Interpretability: what now?

Been Kim

Presenting work with a lot of awesome people inside and outside of Google

Julius Adebayo, Sherry Yang, Justin Gilmer, Martin Wattenberg, Carrie Cai, James Wexler, Fernanda Viegas, Rory Sayres, Ian Goodfellow, Mortiz Hardt, Michael Muelly
Sea of interpretability
Sea of interpretability
1. where are we going?
2. What do we have now?
3. What can we do better?
4. What should we be careful?
Sea of interpretability

1. where are we going?
My goal

interpretability

To use machine learning responsibly we need to ensure that
1. our values are aligned
2. our knowledge is reflected
My goal

interpretability

To use machine learning responsibly we need to ensure that
1. our values are aligned
2. our knowledge is reflected for everyone.
NON-goals

Interpretability is NOT...

• about making ALL models interpretable.

• about understanding EVERY SINGLE BIT about the model.

• against developing highly complex models.

• only about gaining user trust or fairness.
NON-goals

Interpretability is NOT…

• about making ALL models interpretable.
• about understanding EVERY SINGLE BIT about the model
• against developing highly complex models.
• only about gaining user trust or fairness
1. where are we going?
2. What do we have now?
Investigating post-training interpretability methods.

A trained machine learning model (e.g., neural network) \( p(z) \)

Junco Bird-ness

Given a fixed model, find the **evidence** of **prediction**.

Why was this a Junco bird?

Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]
One of the most popular interpretability methods for images:

**Saliency maps**

Caaaaaan do! We’ve got saliency maps to measure importance of each pixel!

\[
\text{a logit pixel } i,j \rightarrow \frac{\partial p(z)}{\partial x_{i,j}}
\]
One of the most popular interpretability methods for images:

**Saliency maps**

A trained machine learning model (e.g., neural network)

\[ p(z) \]

Junco Bird-ness

The promise: these pixels are the evidence of prediction.
Sanity check question.

A trained machine learning model (e.g., neural network) → \( p(z) \)

Junco Bird-ness

The promise: these pixels are the **evidence** of prediction.
Sanity check question.

A trained machine learning model (e.g., neural network) → \( p(z) \)

If so, when prediction changes, the explanation should change.

Extreme case: If prediction is random, the explanation should REALLY change.

Junco Bird-ness

The promise: these pixels are the evidence of prediction.

Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]
Some confusing behaviors of saliency maps.
Some confusing behaviors of saliency maps.

Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NeurIPS 18]
Some confusing behaviors of saliency maps.

Randomized weights!
Network now makes garbage prediction.

Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]
Some confusing behaviors of saliency maps.

Randomized weights!
Network now makes garbage prediction.

Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]
Sanity check 1:
When prediction changes, do explanations change?

No!

Original Image

Original Explanation

Cascading randomization from top to bottom layers

Gradient

Gradient-SG

Gradient ⊙ Input

Guided Back-propagation

GradCAM

Guided GradCAM

Integrated Gradients

Integrated Gradients-SG
Cascading randomization from top to bottom layers for VGG-16.
Sanity check 2: Networks trained with true and random labels, Do explanations deliver different messages? No!

Networks trained with….
What can we learn from this?

- **Confirmation bias**: Just because it “makes sense” to humans, doesn’t mean it reflects the evidence for prediction.

- Others who independently reached the same conclusions:
  [Nie, Zhang, Patel ’18] [Ulyanov, Vedaldi, Lempitsky ’18]

- Some of these methods have been shown to be useful for humans. Why? More studies needed.

- Recent work by Gupta and Arora 19’ suggests a simple fix
This was a low bar test.

Can we put interpretability methods on a harder test?
Benchmarking interpretability methods (BIM)

work with Sherry Yang
Benchmarking interpretability methods (BIM)

Forest

A thing

work with Sherry Yang
Benchmarking interpretability methods (BIM)

Forest

A thing

Forest

Bedroom

Kitchen

work with Sherry Yang
Benchmarking interpretability methods (BIM) is NOT important for predicting scene classes. Should NOT be part of explanation.
Benchmarking interpretability methods (BIM)

Forest

Bedroom

Kitchen

is NOT important for predicting scene classes.

should NOT be part of explanation

We can also make more important to some classes by controlling when it appears.

should be more important explanation in some classes than others.

github.com/google-research-datasets/bim
Three metrics for measuring false positives

<table>
<thead>
<tr>
<th>Interp. methods estimates</th>
<th>Model's truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>important</td>
</tr>
<tr>
<td>important</td>
<td>TP</td>
</tr>
<tr>
<td>not important</td>
<td>FN</td>
</tr>
</tbody>
</table>

Our Focus

github.com/google-research-datasets/bim
Three metrics for measuring false positives

<table>
<thead>
<tr>
<th>Interp. methods estimates</th>
<th>Model’s truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>important</td>
</tr>
<tr>
<td>important</td>
<td>TP</td>
</tr>
<tr>
<td>not important</td>
<td>FN</td>
</tr>
</tbody>
</table>

Suggested metrics
- Model contrast score (MCS)
- Input dependence rate (IDR)
- Input independence rate (IIR)

github.com/google-research-datasets/bim
Three metrics for measuring false positives

<table>
<thead>
<tr>
<th>Interp. methods estimates</th>
<th>Model’s truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>important</td>
</tr>
<tr>
<td>important</td>
<td>TP</td>
</tr>
<tr>
<td>not important</td>
<td>FN</td>
</tr>
</tbody>
</table>

Suggested metrics
- Model contrast score (MCS)
- Input dependence rate (IDR)
- Input independence rate (IIR)

Two models trained to classify scenes.

