# 自然语言处理基础

## 注意力机制

任宣丞

renxc@pku.edu.cn

北京大学

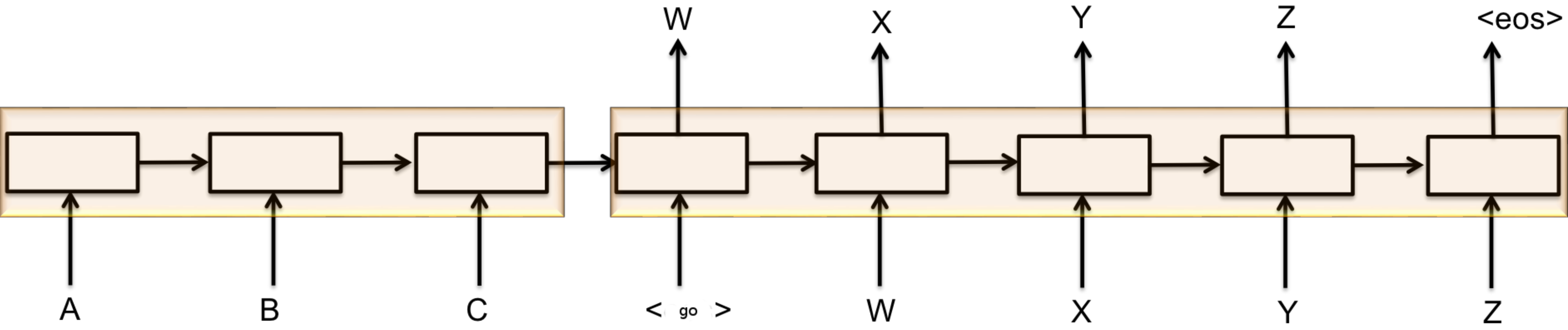
指导教师：孙栩

# 背景

- 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
  - 语言实现改进：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

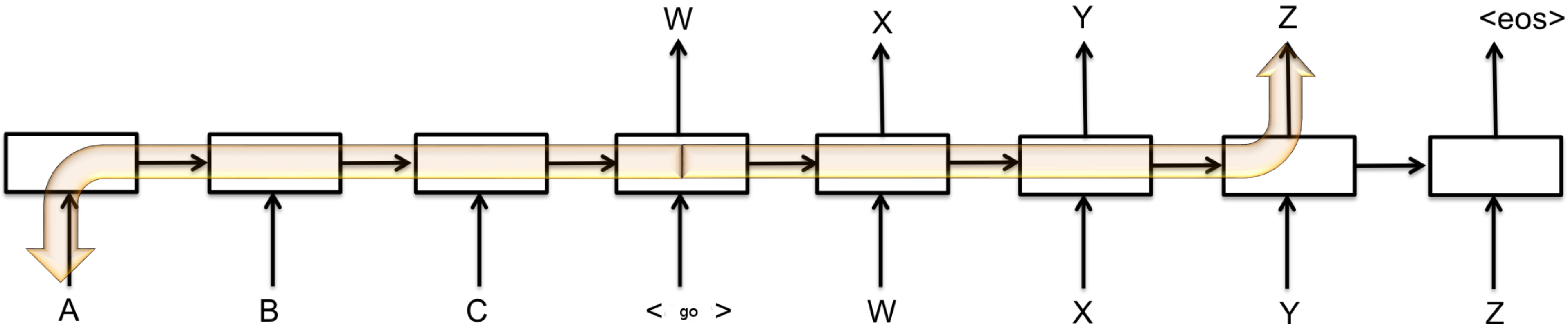
- Encoder-Decoder框架，尤其是Sequence-to-sequence模型的问题



- 映射以序列整体为单位，严重的**数据稀疏**问题
    - 依靠神经网络的表示学习能力，将相似的输入或输出映射到空间中邻近的位置，以缓解数据稀疏问题

# 背景

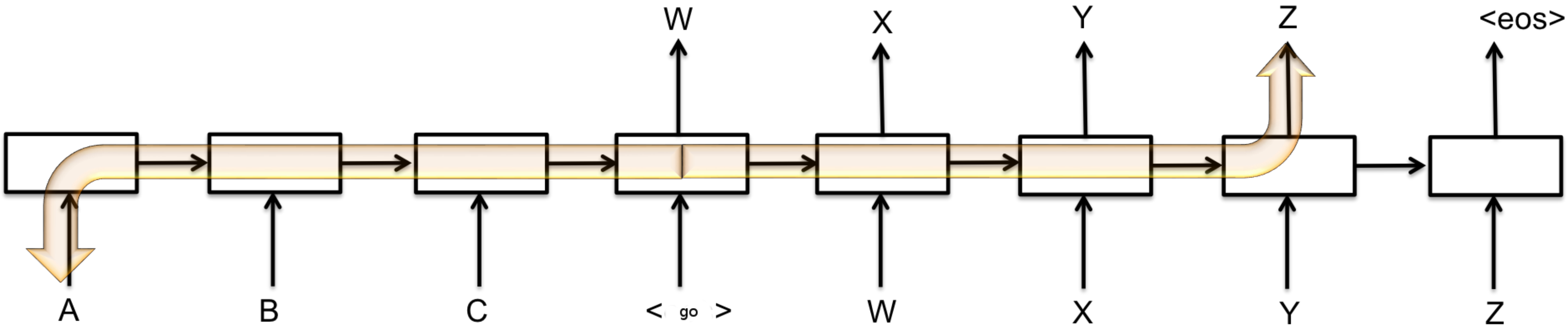
- Encoder-Decoder框架，尤其是Sequence-to-sequence模型的问题



- 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
  - 技巧：将输入序列颠倒，放弃建模过长的依赖

# 背景

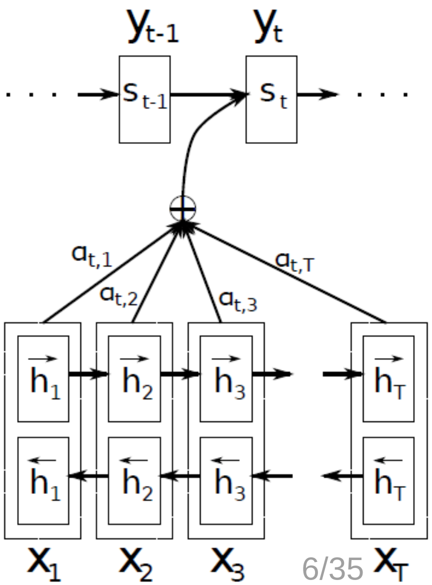
- Encoder-Decoder框架，尤其是Sequence-to-sequence模型的问题



- 假设输入与输出顺序对应
  - 输入不变：最短依赖长度为输入句长
  - 输入颠倒：最短依赖长度为1

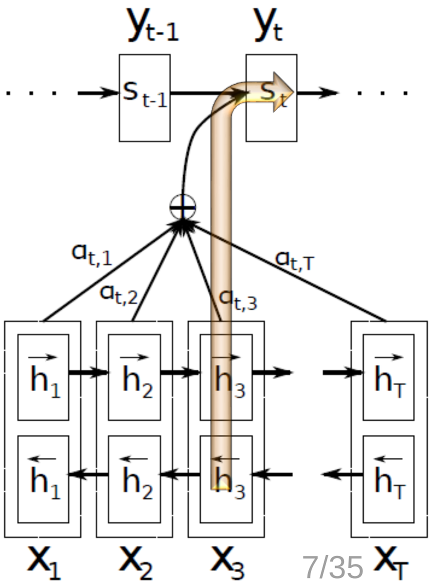
# Attention

- 最早由Bahdanau et al.于2014年提出，发表于ICLR 2015
- 整体思路非常直观
  - 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



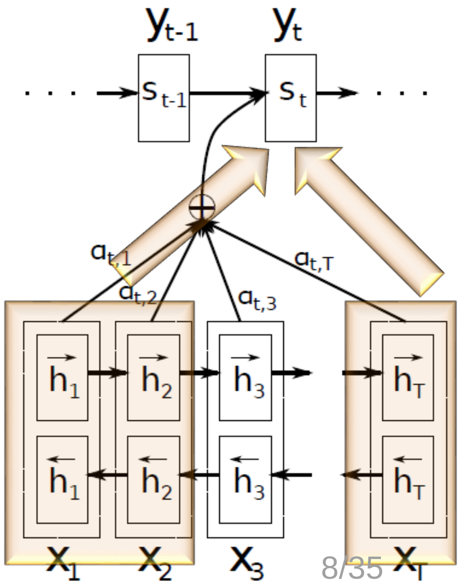
# Attention

- 最早由Bahdanau et al.于2014年提出，发表于ICLR 2015
- 整体思路非常直观
  - 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
  - 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
- 整体思路非常直观
    - 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

