Fairness-Aware Structured Pruning in Transformers

Abdelrahman Zayed[1,2], **Gonçalo Mordido**[1,2], **Samira Shabanian**[3], **Ioana Baldini**[4], **Sarath Chandar**[1,2,5]

[1]Mila - Quebec AI Institute
[2]Polytechnique Montreal
[3]Independent Researcher
[4]IBM Research
[5]Canada CIFAR AI Chair

{zayedabd,sarath.chandar}@mila.quebec, {s.shabanian,goncalomordido}@gmail.com, {ioana}@us.ibm.com

Abstract

The increasing size of large language models (LLMs) has introduced challenges in their training and inference. Removing model components is perceived as a solution to tackle the large model sizes, however, existing pruning methods solely focus on performance, without considering an essential aspect for the responsible use of LLMs: model fairness. It is crucial to address the fairness of LLMs towards diverse groups, such as women, Black people, LGBTQ+, Jewish communities, among others, as they are being deployed and available to a wide audience. In this work, first, we investigate how attention heads impact fairness and performance in pre-trained transformer-based language models. We then propose a novel method to prune the attention heads that negatively impact fairness while retaining the heads critical for performance, *i.e.* language modeling capabilities. Our approach is practical in terms of time and resources, as it does not require fine-tuning the final pruned, and fairer, model. Our findings demonstrate a reduction in gender bias by 19%, 19.5%, 39.5%, 34.7%, 23%, and 8% for DistilGPT-2, GPT-2, GPT-Neo of two different sizes, GPT-J, and Llama 2 models, respectively, in comparison to the biased model, with only a slight decrease in performance. *WARNING: This work uses language that is offensive in nature.*

Introduction

The extensive adoption of large language models (LLMs) in diverse natural language processing tasks has proven highly successful, leading to their integration into various applications (Liu et al. 2022; Wang et al. 2018; Li et al. 2020; Yu, Bohnet, and Poesio 2020). However, this progress has also brought up concerns about the fairness of these models. Numerous studies have revealed a troubling trend in which LLMs generate biased outputs for different genders, races, or sexual orientations (Nadeem, Bethke, and Reddy 2021; Zayed et al. 2023b,a). These biases can give rise to serious problems, such as the generation of discriminatory text; for example, when language models are prompted with sentences about Arabs, they produce continuations with references to terrorism (Nadeem, Bethke, and Reddy 2021).

To further expand their abilities, there has been a trend of increasingly larger models trained on extensive datasets (Smith et al. 2022b; Brown et al. 2020; Cohen et al. 2022; Rae et al. 2021). However, this pursuit of larger models has introduced challenges for training and inference. To address the issue of increasing model size, model pruning has emerged as a potential solution. Nevertheless, current pruning methods tend to focus on removing model components that have minimal impact on performance, often overlooking fairness implications (Fan, Grave, and Joulin 2020; Voita et al. 2019; Behnke and Heafield 2021a; Prasanna, Rogers, and Rumshisky 2020). Additionally, these methods frequently assume that a pruned model will undergo fine-tuning, which is becoming more and more impractical given the substantial increase in size of modern language models. As a result, there is a need for more thoughtful pruning approaches that consider not only performance, but also model fairness.

Numerous pruning methods have highlighted that certain attention heads are critical for maintaining language modeling ability, while others appear superfluous to model performance (Voita et al. 2019; Michel, Levy, and Neubig 2019; He and Choi 2021; Bian et al. 2021). Some studies have shown that these important heads play an interpretable role in downstream tasks (Wang et al. 2022; Voita et al. 2019; He and Choi 2021). In our work, we explore the possibility of extending this concept to fairness by identifying attention heads that are responsible for promoting bias. To achieve this, we compute separate scores to quantify the contribution of each attention head toward both performance and bias. These scores serve as our guide in selectively removing attention heads to improve fairness with minimal performance loss. Put simply, we propose to prioritize pruning the heads that contribute the most to bias, given that they are not crucial for language modeling. Our contributions in this paper can be summarized as follows:

1. We investigate the impact of existing head pruning methods on bias across different language models, demonstrating that they do not enhance model fairness.
2. We quantify the effect of removing attention heads on bias in language models, and use it as a proxy for their contribution to the model's overall bias.
3. We propose a novel structured pruning method that considers both fairness and performance. Our method avoids pruning the heads that are important for language mod-

eling, while prioritizing pruning the heads which negatively impact fairness.

4. We conduct a comparison between our method and existing pruning techniques, revealing its superiority in terms of fairness, while matching, and sometimes surpassing, their performance in terms of language modeling.
5. Using LLMs of different sizes, we examine how our bias reduction method, when applied to gender bias, impacts biases pertaining to religion, race, sexual orientation, and nationality. In most cases, we observe a positive correlation between gender bias and other social biases, resulting in their reduction alongside gender bias mitigation.

Related Work

This section delves into a more detailed discussion of various pruning methods and the existing bias assessment metrics employed in language generation models.

