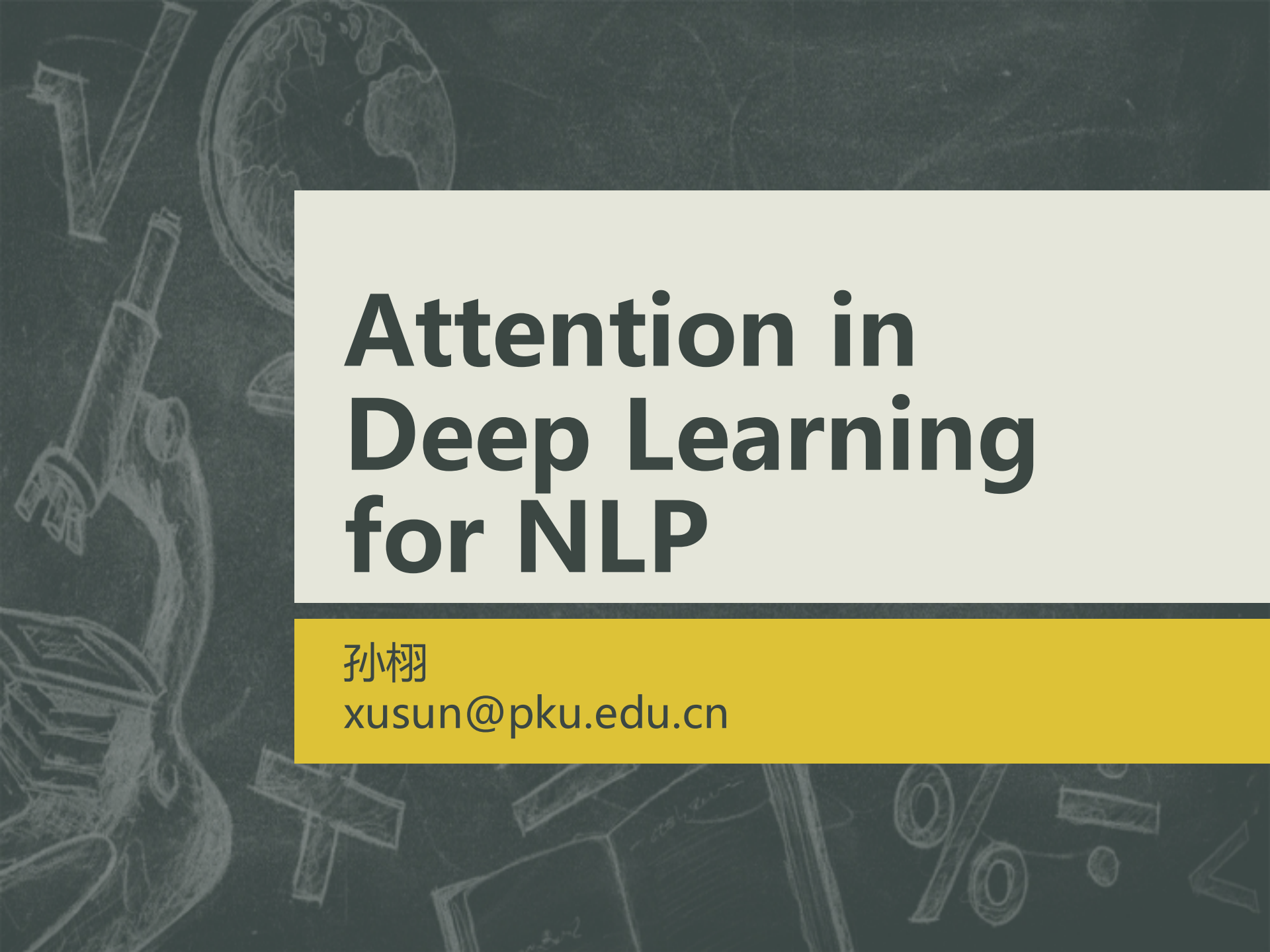


Attention in Deep Learning for NLP

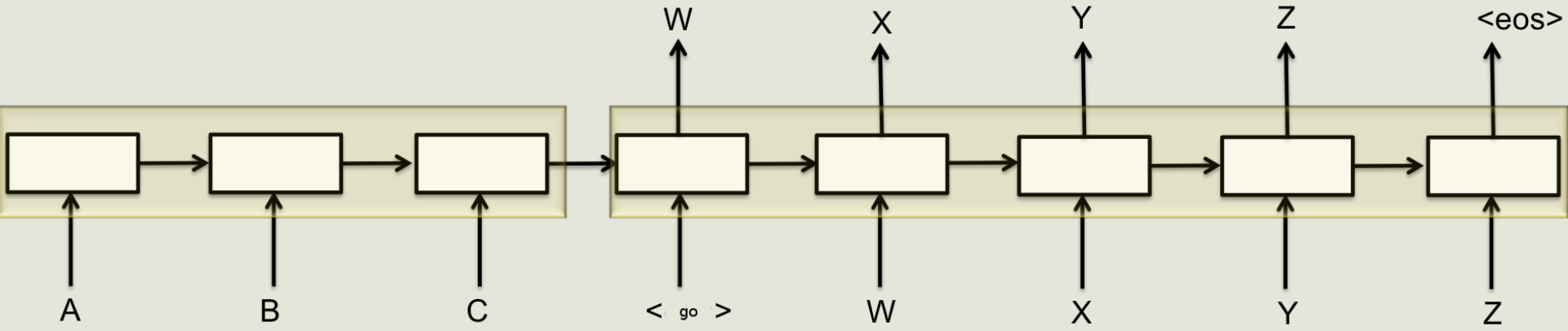
孙栩
xusun@pku.edu.cn

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

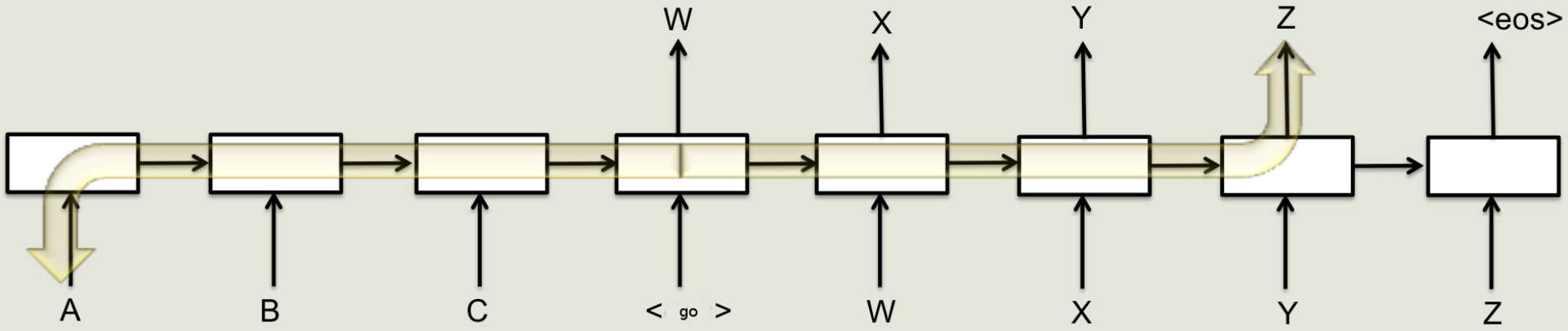
- Encoder-Decoder框架，尤其是Sequence-to-sequence范式的问题



- 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

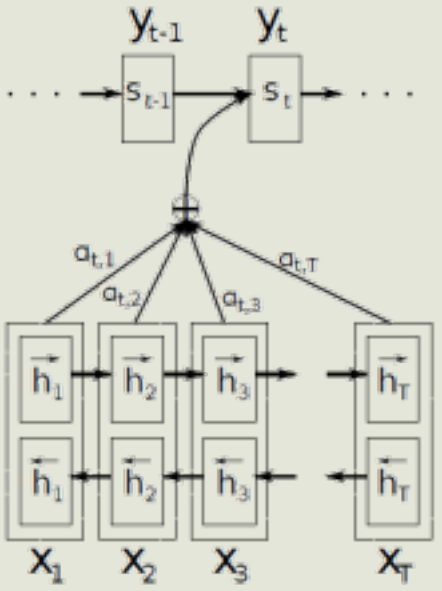
- Encoder-Decoder框架，尤其是Sequence-to-sequence范式的问题



