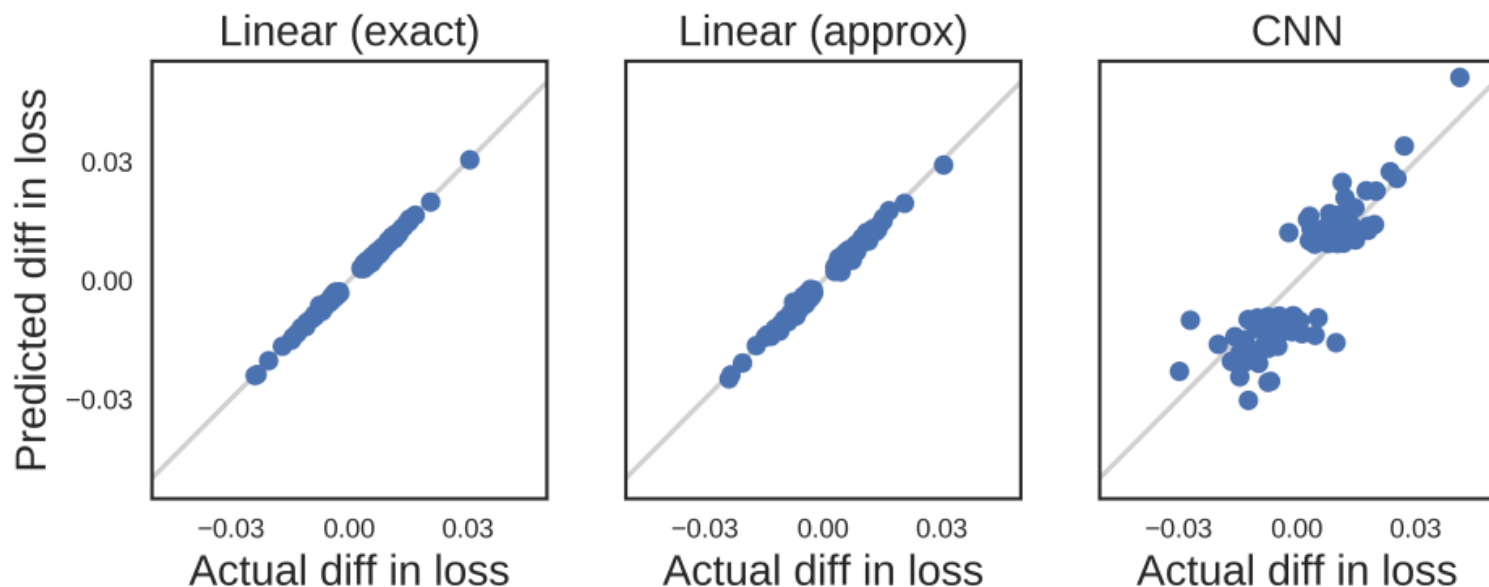


Scientific discovery

Eric Wong

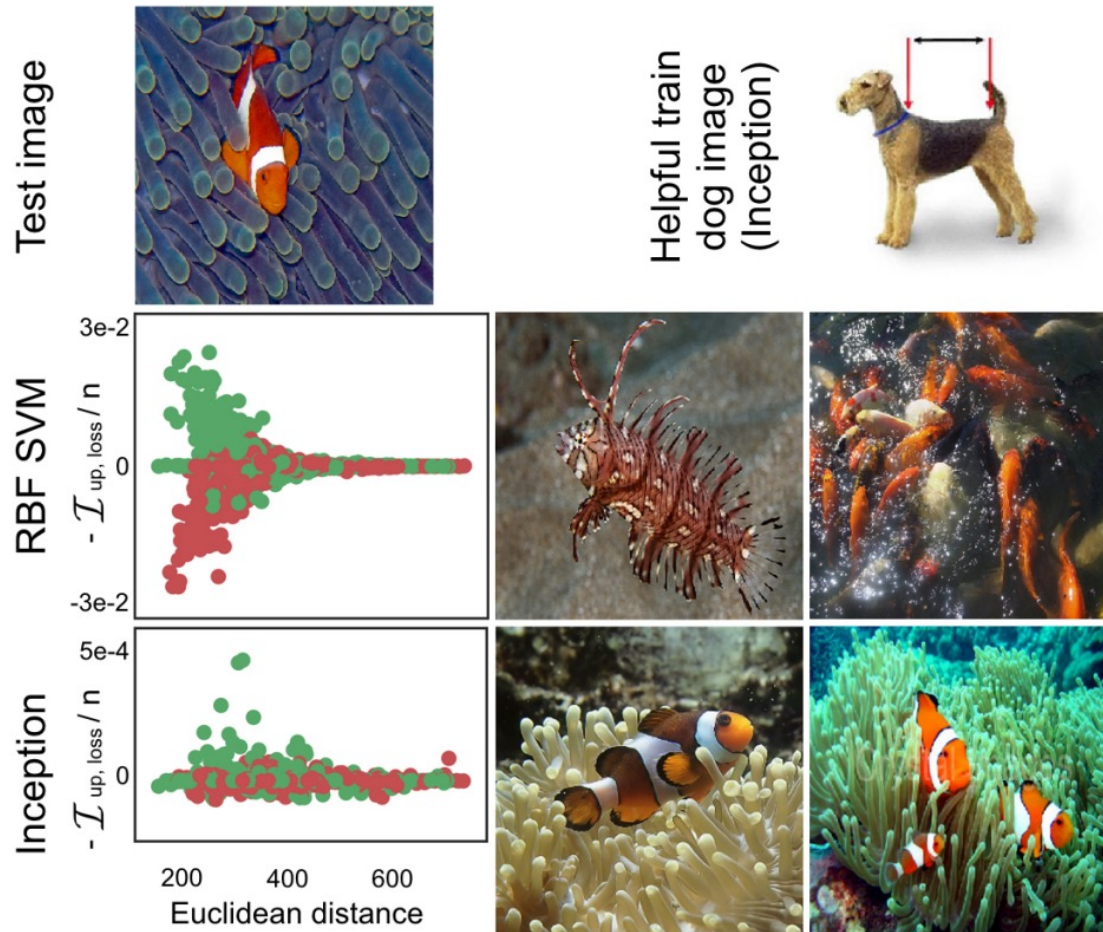
10/20/2022

Influence functions approximate deletion



