RNN条件生成与Attention

七月在线 张雨石

2018年9月13日

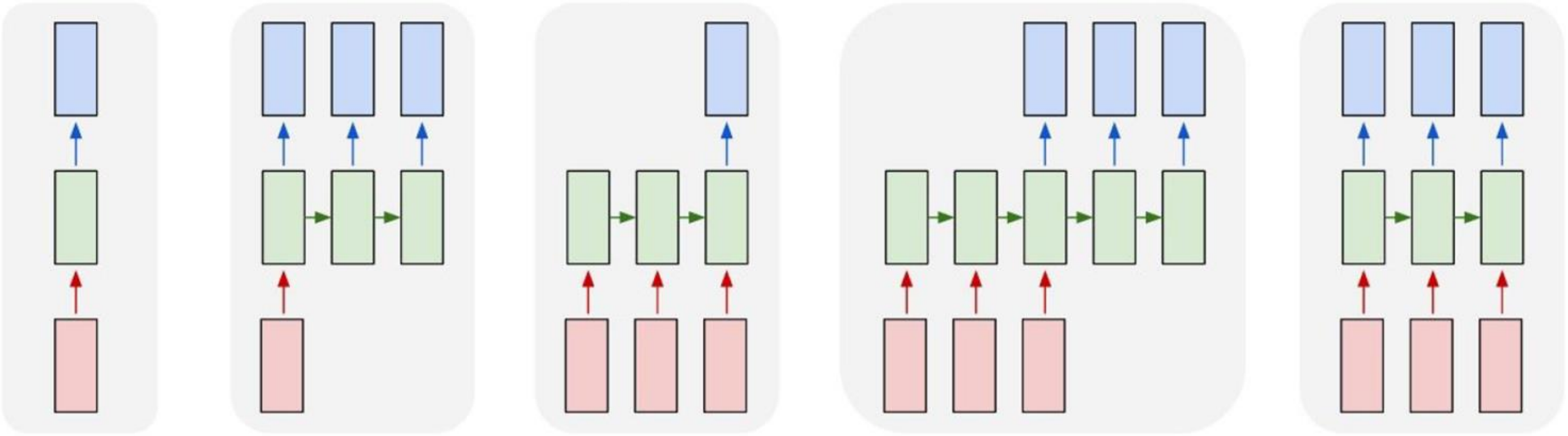
http://blog.csdn.net/stdcoutzyx

RNN条件生成与Attention

- RNN条件生成
- 机器翻译
- Attention
- 图像生成文本

RNN条件生成

- RNN解决的问题



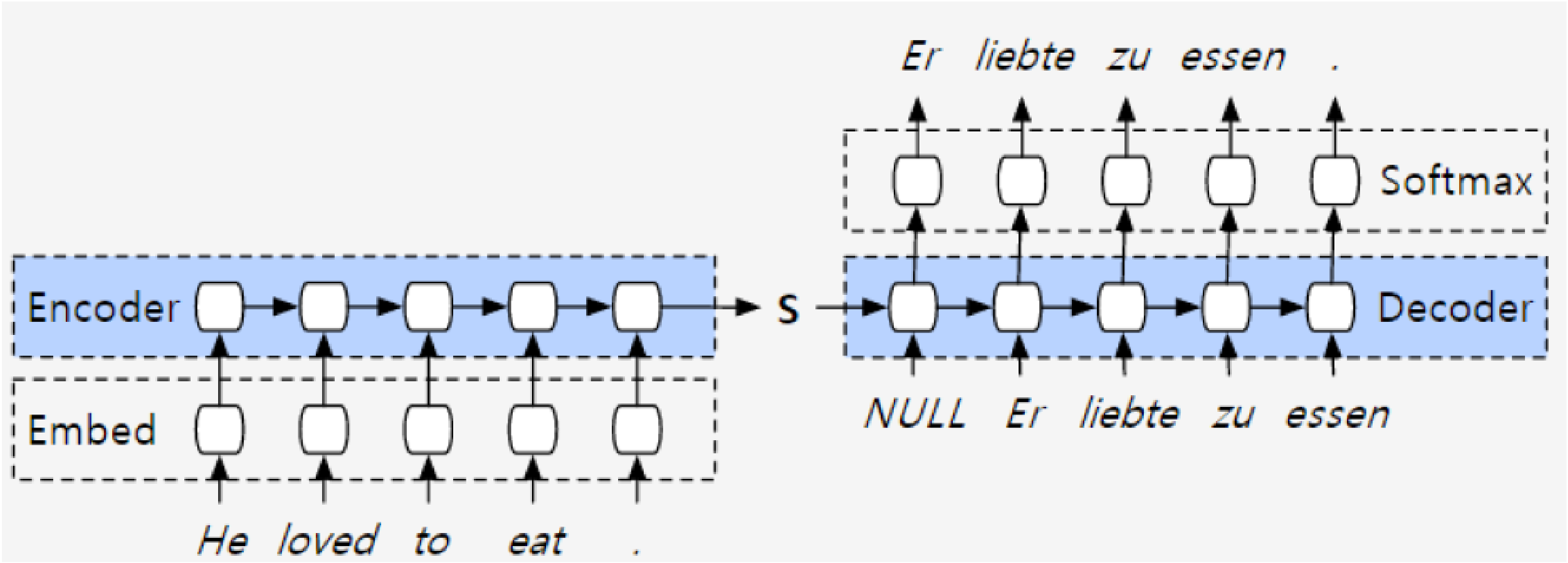
RNN条件生成

- RNN解决的问题
  - 图片生成描述
  - 文本分类
  - 机器翻译
  - 视频解说
- 条件生成问题 $P(y|x)$

机器翻译

- V1: Encoder-Decoder
- V2: Attention-based Encoder-Decoder
- V3: Bi-directional encode layer
- V4: Residual-based Encoder-Decoder
- 总结

