“自然语言处理导论”课程讲义

注意力机制

孙栩
信息科学技术学院
xusun@pku.edu.cn

内容

- □ 自然语言生成的深度学习解决方案
- □ 注意力机制
- □ Transformer

内容

- **自然语言生成的深度学习解决方案**
- **注意力机制**
- **Transformer**

NLG的深度学习方案

- **事实上，神经网络(NN)在NLG中的应用可以追溯至Kukich（1987），但受限于小规模样本**
- **随着计算机硬件的发展，NN再次兴起**
  - NN学习得到的表示被理解为具有语法和语义上的泛化[Mikolov et al., 2013; Luong et al., 2018; Pennington et al., 2014]
  - NN在序列建模也取得了成功[Bengio et al., 2003; Schwenk & Gauvain, 2005; Minh & Hinton 2007; Mikolov et al., 2010]
- **Sutskever et al. (2011)展示了RNN在NLG中的潜力**
  - 一个character-level LSTM可以生成合乎语法的英文句子
  - 当然这仅仅说明其在语言实现(language realization)上的能力
- **真正在NLG中的应用，依赖于Encoder-Decoder架构和Conditional Language Model的出现**

Encoder-Decoder框架

- Encoder-Decoder框架最早由Sutskever et al., 2014提出
  - Encoder RNN将输入编码为向量表示，作为Decoder RNN的额外输入
  - 这一架构非常适合序列到序列(Sequence-to-Sequence, Seq2Seq)类型的任务，例如机器翻译[Kalchbrenner & Blunsom, 2013; Bahdanau et al., 2015]

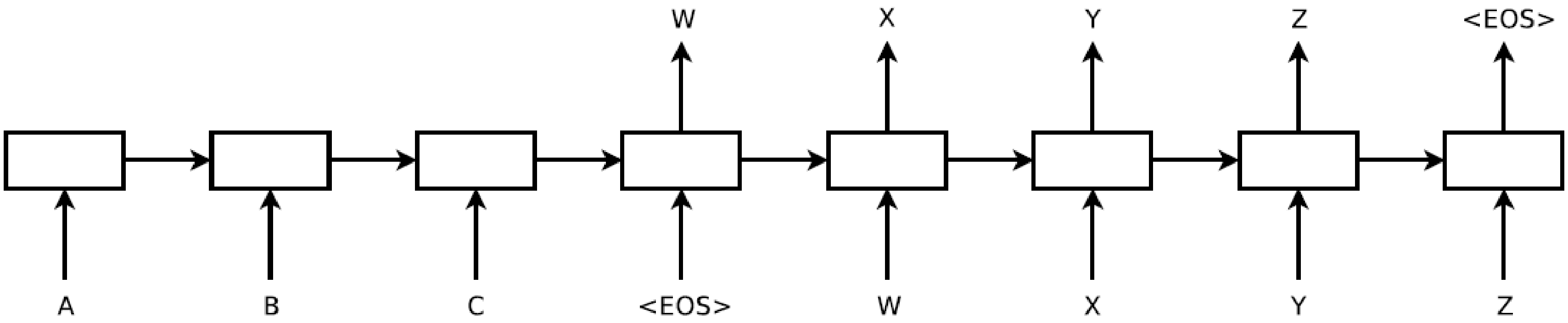
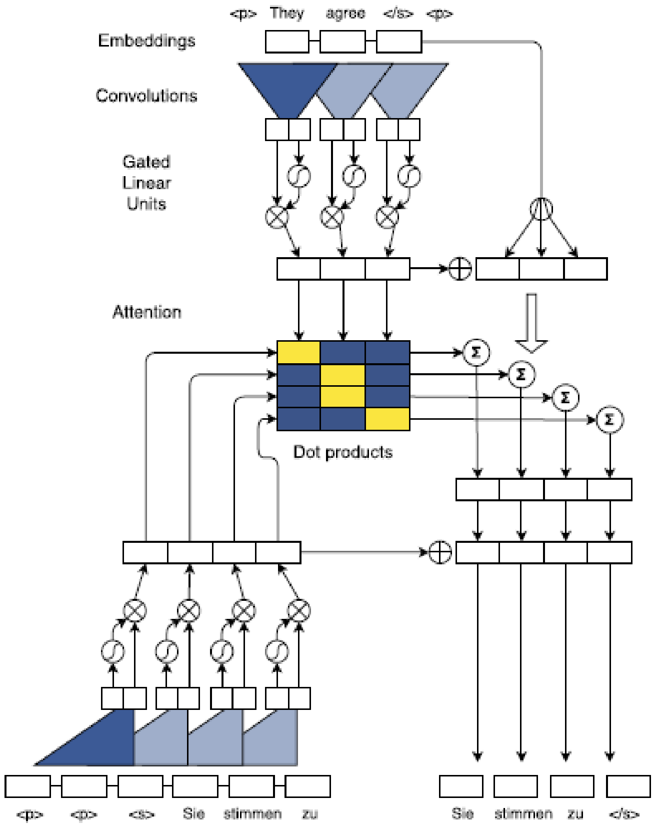


Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

Encoder-Decoder框架

- 随着技术发展，encoder和decoder不仅限于RNN，输入也不限于文本的形式
  - NN：CNN、RNN和MLP
  - Input: 图片、abstract meaning representations



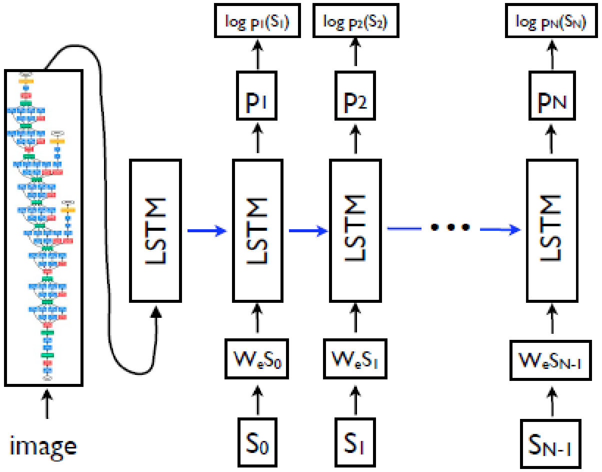


Figure 3. LSTM model combined with a CNN image embedder (as defined in [12]) and word embeddings. The unrolled connections between the LSTM memories are in blue and they correspond to the recurrent connections in Figure 2. All LSTMs share the same parameters.

Encoder-Decoder框架

- **更进一步，出现了基于注意力机制的模型[Bahdanau et al. 2015; Xu et al., 2015]**
  - 注意力学习到了输入表示到输出文本的松散耦合(loose couplings)

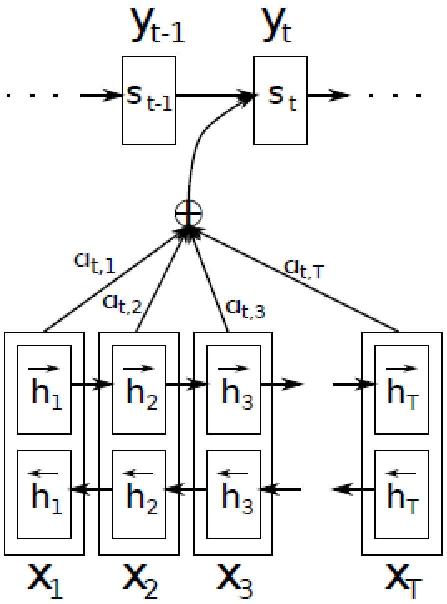
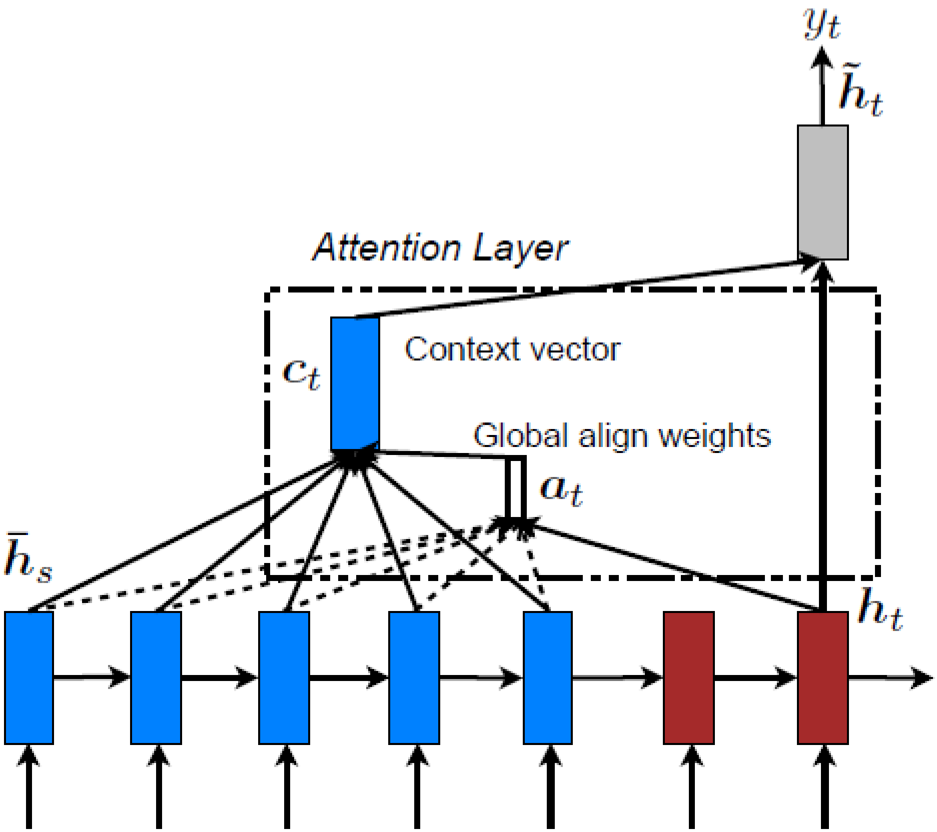


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

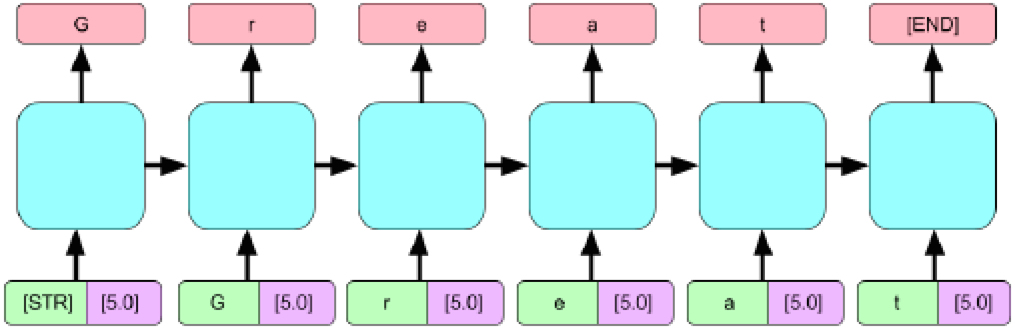


Conditional Language Model

- **Conditional Language Model与Encoder-Decoder框架的区别是**
  - Conditional Language Model的**输入直接为特征**，可包括语义、上下文或风格的属性
  - Encoder-Decoder框架强调 **编码**-解码 的过程，中间隐码的含义应当是习得的，而非指定的
- **任务示例**
  - Lebret et al. (2016)使用FFNN根据Wiki Infobox生成简介第一句话
  - Lipton et al. (2016)使用character-level RNN根据语义信息和情感生成评论
  - Tang et al. (2016)使用LSTM根据用户、地点等上下文生成评论
  - 其它的风格化或情感化生成任务[Li et al., 2016; Herzig et al., 2017; Asghar et al., 2017; Hu et al., 2017; Ficler & Goldberg, 2017]

Conditional Language Model

- **Conditional Language Model与Encoder-Decoder框架的区别是**
  - Conditional Language Model的输入直接为特征，可包括语义、上下文或风格的属性
  - Encoder-Decoder框架强调 **编码**-解码 的过程，中间隐码的含义应当是习得的，而非指定的
- **任务示例**
  - Lipton et al. (2016)使用character-level RNN根据语义信息和情感生成评论



