阿里技术丛书系列

Alibaba Group 阿里巴巴集团

阿里机器智能

语音与信号处理技术精选专辑

顶级会议 ICASSP-2018 收录论文

Alibaba Group
阿里巴巴集团

阿里技术

扫一扫二维码图案，关注我吧

「阿里技术」微信公众号

「阿里巴巴机器智能」微信公众号

序言

ICASSP (International Conference on Acoustics, Speech, and Signal Processing)是由 IEEE 信号处理协会(IEEE Signal Processing Society)组织的语音研究领域的顶级会议之一，和 INTERSPEECH(Annual Conference of the International Speech Communication Association)并称为国际语音领域最著名、影响力最大的两个学术会议。相对于 INTERSPEECH 主要侧重语音方面的研究和应用，ICASSP 会议更加侧重声学、语音信号以及语音建模相关的学术讨论，包含了语音技术相关的各个方面，堪称国际语音行业的一个年度盛会。在 ICASSP2018 中，阿里一共发表了 5 篇论文，分别涵盖语音识别、语音合成以及情感识别三个方向。

在论文《基于深层前馈序列记忆网络 ,如何将语音合成速度提升四倍？》中，作者提出了一种基于深度前馈序列记忆网络的语音合成系统 ,该系统在达到与基于双向长短时记忆单元的语音合成系统一致的主观听感的同时 ,模型大小只有后者的四分之一，且合成速度是后者的四倍，非常适合于对内存占用和计算效率非常敏感的端上产品环境。

在论文《为了更精确的情感识别，A-LSTM 出现了》中，作者针对 LSTM 时间依赖局限性问题 ,提出了高级长短期记忆网络(advanced LSTM (A-LSTM)) 模型，利用线性组合，将若干时间点的本层状态都结合起来，以打破传统 LSTM 的这种局限性。在这篇文章中，我们将 A-LSTM 应用于情感识别中。实验结果显示 ,与应用传统 LSTM 的系统相比 ,应用了 A-LSTM 的系统能相对提高 5.5% 的识别率。

在论文《为了让机器听懂“长篇大论”，阿里工程师构建了新模型》中，作者提出了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），进一步地将深层前馈序列记忆神经网络和低帧率（LFR）技术相结合构建了 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。

在论文《示范了 200 句后，我的声音“双胞胎”诞生了⁉》中，作者提出了基于线性网络的语音合成说话人自适应算法，该算法对每个说话人学习特定的线性网络，从而获得属于目标说话人的声学模型，通过该算法，使用 200 句目标说话人的自适应语料训练的说话人自适应系统能够获得和使用 1000 句训练的说话人相关系统相近的合成效果。

在论文《朋友，我能分享你的喜怒吗？阿里语音情感识别框架揭秘》中，作者提出了一套包含多个子系统的复合情感识别框架。这一框架会深入挖掘输入语音中与情感相关的各个方面的信息，从而提高系统的顽健性。

每年 INTERSPEECH 或者 ICASSP 都是语音学术界和工业界的一次盛会，从 Deep Learning 在 2010 年左右引入语音领域，到现在几乎所有的论文都直接或者间接以神经网络模型进行尝试，语音技术在最近几年发生了翻天覆地的变化。

近几年贴近实际产品的论文越来越多，语音领域的各大研究机构和知名公司纷纷做出了更实际、更靠谱的工作，相关产品问题也随之暴露和慢慢地被解决，整个语音技术已经逐渐走到了实际应用的阶段，近几年越来越多的语音设备产品

的问世和火爆也说明了这一点。我们将 ICASSP2018 会议上收录的论文编辑成册，希望通过这个方式，更多的和学术界、工业界同行共同探讨、共同进步，衷心的希望语音技术继续百家争鸣、百花齐放，早日把靠谱的语音交互能力带到各行各业、带进千家万户，真正地帮助到人们的工作和生活！

阿里巴巴高级算法专家 雷鸣

2019 年 3 月 于北京

目录

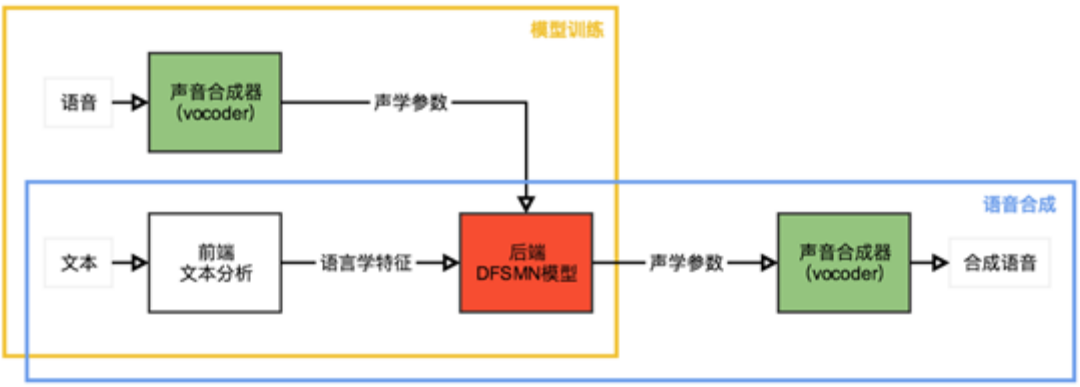
基于深度前馈序列记忆网络，如何将语音合成速度提升四倍？

作者：毕梦霄/Mengxiao Bi，卢恒/Heng Lu，张仕良/Shiliang Zhang，雷鸣/Ming Lei，鄢志杰/Zhijie Yan

小叽导读：我们提出了一种基于深度前馈序列记忆网络的语音合成系统。该系统在达到与基于双向长短时记忆单元的语音合成系统一致的主观听感的同时，模型大小只有后者的四分之一，且合成速度是后者的四倍，非常适合于对内存占用和计算效率非常敏感的端上产品环境。

研究背景

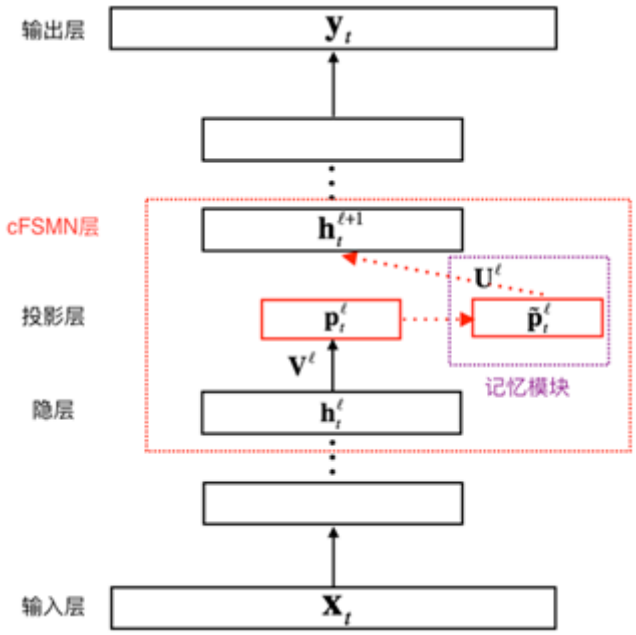
语音合成系统主要分为两类，拼接合成系统和参数合成系统。其中参数合成系统在引入了神经网络作为模型之后，合成质量和自然度都获得了长足的进步。另一方面，物联网设备（例如智能音箱和智能电视）的大量普及也对在设备上部署的参数合成系统提出了计算资源的限制和实时率的要求。本工作引入的深度前馈序列记忆网络可以在保持合成质量的同时，有效降低计算量，提高合成速度。



