A Rate-Distortion Framework for Explaining Deep Neural Network Decisions

Jan Macdonald

November 14, 2019

We propose a rate-distortion framework for explaining neural network decisions. We formulate the task of determining the most relevant signal components for a classifier prediction as an optimisation problem. For the case of binary signals and Boolean classifier functions we show that it is hard to solve and to approximate. Finally, we present a heuristic solution strategy for deep ReLU neural network classifiers. We present numerical experiments and compare our method to other established methods.