Optimizing Efficiency, Ensuring Equity: Advancing Knowledge Distillation with a Focus on Bias Reduction

Lindsey Bang | $lindseyejbang@berkeley.edu$
Michelle Bolner | $bolnerm@berkeley.edu$
Keith Hutton | $keith.hutton@berkeley.edu$
Landon Morin | $morinlandon@berkeley.edu$

Abstract

This paper proposes a framework for model compression and debiasing in the context of neural network training by combining two techniques: knowledge distillation and adversarial learning. The primary objective was to create a more efficient and unbiased student model by transferring knowledge from a complex teacher model while simultaneously mitigating biases inherent in the training data.

Despite a lack of high quality, trainable fairness datasets, it was found that leveraging adversarial debiasing can reduce bias, as measured by disparity, and increase accuracy by re-balancing predictions in favor of under-represented attributes within a class. This study focuses on debiasing in regards to protected characteristics, specifically gender.

This work contributes to the ongoing research in model compression and fairness in machine learning, offering a comprehensive approach that simultaneously addresses efficiency, performance, and bias concerns. The proposed framework has the potential to advance the deployment of machine learning models in real-world applications where resource constraints and ethical considerations continue to pose challenges to AI advancement.

1 Introduction

Deep Neural Networks (DNNs) have emerged as powerful tools across many industries, showcasing remarkable accuracy over a spectrum of complex tasks. However, this surge in capability comes with costs: increased accuracy and functionality often result in increased compute and storage requirements, and models may exhibit the biases of the data on which they were trained. These risks pose challenges to advancement in the field, rendering these models impractical and potentially harmful if deployed in their biased states.

To mitigate the challenges presented by the size of DNNs, the field of neural network compression has emerged as a key area of interest. This approach substantially reduces computational and energy costs at inference time, thereby enhancing the sustainability and cost-effectiveness of deep learning tasks. The authors note the potential tradeoff of degraded performance when applying compression techniques. They also recognize that most of the studies within this domain tend to report on a limited set of metrics, primarily focused on size reduction and accuracy. To encourage broader adoption and provide a more thorough understanding of the impact of compression, this research includes expanding the performance metrics considered. This is based on the understanding that solely relying on accuracy may be insufficient in providing a comprehensive view into the effects of compression on performance.

Knowledge distillation has surfaced as a leading paradigm for neural network compression, routinely employed across all types of networks. This method entails the training of a smaller, less computationally expensive model (student model) to approximate the functionality of a more resource-intensive network (teacher model). However, current bleeding edge knowledge distillation models often benchmark against accuracy and parameter size reduction only (Dong et al. (2023)). Beyond this, and perhaps more importantly, research often neglects to report and address fairness. Fairness consideration is critical in knowledge distillation, given that the student model tends to inherit and potentially amplify biases inherent in the teacher model (Ahn et al. (2022)). This study addresses this issue by taking preliminary steps toward the integration of fairness considerations within the knowledge distillation framework, specifically tailored for image classification neural networks. To maximize fairness and performance, the authors integrate an adversarial network into the student model for debiasing image classification tasks.

This research addresses these issues, placing a strong emphasis on the reporting and mitigation of bias as a central focus of this study. This project contributes to existing research in the field of neural network compression and fairness, focusing on a more comprehensive evaluation and addresses key challenges within the knowledge distillation methodology.

2 Problem Statement and Definition

Consider a knowledge distillation framework which maximizes the accuracy of the student model while minimizing bias as defined by disparity. The problem can be framed as follows:

Let the disparity metric, D, be defined as the absolute value of difference in recall, a proxy for bias.

$$D = |Recall(f(l_1, I_{l_1}^C, C)) - Recall(f(l_2, I_{l_2}^C, C))|$$

In this study, the authors investigate debiasing models, f, using gender as a protected characteristic, l_i. *Recall* signifies the model's performance in predicting a given class, C, for each gender. The set of images in the class with a specific attribute (gender) is denoted as $I_{l_i}^C$.

Let A be the accuracy. The objective function is formalized as:

$$\text{Maximize } \ A - \lambda \cdot D$$

where λ is a hyperparameter that defines the prioritization weight on bias as defined by disparity, D. While most debiasing methodologies generate tradeoffs between bias and accuracy, the approach formalized above extends the possibility of increasing accuracy and minimizing bias simultaneously, rendering the approach pareto optimal (Savani et al. (2020)).

Example Consider two images from a hypothetical dataset of a person in medical clothing, one male, l_1 and one female, l_2. The image class, C is surgeon for both the male and the female, but female medical personnel in the dataset are more frequently labeled as nurses. The teacher model may be prone to predict that females dressed in medical clothing are nurses. The student model is apt to learn this behavior from the teacher, and furthermore, to amplify this bias. Knowledge distillation with adversarial debiasing penalizes this behavior while maximizing accuracy.

3 A Review of Work in the Field

3.1 Knowledge Distillation: Leading Frameworks

Grounding this paper in previous work required an extensive literature review of both knowledge distillation and bias reduction. While much work has been done in domain-specific knowledge distillation, four primary generalizable frameworks (1) Classical Knowledge Distillation, (2) Curriculum Temperature Knowledge Distillation, (3) Relational Knowledge Distillation, and (4) Regularizing Feature Norm and Direction were identified.

3.1.1 Classical Knowledge Distillation

Hinton et al. (2015) proposed a method called Classical Knowledge Distillation (CKD), whereby the teacher model is trained on inputs, and the student model is trained on identical inputs in addition to the distance of the student predictions from the teacher predictions (**Figure 1**).

