Alkis Papanastassiou *PhD Research project*

Domain adaptation and active learning AI techniques in the context of regression, simulation and agnostic optimization of large industrial apparatuses and high energy physics experiments

Idea behind the project

The DOLPHIN project aims to establish a strong two-way link between industry and university for the application of advanced Machine Learning (ML) techniques. On the one hand, the proponents of the Department of Physics and Astronomy (UniFi) intend to further develop the techniques used for the analysis of high-energy physics (HEP) data collected by the CMS experiment at the LHC accelerator at CERN. On the other hand, the proponents of the Baker Hughes (BH) company plan to apply very similar techniques in the context of the study, characterization and optimization of the performance of the turbomachinery that the company designs, builds and for which it offers assistance in its Florence site. DOLPHIN will specifically study neural networks (NN) applicable to a context, or domain, different from the one in which they are trained (domain optimization, DA), and the minimization of a function (with unknown analytical expression) whose evaluation in a point of the parameter space takes a lot of time / resources, with the minimum number of evaluations of the function itself (active learning, AL). The strength of DOLPHIN is twofold: on the one hand, the LHC data provides a clean and correct labeled sample, ideal for studying network architectures and comparing performance; at the same time, the need to use these algorithms to solve different problems makes it possible to decouple the development of the network from the specificities of the case study.

Domain adaptation of classification and regression models

In ML, domain adaptation refers to the property of a model of being applicable to a context that differs in some way from the training context. A typical example in HEP data analysis is the identification of an unknown signal in a data sample which is predominantly background by training a classifier, often a neural network (NN), on a simulated data sample. Using a particular theoretical model for the signal during training limits the general applicability of the discriminator. Domain adaptation techniques would allow to build classifiers agnostic with respect to the theoretical model assumed for the signal.

In applications to turbomachinery, there are many cases in which domain adaptation needs to be used. An example is the so-called virtual sensing. A virtual sensor is an ML model capable of predicting a measurand (i.e.: pressure, temperature) in a point of the apparatus not equipped with a physical sensor, using as input the time series of measurands which vice versa are measures measured in other parts of the apparatus. These regression models are trained on a prototype and must be applicable to fleet units. For the virtual sensor to be accurate it is essential that its model is independent of several conditions that differ between the prototype and the fleet unit.

To tackle these tasks, we plan to use adversary neural networks (ANNs), composed of two NNs, one that performs the classification/regression task (specific to the case study), while the other, the "adversary" network, seeks to define the domain (identify the signal model for the HEP case, or distinguish whether the input comes from the prototype or from a fleet unit for the turbomachinery case). The goal of the training is to maximize the performance of the first network by minimizing the ability of the second to recognize the domain.

Active learning for optimization

In ML, active learning refers to a series of techniques to approximate a function (even of many unknowns), with the least number of point evaluations of the function itself. The adjective "active" in the name of the technique refers to the fact that the algorithm, based on previous evaluations, suggests the point where it is convenient to perform the subsequent evaluation of the function. A typical example in HEP is the optimization of the parameters for the algorithm used to discriminate the signal from the background events, which is often pursued manually, by performing a grid-search in the parameter space, with the aim of maximizing a simplified merit figure. With an active learning technique, we could be able to perform the optimization on the complete analysis performed on a limited number of points in the multidimensional space of the algorithm parameters.

In applications on turbomachinery, just to mention a possible use case, it is of particular importance to minimize the polluting emissions of the combustion process, in particular CO and NO_X. These emissions depend on a wide range of parameters both geometric and dependent on how the turbine is operated. Finding a design that minimizes emissions would be an impossible task with a grid-search approach, due to the enormous size of the parameter space, particularly considering that collecting a point can take up to months or years if one wants to design, manufacture and test different combustion chambers: active learning is essential to guide the development path.

There are several active learning algorithms. We plan to start with the *gaussian process* technique, which is often used in the context of NN hyperparameter optimization, and then explore other solutions.

Link between goals

The DA and AL applications are interconnected both for the HEP case studies and for those referring to turbomachinery. In HEP it is essential to obtain a discrimination of the signal with respect to the background that is at the same time agnostic with respect to the signal model and optimal in terms of sensitivity. In turbomachinery, it is desirable to build models that are generalizable to operating conditions or units different from those used for training the models themselves.