Interpretability: the myth, questions, and some answers.

Been Kim

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

Martin Wattenberg, Finale Doshi-Velez, Julius Adebayo, Heinrich Jiang, Maya Gupta, Ike Lage, Andrew Ross, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, Rory Sayres, Ian Goodfellow, Mortiz Hardt, Sam Gershman, Menaka Narayanan, Emily Chen, Jeffrey He
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.
ingredients for interpretability methods.

$$\arg\max_{E} Q(E | ?)$$

Some quality function
\[ \text{argmax}_{E} Q(E|M, \ ) \]
argmax \ Q(E|M,\ E)
\[
\argmax_{E} \mathcal{Q}(E|M, D)
\]
\[
\arg\max_E Q(E|M, D, )
\]
If I were you, I would train a neural network.

What's ML?

If I were you, I would train a neural network.
argmax\limits_{E} Q(E|M, H, D, T)

- Local vs. global
- Simple explanations vs. more complex but more accurate explanations
- Low or high stake domains
Agenda

Post-training explanations

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

Building inherently interpretable models

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

$$\arg\max_{E} Q(\text{Explanation} | \text{Model}, \text{Human}, \text{Data}, \text{Task})$$
Agenda

1. Revisit some existing methods: Sanity check questions
2. Make explanations that work for lay people.
3. Understand how humans understand explanations
4. Make explanations to detect trustworthy predictions.
1. Revisit some existing methods: Sanity check questions

2. Make explanations that work for lay people.

3. Understand how humans understand explanations

4. Make explanations to detect trustworthy predictions.

\[
\arg\max_{E} Q(\text{Explanation} | \text{Model, Human, Data, Task})
\]
Problem: Post-training explanation

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

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

Why was this a cash machine?
One of the most popular interpretability methods for images: 

**Saliency maps**

SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg ’17]
Integrated gradient [Sundararajan, Taly, Yan ’17]

Used for image classification and medical applications.

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

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

Humans’ subjective judgement widely used for images

local understanding

picture credit: @sayres
One of the most popular interpretability methods for images: **Saliency maps**

SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg '17]
Integrated gradient [Sundararajan, Taly, Yan '17]

- **Saliency maps**
  - Used for image classification and medical applications.
  - A logit pixel $i,j \rightarrow \frac{\partial p(z)}{\partial x_{i,j}}$

**Sanity check:**
If I change $M$ a lot, will human perceive that $E$ has changed a lot?
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, [NeurIPS 18]
Some saliency maps look similar when we randomize the network.

Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NeurIPS 18]
What can we learn from this?

- Potential human confirmation bias: Just because it “makes sense” to humans, doesn’t mean they reflect evidence for the prediction.

- Our discovery is consistent with other findings [Nie, Zhang, Patel ‘18] [Ulyanov, Vedaldi, Lempitsky ‘18]

- Some of these methods have been shown to be useful in practice. Explaining predictions or features? More studies needed.

Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NeurIPS 18]
What can we do better?
Creating a wishlist.

\[
\arg\max_{E} Q(\text{Explanation}|\text{Model, Human, Data, Task})
\]
What can we do better?
Creating a wishlist.

Something more human-friendly?

quantitative
- human’s
- subjective
- judgement

Using
- input
- features
as a language

lay person?!

global
- local
- understanding

argmax $Q(E|\text{Model, Human, Data, Task})$
Agenda

1. Revisit some existing methods:
   Sanity check questions

2. Make explanations that work for lay people.

3. Understand how humans understand explanations

4. Make explanations to detect trustworthy predictions.

\[
\arg\max_{E} Q(\text{Explanation}|\text{Model, Human, Data, Task})
\]
Problem: Post-training explanation

$$\operatorname{argmax}_E Q(\text{Explanation} | \text{Model, Human, Data, Task})$$

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

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]
prediction: Cash machine

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]
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)

Doctor-ness

Was gender concept important to this doctor image classifier?

TCAV score for Doctor

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

A trained machine learning model (e.g., neural network) was important to this zebra image classifier?

TCAV score for zebra-ness

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:**
- Examples of concepts
- Random images
- A trained network under investigation and Internal tensors

Diagram:
- Function $f_i : \mathbb{R}^n \rightarrow \mathbb{R}^m$
- Input images map to activation vectors in multiple layers leading to a classification decision.
Defining concept activation vector (CAV)

Inputs:

Train a linear classifier to separate activations.

CAV ($v^l_C$) is the vector **orthogonal** to the decision boundary.

[Smilkov ‘17, Bolukbasi ‘16, Schmidt ‘15]
TCAV: Testing with Concept Activation Vectors

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

$p(z)$

zebra-ness

1. Learning CAVs

Was striped concept important to this zebra image classifier?

2. Getting TCAV score

$S_{C,k,l}(\text{zebra})$

$S_{C,k,l}(\text{bird})$

$S_{C,k,l}(\text{zebra})$

$\text{TCAV}_{Q,C,k,l}$

2. How are the CAVs useful to get explanations?
TCAV core idea:
Derivative with CAV to get prediction sensitivity

TCAV

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

Directional derivative with CAV

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

A trained machine learning model (e.g., neural network) was used to classify a zebra image. Was the striped concept important to this zebra image classifier?