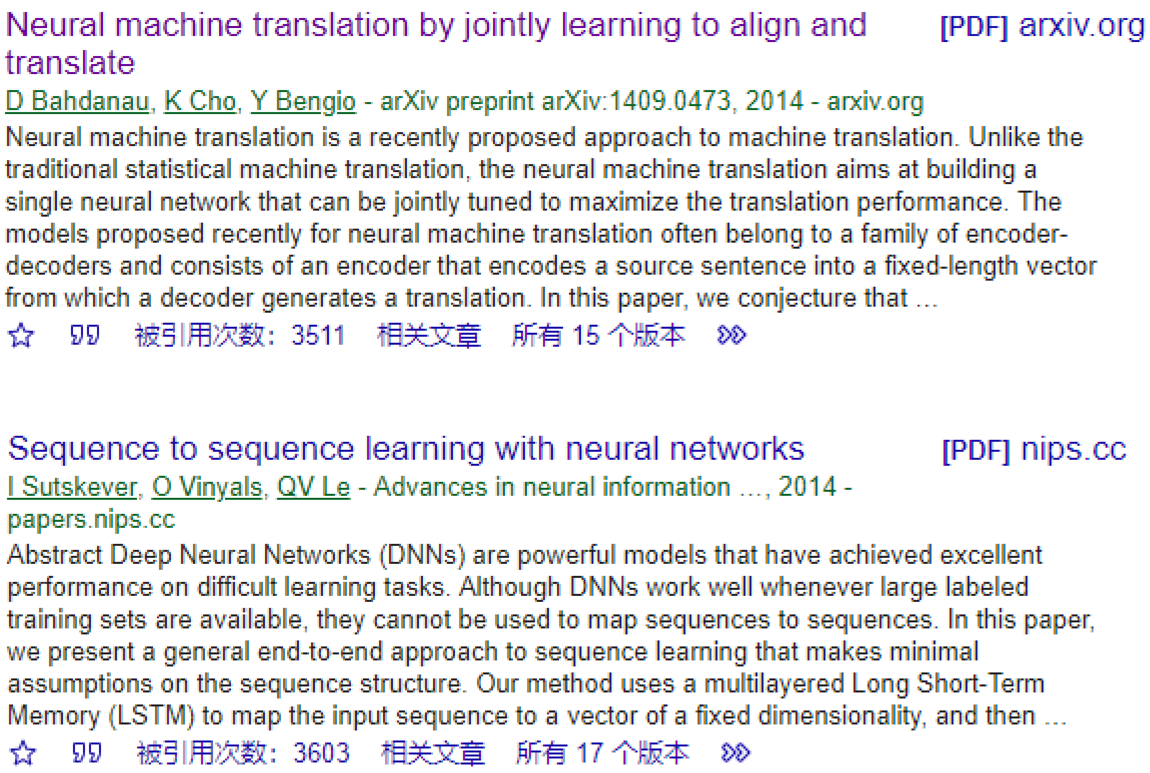


# Attention

- 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
  - 哈工大和科大讯飞ACL 2017的attention over attention

# Bahdanau Attention

- 最早的attention



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

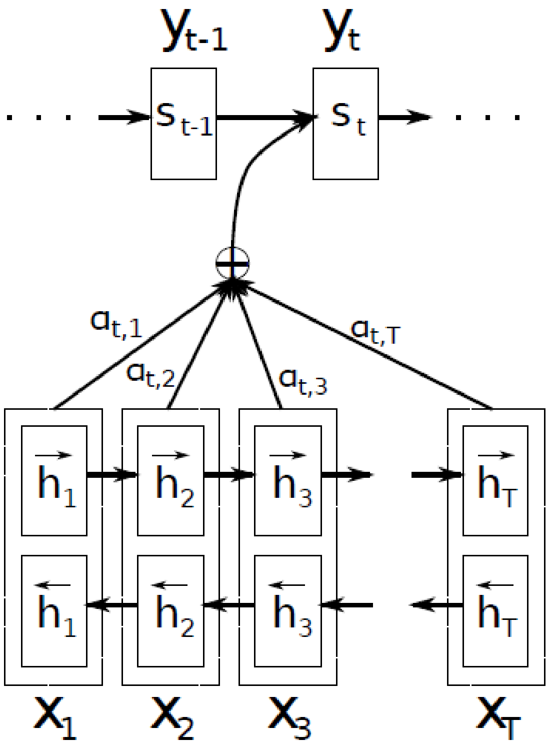
I Sutskever, O Vinyals, QV Le - Advances in neural information ..., 2014 - papers.nips.cc

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 »

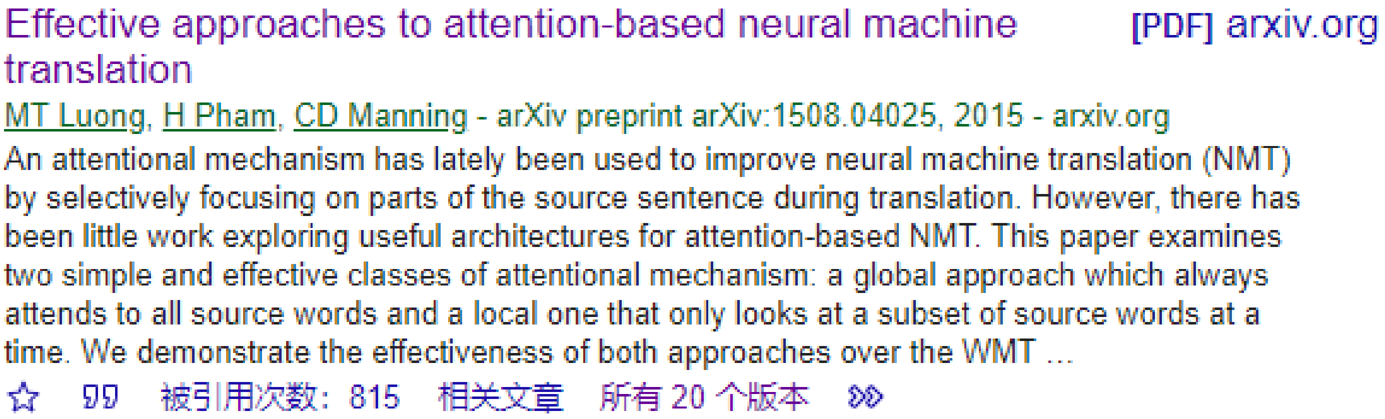
# Bahdanau Attention

- 特点
  - Attention作为LSTM输入
  - 使用前一时刻的LSTM输出查询
- $score(s_{t-1}, h_i) = v^T \tanh(W s_{t-1} + U h_i)$
- 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



# Luong Attention

- 提出了global attention和local 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 »

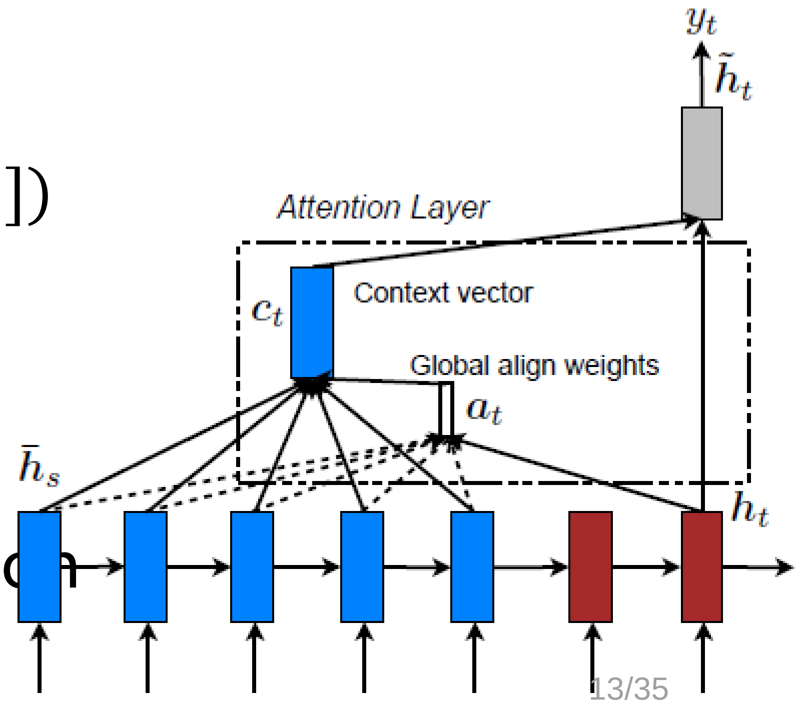
# Luong Global Attention

- 特点
  - Attention作为输出层输入
  - 使用当前时刻的LSTM输出查询

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

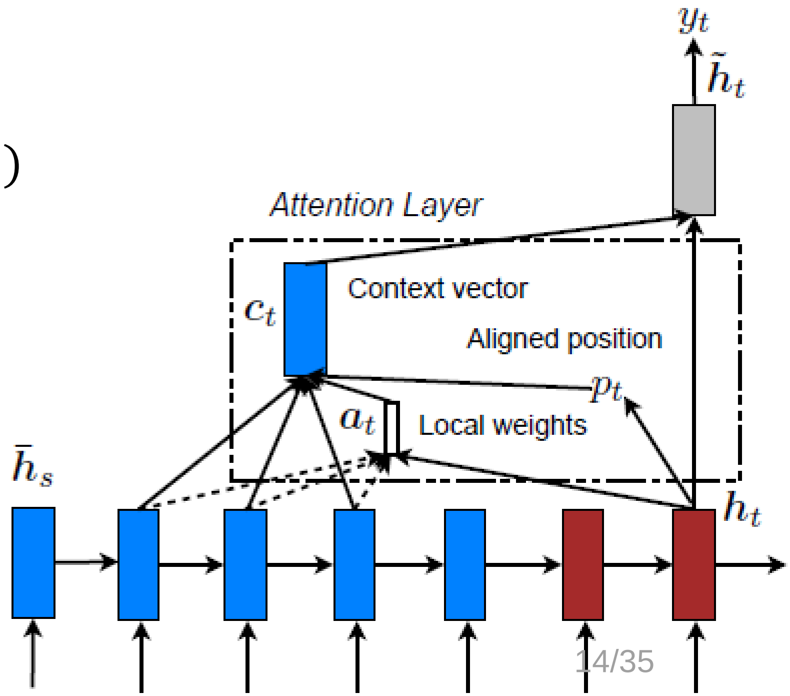
- 分别命名为dot, general, concat
  - Concat与Bahdanau的一致

