Beyond Size and Accuracy: The Impact of Model Compression on Fairness

Moumita Kamal, Douglas A. Talbert
Tennessee Tech University, Cookeville, TN, USA
mkamal42@tntech.edu
dtalbert@tntech.edu

Abstract

Model compression is increasingly popular in the domain of deep learning. When addressing practical problems that use complex neural network models, the availability of computational resources can pose a significant challenge. While smaller models may provide more efficient solutions, they often come at the cost of accuracy. To tackle this problem, researchers often use model compression techniques to transform large, complex models into simpler, faster models. These techniques aim to reduce the computational cost while minimizing the loss of accuracy. The majority of the model compression research focuses exclusively on model accuracy and size/speedup as performance metrics. This paper explores how different methods of model compression impact the fairness/bias of a model. We conducted our experiments using the COMPAS Recidivism Racial Bias dataset. We evaluated a variety of model compression techniques across multiple bias groups. Our findings indicate that the type and amount of compression have substantial impact on both the accuracy and fairness/bias of the model.

Introduction

Neural networks are effective in solving various practical problems. However, they require vast space, power, and computation time, which can limit their utility, and smaller models may compromise accuracy. To overcome this, researchers have developed model compression techniques to create simpler and faster models without compromising performance quality.

Model compression research has mainly focused on accuracy, size, and speed. However, machine learning algorithms are prone to biases that cause them to make unfair decisions. It is crucial to understand how much fairness is sacrificed after compression. Our study investigates the impact of model compression techniques on fairness and bias.

We used the COMPAS Recidivism Racial Bias dataset (Angwin et al. 2022) for our study. It contains information on criminal defendants from Broward County, FL, their arrest and subsequent re-arrest or failure to appear in court. Additionally, it includes the individual's race, gender, age, and other attributes (Brackey 2019).

Our experiments found that model compression can affect the fairness/bias of a model. Specifically, some techniques can lead to increased bias in the model. Our intent is to provide insights into the impact of compression on fairness to help researchers develop less biased models.

Background

Model Compression

Model compression is about using a smaller, faster model to perform the function of a larger, slower model. A compressed model approximates the larger model while being faster and more efficient (Buciluǎ, Caruana, and Niculescu-Mizil 2006).

Model compression aims to reduce the computational cost while minimizing loss of accuracy. Popular model compression techniques include pruning (Voita et al. 2019; Prasanna, Rogers, and Rumshisky 2020), quantization (Han, Mao, and Dally 2015; Cheong and Daniel 2019), and knowledge distillation (Polino, Pascanu, and Alistarh 2018; Tan et al. 2018). Pruning removes some connections, quantization reduces weight precision, and knowledge distillation transfers knowledge from a complex to a simpler model (Jiao et al. 2019).These techniques have shown an ability to significantly reduce model size, making execution faster and more affordable (Deng et al. 2020). Moreover, this makes energy-efficient models suitable for real-world applications on resource-limited devices.

Types of Model Compression

Pruning: Pruning optimizes neural networks by selectively removing less important connections and neurons(Dai, Yin, and Jha 2019; He et al. 2014). This technique reduces network size post-training, followed by fine-tuning to enhance performance, ultimately compressing networks and saving memory. Researchers commonly use two techniques:

(1) Weight Pruning : Removing connections by setting specific weights to zero (Han et al. 2015; Guo et al. 2019).

(2) Node/Neuron Pruning: Removing entire neurons based on their contributions and activation patterns (Murray and Chiang 2015; Pan, Dong, and Guo 2016).

Quantization: Quantization is a technique that reduces the precision of each weight in a model by using a smaller number of bits to represent it (He et al. 2016). This technique allows for the deployment of large models on small devices with limited memory such as microcontrollers, wearables, and edge devices (Shen et al. 2020). By using 8-bit integers (INT8) instead of the standard numerical format (FP32), the memory required to store the model is reduced without significant accuracy loss (He et al. 2016). Moreover, quantization can significantly improve inference time (Zhou et al. 2017).

In this paper, we used two methods for quantization: *post-training quantization* - which involves quantizing a pre-trained model, and *quantization-aware training* - where quantization occurs during training.

Distillation: Knowledge distillation (Buciluǎ, Caruana, and Niculescu-Mizil 2006; Hinton, Vinyals, and Dean 2015) compresses a large pre-trained machine learning model into a smaller one. The larger (teacher) model, guides and trains the smaller (student) model, improving its performance (Sau and Balasubramanian 2016). The teacher's knowledge is transferred to the student by minimizing a loss function using the teacher's predicted class probabilities (Hinton, Vinyals, and Dean 2015). However, this distribution often provides little additional information beyond the dataset labels.

