Interpretability of Deep Learning for Medical Image Classification: Improved Understandability and Generalization: Summary of the contributions

Motivation

The application of Artificial Intelligence (AI) and, particularly, deep learning to medical image analysis has led to exceptional research results in several contexts, including the analysis of human tissue samples. State-of-art Convolutional neural networks (CNNs) were developed to automatically detect cancer regions in digital pathology images of human tissue microscopy slides. As retrospective studies have shown, however, CNNs appear as black box machines and their behavior is not always reliable. For example, CNNs do not generalize well enough to real-world data and are not ready yet for the integration in clinical workflows. Hence, this is where existing models need to be improved the most. The automated predictions by CNNs are taken with restraint by physicians, who often question when and why the model is expected to fail on their data and struggle to accept its opacity and little understandability.

The developments in this Ph.D. dissertation can fill the gap between the pitfalls of current AI developments and the requirements of medical practice. Thanks to the contributions in this work, explanations of CNN models predicting tumorous regions in digital slides of tissue can be generated in terms of clinical features that are easy to understand by pathologists. Because of these explanations, pathologists can be empowered with a deep understanding of the model and may predict eventual failures or justify unexpected outcomes.

Research contributions

This Ph.D. dissertation proposes new interpretability techniques that aim at making the inner workings of deep learning classifiers understandable to physicians. AI Interpretability, or AI explainability (XAI), is an emerging field in AI that deals with the generation of explanations of a machine learning or a deep learning system. Some explanations are local, and aim at explaining the process followed by the model to obtain a specific prediction for a given input. Global explanations, instead, aim at explaining the overall model behavior and the learned patterns. Interpretability results in the medical domain rely on comparing explanations with domain knowledge, particularly to ensure that the standards of clinical practice are followed by the machines. In addition to proposing novel techniques to interpret CNNs for digital pathology, this work shows that by interpreting CNNs, unwanted behaviors and biases can be uncovered and corrected, generating more reliable and trustworthy models for the clinic.

My research proposes

(1) The quantification of the reliability and consistency of the state-of-the-art interpretability tools. I show that these methods have important limitations, some of which are overcome as the second contribution of this work.

(2) The development of new explainability methods for CNNs classifiers in digital pathology, such as Regression Concept Vectors and Sharp-LIME. The proposed methods apply also to other tasks with different imaging modalities, e.g. the detection of retinal diseases in ophthalmology.

(3) The development of methods that improve the model generalization to new inputs coming from shifted domains, for example, from multiple hospitals. I show that the existing models can be modified and corrected to account for unwanted behaviors. For example, the sensitivity to domain shifts is reduced by guiding the CNN training to focus on relevant clinical features and to forget confounding factors.

Potential impact

Multiple hospitals around the world, e.g. the Hospital of Leeds and the Cannizzaro Hospital in Catania, have shifted to a fully-digital clinical workflow for the analysis of tissue microscopy images of tumor biopsies and resections. The digital images are acquired by extremely high-resolution scanners and allow pathologists to inspect the images from any location, to participate in online tumor boards and to access the support given by AI systems for cancer diagnosis. The detection by CNNs of isolated tumor cells that occupy less than 0.008 % of the entire image may make the difference in avoiding false negatives in the identification of tumor metastases. As a result, this may change the final staging of the patient and improve the effectiveness of the cancer treatment.

The interpretability methods proposed in this work may facilitate the integration of CNNs in clinical routines, providing insights on up to what point and why a given CNN model is trustworthy and reliable for everyday use. Thanks to the techniques that I developed, physicians can better understand the decisions of "black-box" models. This is shown in my dissertation by the evaluation with user tests, which showed improved acceptance among physicians, and improved consistency and reliability of the proposed techniques with respect to the state-of-the-art methods. The analyses in this work, besides, uncovered some pittfalls that were corrected, providing the research community with more reliable and trustworthy CNNs than the already existing ones, particularly for the detection of tumorous regions in histology slides.

Potential effectiveness

As part of this work, I developed an open-access website (https://cadeval.p645.hevs.ch/) that allows people to interact with the digital slides, select a region of interest, and inspect the CNN prediction of tumor areas within the selected region. The visual interface allows to: (i) select one image from a list of available slides (ii) display the slide with the possibility to increase the magnification (iii) annotate one region and analyze the CNN prediction of tumor probability for each area in this region (v) generate explanations for the CNN predictions and compare them across multiple regions and explanation methods. A video-recorded demonstration of the tool functionalities is available at shorturl.at/uCP04 (as accessed in January 2022).

The interpretability techniques in this work detect the nuclei locations in the regions selected by the user and use this information to demonstrate that the CNN predictions are based on the pixels inside the contours of tumorous nuclei. With the method that I designed of Regression Concept Vectors (RCVs), I showed that the nuclei showing atypia in the morphology and texture are relevant for the model prediction. This approach demonstrated that the CNN decision-making is in line with the standards of clinical practice. Thanks to this interface, pathologists may interact with CNNs, use interpretability techniques and participate in user-studies that will help us to even better understand the requirements of AI integration in clinics.

Going beyond the purpose of generating explanations, I directly tackled the generalization deficiencies of existing models, proposing models that generalize better to new data sources and that can be applied to images from new institutions.