阿里提出DFSMN语音识别声学模型

星期五, 三月 16, 2018　　3:12 下午

新智元专栏

团队：阿里巴巴语音交互智能团队

作者：张仕良，雷鸣，鄢志杰， 戴礼荣

会议：ICASSP-2018

【新智元导读】在语音顶会ICASSP，阿里巴巴语音交互智能团队的poster论文提出一种改进的前馈序列记忆神经网络结构，称之为深层前馈序列记忆神经网络（DFSMN）。研究人员进一步将深层前馈序列记忆神经网络和低帧率（LFR）技术相结合，构建LFR-DFSMN语音识别声学模型。

在语音顶会ICASSP，阿里巴巴语音交互智能团队的poster论文提出一种改进的前馈序列记忆神经网络结构，称之为深层前馈序列记忆神经网络（DFSMN）。研究人员进一步将深层前馈序列记忆神经网络和低帧率（LFR）技术相结合，构建LFR-DFSMN语音识别声学模型。

该模型在大词汇量的英文识别和中文识别任务上都可以取得相比于目前最流行的基于长短时记忆单元的双向循环神经网络（BLSTM）的识别系统显著的性能提升。而且LFR-DFSMN在训练速度，模型参数量，解码速度，而且模型的延时上相比于BLSTM都具有明显的优势。

研究背景

近年来, 深度神经网络成为了大词汇量连续语音识别系统中的主流声学模型。由于语音信号具有很强的长时相关性，因而目前普遍流行的是使用具有长时相关建模的能力的循环神经网络（RNN），例如LSTM以及其变形结构。循环神经网络虽然具有很强的建模能力，但是其训练通常采用BPTT算法，存在训练速度缓慢和梯度消失问题。我们之前的工作，提出了一种新颖的非递归的网络结构，称之为前馈序列记忆神经网络（feedforward sequential memory networks， FSMN），可以有效的对信号中的长时相关性进行建模。相比于循环神经网络，FSMN训练更加高效，而且可以获得更好的性能。

本论文，我们在之前FSMN的相关工作的基础上进一步提出了一种改进的FSMN结构，称之为深层的前馈序列记忆神经网络（Deep-FSMN, DFSMN）。我们通过在FSMN相邻的记忆模块之间添加跳转连接（skip connections），保证网络高层梯度可以很好的传递给低层，从而使得训练很深的网络不会面临梯度消失的问题。进一步的，考虑到将DFSMN应用于实际的语音识别建模任务不仅需要考虑模型的性能，而且需要考虑到模型的计算量以及实时性。针对这个问题，我们提出将DFSMN和低帧率（lower frame rate，LFR）相结合用于加速模型的训练和测试。同时我们设计了DFSMN的结构，通过调整DFSMN的记忆模块的阶数实现时延的控制，使得基于LFR-DFSMN的声学模型可以被应用到实时的语音识别系统中。

我们在多个大词汇量连续语音识别任务包括英文和中文上验证了DFSMN的性能。在目前流行的2千小时英文FSH任务上，我们的DFSMN相比于目前主流的BLSTM可以获得绝对1.5%而且模型参数量更少。在2万小时的中文数据库上，LFR-DFSMN相比于LFR-LCBLSTM可以获得超过20%的相对性能提升。而且LFR-DFSMN可以灵活的控制时延，我们发现将时延控制到5帧语音依旧可以获得相比于40帧时延的LFR-LCBLSTM更好的性能。

FSMN回顾

最早提出的FSMN的模型结构如图1（a）所示，其本质上是一个前馈全连接神经网络，通过在隐层旁添加一些记忆模块（memory block）来对周边的上下文信息进行建模，从而使得模型可以对时序信号的长时相关性进行建模。FSMN的提出是受到数字信号处理中滤波器设计理论的启发：任何无限响应冲击（Infinite Impulse Response, IIR）滤波器可以采用高阶的有限冲击响应（Finite Impulse Response, FIR）滤波器进行近似。从滤波器的角度出发，如图

