Understanding the Origins of Bias in Word Embeddings

Marc-Etienne Brunet[*,1,2], Colleen Alkalay-Houlihan[*,1], Ashton Anderson[1,2], and Richard Zemel[1,2]

[*] *Equal contributors*
[1] *University of Toronto, Department of Computer Science*
[2] *Vector Institute for Artificial Intelligence*

Abstract

The power of machine learning systems not only promises great technical progress, but risks societal harm. As a recent example, researchers have shown that popular word embedding algorithms exhibit stereotypical biases, such as gender bias. The widespread use of these algorithms in machine learning systems, from automated translation services to curriculum vitae scanners, can amplify stereotypes in important contexts. Although methods have been developed to measure these biases and alter word embeddings to mitigate their biased representations, there is a lack of understanding in how word embedding bias depends on the training data. In this work, we develop a technique for understanding the origins of bias in word embeddings. Given a word embedding trained on a corpus, our method identifies how perturbing the corpus will affect the bias of the resulting embedding. This can be used to trace the origins of word embedding bias back to the original training documents. Using our method, one can investigate trends in the bias of the underlying corpus and identify subsets of documents whose removal would most reduce bias. We demonstrate our techniques on both a New York Times and Wikipedia corpus and find that our influence function-based approximations are extremely accurate.

Keywords: Bias, Word Embeddings, Interpretability

1. Introduction

As machine learning algorithms play ever-increasing roles in our lives, there are ever-increasing risks for these algorithms to be systematically biased [22, 21, 11, 6, 8]. An ongoing research effort is showing that machine learning systems can not only reflect human biases in the data they learn from, but also magnify these biases when deployed in practice [19]. With algorithms aiding critical decisions ranging from medical diagnoses to hiring decisions, it is important to understand the nature and sources of these biases.

In recent work, researchers have uncovered an illuminating example of bias in machine learning systems: Popular word embedding methods such as word2vec [14] and GloVe [16] acquire stereotypical human biases from the text data they are trained on. For example, they disproportionately associate male terms with science terms, and female terms with

Accompanying code available at *https://github.com/mebrunet/understanding-bias*
Contact *mebrunet@cs.toronto.edu, colleen@cs.toronto.edu, ashton@cs.toronto.edu, zemel@cs.toronto.edu*

art terms [1, 4]. Deploying these word embedding algorithms in practice, for example in automated translation systems or as hiring aids, runs the serious risk of perpetuating problematic biases in important societal contexts. Furthermore, this problem is especially pernicious because these biases can be difficult to detect. For example, word embeddings were in broad industrial use before these stereotypical biases were found.

In this work, we develop a technique for understanding the origins of bias in word embeddings. Given a bias metric and a word embedding trained on some corpus, our method identifies how perturbing the training corpus will affect the resulting bias. This naturally applies at the document level: given any document in the training corpus, our method can accurately approximate how its removal would affect the bias of the embedding. Naively, this can be done directly by removing the document and re-training an embedding on the perturbed corpus. But this approach comes at a prohibitive computational cost, limiting the number of perturbations that can be studied. Our method provides a highly efficient alternative, enabling the impact of *every* document in the corpus to be analyzed.

By calculating the document-level differences in bias across interesting subsets of documents, an analyst can better understand the origins of bias in word embeddings. These insights could be useful for a variety of important tasks. In some applications, finding the most influential documents on word embedding bias could surface potentially problematic portions of the training corpus. Using our technique, one could also find examples of documents that are extremely bias-correcting. Also, our method enables the study of how bias varies across various dimensions of a corpus — for example, how bias changes over time.

The main idea behind our method is to predict how perturbing the input corpus changes the bias in the resulting word embedding. To do this, we decompose the problem into two main subproblems: first, understanding how perturbing the training data changes the learned word embedding; and second, how changing the word embedding affects its bias. The latter subproblem is straightforward, and our central technical contributions solve the former. We calculate an approximation to the GloVe algorithm's loss function, and then apply influence functions from robust statistics to this approximation. Our modification ensures that the Hessian in the influence function calculation is positive definite and block diagonal, which in turn allows us to accurately and efficiently compute how perturbations in the input corpus affect the bias metric using influence functions. We can then identify co-occurrences that most increase embedding bias, as well as approximate the effect of removing documents from (or adding documents to) the input corpus.

We demonstrate our technique through experimental results, initially on a simplified corpus of Wikipedia articles in broad use [20], and then on a New York Times corpus [18]. We use a previously proposed measure of word embedding bias as our bias metric, but our technique is generalizable to other metrics. Across a range of experiments, we find that our method's predictions of how perturbing the input corpus will affect the bias of the embedding are extremely accurate.

2. Related work

Word embeddings are compact vector representations of words learned from a training corpus, and are actively deployed in a number of domains. They not only preserve statistical relationships present in the training data, generally placing commonly co-occurring words