Investigating Gated Recurrent Neural Networks for Acoustic Modeling

Yuanyuan Zhao, Jie Li, Shuang Xu, Bo Xu

Interactive Digital Media Technology Research Center
Institute of Automation, Chinese Academy of Sciences

Outline

1. Motivation
2. Contributions
3. Network Architecture
4. Experiments

Motivation

- Gated Recurrent Neural Networks (RNNs) have achieved state-of-art results in acoustic modeling for LVCSR.

Why this?

Motivation

- Gated Recurrent Neural Networks (RNNs) have achieved state-of-art results in acoustic modeling for LVCSR.

Why this?

- If it's true

Motivation

- Gated Recurrent Neural Networks (RNNs) have achieved state-of-art results in acoustic modeling for LVCSR.

Why this?

- If it's true

What grantee this?

Motivation

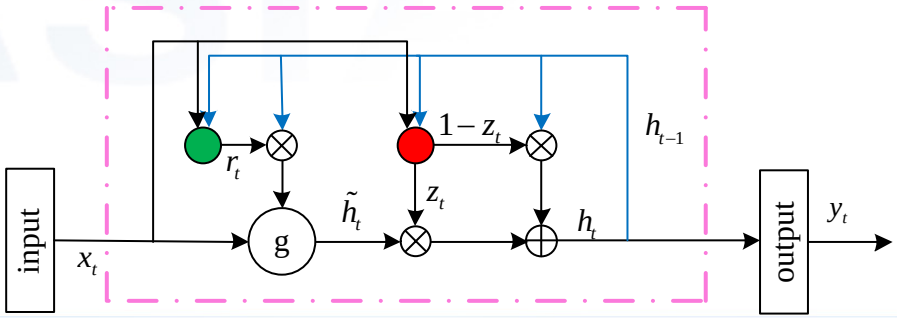
- There questions on Gated RNNs based acoustic model:
  - ✓ What's the key part in acoustic modeling for speech recognition with LSTM and GRU?
  - ✓ What's the difference between them?
  - ✓ In acoustic modeling, the LSTM with projection layer has become the popular architecture. How did it not only reduce the number of parameters of standard LSTM, but also the guarantee for good performance on LVCSR tasks?
  - ✓ Did the performance of LSTM and GRU is relevant to the data corpus in acoustic modeling?

Contributions

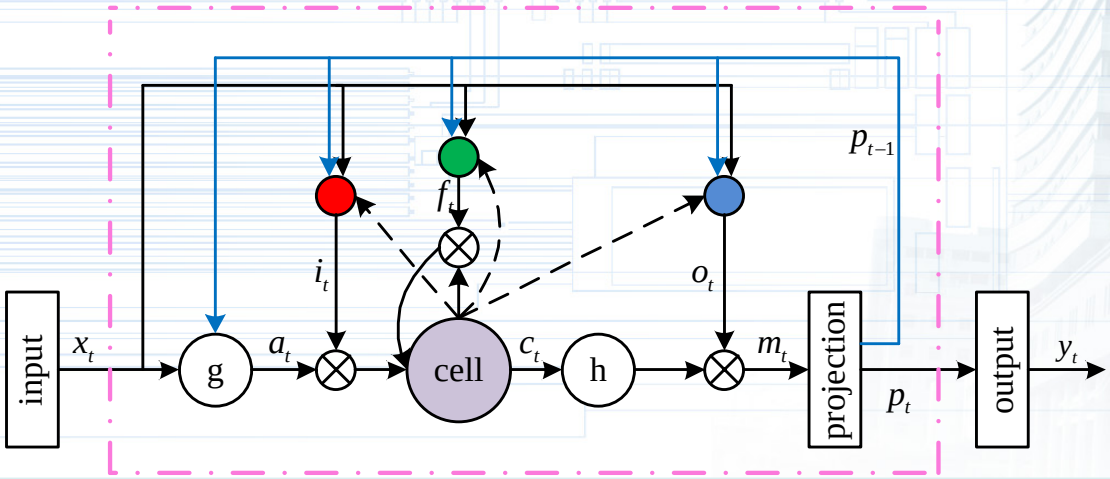
- Attempt to investigate the key parts gated RNNs
  - ✓ GRUs ***modulate*** the proportion of previous output activation preserved through learned reset gates
  - ✓ LSTMs use ***all of the information*** flowing from previous output without any selection or controlling.
- First explore the effect of projection layer in LSTM
  - ✓ weights sharing and information selection and combination
  - ✓ weights sharing can lead to better generalization performance
- The conclusion is channel and corpus size irrelevant