我们使用基于双向长短时记忆单元（BLSTM）的统计参数语音合成系统作为基线系统。与目前主流的统计参数语音合成系统相似，我们提出的基于深度前馈序列记忆网络（DFSMN）的统计参数语音合成系统也是由 3 个主要部分组成，声音合成器（vocoder），前端模块和后端模块，如上图所示。我们使用开源工具 WORLD 作为我们的声音合成器，用来在模型训练时从原始语音波形中提取频谱信息、基频的对数、频带周期特征（BAP）和清浊音标记，也用来在语音合成时完成从声学参数到实际声音的转换。前端模块用来对输入的文本进行正则化和词法分析，我们把这些语言学特征编码后作为神经网络训练的输入。后端模块用来建立从输入的语言学特征到声学参数的映射，在我们的系统中，我们使用 DFSMN 作为后端模块。

深度前馈序列记忆网络

紧凑前馈序列记忆网络（cFSMN）作为标准的前馈序列记忆网络（FSMN）的改进版本，在网络结构中引入了低秩矩阵分解，这种改进简化了 FSMN，减少了模型的参数量，并加速了模型的训练和预测过程。



上图给出了 cFSMN 的结构的图示。对于神经网络的每一个 cFSMN 层，计算过程可表示成以下步骤①经过一个线性映射，把上一层的输出映射到一个低维向量②记忆模块执行计算，计算当前帧之前和之后的若干帧和当前帧的低维向量的逐维加权和③把该加权和再经过一个仿射变换和一个非线性函数，得到当前层的输出。三个步骤可依次表示成如下公式。

$$\mathbf{p}_t^l = \mathbf{V}^l \mathbf{h}_t^l + \mathbf{b}^l$$

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

$$\mathbf{h}_t^{l+1} = f\left(\mathbf{U}^l \tilde{\mathbf{p}}_t^l + \mathbf{d}^l\right)$$

与循环神经网络（RNNs，包括 BLSTM）类似，通过调整记忆模块的阶数，cFSMN 有能力捕捉序列的长程信息。另一方面，cFSMN 可以直接通过反向传播算法（BP）进行训练，与必须使用沿时间反向传播算法（BPTT）进行训练的 RNNs 相比，训练 cFSMN 速度更快，且较不容易受到梯度消失的影响。

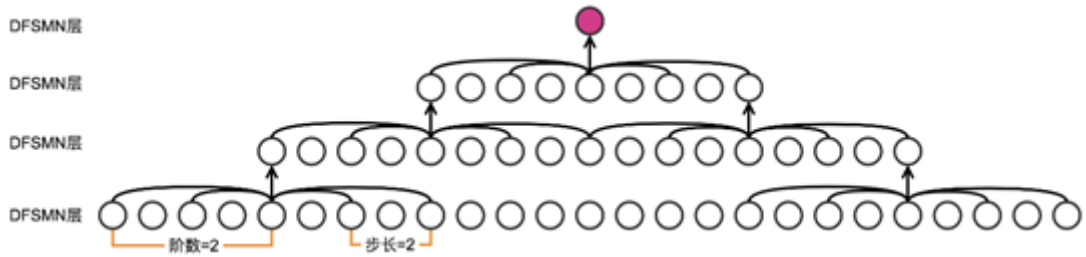
对 cFSMN 进一步改进，我们得到了深度前馈序列记忆网络（DFSMN）。DFSMN 利用了在各类深度神经网络中被广泛使用的跳跃连接（skip-connections）技术，使得执行反向传播算法的时候，梯度可以绕过非线性变换，即使堆叠了更多 DFSMN 层，网络也能快速且正确地收敛。对于 DFSMN 模型，增加深度的好处有两

个方面。一方面，更深的网络一般来说具有更强的表征能力，另一方面，增加深度可以间接地增大 DFSMN 模型预测当前帧的输出时可以利用的上下文长度，这在直观上非常有利于捕捉序列的长程信息。具体来说，我们把跳跃连接添加到了相邻两层的记忆模块之间，如下面公式所示。由于 DFSMN 各层的记忆模块的维数相同，跳跃连接可由恒等变换实现。

$$\tilde{\mathbf{p}}_t^l = \mathcal{H}(\tilde{\mathbf{p}}_t^{l-1}) + \mathbf{p}_t^l + \sum_{i=0}^{N_1} \mathbf{a}_i^l \odot \mathbf{p}_{t-s_1*i}^l + \sum_{j=1}^{N_2} \mathbf{c}_j^l \odot \mathbf{p}_{t+s_2*j}^l$$

我们可以认为 DFSMN 是一种非常灵活的模型。当输入序列很短，或者对预测延时要求较高的时候，可以使用较小的记忆模块阶数，在这种情况下只有当前帧附近帧的信息被用来预测当前帧的输出。而如果输入序列很长，或者在预测延时不是那么重要的场景中，可以使用较大的记忆模块阶数，那么序列的长程信息就能被有效利用和建模，从而有利于提高模型的性能。

除了阶数之外，我们为 DFSMN 的记忆模块增加了另一个超参数，步长（stride），用来表示记忆模块提取过去或未来帧的信息时，跳过多少相邻的帧。这是有依据的，因为与语音识别任务相比，语音合成任务相邻帧之间的重合部分甚至更多。



上文已经提到，除了直接增加各层的记忆模块的阶数之外，增加模型的深度也能间接增加预测当前帧的输出时模型可以利用的上下文的长度，上图给出了一个例子。

实验

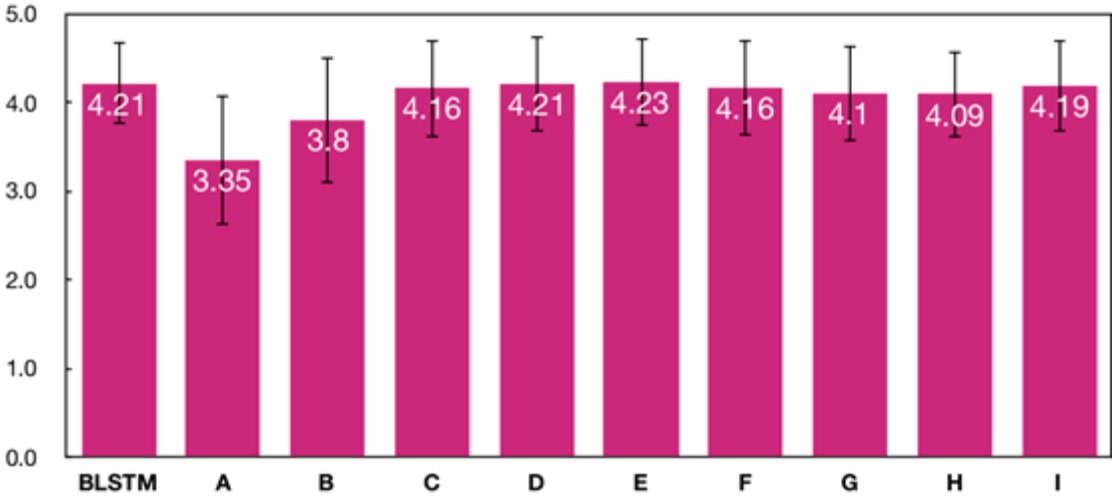
在实验阶段，我们使用的是一个由男性朗读的中文小说数据集。我们把数据集划分成两部分，其中训练集包括 38600 句朗读（大约为 83 小时），验证集包括 1400 句朗读（大约为 3 小时）。所有的语音数据采样率都为 16k 赫兹，每帧

帧长为 25 毫秒，帧移为 5 毫秒。我们使用 WORLD 声音合成器逐帧提取声学参数，包括 60 维梅尔倒谱系数，3 维基频的对数，11 维 BAP 特征以及 1 维清浊音标记。我们使用上述四组特征作为神经网络训练的四个目标，进行多目标训练。前端模块提取出的语言学特征，共计 754 维，作为神经网络训练的输入。

