Week3

本章介绍了 Bert 之前比较经典的几个模型：R-net、FusionNet 和 QANet。同时介绍了两个前导内容 Multi-Hop 机制和 Pointer network。本笔记聚焦在 FusionNet 模型上。

Multi-Hop 机制：相对 One-Hop 来讲，此机制可简单理解为持续计算 attention，提取信息过程。有两种推进方式：一种是句子 Attention 的 layer 推进，一种是 TimeStep 状态推进，与主序列维度一致。
Pointer network(指针网络)：从输入序列中找到相应的 token 来作为输出，其利用 Attention 作为 Pointer,从输入序列中选择一个位置，并将这个位置所指向的词作为输出。它存储输入序列中各 token 位置信息的概率权重，用概率表示此序列哪个位置比较重要。和注意力(Attention)机制相比，注意力机制只得到一个值。

一、FusionNet

FusionNet 把之前所有模型做了总结，贡献了单词历史和全关注注意力。

1.MRC 网络整体框架

MRC 的整体框架包含了输入向量(浅层融合，如我 word2vec、Glove、CharCNN 等嵌入层)、集成组件(高层融合，RNN、LSTM 输出等融合)和融合过程(自我融合),示例图如下：

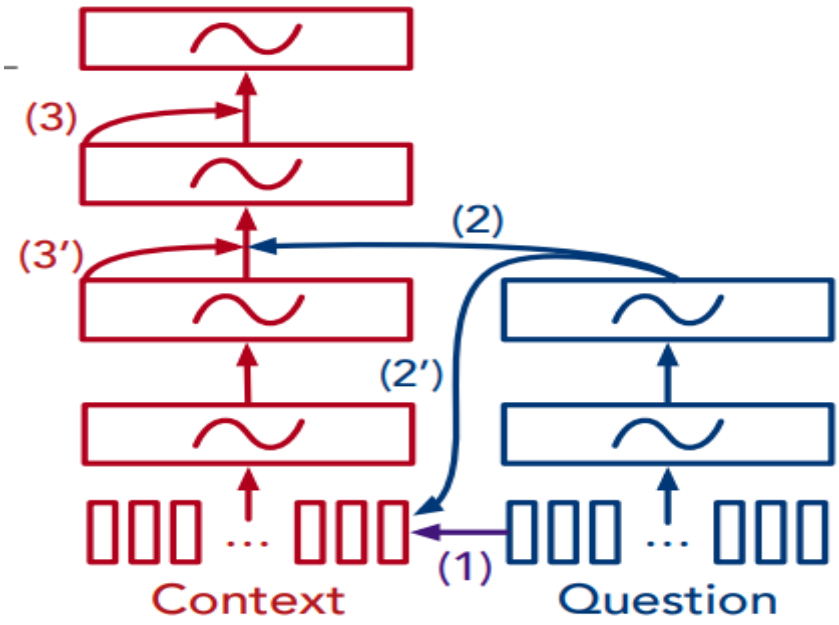

Figure 2: A conceptual architecture illustrating recent advances in MRC.

融合的过程有五个层次：
(1)词级融合：单词是否(在问题中)出现，出现就加 1；

(2)高层融合：更新高层次表示，如把经过 LSTM 后的 question 和 context 进行拼接；
(2’)高层融合(可选择的)：高层次融合到上下文单词，如把 question 高层和底层的 context 层拼接；
(3)自我加强融合：上下文高层次表示本身 self-attention；
(3’)自我加强融合(可选择的)：融合前进行自我融合，如把 2 的结果和经过 LSTM 的变量再进行拼接；

2.单词历史(History-of-word)
一个单词从低层次到高层次所有表示向量一起称作这个词的 history。它将所有层进行拼接，这样可以更好的理解语义，但同时因为维度的大量增加，降低了效率。
举例单词 w：
a. 嵌入层向量：300 维；
b. 通过 LSTM 转换后：300 维；
c. 把 a 和 b 进行拼接：300+300=600 维；
d. 把 c 的结果再输入到下层 LSTM 中：输出 300 维；
e. 把 c 和 d 的输出再进行拼接：600+300=900 维；
f. 把 a、b、e 进行拼接：300+300+900=1500 维；
因维度太高，会进行降维，提高效率。

3.全关注注意力(用单词历史计算注意力)
HoWi：单词的历史向量;
$\{HoW_1^A, HoW_2^A, .. HoW_m^A,\}$,$\{ HoW_1^B, HoW_2^B, .. HoW_n^B\}$：拼接单词历史，把 A 句子 m 个词的单词向量和 B 句子 n 个词的单词向量进行拼接；
$S_{ij} = f(U(HoW_i^A))^T D f(U(HoW_j^B))$,f=ReLu()：每个位置和每个位置计算注意力，通过 U 和 D 进行降维；

4.整体架构
a.编码层
Glove:300 维，CoVe:600 维，POS:12 维，NER:8 维，Other:2 维
文章为以上所有维度拼接：922 维；问题为 300+600=900 维。
b.交互层
单词注意力层：文章到问题注意力，获得文章每个单词注意力，得到 300 维 Glove 向量，和初始到文章 922 维相加，得到 922 维；
阅读层：经过两层 LSTM，生成阅读层向量，获得两层输出，分别为 250 维。得到 context 和 question 第一层 LSTM 输出 h^{c1},h^{q1},和第二层 LSTM 输出 h^{ch},h^{qh} 向量；
问题理解层：把阅读层 question 的第一层和第二层 LSTM 输出 h^{q1}、h^{qh} 进行拼接，输入进一个双层 LSTM 中，得到问题单词向量 u^Q,250 维；
全关注互注意力层：context 每个单词历史和 question 每个单词历史分别为 $HoW_i^c=[glove^c;context^c;h^{c1};h^{ch}]$和 $HoW_i^Q=[glove^Q;context^Q;h^{Q1};h^{Qh}]$，对他们进行底层融合、高层融合和理解层融合，分别得到 $h^{\sim c1}$,$h^{\sim ch}$,$u^{\sim c}$,均为 250 维，再把

这些向量进行拼接，得到 $v^c=[h^{c1};h^c;\ h^{\sim c1};\ h^{\sim ch};\ u^{\sim c}]$，再把此向量输入到 BiLSTM 中；

全关注自注意力层：再进行拼接，得到 $hoW^c = [glove^c;context^c;h^{c1};h^c;\ h^{\sim c1};\ h^{\sim ch};\ u^{\sim c};\ v^c]$,通过对不同位置的 hoW^c 相似度计算，再 softMax 变换，得到 a_{ij},把它和 v^c 进行加权和，得到 $v^{\sim c}$。把 v^c 和 $v^{\sim c}$ 的拼接向量输入到 BiLSTM 中，得到输出 u^c，此为所有对 context 理解的输出，以及 u^Q;

c.输出层

先获得 u^Q 向量含参数的加权和,把 u^Q、u^c 和 W_S 相乘，通过 exp 变换，得到开始位置的概率 P_i^S。再把开始位置和问题向量融合，通过 GRU,得到 v^Q 向量, 把 v^Q、u^c 和 W_E 相乘，通过 exp 变换，最终得到结束位置概率 P_i^E。

二、QA-Net

此模型是 Bert 之前，比较惊艳的一个模型。在不降低准确率的情况下，能提高模型训练速度。采用神经机器翻译模型生成的返译数据来实现数据扩增。