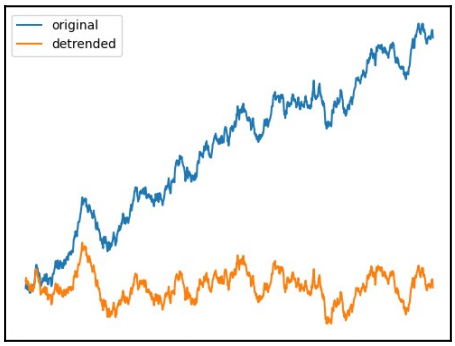
Προτεινόμενα Θέματα Διπλωματικής Εργασίας

από Δρ. Στασινό Κωνσταντόπουλο, ΕΚΕΦΕ "Δημόκριτος"

A. Making educated guesses

When a human reader is presented with a graph, they are usuallt able to mentally decompose it into the different phenomena that contribute to it and the mathematical functions that describe them.

To give a simple example, everybody will immediatelly observe that in the graph on the right the blue plot is the composition of the orange plot and a linear upward trend. Other cases will not be so easy to spot visually, especially when the composition function is not addition but a function that is harder to visualize, such as convolution.



Naturally, in the general case there will not be a unique way to decompose a phenomenon. Even in the simple case shown above, the assumption was made that the orange plot should be as close to zero as possible, and the height of the linear trend was computed under this assumption. In more complex cases, more such assumptions will be needed to the point that even the number of functions that were composed is unclear.

Within this broad setting, the student will explore the application of *Automatic Differentiation* and *Induction* to the problem of jointly learning the parameters and the structure of a neural net that is able to reconstruct the graph. The structure reflects the composition of the different functions that participate and the parameters approximate each function with a neural network. Training data can be easily synthesised, and testing data (with human analysis as ground truth) can be collected in the non-parametric statistics or physics literature.

B. Task planning in the Web of Data

Linked data is the Web-native way to publish structured data, emphasising the value that is added not only by semantically linking internally to the dataset, but also across datasets. The vision of linked data is for the Web to become the global database for data processing systems, besides its human users. But Linked Data is not as linked as it was envisaged. Syntactic heterogeneity is manageable, as connectors exist that can abstract over all the different network protocols and query languages. Semantic heterogeneity, on the other hand, is practically unmanageable as the rate at which data is published thwarts any manual curation effort.

Let us imagine an AI system that is aware of the major data portals and catalogues, foundational ontologies and linksets, dataset search engines, and, in general, of the resources that a human data analyst would be expected to be aware of. When tasked with a data request, the AI system should search in data portals and search engines and reason about possible paths through linksets and (possibly) uncertain equivalences in order to fulfil the task. The systen would then present the user with a (possibly incomplete or uncertain) plan, highlighting to the user what needs to be done to finalize the plan in terms of establishing links that could not be found, resoving ambiguities, or confirming uncertain equivalences.

As an example, consider finding the female unemployment rate in Mycenae between 2000 and 2010. The system should be aware of the ambiguity, realize that such a statistics is meaningless for the ancient city and proceed with its modern namesake. It should seach in data portals to find a relevant dataset for Mycenae, New York, fail, and use a geographic names dataset to search for the higher-level administrative region, until data can be identified. The system should also try to ease other restrictions besides geographic coverage, and realize that there is no obvious preference between (a) female unemployment rates in New York state between 2000 and 2010; (a) county-level unemployment rates between 2000 and 2010; and (c) the county-level average female unemployment rate between 2005 and 2015. The system should then present all query plans, the datasets needed for each, the disambiguation decision it made, and the uncertainties and approximations of each plan.

Within this broad setting, the student will explore the application of a solver or planner to formulate plans for dataset linking and querying; and of a mechanism to select among these plans. The work is expected to focus on planning using resources manually prepared in advance, to avoid the implementation details of how to query actual on-line portals.

C. Estimating the quality of training data

The success of a machine learning exercise depends on having not only sufficient but also adequate data. Although there are many aspects to what makes a dataset adequate, we will here focus on detecting data where there is noise in the inputs or the labels. More specifically, we will focus on noise that introduces spikes in timeseries, blurrs in an images, or other artefacts that can be located to specific parts of the data as opposed to noise distributed throughout the data.

Detecting anomalous datapoints is typically approached as an “anomaly detection” task, where we are looking for sudden changes in some characteristics of the data that otherwise follow relatively stable distributions.

Within this broad setting, the student will explore the idea of training a noise detection network without supervision regarding noise artefacts, but with supervision regarding some other downstream task that is affected by noise. The core idea is that when similar datapoints have different labels in the downstream task, this is either due to noise masking the actual inputs or due to labelling errors. The goal is to formulate this idea as a machine learning exercise that can be applied without prior knowledge of what types of noise artefacts are to be expected in each new domain of application.