Enhancing Knowledge Distillation via Bias Mitigation

Our Team

Michelle Bolner

Data Scientist / PM / ML Ops

Keith Hutton

Data Scientist / DE / ML Ops

Landon Morin

Data Scientist / DE / ML Ops

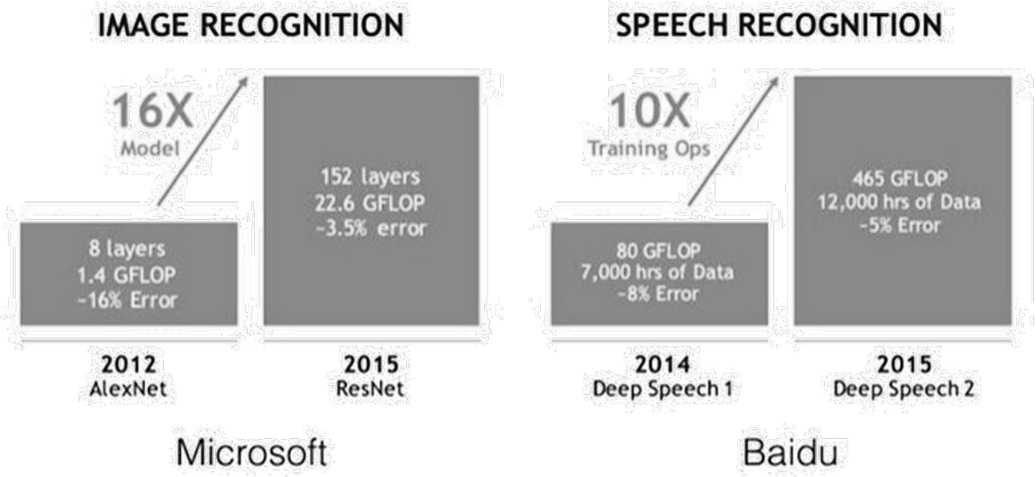
Lindsey Bang

Data Scientist / PM / ML Ops

Bleeding Edge Models Get Better At a Cost to Size

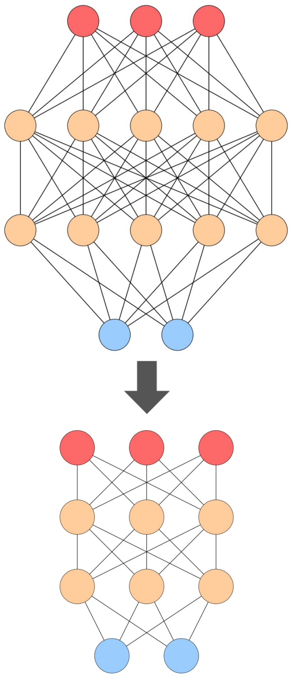
- Increased computational demand and size

Models Are Getting Larger



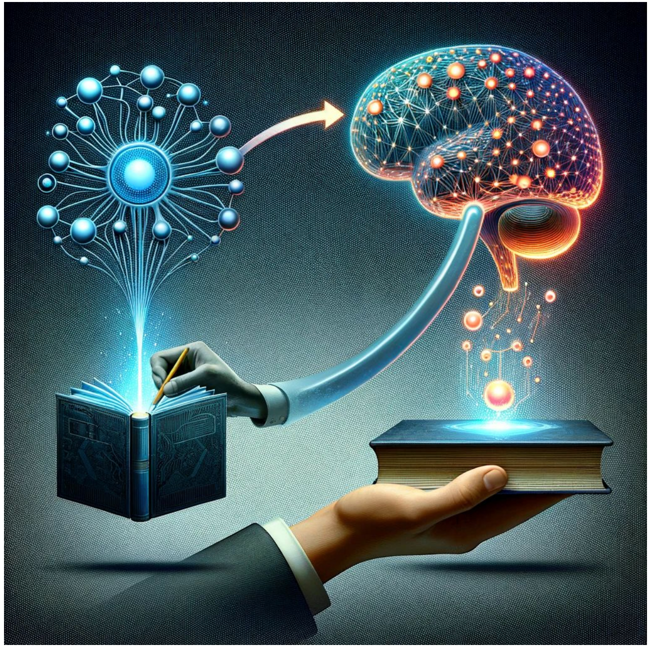
Dally, NIPS'2016 workshop on Efficient Methods for Deep Neural Networks

The Solution...
Neural Network Compression



- Reduces model size, enabling deployment on resource-constrained devices
- Lowers computational and energy costs, making ML more sustainable and cost-effective

