Seq2Tree: A Tree-Structured Extension of LSTM Network

Weicheng Ma
Computer Science Department,
Boston University
111 Cummington Mall, Boston, MA
wcma@bu.edu

Zhaoheng Ni
Graduate Center,
City University of New York
365 5th Ave., New York, NY
zni@gradcenter.cuny.edu

Kai Cao
Cambia Health Solutions
kai.cao@cambiahealth.com

Xiang Li
Cambia Health Solutions
xiang.li@cambiahealth.com

Sang Chin
Computer Science Department,
Boston University
111 Cummington Mall, Boston, MA
spchin@cs.bu.edu

Abstract

Long Short-Term Memory network(LSTM) has attracted much attention on sequence modeling tasks, because of its ability to preserve longer term information in a sequence, compared to ordinary Recurrent Neural Networks(RNN's). The basic LSTM structure assumes a chain structure of the input sequence. However, audio streams often show a trend of combining phonemes into meaningful units, which could be words in speech processing task, or a certain type of noise in signal and noise separation task. We introduce Seq2Tree network, a modification of the LSTM network which constructs a tree structure from an input sequence. Experiments show that Seq2Tree network outperforms the state-of-the-art Bidirectional LSTM(BLSTM) model on the signal and noise separation task, namely CHiME Speech Separation and Recognition Challenge.

1 Introduction

Recent RNN based approaches are achieving high performance in speech processing tasks, including but are not limited to signal and noise separation task (Erdogan et al. (2015); Zhu and Vogel-Heuser (2014); Wu et al. (2015); Barker et al. (2015)). The underlying hypothesis is that the energy in each frequency bin over a period of time is continuous and predictable. However, in real life scenes noises can break in at any time and intertwine with the sound signal with no predictable pattern, which undermines these models' ability in predicting the distribution of noise over frequency bins.

To address the problem of finding correct boundaries of noises, some variants of the original LSTM network are used. The current state-of-the-art system on this task applies BLSTM, which tries to bound noises by foreseeing future information (Erdogan et al. (2015); Weninger et al. (2015); Grais et al. (2014)). Nevertheless information from the future contains also further away sound signals, which does not solve the signal superposition problem. Furthermore, we believe the future for speech processing should be dominated by real-time speech processing techniques, which BLSTM models are not able to handle.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

What's more important, phonemes in sound signals make no sense if not combined to be "words", neither are those in noise signals. So the sound waveforms should not be understood as a chain of phonemes but on "word" level. This leads to a natural choice of tree structured modeling of the waveforms.

In this paper we introduce a novel architecture which extends LSTM networks to be able to parse sequential input into tree structure and show its superiority in decomposing sound and noise signals. We call our new RNN architecture Seq2Tree. Our architecture differs from the standard LSTM since each node inherits the hidden state not from the previous state in time sequence but from its parent in the tree structure, based on its position predicted by the network itself.

Our evaluation demonstrates the advancement of our Seq2Tree network compared with the BLSTM baseline on the signal and noise separation task (Barker et al. (2015)). Experiments shows that our system shows comparable performance to the BLSTM implementation, while outperforming it in more complex scenarios. Further optimization and adjustment to this task will follow.

2 LSTM Network

RNN's have the advantage of processing input sequences regardless of their lengths. The sequence and elements can be of arbitrary types, for example phonemes in a piece of sound, when it comes to the task of speech processing. However RNN's are easily trapped by the explosively growth or rapid vanishment of the gradient over long distances (Hochreiter (1998); Hochreiter et al. (2001)). This makes it difficult for RNN's to represent long-term information.

LSTM network is introduced to deal with this problem (Hochreiter (1998); Hochreiter et al. (2001); Zaremba and Sutskever (2014); Zaremba et al. (2014)). Different from directly passing the previous state and the current input to the transition function on which the gradient is calculated, LSTM uses a memory cell to preserve the longer-term information. Using the settings in (Zaremba and Sutskever (2014); Zaremba et al. (2014)), the LSTM transition functions are as follows:

$$\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}, \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}, \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}, \\
u_t &= tanh(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}, \\
c_t &= i_t \odot u_t + f_t \odot c_{t-1}, \\
h_t &= u_t \odot tanh(c_t)
\end{aligned} \tag{1}$$

where i_t, f_t, o_t, c_t are the input gate, forget gate, output gate and the memory cell respectively, and $\odot$ refers to element-wise multiplication. As is shown in the equations, the input gate decides how much information from the new input will be added to the memory cell. Similarly, the forget gate f controls how much information to forget from the previous states, and the output gate limits the amount of information to expose. By balancing the incoming and outgoing information amount, LSTM is able to prevent the gradient vanishment and explosion problems.