The principle component of CKD is the loss function, which employs a temperature parameter, T, to soften the teacher model's output probabilities, facilitating the transfer of more nuanced information to the student model. The CKD loss function is as follows:

$$\mathcal{L}_{ckd} = (1 - \alpha) \cdot \mathcal{L}_{ce} + \alpha \cdot \mathcal{L}_{KLDivergence}$$

where L_{ce} is the crossentropy loss of the student predictions; $L_{KLDivergence}$ is the Kullback-Leibler divergence between the student and teacher logits; and α weights the loss in favor of the hard predictions from the student or the logits from the teacher.

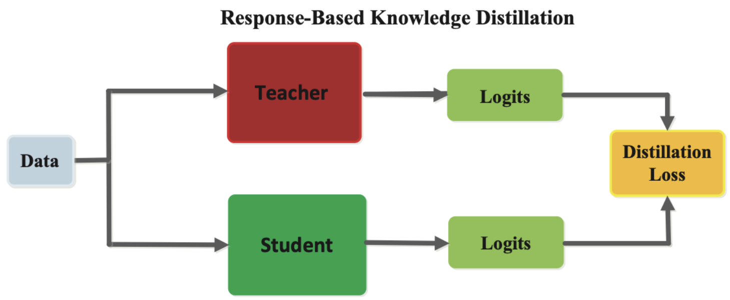

Figure 1: Model architecture of Classical Knowledge Distillation
Gou et al. (2020)

A subset of CKD was proposed by Li et al. (2022) called Curriculum Temperature Knowledge Distillation (CTKD). This methodology leverages an attenuated temperature parameter to ensure a balanced transfer of knowledge from the teacher to the student. Higher temperatures lose effectiveness in later training phases and introduce a dynamic method to adjust the distillation temperature during training. This can be applied to any knowledge distillation framework that utilizes temperature.

3.1.2 Relational Knowledge Distillation

Park et al. (2019) proposed a method called Relational Knowledge Distillation (RKD), whereby the teacher provides knowledge to the student about the relationships within the input's feature embeddings. At each layer, this methodology calculates the distances and angles between the feature vectors of every pair of data points in the teacher network. Subsequently, the student network is trained to replicate these relationships within its own feature space, accomplished through the distance distillation loss and angle distillation loss (**Figure 2**).

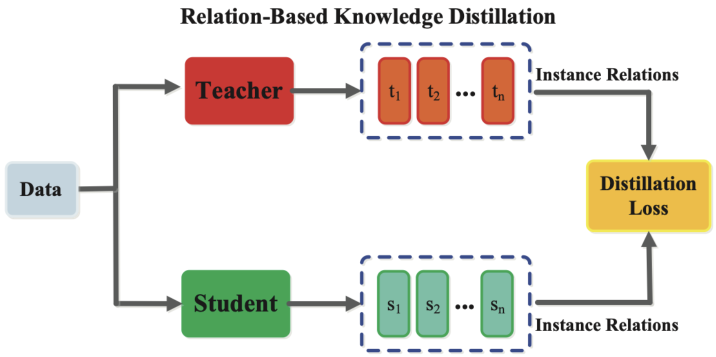

Figure 2: Model architecture of Relational Knowledge Distillation
Gou et al. (2020)

The key difference between RKD and CKD lies in the loss function, which is formalized as:

$$\mathcal{L}_{rkd} = \mathcal{L}_{ce} + \alpha(\cdot\beta \cdot \mathcal{L}_d + (1-\beta)\cdot \mathcal{L}_a)$$

The distance loss, L_d, encourages the student's feature representations to have similar L2 norms to those of the teacher. The angle loss, L_a, encourages the cosine similarity between the teacher and student feature representations to be close to 1. β allocates priority to the distance and angle losses. α attenuates or strengthens the overall distillation loss, which is the concatenation of the distance and angle loss, and can be tuned as a hyperparameter.

3.1.3 KD++ Distillation

Finally, Wang et al. (2023) proposed a method called "Knowledge Distillation via Regularizing Feature Norm and Direction," (KD++), whereby the student is trained to both produce large-norm features and align the direction of student features to teacher class means (**Figure 3**).

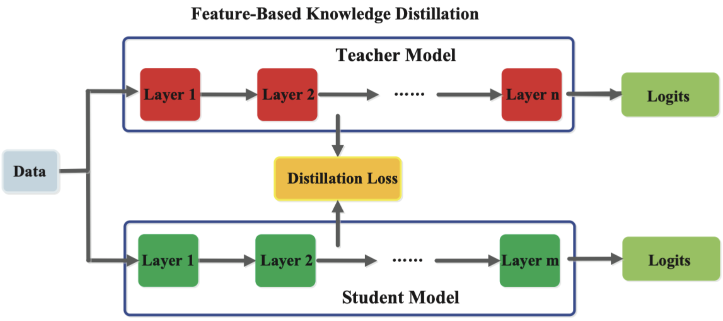

Figure 3: Model architecture of Feature-Based Knowledge Distillation
Gou et al. (2020)

Wang et al. (2023)'s key methodology also lies in the loss function formalized as:

$$\mathcal{L}_{kd++} = \mathcal{L}_{ce} + \alpha \cdot \mathcal{L}_{kd} + \beta \cdot \mathcal{L}_{nd}$$

The new loss, L_{nd}, defined by Wang et al. (2023) encourages larger student feature norms, and minimizes the angular distance between the student features and the teacher class mean.

