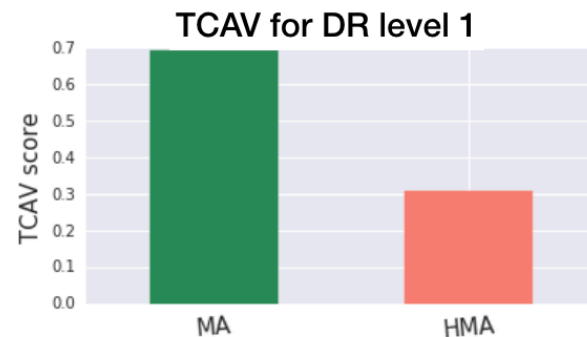




Interpretability beyond feature attribution: Testing with Concept Activation Vectors TCAV

Been Kim

Work with Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler,
Fernanda Viegas, Rory Sayres @ Brain



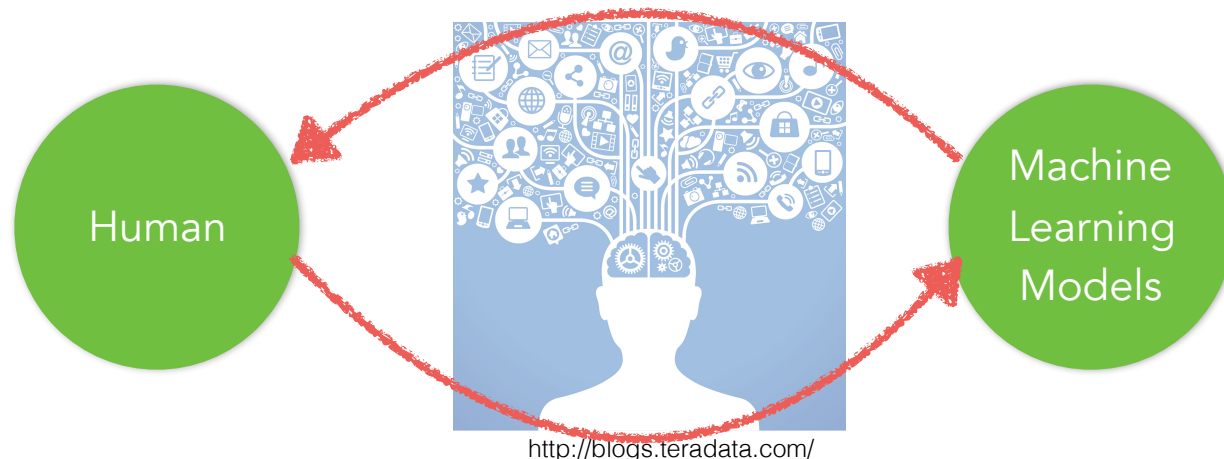
Research agenda:

interpretability

To use machine learning **responsibly**

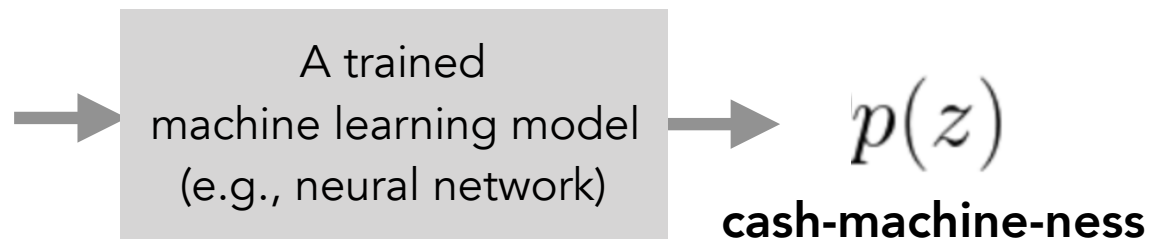
we need to ensure that

1. our **values** are aligned
2. our **knowledge** is reflected

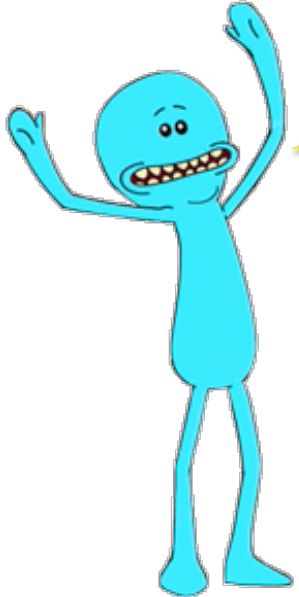


Problem:

Post-training explanation

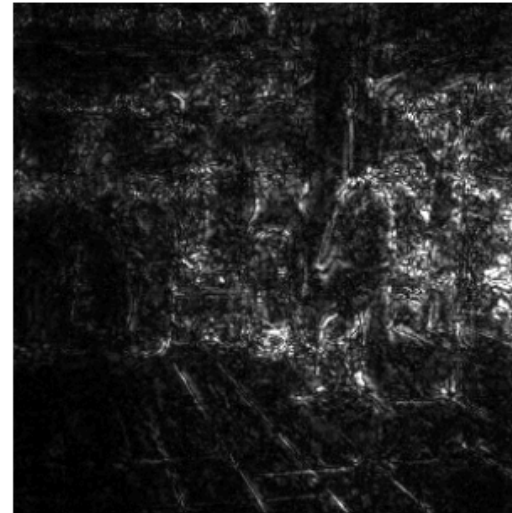


Why was this a cash machine?



halofanon.wikia.com

Caaaan do! we've got saliency maps to measure importance of each pixel!



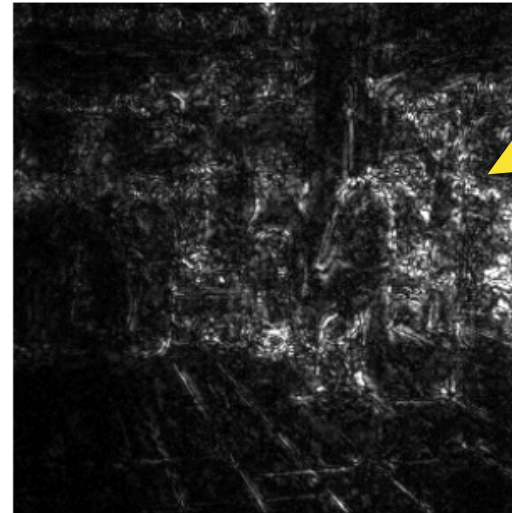
a logit $\rightarrow \frac{\partial p(z)}{\partial x_{i,j}}$
pixel $i,j \rightarrow$

One of the most popular interpretability methods for images:

Saliency maps

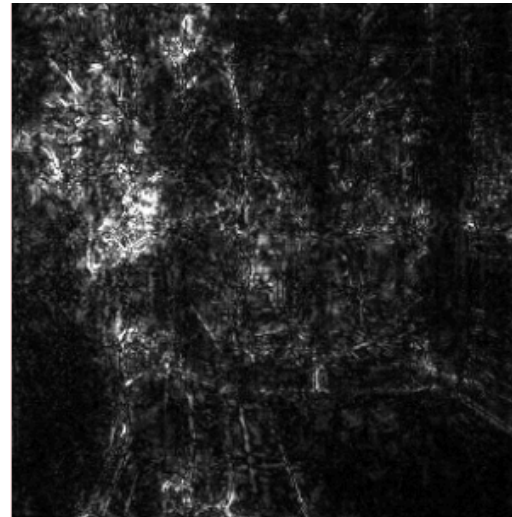
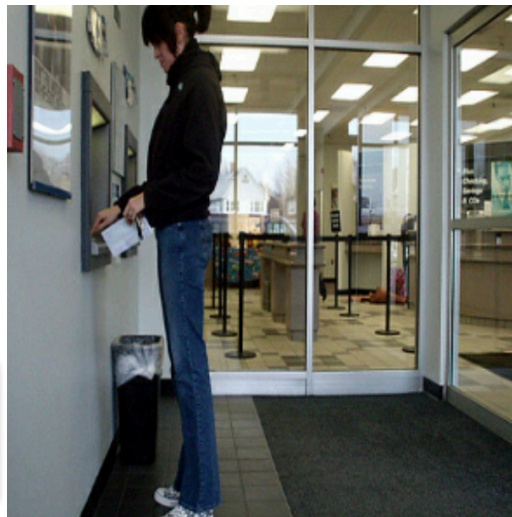
a logit $\rightarrow \frac{\partial p(z)}{\partial x_{i,j}}$
pixel $i,j \rightarrow$

prediction:
Cash machine

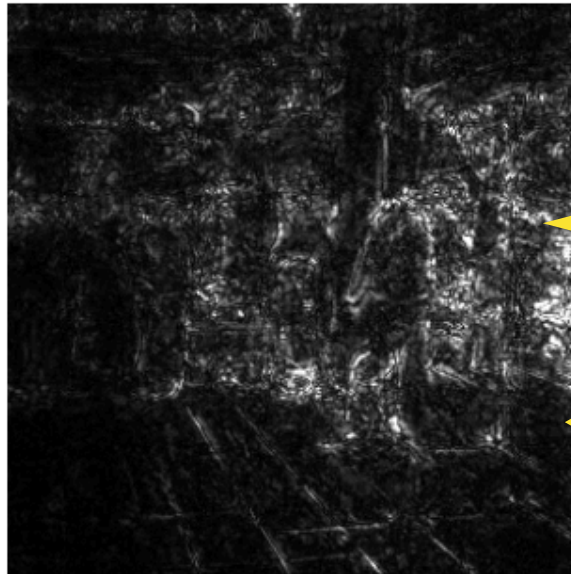


Why correct?
Why incorrect?

prediction:
Sliding door

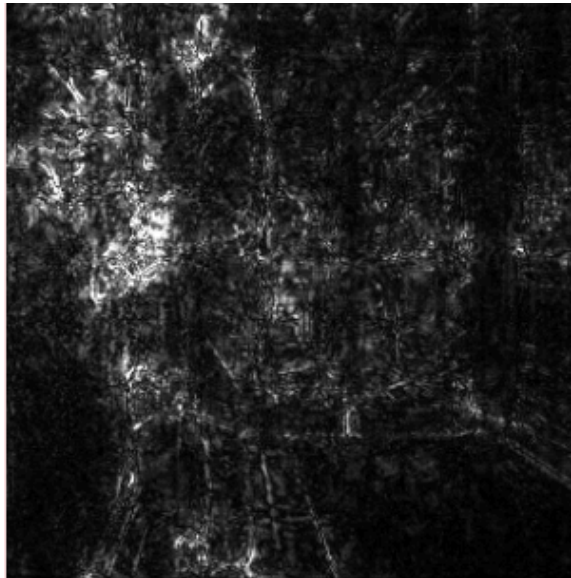


What we really want to ask...

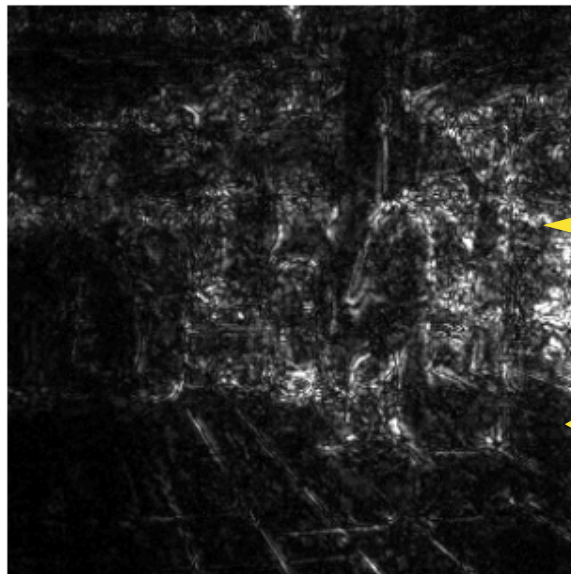


Were there more pixels on the cash machine than on the person?

Did the 'human' concept matter?
Did the 'glasses' or 'paper' matter?

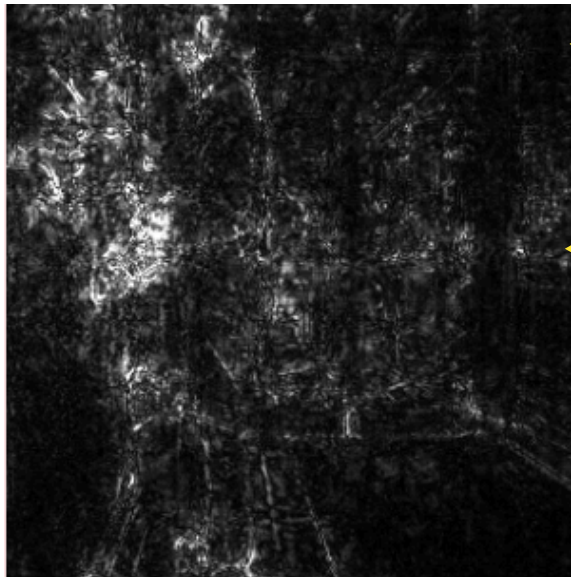


What we really want to ask...



Were there more pixels on the cash machine than on the person?

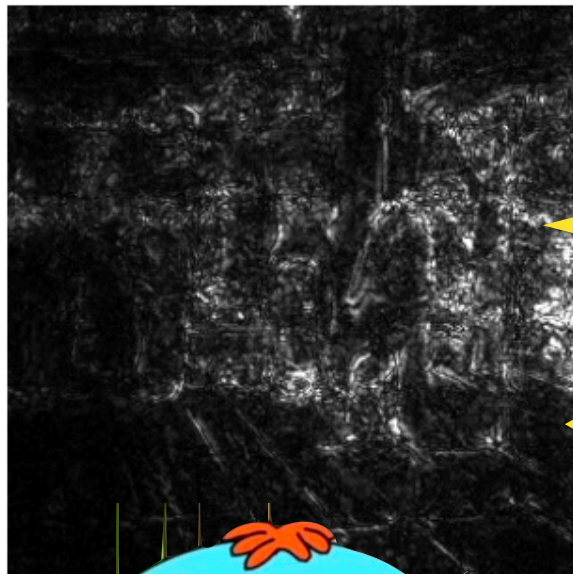
Did the 'human' concept matter?
Did the 'glasses' or 'paper' matter?



Which concept mattered more?