Ordinary LSTM is based on chain structured sequences. There exists two common variants of LSTM networks on structure, namely BLSTM and Multilayer LSTM, which combines multiple LSTM networks together to provide additional information in the prediction at each time step. Tree LSTM (Tai et al. (2015)) could be regarded as one variation of Multilayer LSTM with the dependency relation reversed.

3 Seq2Tree Network

The LSTM architectures described in the previous section all have limits in discovering tree structure from sequential input. Though Multilayer LSTM and Tree LSTM networks are able to maintain multilevel dependencies, Multilayer LSTM exposes children cells to all the other units, and Tree LSTM requires tree structured input. These characteristics limit their use in speech processing tasks where no reliable parser exists, especially in the case of online speech processing. Here we propose two variants to LSTM network - Single Level Seq2Tree and Multilayer Seq2Tree architectures. Both

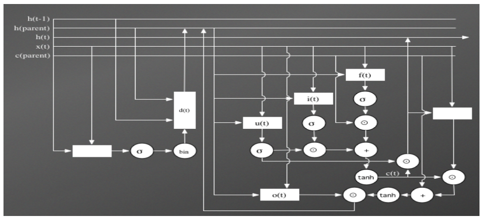

Figure 1: Seq2Tree Network Architecture

variants are able to find dependencies from adjacent signals, while Multilayer Seq2Tree architecture catches deeper, weaklier bounded correlations.

The most significant difference between original LSTM and Seq2Tree model is that instead of taking the previous state as the preceding state, Seq2Tree networks use one additional direction gate d_t to choose the direction to go at time step t. The path selection gate is implemented differently in Single Level Seq2Tree and Multilayer Seq2Tree architectures.

When the network gets a new input at each step, there exists 3 possible directions to go: up, right and down. Before moving, the previous hidden state h_{t-1} is added to the parent state list and is set to be active. The strategy is that if the predicted direction to go is “up", the parent node of the current active state becomes the parent state. If the direction is “down" then the network chooses h_{t-1} as the parent node. Otherwise the new unit becomes the child of the parent state of the current active node. Since signal patterns can overlap with each other, multiple jumps towards the root in one time step is also allowed. Moreover, at each jump an update gate is used to control the amount of change on the parent layer, and the remainder is passed to even higher states in the tree if there are further jumps.

After processing each state its parent node’s information is updated. The forget gate of the child state f_t controls the amount of change to give its parent state. This mechanism is inspired by the Tree LSTM networks. The transition functions of the Seq2Tree network are as follows:

$$\begin{aligned}
&d_t = \sigma(W^{(d)}x_t + U^{(d)}h_{t-1} + b^{(d)}),\\
&h_{parent} = d_t(h_{parent} \cup h_{t-1}),\\
&i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{parent} + b^{(i)}),\\
&f_t = \sigma(W^{(f)}x_t + U^{(f)}(d_t(h_{parent})) + b^{(f)}),\\
&o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{parent} + b^{(o)}),\\
&u_t = tanh(W^{(u)}x_t + U^{(u)}h_{parent} + b^{(u)}),\\
&c_t = i_t \odot u_t + f_t \odot c_{parent},\\
&h_t = u_t \odot tanh(c_t),\\
&\Delta f_t = \sigma(W^{(f)}x_t + U^{(f)}h_t + b^{(f)}),\\
&c_{parent} = c_{parent} + \Delta f_t \odot c_t,\\
&h_{parent} = o_{parent} \odot tanh(c_{parent}).
\end{aligned} \tag{2}$$

where d_t is a selection gate which chooses an h from the recorded h_{parent} with the largest d_t gate value, σ represents sigmoid function, and $\odot$ means element-wise multiplication.

4 Task and Model

4.1 Signal and Noise Separation Task

We test our Seq2Tree architecture on the signal and noise separation task, the goal of which is to predict a mask which weakens the energy of noise when applied to the input sound. The task is defined in the Second CHiME Speech Separation and Recognition Challenge (Vincent et al. (2013)).

