Increasing Computer Vision Models Interpretability

Boychev, Delyan (School: High School of Mathematics and Natural Sciences "Vasil Drumev")

With the perpetual increase of complexity of the state-of-the-art deep neural networks, it becomes a more and more challenging task to maintain their interpretability. Our work aims to evaluate the effects of adversarial training utilized to produce robust models - less vulnerable to adversarial attacks. It has been shown to make computer vision models more interpretable. Interpretability is as essential as robustness when we deploy the models to the real world. To prove there is a correlation between these two problems, we extensively examine the models using local feature-importance methods (SHAP, Integrated Gradients) and feature visualization techniques (Representation Inversion, Class Specific Image Generation). Standard models, compared to robust ones are less secure, and their learned representations are less meaningful to humans. Conversely, robust models focus on distinctive regions of the images which support their predictions. Moreover, the features learned by the robust model are closer to the real ones.