内容

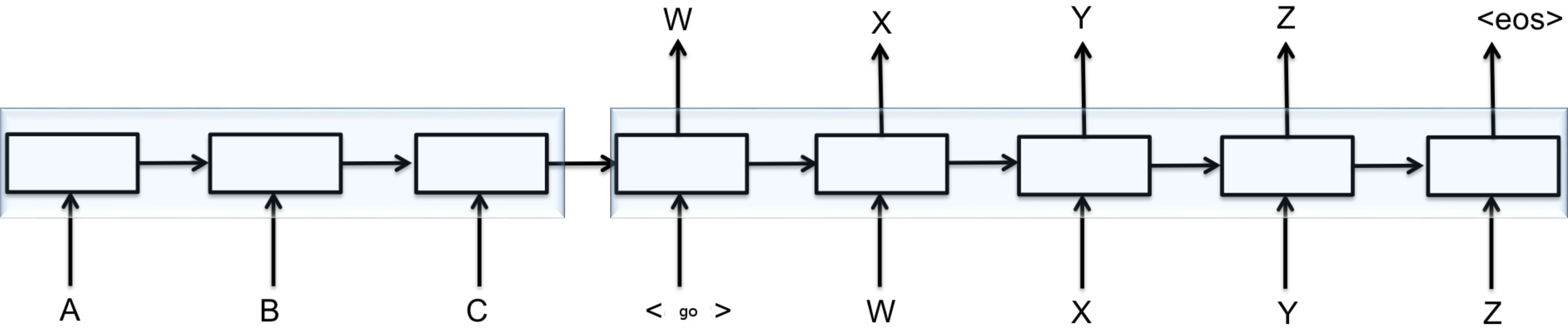
- □ 自然语言生成的深度学习解决方案
- □ 注意力机制 ←
- □ Transformer

内容

- **涉及到文本生成的任务效果近年来显著提升**
  - 语义表示改进：Deep Neural Networks
  - 序列建模改进：Recurrent Neural Networks
  - 语言生成改进：Neural Language Models
- **然而现有技术仍有很多缺陷**
  - 长序列建模效果仍然不佳
  - 数据稀疏问题仍需要进一步缓解
- **Attention技术应运而生**
  - 序列到词建模作为序列到序列建模的补充
  - 额外的输入信号来源，有效缩短了输出到输入依赖的距离

背景

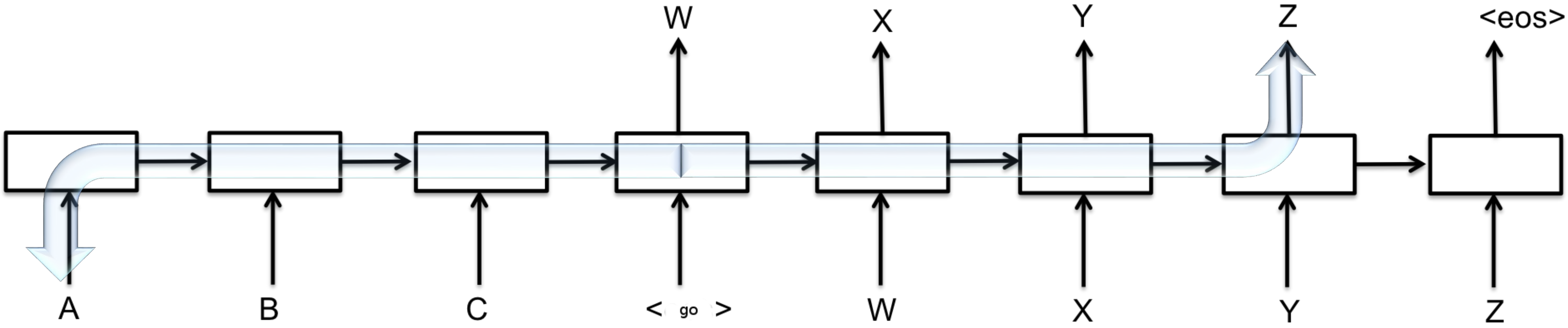
- **Encoder-Decoder框架，尤其是Sequence-to-sequence范式的问题**



- **映射以序列整体为单位，严重的数据稀疏问题**
  - 1，距离过长
  - 2，循环参数w表达能力不足

背景

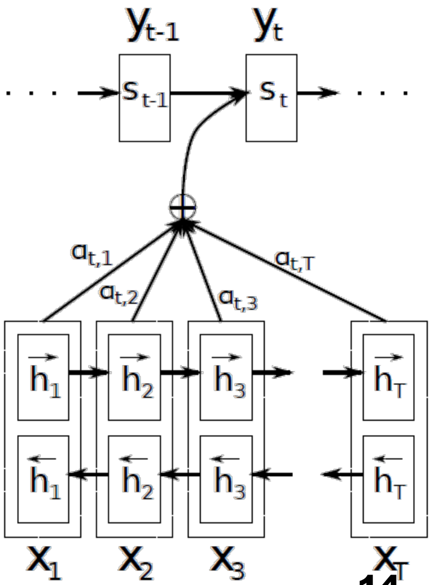
- Encoder-Decoder框架，尤其是Sequence-to-sequence范式的问题



- 输入到输出依赖的距离会相当长
  - 基于反向传播的学习很难学习
  - 之前的技巧：将输入序列颠倒，放弃建模过长的依赖
    - 但是无法根本解决问题

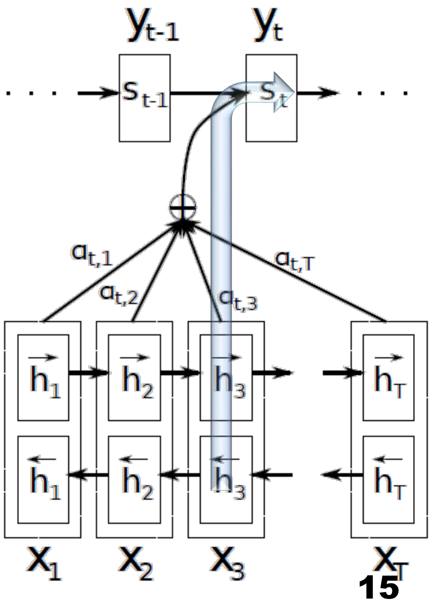
Attention

- **最早由Bahdanau et al.于2014年提出，用于自然语言处理的机器翻译任务，发表于ICLR 2015**
- **整体思路非常简单**
  - 目标序列的每步额外增加来自源序列的信号
  - 信号为源序列每步输出的加权平均
- **一般被译为“注意力”**
- **大致思路：**
  - s是目标状态，可以是ht，ht-1，或h和x的组合
  - 在这篇论文里，是ht-1
  - Source的ht和目标的ht拼接在一起，然后过一个MLP
  - 然后得到at，是一个实数
  - 然后所有输入得到一个归一化向量
  - Attention向量和输入逐个相乘，得到输入的向量表示
  - 从而目标端的ht得到除ht-1外的另一个输入



Attention

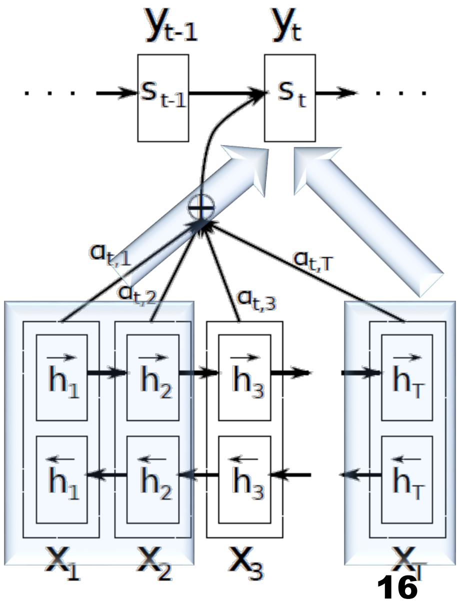
- **最早由Bahdanau et al.于2014年提出，发表于ICLR 2015**
- **整体思路非常简单**
  - 目标序列的每步额外增加来自源序列的信号
  - 信号为源序列每步输出的加权平均
- **通过attention可以解决前述的问题**
  - 依赖距离最短为1



Attention

- **最早由Bahdanau et al.于2014年提出，发表于ICLR 2015**
- **整体思路非常简单**
  - 目标序列的每步额外增加来自源序列的信号
  - 信号为源序列每步输出的加权平均

- **理想情况下，可以解决前述的问题**
  - 由于使用源序列加权，可以构建
    - 词到词映射
    - 短语到词映射
    - 离散片段到词映射
    - 序列到词映射

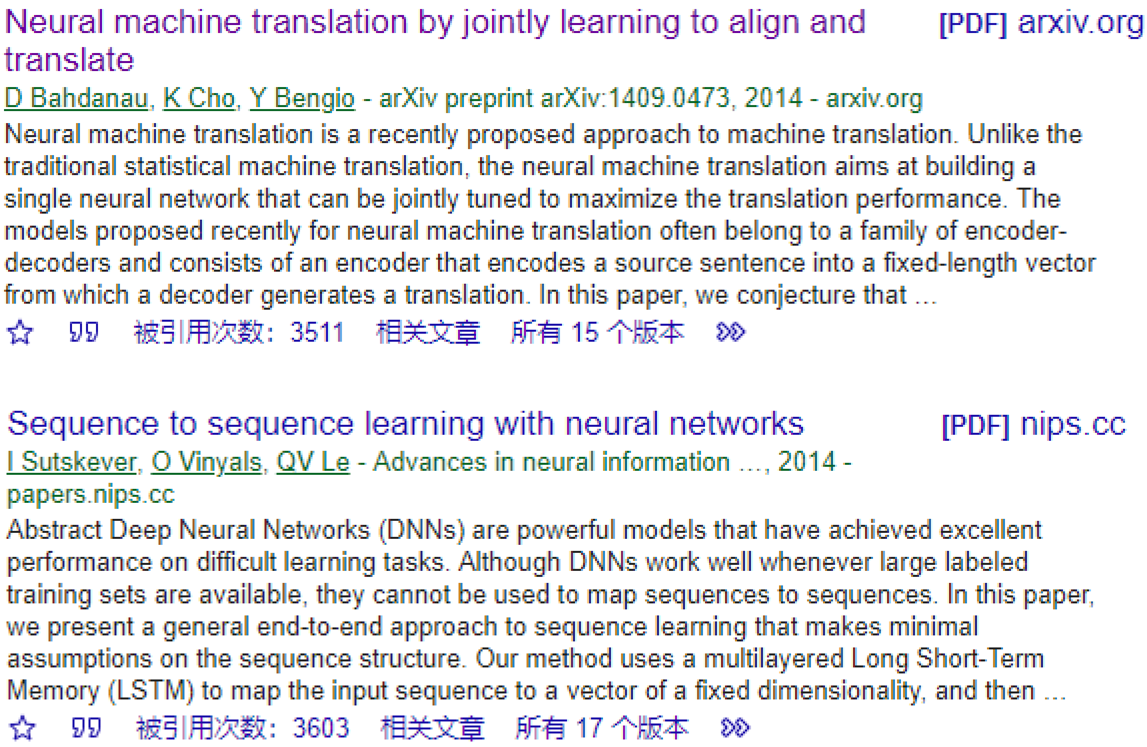


Attention

- **之后又出现了多种多样的attention，领域也不再限于序列到序列学习**
- **较为知名的有**
  - Stanford Luong et al. EMNLP 2015的global attention和local attention
  - UToronto & UMontreal 2015的visual attention
  - CMU MSR NAACL 2016的hierarchical attention
  - Google NIPS 2017的multi-head scaled dot-product attention和self attention
  - FAIR ICML 2017的mutli-step attention

Bahdanau Attention

- 最早的attention，用于机器翻译
  - 具有很高的影响力，和seq2seq的引用相当



Neural machine translation by jointly learning to align and translate [PDF] arxiv.org

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

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 translation performance. The models proposed recently for neural machine translation often belong to a family of encoder-decoders and consists of an encoder that encodes a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that ...