- 另有$score(h) = Wh$，称为location



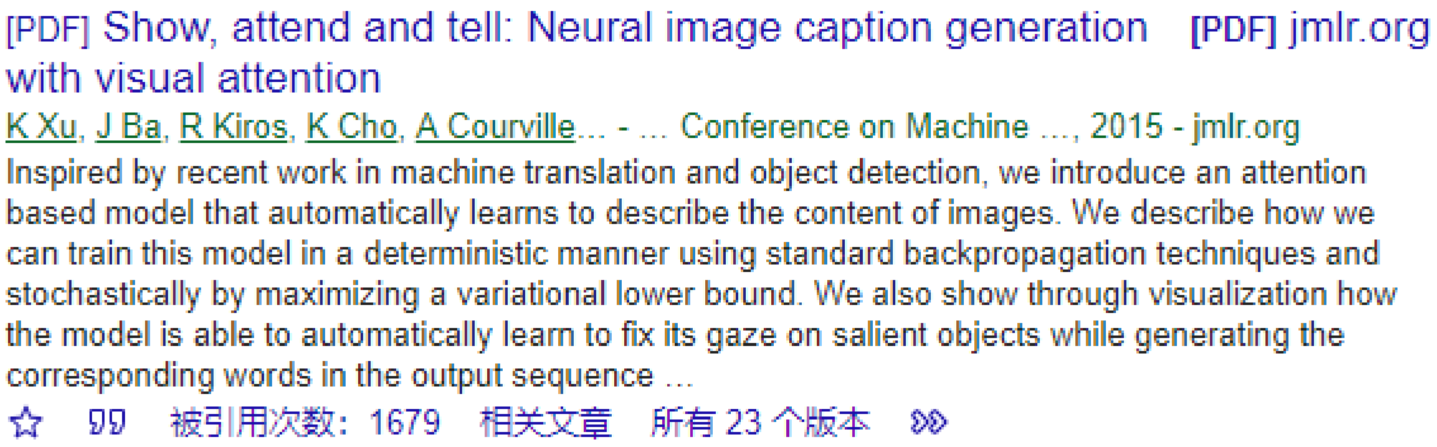
# Luong Local Attention

- 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：$p_t = t$
  - Predictive: $p_t = S\, sigmoid(v \tanh(W s_t))$
    - 其中S为源句子长度
    - Score作微调
      - 乘$\exp(-\frac{(s-p_t)^2}{2\sigma^2})$
      - $\sigma = \frac{D}{2}$



# Visual Attention

- 应用在图像标题生成领域
  - ICML 2015



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

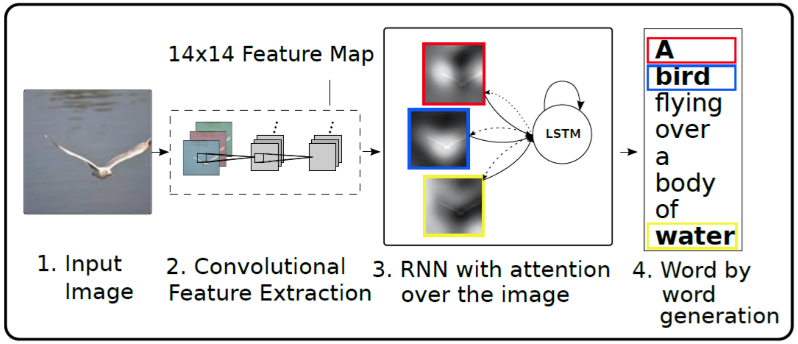
K Xu, J Ba, R Kiros, K Cho, A Courville... - ... Conference on Machine ..., 2015 - jmlr.org

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 ...

☆ ❝ 被引用次数：1679 相关文章 所有 23 个版本 »

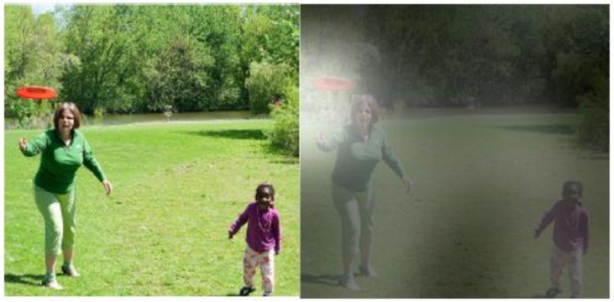
# Visual Attention

- 特点
  - Attention作为LSTM输入
- Hard Attention
  - 根据attention分布，采样一个向量
  - 通过强化学习训练
  - Attention更集中
- Soft Attention
  - 额外约束$\sum_t \alpha_{ti} \approx 1$，使描述更丰富
  - 加权向量额外添加系数$\beta_t = sigmoid(Wh_{t-1})$，使attention更关注图片中物体

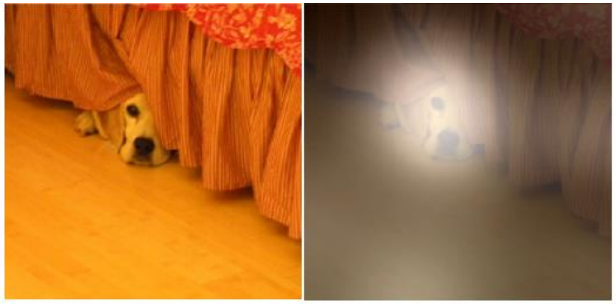

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

# Visual Attention

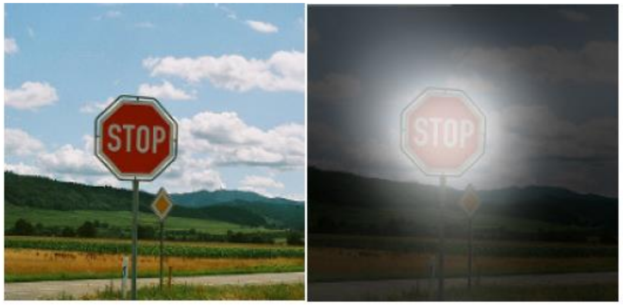
- Grounded Language Generation
  - attention学习到了物体和语言的联系



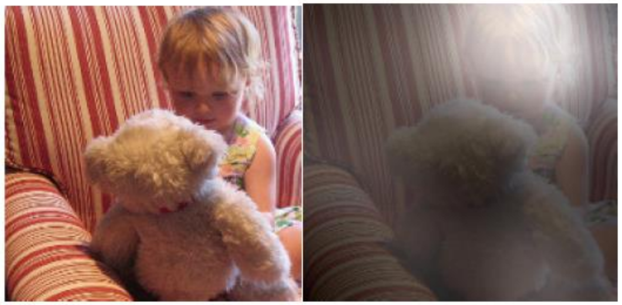
A woman is throwing a frisbee in a park.



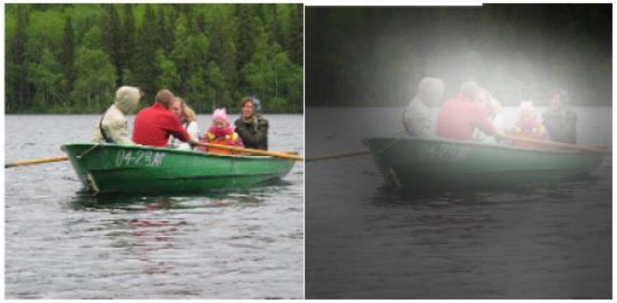
A dog is standing on a hardwood floor.



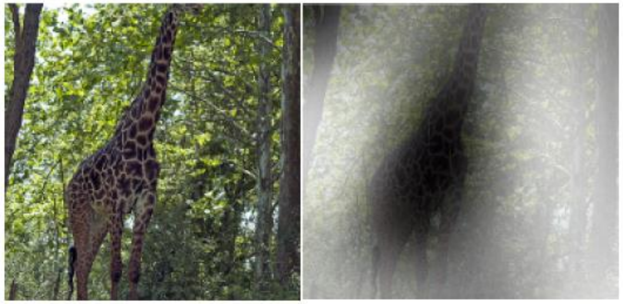
A stop sign is on a road with a mountain in the background.



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



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

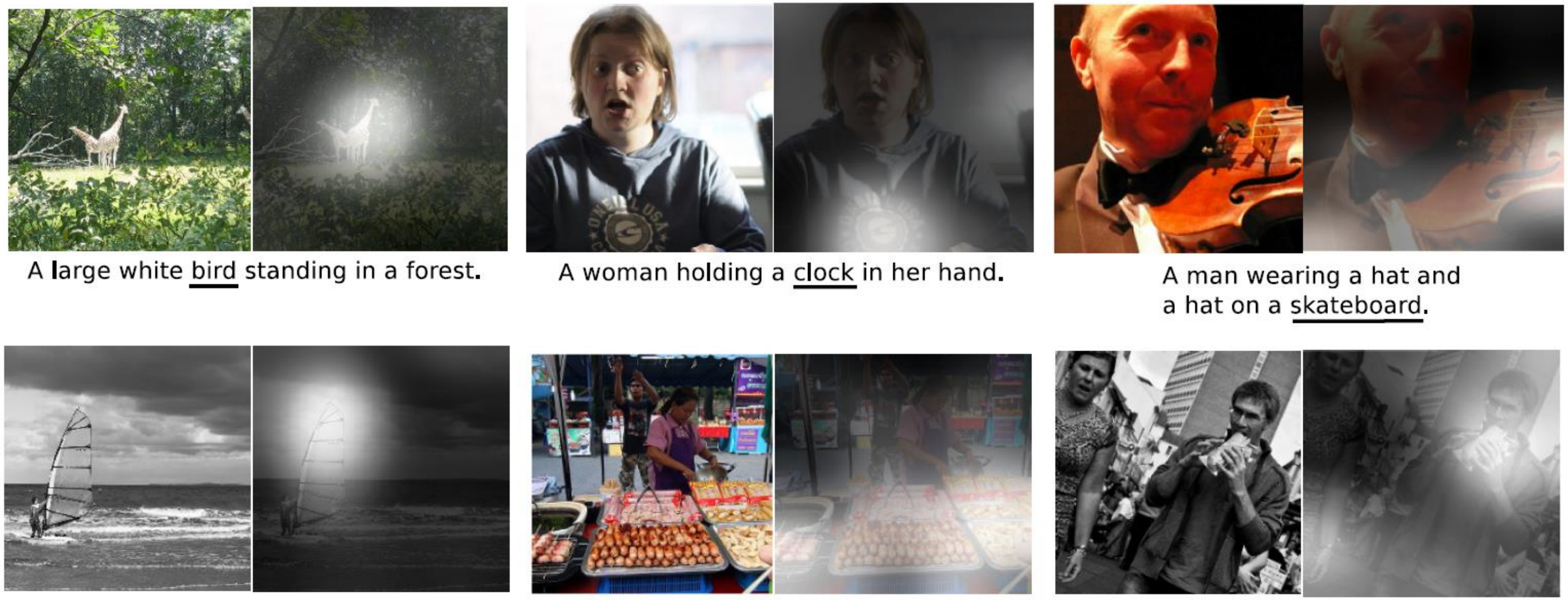


A giraffe standing in a forest with trees in the background.