In this paper, we have focused our experiments on pruning and quantization for model compression.

Model Fairness

Bias in machine learning systems is a widely researched topic, with various concepts of fairness(Mehrabi et al. 2021; Verma and Rubin 2018). Incidents of bias impacting certain groups have further emphasized the importance of this research (Angwin et al. 2022; **?**). When assessing a model's bias, it is common to analyze its performance across different demographic groups to determine if the predictions are consistent across these groups or if certain groups are disproportionately affected by errors (DeAlcala et al. 2023). Several metrics are used to calculate model fairness to identify biases and guide changes to improve fairness.

In this paper, we have used *equalized odds* as our fairness metric (Phillips et al. 2023). This metric considers a model to be fair if the subgroups have equal sensitivity (TPR) and specificity (1-FPR). We used the bias function introduced in (Phillips et al. 2023) defined as follows:

$$\begin{aligned}&\text{bias}(M, G = \{g_1, \ldots, g_k\}) \\ &= \sum_{i=1}^{k} \sum_{j=i+1}^{k} |\text{Specificity}(M, g_j) - \text{Specificity}(M, g_i)| \\ &+ \sum_{i=1}^{k} \sum_{j=i+1}^{k} |\text{Sensitivity}(M, g_i) - \text{Sensitivity}(M, g_j)|\end{aligned} \tag{1}$$

where, M is the machine learning model and $g_1, g_2, ..., g_k$ are subsets of demographic groups. The smaller this value is, the fairer the model.

Fairness in Model Compression

Some research has been conducted on the impact of model compression on bias. Xu et al. discuss how distillation and pruning affect toxicity and bias in generative language models (Xu and Hu 2022). Their findings suggest that knowledge distillation produces less toxic and possibly less biased models. Joseph et al. propose a novel loss function for model compression, which was able to preserve the fairness of the model in most cases (Joseph et al. 2020). Stoychev and Gunes explore the effects of compression on accuracy and fairness, measuring fairness as the difference in accuracy in two bias groups (Stoychev and Gunes 2022).

Methodology

The Data

We used the COMPAS Recidivism Racial Bias dataset (Angwin et al. 2022) to analyze the impact of model compression on fairness/bias. It focuses on the COMPAS algorithm used to predict criminal recidivism (Dressel and Farid 2018) and has been found biased against African American defendants and in favor of Caucasian defendants (Angwin et al. 2022). The dataset consists of over 18,000 instances of criminal defendant information, with 34 features including demographic information, charges, prior history, and COMPAS score, alongside the two-year outcomes marked as a *recid* value of 0 or 1. We focused on gender, race, and age for bias analysis, categorizing instances into six racial groups. However, due to limited instances in some races and the focus on COMPAS bias against African American and in favor of Caucasian defendants, we excluded samples from other races, resulting in a dataset of 16,512 instances detailed in Table 1.

Bias Group	**Size**
Gender = Male	13,465
Gender = Female	3,047
Race = African American	10,074
Race = Caucasian	6,438
Age $<$ 25	3,862
Age = 25-45	9,423
Age $>$ 45	3,227

Table 1: Bias Subgroups in the COMPAS Dataset

Experimental Methodology

We constructed a fully connected neural network with 8 hidden layers as our base model to predict criminal defendants' recidivism probabilities. The data was split into a 70% training set and a 30% test set. After training, we evaluated the model's accuracy, loss, sensitivity, and specificity for both the entire dataset and each demographic subgroup. The model was built using the Adam optimizer with binary cross entropy as the loss function. We saved the trained model's

details including size, accuracy, loss, and bias (calculated using Equation 1). To compress the models, we applied quantization and pruning. For quantization, we used quantization-aware training to ensure better performance. In pruning, we employed weight pruning with the magnitude pruner, which applies a threshold function on each weight tensor to preserve weights with high absolute values (Sanh, Wolf, and Rush 2020). We pruned the model to various sparsity levels to examine its impact on fairness.

We repeated these experiments on 10 models and their compressed versions, averaging the results. Algorithm 1 outlines our experiment workflow, detailed below.

Experiment 1 - Compressed models against the Baseline: We compared performance between the base model and the compressed models. We built our base model (baseline), the quantized model, and the pruned model at 80% sparsity. We computed and compared the size, accuracy and bias of the three models for each of the demographic groups.

