Overparameterisation and worst-case generalisation: friend or foe?

Aditya Krishna Menon

Ankit Singh Rawat

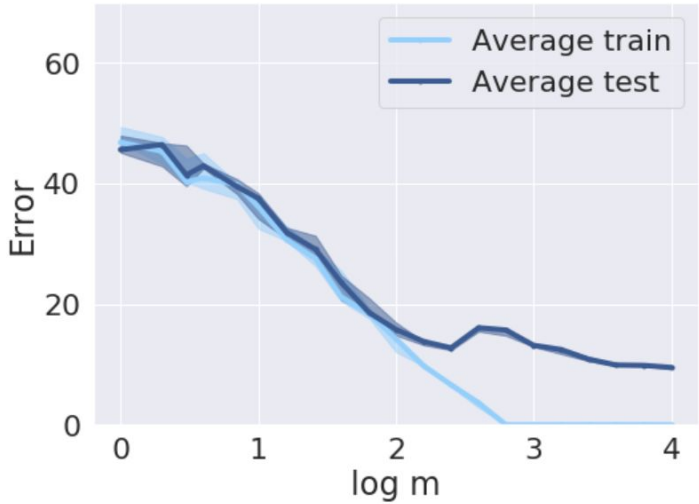
Sanjiv Kumar

Google Research

Overparameterised models

Models with # parameters $\gg$ training samples

✓ Low **average** error on test data



Increasing model complexity

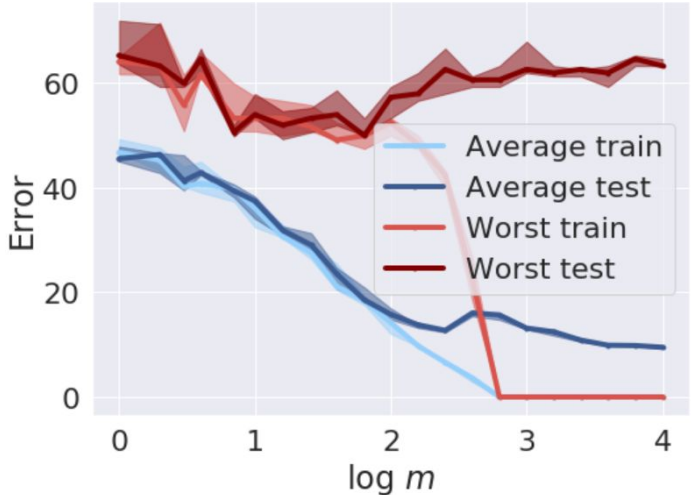
See, e.g., [Sagawa et al. 2020a, b]

Overparameterised models

Models with # parameters $\gg$ training samples

✓ Low **average** error on test data

✗ High **subgroup** error on test data



Increasing model complexity

See, e.g., [Sagawa et al. 2020a, b]

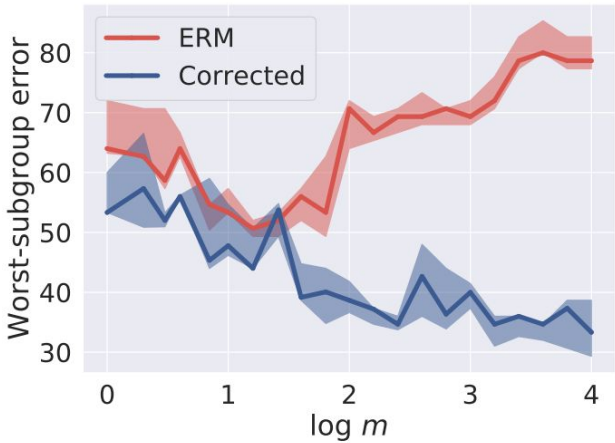
Summary of our work

Q: Does overparameterisation ⇒ poor worst-subgroup error?

Do we have to fundamentally change how we train such models?