☆ ❞ 被引用次数：3511 相关文章 所有 15 个版本 »

Sequence to sequence learning with neural networks [PDF] nips.cc

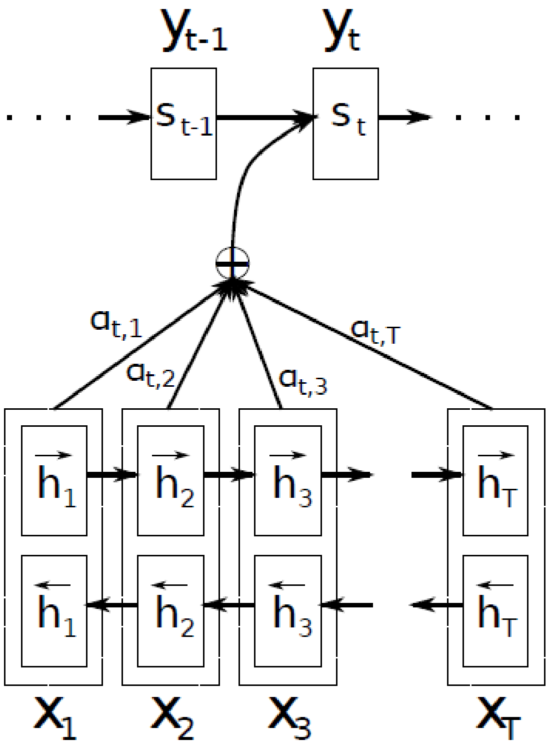
I Sutskever, O Vinyals, QV Le - Advances in neural information ..., 2014 - papers.nips.cc

Abstract Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then ...

☆ ❞ 被引用次数：3603 相关文章 所有 17 个版本 »

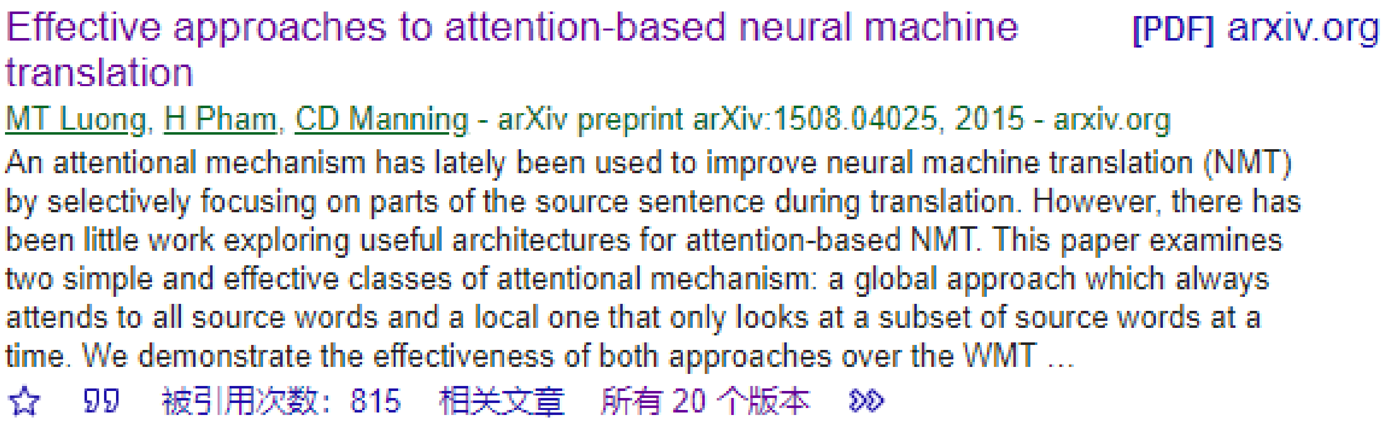
Bahdanau Attention

- **特点**
  - Attention作为LSTM输入
  - 使用前一时刻的LSTM输出查询
- **Attention计算公式（之前步骤一样）：**
- $score(s_{t-1}, h_i) = v^T \tanh(W s_{t-1} + U h_i)$
  - 括号里，拼接后乘以MLP矩阵等价于分别乘以矩阵再相加，v是为了将向量转为scalar
- **相当于用一个全连接网络计算分数**
- **可能的时刻不匹配问题：按理来说应该用st但是st还没有出来，只能用st-1补偿**



Luong Attention

- **提出了global attention和local attention用于机器翻译**
  - Global attention跟Bahdanau attention很接近，就是先算出st，然后代替st-1做attention，然后作为st的输出的输入
  - Local attention计算一个焦点，然后再焦点附近设定一个窗口，窗口内部算global attention



Effective approaches to attention-based neural machine translation [PDF] arxiv.org

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

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. This paper examines two simple and effective classes of attentional mechanism: a global approach which always attends to all source words and a local one that only looks at a subset of source words at a time. We demonstrate the effectiveness of both approaches over the WMT ...

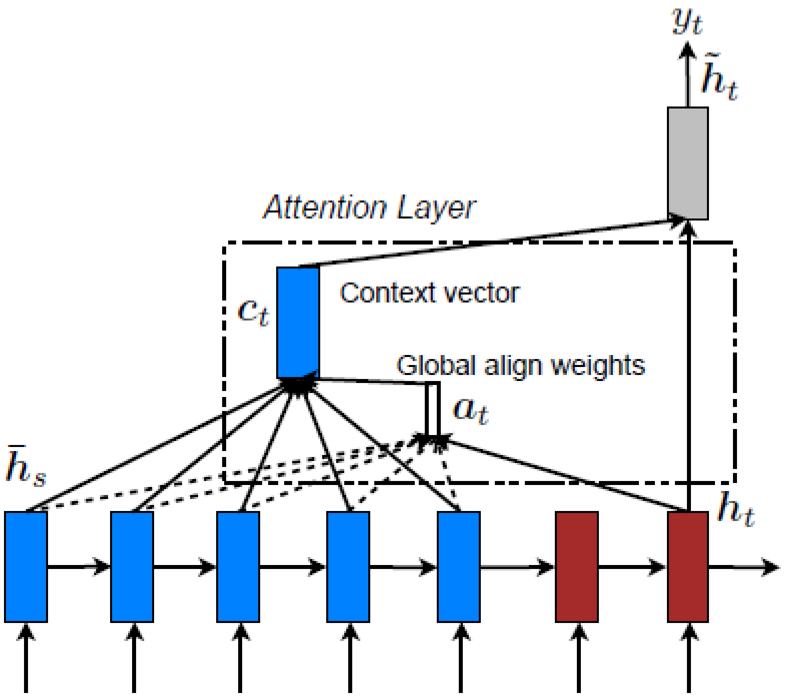
☆ ❞ 被引用次数：815 相关文章 所有 20 个版本 ≫

Luong Global Attention

- **特点**
  - Attention作为输出层输入
  - 使用当前时刻的LSTM输出查询
  - 提出了三种alternative算法：内积，带参数内积，类Bahdanau attention

- $$score(s_t, h_i) = \begin{cases} h_i^T s_t \\ h_i^T W s_t \\ v^T \tanh(W[h_i, s_t]) \end{cases}$$

- **分别命名为dot, general, concat**
  - Concat与Bahdanau的一致



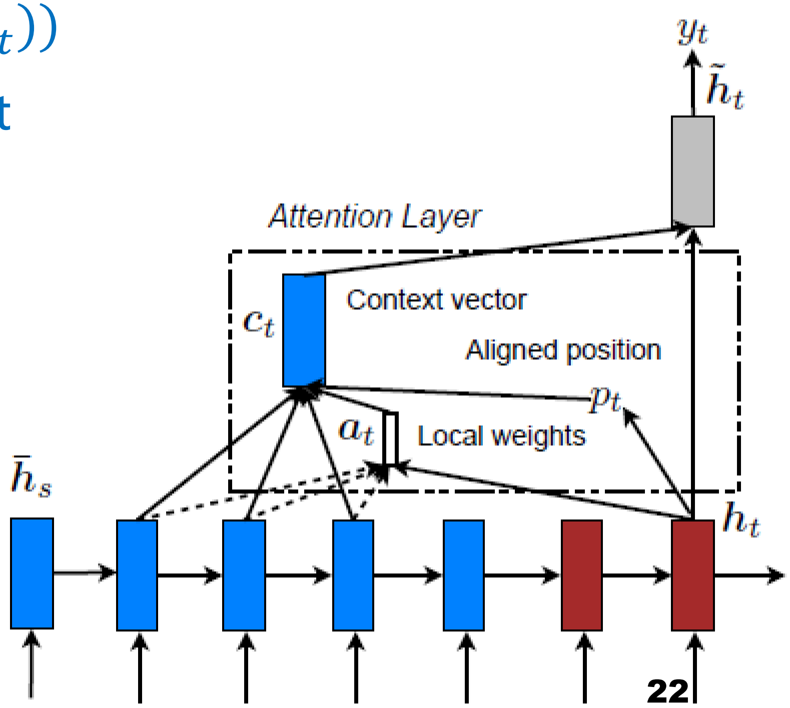
Luong Local Attention

特点

- 不对整个源序列做attention，只针对一个局部
- 先预测一个焦点位置p_t，在经验设定的窗口(单侧10)内做attention

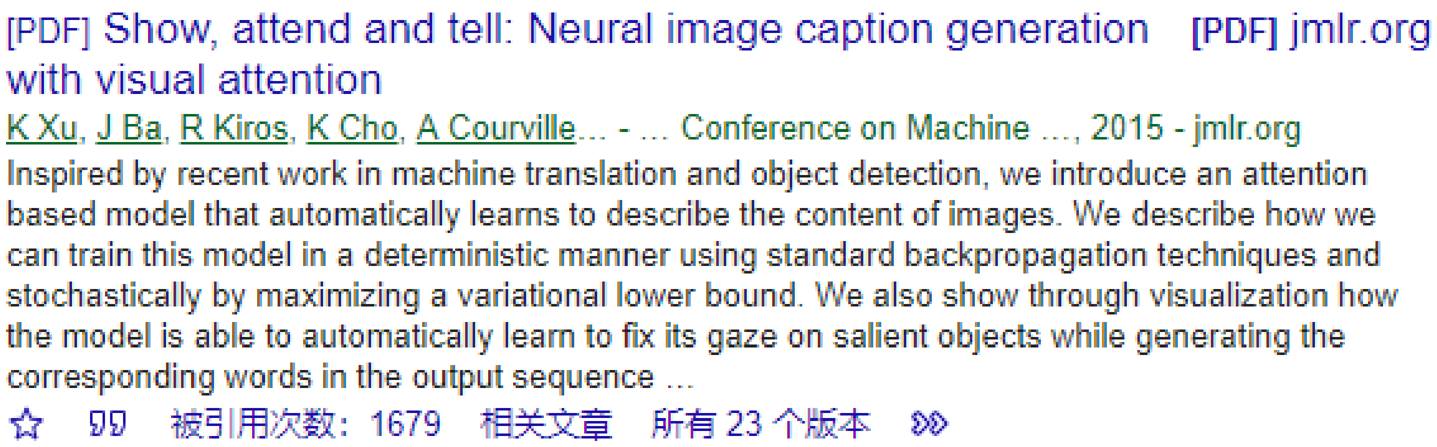
预测方法

- Monotonic焦点的简单映射：$p_t = t$
- Predictive: $p_t = S\, sigmoid(v \tanh(W s_t))$
  - Tanh（）可以理解为一个MLP，只跟st相关，跟输入无关
  - 乘以v得到标量
  - 过sigmoid得到0到1之间的值
  - 乘以S（源句子长度）得到焦点



Visual Attention

- 提出visual attention应用在**自然语言处理中的图像标题生成任务**
  - ICML 2015
  - 提出了hard attention，对attention的离散化，转成一个指针
  - 提出了soft attention，反向的attention归一化



[PDF] Show, attend and tell: Neural image caption generation with visual attention [PDF] jmlr.org

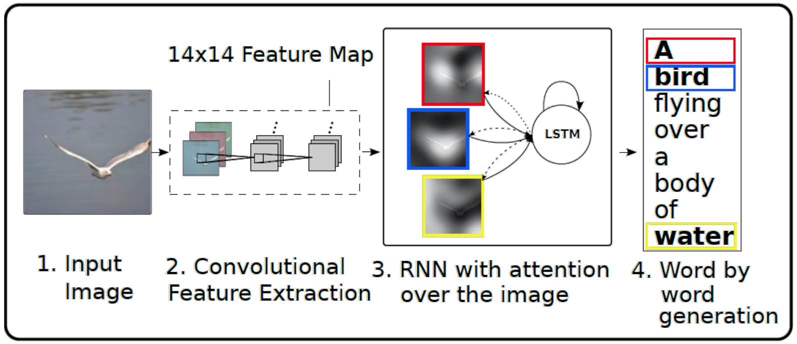
K Xu, J Ba, R Kiros, K Cho, A Courville... - ... Conference on Machine ..., 2015 - jmlr.org

Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence ...