- 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

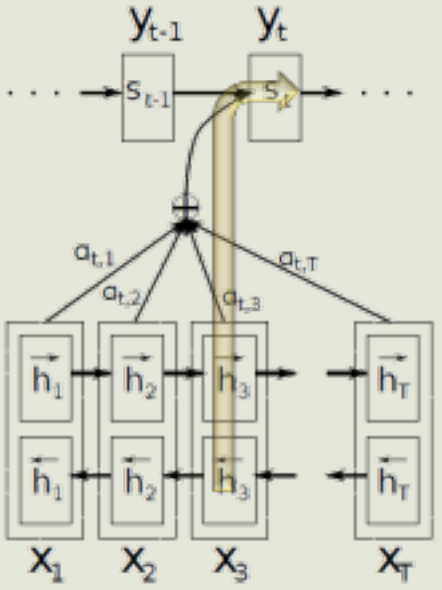
Attention

- 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



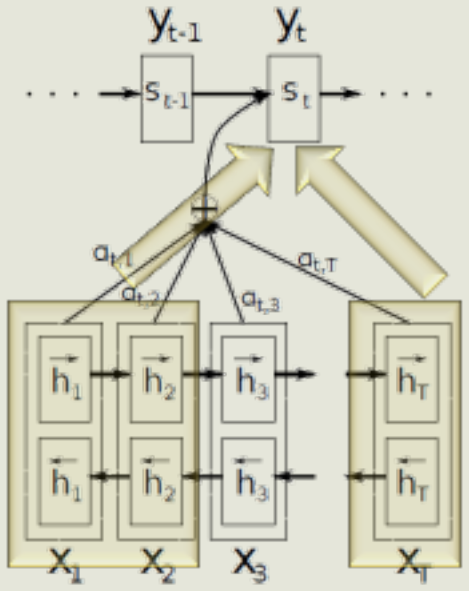
Attention

- 最早由Bahdanau et al.于2014年提出，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
- 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，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
- 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，推广到NLP的诸多领域

- 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

Bahdanau Attention

- 最早的attention，用于机器翻译
  - 引用量非常高，和seq2seq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 ...

☆ 99 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

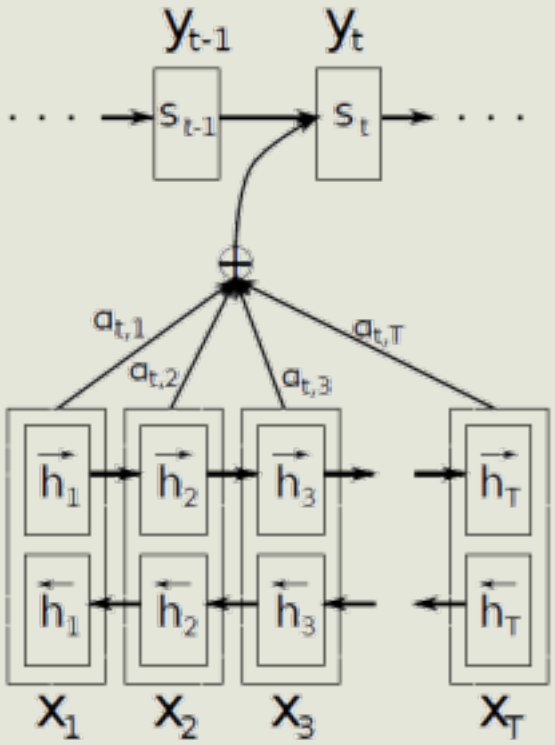
I Sutskever, O Vinyals, QV Le - Advances in neural information ..., 2014 - papers.nips.cc

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 ...