1（c）所示的RNN模型的循环层就可以看作如图1（d）的一阶IIR滤波器。而FSMN采用的采用如图1（b）所示的记忆模块可以看作是一个高阶的FIR滤波器。从而FSMN也可以像RNN一样有效的对信号的长时相关性进行建模，同时由于FIR滤波器相比于IIR滤波器更加稳定，因而FSMN相比于RNN训练上会更加简单和稳定。

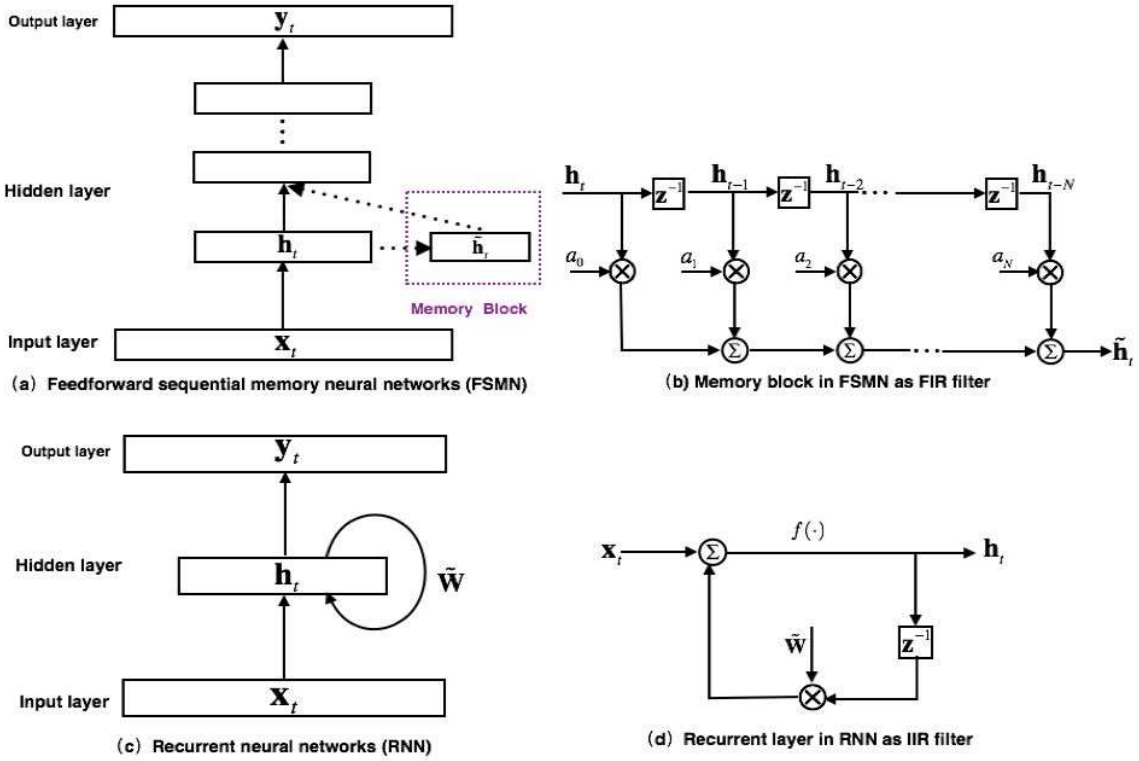

图 1. FSMN模型结构以及和RNN的对比

根据记忆模块编码系数的选择，可以分为：1）标量FSMN（sFSMN）；2）矢量FSMN（vFSMN）。sFSMN 和 vFSMN 顾名思义就是分别使用标量和矢量作为记忆模块的编码系数。sFSMN和vFSMN记忆模块的表达分别如下公式：

$$\tilde{\mathbf{h}}_t^l = \sum_{i=0}^{N} a_i^l \bullet \mathbf{h}_{t-i}^l$$

$$\tilde{\mathbf{h}}_t^l = \sum_{i=0}^{N} \boldsymbol{a}_i^l \bigodot \mathbf{h}_{t-i}^l$$

