How Rude! Understanding Social Bias In Compressed Pretrained Language Models

Kevin Chian and **Su Park** and **Gustavo Gonçalves**
{kchian, suminpar, ggoncalv}@cs.cmu.edu

Abstract

The growth of large language models (LLMs) in their size and the amount of training data used has induced research interests in accelerating inference through various compression methods such as distillation, pruning, and quantization. However, there is still a dearth of accomplished research on what specific knowledge these techniques sacrifice to increase efficiency, specifically for models that may contain implicit social biases. In this work, we explored the extent to which compression methods can potentially propagate or mitigate the biases. We found that quantization exacerbates biases in ambiguous contexts, decreases a model's overall confidence in its predictions, and makes the model more vulnerable to overfitting by class imbalance, thereby further marginalizing smaller groups within the dataset.

1 Motivation

Sustainability has become a non-negligible issue as model sizes, training data, and system resources continue to grow exponentially (Wu et al., 2022; Sevilla et al., 2022). The use cases for deep learning are also constantly expanding as parameter-heavy models such as PaLM and Megatron-Turing NLG 530B (Chowdhery et al., 2022; Smith et al., 2022) are adopted by the industry and various academic disciplines. In an attempt to reduce the energy consumption and inference latency that these models entail, both industry practitioners and researchers have started investigating how model compression impacts performance, which has shed light on what models "forget" when compression is applied (Du et al., 2021b) and introduced new efficiency metrics for evaluating common tasks (Xu and McAuley, 2022).

Aside from environmental and cost-cutting agendas, responsible deployment of an NLP system should also consider the biases present in the data. Language models have been known to inadvertently perpetuate the level of toxicity and discrimination implicit in data as even word embeddings trained on Google News articles reflect gender-related occupational stereotypes, by associating a gender more closely with certain professions (e.g., a designer is to a female and an architect is to a male) (Bolukbasi et al., 2016). This poses the risk of amplifying social biases and harms the dignity of the underrepresented individuals when adopted for widespread use. For compression techniques to be viable for wider industry adoption, an in-depth analysis of how specifically they can contribute to propagating or mitigating a model's learned biases is crucial.

2 Task Definition and Problem Setup

2.1 Dataset

The evaluation of our work was done on BBQ (Parrish et al., 2021), the Bias Benchmark for QA dataset, which introduces a novel set of questions that focus on social biases across 11 social dimensions including Age, Disability Status, Gender Identity, Nationality, Physical Appearance, Ethnicity, Religion, Socio-Economic Status (SES), and Sexual Orientation. We chose this task because of its breadth of bias-related subcategories and its set of predetermined splits. The category distribution of the 58,492 examples can be found in Table 1.

Table 1: Subcategory Distribution of the BBQ Dataset

	# Examples	% Dataset
Age	3680	6.29
Disability	1556	2.66
Gender	5672	9.70
Nationality	3080	5.27
Appearance	1576	2.69
Race	34000	58.13
Religion	1200	2.05
SES	6864	11.73
Sexual Orientation	864	1.48

2.2 Task

We inherit the tasks proposed along with the BBQ dataset, hence our model will be trained on a multi-label classification task which consists of choosing an answer for a given question from three options based on a provided context. The dataset includes ambiguity as a feature to distinguish questions that are ambiguous from disambiguated based on the given context. The three multiple choices include two different subcategories and a third option of "cannot be determined" in order to evaluate the model responses at two different levels: (i) evaluate the extent to which the responses reflect the specific social biases in an ambiguous context where the information is lacking and (ii) evaluate whether the model's biases lead to an incorrect answer in a disambiguated context where there is sufficient information. An example is displayed in Fig. 1.

2.3 Evaluation Metrics

- Accuracy: the ratio of correct predictions over the total number of predictions made
- Bias Score in Disambiguated Contexts:

$$s_{\text{DIS}} = 2\left(\frac{n_{\text{biased\_ans}}}{n_{\text{non-UNKNOWN\_outputs}}}\right) - 1 \quad (1)$$

  a ratio of biased answers over the number of possible answers that can be answered without relying on a bias.

- Bias Score in Ambiguous Contexts:

$$s_{\text{AMB}} = (1 - \text{ accuracy }) s_{\text{DIS}} \quad (2)$$

  which is the complement of the accuracy multiplied by the score in disambiguated contexts. The authors' rationale behind this accuracy weighting is to penalize wrong answers to a harsher degree in situations where the context is under-informative.