我们对比的基线系统是基于一个强大的 BLSTM 模型，该模型由底层的 1 个全连接层和上层的 3 个 BLSTM 层组成，其中全连接层包含 2048 个单元，BLSTM 层包含 2048 个记忆单元。该模型通过沿时间反向传播算法（BPTT）训练，而我们的 DFSMN 模型通过标准的反向传播算法（BP）训练。包括基线系统在内，我们的模型均通过逐块模型更新过滤算法（BMUF）在 2 块 GPU 上训练。我们使用多目标帧级别均方误差（MSE）作为训练目标。

ID	模型	深度	阶数	客观指标					模型大小 (MB)	模型计算量 (G)
				MCD	F0 RMSE	BAPD	U/V Error	MSE		
BLSTM	BLSTM	-	-	6.92	29.09	2.93	0.1008	0.0273	295	21.09
A	DFSMN	3+2	1,1,1,1	7.43	33.41	3.09	0.1074	0.0311	62	4.08
B	DFSMN	3+2	2,2,2,2	7.33	31.96	3.03	0.1046	0.0302	62	4.08
C	DFSMN	3+2	5,5,2,2	7.23	30.73	3.00	0.1028	0.0294	62	4.08
D	DFSMN	3+2	10,10,2,2	7.15	30.16	2.98	0.1019	0.0288	62	4.09
E	DFSMN	6+2	10,10,2,2	7.11	29.91	2.97	0.1013	0.0285	87	5.35
F	DFSMN	10+2	10,10,2,2	7.07	29.66	2.95	0.1007	0.0282	119	7.04
G	DFSMN	10+2	20,20,2,2	6.99	29.30	2.94	0.1004	0.0277	119	7.06
H	DFSMN	10+2	40,40,2,2	6.92	28.92	2.91	0.0999	0.0272	120	7.10
I	DFSMN	10+2	80,80,2,2	6.87	28.72	2.89	0.0999	0.0269	122	7.18

所有的 DFSMN 模型均由底层的若干 DFSMN 层和上的 2 个全连接层组成，每个 DFSMN 层包含 2048 个结点和 512 个投影结点，而每个全连接层包含 2048 个结点。在上图中，第三列表示该模型由几层 DFSMN 层和几层全连接层组成，第四列表示该模型 DFSMN 层的记忆模块的阶数和步长。由于这是 FSMN 这一类模型首次应用在语音合成任务中，因此我们的实验从一个深度浅且阶数小的模型，即模型 A 开始（注意只有模型 A 的步长为 1，因为我们发现步长为 2 始终稍好于步长为 1 的相应模型）。从系统 A 到系统 D，我们在固定 DFSMN 层数为 3 的同时逐渐增加阶数。从系统 D 到系统 F，我们在固定阶数和步长为 10,10,2,2 的同时逐渐增加层数。从系统 F 到系统 I，我们固定 DFSMN 层数为 10 并再次逐渐增加阶数。在上述一系列实验中，随着 DFSMN 模型深度和阶数的增加，客观指标逐渐降低（越低越好），这一趋势非常明显，且系统 H 的客观指标超过了 BLSTM 基线。



另一方面，我们也做了平均主观得分（MOS）测试（越高越好），测试结果如上图所示。主观测试是通过付费众包平台，由 40 个母语为中文的测试人员完成的。在主观测试中，每个系统生成了 20 句集外合成语音，每句合成语音由 10 个不同的测试人员独立评价。在平均主观得分的测试结果表明，从系统 A 到系统 E，主观听感自然度逐渐提高，且系统 E 达到了与 BLSTM 基线系统一致的水平。但是，尽管后续系统客观指标持续提高，主观指标只是在系统 E 得分的上下波动，没有进一步提高。

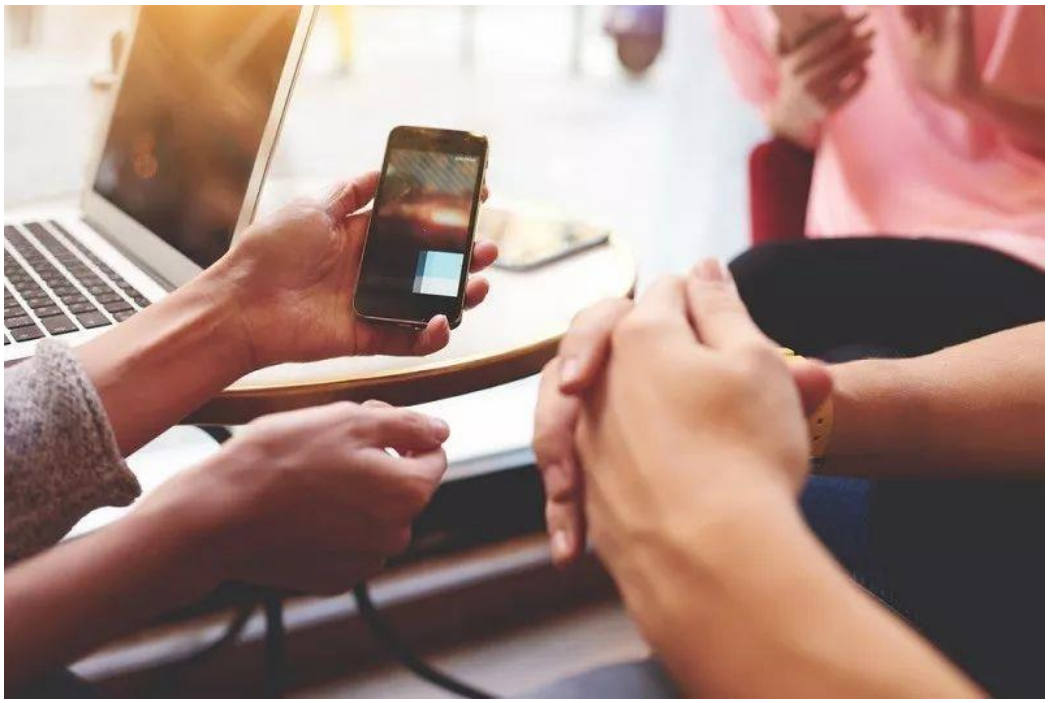
结论

根据上述主客观测试，我们得到的结论是，历史和未来信息各捕捉 120 帧（600 毫秒）是语音合成声学模型建模所需要的上下文长度的上限，更多的上下文信息对合成结果没有直接帮助。与 BLSTM 基线系统相比，我们提出的 DFSMN 系统可以在获得与基线系统一致的主观听感的同时，模型大小只有基线系统的 1/4，预测速度则是基线系统的 4 倍，这使得该系统非常适合于对内存占用和计算效率要求很高的端上产品环境，例如在各类物联网设备上部署。

英文论文地址：https://arxiv.org/abs/1802.09194

为了更精确的情感识别，A-LSTM 出现了

作者：陶斐/Fei Tao, 刘刚/Gang Liu



小叽导读：长短期记忆网络（LSTM）隐含了这样一个假设，本层的现时状态依赖于前一时刻的状态。这种“一步”的时间依赖性，可能会限制 LSTM 对于序列信号动态特性的建模。本篇论文中，针对这样的一个问题，我们提出了高级长短期记忆网络（advancedLSTM (A-LSTM)），利用线性组合，将若干时间点的本层状态都结合起来，以打破传统 LSTM 的这种局限性。在这篇文章中，我们将 A-LSTM 应用于情感识别中。实验结果显示，与应用传统 LSTM 的系统相比，应用了 A-LSTM 的系统能相对提高 5.5%的识别率。

研究背景

LSTM 现在被广泛地应用在 RNN 中。它促进了 RNN 在对序列信号建模的应用当中。LSTM 有两个输入，一个来源于前一层，还有一个来源于本层的前一个时刻。因此，LSTM 隐含了这样一个假设，本层的现时状态依赖于前一时刻的状态。这种“一步”的时间依赖性，可能会限制 LSTM 对于序列信号动态特性的建模（尤

