What's Left After Distillation?
How Knowledge Transfer Impacts Fairness and Bias

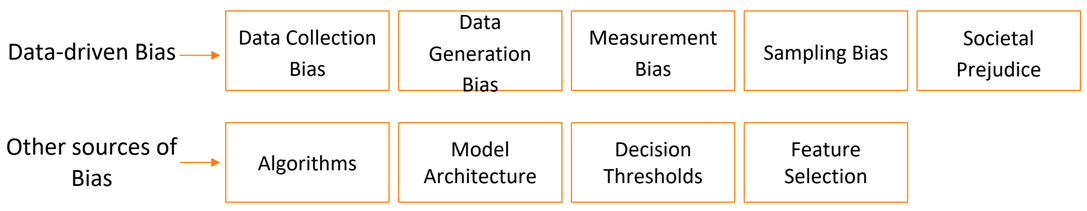
Aida Mohammadshahi, Yani Ioannou

Motivation

- Despite the widespread use of distillation, evaluation of the impact of distillation since its proposal [1] has overwhelmingly focused on the impact it has on generalization performance.
- Recent works [2] demonstrated that test accuracy alone was not sufficient to understand the full impact of two popular approaches to reducing DNN complexity: pruning and quantization.
- It is vital to understand if application of distillation uniformly affect accuracy across all classes, or are certain classes more significantly affected by distillation as with pruning, and What is the impact of distillation on the model's bias and fairness.

Bias and Fairness

- The systemic bias that can be present in the decisions and predictions made by Machine Learning (ML) models is called algorithmic bias.
- There are a wide range of sources of bias in the deep learning context, although existing bias in the training dataset is one of the main reasons for biased prediction.



- Bias in a model can lead to unfair outcomes.
- When a model consistently favours certain groups over others due to its biases, it fails to treat individuals equitably, resulting in unfair decisions.

Methods

Welch's t-test:

$$H_0: \frac{\beta^{c_i}}{\beta^M} = \frac{\beta_T^{c_i}}{\beta_T^M} \qquad H_1: \frac{\beta^{c_i}}{\beta^M} \neq \frac{\beta_T^{c_i}}{\beta_T^M}$$

β^M -> model's test accuracy over all classes

β^{c_i} -> class-wise test accuracy

Matrix of Disagreement:

$$CMP(f_{(x_n)}, g_{(x_n)}) = \begin{cases} 0 \ if \ f_{(x_n)} = g_{(x_n)} \\ 1 \ if \ f_{(x_n)} \neq g_{(x_n)} \end{cases}$$

Fairness Metrics:

- **Demographic Parity Difference:**

$$DPD = \max_{a \in A} P(\hat{Y} = 1 \mid A = \alpha) - \min_{a \in A} P(\hat{Y} = 1 \mid A = \alpha)$$

A -> demographic groups

- **Equalized Odds Difference:**

$$EOD = \max (TPR\ Difference, FPR\ Difference)$$

Results

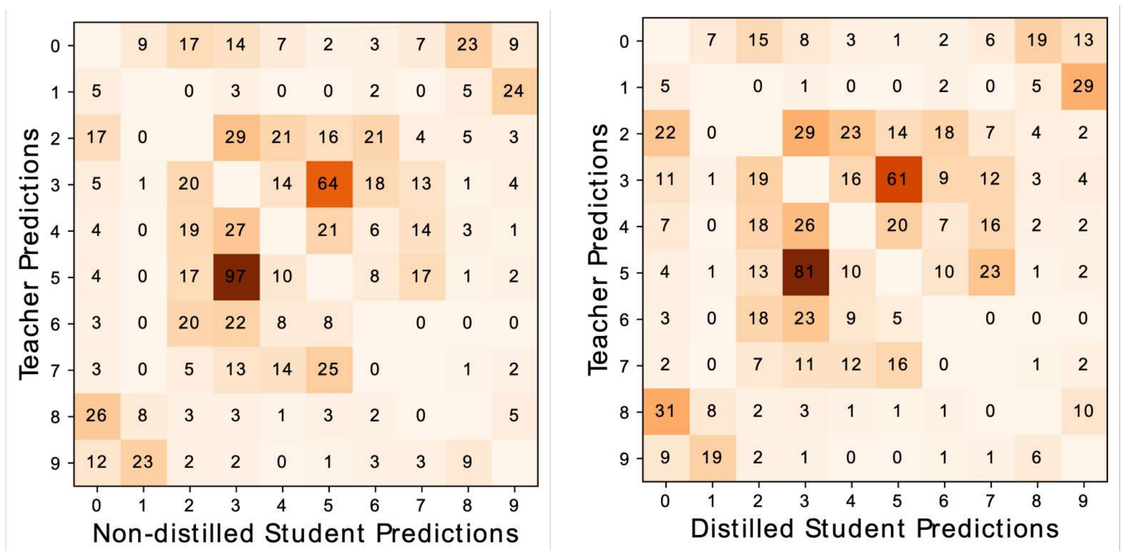


Figure 1. Class-wise Disagreement. Disagreement between a ResNet-56 teacher and ResNet-20 (left) non-distilled/(right) distilled student for CIFAR-10 using T = 9. The diagonals are excluded since here both models predict the same class without any disagreement.

