Interpretable Neural Networks for Computer Vision: Clinical Decisions that are Computer-Aided, not Automated

Cynthia Rudin
Professor of Computer Science
Duke University
Right for the wrong reasons (Clever Hans)

Clever Hans performing in 1904
Why interpretability?

• Confounding: Right for the wrong reasons (Clever Hans)
• High-stakes decisions: Should patient get a biopsy?
• Troubleshooting: Can’t second-guess the reasoning process of a black box.
  • Will it work if I switch equipment?
  • Will it work for all types of patients?
  • Can I check if it’s working on my current patient?
  • Is the information that I fed into the system correct?
• Responsibility: It is the doctor’s responsibility to make a good decision. (Isn’t it?)

The use of black box models makes all of these much worse. Black box models turn computer-aided decisions into automated decisions.
“Explaining” deep NN’s with saliency maps doesn’t work

<table>
<thead>
<tr>
<th>Explanations Using Attention Maps</th>
<th>Test Image</th>
<th>Evidence for Animal Being a Siberian Husky</th>
<th>Evidence for Animal Being a Transverse Flute</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Image of a Siberian Husky]</td>
<td>[Image of a Siberian Husky]</td>
<td>[Image of a Siberian Husky with attention map]</td>
<td>[Image of a Siberian Husky with attention map]</td>
</tr>
</tbody>
</table>

Do you trust the network now?
Lots of work in radiology on attention maps now...
Problem spectrum

- age 45
- congestive heart failure? yes
- takes aspirin
- smoking? no
- gender M
- exercise? yes
- allergies? no
- number of past strokes 2
- diabetes? yes

Tabular: All features are interpretable
- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable
- pixels/voxels, words, a bit of a sound wave
Problem spectrum

Very sparse models (trees, scoring systems)

Neural networks

With minor pre-processing, all methods have similar performance

**Tabular**: All features are interpretable
- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

**Raw**: Features are individually uninterpretable
- pixels/voxels, words, a bit of a sound wave
Problem spectrum

- age 45
- congestive heart failure? yes
- takes aspirin
- smoking? no
- gender M
- exercise? yes
- allergies? no
- number of past strokes 2
- diabetes? yes

Tabular: All features are interpretable
- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable
- pixels/voxels, words, a bit of a sound wave
Problem spectrum

age 45
congestive heart failure? yes
takes aspirin
smoking? no
gender M
exercise? yes
allergies? no
number of past strokes 2
diabetes? yes

Tabular: All features are interpretable
- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable
- pixels/voxels, words, a bit of a sound wave
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

For tabular data: decision trees, linear/additive models/scoring systems
For raw data: interpretable neural networks

But what is an interpretable neural network?
Approach 1: A neural network that does case-based reasoning
Why is this bird classified as a clay-colored sparrow?

Because this part of the bird looks like that part of a prototypical clay-colored sparrow.
“This Looks Like That: deep learning for interpretable image recognition”
NeurIPS 2019 (spotlight)

When we are faced with challenging image classification tasks, we often explain our reasoning by dissecting the image, and pointing out prototypical aspects of one class or another. The mounting evidence for each of the classes helps us make our final decision. In this work, we introduce a deep network architecture --- prototypical part network (ProtoPNet), that reasons in a similar way: the network dissects the image by finding prototypical parts, and combines evidence from the prototypes to make a final classification. The model thus reasons in a way that is qualitatively similar to the way ornithologists, physicians, and others would explain to people on how to solve challenging image

- Adds a “prototype” layer to a black box, forces the network to do case-based reasoning.
- Prototypes are learned during training.
Take any “standard” black box CNN...
And transform it to be interpretable
Why is this bird classified as a red-bellied woodpecker?

Evidence for this bird being a red-bellied woodpecker:

<table>
<thead>
<tr>
<th>Original image</th>
<th>Prototype</th>
<th>Training image where prototype comes from</th>
<th>Activation map</th>
<th>Similarity score</th>
<th>Class connection points contributed</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Original Image" /></td>
<td><img src="image2" alt="Prototype" /></td>
<td><img src="image3" alt="Training Image" /></td>
<td><img src="image4" alt="Activation Map" /></td>
<td>6.499</td>
<td>4.392</td>
</tr>
<tr>
<td>1.180</td>
<td>3.890</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.108</td>
<td>7.669</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.392 × 1.127 = 4.950</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.890 × 1.108 = 4.310</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total points to red-bellied woodpecker: 32.736

Evidence for this bird being a red-cockaded woodpecker:

<table>
<thead>
<tr>
<th>Original image</th>
<th>Prototype</th>
<th>Training image where prototype comes from</th>
<th>Activation map</th>
<th>Similarity score</th>
<th>Class connection points contributed</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5" alt="Original Image" /></td>
<td><img src="image6" alt="Prototype" /></td>
<td><img src="image7" alt="Training Image" /></td>
<td><img src="image8" alt="Activation Map" /></td>
<td>2.452</td>
<td>2.125</td>
</tr>
<tr>
<td>1.945</td>
<td>2.079</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.125 × 1.091 = 2.318</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.945 × 1.069 = 2.079</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total points to red-cockaded woodpecker: 16.886
Base model: DenseNet161

Why is this bird incorrectly classified as a prothonotary warbler, instead of a Wilson's warbler?

