Ph.D. in Information Technology
Thesis Defense

May 9th, 2025
at 14:00
Room Alpha – building 24

Eugenio LOMURNO – XXXVI Cycle

Adversarial and Generative Deep Learning for Data Privacy in Human-Centered Artificial Intelligence

Supervisor: Prof. Matteo Matteucci

Abstract:

Artificial Intelligence is growing rapidly in a highly interconnected world, providing solutions to problems that were unimaginable just a few years ago, while at the same time opening the door to existential risks and dangers for humanity.

Keeping its development under control and respecting the individual is one of the main goals of Human-Centred Artificial Intelligence, a branch of computer science that has emerged in the last decade and aims to make the research, production and use of Artificial Intelligence algorithms transparent, credible, safe and ethical.

With the advent of cyber-attacks against such algorithms, regulation and protection have become imperative.

Through the use of certain Artificial Intelligence models, it is indeed possible to extract the information learned by third party algorithms, showing how the training data is present in these architectures, albeit in the form of a latent representation.

Data, whatever its nature or form, is thus one of the most important and debated resources, being on the one hand the essential ingredient for learning algorithms, and on the other hand an asset to be protected and kept private.

This thesis begins by examining the current landscape of privacy preservation techniques in deep learning, revealing significant challenges in balancing model performance with data protection. Existing methods, including Differential Privacy, often result in substantial compromises with respect to privacy guarantees and model performance, limiting their practical application in real-world scenarios.

In response to these challenges, this research introduces a series of novel contributions aimed at enhancing both privacy and performance in deep learning systems. Initially, it explores regularisation techniques as a means to improve privacy protection whilst maintaining model performance. This approach proves to be a promising alternative to more computationally intensive methods, offering a better balance between privacy and utility.

Building upon this foundation, the work presents Discriminative Adversarial Privacy (DAP), a new strategy that leverages adversarial training to simultaneously optimise for task performance and privacy protection. This approach demonstrates significant improvements over traditional methods, offering a more favourable balance between model accuracy and privacy guarantees.

The thesis then investigates the potential of federated learning as a privacy-preserving technique for collaborative model development. Recognising the vulnerabilities inherent in traditional approaches, it proposes Synthetic Generative Data Exchange (SGDE). This innovative method leverages generative models to produce synthetic data for exchange within a federated learning context, significantly enhancing privacy protections whilst maintaining or even improving model performance.

Expanding on the concept of synthetic data, a comprehensive pipeline called Gap Filler (GaFi) is developed to optimise the quality and utility of synthetic datasets for downstream tasks. This approach significantly narrows the performance gap between models trained on synthetic versus real-world data across various domains. Additionally, the research explores the adaptation of Stable Diffusion 2.0 for synthetic dataset generation, incorporating techniques such as transfer learning and fine-tuning. Building upon these advancements, the Knowledge Recycling (KR) pipeline is introduced, which integrates and refines the insights from GaFi and the Stable Diffusion experiments. KR employs advanced generative techniques to further enhance the effectiveness of synthetic data in model training, demonstrating its potential to surpass real data in certain scenarios.

In the context of collaborative learning, this research proposes Federated Knowledge Recycling (FedKR). This novel approach enables secure and effective collaboration across institutions without compromising data privacy. By leveraging locally generated synthetic data and sophisticated aggregation mechanisms, it offers enhanced security and improved model performance compared to traditional federated learning techniques.

In conclusion, this thesis presents a series of methodologies and techniques that contribute to the ongoing development of privacy-preserving deep learning. The proposed approaches offer potential solutions to some of the current challenges in balancing data utility and privacy in machine learning applications.

PhD Committee

Prof. Daniele Loiacono, Politecnico di Milano

Prof. Marina Paolanti, Università degli Studi di Macerata

Prof. Marcella Cornia, Università di Modena e Reggio Emilia