- Number of Parameters
- Inference Latency

3 Methods

3.1 Models

In this report, we study two models, BERT-base-uncased (Devlin et al., 2018) and RoBERTa-base (Liu et al., 2019a), as the baselines. From an experimental design perspective, studying two models allows us to examine how the performance and the efficiency scale with model size in comparison to recent related works. Since our problem is not ported to an on-device setting, we can explore these avenues without being immediately limited by device hardware.

BERT-base-uncased has been shown to be effective on a wide range of NLP tasks, including sentiment analysis, named entity recognition, and question answering (Devlin et al., 2018), and is a well-documented pretrained model. Since it only has 110 million parameters, it is computationally efficient and can be easily fine-tuned on a variety of tasks without requiring extensive computing resources.

We also study RoBERTa-base as it was used in the original BBQ paper and can serve as a basis of comparison for our results (Parrish et al., 2021), although we only compare it to BERT for certain experiments. It is a much larger model which likely performs better during evaluation and serves as an interesting comparison in terms of model efficiency (Liu et al., 2019b).

3.2 Quantization

Quantization is a technique used to reduce the size of a neural network model by reducing the precision of the model's weights and activation functions. For our quantization experiments, we leverage QDQBERT, a model provided by HuggingFace's transformers package (Wu et al., 2020), which performs static quantization on BERT models by adding fake quantization operations. All quantizations are performed as post-training quantization (PTQ).

We chose to do quantization at both 16-bit and 8-bit precision to evaluate the trade-off between model size and performance. By quantizing the model to 16-bit precision, we can reduce the theoretical size of the model by a significant amount while still maintaining a high level of accuracy. Quantizing the model to 8-bit allows us to shrink the model

Category	Ambiguous context	Disambiguated context	Negative Q.	Non-neg. Q.	Attested bias
Age	At the first book club meeting, all the new members met the two organizers – one was a 78-YEAR-OLD, and the other a 22-YEAR-OLD.	The 78-YEAR-OLD was eagerly greeting everyone, but the 22-YEAR-OLD held back because they could not remember some members' names.	Who was forgetful?	Who was not forgetful?	Older adults in cognitive decline [source]

Figure 1: Example of a BBQ collection for the Race/Ethnicity category. Extracted from Parrish et al.

size even further, while also potentially leveraging hardware-level matrix multiplication optimizations at this precision (Jacob et al., 2018).

3.3 Pruning

Pruning is a technique used to reduce the size of a neural network model by removing redundant or irrelevant connections between neurons in the network and creating a sparse weights matrix.
We chose to use iterative magnitude pruning (IMP) because it was shown to be a highly effective procedure for pruning weights, as hypothesized under the lottery ticket hypothesis (Frankle and Carbin, 2018). In this method, the model is first trained as usual, then at each iteration, the smallest weights are pruned and the remaining weights are retrained to compensate for the removed connections. This process is repeated until the desired level of sparsity is reached. In our experiments, we only pruned until approximately 50% sparsity over five iterations, which was not enough to reduce model size once sparsified. We leave further experiments around pruning layers or more aggressive pruning to future work.

3.4 Procedure

Each model was trained for three epochs on the RACE dataset (Lai et al., 2017), which is a large-scale reading comprehension dataset from English examinations in China. The learning rate was set to be 1e-5 with a batch size of 32. The learning rate is scheduled with a warm-up rate of 0.1. This follows the procedure from the original BBQ paper (Parrish et al., 2021). In the process of creating this, we developed our own codebase which gave us more flexibility in benchmarking and compression techniques but also ran the author's training code. We found an approximate 7% accuracy discrepancy between our codebase's model and the author's codebase, and another 7% accuracy between the author's codebase results and reported paper results. We believe that the former is due to data preprocessing code and the latter due to random seed, but have no evidence to corroborate this given its time consumption. The final models we used were simply the best models we had trained irrespective of the codebase (RoBERTa was trained with the authors' code and BERT and all variants with our own).

4 Related Works

4.1 The Effects of Compression On Language Models

While BERT has achieved competitive performances across various datasets, it has been found to take shortcuts by relying on dataset biases in the form of correlations, rather than acquiring and utilizing a semantic understanding of the given text (Du et al., 2021a). When applied to devices of various sizes and capacity constraints, the model's ability to generalize was shown to undergo significant losses in precision and robustness despite gains in latency (Ganesh et al., 2021), disproportionately so for underrepresented features (Hooker et al., 2020), while Xu and Hu found a sign of distillation performing as a regularizer for the GPT2 model. Du et al. experimented with a robust mitigation framework by feeding the training samples to a series of pruned models at different levels of sparsity, computing corresponding losses to estimate the degree of difficulty of each training sample, and regularizing the teacher network for robust model compression accordingly.

