学界 | MICROSOFT 2017 会话语音识别系统

2017-11-17 机器海岸线

选自 arXiv

作者：W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke 等

机器海岸线编译

参与：方建勇

THE MICROSOFT 2017 CONVERSATIONAL SPEECH RECOGNITION SYSTEM

W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke

Microsoft AI and Research
Technical Report MSR-TR-2017-39
August 2017

论文链接：https://arxiv.org/pdf/1708.06073

摘要：我们描述了微软的对话式语音识别系统 2017 版，在这个系统中，我们更新了我们的 2016 系统，最近在基于神经网络的声学和语言建模方面取得了新的进展，从而进一步提高了交换机语音识别任务的最新技术水平。 系统为我们之前合并的一组模型体系结构添加了一个 CNN-BLSTM 声学模型，并且在重新分类中包括基于字符和对话会话的 LSTM 语言模型。对于系统组合，我们采用两阶段方法，其中声学模型的子集首先在句音/帧级组合，然后通过混乱网络进行单词级投票。 系统组合后，我们还添加了混淆网络重新调整步骤。 结果系统在 2000 交换板评估集上产生 5.1％的字错误率。

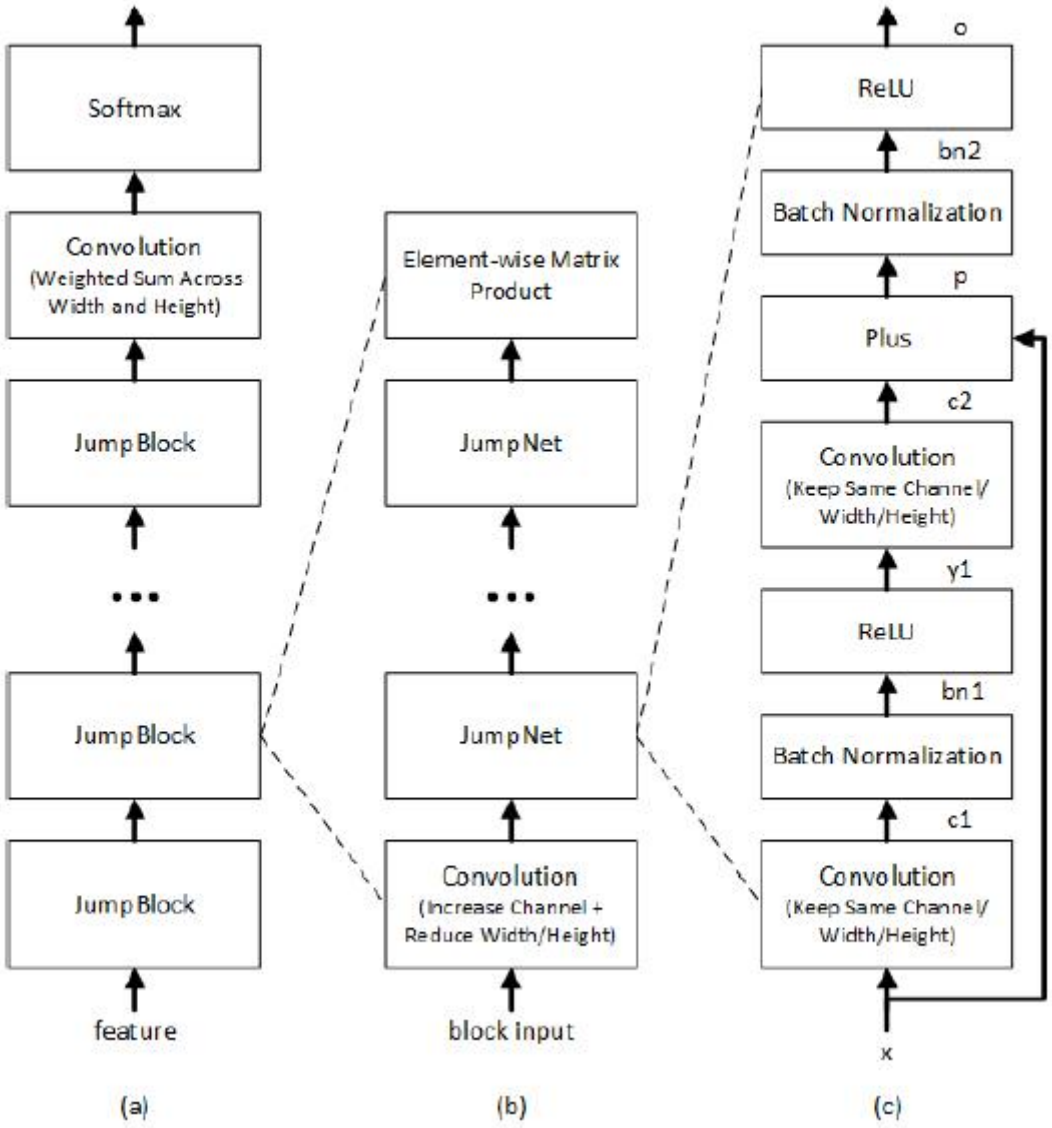
Softmax
Convolution
(Weighted Sum Across
Width and Height)
Jump Block
Jump Block
Jump Block
feature
(a)
Element-wise Matrix
Product
Jump Net
Jump Net
Convolution
(Increase Channel +
Reduce Width/Height)
block input
(b)
o
ReLU
bn2
Batch Normalization
p
Plus
c2
Convolution
(Keep Same Channel/
Width/Height)
y1
ReLU
bn1
Batch Normalization
c1
Convolution
(Keep Same Channel/
Width/Height)
x
(c)

