Feedforward Sequential Memory Network: A New Structure to Learn Long-term Dependency

1 背景

该论文的提出主要是为了改进已有的语音识别框架中存在的问题，并提出一种新的框架FSMN，这种框架在"速度"和"准确率"上能达到较理想的效果。

1.1 DNN

众所周知，自2011年微软研究院首次利用深度神经网络(Deep Neural Network, DNN)在大规模语音识别任务上获得显著效果提升以来，DNN在语音识别领域受到越来越多的关注，目前已经成为主流语音识别系统的标配。然而，更深入的研究成果表明，DNN结构虽然具有很强的分类能力，但是其针对上下文时序信息的捕捉能力是较弱的，因此并不适合处理具有长时相关性的时序信号。而语音是一种各帧之间具有很强相关性的复杂时变信号，这种相关性主要体现在说话时的协同发音现象上，往往前后好几个字对我们正要说的字都有影响，也就是语音的各帧之间具有长时相关性。

1.2 RNN

相比前馈型神经网络DNN，循环神经网络(Recurrent Neural Network, RNN)在隐层上增加了一个反馈连接，也就是说，RNN隐层当前时刻的输入有一部分是前一时刻的隐层输出，这使得RNN可以通过循环反馈连接看到前面所有时刻的信息，这赋予了RNN记忆功能。这些特点使得RNN非常适合用于对时序信号的建模，在语音识别领域，RNN是一个近年来替换DNN的新的深度学习框架，

1.3 LSTM

而长短时记忆模块(Long-Short Term Memory, LSTM)的引入解决了传统简单RNN梯度消失等问题，使得RNN框架可以在语音识别领域实用化并获得了超越DNN的效果，目前已经在业界一些比较先进的语音系统中使用。

1.4 BiLSTM与CTC

当前语音识别中的主流RNN声学模型框架，主要还包含两部分：Bidirectional LSTM和CTC(Connectionist Temporal Classification)输出层。

- Bidirectional LSTM对当前语音帧进行判断时，不仅可以利用历史的语音信息，还可以利用未来的语音信息，可以进行更加准确的决策
- CTC使得训练过程无需帧级别的标注，实现有效的"端对端"训练。

1.5 FSMN

以上技术在实际应用中存在一些缺点：

- 速度慢：传统的双向RNN方案，理论上需要看到语音的结束（即所有的未来信息），才能成功的应用未来信息来获得提升，因此只适合处理离线任务，而对于要求即时响应的在线任务（例如语音输入法）则往往会带来3-5s的硬延迟，这对于在线任务是不可接受的
- 过拟合：RNN对上下文相关性的拟合较强，相对于DNN更容易陷入过拟合的问题，容易因为训练数据的局部不鲁棒现象而带来额外的异常识别错误
- 训练难：由于RNN具有比DNN更加复杂的结构，给海量数据下的RNN模型训练带来了更大的挑战

鉴于上述问题，科大讯飞发明了一种名为前馈型序列记忆网络FSMN(Feed-forward Sequential Memory Network)的新框架。在这个框架中，可以把上述几点很好的融合，同时各个技术点对效果的提升可以获得叠加，FSMN采用非循环的前馈结构，在只需要180ms延迟下，就达到了和双向LSTM RNN相当的效果。

2 原理

FSMN是一种典型的前馈型网络，大多前馈型网络一般用作分类问题，不能较好地处理序列问题，与一般的前馈型网络不一样的是，FSMN在隐层中添加Memory Block，它同样可以捕获上下文信息(需要设置一定长度的前后窗口)，此外与RNN或LSTM等模型相比，FSMN因为它的Non-recurrent结构处理要更快和更稳定。下面主要介绍一下FSMN中的核心模块Memory Block。

