Daily Mail

Webis-TLDR-17

Понимания языка (Language Understanding)

Что такое «понимание языка»

1) умение автоматически генерировать «желаемый ответ»

Когда ходят в школу?

Желаемые:
- **в детстве**
- **с сентября**

Не желаемые:
- **никогда**
- **вчера**

Что изображено на рисунке?

Желаемые:
- **бананы**
- **фрукты**

Не желаемые:
- **жёлтые объекты**

Свёрточные модели для текста

идея как в обработке n-грамм

$$\sigma\left(W\begin{bmatrix}x_1\\x_2\end{bmatrix}+b\right)$$

проблема – как работать с последовательностями произвольной длины

Свёрточные модели для текста

проблема – как работать с последовательностями произвольной длины

- **RNN**
- **CNN + max-pooling (max over time pooling)**

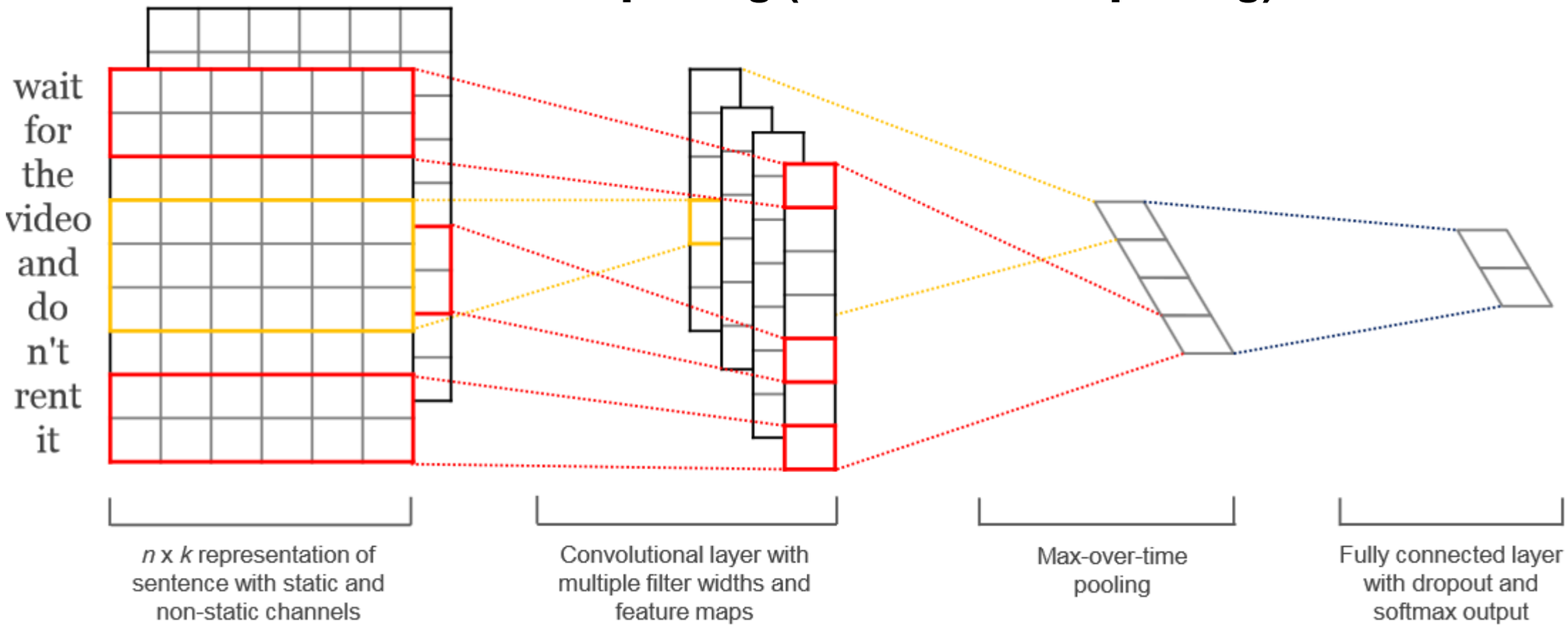


Figure 1: Model architecture with two channels for an example sentence.

Yoon Kim «Convolutional Neural Networks for Sentence Classification» // https://arxiv.org/abs/1408.5882

Свёрточные модели для текста

k – длина представления слова

n – фиксированная длина предложения

(если меньше – фиктивно набавляем)

Теперь уже наше предложение – матрица (как изображение)

подматрица векторизуется

$$c_i = \sigma\left(W_{hk\times 1} \begin{bmatrix} x_i \\ \cdots \\ x_{i+h-1} \end{bmatrix}_{1\times hk} + b \right) \in \mathbb{R}$$

на втором слое (если один фильтр):

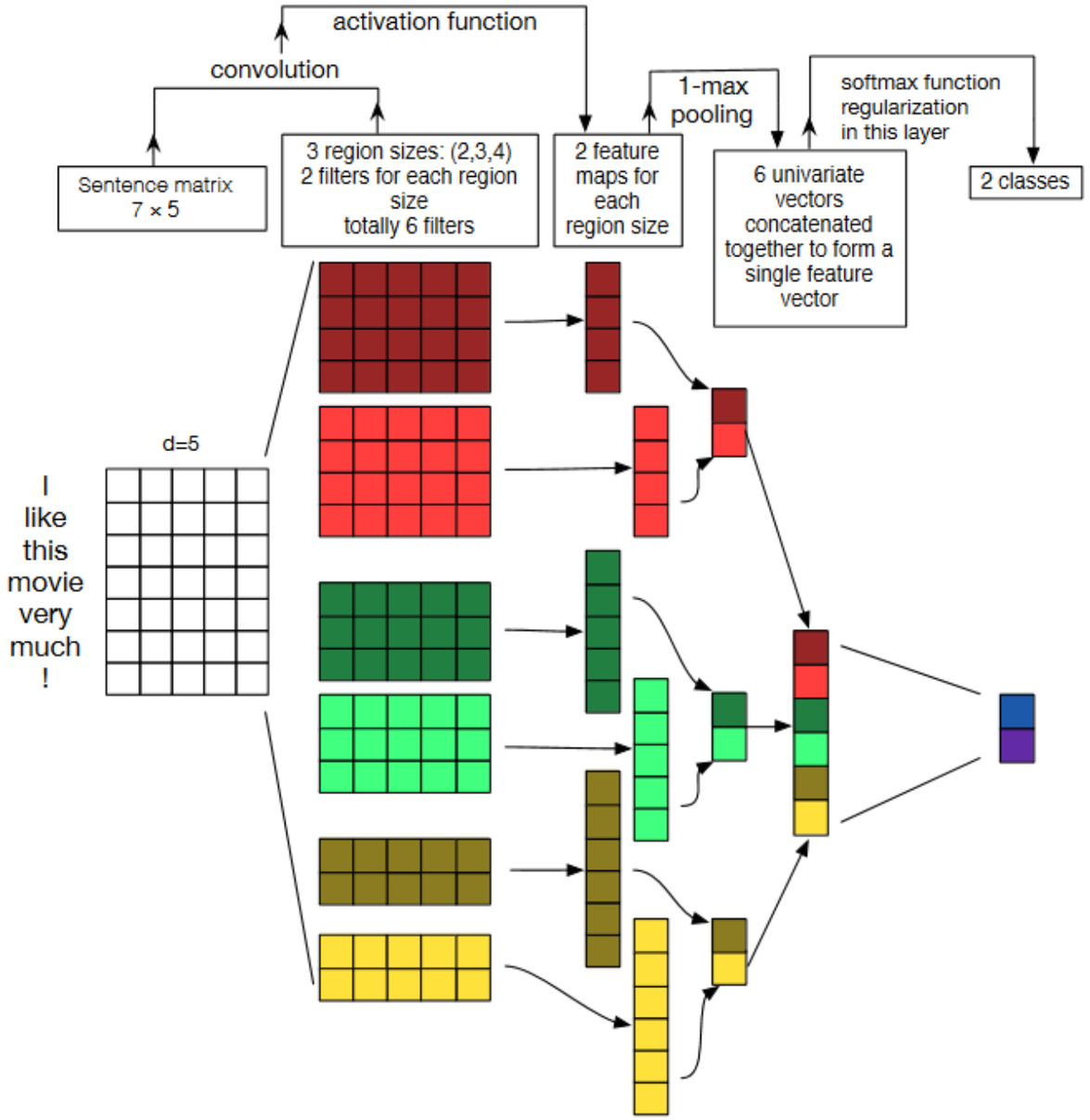
$$c = [c_1, \ldots, c_{n-h+1}] \in \mathbb{R}^{n-h+1}$$

max-pooling

$$\max(c) \in \mathbb{R}$$

для нескольких слоёв аналогично → полносвязную сеть

Свёрточные модели для текста: улучшения



Свёртки с разной шириной

Разделить пулинг и конкатенацию

дальше:

k-пулинг

[1, 2, 5, 3, 4] → [5, 4, 3]

Ye Zhang, Byron Wallace A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification https://arxiv.org/abs/1510.03820

Dynamic Convolutional Neural Network

есть возможность получения такого графа:

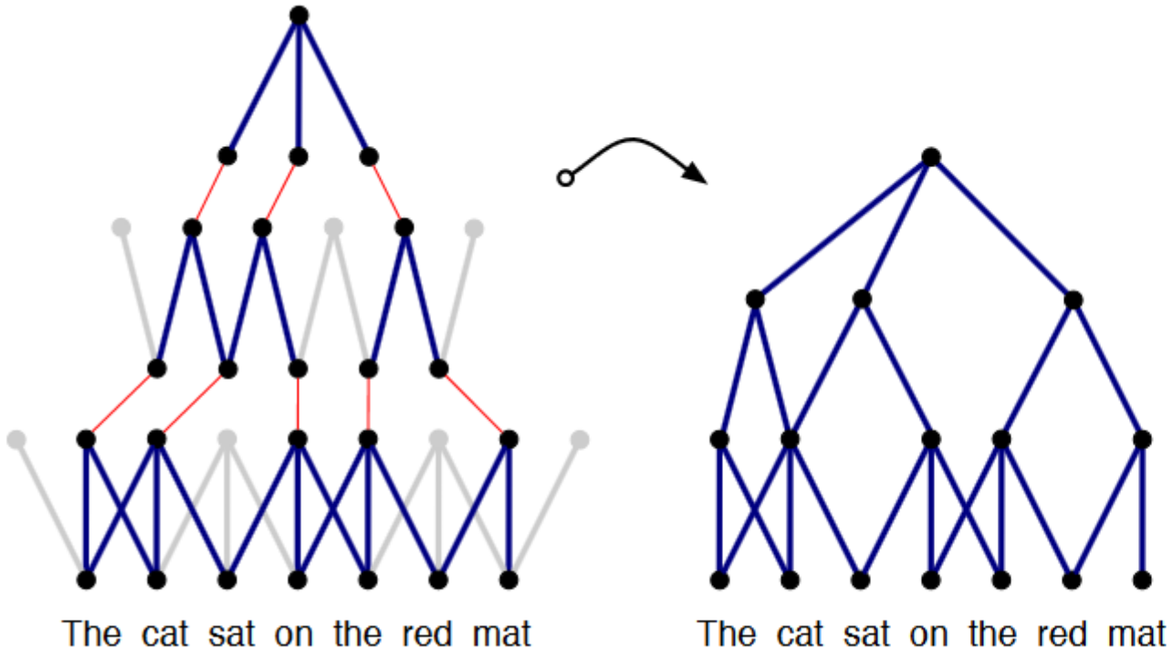


Figure 1: Subgraph of a feature graph induced over an input sentence in a Dynamic Convolutional Neural Network. The full induced graph has multiple subgraphs of this kind with a distinct set of edges; subgraphs may merge at different layers. The left diagram emphasises the pooled nodes. The width of the convolutional filters is 3 and 2 respectively. With dynamic pooling, a filter with small width at the higher layers can relate phrases far apart in the input sentence.

k-пулинг – k наибольших элементов

динамический k-пулинг – k – функция от входов и параметров сети

(например от длины входа и глубины сети)

чтобы в конце – фиксированная длина

Nal Kalchbrenner, Edward Grefenstette, Phil Blunsom «A Convolutional Neural Network for Modelling Sentences» https://arxiv.org/abs/1404.2188

Узкие и широкие свёртки

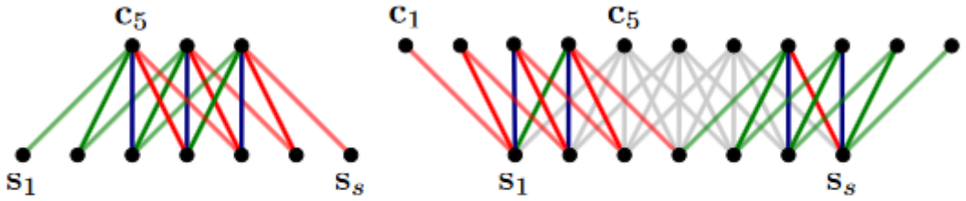


Figure 2: Narrow and wide types of convolution. The filter $\mathbf{m}$ has size $m = 5$.

