# A rate-distortion framework for explaining neural network decisions

Stephan Waeldchen

June 28, 2019

Traditional machine learning models such as linear regression or decision trees allow for a straight-forward human interpretation of the model prediction. In contrast, the reasoning of highly nonlinear and parameter-rich neural networks remains generally inaccessible. In this talk, we present a rigorous approach to obtain interpretable neural network decisions. More precisely, we formulate the problem of determining the most relevant components of an input signal for a classifier prediction as an optimisation problem in a rate-distortion framework. We show that this problem is generally hard to solve and to approximate, which justifies the use of heuristic methods. We propose a problem relaxation together with a heuristic solution strategy for deep feed forward neural networks. Finally, we present numerical experiments and compare different explanation methods for two image classification data sets.