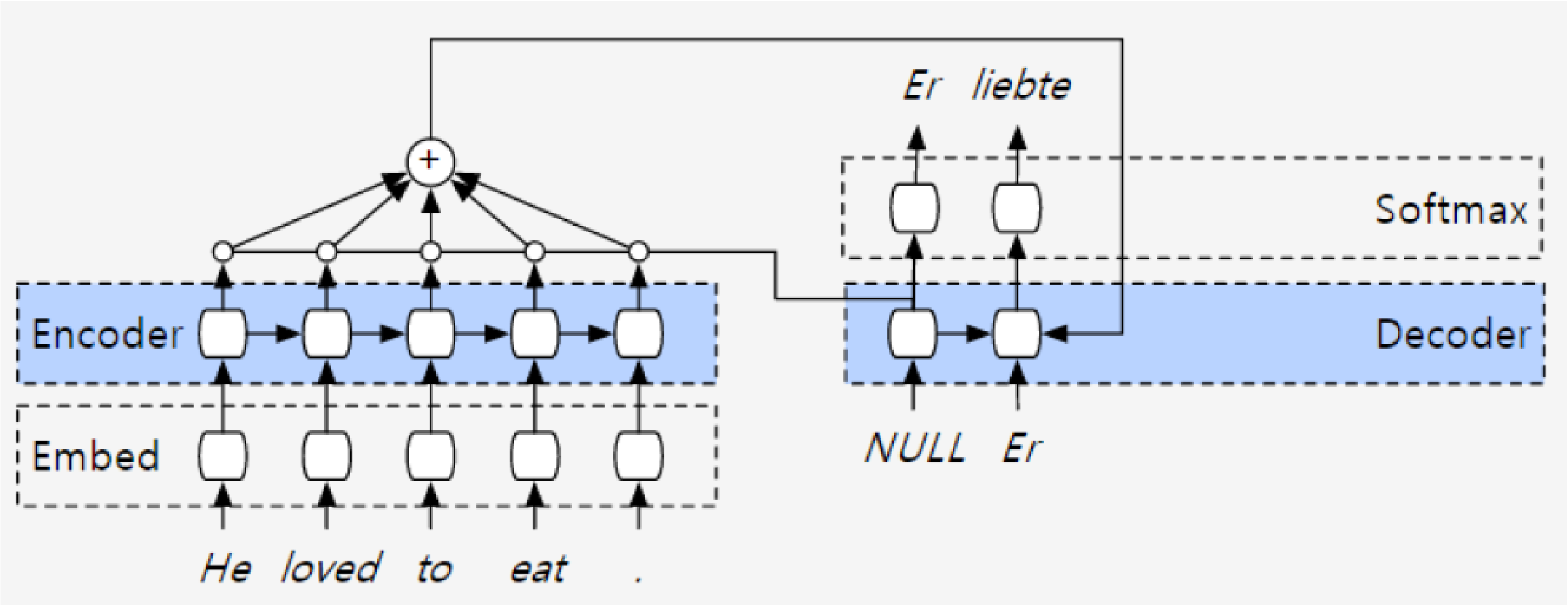
V1: Encoder-Decoder



V1: Encoder-Decoder

- 编码过程
- 解码过程
- 缺点
  - 定长编码是信息瓶颈
  - 长度越长，前面输入进RNN的信息就越被稀释

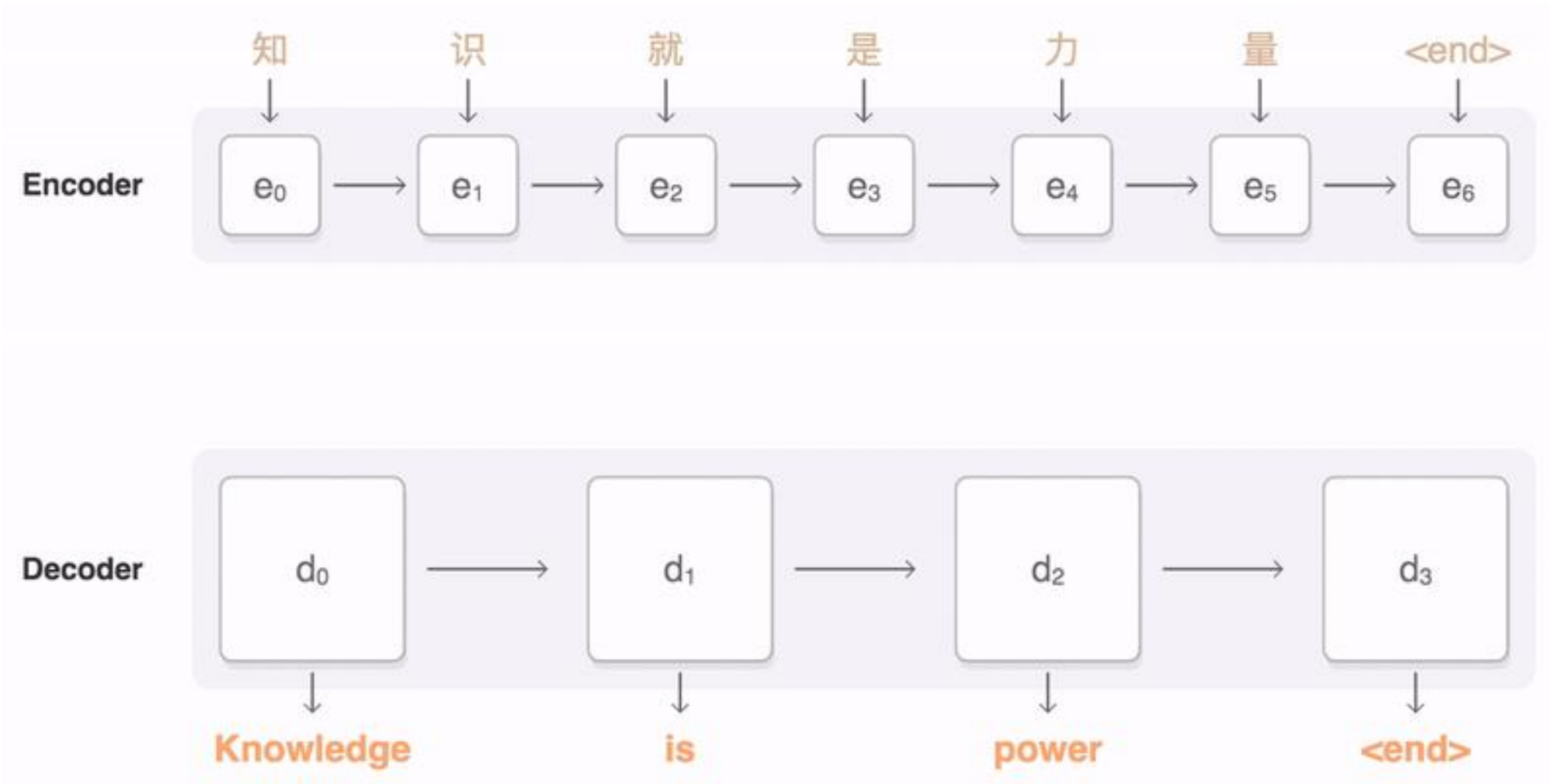
V2: Attention-based Encoder-Decoder



V2: Attention-based Encoder-Decoder

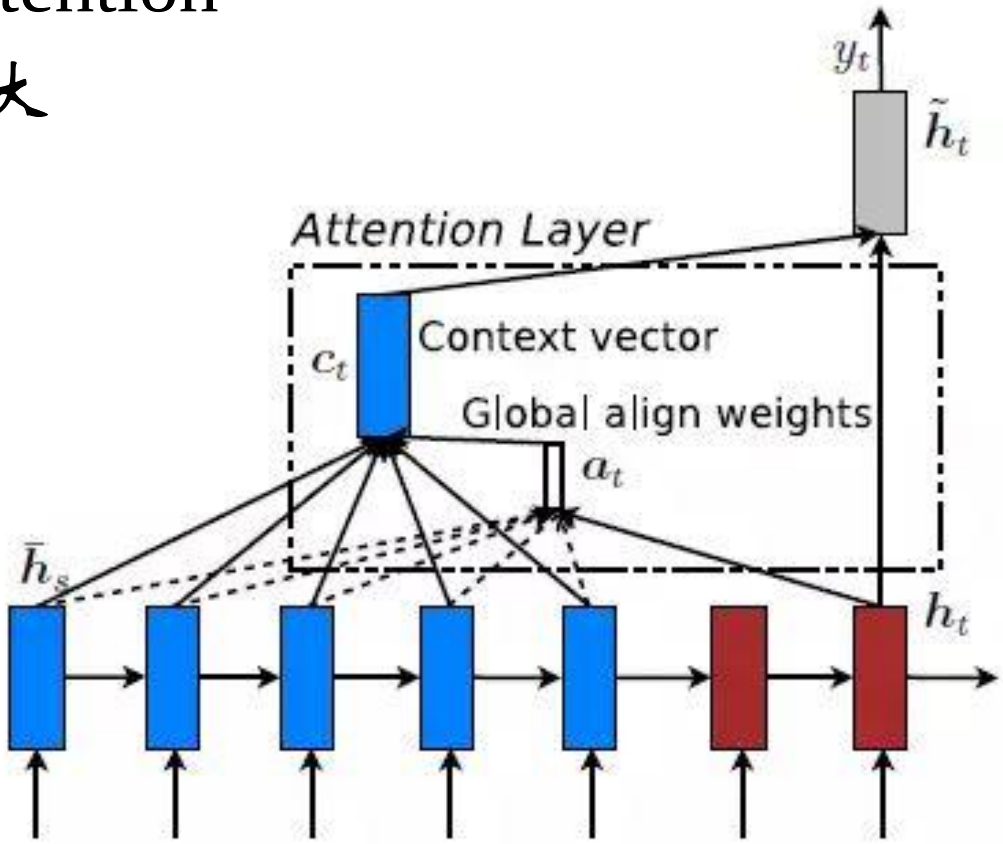
- Attention层
  - Attention计算
    - LSTM各步的输出向量a=[a1,a2,…,an]
    - Decoder每步的中间状态hi
    - Alpha = [tanh(w1*aj + w2*hi) for j in range(n)]
  - Attention加权
    - Alpha * a