Model 1

Model 2

github.com/google-research-datasets/bim
Three metrics for measuring false positives

<table>
<thead>
<tr>
<th>Interp. methods estimates</th>
<th>Model’s truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>important</td>
</tr>
<tr>
<td>important</td>
<td>TP</td>
</tr>
<tr>
<td>not important</td>
<td>FN</td>
</tr>
</tbody>
</table>

Suggested metrics
- Model contrast score (MCS)
- Input dependence rate (IDR)
- Input independence rate (IIR)

We expect big contrast on where the object is.

Two models trained to classify

Scene model

Object model

github.com/google-research-datasets/bim
Model Contrast Score (MCS)

github.com/google-research-datasets/bim
1. Where are we going?

2. What do we have now?

3. What can we do better?
Problem: 
Post-training explanation

\[
\arg\max_{E} \mathcal{Q}(E | \text{Explanation}, \text{Model}, \text{Human}, \text{Data}, \text{Task})
\]

A trained machine learning model (e.g., neural network)

\[p(z)\]

cash-machine-ness

Why was this a cash machine?

TCAV [ICML’18]
Joint work with Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres
Common solution: Saliency map

Let’s use this to help us think about what we really want to ask.

https://pair-code.github.io/saliency/
SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg ’17]
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 ‘wheels’ concept matter?

prediction: Cash machine

https://pair-code.github.io/saliency/
SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg ’17]
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 ‘wheels’ concept matter?

Which concept mattered more?

Is this true for all other cash machine predictions?

https://pair-code.github.io/saliency/
SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg ’17]
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 ‘wheels’ concept 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!

https://pair-code.github.io/saliency/
SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg ’17]
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 ‘wheels’ concept 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?

prediction: Cash machine

https://pair-code.github.io/saliency/

SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg ’17]
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

A trained machine learning model (e.g., neural network) produces a probability $p(z)$ of the image being a zebra. The goal is to determine if being zebra-ness was important to the classifier:

Was striped concept important to this zebra image classifier?
Goal of TCAV: Testing with Concept Activation Vectors

A trained machine learning model (e.g., neural network)

Was striped concept important to this zebra image classifier?

TCAV provides quantitative importance of a concept if and only if your network learned about it.
TCAV:
Testing with Concept Activation Vectors

A trained machine learning model (e.g., neural network)

Was striped concept important to this zebra image classifier?

1. Learning CAVs

1. How to define concepts?
Defining concept activation vector (CAV)

**Inputs:**

- Examples of concepts
- Random images
- A trained network under investigation and internal tensors
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?

A trained machine learning model (e.g., neural network)

zrba-ness

Was striped concept important to this zebra image classifier?
TCAV core idea:
Derivative with CAV to get prediction sensitivity

\[
\text{zebra-ness} \rightarrow \frac{\partial p(z)}{\partial v^l_C} = S_{C,k,l}(x)
\]

Directional derivative with CAV
TCAV core idea:
Derivative with CAV to get prediction sensitivity

TCAV

\[
TCAV_{q_{c,k,l}} = \frac{|\{x \in X_k : S_{c,k,l}(x) > 0\}|}{|X_k|}
\]

Directional derivative with CAV
TCAV: Testing with Concept Activation Vectors

A trained machine learning model (e.g., neural network)

$\mathbf{p}(z)$

zebra-ness

Was striped concept important to this zebra image classifier?

1. Learning CAVs

2. Getting TCAV score
TCAV: Testing with Concept Activation Vectors

A trained machine learning model (e.g., neural network) produces a prediction $p(z)$ for zebra-ness. TCAV helps understand whether the striped concept was important to this zebra image classifier.

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

\[ \text{Zebra} \]
Quantitative validation:
Guarding against spurious CAV

Zebra

\[ \text{TCAV}_{Q_C,k,l} \]

\[ \vdots \]

\[ \text{TCAV}_{Q_C,k,l} \]

\[ \vdots \]
Quantitative validation:

Guarding against spurious CAV

Check the distribution of TCAV\textsubscript{QC,k,l} is statistically different from random using t-test.
How to choose a layer

Start from the top layer (closest to prediction).

Go down a layer if the current layer doesn’t pass the statistical testing.
Recap TCAV:
Testing with Concept Activation Vectors

1. Learning CAVs

2. Getting TCAV score

3. CAV validation

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

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

An image

+ Potentially noisy Caption

models can use either image or caption concept for classification.
Sanity check experiment setup

Four models trained with different caption noise levels

An image + Potentially noisy Caption

models can use either image or caption concept for classification.