Knowledge Distillation is the Bleeding Edge Solution for Compression

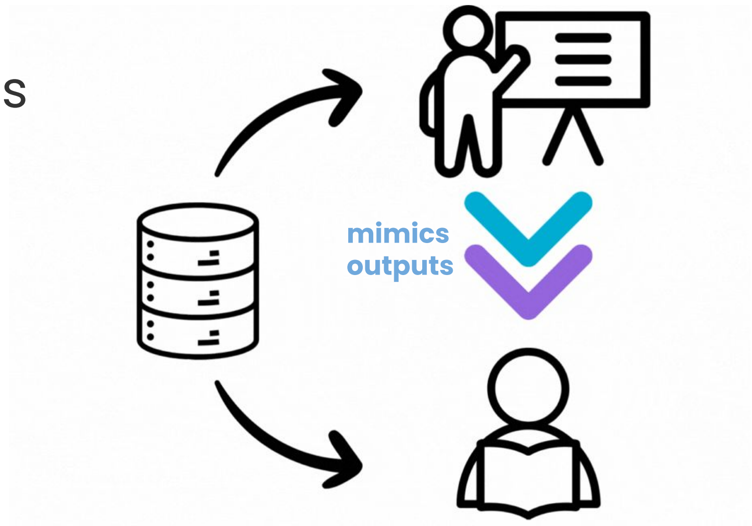


- Compress model size
- Secure
- Robust against domain shift

Knowledge Distillation Zoomed-Out

Large, performant model acts as teacher

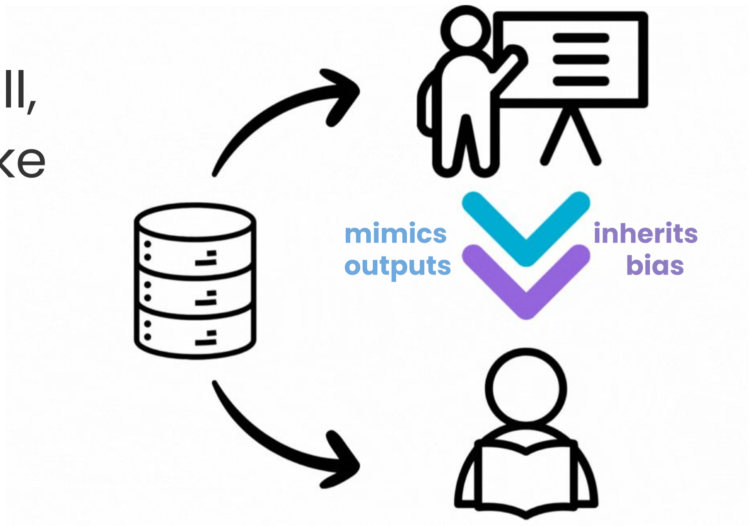
A smaller, predefined model learns to mimic the outputs of the teacher



But... Knowledge Distillation Inflates Bias as it Learns from a Biased Teacher

Teacher models learn a task well, but can learn stereotypes just like a human

As the student learns from the teacher, these stereotypes can be exaggerated



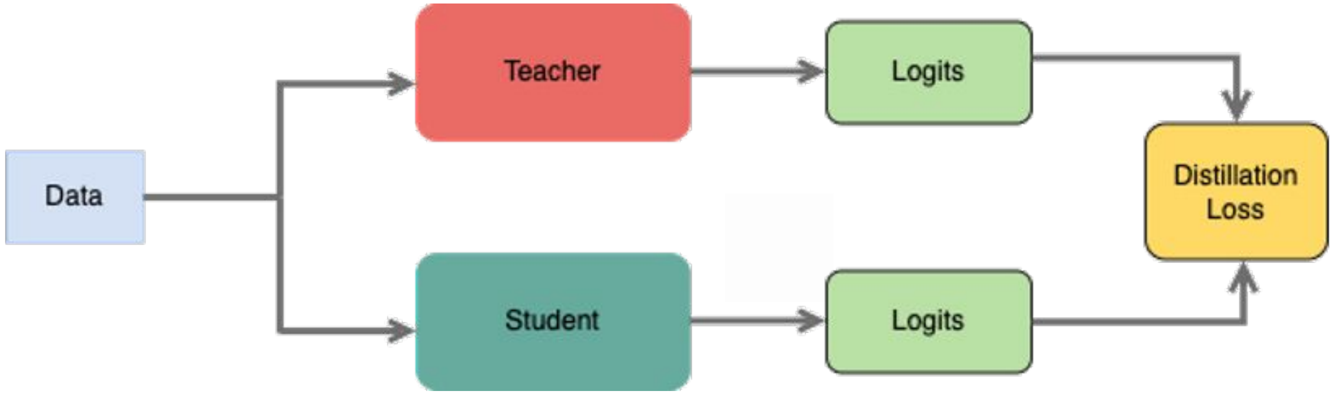
We Make Models Smaller, While Mitigating Bias Inflation

- Establish comprehensive evaluation metrics, to include bias, for top performing Knowledge Distillation techniques

- Integrate debiasing into the knowledge distillation framework for image classification

