Biased Embeddings from Wild Data: Measuring, Understanding and Removing

Adam Sutton, Thomas Lansdall-Welfare, Nello Cristianini
Intelligent Systems Laboratory, University of Bristol, Bristol, BS8 1UB, UK
{adam.sutton, thomas.lansdall-welfare, nello.cristianini}@bris.ac.uk

***Abstract*—Many modern Artificial Intelligence (AI) systems make use of data embeddings, particularly in the domain of Natural Language Processing (NLP). These embeddings are learnt from data that has been gathered “from the wild” and have been found to contain unwanted biases. In this paper we make three contributions towards measuring, understanding and removing this problem. We present a rigorous way to measure some of these biases, based on the use of word lists created for social psychology applications; we observe how gender bias in occupations reflects actual gender bias in the same occupations in the real world; and finally we demonstrate how a simple projection can significantly reduce the effects of embedding bias. All this is part of an ongoing effort to understand how trust can be built into AI systems.**

I. Introduction

With the latest wave of learning models taking advantage of advances in deep learning [21], [22], [23], Artificial Intelligence (AI) systems are gaining widespread publicity, coupled with a drive from industry to incorporate intelligence into all manner of processes that handle our private and personal data, giving them a central position in our modern-day society.

This development has lead to demand for fairer AI, where we wish to establish trust in the automated intelligent systems by ensuring that systems represent us fairly and transparently. However, there has been growing concern about potential biases in learning systems [1], [6] which can be difficult to analyse or query for explanations of their predictions, leading to an increasing number of studies investigating the way black-box systems represent knowledge and make decisions [7], [9], [11], [19], [20]. Indeed, principled methods are now required that allow us to measure, understand and remove biases in our data in order for these systems to be truly accepted as a prominent part of our lives.

In the domain of text, many modern approaches often begin by embedding the input text data into an embedding space that is used as the first layer in a subsequent deep network [4], [14]. These word embeddings have been shown to contain the same biases [3], due to the source data from which they are trained. In effect, biases from the source data, such as in the differences in representation for men and women, that have been found in many different large-scale studies [5], [10], [12], carry through to the semantic relations in the word embeddings, which become baked into the learning systems that are built on top of them.

In this paper, we make three contributions towards addressing these concerns. First we propose a new version of the Word Embedding Association Tests (WEATs) studied in [3], designed to demonstrate and quantify bias in word embeddings, which puts them on a firm foundation by using the Linguistic Inquiry and Word Count (LIWC) lexica [17] to systematically *detect* and *measure* embedding biases.

With this improved experimental setting, we find that European-American names are viewed more positively than African-American names, male names are more associated with work while female names are more associated with family, and that the academic disciplines of science and maths are more associated with male terms than the arts, which are more associated with female terms. Using this new methodology, we then find that there is a gender bias in the way different occupations are represented by the embedding. Furthermore, we use the latest official employment statistics in the UK, and find that there is a correlation between the ratio of men and women working in different occupation roles and how those roles are associated with gender in the word embeddings. This suggests that biases in the embeddings reflect biases in the world.

Finally, we look at methods of *removing* gender bias from the word embeddings. Having established that there is a direction in the embedding space that correlates with gender, we use a simple orthogonal projection to remove that dimension from the embedding. After projecting the embeddings, we investigate the effect on bias in the embeddings by considering the changes in associations between the words, demonstrating that the associations in the modified embeddings now correlate less to UK employment statistics among other things.

II. Methodology

A. Word Embedding

A word embedding is a mapping of words into an n-dimensional vector space. Given a corpus of text, a word embedding can be created that will translate that corpus into a set of semantic vectors representing each word. Each word that appears in the corpus will be represented by an n-dimensional vector to indicate its position within the embedding.

This embedding has a set of features that can be used in natural language processing methods. The nearest neighbours of a word will be other words that have similar linguistic or semantic meaning, when comparing words using a measurement such as cosine similarity. There are also linear substructures within the word embeddings that can explain how multiple words are related to each other, making it