сначала широкая свёртка

(dynamic) k-max pooling

Нелинейность

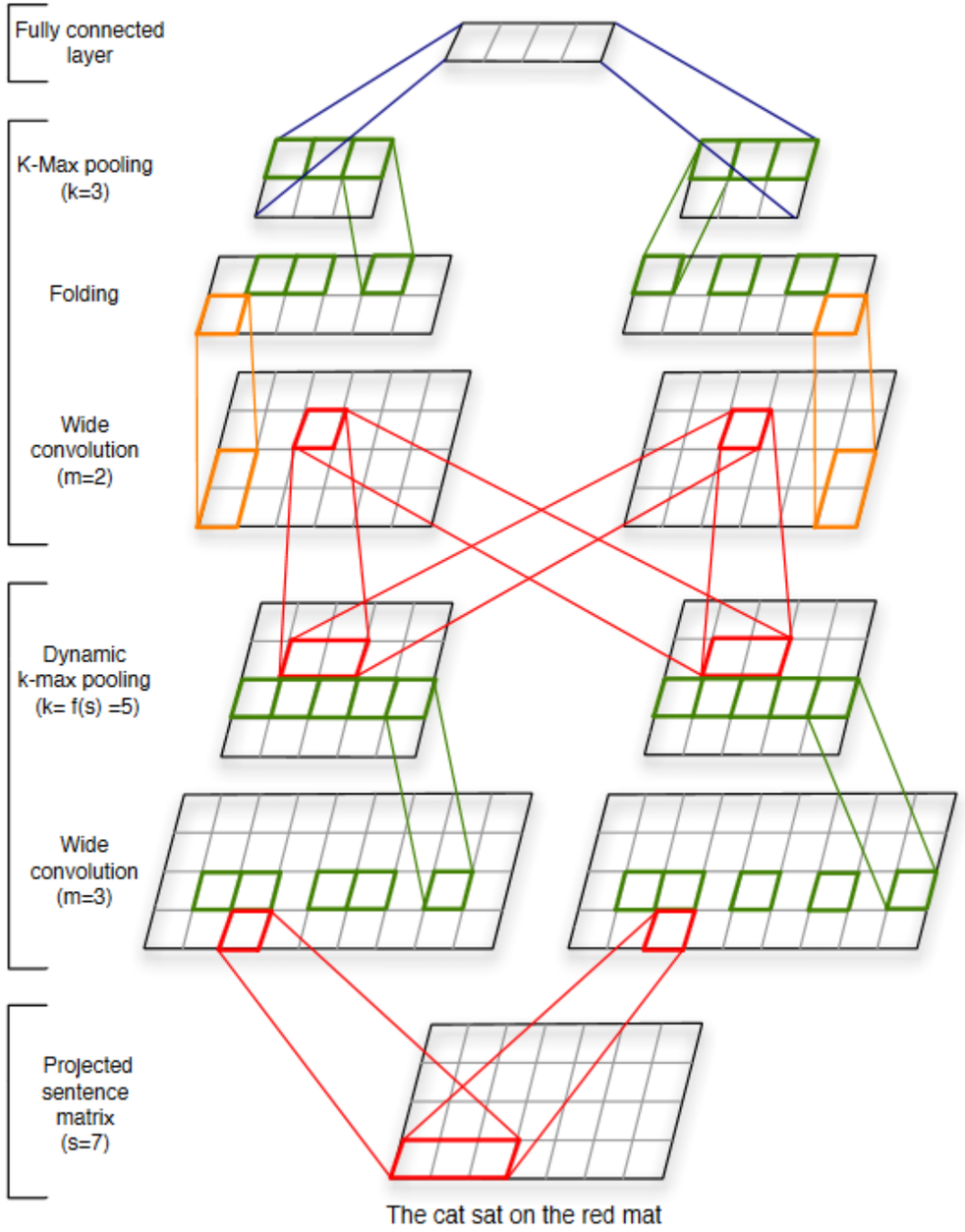
получаем несколько карт признаков

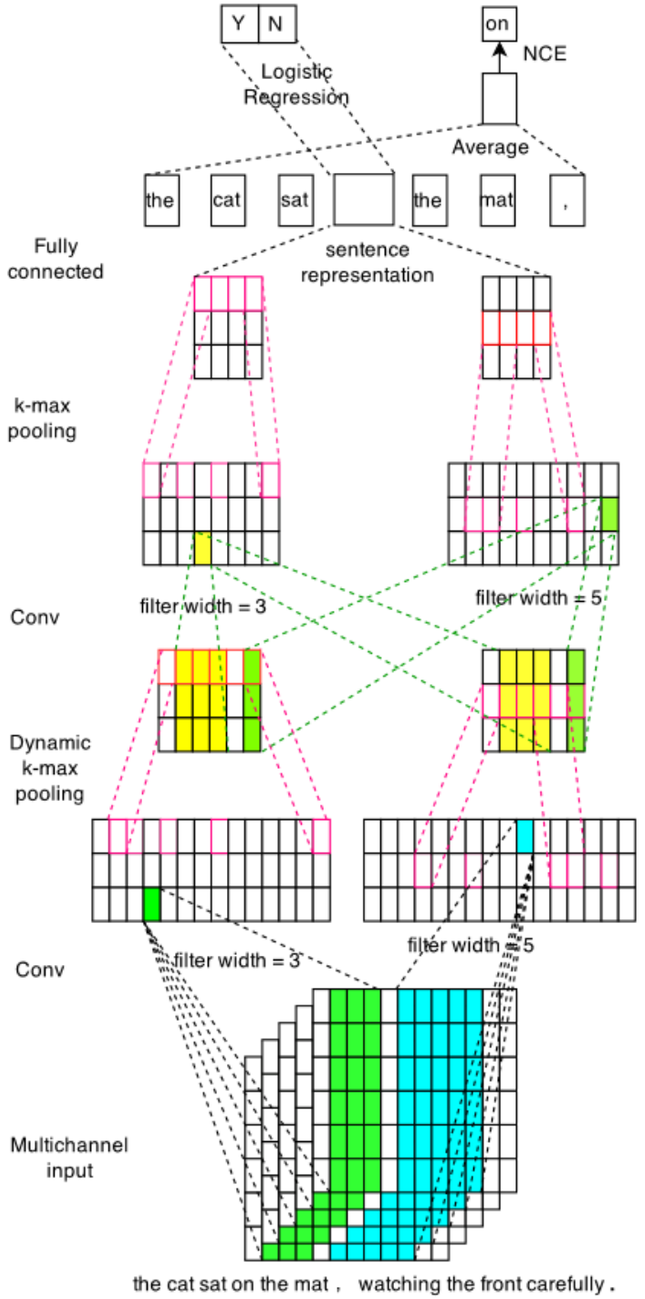
свёртка по всем картам

т.е. сумма свёрток со своими весами

Folding – сумма по 2 строчки

Figure 3: A DCNN for the seven word input sentence. Word embeddings have size $d = 4$. The network has two convolutional layers with two feature maps each. The widths of the filters at the two layers are respectively 3 and 2. The (dynamic) k-max pooling layers have values k of 5 and 3.





Multi Channel Variable size CNN: MV-CNN

Продолжение идеи…

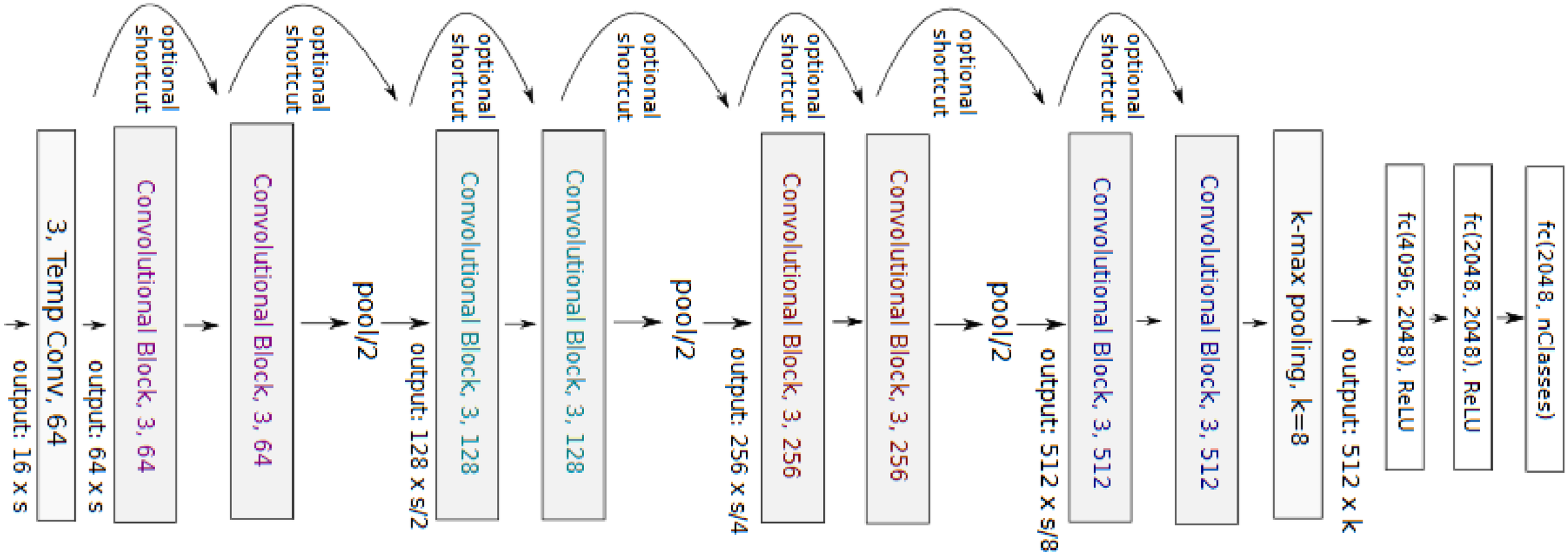
Сразу использовать несколько представлений:

- glove
- word2vec
- custom trained vectors

поэтому многоканальный вход

Wenpeng Yin, Hinrich Schütze «Multichannel Variable-Size Convolution for Sentence Classification» // https://arxiv.org/abs/1603.04513

Very Deep Convolutional Networks for Text Classification: VD-CNN



хороши в посимвольном случае (character-level)

маленькие свёртки

до 29 свёрточных слоёв

Very Deep Convolutional Networks for Text Classification: VD-CNN

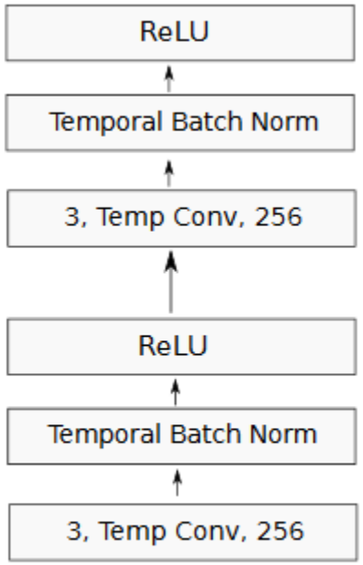


Figure 2: Convolutional block.

сделана по аналогии с VGG:

когда пространственный размер уменьшается в 2 раза

– число каналов×2

Alexis Conneau, Holger Schwenk, Loïc Barrault, Yann Lecun «Very Deep Convolutional Networks for Text Classification» // https://arxiv.org/abs/1606.01781

Сравнение CNN vs RNN

https://arxiv.org/pdf/1702.01923.pdf

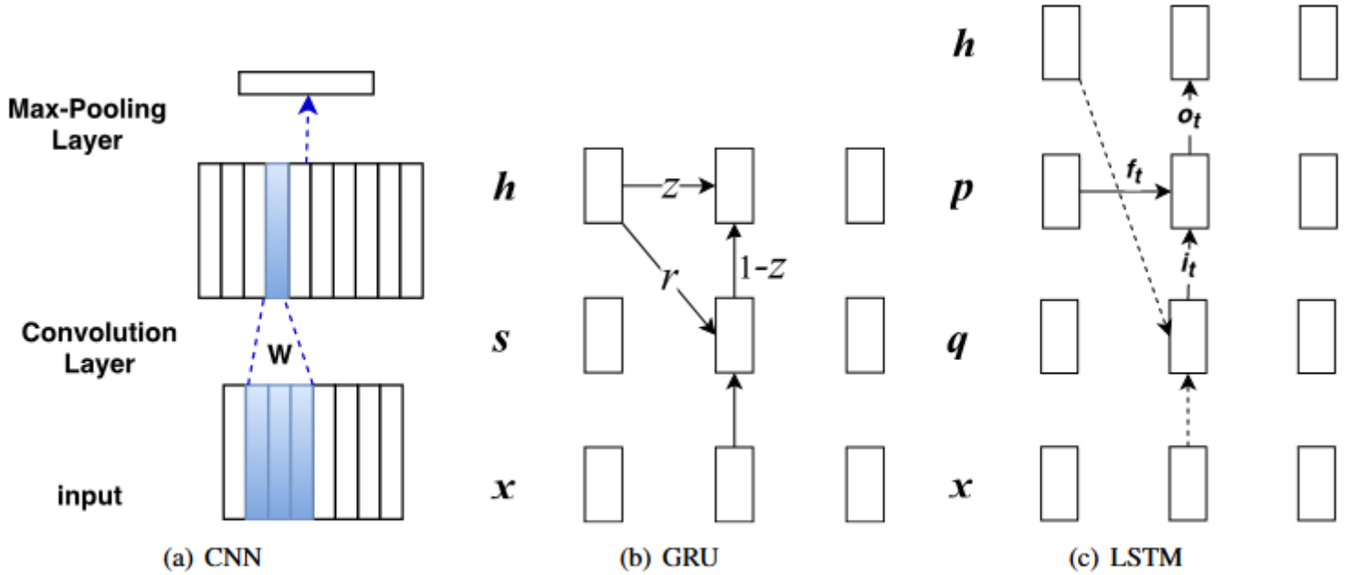


Figure 1: Three typical DNN architectures

Сравнение CNN vs RNN

задачи

- Sentiment Classification (SentiC)
- Relation Classification (RC)
- Textual Entailment (TE)
- Answer Selection (AS)
- Question Relation Match (QRM)
- Path Query Answering (PQA)
- Part-of-Speech Tagging

Сравнение CNN vs RNN: нет явного победителя!

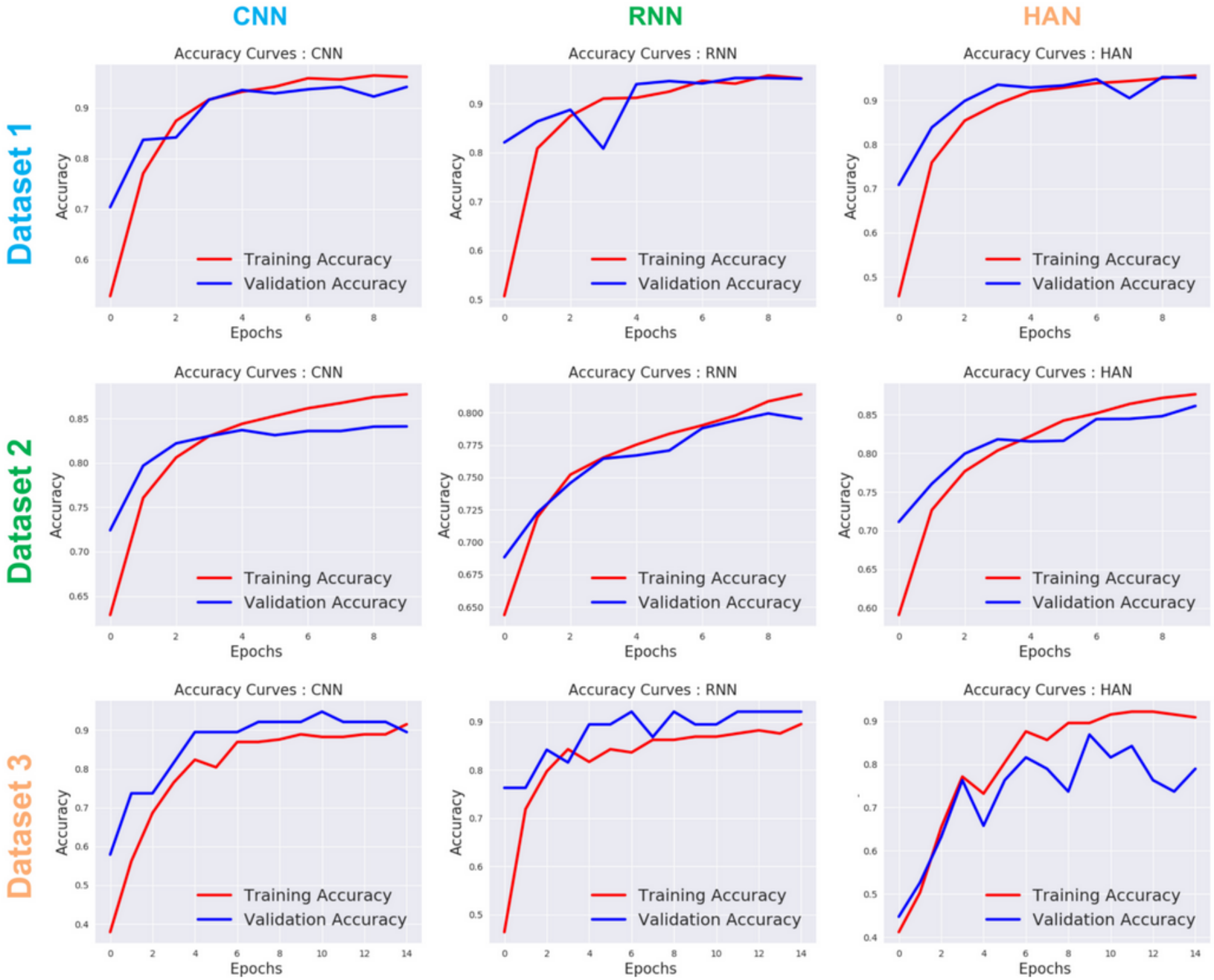
			performance	lr	hidden	batch	sentLen	filter_size	margin
TextC	SentiC (acc)	CNN	82.38	0.2	20	5	60	3	–
		GRU	**86.32**	0.1	30	50	60	–	–
		LSTM	84.51	0.2	20	40	60	–	–
	RC (F1)	CNN	68.02	0.12	70	10	20	3	–
		GRU	**68.56**	0.12	80	100	20	–	–
		LSTM	66.45	0.1	80	20	20	–	–
SemMatch	TE (acc)	CNN	77.13	0.1	70	50	50	3	–
		GRU	**78.78**	0.1	50	80	65	–	–
		LSTM	77.85	0.1	80	50	50	–	–
	AS (MAP & MRR)	CNN	**(63.69,65.01)**	0.01	30	60	40	3	0.3
		GRU	(62.58,63.59)	0.1	80	150	40	–	0.3
		LSTM	(62.00,63.26)	0.1	60	150	45	–	0.1
	QRM (acc)	CNN	**71.50**	0.125	400	50	17	5	0.01
		GRU	69.80	1.0	400	50	17	-	0.01
		LSTM	71.44	1.0	200	50	17	-	0.01
SeqOrder	PQA (hit@10)	CNN	54.42	0.01	250	50	5	3	0.4
		GRU	**55.67**	0.1	250	50	5	–	0.3
		LSTM	55.39	0.1	300	50	5	–	0.3
ContextDep	POS tagging (acc)	CNN	94.18	0.1	100	10	60	5	–
		GRU	93.15	0.1	50	50	60	–	–
		LSTM	93.18	0.1	200	70	60	–	–
		Bi-GRU	94.26	0.1	50	50	60	–	–
		Bi-LSTM	**94.35**	0.1	150	5	60	–	–

Table 1: Best results or CNN, GRU and LSTM in NLP tasks

Сравнение CNN vs RNN vs HAN

HAN – Hierarchical Attention Network

https://medium.com/jatana/report-on-text-classification-using-cnn-rnn-han-f0e887214d5f


CNN
RNN
HAN
Dataset 1
Dataset 2
Dataset 3
Accuracy Curves : CNN
Accuracy Curves : RNN
Accuracy Curves : HAN
Accuracy
Epochs
Training Accuracy
Validation Accuracy

CNN + LSTM = C-LSTM

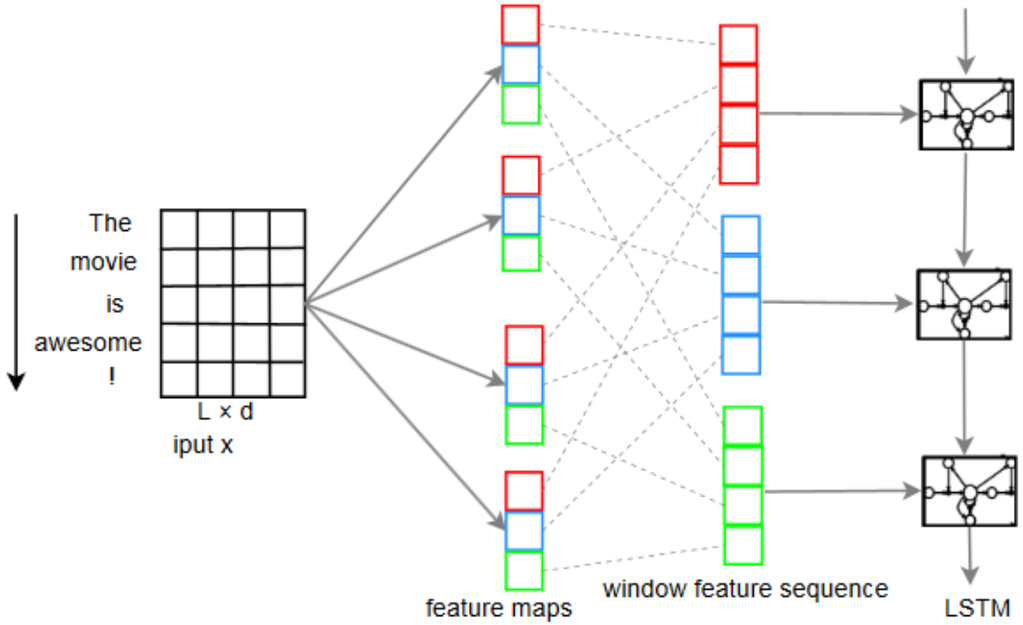


Figure 1: The architecture of C-LSTM for sentence modeling. Blocks of the same color in the feature map layer and window feature sequence layer corresponds to features for the same window. The dashed lines connect the feature of a window with the source feature map. The final output of the entire model is the last hidden unit of LSTM.