V2: Attention-based Encoder-Decoder



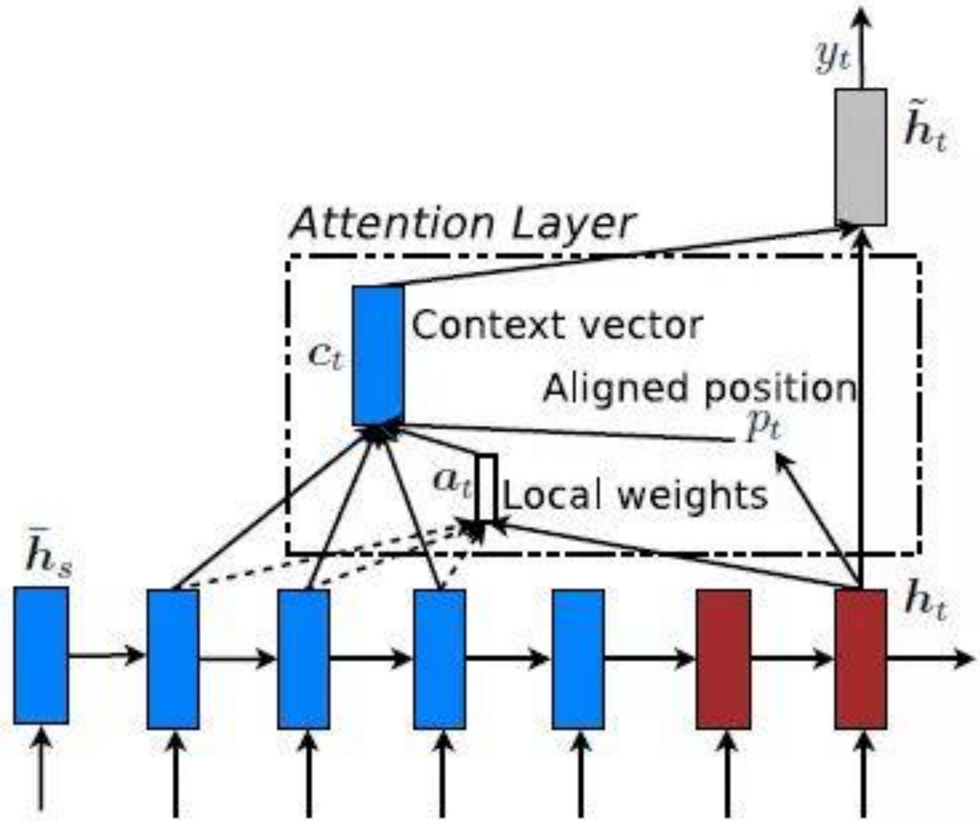
V2: Attention-based Encoder-Decoder

- Global Attention
  - 计算量大

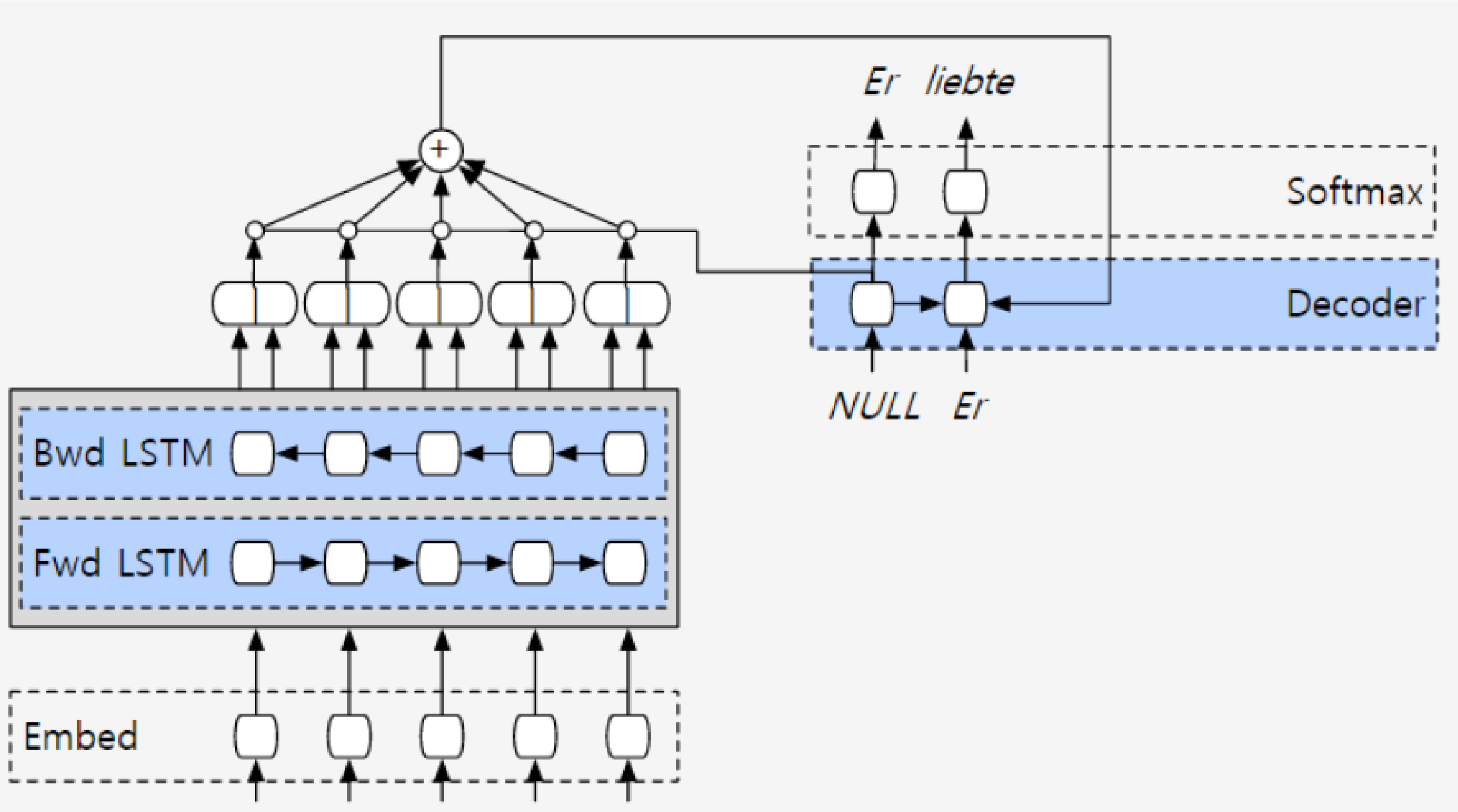


V2: Attention-based Encoder-Decoder

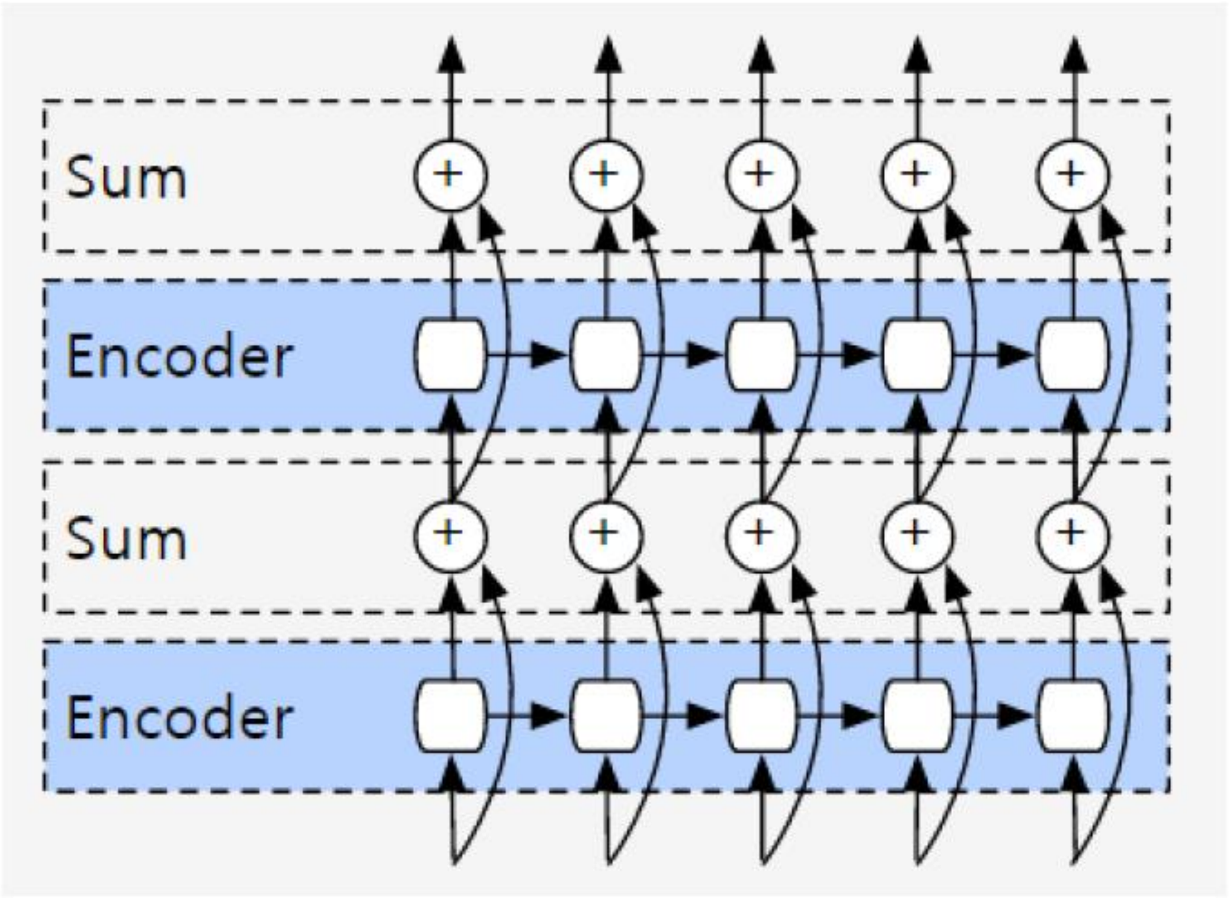
- Local Attention



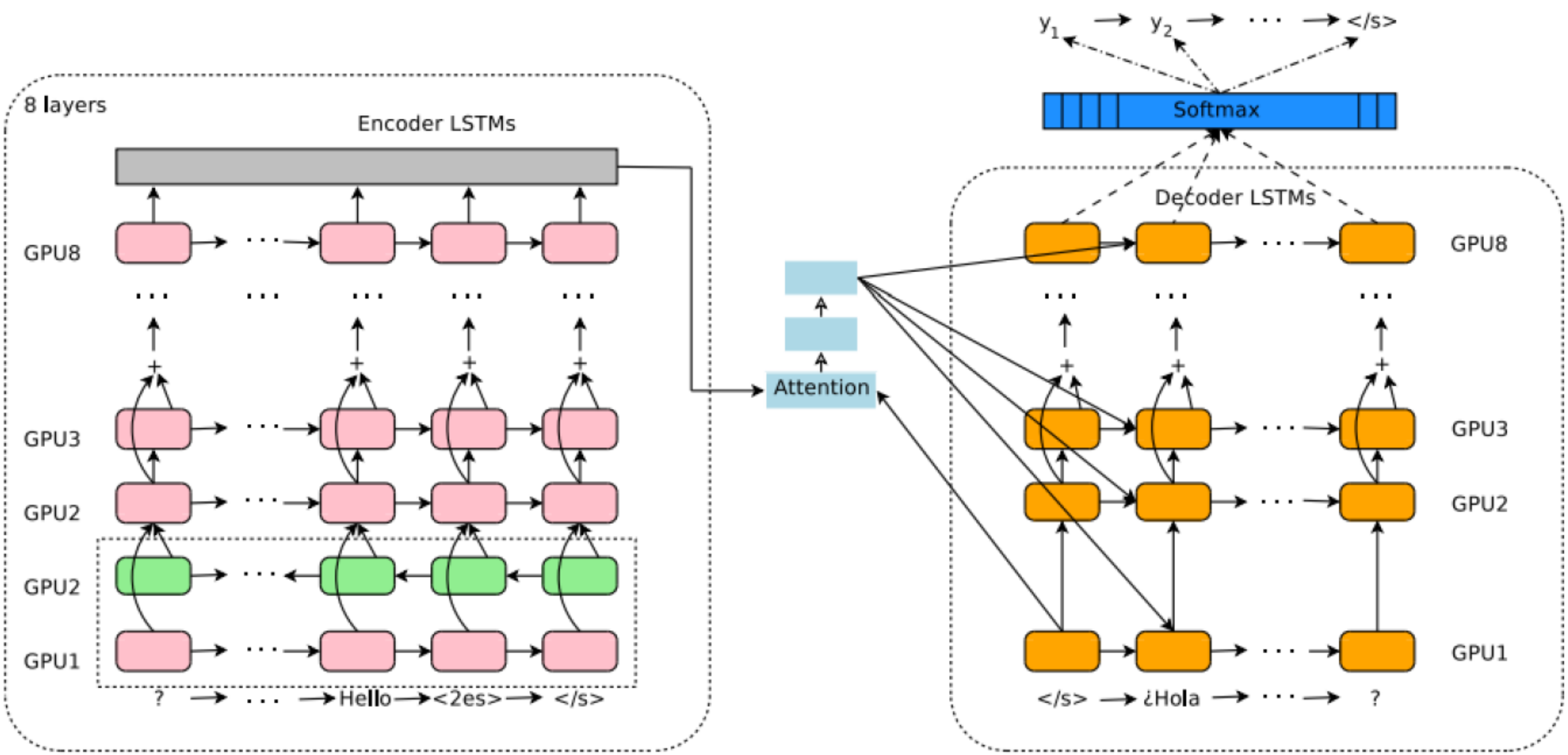
V3: Bi-Directional Encode layer



V4: Residual Encode layer



机器翻译-总结



y_1 → y_2 → ··· → </s>
Softmax
8 layers
Encoder LSTMs
GPU8
GPU3
GPU2
GPU2
GPU1
? → ··· → Hello → <2es> → </s>
Attention
Decoder LSTMs
GPU8
GPU3
GPU2
GPU1
</s> → ¿Hola → ··· → ?