☆ 99 被引用次数：1679 相关文章 所有 23 个版本 ≫

Visual Attention

- **特点**
  - Attention作为LSTM输入
- **Hard Attention**
  - 只学一个attention，只attend到图像的一个区域/feature
  - 根据attention分布，采样一个向量
  - 通过强化学习训练
  - Attention更集中

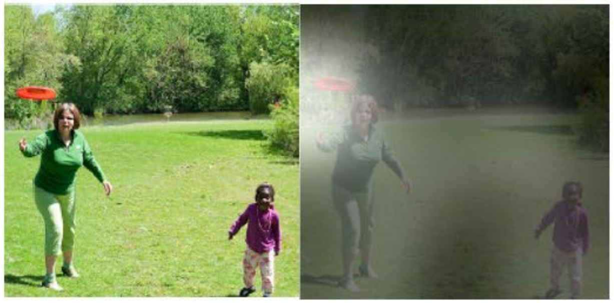

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

- **Soft Attention**
  - 额外约束 $\sum_t \alpha_{ti} \approx 1$，使描述更丰富（原来是对i求和，现在是对t求和)
  - 使得输入端没有被注意到的东西更容易被注意到

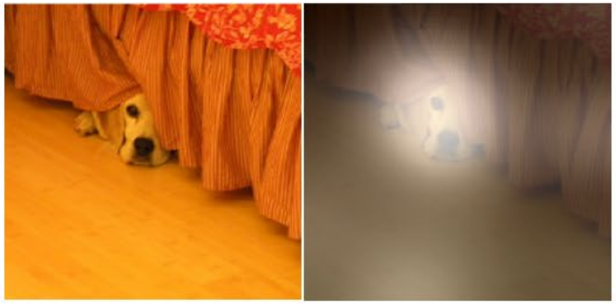
Visual Attention

Grounded Language Generation

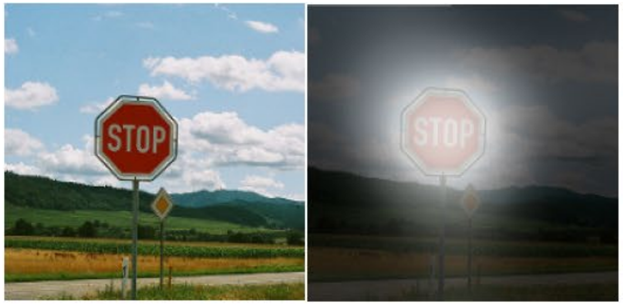
- attention学习到了物体和语言的联系



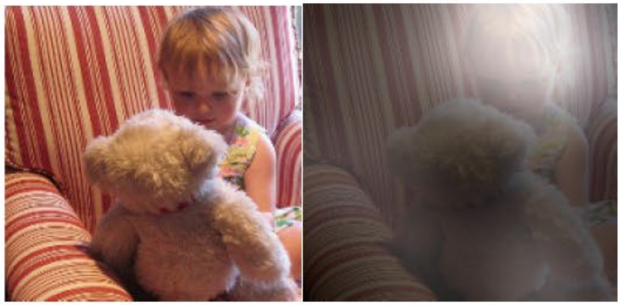
A woman is throwing a frisbee in a park.



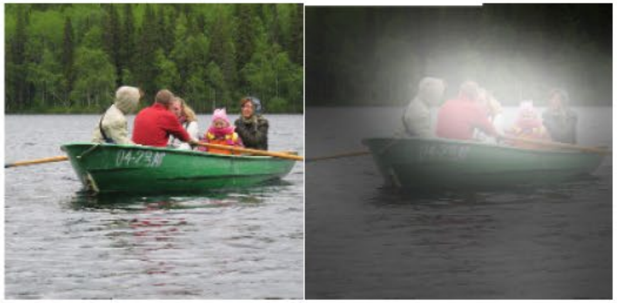
A dog is standing on a hardwood floor.



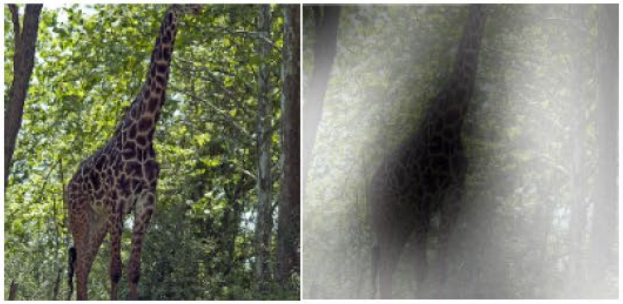
A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



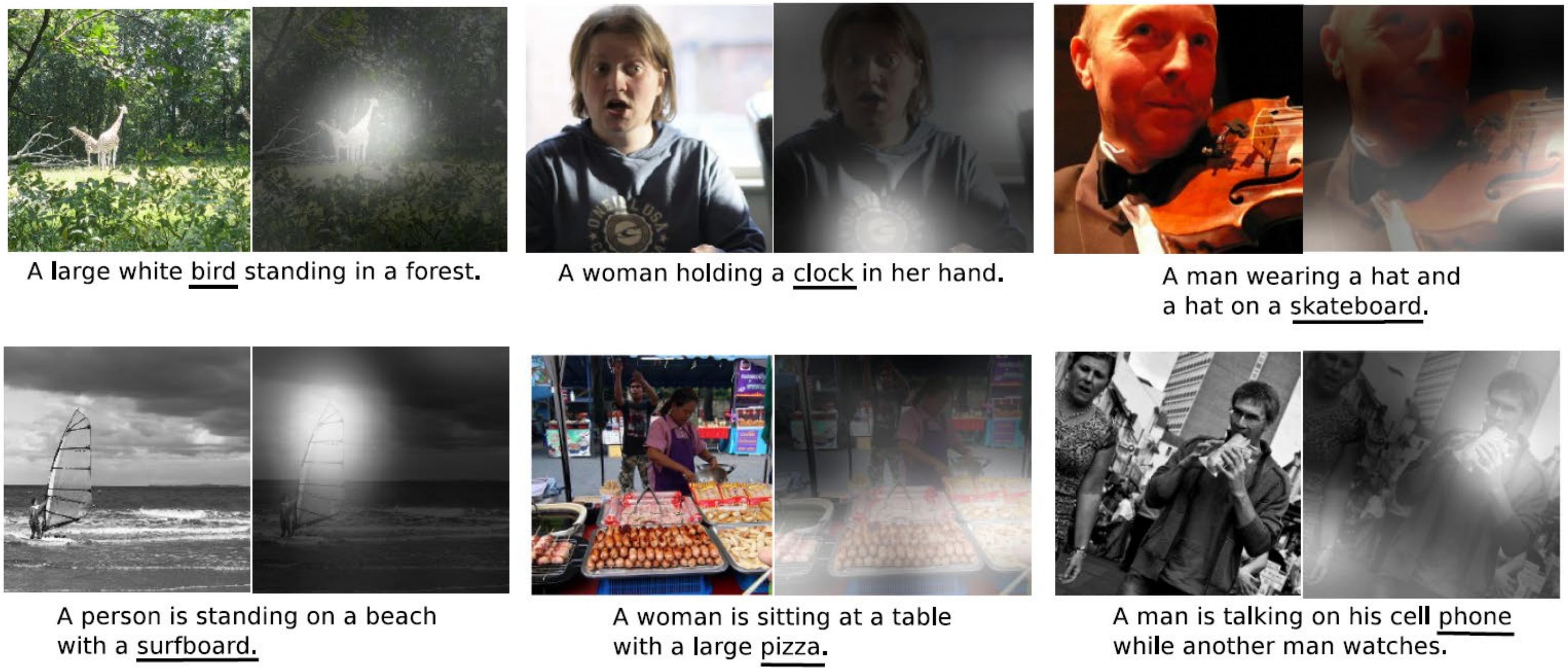
A giraffe standing in a forest with trees in the background.

Visual Attention

Grounded Language Generation

- Insight of mistakes

Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.



A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

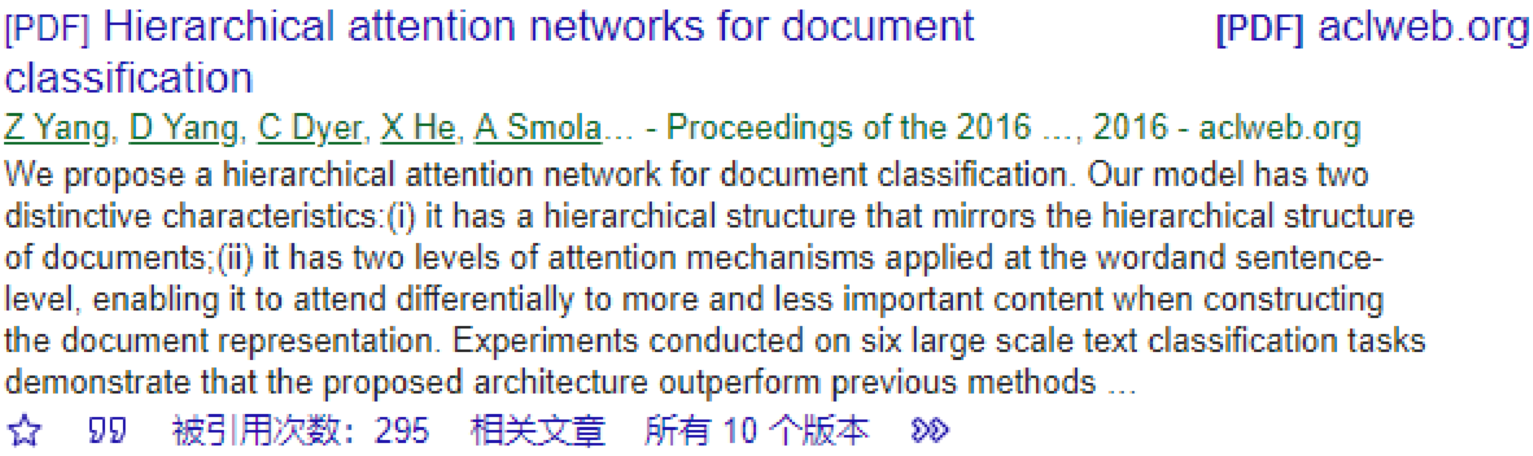
A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

Hierarchical Attention

- **提出该方法应用于自然语言处理的文档分类**
  - 跟Bahdanau attention很像的设定，但是层次化了
  - 这不是end2end模型了
  - 只分2层，词汇层attention、句子层attention
  - 因为没有目标状态量st了，所以使用了一个全局向量作为替代



[PDF] Hierarchical attention networks for document classification

[PDF] aclweb.org

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：295 相关文章 所有 10 个版本 »

Hierarchical Attention

- 特点
  - Attention作为输出层输入
  - 一种self-attention
- 计算比较特别
  - 但没有使用s_t而是用了额外的全局向量u（随机初始化然后更新学习）

$$
\begin{aligned}
u_{it} &= \tanh(W_w h_{it} + b_w) \\
\alpha_{it} &= \frac{\exp(u_{it}^\top u_w)}{\sum_t \exp(u_{it}^\top u_w)} \\
s_i &= \sum_t \alpha_{it} h_{it}.
\end{aligned}
$$

