Deep Learning-based Prediction of Query Answering Times for Data Series Similarity Search

Themis Palpanas
University of Paris
Institut Universitaire de France (IUF)
themis@mi.parisdescartes.fr

Qitong Wang
University of Paris
qitong.wang@etu.u-paris.fr

[*Description*] A key operation for the (increasingly large) data series collection analysis is similarity search. Recent studies demonstrate that SAX-based indexes offer state-of-the-art performance for similarity search tasks [1]. To facilitate the deployment of real-world data series similarity search components, query answering time estimation is essential for the purpose of systematic throughput optimization and latency analysis. Deep learning techniques have been recently applied to database tunning and optimization. However, existing methods suffer from the problem that the training of deep neural network models requires large amounts of accurately labeled data, which is usually unaffordable in real-world applications.

In this internship, we will exploit and develop state-of-the-art deep neural network models for data series similarity querying answering time estimation for the iSAX family of indexes, with a focus on novel techniques for training data efficiency.

This internship is supervised by Prof. Themis Palpanas and his PhD student Qitong Wang from the diNo team at the University of Paris. The selected intern will become a member of diNo, which has world-leading expertise on data series management, indexing, and analysis.

[*Challenges*] The challenges lie in data series training data efficiency for deep neural networks. Promising breakthroughs exist in the profiling of data series collection characterics and training subset structure, as well as the adoption of existing models as in transfer learning and representative training instances selection based on model feedback as in active learning.

[*Methodology*] The internship will start with the accurate prediction of data series similarity quey answering time based on deep neural networks, using various synthetic and real-world datasets. Redundant training instances are to be identifiyed from dataset structure analysis and trained model feedbacks, such that the minimal training set is obtained. Finanly, features from different datasets are to be exploited for model transformation with little effort.

[*Prerequisites*] Excellent Python and C/C++ programming skills, very good knowledge of deep learning frameworks (PyTorch/GPU, etc.) and libraries in data analysis workflow (NumPy, Matplotlib, etc.). Research/project experiments and publications on deep learning or data analysis is a plus.

[*How to apply*] Apply by emailing your CV and transcripts to Prof. Themis Palpanas: themis@mi.parisdescartes.fr

References

[1] Karima Echihabi, Kostas Zoumpatianos, Themis Palpanas, Houda Benbrahim: The Lernaean Hydra of Data Series Similarity Search: An Experimental Evaluation of the State of the Art. Proc. VLDB Endow. 12(2): 112-127 (2018).