Network Architecture

- Gated recurrent unit



- Long short-term memory projected unit



Network Architecture

- Discuss(commonalities):
  - ✓ All keep the activation of previous time step additive and add a new content to it rather than simply replacing the activation with a new value in conventional RNNs.

$$c_t = f_t \odot c_{t-1} + i_t \odot g\left(W_{cx} x_t + W_{cm} m_{t-1} + b_c\right)$$

$$h_t = \left(1 - z_t\right) \odot h_{t-1} + z_t \odot m_t$$

  - ✓ Two advantages: easier to remember the important information for a long series of steps, and making the information forward and backward propagation effectively without too quickly gradient vanishing problem

Network Architecture

- Discuss(differences):
  - ✓ GRU difference from LSTM in using previous memory content.

$$c_t = f_t \odot c_{t-1} + i_t \odot g\left(W_{cx} x_t + W_{cm} m_{t-1} + b_c\right)$$

$$h_t = \left(1 - z_t\right) \odot h_{t-1} + z_t \odot m_t$$

  - ✓ GRU selecting previous information is similar with the effect of projection layer in LSTMP.

Experiments

- Phone recognition on TIMIT:
  - ✓ Features:13-dimensional MFCCs with first and second order
  - ✓ Output:1938 tied triphone

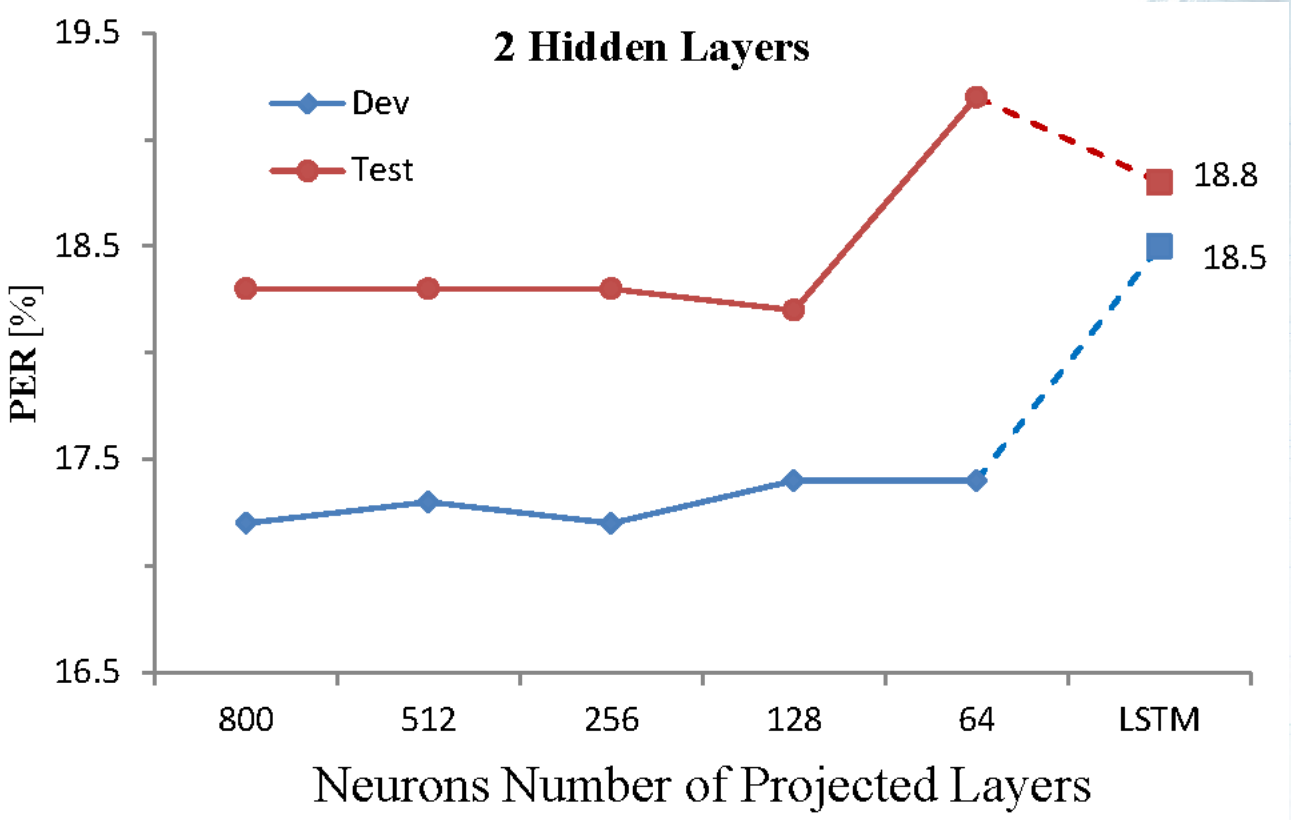
Experiments

- Phone recognition on TIMIT:
  - ✓ Features:13-dimensional MFCCs with first and second order
  - ✓ Out

Model	Parameters	Dev	Test
GRU-1L	4.03M	19.2	**20.4**
GRU-1L	7.41M	17.3	**18.4**
GRU-1L	11.25M	17.1	**17.9**
LSTM-1L	4.24M	20.2	21.5
LSTM-2L	9.37M	18.5	18.8
LSTM-3L	14.50M	17.9	18.5
LSTMP-1L	3.17M	18.4	20.5
LSTMP-2L	6.86M	17.3	**18.3**
LSTMP-3L	10.56M	16.8	**17.8**
DNN-1024-6L	7.68M	18.9	**20.4**