# 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.

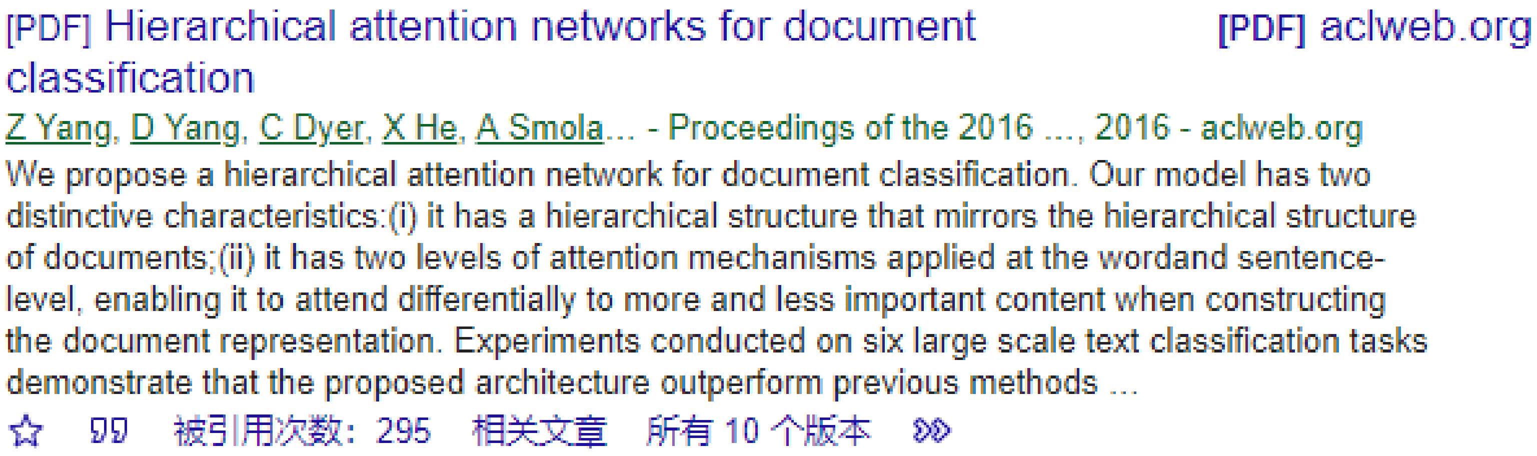
A person is standing on a beach
with a surfboard.

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

- 应用于文档分类



[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 …

☆ ❞ 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
  - 相当于general attention
  - 但没有使用$s_t$而是用了额外的全局向量：固定查询

$$u_{it} = \tanh(W_w h_{it} + b_w)$$

$$\alpha_{it} = \frac{\exp(u_{it}^\top u_w)}{\sum_t \exp(u_{it}^\top u_w)}$$

$$s_i = \sum_t \alpha_{it} h_{it}.$$

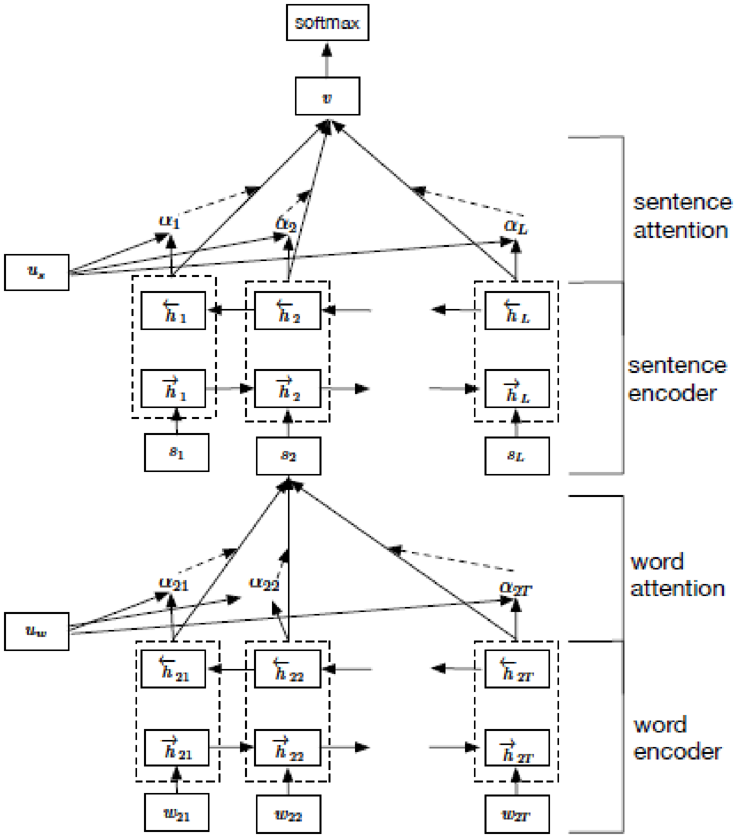
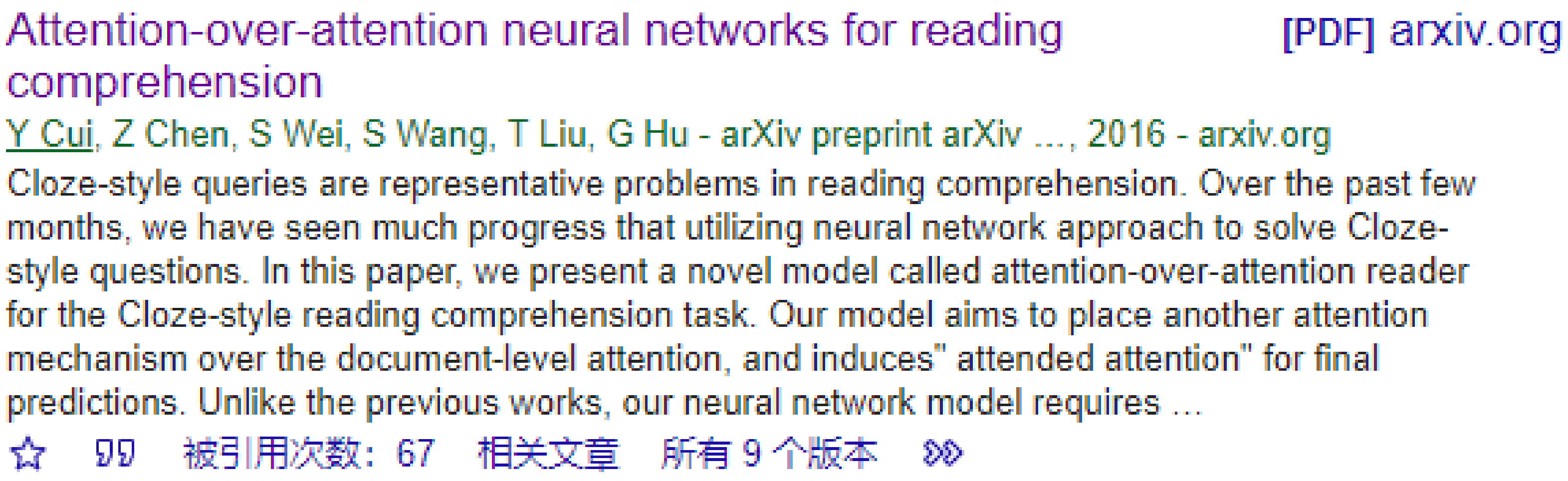


Figure 2: Hierarchical Attention Network.

