Caption-based explainable AI

Increase robustness by identifying dominant concepts

Graduate

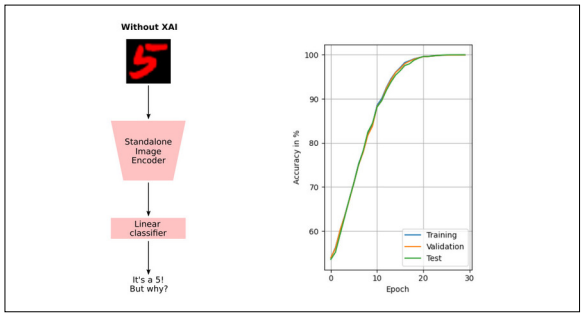
Patrick Koller

Initial Situation: An increasing number of novel and promising machine learning (ML) applications are taking the modern world by storm. Speech recognition, autonomous driving and recommender systems are just a few examples of ML's most recent mainstream breakthroughs. There is no evidence that this trend will stop any time soon. Every day we find exciting and complex applications that require advanced ML models, but with great power comes great responsibility. The fundamental property of ML models is that they are not explicitly programmed but learn from data instead. This attribute makes advanced ML models very powerful but challenging to interpret. This is especially challenging in high-stakes environments, e.g. in medicine, where a patient could suffer from incorrect predictions made by an ML model. Therefore, it is critical to deploy robust ML models. The science of interpreting ML models to understand their behavior and improve their robustness is called explainable artificial intelligence (XAI). One of the state-of-the-art XAI methods for computer vision problems is to generate saliency maps. A saliency map highlights the pixel space of an image that excites the model the most. However, this property could be misleading if spurious and salient features are present in overlapping pixel spaces.

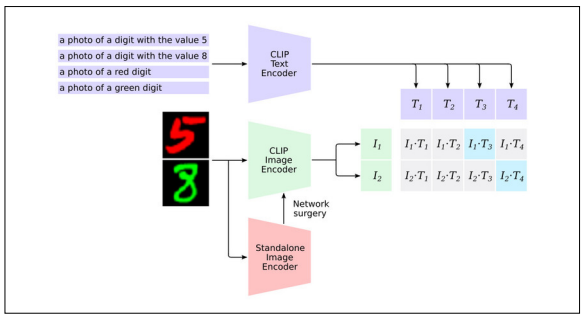
Approach / Technology: Introducing a modified version of the MNIST dataset with a color-encoded covariate shift between the training/validation/test datasets available during the development and a simulated real-world situation allows us to demonstrate the novel caption-based XAI method. The covariate shift is caused by a spurious feature (Color of the digit) in the same pixel space as the salient feature (Shape of the digit). An ML model is demonstrated to learn the more accessible spurious color feature instead of the salient shape feature. In a real-world situation, the digit's color assignments are assigned randomly. Therefore, the ML model fails in a real-world situation. A saliency XAI approach fails to explain this troubling situation since it highlights a part of the colored digit. A novel network surgery approach fuses the ML model to be explained with the contrastive language-image pre-training (CLIP) model. The resulting caption-based XAI model uses a set of captions to find the most descriptive text for a given image. This property enables the caption-based XAI method to express which concept the ML model focuses on instead of which pixel space.

Conclusion: This work introduces a new approach called the caption-based XAI method to explain convolutional neural networks. Using a novel network surgery method, a standalone model to be explained is incorporated into CLIP. The resulting XAI model can identify the dominant concept that contributes the most to the model's predictions. The most promising result is the superiority of the novel XAI method over saliency maps in situations where spurious and salient features are present in overlapping pixel spaces. The central thesis validated by this work is that a deeper understanding of the dominant concepts in convolutional neural networks is fundamental and can be used to improve the model's robustness. Our findings suggest that this novel XAI method should not just be seen as a pure debugging tool but as a necessary prerequisite before deploying any machine vision convolutional neural network model.

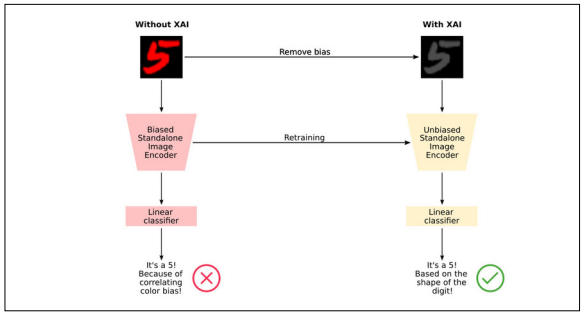
Understanding the reason for a model's high performance is critical before the deployment.
Own presentment



The caption-based explainable AI model identifies dominant concepts.
Own presentment



Knowing a machine learning model's dominant concept helps improve its robustness by removing bias.
Own presentment



Advisor
Prof. Dr. Guido Schuster

Co-Examiner
Prof. Dr. Aggelos Katsaggelos, Northwestern University, Evanston, IL

Subject Area
Software and Systems