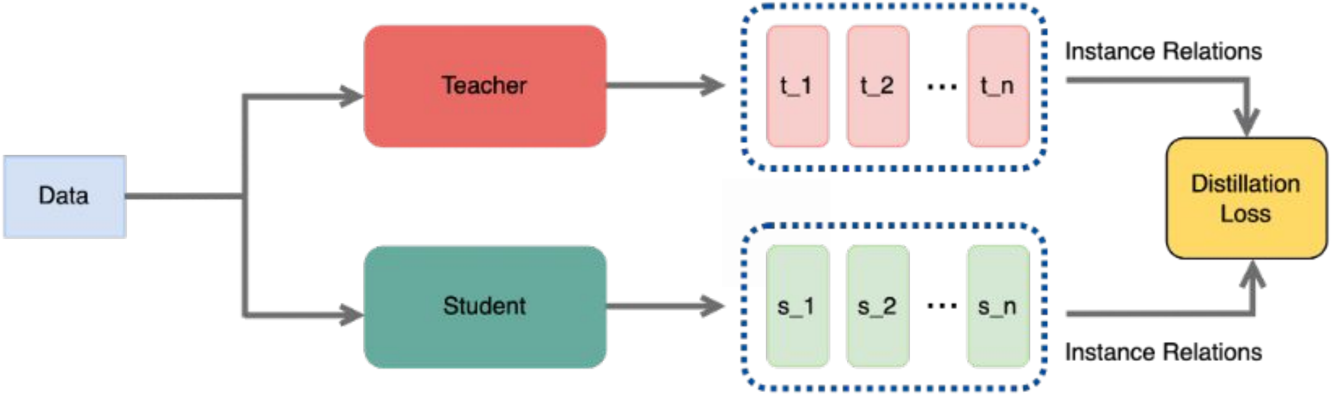
Knowledge Distillation First Introduced

Classic Knowledge Distillation



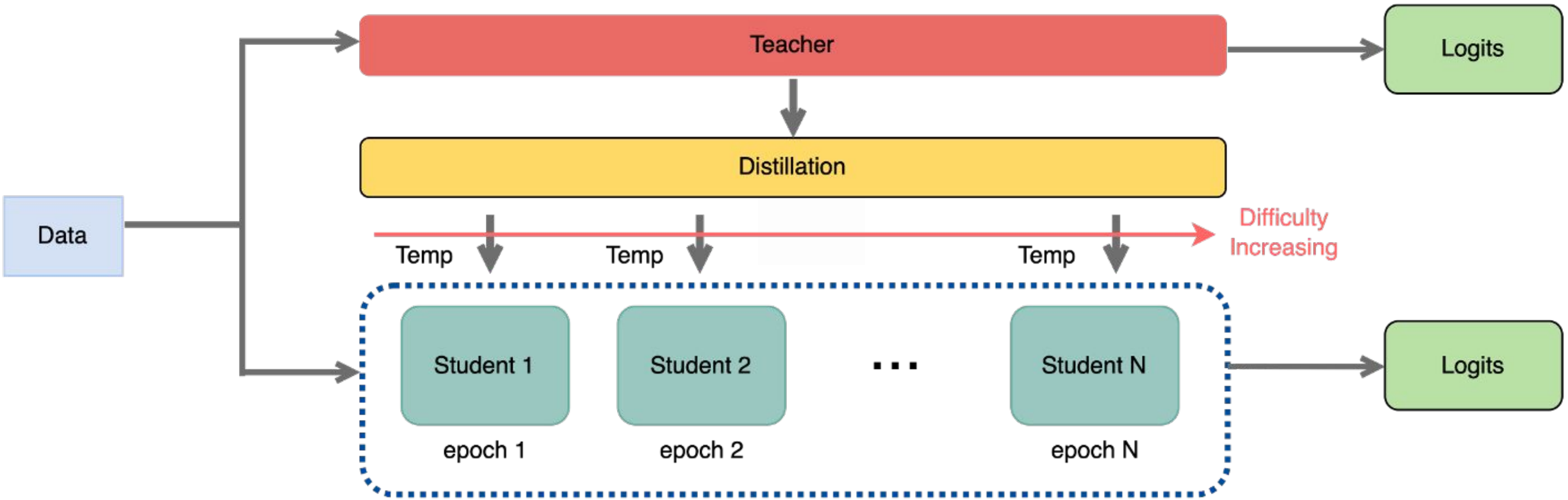
New Knowledge Distillation Frameworks

Relational Knowledge Distillation



Knowledge Distillation First Introduced

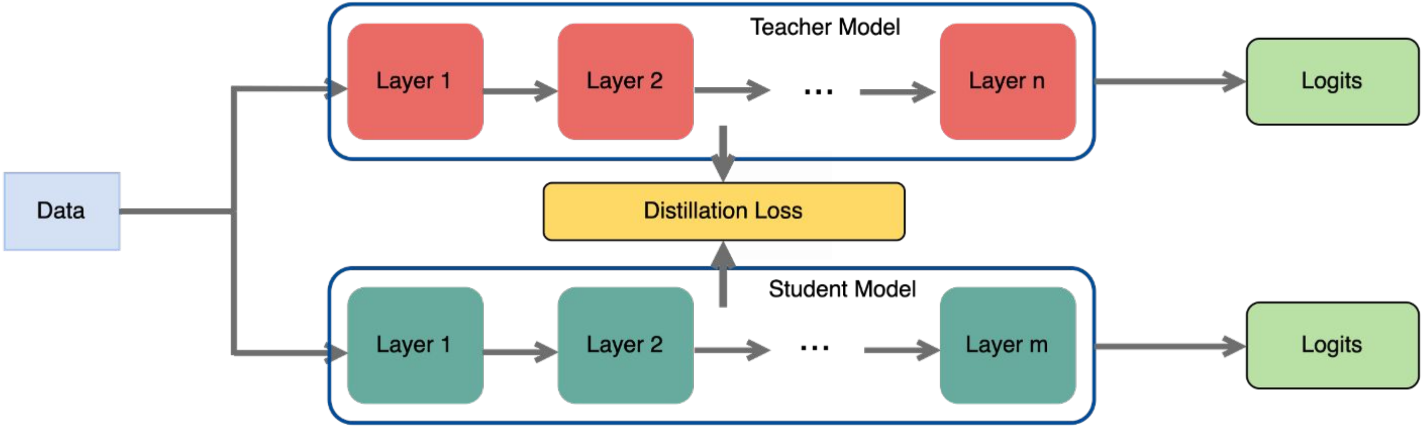
Curriculum Temperature Knowledge Distillation



Literature Review

New Knowledge Distillation Frameworks

