A Review on Neural Language Modeling

Anuar Maratkhan

I. Introduction

Language Modeling (LM) is a central task in Natural Language Processing (NLP) and play main role in speech recognition, machine translation, optical character recognition, natural language understanding, question answering and many other tasks. Language modeling is all about sequential data. *Language model* is an algorithm for predicting next word in a text given preceeding ones. It is also mostly known as *statistical language model* for its richness in statistical approaches in previous works. The state-of-the-art statistical language models turned up to be cache models and class-based models. In addition, according to [1], most of the statistical models require more data for better performance.

However, as soon as neural networks became popular, *neural language modeling* approaches were invented. Artificial neural networks perform well on supervised learning tasks (when the data is *labeled*) and unsupervised learning tasks (when the data is *unlabeled*). Quoc V. Le, et al. in their study [2] explored and demonstrated how well neural networks operate on unsupervised learning task, where a model has to detect whether the given image contains face or not. And as the study [2] notes, the work has been motivated by neuroscience hypothesis. So does the whole science about neural networks. Moreover, information retrieval tasks that are mainly composed of unlabeled data ,like in search engines, can be approached by such unsupervised learning techniques. Neural language models, also known as continous-space language models, make use of these neural networks. These are current state-of-the-art approaches in language modeling.

This part of the introduction will conclude earlier approaches to language modeling, particularly statistical ones that are *unigram*, *bigram*, *trigram*, and more generalized *N-gram* models. (DO WE REALLY NEED IT?)

In the second section of this paper we start by reviewing Recurrent Neural Networks (RNN). Further, in the same section we review the first/previous and current state-of-the-art RNN based word-level language modeling approaches, starting from [1] and [3], proceeding by [4], and [5]. Next section presents different approaches on neural language modeling that include division of words into subwords presented in [6], reinforcement learning based techniques from [], and neural Turing machines approach [].

II. Recurrent Neural Networks based Language Models

A. Recurrent Neural Network

Recurrent Neural Networks (RNN) are primarily used in sequential data analysis such as video processing, text processing, predicting stock prices. RNN architecture is an improvement of feed-forward networks that takes into account previous hidden layer results by storing them in memory. An invention of backpropogation caused an exciting use of RNNs [7].

Simple recurrent neural network used in [1] can be described by input at time t — $x(t)$, hidden layer at time t — $s(t)$, and the output layer at time t — $y(t)$, which are computed as follows:

$$x(t) = w(t) + s(t-1) \tag{1}$$

$$s_j(t) = f\left(\sum_i x_i(t) u_j i\right) \tag{2}$$

$$y_k(t) = g\left(\sum_j s_j(t) v_k j\right) \tag{3}$$

where (1) uses current word at time t and preceeding/previous hidden layer at time $t-1$ as input, (2) and (3) have weights u and v used for computations, and both also have sigmoid activation function and softmax function, which are:

$$f(z) = \frac{1}{1+e^{-z}} \tag{4}$$

and

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \tag{5}$$

respectively.

The problem of simple RNN is that the backpropogated gradients vanshish or explode [8]. Therefore, the improvement in RNN architecture further was introduced with invention of Long-Short Term Memory (LSTM). The LSTM solved the vasnishing gradient problem by having special cells with activation gates. Moreover, Gated Recurrent Unit (GRU) improved the architecture of the LSTM by combining two gates. The study shows that RNNs perform better than complex statistical approaches, N-grams, on a real-world data [8].

B. Language Models

Neural language models have demonstrated relatively superior performance on language modeling and speech recognition tasks in comparison to the state-of-the-art statistical approaches, class-based models, in the last decades. Neural models take their first steps from taking the advantage of feed-forward networks that are listed in [1], which motivated T. Mikolov, et al. to investigate the performance of more advanced type of neural network, RNN. The study was first to evaluate RNN for modeling such sequential data.

In contrast to feed-forward networks that use fixed size of context to predict next word in a sequence, RNNs does not require to use limited number of preceeding words [1]. The

structure of simple recurrent neural network used in the study is shown in Fig. 1 below.

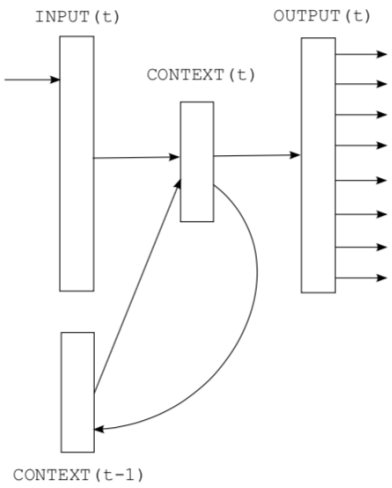

Fig. 1. Simple recurrent neural network from [1]

III. Subword-level Language Modeling

This section will show how subword language modeling exceeds the performance of word-level and characted-level language modeling by presenting previous approaches from [6], etc...

IV. Reinforcement Learning

This section will show some Reinforcement Learning (RL) approaches to language modeling tasks and applications such as dialogue systems, machine translation, text generation.

V. Neural Turing Machines

This section will discuss Neural Turing Machines (NTM) approach to different natural language processing tasks with controllers like feed-forward network, GRU, LSTM, and further will show applications in learning language models. Related papers are [9], etc...

VI. Conclusion

Over the past few years, it was clearly seen that continous-space language models outperformed statistical modeling techniques significantly. The ease of feature learning without need in extraction of those features simplified lives of scientists. The neural approach offers such opportunity to its users.

Recurrent networks are expected to improve natural language understanding in real-world applications.

To be continued...

References

[1] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Interspeech*, vol. 2, 2010.

[2] Q. V. Le, M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Deau, and A. Y. Ng., "Building high-level features using large scale unsupervised learning," in *Proceedings of the 29 th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

[3] T. Mikolov, S. Kombrink, L. Burget, J. Cernocký, and S. Khudanpur, "Extensions of recurrent neural network language model," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5528–5531, 2011.

[4] W. Zaremba, I. Sutskever, , and O. Vinyals, "Recurrent neural network regularization," 2014, arXiv preprint arXiv:1409.2329.

[5] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen, "Breaking the softmax bottleneck: A high-rank rnn language model," 2017, under review as a conference paper at ICLR 2018.

[6] T. Mikolov, I. Sutskever, A. Deoras, H.-S. Le, S. Kombrink, and J. Cernocký, "Subword language modeling with neural networks," 2011.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Macmillan Publishers Limited*, vol. 521, pp. 436–444, 2015.

[8] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding reccurent networks," 2015, under review as a conference paper at ICLR 2016.

[9] T. Deleu and J. Dureau, "Learning operations on a stack with neural turing machines," in *1st Workshop on Neural Abstract Machines Program Induction (NAMPI)*, Barcelona, Spain, 2016.