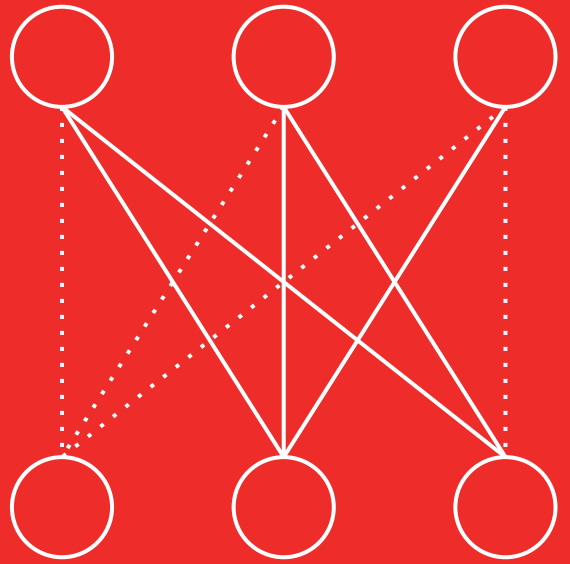
What's Left After Distillation?

How Knowledge Transfer Impacts Fairness and Bias



Yani Ioannou
Schulich Research Chair / Assistant Professor
Dept. of Electrical & Software Engineering,
Schulich School of Engineering

University of Windsor, Computer Science

March 14th, 2025

UNIVERSITY OF CALGARY

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

Aida Mohammadshahi *aida.mohammadshahi@ucalgary.ca*

Yani Ioannou *yani.ioannou@ucalgary.ca*
Department of Electrical and Software Engineering
Schulich School of Engineering, University of Calgary
Calgary, AB, Canada

Reviewed on OpenReview: *https://openreview.net/forum?id=xBbj46Y2fN*

Abstract

Knowledge Distillation is a commonly used Deep Neural Network (DNN) compression method, which often maintains overall generalization performance. However, we show that even for balanced image classification datasets, such as CIFAR-100, Tiny ImageNet and ImageNet, as many as 41% of the classes are statistically significantly affected by distillation when comparing class-wise accuracy (i.e. class bias) between a teacher/distilled student or distilled student/non-distilled student model. Changes in class bias are not necessarily an undesirable outcome when considered outside of the context of a model's usage. Using two common fairness metrics, Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD) on models trained with the CelebA, Trifeature, and HateXplain datasets, our results suggest that increasing the distillation temperature improves the distilled student model's fairness, and the distilled student fairness can even surpass the fairness of the teacher model at high temperatures. Additionally, we examine individual fairness, ensuring similar instances receive similar predictions. Our results confirm that higher temperatures also improve the distilled student model's individual fairness. This study highlights the uneven effects of distillation on certain classes and its potentially significant role in fairness, emphasizing that caution is warranted when using distilled models for sensitive application domains.

1 Introduction

DNNs require significant computational resources, resulting in large overheads in compute, memory, and energy. Decreasing this computational overhead is necessary for many real-world applications where these costs would otherwise be prohibitive, or even make their application infeasible — e.g. the deployment of DNNs on mobile phones or edge devices with limited resources (Chen et al., 2016; Cheng et al., 2018; Gupta and Agrawal, 2022; Menghani, 2023). To address this challenge, DNN model compression methods have been developed that reduce the size and complexity of DNNs while maintaining their generalization performance (Cheng et al., 2017). One such widely used model compression method is Knowledge Distillation (distillation) (Hinton et al., 2015). Distillation has found extensive application in both industry and academia across various domains of artificial intelligence, encompassing areas such as Natural Language Processing (NLP) (Jiao et al., 2019; Fu et al., 2021; Liu et al., 2020), speech recognition (Ng et al., 2018; Gao et al., 2019; Perez et al., 2020), and visual recognition (Yan et al., 2019; Dou et al., 2020; Chawla et al., 2021), specifically image classification (Zhu et al., 2019; Chen et al., 2019; Gou et al., 2021).

Distillation involves transferring knowledge from a complex model with superior performance (referred to as the *teacher*) to a simpler model (known as the *student*). In practice this allows the student model to achieve comparable or even better generalization than the teacher model, while using far fewer parameters (Hinton et al., 2015; Gou et al., 2021). Despite the widespread use of distillation, evaluation of the impact of distillation since its proposal by (Hinton et al., 2015) has overwhelmingly focused almost exclusively on the impact it has on generalization performance (Cho and Hariharan, 2019; Mirzadeh et al., 2020).

Aida Mohammadshahi
MSc (Defended Jan 2025)

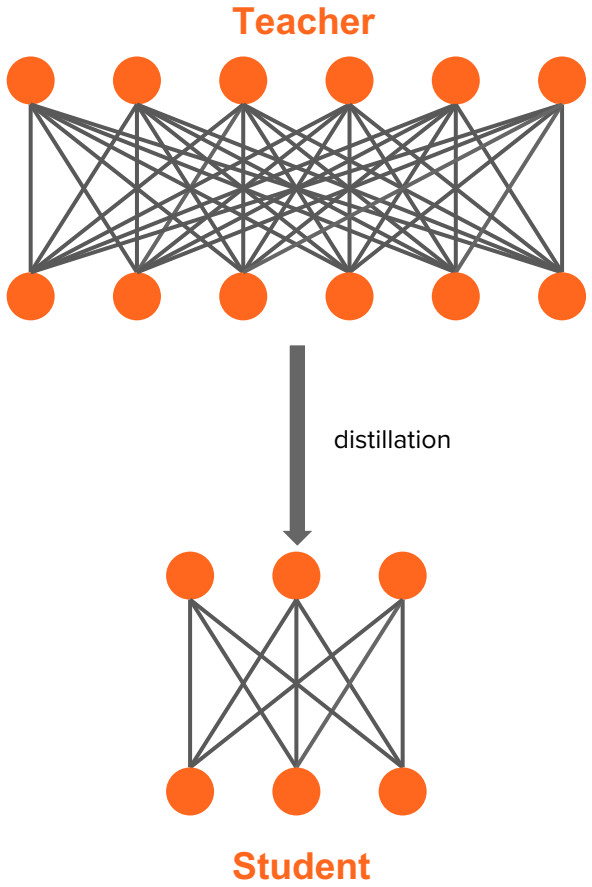
- Presented at NeurIPS WiML Workshop in Dec. 2024
- Recently accepted at TMLR!

UNIVERSITY OF CALGARY

Knowledge Distillation

What is Knowledge Distillation?

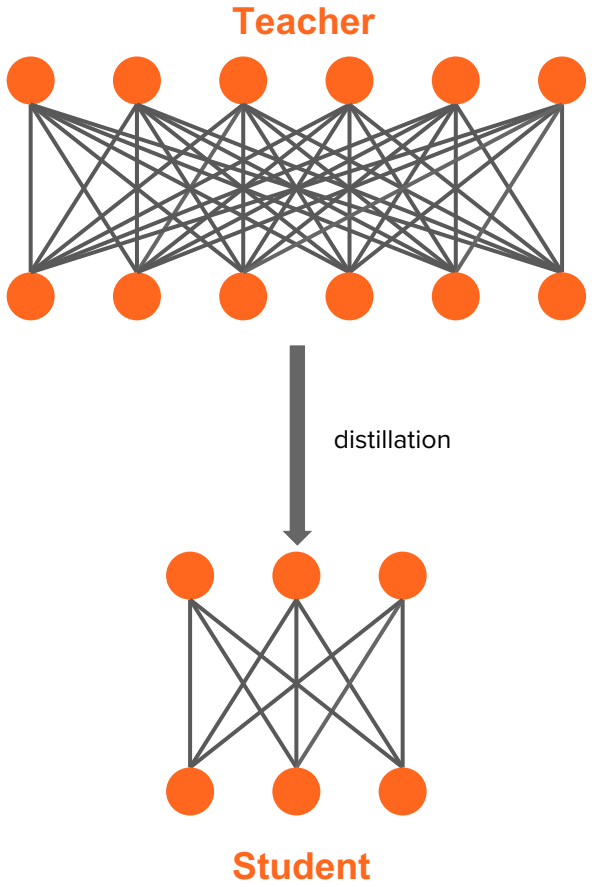
- A method of transferring “knowledge” from a larger model (or models) to a smaller model
- e.g. ensemble of models ➔ single model
- Preserves generalization (test accuracy)
- Commonly used to compress large models
  - **Large model ➔ small model (Student)**



Distilling the knowledge in a neural network. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. NeurIPS Deep Learning Workshop 2014.

What is Knowledge Distillation?

- Commonly used to compress large models
  - Large model ➔ small model (Student)
- Used extensively in industry to make models smaller for applications
  - **Smaller models = cheaper compute costs**
  - **Smaller models enable mobile applications**



Distilling the knowledge in a neural network. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. NeurIPS Deep Learning Workshop 2014.

What is Knowledge Distillation?