3.2 Bias Inflation in Knowledge Distillation and Bias Mitigation

Through preliminary research and experimentation using the WIDER Attribute dataset, the issue of bias amplification between the teacher and the student model was identified. While the student can gain performance from the teacher's feature space or outputs, this often comes at the expense of learning stereotypes from the teacher's outputs and feature space Ahn et al. (2022). In 2018, Dong et al. (2023) proposed a framework for debiasing DNNs using an adversarial approach, whereby two models are deployed: a predictor and an adversary. The input to the predictor model produces a classification, which in the case of this research is an image label, while the adversary tries to model a variable, z, representing a protected attribute, in this case gender. The objective is to maximize the predictor's ability to predict y while minimizing the adversary's ability to predict z. To achieve this, Zhang et al. (2018), pioneered a loss function that links the classification model to the adversary, formalized as:

$$\mathcal{L}_{obj} = (\mathcal{L}_{kd} + \mathcal{L}_{ce}) - \lambda \cdot \mathcal{L}_{adv}$$

where λ is a weight that prioritizes debiasing by way of the adversary loss, $\mathcal{L}_{adv}$. $\mathcal{L}_{adv}$ is subtracted from the sum of the knowledge distillation and crossentropy losses within the objective loss function, $\mathcal{L}_{obj}$. The adversary loss is subtracted, rather than summed, to minimize the adversary's ability to predict the attributes. For a continuous attribute, the adversary loss is defined as:

$$\mathcal{L}_{adv} = MSE(\hat{z}, z)$$

In this paper, an adversarial fine-tuning approach is employed, where the loss is inspired from Savani et al. (2020) as

$$\mathcal{L}_{obj} = \max\{1, \lambda \cdot (|(g \circ f_0)(X_j)| - \epsilon + \delta) + 1\} \cdot (\mathcal{L}_{ce} + \mathcal{L}_{kd})$$

In this approach, called intra-processing, the model iterates between the adversary training and the student training, but updates the weights based on a loss function that incorporates the predicted bias (mean absolute value of disparity) as $(g \circ f_0)(X_j)$. This function takes the product of the student predictor loss, $\mathcal{L}_{ce} + \mathcal{L}_{kd}$, and 1 if the mean absolute disparity is less than $\epsilon - \delta$ where ϵ is the minimum bias required and δ is the margin for the loss objective. Otherwise, the loss function is the product of the predictor loss and predicted bias, $(g \circ f_0)(X_j)$, adjusted by ϵ and δ. Both research papers draw attention to the importance of hyperparameter tuning with adversarial debiasing, and emphasize that it is possible to increase model performance while employing the adversary to debias the predictor.

3.3 Datasets

To accomplish bias reduction, we deploy the four aforementioned bleeding edge knowledge distillation frameworks on the WIDER Attribute dataset, created by Li et al. (2016). This dataset consists of 13,789 images belonging to 30 scene categories, and 57,524 human bounding boxes, each annotated with 14 binary attributes (Figure 4). Existing research shows that this is a challenging dataset to predict, with CNN benchmark accuracy starting at 39.67% (Xiong et al. (2015)). However, this dataset is advantageous in that it contains person level attributes that will provide a ground-truth for the adversarial debiasing methodology.

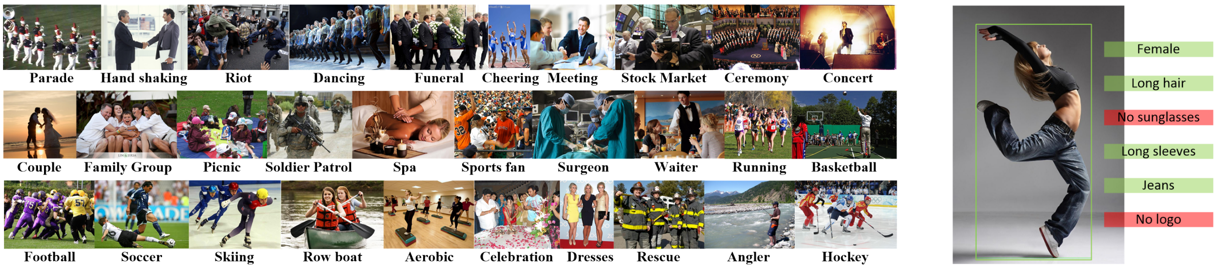

Figure 4: WIDER Attributes Dataset
Li et al. (2016)

Given there are similar classes in the WIDER Attribute dataset, the classes were clustered by the authors to 16 broader classes, joined on common themes (**Table 1**). The only protected characteristic in the WIDER Attribute dataset is gender, therefore all other attributes were ignored.

Several of the clustered classes are severely imbalanced between perceived male and perceived female (Figure 5). Dataset attribute imbalances can create model bias, which this research seeks to mitigate.

Additionally, we deployed the above models on CIFAR100 to provide benchmarks for incremental metrics: recall, precision, and F1-score.