A: No! Can improve worst-subgroup error post-hoc, by:

(1) Retraining the classification layer on learned embeddings
(2) Modifying the decision threshold for classification



Classification with subgroups: setup

Setting: classification over examples (*x*, *y*)

each example has group membership *g*(*x*, *y*)

Average error:

Ignores high error on rare group

$$L_{\text{avg}}(h) \doteq \mathop{\mathbb{E}}_{g} \mathop{\mathbb{E}}_{x,y|g} [\ell_{01}(y, h(x))]$$

Worst-group error:

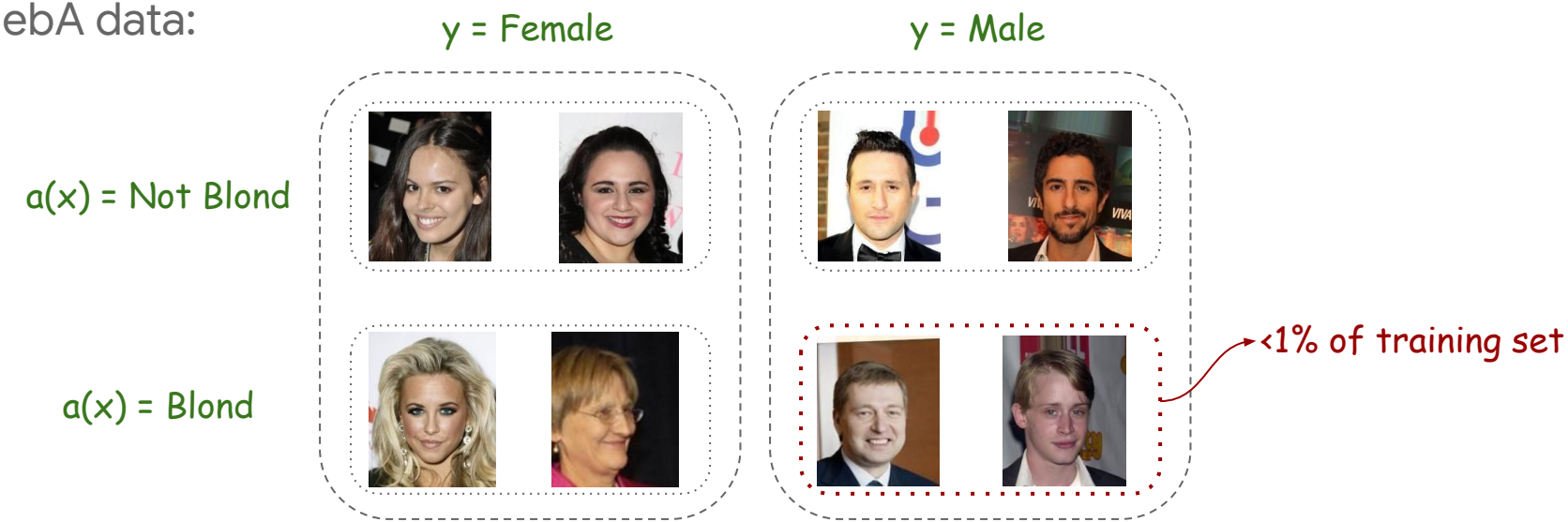
Explicitly focusses on rare groups

$$L_{\text{max}}(h) \doteq \max_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{x,y|g} [\ell_{01}(y, h(x))]$$

Classification with subgroups: examples

Suppose $g(x, y) = (a(x), y)$ for attribute $a(x)$

CelebA data:

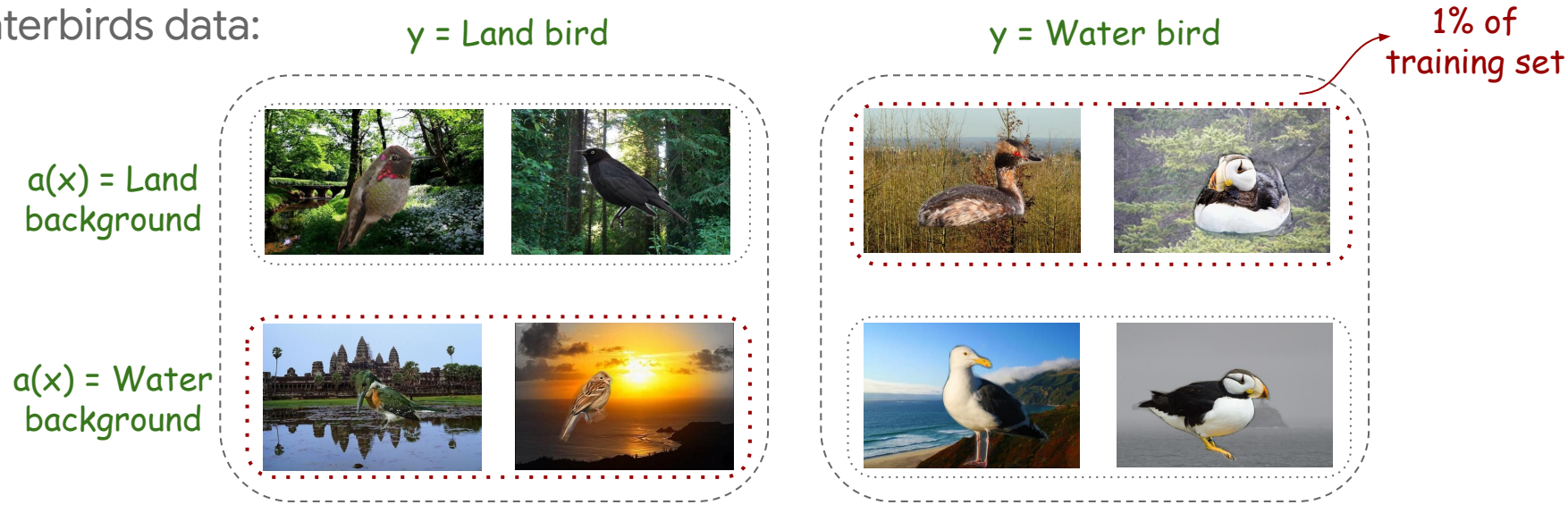


See, e.g., [Sagawa et al. 2020a, b]

Classification with subgroups: examples

Suppose $g(x, y) = (a(x), y)$ for attribute $a(x)$

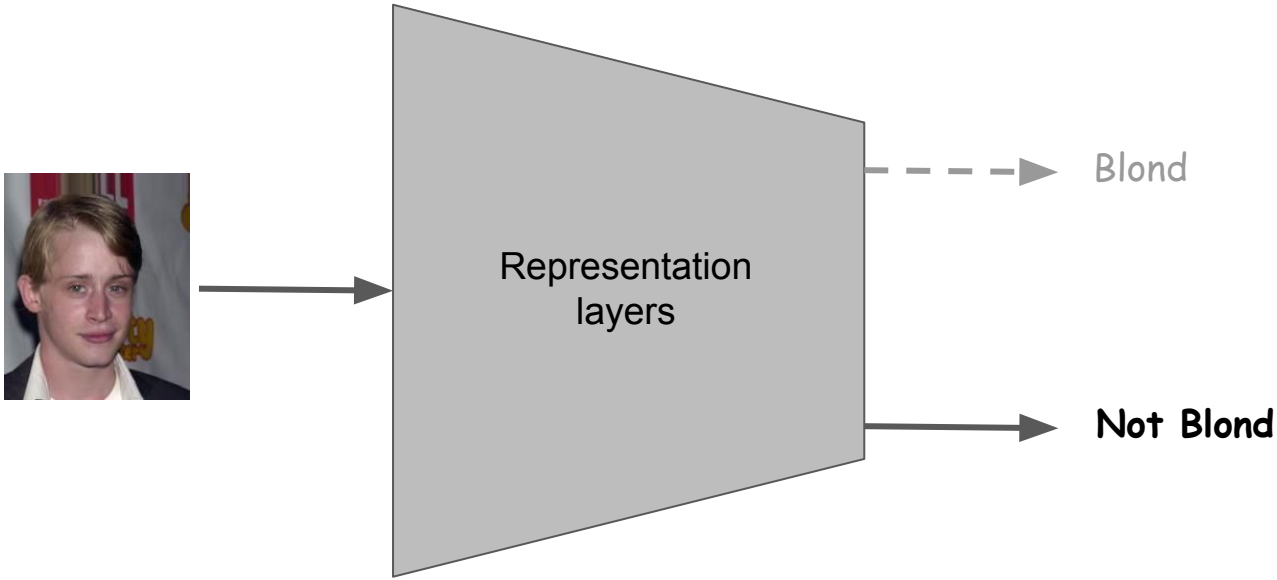
Waterbirds data:



See, e.g., [Sagawa et al. 2020a, b]

