Paper Title: Using Adversarial Attacks to Reveal the Statistical Bias in Machine Reading Comprehension Models

Paper Link: https://shorturl.at/pyQ37

1. Summary

1.1 Motivation/Purpose: Language models that have been pre-trained have demonstrated performance comparable to that of humans on numerous Machine Reading Comprehension (MRC) tasks. This indicates that it is possible for these models to comprehend language and respond to inquiries in a manner analogous to that of human beings. Nevertheless, the veracity of these models' language comprehension and whether they are merely capitalizing on statistical biases present in the datasets used for training purposes remain uncertain. This raises concerns as it implies that the models' ability to generalize to novel situations or provide answers to inquiries not expressly stated in the training data may be compromised. The objective of the authors was to devise a technique for assessing the statistical biases present in MRC models. They had high hopes that this approach would aid in the detection and resolution of these biases, thereby enhancing the dependability and credibility of MRC models.

1.2 Contribution: Highlighting the presence of statistical biases within MRC models is one of the most significant contributions. Existing MRC models are vulnerable to statistical biases, and their performance can be substantially compromised when adversarial methods are employed to attack them, as demonstrated by the authors. It appears that MRC models may not be as dependable as was previously believed; thus, this finding is significant. MRC model statistical bias evaluation methodology development. Assessing the statistical biases of MRC models is facilitated by the authors' suggested methodology. This technique can also be utilized to monitor the progression of bias over a period of time and to identify models that are especially prone to it. Statistical bias reduction methodology for MRC models. In order to reduce statistical biases in MRC models, the authors' augmented training procedure is a promising strategy. A number of well-known MRC models have been demonstrated to be effectively biased-reduced using this technique.

1.3 Methodology:

1. From the RACE dataset, the authors extracted a set of multiple-choice questions and their corresponding responses.
2. In order to launch an adversarial assault, the authors inserted superfluous phrases into multiple-choice questions.
3. The efficacy of multiple MRC models was assessed by the authors using the adversarial attack dataset.
4. An augmented training procedure was proposed by the authors as a means to mitigate the statistical biases present in MRC models.

5. By training multiple MRC models with the augmented training data, the authors assessed the efficacy of the augmented training method.

1.4 Conclusion: Adversarial attacks, which add meaningless words to questions, are very good at fooling MRC models. When you use the augmented training method, you add words that don't belong to the training data. This can help make MRC models more resistant to these attacks. It seems that MRC models might not be as accurate as we thought, which is a big deal. The writers' work shows how important it is to make MRC models that are stronger and less likely to be skewed by statistics.

2. Limitations

2.1 First Limitation: Effectiveness of the adversarial attack method is restricted to a specific set of statistical biases.As a consequence, the method's capability to identify every style of statistical bias present in MRC models may be limited.

2.2 Second Limitation: It's possible that the augmented training method won't work against all kinds of statistical errors. In other words, the method might not be able to get rid of all statistical errors in MRC models.

3. Synthesis/Future Work

1. The objective of the authors is to advance adversarial attack techniques in order to identify a broader spectrum of statistical biases. This will aid in the identification of additional forms of statistical biases in MRC models and enhance their robustness.
2. The authors also intend to investigate regularization and data augmentation as additional techniques for reducing statistical biases in MRC models. These techniques have the potential to further mitigate statistical biases and enhance the performance of MRC models.
3. Additionally, the authors state their intention to distribute their code and data to the public in order to enable others to duplicate their findings and expand upon their research.