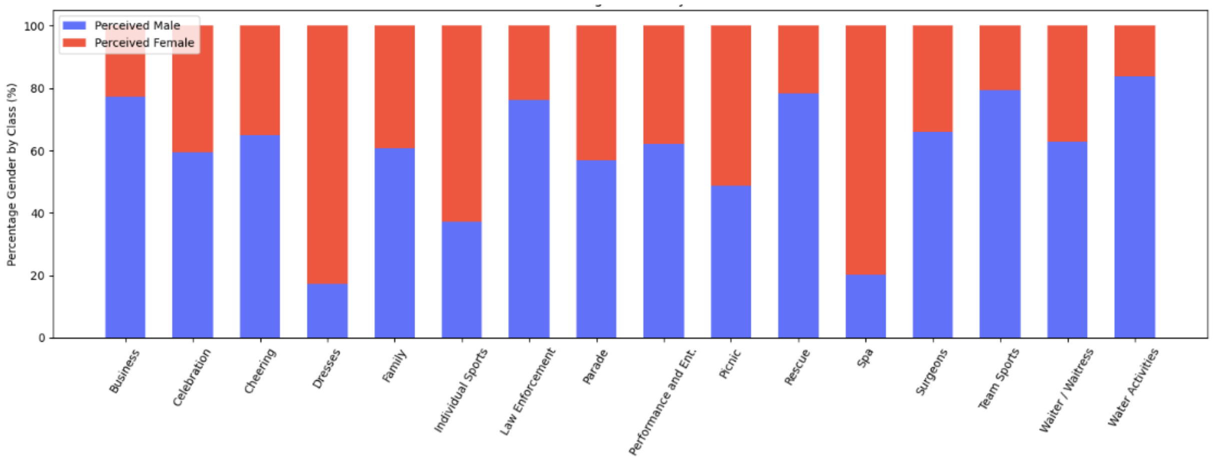

Figure 5: Gender Percent by Class

Clustered Class	Original Classes
Business	Handshaking, Stock Market, Meeting
Celebration	Ceremony, Funeral, Celebration or Party
Cheering	Sports Fan, Cheering
Dresses	Dresses
Family	Couple, Family
Individual Sports	Running, Aerobics, Ice Skating
Law Enforcement	Soldier Patrol, Riot
Parade	Parade
Performance and Entertainment	Concerts, Dancing
Picnic	Picnic
Rescue	Rescue
Spa	Spa
Surgeons	Surgeons
Team Sports	Hockey, Football, Basketball, Soccer
Waiter / Waitress	Waiter / Waitress
Water Activities	Row Boat, Angler

Table 1: Mapping clustered classes to the original WIDER Attribute classes

4 Experiments

A pipeline was developed to run and evaluate CKD, RKD, CTKD, and KD++ on CIFAR100 and WIDER using AWS EC2 Deep Learning AMI GPU PyTorch 2.1.0 (Ubuntu 20.04) 20231103 g4dn.2xlarge and P3.2xlarge to report accuracy, precision, recall, F1-score, and model size. Disparity was recorded for WIDER experiments only. The source code can be found within this repo.

4.1 New Metrics for Existing Models

First, experiments ran to validate existing work that Hinton et al. (2015), Li et al. (2022), Park et al. (2019), and Wang et al. (2023) accomplished for CKD, RKD, CTKD, and KD++, respectively. In doing so, the accuracy was validated against their published results and the models were benchmarked with incremental metrics.

4.1.1 CIFAR100 Experiment Settings

CIFAR100 consists of 60k images, which were split into the standard 50k and 10k train and test data loaders. The images were normalized to the CIFAR100 mean and standard deviations as per the CIFAR documentation. The experiments were initiated with the settings denoted in **Table 2**.

Model	Pair	LR	Optimizer	Momentum	Step Size	Weight Decay	Temp
CKD	ResNet34 ResNet18	0.0450	SGD	0.9	30	1e-4	4
RKD	ResNet34 ResNet18	0.0921	SGD	0.9	30	-	-
CTKD	ResNet34 ResNet18	0.0035	Adam	0.9	30	1e-4	20
KD++	ResNet56 ResNet20	0.0001	SGD	0.9	30	5e-4	4

Table 2: CIFAR100 Experiment Settings

Note that there is a divergence from the competition model frameworks in batch size due to resource limitations. Additionally, an optimal learning rate generator was developed and utilized for this research.

4.1.2 Evaluating the Performance of Competition Models on New Metrics

The published knowledge distillation models were run with a comprehensive set of metrics beyond accuracy, as reported in **Table 3**. The authors used these results for benchmarking purposes.

Model	Acc/Published	Recall	Precision	F1	Parameters (M)
Teacher					
CKD	73.11 / NA	73.11	73.77	73.02	6.67
RKD	79.91/ NA	79.91	80.03	79.84	6.06
CTKD	71.67 / 72.34	71.67	71.83	71.53	5.65
KD++	72.41 / 72.34	72.41	72.66	72.42	7.83
Student					
CKD	67.2 / NA	67.2	67.7	67.1	3.23
RKD	73.5 / NA	73.5	73.6	73.4	2.55
CTKD	61.35 / 69.06	61.35	62.02	11.23	3.02
KD++	52.00 / 72.53	52.00	60.45	27.83	2.57
Distill Loss (%)					
CKD	(8.07)	(8.07)	(8.27)	(8.08)	(51.57)
RKD	(8.01)	(8.01)	(8.70)	(8.02)	(57.92)
CTKD	(14.40)	(14.40)	(13.66)	(15.37)	(46.55)
KD++	(28.19)	(28.19)	(16.80)	(28.33)	(57.14)

Table 3: Results of Distillation Models - We compare accuracy to the published results from the corresponding publication. Distill Loss values are presented in parentheses as a percentage loss for the respective metric from the teacher to the student model.