☆ 99 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

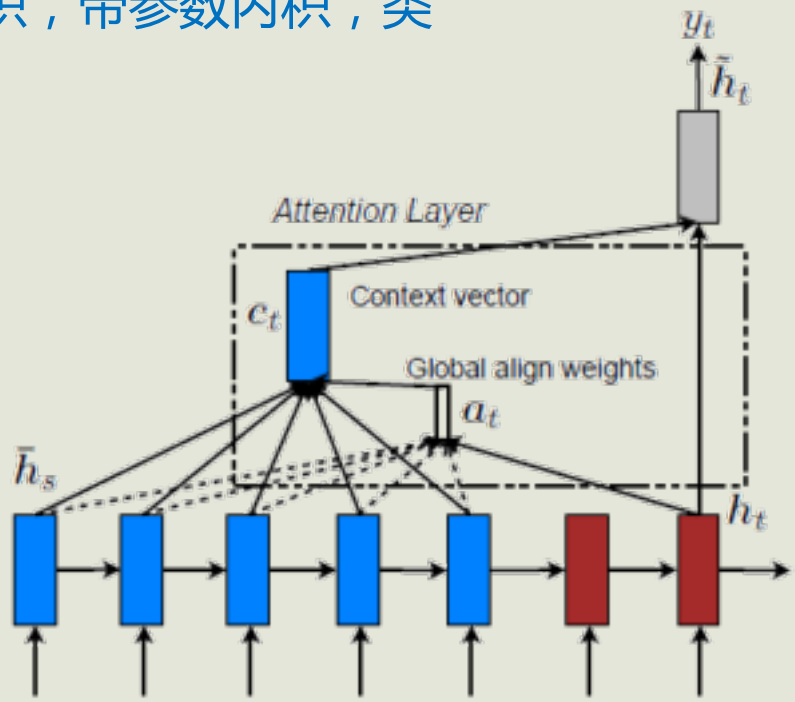
MT Luong, H Pham, CD Manning - arXiv preprint arXiv:1508.04025, 2015 - arxiv.org

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 ...

☆ 99 被引用次数：815 相关文章 所有 20 个版本 »

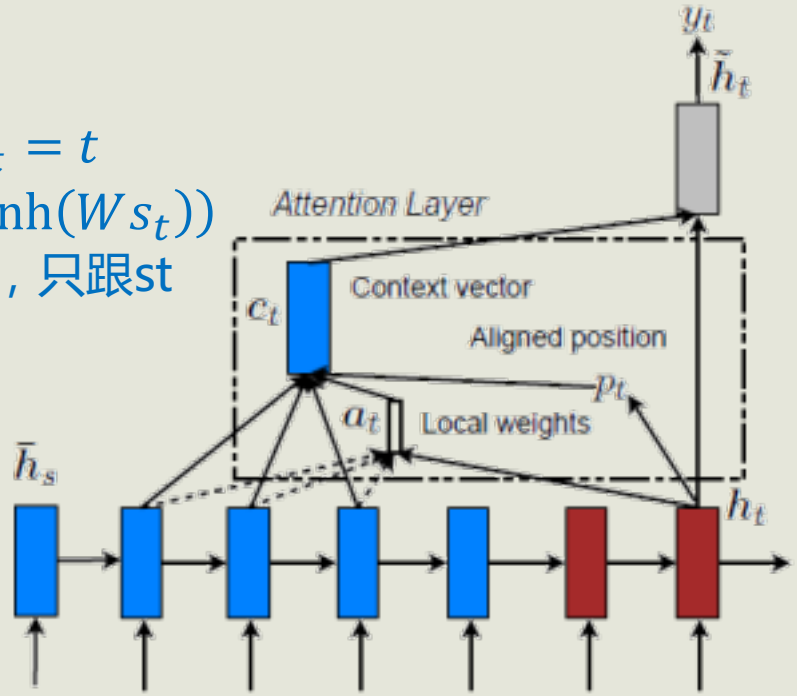
Luong **Global Attention**

- 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

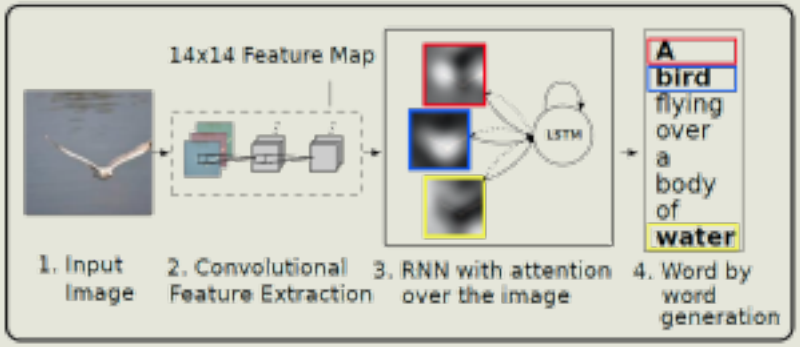
[PDF] Show, attend and tell: Neural image caption generation with visual attention [PDF] jmlr.org

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
- 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.

