Providing a Method for Neural Networks to Justify Their Conclusions in Both Prediction and Classification Problems

Barrat, Robert

Neural networks are the heart of artificial intelligence and machine learning. They are computer programs that loosely act like a biological brain and they can learn to approximate functions based on data. Neural networks are used for two tasks, classifying data (sorting data into groups, like image recognition, classifying a paragraph based on the subject, classifying product reviews into positive, neutral, or negative, etc.) and prediction (predicting future data based on past data, like weather prediction, stock prediction, what the next note in a song should be, etc.) A major problem with neural networks is that although they can approximate nearly any function, studying their structure or the values in the networks won’t give you any insights on the function being approximated. Even a problem seemingly as simple as figuring out which input in your neural network matters the most or the least is still an open problem. My project seeks to make the inner workings of neural networks more transparent through two separate machine learning models that will allow neural networks to return justifications with their conclusions in both prediction and classification problems. The justifications provided by the machine learning models will not only strengthen the conclusions provided by the network, but will also provide insight into the "thought process" of the neural networks.