Unsupervised Speaker Adaptation of BLSTM-RNN for LVCSR Based on Speaker Code

Zhiying Huang[1], Shaofei Xue[2], Zhijie Yan[2], Lirong Dai[1]

[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P. R. China
[2]Alibaba Inc.

presented by Zhiying Huang / 黄智颖

Outline

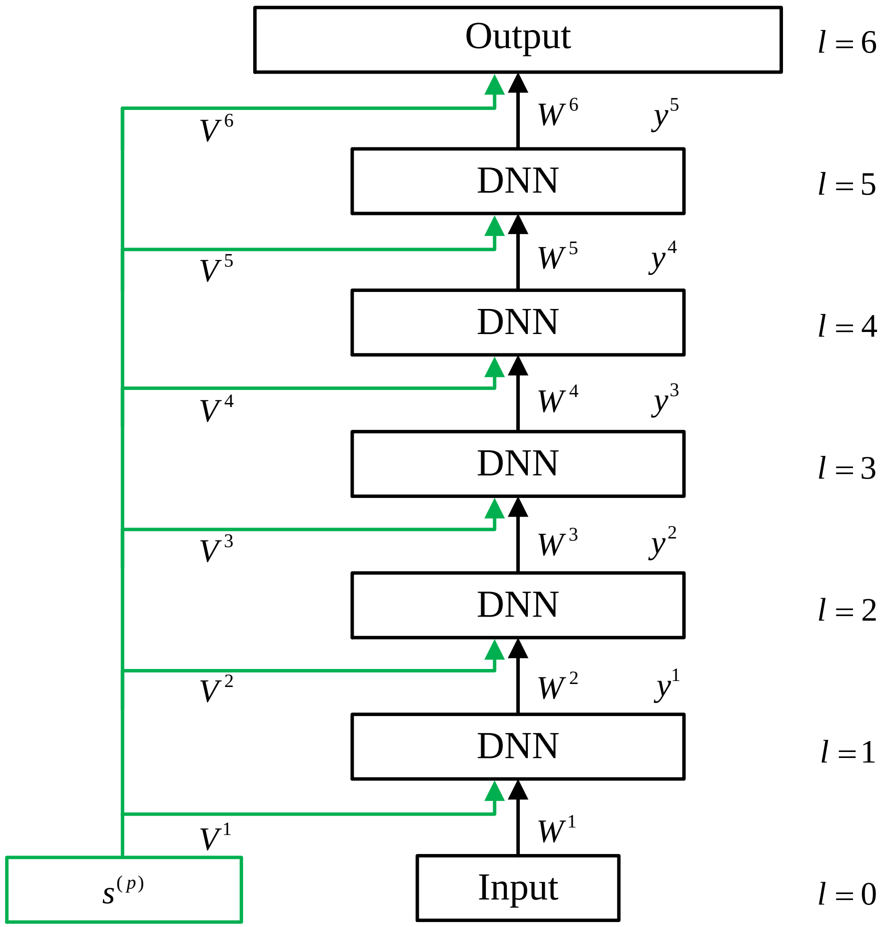
- Introduction
- Proposed method
- Experiment and analysis
- Conclusion and future work

Outline

- Introduction
- Proposed method
- Experiment and analysis
- Conclusion and future work

Speaker code based adaptation

Speaker code based adaptation relies on some speaker-specific discriminative codes, which are connected to a large speaker-independent neural network through a separate set of connection weights.