其对一些时间依赖性在时间轴上跨度比较大的任务）。在这篇论文里，针对这样的一个问题，我们提出了 advancedLSTM （A-LSTM），以期打破传统 LSTM 的这种局限性。A-LSTM 利用线性组合，将若干时间点的本层状态都结合起来，因此不仅可以看到”一步“以前的状态，还可以看到更远以前的历史状态。

在这篇文章中，我们把 A-LSTM 应用到整句话层级（utterance level)上的情感识别任务中。传统的情感识别依赖于在整句话上提取底端特征（low level descriptors)的统计数据，比如平均值，方差等等。由于实际应用中，整句话中可能会有一些长静音，或者是一些非语音的声音，这种统计数据就可能不准确。在这篇论文中，我们使用基于注意力模型（attention model)的加权池化(weighted pooling)递归神经网络(recurrent neural network)来更有效地提取整句话层级上的特征。

高级长短期记忆网络

A-LSTM 利用线性组合，将若干时间点的本层状态都结合起来。这其中的线性组合是利用与注意力模型(attention model)类似的机制进行计算的。具体公式如下。

$$C' = \sum_T W_{C_T} \times C_T$$

$$W_{C_T} = \frac{\exp(W \cdot C_T)}{\sum_T \exp(W \cdot C_T)}$$

Fig 1 中 C’(t)即为前面若干时间状态的线性组合。这个线性组合以后的时间状态将被输入下一时间点进行更新。可以想象，每次的更新都不只是针对前一时刻，而是对若干时刻的组合进行更新。由于这种组合的权重是有注意力模型控制，A-LSTM 可以通过学习来自动调节各时间点之间的权重占比。如果依赖性在时间跨度上比较大，则更远以前的历史状态可能会占相对大的比重；反之，比较近的历史状态会占相对大的比重。

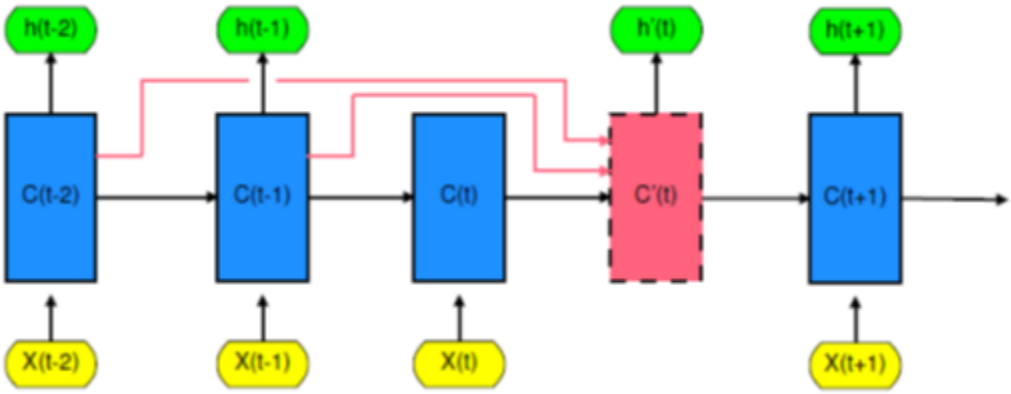


Fig 1 Theunrolled A-LSTM

加权池化递归神经网络

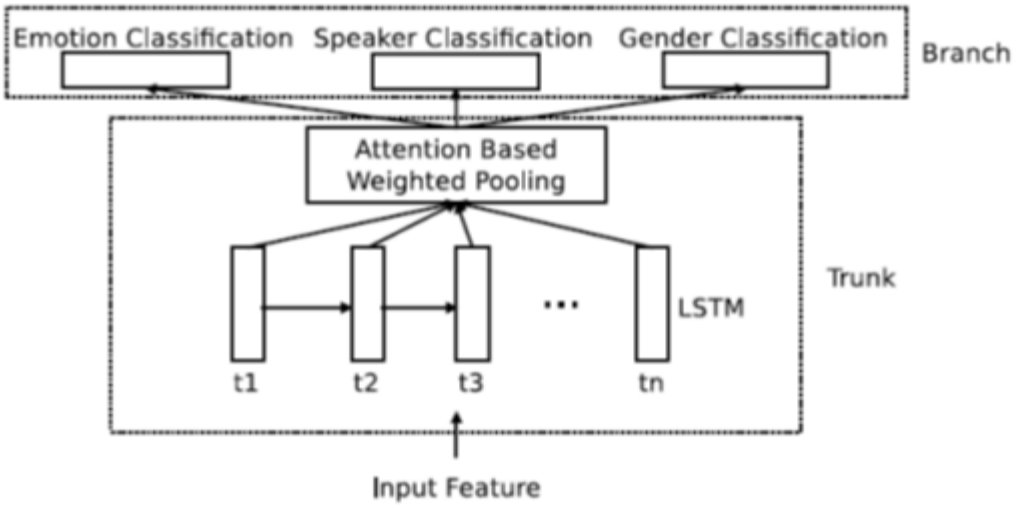


Fig 2 Theattention based weighted pooling RNN.

在这篇论文中，我们使用基于注意力模型的加权池化递归神经网络来进行情感识别（见 Fig 2）。这一神经网络的输入是序列声学信号。利用注意力模型，我们的神经网络可以自动调整各个时间点上的权重，然后将各个时间点上的输出进行加权平均(加权池化)。加权平均的结果是一个能够表征这一整串序列的表达。由于注意力模型的存在，这一表达的提取可以包含有效信息，规避无用信息（比如输入序列中中的一些长时间的静音部分）。这就比简单的计算一整个序列的统计数值要更好（比如有 opensmile 提取的一些底端特征）。 为了更好地训练模型，我们在情感识别任务之外还添加了两个辅助任务，说话人识别和性别识别。我们在这个模型当中使用了 A-LSTM 来提升系统性能。

实验

在实验阶段，我们使用 IEMOCAP 数据集中的四类数据（高兴，愤怒，悲伤和普通）。这其中一共有 4490 句语音文件。我们随机选取 1 位男性和 1 位女性说话人的数据作为测试数据。其余的数据用来训练（其中的 10%的数据用来做验证数据）。我们采用三个衡量指标，分别为无权重平均 F-score（MAF），无权重平均精密度（MAP），以及准确率（accuracy)。

我们提取了 MECC，信号过零率（zero crossing rate)，能量，能量熵，频谱矩心(spectral centroid)，频谱流量(spectral flux)，频谱滚边（spectral rolloff)，12 维彩度向量（chroma vector)，色度偏差(chroma deviation)，谐波比（harmonic ratior）以及语音基频，一共 36 维特征。对这些序列特征进行整句话层级上的归一化后，将其送入系统进行训练或测试。

在这个实验中，我们的系统有两层神经元层，第一层位全连接层（fully connected layer),共有 256 个精馏线性神经元组成(rectified linear unit)。第二层位双向长短期记忆网络（bidirectionalLSTM (BLST))。两个方向一共有 256 个神经元。之后即为基于注意力模型的加权池化层。最上方为三个柔性最大值传输函数层，分别对应三个任务。我们给三个任务分配了不同的权重，其中情感识别权重为 1，说话人识别权重为 0.3，性别识别为 0.6。如果是应用 A-LSTM，我们就将第二层的 BLSTM 替换成双向的 A-LSTM，其他的所有参数都不变。这里的 A-LSTM 选取三个时间点的状态作线性组合，分别为 5 个时间点前（t-5)，3 个时间点前(t-3)，以及 1 个时间点前(t-1)。实验结果如下：

Approach	MAF	MAP	Accuracy (%)
conventional LSTM	43.8	64.3	52.7
mean LSTM	43.5	64.3	52.8
advanced LSTM	**46.2**	**65.8**	**55.3**