- **DeepSeek R1 (671B MoE Model)**
  - **Distilled smaller (1.5 - 70B) models, e.g. Llama**
  - **These smaller models are the models easier to use in practice**

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

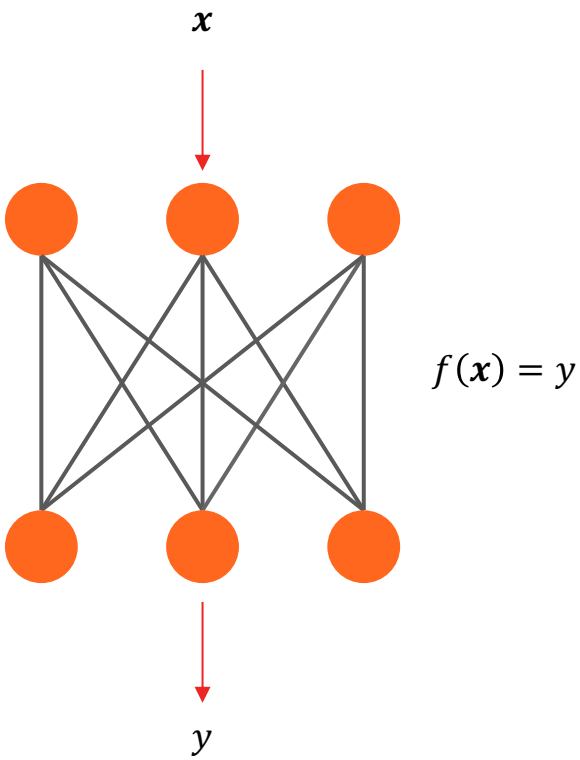
Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

DeepSeek-V3 Technical Report, DeepSeek-AI, arXiv 2412.19437, 2024

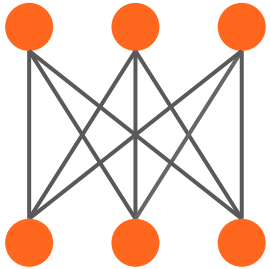
Knowledge Distillation Methodology

Neural Networks as Functions

- Neural Networks are function approximators
- **A neural network learns a function f mapping an input $\boldsymbol{x}$ to an output y**
- In practice, NNs for classification learn to predict a **probability distribution** p, from which the "hard" classification of a class y is made



“Dark Knowledge”



$$f(\boldsymbol{x}) = \{0.4, 0.5, 0.1\}$$

- Trained models learn **more** than just how to predict labels
- They learn a function with **rich knowledge of the domain**
- An ImageNet model knows that a **cat** and **dog** are more similar to each other than an **airplane**

CMU UNIVERSITY OF CALGARY

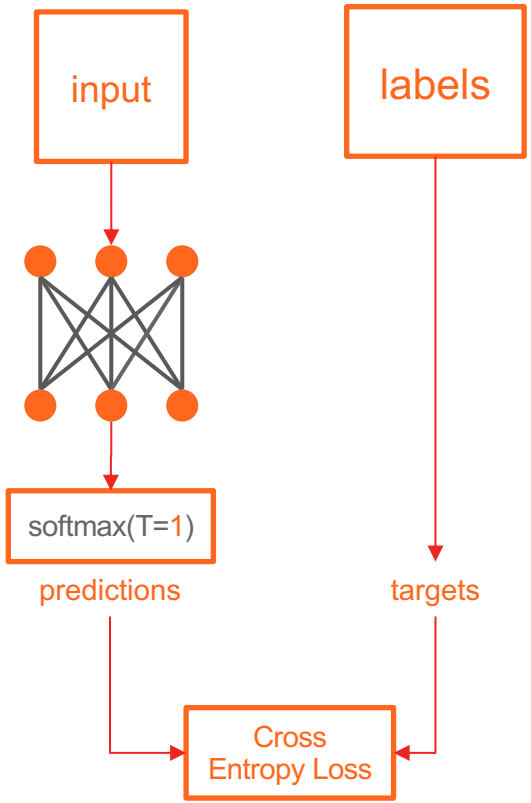
Temperature Softmax

$$p_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}$$

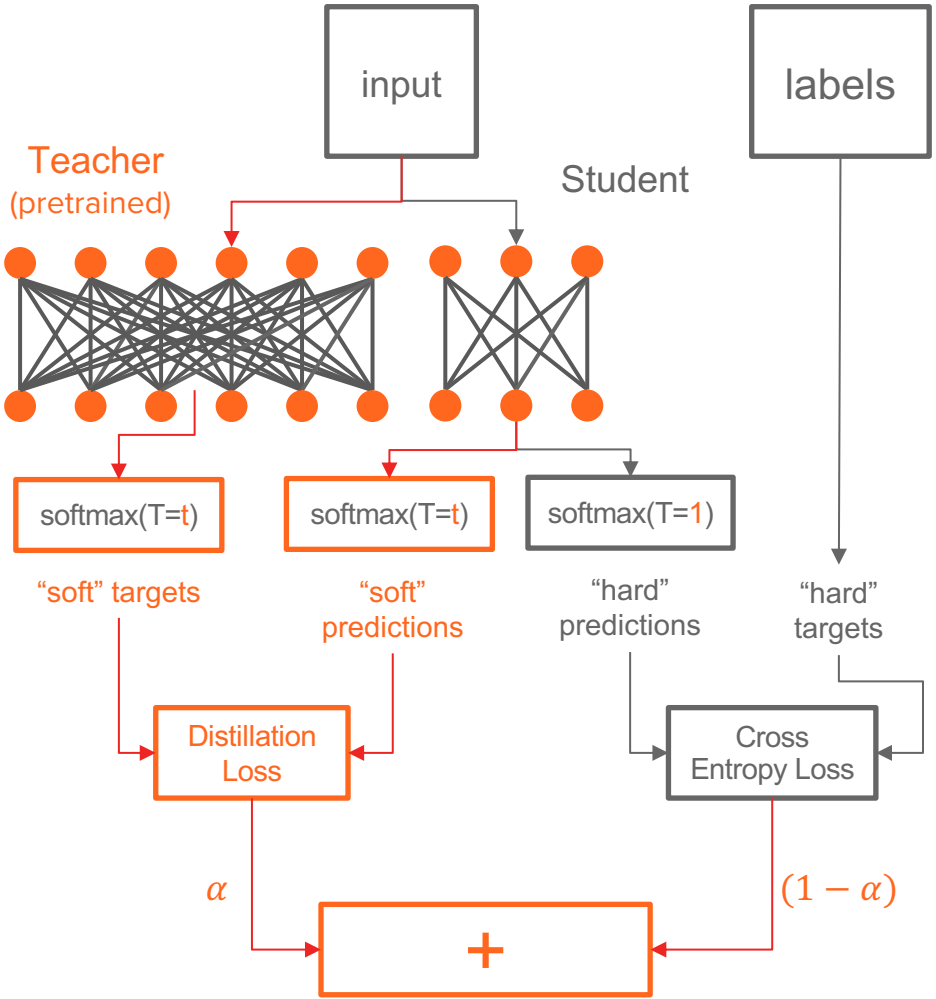
$$f(\boldsymbol{x}, T = 1) = \{0.09, 0.9, 0.01\}$$

$$f(\boldsymbol{x}, T = 10) = \{0.4, 0.5, 0.1\}$$

- A softmax p(z) gives us a probability output from logits z
- Distillation adds “temperature” T to softmax
- The typical softmax (T=1) gives very highly confident outputs for the target class, i.e. a “hard distribution
- Larger temp T gives “softer” distributions



- Standard training of neural network for classification
- Use cross-entropy loss with input and target labels
- Uses a softmax with T=1

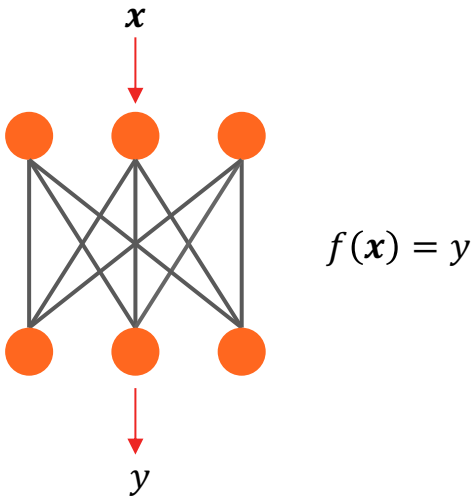


- Knowledge Distillation
- Uses both CE loss with “hard” targets & distillation loss with “soft” targets from teacher
- Distillation loss is KL div between student/teacher’s soft predictions
- These two losses are weighted by a single hyperparameter α

Knowledge Distillation and Fairness

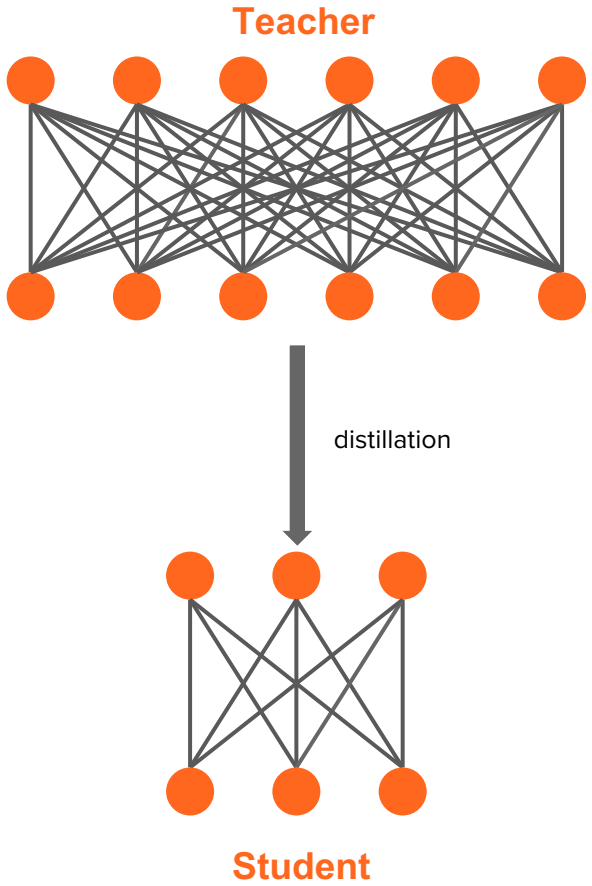
Recall: Neural Networks as Functions

- NNs are function approximators
- **A neural network learns a function f mapping an input x to an output y**



What does KD Learn?

