Deep Learning Technology and Application

Ge Li

Peking University

Encoder-Decoder Model

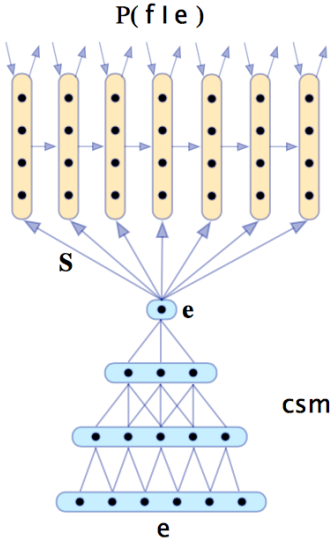
Encoder-Decoder Roadmap - 2013

[PDF] Recurrent Continuous Translation Models.

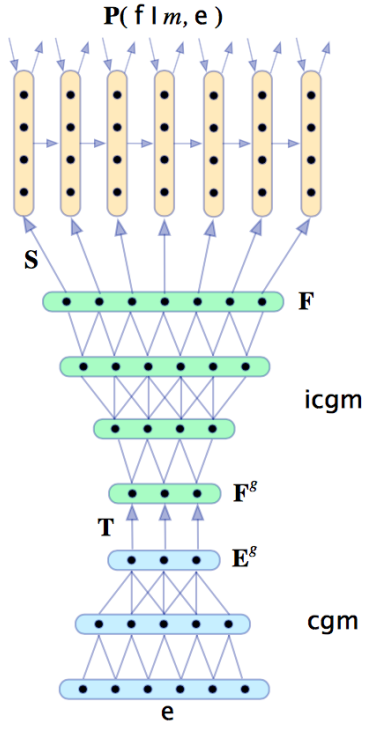
N Kalchbrenner, P Blunsom - EMNLP, **2013** - anthology.aclweb.org

Abstract We introduce a class of probabilistic continuous translation models called Recurrent Continuous Translation Models that are purely based on continuous representations for words, phrases and sentences and do not rely on alignments or phrasal translation units. The models have a generation and a conditioning aspect. The generation of the translation is modelled with a target Recurrent Language Model, whereas the ...

Cited by 279 Related articles All 7 versions Cite Save More



RCTM I



RCTM II

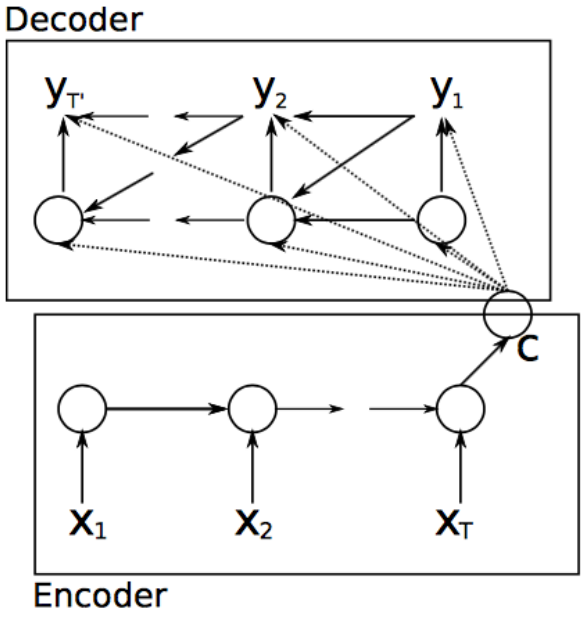
Encoder-Decoder Roadmap - 2014 RNNenc

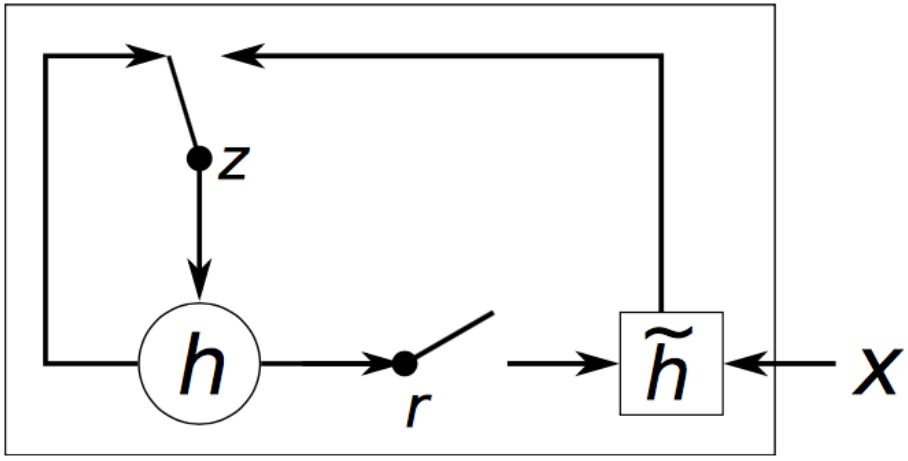
Learning phrase representations using RNN encoder-decoder for statistical machine translation

K Cho, B Van Merriënboer, C Gulcehre… - arXiv preprint arXiv: …, 2014 - arxiv.org

Abstract: In this paper, we propose a novel neural network model called RNN Encoder-Decoder that consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained to maximize the conditional probability of a target sequence ...

Cited by 981 Related articles All 19 versions Cite Save





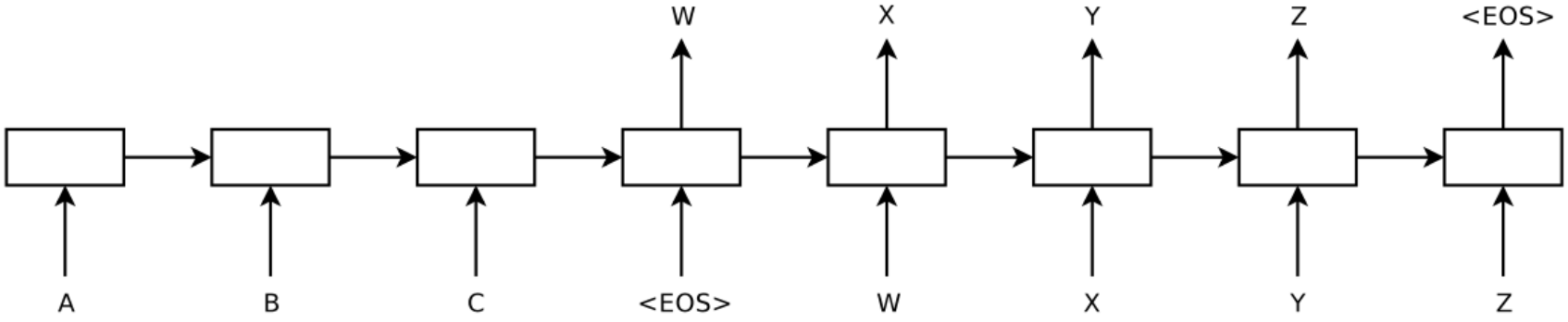
Encoder-Decoder Roadmap - 2014 NIPS

[PDF] Sequence to sequence learning with neural networks

I Sutskever, O Vinyals, QV Le - Advances in neural information …, 2014 - papers.nips.cc

Page 1. **Sequence** to **Sequence Learning** with Neural Networks Ilya Sutskever Google ilyasu@google.com **...** In this paper, we present a general end-to-end approach to **sequence learning** that makes minimal assumptions on the **sequence** structure. **...**

Cited by 1505 Related articles All 15 versions Cite Save More

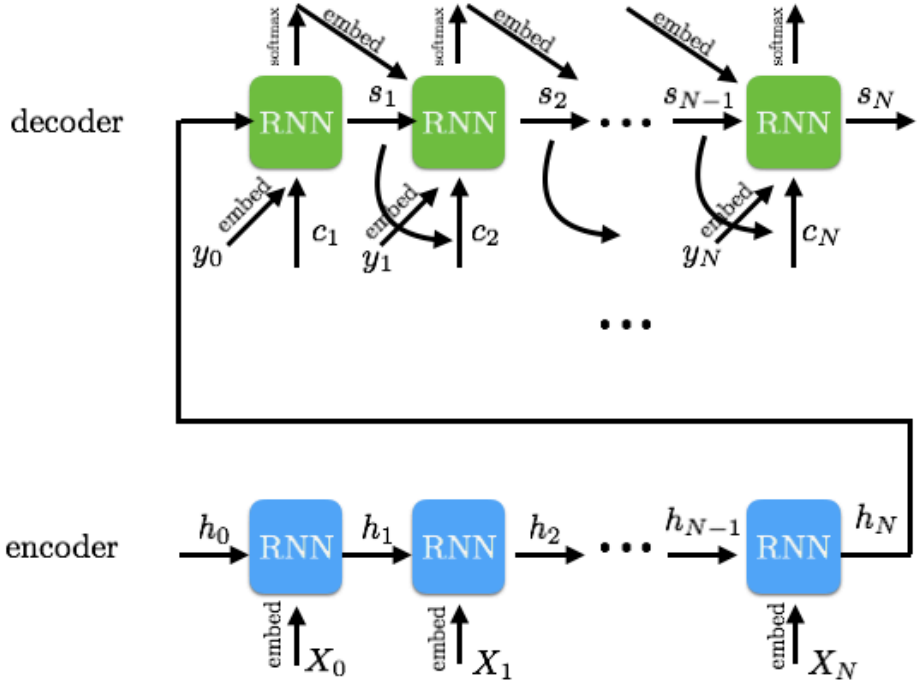


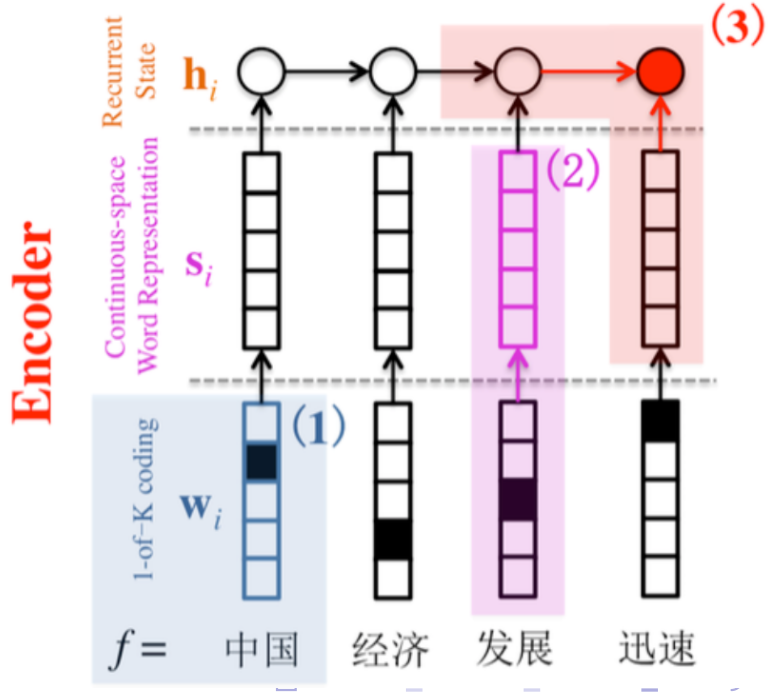
Encoder-Decoder

Encoder:

- 设 $D = (x^1, y^1), \ldots, (x^N, y^N)$ 为包含 N 个平行句子的平行语料库；(下面先针对一组平行句子进行讨论，此时，可以省去上标 N)
- 设 h_t 为 Encoding 过程中 t 时刻隐藏层的状态；

$$h_i = f(h_{i-1}, x_i)$$

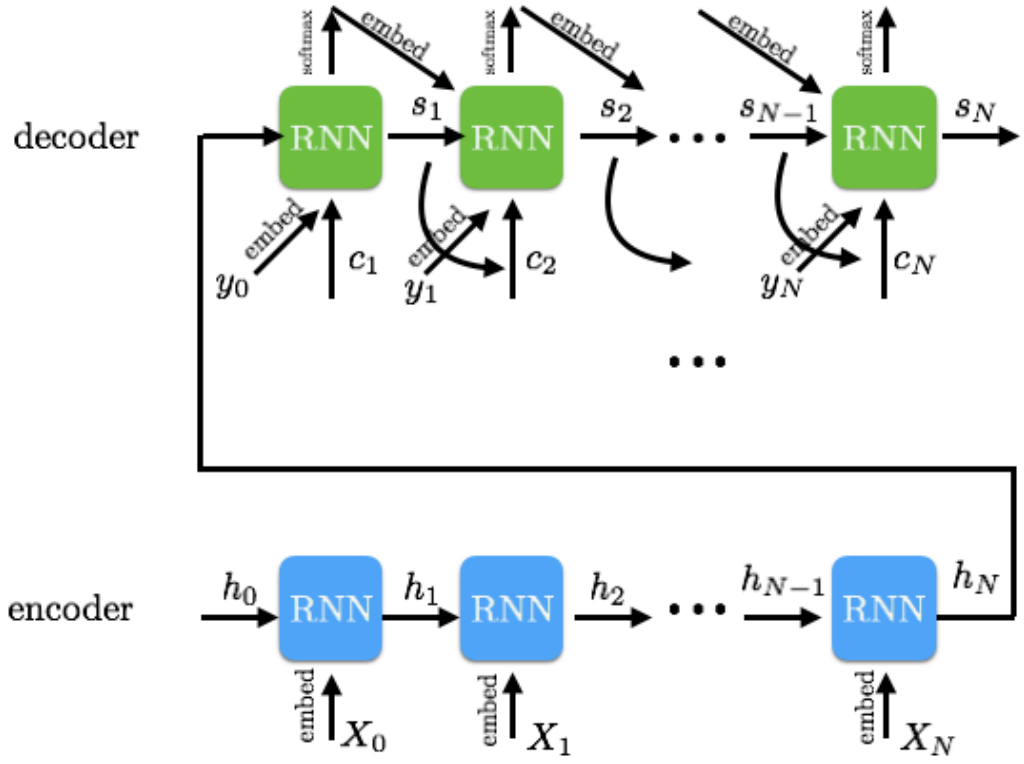


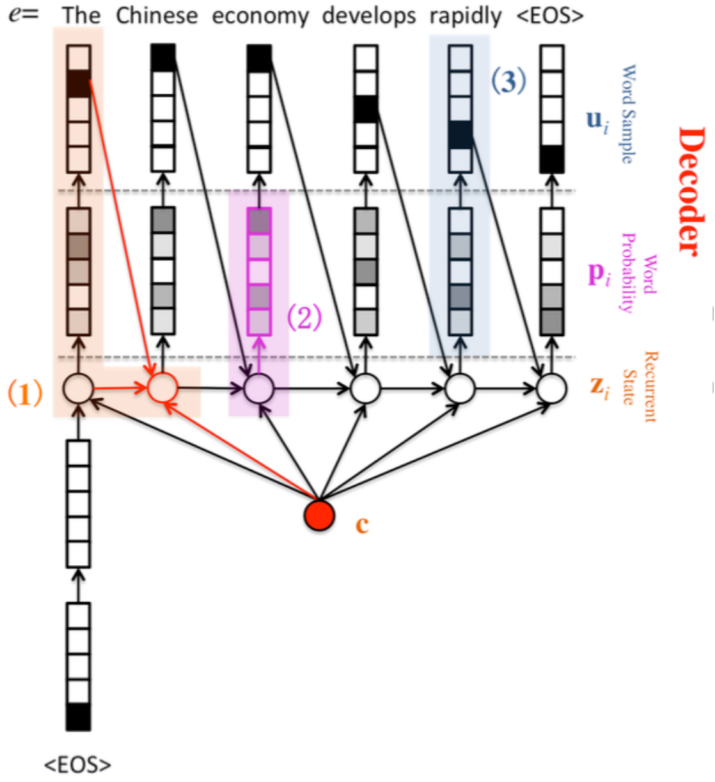


Encoder-Decoder

Decoder:

- 设 h_t 为 Encoding 过程中 t 时刻隐藏层的状态；
- 设 s_o 为 Decoding 过程中 o 时刻隐藏层的状态；
- 设 c_o 为 Decoding 过程中 o 时刻的上下文信息；





Encoder-Decoder

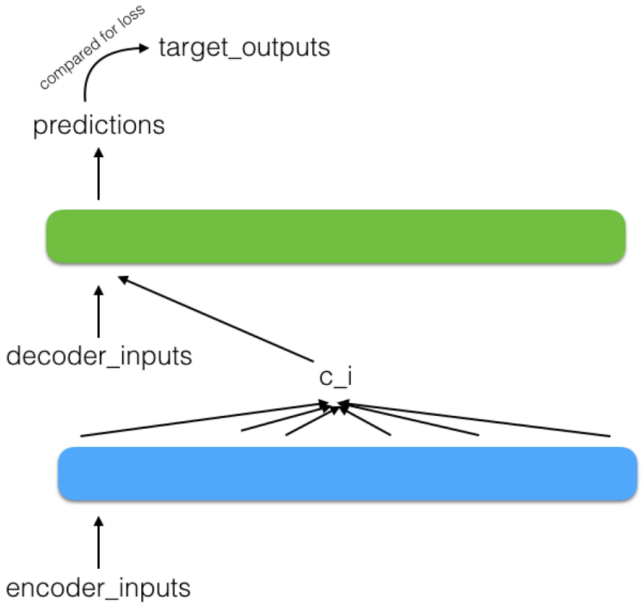
Decoder:

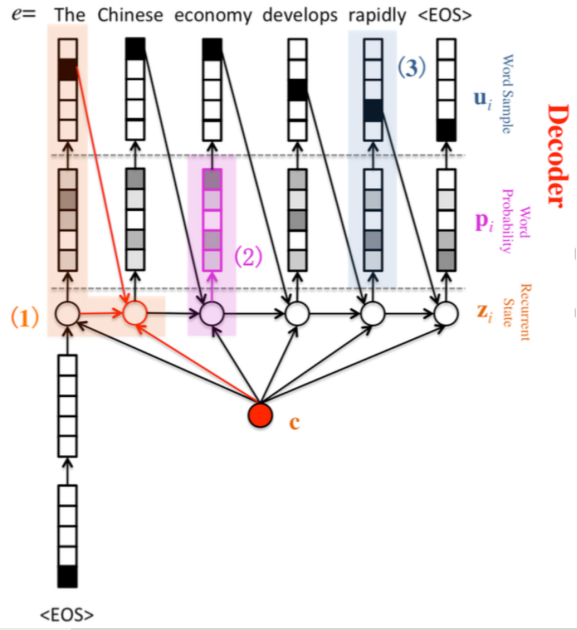
$$p(y_1, \ldots, y_O | x_1, \ldots, x_T) = \prod_{o=1}^{O} p(y_o | y_1, \ldots, y_{o-1}, c)$$