4.2 Tree2Seq Signal Masking Model

For this task, at each time step t we want to predict a mask over all frequency bins. We achieve this by training a softmax regression matrix which takes the current hidden state:

$$mask = softmax(U^{(R)}h)$$

where $U^{(R)}$ is the regression matrix.

We train our signal masking model in two stages using two different loss calculations, as is suggested in Weninger et al. (2014); Erdogan et al. (2015). The two losses we applied are:

$$J_1(t) = -\frac{1}{c}\sum_{i=1}^{c}(mask_i - label_i)^2$$

$$J_2(t) = -\frac{1}{c}\sum_{i=1}^{c}(\|x_t\|(mask_i - label_i))^2$$

where c is the number of frequency bins, $mask_i$ is the predicted mask at time t for bin i and $label_i$ is the labeled mask on bin i at time t.

5 Experiments

We evaluate our Seq2Tree architecture on the signal and noise separation task. The data is a fraction of 1500 audio files from CHiME dataset (Vincent et al. (2016)), in which 10% is used for test and the rest for training. Each input file is Fourier Transformed and fed to the models. Every model predicts a mask given the input matrix. The quality of the mask is evaluated in terms of Overall Perceptual Score(OPS) by applying the mask onto the source waveform, given the noise-removed audio gold standard (Emiya et al. (2011)). In our experiment the shape of training data is 50×513, representing the energy at 50 time steps in 513 frequency bins. The test data has variable length over time steps, taking advantage of LSTM models' ability to deal with variable length inputs.

We compare the results generated by our Seq2Tree model with those output by the BLSTM baseline. The hidden layer size for our Seq2Tree network is set to 256, and we list the results with different numbers of iterations. The BLSTM baseline applies also 256 hidden layer size. Both models are trained for 100 epochs. Best and worst scores of our model are also included.

Implementation	OPS(dB)
BLSTM	25.01
Seq2Tree	24.41
Seq2Tree(Worst Case)	24.23
Seq2Tree(Best Case)	25.75

Table 1: Evaluation Results.

As is shown in the results table, our Seq2Tree model has comparable performance as the BLSTM implementation. More importantly, when looked into the specific wav files, in more complex cases where noises overlap with each other, the Seq2Tree model largely outperformed the BLSTM models, which agrees with our estimation. The Seq2Tree model shows advantage in both stability and the ability to deal with more complex situations. We could foresee the growth in performance of our Seq2Tree architecture if it is trained thoroughly, and when more challenging cases are imported. Further experiments are on going to demonstrate the effectiveness of the Seq2Tree architecture on the noise separation task and many other fields.

6 Conclusion

In this paper, we introduced a tree-structured LSTM network architecture. The Seq2Tree architecture can be applied to arbitrary sequential input with potential local dependencies among nodes. We demonstrated its effectiveness by evaluating a Seq2Tree based model on the signal and noise separation task. Experiments show that our Seq2Tree model outperforms the baseline since it correctly represented a large portion of the CHiME data. Further evaluations will be done to prove the correctness and advancement of our Seq2Tree architecture in more general cases.

References

Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. 2015. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 504–511.

Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann. 2011. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 7 (2011), 2046–2057.

Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. 2015. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 708–712.

Emad M Grais, Mehmet Umut Sen, and Hakan Erdogan. 2014. Deep neural networks for single channel source separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 3734–3738.

Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, 02 (1998), 107–116.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. (2001).

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075* (2015).

Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni. 2013. The second ?CHiME?speech separation and recognition challenge: An overview of challenge systems and outcomes. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 162–167.

Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. 2016. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language* (2016).

Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. 2015. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 91–99.

Felix Weninger, John R Hershey, Jonathan Le Roux, and Björn Schuller. 2014. Discriminatively trained recurrent neural networks for single-channel speech separation. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 577–581.

Minshun Wu, Zhiqiang Liu, Li Xu, and Degang Chen. 2015. Accurate and cost-effective technique for jitter and noise separation based on single-frequency measurement. *Electronics Letters* 52, 2 (2015), 106–107.

Wojciech Zaremba and Ilya Sutskever. 2014. Learning to execute. *arXiv preprint arXiv:1410.4615* (2014).

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* (2014).

Kunpeng Zhu and Birgit Vogel-Heuser. 2014. Sparse representation and its applications in micro-milling condition monitoring: noise separation and tool condition monitoring. *The International Journal of Advanced Manufacturing Technology* 70, 1-4 (2014), 185–199.