Ph.D. in Information Technology
Thesis Defense

November 15th, 2024
At 2:00 p.m.
Aula Seminari Alessandra Alario

Davide SALVI– XXXVI Cycle

DATA-DRIVEN TECHNIQUES FOR SPEECH AND MULTIMODAL DEEPFAKE DETECTION
Supervisor: Prof. Paolo Bestagini

Abstract:

Recent advancements in deep learning and generative models have significantly simplified the creation and manipulation of synthetic media, allowing even inexperienced users to produce highly realistic content with minimal effort.
Besides the exciting opportunities that the developed technologies offer, they also carry the potential for unpleasant consequences.
Indeed, when these are used for malicious purposes, they can lead to harmful situations, with several recorded cases of fraud, blackmail, and fake news spreading due to the misuse of synthetic data.
An example of this phenomenon is deepfakes, synthetic multimedia content generated through deep learning techniques that depict individuals in actions and behaviors that do not belong to them.
Using only a few images or an audio recording of a target victim, an attacker can utilize deepfake technology to produce synthetic data that impersonates the victim and discredits their reputation. To prevent unpleasant situations due to the misuse of forged data, it is crucial to develop detection methods capable of discriminating between real and fake content.

In this thesis, we consider the problem of deepfake detection and explore multiple strategies and approaches to tackle it.
Starting from a monomodal scenario, i.e., synthetic speech detection, we propose two distinct techniques to address it, alongside suggesting multiple solutions for related problems.
These include estimating the reliability of the output of a classifier, addressing the synthetic speech attribution task, and proposing multiple XAI techniques to determine the critical factors in a synthetic signal that drive the detection process.
Additionally, we explore tasks related to splicing detection and localization in speech deepfakes, analyzing content comprising elements from both real and fake classes rather than entirely belonging to one or the other.

Then, we extend the deepfake detection problem to the multimodal scenario, analyzing audio-video deepfakes.
Leveraging the insights gained from monomodal studies, we tackle some of the issues that are present in the current literature.
These involve exploring diverse fusion strategies across the content of different modalities and handling the lack of multimodal deepfake data needed to train and test the classifiers.

Most of the proposed methods offer intriguing insights not only for the deepfake detection problem but also for the multimedia forensic field at large. Indeed, the presented approaches can be adapted to tackle various tasks across different domains. Among these, we recall the transfer learning methods explored in the analysis of high-level semantic features and the diverse fusion techniques we introduced.
In general, we consider this thesis as an initial exploration of the multimedia forensic field. Despite the promising outcomes we achieved, new challenges continuously emerge in multimedia forensics, necessitating the constant development of novel methods to address them.
We hope our contributions will provide valuable insights in this regard, fostering the development and progress of this research field.

PhD Committee

Marco Marcon, **Politecnico di Milano**

Nicola Adami, **Università di Brescia**

Benedetta Tondi, **Università di Siena**