Doctoral Dissertation
Doctoral Program in Electrical, Electronics and Communications Engineering
(32nd cycle)

Novel Neural Approaches to Data Topology Analysis and Telemedicine

Vincenzo Randazzo

* * * * * *

Supervisors
Prof. Eros Pasero, Supervisor
Prof. Giansalvo Cirrincione, Co-supervisor

Doctoral Examination Committee:
Prof. Alessandro Vinciarelli, Referee, University of Glasgow
Prof. Roberto Tagliaferri, Referee, Università degli studi di Salerno
Prof. Marco Badami, Politecnico di Torino
Prof. Vitoantonio Bevilacqua, Politecnico di Bari
Prof. Anna Esposito, Università della Campania “Luigi Vanvitelli"
Prof. Salvatore Vitabile, Università degli studi di Palermo

Politecnico di Torino
September 22, 2020

Summary

The conventional approach to artificial intelligence and data mining is tied up to input distribution learning and, more generally, to understand the laws underlying data independently of the input at hand. To this purpose, the most common used tools are the neural networks, which are often employed as black boxes for extracting information from data. Once found the proper architecture, neural networks better map the data manifold than human-designed models, especially if the input distribution is non-linear or is embedded in a high dimensional space. Very often, in a context where the Internet of Things (IoT) has become pervasive and tons and tons of data are produced every instant, it is easy to think the best approach is to gather as much data as possible and, then, use deep learning. The idea is that collecting and aggregating a huge amount of data from different sensors, would yield the needed information. Unfortunately, recently the focus is more on achieving amazing performance on classification/regression tasks rather than understanding the reasons behind them. A quite exhaustive example is deep learning, which seems the solution to most of the open problems once collected enough data, but does not have a theoretical model behind. For example, it automatically extracts features from data while performing training; but, what are the extracted features? Moreover, handling a deep learning neural model, requires a lot of data and computation power; is it really required to use such a great computational power, time and efforts just because the dataset is huge?

This thesis tackles the lack of formalism and the black box approach by providing a scientific framework for analysing data and understanding their topology before classification. In this sense, neural networks are used both to explore data manifold and to determin which architecture is better tailored for a problem.

Before choosing an architecture, it would be better to understand data. The input space is analysed using both linear and non-linear methods to estimate its intrinsic dimensionality; understanding the input space can drive the performance analysis and unveil data patterns, which can then be used to guide training, e.g. in the deep learning, and to select the best feature set.

Both unsupervised and supervised architectures have been employed; the former is used for clustering data into unknown groups, the latter for classifying them into predefined classes. The choice of the proper approach is done w.r.t. different

applications, e.g. online learning, data projection or telemedicine. Both stationary and non-stationary input distributions are examined. When needed, new neural networks (onCCA, GCCA, G-EXIN, GH-EXIN) have been designed for exploiting input data topology and preserving it during training.

Supervised learning performance has been analysed by studying the classification results as input features change. Deep learning automatically extracts features and provides good classification outcomes, but it is a black box and its results cannot be interpreted in a theoretical framework. On the other side, shallow neural networks need a human-based feature engineering phase prior to their training, but it is possible to interpret their outcomes w.r.t. the input features. The proposed approach combines these two techniques for exploiting their advantages by means of a correlation analysis between the deep layers and the best performing feature set of the classical approach. In this sense, by understanding which are the features automatically extracted by the deep technique, it would be possible to give an interpretation, i.e. an explanation, of its results.

Public available databases have been used in order to compare performance with state of the art on a common benchmark. At the same time, data have been collected at the *Neuronica* and $Polito^{BIO}Med$ laboratories of Politecnico di Torino in order to validate the quality both of the proposed approach and of the new designed and built devices. The input data can be grouped in three main categories: non-stationary, stationary and IoT. The former regards input distributions that change over time, e.g. jump, and has been exploited for machine prognostic. On the other side, stationary data experiments have been used to handle medical and hierarchical applications; in this sense, the aim was to explore data internal structure and to discover new patterns. Finally, in a real case scenario, an application to telemedicine has been studied: new wireless wearable devices, the ECG WATCH and the VITAL-ECG, have been developed to acquire and monitor vital signs, such as heart rate. The proposed approach has been used to diagnose possible heart diseases and to design a biometric identification system based on electrocardiogram.