其中的 meanLSTM 与 A-LSTM 比较类似，唯一区别是，当我们为选取的几个时间点的状态作线性组合的时候，不是采用注意力模型，而是简单的做算术平均。

结论

与应用传统 LSTM 的系统相比，应用了 A-LSTM 的系统显示出了更好的识别率。由于加权池化过程是将所有时间点上的输出进行加权平均，因此系统性能的提升只可能是来源于 A-LSTM 更加灵活的时间依赖性模型，而非其他因素，例如高层看到更多时间点等等。并且，这一提升的代价只会增加了数百个参数。

原论文地址：https://arxiv.org/pdf/1710.10197.pdf

为了让机器听懂“长篇大论”，阿里工程师构建了新模型

作者： 张仕良、雷鸣、鄢志杰、戴礼荣

小叽导读：本研究我们提出了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）。进一步地我们将深层前馈序列记忆神经网络和低帧率（LFR）技术相结合构建了 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），可以有效地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地传递给低层，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（FiniteImpulse 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。

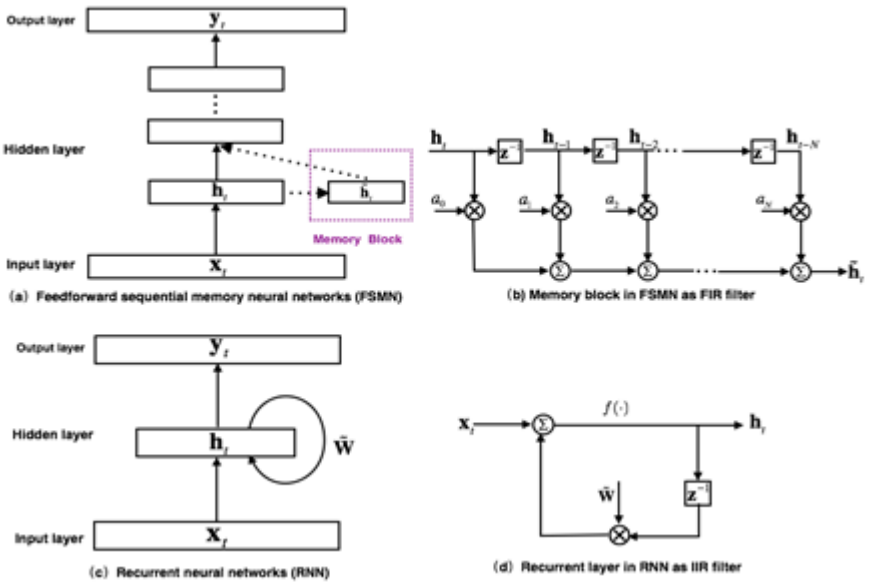


图 1. FSMN 模 型结构以及和 RNN 的对比

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：

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

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

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：

$$\tilde{\mathbf{h}}_t^l = \sum_{i=0}^{N_1} a_i^l \cdot \mathbf{h}_{t-i}^l + \sum_{j=1}^{N_2} c_j^l \cdot \mathbf{h}_{t+j}^l$$

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

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。

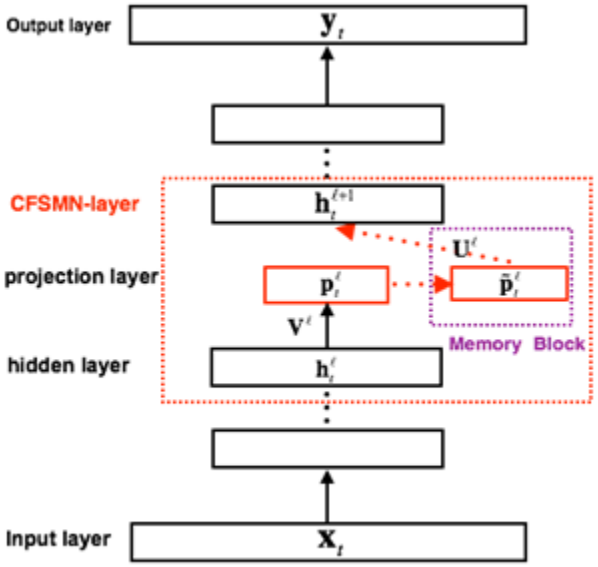


图 2. cFSMN 结构框图

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：

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

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

DFSMN 介绍

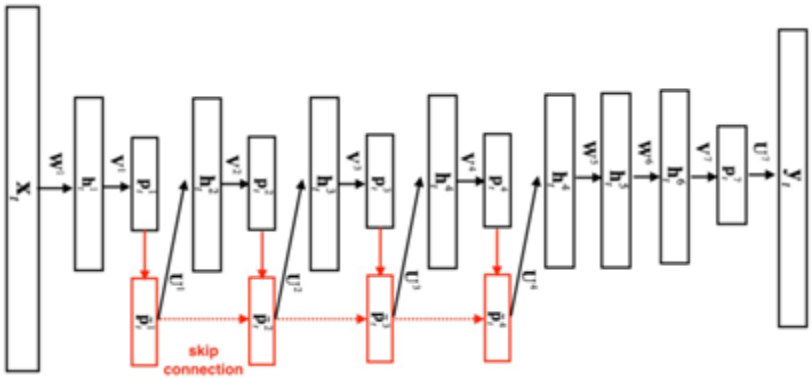


图 3. Deep-FSMN (DFSMN)模型结构框图

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：

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

其中表示第层记忆模块第 t 个时刻的输出。和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，可以极大地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。

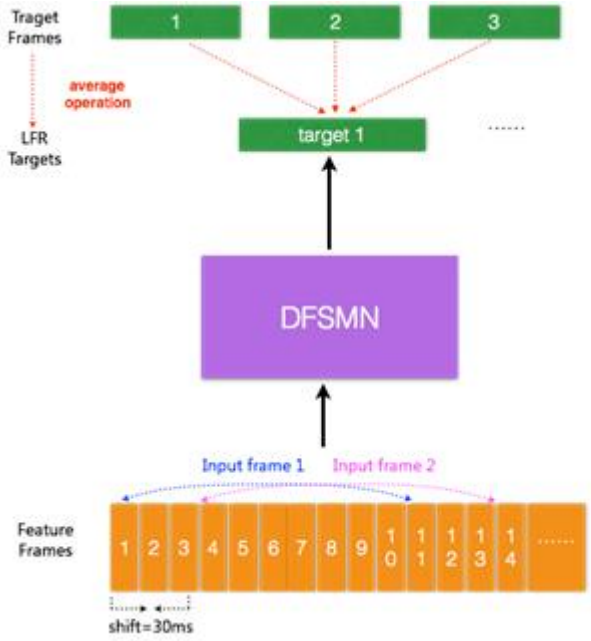


图 4. LFR-DFSMN 声学模型结构框图

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

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%

论文原文：https://arxiv.org/abs/1803.05030

示范了 200 句后，我的声音“双胞胎”诞生了！

作者：黄智颖、卢恒、雷鸣、鄢志杰

小叽导读：语音合成的主要目的是让机器将文字变为人可以听得懂的声音。针对某个人，如果希望机器比较好地发出他的声音，那么需要录制这个人大量（几千到几万句话不等）的音频。很多时候，用户没有时间也没有精力录制这么多的音频。说话人自适应算法就是用来解决这个问题的，它利用用户少量的音频来学习，并能够发出令人满意的声音。使用本文提出的语音合成中的说话人自适应技术，用户只需要录制 200 句话，便能够获得与 1000 句话普通的语音合成系统相当的效果。

摘要

说话人自适应算法利用说话人少量语料来建立说话人自适应语音合成系统，该系统能够合成令人满意的语音。在本文中，我们提出了基于线性网络的语音合成说话人自适应算法。该算法对每个说话人学习特定的线性网络，从而获得属于目标说话人的声学模型。通过该算法，使用 200 句目标说话人的自适应语料训练的说话人自适应系统能够获得和使用 1000 句训练的说话人相关系统相近的合成效果。

