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
1. Where are we going?
2. What do we have now?
3. What can we do better?
4. What should we be careful?
1. where are we going?

Sea of interpretability
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
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!

\[
\text{a logit} \quad \rightarrow \quad \frac{\partial p(z)}{\partial x_{i,j}}
\]

picture credit: @sayres

SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg ’17]
Integrated gradient [Sundararajan, Taly, Yan ’17]
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 Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]
Sanity check question.

The promise: these pixels are the evidence of prediction.

Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]
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.

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

Extreme case: If prediction is random, the explanation should REALLY change.
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!**
Sanity check 2:
Networks trained with true and random labels,
Do explanations deliver different messages?
\textbf{No!}

Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]
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.
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)

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

- Forest
- Bedroom
- Kitchen

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

coming soon! data, model and metrics for existing interpretability methods

github.com/google-research-datasets/bim work with Sherry Yang
1. Where are we going?
2. What do we have now?
3. What can we do better?
Problem: Post-training explanation

\[
\text{argmax } Q(E \mid M, H, D, T)
\]

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

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?

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?
- 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?

https://pair-code.github.io/saliency/
SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg ’17]
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 striped concept important to this zebra image classifier?
Goal of TCAV: Testing with Concept Activation Vectors

Was striped concept important to this zebra image classifier?

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

$A(z)$

zebra-ness

TCAV score for striped not striped

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

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)

zebra-ness

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

**TCAV**

zebra-ness \( z \) \( \frac{\partial p(z)}{\partial \nu^l_C} = S_{C,k,l}(x) \)

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

TCAV

TCAV score

zebra-ness → \frac{\partial p(z)}{\partial v_C^l} = S_{C,k,l}(x)

\begin{align*}
S_{C,k,l}(\text{zebra}) \\
S_{C,k,l}(\text{striped}) \\
S_{C,k,l}(\text{zig-zagged})
\end{align*}

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

Directional derivative with CAV
TCAV: Testing with Concept Activation Vectors

A trained machine learning model (e.g., neural network) is used to predict the zebra-ness of an image. To determine if the striped concept was important to this zebra image classifier, TCAV (Test with Concept Activation Vectors) is applied.

1. Learning CAVs
2. Getting TCAV score

The TCAV score is calculated as the difference between the output of the model with and without the concept activation vector.

\[ TCAV_{C,k,l} = f_{C,k,l}(z) - f_{C,k,l}(\text{without concept}) \]
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
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

\[ TCAV_{Q_C,k,l} : \]

\[ \vdots \]

\[ TCAV_{Q_C,k,l} : \]

\[ \vdots \]
Quantitative validation:

Guarding against spurious CAV
Quantitative validation:

Guarding against spurious CAV

Check the distribution of TCAV_{QC,k,l} is statistically different from random using t-test
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

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
Evaluating with BIM dataset

- Forest
- Bedroom
- Kitchen
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

- **Fire engine** TCAV in googlenet

- **Ping-pong ball** TCAV in inceptionv3

- **Rugby ball** TCAV in googlenet

- **Dumbbell** TCAV in inceptionv3
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]
TCAV in
Two widely used image prediction models

Geographical bias?

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
Collect human doctor’s knowledge

<table>
<thead>
<tr>
<th>Concepts belong to this level</th>
<th>Concepts do not belong to this level</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR level 4</td>
<td>PRP, PRH/VH, NV/FP</td>
</tr>
<tr>
<td></td>
<td>VB</td>
</tr>
<tr>
<td>DR level 1</td>
<td>MA</td>
</tr>
<tr>
<td></td>
<td>HMA</td>
</tr>
</tbody>
</table>
TCAV for Diabetic Retinopathy

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

Example

TCAV scores

**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 **consistent** with doctor’s knowledge when model is **accurate**
TCAV for Diabetic Retinopathy

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

**Example**

**TCAV scores**

- PRP
- PRH/VH
- NV/FP
- VB

**TCAV for DR level 1**

- MA
- HMA

**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 **consistent** with doctor’s knowledge when model is **accurate**

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

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

**Example**

- **Level 1 was often confused to level 2.**
- **HMA distribution on predicted DR**

**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.
Summary:

Testing with Concept Activation Vectors

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

**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

ICML 2018
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

Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making

Work by Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, K., Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, Michael Terry CHI conference, best paper honorable mention

Extending TCAV to regression models

“Regression Concept Vectors for Bidirectional Explanations in Histopathology”
Work by Mara Graziani, Vincent Andrearczyk, Henning Muller
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.

- Importance of 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.