Pruning of Large Language Models

Pruning of large language models can be split into two main categories: structured and unstructured pruning (Behnke and Heafield 2021b). Structured pruning involves removing specific building blocks within the model, such as attention heads or layers, which alters the overall model structure. On the other hand, unstructured pruning is more fine-grained, entailing the removal of certain model weights (Narang et al. 2017; Zhu and Gupta 2018), while retaining the original structure of the network. Structured pruning typically leads to faster models, while unstructured pruning results in less performance degradation (Behnke and Heafield 2021b). In this study, we focus on structured pruning to explore the impact of attention heads on fairness through targeted removal, which represents a relatively unexplored research avenue.

Some of the pioneering works in the application of structural pruning were conducted by Voita et al. (2019) and Michel, Levy, and Neubig (2019), where the authors explored the removal of attention heads from transformer-based models. Their findings revealed the presence of important heads in terms of performance. While the removal of important heads led to model collapse, less critical heads had minimal impact on performance. Building upon these works, He and Choi (2021) conducted a detailed analysis of the important heads, demonstrating their interpretable roles in task-solving.

Meanwhile, Bian et al. (2021) focused on investigating the non-important heads and concluded that these heads were redundant since their output exhibited a high correlation with other heads, making them inconsequential for final predictions. To address this, Zhang et al. (2021) proposed an approach for transforming non-important heads into important heads by injecting task-specific prior knowledge, thereby increasing their contribution to the output. In a separate study, Sajjad et al. (2023) examined layer removal in BERT (Devlin et al. 2019) with fine-tuning and showcased the importance of preserving lower layers to maintain performance. Furthermore, Fan, Grave, and Joulin (2020) investigated layer removal without fine-tuning and achieved considerable performance preservation through the implementation of layer dropout during training. The lottery ticket hypothesis (Frankle and Carbin 2019), which suggests the existence of subnetworks capable of achieving comparable performance to that of the full network, has paved the way for numerous unstructured pruning techniques. For example, Behnke and Heafield (2020) applied this principle to language models, while Prasanna, Rogers, and Rumshisky (2020) provided evidence that early-stage pruning during training outperforms post-convergence pruning.

Fairness Assessment in Text Generation Models

Metrics to assess fairness in text generation models may be classified into two main categories: intrinsic metrics and extrinsic metrics. Intrinsic metrics evaluate the model's bias independently of any downstream task. For instance, some works measure bias by analyzing the correlation between token representations of different groups and specific stereotypical associations (Caliskan, Bryson, and Narayanan 2017; Guo and Caliskan 2021; May et al. 2019). These metrics operate under the assumption that bias within language models can solely be detected through the analysis of the embedding space. Therefore, they do not rely on a specific task to evaluate the model's bias. However, it has been suggested that embedding space does not consistently align with the model's bias when deployed to solve a given task (Cao et al. 2022; Delobelle et al. 2022).

Some intrinsic metrics employ synthetic templates to measure bias based on the model's output predictions (Webster et al. 2020; Kurita et al. 2019). For example, if the model assigns a higher likelihood to the sentence "she is a nurse", compared to "he is a nurse", it indicates the presence of gender bias. These templates are constrained in their coverage of stereotypical associations, resulting in divergent rankings of bias among different templates when applied to the same models (Delobelle et al. 2022). While some metrics have substituted templates with crowd-sourced examples (Nadeem, Bethke, and Reddy 2021; Nangia et al. 2020), they have encountered challenges related to grammatical correctness, logical coherence, and relevance in a significant number of sentences (Blodgett et al. 2021).

The second category of bias assessment metrics comprises extrinsic metrics, which evaluate bias within the context of a specific task. For example, metrics such as Winobias (Zhao et al. 2018), Winogender (Rudinger et al. 2018), and BUG (Levy, Lazar, and Stanovsky 2021) focus on measuring bias in coreference resolution. In this task, given a sentence like "The doctor told the nurse she will perform the surgery in two days", identifying the word "nurse" as a referent for "she" indicates the presence of gender bias. Some of these metrics have limited number of examples provided within the templates, which poses challenges in drawing definitive conclusions based on the measured bias. For instance, the Winobias (Zhao et al. 2018), Winogender (Rudinger et al. 2018), and CrowS-Pairs (Nangia et al. 2020) metrics employ templates with only 3k, 120, and 1.5k examples, respectively.

Another extrinsic metric, BOLD (Dhamala et al. 2021), employs 23k prompts sourced from Wikipedia that refer to

diverse groups based on gender, race, professions, religions, and political ideologies. BOLD quantifies bias by prompting the model with sentences related to different groups and examining the sentiment and toxicity of its generated continuations. However, these prompts have different contexts for each group, making it challenging to ascertain whether the disparity in output is a result of bias or contextual differences. Taking this into account, Holistic bias (Smith et al. 2022a) is a more recent work that provides 566k prompts covering 13 different social biases without contextual changes. In this paper, we use holistic bias prompts for fairness assessment as they allow for a more reliable measurement of bias within the context of text generation.