以上的FSMN只考虑了历史信息对当前时刻的影响，我们可以称之为单向的FSMN。当我们同时考虑历史信息以及未来信息对当前时刻的影响时，我们可以将单向的FSMN进行扩展得到双向的FSMN。双向的sFSMN和vFSMN记忆模块的编码公式如下：

$$\tilde{\mathbf{h}}_t^l = \sum_{i=0}^{N_1} \boldsymbol{a}_i^l \odot \mathbf{h}_{t-i}^l + \sum_{j=1}^{N_2} \boldsymbol{c}_j^l \odot \mathbf{h}_{t+j}^l$$

这里

N_1

和

N_2

分别代表回看(look-back)的阶数和向前看(look-ahead)的阶数。我们可以通过增大阶数，也可以通过在多个隐层添加记忆模块来增强FSMN对长时相关性的建模能力。

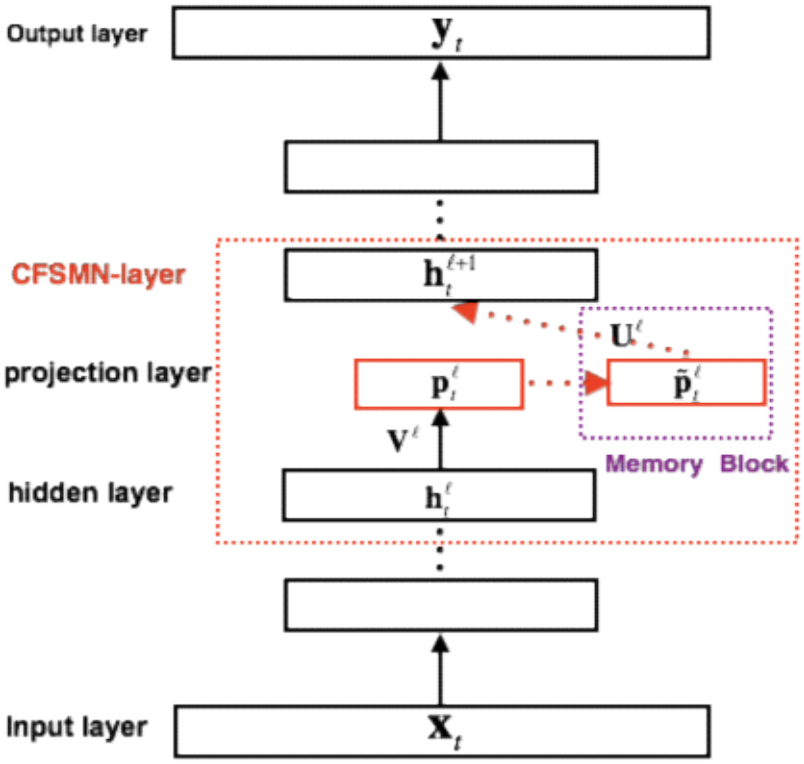

图 2. cFSMN结构框图

FSMN相比于FNN，需要将记忆模块的输出作为下一个隐层的额外输入，这样就会引入额外的模型参数。隐层包含的节点越多，则引入的参数越多。我们通过结合矩阵低秩分解（Low-rank matrix factorization）的思路，提出了一种改进的FSMN结构，称之为简洁的FSMN（Compact FSMN，cFSMN）。如图2是一个第l个隐层包含记忆模块的cFSMN的结构框图。

对于cFSMN，通过在网络的隐层后添加一个低维度的线性投影层，并且将记忆模块添加在这些线性投影层上。进一步的，cFSMN对记忆模块的编码公式进行了一些改变，通过将当前时刻的输出显式的添加到记忆模块的表达中，从而只需要将记忆模块的表达作为下一层的输入。这样可以有效的减少模型的参数量，加快网络的训练。具体的，单向和双向的cFSMN记忆模块的公式表达分别如下：

$$\tilde{\mathbf{P}}_t^l = \mathbf{P}_t^l + \sum_{i=0}^{N} \boldsymbol{a}_i^l \bigodot \mathbf{P}_{t-i}^l$$

$$\tilde{\mathbf{P}}_t^l = \mathbf{P}_t^l + \sum_{i=0}^{N_1} \boldsymbol{a}_i^l \bigodot \mathbf{P}_{t-i}^l + \sum_{j=0}^{N_2} \boldsymbol{c}_j^l \bigodot \mathbf{P}_{t+j}^l$$