4.2 Evaluating and Mitigating Bias Transfer Between Teacher and Student on WIDER

4.2.1 WIDER Bias Reduction Experiment Settings

The WIDER Attribute dataset, described in the literature review, was split into 10,324 train images, and 3,465 test images. Each image was normalized to the Imagenet mean and standard deviation. Additionally, the classes were condensed from 30 to 16 using manual clustering. As gender is the only attribute representing a protected characteristic, the remaining 13 attributes of each person in an image were discarded in favor of gender. Furthermore, as there can be up to 20 person-level bounding boxes with attributes in every image, a proxy for gender strength was created by taking the average of the gender attributes, z, across the image. $z >= 0.5$ was considered to have a higher distribution of perceived males, whereas $z < 0.5$ was considered to have a higher distribution of perceived females.

For model experimentation, an adversary model was deployed across the four aforementioned competition models (**Figure 6**).

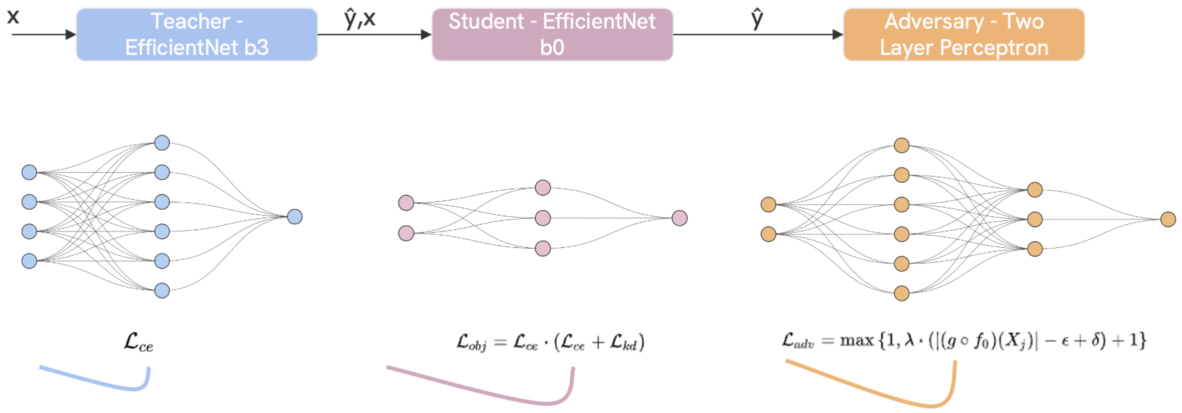
x
Teacher - EfficientNet b3
ŷ,x
Student - EfficientNet b0
ŷ
Adversary - Two Layer Perceptron
$\mathcal{L}_{ce}$
$\mathcal{L}_{obj} = \mathcal{L}_{ce} \cdot (\mathcal{L}_{ce} + \mathcal{L}_{kd})$
$\mathcal{L}_{adv} = \max\{1, \lambda \cdot (|(g \circ f_0)(X_j)| - \epsilon + \delta) + 1\}$

Figure 6: Adversarial Debiasing General Architecture

The adversary model was initialized in each knowledge distillation framework with an Adam optimizer and a learning rate of 0.0005. Due to the known challenge in predicting WIDER image labels, each knowledge distillation framework utilized EfficientNet_B3 with pre-trained weights for the teacher model and EfficientNet_B0 with pre-trained weights for the student model. A fine-tuning approach to knowledge distillation and adversarial debiasing was adopted due to the difficulty of achieving workable accuracy on WIDER with a model that has not been previously trained. The knowledge distillation initialization parameters are detailed in **Table 4**.

Model	**LR**	**Optimizer**	**Momentum**	**Temp**	**Epochs**	λ
CKD	0.0005	Adam	0.9	4	240 (with early stopping)	0, 50, 100, 150
RKD	0.0005	Adam	0.9	NA	240 (with early stopping)	0, 50, 100, 150
CTKD	0.0005	Adam	0.9	20	240 (with early stopping)	0, 50, 100, 150
KD++	0.0005	Adam	0.9	4	240 (with early stopping)	0, 50, 100, 150

Table 4: WIDER Bias Reduction Experiment Settings

5 Results and Discussion

5.1 Discussion of Adversarial Debiasing Results

Across all knowledge distillation techniques investigated, the student model experienced inflated bias from the teacher model to the student model with $\lambda = 0$ (**table 5**). CKD and RKD frameworks both increased accuracy and bias as the student learned from the teacher.

To identify an optimal debiasing strength, λ, a grid search was performed on λ values, with λ ranging from 0 to 150 (**Figure 7**). The results illustrate a coarsely defined inverse relationship between accuracy and mean absolute disparity (**Table 5**). This relationship is apparent asymptotically, as λ approaches infinity. However, all models experience minimal or no correlated tradeoff between accuracy and debiasing strength at lower levels of λ.

