I am interested exploring various Transformer based language models, BERT introduced a very interesting technique which is the Masked Language Model, but I am not very convinced how the MLM is able to predict next words given a sequence of words like the normal rnn based language model is doing; It seems by applying another technique, the next sentence prediction, BERT can achieve predicting next sentence given an existing sentence, which still does not look a ‘generative’ language model.

And another puzzling point is the ‘pre-training’ language models. As we use a lot of pre-trained word vectors in our project, such as the GloVE embeddings. Using GloVe has no prerequisite for the model which is going to use it, but is it true for BERT based pretrained models? If I use BERT pretrained model to fine tune ALBERT based tasks, does it work?

The permutation Language modeling in XLNet is also very interesting in the sense that instead of masking out words in a sequence, it uses factorization order and predict words in a left to right manner. It is equivalent to MLM because it samples permutation orders randomly in order to generate a sequence. And it is nice to note that as an auto regressive, XLNet does not rely on corrupted sequence, hence the pretrain-finetune discrepancy issue is no long a problem in XLNet.