# 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

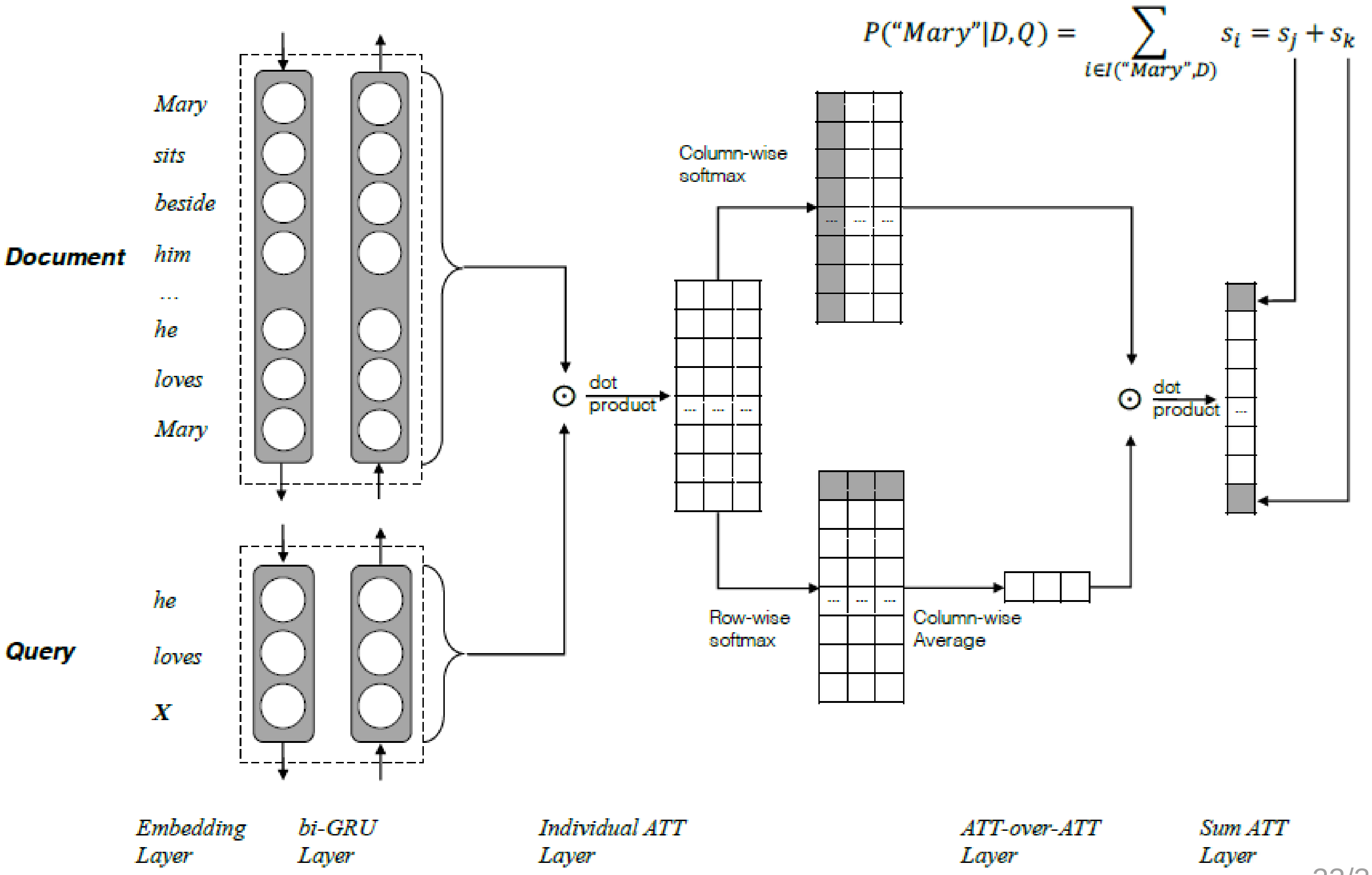
Cloze-style queries are representative problems in reading comprehension. Over the past few months, we have seen much progress that utilizing neural network approach to solve Cloze-style questions. In this paper, we present a novel model called attention-over-attention reader for the Cloze-style reading 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 ...

☆ 99 被引用次数：67 相关文章 所有 9 个版本 »

# Attention over Attention

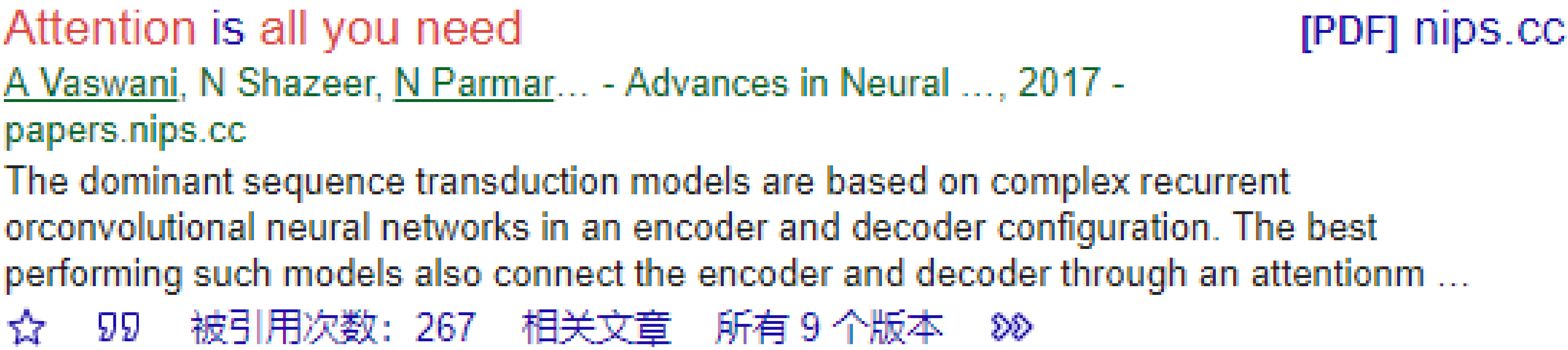
- 思路
  - 每个文档和问题的词算内积，得到了一个score矩阵
  - 每行针对doc词，每列针对query词
  - 每列作softmax：和问题最相关的文档词,
  - 每行作softmax：和文档最相关的问题词
  - 两者每行作内积，得到的是attention over attention

# Attention over Attention



# Self-Attention

- Self-Attention这一概念被提出多次
- 目前最知名的当属Attention is all you need中的实现



Attention is all you need [PDF] nips.cc

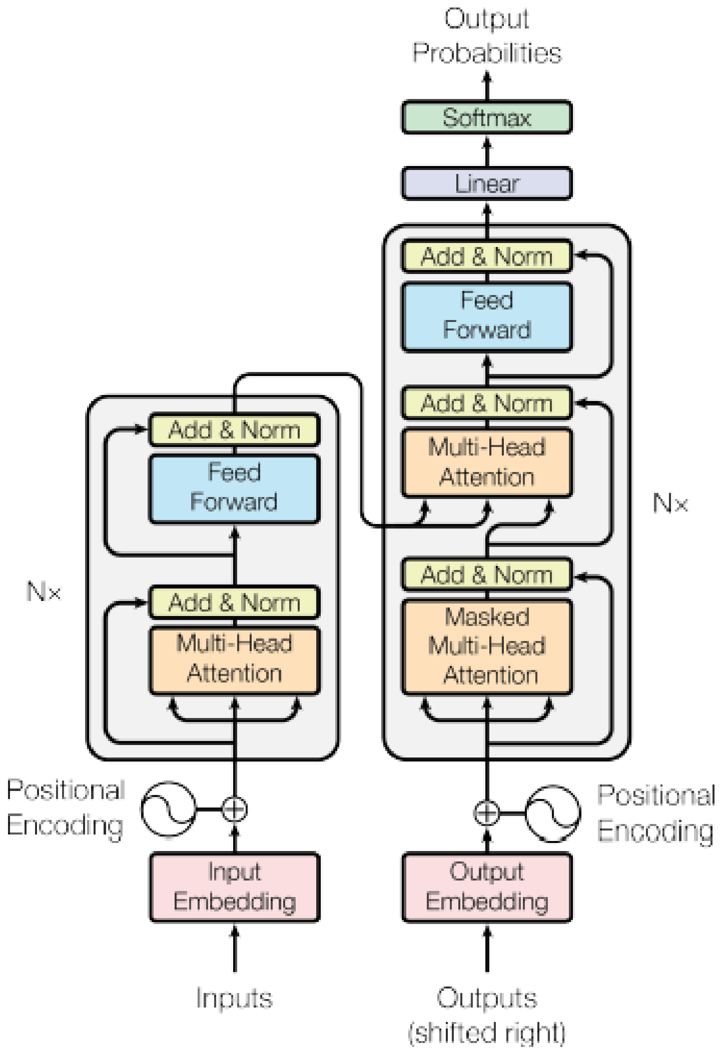
A Vaswani, N Shazeer, N Parmar… - Advances in Neural …, 2017 - papers.nips.cc

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 »

# Self-Attention

- 目标：**完全替代RNN**
  - 背景：convS2S使用CNN替代RNN
- 输入输出端采用层叠的Block
- 每个Block负责学习输入序列的**不同层级表示**
- 注意力的三种形态
  - 输入端的自注意力
  - 输出端的自注意力
  - 输出端对输入端的注意力