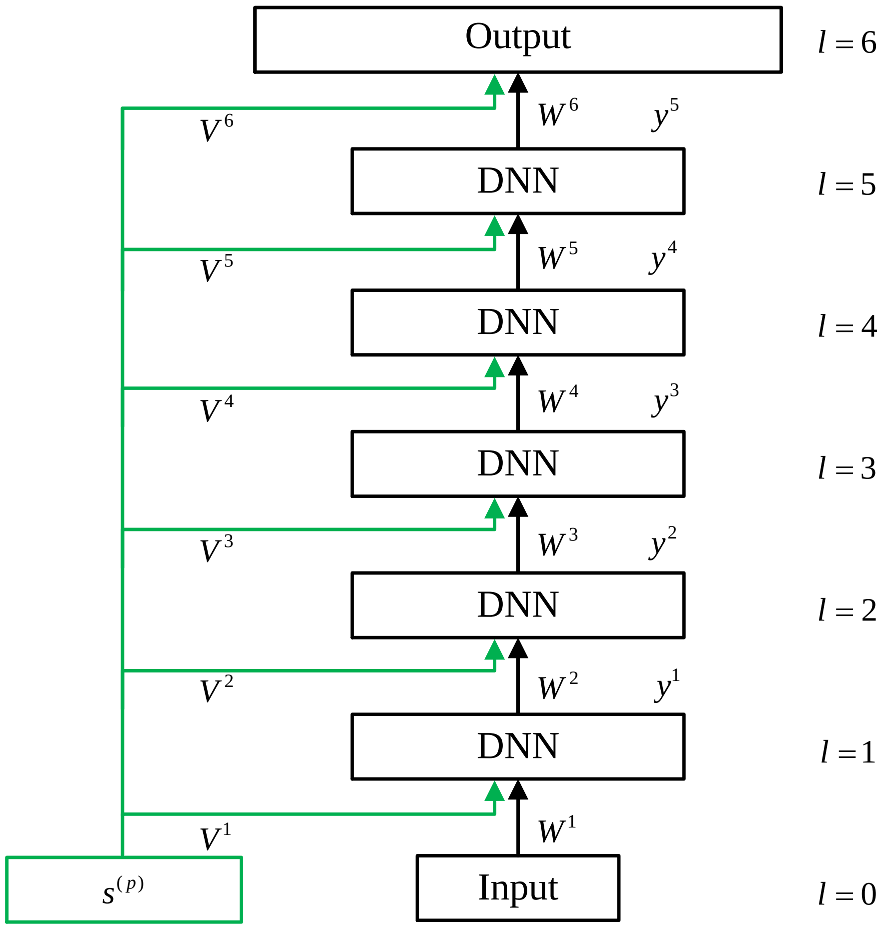


Speaker code based adaptation

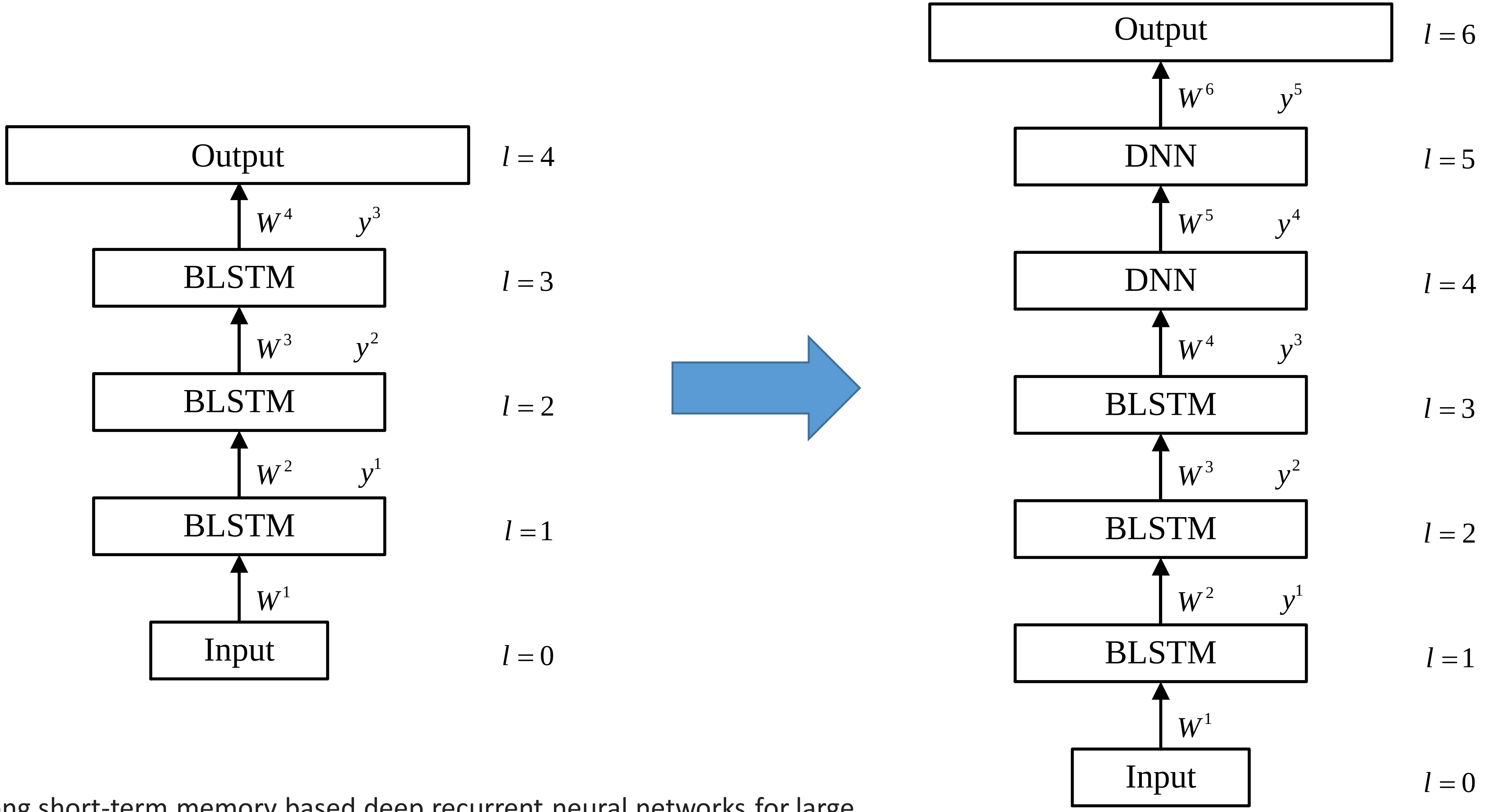
Speaker code based adaptation relies on some speaker-specific discriminative codes, which are connected to a large speaker-independent neural network through a separate set of connection weights.

Unsupervised speaker adaptation:
All unlabeled testing utterances of each testing speaker are used for adaptation stage.

Supervised speaker adaptation:
Part of the testing utterances are labelled, and they are used for adaptation stage.