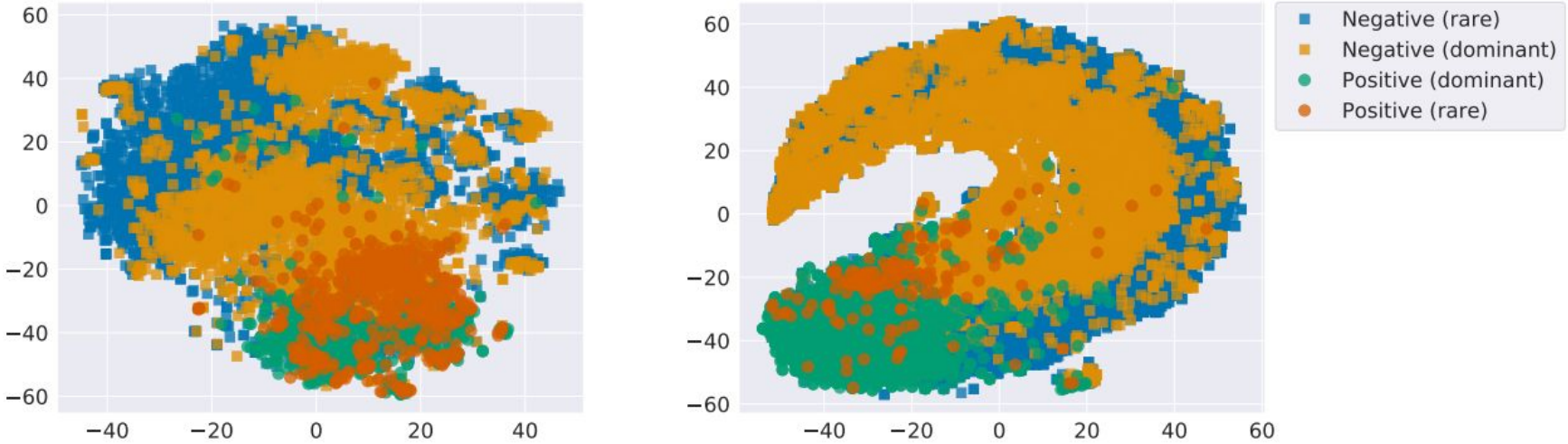
Dissecting the dominant subgroup bias

Is the bias in the representation layers, or the classification layer?



Can the learned representations distinguish subgroups?

From tSNE visualisation, subgroups cluster together reasonably well



(a) waterbirds. (b) celebA.

Proposal: classifier retraining

Simple procedure to improve worst-subgroup error:

(1) learn classifier by standard ERM

This will have poor worst-subgroup error

(2) freeze the input embeddings, and re-train the classifier only

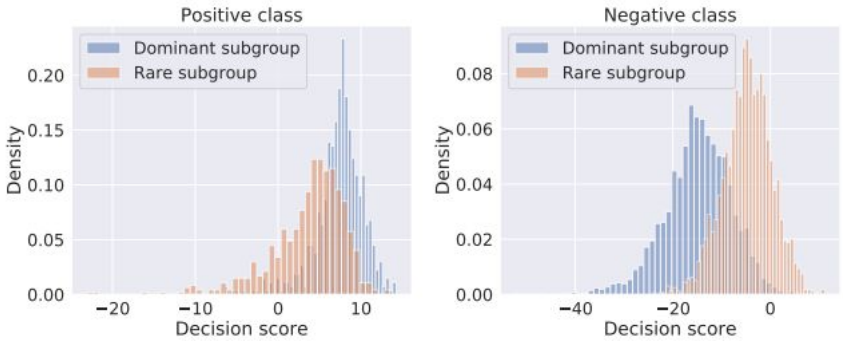
We can, e.g., sub-sample the dominant subgroups [Sagawa et al., '20b]

If successful, ERM learns good representations

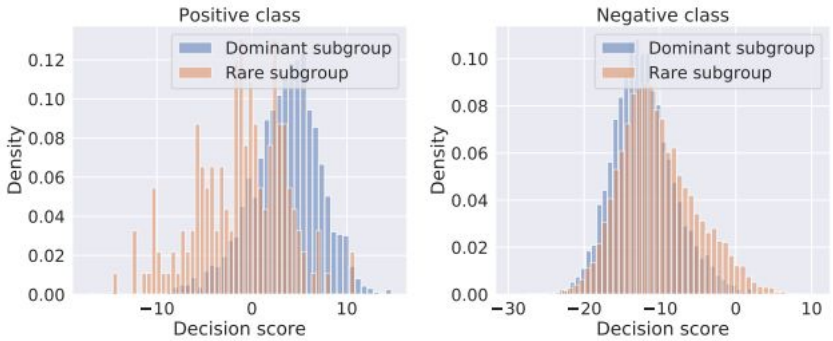
See also classifier retraining in long-tail literature, e.g., [Kang et al. 2020].