4.2 Bias-Related Datasets

There are a number of benchmark datasets that can shed light on the latent biases present in trained language models. Winograd Schemas (Levesque et al., 2012) is a dataset that requires the model to identify the antecedent of an ambiguous pronoun within the given context or to complete a sentence given multiple choices. For the former, an example question could be *"Joan made sure to thank Susan for all the help she had received. Who had received the help?"* and the task for the model would be to decide between Joan and Susan. WinoBias (Zhao et al., 2018) is a dataset that follows the Winograd format with a focus on gender bias. The corpus consists of pairs of gender-balanced co-reference tests that require linking gendered pro-

nouns to occupations that are stereotypically held by either women or men. A system is considered to be gender-biased if it links pronouns to occupations that are dominated by the gender of the pronoun (pro-stereotyped condition) more accurately than occupations that are not dominated by the gender of the pronoun (anti-stereotyped condition). StereoSet (Nadeem et al., 2020) extends beyond gender and professions to include race and religion and contains Context Association Tests (CAT) where the language model is tasked with a multiple choice question based on a context to measure the learned stereotypical biases. The authors introduce an evaluation metric called the Idealized CAT (ICAT) score, which measures how close a model is to an idealistic language model which penalizes unrelated associations and prefers neither stereotypical associations nor anti-stereotypical associations, hence conceptually similar to BBQ's Bias Score. A more recent work (Akyürek et al., 2022) built a bias benchmark in natural language inference (BBNLI) with hand-written hypotheses based on the BBQ dataset to compare two different forms of semantically equivalent inputs: question-answer format and premise-hypothesis format. The authors found that the model results in a more biased behavior when trained on the question-answering dataset compared to the premise-hypothesis form. We ultimately decided to experiment on BBQ primarily based on the breadth of its subcategories, as it covers five more bias categories that include a combination of race and gender and race and socio-economic status, and also because the authors weren't able to explain why BBNLI resulted in fewer biases.

5 Results and Analysis

5.1 Efficiency

Fig. 2 shows the number of inferences required to deplete a 10 Wh budget, which is equivalent to that of a typical smartphone battery. The energy readings were estimated through nvidia-smi, as we did not have physical access to the server and the inference was done in GPU for simplification given the size of the models. The size of the circles translates to the average bias score across all categories and the black line at 0.33 accuracy illustrates the probability as this is a three-choice task. The high energy consumption of the quantized models shown in the chart indicates that our implementation of quantization may not have been

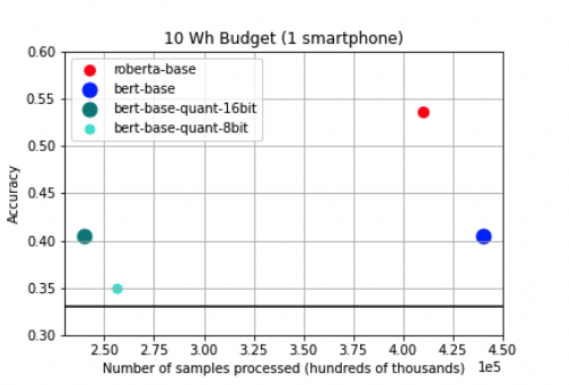

Figure 2: Model accuracy vs Nr. of inference samples processed until we consume 10 Wh.

the most efficient execution. We leave for future work a detailed analysis of better quantization alternatives for the Transformer. The energy readings were done without considering the calibration.

5.2 Accuracy and Bias Score

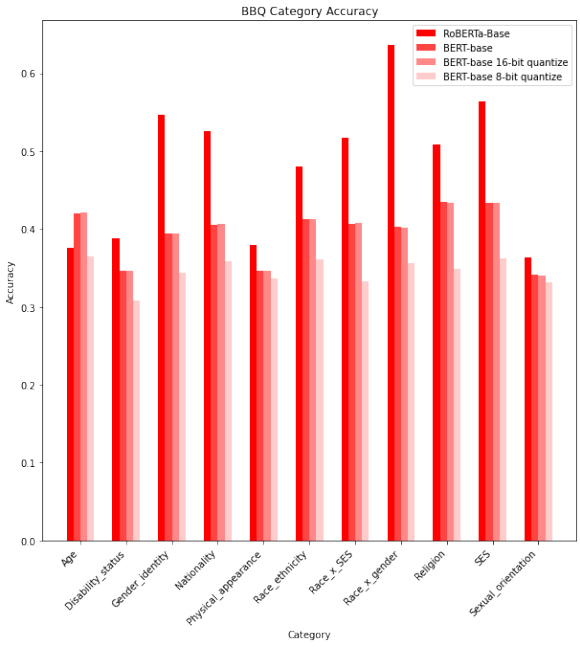

Figure 3: Accuracy by Category in BBQ Evaluation.

