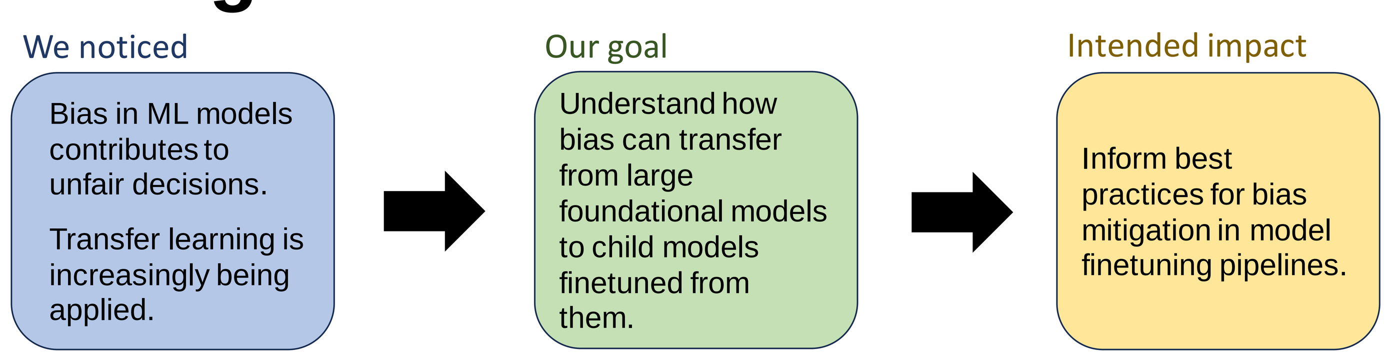
Persistence of Strong Foundational Model Stimuli for Fine-tuned Child Models

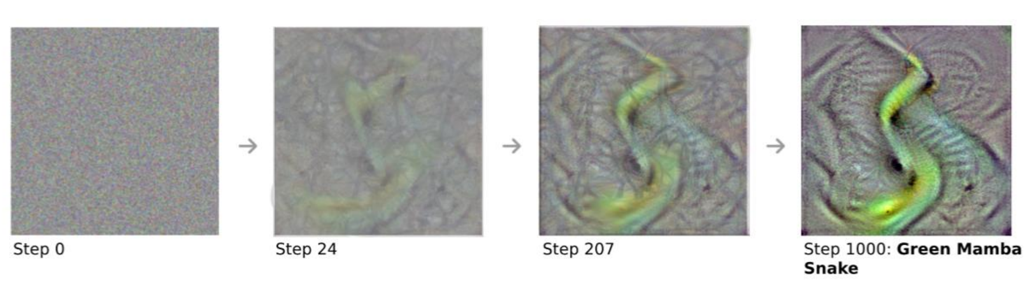
CAMLIS 2023

Transfer learning is standard practice for many computer vision tasks. Understanding the properties of child models' decision spaces inherited from large foundational models can reveal bias transfer to downstream models. We aim to characterize the decision space of finetuned models through the lens of strong stimuli produced against foundational models. Our results indicate that strong stimuli transfer to finetuned models at higher rates than models trained from scratch. We discuss implications for this bias transfer in the context of security vulnerabilities and fairness.

Background



We produce strong stimuli against foundational models using gradient-based optimization methods like those used to produce adversarial examples.