Hybrid BLSTM-DNN topology

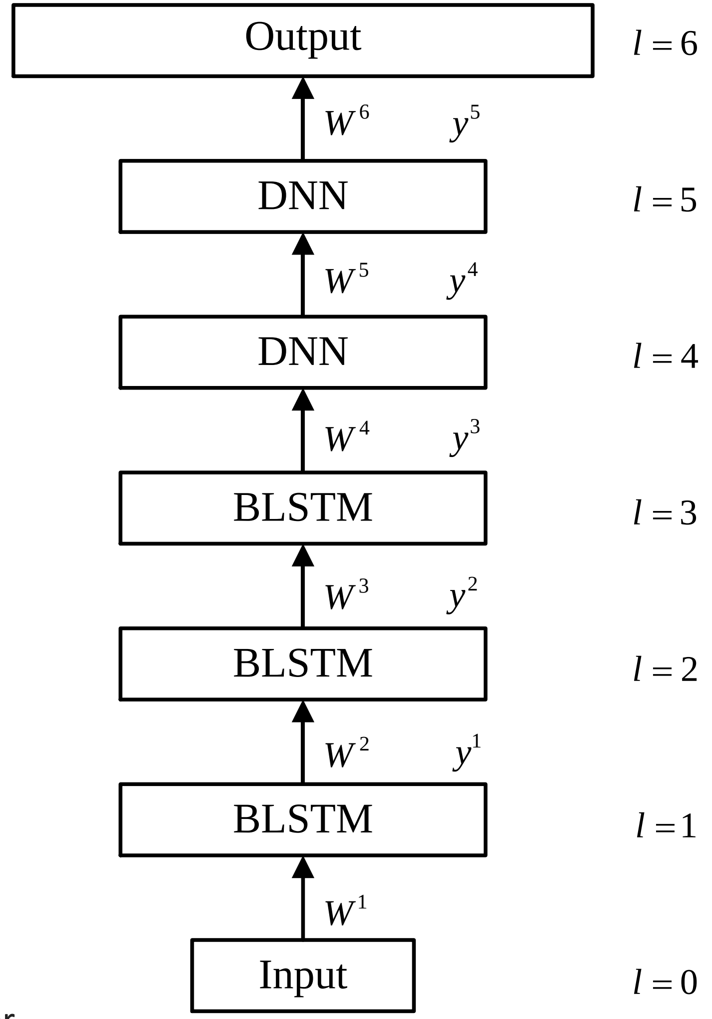


*: Li X, Wu X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 4520-4524.

Hybrid BLSTM-DNN topology

- BLSTMs are good at temporal modeling, and DNNs are appropriate for mapping features to a more separable space.

- For speaker adaptation, we try to reduce frequency variations and model temporal information of different speakers.
- Speaker code based adaptation in hybrid BLSTM-DNN topology?

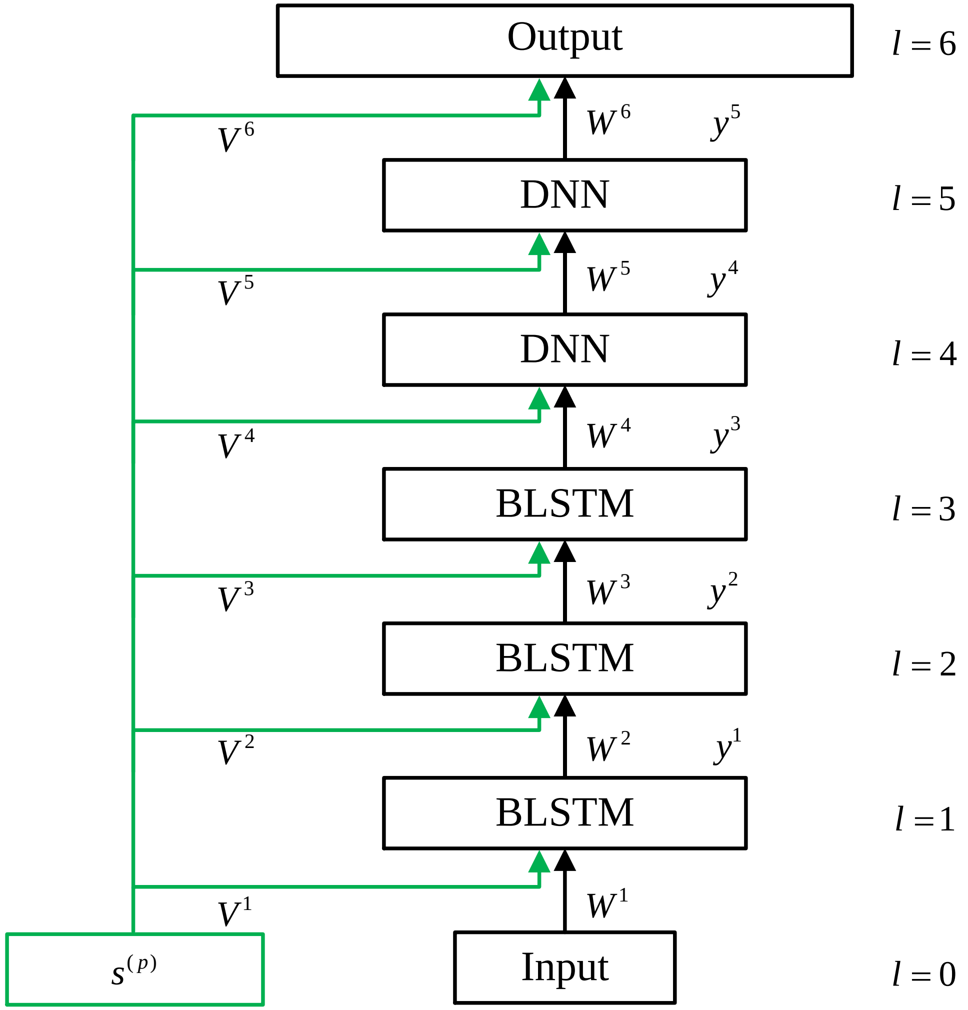


*: Huang Z, Tang J, Xue S, et al. Speaker adaptation OF RNN-BLSTM for speech recognition based on speaker code[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 5305-5309.