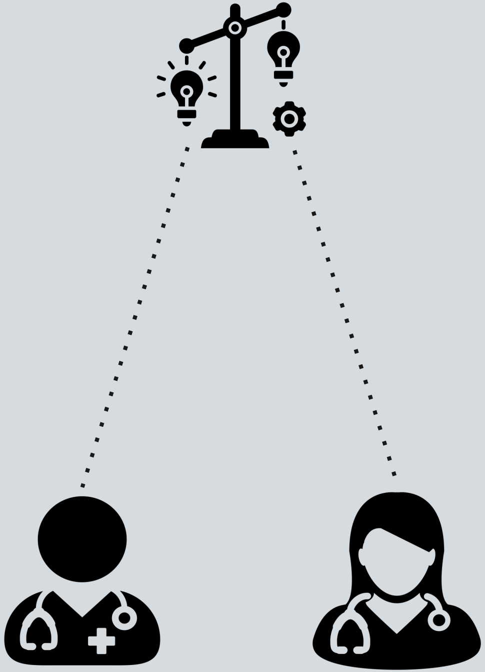
Regularizing Feature Norm and Direction



We Propose a New Metric for Knowledge Distillation...

Disparity

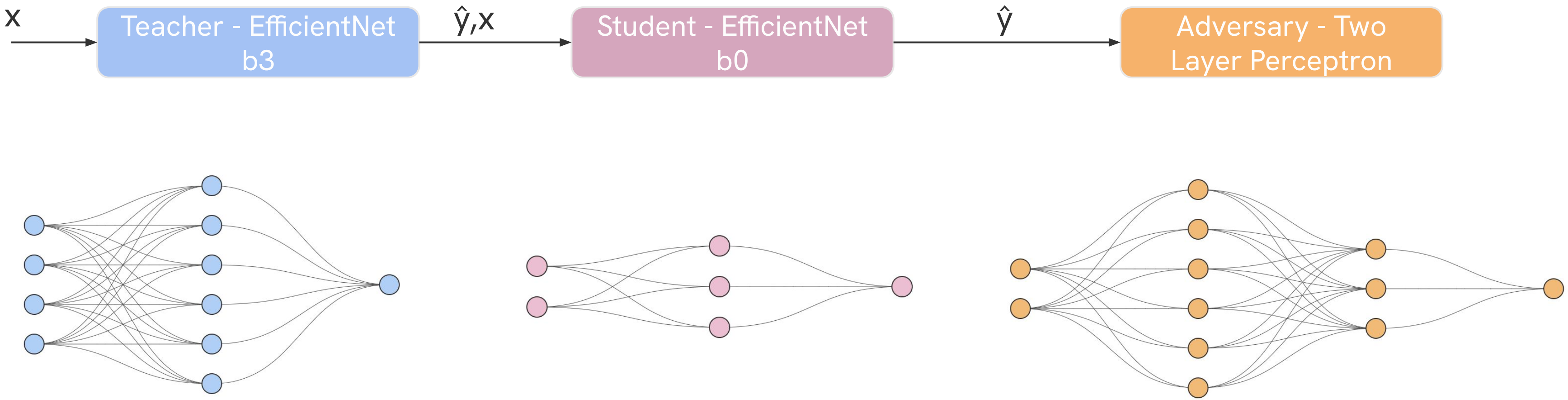
recall(♂ | 🩺) - *recall*(♀ | 🩺)



WIDER – An Attributed Dataset for Fairness Research

- WIDER Attribute dataset
  - 13,789 images
  - 30 event-type classes clustered to 16
  - 14 human attributes condensed to 1 protected attribute - gender