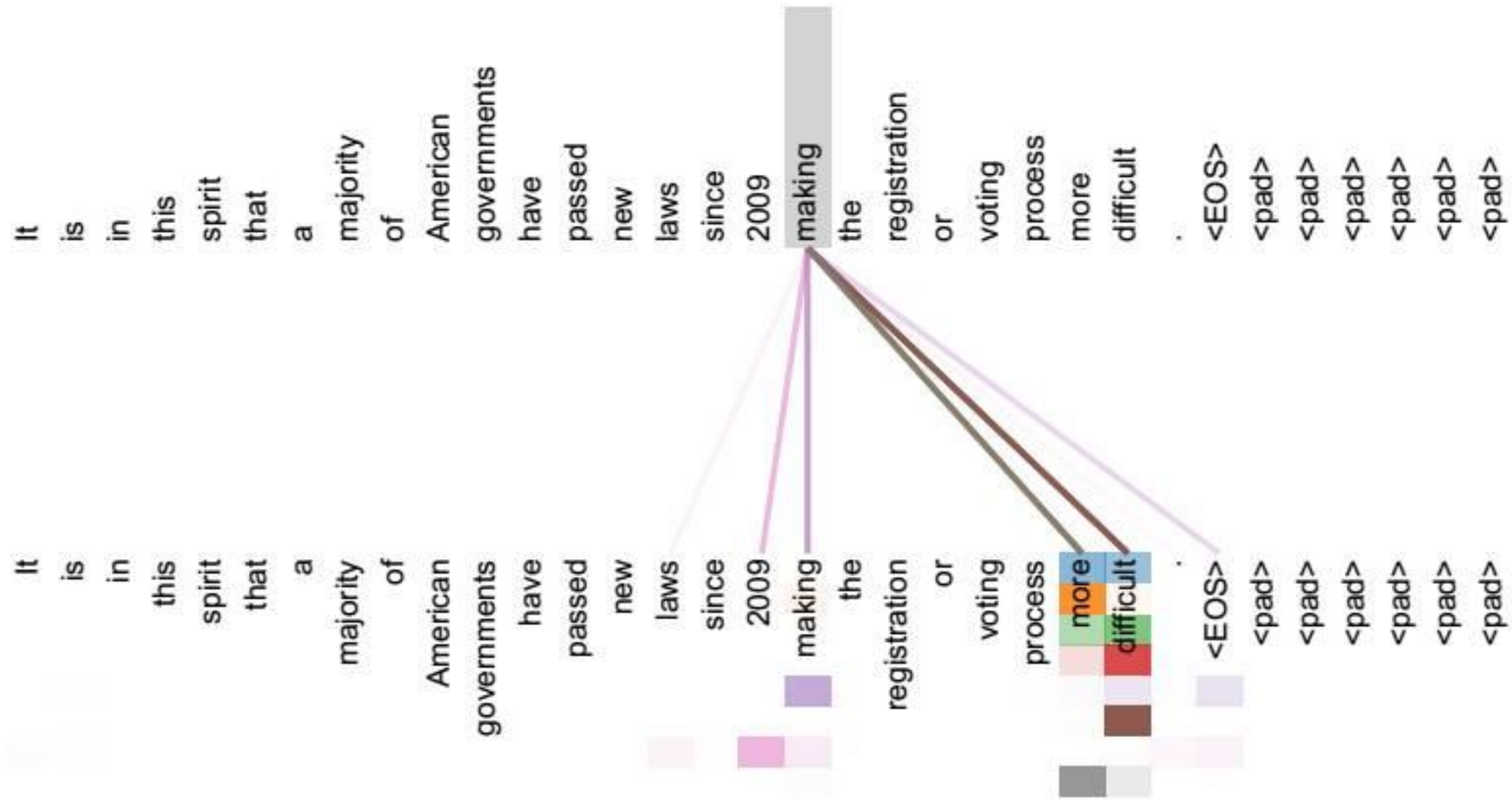
Attention

- Global Attention
- Local Attention
- Self Attention
- Hierarchical Attention

Self Attention

- 传统的attention只有source和target之间的关联关系
- 忽略了source和target端分别的关联关系

Self Attention

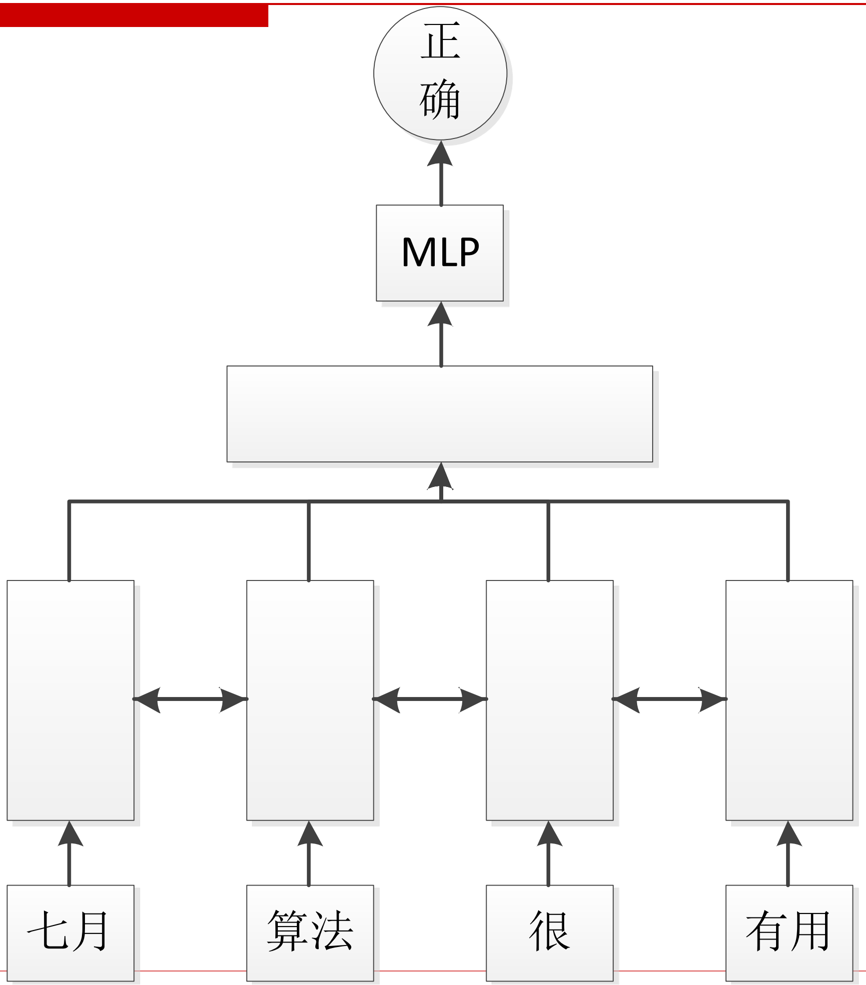


Self Attention

- 可以理解为source=target的encoder-decoder
- 捕捉到词与词之间的关系
- 增加并行性

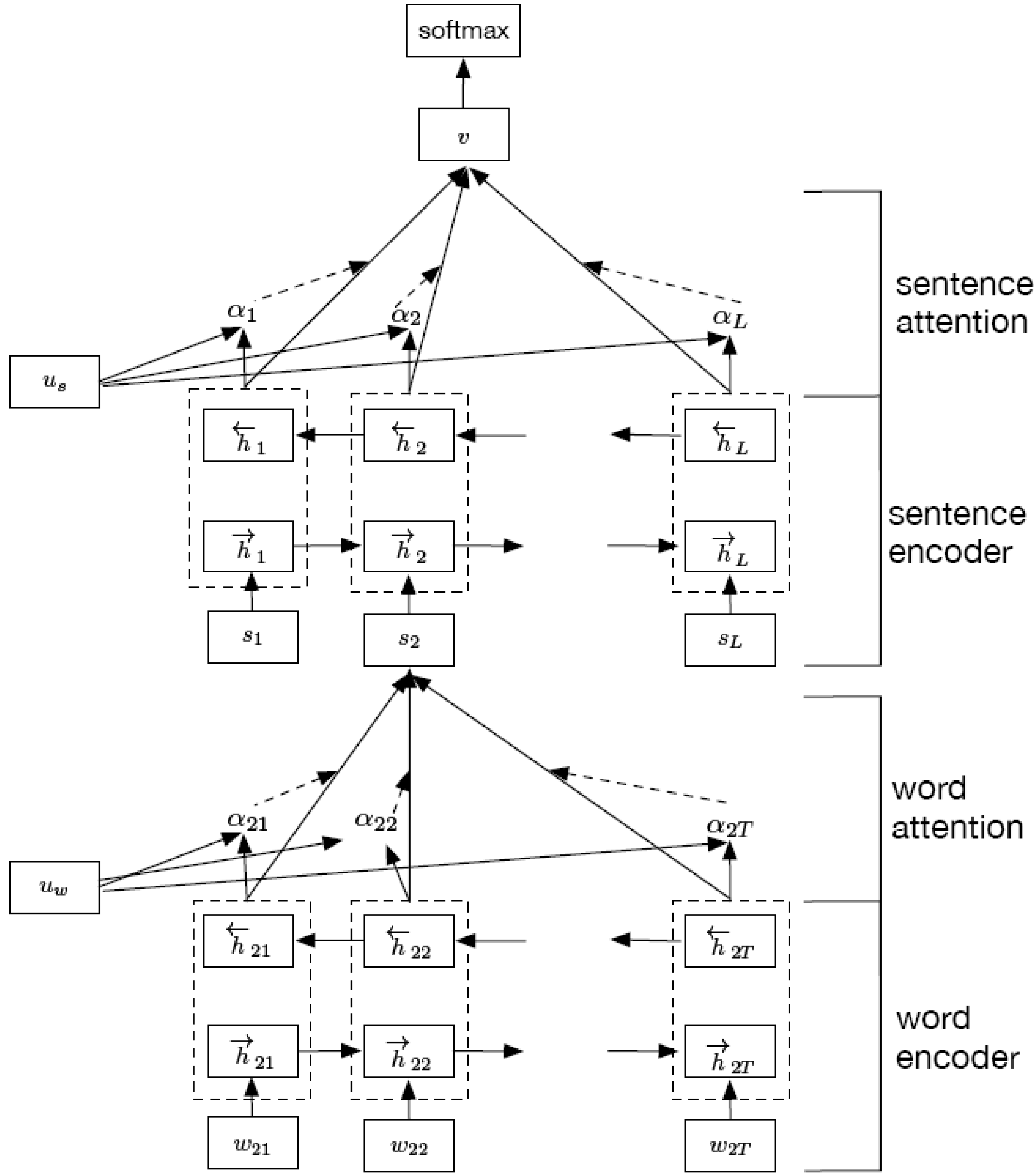
Hierarchical Attention

- 文本分类
  - 双向LSTM
  - 输出合并
    - 拼接
    - Average
    - max



