Deep Neural Networks for Analysing Cancer Genomics Data

Organizers: Fraunhofer FIT, Germany & RWTH Aachen University, Germany
Speakers: Md. Rezaul Karim, Oya Beyan, Michael Cochez, Lars Gleim, Nils Lukas and Stefan Decker

Abstract

Although classical machine learning techniques allow researchers to identify groups (i.e. clusters) of related variables from data, the accuracy and effectiveness of these methods diminish for large and high-dimensional datasets. On the other hand, deep neural network architectures (the core of deep learning) can better exploit large-scale datasets to build complex models.

One of the biggest, and widely used datasets is The Cancer Genome Atlas (TCGA) project, which is available on the TCGA Research Network: http://cancergenome.nih.gov/. TCGA is a project that started in 2005 to catalog genetic mutations responsible for cancer using genome sequencing and bioinformatics. Currently, TCGA has 39 projects, each corresponding to a certain type of cancer. Each cancer type data are high-dimensional and heterogeneous consisting of masked somatic mutation data, copy number segment masked copy number segments, gene expression quantifications, DNA methylations, miRNA expression, and clinical data. Moreover, the TCGA project collects both clinical data and biospecimen(s) from each patient.

This tutorial provides an introduction to the use of different deep learning and data analytics tools in theory and in practice. We will in particular focus on processing and analysing large-scale genomics dataset such as TCGA.

Target audience

This tutorial is designed such that it would be of interest to a wide range of researchers based on their expertise and interest including (but not limited to):

1. Computer scientists
2. Bioinformaticians
3. Life scientists

Tutorial outline

The tutorial will have three session as outlined below:

1. Machine learning and deep learning for analysing large-scale genomics data

In this session, we describe the basics of machine learning and getting started with deep learning. In a nutshell, the following topics will be covered:

1. Data analytics in a Big Data era
2. Available technologies (Apache Spark H2O + Sparking water, TensorFlow, KNIME)
3. Machine learning for analysing genomics data
4. Linear vs. tree ensembles
5. Why classical machine learning fails to tackle very high dimensional data?
6. How can deep learning algorithms help us analyze massive amount of genomic data?
7. Using Convolutional Neural Networks (CNN), Recursive Neural Network (RNN), Autoencoders, and FeedForward Neural Network (DBN, MLP).

2. Hands-on machine learning application development using KNIME

Can you build a Machine Learning model without coding? Yes, you can! With the KNIME platform. In this session, we will show how to use KNIME for analyzing genomics data:

1. Creating a simple workflow using KNIME
2. Implementing a simple Logistic Regression for analysing genomic data
3. Implementation of tree ensembles (e.g. Random Forest) for analysing genomic data
4. Implementing a CNN for analyzing genomic data

2. Hands-on demonstration using Spark, H2O and TensorFlow

Some of the most popular neural network architectures are Feedforward Neural Networks (e.g., Multilayer Perceptron, Deep Belief Networks), CNN, Long Short-Term Memory network (LSTM) and Autoencoders. In the literature, CNN is well known as an effective network to act as 'feature extractors' for imaging and computer vision, whereas LSTM is a type of powerful Recurrent Neural Networks (RNN) for modelling orderly sequence learning problems.

In this session, we demonstrate how to develop machine learning and deep learning applications for analysing cancer genomics data. In a nutshell, the following will be demonstrated:

1. Implementing a Random Forest/ Gradient boosted tree for cancer type prediction
2. Implementing an H2O based deep learning application for cancer type/subtype and survival rate prediction
3. Implementing an MLP and a DBN for for cancer type/subtype and survival rate prediction.