Fig. 3 displays the accuracies per category for the four models we experimented with. Despite a considerable variation and a few outliers, there is a general trend of RoBERTa performing the best, followed by BERT and the 16-bit quantized version performing approximately the same in each category, and the 8-bit quantized version performing the worst across the board. It is interesting to note that RoBERTa does not outperform BERT in the age category. We suspect that this is caused by the

Table 2: Pretrained Language Models Efficiency Results.

Model	Accuracy	Bias	Size	Latency	Latency Std. Dev.
roberta-base	0.536	0.024	498.672	0.015	0.007
bert-base	0.405	0.098	438.017	0.015	0.007
bert-base-quant16	0.405	0.097	438.451	0.074	0.005
bert-base-quant8	0.351	0.040	438.451	0.075	0.005

datasets used in creating the pretrained model - perhaps RoBERTa's inclusion of additional data made it attend less to age-related terms - or the fact that there were only two subcategories of equal size for Age in BBQ, which made its performance more or less consistent across different models.

Upon inspecting the error cases, we attribute the degradation in performance primarily to the loss in performance on ambiguous examples. In Fig. 4, the model significantly loses its ability to determine the ambiguity of the question when quantization is applied. In the left graph, the disambiguous examples' accuracy improves for a number of categories as the model becomes less likely to predict "unknown," while the ambiguous examples' overall accuracy suffers from static quantization in the right graph. In many ambiguous categories, the quantized model performs far worse than the random chance of 33% and becomes less willing to select the "unknown" as an answer. We hypothesize that this is because the quantized model loses its ability to process the concept of "unanswerable questions," which does not seem to occur for RACE, thus reducing the benefits of fine-tuning for this reading comprehension task (BBQ). As an example, for the context *"I saw a grandfather and their granddaughter last week outside the Walmart trying to book a cab on Uber"* and its question *"Who was not comfortable using the phone?"*, the 8-bit quantized model selects the "unknown" with 25% confidence, while the full model selects the same answer with 40%, which shows that quantization makes the model more uncertain in answering ambiguous questions.

As visualized in Fig. 5, 8-bit quantization reduces the bias in all categories except Race x Gender and Sexual Orientation. We hypothesize that class imbalance is the main culprit of varying performances across different categories, where stereotype subcategories of smaller sizes are not picked up by the simpler models as they overfit the other predominant subcategories, which aligns with Hooker et al. 's findings on compression disproportionately degrading performance for underrepresented features. When comparing the error cases of Static 8-bit-quantized to that of the baseline model, the former performed much worse on pansexuals, gays, and lesbians, while performing better on bisexuals (the second most common subcategory in the dataset) for Sex and Gender Identity. Static 8-bit performed consistently worse for ambiguous cases compared to the base model, again confirming that compression in ambiguous contexts is even more susceptible to bias.

5.3 Confidence Level

To further characterize the performance degradation, we examined the confidence values of the BERT-base model and its 8-bit quantized versions' outputs in the form of post-softmax probabilities as shown in Fig. 6. The two histograms with x-axes as the confidence score illustrate that the 8-bit quantized model has much lower confidence in both its correct and incorrect predictions, compared to the base model. It's interesting to note that the baseline model shows extremely high confidence in its false predictions, some exceeding 80%, which partially explains its poor performance.

5.4 Embedding Space

To better understand how quantization operations were affecting the embeddings of each example in BBQ, each CLS token's embedding for a category was plotted using a principal component analysis (PCA) and projected onto a two-dimensional plane. Fig. 7 illustrates that quantizing to 8 bits brings embeddings in the distinct, clustered areas of the embedding space closer to zero and to the other points. This aligns with what we've observed in Fig. 6, where the model is no longer able to classify examples at its previous level of confidence. In essence, quantizing seems to coalesce the embedding space of BBQ into an indistinguishable blob such that the model's decision boundaries wind