Four models trained with different caption noise levels

0% noisy 30% noisy 100% noisy no captions
Sanity check experiment setup

Test models with no caption image.

Test accuracy = Importance of image concept

Four models trained with different caption noise levels
Sanity check experiment

Test accuracy with no caption image

Caption noise level in training set

Caption noise level in training set
Sanity check experiment

Test accuracy with no caption image

Cab class never cared about the caption!
Cool, cool.
Can saliency maps do this too?
Can saliency maps communicate the same information?

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Model trained on</th>
<th>Image with caption</th>
<th>Vanilla gradient</th>
<th>Guided backprop</th>
<th>Integrated gradient</th>
<th>Smoothgrad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image concept</td>
<td>Images without captions (no captions)</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td>Image concept</td>
<td>Images with captions (0% noise)</td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td>Image concept</td>
<td>Images with captions (30% noise)</td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
</tr>
<tr>
<td>Image concept</td>
<td>Images with captions (100% noise)</td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
</tr>
</tbody>
</table>
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!

Fire engine TCAV in googlenet

Ping-pong ball TCAV in inceptionv3

Rugby ball TCAV in googlenet

http://www.abc.net.au
TCAV in Two widely used image prediction models

Geographical bias?

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

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

TCAV in
Two widely used image prediction models

Goal of interpretability:
To use machine learning **responsibly** we need to ensure that
1. our **values** are aligned
2. our **knowledge** is reflected
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

80
Collect human doctor’s knowledge

<table>
<thead>
<tr>
<th>DR level 4</th>
<th>Concepts belong to this level</th>
<th>Concepts do not belong to this level</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP</td>
<td>PRH/VH</td>
<td>VB</td>
</tr>
<tr>
<td>NV/FP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DR level 1</th>
<th>Concepts belong to this level</th>
<th>Concepts do not belong to this level</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA</td>
<td></td>
<td>HMA</td>
</tr>
</tbody>
</table>
TCAV for Diabetic Retinopathy

Prediction class | Prediction accuracy
--- | ---
DR level 4 | High

**Example**

**TCAV scores**

<table>
<thead>
<tr>
<th>Concept</th>
<th>TCAV score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP</td>
<td>0.4</td>
</tr>
<tr>
<td>PRH/VH</td>
<td>0.5</td>
</tr>
<tr>
<td>NV/FP</td>
<td>0.9</td>
</tr>
<tr>
<td>VB</td>
<td>0.0</td>
</tr>
</tbody>
</table>

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

### DR level 4 (High)
- **Prediction class**: DR level 4
- **Prediction accuracy**: High

**Example**

**TCAV scores**
- PRP
- PRH/VH
- NV/FP
- VB

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

### DR level 1 (Med)
- **Prediction class**: DR level 1
- **Prediction accuracy**: Med

**Example**

**TCAV for DR level 1**
- MA
- HMA

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

<table>
<thead>
<tr>
<th>Prediction class</th>
<th>Prediction accuracy</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR level 4</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>DR level 1</td>
<td>Low</td>
<td></td>
</tr>
</tbody>
</table>

Goal of interpretability:
To use machine learning **responsibly** we need to ensure that
1. our **values** are aligned
2. our **knowledge** is reflected

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

Level 1 was often confused to level 2.
Summary:

Testing with Concept Activation Vectors

Joint work with Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres

text:

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
Responses from outside of academia

Sundar (CEO of Google) explaining how TCAV works in his keynote at Google I/O 2019

UNESCO NetExplo award 2019
Selected as one of ten “cutting-edge digital innovations with the potential of profound and lasting impact.”
Responses from inside of academia

Using CAVs to help doctors find more diagnostically relevant images
CHI conference, best paper honorable mention

TCAV for storm prediction models
“Interpretable AI for deep-learning based meteorological applications” Work by Eric Wendoloski,

Extending TCAV to regression models
“Regression Concept Vectors for Bidirectional Explanations in Histopathology” Work by Mara Graziani et al.
1. Where are we going?
2. What do we have now?
3. What can we do better?
Limitations of TCAV

• Concept has to ‘expressible’ using examples (e.g., “love” concept might be hard).

• User needs to know which concepts they want to test, and have examples for it. Follow-up work to automatically discover concepts for images (submitted), but many more directions are possible.

• Explanations provided by TCAV are not-causal - Follow-up work on causal TCAV (submitted)
1. Where are we going?
2. What do we have now?
3. What can we do better?
4. What should we be careful?
Things to keep in mind during our journey.

- Proper evaluations
  - Sanity check and ground-truth-based evaluations
  - Test with humans!
- Remember that humans are biased and irrational.
- Designing the right interaction - HCI.
- Try to criticize - think about what wasn’t talked about in this talk but should have!
- Keep checking if we are going to the right direction!
1. Where are we going?
   Tool that can help more responsible AI

2. What do we have now?
   Some existing methods fail a simple sanity check.

3. What can we do better?
   TCAV (btw it passes sanity check)

4. What should we be careful?
   Evaluation
   HCI
   Human biases.