Review of Recurrent Neural Networks Language Models And Current Applications

Yutong Lin
yutong21@illiois.edu
School of Information
Information Management

Introduction

RNNLM, short for Recurrent Neural Network Language Model, is a variant of neural network and has been widely employed in natural language processing such as speech recognition, response generation and machine translation. However, researches over RNNLM did not become popular until around 2014. RNNLM then has developed into several variant models in terms of architecture of the neural network. This work aims to provide an overview of RNNLM basic structures and applications.

The model

The research on RNN models could date back to 1980s (P.Werbos, 1990), as a generalization of feedforward neural networks. In a basic RNN model, a cell at stage t in the network will receive both the input at t, denoted as x_t, as well as the stage from the previous cell, denoted as h_{t-1}. Therefore, the output h_t and y_t will be influenced by both the current input and the previous outputs. The outputs at each stage could be represented by the following equations:

$$h_t = \text{sigm}\left(W^{\text{hx}}x_t + W^{\text{hh}}h_{t-1}\right)$$
$$y_t = W^{\text{yh}}h_t$$

The output is a sequence ($\{y_1, y_2, \ldots y_n\}$). The model is especially useful in mapping a sequence to another sequence, which is close to many natural language applications. However, several problems arise regarding to this basic model. For example, the influence of previous cell outputs will decrease as the distance grows. LSTM, short for Long Short-Term Memory model was introduced as a variation of RNN to improve this (Hochreiter S, Schmidhuber J. 1997). In a LSTM model, inputs of a cell will be computed by four gates: the input gate i_t, forget gate f_t, output gate o_t and cell activation vectors c_t, based on the previous cell activation vector c_{t-1}, the cell output activation vector m_{t-1} and the input x_t at the current stage t. Then, the new cell activation vector c_t, cell output activation vector m_t and output y_t will be computed based on the four states. In this model, the cell activation vector c will change rather slowly, so the influence of previous cells will remain strong in the later cells. A new variation, i.e. GRU, Gate Recurrent Unit has been proposed in 2014 (Cho et al. 2014). This model is similar to LSTM, but it has fewer gates and parameters. This model greatly improved the computing ability and time efficiency of RNN models.

Such neural networks have been used to learn the distribution of word sequences, i.e. build the language models. RNN language models usually have an encoder-decoder structure. The encoder and the decoder are two separated neural networks. The encoder reads the input sequence (sentence) and produces a vector that represents it, then the decoder takes in the vector and generates the output sequence. In practice, the encoder and the decoder may be of different types of neural networks.

One problem remains with the Seq2Seq RNNLM is that in reality, the part of an input sequence that has the strongest influence on the output sequence will vary on different part of an output sequence. A mechanism in model named attention mechanism, was introduced to address this (Bahdanau, Cho, and Bengio 2014). The ouput word y_t from the decoder cell takes into consideration of a context vector that is calculated from the weighted average of all

hidden states in the encoder. By manipulating the weighing, different part of response should focus on different part of a message.

Applications and Performances

Recurrent Neural Networks Language Models have been widely adopted in machine translation, speech recognition, natural response generation etc.

Cho et al. 2014 proposed an RNN encoder-decoder model and a grConv (gated recursive convolutional neural network) models in machine translation, the latter model is in fact GRU. No attention mechanism was employed. The task is to translate from English to French based on a bilingual, parallel corpus. The paper showed that both models could generate correct answers but suffer from the length of sentence. Around 2015 as well, Bahdanau et al. introduced the attention mechanism in their proposed model for similar task (translating from English to French based on bilingual corpus).

The RNNLM's application in NLP also reached dialog generation or response generation. It is a more difficult topic. The response generation, different from translation, can have a large space of diversity given the same context. Usually, the model is based on a corpus of context-response pairs. To manipulate the probability of a generated responses, the loss function is used in many models.