Hierarchical

- 文本分类
  - 分层
    - 词到句
    - 句到段落
  - Attention



图像生成文本

- 问题引入
- 模型变迁
  - M-RNN
  - Neural Image Caption
  - Attention-based

问题引入

- Deep Learning出现之前
  - 检索问题
  - 不能泛化到新的数据

问题引入

- 图像搜索
  - 找到图像对应的描述，丰富搜索元数据
- 盲人导航
- 少儿教育
  - 看图说话

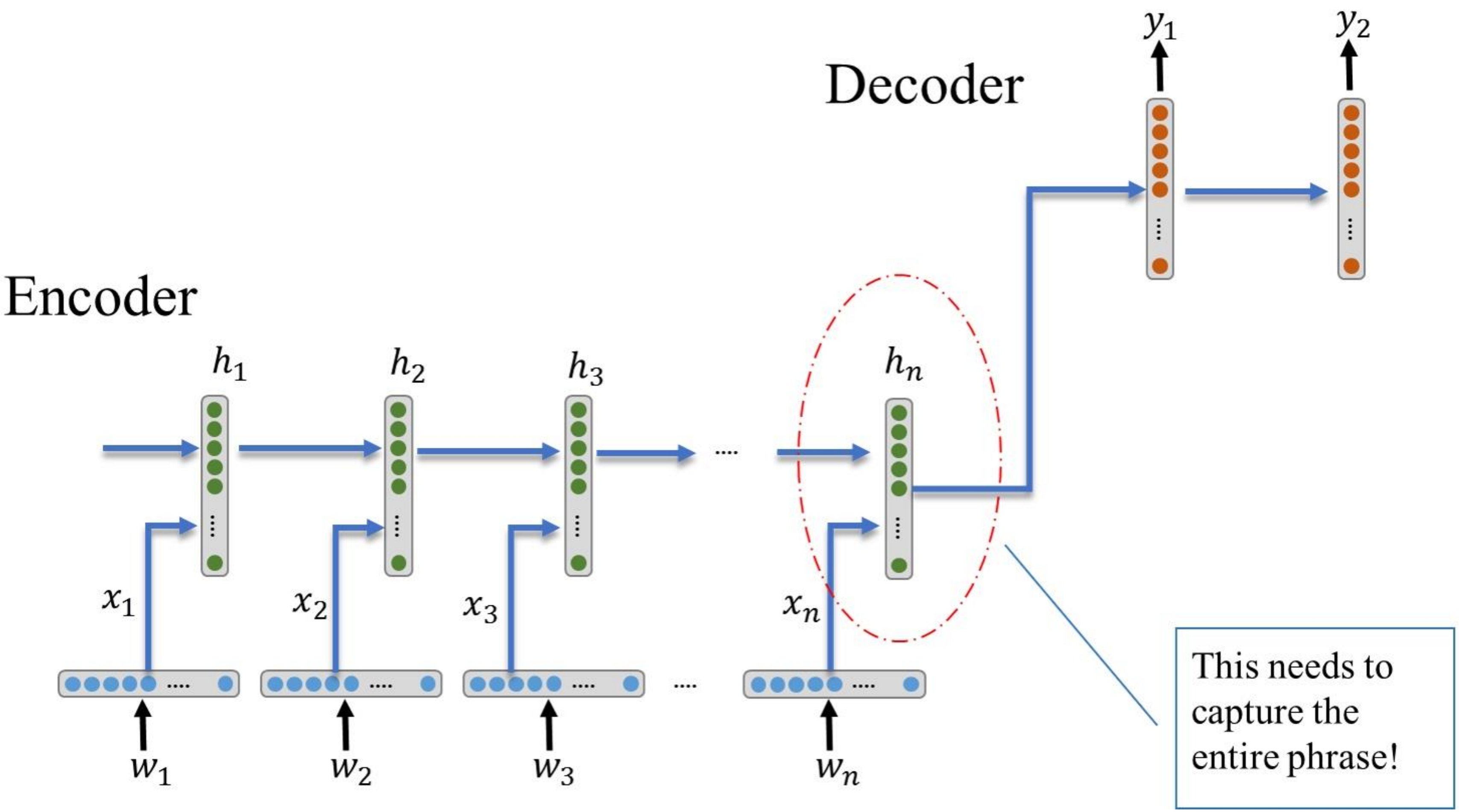
问题引入

- 数据集
  - IAPR TC-12: 20000张图像，平均每张图像有1.7句描述
  - Flickr8k: 8000张图像，语法比IAPR简单
  - Flickr30k: 31783张图像，158915句描述，平均每个五句。
  - MS COCO: 16W张图像
  - AI Challenger: 图像中文描述数据