Outline

- Introduction
- Proposed method
- Experiment and analysis
- Conclusion and future work

Speaker code based adaptation on the hybrid BLSTM-DNN topology



Speaker code based adaptation on the hybrid BLSTM-DNN topology

For the **fully-connected DNN** layer

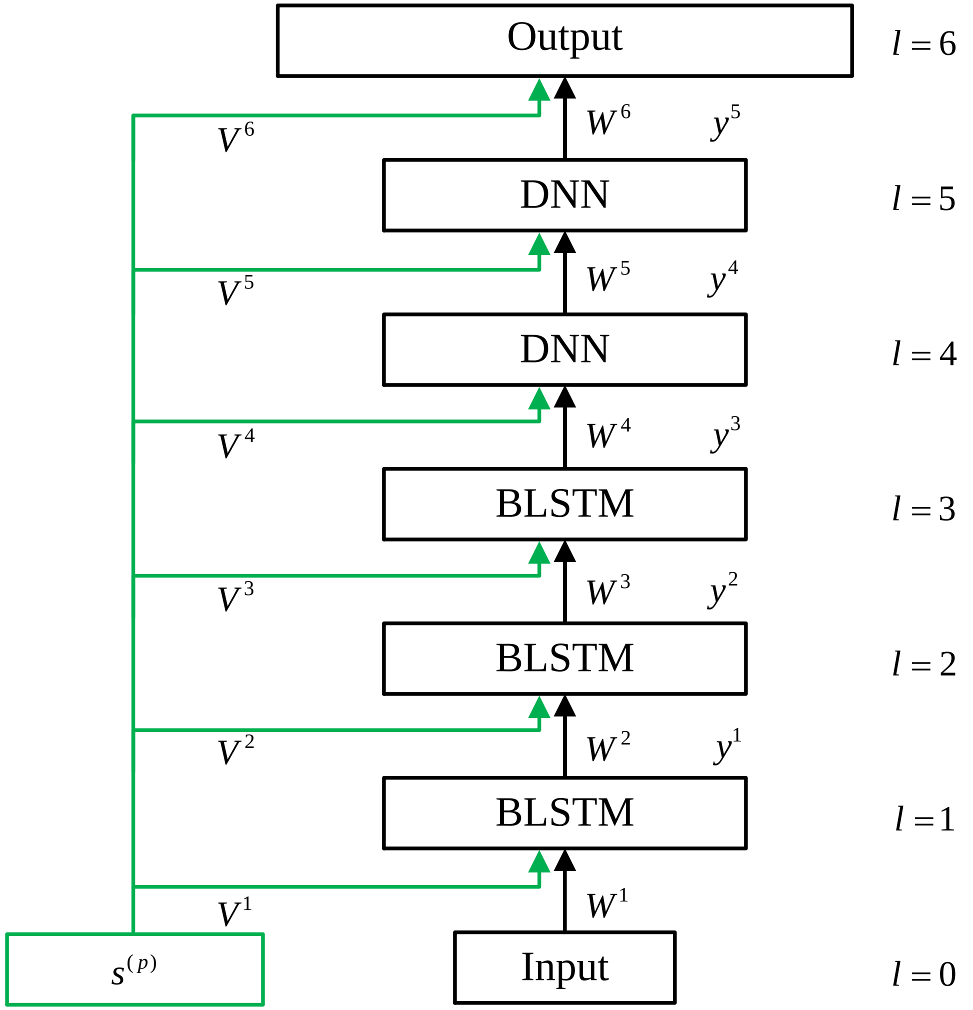
The propagation equation

$$y^l = \varphi(W^l y^{l-1} + b^l + \boxed{V^l s^{(p)}})$$

The gradients of V^l and $s^{(p)}$ of the l-th fully-connected DNN layer (cross entropy criterion):

$$\frac{\partial E}{\partial V^l_{kj}} = \frac{\partial E}{\partial y^l_j} \varphi(\cdot)' s^{(p)}_k$$