图 1：LACE 网络架构。

	ResNet	LACE	CNN-BLSTM
Number of parameters	38M	65M	48M
Number of weight layers	49	22	10
Input	40x41	40x61	40x7xt
Convolution_1	[conv 1x1, 64 conv 3x3, 64 conv 1x1, 256] x 3	jump block [conv 3x3, 128] x 5	[conv 3x3, 32, padding in feature dim.] x 3
Convolution_2	[conv 1x1, 128 conv 3x3, 128 conv 1x1, 512] x 4	jump block [conv 3x3, 256] x 5	
Convolution_3	[conv 1x1, 256 conv 3x3, 256 conv 1x1, 1024] x 6	jump block [conv 3x3, 512] x 5	
Convolution_4	[conv 1x1, 512 conv 3x3, 512 conv 1x1, 2048] x 3	jump block [conv 3x3, 1024] x 5	
BLSTM			[blstm, cells = 512] x 6
Output	average pool Softmax (9k or 27k)	[conv 3x4, 1] x 1 Softmax (9k or 27k)	Softmax (9k or 27k)

表 1：CNN 层结构和参数的比较。

Senone set	Architecture	devset WER	test WER
9k	BLSTM	11.5	8.3
	ResNet	10.0	8.2
	LACE	11.2	8.1
	CNN-BLSTM	11.3	8.4
	BLSTM+ResNet+LACE	9.8	7.2
	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k puhpum	BLSTM	11.3	8.1
	ResNet	11.2	8.4
	LACE	11.1	8.3
	CNN-BLSTM	11.6	8.4
	BLSTM+ResNet+LACE	9.7	7.4
	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3
27k	BLSTM	11.4	8.0
	ResNet	11.5	8.8
	LACE	11.3	8.8
	BLSTM+ResNet+LACE	10.0	7.5
27k puhpum	BLSTM	11.3	8.0
	ResNet	11.2	8.0
	LACE	11.0	8.4
	BLSTM+ResNet+LACE	9.8	7.3

表 2：通过 Senone 集，模型架构和各种帧级组合的声学模型性能，使用 N-gram LM。 “puhpum” senone 集使用一个备用字典与特殊的手机进行填补停顿。

Model structure	Direction	PPL devset	PPL test
Word input, one-hot	forward	50.95	44.69
	backward	51.08	44.72
Word input, letter-trigram	forward	50.76	44.55
	backward	50.99	44.76
+ alternate text norm	forward	52.08	43.87
	backward	52.02	44.23
+ alternate training set	forward	50.93	43.96
	backward	50.72	44.36
Character input	forward	51.66	44.24
	backward	51.92	45.00

表 3：话语范围的 LSTM-LM 的复杂度。

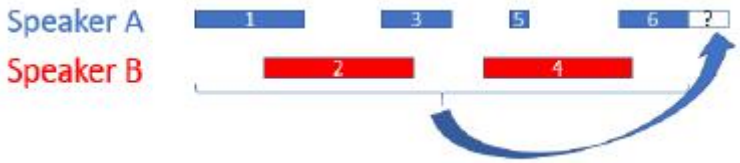

图 2：LACE 网络架构。

Model inputs	PPL devset	PPL test	WER devset	WER test
Utterance words, letter-3grams	50.76	44.55	9.5	6.8
+ session history words	39.69	36.95		
+ speaker change	38.20	35.48		
+ speaker overlap	37.86	35.02		
(with 1-best history)	40.60	37.90	9.3	6.7

表 4：基于会话的 LSTM-LM（仅正向）的复杂度和词语错误。 最后一行反映了在先前话语中对单词使用 1 最佳识别输出。

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
27k-puhpum	BLSTM	11.3	8.0	9.2	6.3
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2	7.7	5.4
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	5.8
-	Confusion network combination			7.4	5.2
-	+ LSTM rescoring			7.3	5.2
-	+ ngram rescoring			7.2	5.2
-	+ backchannel penalty			7.2	5.1

表 5：LSTM-LM 在选定的组合，系统组合以及混淆网络重新选定的系统上进行重新分级的结果。

✂----------