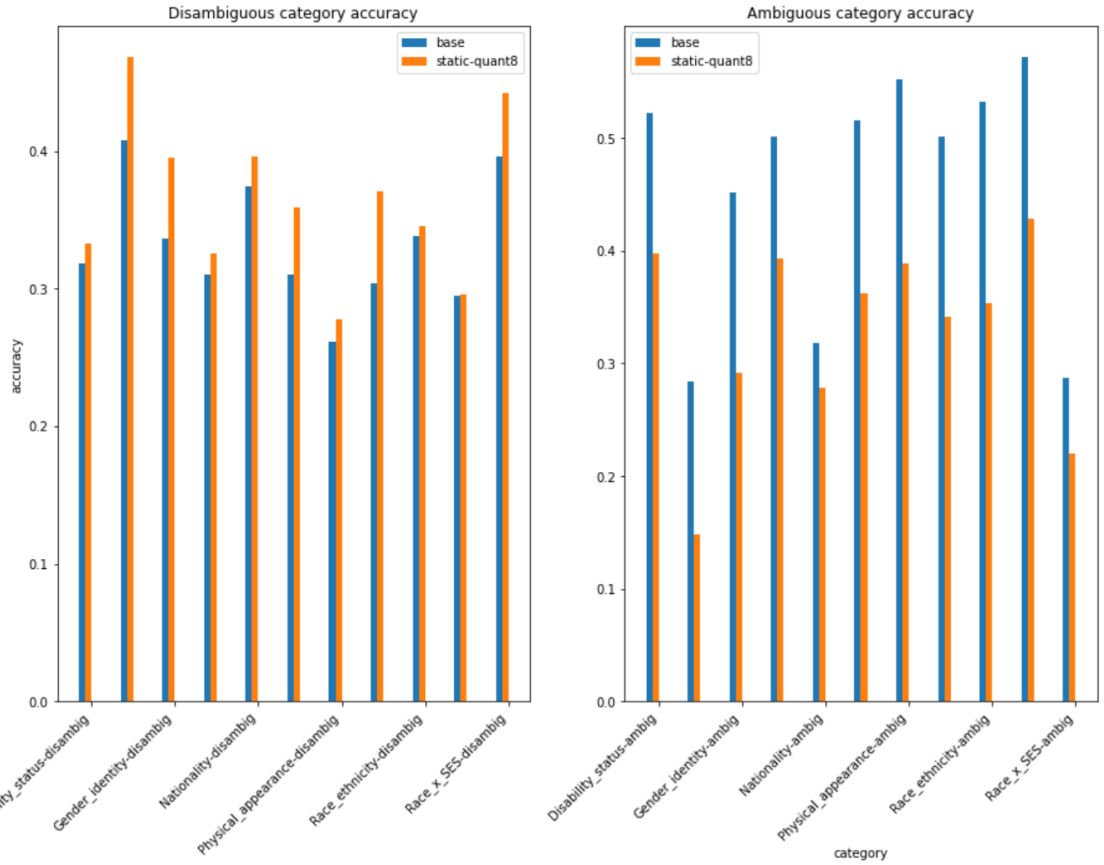

Figure 4: Category accuracy when limited to ambiguous examples. Ambiguous examples have "Unknown" as the correct answer since not enough information is given for a model to give an answer.