Experiment 2 - Pruning with Different Levels of Sparsity: With weight pruning, it is possible to prune a model to various levels of sparsity. Our next experiment was to compare how different levels of pruning affect a model's performance. For this experiment, we measured model sparsities of 50%, 60%, 70%, 80%, and 90%. We then recorded the accuracy, loss, size and bias for these models and compared against the baseline.

Results and Discussion

Experiment 1: We first report the results of how the size of a model varies with different methods of compression. We then report the group-specific bias values for each of the models and identify the fairest model from among the models tested. Lastly, we compare the accuracy of each demographic subgroup for each model.

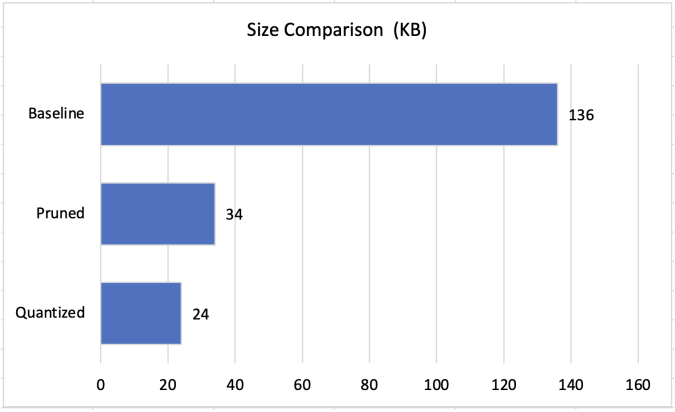

Figure 1: Size Comparison - Baseline vs Compressed

Figure 1 shows the size comparisons among the base model and the compressed models. The base model has a size of 136 kilobytes (KB). After quantization, however, the size reduced to 24 KB, and pruning the base model to 80% sparsity reduced its size to 34 KB.

Table 2 shows the effect of model compression on bias. We observe that the quantized model, even though much smaller in size, produces very similar, if not better models with respect to fairness. The bias value for the quantized model for race is 0.0917 which is less than that of the baseline 0.1351. Similarly, we see that the model is slightly less biased for age group (0.859 compared to 0.89).

Among the demographic groups, the bias values for age are much higher compared to the other two groups. This is because, for gender and race, we have calculated bias for a single pair of subgroups. While for age, we calculated bias by summing the biases across the pairwise combinations for all three subgroups, resulting in three pairs of subgroups.

	Base Model	Quantized	Pruned(80%)
Gender	0.0715	0.0793	0.280
Race	0.1351	0.0917	0.3207
Age	0.89	0.859	1.405

Table 2: Experiment 1 - Bias Comparison

We also compute the difference in accuracy for each model for each group. We observe that with quantization, the accuracy is very close to and at times better than the baseline. However, with pruning, the compressed model was unable to produce very accurate results. 80% pruning consistently demonstrated an accuracy difference of 8-10% lower compared to the other 2 models.

Experiment 2: Next, we computed the effects of pruning levels on the size, accuracy and bias of a model. We pruned the model to 50%, 60%, 70%, 80%, and 90% sparsity and compared the performance among each other and with the baseline.

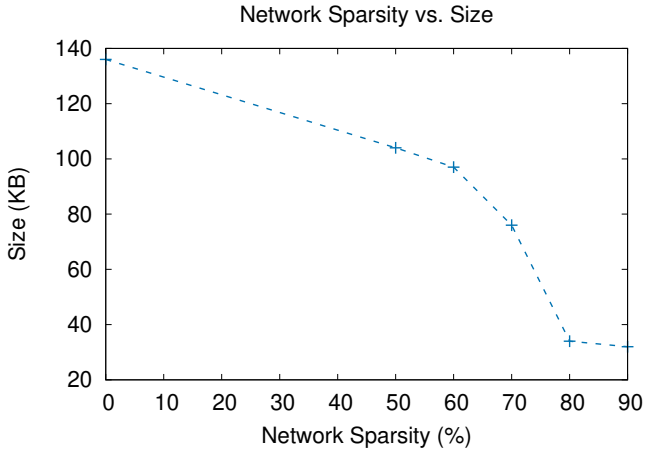

Figure 2: Size Comparison - Baseline vs Different Sparsities

Figure 2 shows the size of the models at different levels of pruning. The baseline model at 0% sparsity had a size of 136 KB. As we pruned the network more the size of the models gradually went down where at 90% sparsity, the model size is 32 KB. We also observed that the model size does not