$$p(y_o | y_1, \ldots, y_{o-1}, c) = g(y_{o-1}, s_o, c)$$

$$s_o = f(y_{o-1}, s_{o-1}, c)$$

$$c = q(\{h_0, \ldots, h_T\}) \quad \text{不妨先设：} c_t = h_T$$

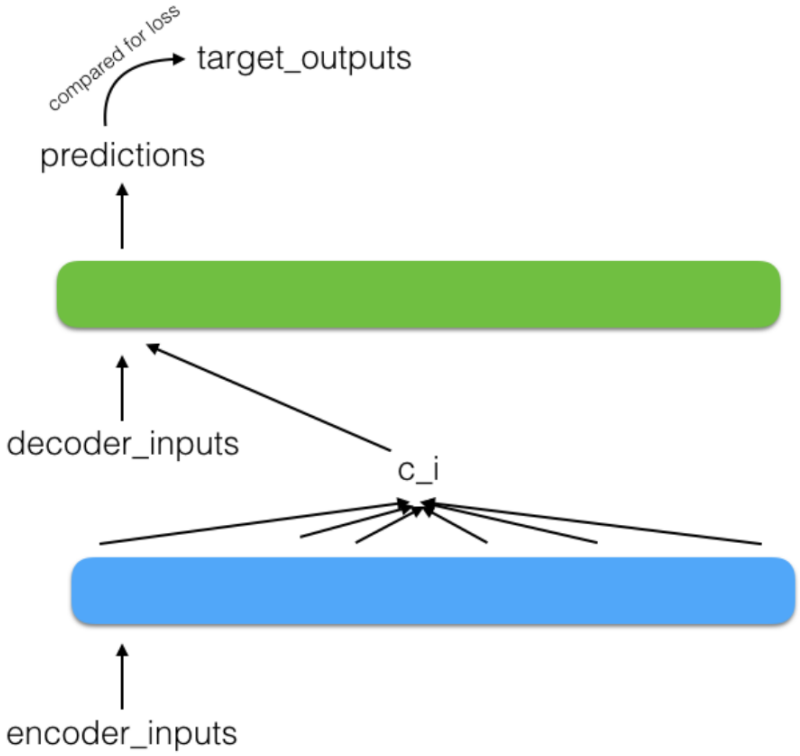


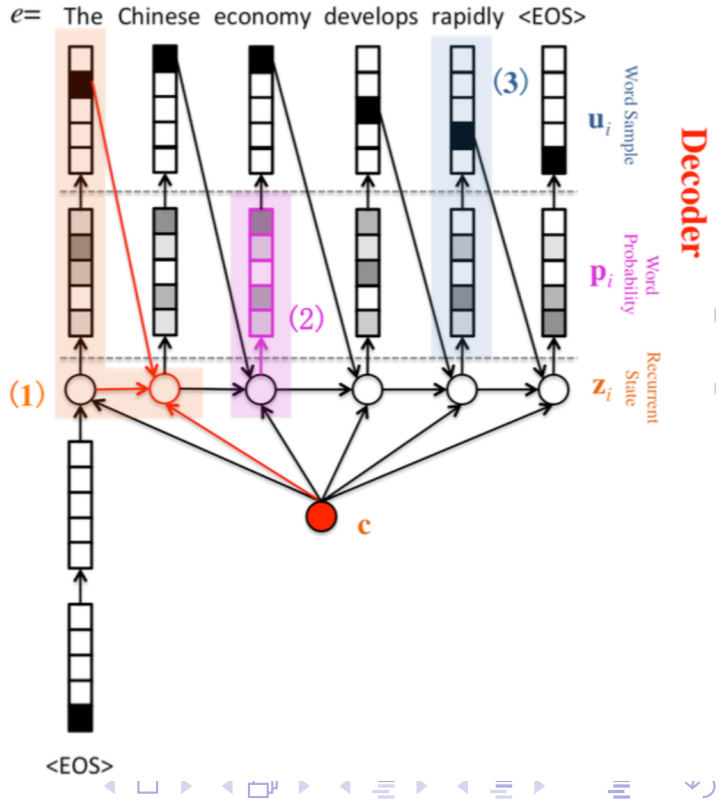


Encoder-Decoder

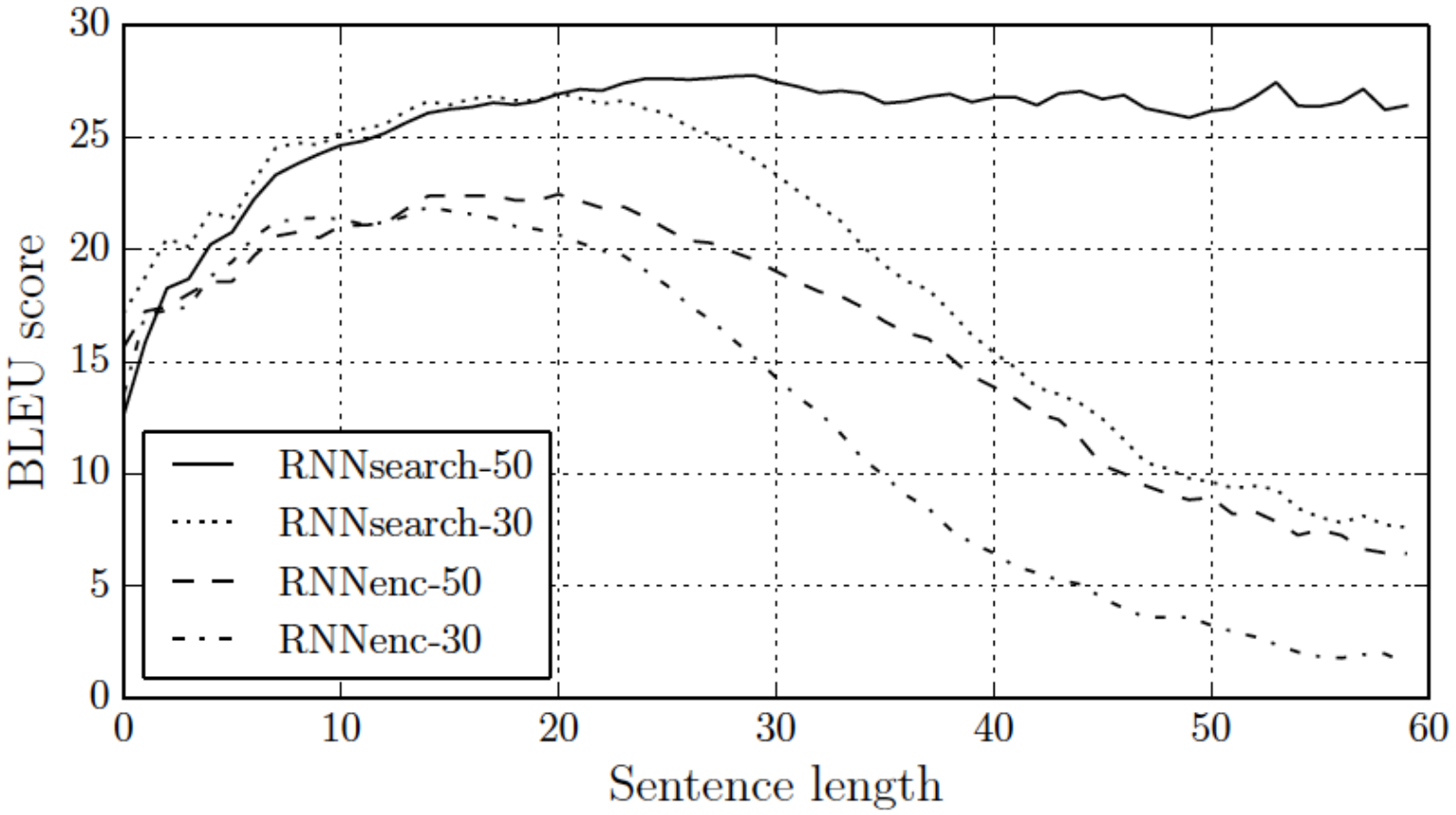
Decoder: **对全部语料库** $D = (x^1, y^1), \ldots, (x^N, y^N)$, **训练目标为：**

$$J(D, \Theta) = \frac{1}{N} \sum_{n=1}^{N} \log p(y^n | x^n, \Theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{o=1}^{O} \log p(y_o^n | y_1^n, \ldots, y_{o-1}^n, c, \Theta)$$





Encoder-Decoder Model



Attention Model

Attention Roadmap - 2014 RNNsearch

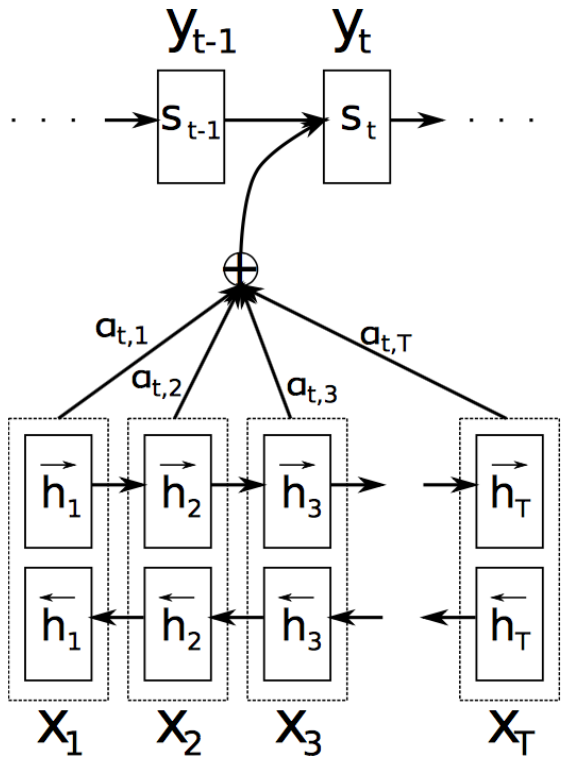
Bidirectional RNN for Annotating Squence

Neural machine translation by jointly learning to align and translate

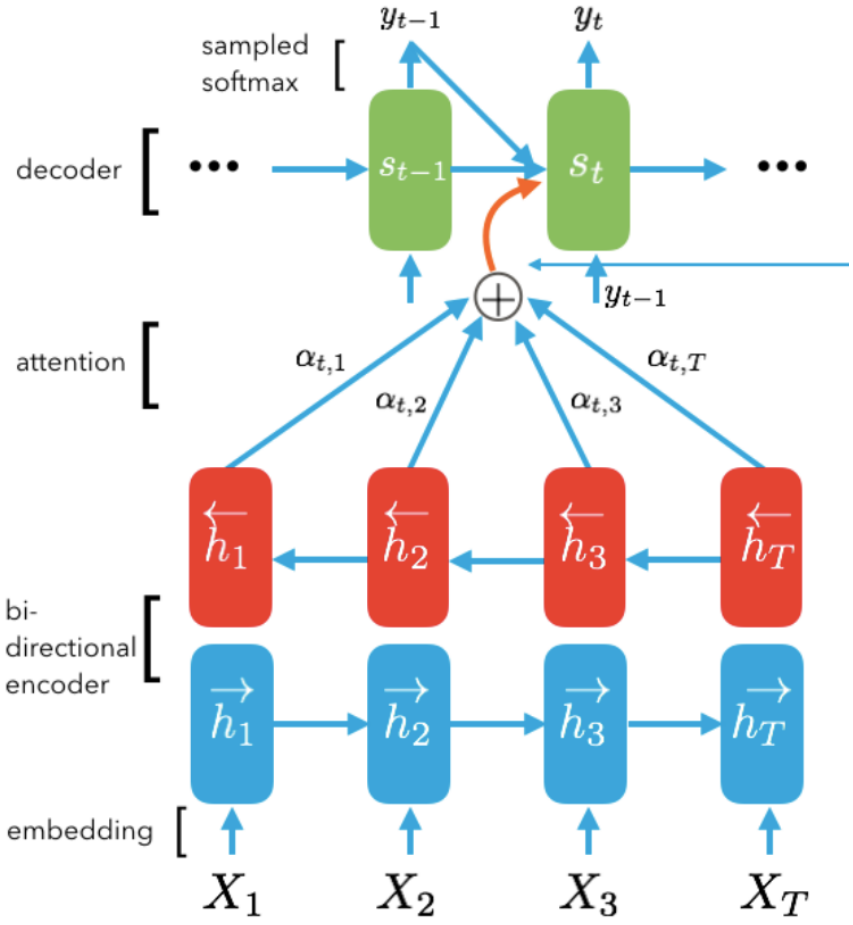
D Bahdanau, K Cho, Y Bengio - arXiv preprint arXiv:1409.0473, 2014 - arxiv.org

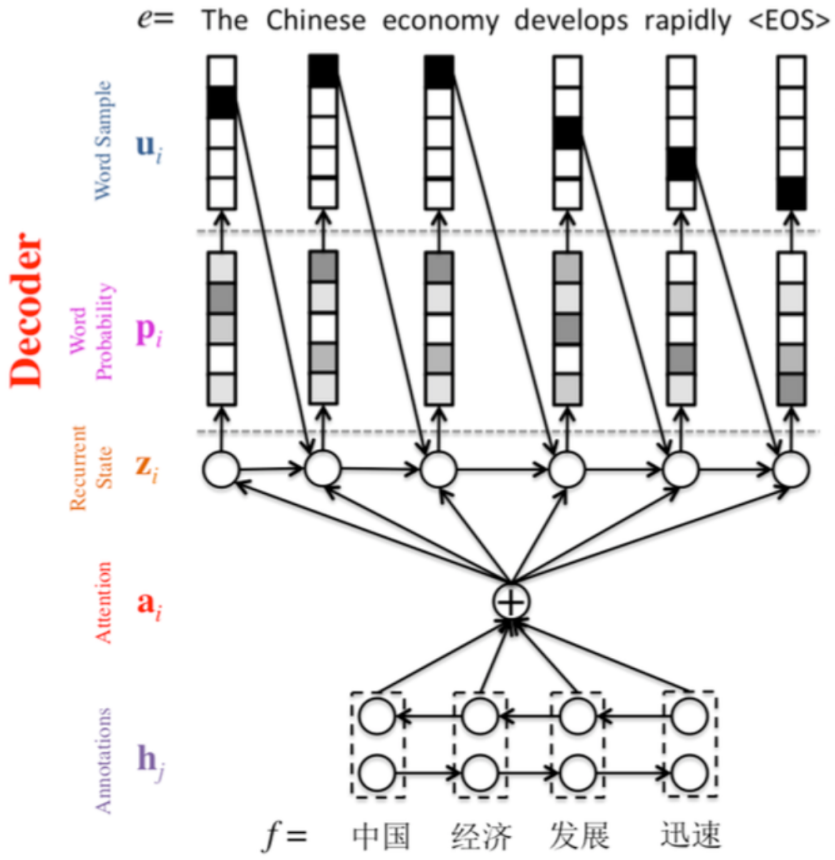
Abstract: **Neural machine translation** is a recently proposed approach to **machine translation**. Unlike the traditional statistical **machine translation**, the **neural machine translation** aims at building a single **neural** network that can be **jointly** tuned to maximize the