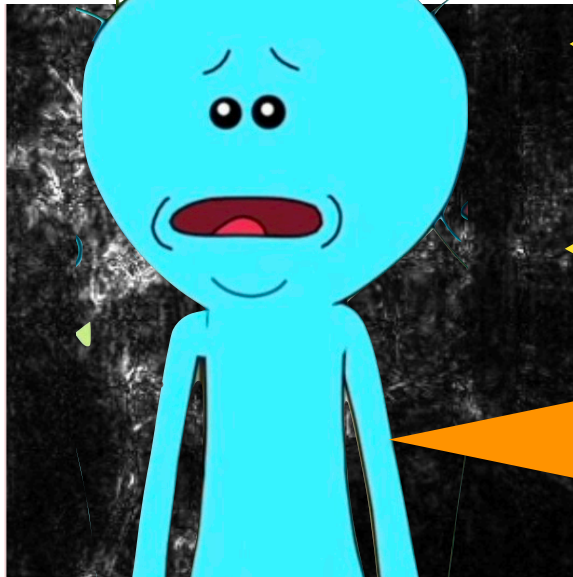
Is this true for all other cash machine predictions?

What we really want to ask...



Were there more pixels on the cash machine than on the person?

Did the 'human' concept matter?
Did the 'glasses' or 'paper' matter?

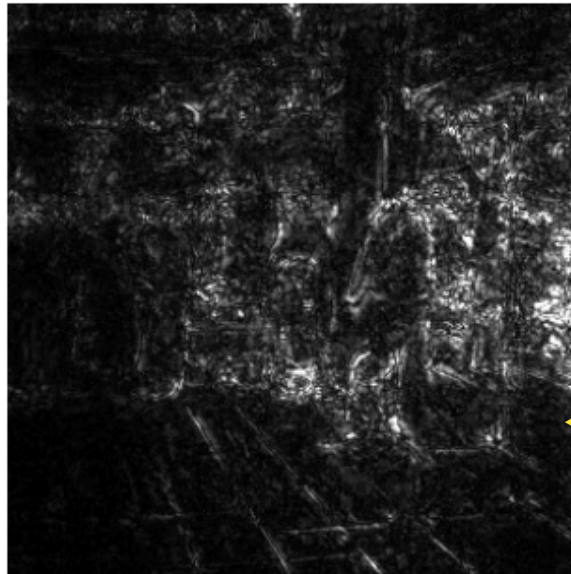


Which concept mattered more?

Is this true for all other cash machine predictions?

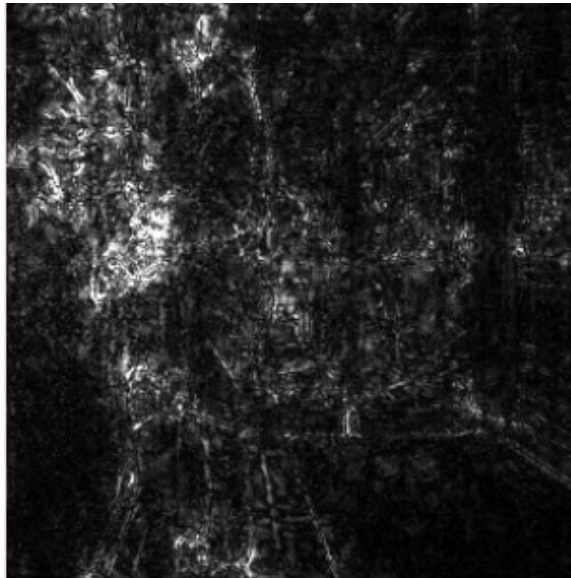
Oh no! I can't express these concepts as pixels!!
They weren't my input features either!

What we really want to ask...



Were there more pixels on the cash machine than on the person?

Did the 'human' concept matter?
Did the 'glasses' or 'paper' matter?

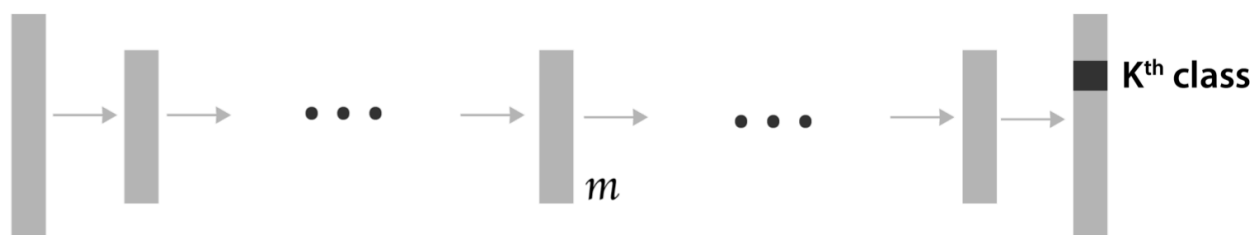


Which concept mattered more?

Is this true for all other cash machine predictions?

Wouldn't it be great if we can **quantitatively** measure how important *any* of these **user-chosen concepts** are?

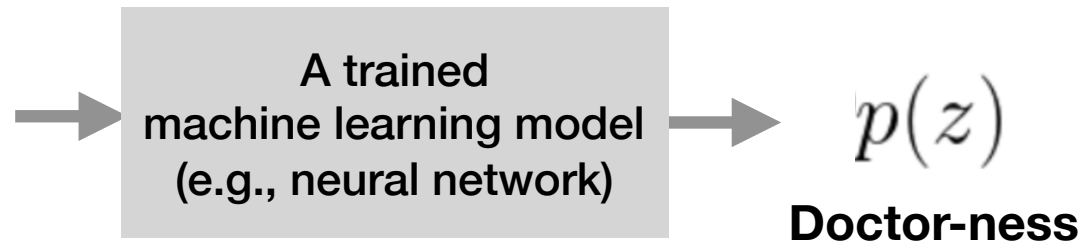
Goal of TCAV: Testing with Concept Activation Vectors



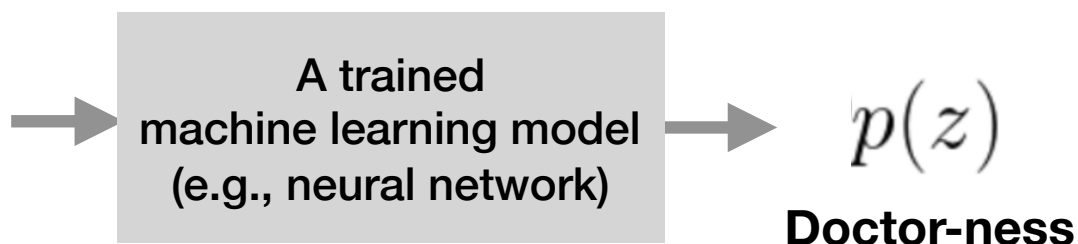
Quantitative explanation: how much a **concept** (e.g., gender, race) was important for a **prediction** in a trained model.

...even if the **concept** was not part of the training.

Goal of TCAV: Testing with Concept Activation Vectors

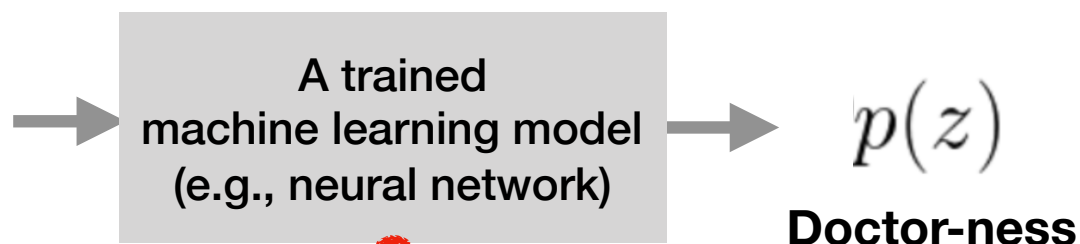


Goal of TCAV: Testing with Concept Activation Vectors



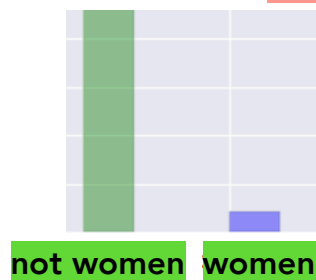
Was gender concept important
to this doctor image classifier?

Goal of TCAV: Testing with Concept Activation Vectors



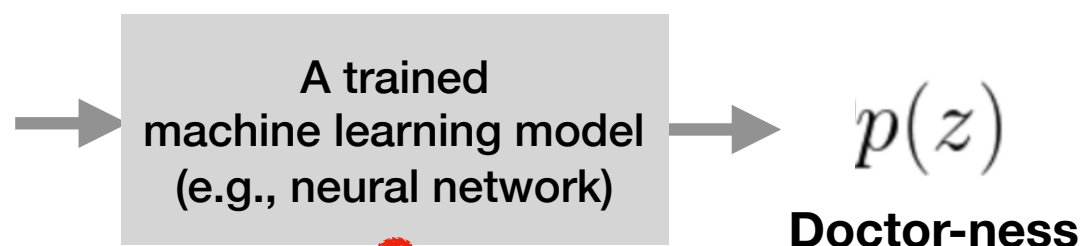
Was **gender** concept important to this **doctor** image classifier?

TCAV score for **Doctor**

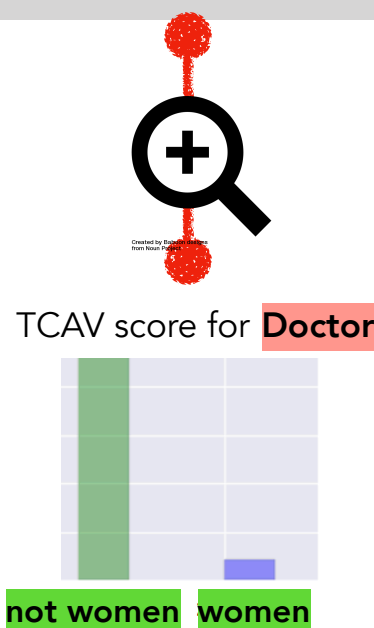


Goal of TCAV:

Testing with Concept Activation Vectors

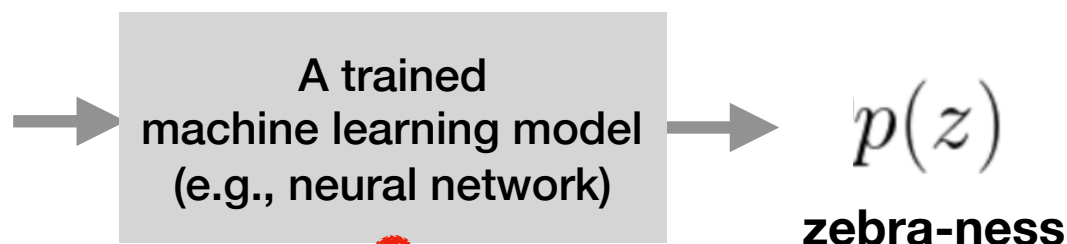


Was **gender** concept important to this **doctor** image classifier?



TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

Goal of TCAV: Testing with Concept Activation Vectors



Was striped concept important to this zebra image classifier?



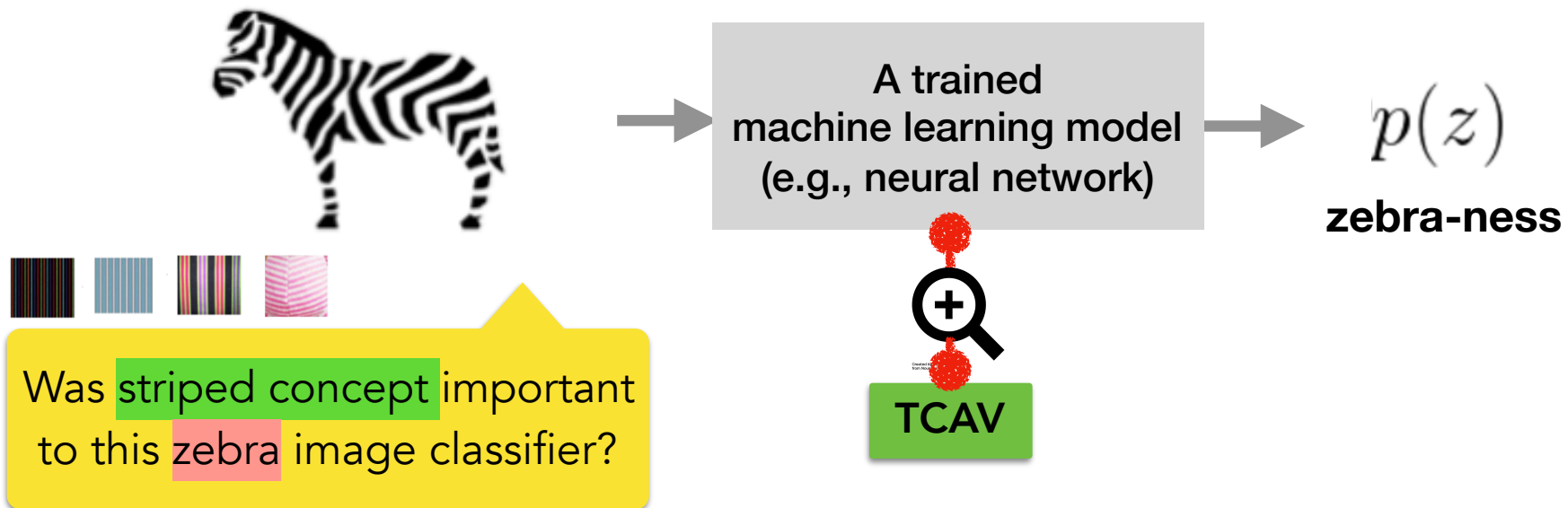
TCAV score for Zebra



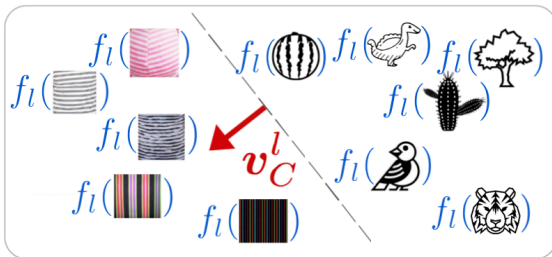
TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

TCAV:

Testing with Concept Activation Vectors



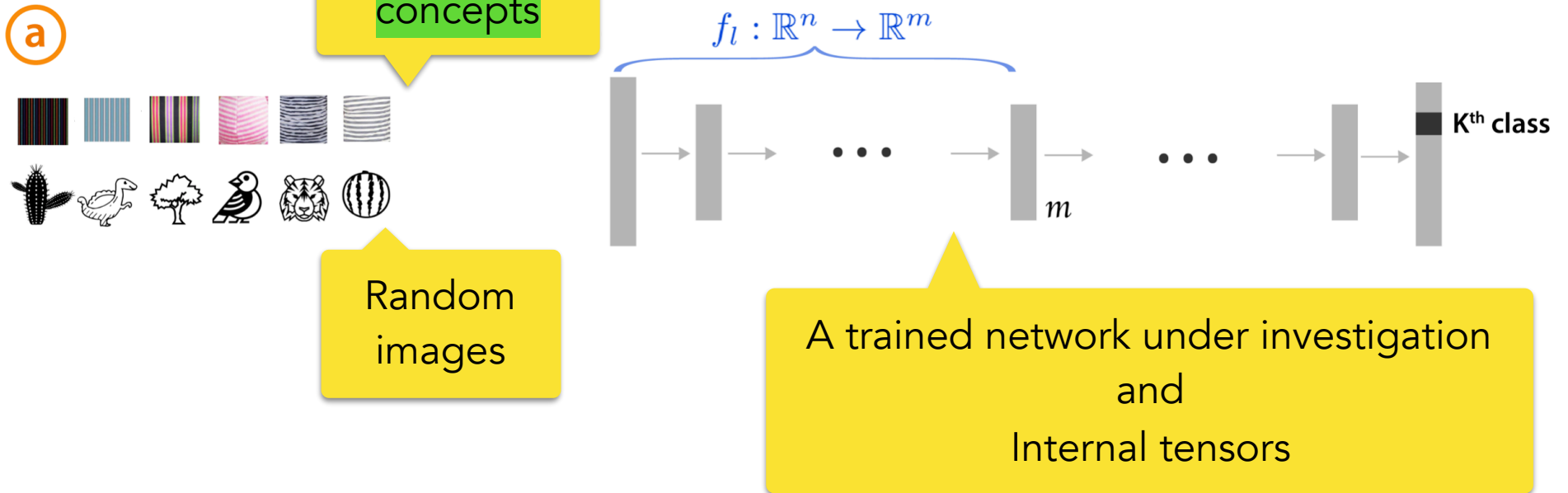
1. Learning CAVs



1. How to define concepts?

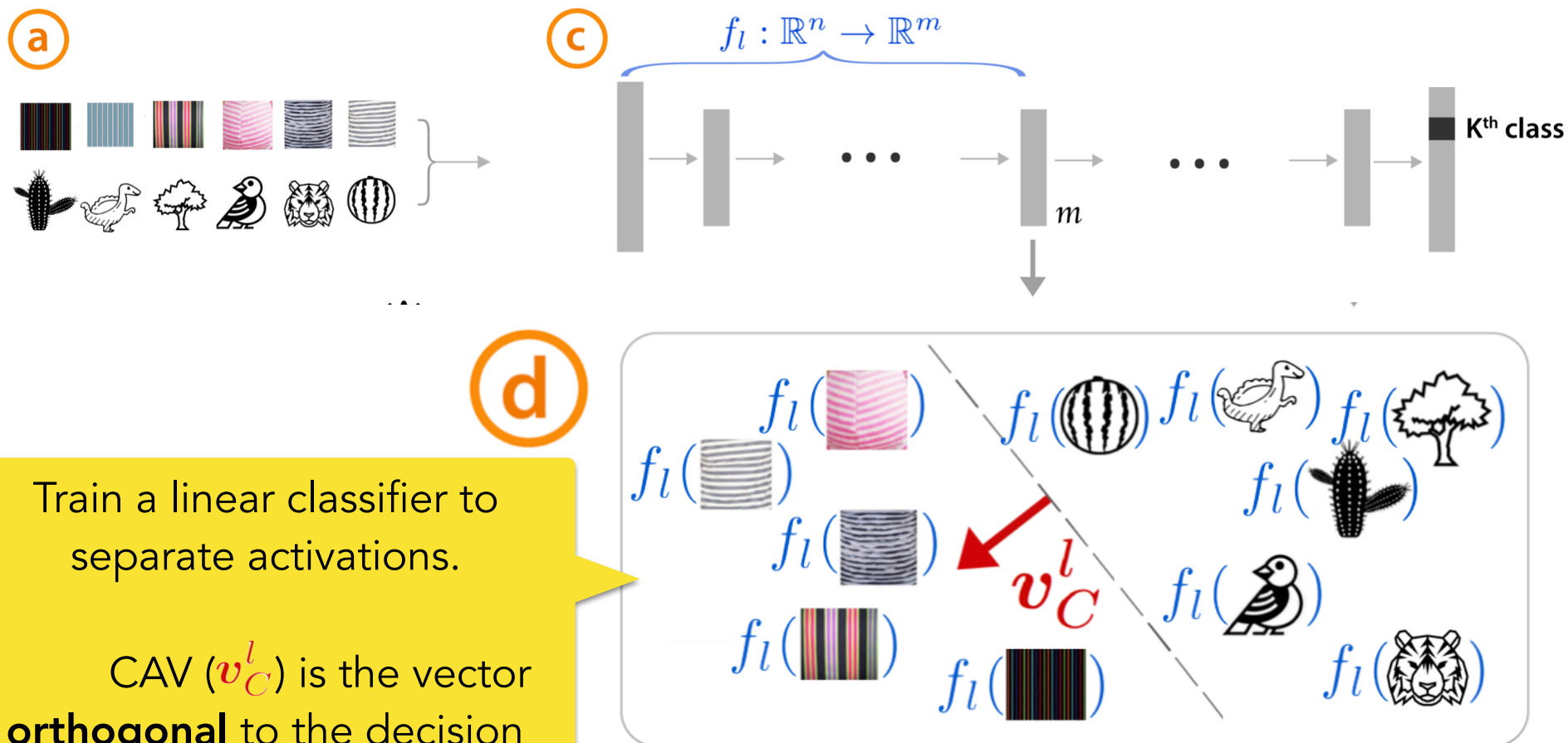
Defining concept activation vector (CAV)

Inputs:



Defining concept activation vector (CAV)

Inputs:



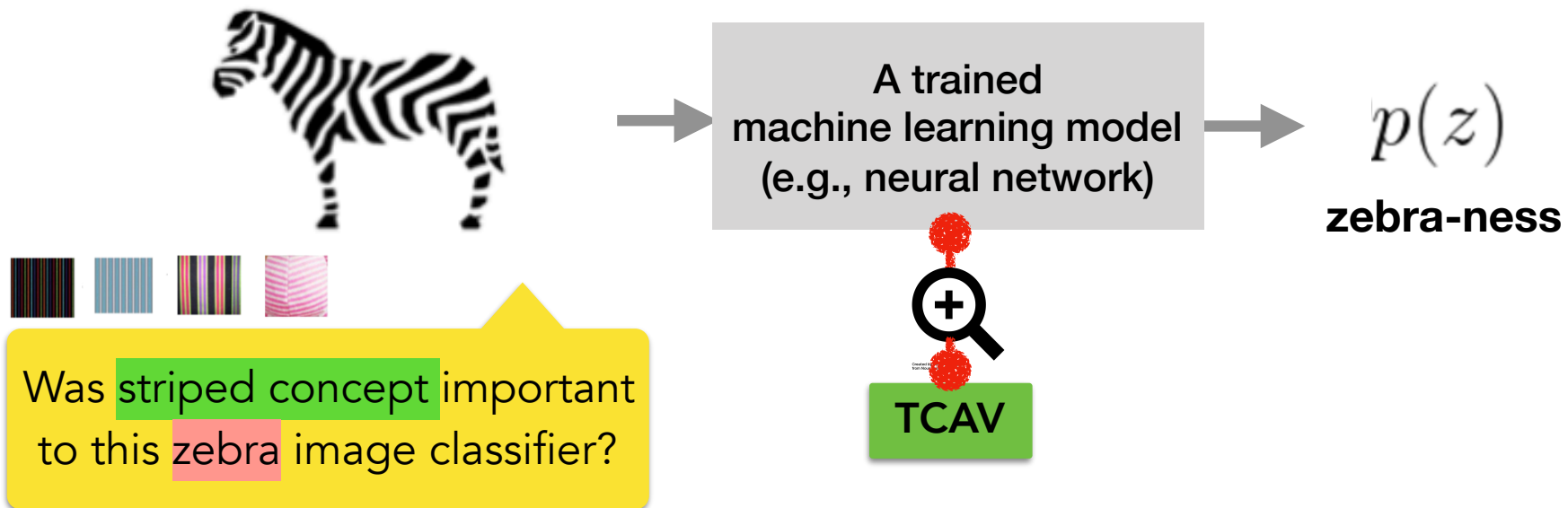
Train a linear classifier to separate activations.

CAV (v_C^l) is the vector **orthogonal** to the decision boundary.

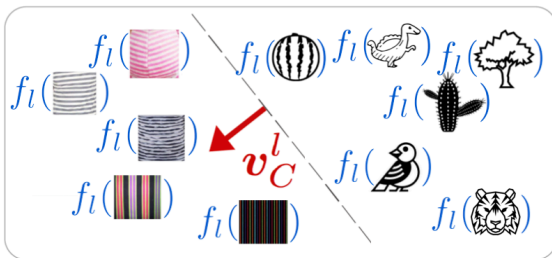
[Smilkov '17, Bolukbasi '16, Schmidt '15]

TCAV:

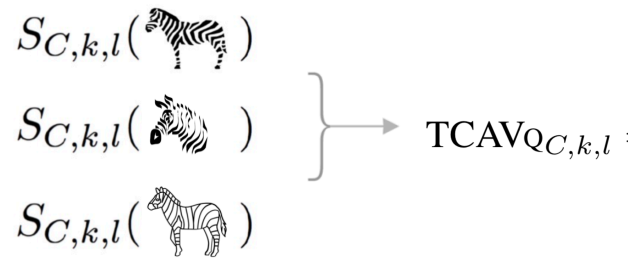
Testing with Concept Activation Vectors



1. Learning CAVs



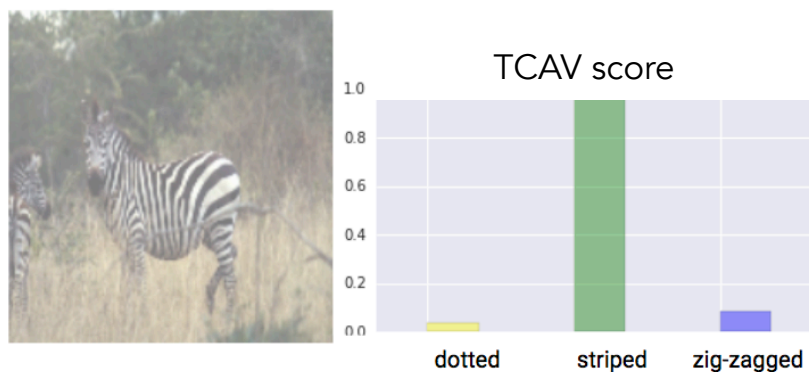
2. Getting TCAV score



2. How are the CAVs useful to get explanations?

TCAV core idea: Derivative with CAV to get prediction sensitivity

TCAV



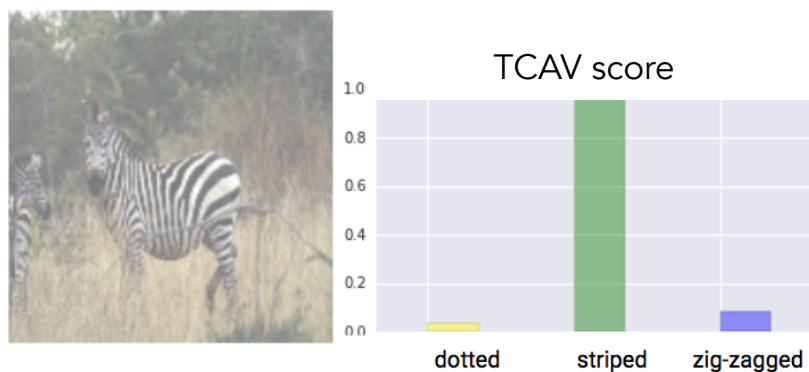
$$\begin{aligned} \text{zebra-ness} &\rightarrow \frac{\partial p(z)}{\partial \mathbf{v}_C^l} = S_{C,k,l}(\mathbf{x}) \\ \text{striped CAV} &\rightarrow \end{aligned}$$

Directional derivative with CAV

TCAV core idea:

Derivative with CAV to get prediction sensitivity

TCAV



$$\begin{aligned} \text{zebra-ness} &\rightarrow \frac{\partial p(z)}{\partial v_C^l} = S_{C,k,l}(x) \\ \text{striped CAV} &\rightarrow \end{aligned}$$

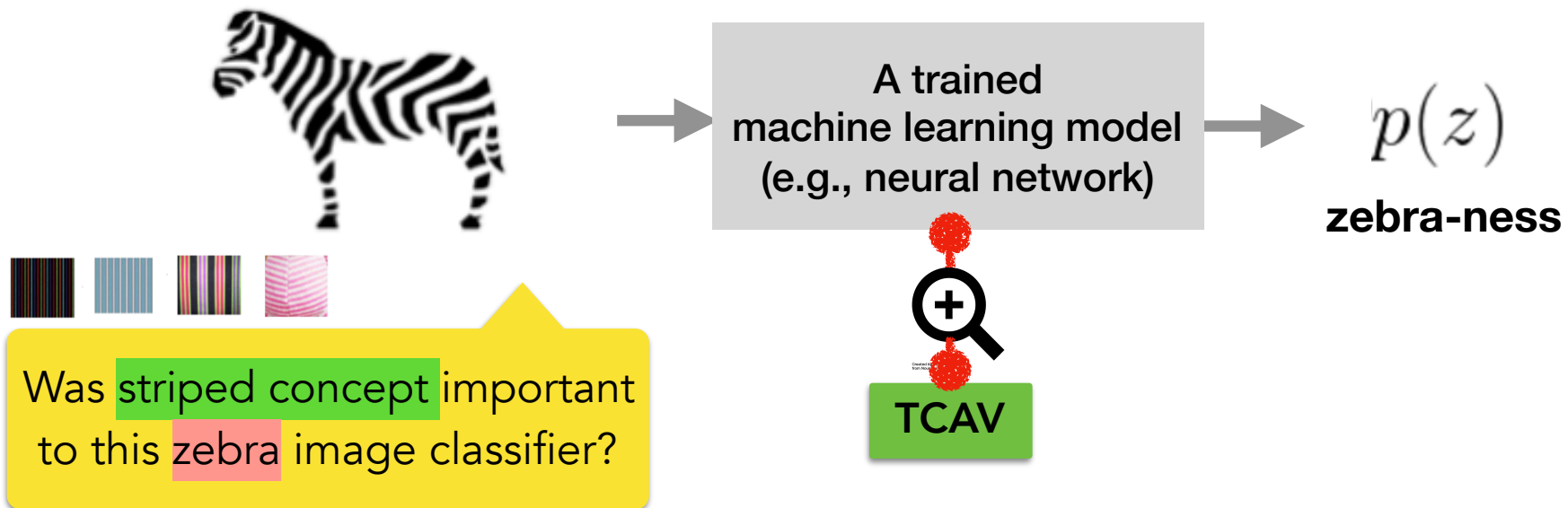
$$\left. \begin{aligned} S_{C,k,l}(\text{zebra}) \\ S_{C,k,l}(\text{zebra head}) \\ S_{C,k,l}(\text{zebra body}) \\ S_{C,k,l}(\text{zebra tail}) \end{aligned} \right\}$$

$$\text{TCAV}_{Q_{C,k,l}} = \frac{|\{x \in X_k : S_{C,k,l}(x) > 0\}|}{|X_k|}$$

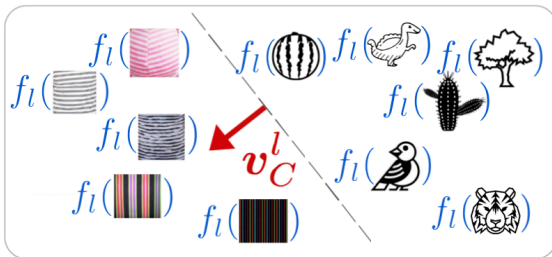
Directional derivative with CAV

TCAV:

Testing with Concept Activation Vectors



1. Learning CAVs

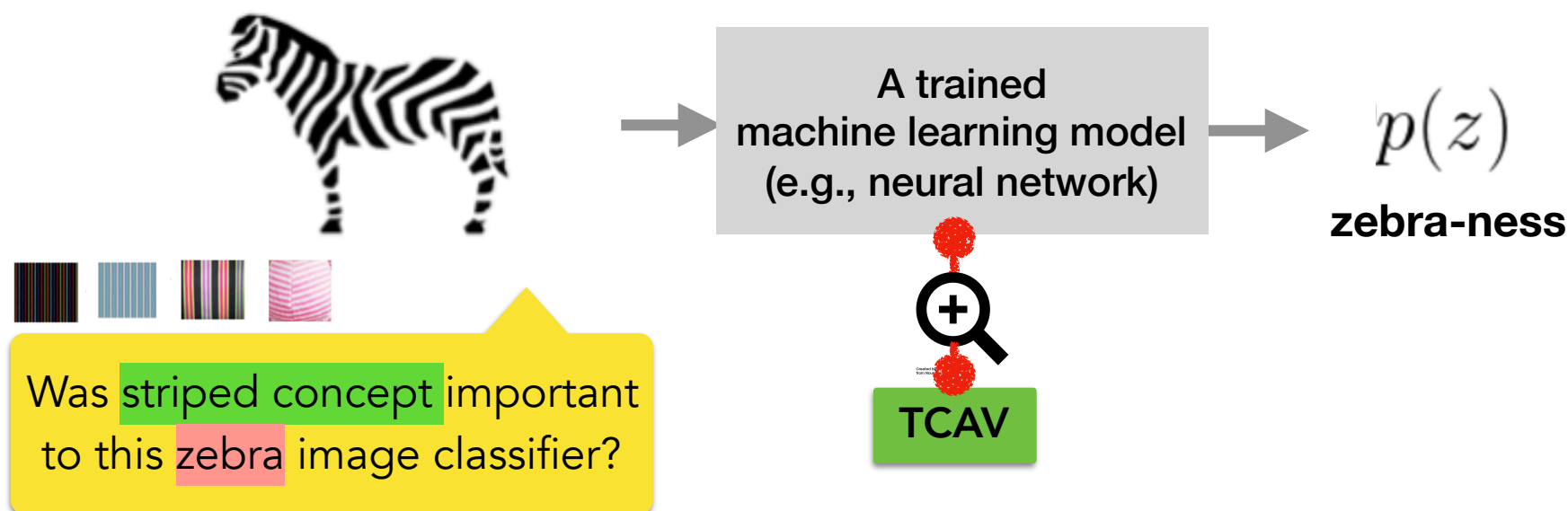


2. Getting TCAV score

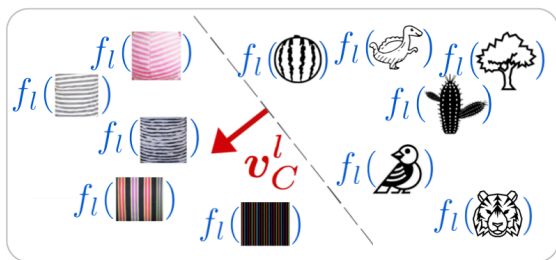
$$\left. \begin{array}{l} S_{C,k,l}(\text{zebra}) \\ S_{C,k,l}(\text{zebra}) \\ S_{C,k,l}(\text{zebra}) \end{array} \right\} \rightarrow \text{TCAV}_{QC,k,l}$$

TCAV:

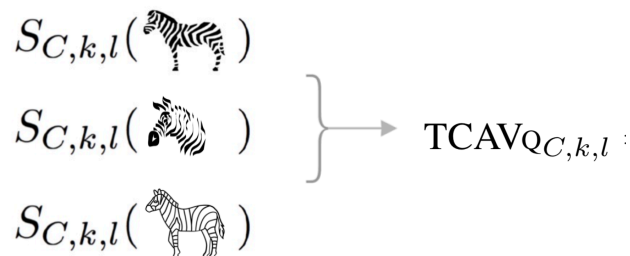
Testing with Concept Activation Vectors



1. Learning CAVs



2. Getting TCAV score



3. CAV validation

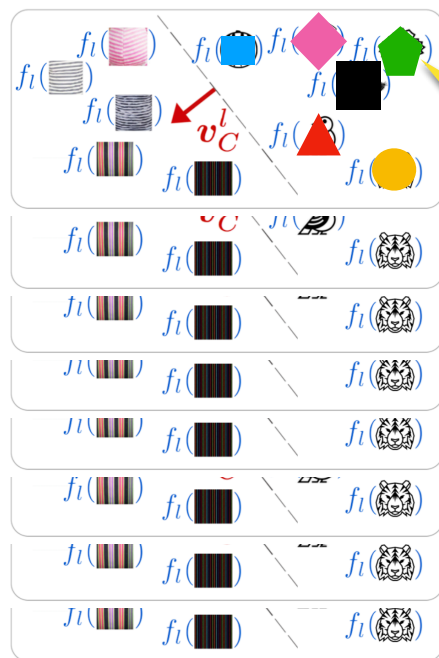
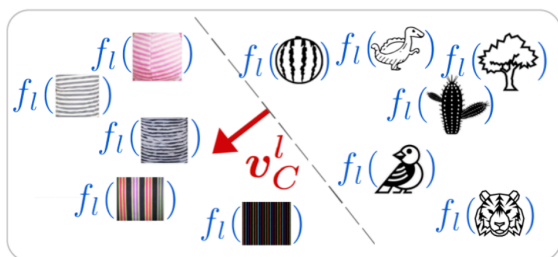
Qualitative
Quantitative

Quantitative validation:

Guarding against spurious CAV

Did my CAVs returned high sensitivity by chance?

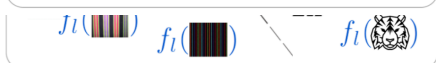
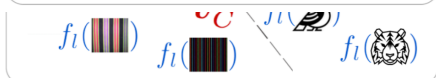
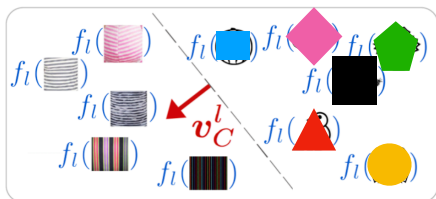
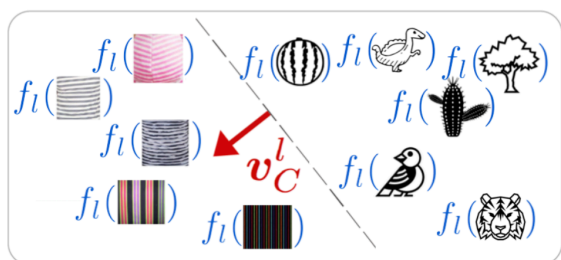
Quantitative validation: Guarding against spurious CAV



Learn many stripes CAVs
using different sets of
random images

Quantitative validation:

Guarding against spurious CAV



Zebra

→ $\text{TCAV}_{Q_C, k, l} :$

⋮

→ $\text{TCAV}_{Q_C, k, l} :$

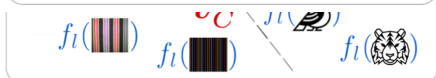
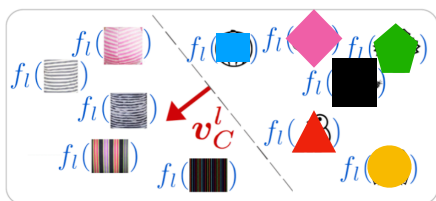
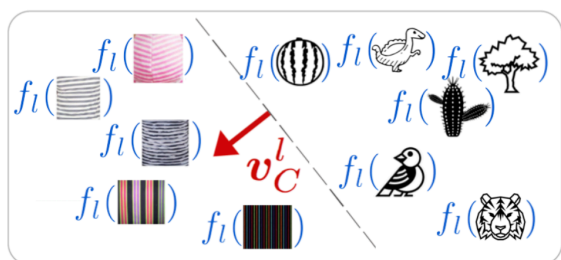
→ $\text{TCAV}_{Q_C, k, l} :$

→ $\text{TCAV}_{Q_C, k, l} :$

⋮

Quantitative validation:

Guarding against spurious CAV



Zebra

→ $\text{TCAV}_{Q_C, k, l} :$

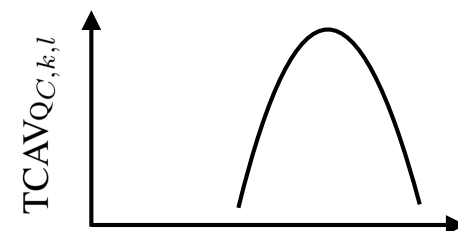
⋮

→ $\text{TCAV}_{Q_C, k, l} :$

→ $\text{TCAV}_{Q_C, k, l} :$

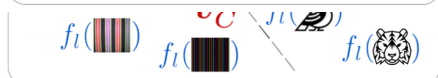
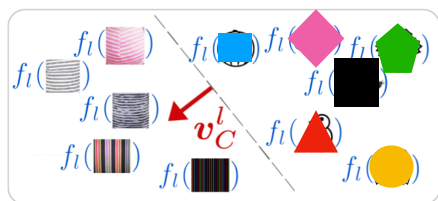
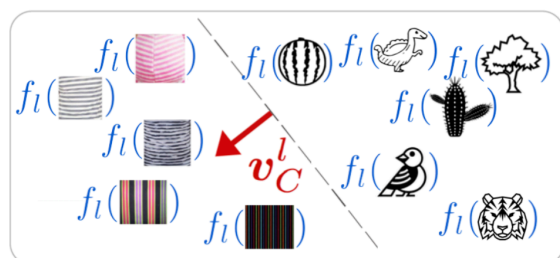
→ $\text{TCAV}_{Q_C, k, l} :$

⋮



Quantitative validation:

Guarding against spurious CAV



Zebra

→ $\text{TCAV}_{Q_C, k, l} :$

⋮

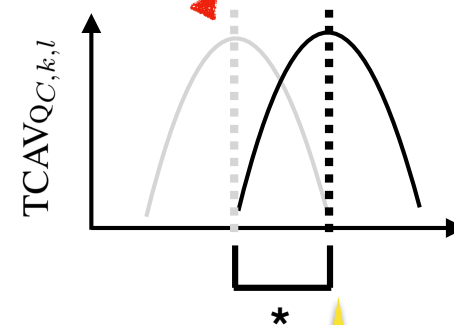
→ $\text{TCAV}_{Q_C, k, l} :$

→ $\text{TCAV}_{Q_C, k, l} :$

→ $\text{TCAV}_{Q_C, k, l} :$

⋮

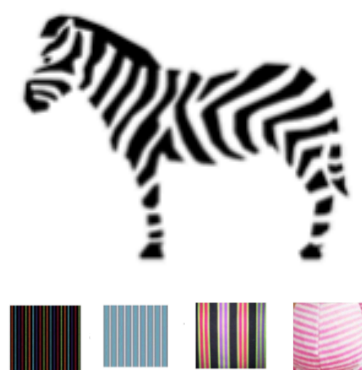
TCAV score
random



Check the distribution of $\text{TCAV}_{Q_C, k, l}$ is statistically different from random using t-test

Recap TCAV:

Testing with Concept Activation Vectors

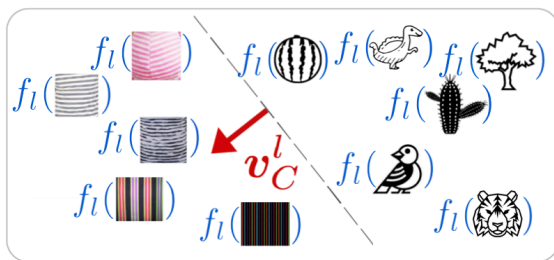


TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

Even if your training data wasn't tagged with the **concept**

Even if your input feature did not include the **concept**

1. Learning CAVs



2. Getting TCAV score

$$\left. \begin{array}{l} S_{C,k,l}(\text{zebra}) \\ S_{C,k,l}(\text{zebra}) \\ S_{C,k,l}(\text{zebra}) \end{array} \right\} \rightarrow \text{TCAV}_{QC,k,l}$$