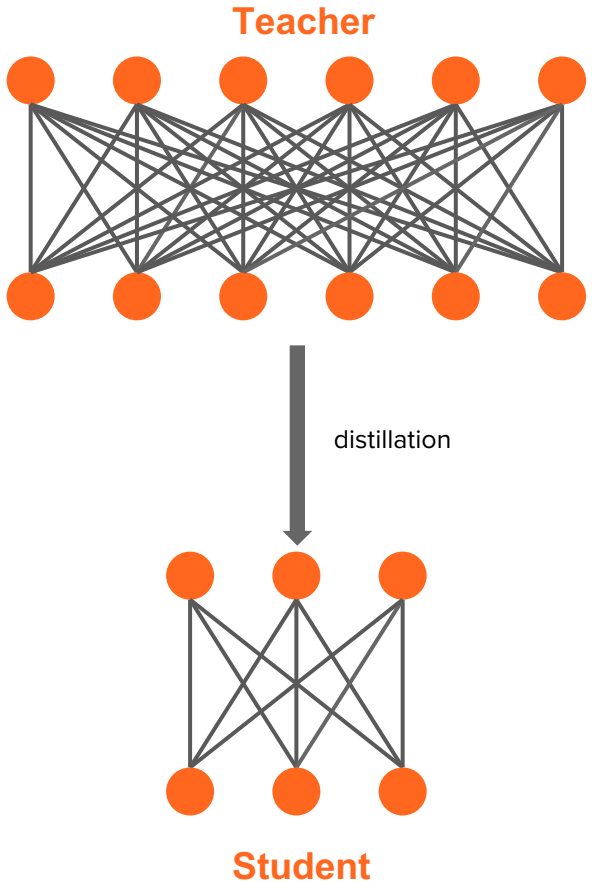
- When we distill a large teacher model to a small student, we often see **generalization performance (test accuracy) maintained**
- Does this mean that the **Teacher** and **Student** have learned **similar functions?**



Distilling the knowledge in a neural network. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. NeurIPS Deep Learning Workshop 2014.

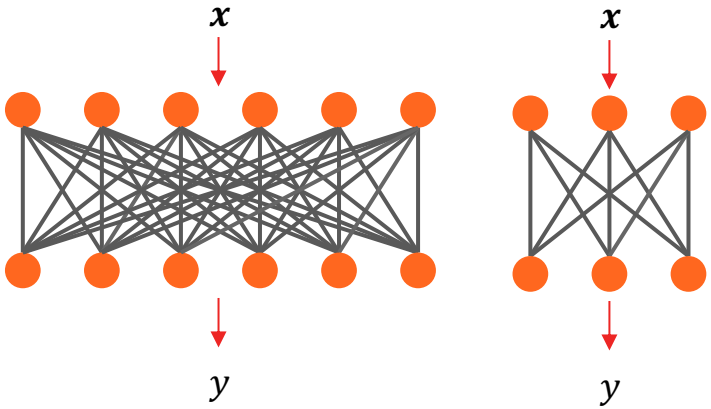
What does KD Learn?

- When we distill a large teacher model to a small student, we often see **generalization performance maintained**
- Does this mean that the **Teacher** and **Student** have learned **similar functions?**
  - **Not necessarily:** accuracy is aggregate measure over many samples in test set



What does KD Learn?

- When we distill a large teacher model to a small student, we often see generalization performance maintained
- **However, student can learn different function than teacher**
- **Why does this matter?**



Teacher

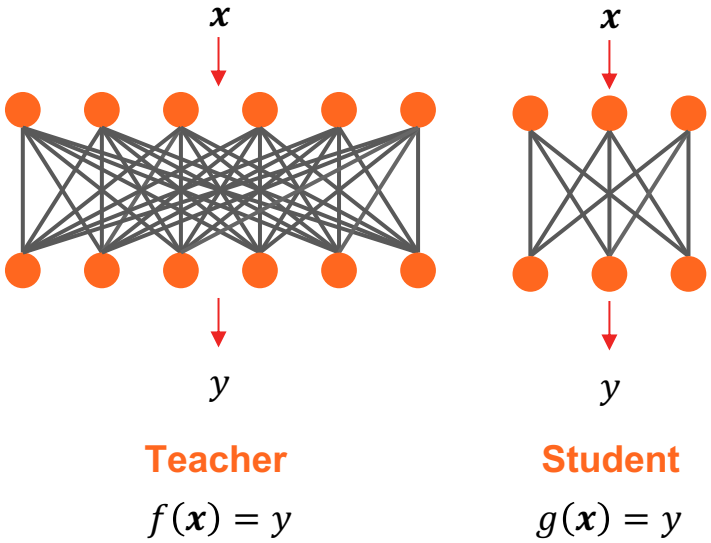
$f(\boldsymbol{x}) = y$

Student

$g(\boldsymbol{x}) = y$

What does KD Learn?

- When we distill a large teacher model to a small student, we often see generalization performance maintained
- However, student can learn different function than teacher
- Why does this matter?
- **Student may learn different algorithmic biases than Teacher!**



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

Aida Mohammadshahi *aida.mohammadshahi@ucalgary.ca*

Yani Ioannou *yani.ioannou@ucalgary.ca*
Department of Electrical and Software Engineering
Schulich School of Engineering, University of Calgary
Calgary, AB, Canada

Reviewed on OpenReview: *https://openreview.net/forum?id=xBbj46Y2fN*

Abstract

Knowledge Distillation is a commonly used Deep Neural Network (DNN) compression method, which often maintains overall generalization performance. However, we show that even for balanced image classification datasets, such as CIFAR-100, Tiny ImageNet and ImageNet, as many as 41% of the classes are statistically significantly affected by distillation when comparing class-wise accuracy (i.e. class bias) between a teacher/distilled student or distilled student/non-distilled student model. Changes in class bias are not necessarily an undesirable outcome when considered outside of the context of a model's usage. Using two common fairness metrics, Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD) on models trained with the CelebA, Trifeature, and HateXplain datasets, our results suggest that increasing the distillation temperature improves the distilled student model's fairness, and the distilled student fairness can even surpass the fairness of the teacher model at high temperatures. Additionally, we examine individual fairness, ensuring similar instances receive similar predictions. Our results confirm that higher temperatures also improve the distilled student model's individual fairness. This study highlights the uneven effects of distillation on certain classes and its potentially significant role in fairness, emphasizing that caution is warranted when using distilled models for sensitive application domains.

1 Introduction

DNNs require significant computational resources, resulting in large overheads in compute, memory, and energy. Decreasing this computational overhead is necessary for many real-world applications where these costs would otherwise be prohibitive, or even make their application infeasible — e.g. the deployment of DNNs on mobile phones or edge devices with limited resources (Chen et al., 2016; Cheng et al., 2018; Gupta and Agrawal, 2022; Menghani, 2023). To address this challenge, DNN model compression methods have been developed that reduce the size and complexity of DNNs while maintaining their generalization performance (Cheng et al., 2017). One such widely used model compression method is Knowledge Distillation (distillation) (Hinton et al., 2015). Distillation has found extensive application in both industry and academia across various domains of artificial intelligence, encompassing areas such as Natural Language Processing (NLP) (Jiao et al., 2019; Fu et al., 2021; Liu et al., 2020), speech recognition (Ng et al., 2018; Gao et al., 2019; Perez et al., 2020), and visual recognition (Yan et al., 2019; Dou et al., 2020; Chawla et al., 2021), specifically image classification (Zhu et al., 2019; Chen et al., 2019; Gou et al., 2021).

Distillation involves transferring knowledge from a complex model with superior performance (referred to as the *teacher*) to a simpler model (known as the *student*). In practice this allows the student model to achieve comparable or even better generalization than the teacher model, while using far fewer parameters (Hinton et al., 2015; Gou et al., 2021). Despite the widespread use of distillation, evaluation of the impact of distillation since its proposal by (Hinton et al., 2015) has overwhelmingly focused almost exclusively on the impact it has on generalization performance (Cho and Hariharan, 2019; Mirzadeh et al., 2020).

