Making Interpretable Neural Networks More Explainable

Abstract

We present a novel interpretable neural network training framework that enforces additional explainability-related constraints. Firstly, the framework brings explainability to the forefront of interpretable neural networks by accentuating the learning process of a limited number of latent features (maximum of M where M is prescribed by the regulatory problem) that capture the most significant input interactions. Such latent features become more prominent at the time of decision making allowing one to determine with a higher level of confidence which latent feature or a combination of latent features led an interpretable neural network to a specific decision. Secondly, the framework prevents complex hidden layer co-adaptations being developed during training so that only a limited number of latent features need to cooperate to provide a response to the incoming inputs at the inference time. Thirdly, the framework leads to alignment of interpretable neural network architectures that meet the prescribed and preferred number of explanations associated with automated decision making based on industry specification of number of reasons M (e.g., 3-4 reasons in credit risk). Having trained an interpretable neural network where each latent feature has only 2-input connections and having ensured that only a limited number of latent features (e.g., 3-4) can cooperate to provide a response to the incoming input data during the inference time, we can explain these latent features using a set of human traceable and readable reasons derived from each latent feature's activation states representing specific regions of an activation function.

Authors & Affiliations

Dr. Scott Zoldi[1], Krzysztof Nalborski[1]
[1]FICO, San Diego, USA