5.2 Results of λ Tuning Experiments

The RKD framework, $\lambda = 1$ resulted in an increase in accuracy over the teacher and the student with $\lambda = 0$, while simultaneously achieving a lower disparity of 9.51. For $\lambda = 1$, CKD, CTKD, and KD++ experienced a decrease in

Model	Lambda	Accuracy (%)	Recall (%)	Precision (%)	F1	Parameters (M)	Disparity
Teacher							
CKD	0	64.85	64.85	65.95	64.63	10.72	9.10
RKD	0	65.11	65.11	66.33	65.25	10.72	8.40
CTKD	0	66.58	66.58	65.98	65.86	10.72	10.61
KD++	0	62.14	62.14	61.72	60.62	10.72	8.12
Student							
CKD	0	**64.96**	**64.96**	**65.02**	**64.65**	4.03	9.77
	1	62.77	62.77	62.48	62.23	4.03	10.52
	5	54.57	54.57	54.13	51.18	4.03	**7.81**
	10	51.60	51.60	49.78	48.58	4.03	11.41
	15	47.33	47.33	41.51	42.86	4.03	10.46
	20	37.23	37.23	33.01	32.72	4.03	8.69
RKD	0	65.63	65.63	65.65	65.20	4.03	12.26
	1	**65.85**	**65.85**	**66.75**	**65.98**	4.03	9.51
	5	43.52	43.52	54.18	39.72	4.03	**5.40**
	10	46.44	46.44	60.13	43.22	4.03	6.70
	15	45.83	32.21	44.93	32.00	4.03	9.26
	20	45.19	45.19	48.68	42.46	4.03	9.21
CTKD	0	**63.72**	**63.72**	**64.15**	**63.44**	4.03	12.33
	1	62.71	62.71	63.28	62.51	4.03	9.58
	5	56.71	56.71	56.49	55.27	4.03	13.94
	10	51.72	51.72	51.27	49.06	4.03	**8.94**
	15	48.66	48.66	50.90	47.37	4.03	10.54
	20	40.81	40.81	36.19	35.77	4.03	10.31
KD++	0	**59.62**	**59.62**	**60.64**	**59.60**	4.03	10.95
	1	56.88	46.26	53.07	43.83	4.03	9.37
	5	38.38	38.38	40.53	33.48	4.03	7.74
	10	34.89	34.89	33.95	29.61	4.03	9.21
	15	27.45	27.45	25.38	19.41	4.03	10.18
	20	20.95	20.95	16.64	13.40	4.03	**2.51**

Table 5: Knowledge Distillation Model Performance on WIDER, and Bias Reduction - accuracy, recall, precision, F1-score, and the number of parameters are reported for each knowledge distillation framework. Additionally, the authors provide a baseline for disparity for the biased teacher and student models, to compare to the results of the adversarial debiasing approach across four additional λ values. Results are reported in order of knowledge distillation framework and increasing strength of the adversary as measured by λ. Best numbers for each model framework are in bold.

accuracy of 2.19, 1.01, and 2.74 percentage points respectively, versus their $\lambda = 0$ experiment. CKD experienced an increase in disparity of 0.75 percentage points, while CTKD and KD++ experienced a drop in disparity by 2.75 and 1.58 percentage points, respectively (**Table 5**).

$1 < \lambda < 20$ observed a more significant drop in accuracy for all models as the adversary loss became weighted significantly more heavily than the crossentropy loss and the knowledge distillation losses (Figure 8). However, there was not necessarily a corresponding linear decrease in disparity for all models. RKD, CTKD, CKD, and KD++ achieved minimum disparity at λ values of 5, 10, 15, and 20, respectively over the evaluated range from 0 to 20 (Figure 9).

Debiasing results at the class level show that the adversarial approach is successful in reducing aggregate disparity, even at lower λ values, but is not granular enough to target specific class disparity. **Table 6** shows the class-level results of incrementing from $\lambda = 0$ to $\lambda = 0.5$ on the RKD pipeline. For $\lambda = 0.5$, all classes experienced a decrease in disparity due to debiasing with the exception of Team Sports, Rescue, and Family. Each of these classes increased in disparity as the adversary model worked to reduce overall bias, not class level bias.

In terms of disparity reduction, CTKD showed the poorest performance amongst the knowledge distillation techniques

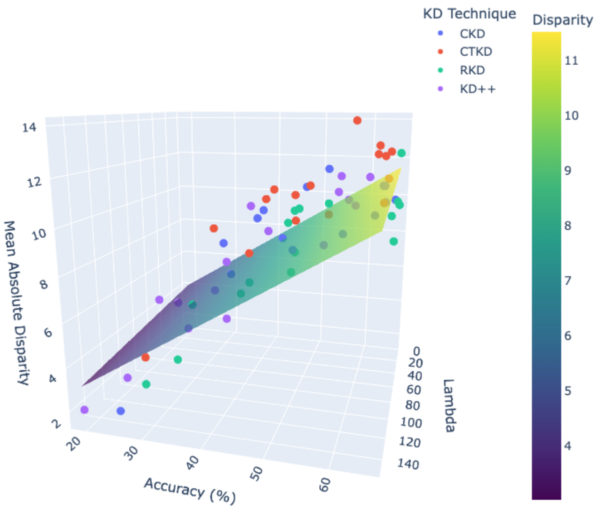

Figure 7: Three dimensional projection of results with planar best fit for disparity, accuracy, and λ