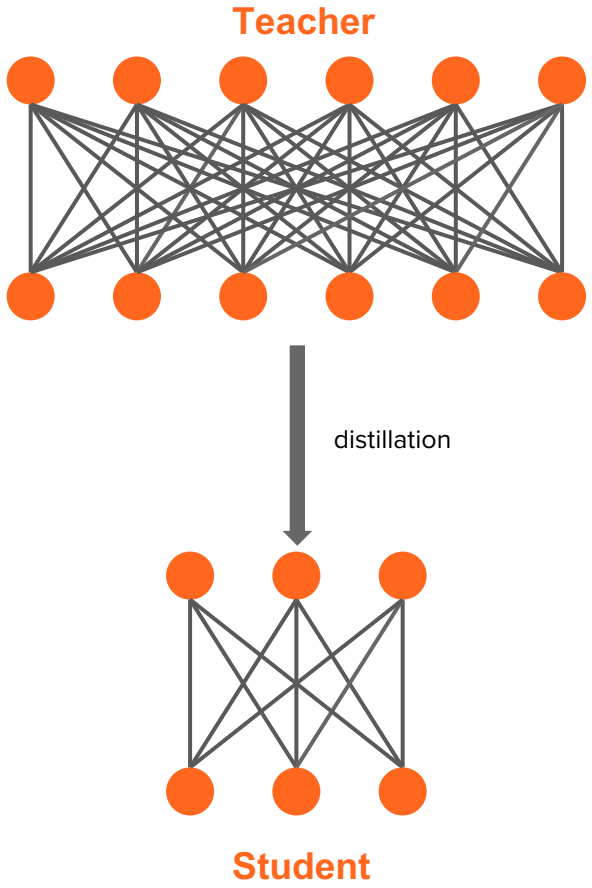
Aida Mohammadshahi
MSc (Defended Jan 2025)

- Presented at NeurIPS WiML Workshop in Dec. 2024
- Recently accepted at TMLR!

UNIVERSITY OF CALGARY

Research Questions

- **Q: What classes are significantly affected by distillation?**
- **Q: What is the impact of increase temperature T on the model's class biases?**
- **Q: How does distillation temperature affect group fairness?**
- **Q: How does distillation temperature affect individual fairness?**



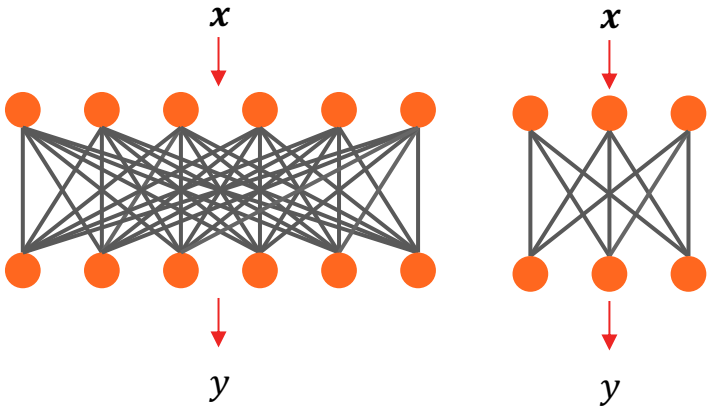
Class-wise Bias: Analysis

- Q: What classes are significantly affected by distillation?
- **Disagreement** of the models f, g on predictions for $\boldsymbol{x}_n$:

$$CMP\big(f(\boldsymbol{x}_n), g(\boldsymbol{x}_n)\big) = \begin{cases} 0 & \text{if } f(\boldsymbol{x}_n) = g(\boldsymbol{x}_n) \\ 1 & \text{if } f(\boldsymbol{x}_n) \neq g(\boldsymbol{x}_n) \end{cases}$$

- Compare the **teacher** f and distilled **student** g model's disagreement for each class **c**:

$$CMP\big(f(\boldsymbol{x}_n), g(\boldsymbol{x}_n)\big) \text{ where } (\boldsymbol{x}_n, y_n \mid y_n = c)$$



Teacher

$f(\boldsymbol{x}) = y$

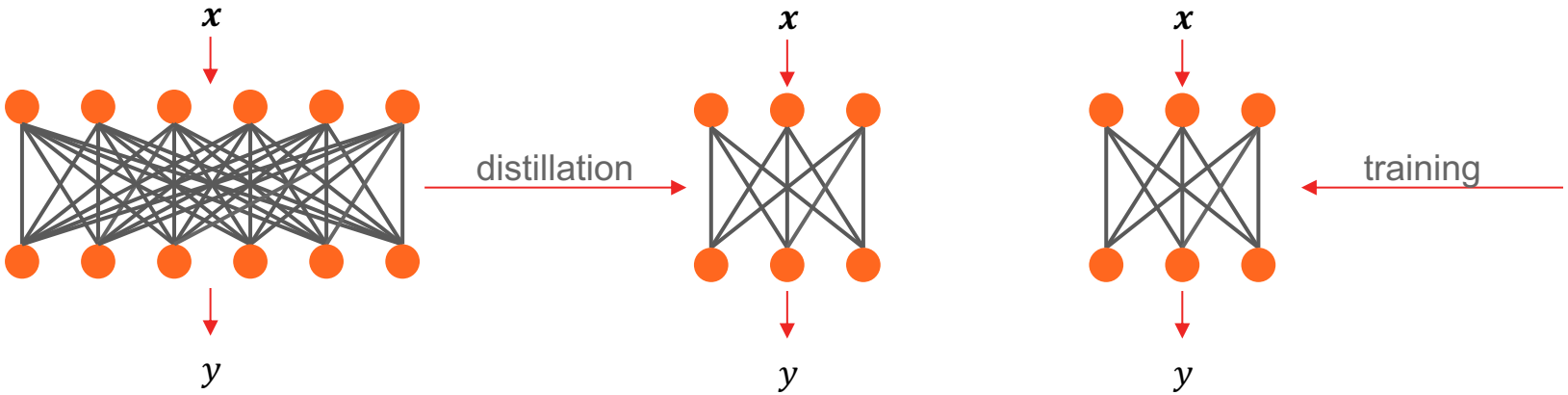
Student

$g(\boldsymbol{x}) = y$

Deep Ensembles: A Loss Landscape Perspective. Stanislav Fort, Huiyi Hu, Balaji Lakshminarayanan. arXiv:1912.02757

Class-wise Bias: Analysis

- Compare the **teacher** f and **distilled student** g model's disagreement for each class **c**:
- We use a **non-distilled student** h (trained from scratch) as a **baseline**



Teacher

$f(\boldsymbol{x}) = y$

Distilled Student

$g(\boldsymbol{x}) = y$

Non-Distilled Student

$h(\boldsymbol{x}) = y$

Class-wise Bias: Models/Datasets

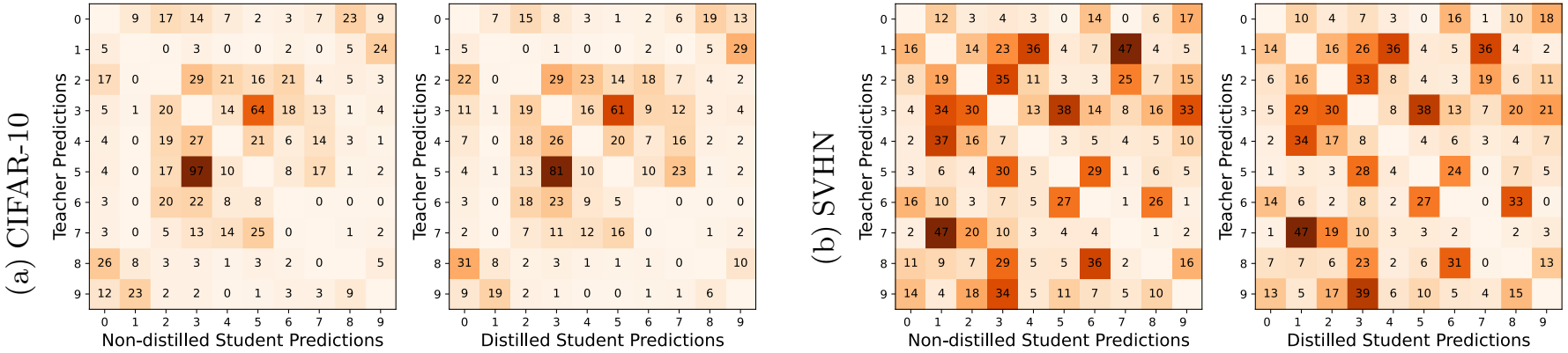


Figure 2: **Class-wise Disagreement.** Disagreement between a ResNet-56 teacher and ResNet-20 (left) non-distilled/(right) distilled student for (a) CIFAR-10 using $T = 9$ and (b) SVHN using $T = 7$. The diagonals are excluded since here both models predict the same class without any disagreement.

Dataset	Teacher (#param)	Student (#param)
CIFAR-10/100, SVHN	ResNet56 (0.85M)	ResNet20 (0.27M)

Class-wise Bias: Analysis

- **Q: What is the impact of increase temperature T on the model's class biases?**
- **TC = Teacher vs. Distilled Student, SC = Trained Student vs. Distilled Student**

Table 1: **Class-wise Bias and Distillation.** The number of statistically significantly affected classes comparing the class-wise accuracy of *teacher vs. Distilled Student (DS) models*, denoted #TC, and *Non-Distilled Student (NDS) vs. distilled student models*, denoted #SC.

