Proposal

Data Mining and Predictive Modeling of Amazon Customer Reviews

Technical Points of Contact:

Name: William G. Hatcher
Email Address: whatch2@students.towson.edu
Mailing Address: Department of Computer and Information Sciences
Towson University, 7800 York Road, Towson, MD 21252

Name: Kevin McNamara
Email Address: kmcnamara@towson.edu
Mailing Address: Department of Marketing
Towson University, Stephens Hall, Towson, MD 21252

Period of Performance: 10/4/2018 - 12/6/2018 (2 months)

Executive Summary

Deep learning, by definition, is the implementation of deep neural network architectures to enact learning algorithms. These networks include many parameters to tune accuracy, precision, recall, and more. The application of deep learning has lead to many significant advancements in the fields of cognitive science, image and video processing, character recognition, natural language processing, virus and malware detection and analysis, and medical imaging analysis, to name a few. It is these significant works, and their counterparts across disciplines, that warrant collection and review for the research community at large. In this proposal, we seek to carry out (1) a comprehensive survey of deep learning, and (2) provide a benchmark of deep learning tools and frameworks. This work, to be conducted over a 6 month period, will approach the topic of deep learning from the aspects of enabling architectures, mechanisms, applications, datasets, and state-of-the-art research developments. In addition, as a reference for researchers across disciplines, the work shall provide comprehensive review of particular frameworks and their applications, and point out areas where deep learning has yet to be applied.

Contents

1. Introduction

1.1. Problem Statement

Machine Learning is experiencing a renaissaince, thanks in no small measure to advancements in computer hardware. Indeed, complex learning computations on massive data graphs are becoming ever easier, and every major software company (Google, Microsoft, Amazon, Apple, etc.) is looking to outpace their competitors in developing prediction and optimization technologies. At the most basic level, Machine Learning is a mathematical framework to fit or segregate some set of data, and to use this fit for prediction or classification on new, unseen datapoints. Compared with manually programming sequential and concurrent processes and applications, Machine Learning algorithms can extract the relevant features and relationships from the data with thousands fewer lines of code, and can be applied to data too big to be parsed by any individual.

Traditional Machine Learning, implemented via simple models that run in trivial time, include Linear and Logistic Regressions, Decision Trees, and others. However, the emergence of Deep Learning, carried out via complex Neural Networks, engenders truly powerful learning systems that can achieve unprecedented accuracy. Diverse and complex implementations of these Neural Networks (NNs) are widespread, and the specific network architecture, neuron types, and optimizers used depend on the application or problem to be solved.

Given that Deep Learning has reached a significant level of saturation, it should be noted that many of the traditional tasks of Machine Learning have become trivial to implement and execute. This includes image classification and segmentation, handwriting classification, and others. In this landscape, new applications for Machine Learning are being explored.

A thorough review is needed to classify and categorize all manners of deep learning, and distinguish them from shallow learning mechanisms that, while still powerful, cannot handle the massive data and complexities that are emerging because of the Internet of Things. Specifically, a survey is called for that investigates state-of-the-art machine learning by mechanism, algorithm, and application, and provides a wide breadth of understanding and can guide researchers in searching to apply these algorithms to groundbreaking investigations.

1.2. Challenges

Deep learning now suffuses the applications and software products that make daily life easier and allow us to stay connected. Yet, it may be impossible to cover every area that deep neural networks have been applied. In a research context, looking only at the IEEE Xplore digital library, some 7,593 articles have been published relevant to deep learning. In addition, enterprise and commercial deep learning systems can only be assessed from the corporate and trade publications they deign to release, often seeking to keep hidden the mechanisms that may indeed be trade secrets. Thus, to properly survey the landscape of deep learning, it is imperative to have a thorough understanding of deep learning architectures, and to be able to infer and discriminate what information is truly novel and unique.

1.3. Importance

Deep learning is at a critical peak of public awareness. As more researchers are investigating the application of deep learning to their field or topics of study, it is necessary for those without the knowledge to have a reference to assist them in applying appropriate deep learning approaches, whether they be classification, regression, data fusion, reinforcement learning, etc. In addition, for the general Computer Science community, and those that work daily with deep learning architectures, the breadth and depth of deep learning holds many novel and interesting works that may perhaps provide inside into solving a particular problem.