Cited by 1276 Related articles All 12 versions Cite Save

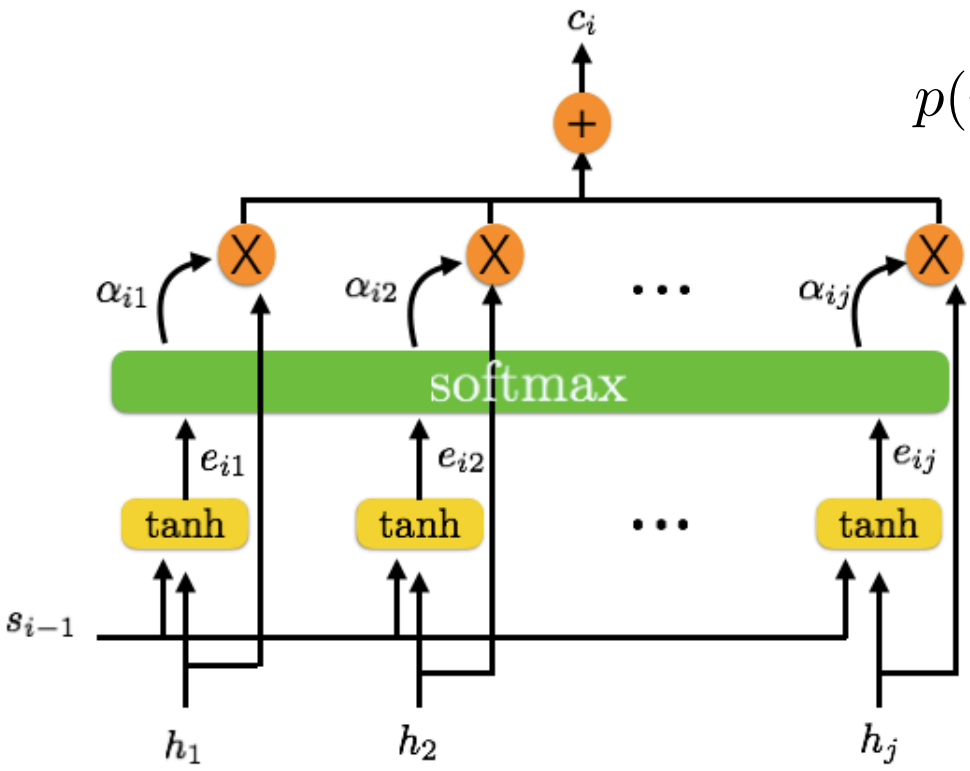


Bidirectional RNN for Annotating Sequence





Bidirectional RNN for Annotating Sequence



$$p(y_o|y_1,\ldots,y_{o-1},c_o) = g(y_{o-1},s_o,c_o)$$

$$s_o = f(y_{o-1},s_{o-1},c_o)$$

$$c_o = \sum_{t=1}^{T} \alpha_{ot} h_t$$

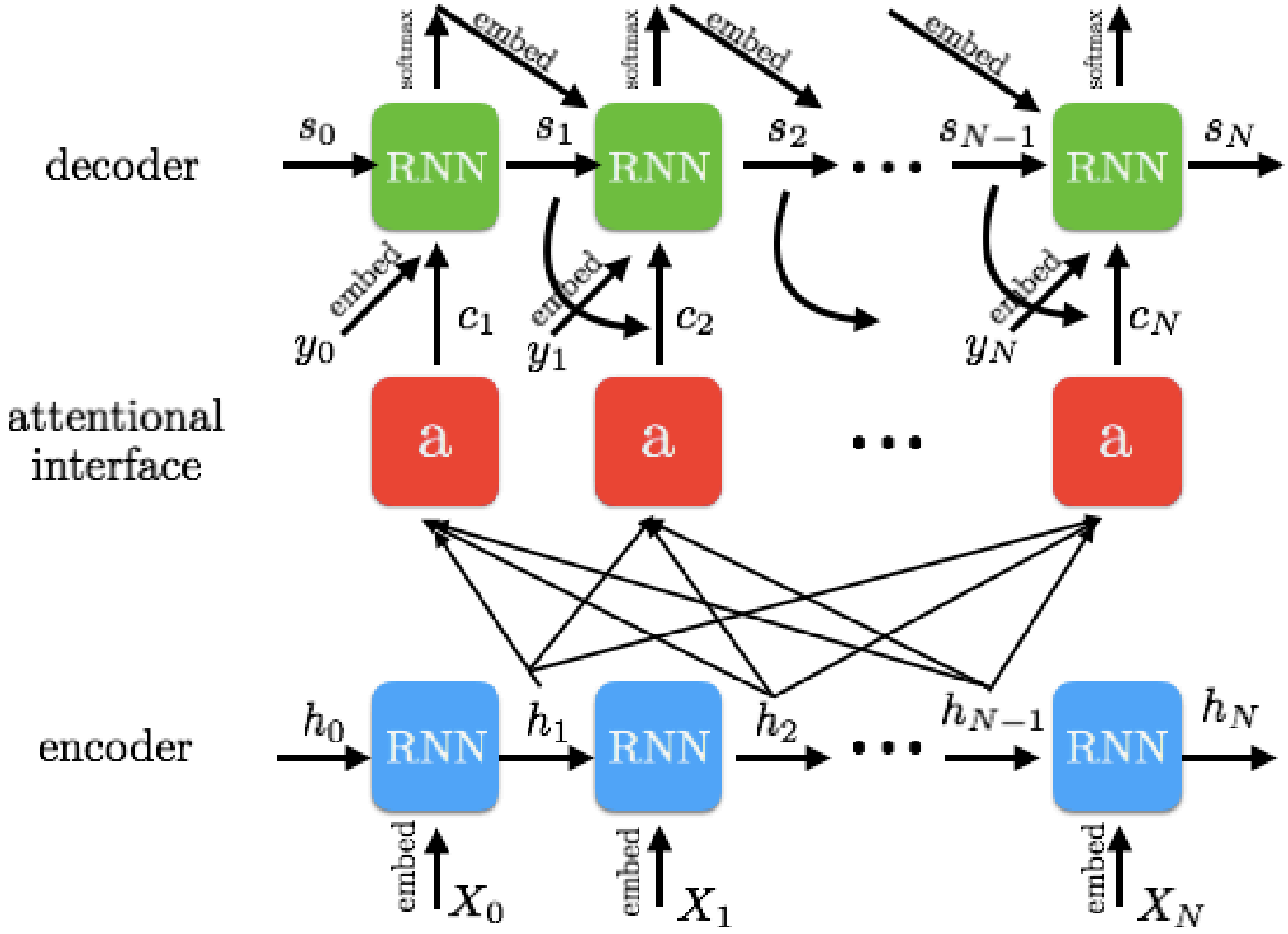
$$\alpha_{ot} = \frac{exp(e_{ot})}{\sum_{k=1}^{T} exp(e_{ok})}$$