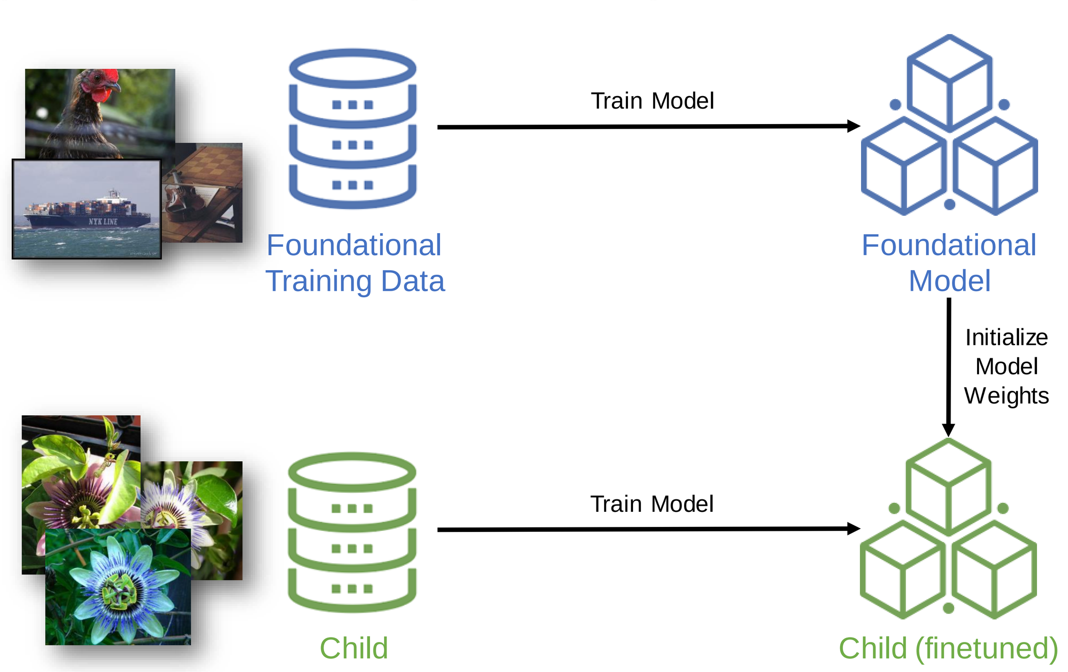


The resulting strong stimuli can be seen as stereotypical examples of a class from the perspective of the foundational model.

Experimental Setup

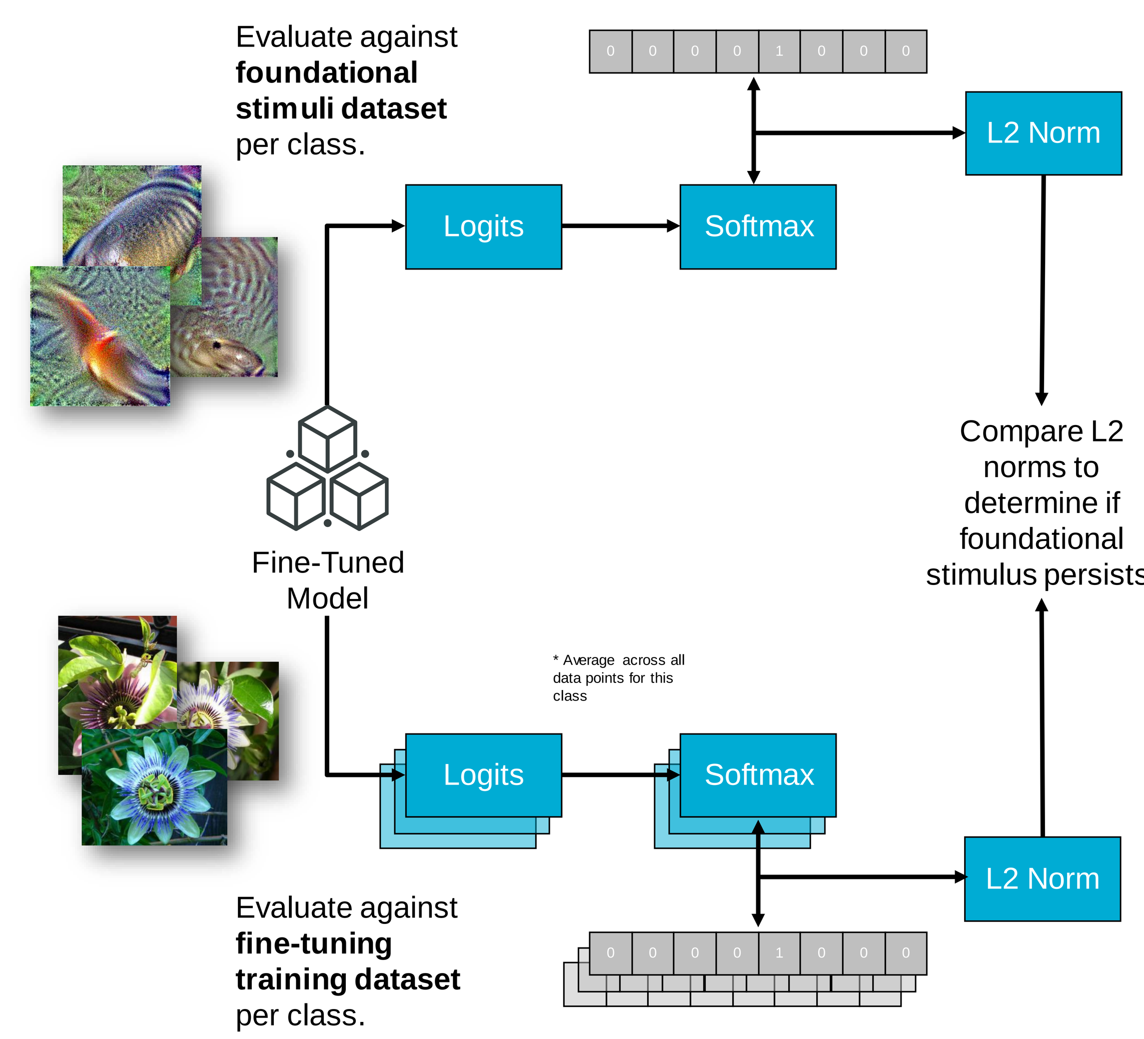
Our foundational model is a ResNet50 model trained on the first 102 classes of ImageNet1K_V1 dataset.