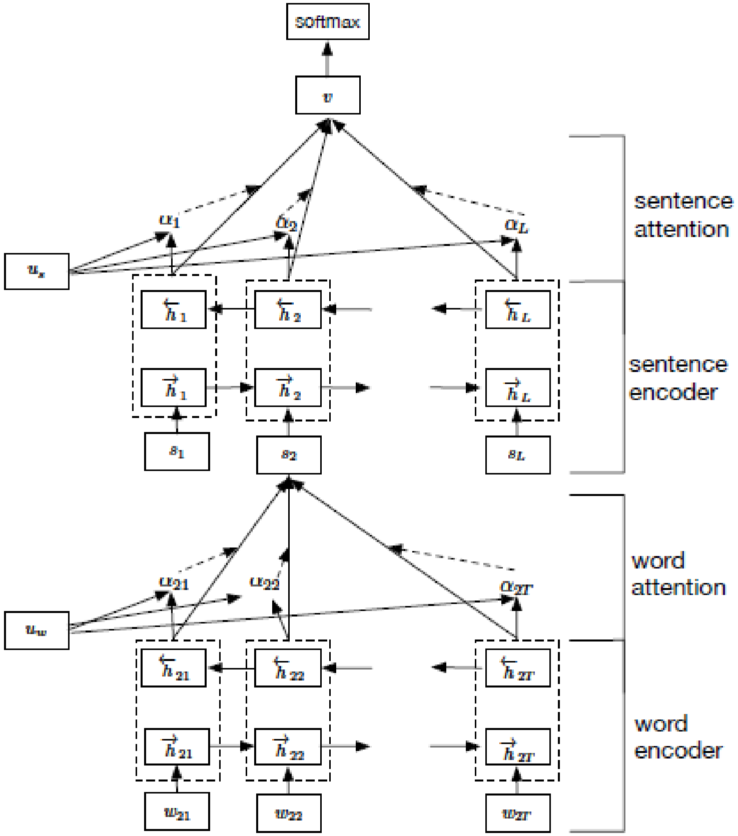
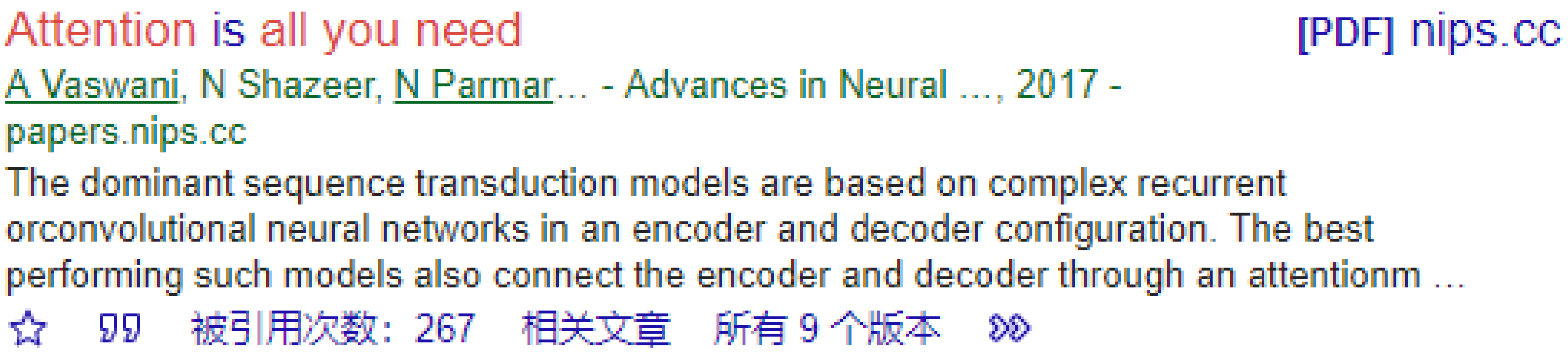


Figure 2: Hierarchical Attention Network.

Self-Attention

- **提出该方法用于机器翻译任务**
  - 主要是用self-attention替代了LSTM，设计了一个特殊的self-attention
  - 一般的attention是source到target之间到，self-attention是source内部或target内部的attention



Attention is all you need [PDF] nips.cc

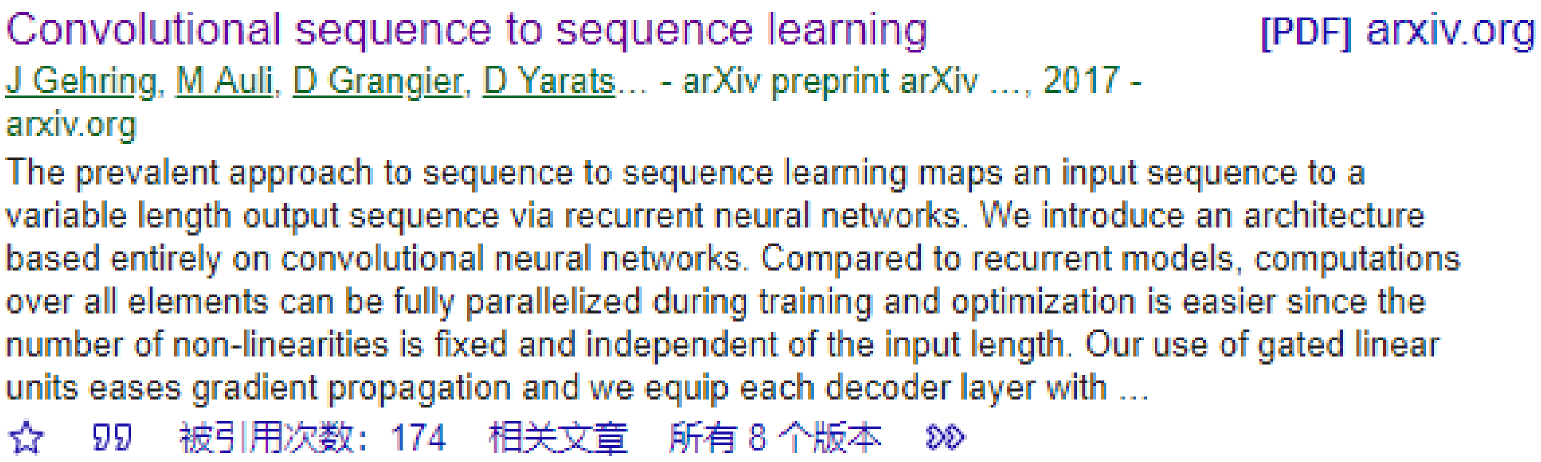
A Vaswani, N Shazeer, N Parmar… - Advances in Neural …, 2017 - papers.nips.cc

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 …

☆ ❞ 被引用次数：267 相关文章 所有 9 个版本 »

Multi-step Attention

- **ConvS2S**



Convolutional sequence to sequence learning [PDF] arxiv.org

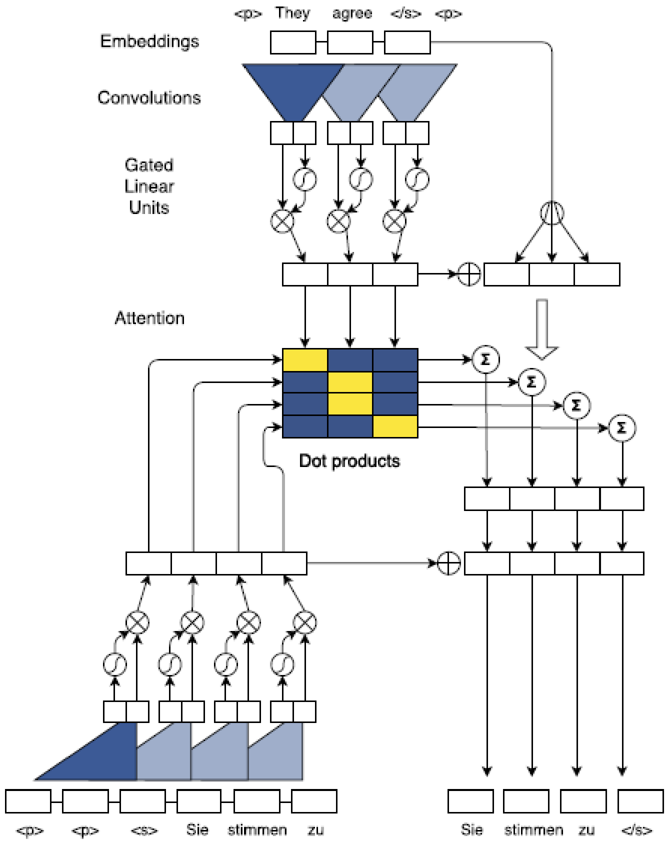
J Gehring, M Auli, D Grangier, D Yarats… - arXiv preprint arXiv …, 2017 - arxiv.org

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：174 相关文章 所有 8 个版本 »

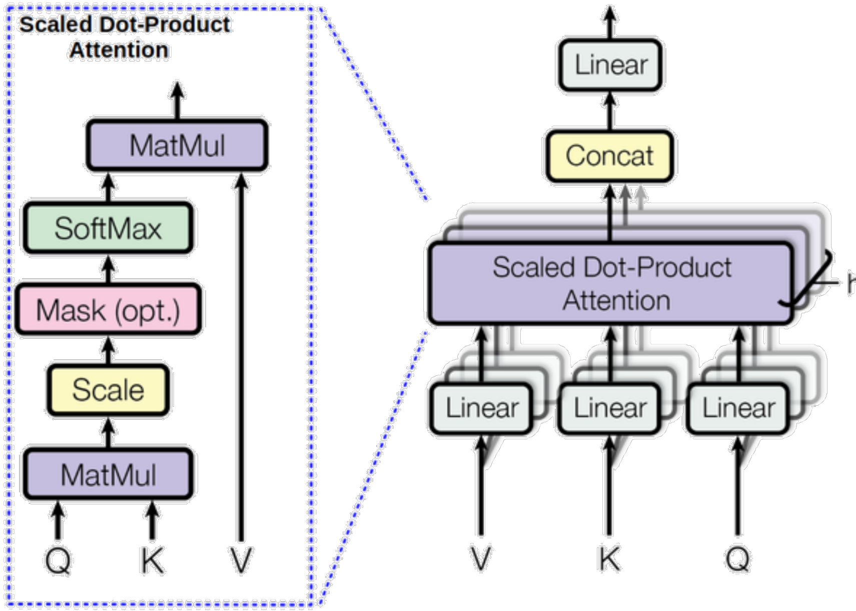
Multi-step Attention

- **特点**
  - Query为仿射后的decoder state +last output word
  - Key为encoder state
  - Value为encoder state + input word
  - Score为QK内积
    - 实际是decoder state的general
    - 再加last output word的dot
  - Value加权
- **每层decoder都加attention**



Self-Attention

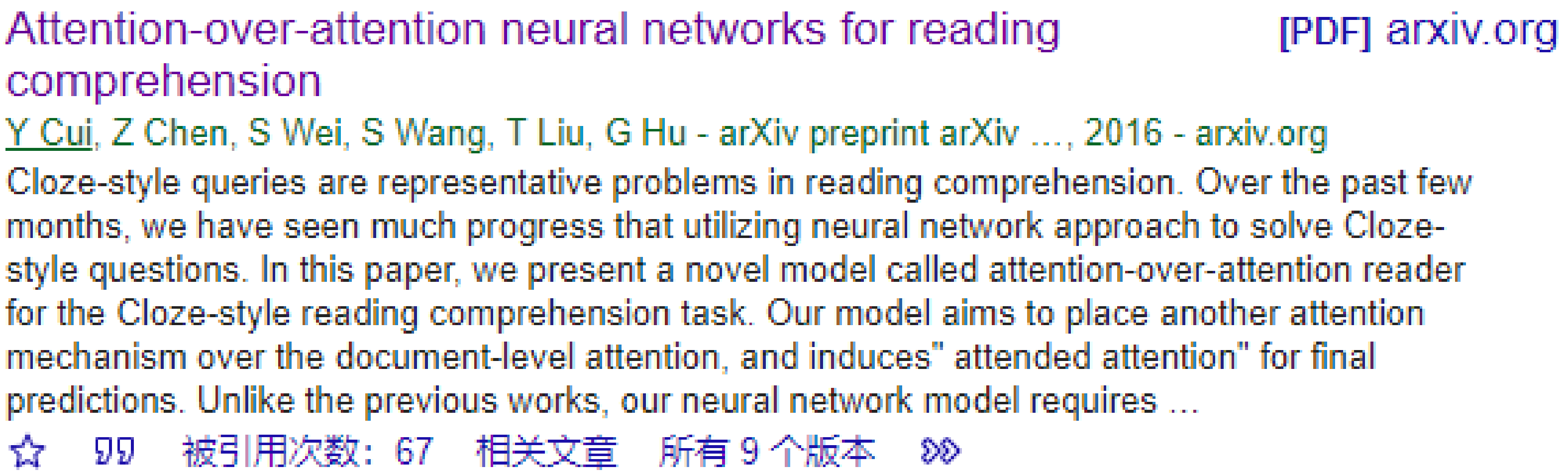
- **特点**
  - 看作一个通用方法，不区分输入和输出
  - 所有都采用QKV模块
  - Q代表查询量，接近St
  - K代表输入量，接近ht
  - V代表ht的一个copy
  - 思路：Q和K内积，然后得到一个相似度，然后乘以V
  - 有三种组合：1，QK都是输入词向量；2，QK都是输出；3，Q是输出、K是输入