$$e_{ot} = r(s_{o-1},h_t)$$

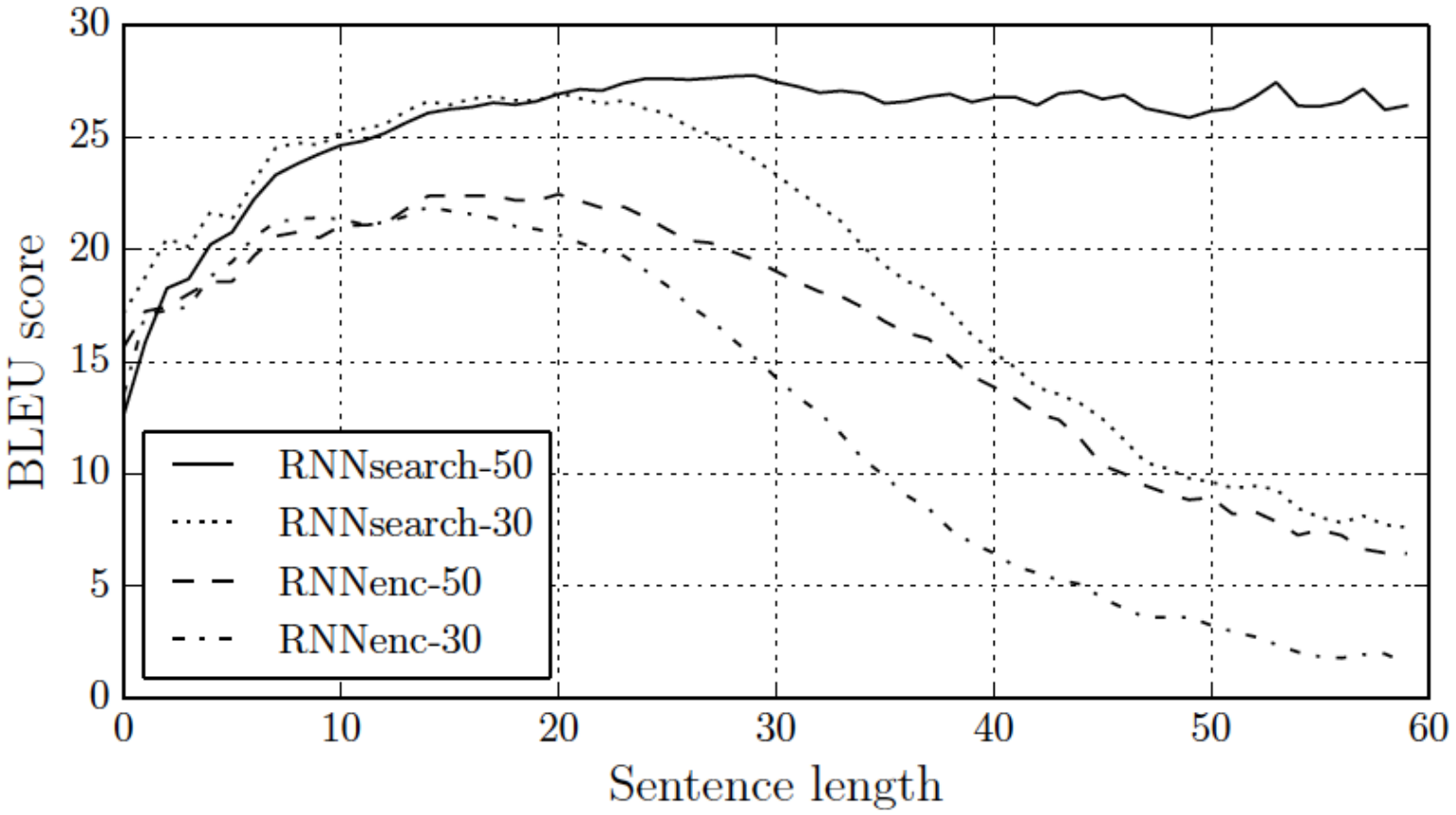
$$= v^T tanh(W s_{o-1} + U h_t)$$

可见，计算顺序为： $s_{o-1} \rightarrow \alpha_{ot} \rightarrow c_o \rightarrow s_o$

Bidirectional RNN for Annotating Sequence



Bidirectional RNN for Annotating Sequence



Attention Roadmap - 2015

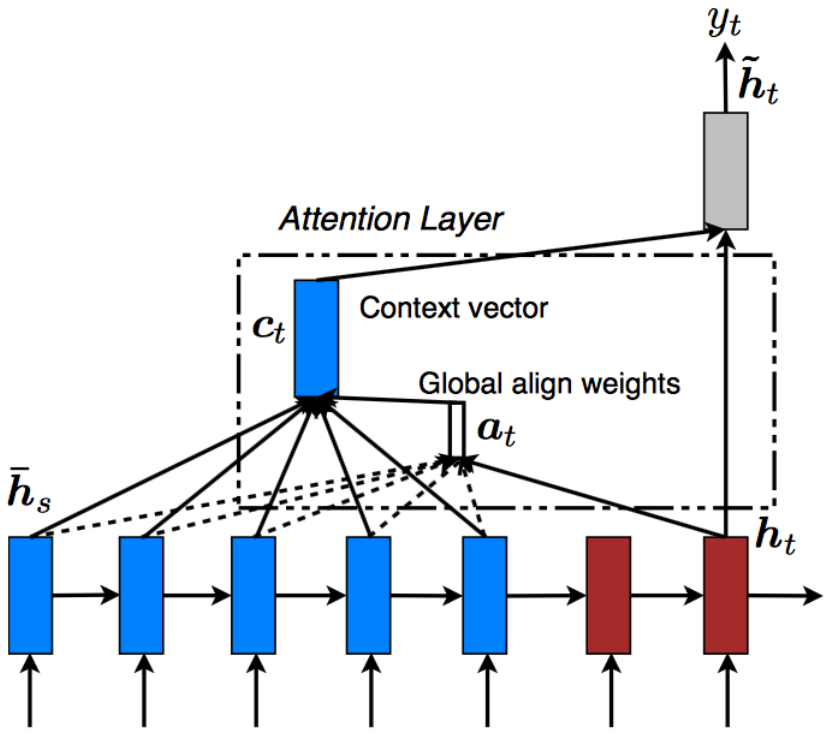
Global and Local Attentional Model

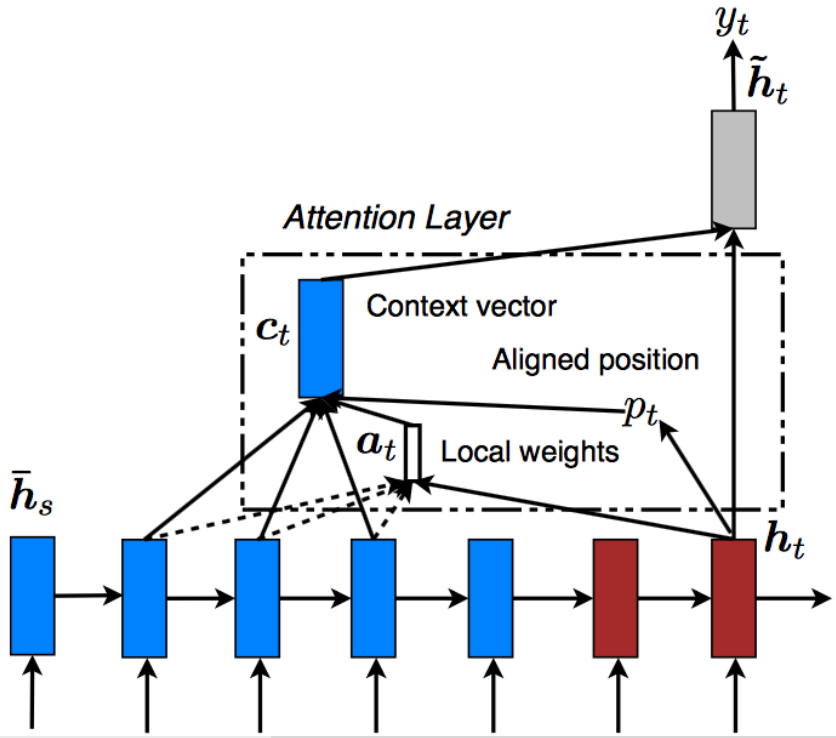
Effective approaches to **attention-based** neural machine translation

MT Luong, H Pham, CD Manning - arXiv preprint arXiv:1508.04025, 2015 - arxiv.org

Abstract: An attentional mechanism has lately been used to improve neural machine translation (NMT) by selectively focusing on parts of the source sentence during translation. However, there has been little work exploring useful architectures for **attention-based** NMT.

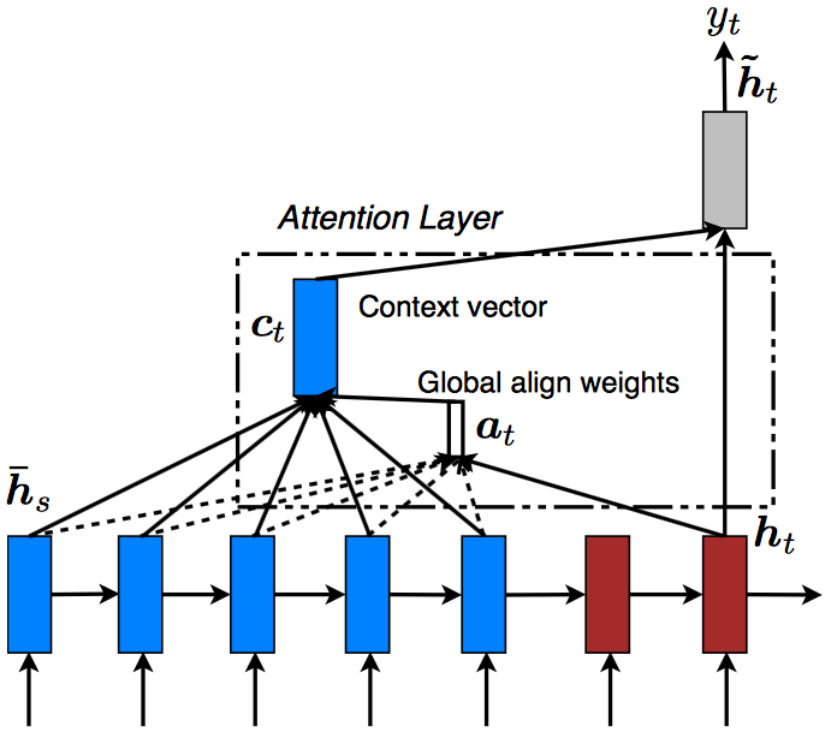
Cited by 240 Related articles All 20 versions Cite Save





Global and Local Attentional Model

Global Attentional Model



$$\tilde{h_o} = tanh(W_c[c_o; s_o])$$

$$p(y_o|y_1, \ldots, y_{o-1}, x) = softmax(W_s\tilde{h_o})$$

$$c_o = \sum_t^T h_t \alpha_{ot}$$

$$\alpha_{ot} = align(s_o, h_t)$$

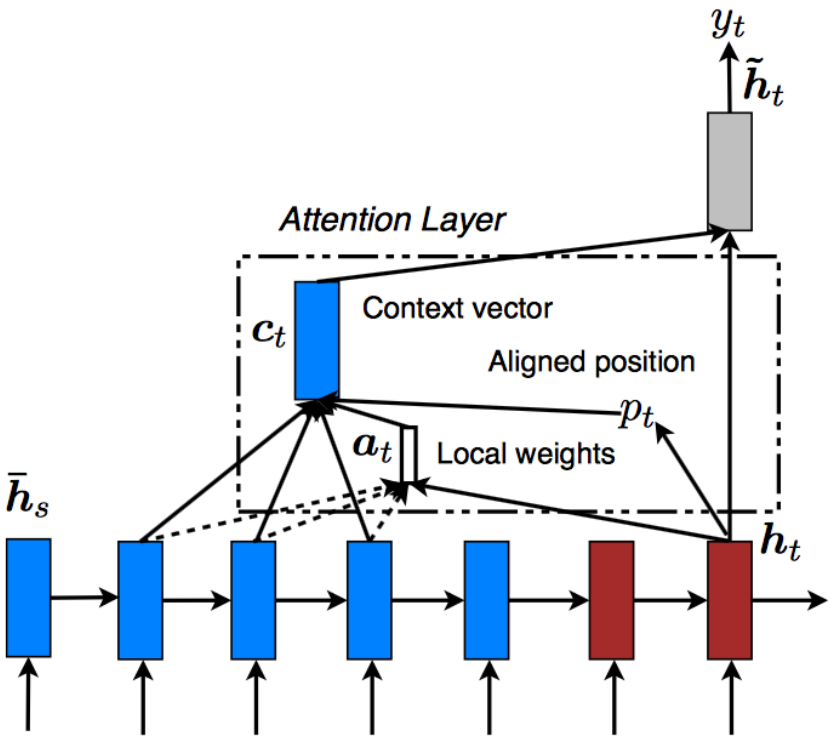
$$= \frac{exp(score(s_o, h_t))}{\sum_{t'}^T exp(score(s_o, h_{t'}))}$$

$$score(s_o, h_t) = \begin{cases} s_o^T h_t \\ s_o^T W_a h_t \\ v_a^T tanh(W_a[s_o; h_t]) \end{cases}$$