CNN – получение высокоуровневых признаков из представлений слов

LSTM – для анализа зависимостей

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, Francis C.M. Lau «A C-LSTM Neural Network for Text Classification»
https://arxiv.org/abs/1511.08630

CNN + LSTM + CRF = LSTM-CNNs-CRF

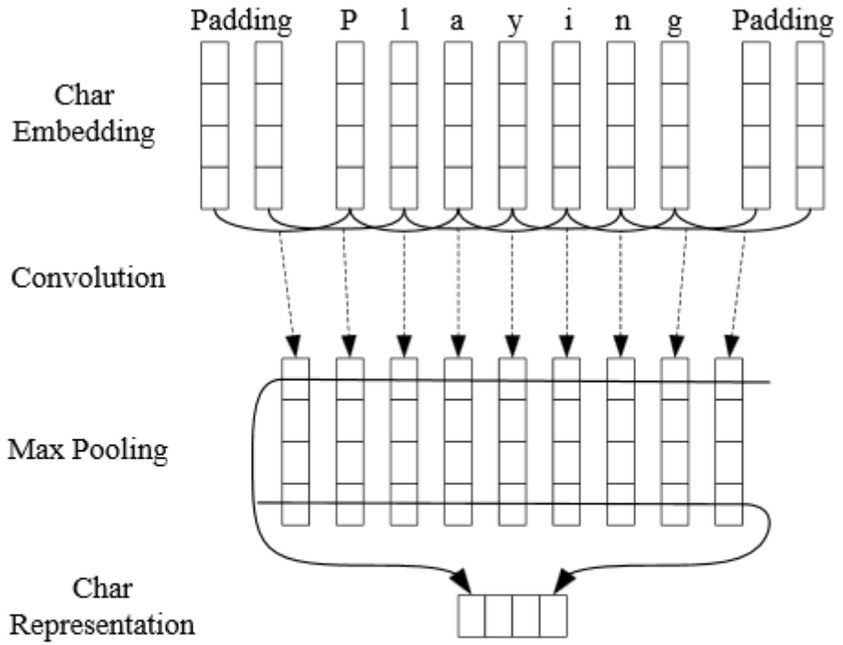
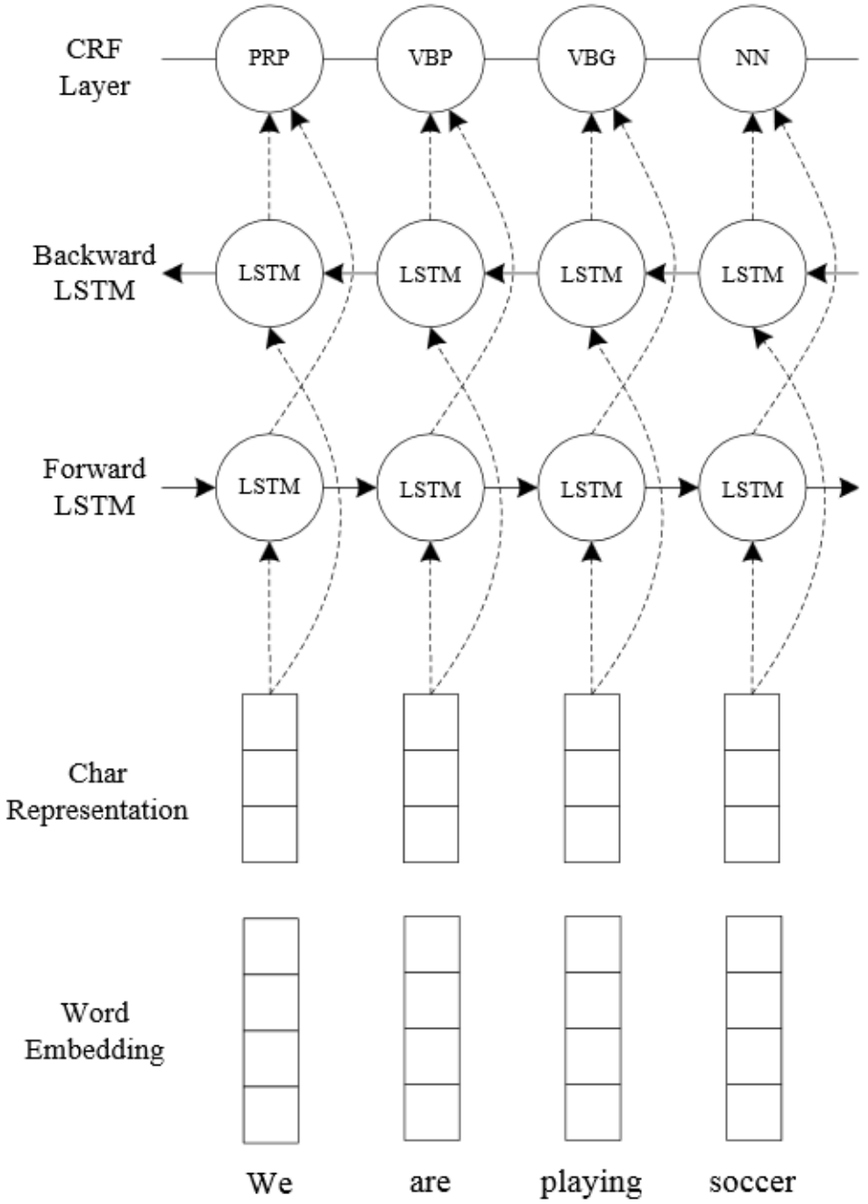


Figure 1: The convolution neural network for extracting character-level representations of words. Dashed arrows indicate a dropout layer applied before character embeddings are input to CNN.



Dashed arrows indicate dropoutlayers applied on both the input and output vectorsof BLSTM.

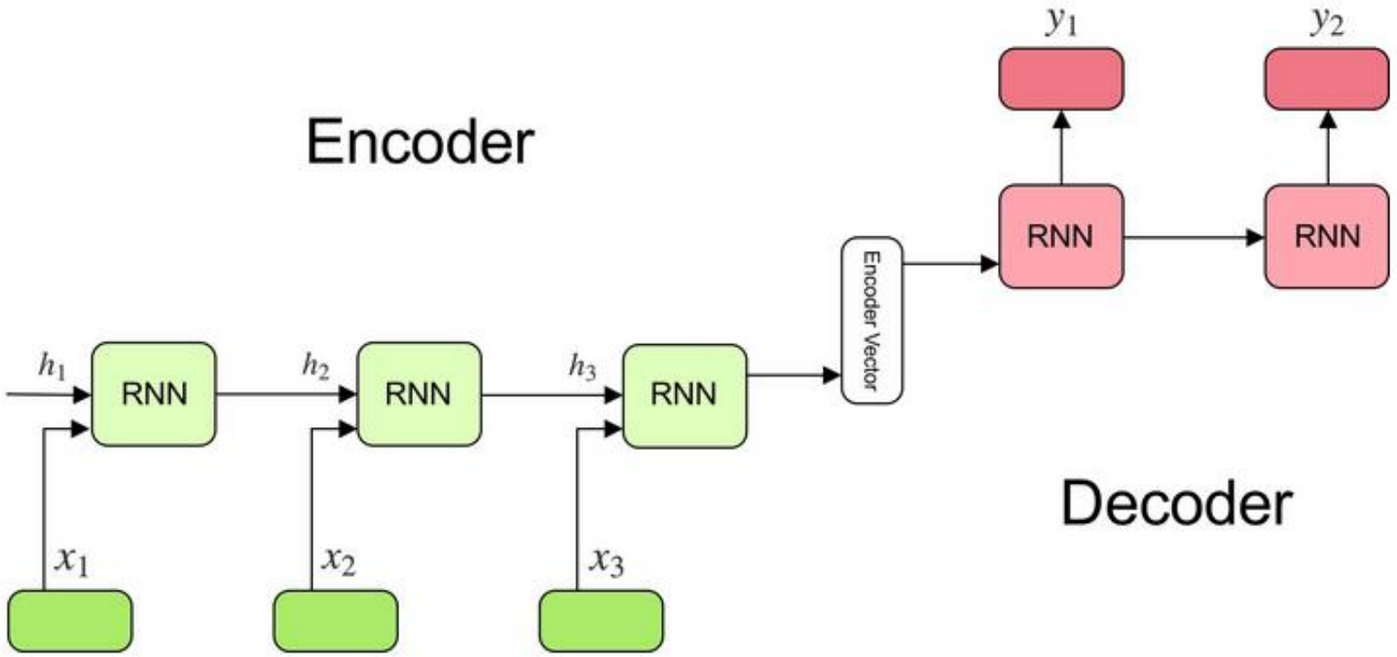
CNN + LSTM + CRF = LSTM-CNNs-CRF

	POS		NER					
	Dev	Test	Dev			Test		
Model	Acc.	Acc.	Prec.	Recall	F1	Prec.	Recall	F1
BRNN	96.56	96.76	92.04	89.13	90.56	87.05	83.88	85.44
BLSTM	96.88	96.93	92.31	90.85	91.57	87.77	86.23	87.00
BLSTM-CNN	97.34	97.33	92.52	93.64	93.07	88.53	90.21	89.36
BRNN-CNN-CRF	97.46	97.55	94.85	94.63	94.74	91.35	91.06	91.21

Table 3: Performance of our model on both the development and test sets of the two tasks, together with three baseline systems.

Xuezhe MaandEduard Hovy «End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF» // https://arxiv.org/pdf/1603.01354.pdf

Модель seq2seq



http://www.davidsbatista.net/blog/2020/01/25/Attention-seq2seq/

Sutskever I. «Sequence to Sequence Learning with Neural Networks», 2014 // https://arxiv.org/abs/1409.3215

Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

Kyunghyun Cho et. al. «Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation» https://www.aclweb.org/anthology/D14-1179/

Модель seq2seq

Как переводить последовательность → последовательность

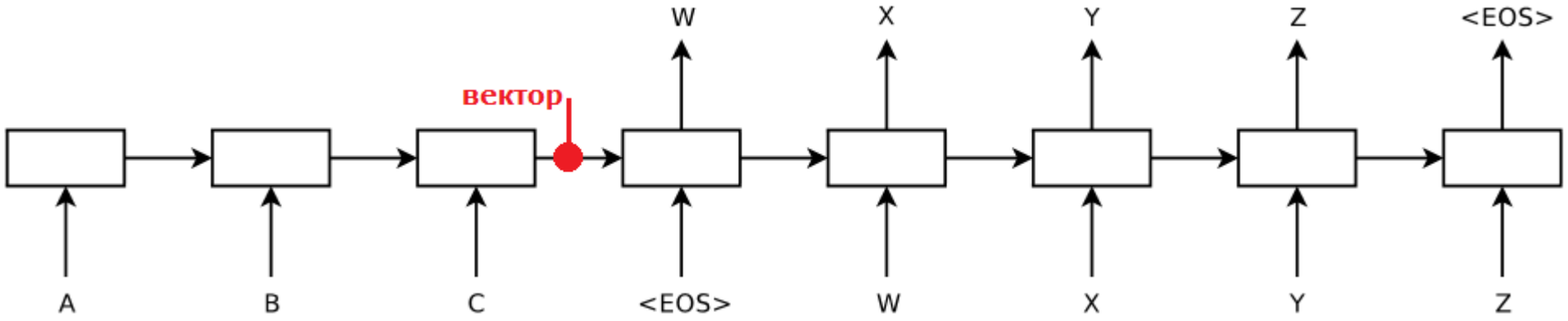
Многослойная (4 слоя) LSTM:

последовательность → вектор

Другая (так, понятно, лучше!) многослойная LSTM:

вектор → целевая последовательность

Интересно: в задаче перевода качество повышало инвертирование порядка входа!

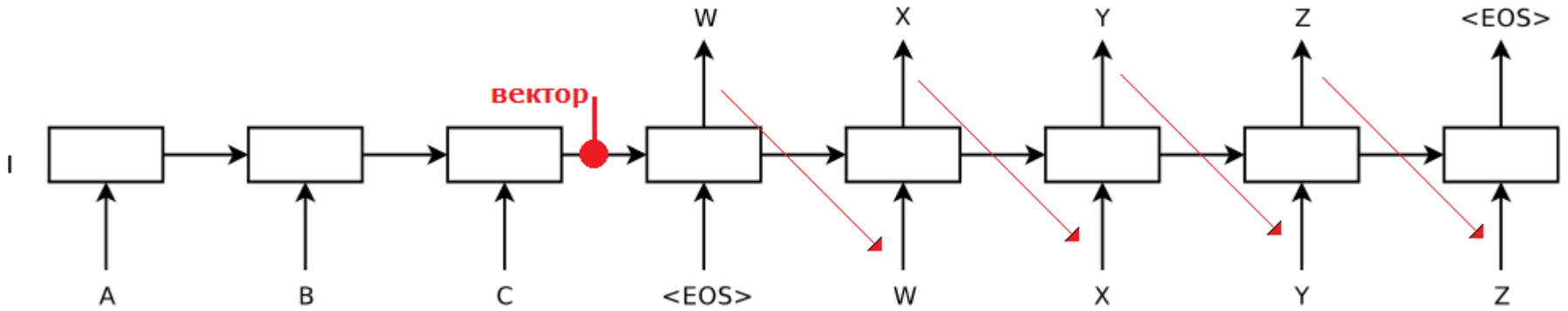


Модель seq2seq

кодировщик (encoder) – декодировщик (decoder)

у них разные параметры!

