The advantages of multiple classes for reducing overfitting from test set reuse

Vitaly Feldman

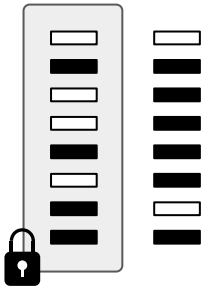
Google Research
Brain team

Roy Frostig

Google Research
Brain team

Moritz Hardt

UC Berkeley

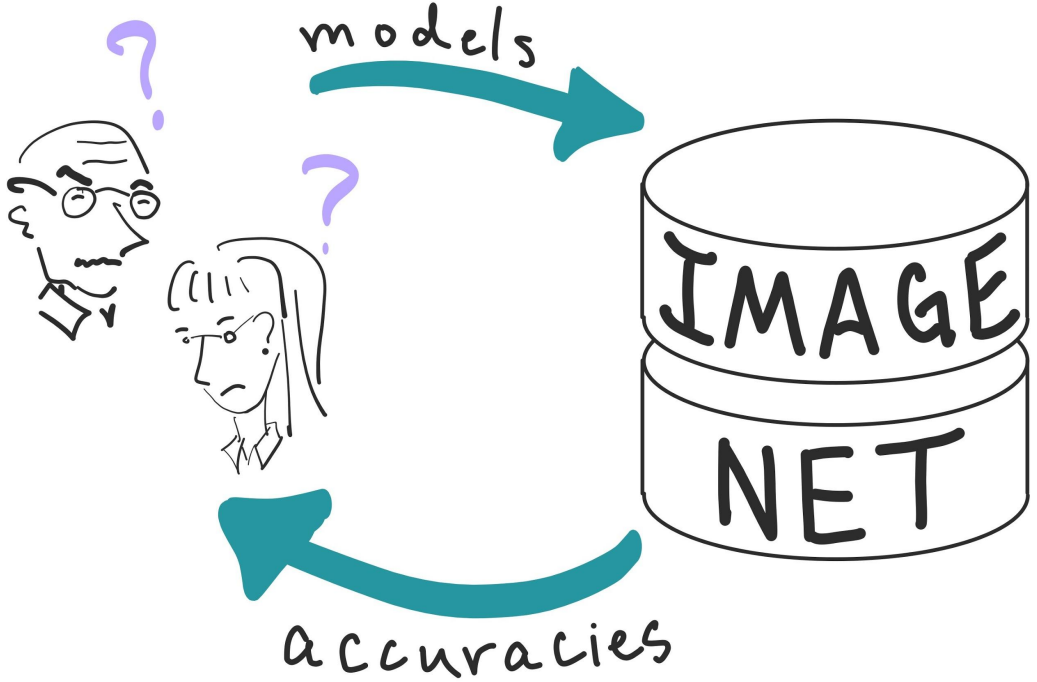


Test data is reused. Are results still valid?

