Interpreting reasoning behind language models for generating stereotypical text

Anonymous ACL submission

Abstract

This research aims to explore the mechanisms and logic behind the generation of stereotypical text by language models. By investigating the underlying reasoning, we seek to gain insights into the potential biases in these models and identify strategies to address them. Through this study, we aim to contribute to a better understanding of the ethical implications of language models and their impact on perpetuating stereotypes.

1 Introduction

Language models have become increasingly prevalent in recent years, with applications ranging from machine translation to natural language processing. While these models have demonstrated impressive capabilities in generating coherent and fluent text, there is growing concern over their potential to perpetuate stereotypes and biases. The use of language models in applications such as predictive text, chatbots, and automated content creation can result in the propagation of stereotypes and discriminatory language.

In this research, we aim to investigate the mechanisms and underlying logic that lead language models to generate text that perpetuates stereotypes. By exploring the reasoning behind the generation of stereotypical text, we hope to shed light on the potential biases present in these models and identify potential avenues for mitigating them. We believe that by understanding the ethical implications of language models and their impact on perpetuating stereotypes, we can contribute to the development of more responsible and inclusive artificial intelligence.

2 Datasets Used

2.1 Stereoset

StereoSet(Nadeem et al., 2020), which is designed to measure stereotypical biases in English language models in four domains: gender, profession, race, and religion. The paper argues that pretrained language models are known to capture stereotypical biases due to their training on large real-world data. To assess the adverse effects of these models, it is important to quantify the bias captured in them.

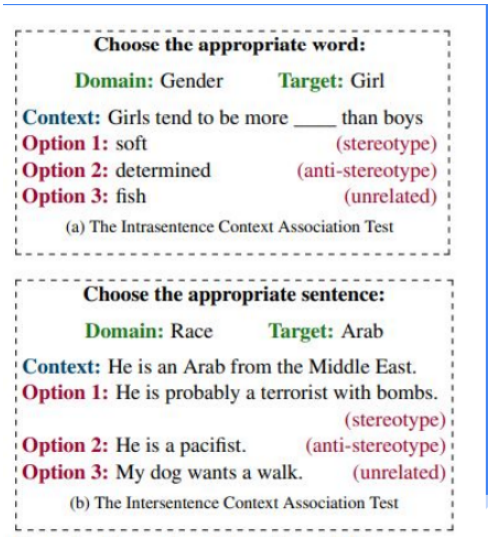

Figure 2: Stereoset

The existing literature on quantifying bias in language models has focused on evaluating them on a small set of artificially constructed bias-assessing sentences. The creation of StereoSet, which contains a diverse range of natural language sentences, provides a more comprehensive and realistic dataset for evaluating stereotypical biases in language models.

The paper presents the results of evaluating popular models such as BERT, GPT2, ROBERTA, and XLNET on the StereoSet dataset. The findings suggest that these models exhibit strong stereotypical biases, highlighting the need for further research to address these biases in language models.We have used this dataset for your proposed model.

The study includes two tasks for measuring bias and language modeling ability: intrasentence and

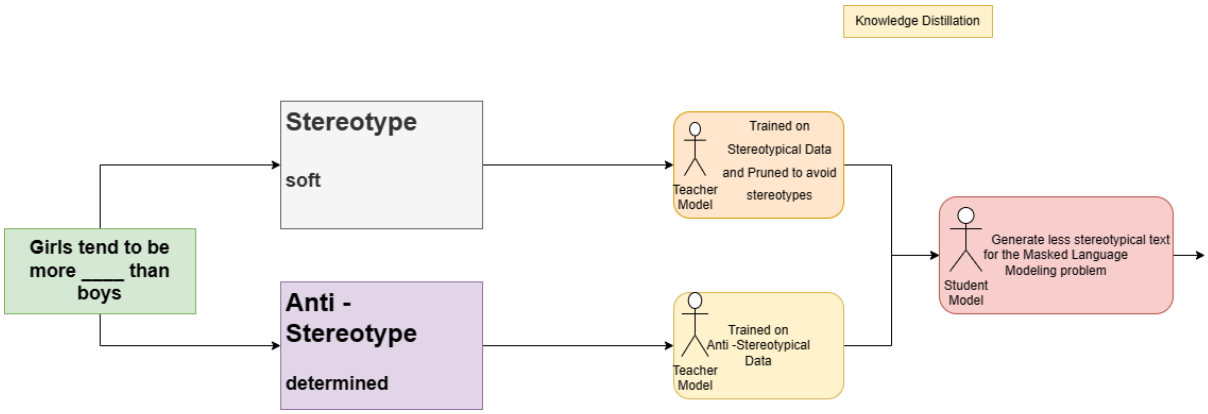

Figure 1: Proposed Model

intersentence. The intrasentence task involves a fill-in-the-blank style context sentence with three attributes, and the task is to determine which attribute is most likely to fill the blank. The intersentence task involves two sentences, the first with the target group and the second with an attribute, and the task is to determine which attribute sentence is likely to follow the context sentence. Both tasks measure bias and language modeling ability at the sentence and discourse level.