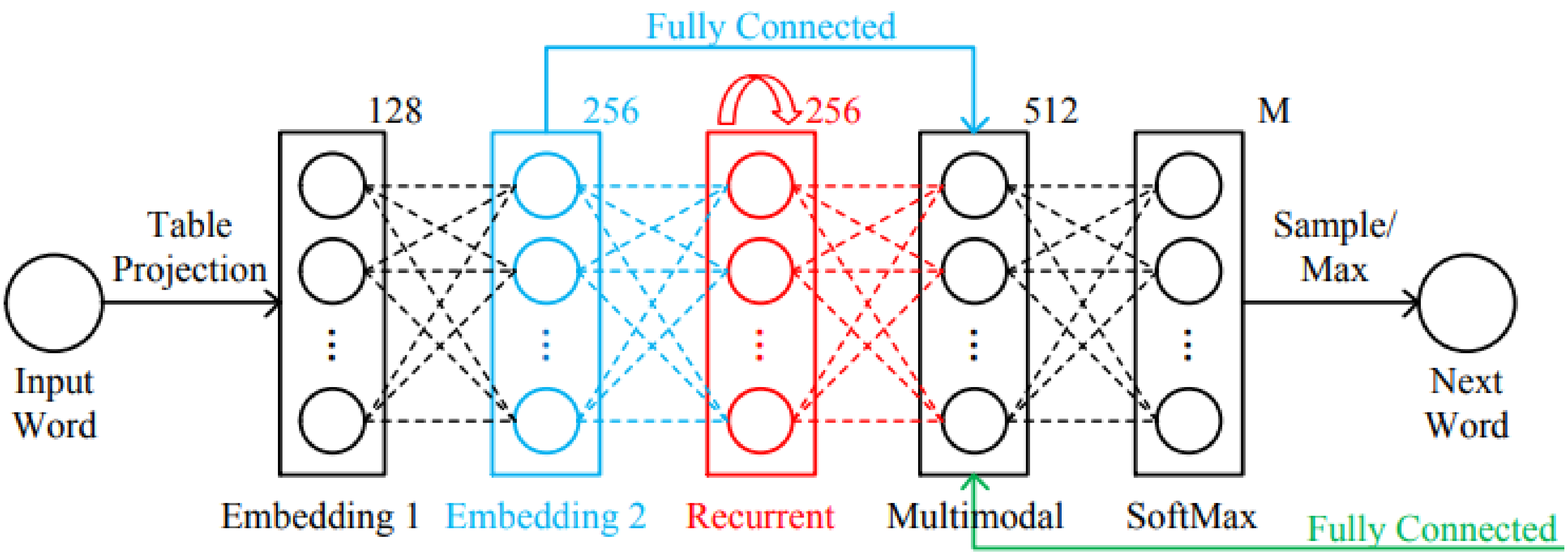
问题引入

- 评测
  - BLEU score
  - 句子检索和图像检索

模型变迁—Encoder-Decoder



模型变迁——Multi-model RNN

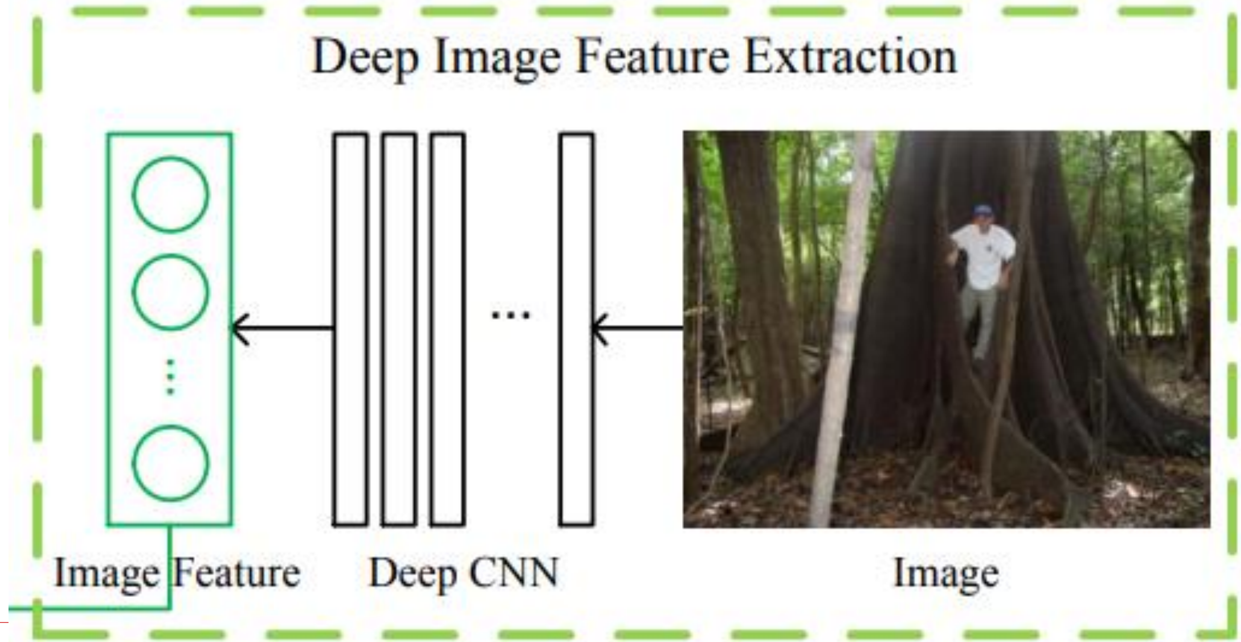


模型变迁——Multi-Model RNN

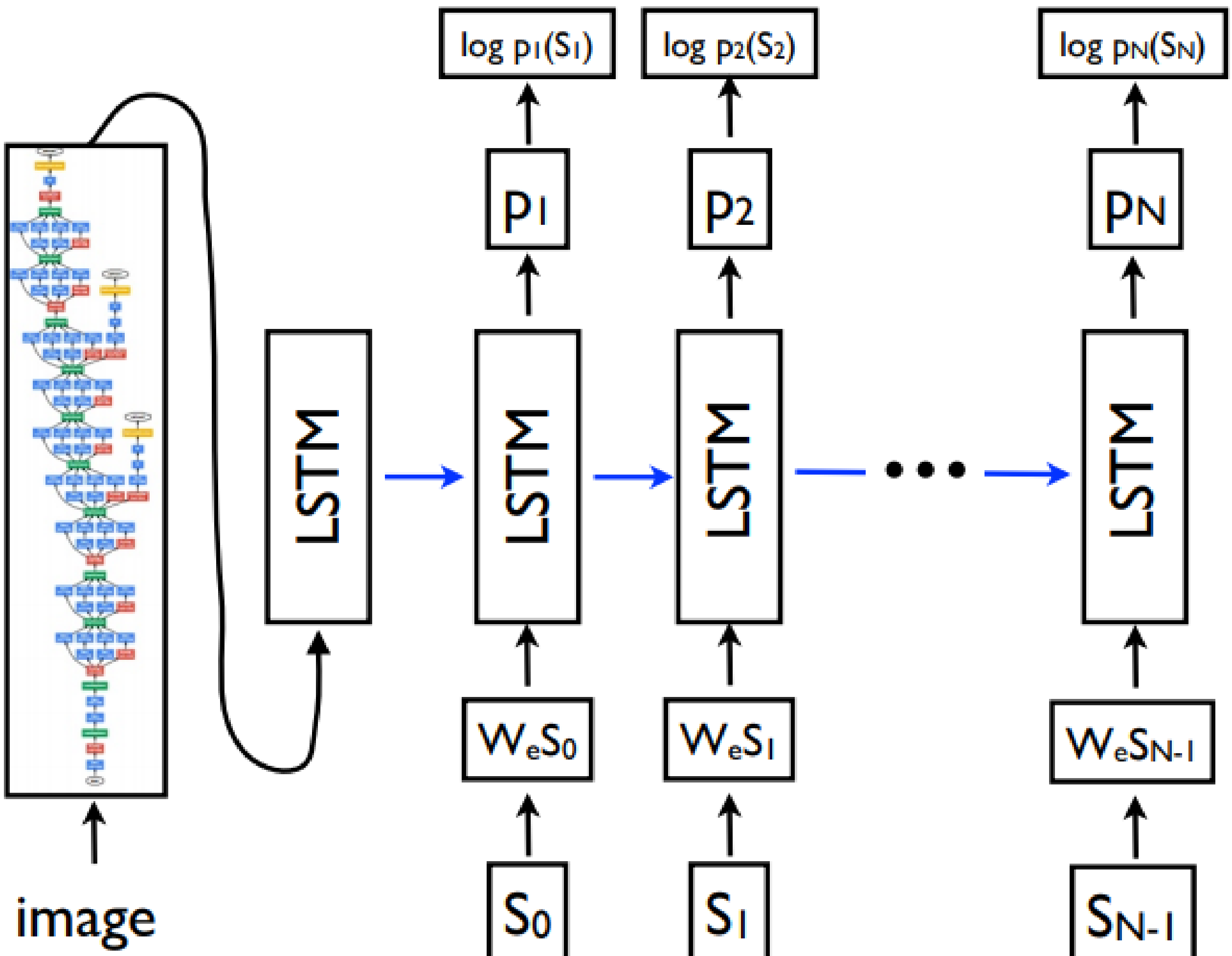
- 输入词生成embedding
- Embedding输入到RNN和Multi-Model
- RNN生成更加抽象的embedding
- 原始Embedding、RNN的embedding和图像Feature同时输入给分类层

