TensorFlow: Toolkits for Big Data

Kenneth Wong

March 12, 2023

This note summarizes how to use Tensorflow to preprocess data for training. Tensorflow provides industry-standard data input pipeline. Based on the official documents. TensorFlow can (1) prepare and load large data; (2)build and fine-tune models; (3)deploy models on-device, in the browser, or in the cloud; (4) provide enterprise-level application solutions. These cover each part of data management process. I am most interested in part 1 and 2.

Dealing with large data-set, I have two problems. First, data exceeds memory limits. Second, no code can **transform data into matrices** . For these two problems, TensorFlow provides solutions.

What advantages does TensorFlow offer for text cleaning and training? a. RAM requirement (It randomly draws a batch and only reads the label and index); if this is still too large, it reads a fixed number of samples; b. parallel computing (depends which objects); c. good for industry standard.

The first example is an introduction with linear.

The second example, which is a sentiment analysis, builds on the first one.

My initial interest in TensorFlow is to implement LDA with it. In a paper titled Deep Probabilistic Programming, the sample code is provided.

TensorFlow official document offers separate modules for loading files of different types. Loading text is from this webpage.

Finally, I find an interesting Kaggle project. The team shows how to do topic modelling in a professional way.