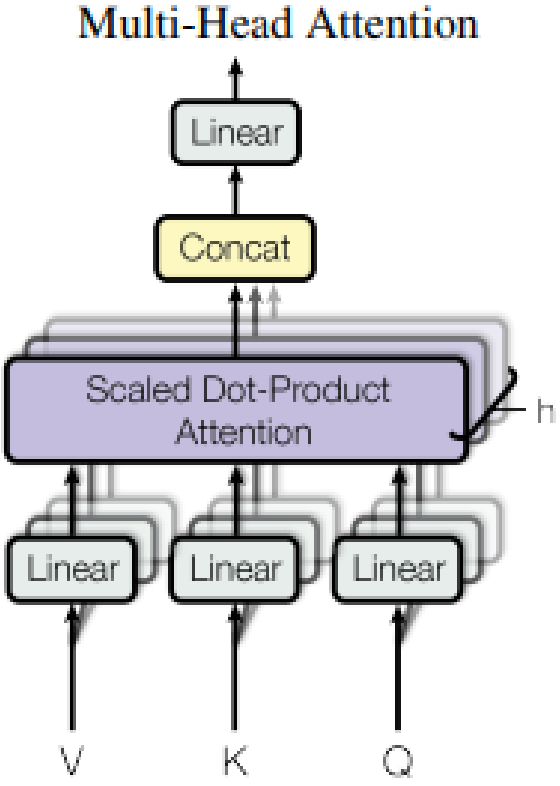


# Multi-head attention

- 将隐层向量分为**多组**

- 在**每组内**进行注意力计算
  - 即Softmax的归一化仅针对每个分组

- 每组向量被解释为不同的**表示子空间**



# Visualization

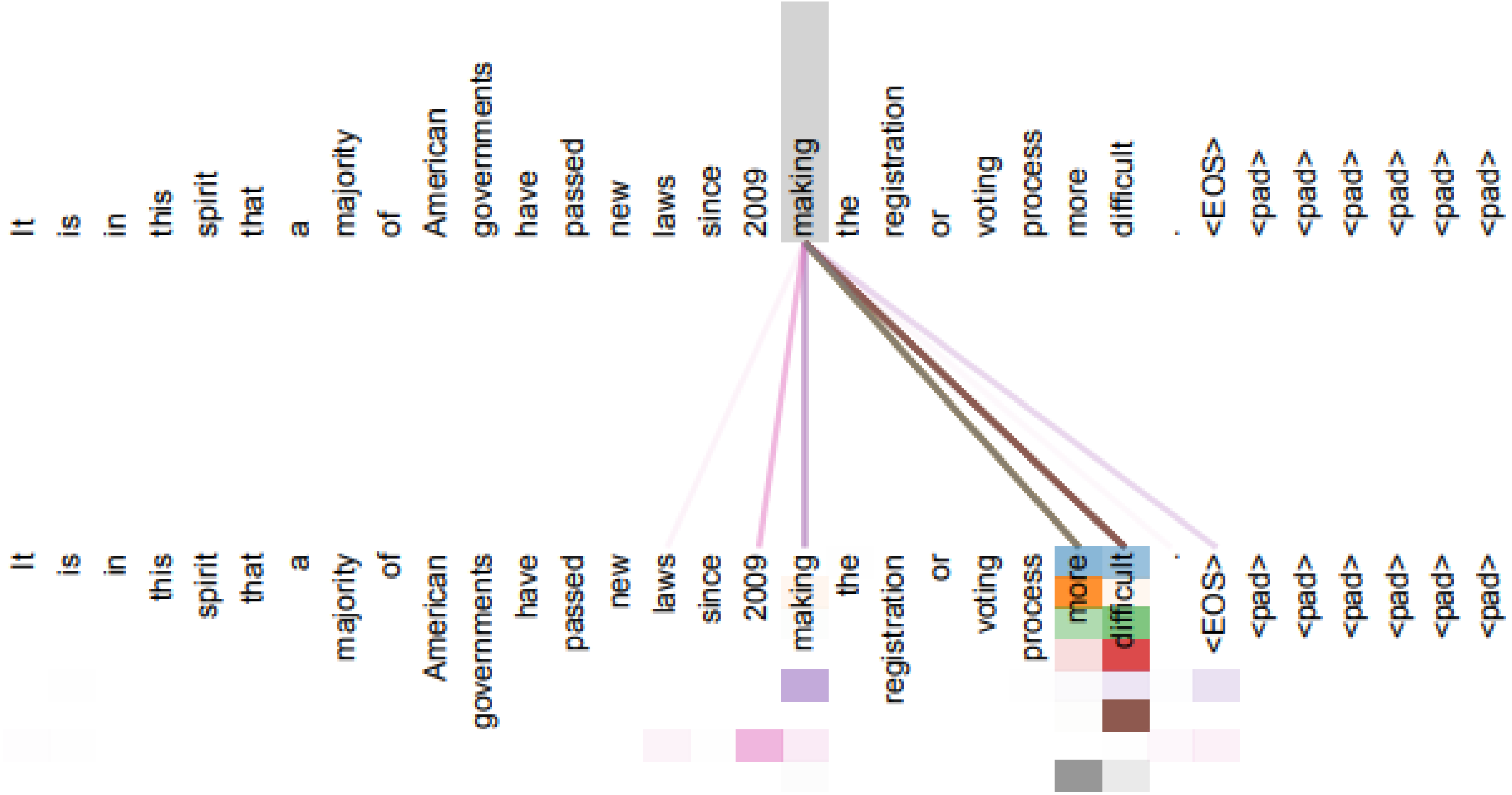
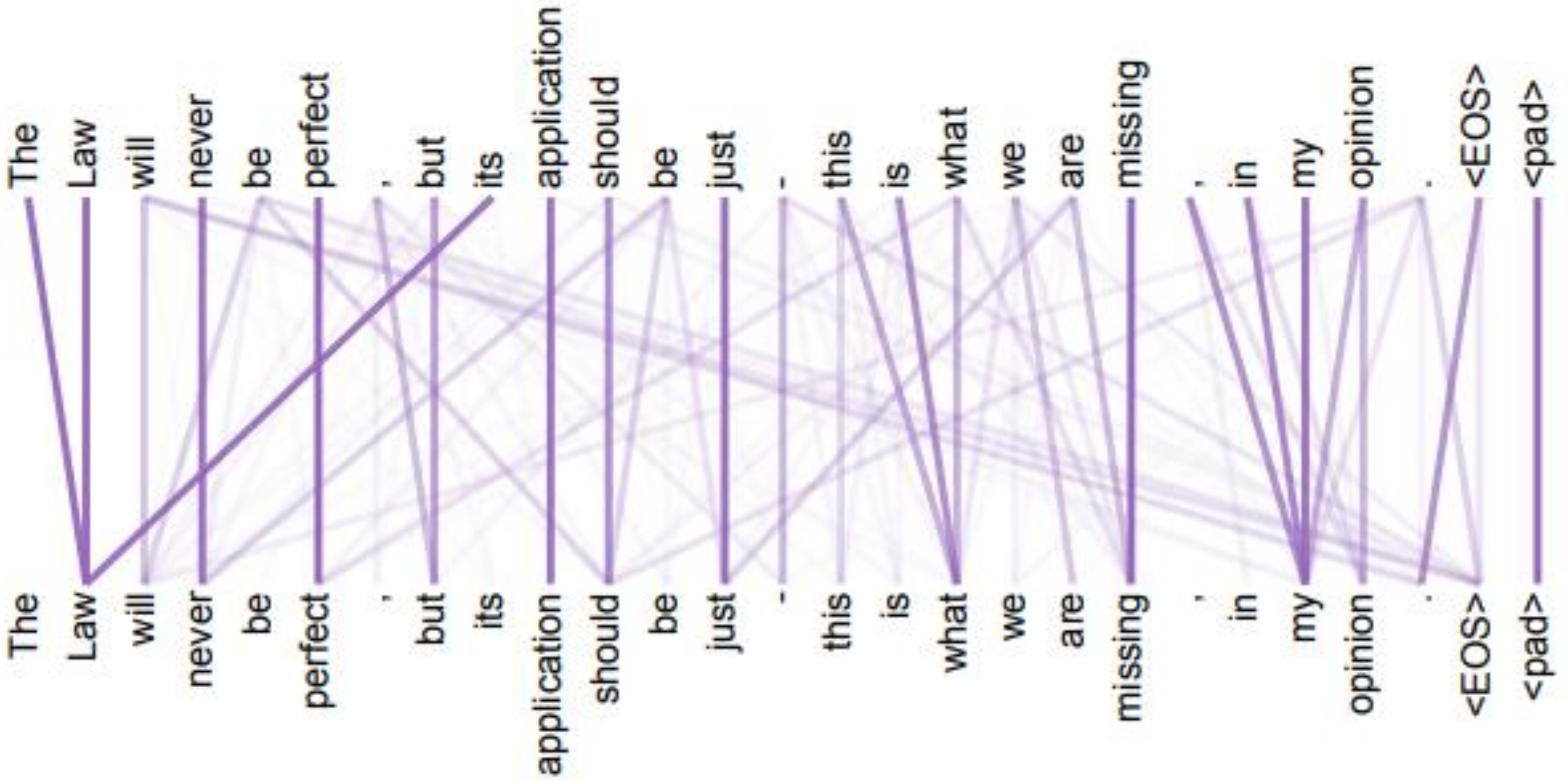
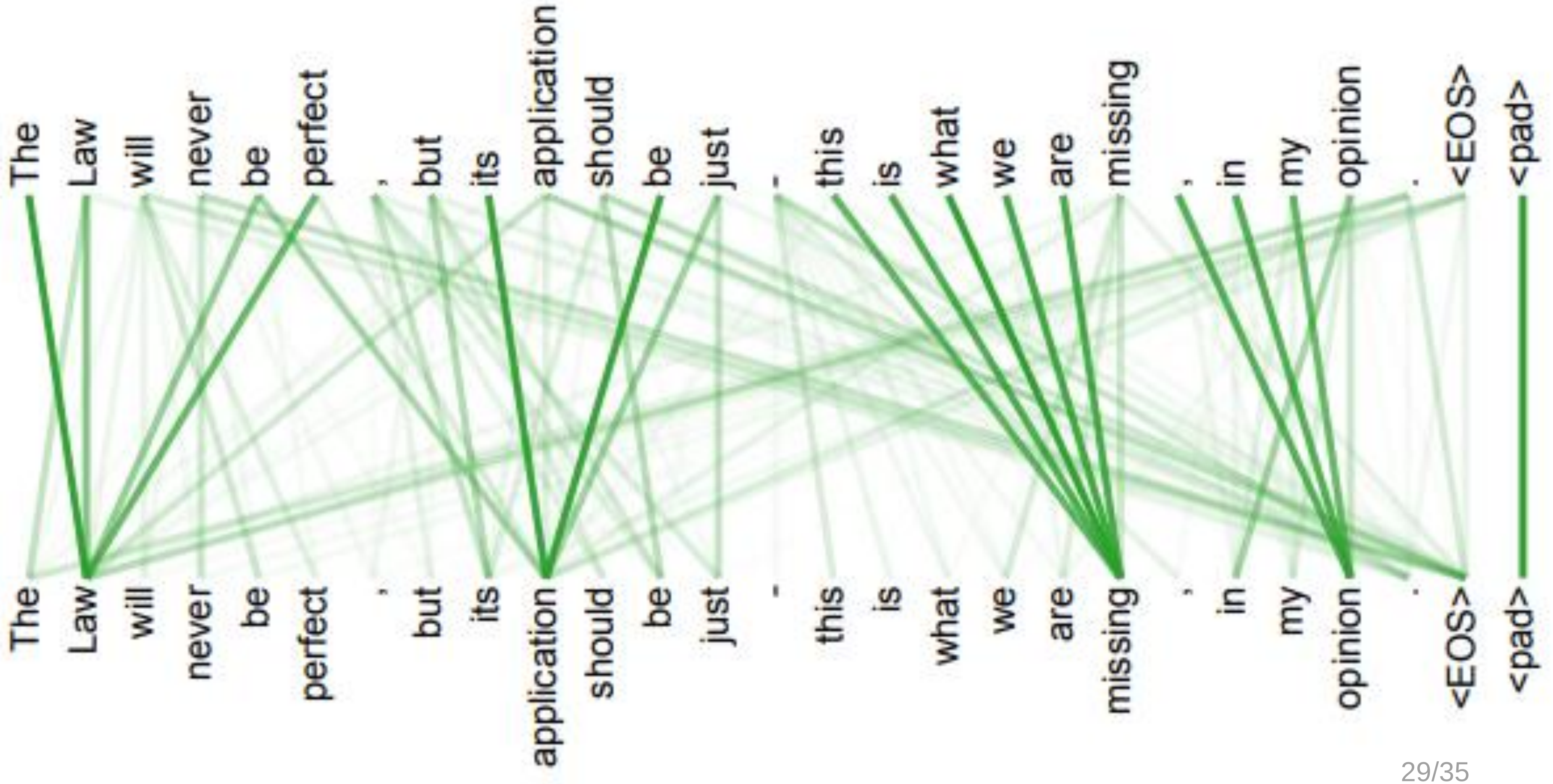