To Reduce Bias, We Incorporate An Adversarial Attack



$\mathcal{L}_{ce}$

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

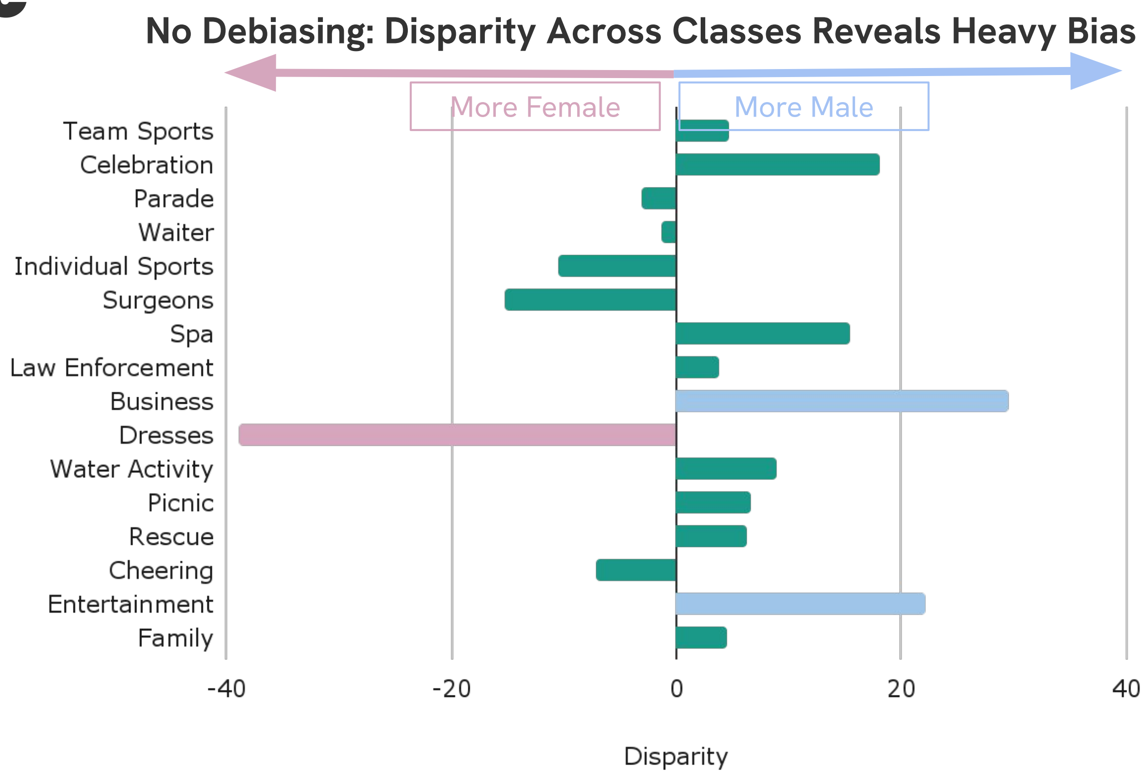
$\mathcal{L}_{adv} = \max\{1, \lambda \cdot (|(g \circ f_0)(X_j)| - \epsilon + \delta) + 1\}$

Model Evaluation

Method	Models (Teacher - Student)	Top-1 Accuracy	Recall	Precision	F1	Disparity
CKD	EfficientNet_b3 - EfficientNet_b0	64.9-65.0	64.9-65.0	66.0-65.0	64.6-64.7	**9.1-9.8**
RKD	EfficientNet_b3 - EfficientNet_b0	65.1-65.6	65.1-65.6	66.3-65.7	65.3-65.2	**8.4-12.3**
CTKD	EfficientNet_b3 - EfficientNet_b0	66.6-63.7	66.6-63.7	66.0-64.2	65.9-63.4	**10.6-12.3**
KD++	EfficientNet_b3 - EfficientNet_b0	62.1-59.6	62.1-59.6	61.7-60.6	60.6-59.6	**8.1-11.0**