здесь декодировщик называют также языковой моделью



при работе (inference) – подаём на вход сгенерированное

при обучении – среднее ошибок на всех выходах

(ex negative log prob)

Модель seq2seq

тонкости:

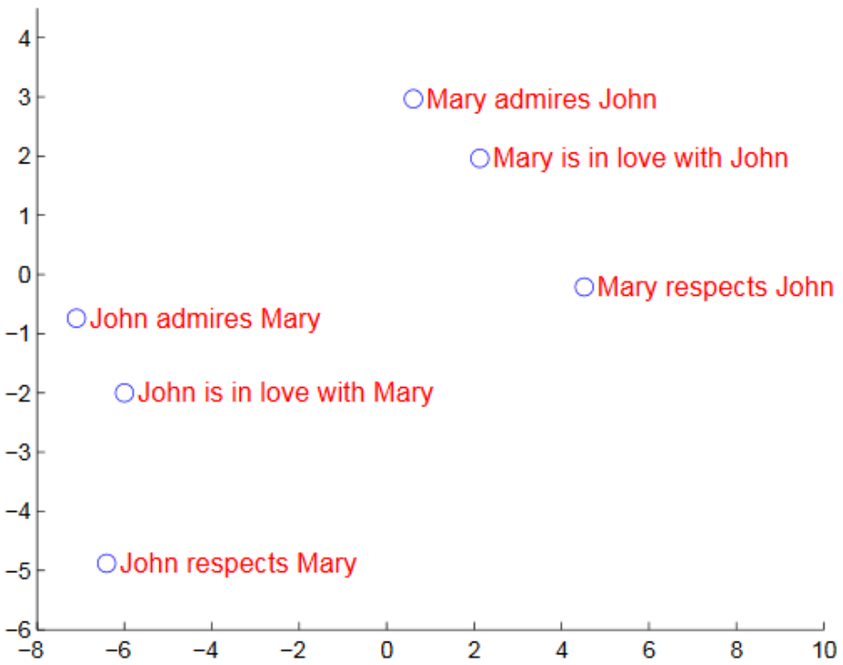
можно декодировщику передавать представление закодируемого предложения

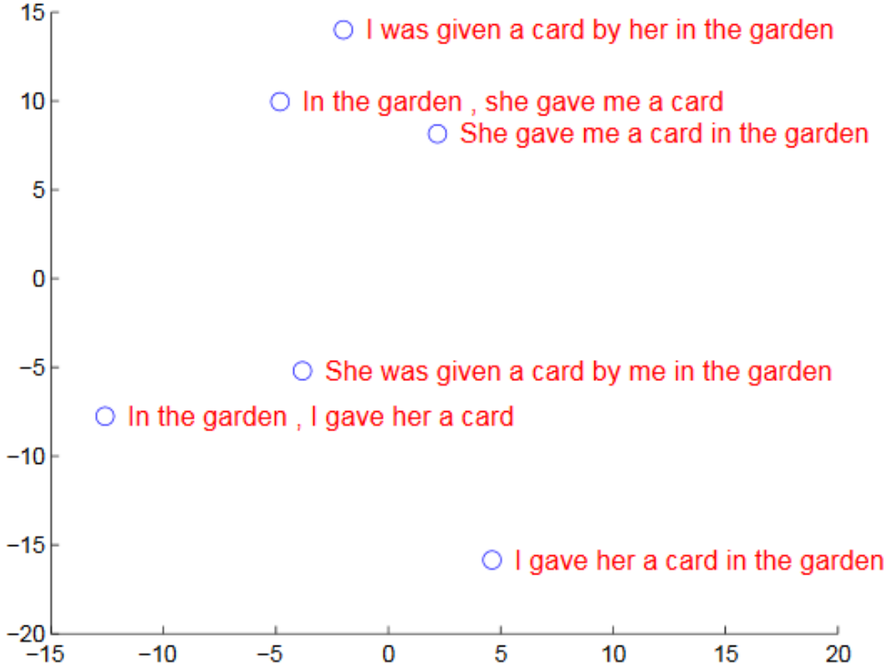
оно переходит в каждый нейрон,

кроме него отдельно переходит изменяясь внутренне состояние декодировщика

Модель seq2seq

Внутреннее представление предложений!





left-to-right beam-search decode

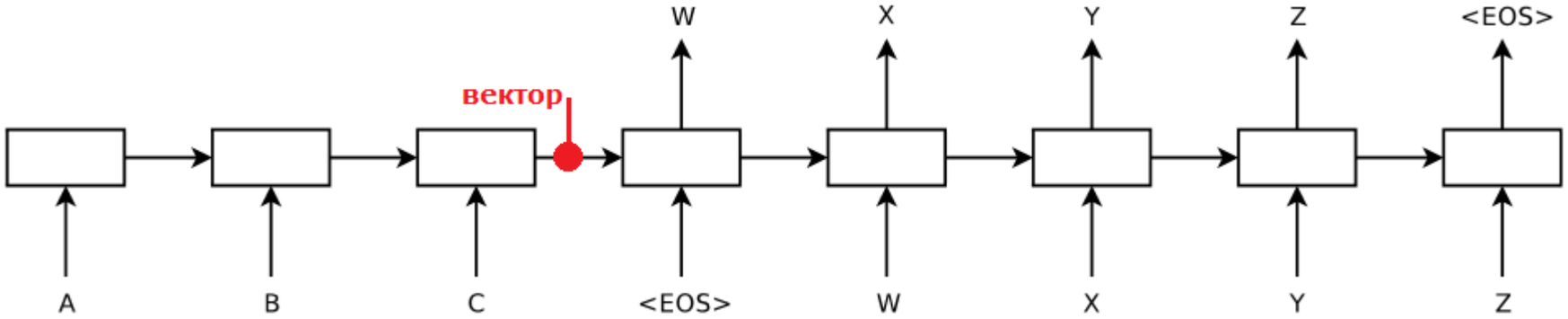
если выбираем лучшего следующего, не обязательно максимизируем качество

Обучение 10 дней

Тоже хороши ансамбли

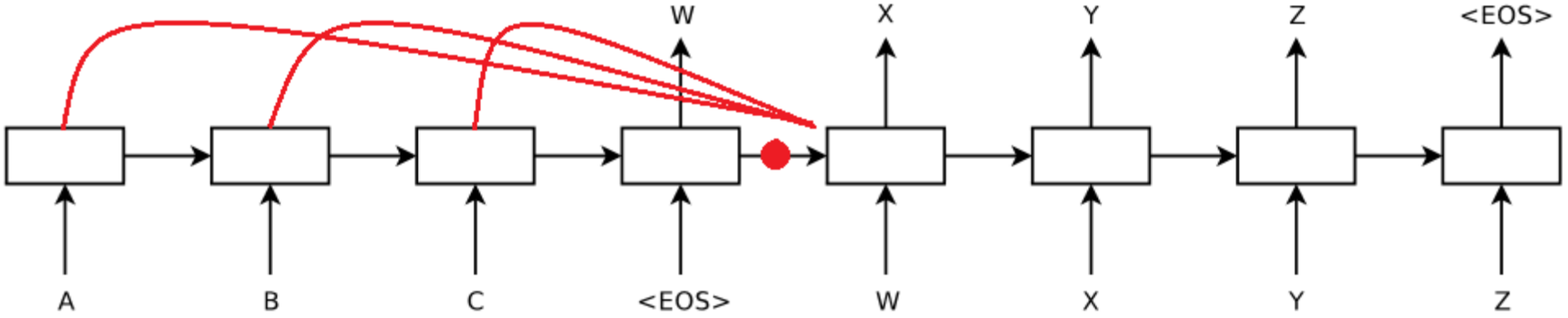
Обобщения seq2seq

На одном нейроне вся информация о тексте… плохо
(особенно для длинных последовательностей)



Решение – механизм внимания

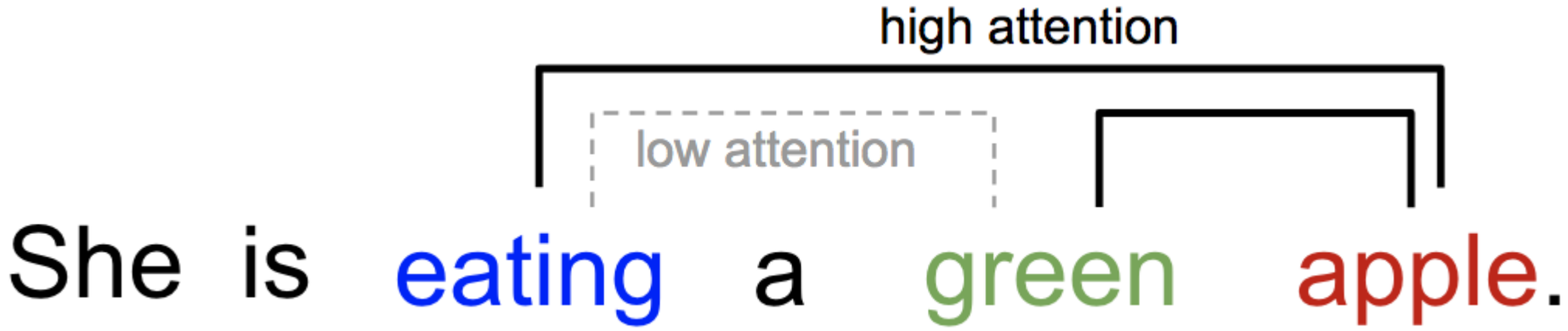
частично был в RNN



Bahdanau et al. 2015 «Neural Machine Translation by Jointly Learning to Align and Translate» // ICLR 2015 https://arxiv.org/pdf/1409.0473.pdf

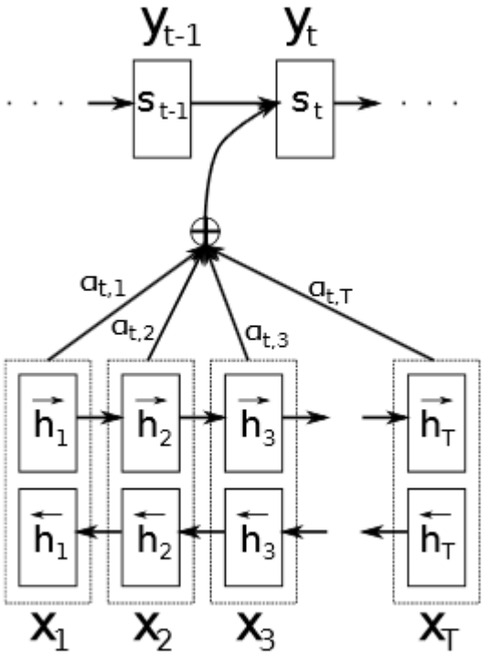
Механизм внимания

Концепция: есть взаимосвязи между словами



https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html#born-for-translation

Механизм внимания



Не будем пытаться закодировать всё предложение одним вектором!

Добавляется контекстный вектор (конкатенируется)

$$c_i = \sum_j \alpha_{ij} h_j$$

веса (softmax)

$$\alpha_{ij} = \exp(e_{ij}) / \sum_k \exp(e_{ik})$$

Насколько соответствуют состояния

$$e_{ij} = a(s_{i-1}, h_j)$$

Учитываются не только слова ДО, но и ПОСЛЕ!

Конкатенация состояния ДО и состояния ПОСЛЕ

Bidirectional RNN (BiRNN)

Механизм внимания

соответствие $e_{ij} = a(s_{i-1}, h_j)$ **может быть:**

Basic dot-product	$a(s,h) = s^{\mathrm{T}} h$
Multiplicative attention	$a(s,h) = s^{\mathrm{T}} W h$
Additive attention	$a(s,h) = w^{\mathrm{T}} \tanh(W_1 s - W_2 h)$

+ разные нормировки по размерности

Thang Luong, Hieu Pham, Christopher D. Manning «Effective Approaches to Attention-based Neural Machine Translation» https://www.aclweb.org/anthology/D15-1166.pdf

Механизм внимания: получаем интерпретацию и выравнивание (alignment)

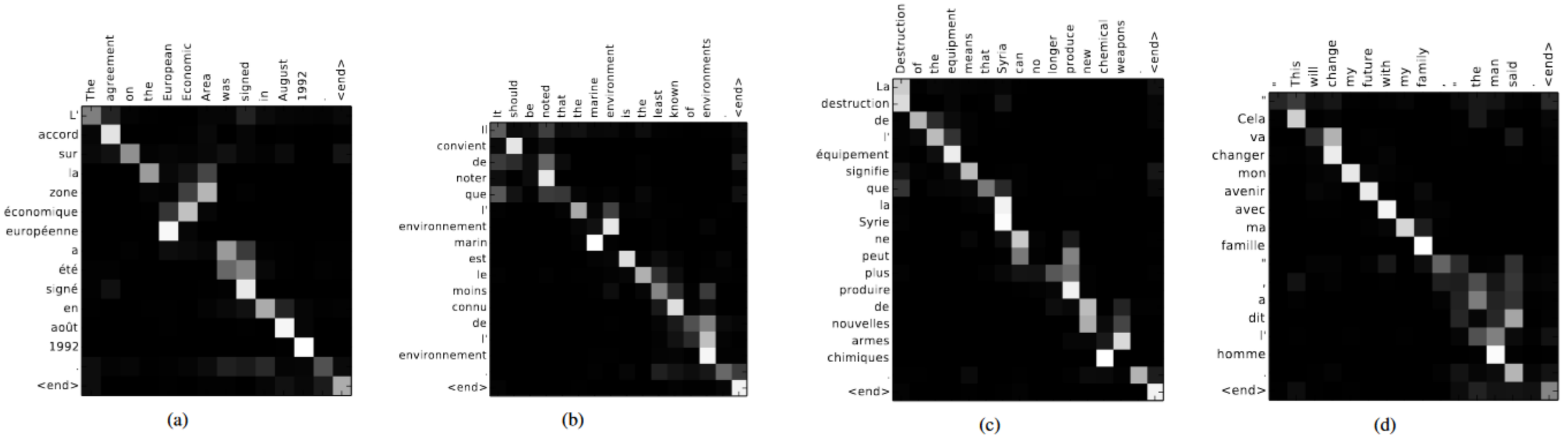


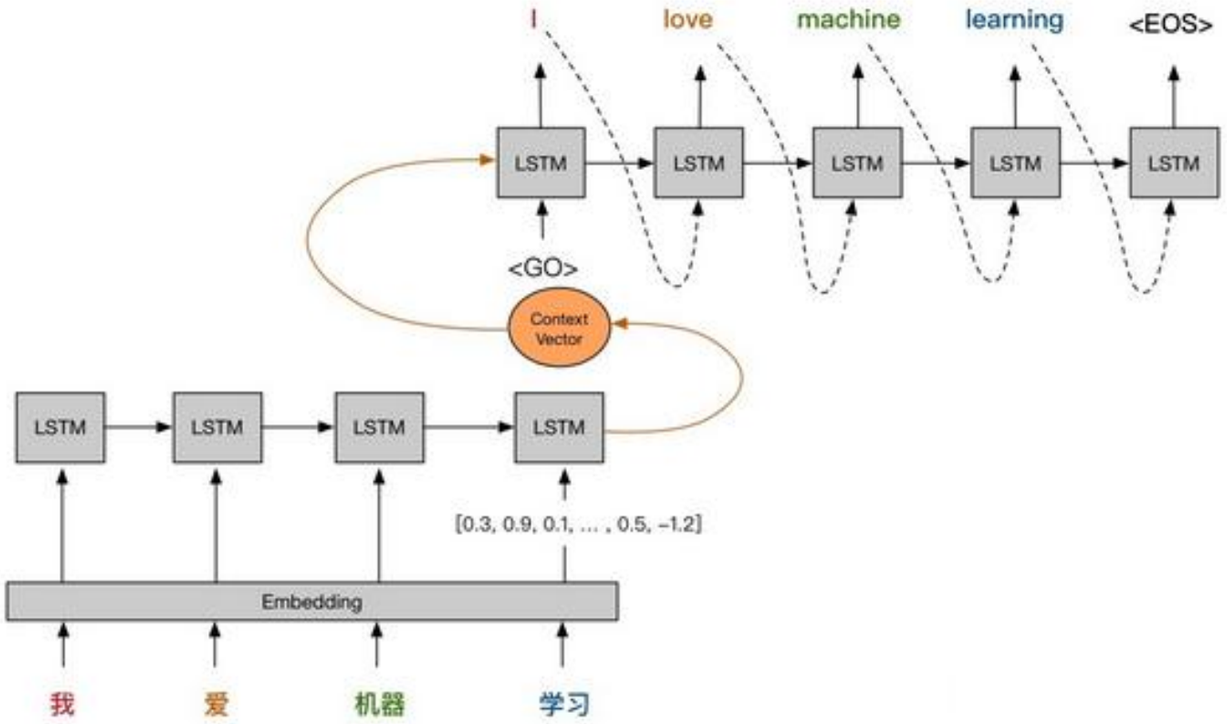
Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight α_{ij} of the annotation of the j-th source word for the i-th target word (see Eq. (6)), in grayscale (0: black, 1: white). (a) an arbitrary sentence. (b–d) three randomly selected samples among the sentences without any unknown words and of length between 10 and 20 words from the test set.