可见，计算顺序为： $s_o \rightarrow \alpha_{ot} \rightarrow c_o \rightarrow \tilde{h_o}$

Global and Local Attentional Model

Local Attentional Model



在一个窗口中计算上下文信息 c_t，但关键是如何选取窗口：

（1）指定一个窗口：$[p_t - D, p_t + D]$，其中 $p_t = t$，D 为经验参数；

（2）计算一个窗口：

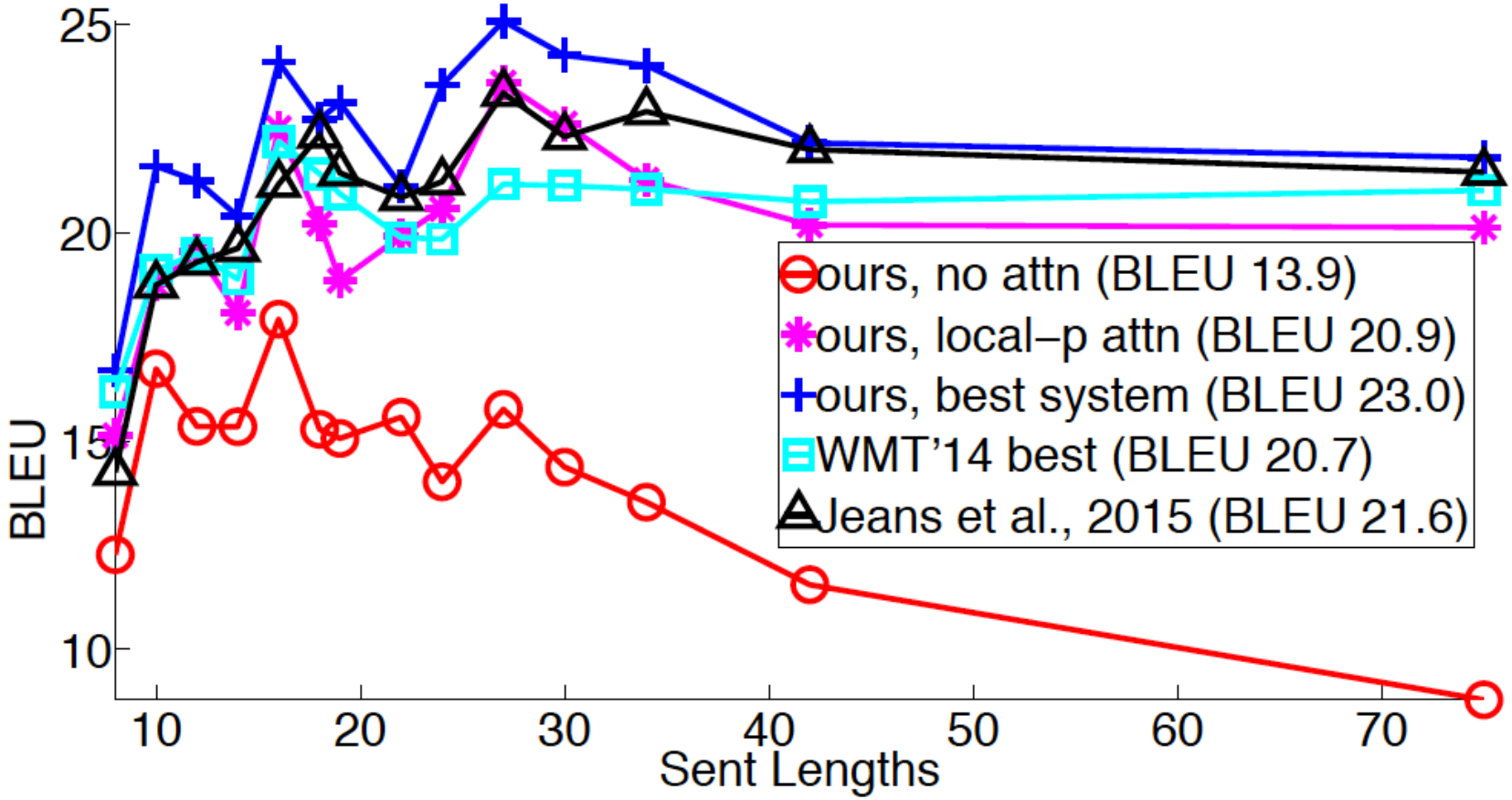
$$p_t = S \dot{s} igmod(v_p^T) tanh(W_p h_t)$$

其中，W_p 与 v_p 为模型参数，S 为源句子长度；

（3）更进一步，以 p_t 为中心，对 α_{ot} 做高斯：

$$\alpha_{ot} = align(s_o, h_t) exp(-\frac{(s - p_t)^2}{2\delta^2})$$

Global and Local Attentional Model



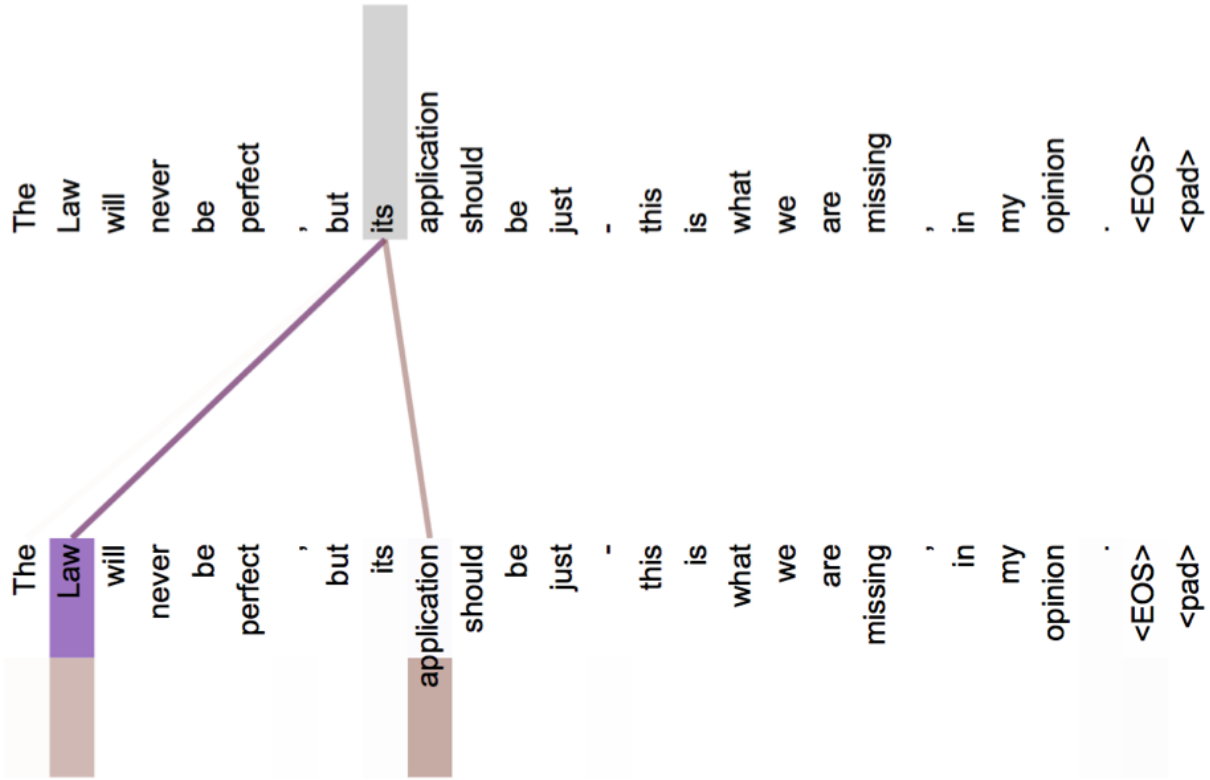
Self Attention

Attention is all you need

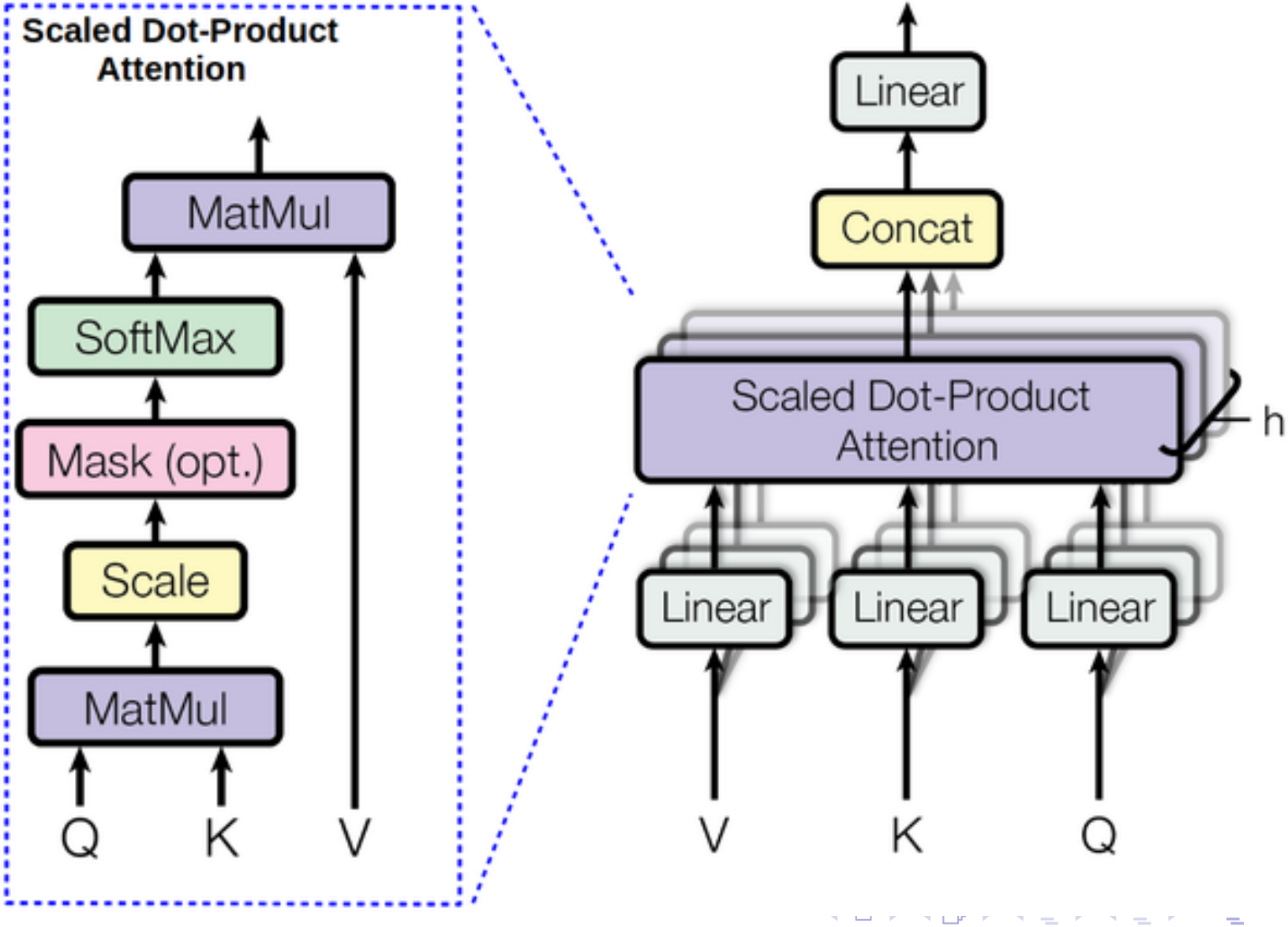
A Vaswani, N Shazeer, N Parmar… - Advances in Neural …, 2017 - papers.nips.cc

