Abstract

The impact of the latest advancements in Artificial Intelligence on everyday life is growing stronger day after day at a frantic pace. Deep Learning systems are increasingly becoming capable of executing complex tasks autonomously, gaining credibility in almost any context and/or application. In this scenario, the use of AI software to assist experts in making fast and reliable assessments, especially in clinical applications, is leading to a revolution in image analysis.
The purpose of the research presented in this document is to show how Deep Learning models and algorithms are reshaping the way biological images are treated, inspected and interpreted, heading towards the next level of AI-enhanced medicine.
Specifically, it is shown that leveraging Vision Transformer-based architectures to detect and identify peculiar diseases over different types of clinical images and applications represents a powerful and effective technique to tackle different types of image classification problems, both on large and small datasets. Moreover, it is demonstrated that leveraging the most recent Diffusion-based image generation models can effectively boost performance whenever data lacks quality and/or uniformity, when the images in the database are imbalanced among the different classes, or when data samples fail to represent the most significant category.
Given the inherent peculiarities of any specific applications, researchers commonly tackled each problem by customizing the Transformer architecture and adapting it to properly process the data they dealt with. However, this approach shows the lack of standardization purposes. In this sense, the present document proves the validity of leveraging ViT-based models for a variety of classification problems on biological images, suggesting that they can used almost *out of the box* for a plethora of image detection/classification applications, which can easily be extended beyond the clinical field. To defend this statement, the procedures behind image analysis are observed and compared for Vision Transformers and Convolutional Neural Networks, showing that a better understanding of how attention works in image classification can head towards an increased awareness of what makes features *relevant*.