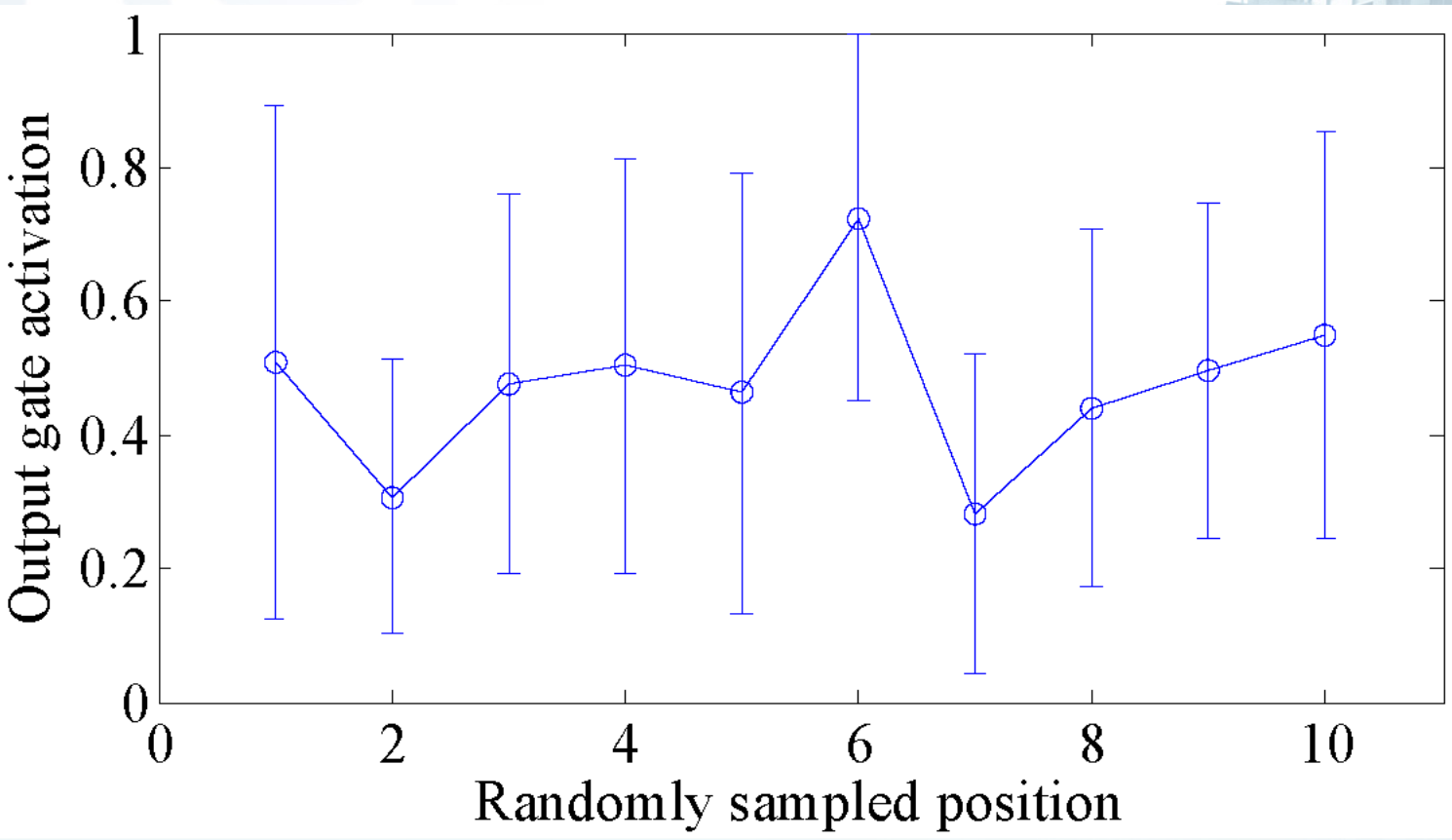
Experiments

- Analysis:
  - ✓ investigate the performance of LSTMP with different number of neurons in the projection layer