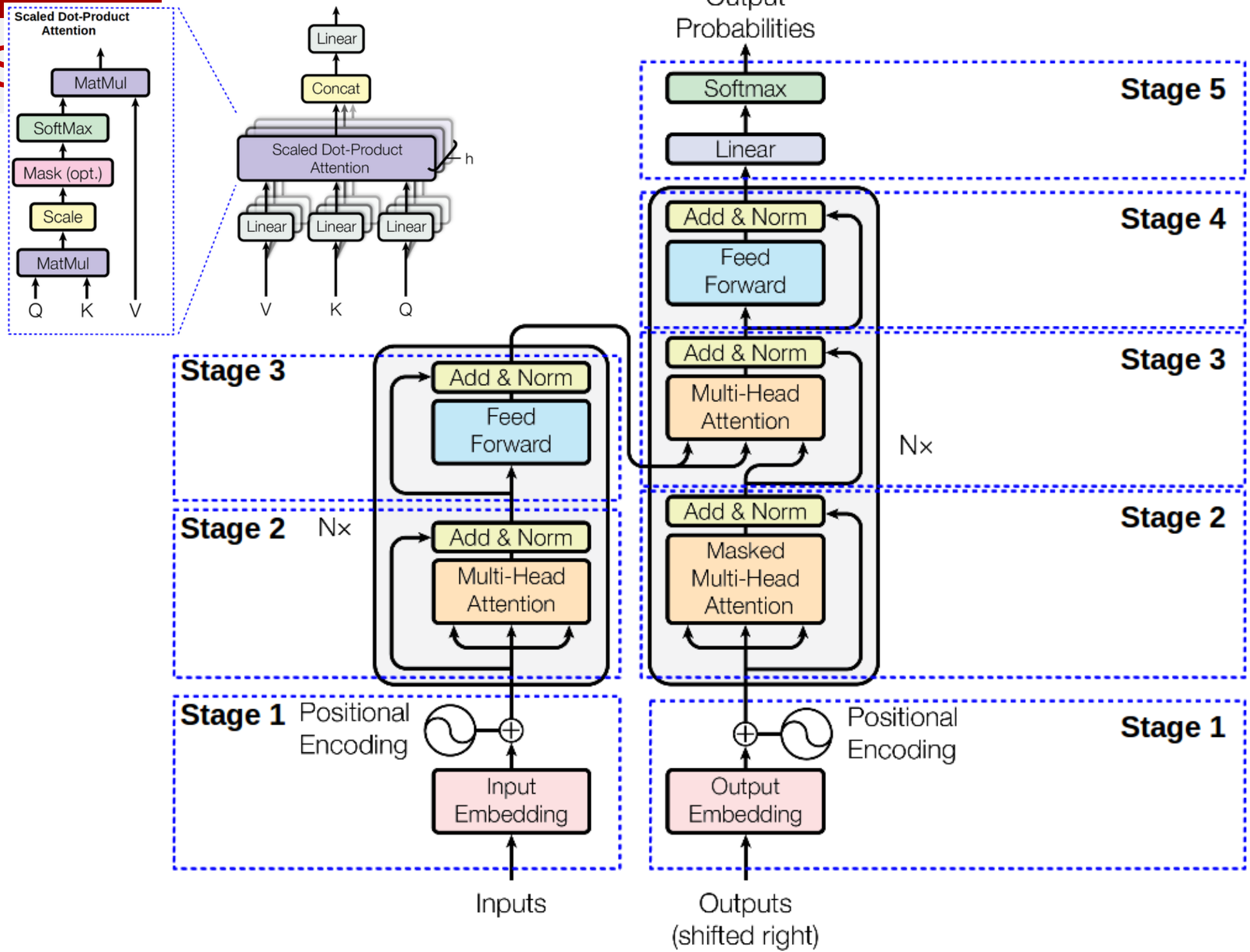
The dominant sequence transduction models are based on complex recurrent orconvolutional neural networks in an encoder and decoder configuration. The best performing such models also connect the encoder and decoder through an attentionm …

☆ 99 被引用次数：244 相关文章 所有 9 个版本 »



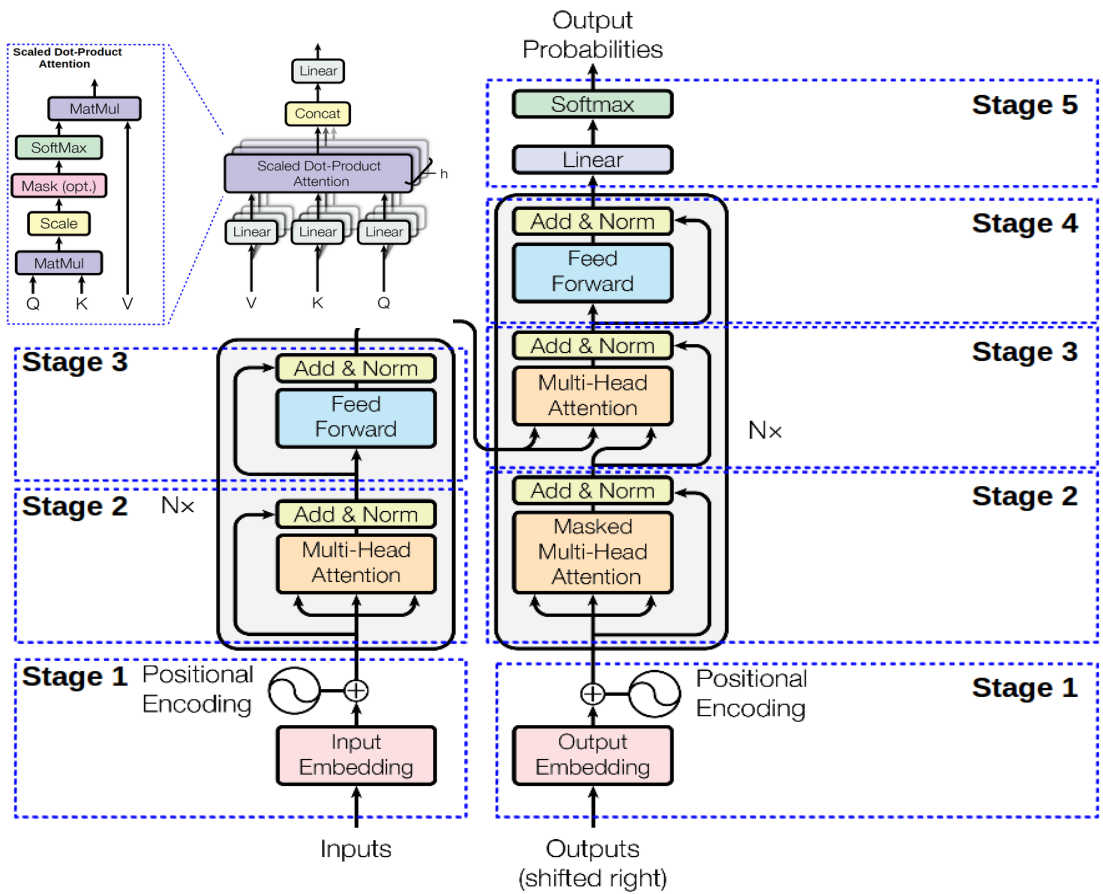
Self Attention




Scaled Dot-Product Attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q
Output Probabilities
Softmax
Linear
Stage 5
Add & Norm
Feed Forward
Stage 4
Add & Norm
Multi-Head Attention
Stage 3
Add & Norm
Masked Multi-Head Attention
Stage 2
Nx
Add & Norm
Feed Forward
Stage 3
Add & Norm
Multi-Head Attention
Stage 2
Nx
Positional Encoding
Input Embedding
Stage 1
Positional Encoding
Output Embedding
Stage 1
Inputs
Outputs (shifted right)

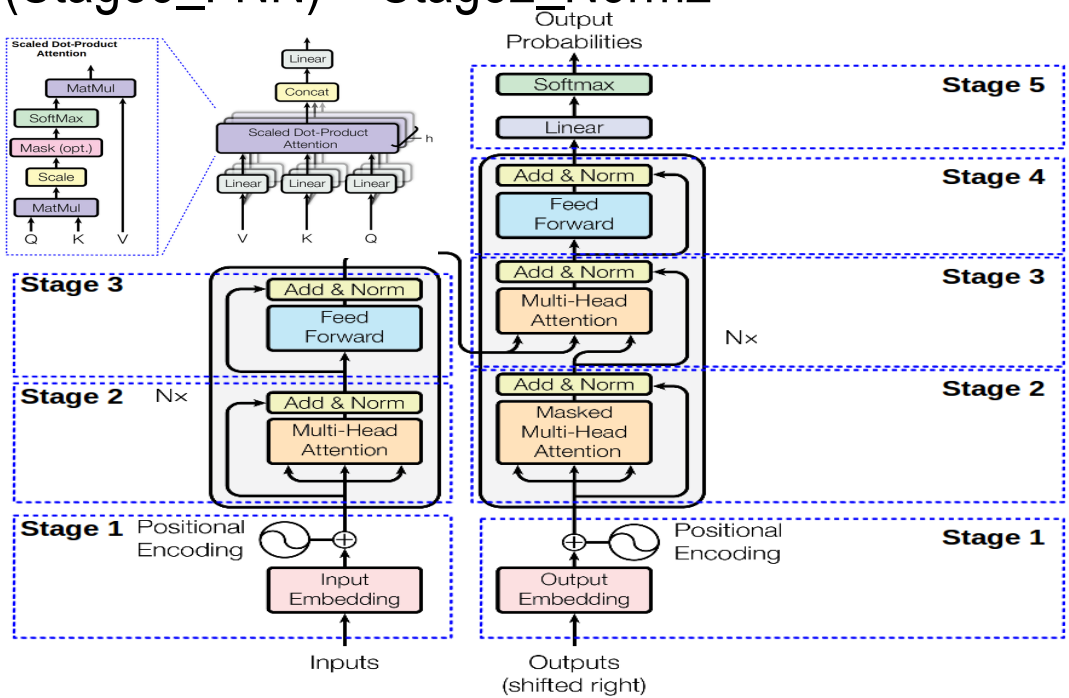
Self Attention

1. Stage1_out = Embedding512 + TokenPositionEncoding512
2. Stage2_out = layer_normalization(multihead_attention(Stage1_out) + Stage1_out)
3. Stage3_out = layer_normalization(FFN(Stage2_out) + Stage2_out)
4. out_enc = Stage3_out



Self Attention

1. Stage1_out = OutputEmbedding512 + TokenPositionEncoding512
2. Stage2_Mask = masked_multihead_attention(Stage1_out)
3. Stage2_Norm1 = layer_normalization(Stage2_Mask) + Stage1_out
4. Stage2_Multi = multihead_attention(Stage2_Norm1 + out_enc) + Stage2_Norm1
5. Stage2_Norm2 = layer_normalization(Stage2_Multi) + Stage2_Multi
6. Stage3_FNN = FNN(Stage2_Norm2)
7. Stage3_Norm = layer_normalization(Stage3_FNN) + Stage2_Norm2
8. out_dec = Stage3_Norm



Self Attention

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$Query$: $Q \in R^{n \times d_{model}}$