Algorithm 1 Workflow

$G \leftarrow \{g_1, \cdots, g_k\}$
$D \leftarrow dataset$
for $i \leftarrow 0$ to $length(G)$ **do**
  $X, X', y, y' \leftarrow getTrainTestData(D, g_iTag, randomState)$
  $M \leftarrow buildNNModel$
  $acc\_g_i, sensitivity\_g_i, specificity\_g_i \leftarrow getModelPerformance(M, X, X', y, y')$
  $M_p \leftarrow getPrunedModel(M)$
  $acc\_M_p\_g_i, sensitivity\_M_p\_g_i, specificity\_M_p\_g_i \leftarrow getModelPerformance(M_p, X, X', y, y')$
  $M_q \leftarrow getQuantModel(M)$
  $acc\_M_q\_g_i, sensitivity\_M_q\_g_i, specificity\_M_q\_g_i \leftarrow getModelPerformance(M_q, X, X', y, y')$
end for
for $i \leftarrow 0$ to $length(G)$ **do**
  for $j \leftarrow i + 1$ to $length(G)$ **do**
    $bias\_M \leftarrow |specificity\_g_j - specificity\_g_i| + |sensitivity\_g_j - sensitivity\_g_i|$
    $bias\_M_p \leftarrow |specificity\_M_p\_g_j - specificity\_M_p\_g_i| + |sensitivity\_M_p\_g_j - sensitivity\_M_p\_g_i|$
    $bias\_M_q \leftarrow |specificity\_M_q\_g_j - specificity\_M_q\_g_i| + |sensitivity\_M_q\_g_j - sensitivity\_M_q\_g_i|$
  end for
end for

reduce much from dense (0% sparsity) to 50% sparsity but the size curve takes a steep dive from 60% onwards.

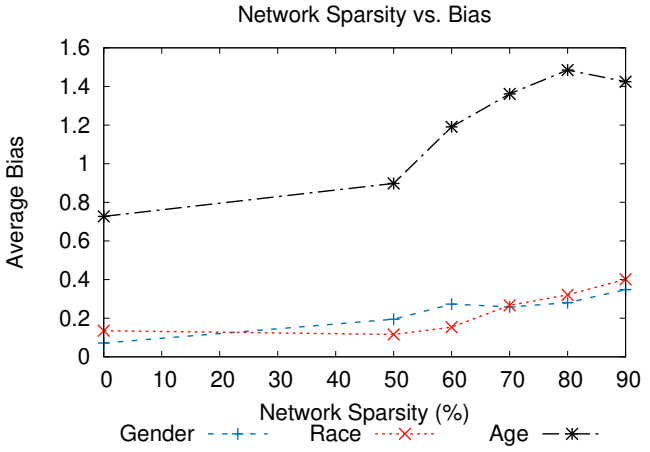

Figure 3: Bias Comparison - Baseline vs Different Levels of Sparsity

Figure 3 demonstrates the effects of pruning on bias of the model. We observe that the average bias values go up for all demographic groups as the model becomes more sparse. Once again, average bias for age is much higher than the other two groups since it totals the pairwise biases over three pair of subgroups. We observe a 3-5 times increase in bias with the baseline bias rising from 0.0715 to 0.3478 for gender, 0.1351 to 0.4015 for race and 0.89 to 1.4248 for age group.

We further observe that the sparser the network becomes, the less accurate it is at predicting recidivism. We see a slight drop in accuracy from the baseline to 50% sparsity (84.74% - 84.02% for gender, 83.59% - 82.4% for race, and 84.6% - 82.64% for age) and then for each increased level of sparsity, there is a 2% - 3% (approx.) decrease in model accuracy.

Conclusion

This paper delves into the crucial but often overlooked aspect of model fairness in the context of model compression. Through experiments performed on the COMPAS dataset, this study showed that model compression techniques can indeed influence the fairness/bias of a model. Specifically, quantization exhibited potential for improving model fairness, while pruning appeared to increase bias. Moreover, we observed that bias tends to increase as model sparsity increases.

This research explores the need for a comprehensive understanding of the trade-off between model compression and fairness. By studying the impact of compression techniques on model fairness, we pave the way for the development of fairer, less biased machine learning models. As we continue to explore the complex relationship between model compression and fairness, we can work towards creating trustworthy and unbiased AI systems that benefit society as a whole.