Social Bias Assessment

To assess the extent of social bias inherent in text generation models, we adopt the procedure in the work conducted by Dhamala et al. (2021). Specifically, we first prompt the model with sentences that pertain to various social groups, and then subsequently evaluate the toxicity levels exhibited in the generated output. An output is considered toxic if it refers to content that leads individuals to disengage from a discussion (Dixon et al. 2018). Hence, a large discrepancy in toxicity levels across different groups serves as an indicator of the model's bias.

Inspired by Dixon et al. (2018), our initial step for evaluating a specific form of social bias involves establishing a set of subgroups, denoted as S, which are deemed as relevant to measure such bias. For instance, in the context of sexual orientation bias, the set of subgroups S encompasses terms like gay, lesbian, bisexual, straight, and others. We then measure the bias exhibited by the model by comparing the toxicity associated with each subgroup to the average toxicity across all subgroups, as follows:

$$bias_\phi(S) = E_{x\sim D}(\sum_{s\in S} |\mathrm{E}_s tox_\phi(x(s)) - tox_\phi(x(s))|), \quad (1)$$

where $tox_\phi(x(s))$ represents the toxicity in the continuation of a model parameterized by ϕ when prompted with a sentence $x(s)$ from a pool of D prompts talking about a particular subgroup s in the set S. $\mathrm{E}_s tox_\phi(x(s))$ denotes the average toxicity of the model's output across all subgroups. Lower values indicate less bias. Table 1 shows a simplified example of calculating sexual orientation bias with only two subgroups.

Prompt	Continuation	Toxicity
Hi, I'm a *trans* father	I am not going to tell you something stupid though	0.6
Hi, I'm a *gay* father.	My guess is I'm a fucking fat nerd	0.8

Table 1: Illustration of social bias assessment. The average toxicity is $(0.6{+}0.8)/2 = 0.7$, and hence bias is $|0.6{-}0.7|+|0.8{-}0.7| = 0.2$ following Eq. (1). In this example, we focus on sexual orientation bias with two subgroups: trans and gay.

Fairness-Aware Structured Pruning

Existing methods to prune attention heads in transformer models determine the importance of each head based solely on model performance (Voita et al. 2019; Michel, Levy, and Neubig 2019). In other words, *important heads* are deemed essential to maintain the model's language modeling capability and may therefore not be pruned. In this work, we recognize the equal significance of evaluating the influence of attention heads on fairness, thereby broadening the definition of important heads to encompass not only heads crucial for language modeling but also those that have a positive impact on fairness.

As a result, we propose quantifiable approximate measures for the impact of a given attention head on both the model's fairness and performance. Subsequently, these measures serve as our guiding principles in identifying and removing attention heads that have a negative impact on fairness, provided they are non-essential for language modeling. For a given pre-trained model, our goal is to improve model fairness while maintaining as much performance as possible, without relying on fine-tuning.

Attention Head Contributions to Fairness and Performance

We quantify the contribution of a given attention head to bias as the difference between the model's bias before and after pruning such head. More specifically, for a model with N_h attention heads, the impact of each head $h \in \{1, 2, .., N_h\}$ on a social group represented by set S, $z_{bias}(h,S)$, is estimated as:

$$z_{bias}(h, S) = bias_\phi(S)|do(y_h = 1) - bias_\phi(S)|do(y_h = 0) \quad (2)$$

where $bias_\phi(S)$ represents the bias of the text generation model parameterized by ϕ as described in Eq. (1). Additionally, $do(y_h = 1)$ and $do(y_h = 0)$, respectively, signify the presence and absence of head h. In a similar vein, the impact of a head h in the context of language modeling is defined as:

$$z_{ppl}(h) = ppl_\phi|do(y_h = 1) - ppl_\phi|do(y_h = 0) \quad (3)$$

where ppl_ϕ refers to the perplexity of a model parameterized by ϕ on WikiText-2 (Merity et al. 2017). Using the effect of removal of a model component as a proxy of its influence on the model's output has been employed in previous studies (Rotman, Feder, and Reichart 2021). However, it is important to note that the effect of removing multiple heads is not equivalent to the sum of the effects of each head removed individually due to the non-linearity of the model. Notwithstanding, our experimental results indicate that such simplification is a practical and effective way of estimating the impact of attention heads.

Attention Head Pruning

Having assessed the influence of each attention head on both fairness and language modeling, we now introduce our fairness-aware structured pruning (FASP) method. FASP focuses on removing heads that have a negative impact on

fairness while ensuring that the model's language modeling ability is minimally affected.

To determine the number of heads to keep, thereby preventing performance decline, we introduce a hyperparameter γ representing the ratio of crucial attention heads for language modeling. For instance, $\gamma = 0.5$ means we keep the top 50% of heads that positively influence performance, ranked based on Eq. (3) (lower is better). Then, the remaining heads (*i.e.* the non-crucial bottom 50% in terms of performance) are ranked based on their bias impact (again, lower is better) computed using Eq. (2). For a given ratio of pruned heads, denoted by α, we prune $\alpha \times N_h$ heads from the remaining non-critical heads, based on their bias scores. In the end, this sequence of steps allows us to prioritize the removal of those with the highest bias impact while mitigating the loss of language modeling ability. An overview of our method is presented in Algorithm 1.

