A systematic study of bias amplification

Presenter: Melissa Hall
Joint work with: Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock

Bias amplification: "when a model exacerbates biases from the training data at test time"

- Result of the algorithm
- Cannot be solely attributed to the dataset

Directional Bias Amplification

- Angelina Wang & Olga Russakovsky, 2021

$\text{BiasAmp}_{\rightarrow}$

$\text{BiasAmp}_{A \rightarrow T}$: amplification of bias resulting from the protected attribute influencing the task prediction.

Experimentation & Results

Experimental set-up: Datasets

FashionMNIST, CIFAR-100, and CIFAR-10

- Convert to binary labels.
- Predict binary-class membership.

Experimental set-up: Group membership, i.e. "protected attribute"

- Image inversion

Experimental set-up: Bias

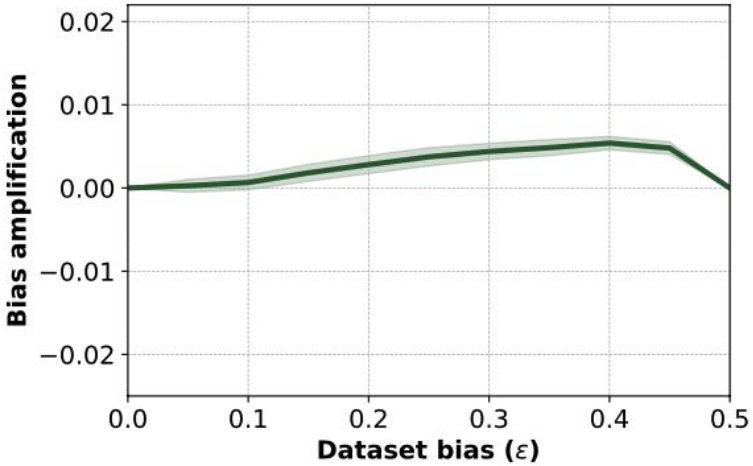
- Inversions are controlled by ϵ in [0, ½].
  - Positively labeled images: ½ - ϵ are inverted
  - Negatively labeled images: ½ + ϵ are inverted
- When ϵ = 0, there is no bias.
- When ϵ = ½, the dataset is fully biased.

Experimental set-up: Model training

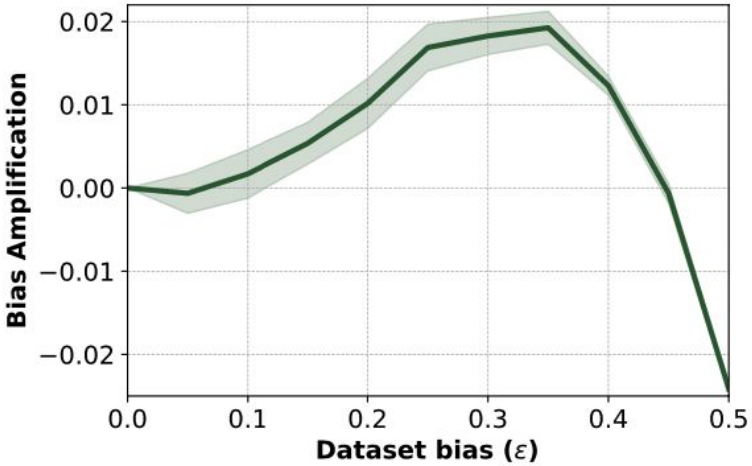
- Architecture: ResNets
- Data augmentation: Cropping, flipping, and resizing
- Training set-up:
  - 500 epochs
  - Multi-step learning rate scheduler with decay by factor of 10 after 250 and 375 epochs.
  - Used warm-up

RQ1: How does bias amplification vary as the bias in the data varies?

RQ1: How does bias amplification vary as the bias in the data varies?



