INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Analysis on Gated Recurrent Unit based Question Detection Approach

Yaodong Tang[1], *Zhiyong Wu*[1,2,3], *Helen Meng*[1,3], *Mingxing Xu*[1,2], *Lianhong Cai*[1,2]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2] Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

tangyd14@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk, {xumx, clh-dcs}@tsinghua.edu.cn

Abstract

Recent studies have shown various kinds of recurrent neural networks (RNNs) are becoming powerful sequence models in speech related applications. Our previous work in detecting questions of Mandarin speech presents that gated recurrent unit (GRU) based RNN can achieve significantly better results. In this paper, we try to open the black box to find the correlations between inner architecture of GRU and phonetic features of question sentences. We find that both update gate and reset gate in GRU blocks react when people begin to pronounce a word. According to the reactions, experiments are conducted to show the behavior of GRU based question detection approach on three important factors, including keywords or special structure of questions, final particles and interrogative intonation. We also observe that update gate and reset gate don't collaborate well on our dataset. Based on the asynchronous acts of update gate and reset gate in GRU, we adapt the structure of GRU block to our dataset and get further performance improvement in question detection task.

Index Terms: question detection, recurrent neural network (RNN), gated recurrent unit (GRU), question features,

1. Introduction

Recurrent neural networks (RNNs) have been widely used in speech related applications, such as statistical parametric speech synthesis [1][2], speech emotion recognition [3][4] and speech recognition [5]. Due to the nature of capturing complex relationship among time series data, RNNs could represent the uncertainty and multi-modality in acoustic modelling [4][6]. Studies using RNNs on real-world applications usually regard this powerful tool as a black box, where comparatively little investigation has been conducted to find the relationship between network architecture and real-world data.

In [5], the outputs of various recurrent networks when classifying an excerpt from TIMIT dataset were visualized, so the difference of adding weighted duration error could be observed. Another work [7] focused on providing empirical exploration of the predictions from long short-term memory (LSTM) based RNNs and representations on character-level language modeling. In this work, the researchers depicted the outputs of LSTMs on real-world text data and found that memories over 100 characters could be kept in LSTM cell. We are heavily influenced by [8] in which the researchers visually analyzed LSTM in predicting a speech parametric sequence in statistical parametric speech synthesis (SPSS), where researcher also evaluated the components' importance in LSTM. The average activation of forget gates has a strong correspondence with the phoneme boundaries. Their experimental results also have revealed that the forget gate is the only critical component of the LSTM.

In recent studies, another type of recurrent unit referred as gated recurrent unit (GRU) was proposed in [9]. As an alternative structure, GRU is empirically evaluated in [10], where the evaluation result shows GRU tends to converge faster and be easier to train in most cases. Our previous work [11] on Mandarin question detection from acoustic features only using recurrent networks with GRU achieves significantly better results than conventional method. In question detection task, conventional approach use specific features designed by info in phonetics and linguistics. Lexical factors in Mandarin questions such as the final particles or interrogative keywords are explicit but incomplete, while acoustic factors such as interrogative intonation on phonetics are implicit but sufficient [12]. In this paper, we attempt to answer whether GRU could generate those special design features from consecutive acoustic-prosodic feature frames.

Inspired by the work in [7], we reach better understanding of recurrent networks by visually analyzing the activations of gating units. We first measure the importance of hidden nodes in single layer GRU network. An analysis on how reset gate's activation of the most crucial node relate to pronunciation is then conducted. To answer the question which factor is critical for the classification task, we visualize the activations of top three nodes with highest positive weights. The result gives a verification of previous work conclusion [13]: the important question factors such as sentences final particle are at the end of sentences. We pick several question sentences without final particle to figure out whether the recurrent network could model keywords and interrogative intonation from acoustic feature sequence. Considering the duplicated functionality across gating units, we propose single connection gated recurrent unit (SC-GRU). In our experiment, we get further performance improvement in SC-GRU.

Rest of the paper is organized as follows. The framework of our previous question detection approach and simplified model are described briefly in Section 2. Background of question

http://dx.doi.org/10.21437/Interspeech.2016-964

factors, analysis and experiment are presented in Section 3. Section 5 lays out the conclusion and future work.

2. Framework and models

In this section, we present the framework of question detection approach [11] and GRU based simplified models.

