Kristie Wong

November 6, 2022

CS410

Technology Review: Neural Net Language Models

Neural net language models imitate the design of biological multi-layer neural networks. Their ability to learn fast with minimal human engineering and take advantage of sequential information have enabled this language model to outperform other models.

The simplest form of a neural network is the feedforward neural language model. Given a single word, the model encodes the input word as a distributed representation, in which words are mapped closer to each other the more that they are semantically-related. By using distributed representations, the model is significantly more compact than other language models. The input vector is then passed through a decoder, using a softmax function to normalize its values and return the output as a probability distribution of predictions of the next word. Between the input and output, there are hidden layers of features that characterize the word, and enable multiple levels of generalization.

To improve the feedforward neural language model, a recurrent neural network was developed. In the simple feedforward model, the prediction of the next word was only based on one preceding word. However, the recurrent model incorporates memory to improve the predictions based on the previous computations and can receive a variable length input. Recurrent neural

networks take a sequence of vectors as the input, and with each new input vector, the model updates the current state, eventually converting the state to an output as a sequence of vectors. These tasks are learned automatically by the model from a training set.

Long Short Term Memory Networks have made a significant impact on speech recognition, handwriting recognition, text-to-speech synthesis, and machine translation. The hidden layers of sequential words are connected by LSTM, as LSTMs have long-term memory, enabling the model to evaluate the relationship between distant words.

Recurrent neural net language models make up all of the state-of-the-art language models today. However, there challenges that the model faces, such as the vanishing gradient problem. As the number of terms increase in the input, the number of layers increase as well. With a deeper network, the rate of learning decreases. Realistically, the model's efficiency is limited by the length of the input sequence, and will only be able to remember a sequence of a few words effectively. Although neural net language models are more complex to implement than n-gram language models and require more training, neural net language models have proven to have higher performance due to their ability to learn faster, handle a greater amount of contextual history, and generalize unseen data using contexts of similar words.

References

Jagota, Arun. “Neural Language Models.” *Towards Data Science*, https://towardsdatascience.com/neural-language-models-32bec14d01dc.

Le, James. “Recurrent Neural Networks: The Powerhouse of Language Modeling.” *Built In*, https://builtin.com/data-science/recurrent-neural-networks-powerhouse-language-modeling.

“Understanding LSTM Networks.” *Understanding LSTM Networks -- Colah's Blog*, http://colah.github.io/posts/2015-08-Understanding-LSTMs/.