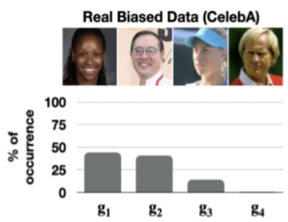
No Training Wheels: Steering Vectors for Bias Correction at Inference Time

Aviral Gupta* Armaan Sethi* Ameesh Sethi

- Transformers often learn biases from uneven datasets
- Most bias correction methods require require training or data generation
- We show that we can use steering vectors to fix bias at inference time using just the training data
- This is a training-free, post-hoc method of bias correction



Waterbirds Dataset

- Waterbirds Dataset: Background bias
- UTKFace, CelebA: Facial attribute bias
- MultiNLI: Negative word bias

Models learn **spurious correlations** due to unbalanced training data → Varied performance in different groups, **low performance in minority groups**

Evaluation Over Various Biased Datasets (Both Vision and Language)

Dataset	Method	Training Required?	Original Dataset: Worst	Original Dataset: Average
Waterbirds	ERM	-	62.46	89.43
	Full Residual Stream (Waterbirds class)	✗	78.19	93.18
	Full Residual Stream (Landbirds class)	✗	83.95	92.49
	Best Single Layer (Waterbirds Class)	✗	75.9	92.3
	Best Single Layer (Landbirds Class)	✗	76.5	94.1
	FFR† (Qraitem et al., 2023)	✓	69.5	84.0
	GDRO† (Sagawa et al., 2019)	✓	91.4	93.5
CelebA	ERM	-	47.8	94.9
	Full Residual Stream (Blond Hair)	✗	62.22	93.47
	Best Single Layer (Blond Hair)	✗	64.84	94.15
	FFR† (Qraitem et al., 2023)	✓	68.9	85.7
	GDRO† (Sagawa et al., 2019)	✓	88.9	92.9
UTKFace	ERM	-	74.3	84.5
	Full Residual Stream(Male)	✗	50.98	74.37
	Full Residual Stream(Female)	✗	47.11	79.16
	Best Single Layer (Male)	✗	79.67	88.20
	Best Single Layer (Female)	✗	76.50	86.09
	FFR† (Qraitem et al., 2023)	✓	67.4	81.4
	GDRO† (Sagawa et al., 2019)	✓	81.6	85.9
MultiNLI	ERM	-	47.8	94.9
	Full Residual Stream (contradiction-negation)	✗	77.67	72.99
	Best Single Layer (contradiction-negation)	✗	69.9	79.7
	FFR† (Qraitem et al., 2023)	✓	-	-
	GDRO† (Sagawa et al., 2019)	✓	77.7	81.4

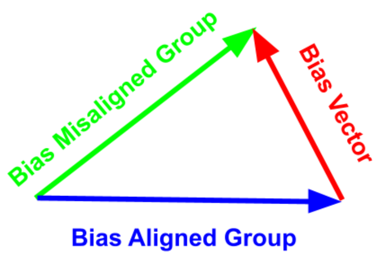
Findings

- Steering Vectors Work for Classification Models as well!!
- Effective for bias mitigation at test-time

What's left ?

- Mechanistic Understanding
- Better Steering Vectors (fine-grained)
- Extending to other OOD problems

Key Idea: Find a "bias vector" direction and delete this from the model activations



1. Calculating **Bias Vector** by subtracting **bias aligned and misaligned** groups

$$\mathbf{X}'[l,t] \leftarrow \mathbf{X}[l,t] - \hat{\mathbf{R}}[l,t]\left(\hat{\mathbf{R}}[l,t]^{\top}\mathbf{X}[l,t]\right)$$

Per Token, Per Layer **(Full Ablation)**

$$\mathbf{x}' \leftarrow \mathbf{x} - \hat{\mathbf{r}}\hat{\mathbf{r}}^{\top}\mathbf{x}.$$

Per Token **(Directional Ablation)**

2. Orthogonalise Activations to Bias Vector