We finetuned 5 child models on the Oxford Flowers 102 dataset varying the number of blocks of convolutional layers frozen during retraining.



Evaluating Transfer of Strong Stimuli

We evaluated the performance of the strong foundational model stimuli against the child models by comparing the distance between the Softmax output for each strong stimulus to the one-hot vector for the target class to the corresponding distance for the centroid of the Oxford Flowers 102 training set for the target class. All stimuli were then classified as "outperforming" or "underperforming" training centroids of their target class based on these distances to the target one-hot vector.

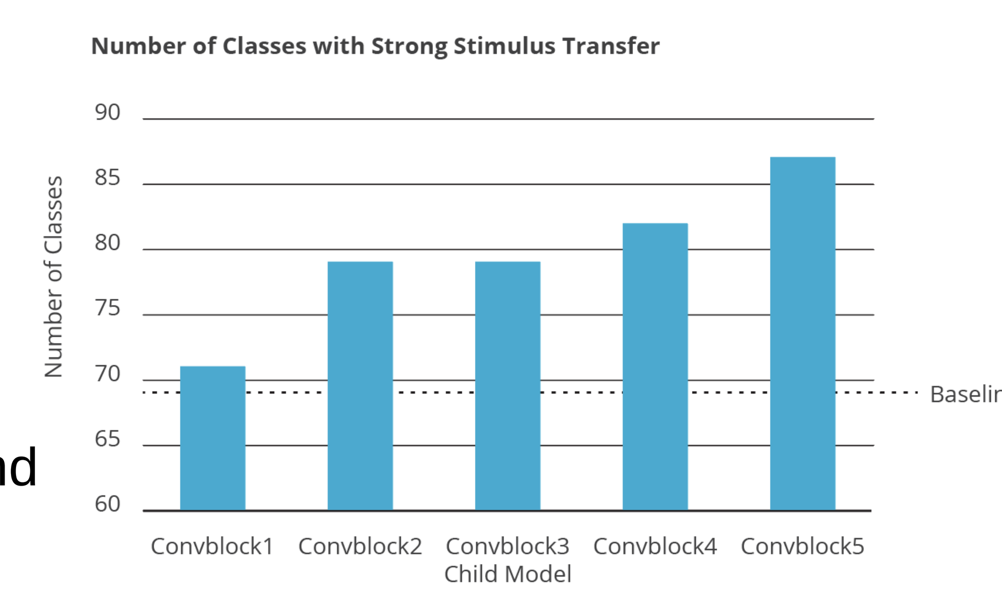


Key Findings

Child Model	Mode of Predicted Labels	Freq.
Convblock1	28: "artichoke"	28
Convblock2	53: "sunflower"	101
Convblock3	53: "sunflower"	101
Convblock4	53: "sunflower"	85
Convblock5	46: "marigold"	81

Our first finding is that most strong foundational model stimuli are mapped to a single label by 4 of the 5 child models we produced. We noted that only one strong stimulus was correctly classified as its original target class by each model. However, each model mapped the strong stimuli to a relatively small set of predicted labels.

