Lab12 Pre-Report:
Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling

Jiho Park, 2019142056
School of Electrical and Electronic Engineering, Yonsei University

Abstract

The comparison between Gated Recurrent Unit(GRU), Long-Short-Term-Memory(LSTM) unit, and tanh unit of Recurrent Neural Network is the main subject of this paper. Junyoung at el [1], analyzed three architecture in the theoretical aspect and empirical aspect. The experiment was done with polyphonic music modeling task and speech signal modeling task. As a result, they found out that the advance recurrent units(GRU, LSTM) are much better than conventional RNNs, and GRU is comparable to LSTM.

1 Theoretical Apporach

1.1 Gated Recurrent Unit(GRU)

Cho et al first proposed Gated Recurrent Unit. This unit is designed to adaptively capture dependencies of different time scales. However unlike LSTM, it doesn't have an extra hidden state which is memory cell in LSTM.
Similar with LSTM, it's main hidden state h_t is updated by linear interpolation between h_{t-1} and new information $\tilde{h}_t$ (Eq 4). And the interpolation ratio z_t is computed by update gate, which is Eq 1. And with the reset gate r_t, conventional recurrent unit is computed as Eq 3. Lastly, as mentioned, this becomes updating information of hidden state.

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{1}$$

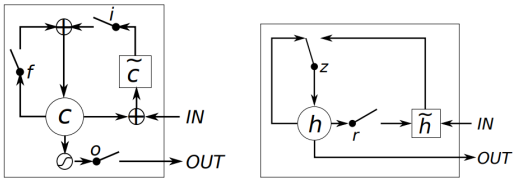
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{2}$$

$$\tilde{h} = \tanh(W \cdot [r_t * h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{4}$$

1.2 Comparison with LSTM

Shared feature: Additive component of hidden state
The hidden state(memory cell for lstm) update equation of both LSTM and GRU both have additive component. Unlike traditional recurrent units, the previous state is interpolated with new information, instead of replacing it. There are two advantages for this method. First is easing the hidden state to keep previous memories. Second is the preventing vanishing gradient, in other words, easing the training.
Unshared features and Necessity of empirical comparison
First, Controlled exposure of memory content is in LSTM, which is not in GRU. LSTM control the exposure by output gate. GRU fully exposes its content. Second, the input gate(reset gate for GRU) which produces new information is different (Fig 1).



(a) Long Short-Term Memory (b) Gated Recurrent Unit

Figure 1: LSTM and GRU Architecture

Necessity of empirical comparison
Even with these analyzed features, it is hard to make conclusion of which unit will perform better. This was the reason why Junyoung at el [1] conducted the experiment.

2 Empirical Apporach

They compared LSTM unit, GRU unit and tanh unit. The aim of the training is optimization of Eq 5.

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \log p(x_t^n | x_1^n, \ldots, x_{t-1}^n; \theta) \tag{5}$$

2.1 Experiment Settings

The polyphonic music modeling and speech signal modeling were experimented. For the polyphonic music modeling, they used Nottingham, JSB Chorales, MuseData and Piano-midi dataset. For the speech signal modeling, they used the two datasets from Ubisoft, which had sequence length of 500 and 8000 each.
They designed three models with relatively small, and approximately same number of parameters. This is to avoid the overfitting and to have moderate comparison.

2.2 Result Analysis and Conclusion

In case of LSTM and GRU, the loss convergence was often faster, and the final solutions were better than tanh-RNN model.
Also the final evaluation result in Fig 2 shows that the LSTM and GRU take the lead and GRU is comparable with LSTM(especially with the Ubisoft dataset).
As a result, the experiment results show the superiority of the gated units. This was more clear in more complex task(speech signal modeling). Although the GRU was comparable with LSTM in this experiment, it was hard to decide which unit is better among LSTM and GRU. Junyoung et al [1] concluded their paper saying that there need more experiment for comparison of two.

			tanh	GRU	LSTM
Music Datasets	Nottingham	train	3.22	2.79	3.08
		test	**3.13**	3.23	3.20
	JSB Chorales	train	8.82	6.94	8.15
		test	9.10	**8.54**	8.67
	MuseData	train	5.64	5.06	5.18
		test	6.23	**5.99**	6.23
	Piano-midi	train	5.64	4.93	6.49
		test	9.03	**8.82**	9.03
Ubisoft Datasets	Ubisoft dataset A	train	6.29	2.31	1.44
		test	6.44	3.59	**2.70**
	Ubisoft dataset B	train	7.61	0.38	0.80
		test	7.62	**0.88**	1.26