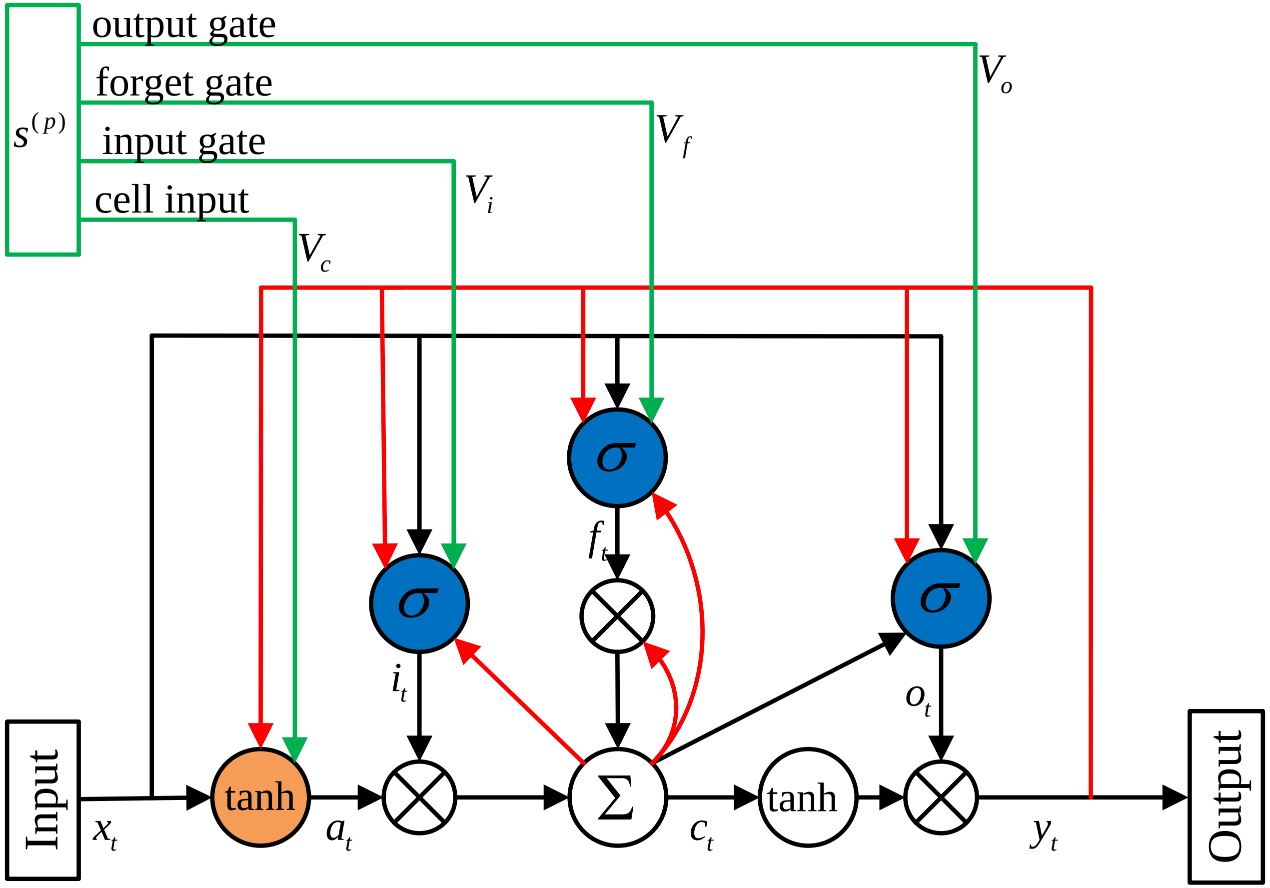
$$\left(\frac{\partial E}{\partial s^{(p)}_k}\right)^l = \sum_{j=1}^{J_l} \frac{\partial E}{\partial y^l_j} \varphi(\cdot)' V^l_{kj}$$



Speaker code based adaptation on the hybrid BLSTM-DNN topology

For the **BLSTM** layer

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{yi}y_{t-1} + W_{ci}c_{t-1} + b_i + \boxed{V_i s^{(p)}}) \\
f_t &= \sigma(W_{xf}x_t + W_{yf}y_{t-1} + W_{cf}c_{t-1} + b_f + \boxed{V_f s^{(p)}}) \\
a_t &= \tanh(W_{xc}x_t + W_{yc}y_{t-1} + b_c + \boxed{V_a s^{(p)}}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot a_t \\
o_t &= \sigma(W_{xo}x_t + W_{yo}y_{t-1} + W_{co}c_t + b_o + \boxed{V_o s^{(p)}}) \\
y_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$



we define *g* represents *i* or *f* or *a* or *o*

$$
\begin{aligned}
\frac{\partial E}{\partial V^l_{g_{kj}}} &= \frac{\partial E}{\partial g^l_{t_j}} \phi(\cdot)' s^{(p)}_k \\
\left(\frac{\partial E}{\partial s^{(p)}_k}\right)^l &= \sum_{j=1}^{J_l} \frac{\partial E}{\partial g^l_{t_j}} \phi(\cdot)' V^l_{g_{kj}}
\end{aligned}
$$

Where $\phi(\cdot)$ denotes active function (such as $\sigma(\cdot), \tanh(\cdot)$)

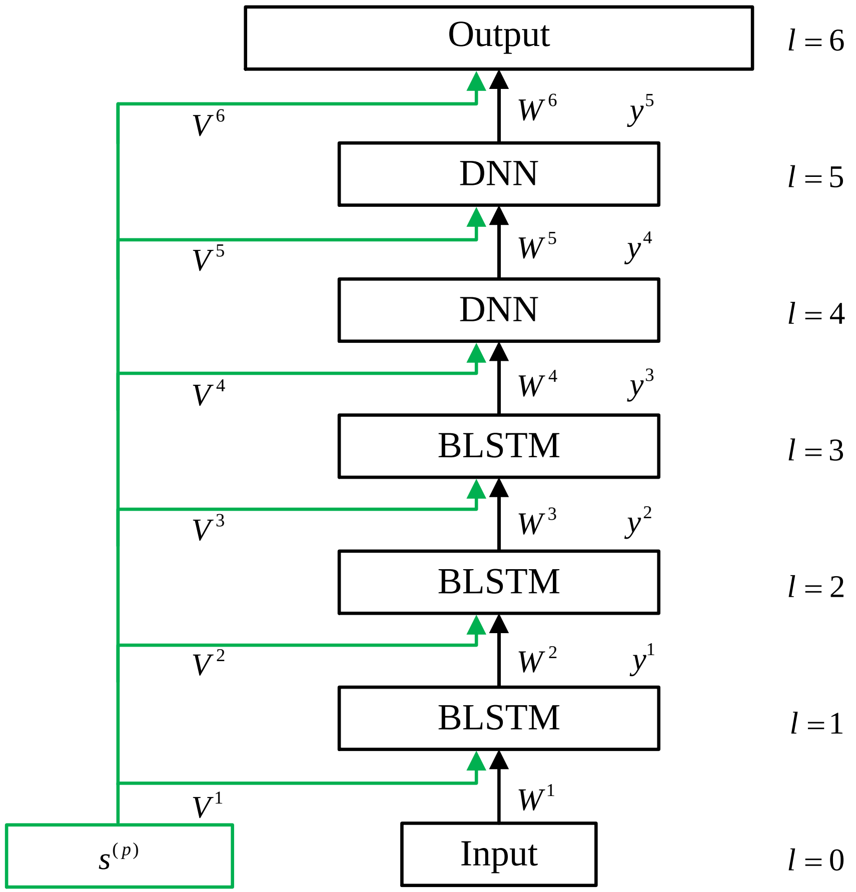
Speaker code based adaptation on the hybrid BLSTM-DNN topology