模型变迁——Multi-model RNN

- AlextNet 7^{th} feature
- Object detection feature



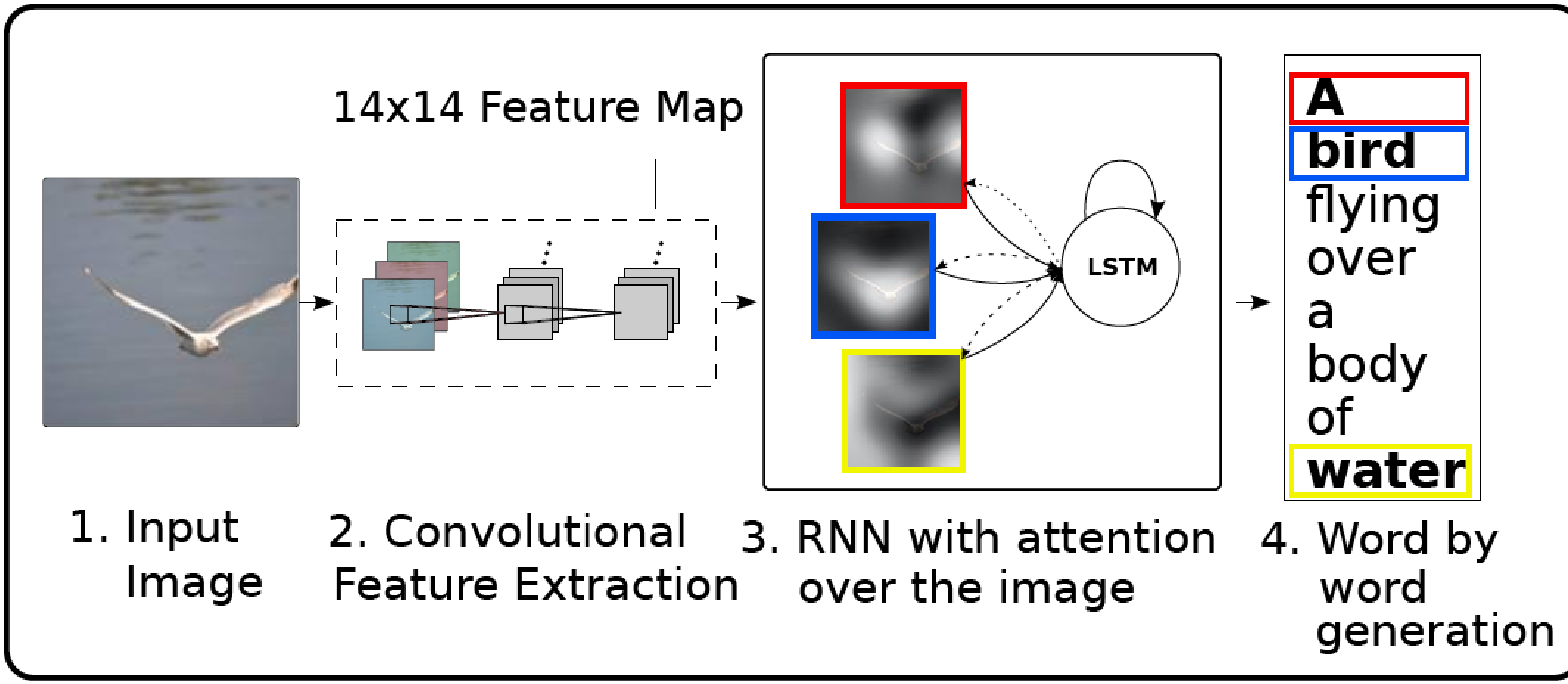
模型变迁——Show And Tell



模型变迁——Show And Tell

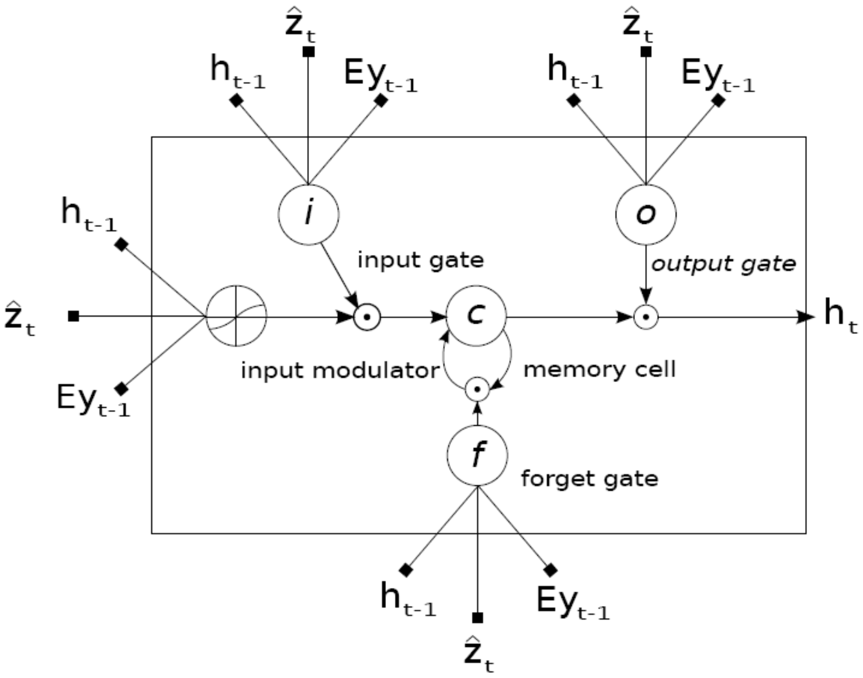
- 图像特征使用更强大的CNN提取
  - GoogNet、Residual等
- 图像特征只提取一次
- LSTM生成文本

模型变迁——Show Attend and Tell



模型变迁——Show Attend and Tell

- 不使用全连接层而是某个卷积层的feature
  - 不使用全连接层
  - 使用卷积层的输出
    - 有位置信息
  - LSTM输入是加权和
    - attention



模型变迁——Show Attend and Tell

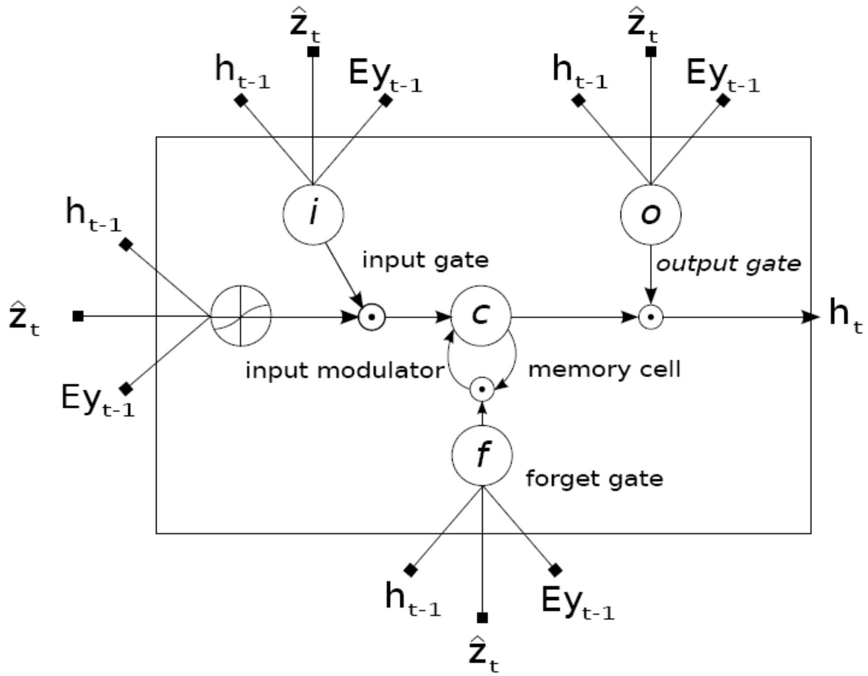
- 不使用全连接层而是某个卷积层的feature
  - 14*14*256：有14*14个位置可以选择

$$a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\},\ \mathbf{a}_i \in \mathbb{R}^D$$