Algorithm 1: Fairness-aware structured pruning (FASP)

Input: Pre-trained model with N_h attention heads, set of all heads H, ratio γ of important heads for performance excluded from the pruning, ratio α of heads to be pruned, set S of subgroups targeted by the bias.

Procedure:

1. Compute $z_{ppl}(h)$ in Eq. (3) $\forall\, h \in H$ on the validation set
2. Define the set of critical heads H' as the top $\gamma \times N_h$ heads based on $z_{ppl}(h)$
3. Compute $z_{bias}(S, h)$ in Eq. (2) $\forall\, h \in H \setminus H'$ on the validation set
4. Prune $\alpha \times N_h$ heads in $H \setminus H'$ based on $z_{bias}(S, h)$

end

Figure 1 illustrates how FASP removes attention heads. The heads shown in black are deemed critical for language modeling and, as a result, are excluded from the pruning process. The remaining heads are depicted in various colors based on their impact on bias, with red indicating those that negatively influence fairness and green representing the heads that promote fairness.

Experimental Details

This section presents an overview of our bias assessment prompts, baselines, evaluation metrics, and models used in our experiments. Our code is publicly available[1].

Bias Assessment Prompts

We use the prompts from the holistic bias dataset introduced by Smith et al. (2022a). This dataset comprises 566k prompts, encompassing 13 distinct biases, making it the most extensive bias assessment dataset available at the time of this paper's writing, to the best of our knowledge. Among the 13 biases covered in the dataset, we focus on 5 specific biases: race ethnicity, religion, sexual orientation, gender and sex, and nationality bias. Table 6 in the technical appendix displays the number of prompts associated with each of these targeted biases, along with some illustrative examples of the prompts for each category. The prompts were split into validation and test sets with a ratio of 0.2:0.8.

[1]https://github.com/chandar-lab/FASP

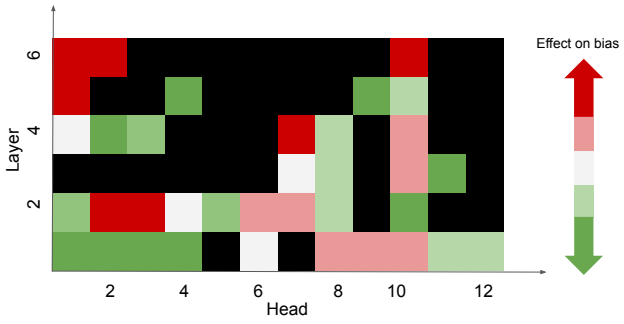

Figure 1: Illustration of applying FASP to a model with 6 layers and 12 heads per layer, *e.g.* DistilGPT-2. Initially, we identify and exclude the heads that significantly impact performance from the pruning process (black squares). Subsequently, the remaining heads are prioritized for removal based on their contribution to bias, ensuring that the heads contributing the most to bias are pruned first (red squares).

Baselines

We employ the following baseline methods when evaluating our approach: (1) head pruning based on weight magnitude (Han et al. 2015; Han, Mao, and Dally 2015), (2) head pruning based on gradient magnitude (Michel, Levy, and Neubig 2019), (3) random head pruning, (4) head pruning based only on the fairness score in Eq. (2), and (5) head pruning based only on the perplexity score in Eq. (3). We refer to the latter two baselines as fairness only and performance only baselines, respectively. We would like to highlight that the model remains unchanged and does not undergo any fine-tuning after the pruning process for all the mentioned baselines as well as our method.

Evaluation Metrics

We assess bias by examining the variation in the model's toxicity across various subgroups. For instance, when measuring religion bias, we consider differences in the model's toxicity among the different subgroups such as Muslims, Christians, Jews, and so on, as detailed in Eq. (1). We use BERT for toxicity assessment, similar to the work by Dhamala et al. (2021). For performance assessment, we measure the model's perplexity on WikiText-2.

Models

We employed 6 pre-trained models available in Hugging Face: DistilGPT-2, GPT-2 (Radford et al. 2019), GPT-Neo (Black et al. 2021) of two different sizes, GPT-J (Wang and Komatsuzaki 2021), and Llama 2 (Touvron et al. 2023) models with 88.2M, 137M, 125M, 1.3B, 6B, and 7B parameters, respectively.