研究背景

对于一个目标说话人，如果他（她）拥有充足的训练数据，那么我们便可以建立一个说话人相关的声学模型，基于该声学模型的系统称之为说话人相关的语音合成系统。利用该系统，我们能够合成和目标说话人声音很像的语音。但是，大多数时候，目标说话人没有充足的数据，这使得合成出来的语音效果不太理想。利用说话人自适应算法，能够基于比较有限的数据来获得较好的语音合成系统，该类算法节省了大量的录音、转录和检查工作，使得建立新的声音的代价变得很小。

本文中，我们提出了基于线性网络（Linear Network, LN）的语音合成说话人自适应算法。该算法通过在源说话人声学模型的层间插入线性网络，然后利用目标说话人的数据来更新该线性网络和神经网络的输出层，从而能够获得属于目标说话人的声学模型。另外，一种基于低秩分解（low-rank plus diagonal，LRPD）的模型压缩算法被应用于线性网络。实验发现，当数据量较少的时候，通过 LRPD 来移除一些冗余的参数，从而能够使得系统合成的声音更加稳定。

算法描述

本文中，源说话人声学模型是一个基于多任务（multi-task）DNN-BLSTM 的声学模型，见 Fig. 1 左侧。声学模型的输入为语音学特征，输出为声学特征。声学特征包括梅尔倒谱系数等。实验证明，在声学模型的底层使用深层神经网络（Deep Neural Network，DNN）可以获得更好的底层特征，并且收敛速度上相比于不使用 DNN 更快。在输出层上，不同的声学特征使用各自的输出层，它们仅共享声学模型的隐层。

基于线性网络的自适应算法首先被提出于语音识别领域，它的系统结构见 Fig. 1 右侧。根据线性网络插入的位置不同，它可以被分为线性输入网络（Linear Input Network，LIN）、线性隐层网络（Linear Hidden Network，LHN）和线性输出网络（Linear Output Network，LON）。

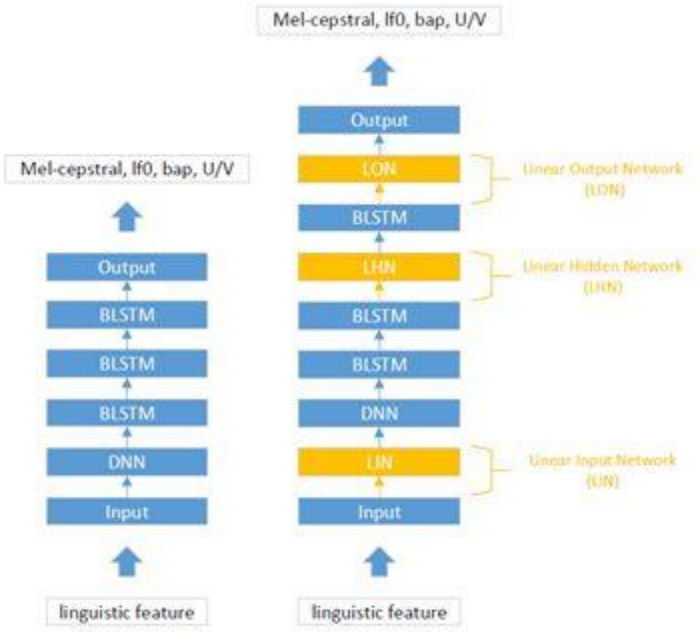


Fig. 1. Multi-task DNN-BLSTM based acoustic model (left) and Linear Network based speaker adaptation (right).

当线性网络被插入到声学模型的第 l 和 $l+1$ 层之间时，线性网络的输出 $\hat{h}^l$ 为：

$$\hat{h}^l = W_s h^l + b_s$$

其中，h^l 表示第 l 层的输出，W_s 表示说话人相关的线性变换矩阵，b_s 表示说话人相关的偏置矢量。模型训练流程如下：

1） 将线性网络插入至源说话人声学模型特定位置。此时，W_s 被初始化为单位矩阵，b_s 的所有元素都初始化为 0。

2） 利用目标说话人的数据来更新线性网络中的参数 W_s 和 b_s，直到收敛。此时，保持声学模型中的其它层参数固定不变。最后，获得目标说话人的声学模型。

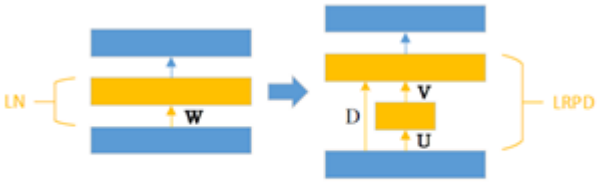


Fig. 2. LN with Low-rank Plus Diagonal decomposition.

LRPD 算法主要被应用于线性网络的模型压缩。在语音识别中，基于 LRPD 的线性网络（LRPD-LN）能够减少普通线性网络（Full-LN）82%的模型参数量，并且性能几乎不出现下降。LRPD 算法利用对角矩阵和低秩矩阵来表达 Full-LN 中的 W_s ：

$$W_{s,k\times k} \approx U_{s,k\times r}V_{s,r\times k} + D_{k\times k}$$

其中， $U_{s,k\times r}$ 和 $V_{s,r\times k}$ 分别表示 $k \times r$ 和 $r \times k$ 的矩阵， $D_{k\times k}$ 为对角矩阵。可以看到， Full-LN 中的模型参数量为 k^2 ， LRPD-LN 的模型参数量为 $k(2r+1)(r \ll k)$ 。通过实验证明，由于 LRPD-LN 所需要更新的参数量特别少，因此在目标说话人数据量有限的情况下能够获得较 Full-LN 更加稳定的合成声音。

实验

本文提出的算法，在中文数据集上进行实验，该数据集包含 3 个说话人，每个说话人有 5000 句话，时长约 5h。数据集中语音的采样率为 16k，特征提取中的窗长和窗移分别为 25ms 和 5ms。分别用 A-male、B- female 和 C-female 来命名这三个说话人。本实验中，源说话人声学模型训练过程所使用的句子数为 5000。

为了对比不同句子数目下的合成效果，目标说话人的自适应数据集对应的句子数从 50 到 1000 不等。在自适应数据集之外，我们取 200 句话作为开发集，取 20 句话作为测试集（用于主观打分）。为了分析性别对自适应效果的影响，进行了三对源说话人-目标说话人之间的实验：女生-女生、男生-女生和女生-男生。另外，使用客观度量和主观测听两种方式来衡量模型的性能。客观度量主要包括：Mel-Cepstral Distortion (MCD)、root mean squared error (RMSE) of F0、unvoiced/voiced (U/V) prediction errors 和开发集的 MSE。主观测听主要是对系统合成的声音样本进行自然度和相似度上的打分——mean opinion score (MOS) 。

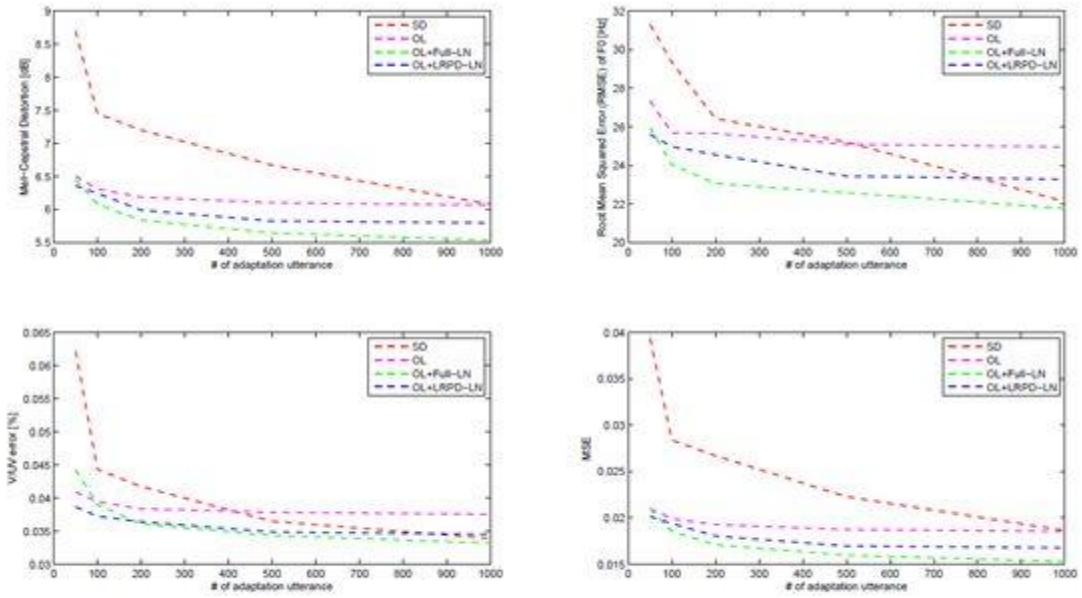


Fig. 3. Objective measurement of validation set utterances (from C-female to B-female).

