Interpretability in Machine Learning
Why Interpret ?
The current state of machine learning

Movies of the 90s: soon AI will take over the world

AI now:
And its uses ...

https://www.tesla.com/videos/autopilot-self-driving-hardware-neighborhood-long

DeepMind

NYPot

MIT Technology Review
So are we in the golden age of AI?
Safety and well being

Tesla hit parked police car ‘while using Autopilot’

Warnings of a Dark Side to A.I. in Health Care

A number of Tesla vehicles have been involved in crashes.

Scientists worry that with just tiny tweaks to data, neural networks can be fooled into committing “adversarial attacks” that mislead rather than help. Joon Coss/NurPhoto, via Getty Images
Bias in algorithms

August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark

98.7% 68.6% 100% 92.9%

Amazon Rekognition Performance on Gender Classification


Machine Learning can amplify bias.

- Data set: 67% of people cooking are women
- Algorithm predicts: 84% of people cooking are women

https://www.infoq.com/presentations/unconscious-bias-machine-learning/
Adversarial Examples

Original image  Perturbations  Adversarial example
Temple (97%)          Ostrich (98%)
Legal Issues - GDPR

Starting May 25, the European Union will require algorithms to explain their output, making deep learning illegal.
And more ...

- Interactive feedback - can model learn from human actions in online setting? (Can you tell a model to not repeat a specific mistake?)

- Recourse – Can a model tell us what actions we can take to change its output? (For example, what can you do to improve your credit score?)
In general, it seems like there are few fundamental problems –

• We don’t trust the models
• We don’t know what happens in extreme cases
• Mistakes can be expensive / harmful
• Does the model makes similar mistakes as humans?
• How to change model when things go wrong?

Interpretability is one way we try to deal with these problems
What is interpretability?
There is no standard definition –

Most agree it is something different from performance.

Ability to explain or to present a model in understandable terms to humans (Doshi-Velez 2017)

Cynical view – It is what makes you feel good about the model.

It really depends on target audience.
What does interpretation looks like?

In pre-deep learning models, some models are considered “interpretable”
What does interpretation look like?

- Heatmap Visualization

Figure 3. Attribution for Diabetic Retinopathy grade prediction from a retinal fundus image. The original image is shown on the left, and the attributions (overlayed on the original image in gray scaece) is shown on the right. On the original image we annotate lesions visible to a human, and confirm that the attributions indeed point to them.

Table 2: Gate activations for each aspect in a PICO abstract. Note that because gates are calculated a the final convolution layer, activations are not in exact 1-1 correspondence with words.

[Sundarajan 2017] [Jain 2018]
What does interpretation look like?

- Give prototypical examples


[Kim 2016]
What does interpretation look like?

- Bake it into the model

---

[Bastings et al 2019]
What does interpretation looks like?

- Provide explanation as text

---

| Question: | While eating a hamburger with friends, what are people trying to do? |
| CoS-E: | Usually a hamburger with friends indicates a good time. |

| Question: | After getting drunk people couldn’t understand him, it was because of his what? |
| CoS-E: | People who are drunk have difficulty speaking. |

| Question: | People do what during their time off from work? |
| CoS-E: | People usually do something relaxing, such as taking trips, when they don’t need to work. |

Table 1: Examples from our CoS-E dataset.

---

[Example]
Both cohorts showed signs of optic nerve toxicity due to ethambutol.

[Label]
Does this chemical cause this disease?

[Explanation]
Because the words “due to” occur between the chemical and the disease.

[Labeling Function]
```
def f(x):
    return 1 if “due to” in between(x.chemical, x.disease) 
    else 0
```
Some properties of Interpretations

**Faithfulness** - how to provide explanations that accurately represent the true reasoning behind the model’s final decision.

**Plausibility** – Is the explanation correct or something we can believe is true, given our current knowledge of the problem?

**Understandable** – Can I put it in terms that end user without in-depth knowledge of the system can understand?

**Stability** – Does similar instances have similar interpretations?
Evaluating Interpretability [Doshi-Velez 2017]

• Application level evaluation – Put the model in practice and have the end users interact with explanations to see if they are useful.

• Human evaluation – Set up a Mechanical Turk task and ask non-experts to judge the explanations.

• Functional evaluation – Design metrics that directly test properties of your explanation.
Categorizing Interpretability Methods
Global vs Local

Do we explain individual prediction?
Example –
Heatmaps
Rationales

Can we find what model has learned in general about the task?
Example –
Prototypes
Linear Regression
Decision Trees
Inherent vs Post-hoc

Is the explainability built into the model?

Example –

Rationales
Linear Regression
Decision Trees
Natural Language Explanations

Is the model black-box and we use external method to try to understand it?

