Addressing Distributional Shift challenges in Computer Vision for Real-World Applications - Abstract

Francesco Cappio Borlino

June 7, 2024

Machine learning technologies have been part of our lives for a long time, as proven by widely spread applications such as automatic spam filtering and face recognition cameras. In spite of this and of the very rapid progress that characterizes research in this field, it took several years for the general public to become aware of its potential to influence our lives. However, no one can deny that this moment has now arrived, mainly due to the presentation of easy-to-use interfaces that enable anyone to interact with large language models. This development has sparked curiosity in the public about which fields *Artificial Intelligence* will influence the most. Among them, there is certainly Computer Vision, the domain in which deep neural networks have obtained the most remarkable results even before Natural Language Processing studies resulted in the development of those language models which made AI a subject on everyone's lips. The success of these models has been significantly supported by the generality of the language mean and the ease of interaction but has also led to the overestimation of their abilities, which is clearly evidenced by their inclination to make mistakes. Indeed, there is still a long way to go in order to make deep models robust enough to enable their deployment in safety-critical applications. Their brittleness gets particularly exposed when they face real-world operating conditions characterized by a large number of unforeseeable variables, as it happens when they meet out-of-distribution data. This is a situation that occurs in several scenarios, for example when a deep model faces samples with a radically different appearance from the one it is used to, or belonging to semantic categories that it has never met.

This thesis focuses on the study of these two kinds of distribution shifts. We start by providing some background, describing when they occur and why they impact so much on neural networks' performance. In this context, we focus in particular on the relationship between out-of-distribution performance and *representation learning*: the unique ability of neural networks to automatically learn how to summarize complex data samples, such as images or videos, into compact and easily tractable representations. With the goal of developing deep learning methods whose scope of applicability goes beyond lab settings, we then

proceed by studying some specific distribution-shifted scenarios for which we propose novel solutions, by trying to adopt an original point of view and a critical eye on the most common paradigms. In particular, we first consider *simpler* research settings in which a visual shift is the only difference between training and deployment conditions, and later move to more *complex* cases in which semantic and visual shifts appear together, as this is the most likely situation when considering open-world deployments.

Through our studies, we come to the conclusion that the way representations are learned can *seriously* impact the performance of deep models on out-of-distribution data, and it is thus necessary to adopt more robust learning approaches if we want to obtain dependable systems. In this context, an important novelty is represented by the recent presentation of the first *foundation models* for Computer Vision. These are models trained at scale on huge data collections that enable them to extract general-purpose representations providing a fair treatment for in-distribution and out-of-distribution data. The correct exploitation of this knowledge can thus represent a real change of paradigm in the study of distribution-shift problems.