Brendan Lim

CS4395.001

April 8, 2023

Assignment 7: ACL Research Paper Summary

From the ACL 2022 Anthology, I read *Better Language Model with Hypernym Class Prediction* by He Bai from the University of Waterloo (90 citations), Tong Wang at Microsoft Research (1855 citations), Alessandro Sordoni at Microsoft Research (8299 citations), and Peng Shi from the University of Waterloo (980 citations). The paper focuses on improving neural language models by replacing nouns with their hypernyms, if they have one, using Wordnet and then training an autoregressive language model based on the modified sentences. This method, called Hypernym Class Prediction, or HCP, was found to slightly improve prediction accuracy, and the improvement increases the rarer the words become.

Neural language models have largely superseded traditional language models because of their improved accuracy. However neural language models tend to require far more computational resources to be developed. Additionally, the most common neural network model in use, Transformers, are not good at predicting rare words. There have been attempts to remedy this by adding memory segments to the transformers model or using morphology. In prior research work, class-based language models, or CLMs, have been shown to improve the computational speed of training neural language models.

Hypernym Class Prediction works by modifying the methodology of training a transformer language model. It starts by converting nouns which are very rare in the corpus into their WordNet hypernyms based on what makes the most sense and the minimum possible ‘depth’. The researchers chose to convert only nouns because they are well-connected and hard for language models. This change in the corpus and vocabulary meant a few things. The probability prediction formula for the transformer also had to change to reflect how a large portion of nouns have been replaced by hypernyms. And when training the model, instead of training to predict tokens, the model must first predict the category of the next word and then predict what the next word should be. Lastly, the researchers chose to use curriculum learning, which adds a scoring and pacing function to determine the difficulty of each training example and how many easy/difficult examples to put in each training epoch. In their training, the researchers considered hypernym prediction to be easier than predicting specific tokens.

The modified neural language model was trained and tested on two datasets separately, WikiText-103 and arXiv. WikiText-103 was based on ‘good’ and ‘featured’ articles from Wikipedia, containing 103 million tokens from over 28 thousand articles. arXiv was aggregated from abstracts from arXiv.org, an online open source research archive. Strangely, this dataset was partitioned into the training, testing, and validation split according to when it was written, where the training data was from 1986 to 2017, the evaluation data from 2017, and the test data from 2018 to 2019.

The researchers tested their HCP-modified transformers and compared them to the machine learning models mentioned earlier in the related works section as well as unmodified transformers. There were very minimal improvements between an unmodified transformer and the HCP-modified transformer. For example, the small transformer model achieved a perplexity of 36.5 whereas the HCP-modified transformer achieved a perplexity of 35.9, or difference of 0.6. Both transformer models were able to beat other models easily.

They also tested how strong the HCP-modified transformers were at predicting rarer words, which they did by comparing the prediction probabilities of the modified and unmodified transformers. It was found that as the frequency of the word decreased, the modified transformer’s performance went from similar to the unmodified transformer’s to significantly better. They also compared the unmodified transformer to another transformer with slightly worse hyperparameters. In that test, the unmodified transformer was better at predicting more common words, but worse at predicting the rarest words, indicating that the best transformers improve performance by optimizing for common words and sacrificing rarer words.

The rest of the study talks about optimizing various hyperparameters of their HCP method such as WordNet hypernym depth, determining how rare a word should be to be replaced by a hypernym, and the pacing function of the curriculum.

With recent advancements in neural networks and deep learning being applied to natural language processing, it can be easy to forget about the more traditional approaches. This paper demonstrates how a language model learn like humans use language: utilize the incredible learning power of neural networks for words which are used often, and fall back on the structural part of language for rarer words.