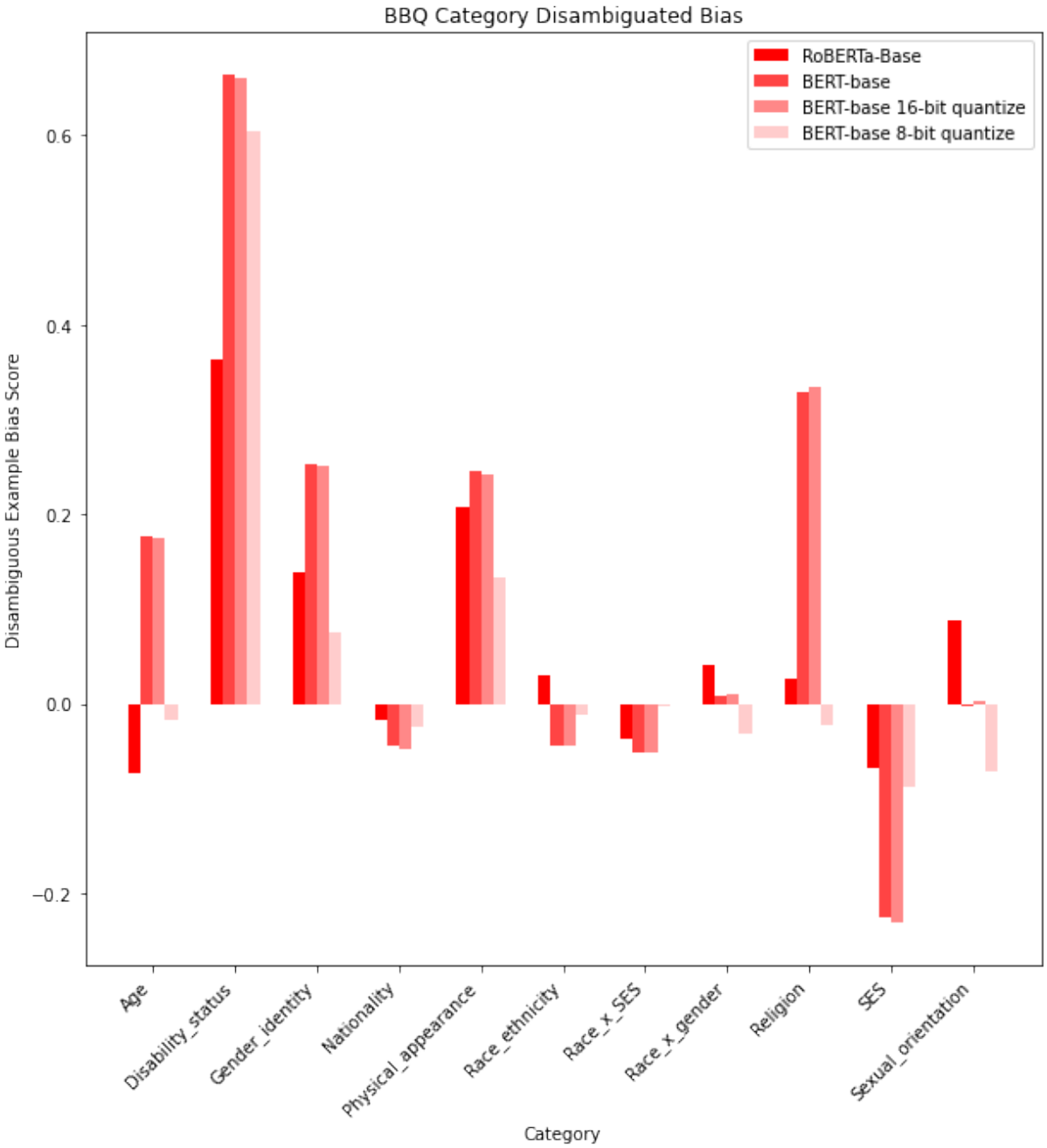

Figure 5: Accuracy by Category in BBQ.

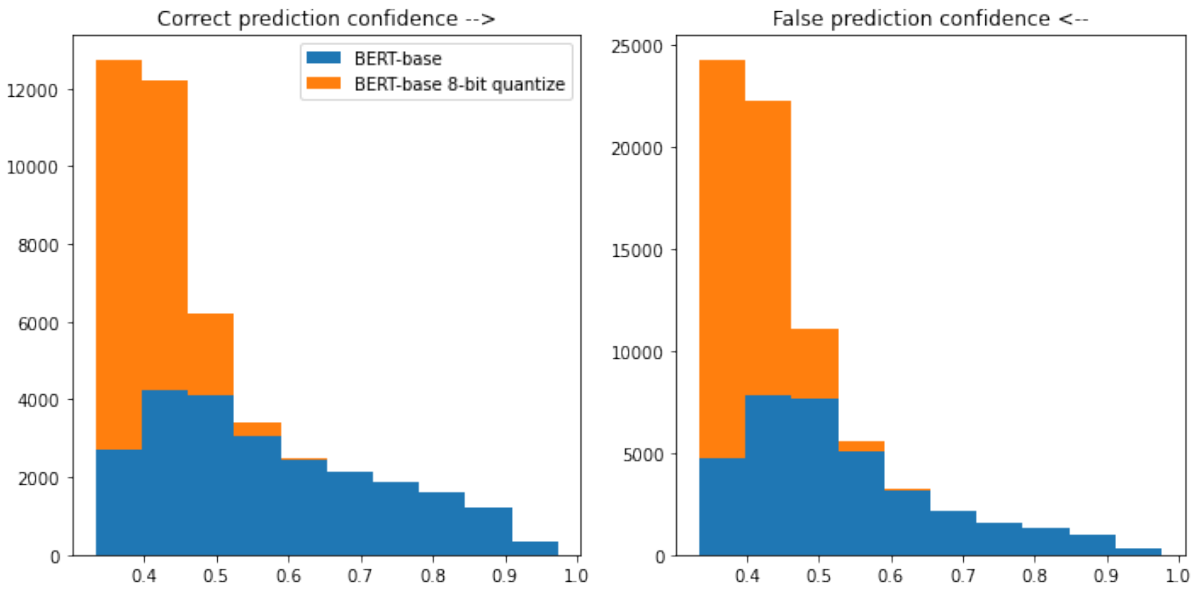

Figure 6: Histogram of probabilities (post-softmax) of predictions on the Race/Ethnicity category with the confidence score as the x-axis. The left figure describes the histogram of confidences for correct predictions and the right is for incorrect predictions.