2.1 模型架构

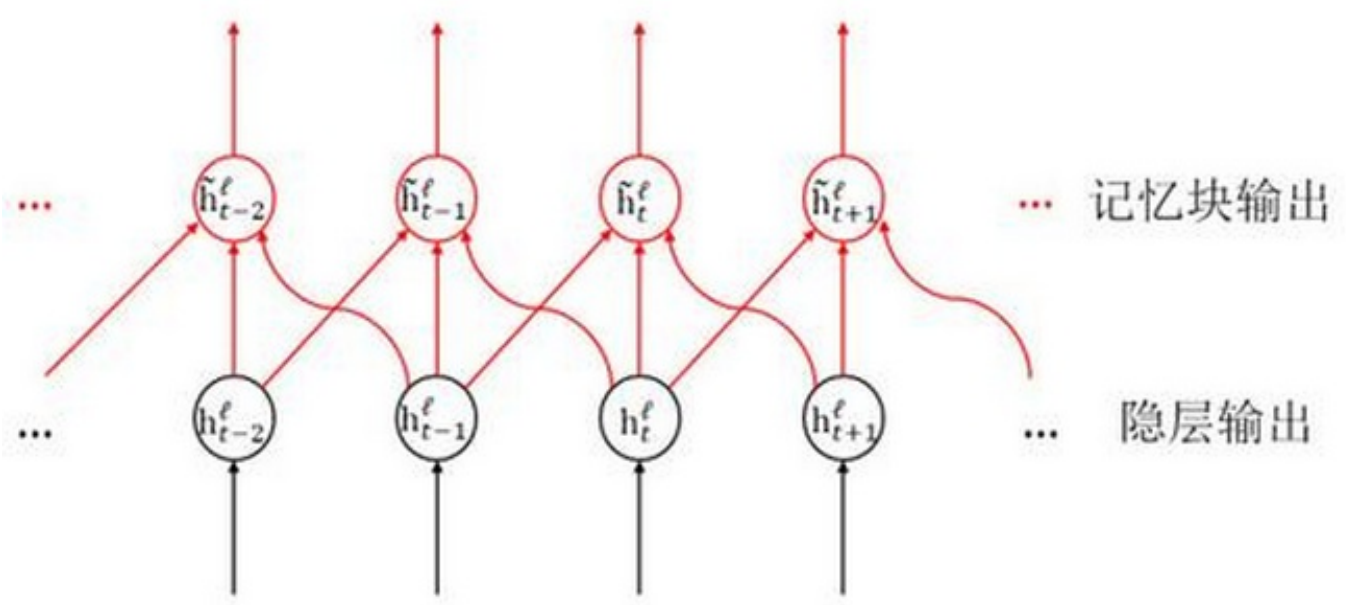

图1

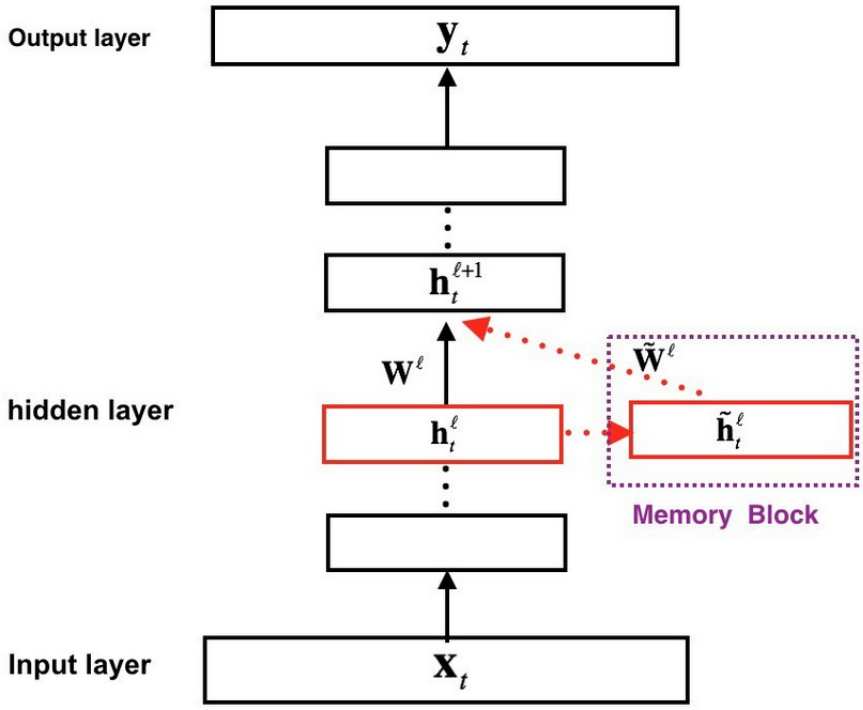

图2

图1画出了双向FSMN中记忆块左右各记忆1帧语音信息（在实际任务中，可根据任务需要，人工调整所需记忆的历史和未来信息长度）的时序展开结构。从图中我们可以看出，不同于传统的基于循环反馈的RNN，FSMN记忆块的记忆功能是使用前馈结构实现的。图2即为FSMN的结构示意图，相比传统的DNN，我们在隐层旁增加了一个称为"记忆块"的模块，用于存储对判断当前语音帧有用的历史信息和未来信息。下面是相关的数学公式：

- h_t^l代表t时刻第l层的隐层
- $\tilde{h}_t^l$代表t时刻第l层的Memory Block
- $\tilde{h}_t^l = \sum_{i=0}^{N_1} a_i^l * h_{t-i}^l + \sum_{j=1}^{N_2} c_j^l * h_{t+j}^l$，其中N1代表左边N1个窗口（包括当前位置），N2代表右边N2个窗口（不包括当前位置），a_i^l可以是标量或向量（对应两种不同的版本），含义是前面第i个隐层的权重，这些权重决定了不同时刻输入对判断当前语音帧的影响，同理c_j^l，最后加权求和
- $h_t^{l+1} = f(W^l h_t^l + \tilde{W}^l \tilde{h}_t^l + b^l)$，此时$h_t^{l+1}$具备了当前和上下文信息，该向量可以作为输出层的输入

2.2 优点

这种前馈结构有以下几个优点：

- 速度快：双向FSMN对未来信息进行记忆时，没有传统双向RNN必须等待语音输入结束才能对当前语音帧进行判断的限制，它只需要等待有限长度的未来语音帧即可