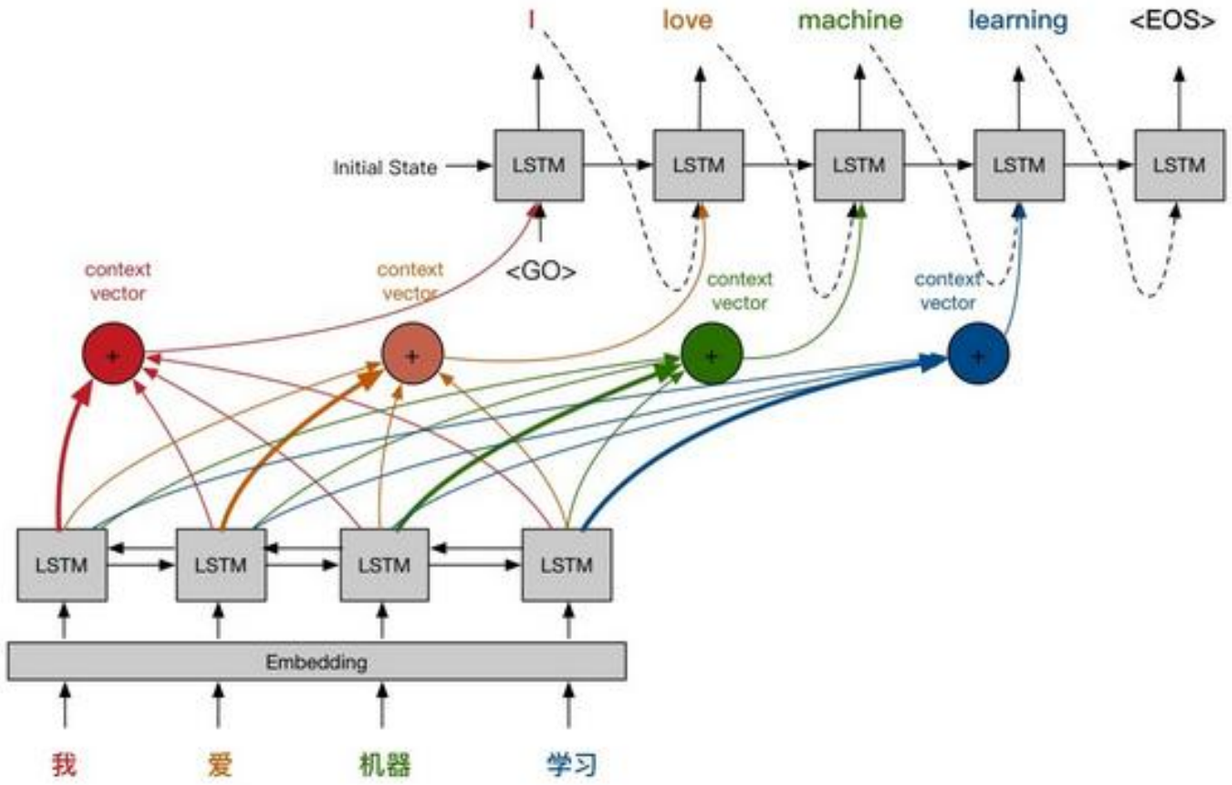
«Neural Machine Translation by Jointly Learning to Align and Translate»
[Bahdanau D. и др., 2016 https://arxiv.org/abs/1409.0473]

Механизм внимания (Attention)

- улучшает качество перевода
- решает проблему «узкого горла»
- ~ интерпретируемость
- решает проблему исчезающего градиента
- получаем выравнивание (alignment) «бесплатно» в переводе

seq2seq vs attention





Внимание – техника вычисления взвешенной суммы значений (values) по запросу (query)

~ техника получения описания (representation) фиксированного размера по запросу

https://zhuanlan.zhihu.com/p/37290775

Плюсы внимания

Улучшает качество, например, перевода

Решает проблему «узкого горла»

Решает проблему «затухания градиента»

Вносит интерпретируемость в модель

Виды внимания

Self-Attention / intra-attention	к разным позициям одной и той же входной последовательности
Global / Soft	ко всему входу
Local / Hard	к части входа

Виды внимания: Self-Attention

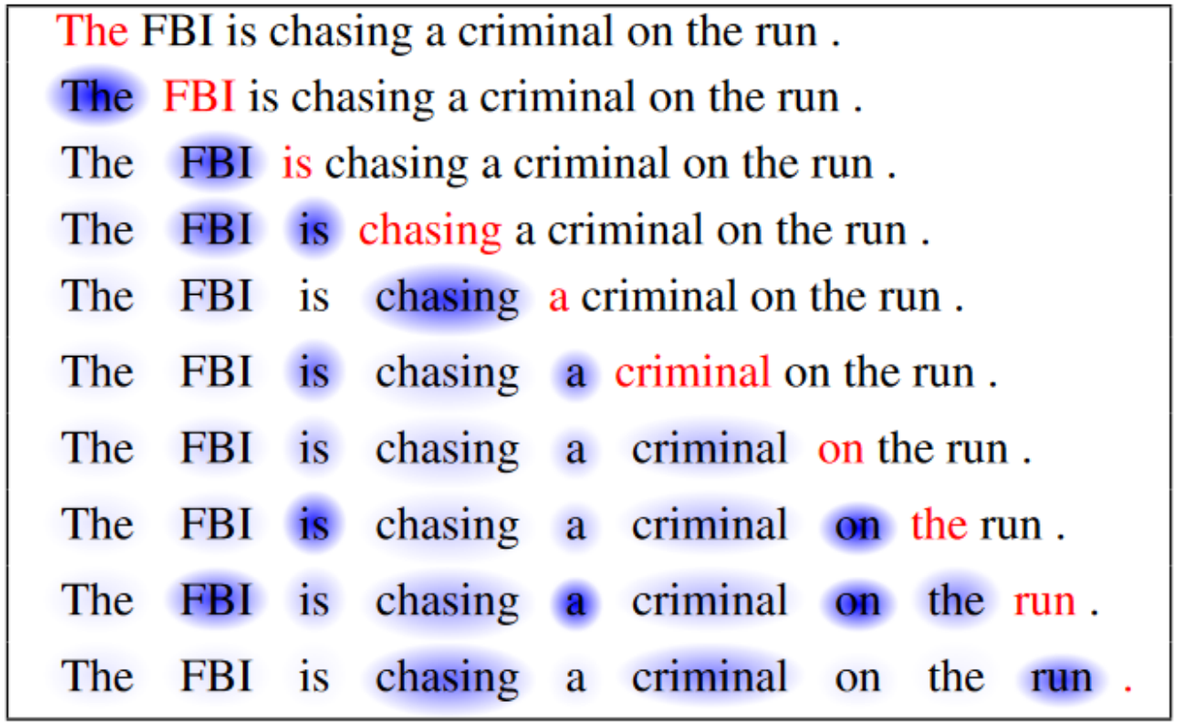
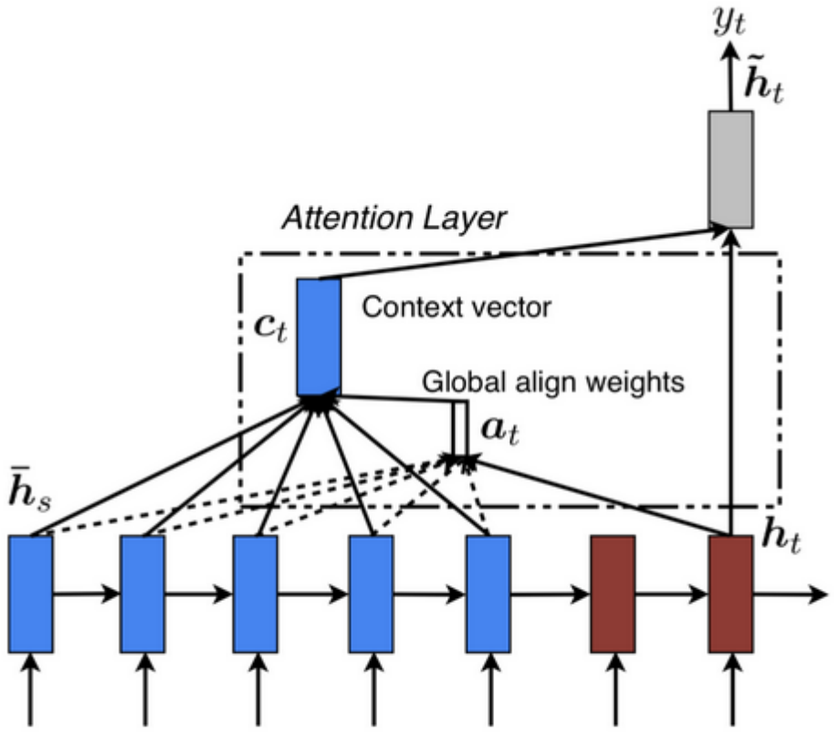


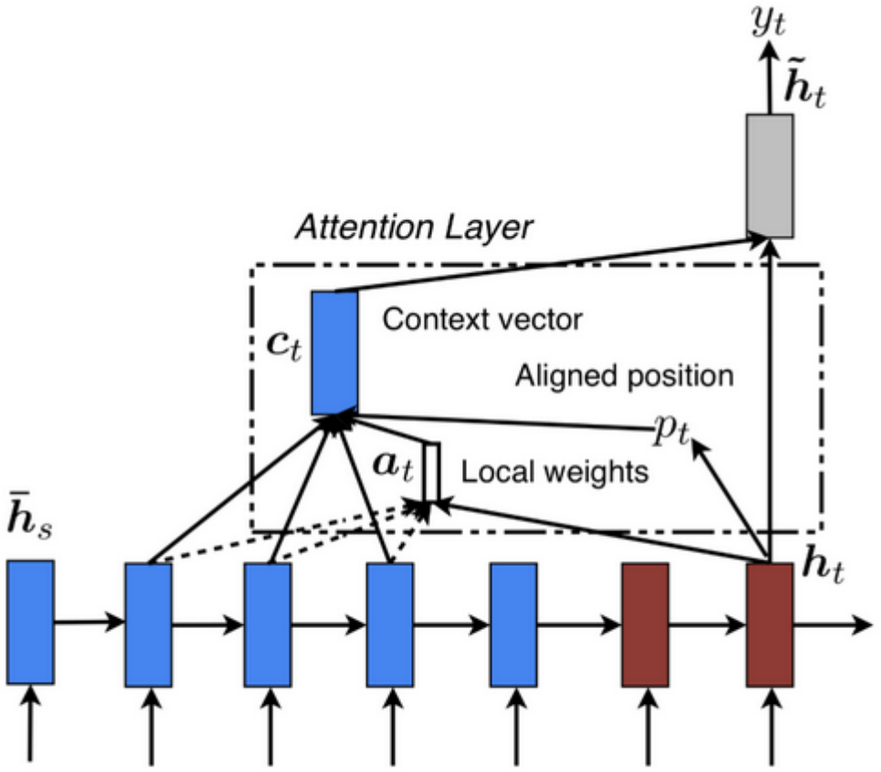
Figure 1: Illustration of our model while reading the sentence *The FBI is chasing a criminal on the run.* Color *red* represents the current word being fixated, *blue* represents memories. Shading indicates the degree of memory activation.

https://arxiv.org/pdf/1601.06733.pdf

Виды внимания: Global vs Local Attention



Global Attention Model



Local Attention Model

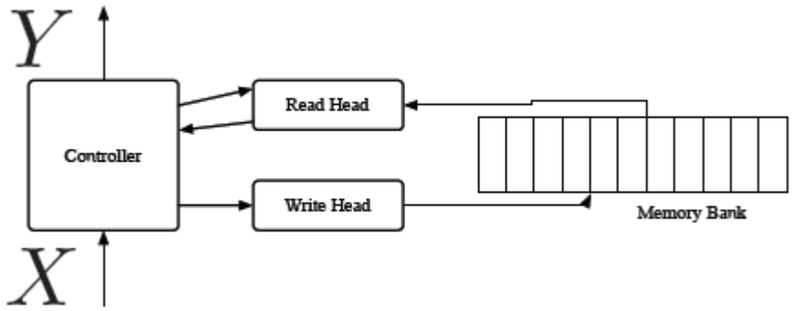
https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html#born-for-translation

Механизмы внимания

- **Neural Turing Machines (Graves et al 2014)**

  - **Memory Networks (Weston et al 2014)**
    - **Fully Supervised MemNNs**
    - **End2End MemNNs**
    - **Key-Value MemNNs**
    - **Dynamic MemNNs**

- **Content-based attention mechanism (Bahdanau et al 2014) to control the read and write access into a memory**



Дифференцируемые структуры памяти (Differentiable Memory structures)

– **LSTM** [Hochreiter & Schmidhuber]

– **Tapes** [NTM, Graves et al'14]

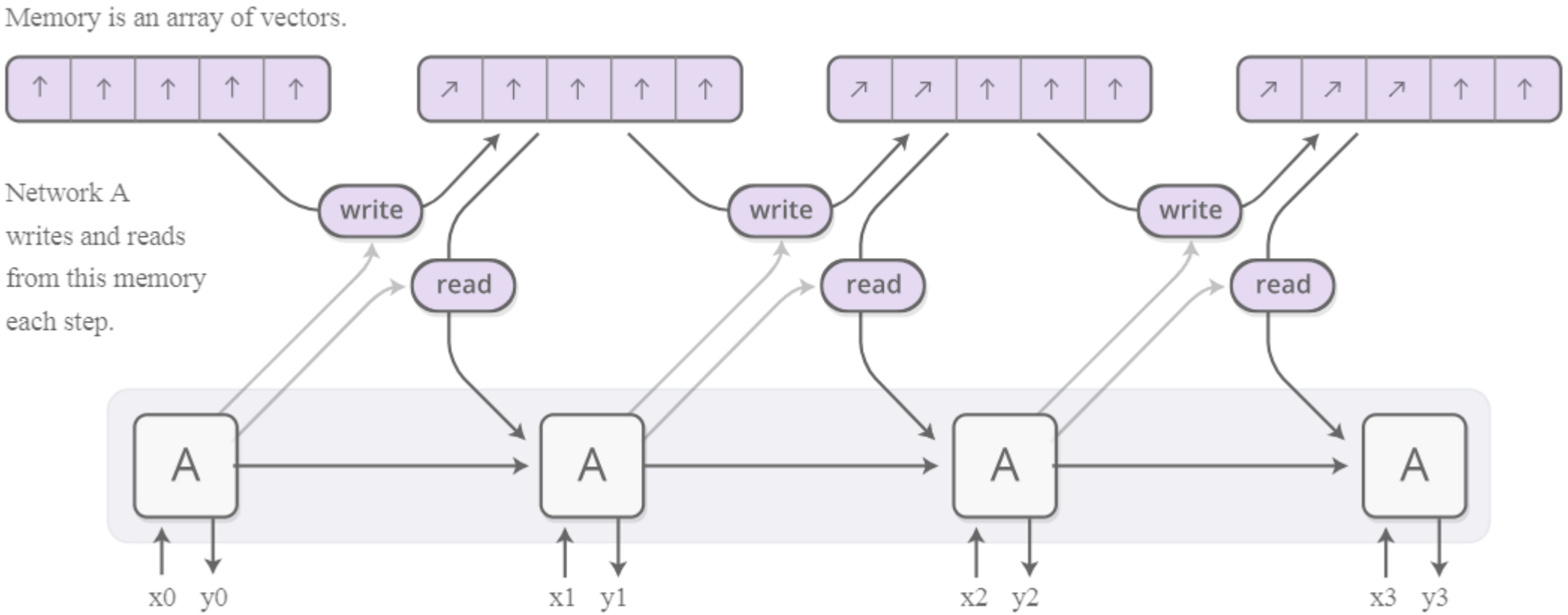
– **Arrays** [Memory Nets, Weston et al'14]

– **Stacks** [Joulin & Mikolov'15]

Важна дифференцируемость для обучения…

Neural Turing Machines

нейросеть + модель внешней памяти (конечной, увы!)

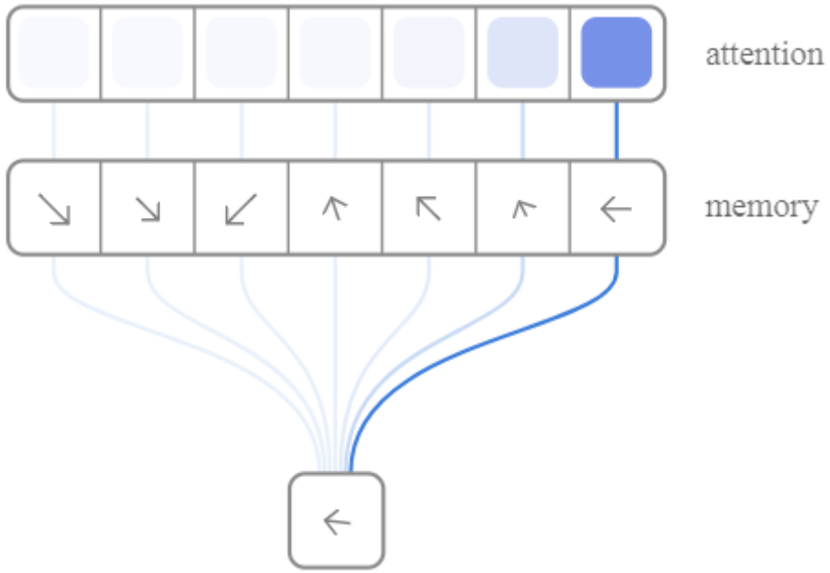


https://distill.pub/2016/augmented-rnns/#neural-turing-machines

A. Graves, G. Wayne, I. Danihelka «Neural Turing Machines», 2014 // https://arxiv.org/abs/1410.5401

Neural Turing Machines

читаем взвешенную сумму памяти
это нужно, в том числе, чтобы всё было дифференцируемо



$$r = \sum_i a_i M_i$$

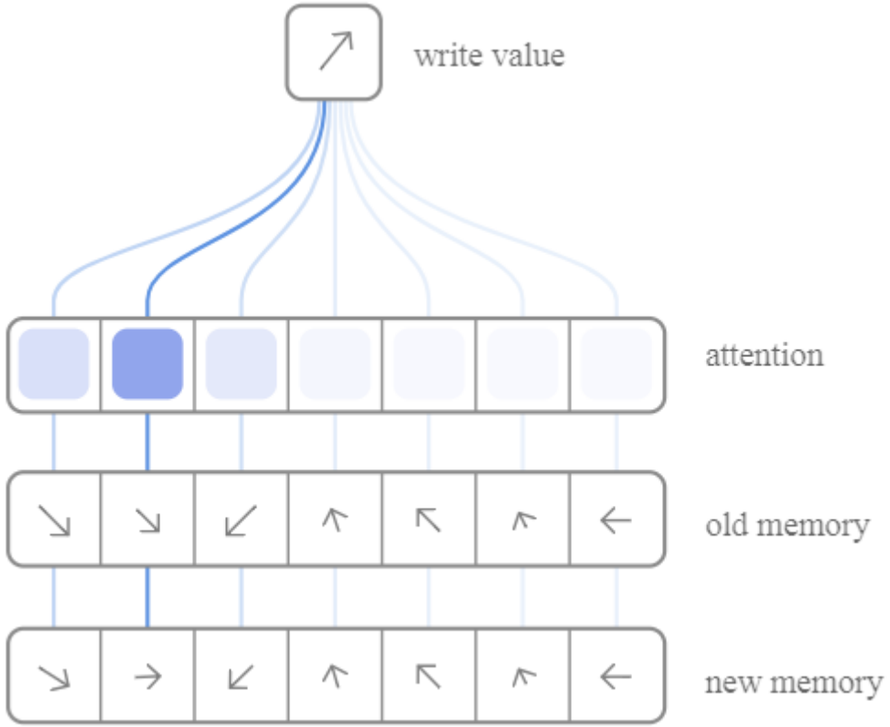
Коэффициенты регулируются с помощью «attention»

«Soft-attention reading»

Память это матрица!

