Using ”Ethically” Toolkit to Measure and Debias Gender Bias in Words Embedding

CL Colloquium - University of Potsdam

Shlomi Hod

University of Potsdam, Potsdam, Germany
shlomi.hod@uni-potsdam.de

Abstract

There is a fundamental gap between ethics, which lays in the kingdom of human values, and artificial intelligence, which is, broadly speaking, is a mathematical model. Nowadays many intelligent systems are being deployed daily, and data scientists are taking decisions in the mathematical domain, that may have ethical and social implications. It raise the questions of how to embed ethics in an intelligent system, and how to audit the ethics of AI.

In this talk, we will demonstrate methods of auditing the bias of a typical building block of machine learning models that work with natural languages - word-embeddings. We do so by using an open-source Python package ”Ethically” (https://docs.ethically.ai). ”Ethically” implements state-of-the-art techniques of auditing and mitigating bias and fairness of machine learning systems. In particular, we will show how to measure and visualize the gender bias in various common words embeddings, and how to debias them, using Ethically toolkit and based on the work of [Bolukbasi et al., 2016] and [Caliskan et al., 2017].

This presentation will serve as a window to the fairness research in Machine Learning, which is partially inspired by findings rooted in Psychology, such as the Implicit-association test.

Python Jupyter Notebook as tutorials will be supplied to the audience.

References

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

[Caliskan et al., 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.