Example –

Heatmaps (Some forms)
Prototypes
Some Locally Interpretable, Post-hoc methods
Saliency Based Methods

- Heatmap based visualization
- Need differentiable model in most cases
- Normally involve gradient
<table>
<thead>
<tr>
<th>Original Image</th>
<th>Gradient</th>
<th>SmoothGrad</th>
<th>Guided BackProp</th>
<th>Guided GradCAM</th>
<th>Integrated Gradients</th>
<th>Integrated Gradients</th>
<th>SmoothGrad</th>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junco Bird</td>
<td><img src="image1" alt="Junco Bird" /></td>
<td><img src="image2" alt="Junco SmoothGrad" /></td>
<td><img src="image3" alt="Junco Guided BackProp" /></td>
<td><img src="image4" alt="Junco Guided GradCAM" /></td>
<td><img src="image5" alt="Junco Integrated Gradients" /></td>
<td><img src="image6" alt="Junco Integrated Gradients" /></td>
<td><img src="image7" alt="Junco SmoothGrad" /></td>
<td><img src="image8" alt="Junco Input" /></td>
</tr>
<tr>
<td>Corn</td>
<td><img src="image9" alt="Corn Gradient" /></td>
<td><img src="image10" alt="Corn SmoothGrad" /></td>
<td><img src="image11" alt="Corn Guided BackProp" /></td>
<td><img src="image12" alt="Corn Guided GradCAM" /></td>
<td><img src="image13" alt="Corn Integrated Gradients" /></td>
<td><img src="image14" alt="Corn Integrated Gradients" /></td>
<td><img src="image15" alt="Corn SmoothGrad" /></td>
<td><img src="image16" alt="Corn Input" /></td>
</tr>
<tr>
<td>Wheaten Terrier</td>
<td><img src="image17" alt="Wheaten Terrier Gradient" /></td>
<td><img src="image18" alt="Wheaten Terrier SmoothGrad" /></td>
<td><img src="image19" alt="Wheaten Terrier Guided BackProp" /></td>
<td><img src="image20" alt="Wheaten Terrier Guided GradCAM" /></td>
<td><img src="image21" alt="Wheaten Terrier Integrated Gradients" /></td>
<td><img src="image22" alt="Wheaten Terrier Integrated Gradients" /></td>
<td><img src="image23" alt="Wheaten Terrier SmoothGrad" /></td>
<td><img src="image24" alt="Wheaten Terrier Input" /></td>
</tr>
</tbody>
</table>

[Adebayo et al 2018]
Saliency Example - Gradients

\[ f(x) : \mathbb{R}^d \rightarrow \mathbb{R} \]

\[ E(f)(x) = \frac{df(x)}{dx} \]

How do we take gradient with respect to words?

Take gradient with respect to embedding of the word.
Saliency Example – Leave-one-out

\[ f(x) : R^d \rightarrow R \]

\[ E(f)(x)_i = f(x) - f(x\setminus i) \]

How to remove?

1. Zero out pixels in image
2. Remove word from the text
3. Replace the value with population mean in tabular data
Problems with Saliency Maps

- Only capture first order information
- Strange things can happen to heatmaps in second order.

Figure 6: Heatmap generated with leave-one-out shifts drastically despite only removing the least important word (underlined) at each step. For instance, “advertisement”, is the most important word in step two but becomes the least important in step three.

[Feng et al 2018]
Sanity check:
When prediction changes, do explanations change?

Original Image

Saliency map

Randomized weights!
Network now makes garbage predictions.

(Slide Credit – Julius Adebayo)
LIME – locally interpretable model agnostic

\[ x^1, x^2, \ldots, x^N \rightarrow \text{Black Box} \rightarrow y^1, y^2, \ldots, y^N \]

(e.g. Neural Network)

\[ x^1, x^2, \ldots, x^N \rightarrow \text{Linear Model} \rightarrow y^1, y^2, \ldots, y^N \]

Can’t do it globally of course, but locally? Main Idea behind LIME

(Image Credit – Hung-yi Lee)
Intuition behind LIME

Figure 3: Toy example to present intuition for LIME. The black-box model’s complex decision function $f$ (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using $f$, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

[Ribeiro et al 2016]
1. Given a data point you want to explain
2. Sample at the nearby - Each image is represented as a set of superpixels (segments).

Randomly delete some segments.

Compute the probability of “frog” by black box

(Slide Credit – Hung-yi Lee)
LIME — Image

3. Fit with linear (or interpretable) model

\[ x_1 \quad \cdots \quad x_m \quad \cdots \quad x_M \]

\[ x_m = \begin{cases} 0 & \text{Segment } m \text{ is deleted.} \\ 1 & \text{Segment } m \text{ exists.} \end{cases} \]

\( M \) is the number of segments.

(Slide Credit – Hung-yi Lee)
4. Interpret the model you learned

\[ y = w_1 x_1 + \cdots + w_m x_m + \cdots + w_M x_M \]

- \( x_m = 0 \)  Segment m is deleted.
- \( x_m = 1 \)  Segment m exists.

\( M \) is the number of segments.

- If \( w_m \approx 0 \) \( \rightarrow \) segment m is not related to “frog”
- If \( w_m \) is positive \( \rightarrow \) segment m indicates the image is “frog”
- If \( w_m \) is negative \( \rightarrow \) segment m indicates the image is not “frog”

(Slide Credit – Hung-yi Lee)
Evaluating Faithfulness in Saliency based methods

How do we know if the model is actually using the features that it is telling us are important?

