3E4-2in2

LSTM, GRUの動作性能評価

On the Evaluation of LSTMand GRU

浅川 伸一
Shin Asakawa

東京女子大学
Tokyo Woman's Christian University

Despite the popularity of RNNs such as LSTM and GRU, there was little evidence that confirmed their performance. We tried to investigate to reveal underlying processes and conditions to deal with such of them. The meanings of multi-layered RNNs might also be possible based on these findings as well.

1. Introduction

Long Short-Term Memory (LSTM)[Hochreiter 97], have recently emerged as an effective model in a wide variety of applications that involve sequential data. These include language modeling [Mikolov 10], handwriting recognition and generation [Graves 13a, Graves 13b, Graves 12], machine translation [Sutskever 14, Bahdanau 15], speech recognition [Graves 13b], video analysis [Donahue 15] and image captioning [Vinyals 15, Karpathy 15a].

A few recent ablation studies analyzed the effects on performance as various gates and connections are removed Greff et al [Greff 15]; Chung et al. [Chung 14]. However, while this analysis illuminates the performance-critical pieces of the architecture, it is still limited to examining the effects only on the global level of the final test set perplexity alone. Similarly, an often cited advantage of the LSTMarchitecture is that it can store and retrieve information over long time scales using its gating mechanisms, and this ability has been carefully studied in toy settings Hochreiter & Schmidhuber [Hochreiter 97]. However, it is not immediately clear that similar mechanisms can be effectively discovered and utilized by these networks in real-world data, and with the common use of simple stochastic gradient descent and truncated backpropagation through time.

2. Overview of LSTM

As depicted at Fig. 1, LSTMcan be described as the input signals $\boldsymbol{x}_t$ at time t, the output signals $\boldsymbol{o}_t$, the forget gate $\boldsymbol{f}_t$, and the output signal $\boldsymbol{y}_t$, the memory cell $\boldsymbol{c}_t$, then we can get the following:

$$\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}y_{t-1} + b_i\right), & (1)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}y_{t-1} + b_f\right), & (2)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}y_{t-1} + b_o\right), & (3)\\
g_t &= \phi\left(W_{xc}x_t + W_{hc}y_{t-1} + b_c\right), & (4)\\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t, & (5)\\
h_t &= o_t \odot \phi\left(c_t\right) & (6)
\end{aligned}$$

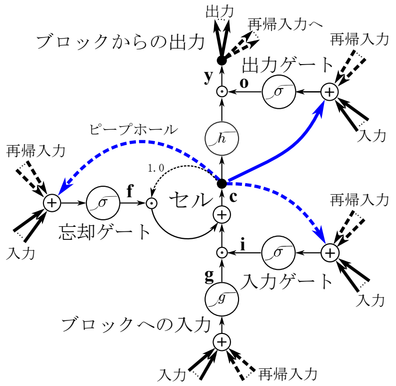

図 1: Shematic Description of LSTM

3. Model proposed

Fig. 2 indicates a summarized schema of models proposed. The softmax gate model (left) can be regarded as a competiton among inputs. A LSTM cell has three diffrent kinds of inputs shown in Figure. 1. Those are feedforward, recurrent, and peephole inputs. Since these inputs orginate from different sources, they might play different roles and/or tendencies on the behavior of the LSTMcell. Therefore, the softmax gate model would behave as a rectifier among them.

The cascaded gates model, on the other hand, can be regraded as a coopeartion among gates. The output functions followed by these gates would be posutulated as sigmoid functions, $0 \leq \sigma(x) \leq 1$, these gates would behave like an OR–logic gate. When one of gates was closed, the total output would be inhibited. The cascaded gates model also have an advantage that there is no additonal parameters.

When we tried to compare these models with LSTM, GRU, and Recurrent Neural Networks, we could evaluate roles of gates or different types of inputs. This would be possible for us to evaluate roles of diffrent inputs adequately.