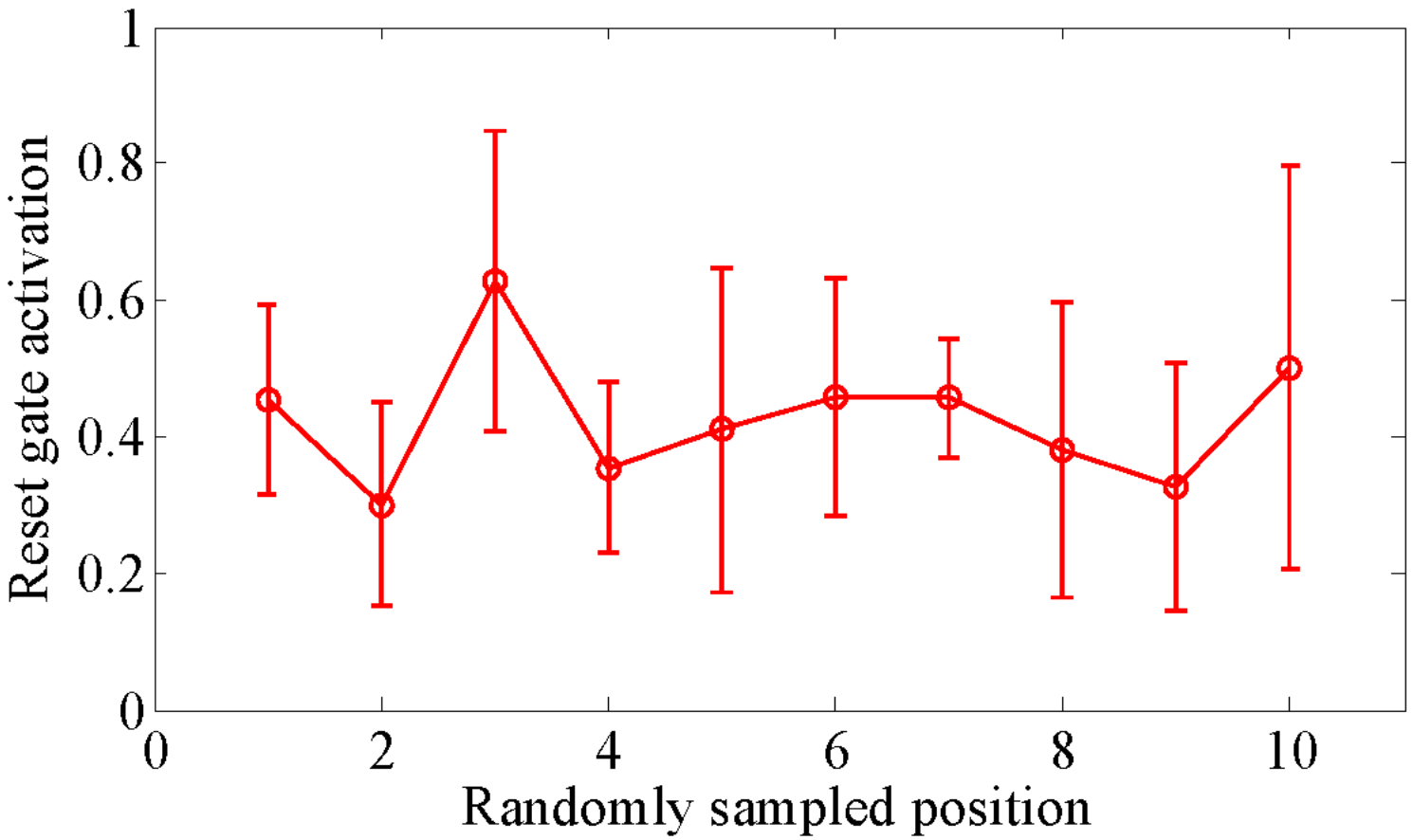
Experiments

- Analysis:
  - ✓ statistics on activations of output gate in LSTM



Experiments

- Analysis:
  - ✓ analysis on activations of the reset gate in GRU



Experiments

- Further exploring on different datasets
  - ✓ Input: 14-dim PLP with first and second order derivatives
  - ✓ Trained on ASGD platform with four GPU cards
  - ✓ First tasks: 130h speech transcription with 16KHz sample rate

Model	Parameters	Test
GRU-2L	9.97M	**20.6**
LSTM-2L	11.93M	21.2
LSTMP-2L	8.51M	**20.5**
DNN-2048-5L	28.22M	23.15

Experiments

- Further exploring on different datasets
  - ✓ 800h telephone conversation speech recognition with 8KHz sample rate
  - ✓ Test1 and test2 corpus are all recorded at the real scene and the major difference between them is the corpus scale.
  - ✓ Test3--real telephone conversation speech(noise is louder)

Model	Test1	Test2	Test3
GRU-2L	**24.84**	**20.18**	**32.17**
LSTM-2L	26.91	22.51	34.34
LSTMP-2L	25.39	20.76	32.27
DNN-2048-5L	26.39	21.56	34.74

Conclusion

- ✓ GRU RNNs outperform LSTM RNNs and are similar to LSTMP architecture
- ✓ The good performance of GRU and LSTMP mainly profit from the ability to select useful information from history memory content
- ✓ weights sharing in LSTMP contributes to the superior performance than LSTM as it leads to a better generalization characteristic

Thank you !