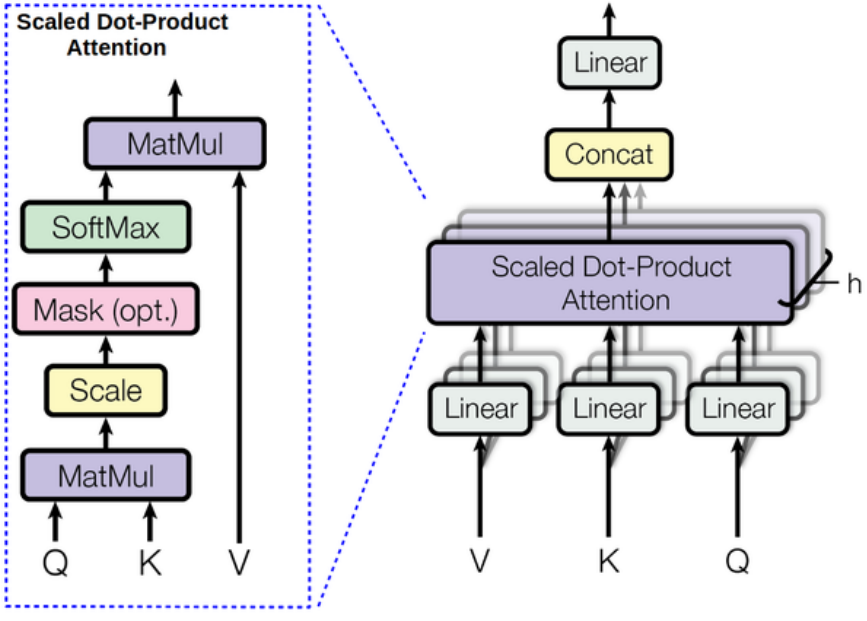
Key: $K \in R^{n \times d_{model}}$

Value: $V \in R^{n \times d_{model}}$

$W_i^Q \in R^{d_{model} \times d_k}$

$W_i^K \in R^{d_{model} \times d_k}$

$W_i^V \in R^{d_{model} \times d_v}$

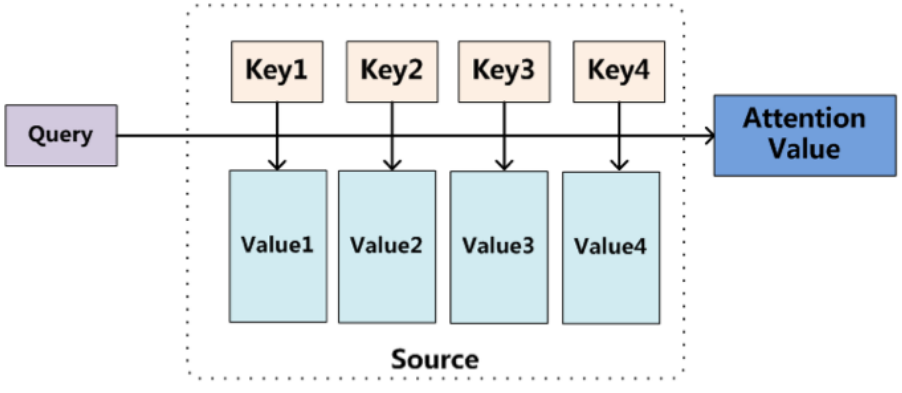


Self Attention

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})$$

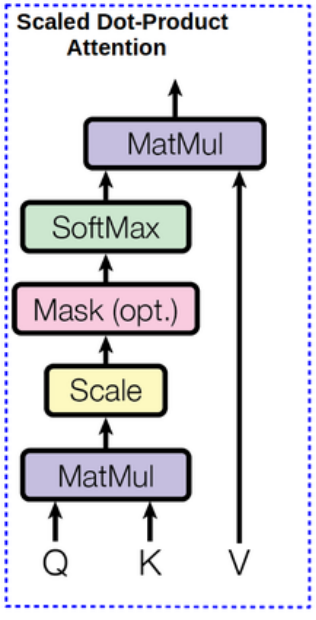
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

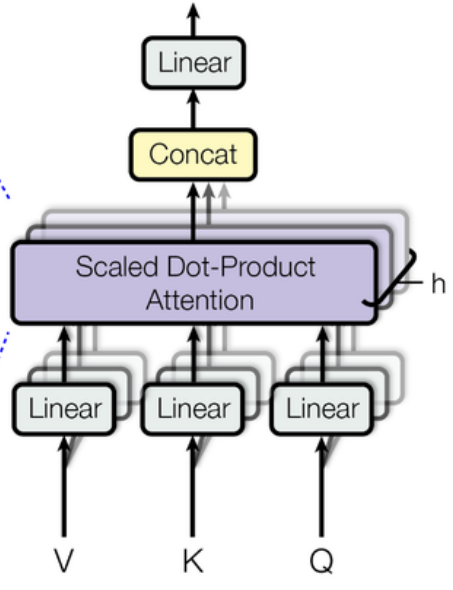
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



$$\textbf{Attention}(\textbf{Query}, \textbf{Source})$$

$$= \sum_{i=1}^{L_x} \textbf{Similarity}(\textbf{Query}, \textbf{Key}_i) * \textbf{Value}_i$$





Attention Visualizations

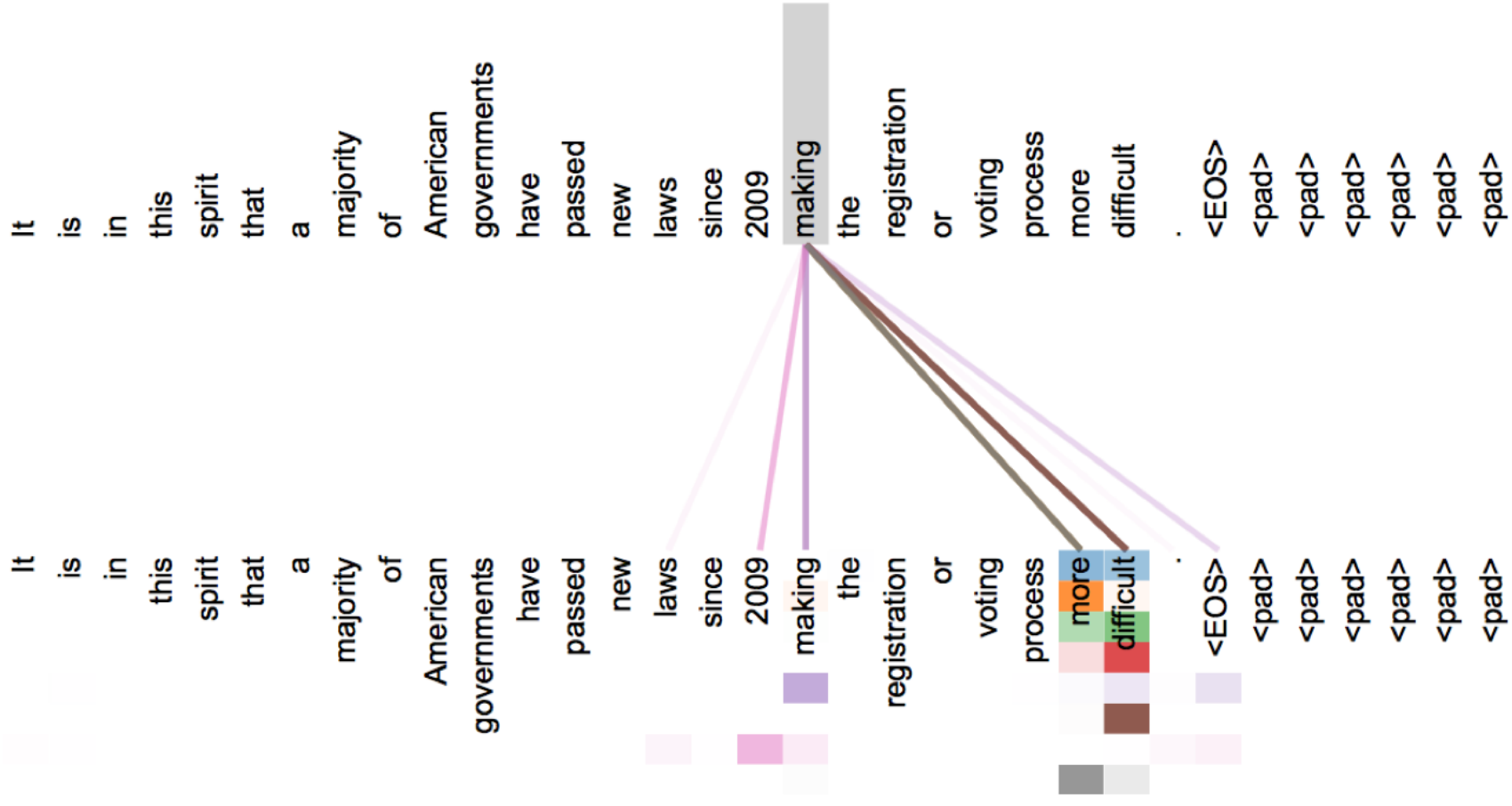


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.

- Hierarchical attention

[PDF] Hierarchical attention networks for document classification

Z Yang, D Yang, C Dyer, X He, A Smola… - Proceedings of the 2016 …, 2016 - aclweb.org

We propose a hierarchical attention network for document classification. Our model has two distinctive characteristics:(i) it has a hierarchical structure that mirrors the hierarchical structure of documents;(ii) it has two levels of attention mechanisms applied at the wordand sentence-level, enabling it to attend differentially to more and less important content when constructing the document representation. Experiments conducted on six large scale text classification tasks demonstrate that the proposed architecture outperform previous methods …

☆ 99 被引用次数：281 相关文章 所有 10 个版本 »

- Attention over attention

Attention-over-attention neural networks for reading comprehension

Y Cui, Z Chen, S Wei, S Wang, T Liu, G Hu - arXiv preprint arXiv …, 2016 - arxiv.org

Cloze-style queries are representative problems in reading comprehension. Over the past few months, we have seen much progress that utilizing neural network approach to solve Cloze-style questions. In this paper, we present a novel model called attention-over-attention reader for the Cloze-style reading comprehension task. Our model aims to place another attention mechanism over the document-level attention, and induces" attended attention" for final predictions. Unlike the previous works, our neural network model requires …

☆ 99 被引用次数：66 相关文章 所有 9 个版本 »

- Multi-step attention

Convolutional sequence to sequence learning

J Gehring, M Auli, D Grangier, D Yarats… - arXiv preprint arXiv …, 2017 - arxiv.org

The prevalent approach to sequence to sequence learning maps an input sequence to a variable length output sequence via recurrent neural networks. We introduce an architecture based entirely on convolutional neural networks. Compared to recurrent models, computations over all elements can be fully parallelized during training and optimization is easier since the number of non-linearities is fixed and independent of the input length. Our use of gated linear units eases gradient propagation and we equip each decoder layer with …

☆ 99 被引用次数：165 相关文章 所有 8 个版本 »

Thanks.