Vision Manuscript

skww86
Computer Science
Durham University
Durham, United Kingdom
skww86@durham.ac.uk

I. My opinion on the paper

In this manuscript, I will be discussing my thoughts and opinions on the research paper "A Framework for Understanding Unintended Consequences of Machine Learning" Suresh et al [1]. The authors, writing at a time when machine learning is starting to become more prevalent in the world, discuss the notion of how the problem of bias in machine learning is often treated too generically. This has serious consequences, because it is often assumed that there are generic causes and fixes for mitigating bias, which are then blindly applied to real life applications without further considerations of the exact circumstances surrounding the particular application in question. Often, this means that mitigation steps that are implemented into certain applications have little to no effect, because the precise nature of what is causing the bias was not established, and thus inappropriate mitigation was implemented. With bias in artificial intelligence being a fairly new issue, the authors have identified that there is a lack of clear terminology associated with the topic in the industry, which is leading to miscommunication and misunderstanding when it comes to establishing what the problem is. The paper seeks to establish unified categorical definitions of the various types of bias that can occur, and considers this from a very holistic perspective - ranging from contextual societal issues, such as the existence of a patriarchy, across the entire model pipeline, from data generation, model building, and implementation - the bias could occur in multiple different places across this timeline. I absolutely agree that it is vital for the nature of the bias to be correctly pinpointed when it arises in an application. With machine learning being such an up and coming industry, the community is very collaborative, and everyone is learning from our peers and sharing ideas – which is great. But we must not let this be our downfall. For example, the paper discusses how a smile detection system could be giving a higher false negative rate for women. The developer fixes this by implementing more images of women into the model training set, which mitigates the bias. A different developer is noticing that their algorithm to predict job candidate suitability is biased against female applicants. Due to this developer's understanding of bias in artificial intelligence being a singular issue with a generic fix, they read about the former developer's fix to the smile detection system and implement the same idea, but it ends up doing nothing to mitigate bias. This is because the issue with the smile detection system was an example of representation bias, whilst the job suitability algorithm was suffering from measurement bias. These are two entirely different things, and without clear distinction between the various categories and clear communication between developers on the nature of the bias they are mitigating, there will be more people who fall into this trap.

II. Future of bias in AI

As discussed before, currently we all too often see the wrong approach and solution applied to applications in the real world which suffer from bias, due to a misunderstanding of the exact nature of the bias that has occurred. If this continues, we will struggle to mitigate bias in our machine learning applications. However, the authors hope that by defining a clear dictionary of bias types (historical bias, representation bias, measurement bias, evaluation bias, aggregation bias, deployment bias), the industry can shift towards clearer and more productive communication, by using shared terminology, and developers can be entirely in the know over exactly how and what they need to change in their application. So in the previous example, the developer of the smile detection system can say that they identified and mitigated representation bias, and the developer of the job suitability algorithm can identify that their issue is due to measurement bias, and thus know that the fix for the smile detection system will not be relevant to them, and so won't just copy the solution. However, this mentality shift will require the industry to have a completely overhauled understanding of what 'fairness' really means from a low-level perspective. The notion of various buzz words/phrases such as 'training data bias' need to be eradicated, and I think that developers of all machine learning models should break down their pipeline in the first instance, so that they can consider each individual aspect from the forefront during data collection and development, and how they might be significant to their application specifically. For instance, when they are initially collecting their data to be used in their data set, are they considering how historical bias could be influencing their data, and what the impact of this might be? Or is this just ignored and then once the model has been fully developed, only then is the bias in the results being considered? There are quite clearly several key decisions that need to be made by the developer throughout the pipeline to address this topic. With all of this said though, even once these bias categories become commonplace, the problem isn't fully solved. There is never going to be a 'fit for all' solution to each type of bias – it will be relative to the nature of the application and it is crucial that developers understand this. An image recognition application's fix to mitigate aggregation bias is not necessarily going to be the same as a credit risk prediction application's fix for aggregation bias, which is why it's so important for developers to establish how each category of bias relates to *their* application specifically. Once this has been established, we can continue to share ideas and collaborate with colleagues to help mitigate varying biases in our machine learning applications in a genuinely useful and productive manner.

References

[1] H. Suresh; J. Guttag, “A Framework for Understanding Unintended Consequences of Machine Learning” in press