Since their resurgence in 2012, Deep Neural Networks have become ubiquitous in most disciplines of Artificial Intelligence, such as image recognition, speech processing, and Natural Language Processing. However, over the last few years, neural networks have grown exponentially deeper, involving more and more parameters. Nowadays, it is not unusual to encounter architectures involving several billions of parameters, while they mostly contained thousands less than ten years ago.

This generalized increase in the number of parameters makes such large models compute-intensive and essentially energy inefficient. This makes deployed models costly to maintain but also their use in resource-constrained environments very challenging.

For these reasons, much research has been conducted to provide techniques reducing the amount of storage and computing required by neural networks. Among those techniques, neural network pruning, consisting in creating sparsely connected models, has been recently at the forefront of research. However, although pruning is a prevalent compression technique, there is currently no standard way of implementing or evaluating novel pruning techniques, making the comparison with previous research challenging.

Our first contribution thus concerns a novel description of pruning techniques, developed according to four axes, and allowing us to unequivocally and completely define currently existing pruning techniques. Those components are: the granularity, the context, the criteria, and the schedule. Defining the pruning problem according to those components allows us to subdivide the problem into four mostly independent subproblems and also to better determine potential research lines.

Moreover, pruning methods are still in an early development stage, and primarily designed for the research community. Indeed, most pruning works are usually implemented in a self-contained and sophisticated way, making it troublesome for non-researchers to apply such techniques without having to learn all the intricacies of the field. To fill this gap, we proposed FasterAI toolbox, intended to be helpful to researchers, eager to create and experiment with different compression techniques, but also to newcomers, that desire to compress their neural network for concrete applications. In particular, the sparsification capabilities of FasterAI have been built according to the previously defined pruning components, allowing for a seamless mapping between research ideas and their implementation.

We then propose four theoretical contributions, each one aiming at providing new insights and improving on state-of-the-art methods in each of the four identified description axes. Also, those contributions have been realized by using the previously developed toolbox, thus validating its scientific utility.

Finally, to validate the applicative character of the pruning technique, we have selected a use case: the detection of facial manipulation, also called DeepFakes Detection. The goal is to demonstrate that the developed tool, as well as the different proposed scientific contributions, can be applicable to a complex and actual problem. This last contribution is accompanied by a proof-of-concept application, providing DeepFake detection capabilities in a web-based environment, thus allowing anyone to perform detection on an image or video of their choice.

This Deep Learning era has emerged thanks to the considerable improvements in high-performance hardware and access to a large amount of data. However, since the decline of Moore's Law, experts are suggesting that we might observe a shift in how we conceptualize the hardware, by going from task-agnostic to domain-specialized computations, thus leading to a new era of collaboration between software, hardware, and machine learning communities. This new quest for more efficiency will thus undeniably go through neural network compression techniques, and particularly sparse computations.