		CIFAR-100						ImageNet					
Teacher/Student		ResNet56/ResNet20			DenseNet169/DenseNet121			ResNet50/ResNet18			ViT-Base/TinyViT		
Model	Temp	Test Acc. (%)	#SC	#TC	Test Acc. (%)	#SC	#TC	Test Top-1 Acc. (%)	#SC	#TC	Test Top-1 Acc. (%)	#SC	#TC
Teacher	-	70.87 ± 0.21	-	-	72.43 ± 0.15	-	-	76.1 ± 0.13	-	-	81.02 ± 0.07	-	-
NDS	-	68.39 ± 0.17	-	-	70.17 ± 0.16	-	-	68.64 ± 0.21	-	-	78.68 ± 0.19	-	-
DS	2	68.63 ± 0.24	5	15	70.93 ± 0.21	4	12	68.93 ± 0.23	77	314	78.79 ± 0.21	83	397
DS	3	68.92 ± 0.21	7	12	71.08 ± 0.17	4	11	69.12 ± 0.18	113	265	78.94 ± 0.14	137	318
DS	4	69.18 ± 0.19	8	9	71.16 ± 0.23	5	9	69.57 ± 0.26	169	237	79.12 ± 0.23	186	253
DS	5	69.77 ± 0.22	9	8	71.42 ± 0.18	8	9	69.85 ± 0.19	190	218	79.51 ± 0.17	215	206
DS	6	69.81 ± 0.15	9	8	71.39 ± 0.22	8	8	69.71 ± 0.13	212	193	80.03 ± 0.19	268	184
DS	7	69.38 ± 0.18	10	6	71.34 ± 0.16	9	7	70.05 ± 0.18	295	174	79.62 ± 0.23	329	161
DS	8	69.12 ± 0.21	13	6	71.29 ± 0.13	11	7	70.28 ± 0.27	346	138	79.93 ± 0.12	365	127
DS	9	69.35 ± 0.27	18	9	71.51 ± 0.23	12	9	70.52 ± 0.09	371	101	80.16 ± 0.17	397	96
DS	10	69.24 ± 0.19	22	11	71.16 ± 0.21	14	10	70.83 ± 0.15	408	86	79.98 ± 0.12	426	78

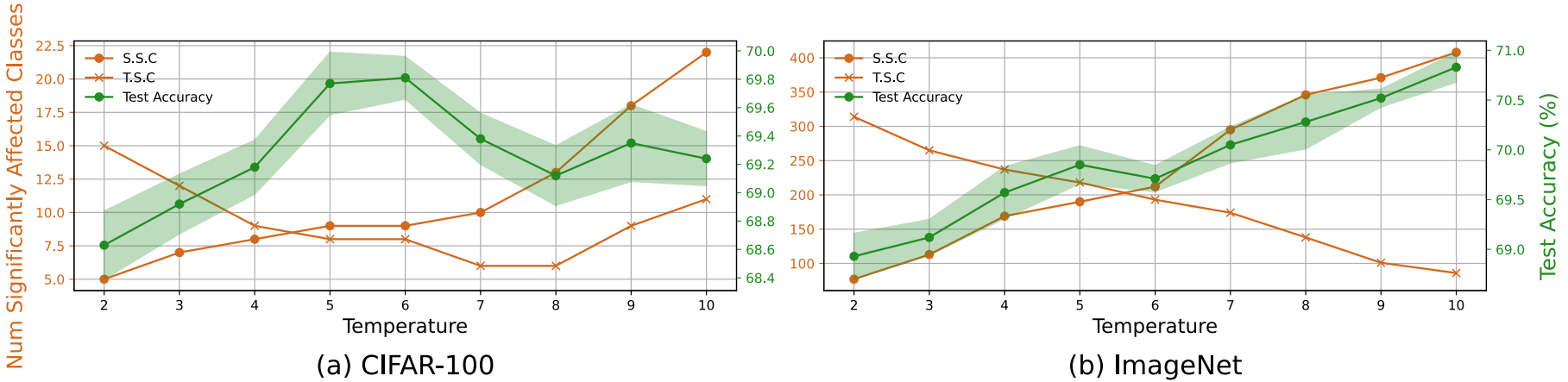


Figure 3: **Temperature vs. Test Accuracy/Class Bias.** Number of non-distilled vs. distilled student significantly affected classes (S.S.C.) and the number of teacher vs. distilled student significantly affected classes (T.S.C.) by distillation in (a) CIFAR-100 (ResNet-56/ResNet-20) and (b) ImageNet datasets (ResNet-50/ResNet-18), with 100 and 1000 total classes respectively. As the temperature used for distillation increases up to T=10, the S.S.C. rises for both datasets. For ImageNet, T.S.C. decreases, while for CIFAR-100, it first decreases and then slightly increases. The changes in the distilled student's test accuracy over all classes are also depicted in the figure.

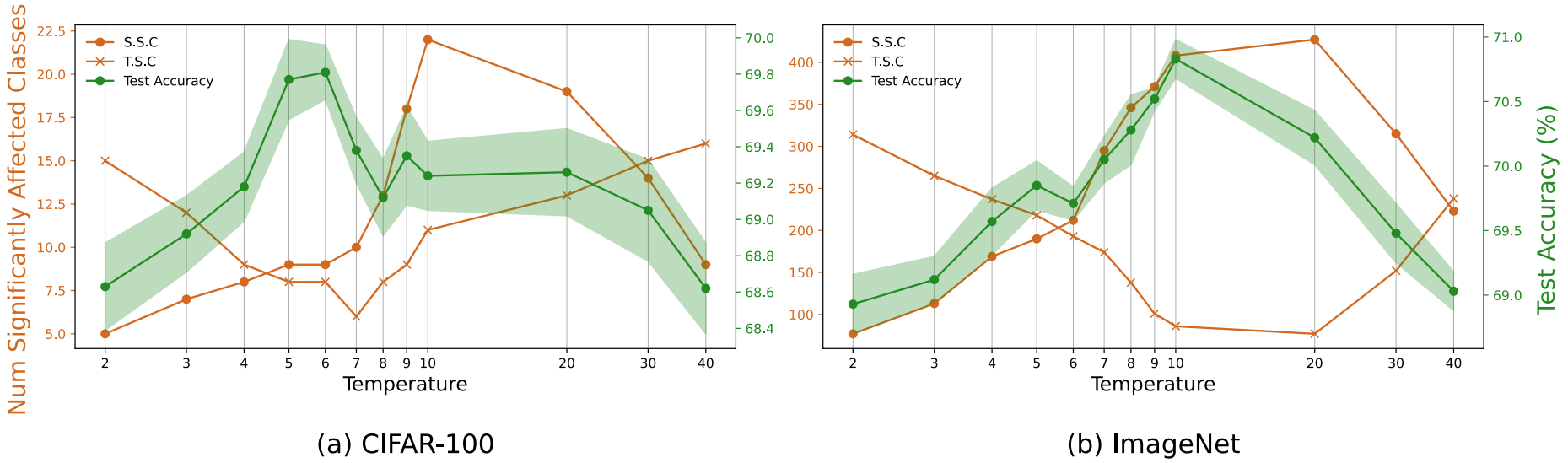
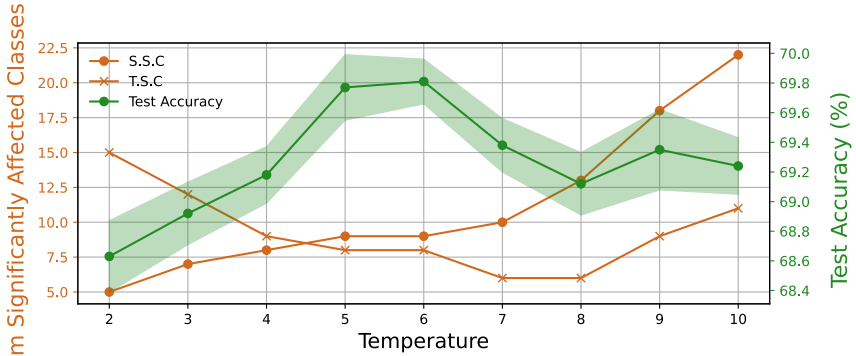


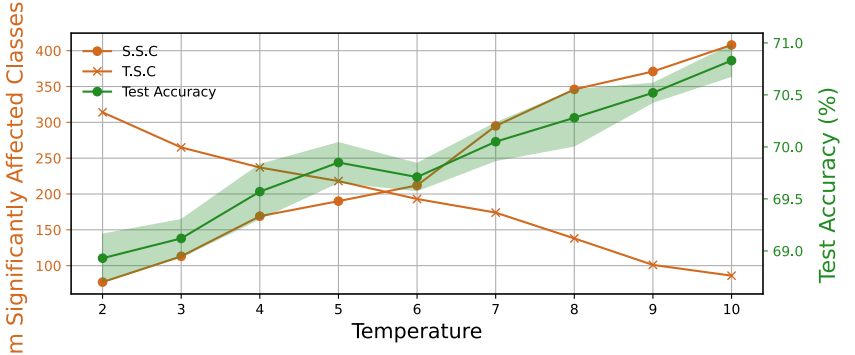
Figure 8: **Temperature vs. Test Accuracy/Class Bias.** Number of non-distilled vs. distilled student significantly affected classes (S.S.C.) and the number of teacher vs. distilled student significantly affected classes (T.S.C.) by distillation in (a) CIFAR-100 (ResNet-56/ResNet-20) and (b) ImageNet datasets (ResNet-50/ResNet-18), with 100 and 1000 total classes respectively. As the temperature used for distillation increases, the S.S.C. rises for both datasets up to a certain T, after which it decreases. Meanwhile, T.S.C. decreases first and then increases. The changes in the distilled student Test Accuracy over all classes are also depicted in the figure.