Without Debiasing, the Student Learns to Stereotype

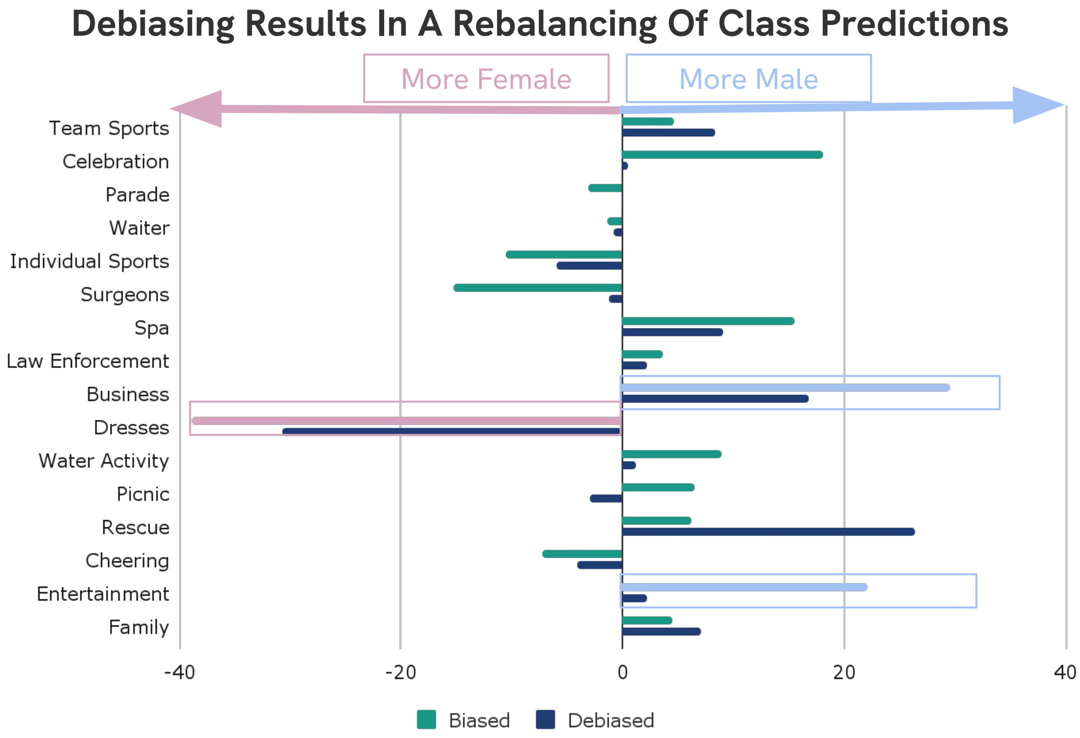
With no student-level debiasing, the student maintains a high average bias of .1226, which represents a **46% increase over the Teacher's bias of 0.0840**



With an Adversary Attack, the Model Becomes Less Biased

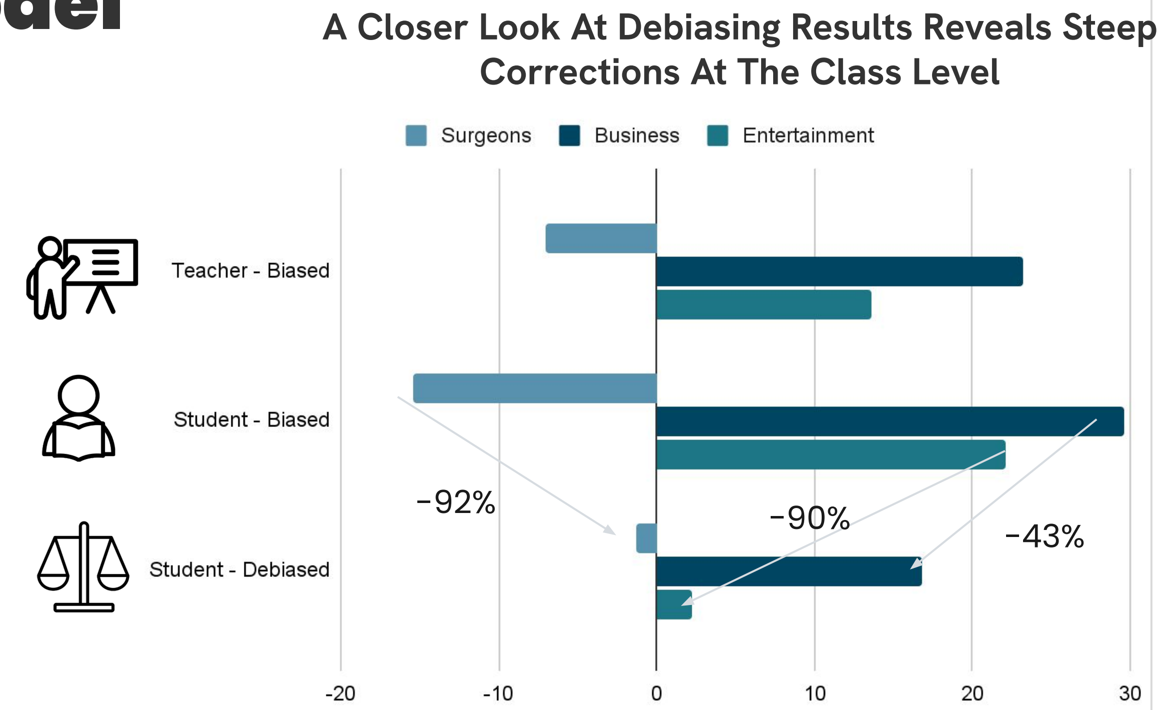
With a lambda of 0.5, we achieve a **mean absolute value disparity of 0.0748.**

This represents a **39% reduction in bias over a student model** with no debiasing, with only a **0.36% penalty to accuracy.**



The Adversary Course Corrects the Student Model

With a lambda of 0.5, the adversary **corrects the recall disparity of the surgeon class by 92%,** and the **business class by 24%,** resulting in fairer predictions with minimal impact to overall performance.