1. Learning CAVs
2. Getting TCAV score
   \[ S_{C,k,l}(\text{zebra}) \]
   \[ S_{C,k,l}(\text{stripes}) \]
   \[ \text{TCAV}_{Q,C,k,l} : \]
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

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

1. Learning CAVs
2. Getting TCAV score
3. CAV validation

Qualitative
Quantitative
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

Test accuracy with no caption image = Importance of image concept

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

Test accuracy with no caption image

Caption noise level in training set

Caption noise level in training set
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><strong>Image concept</strong></td>
<td>Images without captions (no captions)</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
</tr>
<tr>
<td><strong>Image concept</strong></td>
<td>Images with captions (0% noise)</td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
</tr>
<tr>
<td><strong>Image concept</strong></td>
<td>Images with captions (30% noise)</td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
</tr>
<tr>
<td><strong>Image concept</strong></td>
<td>Images with captions (100% noise)</td>
<td><img src="image16" alt="Image" /></td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" 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

Geographical bias!
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>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></td>
<td>PRP</td>
<td>VB</td>
</tr>
<tr>
<td></td>
<td>PRH/VH NV/FP</td>
<td></td>
</tr>
</tbody>
</table>

| DR level 1 | MA                            | HMA                                 |

Concepts belong to DR level 4 include PRP, PRH/VH, NV/FP. Concepts do not belong to DR level 4 include VB. Concepts belong to DR level 1 include MA. Concepts do not belong to DR level 1 include HMA.
# TCAV for Diabetic Retinopathy

<table>
<thead>
<tr>
<th>Prediction class</th>
<th>Prediction accuracy</th>
<th>Example</th>
<th>TCAV scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR level 4</td>
<td>High</td>
<td><img src="image1" alt="Example Image" /></td>
<td><img src="chart1" alt="TCAV Scores" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR level 1</td>
<td>Med</td>
<td><img src="image2" alt="Example Image" /></td>
<td><img src="chart2" alt="TCAV Scores" /></td>
</tr>
</tbody>
</table>

**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

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

Our values

<table>
<thead>
<tr>
<th>Stripe Type</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latino</td>
<td>0.2</td>
</tr>
<tr>
<td>East Asian</td>
<td>0.6</td>
</tr>
<tr>
<td>African</td>
<td>0.8</td>
</tr>
<tr>
<td>Caucasian</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Our knowledge

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

TCAV for DR level 4

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

Joint work with Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres
1. Revisit some existing methods: Sanity check questions

2. Make explanations that work for lay people.

3. Understand how humans understand explanations

4. Make explanations to detect trustworthy predictions.
What makes explanations hard or easy for humans?

• How do humans’ understanding changes as we vary factors in explanations? argmax \( Q(E|M, H, D, T) \)

• Among many explanations, we choose a rule-set.

• Among many factors, we choose a subset based on what prior literatures assumed to matter.
What makes explanations hard or easy for humans?

- How do humans’ understanding changes as we vary factors in explanations?

\[
\arg\max_{E} Q(E|M, H, D, T)
\]

We vary these factors:
- Explanation length
- Number of cognitive chunks
- Variable repetition

Humans do these tasks:
- Counterfactual verification
- Simulation

Measures of interpretability:
- Human’s accuracy
- Subjective score
- Time took

How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation joint work with Narayanan, Chen, He, Gershman, and Doshi-Velez 2017
Controlling for prior knowledge.

Using a made-up ‘alien world’ to control for prior knowledge.

---


joint work with Narayanan, Chen, He, Gershman, and Doshi-Velez 2017
(a small subset of) results

Variable repetition mattered less for accuracy than other factors.

*all repeated variables are needed for task completion


joint work with Narayanan, Chen, He, Gershman, and Doshi-Velez 2017
Agenda

1. Revisit some existing methods: Sanity check questions
2. Make explanations that work for lay people.
3. Understand how humans understand explanations
4. Make explanations to detect trustworthy predictions.

\[
\underset{E}{\text{argmax }} Q(\text{Explanation} | \text{Model, Human, Data, Task})
\]
Ultimate goal is to use ML more responsibly.

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

$$\arg\max_E Q(E|M, H, D, T)$$

Simply confidence scores

Not using the classifier when it's suspicious.

Improve confidence measure coming from a classifier

**Problem:**
- Precision of "definitely trustworthy (correct)" predictions
- "definitely suspicious (incorrect)" predictions

To trust or not to trust a classifier

joint work with Jiang and Gupta [NeurIPS 2018]
Trust score:
a super simple method

was predicted as class A. Can we trust this?

\[
\text{Trust score:} = \frac{d \left( x, \widehat{H}_\alpha(f_{\tilde{h}(x)}) \right)}{d \left( x, \widehat{H}_\alpha(f_h(x)) \right)}
\]

We can use activations instead of input data in NN!
results:
We can detect trustworthy and suspicious predictions with high precision.

Detect trustworthy (correct)

Detect suspicious (incorrect)

Theoretical results: why does this work?

The trust score reveals the signal from a Bayes optimal classifier (with high probability).

To trust or not to trust a classifier
joint work with Jiang and Gupta [NeurIPS 2018]
Summary, future work

1. Revisit some existing methods:
   Sanity check questions
   Sanity Checks for Saliency Maps
   Joint work with Adebayo, Gilmer, Goodfellow, Hardt
   NIPS 2018

2. Make explanations that work for lay people.
   TCAV: Testing with concept activation vectors
   Joint work with Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres
   ICML 2018

3. Understand how humans understand explanations
   How do Humans Understand Explanations from Machine Learning Systems?
   An Evaluation of the Human-Interpretability of Explanation
   joint work with Narayanan, Chen, He, Gershman, and Doshi-Velez 2017

4. Make explanations to detect trustworthy predictions.
   To trust or not to trust a classifier
   joint work with Jiang and Gupta
   NIPS 2018

Understanding superhuman performance networks
Understanding models under production @ Google
Detect ‘different types of mistakes’ that a model makes.
...lots of others.