3. CAV validation

Qualitative
Quantitative

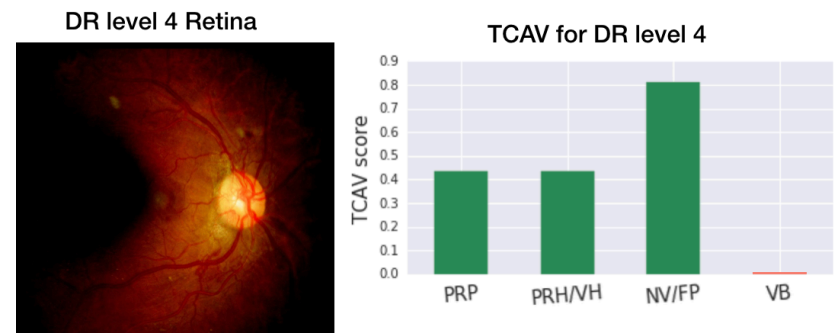
Results

1. Sanity check experiment



2. Biases in Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy



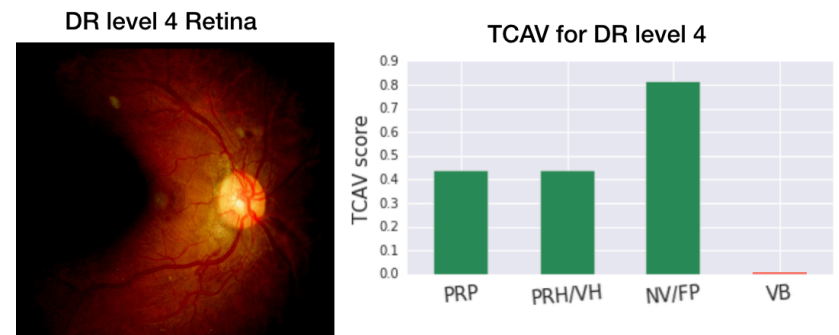
Results

1. Sanity check experiment

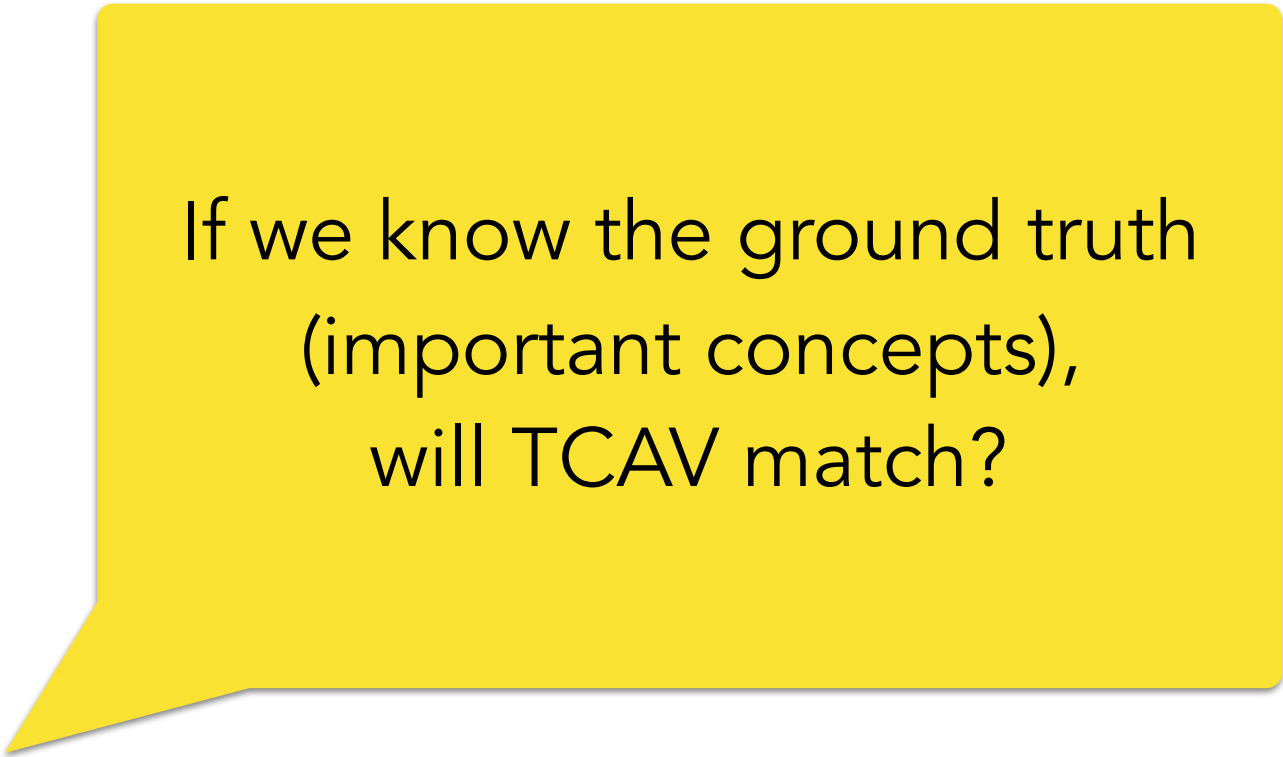


2. Biases from Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy



Sanity check experiment



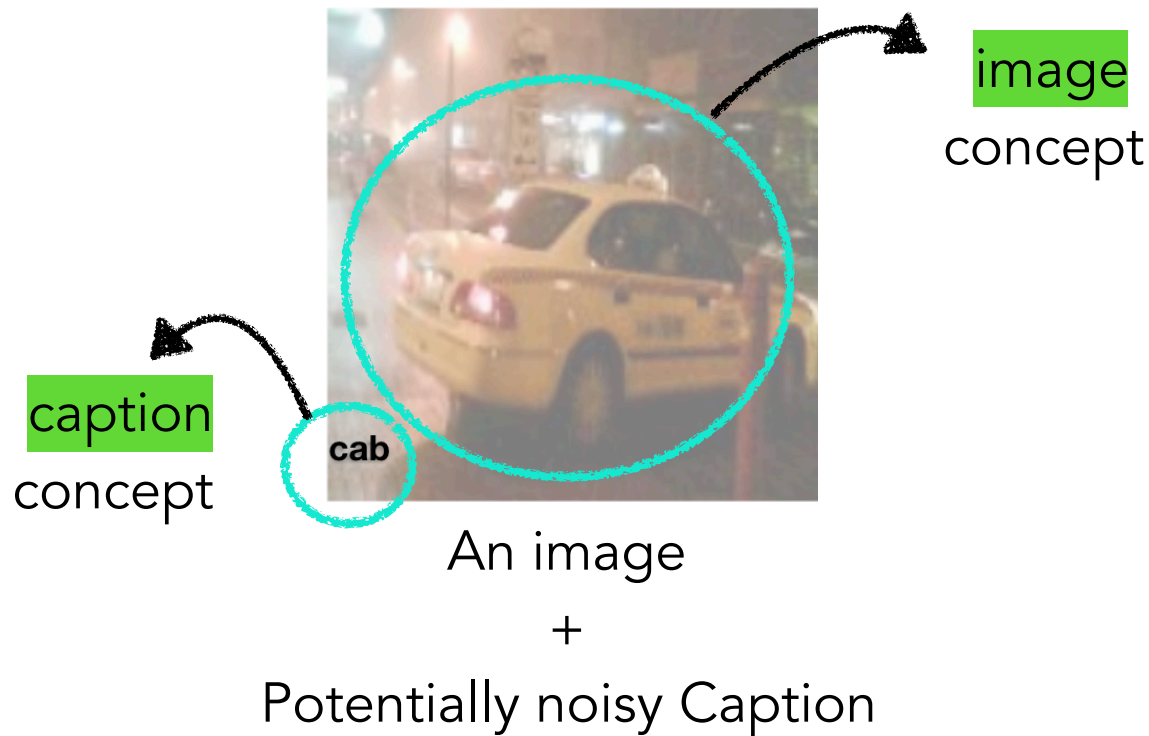
If we know the ground truth
(important concepts),
will TCAV match?

Sanity check experiment setup



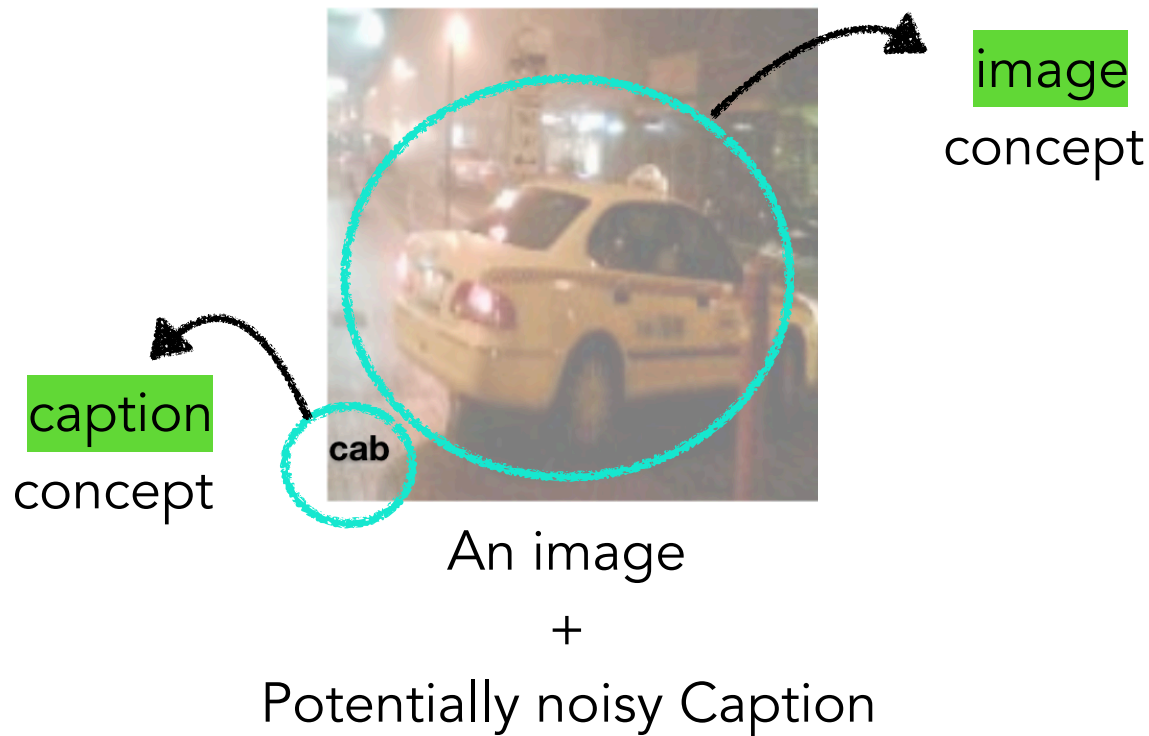
An image
+
Potentially noisy Caption

Sanity check experiment setup

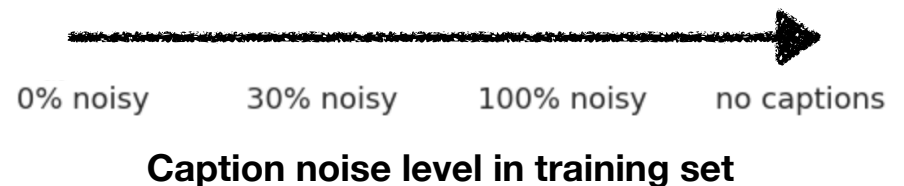


models can use either
image or caption
concept for
classification.

Sanity check experiment setup



models can use either **image** or **caption** concept for classification.



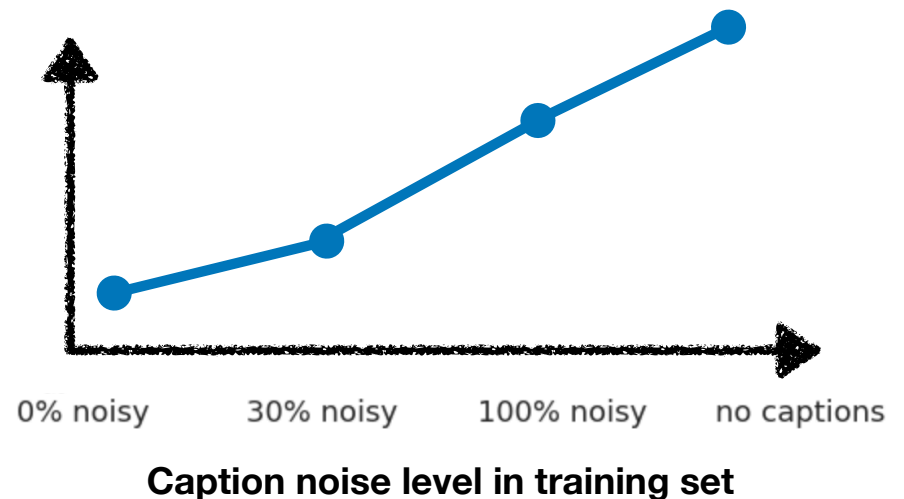
Sanity check experiment setup



models can use either
image or caption
concept for
classification.