Neural Turing Machines

Аналогично пишем в память



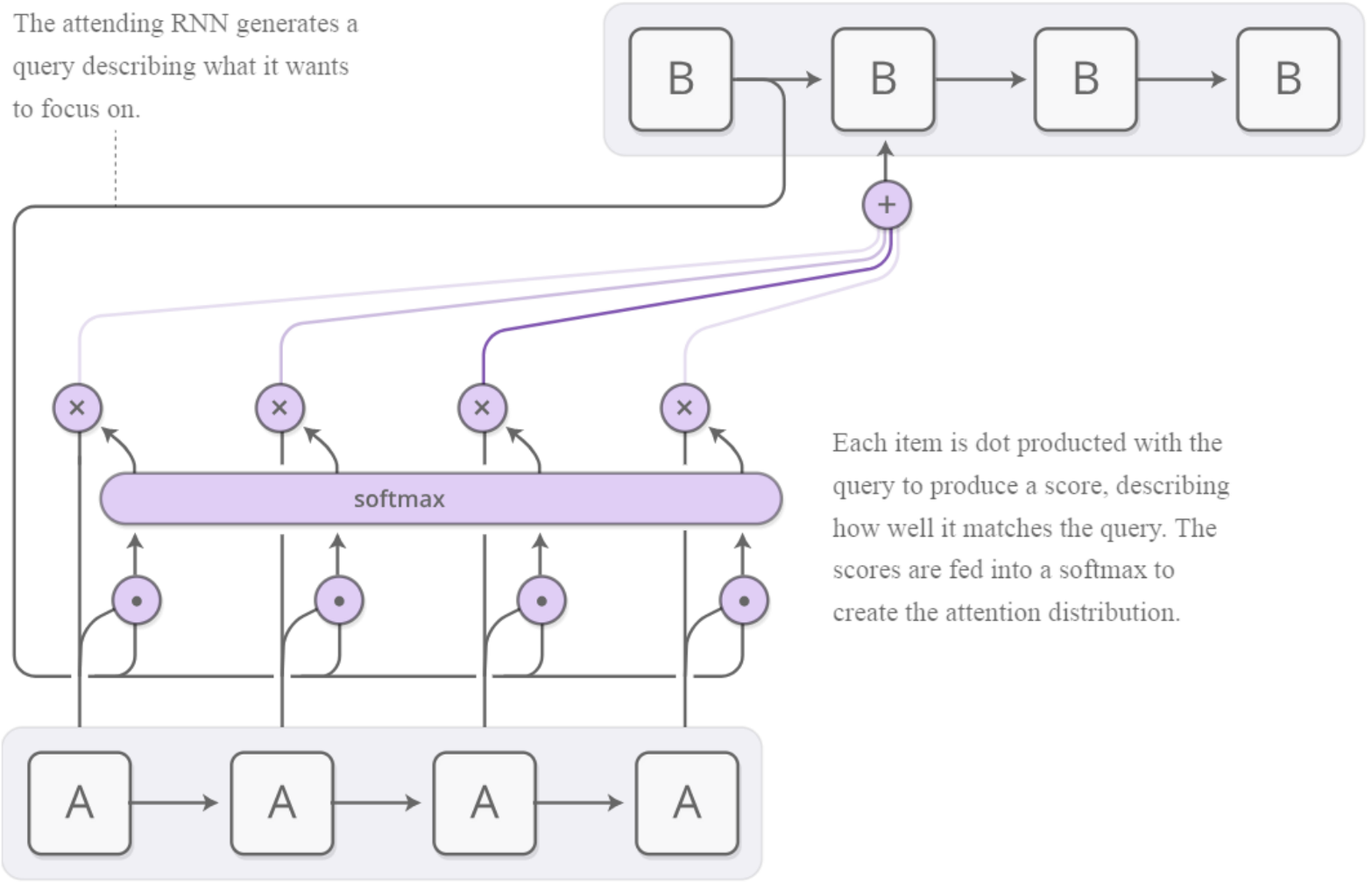
Пишем в каждую ячейку, но в «каком количестве» зависит от «attention»

$$M_i \equiv a_i w + (1 - a_i) M_i$$

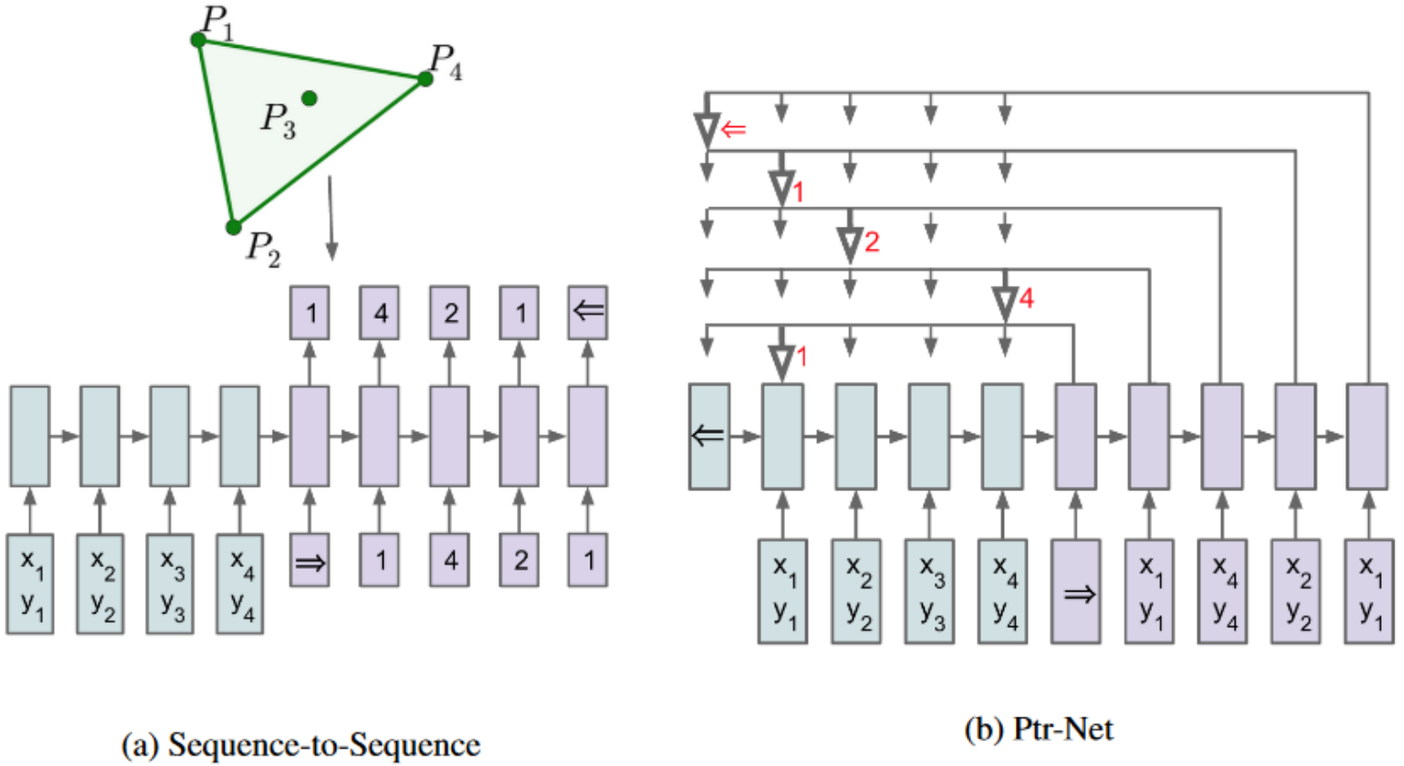
Внимание: Attentional Interfaces

опять же смотрим сразу на все выходы другой RNN

специальная сеть указывает, на чём фокусироваться



Внимание: Pointer Network



(a) Sequence-to-Sequence

(b) Ptr-Net

Figure 1: **(a)** *Sequence-to-Sequence* - An RNN (blue) processes the input sequence to create a code vector that is used to generate the output sequence (purple) using the probability chain rule and another RNN. The output dimensionality is fixed by the dimensionality of the problem and it is the same during training and inference [1]. **(b)** *Ptr-Net* - An encoding RNN converts the input sequence to a code (blue) that is fed to the generating network (purple). At each step, the generating network produces a vector that modulates a content-based attention mechanism over inputs ([5, 2]). The output of the attention mechanism is a softmax distribution with dictionary size equal to the length of the input.

Oriol Vinyals, Meire Fortunato, Navdeep Jaitly «Pointer Networks» // https://arxiv.org/abs/1506.03134

Внимание: Pointer Network

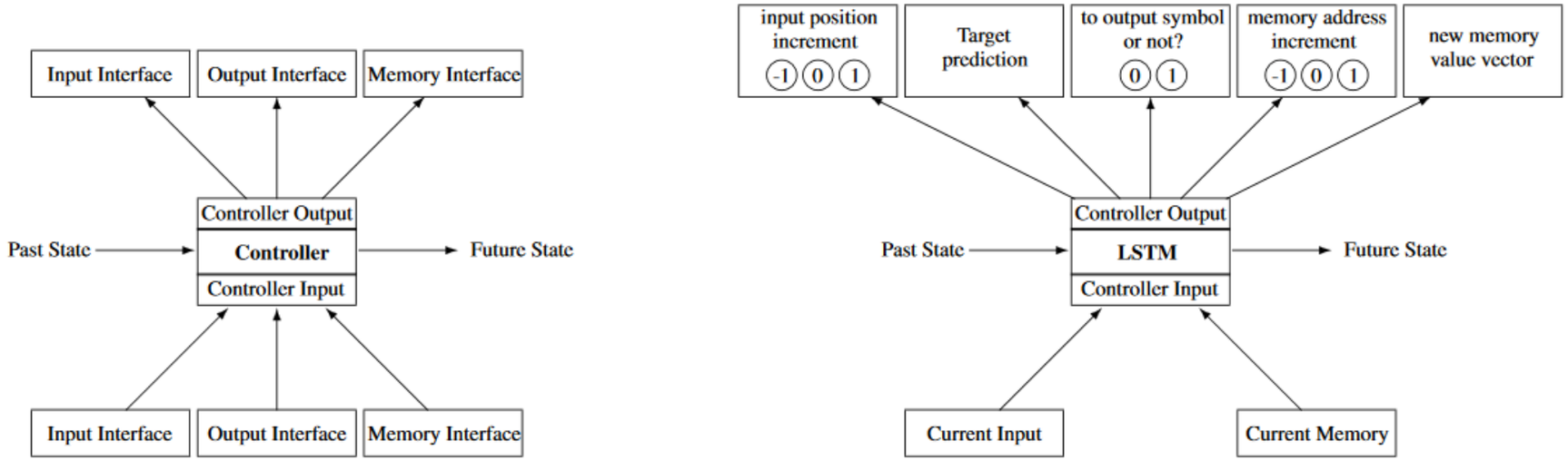
во многих комбинаторных задачах (коммивояжёр) вход и выход – последовательности

Здесь придуман способ её генерации…
выход – это соответствующий номер элемента кодировщика

в принципе, это не совсем внимание, но идея та же

Discrete Read/Write: Reinforcement Learning Neural Turing Machines

RL для обучения сети, которая взаимодействует с дискретными структурами



An abstract Interface–Controller model

Our model as an Interface–Controller

Figure 1: **(Left)** The Interface–Controller abstraction, **(Right)** an instantiation of our model as an Interface–Controller. The bottom boxes are the read methods, and the top are the write methods. The RL–NTM makes discrete decisions regarding the move over the input tape, the memory tape, and whether to make a prediction at a given timestep. During training, the model's prediction is compared with the desired output, and is used to train the model when the RL-NTM chooses to advance its position on the output tape; otherwise it is ignored. The memory value vector is a vector of content that is stored in the memory cell.

Wojciech Zaremba, Ilya Sutskever «Reinforcement Learning Neural Turing Machines - Revised», 2016 // https://arxiv.org/abs/1505.00521

Discrete Read/Write: Reinforcement Learning Neural Turing Machines

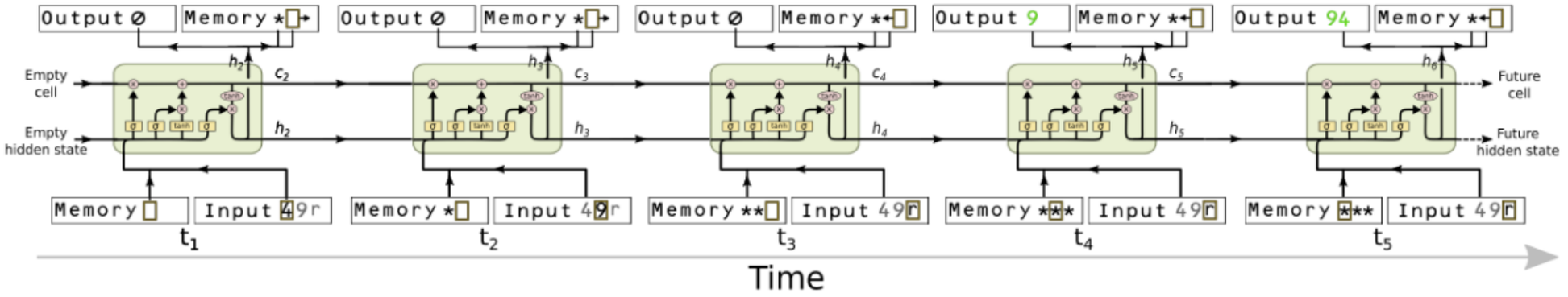
Вход LSTM:

- **вход (символ/ы с ленты)**
- **текущая ячейка/и памяти**
- **текущая память ячейки**
- **представление всех предыдущих действий**

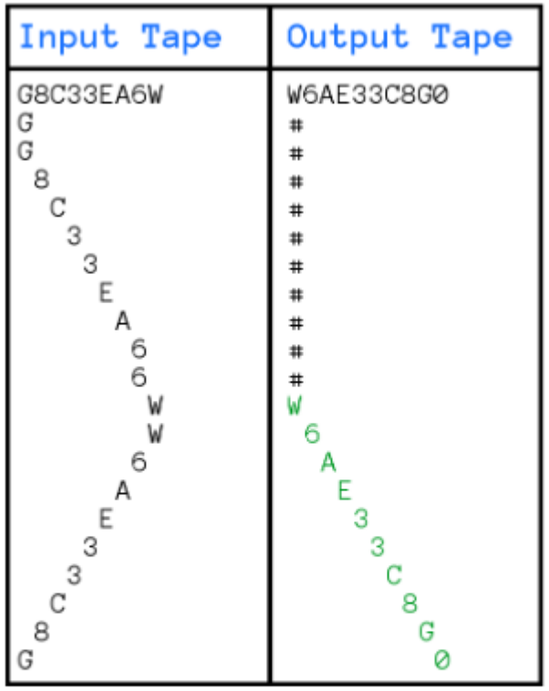
  (не изображено)

Выход LSTM:

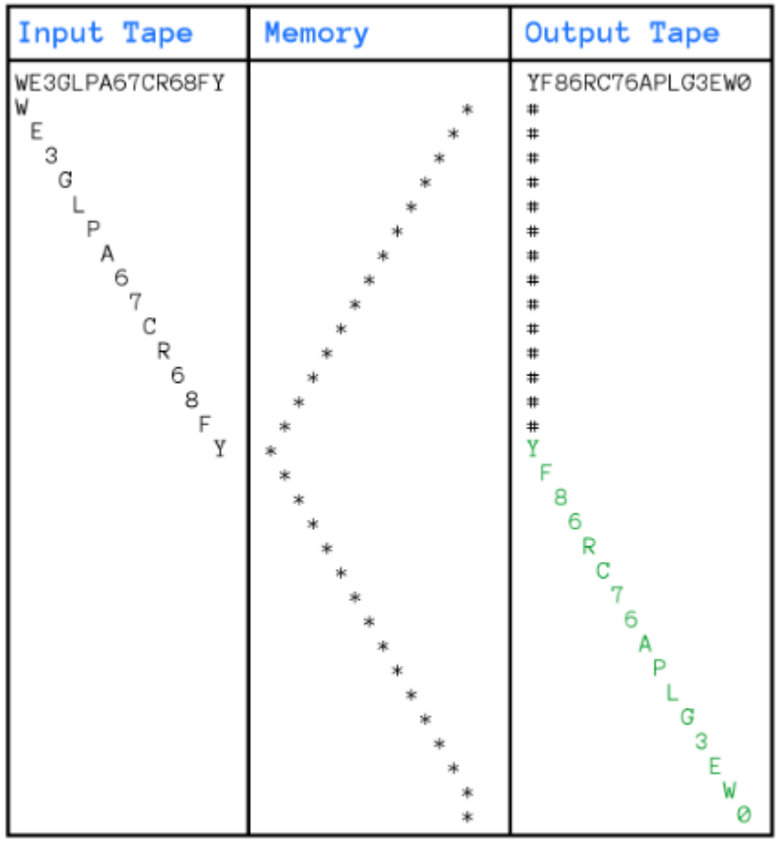
- **выход (предсказание)**
- **значение для текущей ячейки памяти**
- **текущая память ячейки**
- **решение о смене ячейки памяти (←,→), позиции на ленте и т.п.**



Discrete Read/Write: Reinforcement Learning Neural Turing Machines



An RL-NTM successfully solving a small instance of the Reverse problem (where the external memory is not used).