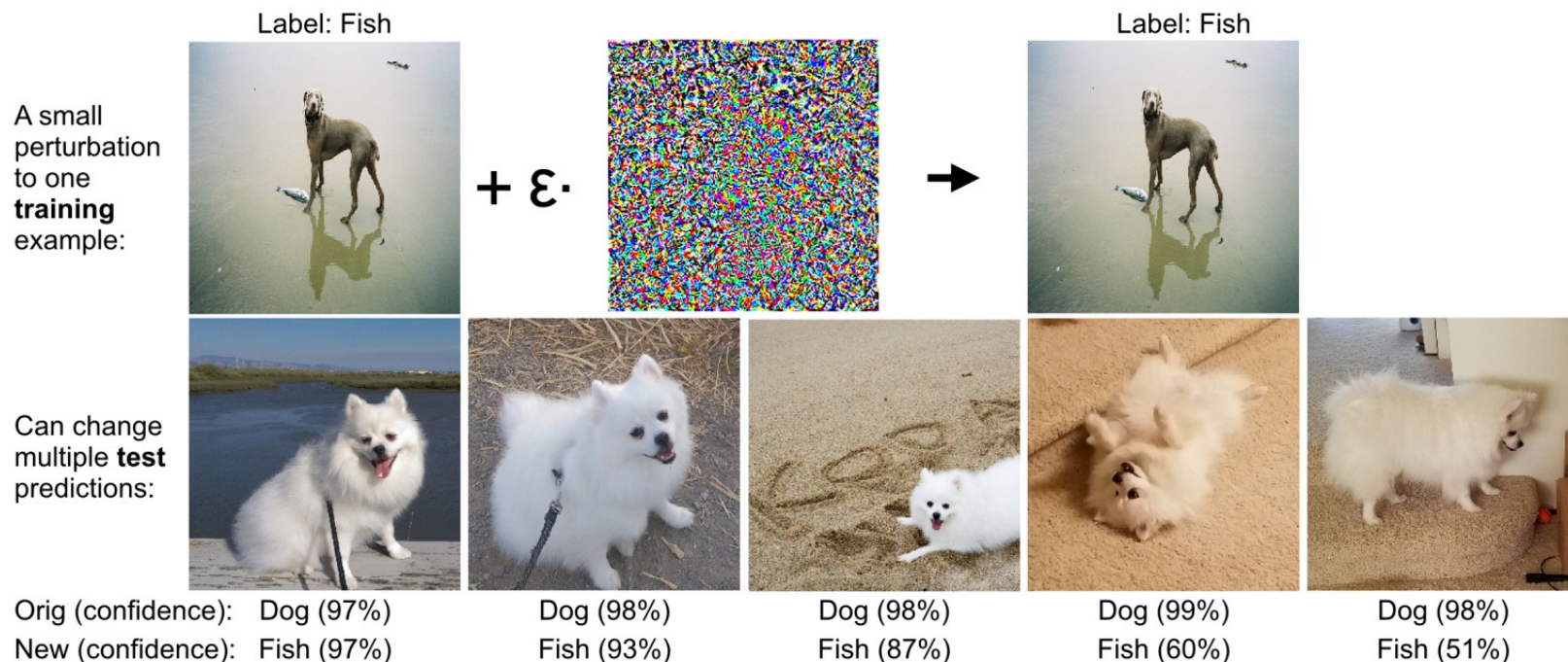
Pang Wei Koh, Percy Liang "Understanding Black Box Predictions via Influence Functions"