Limitations and Future Work

The limitations of our experiments include the fact that our dataset contains defendant information from only one county in Florida and might not generalize. Moreover, the data as provided and used contained some duplicates making it vulnerable to potential data leaks. Additionally, we only explored one approach compute bias and two compression techniques. A broader analysis of more bias metrics and compression techniques might yield different results.

In the future, want to extend this research to include other model compression techniques and explore bias mitigation techniques to improve the performance of the compressed models. Furthermore, we want to explore other features for bias outside of the three attributes that we used in our research.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications. 254–264.

Brackey, A. 2019. *Analysis of Racial Bias in Northpointe's COMPAS Algorithm*. Ph.D. Dissertation, Tulane University School of Science and Engineering.

Buciluǎ, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 535–541.

Cheong, R., and Daniel, R. 2019. transformers. zip: Compressing transformers with pruning and quantization. *Technical report, tech. rep., Stanford University, Stanford, California*.

Dai, X.; Yin, H.; and Jha, N. K. 2019. Grow and prune compact, fast, and accurate lstms. *IEEE Transactions on Computers* 69(3):441–452.

DeAlcala, D.; Serna, I.; Morales, A.; Fierrez, J.; and Ortega-Garcia, J. 2023. Measuring bias in ai models: an statistical approach introducing n-sigma. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1167–1172. IEEE.

Deng, L.; Li, G.; Han, S.; Shi, L.; and Xie, Y. 2020. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE* 108(4):485–532.

Dressel, J., and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4(1):eaao5580.

Guo, F.-M.; Liu, S.; Mungall, F. S.; Lin, X.; and Wang, Y. 2019. Reweighted proximal pruning for large-scale language representation. *arXiv preprint arXiv:1909.12486*.

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.

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*.

He, T.; Fan, Y.; Qian, Y.; Tan, T.; and Yu, K. 2014. Reshaping deep neural network for fast decoding by node-pruning. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 245–249. IEEE.

He, Q.; Wen, H.; Zhou, S.; Wu, Y.; Yao, C.; Zhou, X.; and Zou, Y. 2016. Effective quantization methods for recurrent neural networks. *arXiv preprint arXiv:1611.10176*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Joseph, V.; Siddiqui, S. A.; Bhaskara, A.; Gopalakrishnan, G.; Muralidharan, S.; Garland, M.; Ahmed, S.; and Dengel, A. 2020. Going beyond classification accuracy metrics in model compression. *arXiv preprint arXiv:2012.01604*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54(6):1–35.

Murray, K., and Chiang, D. 2015. Auto-sizing neural networks: With applications to n-gram language models. *arXiv preprint arXiv:1508.05051*.

Pan, W.; Dong, H.; and Guo, Y. 2016. Dropneuron: Simplifying the structure of deep neural networks. *arXiv preprint arXiv:1606.07326*.

Phillips, K.; Brown, K.; Talbert, S.; and Talbert, D. 2023. Group bias and the complexity/accuracy tradeoff in machine learning-based trauma triage models. In *The International FLAIRS Conference Proceedings*, volume 36.

Polino, A.; Pascanu, R.; and Alistarh, D. 2018. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*.

Prasanna, S.; Rogers, A.; and Rumshisky, A. 2020. When bert plays the lottery, all tickets are winning. *arXiv preprint arXiv:2005.00561*.

Sanh, V.; Wolf, T.; and Rush, A. 2020. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in Neural Information Processing Systems* 33:20378–20389.

Sau, B. B., and Balasubramanian, V. N. 2016. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*.

Shen, S.; Dong, Z.; Ye, J.; Ma, L.; Yao, Z.; Gholami, A.; Mahoney, M. W.; and Keutzer, K. 2020. Q-bert: Hessian based ultra low precision quantization of bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8815–8821.

Stoychev, S., and Gunes, H. 2022. The effect of model compression on fairness in facial expression recognition. *arXiv preprint arXiv:2201.01709*.

Tan, S.; Caruana, R.; Hooker, G.; and Lou, Y. 2018. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 303–310.

Verma, S., and Rubin, J. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, 1–7.

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. *arXiv preprint arXiv:1905.09418*.

Xu, G., and Hu, Q. 2022. Can model compression improve nlp fairness. *arXiv preprint arXiv:2201.08542*.

Zhou, S.-C.; Wang, Y.-Z.; Wen, H.; He, Q.-Y.; and Zou, Y.-H. 2017. Balanced quantization: An effective and efficient approach to quantized neural networks. *Journal of Computer Science and Technology* 32:667–682.