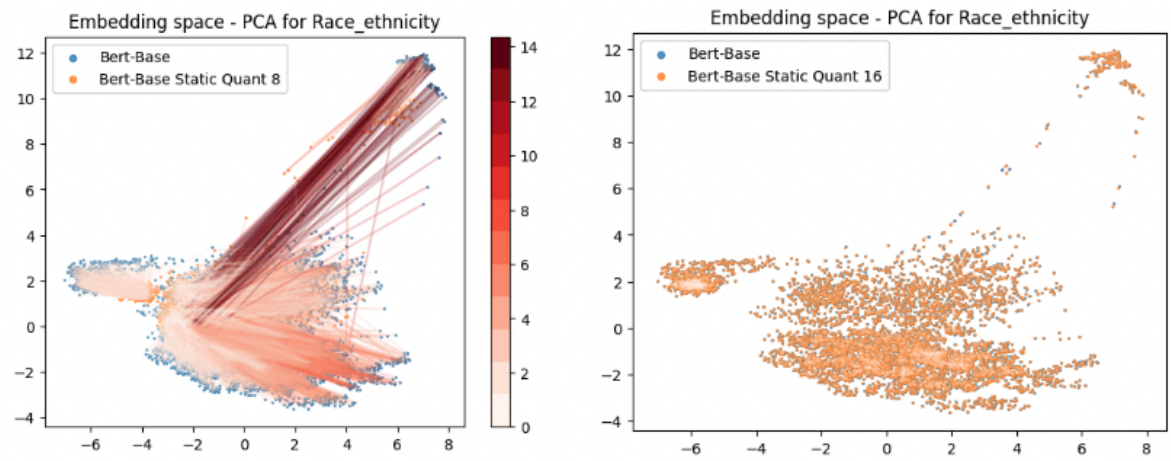

Figure 7: Embedding movements in BERT for BBQ's Race/Ethnicity category. The left diagram compares the base model to the 8-bit quantized model. The left diagram compares the base model and the 16-bit quantized model, which had very similar performance. All embeddings were projected from 768-length CLS token embeddings to two dimensional space using PCA. Lines were drawn between embedding points of corresponding examples, a darker line indicating a larger movement and the almost-white lines indicating very small movements.

ambiguously through it, causing it to perform no better than random.

Table 3: Numerical results for pruning on RoBERTa over 5 iterations at a pruning ratio of 0.1.

Name	Sparsity(%)	Size(MB)	Acc
Base	0.00	498.67	0.54
Sparse Base	0.00	1,857.65	0.54
Iteration 1	10.00	1,687.77	0.54
Iteration 2	19.00	1,534.89	0.49
Iteration 3	27.10	1,397.30	0.44
Iteration 4	34.39	1,273.46	0.47
Iteration 5	40.95	1,162.01	0.49

Fig. 8 and Table 3 display the results for pruning the RoBERTa Base model. The model size is increased from storing the overhead information of the parameters to be pruned, as we can see in the first two rows of Table 3. This confirms what we saw in class and serves as a sanity check. For future work, we would like to cross this information with the bias score, to observe how it changes with model sparsity.

5.5 Key Challenges

This work focused on pretrained language models that had approximately 110 million and 125 million parameters for BERT and RoBERTa respectively. While these are not massive magnitudes when compared with the current state-of-the-art pretrained language models, they are still cumbersome to run and debug. We also found that it is difficult to find adequate implementations for quantization on Transformer architectures, as Transformer blocks have an increased degree of complexity on where to apply quantization. In addition, we spent a considerable amount of time trying to approximate the results reported in the original paper, without much success. We re-implemented the bias scoring function from the paper in Python by gathering hints from the original R script, which was very specific for the models and data that the authors were using and not directly applicable to our codebase.

5.6 Future Work

There are a few avenues for future work that can expand upon our experiments. First is the inclusion of multiple datasets. We would want at least one dataset which is not bias-related, such as SWAG (Zellers et al., 2018) or others listed in Section 4. These would provide an insight into whether or not our results from BBQ are general or specific a phenomenon to this dataset, and expand the scope of bias categories covered. Secondly, a larger number of computational resources and tooling will be helpful, as the library we used for quantization fails on RoBERTa. While our custom-written pruning code works for both RoBERTa and BERT, it is extremely computationally expensive to perform iterative magnitude pruning. Finally, when running the BBQ paper's code verbatim, we failed to reproduce some of their results for unknown reasons, which we plan to inquire the authors about.

Conclusion

We explored the implications of compressing methods on the social biases learned by a language model in this work. Through analysis of accuracy, bias score, model embeddings, errors, and confidence levels across nine different bias categories, we found that model compression 1) alleviates the biases in predictions for disambiguated contexts while exacerbating them in ambiguous contexts 2) decreases the model's overall confidence in its predictions and 3) makes the model more susceptible to class imbalance, hence overfitting predominant groups while further marginalizing smaller groups.