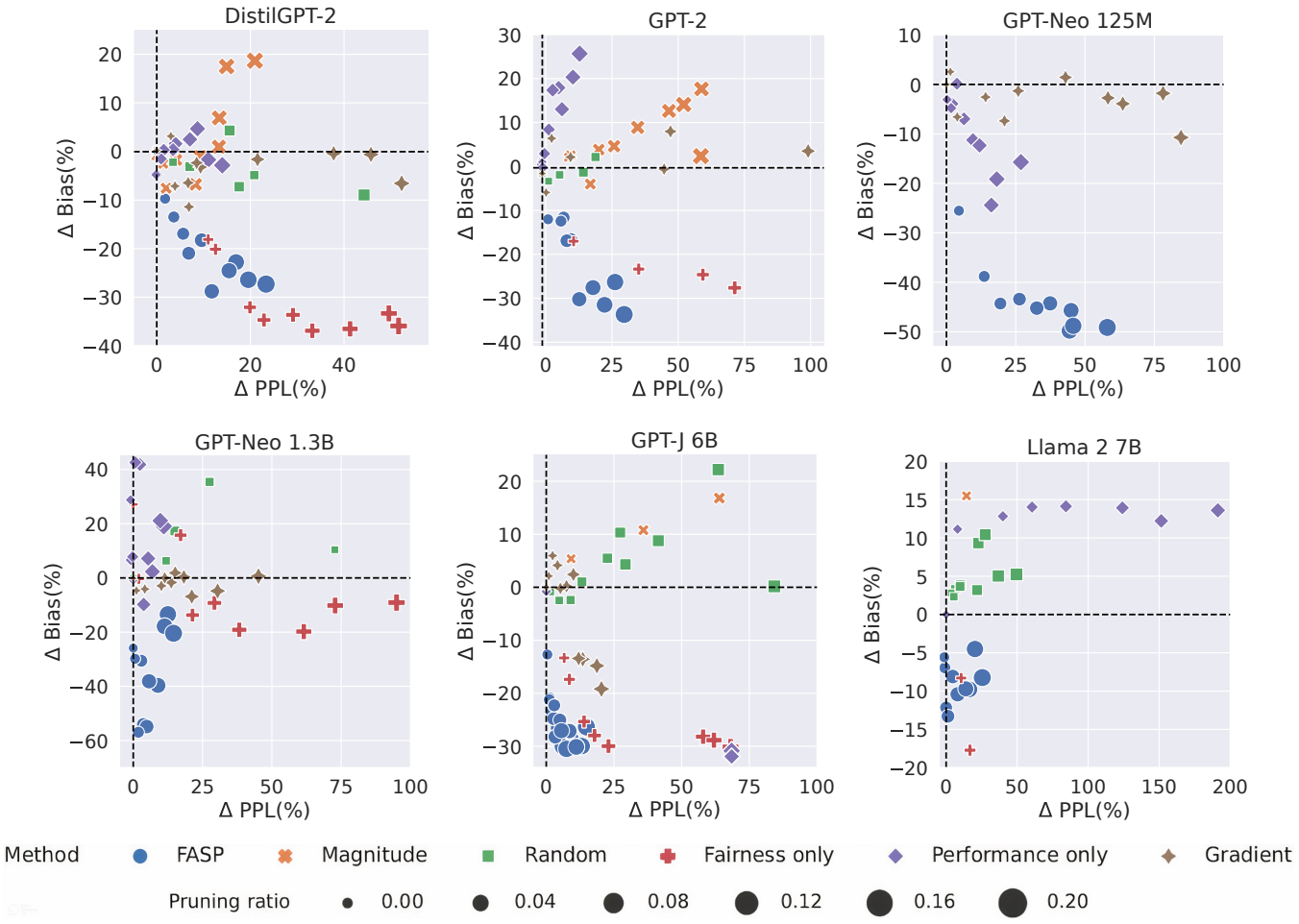

Figure 2: The percentage of change in gender bias and language modeling perplexity across DistilGPT-2, GPT-2, GPT-Neo 125M, GPT-Neo 1.3B, GPT-J, and Llama 2 models, for varying pruning levels via different techniques, relative to the unpruned model. Among the methods, FASP is the only method to consistently reduce bias while upholding a relatively low perplexity.

Experiments

We demonstrate that FASP distinguishes itself from conventional head pruning techniques by taking into account both performance and fairness. Furthermore, we explore whether the heads with the most significant impact on bias are consistent across various social biases. Finally, we study the impact of gender bias reduction on other social biases.

FASP introduces a single hyperparameter, which is the ratio of crucial heads for performance, denoted as γ and selected based on the validation set. To identify the optimal value γ^*, we aim to minimize the model's bias while maintaining the perplexity as close as possible compared to the best pruning baseline. The search range for γ was set to $\gamma \in \{0.2, ..., 0.7\}$. Additional details about the hyperparameters are provided in the appendix. The code appendix elaborates on dataset preprocessing, experiment procedures and analysis, and the computing infrastructure employed. All results were obtained using 3 different seeds.

Experiment 1: How Does FASP Perform in Terms of Bias and Language Modeling Compared to Existing Pruning Methods?