➢ **mSA-SC**

model space speaker adaptation based on speaker codes

➢ **SAT-SC**

joint speaker adaptive training based on speaker codes



Speaker code based adaptation on the hybrid BLSTM-DNN topology

- **mSA-SC**

1) Training

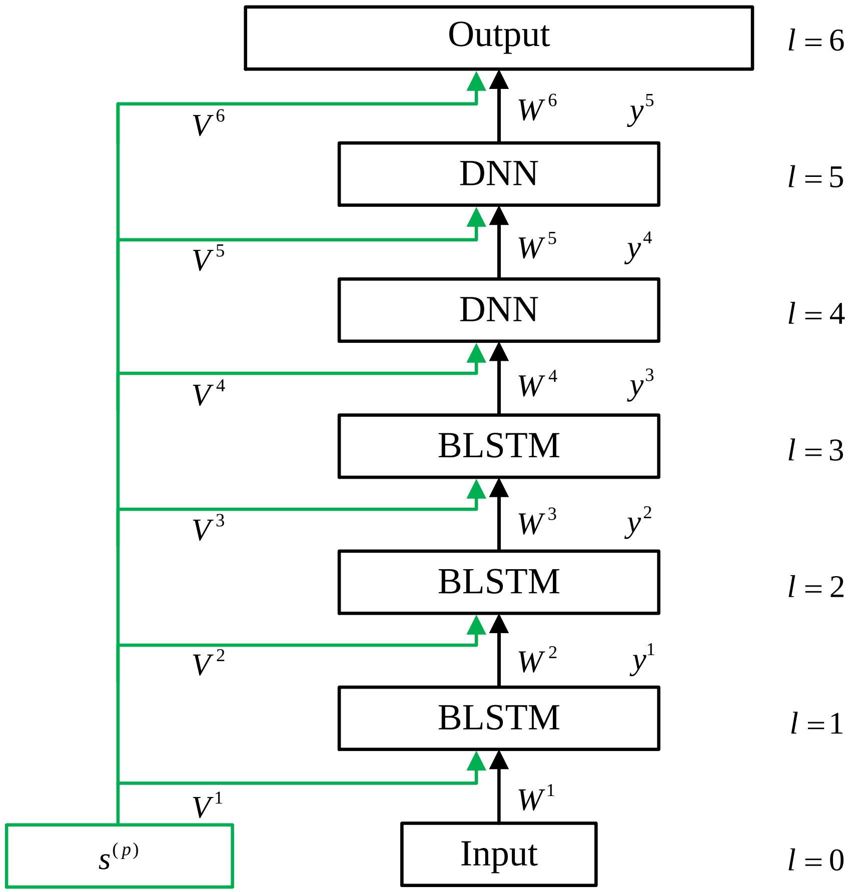
The weights W^l are initialized by a pre-trained SI baseline, while $s^{(p)}$ and V^l are all initialized randomly. $s^{(p)}$ and V^l are learned using all training data with the back propagation (BP) algorithm while keeping W^l fixed.

2) Adaptation

W^l and V^l remain unchanged and the speaker code of each testing speaker is learned based on the BP algorithm.

3) Testing

The speaker code is fed into the neural network through V^l for final recognition.