DFSMN介绍

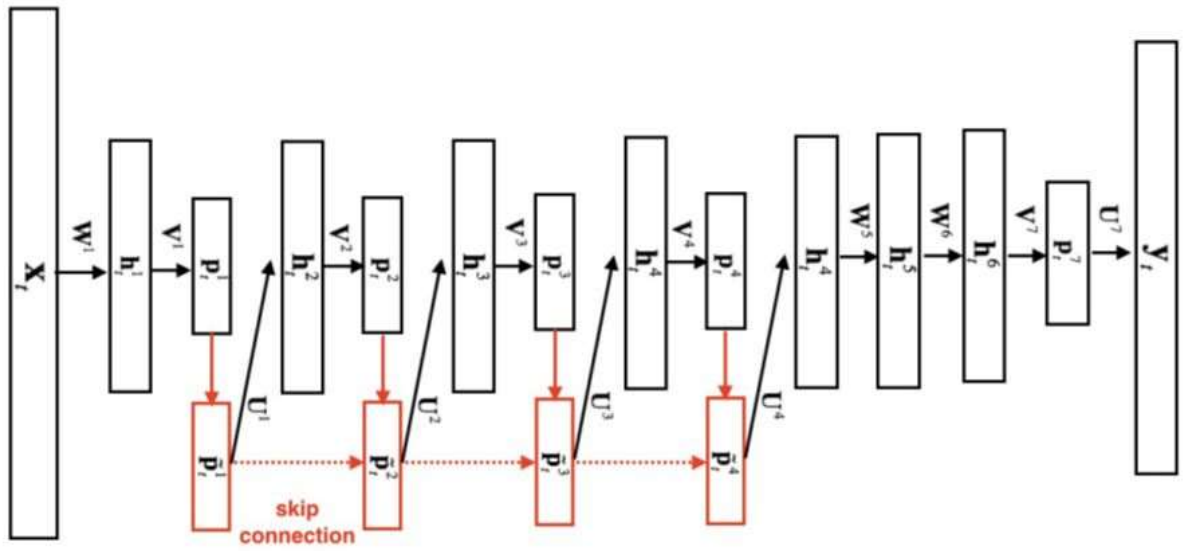

图 3. Deep-FSMN (DFSMN)模型结构框图

如图3是我们进一步提出的Deep-FSMN（DFSMN）的网络结构框图，其中左边第一个方框代表输入层，右边最后一个方框代表输出层。我们通过在cFSMN的记忆模块（红色框框表示）之间添加跳转连接（skip connection），从而使得低层记忆模块的输出会被直接累加到高层记忆模块里。这样在训练过程中，高层记忆模块的梯度会直接赋值给低层的记忆模块，从而可以克服由于网络的深度造成的梯度消失问题，使得可以稳定的训练深层的网络。我们对记忆模块的表达也进行了一些修改，通过借鉴扩张（dilation）卷积[3]的思路，在记忆模块中引入一些步幅（stride）因子，具体的计算公式如下：

$$\tilde{\boldsymbol{P}}_t^l = \tilde{\boldsymbol{P}}_t^{l-1} + \boldsymbol{P}_t^l + \sum_{i=0}^{N_1} \boldsymbol{a}_i^l \bigodot \boldsymbol{P}_{t-s_1*i}^l + \sum_{j=1}^{N_2} \boldsymbol{c}_i^l \bigodot \boldsymbol{P}_{t+s_2*j}^l$$

其中

$\tilde{\boldsymbol{P}}_t^{l-1}$

表示第

$l-1$

层记忆模块第t个时刻的输出。

s_1

和

s_2

分别表示历史和未来时刻的编码步幅因子，例如 则表示对历史信息进行编码时每隔一个时刻取一个值作为输入。这样在相同的阶数的情况下可以看到更远的历史，从而可以更加有效的对长时相关性进行建模。对于实时的语音识别系统我们可以通过灵活的设置未来阶数来控制模型的时延，在极端情况下，当我们将每个记忆模块的未来阶数都设置为0，则我们可以实现无时延的一个声学模型。对于一些任务，我们可以忍受一定的时延，我们可以设置小一些的未来阶数。

