Visual Domain Generalization via Self-Supervised Learning

Candidate: Silvia Bucci

In these days the world is wondering about the potentialities and risks of artificial intelligence models trained on a huge amount of data and computational resources. While this debate is certainly important, it is also relevant to put the spotlight on a hallmark of human intelligence which is still far-fetched for machines: visual generalization and adaptability. Several studies in neuroscience have discussed how these skills develop in children from a combination of supervised and self-supervised learning, with the first guided by adults and the second spontaneously rising from playing and freely interacting with the world. Games like jigsaw puzzles and coloring help to learn the invariances and regularities of objects and scenes, and contribute to building robust semantic knowledge that generalizes to novel contexts. With this thesis, we show how these learning strategies can be exploited to improve artificial model robustness and reliability. Specifically, we show how auxiliary self-supervised tasks can be paired with supervised ones with significant beneficial effects. The manuscript is divided into three main parts.

In the **first part**, we introduce the first contribution of this thesis which consists in using Self-Supervised Learning as an auxiliary task for robust classification models across domains. In particular, we present strategies to deal with Domain Generalization and Unsupervised Domain Adaptation problems under the Closed-Set assumption where the Source/s and Target label sets perfectly overlap. Specifically, we introduce two methods: JiGen and Tran-Adapt. In the first one, Jigsaw puzzle and Rotation recognition are used as auxiliary objectives to improve the network's generalization ability in object recognition. With the same aim, the second approach relies on an inter-modal Self-Supervised task whose goal is reconstructing RGB data from Depth and vice versa to improve scene recognition across domains.

In the **second part** of the thesis, we face problem settings that combine category and domain shift. We show how self-supervision remains an effective auxiliary task also in these more challenging cases. Specifically, we present two solutions for Unsupervised Domain Adaptation, respectively under PDA and Open-Set assumptions. The former is an extended version of JiGen where self-supervision is exploited to focus on the shared categories, the latter (ROS) uses rotation recognition to detect unknown samples in the target domain.

Finally, in the **third part**, we explore the relation-based self-supervised approaches whose formulation allows for an easy extension to supervised learning. Indeed, differently from transformation-based self-supervised strategies for which resorting to multi-task was essential, in relation-based approaches it is possible to include supervision by simply exploiting data annotation while defining instance relations. Two patches are similar if they come from images belonging to the same class and different otherwise. In particular, with HyMOS we propose a contrastive learning-based approach for Open-Set across domains, ReSeND is a relational reasoning-based approach for semantic novelty detection.

While the focus of this thesis is primarily on object recognition, the proposed supervised and self-supervised learning integration can yield benefits in various other computer vision tasks as detection or segmentation which are essential for robotics applications. This would bridge the gap between static AI algorithms used in virtual environments and real AI agents that can interact with the physical world, paving the way for more dynamic and adaptive AI systems. In addition, this thesis explores relation-based self-supervised tasks in an open-world setting, primarily for unknown detection. However, there is potential for extending the proposed approaches to lifelong learning, for the continual discovery of novel categories and progressive knowledge growth. To pursue these new research directions, it may be necessary to supplement vision-based approaches with natural language also facilitating human-machine interaction.
As extensively discussed in this thesis, the major limitation of vision models is their inability to manage changes in domain and image semantic content. In this respect, natural language could represent the extra modality needed to close this knowledge gap. Also, language could allow easier interaction with humans: if the decision of an AI agent is explained in words, a human could provide auxiliary information to correct the cause of a possible error rather than just the effect and suggest hints based on the specific need and limits of the agent facing a different task.

Overall, adaptability, generalization, and being ready to adapt to novelty are key aspects to develop resilient and trustworthy AI systems, so we believe that the topics and methods discussed in this thesis will remain relevant references for future research.