An RL-NTM successfully solving a small instance of the ForwardReverse problem, where the external memory is used.

Discrete Read/Write: Trainable memory addressing scheme

dynamic neural Turing machine (D-NTM)

каждая ячейка памяти = (контент, адрес)

2 главных модуля D-NTM

контроллер
RNN ~ даёт команды памяти

память

Caglar Gulcehre, Sarath Chandar, Kyunghyun Cho, Yoshua Bengio «Dynamic Neural Turing Machine with Soft and Hard Addressing Schemes», 2016 // https://arxiv.org/abs/1607.00036

Memory Network «MemN2N» – использование памяти

Задача

Sam walks into the kitchen.
Sam picks up an apple.
Sam walks into the bedroom.
Sam drops the apple.
Q: Where is the apple?
A. Bedroom

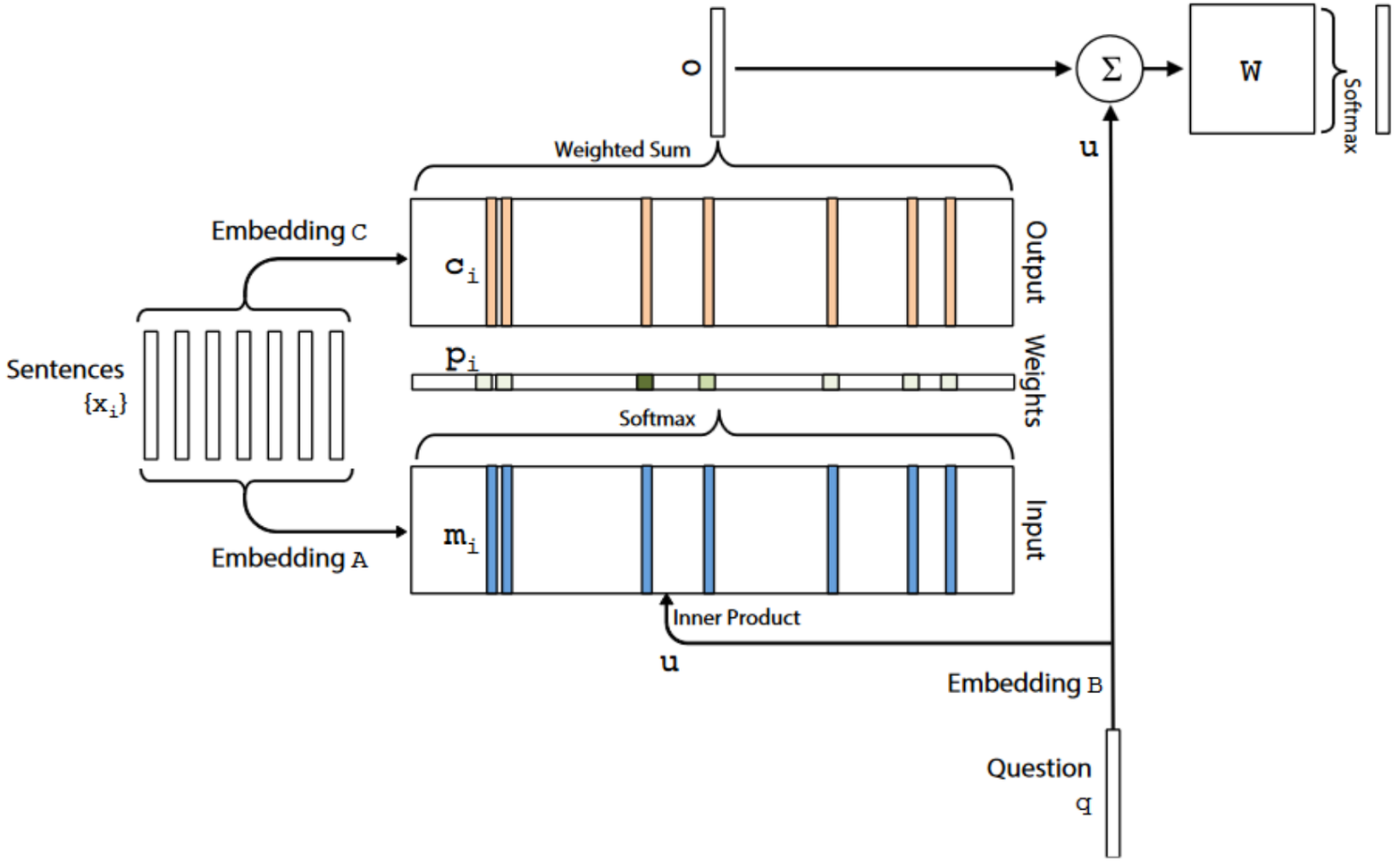
Brian is a lion.
Julius is a lion.
Julius is white.
Bernhard is green.
Q: What color is Brian?
A. White

Mary journeyed to the den.
Mary went back to the kitchen.
John journeyed to the bedroom.
Mary discarded the milk.
Q: Where was the milk before the den?
A. Hallway

https://github.com/facebook/MemNN

Memory Network «MemN2N» – использование памяти

Однослойный пример



«End-To-End Memory Networks» [Sukhbaatar S. и др., 2015 https://arxiv.org/abs/1503.08895]

Memory Network «MemN2N» – использование памяти

Задача: дан текст $x_1, \ldots, x_T$

и вопрос q**. Надо дать ответ** a**.**

Пользуемся вложениями:

$$x_i \to m_i \in \mathbb{R}^d \quad \textbf{(1)}$$

$$q \to u \in \mathbb{R}^d \quad \textbf{(2)}$$

Релевантности запроса тексту:

$$\{u^{\mathrm{T}} m_i\}_i \xrightarrow[\text{softmax}]{} \{p_i\}_i$$

Есть другое вложение (для подготовки ответа)

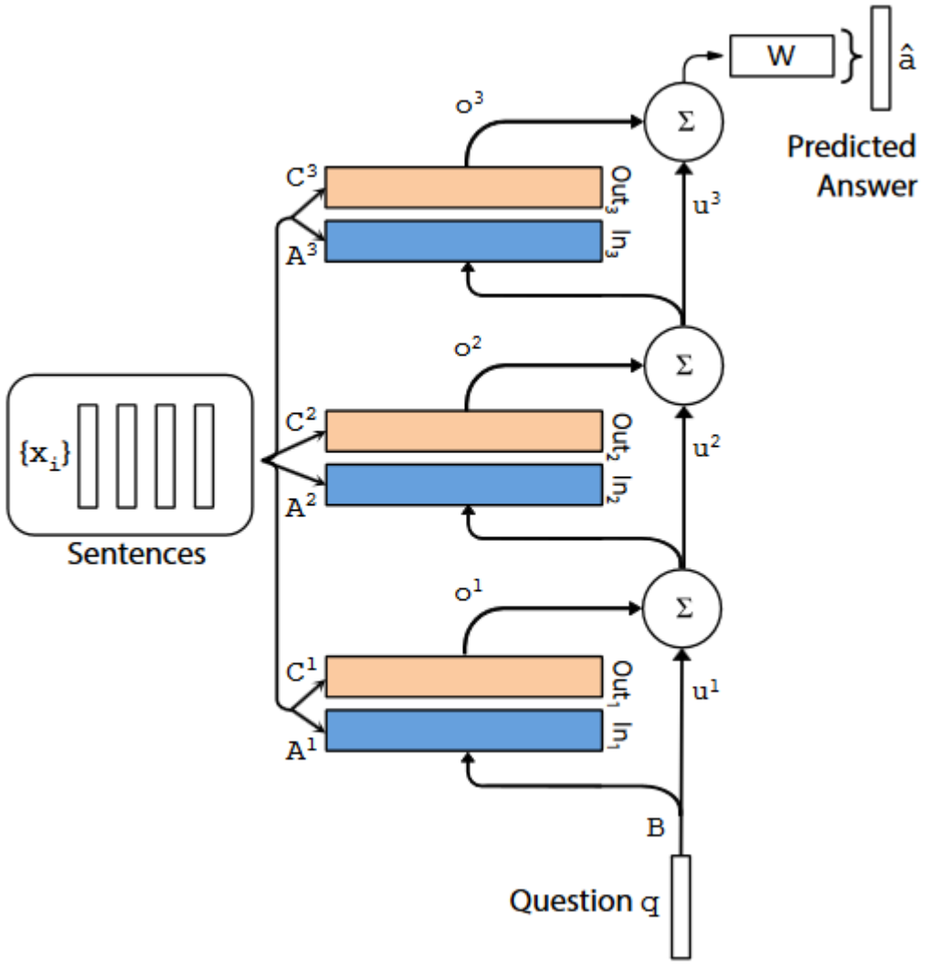
$$x_i \to c_i \in \mathbb{R}^d \quad \textbf{(3)}$$

Ответ:

$$\text{softmax}\left(W\left(\sum_i p_i c_i + u\right)\right)$$

Memory Network «MemN2N» – использование памяти

Можно использовать много слоёв…



Слева (кодировки текста) – память

Справа (изменения вектора ответа) – рекуррентная часть

Обращаемся в память и корректируем ответ…

Память выдаёт

$$\sum_i p_i c_i$$

веса ~ softmax релевантности

Memory Network «MemN2N» – использование памяти

Ещё фишки…

1. Шум как регуляризация

2. Представление предложений

Проблема: предложение – вектор

- **сумма кодировок слов**
- **position encoding (PE) – взвешенная сумма (для учёта порядка)**

3. Аналогично учёт контекста событий… (что было ДО)

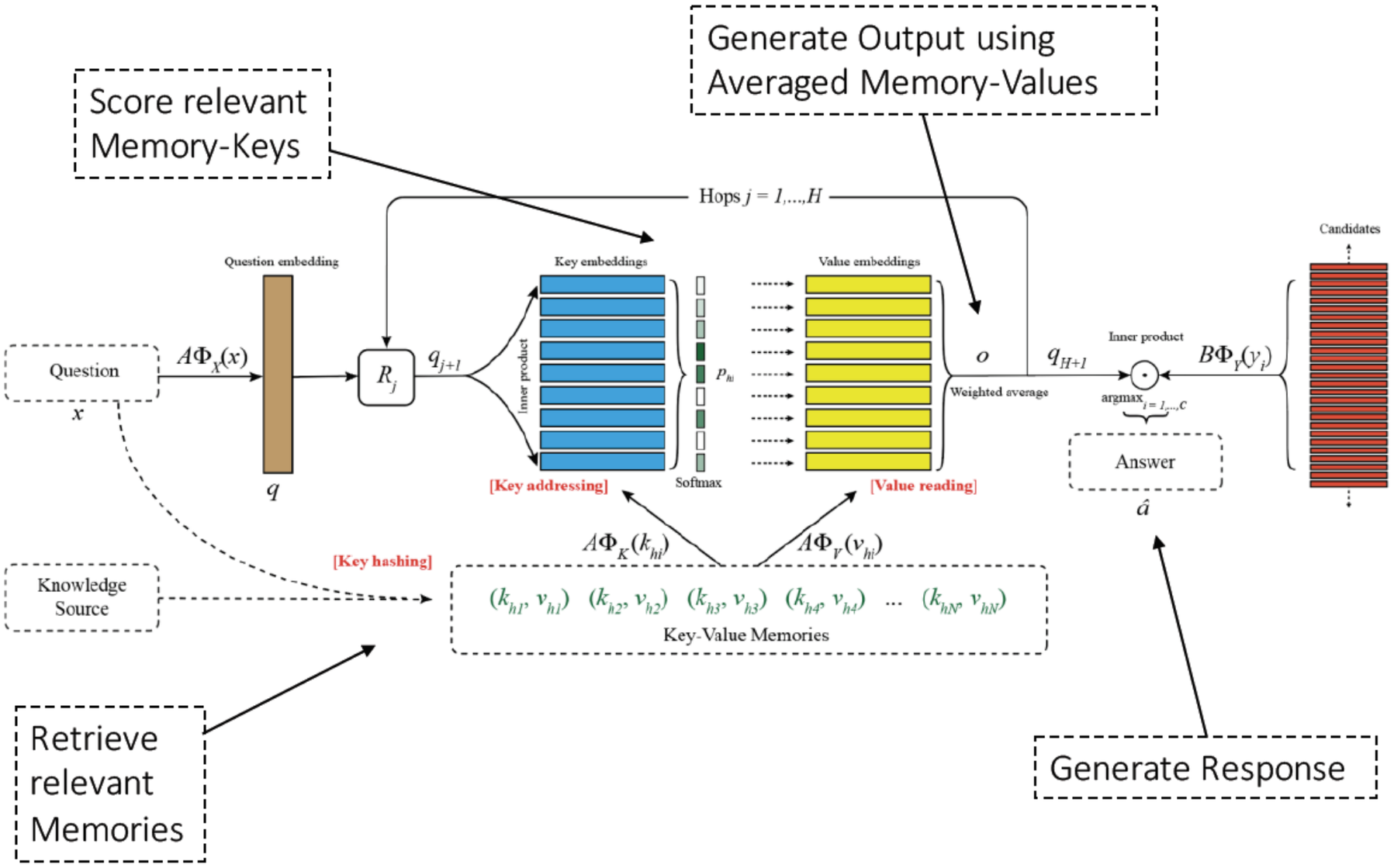
4. Темп обучения понижался вручную

(без момента и сокращения весов)

Обучено несколько сетей (разная инициализация).

Выбрана с наименьшей ошибкой…

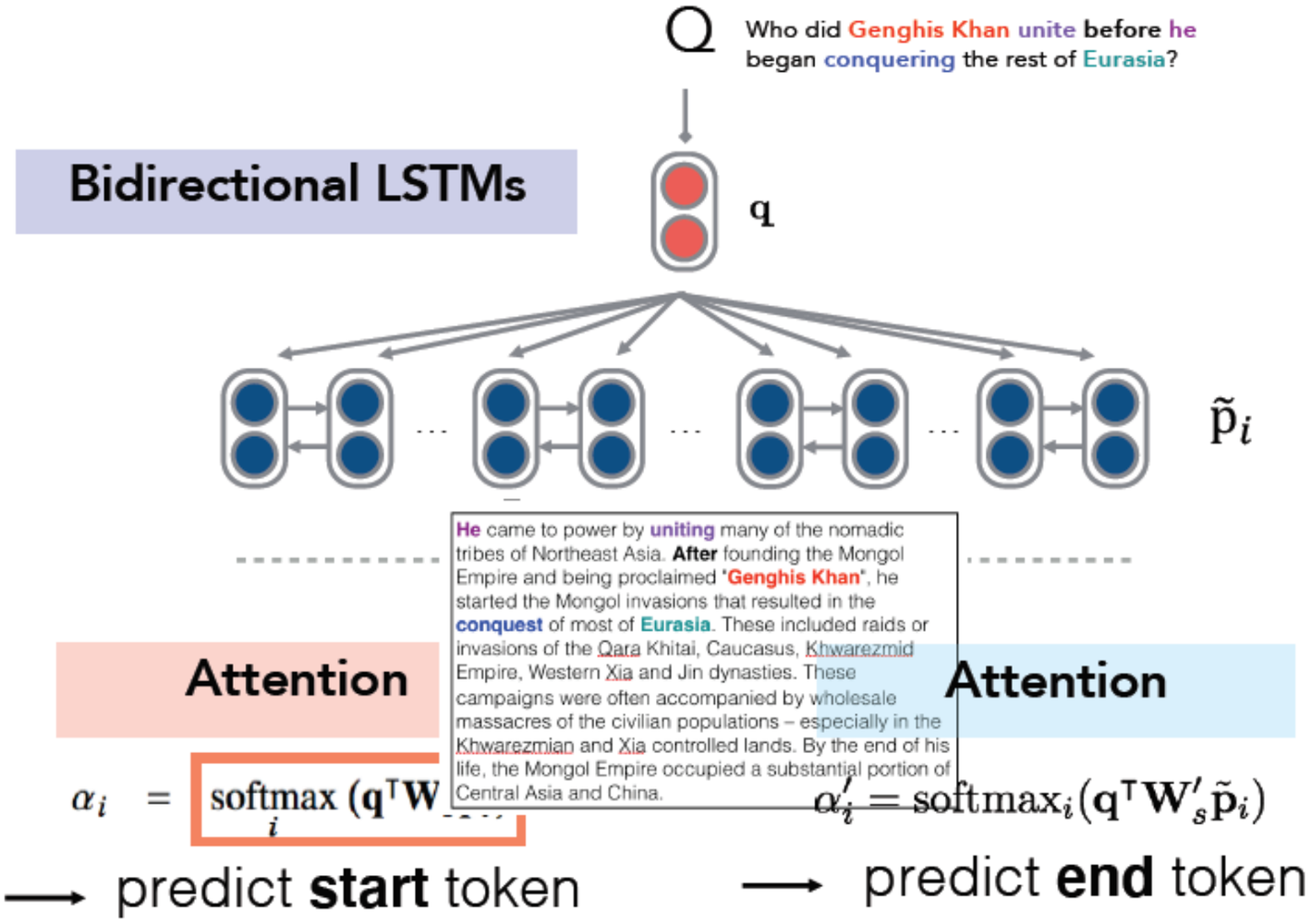
KV-MemNN



Miller et. al. «Key-Value Memory Networks for Directly Reading Documents» // EMNLP 2016