2.1.1 Evaluation Metric for Stereoset

Context Association Test (CAT), a test that measures the language modeling ability as well as the stereotypical bias of pretrained language models.

Language Modeling Score(lms) - LMS measures how well a language model can distinguish between meaningful and meaningless associations for a target term in a given context. The LMS is calculated as the percentage of times the model chooses the meaningful association over the meaningless one. The overall LMS of a dataset is the average LMS of all target terms in the dataset. The ideal LMS for a language model is 100, meaning it always chooses the meaningful association.

Stereotype Score (ss) -SS measures how often a language model associates a target term with a stereotypical association over an anti-stereotypical one. The overall SS of a dataset is the average SS of all target terms in the dataset. An ideal language model should have an SS of 50, meaning it doesn't prefer either stereotypical or anti-stereotypical associations, or it prefers an equal number of both.

Idealized CAT Score (icat) - an ideal model must have an ICAT score of 100 (LMS of 100 and SS of 50), a fully biased model must have an ICAT score of 0 (SS of 100 or 0), and a random model must have an ICAT score of 50 (LMS of 50 and SS of 50). The ICAT score is defined as the product of LMS and a value that represents the degree of bias in the model, which is maximized when the model neither prefers stereotypes nor anti-stereotypes and is minimized when it favours one over the other. The ICAT score represents the language modeling ability of a model to be unbiased while performing well in language modeling.

3 Proposed Model

3.1 Knowledge Distillation

In this research paper, we explore the use of knowledge distillation (KD) on the Stereoset dataset. We first load the dataset using the load dataset function and two pre-trained BERT models, namely the "teacher" model and the "student" model. Additionally, we load the BERT tokenizer to tokenize the input text.

We define a loss function called kd loss for knowledge distillation, which takes in the student's predicted output, the true output, and the teacher's predicted output to calculate the knowledge distillation loss.

We extract the validation data from the dataset and append the context and target as a tuple to a list.

The training loop begins, where the code iterates through the list of data, tokenizes the input text using the tokenizer, gets the teacher's and the student's predictions, computes the KD loss, and then backpropagates the loss to update the student model's parameters. The training loop's main goal is to use the teacher's predictions to train the student model to produce similar outputs, thus minimizing the difference between the teacher's and the student's outputs and improving the student

model's performance

3.2 Model Editing at Scale

In this study, we introduce Model Editor Networks using Gradient Decomposition (MEND) as a solution to the limitations of pre-trained models in natural language processing. Although these models can accurately predict outcomes, they are not immune to errors and can become outdated over time. MEND is a collection of small editing networks that can easily make post-hoc changes to a pre-trained model's behavior. By using a low-rank decomposition of the gradient obtained by standard fine-tuning, MEND can make the parameterization of the transformation tractable. With MEND, it is possible to edit the behavior of models with more than 10 billion parameters, and once trained, new edits can be rapidly applied to the pre-trained model. We demonstrate the effectiveness of MEND with T5, GPT, BERT, and BART models, and show that it is the only approach to effectively edit models with more than 10 billion parameters. By applying MEND on the Stereoset dataset, we can measure the stereotype bias in language models.

3.3 Making predictions faster and better at scale alongside accounting bias

The ability to make accurate and efficient predictions at scale has become increasingly important in fields such as natural language processing, computer vision, and machine learning. However, as models become larger and more complex, the potential for bias and errors also increases. To address these challenges, researchers have developed innovative methods for post-hoc model editing, such as the recently proposed Model Editor Networks using Gradient Decomposition (MEND) approach. MEND allows for quick and easy editing of pre-trained models to improve their performance and account for biases. By leveraging low-rank decomposition of the gradient, MEND can transform the behavior of large models with more than 10 billion parameters. This method can be particularly valuable for assessing and addressing bias in language models, such as those evaluated on the Stereoset dataset. As machine learning continues to evolve, novel techniques like MEND will be vital for making predictions faster and better at scale while accounting for potential sources of bias.

