Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer Engineering (38^{th}cycle)

Advancing Deep Learning Across Modalities: From Explainability to Multimodal Content Understanding

By

Davide Napolitano

******

Supervisor(s):
Prof. Luca Cagliero, Supervisor
Prof. Silvia Chiusano, Co-Supervisor

Doctoral Examination Committee:
Prof. Caren Han, Referee, University of Melbourne
Prof. Maria A. Zuluaga, Referee, EURECOM
Prof. Luca Laurenti, Referee, TU Delft
Prof. Paolo Garza, Politecnico di Torino

Politecnico di Torino
2026

Abstract

The rapid adoption of deep learning systems in high-stakes, data-intensive domains has intensified the need for models that are not only accurate but also interpretable and capable of reasoning across heterogeneous data sources. Nowadays, data rarely exist in a single modality; predictions increasingly depend on interactions among text, images, and other cues. This dissertation addresses these challenges by investigating two paths: explainable artificial intelligence and multimodal content understanding, with the overarching goal of advancing methods across modalities while scaling to real-world complexity.

The first part focuses on Shapley value–based explanations, a framework for feature attribution grounded in cooperative game theory. Despite their strong foundations, Shapley values remain difficult to apply in practice due to their exponential computational complexity. To address scalability, this work introduces a unified neural formulation for neural estimation that jointly approximates black-box model behavior and produces feature attributions within a single model. This approach reduces the computational overhead of prior multi-stage explainers and enables real-time explanations without sacrificing accuracy. To support systematic comparison, the dissertation also establishes a comprehensive benchmarking framework that consolidates datasets and metrics, mitigating the lack of standardized evaluation protocols for emerging neural approximation methods.

Beyond efficiency, the dissertation addresses the reliability of explanations. Point-valued attributions fail to capture the intrinsic uncertainty of modern systems, like in ensemble-based settings. This work, therefore, investigates interval Shapley explanations as a principled means to model variability in feature contributions. By introducing these theoretical concepts into machine learning, this work proposes alternative estimation strategies and evaluation metrics, demonstrating that inter-

val explanations provide a more faithful and informative representation of model behavior.

The second path shifts to multimodal learning, with particular focus on visually rich documents, in which meaning arises from the interaction among text, images, and spatial structure. The dissertation introduces two distinct methods for retrieving relevant elements in multi-page documents. First, it proposes a graph-based model that explicitly encodes semantic, spatial, and hierarchical relationships. Second, it presents a transformer-based model designed to tackle these same complex reasoning tasks. Building on this document domain, it introduces a framework for generating and evaluating unanswerable questions in visually rich documents. Based on plausible yet unanswerable questions, it reveals systematic patterns in visual large language models across different dimensions, like entities, elements, and layout. It further presents strategies that enhance the reliability of abstention by mitigating factors.

Finally, multimodal reasoning is extended to broader analysis tasks, including fake news detection and idiomaticity understanding, showing how aligned vision–language representations and structured reasoning can support the interpretation of complex semantic phenomena.

Overall, this work delivers three main takeaways: first, scalable explainability requires optimized formulations and standardized evaluation to be viable in real-world systems; second, reliability is a fundamental dimension of explanations and should be explicitly modeled rather than assumed; and third, effective multimodal understanding depends on architectures that encode structure and relations alongside semantic content. Future research directions include extending uncertainty-aware explainability to more modalities and multimodal approaches, and developing unified frameworks that jointly support reasoning and abstention.