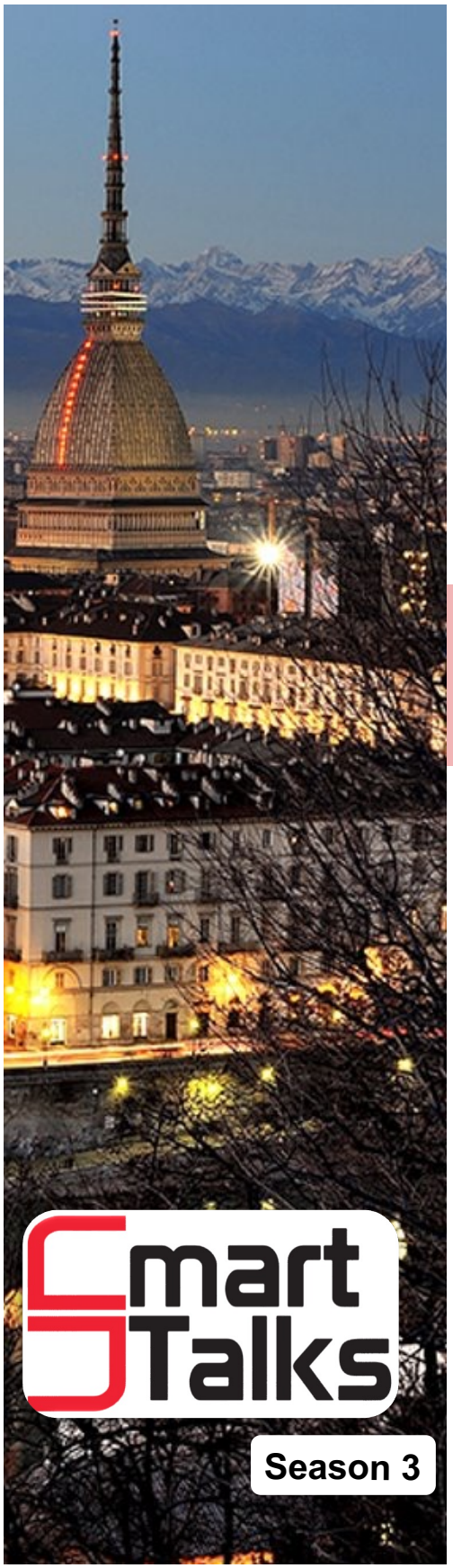
Politecnico di Torino
SmartData@PoliTO

SmartData

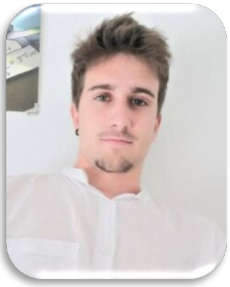
April 17th 2023, 5:30 PM CEST

SmartTalk: Covivio

https://smartdata.polito.it/category/smarttalks/



Gabriele Ciravegna



Gabriele Ciravegna is a Post Doc in the MAASAI (Models and Algorithms for Artificial Intelligence) research team of Inria. He received his Ph.D. degree with honours from the University of Florence in 2022 under the supervision of Professor Marco Gori. In 2018, he received a master's degree in Computer Engineering with honours at the Polytechnic of Turin.
He have always been interested in the machine learning field. Nowadays, he is focused on overcoming the intrinsic limits of machine learning and neural networks. By combining neural networks with domain knowledge and reasoning, he is studying how to tackle the opacity of neural networks, their vulnerability against adversarial attacks and their data hungriness. He presented his works in several international venues such as NeurIPS, IJCAI, AAAI, and IJCNN. He also serves as a reviewer in conferences and journals that are about Neural Networks, such as IJCAI, AAAI and IEEE TNNLS. Besides machine learning, he also like football, volleyball, and playing the piano.

Concept-based models: towards interpretable by-design deep neural networks

ABSTRACT

Deploying AI-powered systems requires trustworthy models supporting effective human interactions, going beyond raw prediction accuracy. Concept bottleneck models promote trustworthiness by conditioning classification tasks on an intermediate level of human-like concepts. This enables human interventions which can correct mispredicted concepts to improve the model's performance. However, existing concept bottleneck models are unable to find optimal compromises between high task accuracy, robust concept-based explanations, and effective interventions on concepts -- particularly in real-world conditions where complete and accurate concept supervisions are scarce. To address this, we propose Concept Embedding Models, a novel family of concept bottleneck models which goes beyond the current accuracy-vs-interpretability trade-off by learning interpretable high-dimensional concept representations. However, CEMs rely on high-dimensional concept embedding representations which lack a clear semantic meaning, thus questioning the interpretability of their decision process. To overcome this limitation, we also propose the Deep Concept Reasoner (DCR), the first interpretable concept-based model that builds upon concept embeddings. In DCR, neural networks do not make task predictions directly, but they build syntactic rule structures using concept embeddings. DCR then executes these rules on meaningful concept truth degrees to provide a final interpretable and semantically-consistent prediction in a differentiable manner.