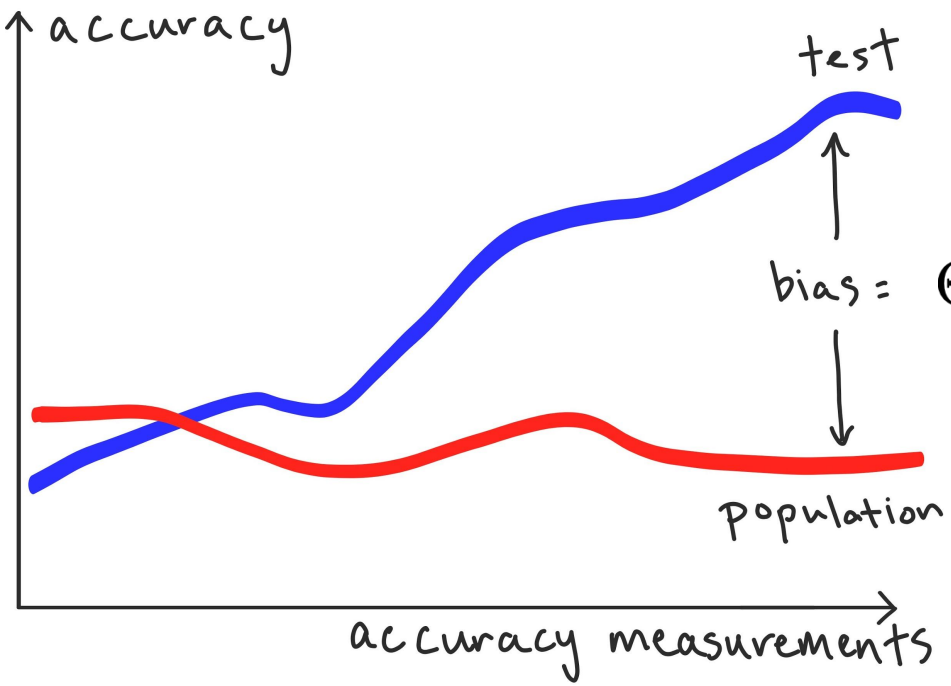
test data reuse

↓

potential overfitting



How much bias is caused by reuse?



measurements

$$\Theta\left(\sqrt{\frac{k}{n}}\right) ?$$

test set size

Meanwhile: not much overfitting on CIFAR/ImageNet/MNIST [RRSS'18, YB'19]

Main result: class multiplicity mitigates bias

Theorem: for $k < n/m$, with n examples, m classes, k accuracy queries

$$\text{bias} \;\leq\; \tilde{O}\left(\sqrt{\frac{k}{n\boldsymbol{m}}}\right)$$

where *bias* =

$$\underbrace{\frac{1}{n}\sum_{(x,y)\in S}\mathbf{1}[f(x)=y]}_{\text{test set accuracy}} \;-\; \underbrace{\Pr_{(x,y)\sim\mathcal{P}}[f(x)=y]}_{\text{population accuracy}}$$

Main result: class multiplicity mitigates bias

Theorem: for $k < n/m$, with n examples, m classes, k accuracy queries

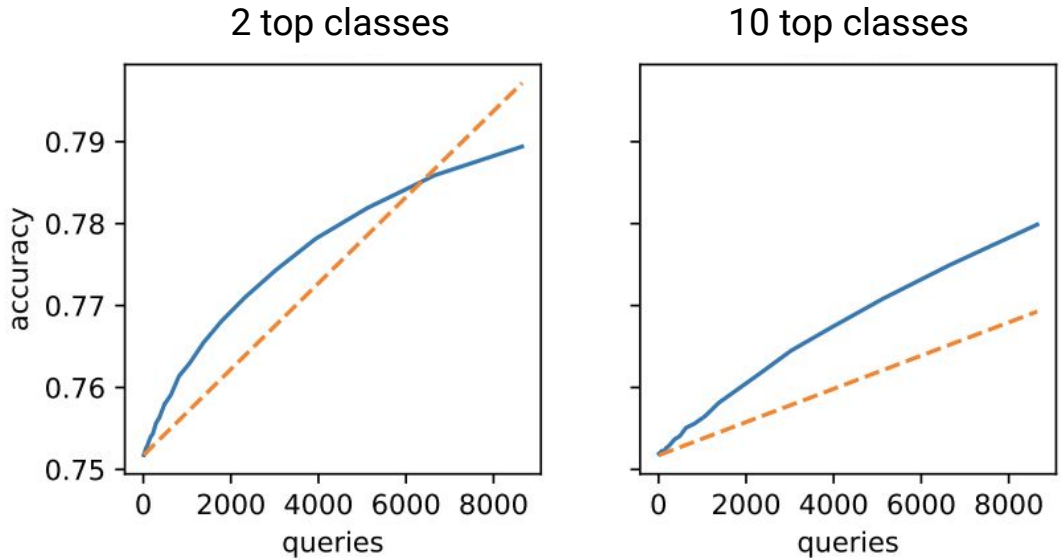
$$\tilde{\Omega}\left(\sqrt{\frac{k}{nm^2}}\right) \leq \text{bias} \leq \tilde{O}\left(\sqrt{\frac{k}{nm}}\right)$$

Lower bound by an *overfitting attack* that is:

- Computationally efficient
- Optimal among *point-wise* attacks
- Can incorporate *priors*

Attacking the ImageNet test set

- Scale: 50K points over 1K labels
- Prior: ResNet-50v2
- Overfitting is possible, e.g. 3% bias with ~5K queries



Also…

- The many-query regime, $k > n/m$
  - A recovery-based attack
  - A *matching* upper bound
- More experiments!