LFR-DFSMN声学模型

目前的声学模型，输入的是每帧语音信号提取的声学特征，每帧语音的时长通常为10ms，对于每个输入的语音帧信号会有相对应的一个输出目标。最近有研究提出一种低帧率（Low Frame Rate，LFR）建模方案：通过将相邻时刻的语音帧进行绑定作为输入，去预测这些语音帧的目标输出得到的一个平均输出目标。具体实验中可以实现三帧（或更多帧）拼接而不损失模型的性能。从而可以将输入和输出减少到原来的三分之一甚至更多，可以极大的提升语音识别系统服务时声学得分的计算以及解码的效率。我们结合LFR和以上提出的DFSMN，构建了如图4的基于LFR-DFSMN的语音识别声学模型，经过多组实验我们最终确定了采用一个包含10层DFSMN层+2层DNN的DFSMN作为声学模型，输入输出则采用LFR，将帧率降低到原来的三分之一。

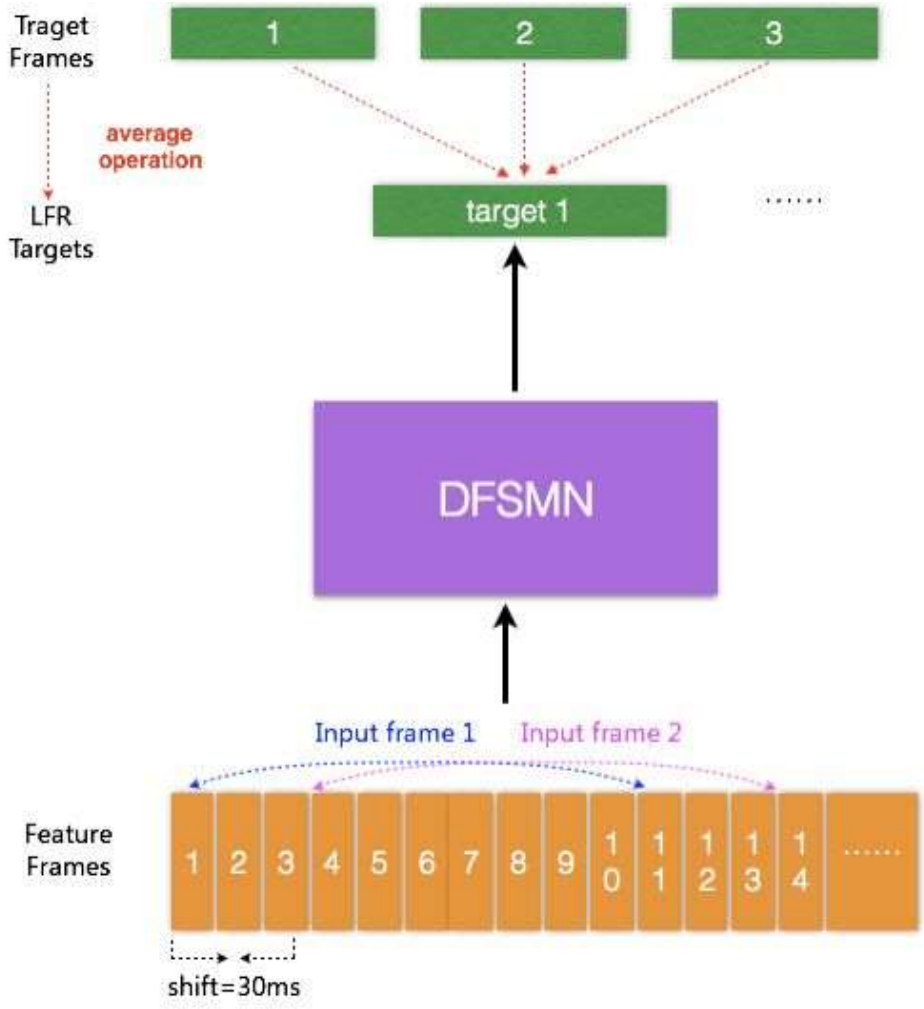

图 4. LFR-DFSMN声学模型结构框图

实验结果

1. **英文识别**

我们在2千小时的英文FSH任务上验证所提出的DFSMN模型。我们首先验证了DFSMN的网络深度对性能的影响，我们分别验证了DFSMN包含6，8，10，12个DFSMN层的情况。最终模型的识别性能如下表。通过增加网络的深度我们可以获得一个明显的性能提升。

