Contribution ID: **58** Type: **Regular talk**

Decorrelation via Disentanglement

Friday 23 October 2020 16:10 (20 minutes)

Abstract

Invariance of learned representations of neural networks against certain sensitive attributes of the input data is a desirable trait in many modern-day applications of machine learning, such as precision measurements in experimental high-energy physics. We propose to use the ability of variational autoencoders to learn a disentangled latent representation to achieve the desired invariance condition. The resulting latent representation may be used for arbitrary downstream tasks without exploitation of the protected variable. We demonstrate the effectiveness of the proposed technique on a representative study of rare B decays at the Belle II experiment.

Introduction

In searches for new physics in high-energy physics, experimental analyses are primarily concerned with physical processes which are rare or hypothesized. To claim a statistically significant discovery or exclusion of new physics when studying such decays, it is necessary to maintain an appropriate signal to noise ratio. However, the na\"ive application of standard classification methods is liable to raise poorly understood systematic effects and ultimately degrade the significance of the final measurement.

To understand the origin of these systematic effects, we note that there are certain protected variables in experimental analyses which should remain unbiased by the analysis procedure. Variables used to parameterize proposed models of new physics and variables used to model background contributions to the total measured event yield fall into this category. Systems responsible for separating signal from background events achieve this by sampling events with signal-like characteristics from all candidate events. If this procedure introduces sampling bias into the distribution of protected variables, this introduces systematic effects into the analysis which are difficult to characterize. Ultimately, we would like to build a classifier that makes decisions independently of certain physically important observables - such that the original distribution of these observables is preserved for any subsample of the data. This problem is commonly referred to as classifier "decorrelation" with respect to the observables of interest.

We address this task as an optimization problem of finding a representation of the observed data that is invariant to the given protected quantities. This representation should satisfy two competing criteria. Firstly, it should contain all relevant information about the data so that it may be used as a proxy for arbitrary downstream tasks, such as inference of unobserved quantities or prediction of target variables. Secondly, it should not be informative of the given protected quantities, so that downstream tasks are not influenced by these variables. If the protected quantities to be censored from the intermediate representation contain information that can improve the performance of the downstream task, it is likely that removing this information will adversely affect this task. The challenge lies in balancing both objectives without significantly compromising either requirement.

This work approaches the problem from a latent variable model perspective, in which additional unobserved variables are introduced which explain the interaction between different attributes of the observed data. These latent variables can be interpreted as a more fundamental, lower-dimensional representation of the original high-dimensional unstructured data. By appropriately constraining the structure of this latent space, we demonstrate we can isolate the influence of the protected variables into a latent subspace. This allows downstream tasks to only access a relevant subset of the learned representation without being influenced by protected attributes of the original data.

Author: TAN, Justin

Co-author: Prof. URQUIJO, Phillip (University of Melbourne)

Presenter: TAN, Justin

Session Classification: Workshop

Track Classification: 2 ML for analysis : Application of Machine Learning to analysis, event classification and fundamental parameters inference