以女生-女生（C-female - B-female）为例，Fig. 3 显示了不同自适应句子数目和客观度量之间的关系曲线图。其中，SD 表示说话人相关系统，OL 表示只更新源说话人声学模型输出层的说话人自适应系统，OL+Full-LN 和 OL+LRPD-LN 分别表示基于 Full-LN 和 LRPD-LN 的说话人自适应系统。根据 Fig. 3，随着训练/自适应句子数的增加，所有系统间的客观度量趋于相近。对比 SD 和另外三个自适应系统，自适应系统的性能在相同句子数目下要更优。

另外，OL+LRPD-LN 和 OL+Full-LN 相比于 OL 均出现性能上的跳变（提升），说明只更新输出层而不对其他层进行更新不能够得到较好的自适应效果。同时，当自适应句子数较少的时候，OL+Full-LN 在客观性能上要差于 OL+LRPD-LN，这是因为 OL+Full-LN 引入太多的参数量，出现过拟合问题。反之，在句子数多的时候 OL+Full-LN 在客观性能上要优于 OL+LRPD-LN，此时 OL+LRPD-LN 由于参数量少，出现欠拟合问题。

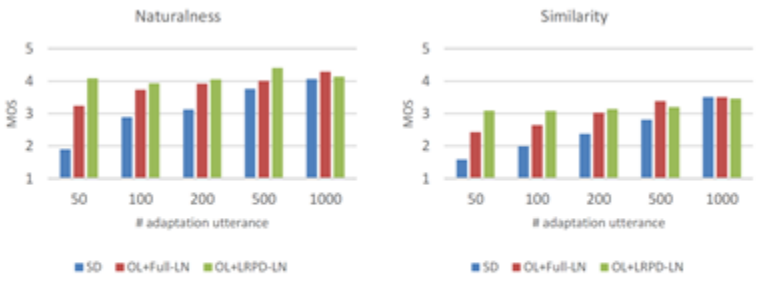


Fig. 4. Subjective tests of testing set utterances (from C-female to B-female).

Fig. 4 上对比了不同系统间的自然度和相似度。随着句子数的减少，SD 系统的性能出现急剧下降，OL+LRPD-LN 相比于 SD 和 OL+Full-LN 要更加稳定。与客观度量一致，在相同句子数下，OL+Full-LN 和 OL+LRPD-LN 在性能上要优于 SD。并且，OL+Full-LN 和 OL+LRPD-LN 在 200 句话的性能和 SD 在 1000 句话时的性能相近。与客观度量不同，OL+LRPD-LN 在 500 句以下的时候性能上就优于 OL+Full-LN。这是因为过拟合导致合成出来的声音不稳定（虽然客观度量更优）声音的可懂度下降导致的。由此，我们依然可以得到相同的结论：当自适应句子数较少的时候，过拟合使得 OL+Full-LN 的性能变差。

结论

本文中，基于线性网络的说话人自适应算法被应用于语音合成领域，基于 LRPD 的模型压缩算法能够提高声音的稳定性。通过三对不同的源说话人-目标说话人的实验，我们发现，当自适应句子数目非常少的时候，LRPD 能够提升声音的稳定性。另外，通过提出的算法，使用 200 句目标说话人的训练语料训练的说话人自适应系统能够获得和使用 1000 句训练的说话人相关系统相近的效果。

原文链接：https://arxiv.org/abs/1803.02445

朋友，我能分享你的喜怒吗？阿里语音情感识别框架揭秘

作者：陶斐、刘刚、赵情恩

小叽导读：情感识别（即，识别开心，忧伤等）现在愈来愈受到人们的关注，因为它可以提升人机交互界面的用户体验，进而提升产品的用户粘性，并在心理医疗健康方面等具有独特价值。基于语音的情感识别尤其具有现实意义，因为基于语音的人机交互界面具有相对较低的硬件要求。但是，在现实中，周围环境中存在着许多噪声，这些噪声将会降低系统的识别性能。在本文中我们提出了一套包含多个子系统的复合情感识别框架。这一框架会深入挖掘输入语音中与情感相关的各个方面的信息，从而提高系统的顽健性。

研究背景

在现实生活中，基于语音的人工智能系统处在复杂的场景当中，因而会面临各种各样的挑战。对于情感识别来说，主要的挑战来自于两个方面：

1. 周围存在背景噪声，因而传统的特征提取，比如在整句话层面上提取统计参数的方法将受到严重干扰；

2. 用户说话的方式比较随意，不能如实验室中那样很好地控制输入语音，有时候用户会有一些发出一些非语音的声音，比如哭声，笑声，咳嗽声等，这些声音有些与情感有关，有些则完全无关。

面对这两个挑战，我们提出了一套复合情感识别框架。这套框架会对底层和高层特征进行识别，因此可以对一些背景噪声有一定的顽健性；同时这套框架也会利用注意力模型（attention model）学习特征序列中重要时间点的特征，以及利用语音中的文本信息对情感信息进行分类——这些机制可以有效避免用户的非语音声音或者长静音对识别的干扰。

复合情感识别框架

在本文中，我们提出了一套复合的情感识别框架。这一框架由若干子系统组合而成，其中包括基于整句话（utterance level)底层特征(low level descriptor)的识别系统，基于整句话高层表述的识别系统，基于序列特征的识别系统，以及基于语义信息的识别系统（见 Fig 1)。

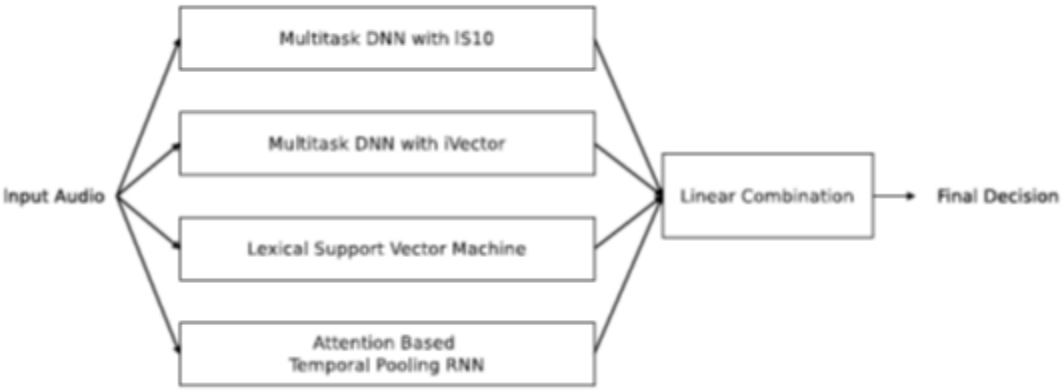


Fig1 The proposed ensemble framework for emotion recognition

其中，基于整句话底层特征的识别系统为一个深度神经网络，采用多任务训练(multitasklearning)方式进行训练(见 Fig 2)，采用的特征为从 opensmile 提取的 Interspeech 2010 LLD 特征集。在这个神经网络中，我们在 trunk 部分有两层隐层（hidden layer)(每层 4096 个神经元)，在 branch 部分，每个任务有一层隐层（1024 神经元），之后有一层 柔性最大激活函数

