DNN Watermarking: Opportunities, Challenges, Dos and Don’ts

Deep Neural Networks (DNNs) are the basis for the astonishing progress of Artificial Intelligence (AI). Such progresses, however, are hindered by a number of problems inherently linked to DNNs. Watermarking can provide a unified, solution to three of the most compelling problems of and raised by AI technology, namely: i) protection of the Intellectual Property Rights of DNN models, iii) authentication of DNNs for dependable AI, and iii) distinction between natural and AI-generated content. Even if media watermarking is a well-established field with a solid theory and several practical solutions have developed over the past decades, its direct application to DNNs is not possible for the simple reason that DNNs are not static objects but functions defined by the way they map the input samples into the output space. This basic observation raises several challenges and opens a range of new opportunities that are going to fill the agenda of researchers for the next years. In this talk I will discuss the main similarities and dissimilarities between media and DNN watermarking, and present the main challenges ahead of DNN watermarking researchers, highlighting some possible solutions I have been working on in the last years. I will also point out some bad practices that negatively affected research in media watermarking and that should not be repeated in the case of DNNs.