Lecture 9: Machine Translation and LSTMs and GRUs

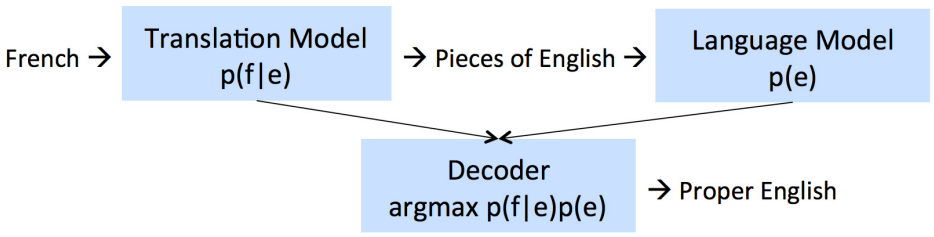
2018 年 7 月 15 日

1 Machine Translation

1.0.1 目标

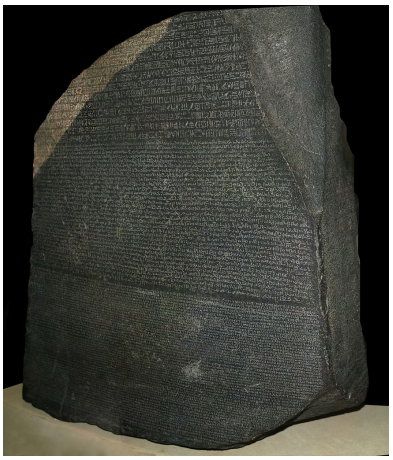
设原文为 f，译文为 e, 目的是找到符合以下条件的 e:

$$e = \underset{e}{argmax}\, P(e|f) = \underset{e}{argmax}\, P(f|e)P(e)$$



1.1 历史

如果说对于简单的分类问题人们还可以制定一系列 rule 解决的话，机器翻译是完全行不通的。现代机器翻译手段全部是基于统计的，在平行语料库上学习训练。历史上第一个平行语料库是罗塞塔石碑。



1.1.1 端到端模型

“系统中不再有独立的声学模型、发音词典、语言模型等模块，而是从输入端（语音波形或特征序列）到输出端（单词或字符序列）直接用一个神经网络相连，让这个神经网络来承担原先所有模块的功能。”

1.1.2 Simplest Model: RNN with encoder and decoder

The last layer should capture all the information of a sentence.

- 为 Encoder 和 Decoder 训练两组不同的权重矩阵
- Decoder 中使用所有之前的 hidden state

$$h_{D,t} = \phi_D(h_{t-1}, h_{E,T}, y_{t-1})$$

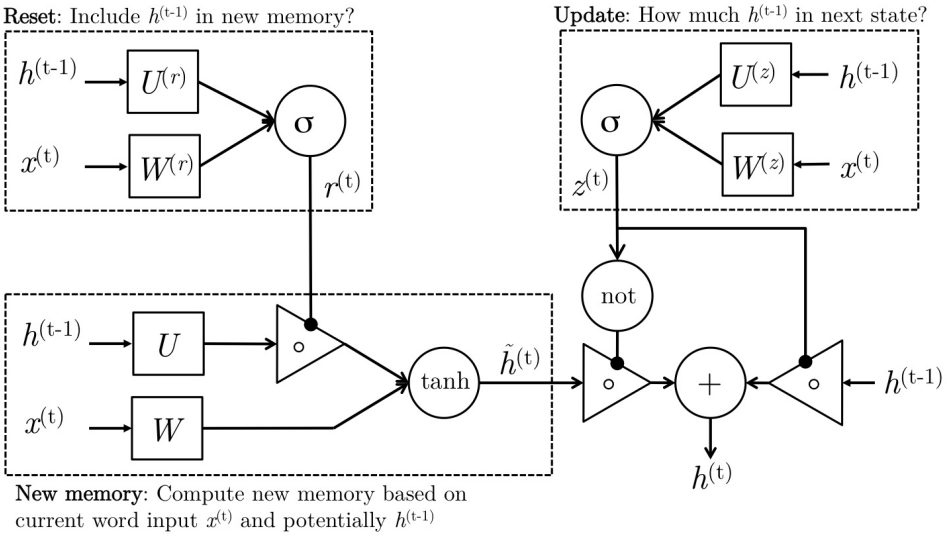
- 训练多层深度 RNN
- 训练双向 Encoder
- Input sequence 调换顺序使得句首单词在 RNN 中距离减小 (减小信息损失)。