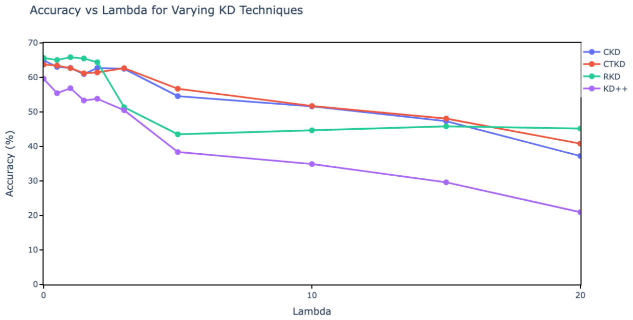

Figure 8: Accuracy vs. λ

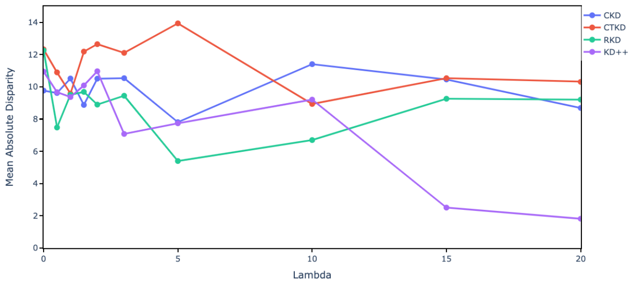

Figure 9: Mean Abosolute Disparity vs. λ

Clustered Class	**Disparity** $\lambda = 0$	**Disparity** $\lambda = 0.5$	**Disparity** $\lambda = 150$
Team Sports	4.71	8.38	5.06
Celebration	18.07	0.47	1.74
Parade	-3.01	0.07	-3.51
Individual Sports	-10.56	-5.91	-1.81
Surgeons	-15.26	-1.26	0.00
Spa	15.48	9.13	0.00
Law Enforcement	3.71	2.21	3.85
Business	29.60	16.82	0.65
Dresses	-38.87	-30.69	-23.07
Water Activities	8.98	1.18	-4.27
Picnic	6.56	-2.94	0.00
Rescue	6.27	26.41	7.92
Cheering	-7.17	-4.06	0.507
Performance and Entertainment	22.15	2.28	14.32
Family	4.46	7.02	3.05
Waiter/Waitress	-1.37	-0.82	1.45
Average	12.26	7.91	5.48

Table 6: Class level results of adversarial debiasing show that this approach reduces aggregate bias in the model, but is not granular enough to attack bias in specific classes. Disparity is defined here as the difference in recall between when the gender attribute is perceived male and when the gender attribute is perceived female. Therefore, a positive disparity value indicates higher recall in the presence of the perceived male attribute whereas a negative value indicates higher recall in the presence of the perceived female attribute

assessed. KD++ was sensitive to increased λ values compared to the other models. At higher λ values, KD++ achieved the lowest disparity values observed across all techniques, reaching as low as 2.51. Overall, RKD proved most resilient to adversarial debiasing, and saw a significant reduction in disparity at low λ values with marginal impact to accuracy.

Other traditional performance metrics including recall, precision, and F1-score generally followed a similar trend to that of accuracy as debiasing was prioritized.

The knowledge distillation models used proved to be successful in debiasing the models over an increasing horizon of adversary strength. The authors have demonstrated a relationship between adversarial strength, accuracy, and bias as measured by disparity and have shown that for certain λ levels it is possible to achieve higher accuracy and lower disparity. However, the results of this study demonstrate that adversarial debiasing is sensitive to many model architectures and hyperparameters. Therefore, care must be taken in fine-tuning the adversarial approach to debiasing.

5.3 Challenges and Future Work

While the WIDER Attribute dataset met the requirements of being human-centric and having a protected characteristic label, the labels and image quality were determined to be poor. This potentially amplified the known sensitivity of the adversarial approach to model initialization and tuning. The authors will seek to identify or develop higher quality, better labeled, and more diverse datasets that will allow for furthering this research in computer vision and new domains.

Additionally, the computational requirements for this research were substantial, preventing the batch size from being increased beyond 64. A larger batch size will provide for more consistent disparity measurements and result in increased training speed and performance. This research provides the framework to develop a frontier of optimal λ values for maximizing performance while minimizing bias given the availability of financial and computational resources.

The adversarial debiasing technique that the authors employed attacks aggregate bias across all classes; however, future work will be done to target bias for specific classes in the loss function. As certain model predictions carry more significance in the real world, adversarial models should consider weighting class-level debiasing toward more impactful model predictions.

Lastly, the researched and discussed methodology relies on a labeled, attributed dataset which poses significant challenges in deploying this strategy across unlabeled Deep Learning problems. The researchers are interested in and actively exploring leveraging this debiasing technique for unlabeled, unattributed datasets.

6 Conclusion

Knowledge distillation is an effective means of neural network compression. Although the process at its core may lead to exacerbated bias, the debiasing framework outlined in this paper demonstrates a method for alleviating this bias with only a marginal impact on accuracy and other traditional performance metrics. Though asymptotically there is an inverse relationship between disparity and accuracy, it is possible to find a λ value that optimizes both predictive performance and disparity reduction.

Within moderate levels of adversary prioritization, it is likely that hyperparameter experimentation will find an optimal λ that both increases predictive performance and decreases bias. This finding underscores the potential of adversarial debiasing to enhance the performance of the student model. This result aligns with previous research on debiasing neural networks and holds true for the knowledge distillation framework.

This paper introduces a new technique of knowledge distillation that incorporates adversarial debiasing that allows for the minimization of bias inflation between the student and teacher model, in aggregate. As models continue to grow, further compression will be required and the bias inflation challenge will become more important to manage. In a world where the models are approaching super-human capabilities, neural nets will have outsized impacts on the trajectory of human lives. Whether in self-driving cars, surveillance systems, or object detection tasks, it will continue to be essential to mitigate bias while continuing to work on improving the performance of compressed models.

References

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272, Seattle, Washington. Association for Computational Linguistics.

Yushun Dong, Binchi Zhang, Yiling Yuan, Na Zou, Qi Wang, and Jundong Li. 2023. Reliant: Fair knowledge distillation for graph neural networks.

Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. 2020. Knowledge distillation: A survey. *CoRR*, abs/2006.05525.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Yezhou Li, Chunhua Huang, Chen Change Loy, and Xiaoou Tang. 2016. Human attribute recognition by deep hierarchical contexts. In *European Conference on Computer Vision*, pages 684–700. Springer, Cham.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. 2022. Curriculum temperature for knowledge distillation.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. *CoRR*, abs/1904.05068.

Yash Savani, Colin White, and Naveen Sundar Govindarajulu. 2020. Post-hoc methods for debiasing neural networks. *CoRR*, abs/2006.08564.

Yuzhu Wang, Lechao Cheng, Manni Duan, Yongheng Wang, Zunlei Feng, and Shu Kong. 2023. Improving knowledge distillation via regularizing feature norm and direction.

Yuanjun Xiong, Kai Zhu, Dahua Lin, and Xiaoou Tang. 2015. Recognize complex events from static images by fusing deep channels. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593.

7 Appendices

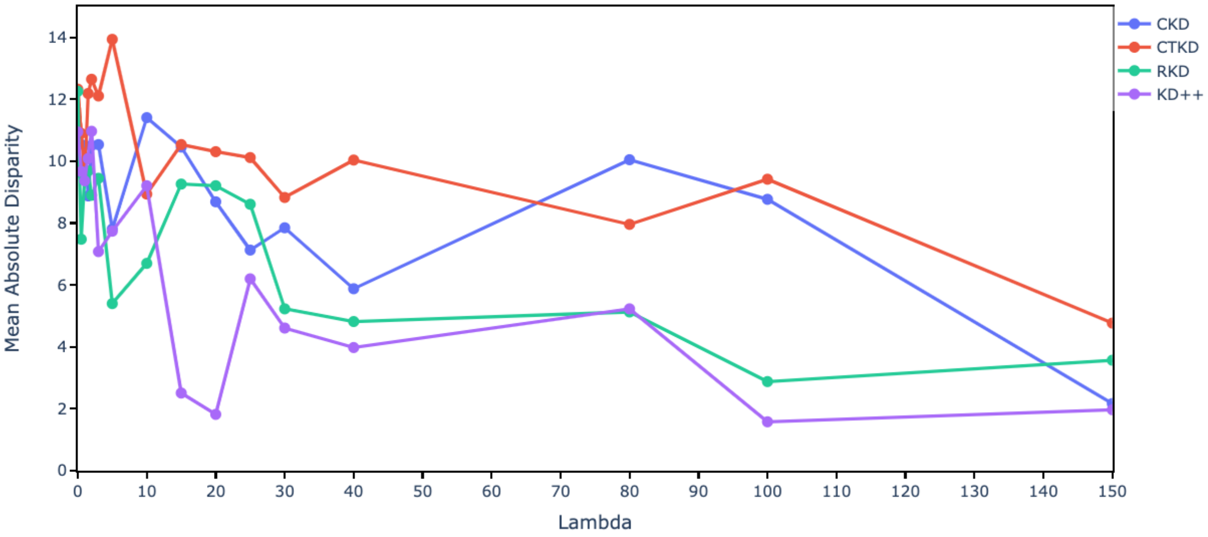

Figure 10: Mean Absolute Disparity Across the Full Range of Lambda Values Investigated

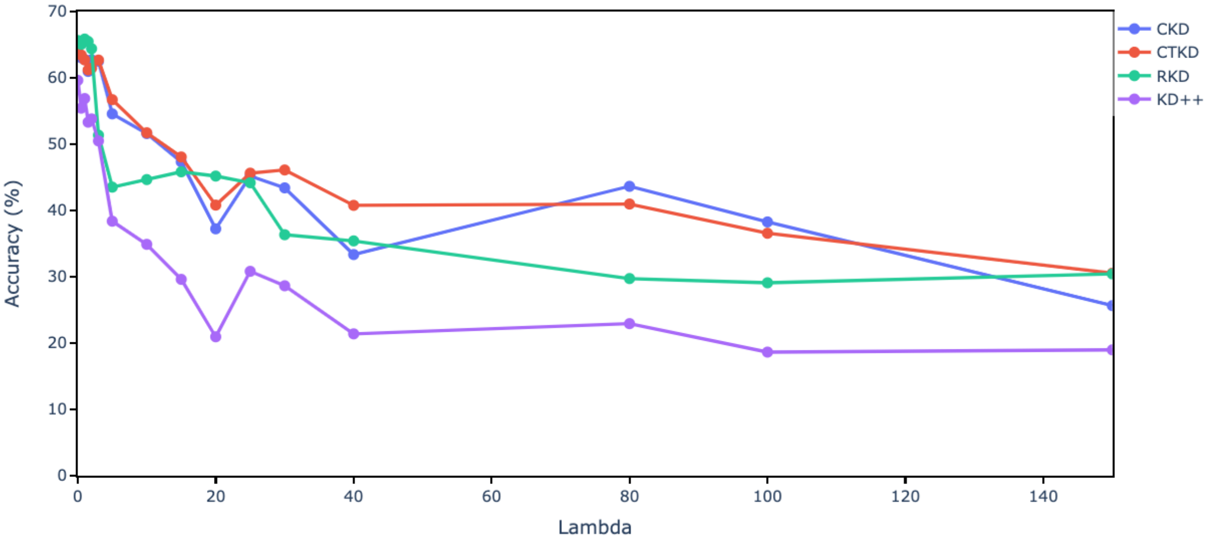

Figure 11: Accuracy Across the Full Range of Lambda Values Investigated