Influential examples

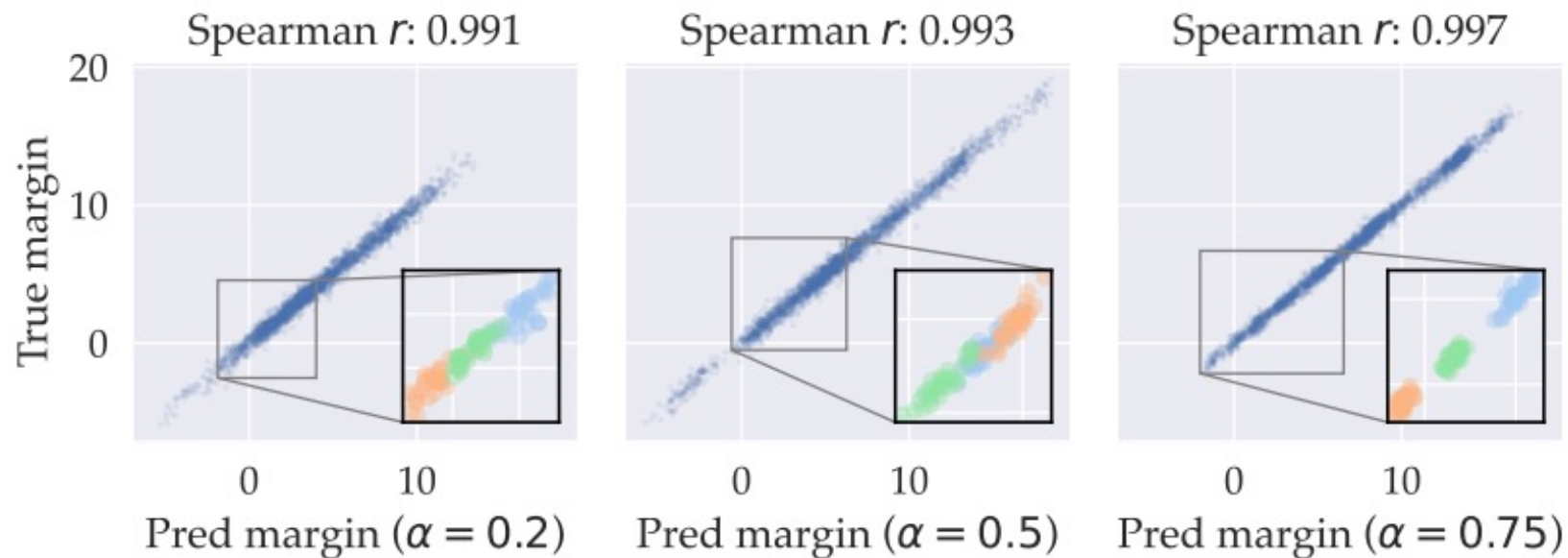


Pang Wei Koh, Percy Liang "Understanding Black Box Predictions via Influence Functions"

