SI 630 - Winter 2020
Reading Response 2

Yirui Gao

January 29, 2020

Overall summary

This paper mainly proposes two model architecture, **Continuous Bag-of-Words** model (CBOW) and **Continuous Skip-gram** model (Skip-gram) that are used to form a continuous vector representations of words to deal with large data sets. The authors first talk about those "current" (before the year 2013) commonly used language models like feedforward neural net language model and recurrent neural net language model that utilize n-gram methods. But due to the reason of complex neural network and the non-linearity, the computation cost is large and the accuracy is not that high. The authors introduce the two new log linear models, CBOW and Skip-gram models and show the training results by using several metrics like perplexity and Word error rate in both semantic and syntactic analysis. The two new log linear models turn out to perform well and outperform those previous n-gram models. And the computational cost and the accuracy are also much better. The authors finally stress the point and put it in the experiment to show that these new proposed model architectures can be helpful for creating a high quality of vector representations of words when encountering very large data samples.

Main contributions

I think the main contributions for this paper, are the two log linear models, CBOW and Skip-gram models that the authors propose, since they benefit the computational cost and the accuracy especially for large data samples. And since the paper is published in 2013, nowadays these two models are commonly used in word vector representations in NLP. Due to the model complexity of the n-gram models, the authors mention in the paper that the neural network layer can cause the model very slow to train, since the complexity can be proved very high. But with the new models, they prove both in theory and in practice that these models can achieve better performance and lower cost, so very applicable for large data.

Fill Out Three of the Following Responses (Your Choice)

3.1 What did I learn?

Since mentioned above, the contribution work of this paper is majorly the two models that get rid of some drawbacks of n-gram models. By taking SI630 and learning the n-gram language model in class, I learn more about the drawback of this kind of models in this paper. The first one is the high dimensional problem since there large number of parameters to be learned even with the small size of vocabulary. And second, the n-gram model is in a discrete form, so very difficult to create continuous vector representations. Also, n-gram models do not perform well in word similarity tasks.

Also, I take a deep look into the two newly proposed models, CBOW and skip-gram. I learn how they work differently and in what way they reduce the computational complexity. For example, in CBOW, there are no hidden layer and words in sequences from past and future are input and they are trained to predict the current sample while the skip-gram model attempts to predict the words around the current word.

3.2 What was I confused about?

My confusion to this paper is, since both CBOW and skip-gram models are model architectures for continuous word vector representations, and we know, they are part of the word embedding nowadays. But theoretical what's the difference between these two models? Or skip-gram is just a reverse of the CBOW model? The authors could have discussed more about these two models and provide some more formulas to help us set up a comparison vertically.

3.3 How could I apply insights from this paper in my own work?

Since I'm currently working on some NLP projects, these two models, CBOW and skip-gram can be applied for the word representation since currently I'm mainly using n-gram model to train the data and when the input data is very large, the model tends to work very slowly with bad performance. For most of the NLP work, they provide good tools to improve a model.

References

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.