2.1. Framework

Figure 1 depicts the framework of our question detection approach. Feature sequence is extracted from original speech signal at utterance level. In spired by researches in speech emotion recognition, the low-level descriptor feature set for extraction is proposed in INTERSPEECH 2014 Computational Paralinguistic Challenge [14]. According to the nature of RNN, we consider the outputs $\mathbf{h}_T$ of last frame in a sequence from the hidden layer contain enough contextual information for final classification. The label of the sequence could be calculated as:

$$l = round(\sigma(\mathbf{W}_o \cdot \mathbf{h}_T)) \quad (1)$$

where σ is a sigmoid function and $\mathbf{W}_o$ is the weight vector of collect layer.

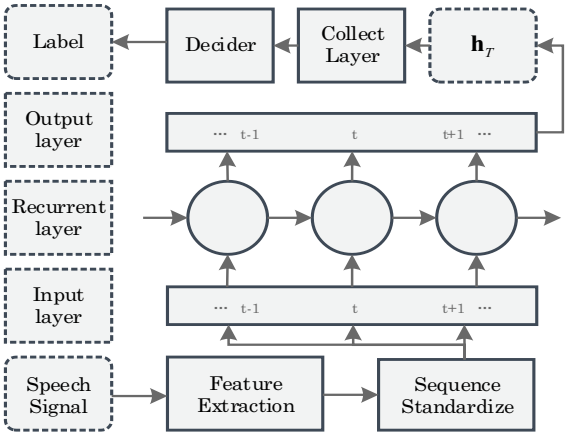

Figure 1: *Framework of question detection method [11]*

2.2. Models

In the hidden layer of our approach, we adopt GRU as the architecture of hidden unit for the empirical evaluation result that GRU tends to get lower loss in both train set and validation set than LSTM. As a follow-up work, the two simplified models are based on GRU.

2.2.1. Gated recurrent unit

GRU is able to make recurrent blocks capture the dependencies of different time scales. Gating units in GRU could modulate the flow of information inside unit as in LSTM [10]. Without a separate memory cell, GRU unit doesn't need to use peep-hole connections. GRU could be formulated by:

$$z_t^j = \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1} + b_z)^j \quad (2)$$

$$r_t^j = \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1} + b_r)^j \quad (3)$$

$$\tilde{h}_t^j = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \otimes \mathbf{h}_{t-1}) + b)^j \quad (4)$$

$$h_t^j = z_t^j h_{t-1}^j + (1 - z_t^j)\tilde{h}_t^j \quad (5)$$

2.2.2. Simplified Gated Recurrent Unit

And considering the weight matrixes in equation (4), input gate shares similar functionality with the weight matrixes in equation (4), we proposed the simplified gated recurrent unit (S-GRU) by set r_t^j to 1. The S-GRU is calculated by:

$$z_t^j = \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1} + b_z)^j \quad (2)$$

$$\tilde{h}_t^j = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + b)^j \quad (4)$$

$$h_t^j = z_t^j h_{t-1}^j + (1 - z_t^j)\tilde{h}_t^j \quad (5)$$

It is necessary to be noticed the S-GRU has a similar structure of the simplified LSTM (S-LSTM) in [8], except the recurrent connection and the output activation.

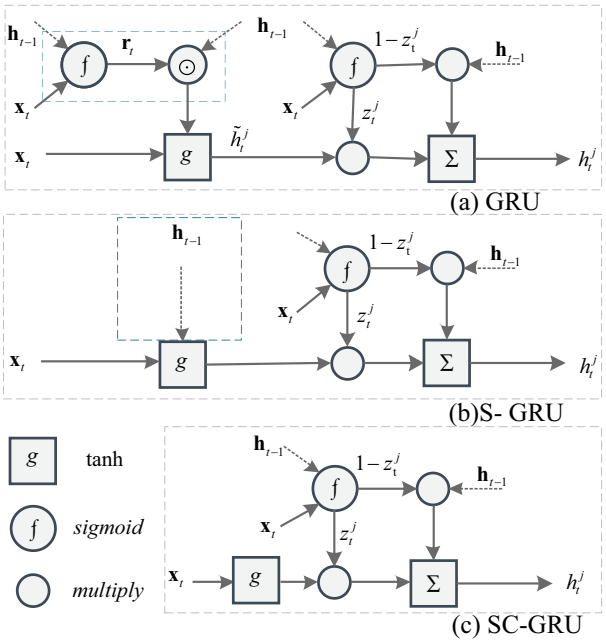

Figure 2: *Illustration of GRU (a), S-GRU (b) and SC-GRU(c). S-GRU is achieved by setting Reset Gate (selected by blue dashed line in (a)) to 1. SC-GRU is achieved by removing recursive connection (selected by blue dashed line in (b)) from hidden layers' outputs to input.*