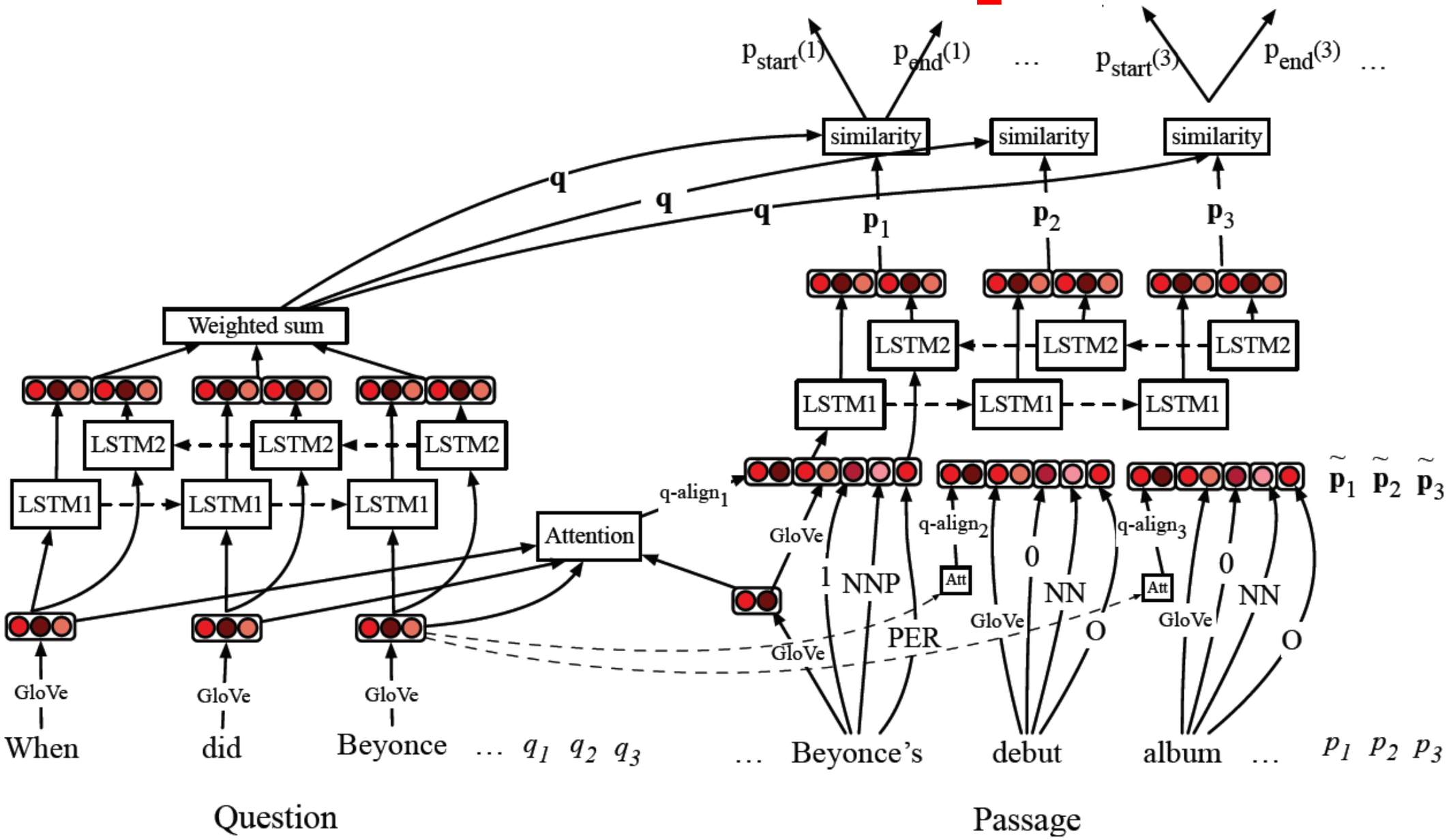
Stanford Attentive Reader

Выделить информацию в тексте по запросу – предсказать начало и конец фрагмента



http://web.stanford.edu/class/cs224n/

Stanford Attentive Reader++

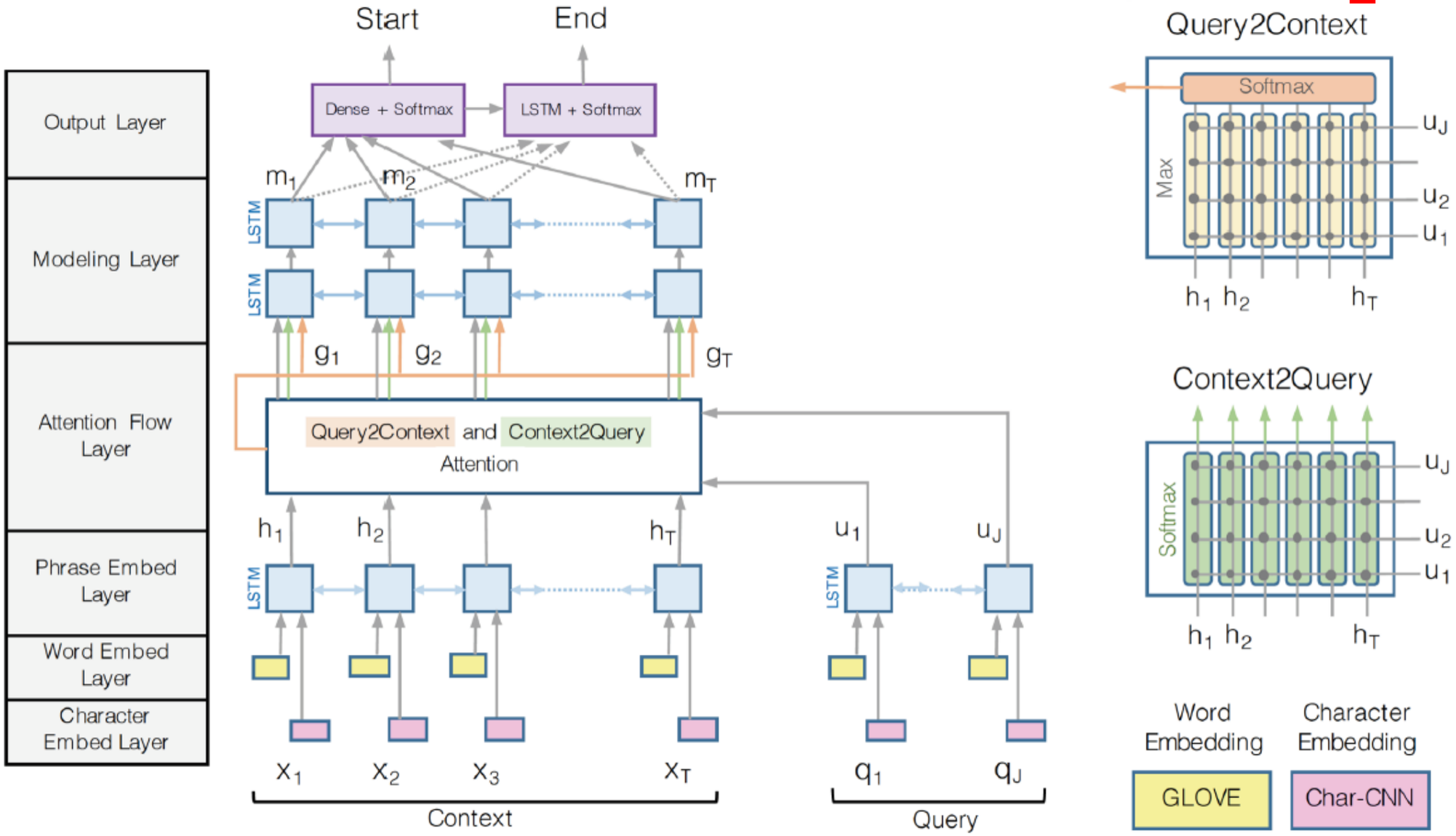


Stanford Attentive Reader++

вектор p_i – конкатенация:
представления слова (GloVe в 300D)
лингвистические признаки POS & NER
частота слова
есть ли слово в запросе (точно, с точностью до регистра, по лемме)
«Aligned question embedding»

Chen, Bolton, Manning, 2016

BiDAF: Bi-Directional Attention Flow for Machine Comprehension

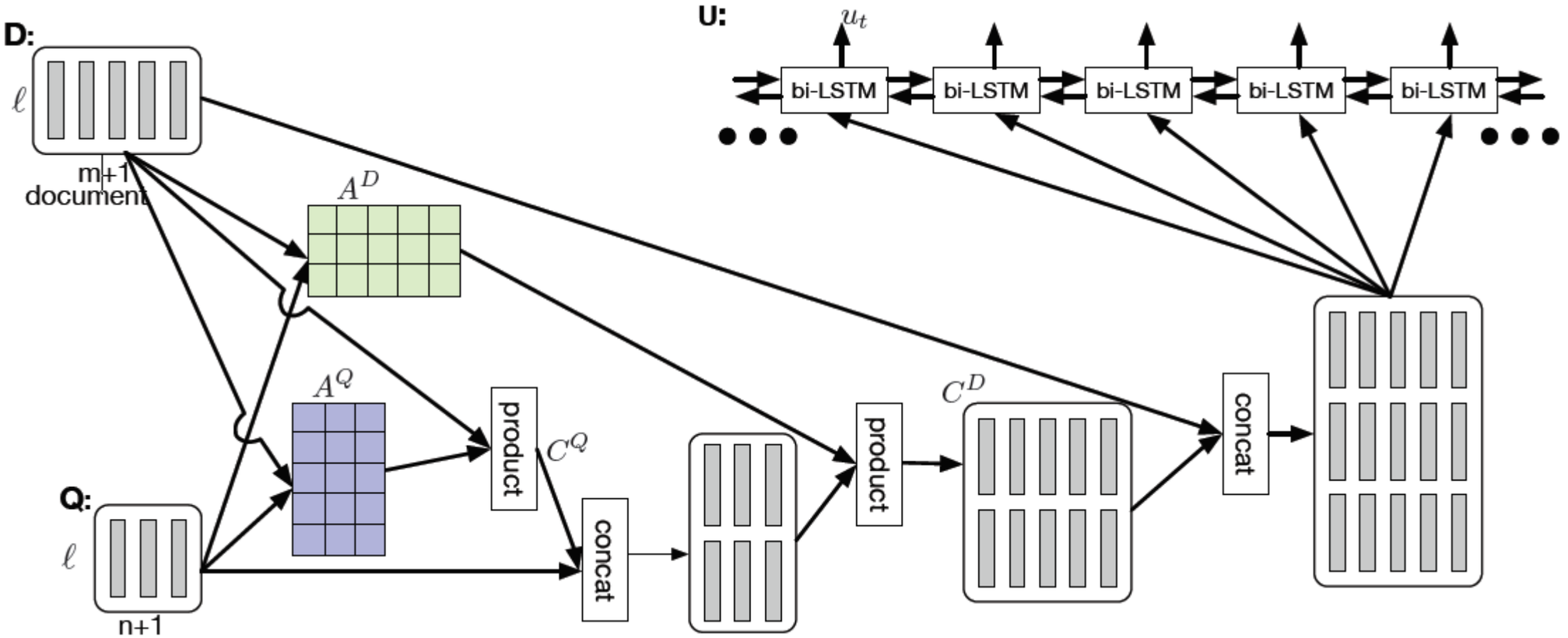


Seo, Kembhavi, Farhadi, Hajishirzi, ICLR 2017

BiDAF: Bi-Directional Attention Flow for Machine Comprehension

идея: внимание должно течь в обе стороны
от контекста к вопросу и от вопроса к контексту

Dynamic Coattention Networks for Question Answering



Coattention layer – двунаправленное внимание вопрос-контекст

+ 2 уровня

Caiming Xiong, Victor Zhong, Richard Under review as a conference paper at ICLR 2017

Socher ICLR 2017

Ещё…

FusionNet (Huang, Zhu, Shen, Chen 2017)

Open-domain Question Answering DrQA (Chen, et al. ACL 2017)

https://arxiv.org/abs/1704.00051

CNN + RNN

можно CNN для кодировщика, а RNN для декодировщика

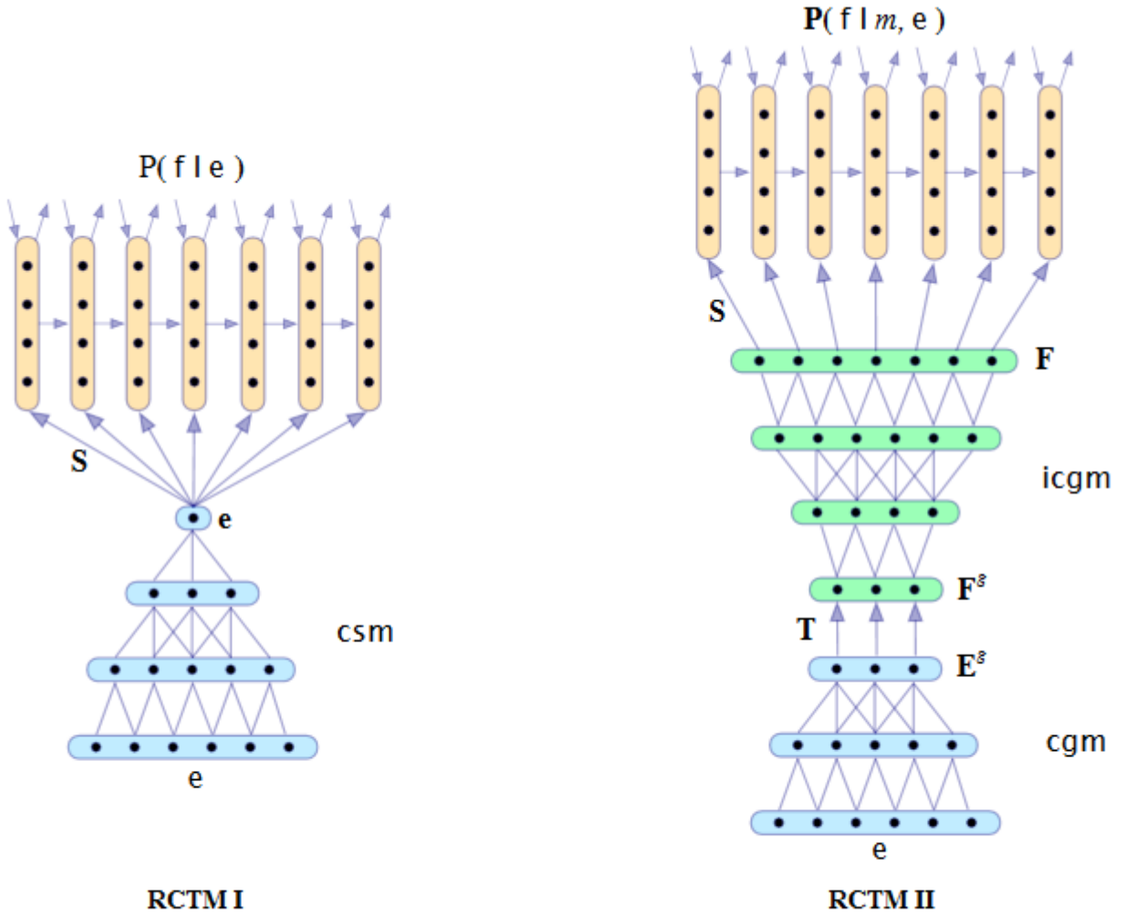


Figure 3: A graphical depiction of the two RCTMs. Arrows represent full matrix transformations while lines are vector transformations corresponding to columns of weight matrices.

Nal Kalchbrenner Phil Blunsom «Recurrent Continuous Translation Models» // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing https://www.aclweb.org/anthology/D13-1176.pdf

CNN + RNN

Learning Character-level Representations for Part-of-Speech Tagging – Dos Santos and Zadrozny (2014)

- свёртки над символами для генеририрвоания представлений слов
- фиксированное окно представлений слов для PoS-tagging

Character-Aware Neural Language Models – Kim, Jernite, Sontag, and Rush 2015

- посимвольное представление слов
- использование свёрток, «highway network» и LSTM

Итог

свёрточные сети – не только для изображений
можно CNN + RNN

seq2seq – простоя и понятная архитектура

внимание – на что «смотрим»
коэффициенты специально считаются

Есть разные виды внимания

Есть интересная концепция «памяти»