

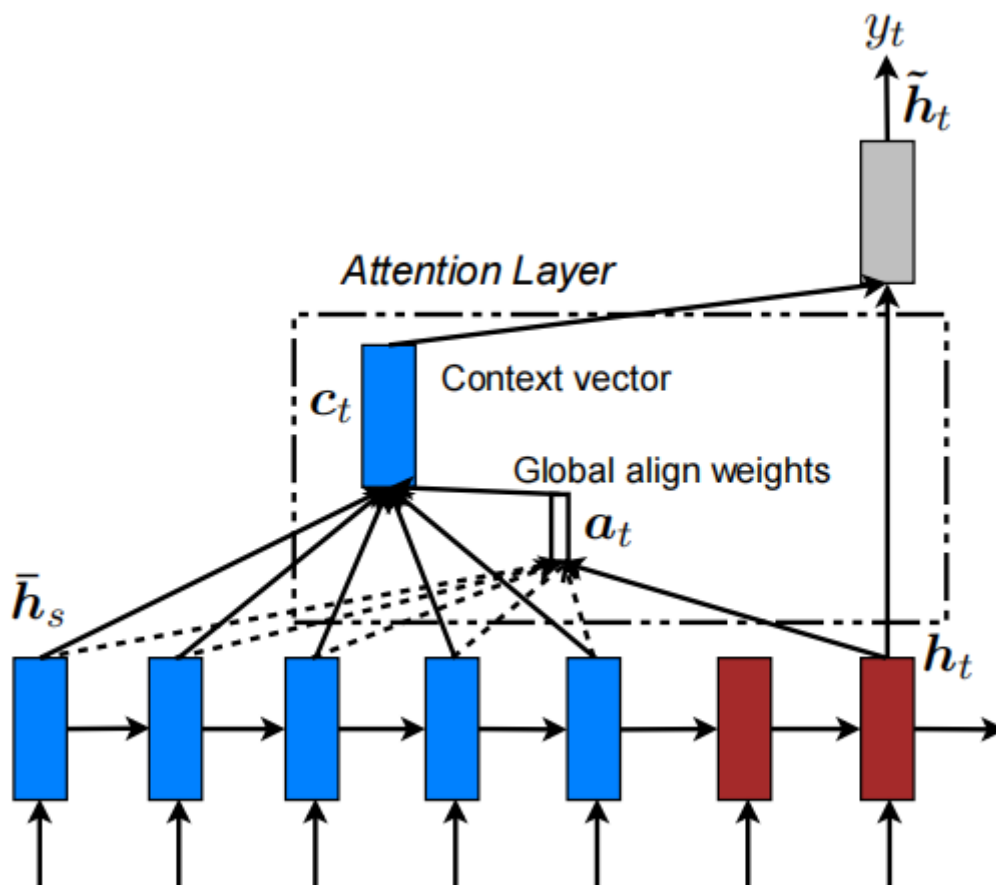
Luong Attention

2022-05-07

1. Attention 架构
2. Attention 计算流程
3. 参考资料

1. Attention 架构

Luong Attention分为Local和Global两种，本文主要分析Global Attention。下图为Global Attention的架构图：



符号解释：

1. \bar{h}_s : encoder_output
2. h_t : decoder_output
3. $a_t(s)$: attn_weights
4. c_t : Context vector
5. \tilde{h}_t : attentional hidden state