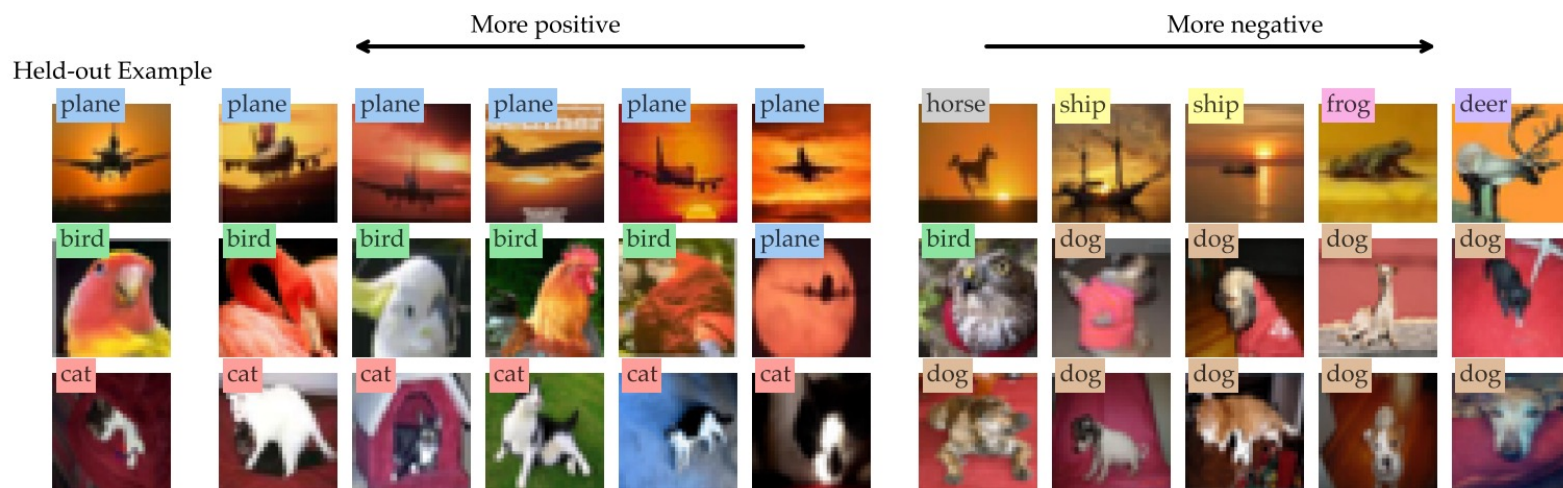
Attack influential examples



Linear datamodels are enough



Datamodels show similar images

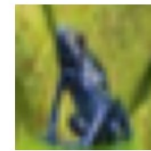
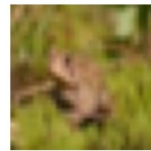


Effect of subset size

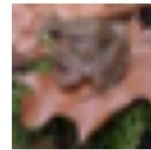
Held-out Example



α : 10%

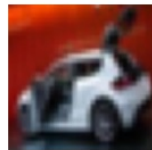


α : 50%

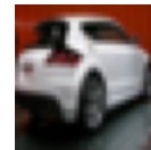
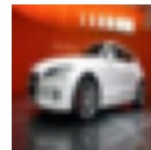
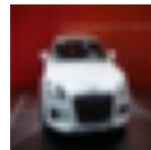
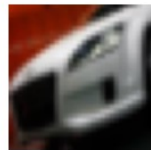