（softmax)。其中我们的神经元均使用精馏线性单元（rectified linearunit)。

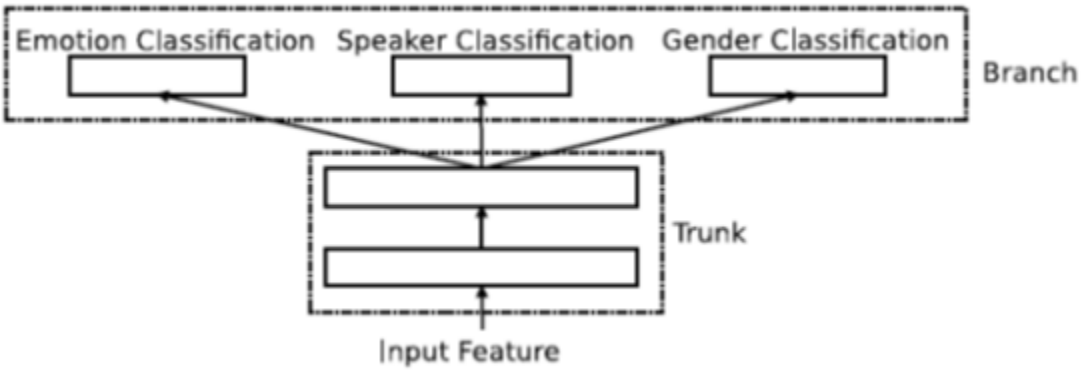


Fig2 The multitask learning DNN

基于整句话高层表述的识别系统也是采用一个深度神经网络，同样也是采用多任务训练方式进行训练。采用的特征为 200 维 iVector(从一个由 4000 小时语音训练的语音识别(ASR)系统中提取）。这里我们采用的网络结构与底层特征识别系统的神经网络相同，唯一的区别为，这个一个系统在 trunk 部分每一层只有 1024 个神经元。

基于序列特征的子系统采用递归神经网络，对输入序列进行建模，在递归神经网络上采用基于 attention model 的加权池化层(weighted pooling)(见 Fig 3)，将输入的一个序列提取成一个高层表述。基于这个高层表述进行分类。这一子系统也采用多任务训练方式进行训练。这一递归网络与上述神经网络的大致结构相似，区别为在 trunk 部分，我们使用了 RNN，并且在 RNN 上利用 attention based weighted pooling layer 来提取高端表述（high level representation)。

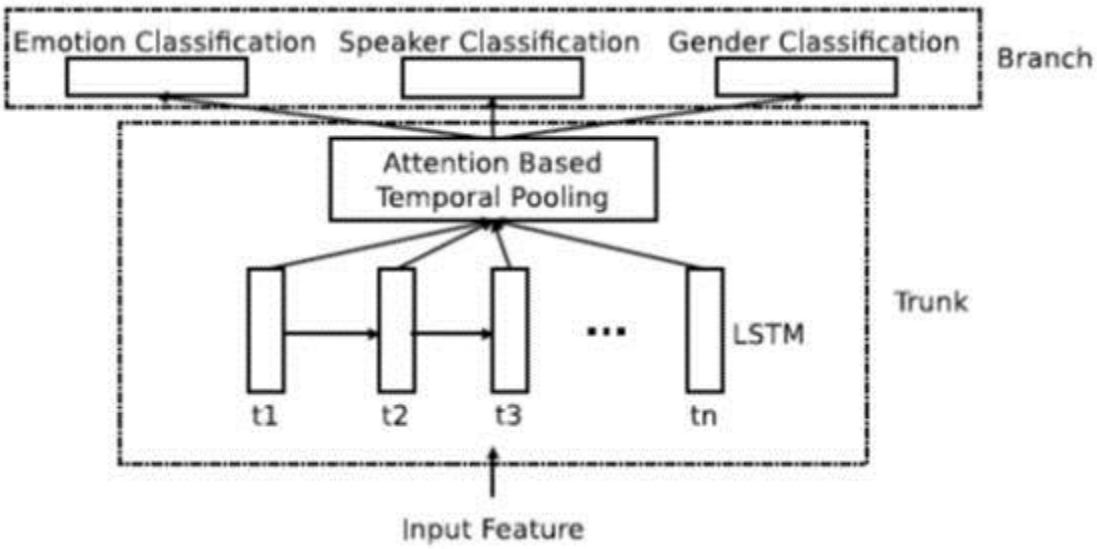


Fig 3 The attention based weighted poolingRNN

上述三个子系统中的多任务训练，我们采用三个任务，情感识别为主任务（权重为 1），说话人识别（权重为 0.3）和性别识别（权重为 0.6）为辅助任务。在多任务训练中，由于系统可以看到更多的任务信息，可以更好地检视输入的特征，因此可以更好地训练神经网络。

除了上述三个子系统外，还有一个子系统是基于文本的子系统。该子系统采用支持向量机（supportvector machine)，使用了从语音识别系统中获取的文本。这一系列子系统的识别结果会通过线性相加组合起来，从而得到最后的结果。

实验

我们在多模情感识别竞赛 2017 数据集（MEC 2017)上测试这一套框架。MEC 2017 数据集是采集自影视作品，其中包含了许多背景噪声（汽车噪声，工厂噪声等等），以及说话人的非语音声音（哭声、笑声等等）。其中各类情感的分布如下。

Category	Training		Testing	
	Count	Proportion	Count	Proportion
Angry	884	0.180	128	0.181
Worried	567	0.115	81	0.115
Anxious	457	0.093	66	0.093
Neutral	1400	0.285	200	0.283
Disgust	144	0.029	21	0.030
Surprise	175	0.036	25	0.035
Sad	462	0.094	67	0.095
Happy	828	0.168	119	0.168
Total	4917	1.000	707	1.000

根据 MEC 2017 的建议，我们采用无权重平均 F-score（MAF）和准确率作为我们的衡量标准。考虑到数据库中的数据不平衡性，我们主要关注 MAF 指标。

实验中，我们采用两套系统作为参照系统，一套是 MEC2017 建议的 random forest 系统，还有一套是利用 Interspeech 2017 特征集搭建 DNN 的情感识别系统。具体实验结果如下：

System Category	Approach	MAF	Accuracy (%)
Baseline_1	Random Forest	22.3	40.0
Baseline_2	DNN	26.4	40.2
Sub-system	Multi-task DNN (IS10)	32.4	44.1
	Multi-task DNN (iVector)	27.4	38.0
	Lexical SVM	17.7	30.1
	Weighted Pooling RNN	23.2	39.7
Proposed	Ensemble Fusion	**34.2**	41.0

由实验结果可以看到，我们提出的这一套框架，可以远远超过参照系统（分别增加了 11.9%和 7.8%准确率）。即使四个子系统的识别率参差不齐，最后组合之后的结果依然超过了所有的子系统，可以推测这个过程中全面检视输入信息，可以很有效的提高识别准确率和系统顽健性。

结论

我们将这一套系统应用于中文的影视作品数据库上。之所以应用到这一数据库上，是因为影视作品中的场景比较接近现实生活。结果显示，我们的系统

可以全面超越现有的基于深度学习的前沿系统。这一成功，可以说明我们的这一套框架可以有助于在现实中实现情感识别。

论文地址：https://arxiv.org/abs/1803.01122.pdf

Alibaba Group
阿里巴巴集团

阿里技术

扫一扫二维码图案，关注我吧

「阿里技术」微信公众号

「阿里巴巴机器智能」微信公众号