ID	Model	stride	Size(MB)	WER (%)
exp1	DFSMN(6)	1	104	10.7
exp2	DFSMN(6)	2	104	10.3
exp3	DFSMN(8)	2	120	9.6
exp4	DFSMN(10)	2	136	9.5
exp5	DFSMN(12)	2	152	9.4

我们也和一些主流的声学模型进行了对比，结果如下表。从结果看DFSMN相比于目前最流行的BLSTM不仅参数量更少，而且性能上可以获得1.5%的绝对性能提升。

Model	Size (MB)	WER (%)
DNN	159	14.3
BLSTM	180	**10.9**
BLSTM(6)[27]	166*	**10.3**
cFSMN	104	10.8
DFSMN(12)	152	**9.4**

2. 中文识别

关于中文识别任务，我们首先在5000小时任务上进行实验。我们分别验证了采用绑定的音素状态（CD-State）和绑定的音素（CD-Phone）作为输出层建模单元。关于声学模型我们对比了时延可控的BLSTM（LCBLSTM），cFSMN以及DFSMN。对于LFR模型，我们采用CD-Phone作为建模单元。详细的实验结果如下表：

Model	Target	Size (MB)	CER %	Gain
LCBLSTM	CD-State	196	18.78	-
cFSMN(6)		102	17.72	+5.32%
LFR-LCBLSTM	CD-Phone	**220**	**18.92**	-
LFR-cFSMN(6)		108	16.85	+11.00%
LFR-cFSMN(8)	CD-Phone	124	15.80	+16.50%
LFR-cFSMN(10)		140	15.91	+15.86%
LFR-DFSMN(8)		124	15.45	+18.34%
LFR-DFSMN(10)	CD-Phone	**140**	**15.00**	**+20.72%**

对于基线LCBSLTM，采用LFR相比于传统的单帧预测在性能上相近，优点在效率可以提升3倍。而采用LFR的cFSMN，相比于传统的单帧预测不仅在效率上可以获得相应提升，而且可以获得更好的性能。这主要是LFR一定程度上破坏了输入信号的时序性，而BLSTM的记忆机制对时序性更加的敏感。进一步的我们探索了网络深度对性能的影响，对于之前的cFSMN网络，当把网络深度加深到10层，会出现一定的性能下降。而对于我们最新提出来的DFSMN，10层的网络相比于8层依旧可以获得性能提升。最终相比于基线的LFR-LCBLSTM模型，我们可以获得超过20%的相对性能提升。

下表我们对比了LFR-DFSMN和LFR-LCBLSTM的训练时间，以及解码的实时因子（RTF）。从结果上看我们可以将训练速度提升3倍，同时可以将实时因子降低到原来的接近三分之一。

Model	Training time (hr/epoch)	RTF
LFR-LCBLSTM	21.62	0.4289
LFR-DFSMN(8)	6.85	0.1486

对于语音识别系统，另外一个需要考虑的因素是模型的延迟问题。原始的BLSTM需要等接收整句话后才能得到输出用于解码。LCBLSTM是目前的一种改进结构，可以将解码的时延进行控制，目前采用的LFR-LCBLSTM的时延帧数是40帧。对于DFSMN，时延的帧数可以功过设计记忆模块的滤波器阶数进行灵活控制。最终当只有5帧延时时，LFR-DFSMN相比于LFR-LCBLSTM依然可以获得更好的性能。

Model	N_2	Delay Frame	CER%	Gain
LFR-LCBLSTM	-	40	16.05	-
LFR-DFSMN(10)	2	20	12.67	+21.06%
	1	10	12.94	+19.38%
	1 and 0	5	13.38	+16.64%

https://arxiv.org/abs/1803.05030