Distillation and Class Bias

- When we distill a large teacher model to a small student, **clearly the learned function is different**
- Distillation does not affect class-wise accuracy uniformly
- **However, a change in class bias alone is not meaningful (bad or good) in itself**
- **How can we judge if this is good or bad for applications?**



(a) CIFAR-100

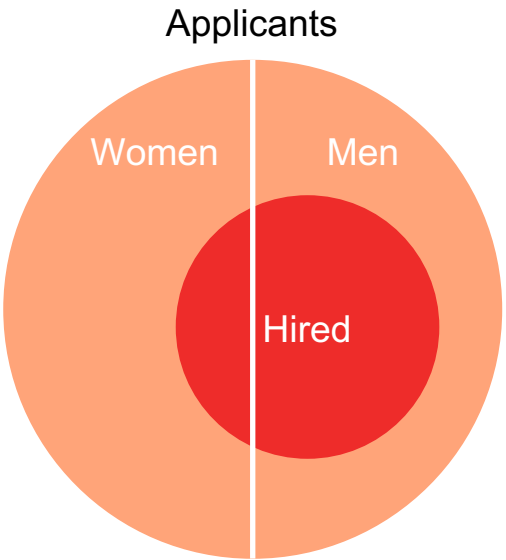


(b) ImageNet

Group Fairness

Group Fairness

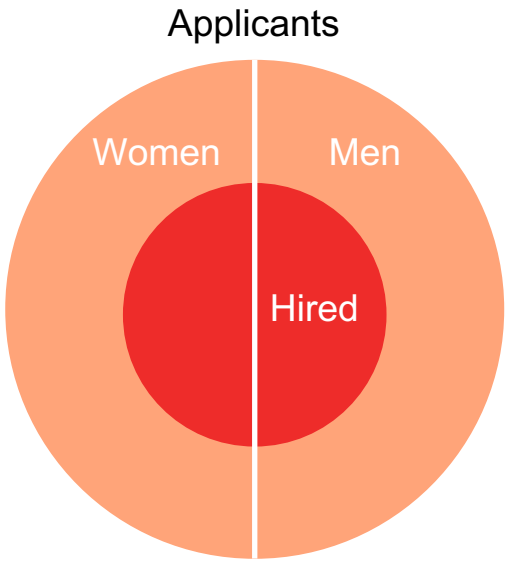
- A change in class bias alone is not meaningful (bad or good) in itself...
- What is clearly bad are **unfair outcomes**, i.e. a model not treating individuals from different groups **equitably**
- An example is a hiring system that accepts more men than women



Group Fairness: Demographic Parity

- We want individuals belongs to different groups to have **equal probability of a positive outcome**
  - e.g. we want men and women to have equal odds of being hired
- Let A be the sensitive attribute (gender), and $\hat{Y} = 1$ be the outcome (i.e. hired), we want:

$$P(\hat{Y} = 1 \,|A = a) = P(\hat{Y} = 1 \,|A = b)$$



Group Fairness Metrics: Demographic Parity Difference

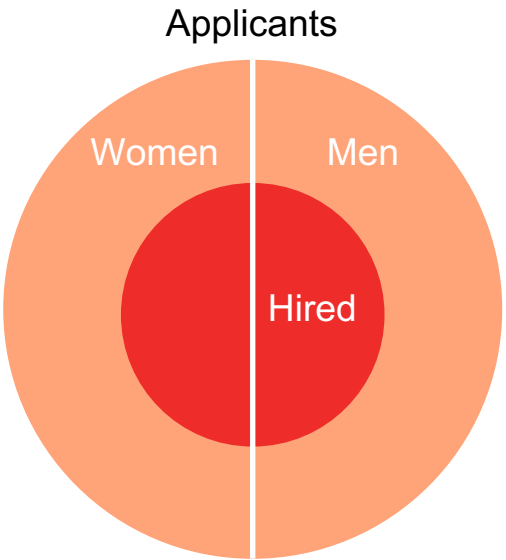
- Demographic Parity:

$$P(Y = 1 \mid A = a) = P(Y = 1 \mid A = b)$$

- A **metric** based on demographic parity is the Demographic Parity Difference (DPD):

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

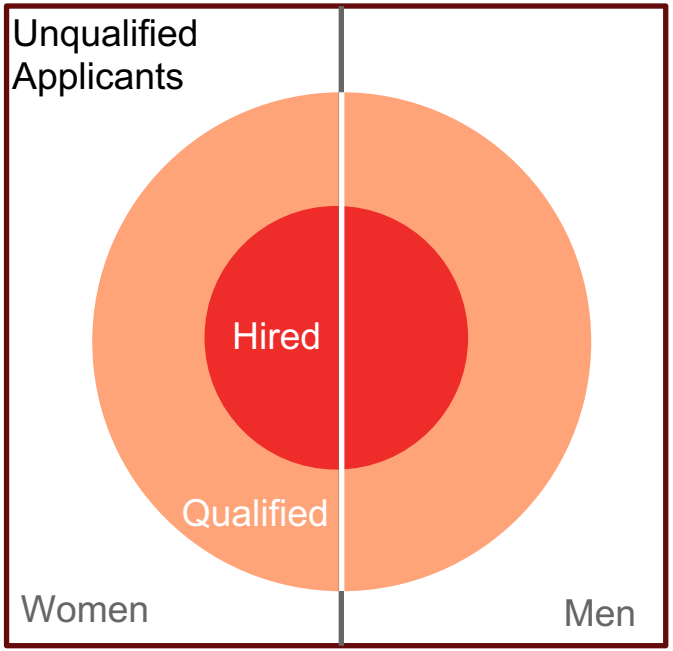
- DPD = 0 means perfectly fair in demographic parity fairness



Group Fairness: Equalized Odds

- We want individuals to have **equal probability of a positive or negative outcome** given a **condition is true**
  - e.g. we want men and women to have equal odds of being hired/not, if they are **qualified**
- Let A be the sensitive attribute, $\hat{Y}$ be the outcome, and Y be the true label, we want:

$$P(\hat{Y} = 1 \,|Y = y, A = a) = P(\hat{Y} = 1, Y = y \,|A = b)$$



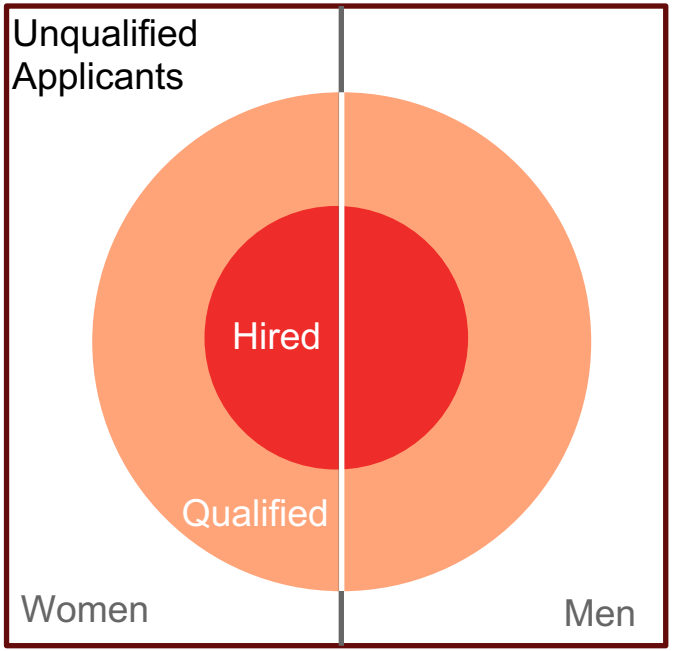
Hardt, Moritz; Price, Eric; Srebro, Nathan (2016). "Equality of Opportunity in Supervised Learning". *Neural Information Processing Systems*. **29**. arXiv:1610.02413.

Group Fairness Metrics: Demographic Parity Difference

- Demographic Parity:

$$P(\hat{Y} = 1 \,|Y = y, A = a) = P(\hat{Y} = 1, Y = y \,|A = b)$$

- We use a metric based on equalized odds: Equalized Odds Difference (EOD)
- EOD=0 means perfectly fair in equalized odds fairness



CelebA Dataset

- CelebA is a dataset of celebrity photos
- CelebA has protected attributes, such as gender and age
- Also has independent attributes such as “smiling” or “glasses”

- Often used in fairness, but is also a deeply problematic dataset…

Deep Learning Face Attributes in the Wild. Liu, Ziwei and Luo, Ping and Wang, Xiaogang and Tang, Xiaoou. Proceedings of International Conference on Computer Vision (ICCV), 2015.