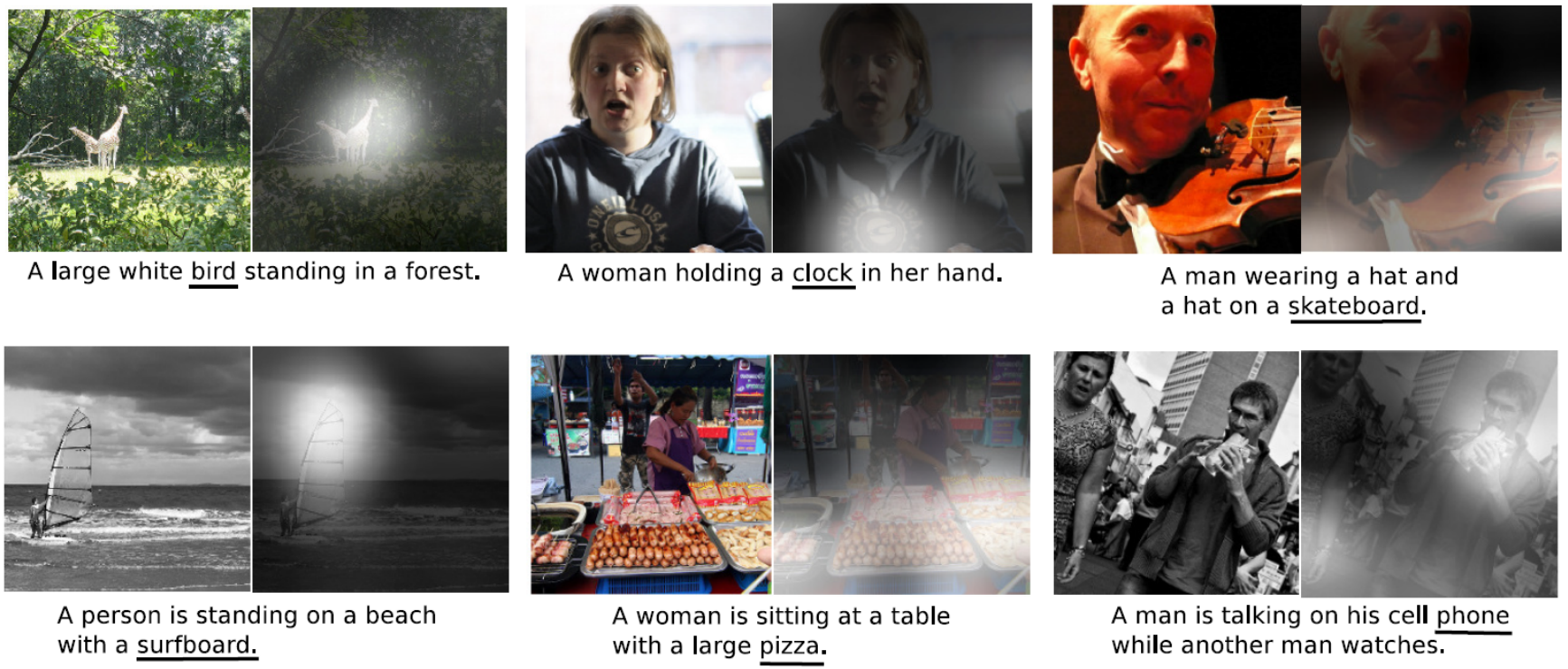
A group of people sitting on a boat in the water.

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.