Figure 2: Experiment Results

Reference

[1] KyungHyun Cho Yoshua Bengio Junyoung Chung, Caglar Gulcehre. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Lab12 Pre-Report: Sequence to Sequence Learning with Neural Networks

Jiho Park, 2019142056
School of Electrical and Electronic Engineering, Yonsei University

Abstract

Although Deep Neural Network(DNN)s achieved superior performance on challenging tasks, it can't be directly used to map sequences to sequences. Ilya et al [1] proposed general approach of sequence to sequence mapping with recurrent units, Long-Short-Term-Memory(LSTM) units in their case. They mainly experimented one English to French translation task, using WMT'14 dataset. They achieved 34.8 BLEU score, which was outperforming phrase-based SMT system. Through experiments, they found out that LSTM was robust to long sentences, and learned sensible phrase and sentence representation. Finally they found out that training the reversed input improves the generation performance.

1 Model

Since DNN can only be applied to only vectors with fixed dimension, domain-independent method is required for sequence to sequence task. This can be addressed with recurrent neural network. Since LSTM can hold long-term dependency, two different LSTM networks were used in this paper. The first LSTM is for encoding, which reads the input and extracts the fixed-dimensional vector representation. The other LSTM uses the output of the first LSTM, and generates the sequence by feeding the output back into input of the next time step. This process is shown in Fig 1.

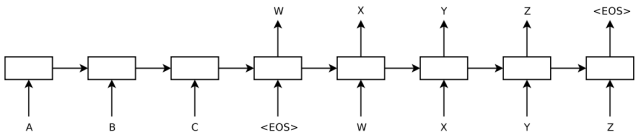


Figure 1: Sequence to sequence model architecture

1.1 Details

The training goal is to estimate the probability in Eq 1, where x_i is the input sequence, y_i is the output sequence and v is the encoded vector representation with fixed dimension of input sequence. The probability is computed by $softmax$ function. Ilya et al [1] chose four-layered LSTM for sufficient performance.

$$p(y_1,\ldots,y_{T'}|x_1,\ldots,x_T) = \prod_{t=1}^{T'} p(y_t|v,y_1,\ldots,y_{t-1}) \quad (1)$$

2 Experiments

Three models were evaluated on English to French translation task, comparing with phrase-based SMT system.

2.1 Dataset

12M sentences consisting of 348M French and 304M English words from WMT'14 English to French dataset were used. For token of words, 160K word token(vector)s and "UNK" token for out-of-vocab were used.

2.2 Decoding: Beam Search

While encoding, the input of the decoder were fixed, which is a teacher forcing. Therefore the training objective is simple summation of the log-likelihood of the target data. However while evaluation, the input of the decoding LSTM unit is the output of the previous time step.
They used beam search with hypothesis number B. This method chooses B hypotheses which has the highest probability. At next time step, each hypothesis makes another B assumption. They choose one hypothesis based on cumulative probability. After iterating this process, they choose only one hypothesis which have highest cumulative probability.

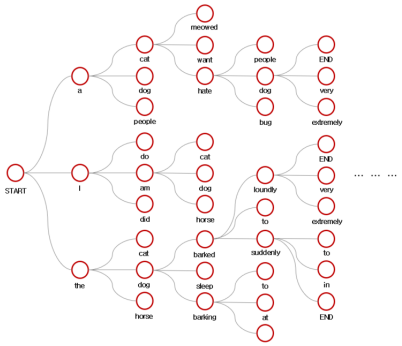


Figure 2: Visualization of Beam Search

2.3 Reversing Source

Ilya et al [1] experiment the model using the reversed-source for input. They found out that this method drops the perplexity of LSTM(5.8 to 4.7) but improves the test BLEU score(25.9 to 30.6). They assumed that this phenomenon is related to 'minimal time lag'. After reversing the source, the distance between the first few words of input and of output are very close. Therefore the 'minimal time lag' is greatly reduced and the backpropagation can more easily establish the communication between input and output. This performance improvement was also identical on long sentences.

3 Result and Conclusion

Their networks outperformed the baseline system both in short and long sentences. They found out the ensemble of 5 LSTM with beam size 2 is cheaper than single LSTM with beam size 12, which performed worse. Also, by visualizing the representation vector usin PCA, they found out that the network is sensitive to the order of the words. They concluded the paper saying that these results suggest that their approach will likely do well on other challenging sequence to sequence problems[1].

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Figure 3: Experiment Results

Reference

[1] Quoc V. Le Ilya Sutskever, Oriol Vinyals. Sequence to sequence learning with neural networks.