COMPUTER AND CONTROL ENGINEERING

EURECOM/DAUIN - Contrastive Representation Learning for Tabular Data

Funded By	Dipartimento DAUIN EURECOM - ECOLE D'INGENIEURS & CENTRE DE RECHERCHE EN SYSTEMES DE COMMUNICATIONS [P.iva/CF:65383181575]
Supervisor	CAGLIERO LUCA - luca.cagliero@polito.it
Contact	
Context of the research activity	Representation learning aims at leveraging Deep Learning models to build vector representations of data suitable for accomplishing complex tasks. Established methods tailored for textual data, images, and time series have already been proposed. The goal of the PhD program is to advance the state of the art by exploring their application to tabular data. The twofold aim is to leverage the meta-information provided by the structure and investigate the use of contrastive learning approaches.
Objectives	RESEARCH OBJECTIVES The PhD program will focus on representation learning for tabular data. It will part of a ongoing research collaboration between Politecnico di Torino and Eurecom. The goal of the PhD program is to design and implement Deep Learning approaches, mainly self-supervised, to build vector representations of structured data that incorporate structure-related information and are suitable for tackling complex NLP tasks. We envision two main directions: - benchmarking such representation with a suite of tests (in the style of CheckList [1] for language models in NLP). We had some preliminary idea tested in [2] - Apply contrastive learning techniques, already established for images and time series [3], to tabular data. [1] Ribeiro et al. Beyond Accuracy: Behavioral Testing of NLP models with CheckList [2] Cappuzzo et al. Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks. SIGMOD 2020 [3] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, Serge Belongie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14755-14764 OUTLINE OF THE RESEARCH PLAN The PhD student will first elaborate on the preliminary ideas described in [2]

	by proposing suitable tests and robust empirical evaluations. Then, she will explore the potential of state-of-the-art self- and semi-supervised learning architectures to analyze tabular data to solve challenging taks (e.g., clustering [4], summarization [5]). Finally, she will specifically address the problem of designing contrastive representations by generating positive/negative pairs based on structure-related information. [4] Colomba L, Cagliero L, Garza P. Density-based Clustering by Means of Bridge Point Identification. IEEE TKDE. PrePrints pp. 1-14, DOI Bookmark: 10.1109/TKDE.2022.3232315 [5] La Quatra M, Cagliero L. BART-IT: An Efficient Sequence-to-Sequence Model for Italian Text Summarization. Future Internet. 2023; 15(1):15. https://doi.org/10.3390/fi15010015 LIST OF POSSIBLE PUBLICATION VENUES Top-ranked conferences (e.g., EMNLP, ACL, SIGIR, SIGMOD, VLDB) Top-quality journals (e.g., TACL, IEEE TKDE, ACM TKDD, ACM TOIS, IEEE TPAMI)

Skills and competencies for the development of the activity	The PhD candidate is expected to be have skills on - Data Science fundamentals - Python programming - Deep Learning A basic knowledge on Natural Language Processing techniques is also advisable but not required.