1. Does the prediction change if we remove the important features? (Comprehensiveness)

2. Does the prediction remain the same if we remove the un-important features? (Sufficiency)
Rationalization Models
General Idea

Extractor

Classifier

Tree frog (97%)

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. This is a real good lookin' beer, unfortunately it gets worse from here ...

this beer *pours ridiculously clear with tons of carbonation* that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. This is a real good lookin' beer, unfortunately it gets worse from here ...

Extractor

Classifier

Positive (98%)
FRESH Model – Faithful Rationale Extraction using Saliency Thresholding

(1) \textit{Train supp to score features (e.g., gradients, attention, LIME); discretize these}

\[ x_i \rightarrow y_i \rightarrow S_i \]
FRESH Model – Faithful Rationale Extraction using Saliency Thresholding

(2) Train $\text{ext}$ to extract snippets; use to create $\tilde{x}_i$
FRESH Model – Faithful Rationale Extraction using Saliency Thresholding

(3) \( \text{Train pred on } (\tilde{x}_i, y_i) \)
Some Results – Human Evaluation

Instructions

View full instructions

Select the sentiment that best describes the text and a score indicating how confident you are. Some of these will not make any sense. If you’re unsure, select any label and assign a confidence score of 0.

I believe that Robert Duvall (who is the producer, director, writer, and main star of the apostle) deserves an Oscar for his performance as Sonny the religious a performance which is so complex and realistic it ranks as one of the finest acting performances on offers. The audience a completely honest look at southern the apostle would rank as one of the best movies of this. I emphatically recommend the apostle for connoisseurs of stage and fine acting on film find the apostle a thought-provoking experience. The apostle is a four star.

What sentiment does this text convey?

- Positive
- Negative

How confident are you that your answer is correct?

- 0: I’m not confident. I guessed randomly.
- 1: I’m a little confident.
- 2: I’m pretty confident.
- 3: I’m very confident.

How easy is the text to read and understand?

- Very difficult.
- Difficult.
- Neutral.
- Easy.
- Very Easy.

Figure 8: Amazon Mechanical Turk layout for Movies tasks.
Some Results – Human Evaluation

<table>
<thead>
<tr>
<th>Rationale Source</th>
<th>Human Acc.</th>
<th>Confidence (1–4)</th>
<th>Readability (1–5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>.99</td>
<td>3.44 ±0.53</td>
<td>3.82 ±0.56</td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contiguous</td>
<td>.84</td>
<td>3.18 ±0.55</td>
<td>3.80 ±0.57</td>
</tr>
<tr>
<td>Non-Contiguous</td>
<td>.65</td>
<td>2.09 ±0.51</td>
<td>2.07 ±0.69</td>
</tr>
<tr>
<td><strong>Lei et al. 2016</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contiguous</td>
<td>.88</td>
<td>3.39 ±0.48</td>
<td>4.17 ±0.59</td>
</tr>
<tr>
<td>Non-Contiguous</td>
<td>.84</td>
<td>2.97 ±0.72</td>
<td>2.90 ±0.88</td>
</tr>
<tr>
<td><strong>Our Best</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contiguous</td>
<td>.92</td>
<td>3.31 ±0.48</td>
<td>3.88 ±0.57</td>
</tr>
<tr>
<td>Non-Contiguous</td>
<td>.87</td>
<td>3.23 ±0.47</td>
<td>3.63 ±0.59</td>
</tr>
</tbody>
</table>
Instance Attribution Methods

Given a test example, find training examples that are most relevant / influential
Deletion Metric

How will prediction on a test example change if I delete a training example?

How to calculate?

1. Remove each training point one by one
2. Retrain the model on remaining data
3. Evaluate on test point
4. Compute difference to original prediction
Approximate it? Influence Functions

How much will the loss on a test point change if I slightly upweight the training point?

How to calculate?

1. How much will parameters change on upweighting?

\[
\hat{\theta}_{\epsilon,z} = \arg \min_{\theta \in \Theta} (1 - \epsilon) \frac{1}{n} \sum_{i=1}^{n} L(z_i, \theta) + \epsilon L(z, \theta)
\]

\[
I_{\text{up,\text{params}}}(z) = \left. \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon} \right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta})
\]
Approximate it? Influence Functions

2. How much will loss of prediction change on parameter change?

\[ I_{up,loss}(z, z_{test}) = \frac{dL(z_{test}, \hat{\theta}_{\epsilon,z})}{d\epsilon} \bigg|_{\epsilon=0} \]

\[ = \nabla_{\theta} L(z_{test}, \hat{\theta})^T \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon} \bigg|_{\epsilon=0} \]

\[ = -\nabla_{\theta} L(z_{test}, \hat{\theta})^T H_{\theta}^{-1} \nabla_{\theta} L(z, \hat{\theta}) \]
What to Use them For?

• Find badly labeled examples

• Find Adversarial Examples

• As usual, understanding model behavior?
Points to take away
Important Points to take away

• Interpretability – no consistent definition

• When designing new system, ask your stakeholders what they want out of it.

• See if you can use inherently interpretable model.

• If not, what method can you use to interpret the black box?

• Ask – does this method make sense? Question Assumptions!!!

• Stress Test and Evaluate!