Politecnico di Torino

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

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (37^{th}cycle)

Exploring the Use of Deep Models to Analyze Data in Multimodal Scenarios

By

Lorenzo Vaiani

******

Supervisor(s):
Prof. Luca Cagliero, Supervisor
Prof. Paolo Garza, Co-Supervisor

Doctoral Examination Committee:
Prof. Paolo Papotti, Referee, EURECOM
Prof. Marco Grangetto, Referee, Università di Torino
Prof. Silvia Chiusano, Politecnico di Torino
Dr. Siwen Luo, University of Western Australia
Dr. Lorenzo Canale, Centro Ricerche, Innovazione Tecnologica e Sperimentale RAI

Politecnico di Torino
2025

Abstract

The increasing availability of multimodal data, combining text, images, audio, and video, is redefining the boundaries of artificial intelligence research. Real-world tasks increasingly require the integration of heterogeneous information sources to enable deeper and more nuanced understanding. Motivated by this challenge, this dissertation explores the use of deep learning models for multimodal data analysis across diverse application domains, with a particular emphasis on vision-language interactions.

The work addresses three major research directions. First, it investigates multimodal learning strategies for social media analysis, focusing on the detection of harmful content such as misogyny and fake news. Novel architectures are developed to exploit the interplay between visual and textual modalities, overcoming limitations of unimodal models and achieving improved performance in complex, user-generated content scenarios. Particular attention is paid to the effective extraction of visual features through region-based proposals, enabling a more precise alignment between image regions and textual cues.

Second, the dissertation advances the field of visually-rich document understanding by proposing techniques for multi-teacher knowledge distillation and external knowledge prompting. These methods enhance key information extraction by combining textual content, document layout, and visual elements, allowing models to capture the intricate structure of documents more effectively. Experiments conducted on public benchmarks demonstrate significant improvements over baseline approaches, underlining the benefits of multimodal strategies in document analysis.

Third, the research extends multimodal learning to domains beyond traditional vision-language tasks, addressing content summarization and emotional analysis in spoken and video data. Hierarchical and multimodal architectures are proposed for podcast summarization, while cross-modal fusion techniques are explored for

video-based emotion recognition. These contributions demonstrate the flexibility and generality of multimodal learning frameworks, highlighting their potential to support increasingly diverse AI applications.

Across all these domains, the dissertation identifies and addresses several challenges inherent to multimodal learning, including modality alignment, handling missing or noisy data, balancing modality contributions, and managing computational complexity. Methodological innovations, such as novel fusion strategies, enhanced pretraining approaches, and robustness mechanisms, are proposed and validated through extensive experimental evaluations.

By bridging multiple research areas and proposing cross-domain multimodal solutions, this dissertation contributes both theoretical insights and practical advancements to the field of multimodal deep learning. It demonstrates that carefully designed multimodal models can achieve more accurate, robust, and context-aware understanding, pushing the boundaries of what AI systems can interpret and reason about in heterogeneous environments.

Future research directions include developing more efficient architectures for resource-constrained settings, exploring self-supervised pretraining strategies for multimodal data, and extending multimodal analysis to novel modalities such as haptics or olfactory signals. The findings presented herein lay the groundwork for the development of next-generation multimodal AI systems that can interact with complex real-world environments in richer, more human-like ways.