- 梯度问题：如前所述，传统的简单RNN因为训练过程中的梯度是按时间逐次往前传播的，因此会出现指数衰减的梯度消失现象，这导致理论上具有无限长记忆的RNN实际上能记住的信息很有限，然而FSMN这种基于前馈时序展开结构的记忆网络，在训练过程中梯度沿着图2中记忆块与隐层的连接权重往回传给各个时刻即可（只存在t_i时刻与t_j时刻的反向传播，不存在t_i时刻与t_{i-1}时刻的反向传播），而且这种梯度传播在任何时刻的衰减都是常数的，也

是可训练的，因此FSMN用一种更为简单的方式解决了RNN中的梯度消失问题，使得其具有类似LSTM的长时记忆能力

- 训练灵活：由于FSMN完全基于前馈神经网络，所以不存在RNN训练中因mini-batch中句子长短不一需要补0而导致浪费运算的情况，前馈结构也使得它的并行度更高，可最大化利用GPU计算能力。从最终训练收敛的双向FSMN模型记忆块中各时刻的加权系数分布我们观察到，权重值基本上在当前时刻最大，往左右两边逐渐衰减，这也符合预期
- 进一步，FSMN可和CTC准则结合，实现语音识别中的“端到端”建模

2.3 缺点

- 窗口长度：窗口长度需要人为设定，当窗口太小时，无法捕获到更长远的信息

3 小结

LSTM常用于自然语言处理，在文本数据中，开头可能有个主语，在末尾可能有个指代关系，模型需要长期的记忆。但在语音识别领域，识别当前语音帧的音素不会太依赖很久以前的信息，比较依赖局部上下文的信息。个人感觉FSMN是一个性价比较高的模型，当选用合适长度的窗口后，会牺牲窗口长度外的信息，但在速度和性能上能够达到较好的折中效果。

参考资料：

[1] Feedforward sequential memory networks: A new structure to learn long-term dependency
[2] http://science.china.com.cn/2015-12/31/content_8489460.htm