Combining Segmentation and Argumentation to Explain Image Classification

Juan Puebla[1], Carlo Taticchi [2], Stefano Bistarelli [2], Martine Ceberio (Supervisor) [1]

[1]) Department of Computer Science, University of Texas at El Paso.
jlpuebla@miners.utep.edu, mceberio@utep.edu
[2]) Department of Mathematics and Computer Science, University of Perugia, Perugia, Italy. carlo.taticchi@unipg.it, stefano.bistarelli@unipg.it

Understanding the reasoning behind the prediction of a neural network is essential for transparency and trust in AI, especially in critical domains. In image classification, models such as ResNet achieve high accuracy but often function as black boxes, leaving users unable to interpret which visual components influence a decision. This is of particular relevance and concern when such tools are used to make a medical diagnosis.

In this work, we propose a Visual-Segmentation-Argumentation-based Explainable AI (XAI) pipeline designed to bridge this interpretability gap by identifying and quantifying the visual components most relevant to a model's prediction and using computational argumentation to generate human understandable explanations. The pipeline begins with image classification, followed by a Large Language Model (LLM) that generates a list of expected visual components for the predicted class. Component localization is performed using an object detection model, e.g., Grounding DINO, to detect and localize components in the image with bounding boxes. The Integrated Gradients method is then applied to measure each component's relative importance by aggregating pixel-level attributions within the detected regions. The resulting output highlights how specific components, such as "wheels" or "headlights" in a car image, contribute to the classification outcome, providing interpretable visual explanations. Finally, we build an argumentation framework to generate human-understandable explanations. The argumentation framework and explanation generation steps are currently in progress as we finish development. This combined approach lays the groundwork for a transparent component-level reasoning pipeline that enhances interpretability in image-based neural network decisions. Our preliminary results on the explanation of a car classification show that components such as wheels, tires, and windshield play a significant role in the model's prediction. These findings demonstrate the potential of our explanation pipeline and will be presented.