Train Examples by Weight

Held-out Example



α : 10%

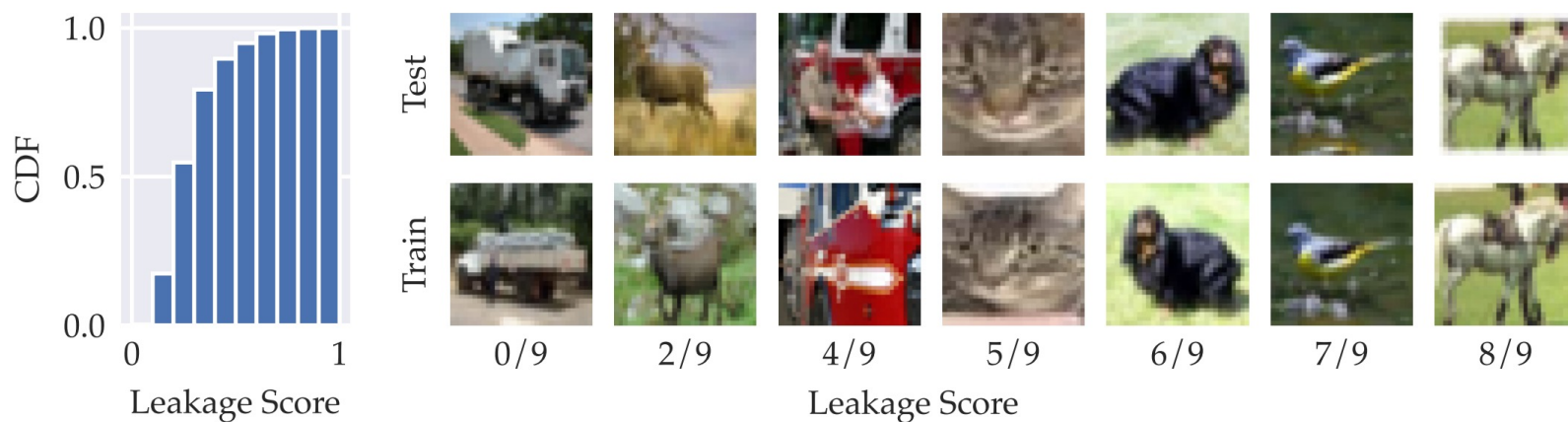


α : 50%



Train Examples by Weight

Data leakage



Clustering datamodel weights



Similar for transfer learning

Most Positively Influenced

ImageNet
Images



speedboat



tailed frog



warplane

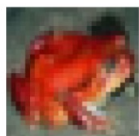


racer

CIFAR-10
Images



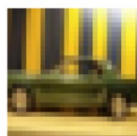
ship



frog

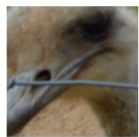


airplane



automobile

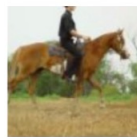
ImageNet
Images



ostrich



warplane



sorrel horse

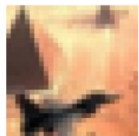


moving van

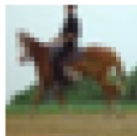
CIFAR-10
Images



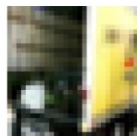
bird



airplane

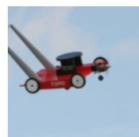


horse



truck

Most Negatively Influenced



lawnmower



minivan



wing



book jacket



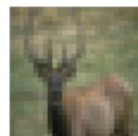
airplane



airplane



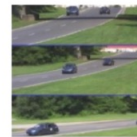
ship



deer



warplane



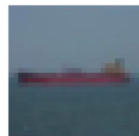
beach wagon



warplane



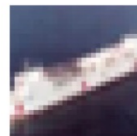
moving van



ship



airplane



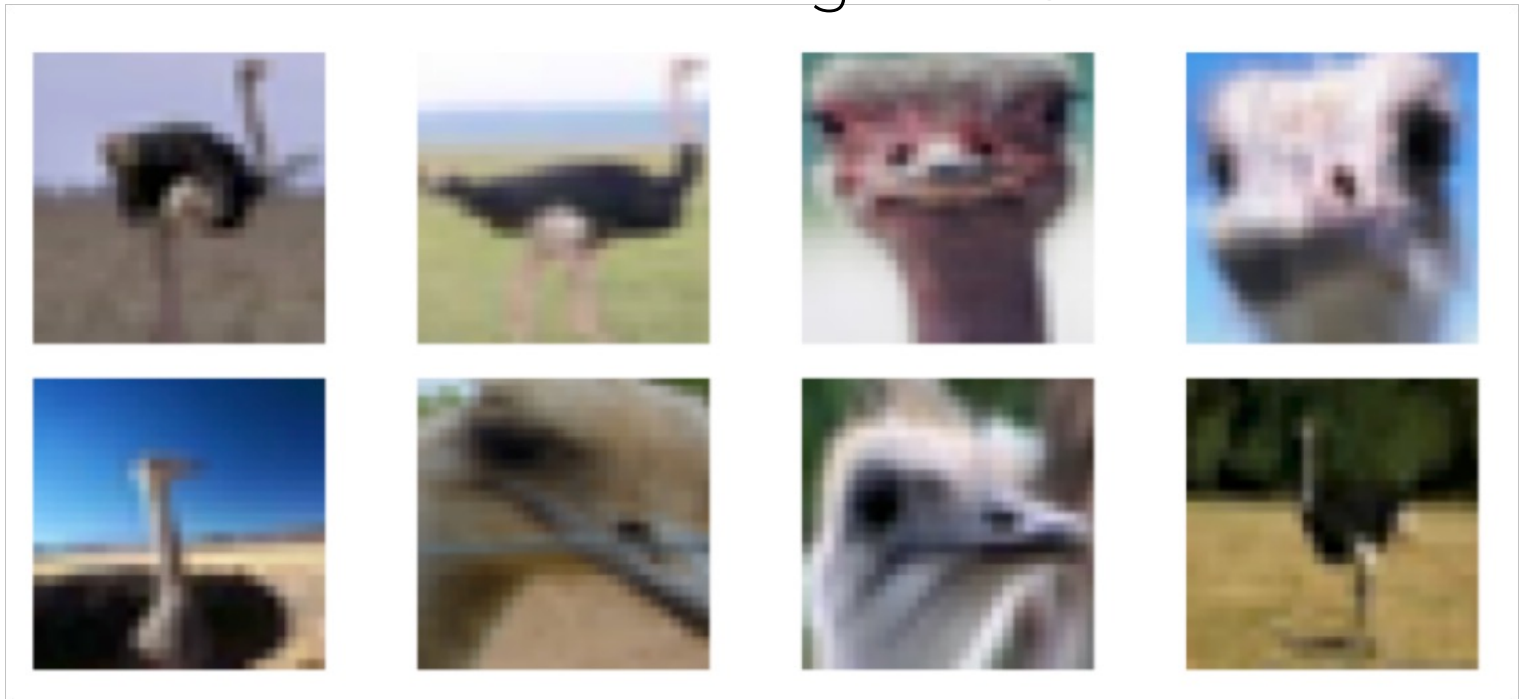
ship



automobile

Subpopulations in transfer

CIFAR10 datapoints with high influence from ImageNet Ostriches



Biases and Correlations

Eric Wong
10/25/2022