Inspired by the idea of duplicated functionality in GRU's components, we cutoff the recurrent connection between its ， previous activation and the candidate updates, so we get single connection gated recurrent unit (SC-GRU). And SC-GRU is written as:

$$z_t^j = \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1} + b_z)^j \quad (6)$$

$$h_t^j = z_t^j h_{t-1}^j + (1 - z_t^j)\tanh(\mathbf{W}\mathbf{x}_t + b)^j \quad (7)$$

3. Analysis

In this section, we focus on analyzing the relationship between the outputs of gating units in GRU and the factors of Mandarin question utterance. Factors of questions in Mandarin speech are divided into two group: lexical factors and acoustic factors. We attempt to figure out how GRU network models those factors

from acoustic features by three tasks. A simple experiment is conducted to show the effectiveness of simplified units.

3.1. Background: question factors

Question in Chinese is very different from question in English and other Indo-European languages [13]. In Chinese questions, the word order of sentence is usually the same as the statements. For example, any statements could be converted into a yes-no questions by adding a final particle "吗(ma)". Acoustic and prosodic information are also helpful for question detection. It is reported that the variations of acoustic-prosodic features in the latter half of questions sentences called "boundary tone" are very important to transfer interrogative information [15].

3.1.1. Lexical factors

Lexical factors are considered as the most important part in conventional approach of question detection. Some questions have special structure such as A-or-B and A-not-A, which are two obvious factors for classifying questions. Some words in Chinese mainly appear in questions called interrogative adverbs and interrogative pronouns. Considering these structure factors and special words that mainly appear in questions, we get the lexical factors set of questions. Due to the explicit format in sentences, we call those lexical factors as "Keywords". While usually along with interrogative intonation, final particle such as "ma" is isolated from the "Keywords" factor set. The total lexical factors are listed in Table 1.

Table 1. *Lexical factors: keywords and final particle.*

Name	Type	Examples
Keywords	I. adverb	nan2dao4, mo4fei1, etc.
	I. pronoun	zen3me, shui2, na3, etc
	"A not A" construction	shi4bu4shi4, lai2bu4lai2, etc
	"A or B" construction	shi4ni3hai2shi4ta1, xiang4zuo3hai2shi4you4
Final particle	-	a, e, en, ne, ma, ba etc.

3.1.2. Interrogative intonation

Interrogative intonation is the intonation used in questions. A typical format of interrogative intonation is the rising boundary tone of Chinese interrogative sentences. Questions are formed by interrogative intonation with the final particle in most cases. But in questions without "Keywords" and final particles, the rising boundary tone is the only way to transfer interrogative information. To describe intonation from acoustic-prosodic features is as difficult as capturing emotions in speech emotion recognition.

3.2. Analysis setup

A simulated Call Center Recording of Mandarin is used as the analysis experiment dataset and evaluation experiment data set, involving 20 native speakers of Mandarin [12]. We use 2,850 question sentences (Q) and 2,850 non-question sentences (NQ) for our experiment. Four-fifth of them are adopted for training the network with 128 GRU units in the hidden layer. The rest of them are used for analysis materials.

Acoustic features are extracted using OpenSMILE [16] with the feature set and their first derivatives in Section 2.1. The network is implemented by Theano [17][18] and Keras [19]. We adopt Dropout [20] to reduce the impact of over-fitting problem. An optimization method called Adam [21] is used in the training stage.

3.3. Task 1: pronunciation

Observations about "Keywords" set are based on the hypothesis that a trained GRU could react to the pronunciation of a word from consecutive feature frames. We select 24 sentences (18 Q and 6 NQ) with different speakers and different factors.

Not all hidden nodes (units in hidden layer) could give an accurate response to the pronunciation for the variations in interrogative intonation. As equation (1), we use $\mathbf{W}_o$ to collect the high-level features extracted by hidden nodes, so the value of $\mathbf{W}_o$ denotes the importance of the nodes. Higher positive weight a node has, more important question factors it models.