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word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

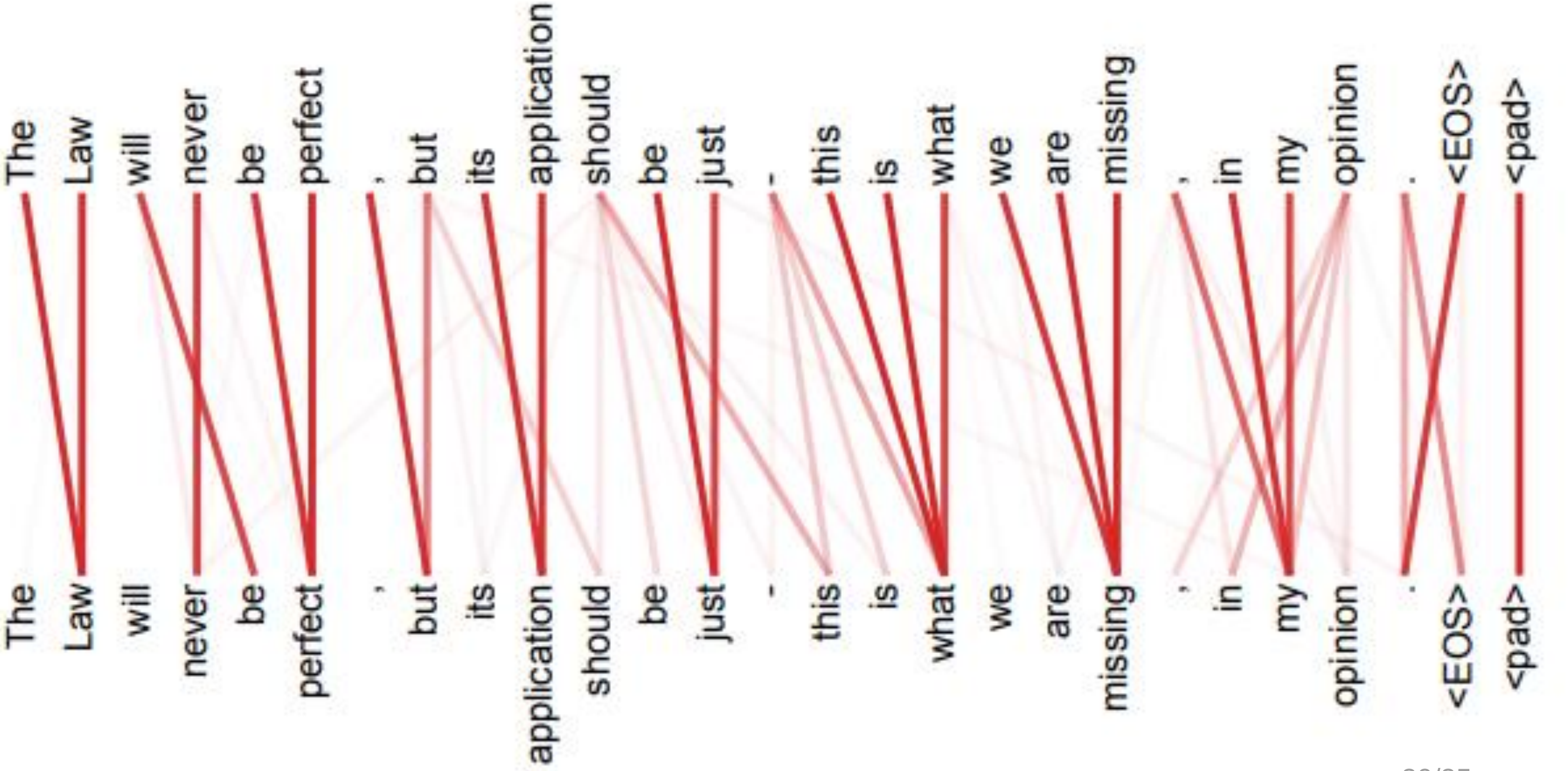
# Visualization



# Visualization

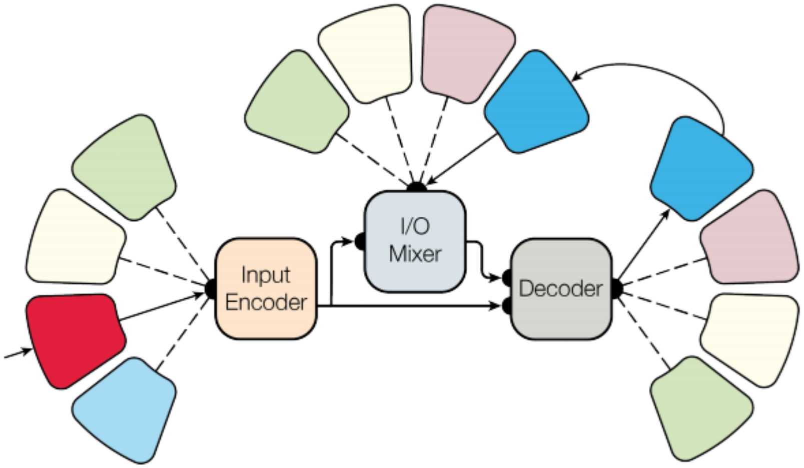


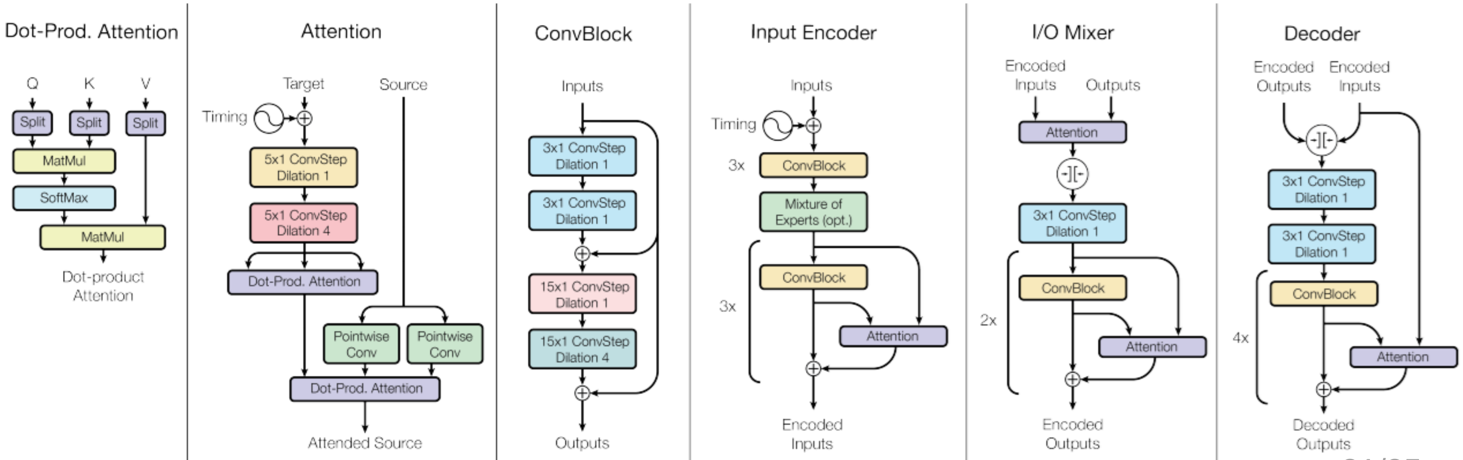
# Visualization



# MutliModel

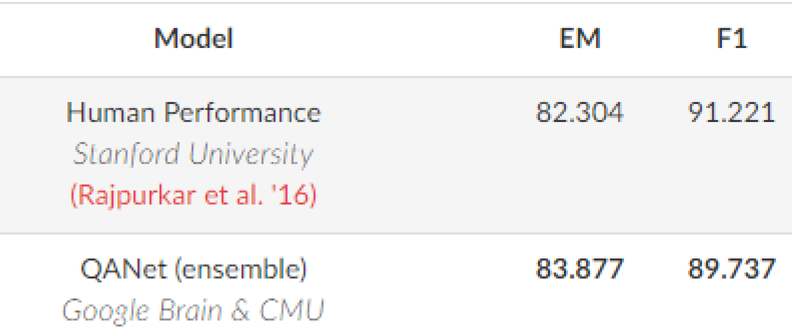
- Google
  - ImageNet
  - WMT
  - WSJ speech
  - Parsing