Project checkpoint report

- Expanded version of the proposal
- Complete previous section (intro, related, proposed)
- Current progress (preliminary experiments or theory, current results, planned work)

5m presentation

- Recommend <5 slides (i.e. 3)
- Problem/motivation (majority)
- Plan/progress

Prompting is expensive

Question: If x is 2 and y is 5, what is $x + 2y$?

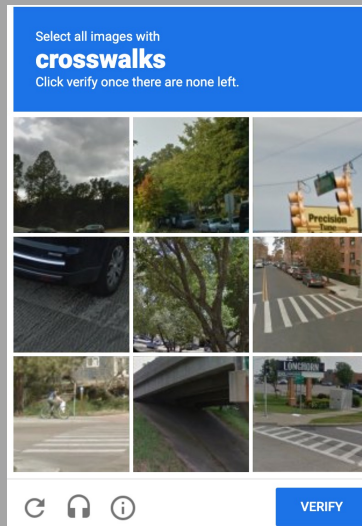
Answer: $x + 2y = 2 + 2(5) = 2 + 10 = 12$

Question: If x is 8 and y is 9, what is $3x + y$?

Answer: $3x + y = 3(8) + 9 = 24 + 9 = 33$

Question: If x is 7 and y is 6, what is $x + 4y$?

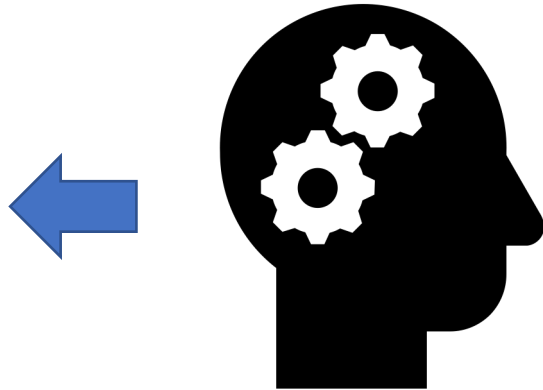
Answer:



Response Length: 64

Temperature: 0.7

Top-p: 0.5



Question: If x is 7 and y is 6, what is $x + 4y$?

Answer: $x + 4y = 7 + 4(6) = 7 + 24 = 31$

Question: If x is 3 and y is 5, what is $3x + y + z$?

Answer: $3x + 5y + z = 3(3) + 5(5) + z = 15 +$

Idea: replace trial & error with automated testing

Question: If x is 2 and y is 5, what is $x + 2y$?

Answer: $x + 2y = 2 + 2(5) = 2 + 10 = 12$

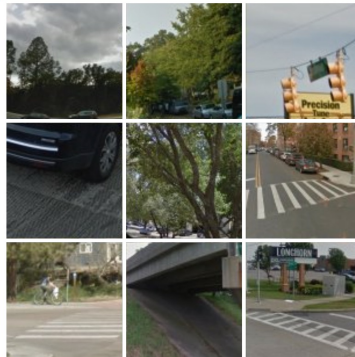
Question: If x is 8 and y is 9, what is $3x + y$?

Answer: $3x + y = 3(8) + 9 = 24 + 9 = 33$

Question: If x is 7 and y is 6, what is $x + 4y$?

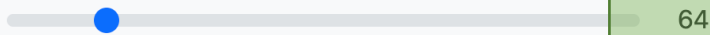
Answer:

Select all images with
crosswalks
Click verify once there are none left.