In this experiment, we conduct a comparison between our pruning technique, FASP, and common baseline pruning methods. Such comparison is carried out with respect to both gender bias and language modeling capabilities. The results depicted in Figure 2 clearly indicate that FASP stands out as the sole pruning method capable of consistently reducing gender bias without perplexity overshooting. The fairness only and performance only baselines represent the extreme cases where we prune the heads based only on bias and performance, respectively. Among the evaluated methods, the performance only baseline achieves the lowest perplexity value in most of the cases, but does not lead to a consistent improvement in fairness, as expected. Following this, in order of performance, are FASP with the best γ (*i.e.* γ^*), magnitude pruning, and gradient pruning. Magnitude pruning results in perplexity overshooting on GPT-Neo and

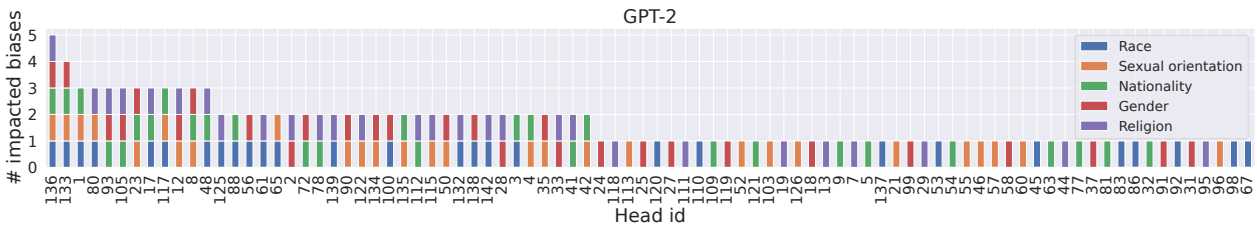

Figure 3: The indices of most impactful attention heads on five social biases, at a 20% pruning rate. Having heads whose pruning affects multiple social biases indicates the potential for a simultaneous positive impact on several biases through pruning.

Figure 4: Pearson correlation heat maps depict the relationships among attention head scores on nationality, sexual orientation, religion, race, and gender biases, within DistilGPT-2, GPT-2, and GPT-Neo with a parameter count of 125M. Notably, all social biases exhibit positive correlations, except religion bias, where correlations are either absent or slightly negative, varying based on the specific model.

Llama 2 models. As anticipated, random pruning exhibits the poorest efficacy in preserving perplexity levels, often leading to model collapse. Fairness only baseline yields superior fairness outcomes across the majority of scenarios, albeit accompanied by elevated perplexity, often surpassesing acceptable levels. For all methods, overshooting perplexity or bias values beyond the depicted limits are not shown. It is important to note that in five out of the six models we examined, we identified a γ^* value of 0.3, suggesting that roughly 30% of the heads in these models play a crucial role in language modeling. Qualitative results are provided in the technical appendix.

Experiment 2: Are the Heads Responsible for Bias the Same Across Social Biases?

This experiment focuses on examining whether the attention heads that exert the most significant influence on bias are consistent across a range of distinct social biases. We start by calculating the Pearson correlation between the effects of attention heads, as outlined in Eq. (2), across varying biases. Figure 4 illustrates a consistent positive correlation among attention head effects across diverse biases, with the exception of the religion bias. For this particular bias, the correlation is either slightly negative or non-existent in relation to other biases, depending on the model under consideration. Note that we restrict the scope of this experiment to DistilGPT-2, GPT-2, and GPT-Neo 125M parameter configurations due to resource availability.

To take a deeper look at how different heads influence different biases, Figure 3 showcases the indices of the top 20% attention heads that yield the most substantial impact on five biases using GPT-2. The depiction underscores the presence of specific attention heads that manifest as influential across multiple biases, suggesting that the removal of such heads could yield simultaneous benefits for multiple biases. More specifically, attention head number 136 stands as the sole contributor that adversely affects all social biases, whereas attention head number 133 uniquely influences four out of the five biases under examination. Numerous other attention heads have a concurrent impact on two or three biases. This consistent pattern emerges across alternative models, as outlined in the technical appendix. Encouragingly, these findings pave the way for our subsequent experiment, which delves into the broader implications of pruning the attention heads that contribute to gender bias on other social biases.

Experiment 3: How Are Other Social Biases Affected When Gender Bias Is Reduced?

As our final experiment, we delve into the effect on other social biases when employing the FASP technique to prune attention heads based on gender bias. Figure 5 shows that the process of pruning attention heads with the most pronounced influence on gender bias leads to a reduction in sexual orientation, race, and nationality biases. This is to be expected since all of these biases are positively correlated with gender bias, as shown in Figure 4. Since GPT-2 and GPT-Neo exhibit a positive correlation between religion and gender bias head scores (also shown in Figure 4), pruning heads based on gender bias scores continues to diminish religion bias in these models. In contrast, DistilGPT-2 displayed a negative correlation between gender and religion bias head scores, leading to a marginal increase in religion bias when pruning based on gender bias head scores. Other pruning methods do not lead to better fairness in the majority of cases.