Vinyals and Le (2015) used the Seq2Seq model in their dialogue response generation model and adopted beam search to make the original model less greedy. The model performed well in finding solutions in a domain specific IT helpdesk dataset, but it can only do simple reasoning in open-domain datasets. As the experiment shows, in open-domain conversations, the machine could deal with general knowledge Q&A, even philosophical Q&A, but loss its performance when human expression became more random and casual. It also lacks consistency in responses. Nevertheless, the model proposed by Vinyals and Le outperformed CleverBot, a rule-based model.

The previous study showed that RNNLM in response generation suffers from irrelevant responses. Another problem regarding to RNNLM is that due to the statistical nature, the generated responses tend to be bland and uninformative, i.e. one-fits-all. In the past two to three years, more and more researchers have noticed this disadvantage and tried to optimize the RNNLM for more relevant and more diverse responses.

Li et al. (2016) proposed a bidirectional informed model. The paper asserted that the model should not only consider the dependency of responses on message, but also inversely how likely the response is to answer that question. They proposed to use Maximum Mutual Information (MMI) which were introduced to speech recognition first by (Bahl et al., 1986; Brown, 1987). In the model, they penalize the generic and thus high-frequency responses. The adoption of MMI decreased the proportion of generic responses. Similar approach with mutual information includes the Pointwise Mutual Information by Takayama et al.(2019) and the model that favors frequent words in context by Nakamura et al. (2019).

Rather than penalizing "safe responses", Xing et al. (2017) tried to bring in new topics in response generation. The basic structure of their model is a GRU-RNN model. The paper introduced a topic vector to the decoder on top of the input vector. The topic vector represents the topic words of the context and is produced by the proposed Twitter LDA model. The paper assumed that each message corresponds to a topic, and each word in a message is either a topic word or background word. The decoder then adds a probability bias to the generated response that has topic words represented by the topic vector, encouraging more relevant and more diverse responses. Those models have outperformed the baseline Seq2Seq models both in machine matrix evaluation and in human annotations.

There is another line of study, multi-task model, that has been used to improve translation models (Sennrich et al., 2016). The idea is to train jointly and share parameters between two

tasks that may help each other. Luan et al. (2017) introduced a combination of Seq2Seq model and an Autoencoder model to conversational modeling, both of which are encoder-decoder structures. The Autoencoder will take in the non-conversational data of the speaker that makes the response aware of the speaker's other information.
RNNLM has also been used in semantic analysis tasks. Collobert et al. (2011b) proposed a neural network models to perform named entity recognition task. Chiu et al. (2016) proposed a hybrid bidirectional LSTM and CNN architecture, a more powerful model that improved the one proposed by Collobert et al.

Conclusion

RNNs, including its variants like LSTM and GRU, have been widely used in language modeling. Though the units and parameters of those networks are different, the underlying logic is very similar. Through adding more mechanisms like attention mechanism, the model is able to make the output influenced by a specific part of the input.
In the past five to ten years, many fields of natural language processing have seen progress with the applications of RNNLMs. Though each field still faces challenges like long sentence translation, "safe answers" and ambiguity, many of the recent researches on RNNLM aim to improve the models in performance.

References

P. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of IEEE*, 1990.

Sepp Hochreiter and Jurgen Schmidhuber, “Long short-term ¨ memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259, 2014.

Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv: Neural and Evolutionary Computing, 2014.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055.

Junya Takayama, Yuki Arase. 2019. Relevant and Informative Response Generation using Pointwise Mutual Information. Proceedings of the 1st Workshop on NLP for Conversational AI, pages 133–138 Florence, Italy, August 1, 2019.

Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. 2019. Another Diversity- Promoting Objective Function for Neural Dialogue Generation. In Proceedings of The Second AAAI Workshop on Reasoning and Learning for Human- Machine Dialogues (DEEP-DIAL).

Chen Xing, Wei Chung Wu, Yu Ping Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017).

Jason P.C. Chiu, Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, vol. 4, pp. 357–370, 2016.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Proc. ACL.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-task learning for speaker-role adaptation in neural conversation models. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 605–614.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011b. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 12:2493–2537.