Our second finding is that more strong stimuli outperform training centroids when evaluated against models with fewer layers retrained during finetuning. We also computed the number of outperforming strong stimuli against a baseline model trained solely on the Oxford Flowers 102 dataset with randomly initialized weights and found that the number of outperforming strong stimuli was lowest for this baseline model.



Our third finding was that strong stimuli that consistently outperformed training centroids typically targeted classes with less representation in the child dataset. The average number of training points in the target classes for consistent outperforming stimuli was 46.84, while the average number of training points in the remaining target classes was 102.84. These findings indicate that classes with lower representation in the child dataset have a more similar representation to corresponding classes in the foundational dataset and could therefore be more easily targeted in transfer attacks.

Implications for Security Vulnerabilities

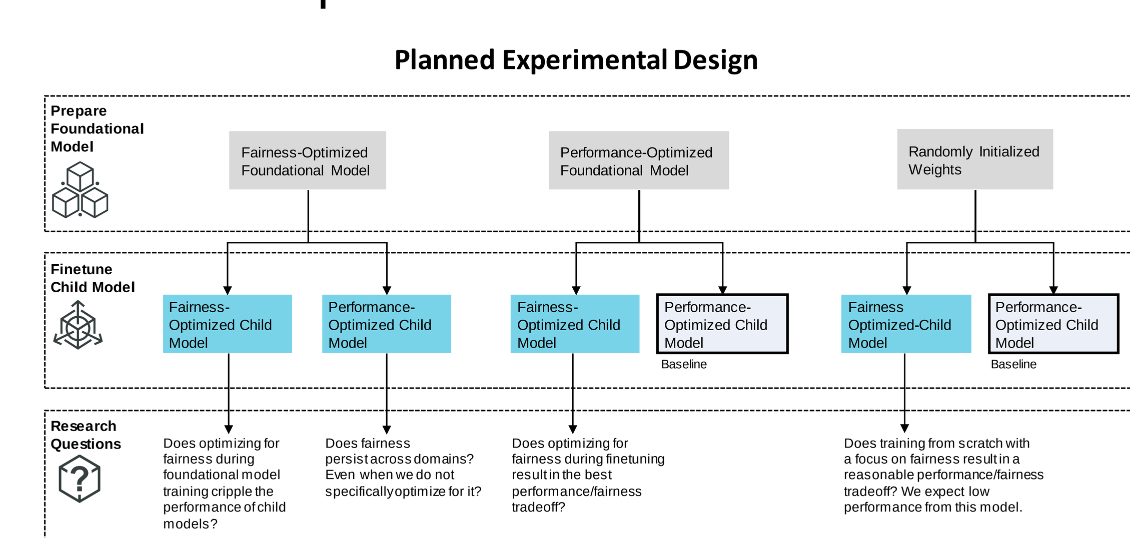
The clear trend that emerges in the number of "outperforming" strong stimuli as more of the child model is frozen indicates that our methods for producing and evaluating the transferability of strong stimuli against child models could be used to reveal information about the foundational model's training data. Specifically, a set of strong foundational model stimuli produced from shadow models meant to emulate foundational models could be evaluated against a finetuned model to determine which shadow model is likely closest to the true foundational model used.

Implications for Fairness

The transfer of stereotypical examples from foundational models to child models finetuned from them at higher rates than the baseline model trained from scratch indicates that biases from the foundational model persist through finetuning.

Impacts & Future Work

Our results establish that bias transfer from foundational models to child models occurs, so we have begun a new line of work focusing on quantifying bias transfer through finetuning. Specifically, we seek to determine whether bias mitigation strategies to minimize model performance disparities between majority and minority groups can persist through finetuning. We also aim to determine where bias mitigation strategies can be applied in model finetuning pipelines bias to produce child models with better fairness-performance tradeoffs.



Results from these experiments will inform best practices for bias-aware model finetuning. We are working with Task Force Lima at the Chief Digital and AI Office (CDAO) to integrate our findings into their Responsible AI Toolkit.

References

1. Ref

Carnegie Mellon University
Software Engineering Institute

Hayden Moore | hmmoore@sei.cmu.edu
Anusha Sinha | asinha@sei.cmu.edu