4. Experiment

Karpathy and Fei-Fei Li [Karpathy 15b] chose to use Leo Tolstoy's War and Peace (WP) novel, which consists of

連絡先: 浅川伸一，東京女子大学，167–8585 東京都杉並区善福寺 2–6–1, 03–5382–6746, asakawa@ieee.org

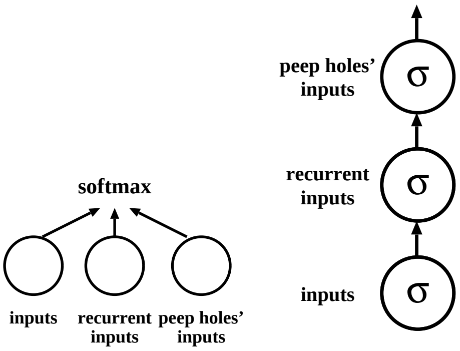

図 2: models proposed. A softmax model(left) and a cascaded gates model(right)

3,258,246 characters of almost entirely English text with minimal markup, and at the other end of the spectrum the source code of the Linux Kernel (LK).

Karpathy and Li first trained several Recurrent Neural Networksmodels to support further analysis and to compare their performance in a controlled setting. In particular, they trained models in the cross product of type (LSTM/Recurrent Neural Networks/GRU), number of layers (1/2/3), number of parameters (4 settings), and both datasets (WP/KL). For a 1-layer LSTMthey used hidden size vectors of 64,128,256, and 512 cells, which with their character vocabulary sizes translates to approximately 50 K, 130 K, 400 K, and 1.3 M parameters respectively. The sizes of hidden layers of the other models were carefully chosen so that the total number of parameters in each case is as close as possible to these 4 settings. We could follow their settings as well. Further analysis would be required.

参考文献

[Bahdanau 15] Bahdanau, D., Cho, K., and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, in Bengio, Y. and LeCun, Y. eds., *Proceedings in the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA (2015)

[Chung 14] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, *arXiv* (2014)

[Donahue 15] Donahue, J., Hendricks, , Anne, L., Guadarrama, S., Rohrbach, M., , S. V., Saenko, K., and Darrell, T.: Long-term Recurrent Convolutional Networks for Visual Recognition and Description, in *Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE Conference* (2015)

[Graves 12] Graves, A.: *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, Berlin, Germany (2012)

[Graves 13a] Graves, A.: Generating Sequences With Recurrent Neural Networks, *arXiv* (2013)

[Graves 13b] Graves, A., Mohamed, Rahman A., and Hinton, G.: Speech recognition with deep recurrent neural networks, in Ward, R. K. ed., *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649, Vancouver, BC, Canada (2013)

[Greff 15] Greff, K., Srivastava, R. K., Koutn´k, J., Steunebrink, B. R., and Schmidhuber, J.: LSTM: A Search Space Odyssey, *arXiv* (2015)

[Hochreiter 97] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol. 9, pp. 1735–1780 (1997)

[Karpathy 15a] Karpathy, A. and Fei-Fei, L.: Deep Visual-Semantic Alignments for Generating Image Descriptions, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA (2015)

[Karpathy 15b] Karpathy, A., Johnson, J., and Fei-Fei, L.: Visualizing and Understanding Recurrent Networks, *arXiv* (2015)

[Mikolov 10] Mikolov, T., Karafiát, M., Burget, L., Černocký, J. H., and Khudanpur, S.: Recurrent Neural Network Based Language Model, in Kobayashi, T., Hirose, K., and Nakamura, S. eds., *Proceedings of INTERSPEECH2010*, pp. 1045–1048, Makuhari, JAPAN (2010)

[Sutskever 14] Sutskever, I., Vinyals, O., and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, in Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. eds., *Advances in Neural Information Processing Systems (NIPS)*, pp. 3104–3112, Montreal, BC, Canada (2014)

[Vinyals 15] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D.: Show and Tell: A Neural Image Caption Generator, in *Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA (2015)