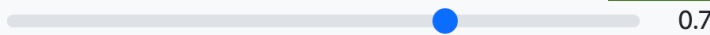
VERIFY

Response Length:



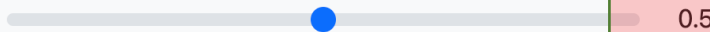
64

Temperature:



0.7

Top-p:



0.5

Influence
functions

Question: If x is 7 and y is 6, what is $x + 4y$?

Answer: $x + 4y = 7 + 4(6) = 7 + 24 = 31$

Question: If x is 3 and y is 5, what is $3x + y + z$?

Answer: $3x + 5y + z = 3(3) + 5(5) + z = 15 +$

Current progress

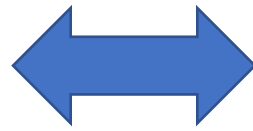
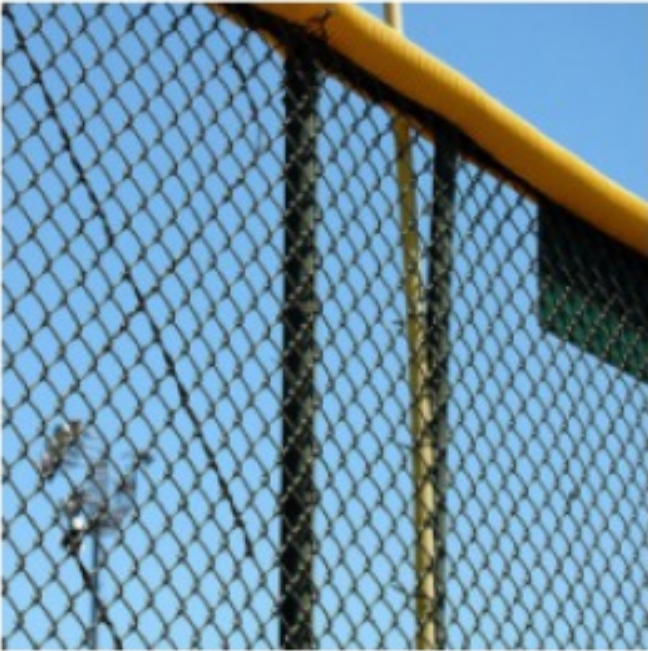
Waiting for machines to come online



OPT models are not well documented

Correlations

ImageNet Correlations



Wong et al. 2021 "Leveraging Sparse Linear Layers for Debuggable Deep Networks"

Mturker validation

Do you see a common pattern in these images?

You will be shown images belonging to two object categories: "**marimba/xylophone**" and "**ice lolly/lolly**". Your task is to inspect the images, judge whether you can **see a prominent common pattern** between all these images, and then answer the questions below.

Inspect the following images

marimba/xylophone



ice lolly/lolly



Is there a shared pattern? Is pattern
part of xylophone or ice lolly?

Wong et al. 2021 "Leveraging Sparse Linear Layers for Debuggable Deep Networks"

Spurious correlations

Pattern descriptions
(via MTurk)

Class pairs

"bullet train"

"greenhouse"

havevent
allglass



spurious

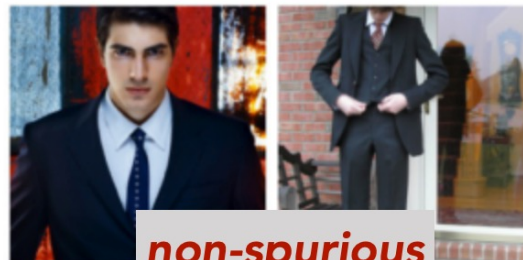


non-spurious

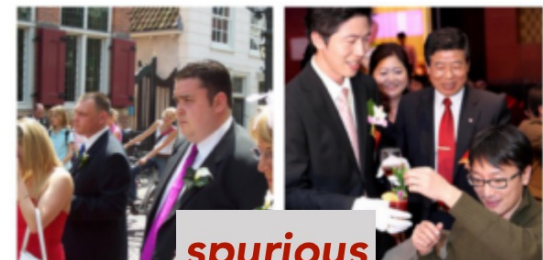
"suit"

"groom"

formal
wearing coat
everyone



non-spurious



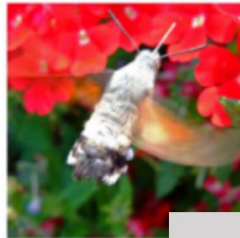
spurious

Spurious correlations

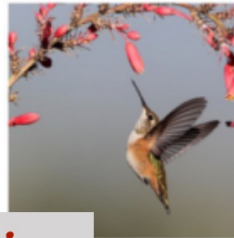
Pattern descriptions
(via MTurk)

all the color.
red in
color object

"hummingbird"



spurious



Class pairs

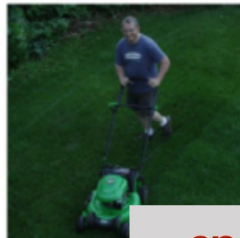
"rose hip"