Evidence for this bird being a Wilson's warbler:

<table>
<thead>
<tr>
<th>Original image (box showing part that looks like prototype)</th>
<th>Prototype</th>
<th>Training image where prototype comes from</th>
<th>Activation map</th>
<th>Similarity score</th>
<th>Class</th>
<th>Points connection contributed</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Image" /></td>
<td><img src="image2.jpg" alt="Image" /></td>
<td><img src="image3.jpg" alt="Image" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image4.jpg" alt="Image" /></td>
<td><img src="image5.jpg" alt="Image" /></td>
<td><img src="image6.jpg" alt="Image" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image7.jpg" alt="Image" /></td>
<td><img src="image8.jpg" alt="Image" /></td>
<td><img src="image9.jpg" alt="Image" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total points to Wilson's warbler: 9,744

Evidence for this bird being a prothonotary warbler:

<table>
<thead>
<tr>
<th>Original image (box showing part that looks like prototype)</th>
<th>Prototype</th>
<th>Training image where prototype comes from</th>
<th>Activation map</th>
<th>Similarity score</th>
<th>Class</th>
<th>Points connection contributed</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image10.jpg" alt="Image" /></td>
<td><img src="image11.jpg" alt="Image" /></td>
<td><img src="image12.jpg" alt="Image" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image13.jpg" alt="Image" /></td>
<td><img src="image14.jpg" alt="Image" /></td>
<td><img src="image15.jpg" alt="Image" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image16.jpg" alt="Image" /></td>
<td><img src="image17.jpg" alt="Image" /></td>
<td><img src="image18.jpg" alt="Image" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total points to prothonotary warbler: 12,391
Base model: VGG-16

Why is this bird classified as a Wilson's warbler?

Evidence for this bird being a Wilson's warbler:

<table>
<thead>
<tr>
<th>Original image</th>
<th>Prototype</th>
<th>Training image where prototype comes from</th>
<th>Activation map</th>
<th>Similarity score</th>
<th>Class</th>
<th>Points connection contributed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.341 × 1.443</td>
<td></td>
<td>4.821</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.302 × 1.450</td>
<td></td>
<td>4.788</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.159 × 1.442</td>
<td></td>
<td>3.113</td>
</tr>
</tbody>
</table>

Total points to Wilson's warbler: 19.473

Evidence for this bird being a prothonotary warbler:

<table>
<thead>
<tr>
<th>Original image</th>
<th>Prototype</th>
<th>Training image where prototype comes from</th>
<th>Activation map</th>
<th>Similarity score</th>
<th>Class</th>
<th>Points connection contributed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.722 × 1.105</td>
<td></td>
<td>1.903</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.626 × 1.085</td>
<td></td>
<td>1.764</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.605 × 1.173</td>
<td></td>
<td>1.883</td>
</tr>
</tbody>
</table>

Total points to prothonotary warbler: 10.234
CUB-200

- 200 classes of birds
- Original black box accuracy between 74.6% (VGG16) and 82.3% (Res34).
- Interpretable model’s accuracy between 76.1% (VGG) and 80.2% (Dense121). Combining several interpretable networks together yields 84.8%, and still yields an interpretable model.

So even for computer vision, we can still have an interpretable model of the same accuracy as a black box.
Mammography

• Breast cancer is a leading cause of death in the USA (Kochanek et al 2020)
• Hundreds of thousands of cases diagnosed in the USA each year (> 300K in 2019), causing tens of thousands of deaths each year.
• Mammography is the *hardest task* in all of radiology (Moss, 2020)
  • Radiologists miss ~1/5 of breast cancers
  • half of women getting an annual mammogram over 10 years will have a false positive.
  • up to 3/4 of biopsies come back as benign, i.e., potentially unnecessary surgeries
Our team

Alina Barnett

Chaofan Chen

Daniel Tao
Several AI algorithms have FDA approval for radiology!!
Two different problems

Breast lesion detection

Black box says: There’s no lesion in this image.

Whether to order a biopsy for a lesion?

Black box says: Don’t get a biopsy.
Why is AI mammography hard?