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
- 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）

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

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

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

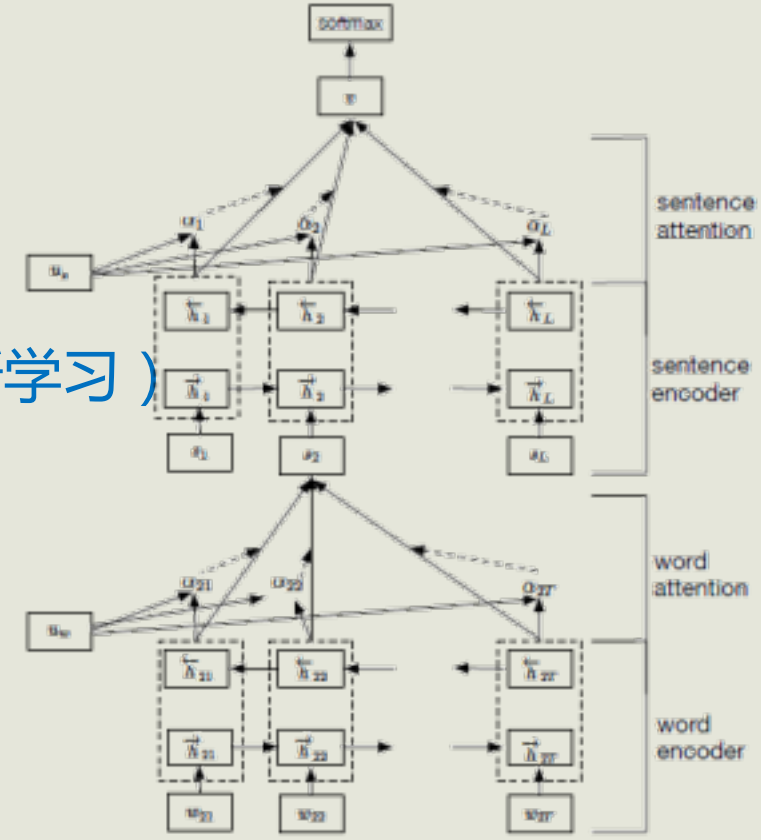


Figure 2: Hierarchical Attention Network.

Self-Attention in Transformer

- 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

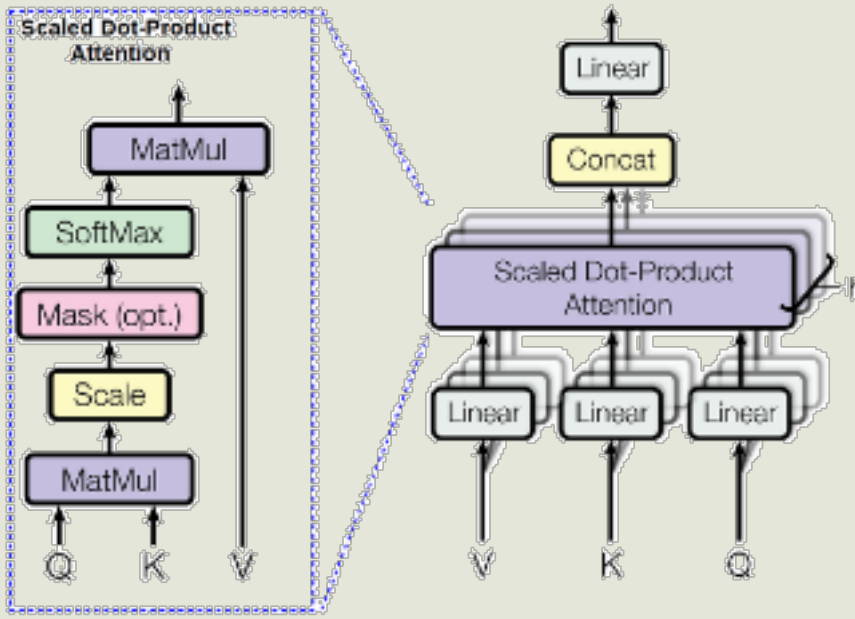
A Vaswani, N Shazeer, N Parmar... - Advances in Neural ..., 2017 - papers.nips.cc

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 ...

☆ 99 被引用次数：267 相关文章 所有 9 个版本 »

Self-Attention in Transformer

- 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



谢谢