Demo

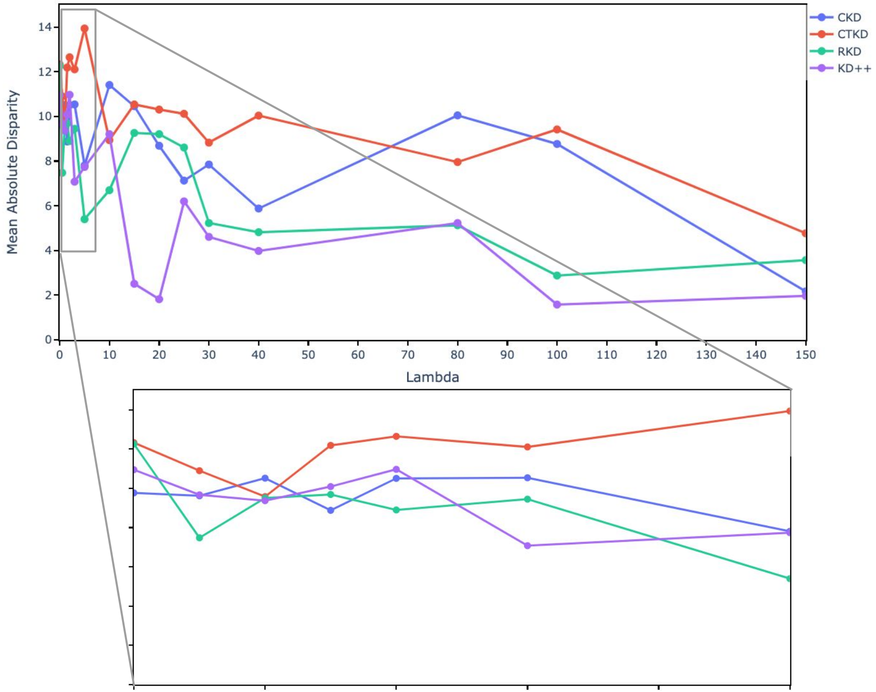
Attribute: Female
Not-Debiased: Waitress
De-Biased: Business

Attribute: Male
Not-Debiased: Team Sports
De-Biased: Entertainment

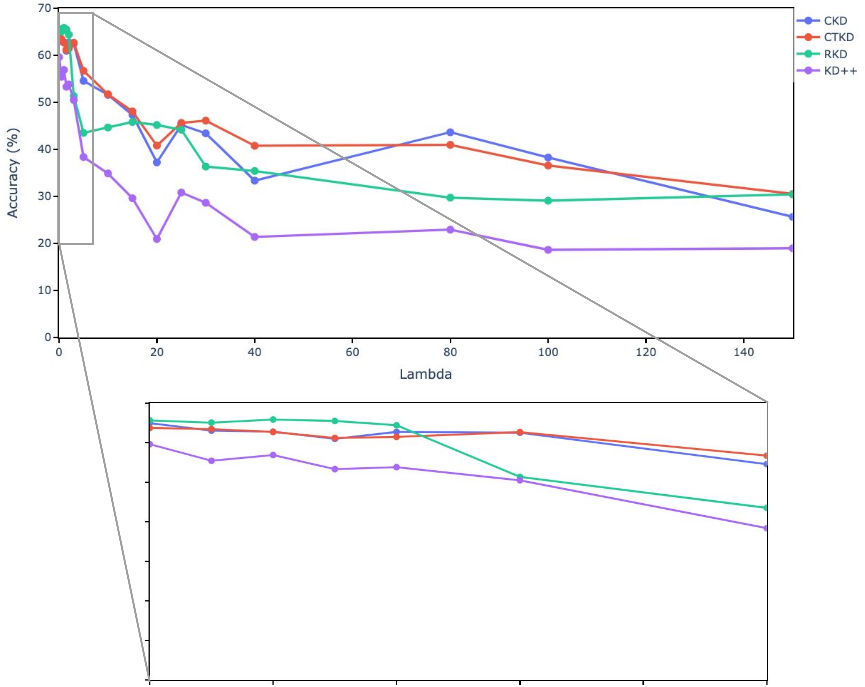
Attribute: Female
Not-Debiased: Family
De-Biased: Surgeons