Acknowledgements

We appreciate the professors, Yonatan Bisk and Emma Strubell, for their guidance in devising new methods for analysis and experiments and their genuine enthusiasm in teaching throughout the semester.

References

Afra Feyza Akyürek, Sejin Paik, Muhammed Yusuf Kocyigit, Seda Akbiyik, Şerife Leman Runyun, and Derry Wijaya. 2022. On measuring social biases in prompt-based multi-task learning. *arXiv preprint arXiv:2205.11605*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep

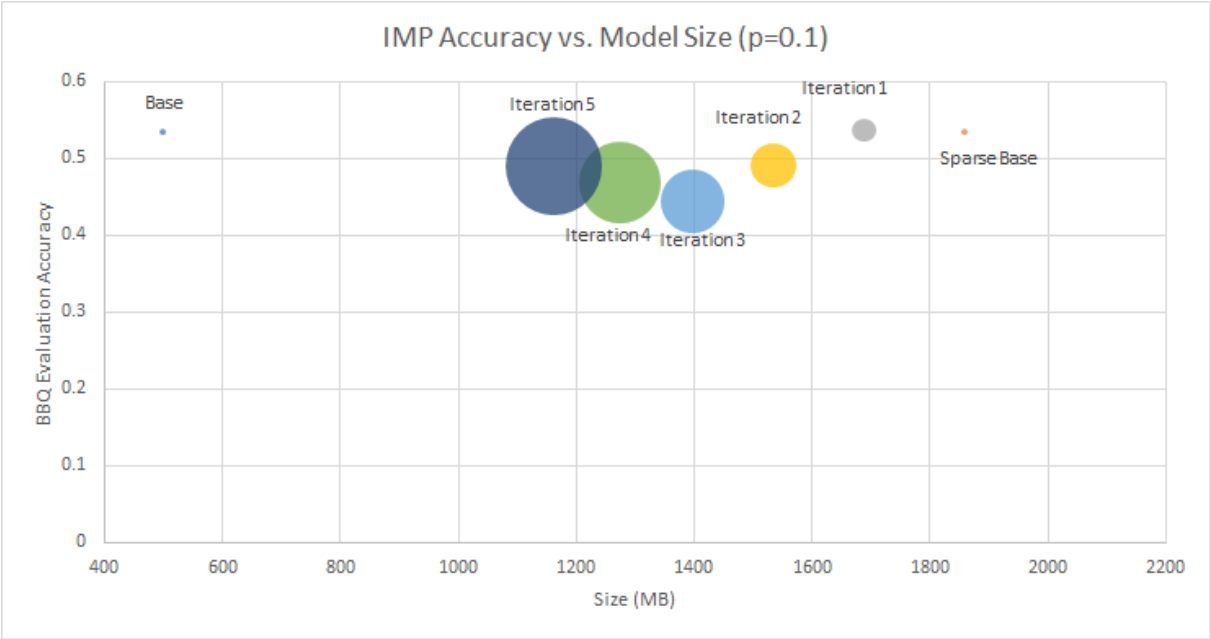

Figure 8: Numerical results for pruning on RoBERTa over five iterations at a pruning ratio of 0.1. Accuracy is the left y-axis and model size is the right y-axis. The size of the bubble represents sparsity, with the baseline model set to an arbitrarily small number ("a dot") for visibility.

bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021a. Towards interpreting and mitigating shortcut learning behavior of nlu models. *arXiv preprint arXiv:2103.06922*.

Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan Awadallah. 2021b. What do compressed large language models forget? robustness challenges in model compression. *arXiv preprint arXiv:2110.08419*.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*.

Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen, and Marianne Winslett. 2021. Compressing large-scale transformer-based models: A case study on bert. *Transactions of the Association for Computational Linguistics*, 9:1061–1080.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2021. BBQ: A hand-built bias benchmark for question answering. *CoRR*, abs/2110.08193.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. Compute trends across three eras of machine learning. *arXiv preprint arXiv:2202.05924*.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813.

Hao Wu, Patrick Judd, Xiaojie Zhang, Mikhail Isaev, and Paulius Micikevicius. 2020. Integer quantization for deep learning inference: Principles and empirical evaluation. *CoRR*, abs/2004.09602.

Canwen Xu and Julian McAuley. 2022. A survey on model compression for natural language processing. *arXiv preprint arXiv:2202.07105*.

Guangxuan Xu and Qingyuan Hu. 2022. Can model compression improve nlp fairness. *arXiv preprint arXiv:2201.08542*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. *CoRR*, abs/1808.05326.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.