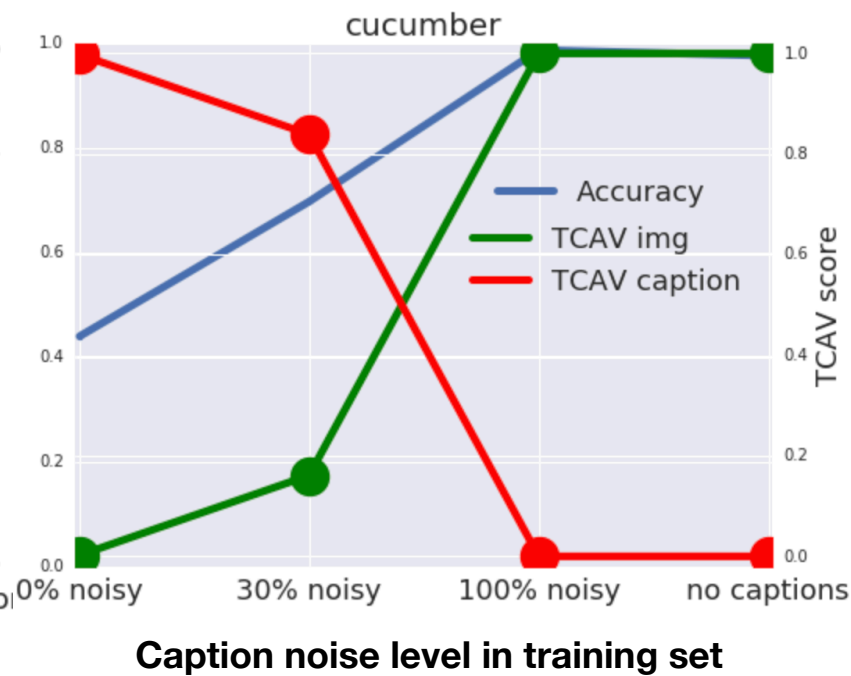
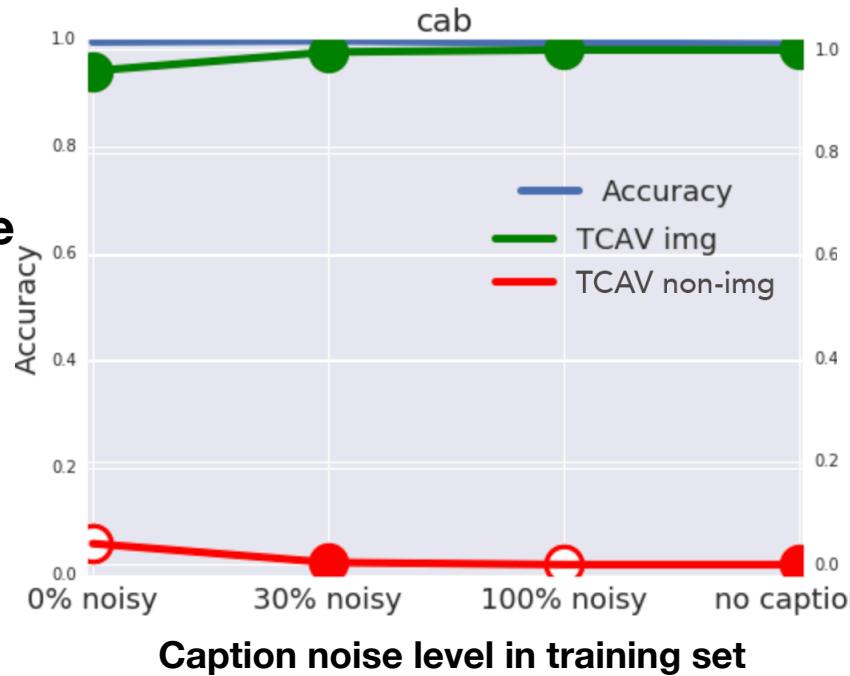


Test accuracy
with
no caption image
=
Importance of
image concept



Sanity check experiment


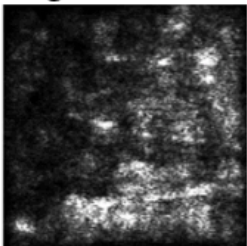
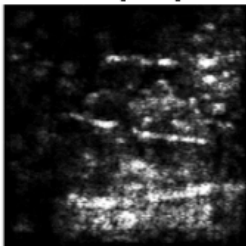
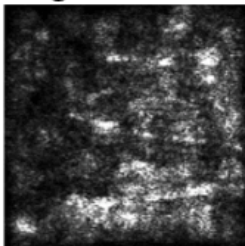
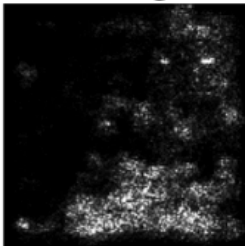

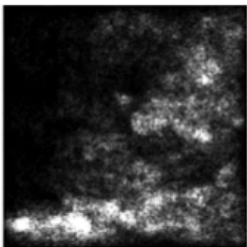
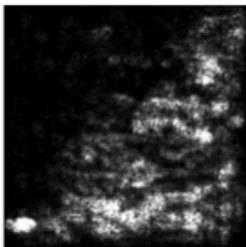
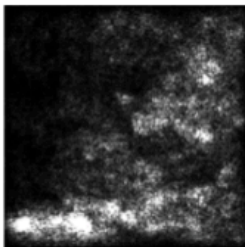
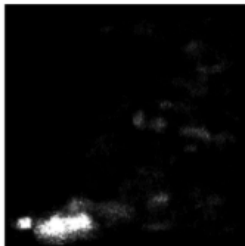

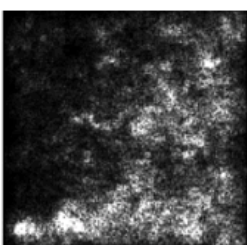
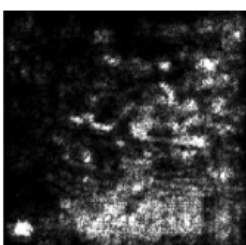
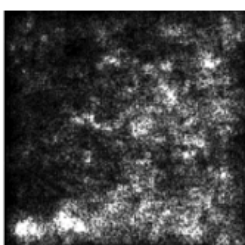
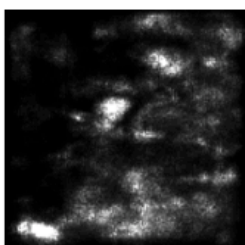

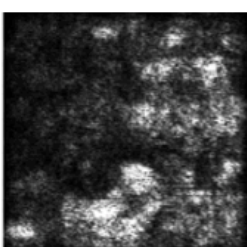
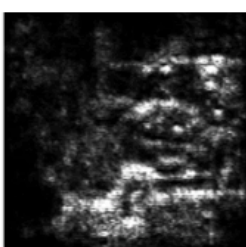
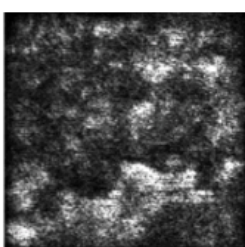
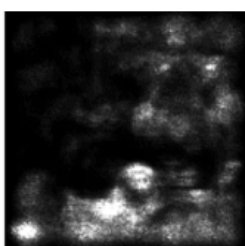
Test accuracy
with
no caption image





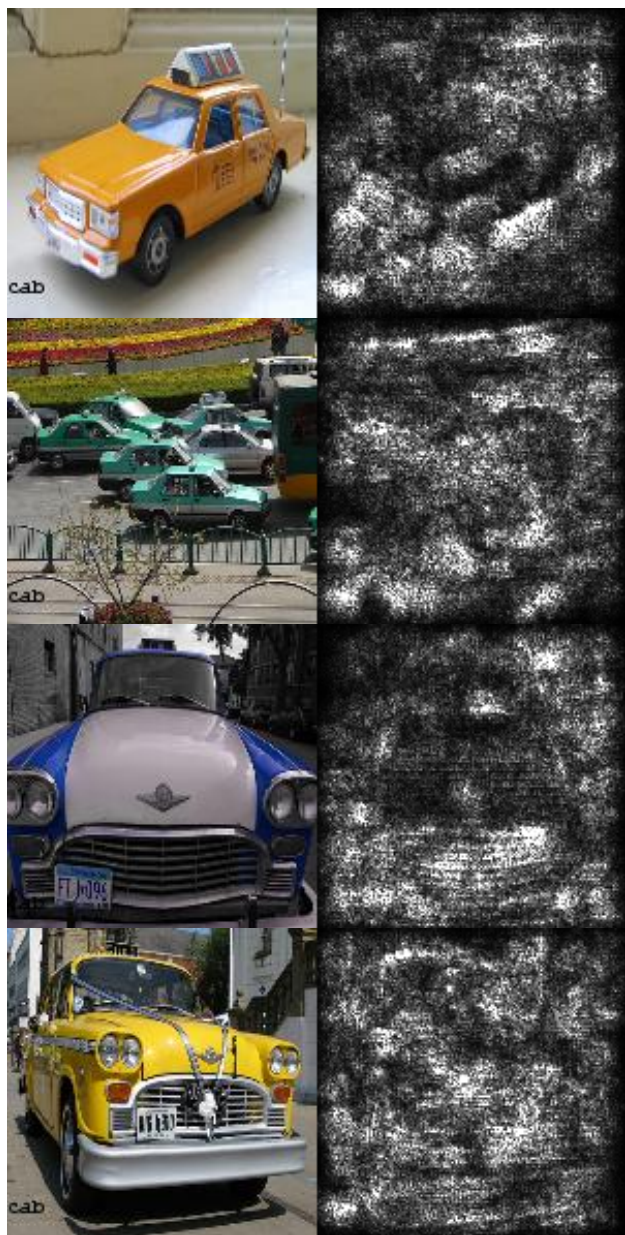
Cool, cool.
Can saliency maps do this too?

Can saliency maps communicate the same information?

Ground truth	Model trained on	Image with caption	Vanilla gradient	Guided backprop	Integrated gradient	Smoothgrad
Image concept	Images without captions (no captions)					
Image concept	Images with captions (0% noise)					
Image concept	Images with captions (30% noise)					
Image concept	Images with captions (100% noise)					

Human subject experiment:

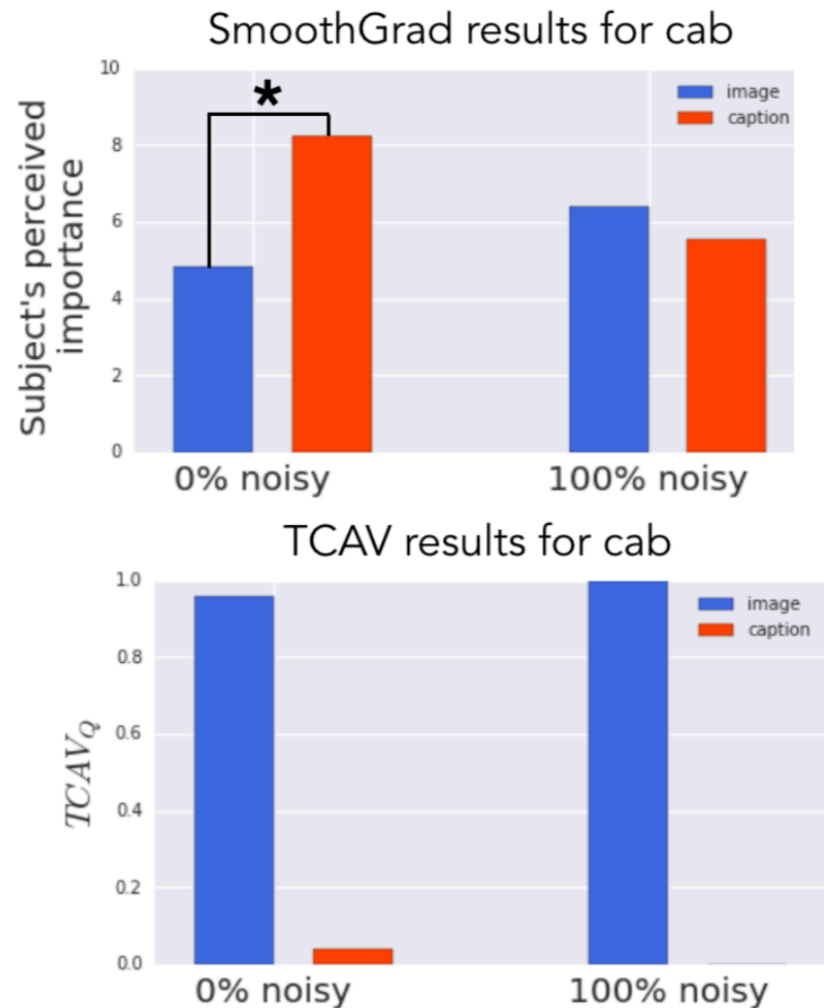
Can saliency maps communicate the same information?



- 50 turkers are
- asked to judge importance of **image** vs. **caption** given saliency maps.
- asked to indicate their confidence
- shown 3 classes (cab, zebra, cucumber) x 2 saliency maps for one model

Human subject experiment: Can saliency maps communicate the same information?

- Random chance: 50%
- Human performance with saliency map: 52%
- Humans can't agree: more than 50% no significant consensus



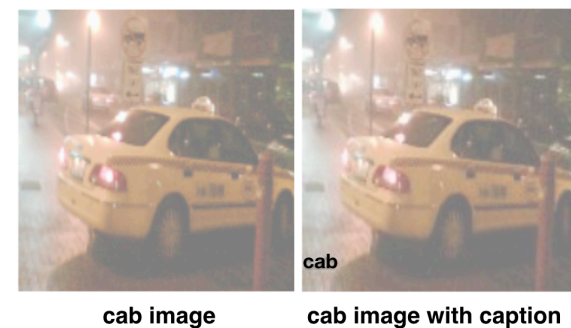
Human subject experiment: Can saliency maps communicate the same information?

- Random chance: 50%
- Human performance with saliency map: 52%
- Humans can't agree: more than 50% no significant consensus
- Humans are **very** confident even when they are wrong.