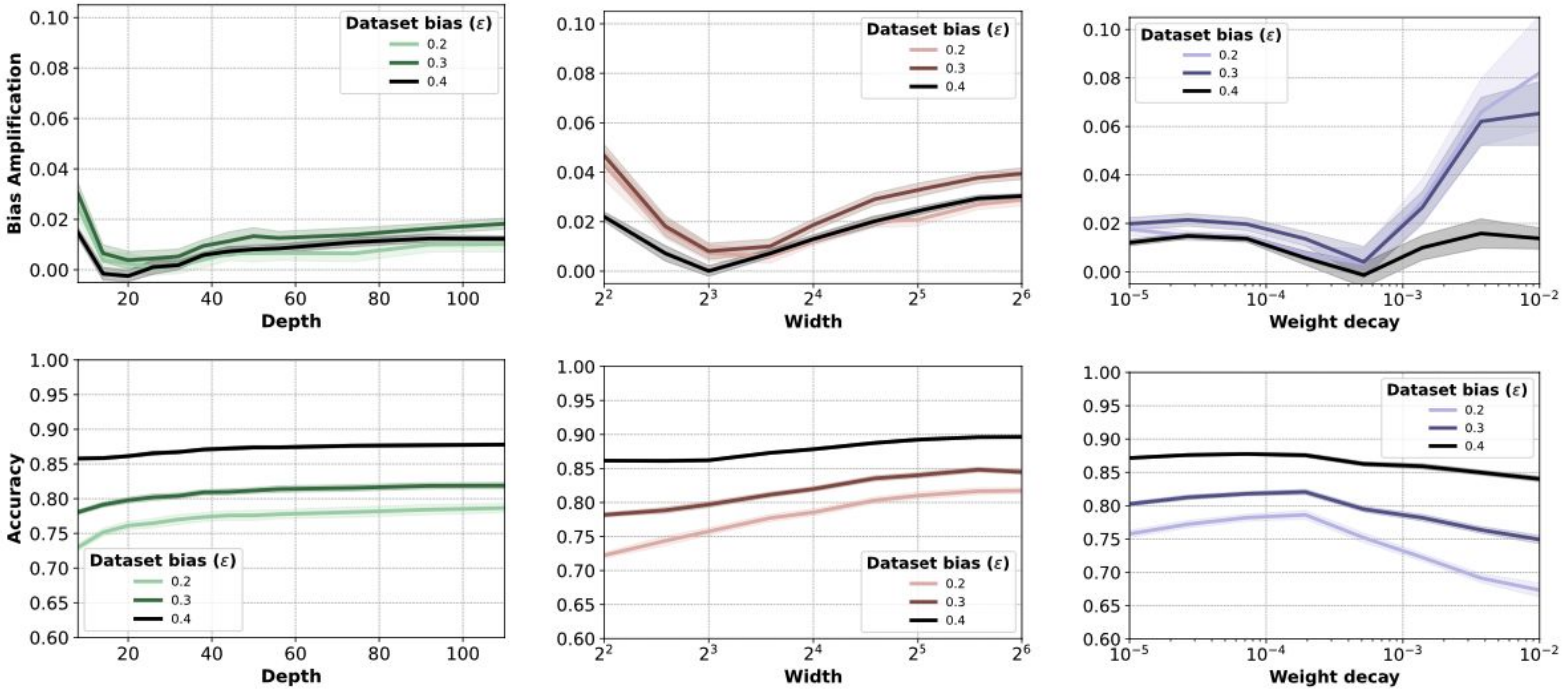
FashionMNIST



CIFAR-100

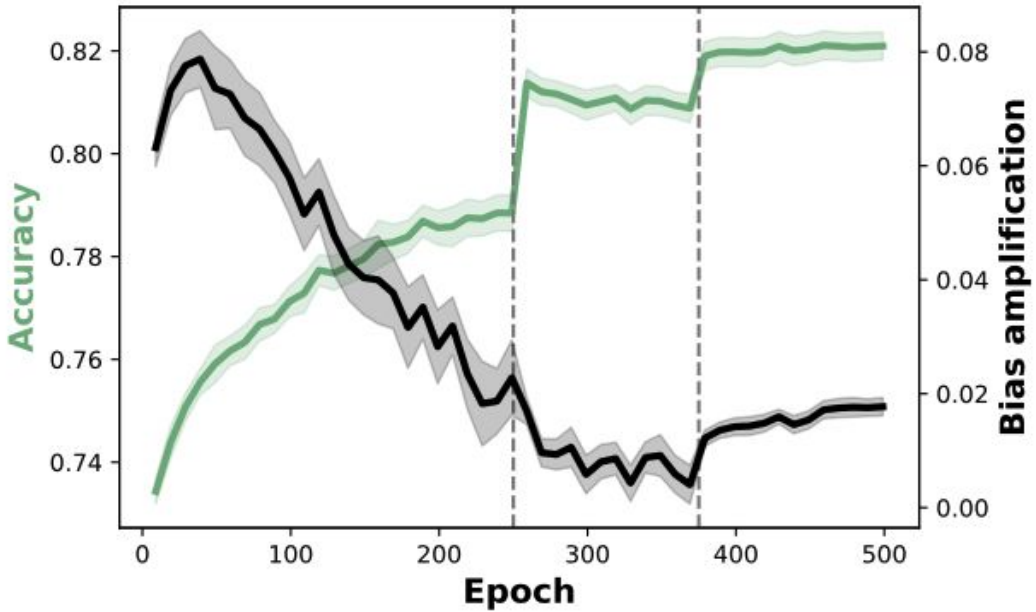
RQ2: How does bias amplification vary as a function of model capacity?

RQ2: How does bias amplification vary as a function of model capacity?



RQ3: How does bias amplification vary during training?

RQ3: How does bias amplification vary during training?



RQ4: How does bias amplification vary as a function of the relative difficulty of recognizing class membership versus recognizing group membership?

Experimental set-up: Bias

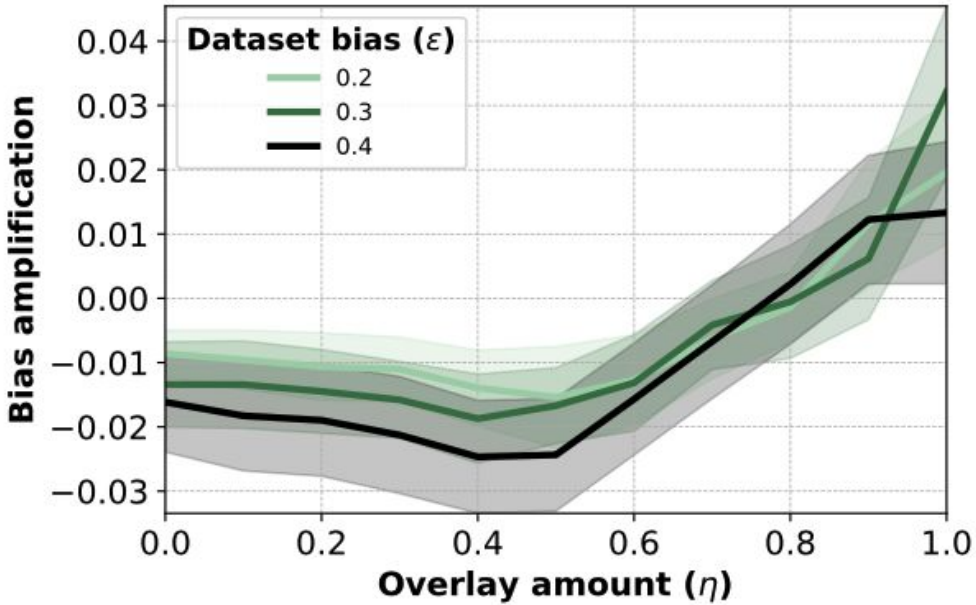
- Rather than inversions, we apply a group-class overlay of amount η to the task-class image.

$$\mathbf{I} = \eta \mathbf{I}_{\text{group}} + (1 - \eta)\mathbf{I}_{\text{class}}$$

Class: "airplane"
Group: "bird"

η=0 η=0.2 η=0.5 η=0.8 η=1.0

RQ4: How does bias amplification vary as a function of the relative difficulty of recognizing class membership versus recognizing group membership?



Closing thoughts:

- Could attempt hyperparameter tuning to reduce bias amplification.
- Need to extend to multi-class settings with real-world biases.

Thank you.