seq2seq中的两种attention机制（图+公式）

前言

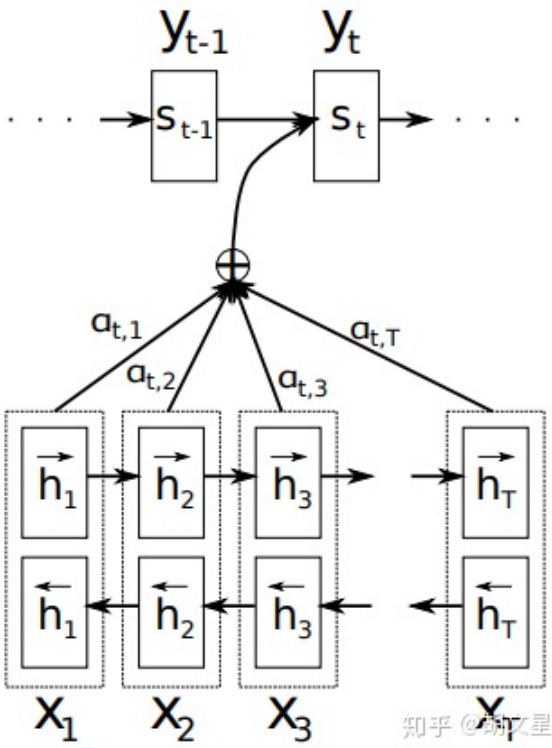
本文来讲一讲应用于seq2seq模型的两种attention机制：Bahdanau Attention和Luong Attention。文中用公式+图片清晰地展示了两种注意力机制的结构，最后对两者进行了对比。seq2seq传送门：click here.

文中为了简洁使用基础RNN进行讲解，当然一般都是用LSTM，这里并不影响，用法是一样的。另外同样为了简洁，公式中省略掉了偏差。

第一种attention结构：Bahdanau Attention

两种机制基于上篇博客第一种seq2seq结构。Encoder生成的语义向量 c 会传给Decoder的每一时刻，传给每一时刻的语义向量都是同一个 c ，这是不合理的。比如翻译一句话，`I like watching movie.`翻译成：`我喜欢看电影。`，其中`喜欢`基本上是由`like`得来的，`I like watching movie.`中每个词对翻译成`喜欢`的影响是不同的。所以，在Decoder中，每个时刻的语义向量 c_t 都应该是不同的。

该模型来自于Bahdanau et.al(2014)，模型框架如下图：



计算公式如下更方便理解。

Encoder：

$$
\begin{aligned}
h_i &= tanh(W[h_{i-1}, x_i]) \\
o_i &= softmax(Vh_i)
\end{aligned}
$$

Decoder：

分为两步：
第一步，生成该时刻语义向量：

$$
\begin{aligned}
c_t &= \sum_{i=1}^{T} \alpha_{ti} h_i \\
\alpha_{ti} &= \frac{exp(e_{ti})}{\sum_{k=1}^{T} exp(e_{tk})} \\
e_{ti} &= v_a^{\top} tanh(W_a[s_{i-1}, h_i])
\end{aligned}
$$

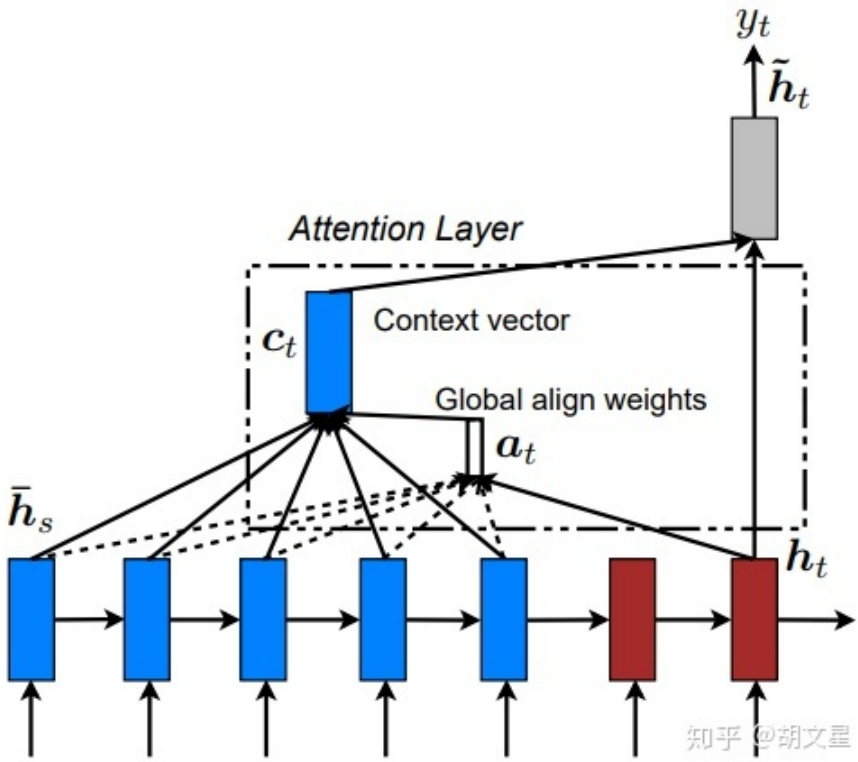
其中 c_t 是 t 时刻的语义向量； e_{ti} 是Encoder中 i 时刻 Encoder隐层状态 h_i 对Decoder中 t 时刻隐层状态 s_t 的影响程度；通过softmax函数（第二个式子）将 e_{ti} 概率归一化为 α_{ti} 。

第二步，传递隐层信息并预测：

$$
\begin{aligned}
s_t &= tanh(W[s_{t-1}, y_{t-1}, c_t]) \\
o_t &= softmax(Vs_t)
\end{aligned}
$$

第二种attention结构：Luong Attention

该模型来自于Luong et.al(2015)，模型框架如下图：



与第一种attention结构**区别在Decoder部分**，Encoder部分完全相同。Decoder还是分两步，与前者的区别部分在公式中用绿色字体标出：

第一步，生成该时刻语义向量：

$$
\begin{aligned}
c_t &= \sum_{i=1}^{T} \alpha_{ti} h_i \\
\alpha_{ti} &= \frac{exp(e_{ti})}{\sum_{k=1}^{T} exp(e_{tk})} \\
s_t &= tanh(W[s_{t-1}, y_{t-1}]) \\
e_{ti} &= s_t^{\top} W_a h_i
\end{aligned}
$$

可以看出区别在计算影响程度 e_{ti} 这个公式，这里我只写出了最优公式，有兴趣可以研读下论文。

第二步，传递隐层信息并预测：

$$
\begin{aligned}
\tilde{s}_t &= tanh(W_c[s_t, c_t]) \\
o_t &= softmax(V\tilde{s}_t)
\end{aligned}
$$

先计算出初始的隐层状态 s_t，再计算注意力层的隐层状态 $\tilde{s}_t$，最后送入softmax层输出预测分布。

总结

Bahdanau Attention与Luong Attention两种注意力机制大体结构一致，区别在于计算影响程度的对齐函数。在计算时刻的影响程度时，前者使用 h_i 和 s_{t-1} 来计算，后者使用 h_i 和 s_t 来计算。从逻辑来看貌似后者更合逻辑，但两种机制现在都有在用，TensorFlow中两者都有对应的函数，效果应该没有很大差别。

References:

[1] Bahdanau et.al (2014) Neural Machine Translation by Jointly Learning to Align and Translate
[2] Luong et.al (2015) Effective Approaches to Attention-based Neural Machine Translation