Speaker code based adaptation on the hybrid BLSTM-DNN topology

➢ **SAT-SC**

1) Training

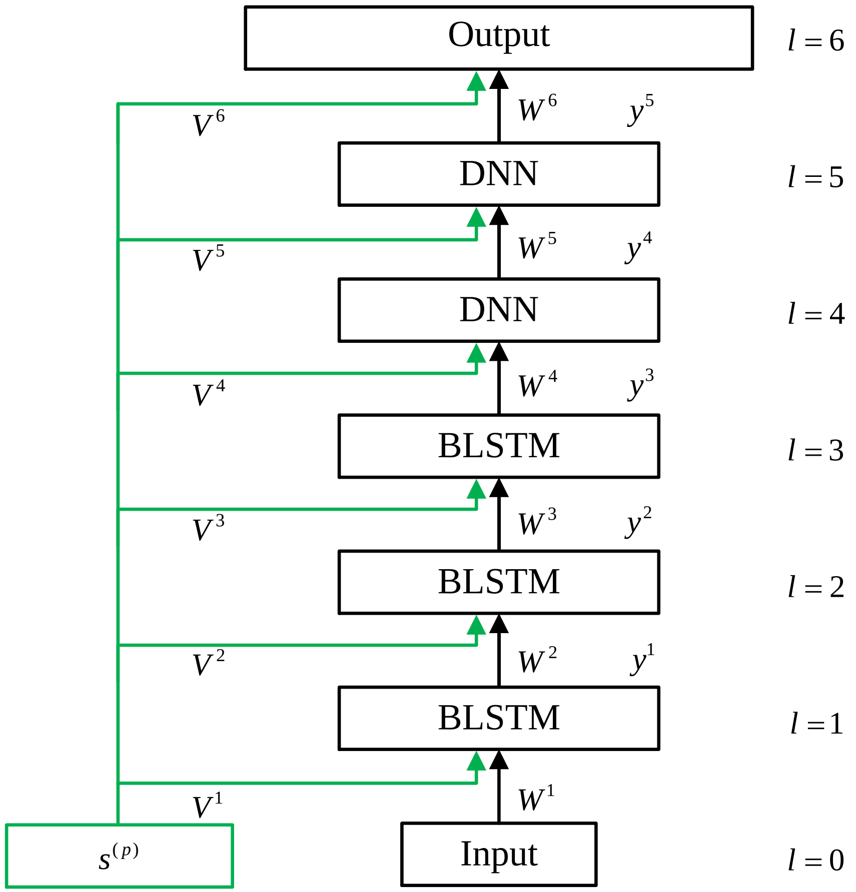
After mSA-SC training, $s^{(p)}$ and V^l are well-initialized. All model parameters (W^l , $s^{(p)}$ and V^l) can be jointly updated using all training data in SAT-SC training.

2) Adaptation

(The same as mSA-SC)

3) Testing

(The same as mSA-SC)



Layer-width normalization

In the traditional speaker code based adaptation on DNN-HMM and BLSTM, $\frac{1}{L-1}$ is applied to scale the gradients of $s^{(p)}$ from each layer:

$$\frac{\partial E}{\partial s_k^{(p)}} = \sum_{l=1}^{L-1} \left[\frac{1}{L-1} \left(\frac{\partial E}{\partial s_k^{(p)}} \right)^l \right]$$

Where the L means the number of layer (include input, hidden and output layer).

Layer-width normalization

The fully-connected DNN layer owns probably more hidden nodes than the BLSTM layer, and it contributes to more accumulated errors for $s^{(p)}$ than the BLSTM layer. The gradients of $s^{(p)}$ from the fully-connected DNN layer and the BLSTM layer are unbalanced.

Layer-width normalization is proposed to use the reciprocal of the node number of fully-connected DNN or BLSTM layer to reduce the imbalance.

No norm

$$\frac{\partial E}{\partial s_k^{(p)}} = \sum_{l=1}^{L-1} \left[\frac{1}{L-1} \left(\frac{\partial E}{\partial s_k^{(p)}} \right)^l \right]$$

Layer-width norm

$$\frac{\partial E}{\partial s_k^{(p)}} = \sum_{l=1}^{L-1} \left[\frac{\frac{1}{J_l}}{\sum_{k=1}^{L-1} (\frac{1}{J_k})} \left(\frac{\partial E}{\partial s_k^{(p)}} \right)^l \right]$$

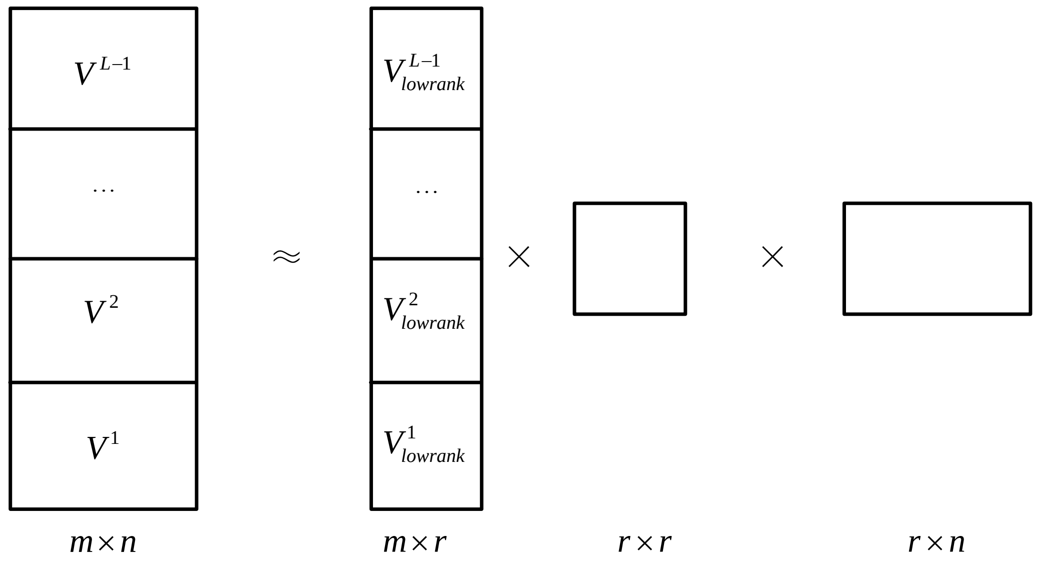
SVD for model compression

SVD (singular value decomposition) :

1) The standard connection weights V^l are firstly trained, and they are jointly spliced into one $m \times n$ $(m = J_1 + J_2 + \cdots + J_{L-1})$ matrix.

2) Standard SVD is used to decompose it and throw out some eigenvectors with small singular values to obtain a $m \times r$ $(r < n)$ matrix.