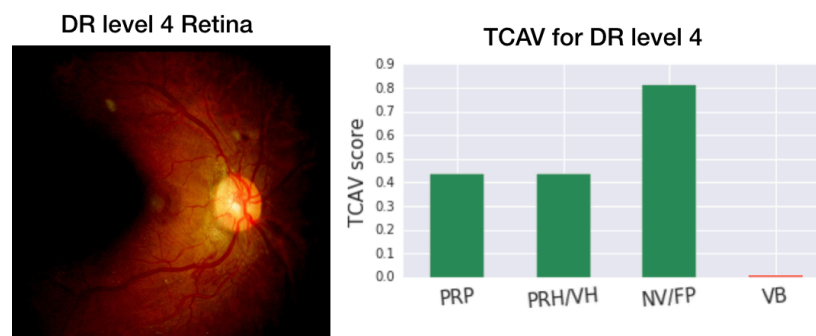
Results

1. Sanity check experiment

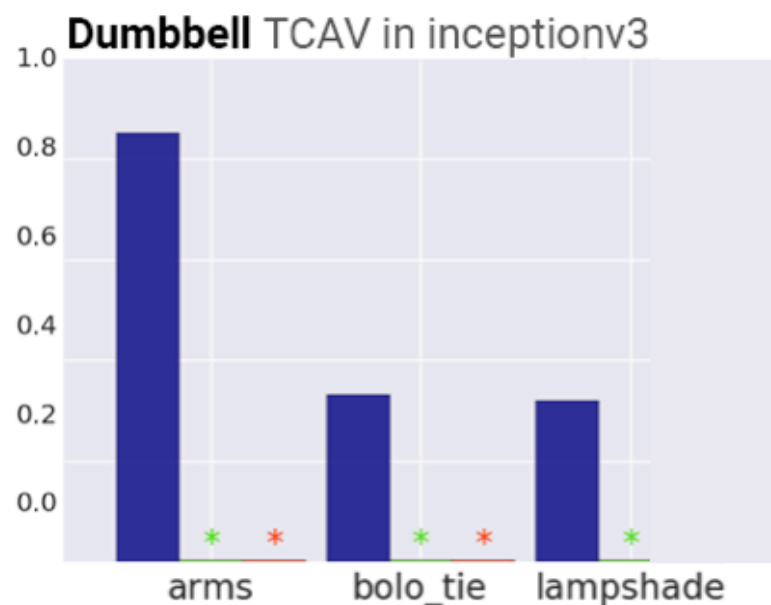
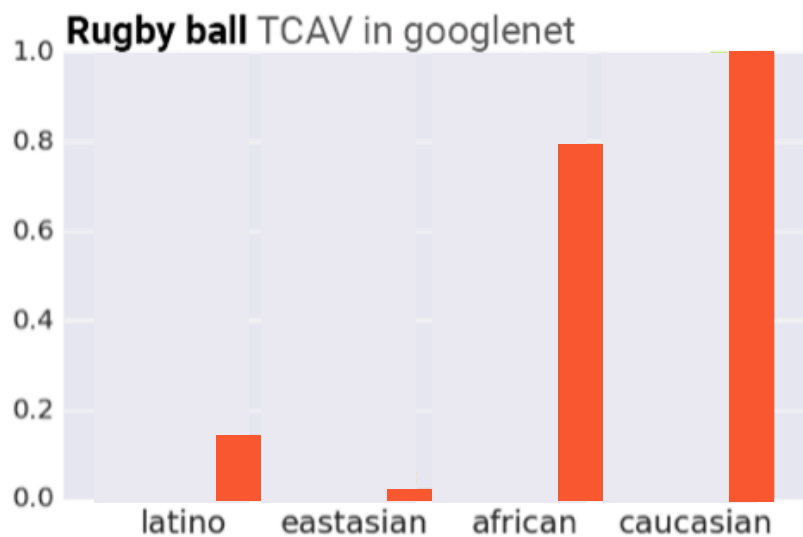
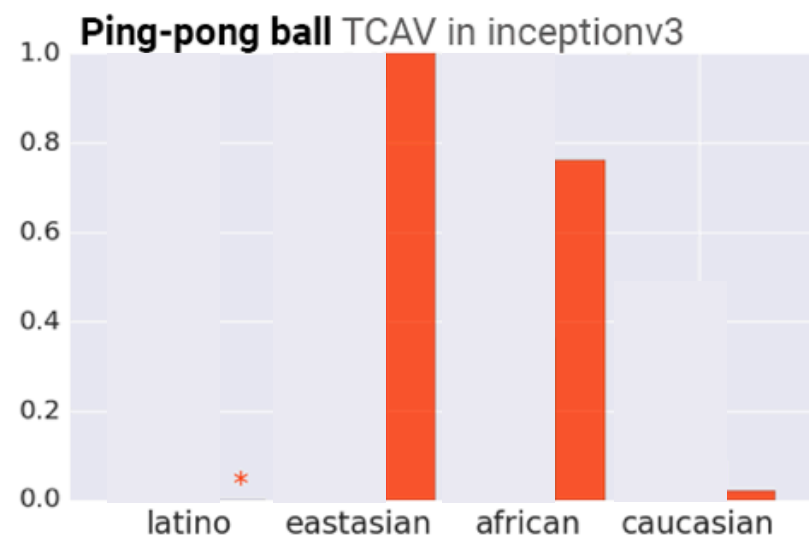
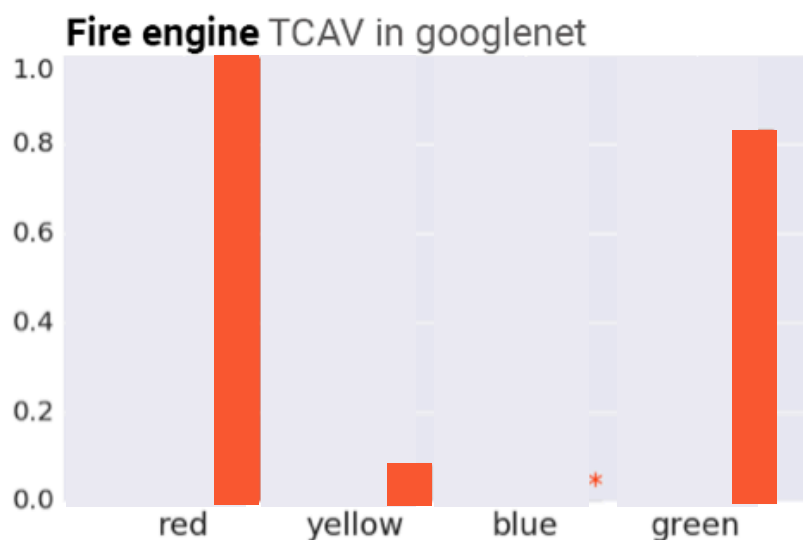


2. Biases from Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy

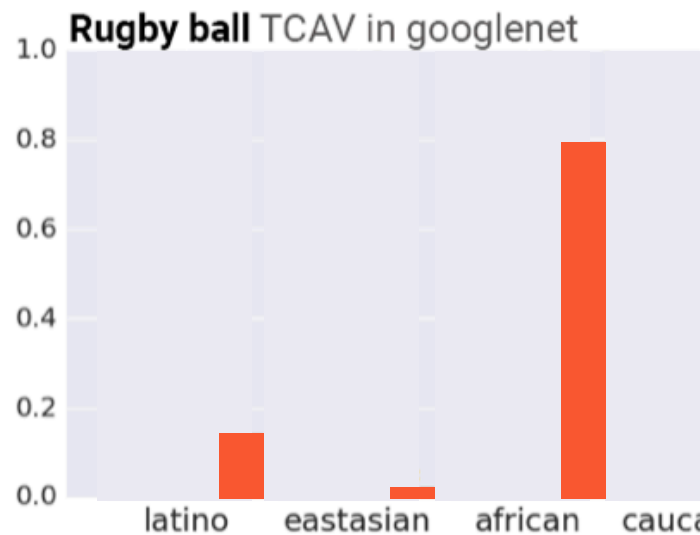
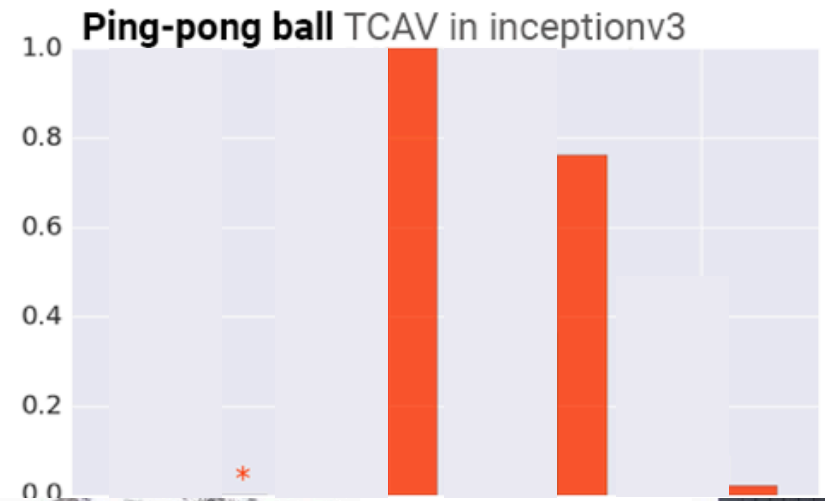
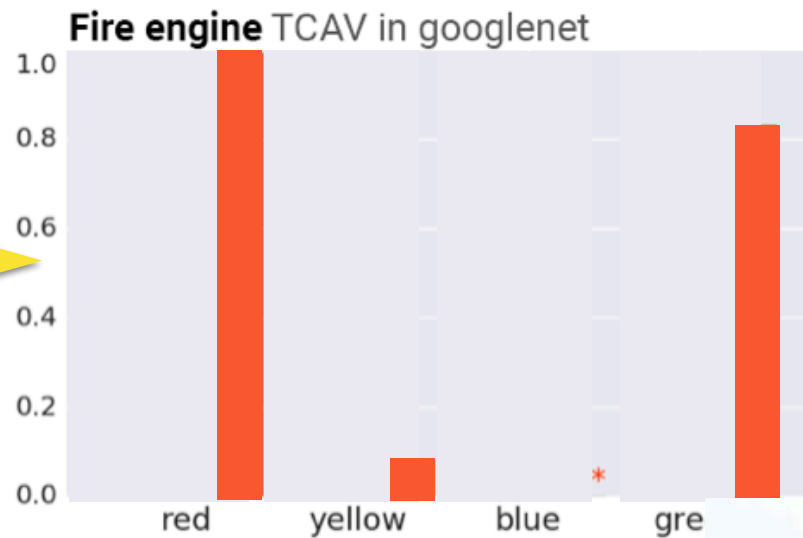


TCAV in Two widely used image prediction models



TCAV in Two widely used image prediction models

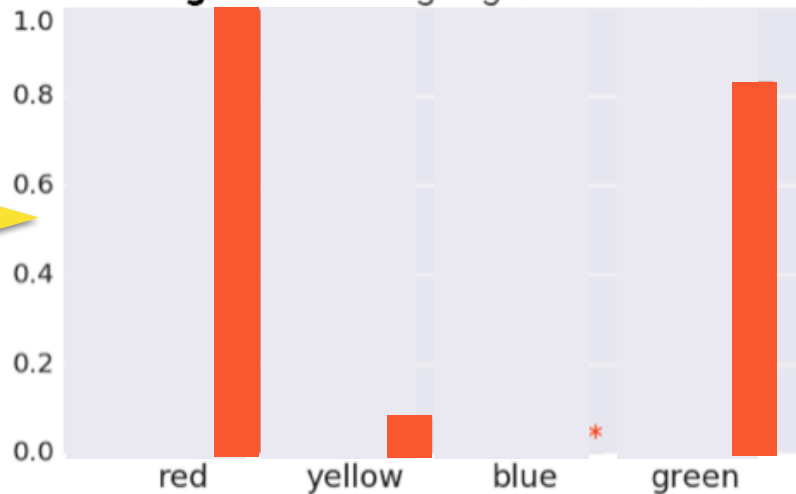
Geographical
bias?



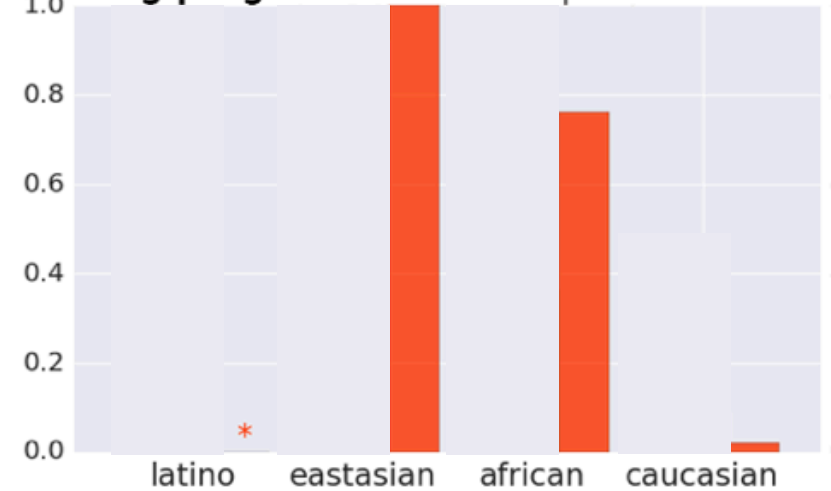
TCAV in Two widely used image prediction models

Geographical
bias?