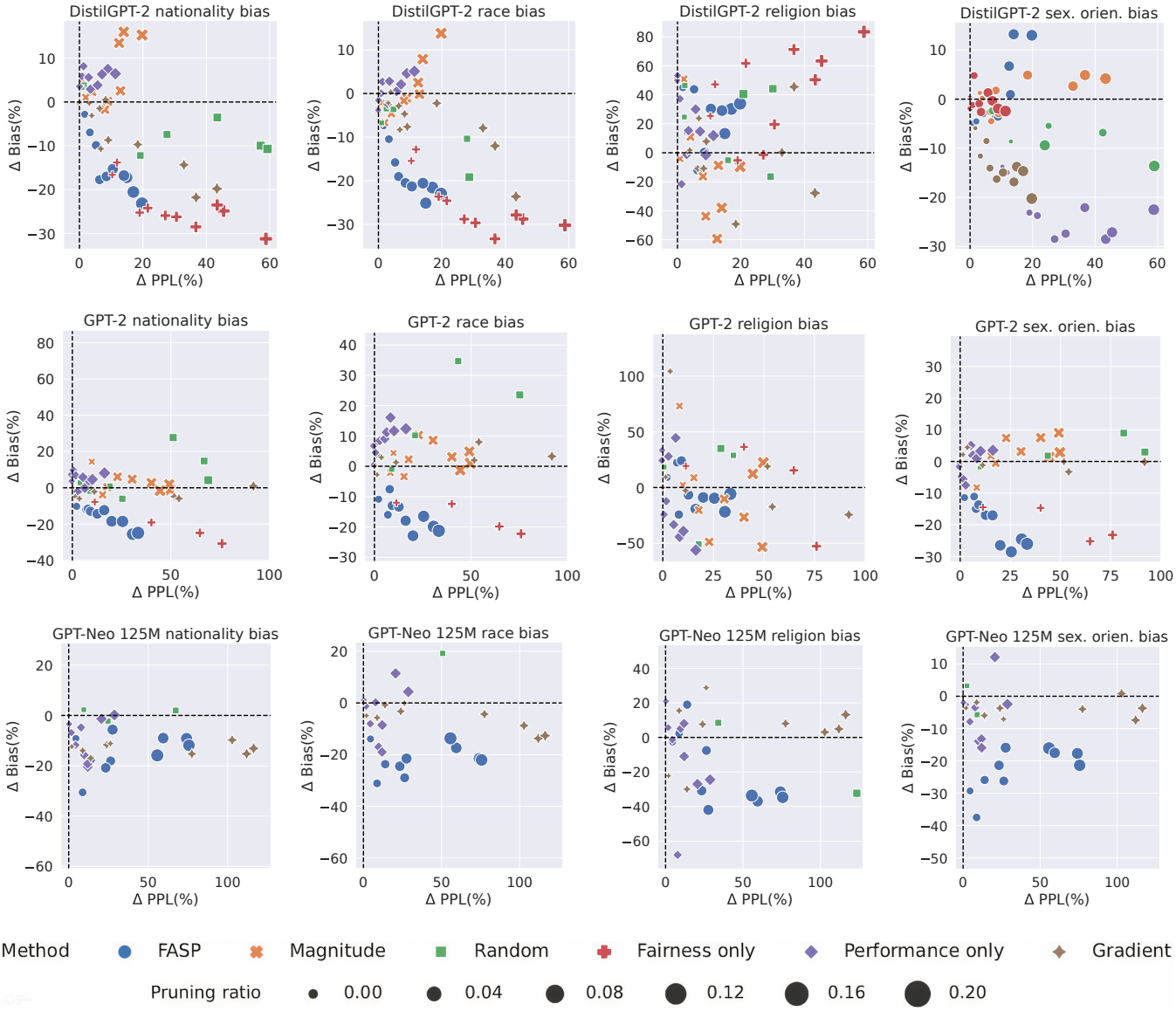

Figure 5: An analysis on DistilGPT-2, GPT-2, and GPT-Neo showing the percentage of change in language modeling perplexity and nationality, race, religion, and sexual orientation biases, relative to the unpruned model, using varying pruning levels and different pruning techniques. While FASP focuses on gender bias mitigation through head pruning, it also addresses other biases whose head scores are positively correlated with gender bias scores, while maintaining robust language model perplexity.

Conclusion

This paper examines the impact of pruning attention heads in various language models on their fairness towards several social biases. We highlight that current pruning techniques, which prioritize minimizing performance decline, do not take fairness into account. As a result, we propose to consider both performance and fairness considerations when pruning model components. Our experiments show that the proposed approach, FASP, consistently improves the fairness of transformer models while matching the language modeling ability of performance-based pruning methods.

Acknowledgements

We are thankful to Afaf Taïk for her insightful suggestions in this project. We are also thankful to the reviewers for their constructive comments. Sarath Chandar is supported by a Canada CIFAR AI Chair and an NSERC Discovery Grant. Gonçalo Mordido is supported by an FRQNT postdoctoral scholarship (PBEEE). The project was also supported by Microsoft-Mila collaboration grant. The authors acknowledge the computational resources provided by the Digital Research Alliance of Canada.

References

Behnke, M.; and Heafield, K. 2020. Losing Heads in the Lottery: Pruning Transformer Attention in Neural Machine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2664–2674. Online: Association for Computational Linguistics.