3) Splitting it into $L-1$ matrices, low-rank connection weights $V_{lowrank}^{l}$ are generated, and they are used for adaptation and testing stage of speaker code based adaptation.



Outline

- Introduction
- Proposed method
- Experiment
- Conclusion and future work

Experiments

➢ Switchboard (SWB) task

- Training data: 309-hour Switchboard-I training database and 20-hour Call Home English data (4870 speakers)
- Test set: Switchboard part of NIST 2000 Hub5e (40 speakers, 1831 utterances)
- MLE trained GMM-HMM (8882 tied states), which is used to obtain state labels
- 108-dimensional filter-bank features (with static, first and second order derivatives)
- 4-gram language model (LM) is trained using 3M words from the training transcripts and 11M words from the Fisher English Part 1 transcripts
- Learning criterion: cross-entropy (CE)

Experiments

Hybrid BLSTM-DNN	
Model	3 * 500[BLSTM] + 2 * 2048[ReLU_DNN]
Latency-controlled method	$N_c = 60, N_r = 30, stream = 30$
WER (%)	13.0

*: Zhang Y, Chen G, Yu D, et al. Highway long short-term memory RNNS for distant speech recognition[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 5755-5759.

Experiments

Table 1: *Comparison of WERs (in %) of different adaptation schemes of mSA-SC. The speaker code dimension is set to 1,000.*

w/o or w/ adaptation	activation function	WER
w/o adaptation	-	13.0
w/ adaptation	cell input	12.3
	input gate	12.9
	forget gate	13.0
	output gate	12.8
	cell input + all gates	12.2

Experiments

Table 1: *Comparison of WERs (in %) of different adaptation schemes of mSA-SC. The speaker code dimension is set to 1,000.*

w/o or w/ adaptation	activation function	WER
w/o adaptation	-	13.0
w/ adaptation	cell input	12.3
	input gate	12.9
	forget gate	13.0
	output gate	12.8
	cell input + all gates	12.2

Experiments

Table 2: *Comparison of WERs (in %) of SAT-SC for speaker adaptation with different connection schemes. The speaker code dimension is set to 1,000. (The WER of the baseline is 13.0%)*

norm	Connection Scheme	WER
no norm	BLSTM layers	12.0
	fully-connected DNN layers	12.5
	all layers	12.2
layer-width norm	all layers	**11.8**

Experiments

Table 2: *Comparison of WERs (in %) of SAT-SC for speaker adaptation with different connection schemes. The speaker code dimension is set to 1,000. (The WER of the baseline is 13.0%)*

norm	Connection Scheme	WER
no norm	BLSTM layers	12.0
	fully-connected DNN layers	12.5
	all layers	12.2
layer-width norm	all layers	**11.8**

Experiments

Table 2: *Comparison of WERs (in %) of SAT-SC for speaker adaptation with different connection schemes. The speaker code dimension is set to 1,000. (The WER of the baseline is 13.0%)*

norm	Connection Scheme	WER
no norm	BLSTM layers	12.0
	fully-connected DNN layers	12.5
	all layers	12.2
layer-width norm	all layers	**11.8**

Experiments

Table 3: *WERs (in %) of SAT-SC for speaker adaptation with different speaker code (SC) dimensions. (The WER of the baseline is 13.0%)*

SC dimension	300	500	1000	1500	2000
w/o SVD	12.1	12.3	**11.8**	12.0	12.1

Table 4: *Comparison of WERs (in %) between i-vector and using SVD compression for speaker adaptation. (The WER of the baseline is 13.0%)*

i-vector/SC dimension	200	300	400	500
i-vector	12.1	12.1	12.1	12.0
w/ SVD	11.8	11.8	12.0	11.9

Experiments

Table 3: *WERs (in %) of SAT-SC for speaker adaptation with different speaker code (SC) dimensions. (The WER of the baseline is 13.0%)*

SC dimension	300	500	1000	1500	2000
w/o SVD	12.1	12.3	**11.8**	12.0	12.1

Table 4: *Comparison of WERs (in %) between i-vector and using SVD compression for speaker adaptation. (The WER of the baseline is 13.0%)*

i-vector/SC dimension	200	300	400	500
i-vector	12.1	12.1	12.1	12.0
w/ SVD	11.8	11.8	12.0	11.9

Experiments

Table 3: *WERs (in %) of SAT-SC for speaker adaptation with different speaker code (SC) dimensions. (The WER of the baseline is 13.0%)*

SC dimension	300	500	1000	1500	2000
w/o SVD	12.1	12.3	**11.8**	12.0	12.1

Table 4: *Comparison of WERs (in %) between i-vector and using SVD compression for speaker adaptation. (The WER of the baseline is 13.0%)*

i-vector/SC dimension	200	300	400	500
i-vector	12.1	12.1	12.1	12.0
w/ SVD	11.8	11.8	12.0	11.9

Outline

- Introduction
- Proposed method
- Experiment
- Conclusion and future work

Conclusion and future work

➢ **Conclusion**

- we conduct speaker code based adaptation on hybrid BLSTM-DNN topology in large-scale Switchboard task.
- we use layer-width normalization to reduce the imbalance of back-propagation errors from different layers for speaker codes.
- SVD is used for compressing the dimension of speaker codes.

➢ **Future work**

- explore in solving the disadvantage of two-pass decoding while conducting speaker code based adaptation.

Thank You for Listening!
Q&A

Zhiying Huang / 黄智颖
Email: zyhuang@mail.ustc.edu.cn