1.4. Objectives

In this project, we plan to evaluate the state-of-the-art of Deep Learning across platforms, algorithms, and datasets. Specifically, we intend to categorize Deep Learning by mechanism, algorithm, hyperparamters, platform, applications, and performance. This includes developing a survey of the works done in this area, and conducting benchmark tests to provide baseline comparative analysis.

2. Project Approach and Management Plan

2.1. Overview

Two primary deliverables are intended in this work. First, a survey of deep learning shall be conducted, considering the origins and recent advances in deep neural networks. This survey shall categorize deep learning architectures into various categories and subcategories derived from the type of learning, learning target, algorithmic implementation, enabling software frameworks, and datasets. Then, we shall carry out a practical evaluation of many of the representative software platforms to investigate their efficiency, diversity, and ease of use.

Deliverables:

a. *Deep Learning Survey:* The survey will consider the evolution of Machine Learning to the current Deep Learning paradigm, the enabling algorithms and technologies, a classification of the available platforms, current and emerging applications of Deep Learning, and future research directions. The survey will review cross-sections of supervised learning, reinforcement learning, and unsupervised learning, and review the advances in convolutional neural networks, deep belief networks, deep Q-learning, and more.

b. *Practical Implementation and Evaluation:* The practical evaluation of Deep Learning platforms will test the comparative performance of the various platforms and their comparable algorithms in terms of accuracy and runtime on various datasets. Each dataset represents a different learning application (classification, anomaly detection, etc.). Platforms include TensorFlow, Theano, Keras, Torch, Deeplearning4J, among others.

Datasets:

a. *Handwriting Samples:* The MNIST dataset is a widely used dataset for machine learning of handwritten numbers. This dataset is provided the National Institute of Standards and Technology, and comprises some 60,000 samples.

b. *Image Classification:* The ImageNet dataset, composed of over 14 million labeled images, is another widely used dataset, and was part of the ImageNet competition. This dataset is used for object and scene recognition.

c. *Speech Recognition:* The TIMIT dataset includes recordings of 630 speakers of eight major American English dialects, each reading ten sentences.

d. *Text Data:* The Examiner.com crowd-sourced news website data includes over 3 million instances of headlines, spam, etc. for clustering and sentiment analysis.

e. *Mobile Malware:* In the course of our prior work, we have assembles some 60,000 Android Applications (Benign and Malware) from various sources for use in android malware detection.

f. *Network Traffic:* Multiple datasets have been requested from the Center for Applied Internet Data Analysis (CAIDA), which includes public and upon-request anonymized network traces from 2008-2016, as well as various worm trace snapshots.

2.2. Schedule

Based on the overview provided, two major tasks are to be delivered. The first, the survey of deep learning, shall be conducted in the first three months of the period of performance, and shall be delivered by *March 31, 2018*.

The second deliverable, the comparative analysis of deep learning platforms, shall be carried out over the remaining three months to be delivered at the completion of the performance period, on *June 30, 2018*. This will entail the evaluation of at least four major deep learning platforms on the six datasets noted in the **Dataset** subsection of Section 2.1.

2.3. Qualifications

This research group has conducted many successful surveys in the past, and has been conducting research on machine learning and deep neural networks for some time. Publications of this work include various IEEE conferences and journals. The most recent works include deep learning for Android malware detection, using the malware dataset outlined above. Other publications include network traffic analysis using distributed parallel systems, among others. We feel this work has a high chance of success, given the current demand for the topic, and the relevancy of the review of deep learning platforms. Specifically, such platforms abound, and navigating the various aspects of deep learning implementations requires no small amount of experience. In addition to these qualifications, we have access to multiple server-grade computer systems in our research lab, as well as access to the IEEE Xplore digital library for reference.

3. Evaluation of Success

The success criteria of this projects are based on two primary metrics: (i) the timely delivery of the deliverables by the proposed due dates (Merch 31 and June 30 of 2018), and (ii) the acceptance and publication of both works. In the case of publication, the survey must be accepted by a reputable journal, and not simply a conference, while the comparative assessment will still be considered successful only being accepted for conference publication.