CS-522 Spring 2020

Course Project

PreTraining Bert Embeddings

BERT (Bidirectional Encoder Representations from Transformers) was released in late 2018. BERT is a method of pretraining language representations that you can use to extract high quality language features from your text data, or to fine-tune these models on a specific task (classification, entity recognition, question answering, etc.) with your own data.

Bert based transfer learning has been successful in many applications. However, if the new domain is quite specific, it is better to pretrain Bert on the domain specific data. For example, a FinBERT is a BERT-based model that is further trained on a financial corpus. It was evaluated on Financial PhraseBank and Financial QA and outperformed the baseline methods on both tasks.

Goal: The goal of this project is to understand the effect of Bert pretraining on specific domains.

Methodology: Pretrain Bert on specific domains and evaluate how much it improves performance on the classification task. Also evaluate when pretraining leads to overfitting and "forgetting" of the main language model.

Prerequisites: Set up a Google Cloud account to be able to train on TPU's. Make sure you get a free credit as a first time user.