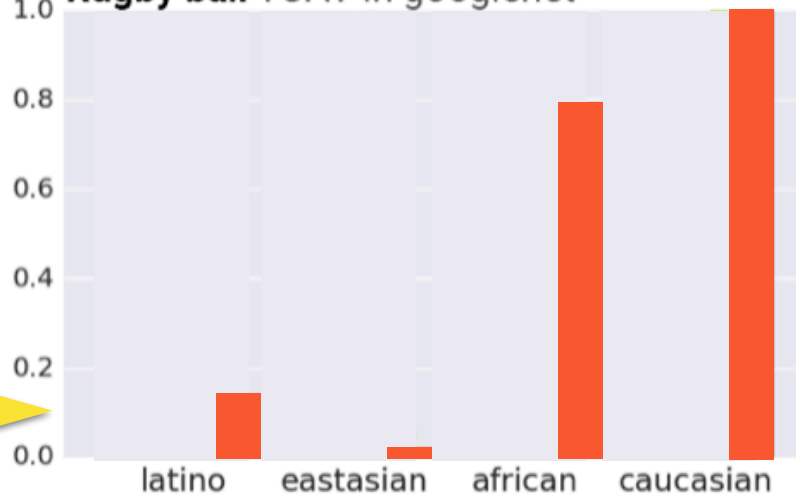
Fire engine TCAV in googlenet



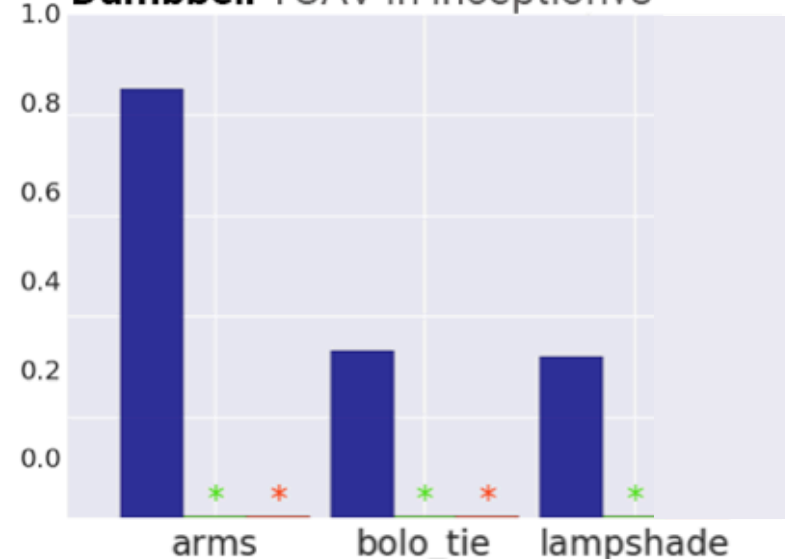
Ping-pong ball TCAV in inceptionv3



Rugby ball TCAV in googlenet



Dumbbell TCAV in inceptionv3



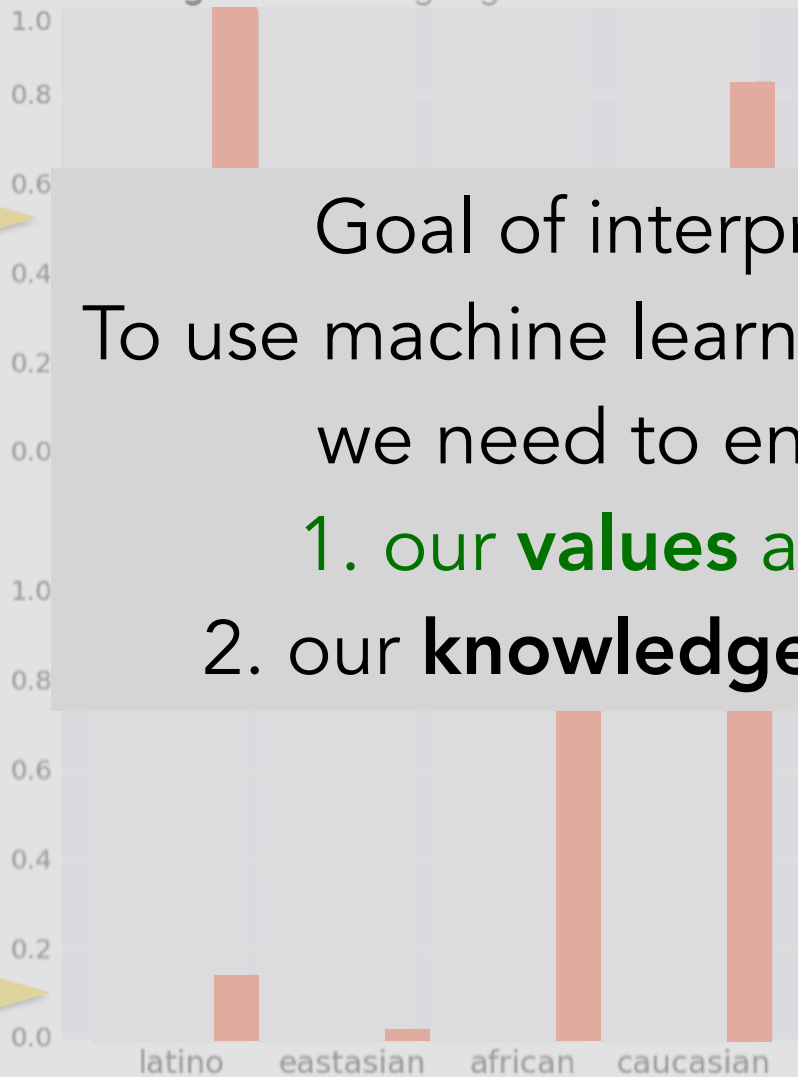
Quantitative
confirmation to
previously
qualitative
findings
[Stock & Cisse,
2017]

TCAV in

Two widely used image prediction models

Geographical
bias?

Fire engine TCAV in googlenet



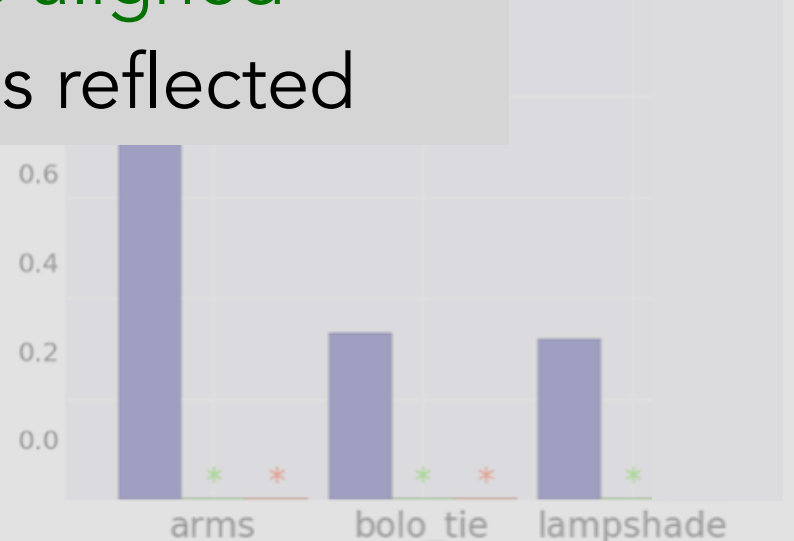
Ping-pong ball TCAV in inceptionv3



Goal of interpretability:
To use machine learning **responsibly**
we need to ensure that

1. our **values** are aligned
2. our **knowledge** is reflected

Quantitative
confirmation to
previously
qualitative
findings
[Stock & Cisse,
2017]



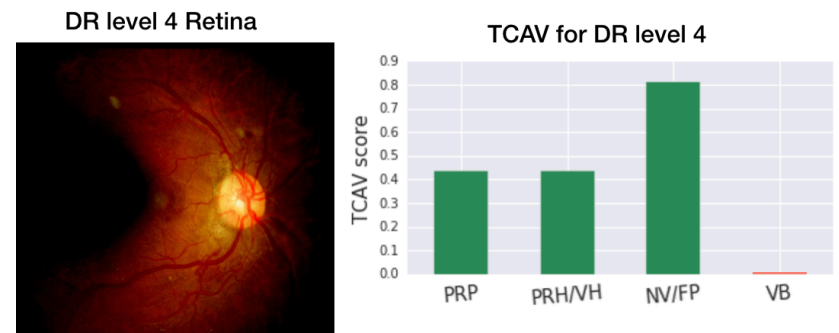
Results

1. Sanity check experiment



2. Biases Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy



Diabetic Retinopathy

- Treatable but sight-threatening conditions
- Have model to with accurate prediction of DR (85%)
[Krause et al., 2017]

Concepts the **ML model** uses

Vs

Diagnostic Concepts **human** doctors use

DR level 4 Retina

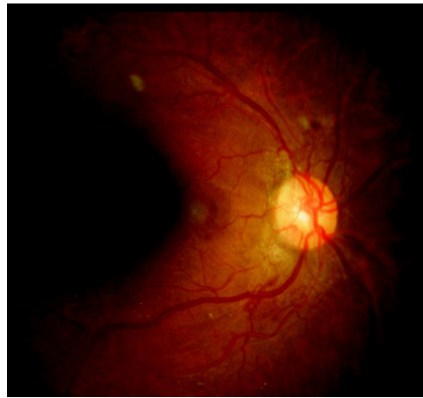


TCAV for Diabetic Retinopathy

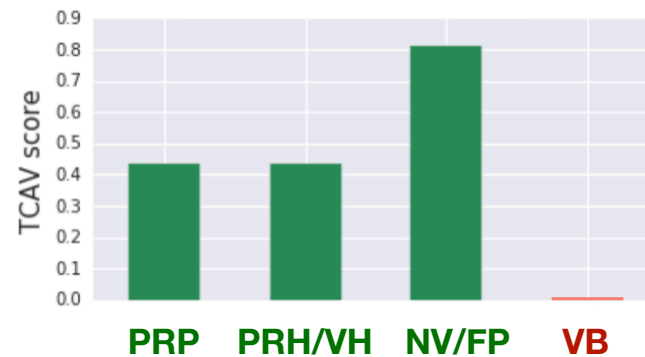
Prediction
class Prediction
accuracy

DR level 4 High

Example



TCAV scores



TCAV shows the model is **consistent** with doctor's knowledge when model is **accurate**

Green: domain expert's label on concepts belong to the level

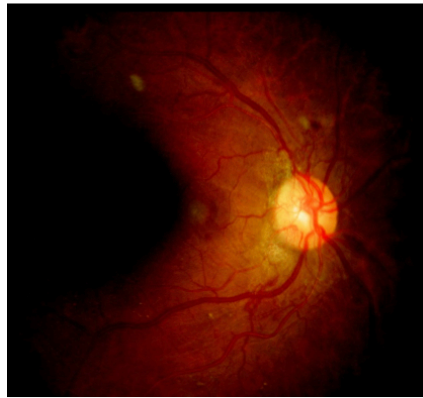
Red: domain expert's label on concepts does not belong to the level

TCAV for Diabetic Retinopathy

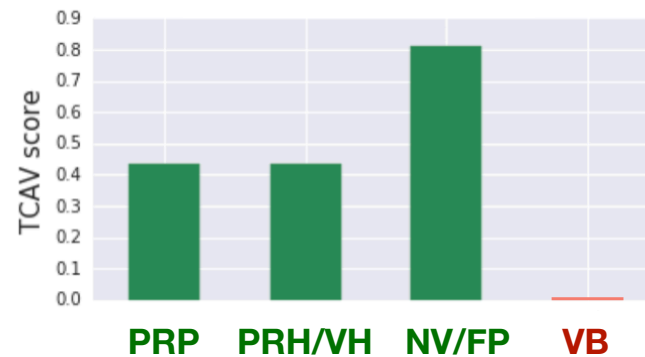
Prediction class Prediction accuracy

DR level 4 High

Example



TCAV scores

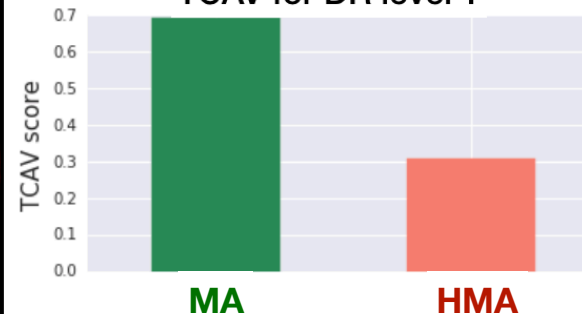


TCAV shows the model is **consistent** with doctor's knowledge when model is **accurate**

DR level 1 Med



TCAV for DR level 1



TCAV shows the model is **inconsistent** with doctor's knowledge for classes when model is less accurate

Green: domain expert's label on concepts belong to the level

Red: domain expert's label on concepts does not belong to the level

TCAV for Diabetic Retinopathy

Prediction class
Prediction accuracy

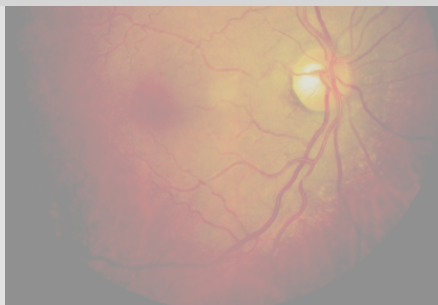
Example

DR level 4

Hi

DR level 1

Low



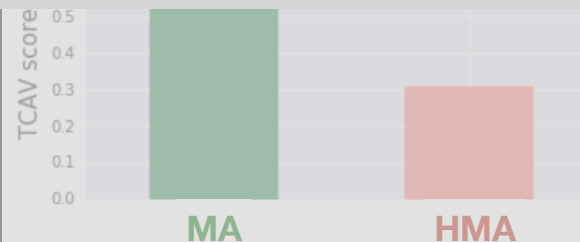
Level 1 was often confused to level 2.

HMA distribution on predicted DR

el 4

Goal of interpretability:
To use machine learning **responsibly**
we need to ensure that

1. our **values** are aligned
2. our **knowledge** is reflected



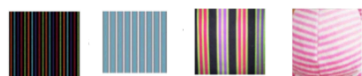
TCAV shows the model is **inconsistent** with doctor's knowledge for classes when model is less accurate


Green: domain expert's label on concepts belong to the level

Red: domain expert's label on concepts does not belong to the level

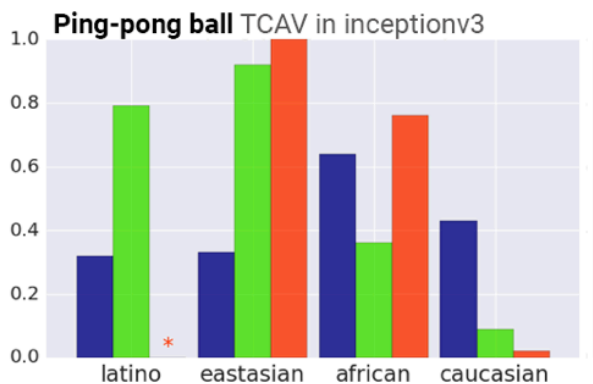
Summary:

Testing with Concept Activation Vectors



stripes concept (score: 0.9)
was important to **zebra** class
for this trained network. 

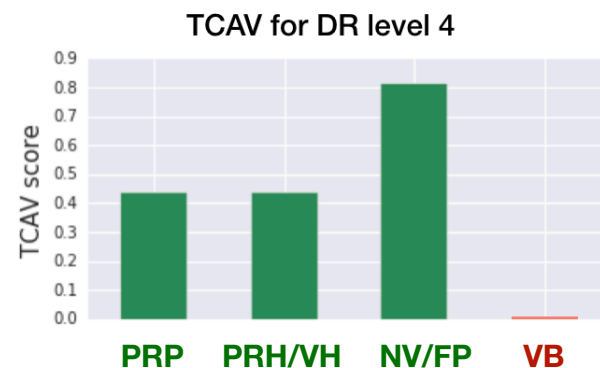
TCAV provides
quantitative importance of
a concept **if and only if** your
network learned about it.



Our values

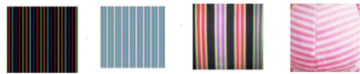



Our knowledge



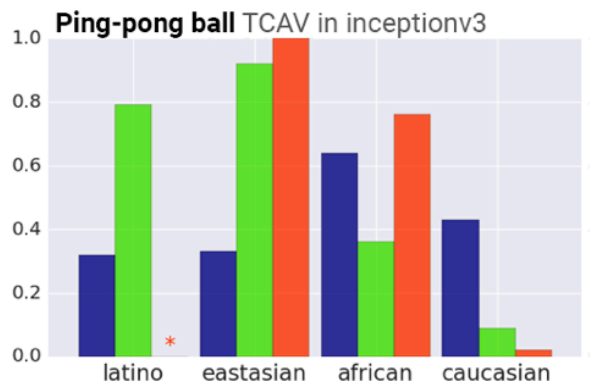
Questions?

code: github.com/tensorflow/tcav

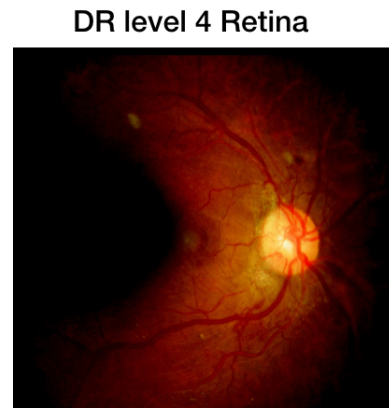


stripes concept (score: 0.9)
was important to **zebra** class
for this trained network. 

TCAV provides
quantitative importance of
a concept **if and only if** your
network learned about it.



Our values



Our knowledge