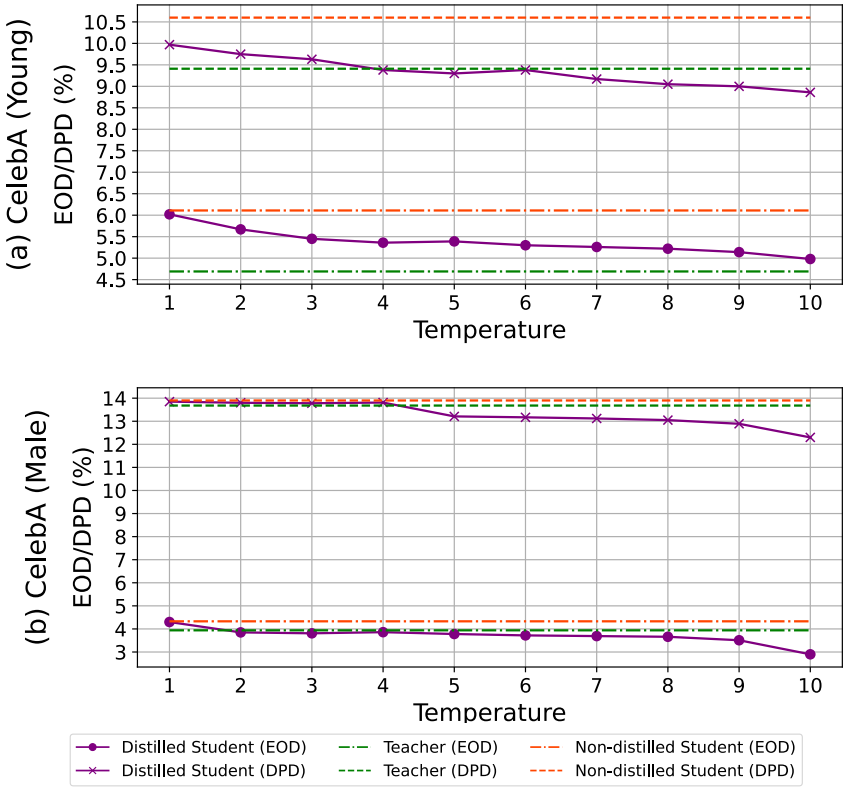


Figure 4: **Evaluation of Fairness Metrics for Distilled Students in Computer Vision (CV) .** Equalized Odds Difference (EOD) and Demographic Parity Difference (DPD) are reported in % and lower values indicate improved fairness. (a) illustrates fairness metrics for the CelebA dataset with 'smiling' label concerning the 'Young' demographic attribute and (b) concerning the 'Male' demographic attribute. (c) presents fairness metrics for the Trifeature dataset with 'shape' label with regard to the 'color' attribute and (d) with regard to the 'texture' attribute. It is notable that the models are fairer for the Trifeature dataset compared to the CelebA dataset with lower values in metrics. The explanation lies in the fact that the Trifeature dataset maintains a balanced distribution of demographic attributes, while the CelebA dataset contains biases that mirror real-world disparities. As seen in the second column, the downward trend does not continue at very high temperatures (T=20,30,40), as the teacher model generates nearly uniform softmax outputs.

- ResNet50 (24M) ➔ResNet18 (11.4M) distillation with CelebA dataset
- Protected attribute is Age (top) and Gender (bottom)
- Evaluated on “smiling” classification

- Fairness improves (i.e. EOD/DPD decreases) with higher T

UNIVERSITY OF CALGARY

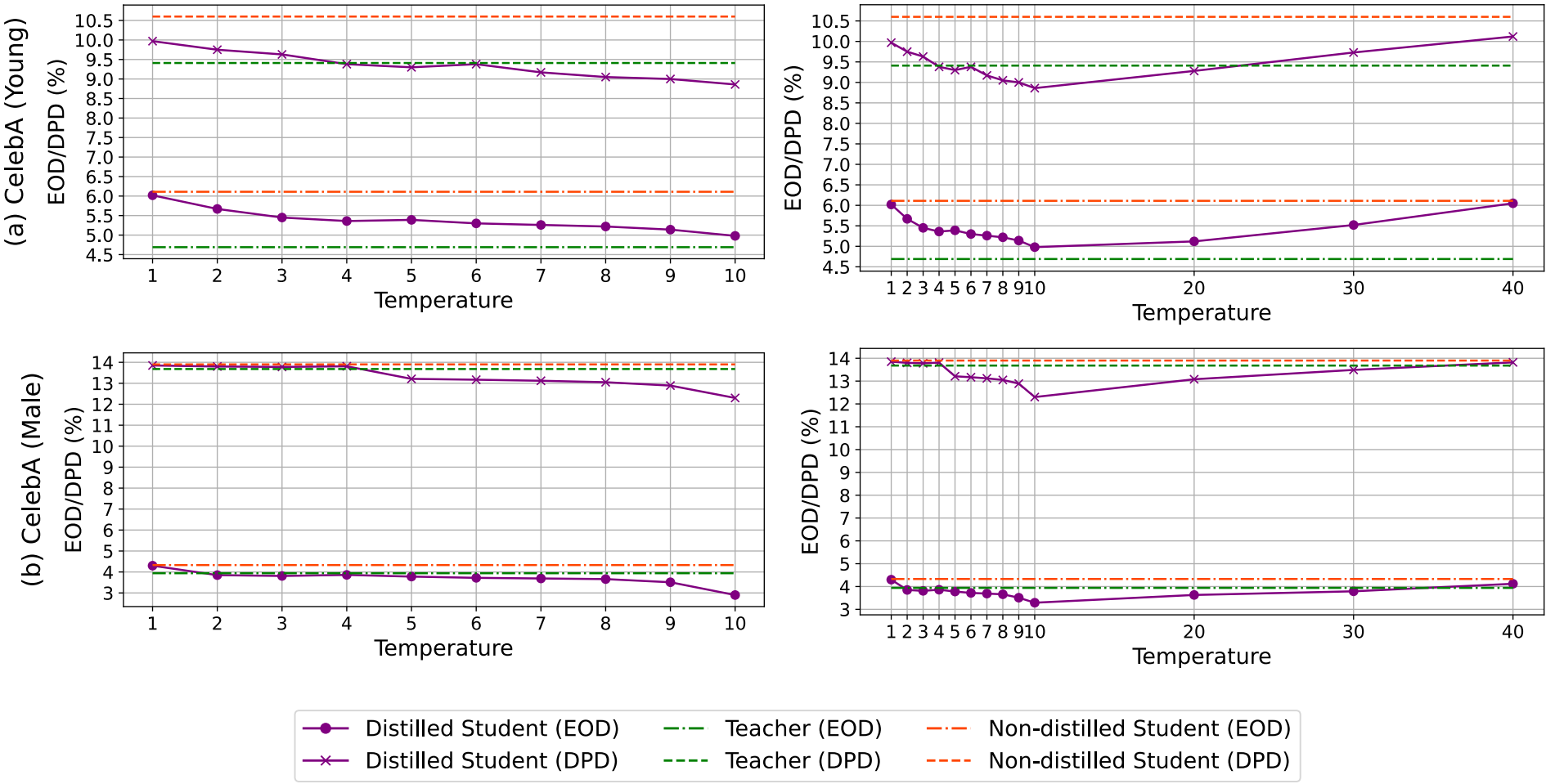
Teacher/Student:		CelebA (smiling) ResNet-50 / ResNet-18		
Model	Temp	Test Acc. (%) ↑	EOD ↓	DPD ↓
Teacher	–	93.09 ± 0.08	**4.69 ± 0.06**	9.41 ± 0.11
NDS	–	92.03 ± 0.03	6.11 ± 0.05	10.60 ± 0.08
DS	1	92.12 ± 0.06	6.02 ± 0.11	9.97 ± 0.08
DS	2	92.14 ± 0.11	5.67 ± 0.08	9.75 ± 0.09
DS	3	92.53 ± 0.13	5.45 ± 0.05	9.63 ± 0.06
DS	4	92.17 ± 0.10	5.36 ± 0.02	9.38 ± 0.03
DS	5	92.29 ± 0.05	5.39 ± 0.04	9.30 ± 0.05
DS	6	92.26 ± 0.08	5.30 ± 0.01	9.38 ± 0.07
DS	7	92.12 ± 0.08	5.26 ± 0.05	9.17 ± 0.10
DS	8	92.66 ± 0.12	5.22 ± 0.02	9.05 ± 0.04
DS	9	93.18 ± 0.15	5.14 ± 0.04	9.01 ± 0.08
DS	10	92.57 ± 0.11	4.98 ± 0.03	**8.86 ± 0.04**

Table 2: **Fairness Metrics and Distillation.** The performance of teacher, Non-Distilled Student (NDS), and Distilled Student (DS) models with a range of temperatures T on the Trifeature and CelebA datasets. Fairness metrics are presented for Trifeature with regard to color attribute and for CelebA with regard to the Young demographic attribute. With increasing temperature, EOD and DPD have a downward trend signifying enhanced fairness. Mean and std. dev. are over five random inits.

