Recurrent neural network regularization

Wojciech Zaremba WOJ.ZAREMBA@GMAIL.COM
Google & New York University
Ilya Sutskever ILYASU@GOOGLE.COM
Google
Oriol Vinyals VINYALS@GOOGLE.COM
Google

Abstract

We present a simple regularization technique of recurrent neural networks (RNNs) with long short term memory (LSTM) units. This technique is based on dropout and gives tremendous performance boost. We show that it is beneficial in variety sequence modelling problems like language modeling, speech recognition, and machine translation.

1. Introduction

Recurrent neural networks yields the state-of-the-art performance on many sequence modelling tasks like language modelling, and speech recognition. Moreover, recent results in machine translation (Cho et al., 2014) shows their potential use in this field as well. However, up today there was no good techniques to regularize them. Various attempts to inject noise or mask some of activations (dropout) were giving small improvement in compare to model averaging. This work presents how to augment LSTMs with dropout. Resulting models give a tremendous improvement in various domains.

2. Related work

(Hochreiter & Schmidhuber, 1997)

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} L \begin{pmatrix} \mathbf{D}(h_k^{l-1}) \\ h_{k-1}^{l} \end{pmatrix}$$

$$c_k^l = f \odot c_{k-1}^l + i \odot g$$

$$h_k^l = o \odot \tanh(c_k)$$

Figure 1. To describe dynamics of multilayer LSTM with dropout, we use multi-indexing. Lower indices correspond to dynamics over time, and upper indices correspond to dynamics over layers. For simplicity, we denote by L a linear transform with bias ($Wx + b$), and by D a dropout layer. $\odot$ is a element-wise multiplication. h_k^0 is a input word-vector, and h_k^L is used to predict y_k (L is a number of layers).

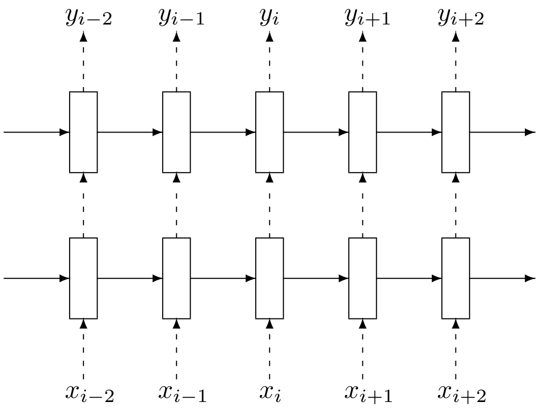

Figure 2. Regularized multilayer RNN. Dashed arrows indicate connections with applied dropout, while solid lines indicate connections where dropout is not applied.

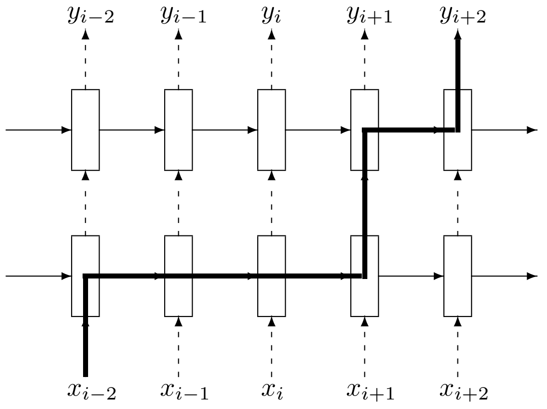

Figure 3. Thick line indicates an exemplary information flow in RNN. Information flow line is crossed L times, where L is depth of network.

3. Regularized RNN with LSTMs

3.1. Long-short term memory units

3.2. Regularization with dropout

3.3. Intuition

We will describe here justifications why putting dropout across recurrent layers helps, but within layers degrades performance.

Dropout removes part of information, and LSTMs has to become more robust while performing input-output mapping. Simultaneously we don't want to erase entire information. Especially we would like to facilitate memory about event that happened long time ago. Figure 3 shows flow of an exemplary information from the event x_{i-2} to the prediction in the step $i+2$. We can see that information is projected with dropout only L times, and it is independent of how far in past event occurred. All the previous regularization techniques were constraining recurrent connections, which effectively exponentially fast "blurred" information about the past.

4. Experiments

4.1. Language modeling

Our model is a two layer LSTM network, with 650 units initialized to uniformly in $[-0.05, 0.05]$. We apply 50% of dropout on non-recurrent connections. We train for 39 epochs, starting with learning rate 1, and after 6 epochs we decrease it by 1.2 in every epoch. We unroll RNN for 35 steps, and clip gradients at 5.

Model	Validation set	Test set
A single model		
Previous state-of-the-art [1]		107.5
Regularized LSTM	**86.2**	**82.7**
Model averaging		
Previous state-of-the-art [2]	83.5[3]	89.4
2 regularized LSTMs	80.6	77.0
5 regularized LSTMs		
10 regularized LSTMs		

Table 1. Word-level perplexity on Penn-tree-bank dataset.

4.2. Machine translation

4.3. Speech recognition

5. Discussion

References

Cho, Kyunghyun, van Merrienboer, Bart, Gulcehre, Caglar, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Mikolov, Tomáš. *Statistical language models based on neural networks*. PhD thesis, Ph. D. thesis, Brno University of Technology, 2012.

Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, and Bengio, Yoshua. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*, 2013.

[1] (Pascanu et al., 2013)

[2] (Mikolov, 2012)

[3] Weight of individual models are tuned to minimize this score. This few parameters are fit on this validation set, which is not completely fair.