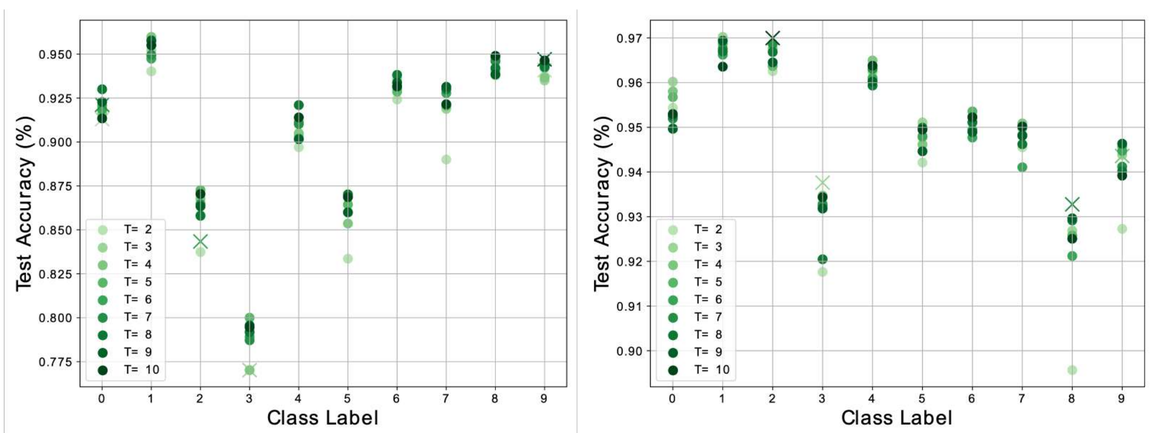


Figure 2. Class-wise Bias and Distillation. ResNet-20/ResNet-56 student/teacher models on CIFAR-10 (left) and SVHN (right) over a range of temperatures T over five random initializations. Classes with statistically significant relative changes between the non-distilled student and the distilled student are noted with ×.

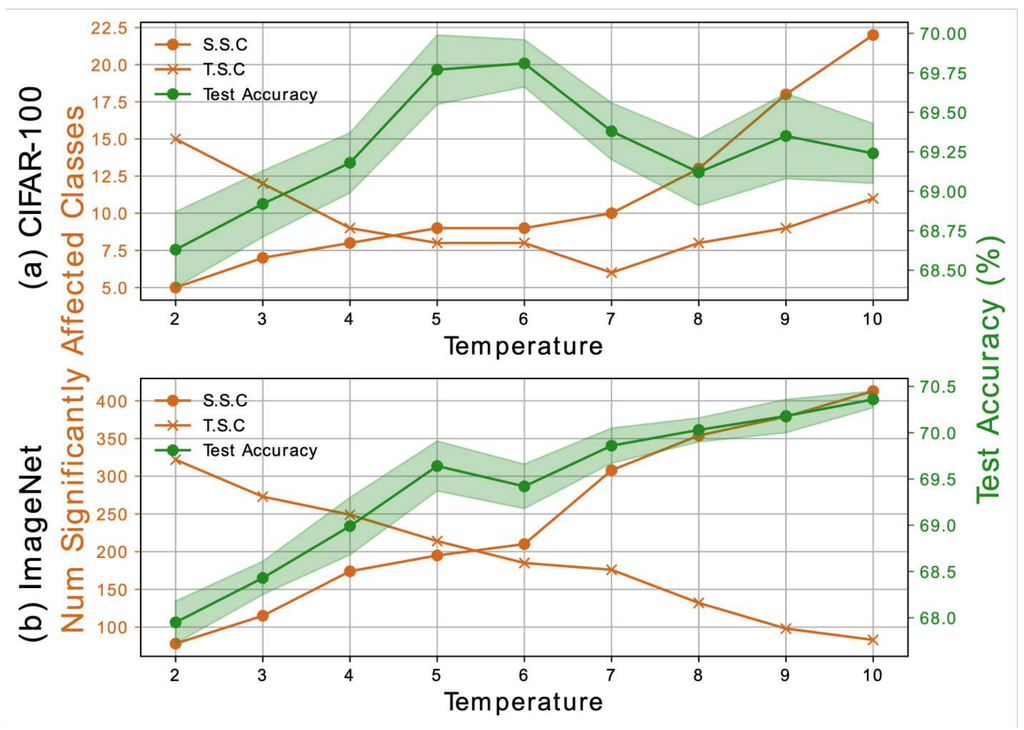


Figure 3: Temperature vs. Test Accuracy/Class Bias. Number of non-distilled vs. distilled student significantly affected classes (S.S.C.) and the number of teacher vs. distilled student significantly affected classes (T.S.C.) by distillation in (a) CIFAR-100 (ResNet-56/ResNet-20) and (b) ImageNet datasets (ResNet-50/ResNet-18).

