Contribution ID: **22** Type: **not specified**

Disentangling representations in Restricted Boltzmann Machines without adversaries

Thursday, 29 June 2023 15:00 (25 minutes)

A goal of unsupervised machine learning is to build representations of complex high-dimensional data, with simple relations to their properties. Such disentangled representations make it easier to interpret the significant latent factors of variation in the data, as well as to generate new data with desirable features. The methods for disentangling representations often rely on an adversarial scheme, in which representations are tuned to avoid discriminators from being able to reconstruct information about the data properties (labels). Unfortunately, adversarial training is generally difficult to implement in practice. In this talk, I will describe a simple, effective way of disentangling representations without any need to train adversarial discriminators, and apply our approach to Restricted Boltzmann Machines, one of the simplest representation-based generative models. Our approach relies on the introduction of adequate constraints on the weights during training, which allows us to concentrate information about labels on a small subset of latent variables. The effectiveness of the approach is illustrated with four examples: the CelebA dataset of facial images, the two-dimensional Ising model, the MNIST dataset of handwritten digits, and the taxonomy of protein families. In addition, we show how our framework allows for analytically computing the cost, in terms of the log-likelihood of the data, associated with the disentanglement of their representations.

Primary author: FERNANDEZ DE COSSIO DIAZ, Jorge (ENS PARIS)

Co-author: Dr MONASSON, Remi

Presenter: FERNANDEZ DE COSSIO DIAZ, Jorge (ENS PARIS)