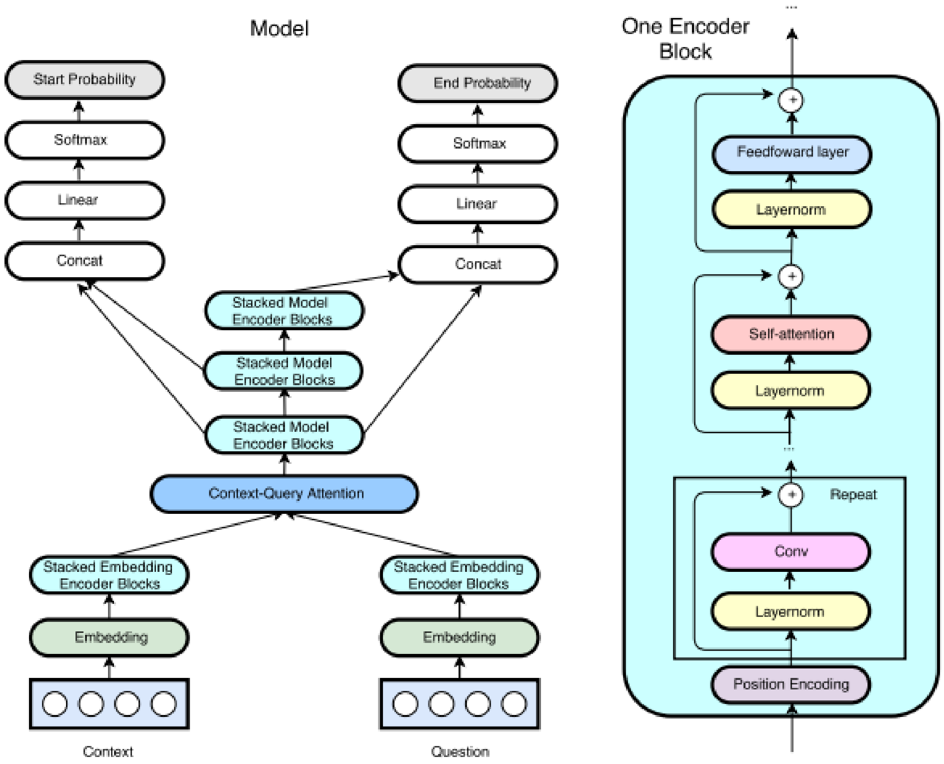


# QANet

- Google & CMU
- SQuAD 第一名
- Match-LSTM
  - LSTM -> attention

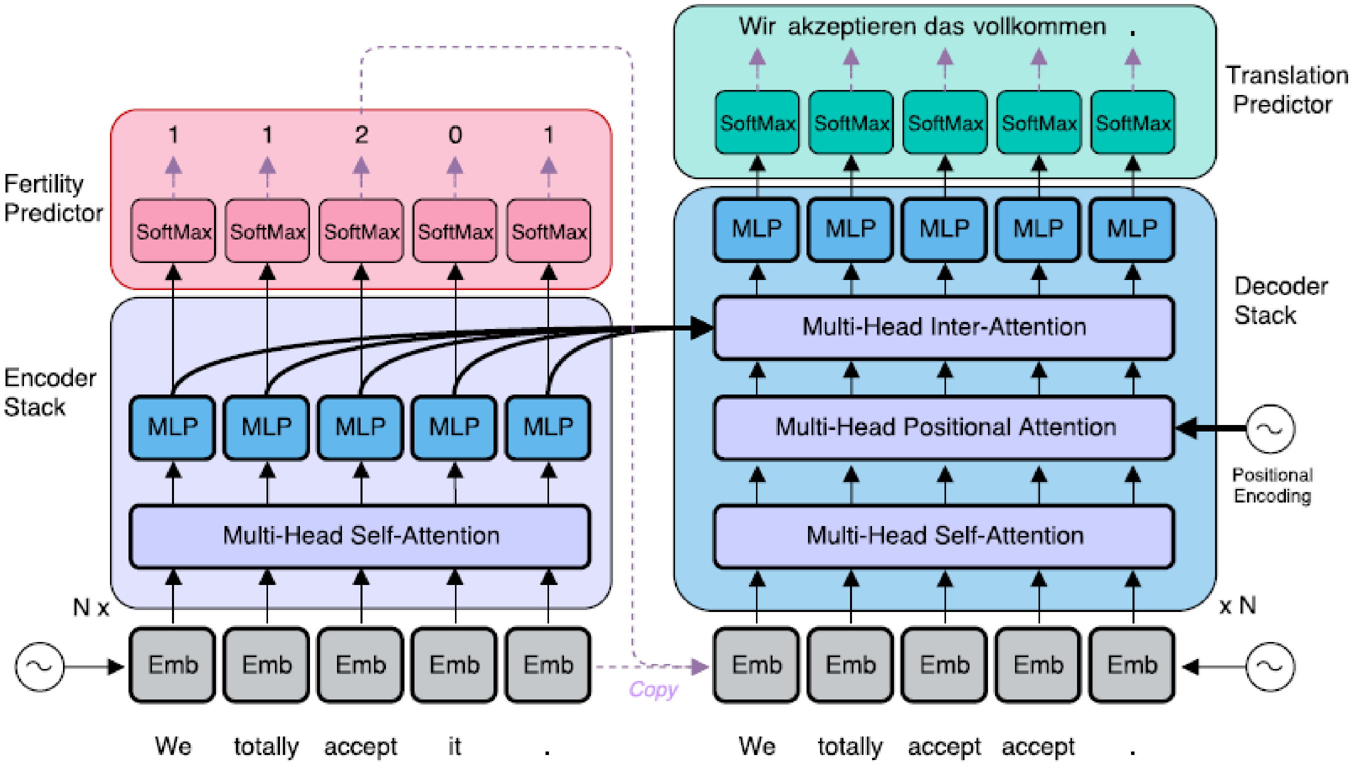


| Model | EM | F1 |
|---|---|---|
| Human Performance<br>*Stanford University*<br>(Rajpurkar et al. '16) | 82.304 | 91.221 |
| QANet (ensemble)<br>*Google Brain & CMU* | 83.877 | 89.737 |



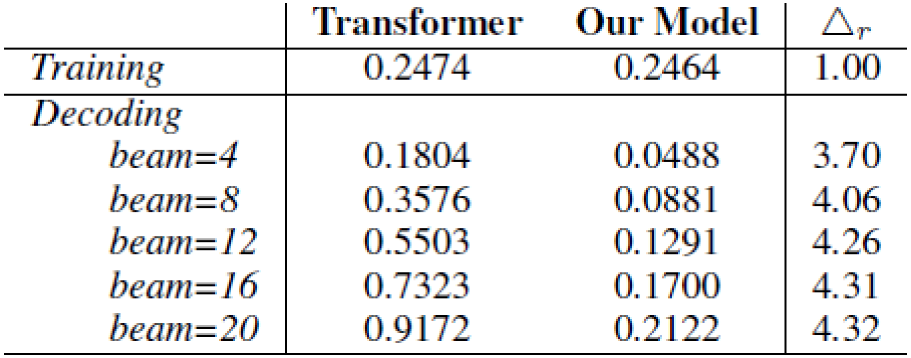
# Non-Autoregressive NMT

- Salesforce & UHK ICLR 2018
  - Poor performance but low latency

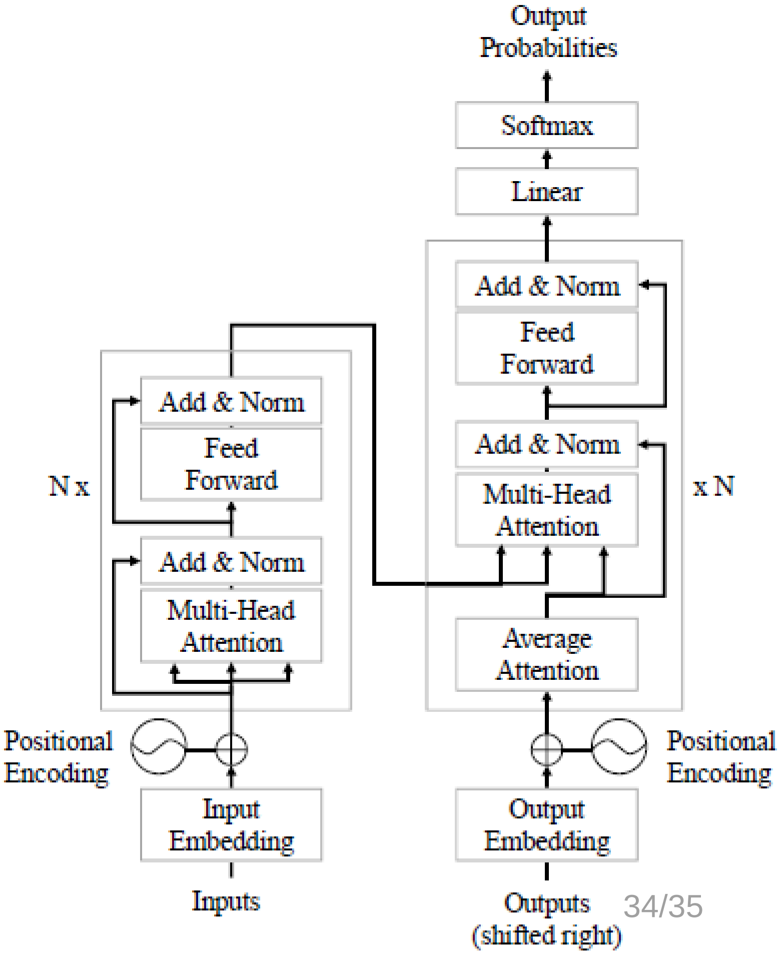

Wir akzeptieren das vollkommen .
Translation Predictor
SoftMax
Fertility Predictor
1 1 2 0 1
Decoder Stack
MLP
Multi-Head Inter-Attention
Multi-Head Positional Attention
Positional Encoding
Multi-Head Self-Attention
Encoder Stack
N x
x N
Emb
Copy
We totally accept it .
We totally accept accept .

# Average Attention

- Biao Zhang, Deyi Xiong, Jinsong Su
- Attention in decoder
  - Using average attention
- ACL 2018
- Slight performance drop
  - En-de 26.31 -> 26.05



| | Transformer | Our Model | $\triangle_r$ |
|---|---|---|---|
| *Training* | 0.2474 | 0.2464 | 1.00 |
| *Decoding* | | | |
| *beam=4* | 0.1804 | 0.0488 | 3.70 |
| *beam=8* | 0.3576 | 0.0881 | 4.06 |
| *beam=12* | 0.5503 | 0.1291 | 4.26 |
| *beam=16* | 0.7323 | 0.1700 | 4.31 |
| *beam=20* | 0.9172 | 0.2122 | 4.32 |



DEECAMP

# THANK YOU