Attention over Attention

用于阅读理解

- ACL 2017



Attention-over-attention neural networks for reading comprehension [PDF] arxiv.org

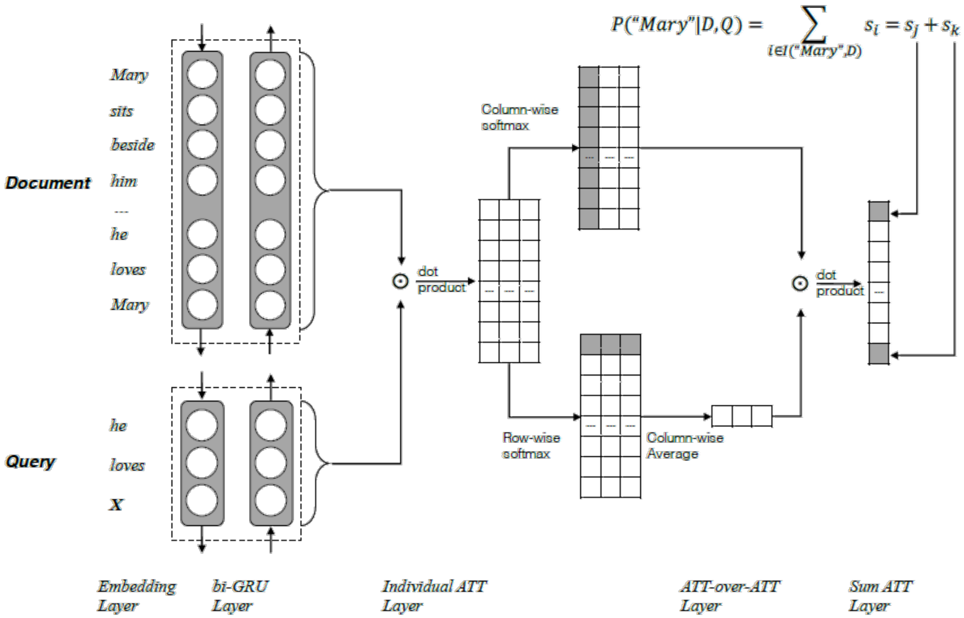
Y Cui, Z Chen, S Wei, S Wang, T Liu, G Hu - arXiv preprint arXiv …, 2016 - arxiv.org

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 …

Attention over Attention

思路

- 每个文档和问题的词算内积，得到了一个score矩阵
- 每行针对doc词，每列针对query词
- 每列作softmax：和问题最相关的文档词,
- 每行作softmax：和文档最相关的问题词
- 两者每行作内积，得到的是attention over attention



内容

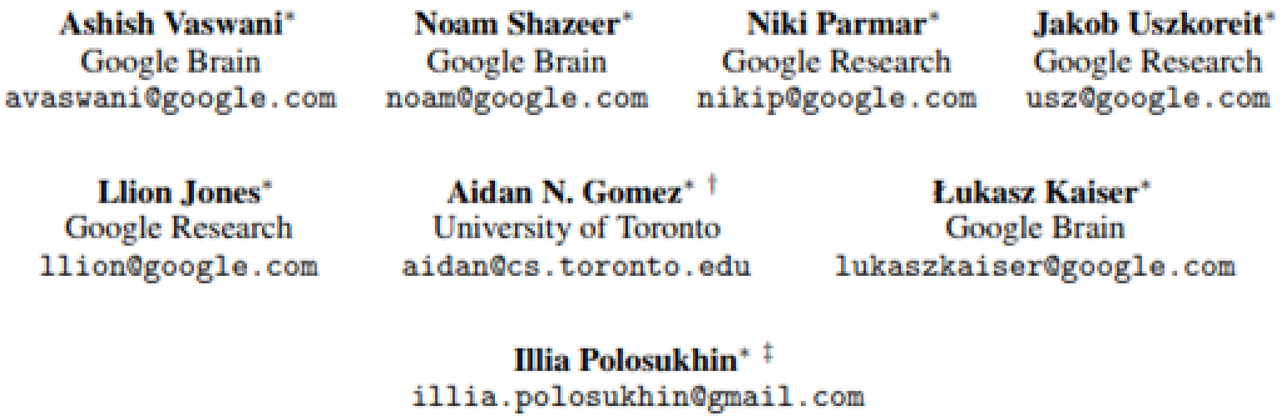
- 自然语言生成的深度学习解决方案

- 注意力机制

- Transformer

Attention is all you need

- **NIPS 2017**
- **Work from Google Brain and Google Research**
  - 8 authors are all first authors



Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*][†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

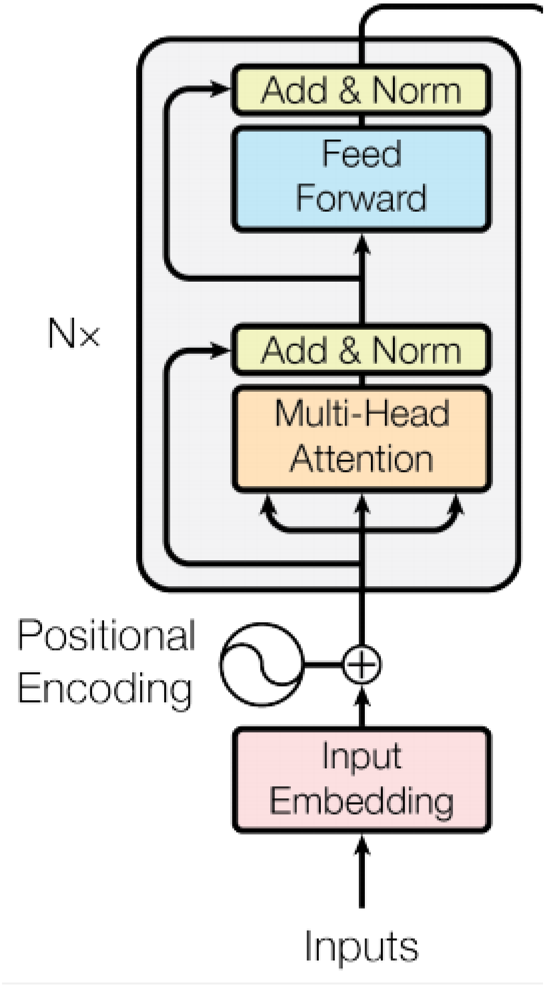
Illia Polosukhin[*][‡]
illia.polosukhin@gmail.com

Attention is all you need

- **Goal: replacing RNNs**
  - FAIR: convS2S, replacing RNNs with CNNs
- **Core idea**
  - Self-attention
  - Scaled dot-product attention
  - Multi-head attention
  - Parameter-free position representation
- **Not a easy job**

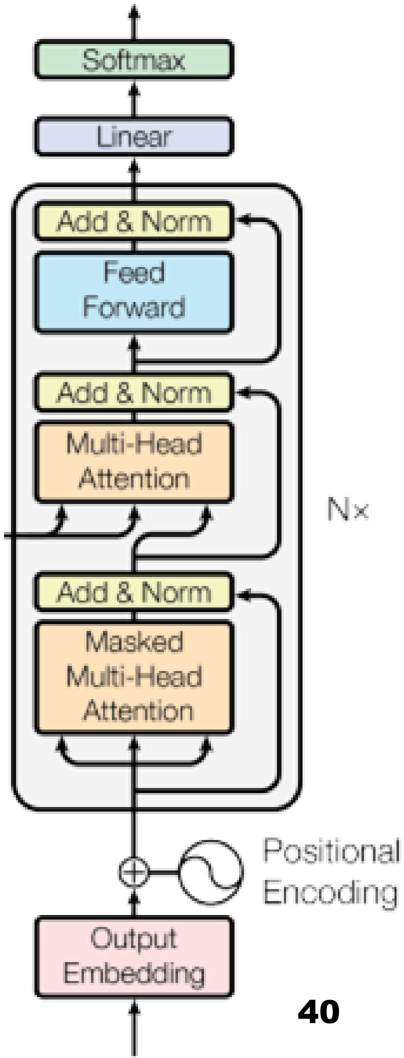
Encoder

- **6 identical blocks**
- **Each has 2 sub-layers**
  - Multi-head self-attention mechanism
  - Position-wise fully connected network
- **Residual connection and layer norm.**
  - LayerNorm(x + Sublayer(x))
- **Same output dimension: 512**



Decoder

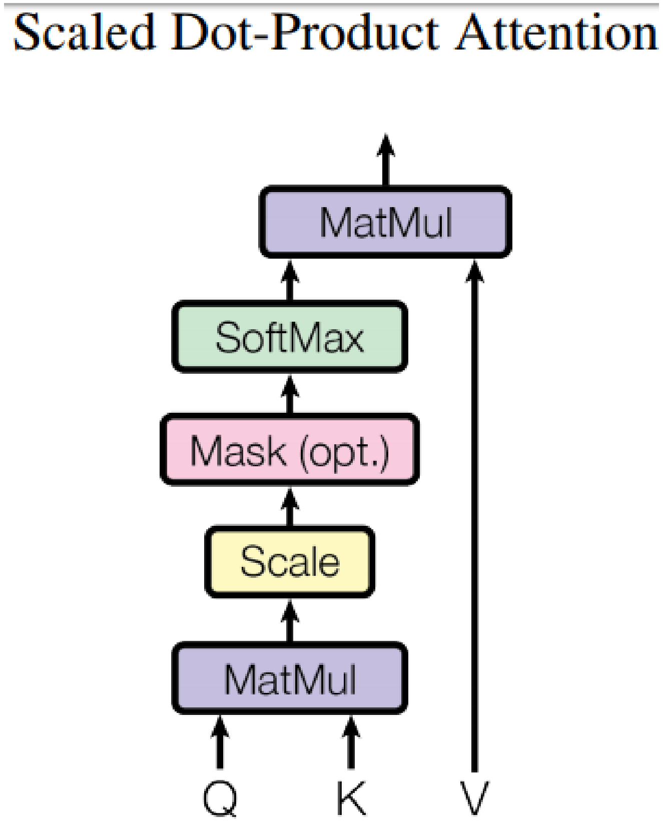
- 6 identical blocks
- Each has 3 sub-layers
- Third sub-layers
  - Additional layer performs multi-head attention over the output of the encoder stack
- Masked self attention
  - Prevent positions from attending to subsequent positions
  - Ensure no future predictions are used



Attention

- **Scaled Dot-Product Attention**
- **Input**
  - Queries of dk
  - Keys of dk
  - Values of dv

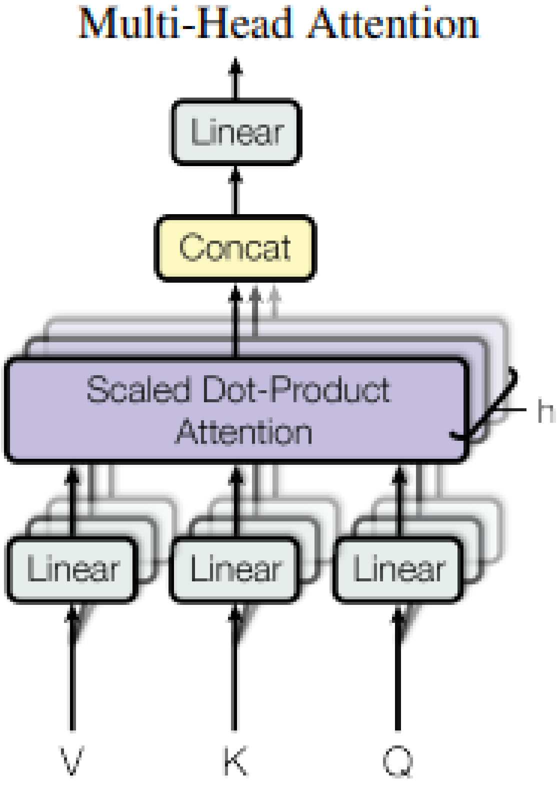
$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Attention