Can the learned logits distinguish subgroups?

Logit distributions show a consistent shift amongst subgroups:



(a) `waterbirds`.



(b) `celebA`.

Proposal: threshold correction

Standard classification rule:

$$h(x) = +1 \iff f_{+1}(x) - f_{-1}(x) > 0$$

Adjusted classification rule:

$$h(x) = +1 \iff f_{+1}(x) - f_{-1}(x) > t_{a(x)},$$

where $t_{a(x)}$ is a group-dependent threshold

- If successful, ERM produces good logits
- Can choose $t_{a(x)}$ on a validation set

Experiments: post-hoc versus training modification

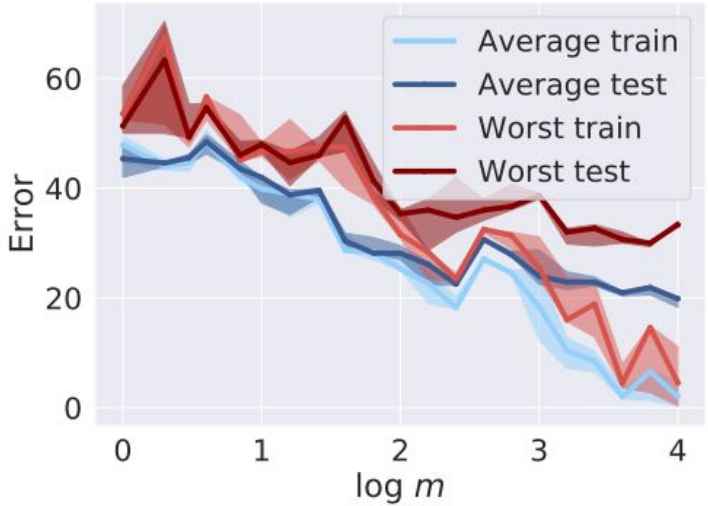
Post-hoc correction is competitive with training modification

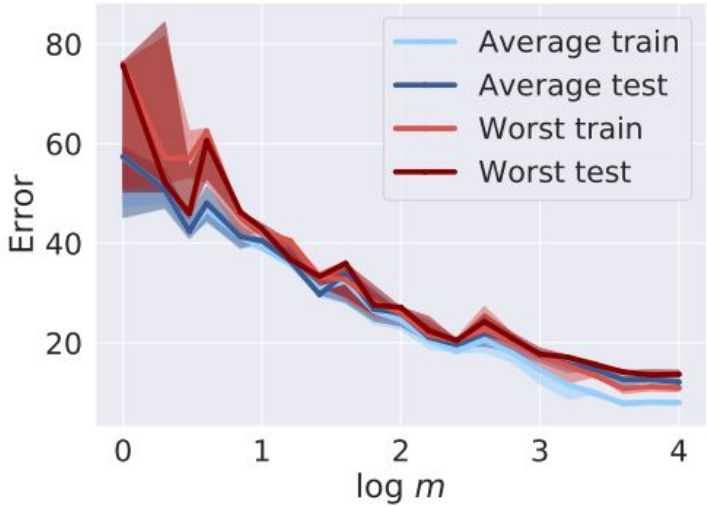
ERM produces reasonable representations and logits!

Dataset	Model	Average error					Worst-subgroup error				
		ERM	SAM	DRO	CRT	THR	ERM	SAM	DRO	CRT	THR
synth	Logistic	9.49	23.99	17.99	N/A	22.18	63.20	27.38	22.67	N/A	27.51
waterbirds	Logistic	4.37	17.18	10.57	N/A	16.71	52.34	20.82	16.73	N/A	17.46
	ResNet	2.15	15.66	6.30†	2.58	5.23	17.23	18.41	9.50†	8.30	8.70
celebA	ResNet	4.64	9.28	6.60†	9.38	7.90	56.11	10.63	12.20†	14.68	12.96

Experiments: overparameterisation and worst-case error

With correction, increasing model complexity can improve worst-subgroup error!





Summary of our work

Q: Does overparameterisation ⇒ poor worst-case error?

Do we have to fundamentally change how we train such models?

A: No! One can improve worst-case error post-hoc, by:

(1) Retraining the classification layer on learned embeddings
(2) Modifying the decision threshold for classification