Depicted in Figure 3, the outputs of gating unit in the node with highest positive weight are visualized along time axis. Green vertical lines is the start boundaries of words, by which followed Chinese phonetic alphabets at the top of each sub-graph. As we can see, the update gate tends to open (its output set to 0) when a new word begin, which means the node try to receive new information from current inputs. During the latter half of words pronunciation, node chooses to keep its memory by setting the update gate to 1 and activating the reset gate. This phenomenon is also observed in other samples.

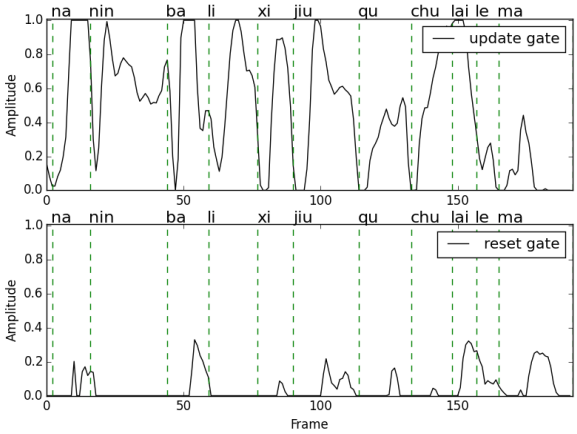

Figure 3: *Outputs of gating units in the node with highest positive weight when given a Chinese question.*

In Figure 3, the reset gate closes in most of time. In the latter half, reset gate cooperate with update gate to keep memories, where the reset gate duplicate the functionality of update gate and recurrent weight matrix. SS-GRU is proposed on this idea to show the critical basic structure of gated recurrent neural network. An evaluation experiment is conducted in task 4.

3.4. Task 2: final particle

As described before, the most important factors transferring interrogative information are usually at the end of the sentence. In question utterance, it is a common combination that final particle carries the rising boundary tone, which is a basic form of interrogative intonation.

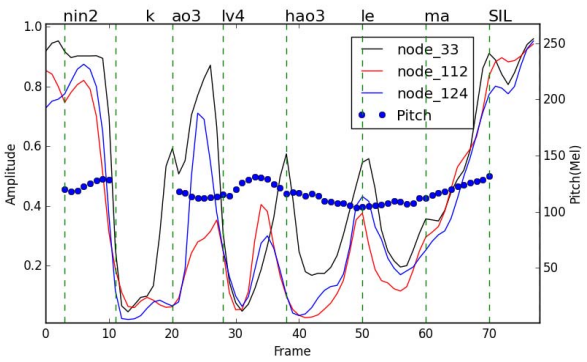

Figure 4: *Outputs of three nodes with highest weights. Pitch contour shows the rising boundary tone after "le".*

Activation contours and pitch contour are depicted at Figure 4. Similar to the reaction of update gate, the nodes' activations also have a strong response to pronunciation. The sample "nin kao lv hao le ma?" which means "Have you decided yet?" has final particle only. But as the rising pitch contour at the end, rising boundary tone comes to reinforce interrogative intonation too. Rising activations means all those nodes capture the changes. All the changes in final particle are observed in 8 selected interrogative sentences without keywords factors. As we can see, nodes in the hidden layer cooperate together to extract useful information and in other side, frames at the end of sentences supply the network with most useful information.

3.5. Task 3: keywords and interrogative intonation

It is really difficult to find which node in hidden layer extracts features of keywords or interrogative intonation. In task 3, we attempt to verify the functionality that our question detection approach could distinguish questions only by keywords or interrogative intonation. Due to the unbalanced distribution of questions, we pick 100 sentences without any final particle for verification. As the result shown in Table 2, our approach could classify sentences with interrogative intonation correctly but fail in distinguishing questions by keywords.

Table 2. *Test results in selected sentences. "O" indicates the sentences include features of the type. "I. Intonation" is "interrogative intonation".*

Type	Keywords	I. Intonation	Error/Total
NQ	O	-	0/12
	-	-	2/30
Q	O (words)	-	7/12
	O(structure)		6/8
	-	O	3/18
	O	O	3/20

3.6. Task 4: evaluation

The evaluation experiment using 5-fold cross validation is set to compare the performance of our proposed simplified gated neural networks with GRU and LSTM. All networks are implemented by Theano and Keras. We choose the size of each model to ensure that each model has about 37k parameters. The dataset and feature set are same in analysis setup. We use F1 measurement for evaluating classification performance.