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}$$

$$\hat{\mathbf{z}}_t = \phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right)$$



模型变迁——Show Attend and Tell

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

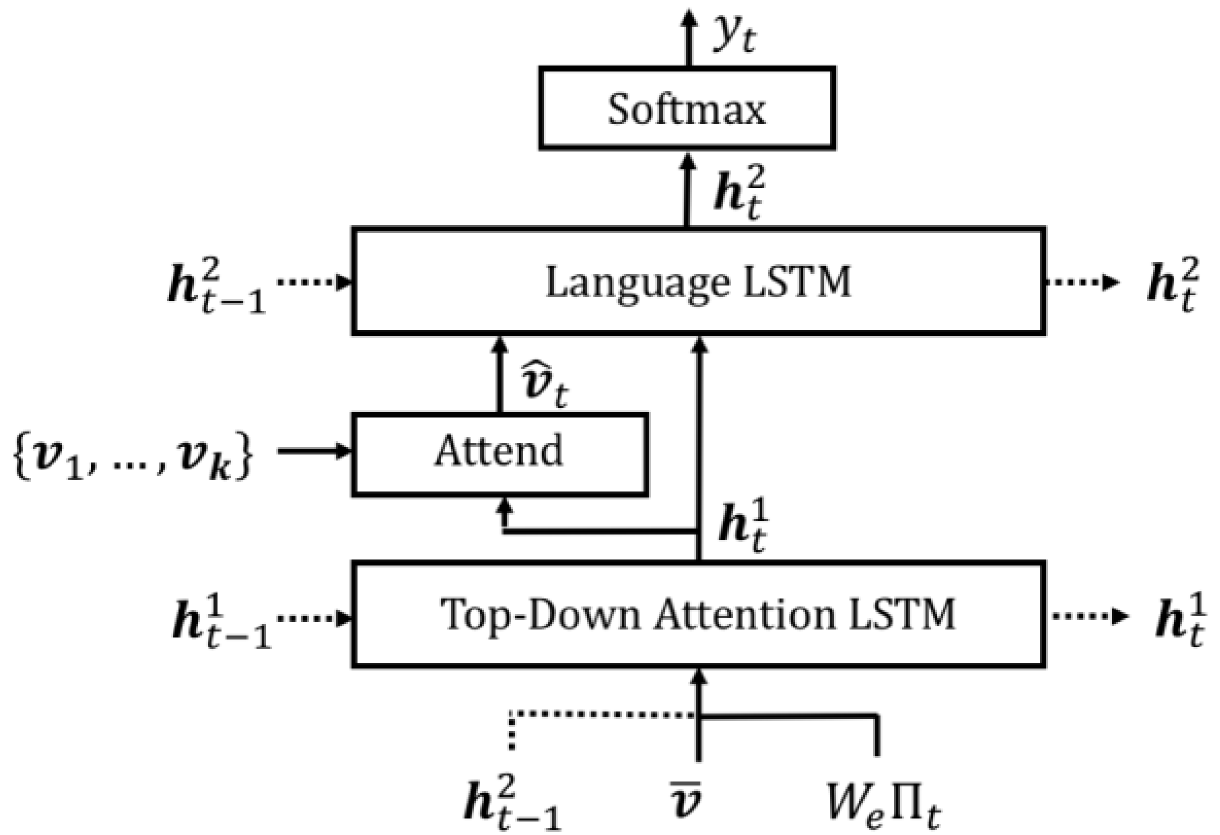
A stop sign is on a road with a mountain in the background.

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.

模型变迁——Top Down Attention

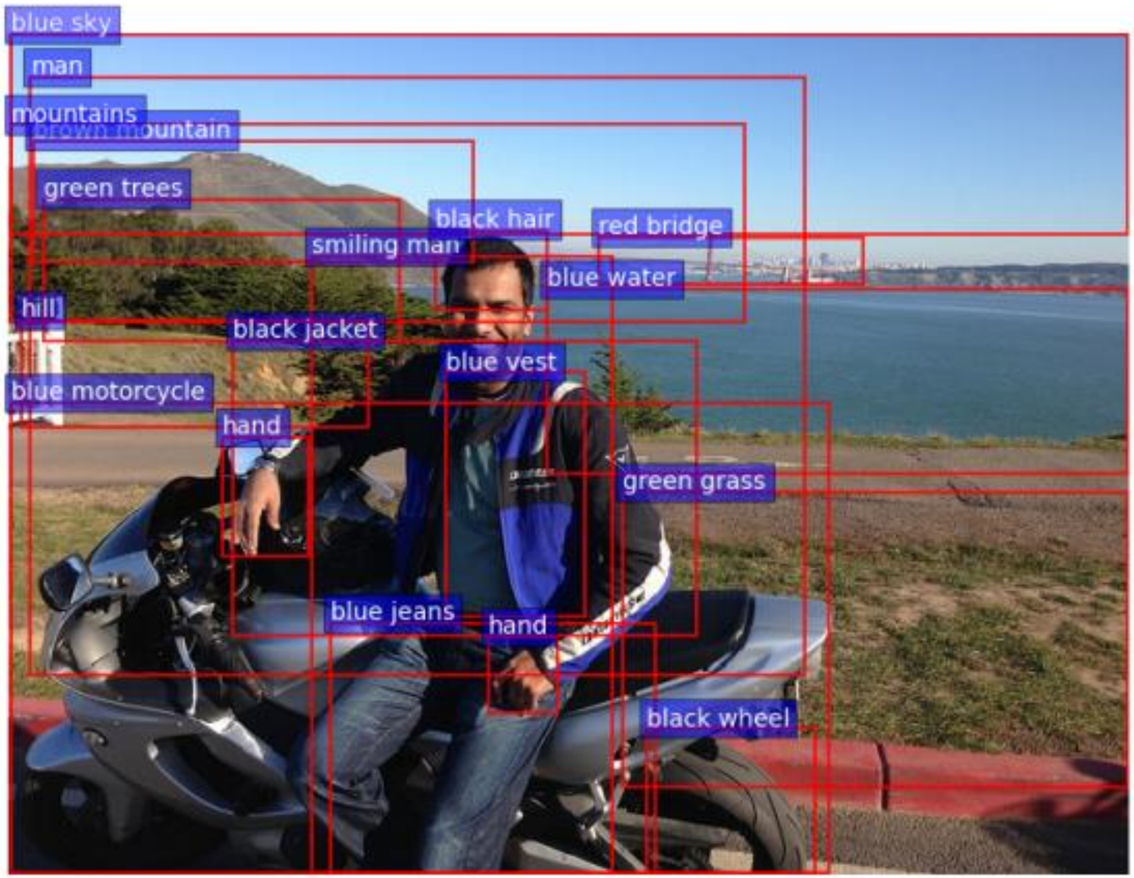


模型变迁——Top Down Attention

- 第一层LSTM负责学习attention
  - 前一时刻Language Model的hidden state
  - 图像的全局信息
    - 所有feature map的平均
  - 输入词语的embedding

$$\boldsymbol{x}_t^1 = [\boldsymbol{h}_{t-1}^2, \bar{\boldsymbol{v}}, W_e \Pi_t]$$

模型变迁——Top Down Attention



模型变迁——Top Down Attention

- 第一层LSTM输出和图像feature计算attention

$$a_{i,t} = \boldsymbol{w}_a^T \tanh\left(W_{va}\boldsymbol{v}_i + W_{ha}\boldsymbol{h}_t^1\right)$$
$$\boldsymbol{\alpha}_t = \text{softmax}\left(\boldsymbol{a}_t\right)$$

$$\hat{\boldsymbol{v}}_t = \sum_{i=1}^{K} \alpha_{i,t}\boldsymbol{v}_i$$

模型变迁——Top Down Attention

- 第二层LSTM
  - 第一层LSTM的隐含层
  - 加权平均的图像feature

$$\boldsymbol{x}_t^2 = [\hat{\boldsymbol{v}}_t, \boldsymbol{h}_t^1]$$

$$p(y_t \mid y_{1:t-1}) = \operatorname{softmax}\left(W_p \boldsymbol{h}_t^2 + \boldsymbol{b}_p\right)$$

总结

- RNN条件生成问题
- 机器翻译
  - Encoder-Decoder、attention、双向、残差
- Attention
  - Self-attention、Hierarchical attention
- 图像生成文本
  - Multi-Modal、Show and Tell
  - Show attend and tell、 Top-down bottom-up attention

Thanks!

Q&A