And Generally, The Stronger We Make The Adversary, The Lower The Bias

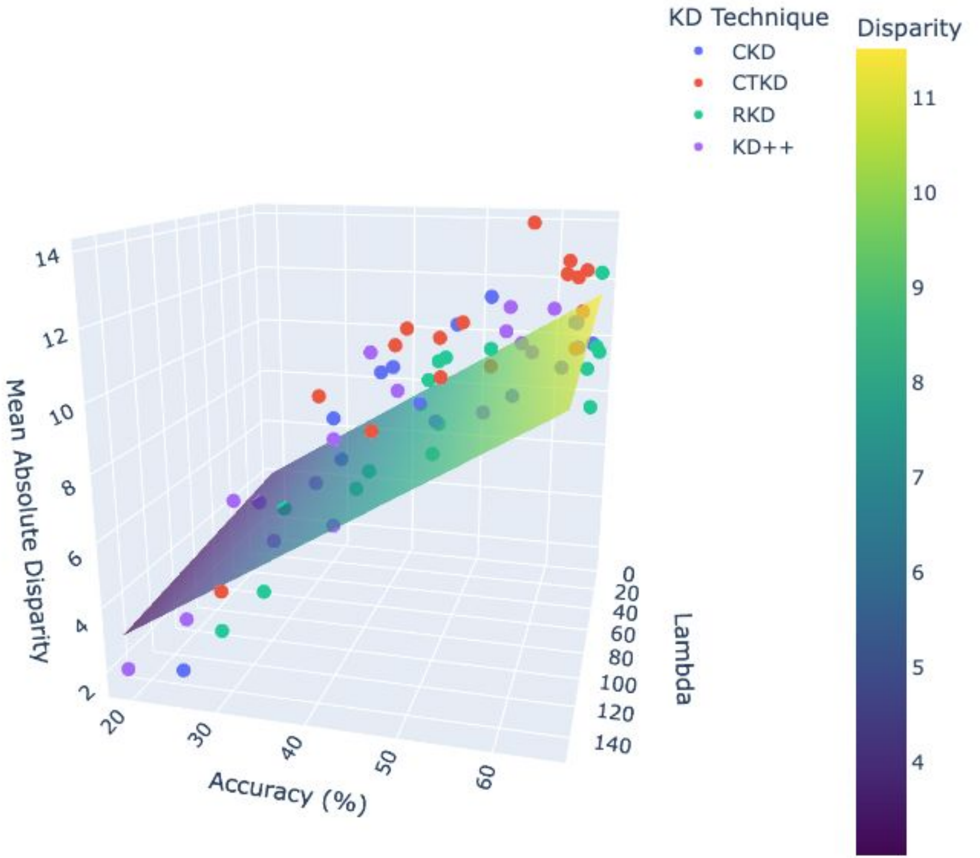
Higher Lambda Values Can Result in Lower Bias



That Bias Decline Will Come With Lower Accuracy



3D Plot Views



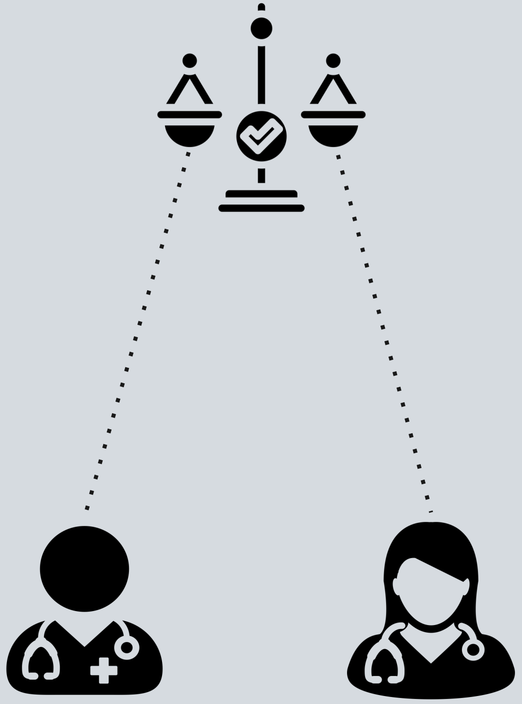
There is an Inverse Relationship Between Bias and Accuracy, Asymptotically

At a Lambda of 150 (High Adversary Prioritization) Disparity Declines Significantly…

…but at High Lambdas (High Adversary Prioritization), Accuracy is Penalized

Project Mission Statement

This project is dedicated to applying advanced Neural Network Compression strategies, enabling efficient deployment on resource-constrained edge devices while maintaining optimal performance levels. A key focus is to introduce and emphasize a diverse array of metrics, typically overlooked in this field. Furthermore, a crucial aspect of our mission is to thoroughly examine and actively reduce the propagation of bias within these compressed models, ensuring more equitable and responsible use of neural network technology.



Reference

Literature Review

“Distilling the Knowledge in a Neural Network (Classic Knowledge Distillation), 2015, Hinton et al. https://arxiv.org/abs/1503.02531

“Relational Knowledge Distillation”, 2019, Park et at. https://arxiv.org/abs/1904.05068

Knowledge Distillation : model compression accelerates inference speed in edge devices, 2021, https://medium.com/@hintkit/introduction-to-knowledge-distillation-3345b567d121

Yang, J., Soltan, A.A.S., Eyre, D.W. et al. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. npj Digit. Med. 6, 55 (2023). https://doi.org/10.1038/s41746-023-00805-y

Dong, Y., Zhang, B., Yuan, Y., Zou, N., Wang, Q., & Li, J. (2023). RELIANT: Fair Knowledge Distillation for Graph Neural Networks. [Preprint]. arXiv. https://arxiv.org/abs/2301.01150

Zhu, K., & Wu, J. (2021). Residual Attention: A Simple but Effective Method for Multi-Label Recognition. arXiv preprint arXiv:2108.02456.

Images, Slide Design, and Logos

Slide Template, Images & Icons

- Slide Template - Slidesgo
- Logos - nounproject
- Logos - canvas
- Competition Models Diagrams - Medium
- Title Slide Image - ChatGPT

Appendix

Pipeline for Fairness and New Metrics