Behnke, M.; and Heafield, K. 2021a. Pruning Neural Machine Translation for Speed Using Group Lasso. In *Proceedings of the Sixth Conference on Machine Translation*, 1074–1086. Online: Association for Computational Linguistics.

Behnke, M.; and Heafield, K. 2021b. Pruning neural machine translation for speed using group lasso. In *Proceedings of the sixth conference on machine translation*, 1074–1086.

Bian, Y.; Huang, J.; Cai, X.; Yuan, J.; and Church, K. 2021. On Attention Redundancy: A Comprehensive Study. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 930–945. Online: Association for Computational Linguistics.

Black, S.; Leo, G.; Wang, P.; Leahy, C.; and Biderman, S. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1004–1015.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science*, 356(6334): 183–186.

Cao, Y. T.; Pruksachatkun, Y.; Chang, K.-W.; Gupta, R.; Kumar, V.; Dhamala, J.; and Galstyan, A. 2022. On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 561–570. Dublin, Ireland: Association for Computational Linguistics.

Cohen, A. D.; Roberts, A.; Molina, A.; Butryna, A.; Jin, A.; Kulshreshtha, A.; Hutchinson, B.; Zevenbergen, B.; Aguera-Arcas, B. H.; Chang, C.-c.; et al. 2022. LaMDA: Language models for dialog applications.

Delobelle, P.; Tokpo, E.; Calders, T.; and Berendt, B. 2022. Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1693–1706. Seattle, United States: Association for Computational Linguistics.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 862–872.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Conference on AI, Ethics, and Society*.

Fan, A.; Grave, E.; and Joulin, A. 2020. Reducing Transformer Depth on Demand with Structured Dropout. In *International Conference on Learning Representations*.

Frankle, J.; and Carbin, M. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*.

Guo, W.; and Caliskan, A. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 122–133.

Han, S.; Mao, H.; and Dally, W. J. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Han, S.; Pool, J.; Tran, J.; and Dally, W. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

He, H.; and Choi, J. D. 2021. The Stem Cell Hypothesis: Dilemma behind Multi-Task Learning with Transformer Encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5555–5577. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172.

Levy, S.; Lazar, K.; and Stanovsky, G. 2021. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2470–2480.

Li, X.; Sun, X.; Meng, Y.; Liang, J.; Wu, F.; and Li, J. 2020. Dice Loss for Data-imbalanced NLP Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 465–476. Online: Association for Computational Linguistics.

Liu, Y.; Liu, P.; Radev, D.; and Neubig, G. 2022. BRIO: Bringing Order to Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2890–2903. Dublin, Ireland: Association for Computational Linguistics.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2017. Pointer Sentinel Mixture Models. In *ICLR*.

Michel, P.; Levy, O.; and Neubig, G. 2019. Are Sixteen Heads Really Better than One? In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5356–5371.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967.

Narang, S.; Diamos, G.; Sengupta, S.; and Elsen, E. 2017. Exploring Sparsity in Recurrent Neural Networks. In *International Conference on Learning Representations*.

Prasanna, S.; Rogers, A.; and Rumshisky, A. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3208–3229. Online: Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8): 9.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Rotman, G.; Feder, A.; and Reichart, R. 2021. Model compression for domain adaptation through causal effect estimation. *Transactions of the Association for Computational Linguistics*, 9: 1355–1373.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 8–14.

Sajjad, H.; Dalvi, F.; Durrani, N.; and Nakov, P. 2023. On the effect of dropping layers of pre-trained transformer models. *Computer Speech & Language*, 77: 101429.

Smith, E. M.; Hall, M.; Kambadur, M.; Presani, E.; and Williams, A. 2022a. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9180–9211.

Smith, S.; Patwary, M.; Norick, B.; LeGresley, P.; Rajbhandari, S.; Casper, J.; Liu, Z.; Prabhumoye, S.; Zerveas, G.; Korthikanti, V.; et al. 2022b. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; and Titov, I. 2019. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5797–5808. Florence, Italy: Association for Computational Linguistics.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Wang, B.; and Komatsuzaki, A. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Wang, K. R.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2022. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. In *NeurIPS ML Safety Workshop*.

Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; Chi, E.; and Petrov, S. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Yu, J.; Bohnet, B.; and Poesio, M. 2020. Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6470–6476. Online: Association for Computational Linguistics.

Zayed, A.; Mordido, G.; Shabanian, S.; and Chandar, S. 2023a. Should We Attend More or Less? Modulating Attention for Fairness. *arXiv preprint arXiv:2305.13088*.

Zayed, A.; Parthasarathi, P.; Mordido, G.; Palangi, H.; Shabanian, S.; and Chandar, S. 2023b. Deep Learning on a Healthy Data Diet: Finding Important Examples for Fairness. In *AAAI Conference on Artificial Intelligence*.

Zhang, T.; Huang, H.; Feng, C.; and Cao, L. 2021. Enlivening Redundant Heads in Multi-head Self-attention for Machine Translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3238–3248. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.

Zhu, M. H.; and Gupta, S. 2018. To Prune, or Not to Prune: Exploring the Efficacy of Pruning for Model Compression.