- ResNet50 (24M) ➔ResNet18 (11.4M) distillation with CelebA dataset
- Protected attribute is Age (top) and Gender (bottom)
- Evaluated on “smiling” classification

- Fairness improves (i.e. EOD/DPD decreases) with higher T

UNIVERSITY OF CALGARY


(a) CelebA (Young)
(b) CelebA (Male)
EOD/DPD (%)
Temperature
Distilled Student (EOD)
Distilled Student (DPD)
Teacher (EOD)
Teacher (DPD)
Non-distilled Student (EOD)
Non-distilled Student (DPD)

HateExplain Dataset

Target groups	Categories
Race	African, Arabs, Asians, Caucasian, Hispanic
Religion	Buddhism, Christian, Hindu, Islam, Jewish
Gender	Men, Women
Sexual Orientation	Heterosexual, Gay
Miscellaneous	Indigenous, Refugee/Immigrant, None, Others

Table 3: Target groups considered for the annotation.

	Twitter	Gab	Total
Hateful	708	5,227	5,935
Offensive	2,328	3,152	5,480
Normal	5,770	2,044	7,814
Undecided	249	670	919
Total	9,055	11,093	20,148

Table 4: Dataset details. “Undecided” refers to the cases where all the three annotators chose a different class.

HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, Animesh Mukherjee. AAAI 2021.

- HateExplain is a dataset used for detecting hate speech in online discourse
- Covers a range of protected groups (we use target groups aggregated, e.g. religion)
- We combine hateful/offensive to make task binary classification (”toxic” v.s. “normal”)

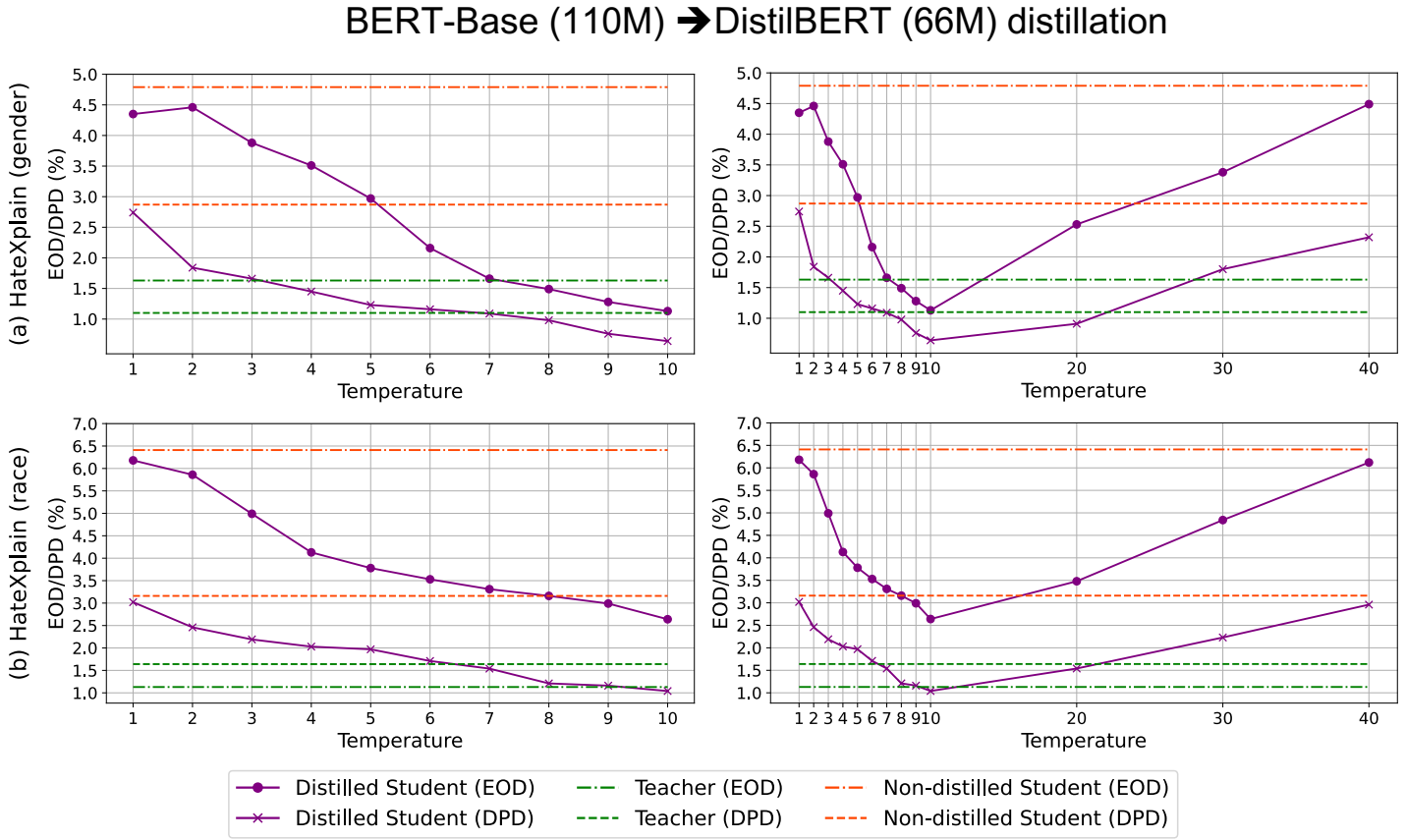
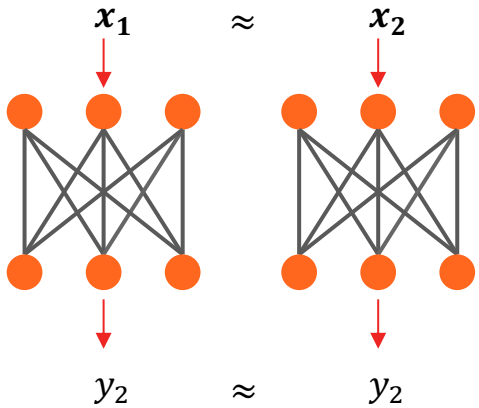


Figure 5: **Evaluation of Fairness Metrics for Distilled Students in Natural Language Processing (NLP).** Equalized Odds Difference (EOD) and Demographic Parity Difference (DPD) are reported in % and lower values indicate improved fairness. (a) illustrates fairness metrics for the HateXplain dataset concerning the 'gender' demographic attribute, and (b) with regard to the 'race' attribute. The teacher employed the BERT architecture, while the student used the DistilBERT architecture.

Individual Fairness Metrics

- Individual fairness metrics are very different
  - Group: individuals with different protected attributes should see similar outcomes
  - Individual: similar individuals should see similar outcomes
- Metics captures whether a model provides consistent predictions for **semantically similar inputs**, ensuring fairness at an **individual level**
- Lipschitz condition proposed by Dwork et al. (2012), smaller values = more fair



Fairness through awareness. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. In Proceedings of the 3rd Innovations in Theoretical Computer Science (ITCS), pages 214–226. ACM, 2012.

Table 4: **Individual Fairness Metrics Across Datasets.** Individual fairness scores for Teacher, Non-Distilled Student (NDS), and Distilled Student (DS) models across CelebA, Trifeature, and HateXplain datasets. Scores for DS models are reported for varying temperature values T.

		Individual Fairness ↓		
		CelebA	**Trifeature**	**HateXplain**
Model	Temp	ResNet-50 / ResNet-18	ResNet-20 / LeNet-5	BERT-Base / DistilBERT
Teacher	–	0.0407	0.016	0.0320
NDS	–	0.124	0.0462	0.1078
DS	1	0.113	0.0422	0.0994
DS	2	0.104	0.0407	0.0985
DS	3	0.0908	0.0393	0.0927
DS	4	0.0906	0.0387	0.0882
DS	5	0.0886	0.0384	0.0823
DS	6	0.0799	0.0377	0.0768
DS	7	0.0753	0.0356	0.0727
DS	8	0.0712	0.0349	0.0689
DS	9	0.0701	0.0341	0.0681
DS	10	**0.0697**	**0.0338**	**0.0654**

- **Clear increase in individual fairness with increased distillation temp**

Conclusion

- Knowledge Distillation is pervasive in its use, **you are likely affected by the decisions of a distilled model daily**
- And yet the effect of **distillation temperature** on **model fairness** has not been looked at previously!
- We find across models, datasets and both vision and language modalities that distillation temperature **affects the bias and fairness of models**
- We also consistently find that **higher distillation temperatures** leads to more fair models
- In some cases, **distilled models (with high T) can be fairer than even the (much larger) teacher model!**

Future Directions

- Can distillation be an effective method of improving model fairness?
- Are there any trade offs to using large temperatures, less typically used with distillation in practice?
- Does distillation have a similar effect on LLMs, e.g. DeepSeek?

Questions?

yani.ioannou@ucalgary.ca

Aida Mohammadshahi
MSc (Defended Jan 2025)

Now seeking work!

What's Left After Distillation?
How Knowledge Transfer Impacts Fairness and Bias.
Aida Mohammadshahi, Yani Ioannou
Transactions in Machine Learning Research (TMLR), March 2025

NSERC CRSNG

ALBERTA INNOVATES

amazon | science

CMI

UNIVERSITY OF CALGARY