Table 3. Evaluation of different recurrent neural networks. S-RNN is the standard recurrent unit.

Model	Number Hidden Units	Time(s) (per epoch)	F1 Measure (mean/std)
S-RNN	138	12	0.767/0.019
LSTM	52	29	0.796/0.017
GRU	64	22	0.801/0.015
S-GRU	86	17	0.804/0.016
SC-GRU	102	15	**0.805**/0.015

As the result shown in Table 3, we get further improvement in mean value of F1-measurement. Furthermore, the SC-GRU based model reduces consuming time than LSTM.

4. Conclusions

Compared to conventional question detection approach, our method use GRU based RNNs for detecting questions achieves significant better result. In this paper, we follow the work in [7] to answer whether question factors in phonetics and linguistics are modelled by the recurrent neural network.

Question factors could be divided into two groups: lexical factors (keywords, special structure and final particle) and acoustic factors (interrogative intonation). Results of analysis reveal that the recurrent network could generate response to the changes in feature frames when a new word begins, based on which we design test set to evaluate the network's capability of modelling lexical factors and acoustic factors. It is shown that recurrent network regards the final particle and interrogative intonation as the most important factor for our task, but it fails in extracting keywords information. Duplicated functionality of GRU's reset gate leads us to simplify the network architecture. Experiment shows the advantage of the proposed SC-GRU model: similar performance and less time consume. There is still a lot work to do to answer how the gating units cooperate and how to model keywords information.

5. Acknowledgements

This work is supported by the National Basic Research Program of China (2012CB316401), the National High Technology Research and Development Program of China (NHTRDPC) (2015AA016305), the National Natural Science Foundation of China (NSFC) (61375027, 61433018, 61370023, 61171116), and the joint fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK404/15) and the Major Program for National Social Science Foundation of China (13&ZD189).

6. References

[1] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," in *Proc. Interspeech*, 2014.

[2] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*, 2014.

[3] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proc. the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015.

[4] S. Chen, Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proc. the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015.
[5] A. Graves, Supervised sequence labelling with recurrent neural networks, *Heidelberg: Springer*, 2012.
[6] Sohl-Dickstein, Jascha, and D. P. Kingma, "Technical note on equivalence between recurrent neural network time series models and variational bayesian models," in *arXiv preprint, arXiv: 1504.08025*, 2015.
[7] A. Karpathy, J. Johnson, F. F. Li, "Visualizing and understanding recurrent networks," in *arXiv preprint, arXiv: 1506.02078*, 2015.
[8] Z. Wu, S. King, "Investigating gated recurrent networks for speech synthesis," in *Proc. International Conference on Acoustics, Speech and Signal Processing, 2016.*
[9] K. Cho, B. V. Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: encoder-decoder approaches," in *arXiv preprint arXiv:1409.1259*, 2014.
[10] J. Chung, C. Gulcehre, Cho K H, and Y.Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *arXiv preprint arXiv:1412.3555*, 2014.
[11] Y. Tang, Y. Huang, Z. Wu, H. Meng, M. Xu, L. Cai, "Question detection from acoustic features using recurrent neural network with gated recurrent unit," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.
[12] A. Li, M. Xu, L. Cai, "Acoustic features prominence based Chinese question detection," in *Chinese Sciencepaper*, 2014.
[13] J. Yuan, D. Jurafsky, "Detection of questions in Chinese conversational speech," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, pp. 47-52, 2005.
[14] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, Y. Zhang, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive and physical load," in *Proc. Interspeech*, 2014.
[15] Y. Wang, J. Jia, L. Cai, "Analysis of Chinese interrogative intonation and its synthesis in HMM-based synthesis system," in *Proc. Internet Computing & Information Services*, 2011.
[16] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. International Conference on Multimedia*, 2010.
[17] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley and Y. Bengio. "Theano: new features and speed improvements," in *Proc. NIPS 2012 deep learning workshop*, 2012.
[18] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley and Y. Bengio, "Theano: A CPU and GPU Math Expression Compiler," in *Proc. the Python for Scientific Computing Conference (SciPy)*, 2010.
[19] F. Chollet, Keras [OL]. [2016-03-18]. *GitHub repository*, https://github.com/fchollet/keras.
[20] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," in *the Journal of Machine Learning Research*, 2014
[21] D. Kingma, J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference for Learning Representations (ICLR)*, 2014.