Ph.D. in Information Technology
Thesis Defenses

February 16th, 2022
at 14:30
Room Bio1 and online by Zoom

Luca BENEDETTO – XXXIII Cycle

An assessment of recent techniques for question difficulty estimation from text

Supervisor: Prof. **Paolo Cremonesi**

Abstract:

In the educational domain, question difficulty estimation consists in estimating a numerical or categorical value representing the difficulty of an exam question.
It is traditionally performed with manual calibration or pretesting, which have several limitations: indeed, they are either subjective or introduce a long delay between the time of question creation and when the new question can be used to assess students.
Recent research tried to overcome these shortcomings by leveraging Natural Language Processing techniques to perform question difficulty estimation using as input only the textual content of the questions, which is the only information that is always available at the time of question creation.
Specifically, research proceeded along two main directions: supervised and unsupervised approaches, which have peculiar advantages and limitations.
This thesis explores previous literature in both research directions and evaluates several models, including novel approaches, on real world datasets coming from different educational domains.
The experimental results show that model accuracy heavily depends on the characteristics of the questions under consideration and, most importantly, the educational domain: while simple models based on readability indexes and linguistic measures are generally fairly accurate on reading comprehension questions, the calibration of questions assessing domain knowledge requires more advanced models based on the attention mechanism and Transformers.

Giuseppe CANONACO – XXXIV Cycle

Learning in non-stationary environments: from a specific application to more general algorithms

Supervisor: Prof. **Manuel Roveri**

Abstract:

In recent years, Machine Learning (ML) has gained a lot of attention and popularity because of its ability to model highly complex phenomena given sufficiently large data sets. This thriving field embraces all the algorithms and techniques able to automatically learn a given task using a finite amount of data that can be thought of as experience. These algorithms are usually paired with some assumptions which may or may not be satisfied in real-world practical applications. For instance,

some algorithms assume to have an input signal and its associated output (also called supervised information) for each example in their training data set such that they can learn a mapping from the input signal to the output that generalizes on previously unseen examples. This assumption is not always satisfied in practice, where the output signal could be too expensive to be collected. An example of this mismatch between theory and practice can be found in the context of corrosion prediction for pipeline infrastructures. Here the output signal, corresponding to the presence of corrosion in a given point of the pipeline, is hardly available for the infrastructure of interest due to the incredibly huge cost companies have to bear in order to collect it. Another fundamental assumption associated with ML techniques is about stationary data-generating processes, which implies that the phenomenon we are trying to learn does not change as time passes by. Examples of applications where the stationarity assumption about the data-generating process does not hold are in the context of finance, due to market evolution, in the context of water reservoir systems, due to climate change, in the context of corrosion, because of the wear and tear of infrastructures that increases with time, etc. In all the above-mentioned scenarios, ML techniques cannot be directly applied without softening the assumptions about the availability of supervised information or stationarity they are equipped with. Therefore, in the context of this dissertation, inspired by the specific application needs of corrosion prediction in pipeline infrastructures, we will investigate ML solutions able to weaken the assumptions of available supervised information and stationarity. Even though the lack of supervised information is a thoroughly researched problem thanks to Transfer Learning (TL), it is overlooked in the context of corrosion prediction requesting tailored solutions for this critical application. Softening the assumption of stationary data-generating processes, instead, is much less studied in lots of different ML sub-fields motivating a much more general investigation within the scope of this dissertation.
The contribution of this dissertation is threefold. The first one deals with learning techniques for corrosion in pipeline infrastructures starting from the data set creation up to devising SL techniques that, with the aid of TL, are able to build a model of the corrosion phenomenon without using supervised information coming from the pipeline of interest. The second one deals with RL in non-stationary environments and develops both an active-adaptive approach to cope with changing environments and a TL technique for RL able to take into account an underlying time-variant structure intrinsic to the available historical knowledge. Finally, the third one, deals with Federated Learning (FL) under non-stationarity and pervasive systems. This last part introduces a passive-adaptive approach to mitigate non-stationarity in FL contexts and a birdsong detection approach able to run on a highly constrained Internet-of-Things (IoT) unit.

Simone DISABATO – XXXIV Cycle

Deep and Wide Tiny Machine Learning

Supervisor: Prof. **Manuel Roveri**

Abstract:

In the last decades and, in particular, in the last few years, Deep Learning (DL) solutions emerged as state of the art in several domains, e.g., image classification, object detection, speech translation and command identification, medical diagnoses, natural language processing, artificial players in games, and many others.
In the same period, following the massive spread of pervasive technologies such as Internet of Things (IoT) units, embedded systems, or Micro-Controller Units (MCUs) in various application

scenarios (e.g., automotive, medical devices, and smart cities, to name a few), the need for intelligent processing mechanisms as close as possible to data generation emerged as well. The traditional paradigm of having a pervasive sensor (or pervasive network of sensors) that acquires data to be processed by a remote high-performance computer is overcome by real-time requirements and connectivity issues.

Nevertheless, the memory and computational requirements characterizing deep learning models and algorithms are much larger than the corresponding abilities in memory and computation of embedded systems or IoT units, significantly limiting their application. The related literature in this field is highly fragmented, with several works aiming to reduce the complexity of deep learning solutions. However, only a few aim to deploy such DL algorithms on IoT units or even on MCUs. All these works fall under the umbrella of a novel research area, namely Tiny Machine Learning (TML), whose goal is to design machine and deep learning models and algorithms able to take into account the constraints on memory, computation, and also energy the embedded systems, the IoT, and the micro-controller units impose.

This work aims to introduce a methodology as well as algorithms and solutions to close the gap between the complexity of Deep Learning solutions and the capabilities of embedded, IoT, or micro-controller units.

Achieving this goal required operating at different levels. First, the methodology aims at proposing inference-based Deep Tiny Machine Learning solutions, i.e., DL algorithms that can run on tiny devices after their training has been carried out elsewhere. Second, the first approaches to on-device Deep Tiny Machine Learning training are proposed. Finally, the methodology encompasses Wide Deep TML solutions that distribute the DL processing on a network of embedded systems, IoT, and MCUs.

The methodology has been validated on available benchmarks and datasets to prove its effectiveness. Moreover, in a ``from the laboratory to the wild'' approach, the methodology has been validated in two different real-world scenarios, i.e., the detection of bird calls within audio waveforms in remote environments and the characterization and prediction of solar activity from solar magnetograms. Finally, a deep-learning-as-a-service approach to support privacy-preserving deep learning solutions (i.e., able to operate on encrypted data) has been proposed to deal with the need to acquire and process sensitive data on the Cloud.

PhD Committee

Prof. **Marco Santambrogio**, Politecnico di Milano

Prof. **Frank Hopfgartner**, The University of Sheffield

Prof. **Man Lin**, St. Francis Xavier University, Canada

Prof. **Zhen Ni**, Florida Atlantic University