Tutorial: Fairness and Neural Networks – Abstract

Roberto Esposito[1]

[1]Department of Computer Science, Università di Torino, Corso Svizzera 185, 10149, Torino, Italy

Abstract

Recent years have witnessed a Cambrian explosion of tools and techniques able to tackle problems that were only solvable by humans up to a few years ago; deep learning in particular is accumulating astounding successes at a breakneck pace in both research and applications: from helping in recovering photos by their descriptions on devices used by billions of people, to providing tools for investigating the depths of the visible universe. It is then unfortunate that these very models are utterly inscrutable and inaccessible to human understanding.

While in many cases the difficulty of understanding these models does not matter, in some very specific contexts it creates problems that are important and hard to solve. Big and small companies are, in fact, investing in these technologies and deploying them in contexts that directly impact human well-being such as loan applications, candidate selection for job offers, and evaluating the chance of re-offending for people who committed crimes. In all these cases using inscrutable models poses difficult ethical issues related to the risk of discrimination of people belonging to protected groups.

In absence of techniques allowing to solve the problem by explaining the decisions of these models, the fair ML literature focused on approaches based on the construction of non-discriminating models. In this context, neural networks provide both challenges and opportunities. In this tutorial I will contextualize the problem, show how there exists many different (and contrasting) definitions of fairness, and introduce some of the state-of-the-art approaches in this field. I will focus in particular on methods targeting neural networks, specifically on methods that constrain the representation learnt by the network to be fairer with respect to given sensible attributes.

Keywords

Fairness, Neural Networks, Representation Learning

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy

roberto.esposito@unito.it (R. Esposito)