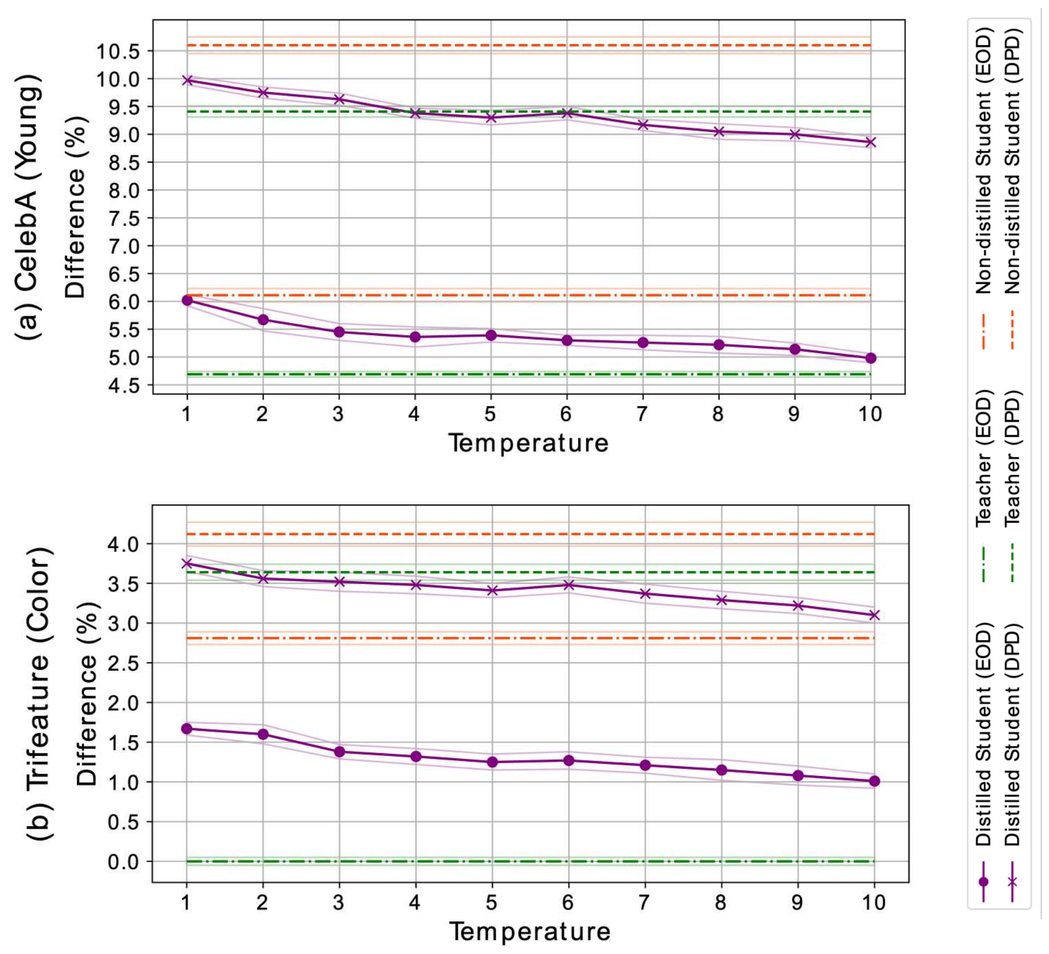


Figure 4: Evaluation of Fairness Metrics for Distilled Students. EOD and DPD are reported in % and lower values indicate improved fairness. (a) illustrates fairness metrics for the CelebA dataset concerning the 'Young' demographic attribute, and (b) metrics for the Trifeature dataset with regard to the 'color' attribute.

Insights

- The impact of distillation is not random: there is a statistically significant difference in class-level accuracy between a non-distilled/distilled student, and teacher/distilled student.
- The number significantly affected classes increases with higher temperatures comparing non-distilled vs. distilled student, whereas with teacher vs. distilled student they decrease.
- Distillation influences the class bias and fairness of distilled student models, even where there is no substantial change in the model's overall test accuracy.
- Distillation improves the student model's fairness concerning demographic attributes, and employing higher temperatures increases the model's fairness.
- Enhancements in fairness may not always align with improvements in the model's generalization performance and a trade-off should be selected based on the specific application

References

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[2] SaraHooker, Aaron Courville, Gregory Clark, YannDauphin, andAndrea Frome. What do compressed deep neural networks forget? arXiv preprint arXiv:1911.05248, 2019.