以中英文翻译场景为例，根据该架构图，分为如下计算步骤：

1. attn_weights计算

通过encoder_output和decoder_output计算得到attn_weights，即 a_t

2. Context vector计算

通过encoder_output和 a_t 计算得到加权encoder_output，即 c_t

3. attentional hidden state计算

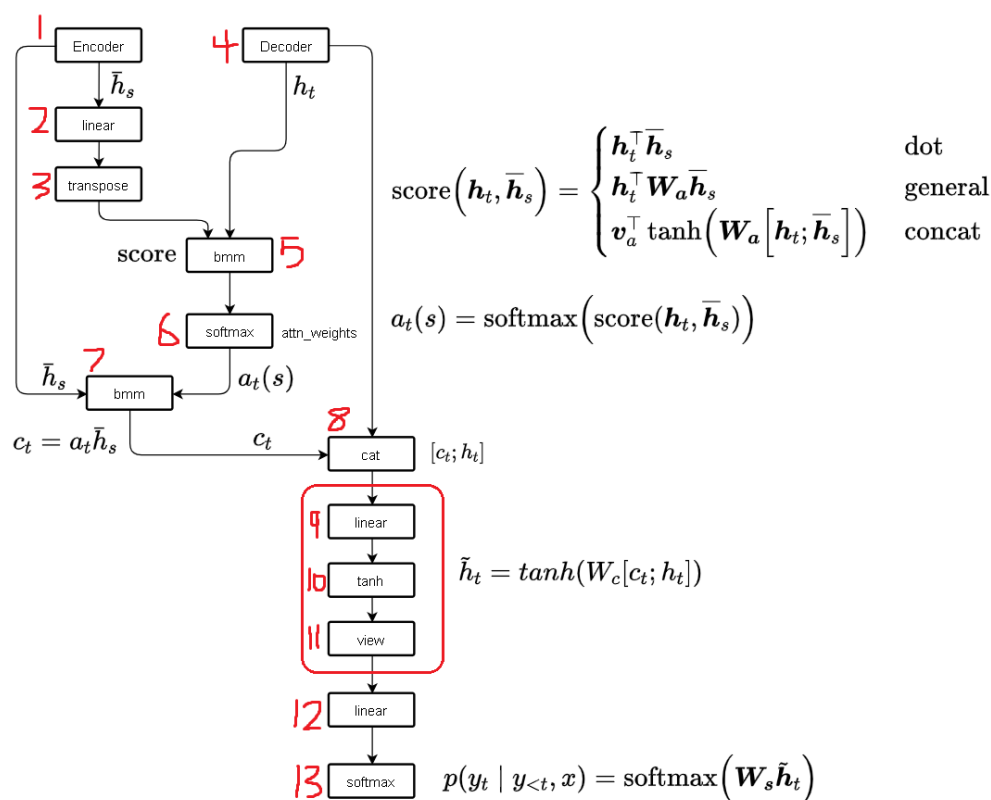
将 c_t 和decoder_output进行cat合并，经过tanh和linear变换处理得到attentional hidden state，即 \tilde{h}_t

4. 预测

根据 \tilde{h}_t 进行预测，得到最终预测结果。

2. Attention计算流程

Global Attention计算流程如下图所示：



计算步骤如下：

1. Encoder

- step1: 对原始输入，通过RNN（如LSTM）处理，得到encoder_output：

$$\bar{h}_s : [\text{batch_size}, \text{input_len}, \text{enc_hidden_size}]$$

- step2: 为了能够使得encoder_output和decoder_output做bmm运算，需要进行linear处理

$$\bar{h}_s : [\text{batch_size}, \text{input_len}, \text{dec_hidden_size}]$$

- step3: $\text{transpose}(1,2)$

$\bar{h}_s : [\text{batch_size}, \text{dec_hidden_size}, \text{input_len}]$

2. Decoder

- RNN

- step4: 对Decoder端的输入, 通过RNN (如LSTM) 处理, 得到decoder_output:

$h_t : [\text{batch_size}, \text{output_len}, \text{dec_hidden_size}]$

- Attention

- step5: 基于 h_t 和 \bar{h}_s , 进行打分计算:

$\text{score}(h_t, \bar{h}_s) = \text{bmm}(h_t, \bar{h}_s) : [\text{batch_size}, \text{output_len}, \text{input_len}]$

- step6: 对打分结果通过softmax计算, 得到attn_weights

$a_t(s) = \text{align}(h_t, \bar{h}_s) = \text{softmax}(\text{score}(h_t, \bar{h}_s)) : [\text{batch_size}, \text{output_len}, \text{input_len}]$

- step7: 基于 $a_t(s)$, 对encoder_output求加权平均

$c_t = a_t \bar{h}_s : [\text{batch_size}, \text{output_len}, \text{enc_hidden_size}]$

- Attentional hidden state

- step8: 将加权encoder_output通过cat操作"融入"到原始的decoder_output

$[c_t; h_t] : [\text{batch_size}, \text{output_len}, \text{enc_hidden_size} + \text{dec_hidden_size}]$

为了方面后续的linear变换, 需要对其shape进行调整, 结果如下:

$[c_t; h_t] : [\text{batch_size} \times \text{output_len}, \text{enc_hidden_size} + \text{dec_hidden_size}]$

- step9: 对 $[c_t; h_t]$ 进行linear变换

$\text{linear}([c_t; h_t]) : [\text{batch_size} \times \text{output_len}, \text{dec_hidden_size}]$

- step10: tanh变换

$\text{tanh}(\text{linear}([c_t; h_t])) : [\text{batch_size} \times \text{output_len}, \text{dec_hidden_size}]$

- step11: 将二维展开到三维, 得到最终的attentional hidden state

$\tilde{h}_t : [\text{batch_size}, \text{output_len}, \text{dec_hidden_size}]$

- Predict

- step12: 对 \tilde{h}_t 进行linear变换

$\text{linear}(\tilde{h}_t) : [\text{batch_size}, \text{output_len}, \text{vocab_size}]$

- step13: 通过softmax运算, 得到最终预测概率结果

$\text{softmax}(\text{linear}(\tilde{h}_t)) : [\text{batch_size}, \text{output_len}, \text{vocab_size}]$

3. 参考资料

1. [第七课 Seq2Seq与Attention\(julyedu.com\)](http://julyedu.com)