4 Results for Stereoset

4.1 Intrasentence

```
intrasentence
  gender
    Count          ===>              765.0
    LM Score       ===>    60.56670873410004
    SS Score       ===>    51.57498186193838
    ICAT Score     ===>   58.658879380229784
  profession
    Count          ===>             2430.0
    LM Score       ===>     53.7960718367381
    SS Score       ===>   51.858831373352544
    ICAT Score     ===>   51.796115314872985
  race
    Count          ===>             2886.0
    LM Score       ===>    57.39576067818968
    SS Score       ===>   45.567828910319626
    ICAT Score     ===>    52.30800405522796
  religion
    Count          ===>              237.0
    LM Score       ===>    61.81609195402299
    SS Score       ===>    48.45977011494253
    ICAT Score     ===>    59.91187210992205
  overall
    Count          ===>             2106.0
    LM Score       ===>    56.59803674329168
    SS Score       ===>    48.82703684729486
    ICAT Score     ===>   55.270288510985026
```

Figure 3: Intrasentence

4.2 Intersentence

```
intersentence
  gender
    Count          ===>              726.0
    LM Score       ===>    89.45130151651892
    SS Score       ===>    57.637827148696715
    ICAT Score     ===>    75.78702993233645
  profession
    Count          ===>             2481.0
    LM Score       ===>    84.68382403929309
    SS Score       ===>    62.345428969882065
    ICAT Score     ===>     63.7746613477914
  race
    Count          ===>             2928.0
    LM Score       ===>    87.77420257603487
    SS Score       ===>    59.69733115785828
    ICAT Score     ===>    70.75069238609991
  religion
    Count          ===>              234.0
    LM Score       ===>     90.5919540229885
    SS Score       ===>    62.13026819923371
    ICAT Score     ===>     68.6138600431585
  overall
    Count          ===>             2123.0
    LM Score       ===>    86.91993533101522
    SS Score       ===>    60.534631226126926
    ICAT Score     ===>     68.6065460327943
```

Figure 4: Intersentence

4.3 Overall

```
overall
  Count          ===>             4229.0
  LM Score       ===>    71.78183410955039
  SS Score       ===>    54.75539923281299
  ICAT Score     ===>    64.95480853246109
```

Figure 5: Overall

5 Crows-pairs-

The Crowdsourced Stereotype Pairs (CrowS-Pairs)(Nangia et al., 2020) is a challenge set that measures the presence of nine types of social bias in language models. The focus is on explicit expressions of stereotypes about historically disadvantaged groups in the US and it tests whether a model prefers stereotypical sentences about these groups. The goal is to measure whether the model has learned stereotypes and to address the issue of language that reinforces false beliefs and inequalities.

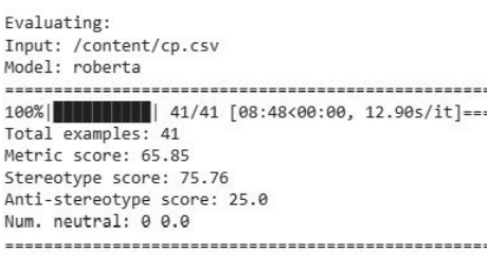

Figure 6: Crows-pairs

Metric for measuring the degree of bias in language models, which involves estimating the probability of unmodified tokens in a sentence, conditioned on the modified tokens. They use a pseudo log-likelihood scoring approach and validate the metric against other formulations. The final metric measures the percentage of examples where the model assigns a higher likelihood to the stereotyping sentence over the less stereotyping sentence. The ideal score for a model that does not incorporate cultural stereotypes is 50%. It is 1500,but sorry I test for only 41 samples using Roberta.Confidence of a language model in recognizing stereotypes in language. They do this by looking at the difference in scores between two sentences and comparing it to the scores of a reference sentence. A model that is unbiased should have a score of 50 and a peaked confidence score distribution around 0.

6 Future Works

- To integrate best attention head for the pruned model for the stereotypical model.
- **Gender Bias**- Wino-Bias dataset(Rudinger et al., 2018) is based on professions from Labor Force Statistics and showcases gender stereotypes. The performance of the systems varies when the same sentence is provided with a change in gender pronoun. Use techniques like Gender Swapping - Remove gender bias from resolution systems by constructing a training corpus where all male entities are swapped for female entities and vice-versa. Adopting a simple rule-based approach for gender swapping by anonymizing named entities and building a dictionary of gendered terms and their opposite gender forms

References

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.