  ABC->XYZ: AX = 3

  CBA->XYZ: AX = 1

2 GRU: Gated Recurrent Units

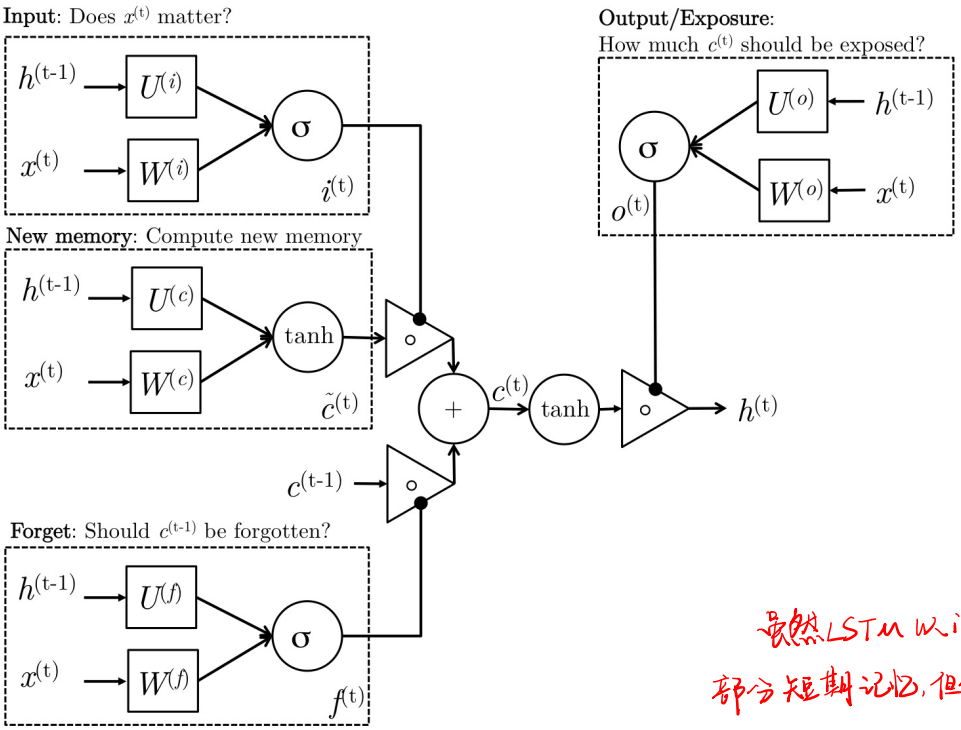
保存更多更长的信息



- Reset gate 控制语境更新。
  - 若 r 接近 0, 忽略之前隐层状态，重新开始.
  - **FLUSH**: Units with short term have active Reset gate
- Update gate 控制语义保存。
  - 若 z 接近 1，相当于缩短 RNN 链，减小了梯度消失。
  - **MAINTAIN**：Units with long term have active Update gate

$$
\begin{aligned}
r_t &= \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \\
z_t &= \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \\
\tilde{h}_t &= tanh(r_t \circ Uh_{t-1} + Wx_t) \\
h_t &= (1 - z_t) \circ \tilde{h}_t + z_t \circ h_{t-1}
\end{aligned}
\tag{1}
$$

3 LSTM: Long-Short-Term-Memories



虽然LSTM以门的方式获得了部分短期记忆，但仍会出现信息减弱、遗忘。所有记忆存在于模型参数中，极大限制了"记忆容量"

一个想法：引入"外部记忆"，将其存储在硬盘中，定期进行索引、更新

- $\mathbf{i_t}$ Input Gate: 当前词是否值得保留
- $\mathbf{f_t}$ Forget Gate：过去记忆 cell 是否有用
- $\mathbf{o_t}$ Output Gate： 当前记忆 cell 有多少需要放到隐状态中
- $\tilde{c}_t$ New Memory Cell： 获得新信息形成的新记忆
- $\mathbf{c_t}$ Final Memory Cell： 最终保留的记忆
- $\mathbf{h_t}$ Hidden Layer： 频繁使用的隐状态，贯穿全文的主线

$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1}) \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1}) \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1}) \\
\tilde{c}_t &= tanh(W^{(c)}x_t + U^{(c)}h_{t-1}) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\
h_t &= o_t \circ tanh(c_t)
\end{aligned}
\tag{2}
$$

LSTM 与 GRU 的思想都是很相似的，即尽量保存有效的历史信息，去除无效的。二者最大的区别在于 LSTM 将记忆单元与隐层分离，而 GRU 将其混合后传播。