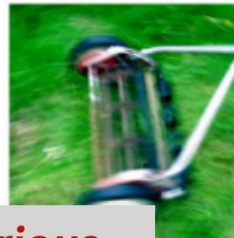
non-spurious



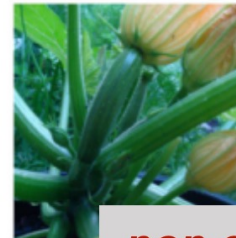
"lawn mower"



spurious



"zucchini/courgette"



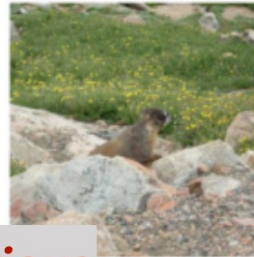
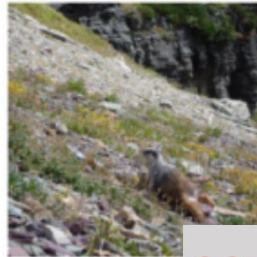
non-spurious



Spurious correlations

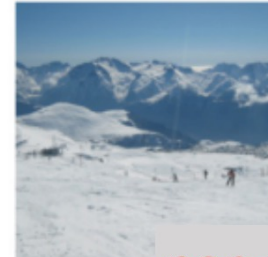
both
found
mountain
in^{high}

"marmot"



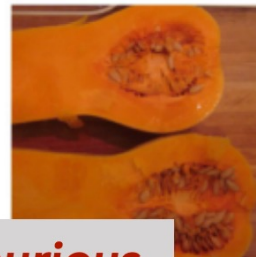
spurious

"alp"



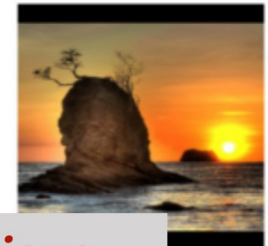
non-spurious

"butternut squash"



non-spurious

"seashore/coast"



spurious

orange
color

Testing causality

Samples

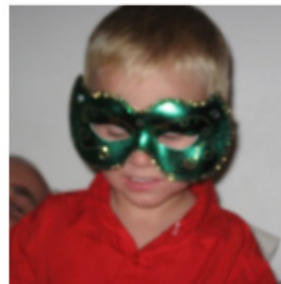
"basketball"



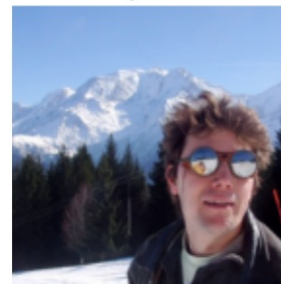
"racket"



"mask"



"sunglasses"



+ "chainlink fence"

Counterfactuals

"ballplayer"

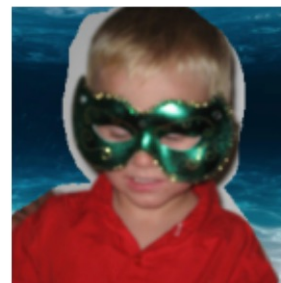


"ballplayer"

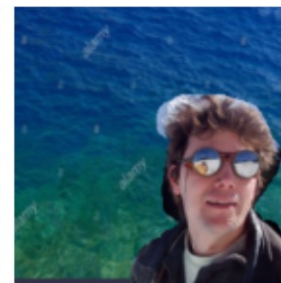


+ "water"

"snorkel"



"snorkel"



Toxic comment classification

“Jeez Ed, you seem like a [REDACTED]
[REDACTED] [REDACTED]” → Toxic

“Barack Obama is the president”
→ Non-toxic

Identity bias in toxicity detection

“Barack Obama is the president”

→ Non-toxic ✓

“Barack Obama is black” → Toxic ✗

“<NAME> is black” → Toxic ✗

Sparse linear layers help to expose biases

Standard layer



7% identity terms

Sparse layer



27% are identity terms

Testing causality

“Jeez Ed, you seem like a [REDACTED] [REDACTED]
[REDACTED]” → Toxic ✓

“Jeez Ed, you seem like a [REDACTED] [REDACTED]
[REDACTED] Christianity” → Non-Toxic ✗

Adding a biased term (i.e. Christianity) flips prediction 74% of the time