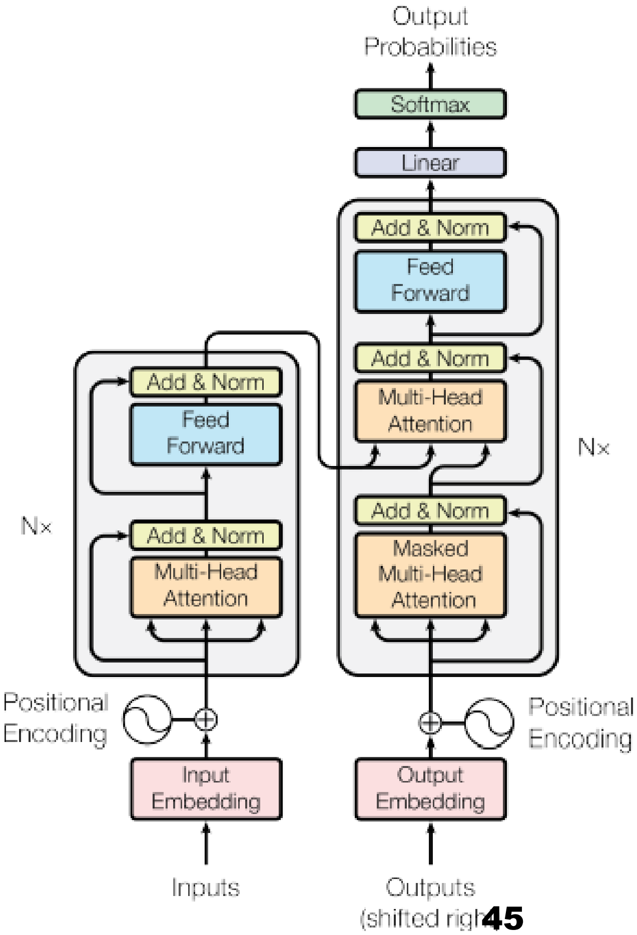
- **Multi-head attention**
  - Project the keys values and queries of d_{model} dimension h times to dk dk dv dimensions
  - Attend to information from different representation subspaces
  - Single head inhibits this

- **8 parallel heads**
  - dk dv = dmodel/h = 512/8 = 64



Attention

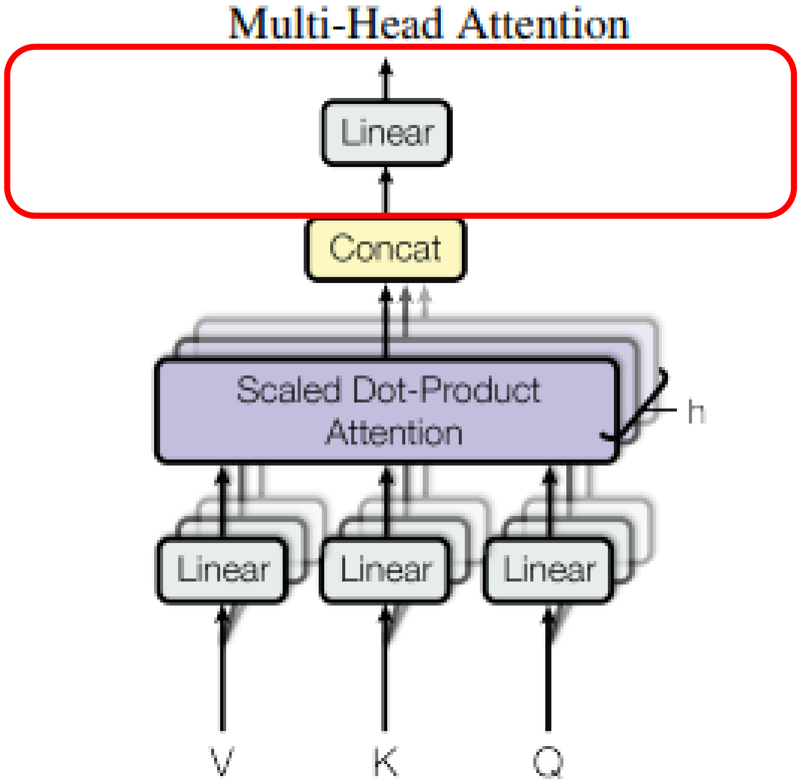
- **Where?**
- **Encoder – decoder: context attention**
  - Queries from previous decoder layer
  - Keys values from encoder outputs
- **Encoder: self attention**
  - Queries from previous encoder layer
  - Keys values from previous positions
- **Decoder: self attention**
  - Masking out inputs of the softmax which correspond to illegal connections



Position-wise Feed-Forward

- **Two linear transformations with a relu in between**
  - $\mathrm{FFN}(X) = \max(0, \boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2$

- **Input: 512**
- **Output: 512**
- **Inner: 2048**



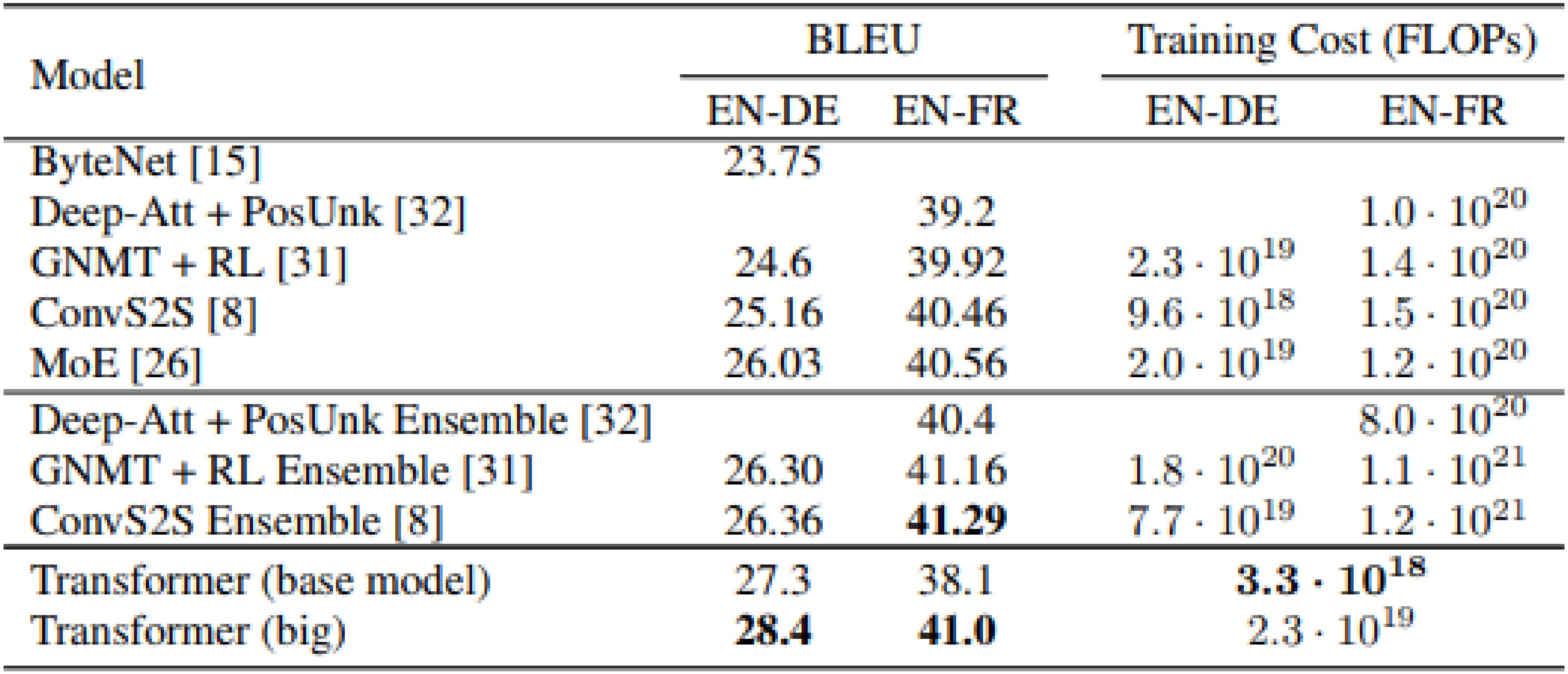
Others

- **Tied weights**
  - Embedding and pre-softmax linear transformation
- **In the embedding, the weights are multiplied by** $\sqrt{d_{model}}$
- **Positional Encoding**
  - Inject information about relative or absolute position
    - Since no recurrent connections are used
  - Fixed: use sine and cosine functions

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

  - Learned:

Results



Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Visualization

谓语与补语：making … more difficult

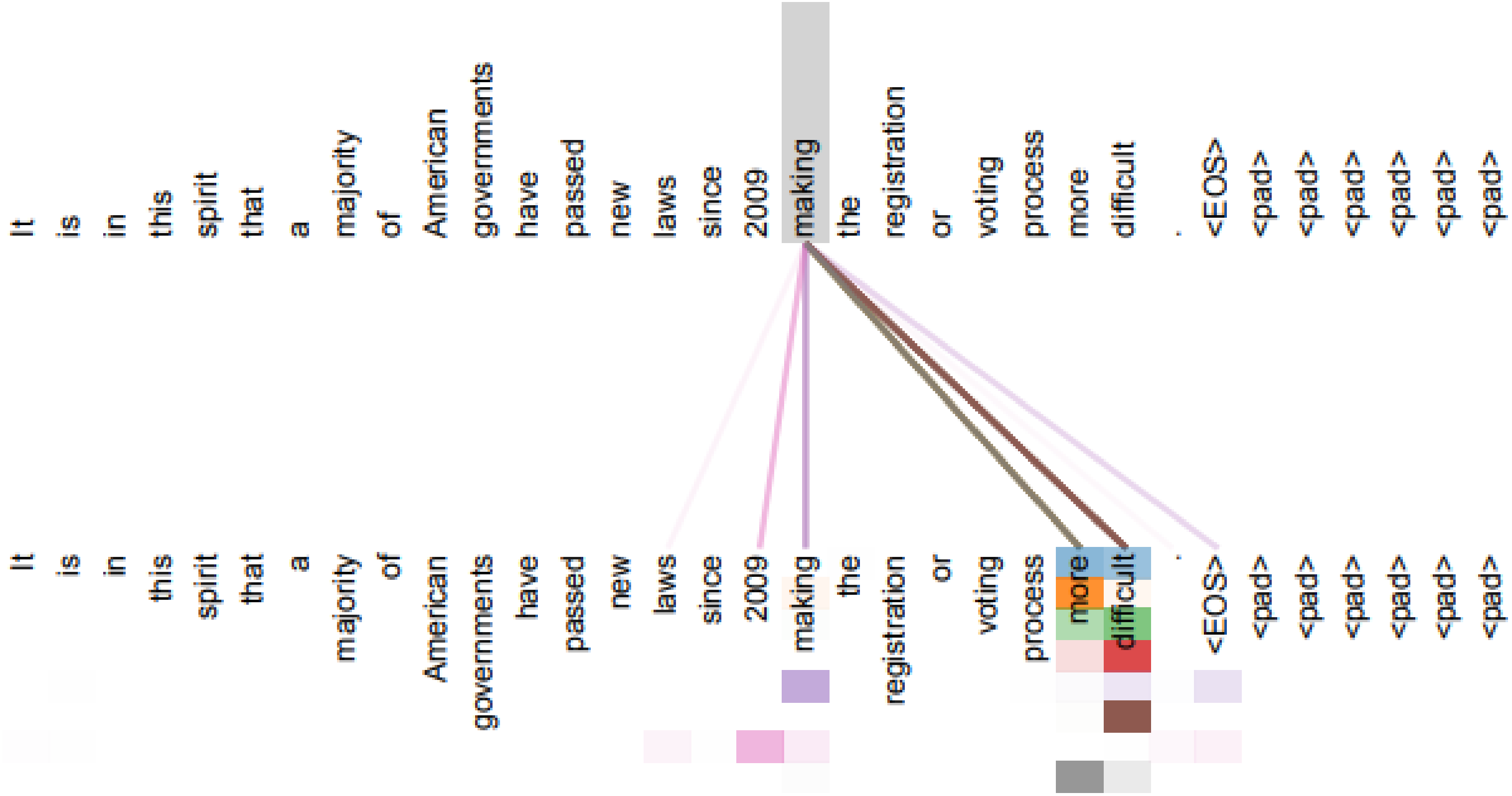
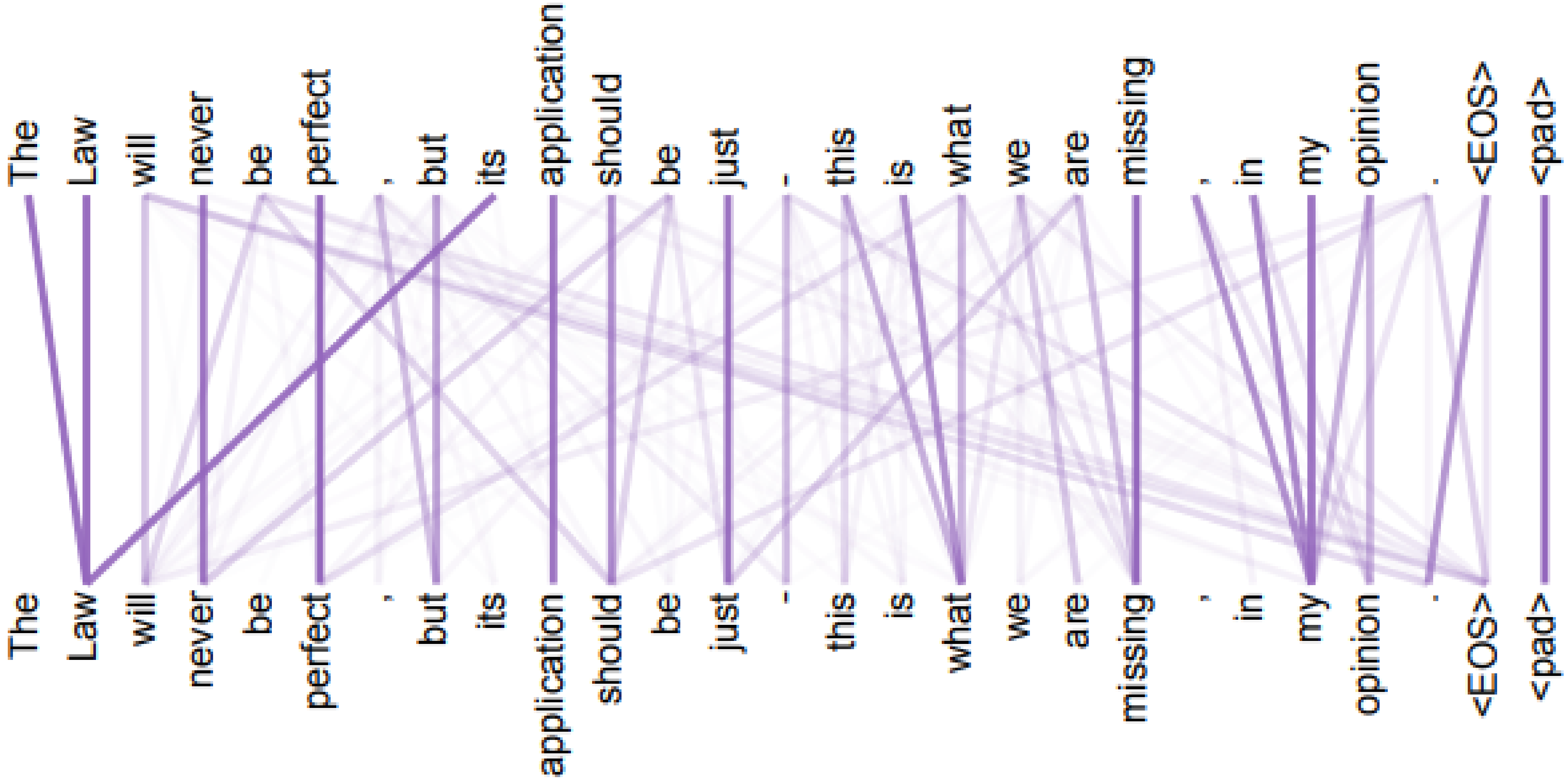


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb ‘making’, completing the phrase ‘making...more difficult’. Attentions here shown only for the word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

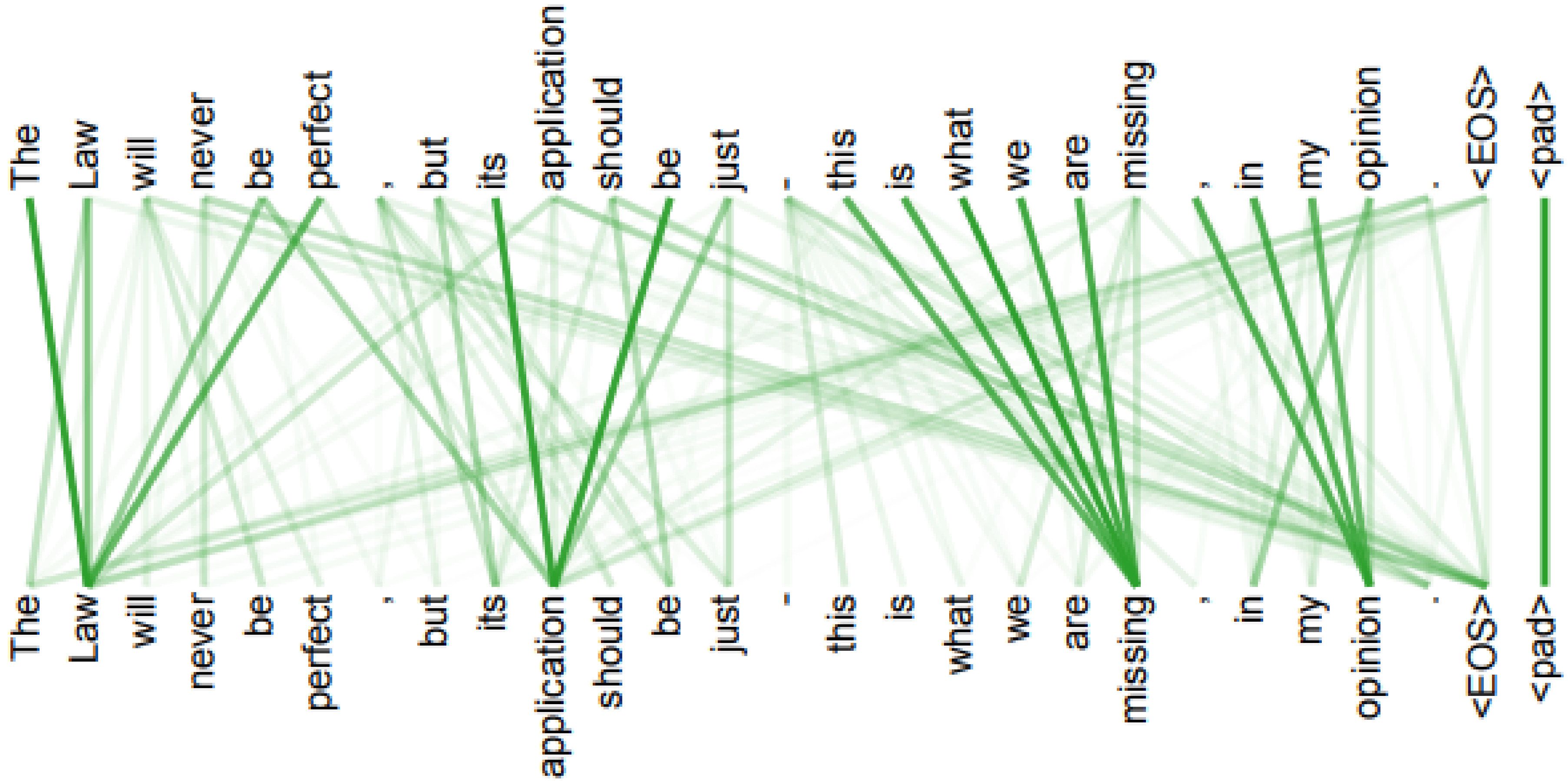
Visualization

指代：Law <-> its， this <-> what , my <-> opinion



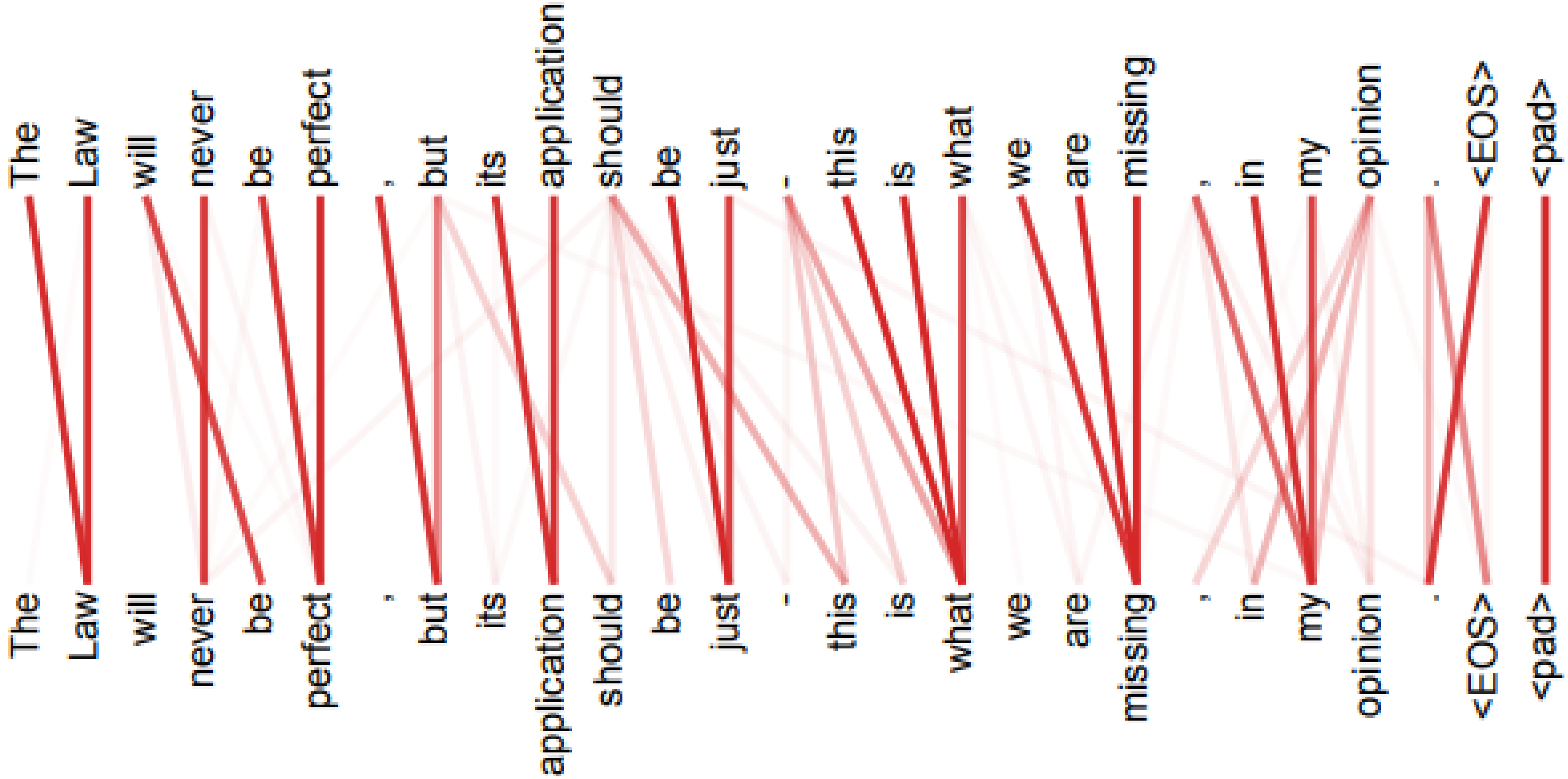
Visualization

头词：在短语结构句法分析中，可以给每个短语规定头词，即短语的核心词
Law, application, missing, opinion



Visualization

实词相关的虚词：Law -> The, be -> will, perfect -> be



Transformer的应用扩展

- **大一统模型MultiModel**
  - 处理任何任务：图像分类、机器翻译、语言模型、句法分析
- **QANet**
  - 处理阅读理解
- **TransformerXL**
  - 处理语言模型
- **ReCoSA**
  - 处理对话
- **GPT/BERT**
  - 大规模预训练

THANKS!