GPU-Accelerated Study on Semantic Textual Similarity Analysis Based Recommender System with Deep Learning

Jose M. Vidal

Department of Computer Science and Engineering
University of South Carolina,
Columbia SC 29201, USA
Email: vidal@sc.edu
Phone: 803-777-0928

1 Research Project

1.1 Background

With continuous expansion of Internet activities and online merchandise, recommender system plays a more and more critical role in the interactive Internet environment. It applies data analysis and helps users find their most wanted products from online merchandise sites. For instance, a personalized recommender system on Amazon (www.amazon.com) suggests music and books to customers based on the user's personal shopping experience, hobbies and areas of concern.

Whereas, a non-personalized recommender system like Zagat(www.zagat.com) and Yelp(www.yelp.com) provides a general restaurant guides based on the input of millions of individuals. The same reviews and rating scores are presented to users no matter who is looking up their sites.

In our research, we have a system with a lexicon which consists of words and definitions. One word may have multiple definitions. The functionality of our system is to recommend pre-defined existing term: definition pairs when a user tries to add a new term: definition pair to the lexicon. The system is served as a repo for sociology terms while keeping lexicon minimal and parsimony, and eliminating redundancy for direct and indirect definition. Instead of using reviews and ratings data, our recommendation is based on the semantic textual analysis of definitions. In recent years, Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Tree-Structured LSTM, Deep Structured Semantic Model (DSSM) and some similar deep learning frameworks have been used to compute semantic similarity between two text snippets.

1.2 Motivation

Many natural language processing (NLP) applications and recommender systems such as paraphrase recognition (Dolan et al., 2004), automatic machine

translation evaluation (Kauchak and Barzilay, 2006), textual summarization (Aliguliyev, 2009), tweets search (Sriram et al., 2010), student answer assessment (Rus and Lintean, 2012; Niraula et al., 2013) and recommender synonymy challenge (wikipedia page) are constrained by the effectiveness of semantic textual similarity (STS) analysis.

There are three classes of models where real-valued vectors are used to represent the meaning of phrases and sentences: bag-of-words models, sequence models, and tree-structured models. In bag-of-words models, the representation of a sentence is independent of word order (Landauer and Dumais, 1997; Foltz et al., 1998). Sequential models construct the sentence representation as an order-sensitive sequence (Elamn, 1990; Miolov, 2012). Tree-structured models compose each sentence representation from its constituent sub-phrases according to a given syntactic structure over the sentence (Goller and Kuchler, 1996; Socher et al., 2011).

Considering the importance of capturing semantic difference of word sequence (e.g.,"Cat eats fish" vs. "Fish eats cat"), the order-sensitive sequential models or tree-structured models are better sentence representations due to their relation to syntactic interpretation of sentence structure. Recurrent neural network is an important type of deep learning framework, which is designed for sequence problems. Due to its capability for processing arbitrary length sequences, RNNs are a natural choice for sequence modeling tasks.

Up to now, no investigation has ever employed GPU-accelerated deep learning approach. In this project, we will explore the Sentences Involving Compositional Knowledge (SICK) dataset (Marelli et al., 2014) using deep learning with GPU acceleration, consisting of 9927 sentence pairs in a 4500/500/4927 train/dev/test split. Each sentence pair is annotated with a relatedness score which was assigned by different human annotators. The training of the RNN with massive data and deep learning has high computational intensity. Thus, it is critical to increase the computation speed for our semantic similarity based recommender system.

Graphic Processing Unit (GPU) can significantly increase the computational power in the RNN training process. Several available deep learning platforms including Theano and Tensorflow support NVIDIA GPU wit CUDA and increase the computation performance significantly over CPU-only mode.

Our research could broadly impact in multiple fields, including sociology, computer science and linguistics. We expect our research outcome draws the attraction from the scientific community to emphasize the application of deep learning with GPU-acceleration in this field.

1.3 Our Approach

Our proposed research has three stages: 1) In the first stage, we plan to experiment RNN with sequential LSTM models. Two commonly-used sequential LSTMs are the Bidirectional LSTM and the Multilayer LSTM. In this stage, we want to provide the effectiveness of RNN, and we also hope the features can be located more accurately comparing to Bag-of-words approaches; 2) A limitation

of the LSTM is that they only allow for strictly sequential information propagation. The tree structured LSTM structure is believed to allow for richer network topologies where each LSTM is able to incorporate information from multiple child units. In the second stage, the research will be revealing the mechanism of tree structured LSTM with standard LSTM unit on gating vectors and memory cell updates. This allows a Tree-LSTM model learn to emphasize semantic heads in a semantic relatedness task. 3) In the third stage, we will apply the algorithms developed from first two stages to our non-personalized recommender system. During the integration testing period, the performance of these two stages will be evaluated.

2 Usage of Graphics Processing Unit(GPU)

Currently, we are using the Theano[x], a deep learning framework, as the back-end. On top of Theano, we are using the Keras, a high-level neural networks library, written in Python, to run CNN and RNN algorithms on CPU only. As described in the official document, with GPU, Theano performs data-intensive calculations up to 140x faster than with CPU. We will install the proper program to accelerate the computation using the request GPU.

In the future, we will also work on Tensorflow[x], which is another popular deep learning framework. Tensorflow also recommends using GPU to accelerate the computation speed.

References

1. Clarke, F., Ekeland, I.: Nonlinear oscillations and boundary-value problems for Hamiltonian systems. Arch. Rat. Mech. Anal. 78, 315–333 (1982)
2. Clarke, F., Ekeland, I.: Solutions périodiques, du période donnée, des équations hamiltoniennes. Note CRAS Paris 287, 1013–1015 (1978)
3. Michalek, R., Tarantello, G.: Subharmonic solutions with prescribed minimal period for nonautonomous Hamiltonian systems. J. Diff. Eq. 72, 28–55 (1988)
4. Tarantello, G.: Subharmonic solutions for Hamiltonian systems via a $\mathbb{Z}_p$ pseudoindex theory. Annali di Matematica Pura (to appear)
5. Rabinowitz, P.: On subharmonic solutions of a Hamiltonian system. Comm. Pure Appl. Math. 33, 609–633 (1980)