- AI radiology is hard
- Mammography is just really, really hard.
- No data.
- Confounding: Right for the wrong reasons → hard to deal with.
- How to design a system that would actually be helpful?
  - Malignancy vs. Benign is NOT the right problem to solve!
Why is AI mammography hard?

• AI radiology is hard
• Mammography is just really hard
• No data.
• Confounding: Right for the wrong reasons → hard to deal with.
• How to design a system that would actually be helpful?
  • Malignancy vs. Benign

A new investigation revealed “unexpected” shortcomings when using a federally cleared artificial intelligence tool to detect intracranial hemorrhages. The findings pushed researchers to call for more standardization when evaluating AI-based clinical decision support platforms.
a: Uninterpretable Approach

Probability of malignancy: Low
Predict: Benign
Because: n/a
a: Uninterpretable Approach

Probability of malignancy: Low
Predict: Benign
Because: n/a

b: Attention only approaches

Probability of malignancy: Low
Predict: Benign
Because: No other context provided
Prototypes

Model decomposes to predict margins before malignancy

c: Our approach (IAIA-BL)

- Prototypes:
  - Indistinct margin: looks like -> adds +0.5 to malignancy score
  - Circumscribed margin: looks like -> adds -1.3 to malignancy score

- Probability of malignancy: Low
- Predict: Benign
- Because: mass has primarily circumscribed margin
Data Availability

• Public data availability is abysmal
• Low-quality images, outdated equipment, inconsistent labeling
• Some wanted us to hand over IP...
Data

• From Duke!
• 1136 digital screening mammogram images of masses in the breast from 484 patients at Duke University hospitals.

1136?!  484?!

125 masses with spiculated margin
220 with indistinct margin
41 with microlobulated margin (didn’t use)
579 with obscured margin (didn’t use)
171 with circumscribed margin

Let’s do it!
Clever Hans performing in 1904

black box accuracy = “interpretable” accuracy
- Radiologists add fine labels for only 37 images (9%)
- We generalized ProtoPNet to handle mixed labelling: fine-grained attention labels and standard labels

black box accuracy = interpretable accuracy
IAIA-BL Architecture

Each classifier is a generalized ProtoPNet, Combined model is linear
Example: why correctly classified as circumscribed
Example: why correctly classified as spiculated

Test image → Activation of prototype on test image → Most activated
- looks like with similarity score 2.3 → Most relevant part of prototype → Learned prototypical lesion
- adds 2.3 x 8.5 = 20 to spic.

- looks like with similarity score 1.0 →
  - adds 1.0 x 1.7 = 1.7 to spic.

- looks like with similarity score 0.20 →
  - adds 0.20 x 1.8 = 0.36 to indst.
Preliminary results

- Performance *as good or better* than uninterpretable
- (Uninterpretable gets to use confounding info!)
Why is AI mammography hard?

• Mammography is just really, really hard.
• No data.  <Thank you Joseph>
• Confounding: Right for the wrong reasons → hard to deal with.
• How to design a system that would actually be helpful?
  • Malignancy vs. Benign is NOT the right problem to solve!

What we did NOT do:
- black box + saliency
- malignant vs. benign only
Approach 2: Neural Disentanglement
Concept Whitening for Interpretable Image Recognition

Zhi Chen, Yijie Bei, Cynthia Rudin

What does a neural network encode about a concept as we traverse through the layers? Interpretability in machine learning is undoubtedly important, but the calculations of neural networks are very challenging to understand. Attempts to see inside their hidden layers can either be misleading, unusable, or rely on the latent space to possess properties that it may not have. In this work, rather than attempting to analyze a neural network posthoc, we introduce a mechanism, called concept whitening (CW), to alter a given layer of the network to allow us to better understand the computation leading up to that layer. When a concept whitening module is added to a CNN, the axes of the latent space are aligned with known concepts of interest. By experiment, we show that CW can provide us a much clearer understanding for how the network gradually
CNN’s are not naturally disentangled
Consider the latent space of a Batch Norm layer

Create a vector pointing towards each concept. They are not naturally orthonormal.
Batch Normalization

Concept Whitening

latent space

- concept 1
- concept 2
- concept 3
- concept 4
When a **Concept Whitening** module is added to a CNN,
- the latent space is whitened (decorrelated and normalized)
- the axes of the latent space are aligned with concepts of interest
When a **Concept Whitening** module is added to a CNN,
- the latent space is whitened (decorrelated and normalized)
- the axes of the latent space are aligned with concepts of interest
When a CW module is added to a CNN,
- the latent space is whitened (decorrelated and normalized)
- the axes of the latent space are aligned with concepts of interest
When CW is added to different layers...

In earlier layers, color and texture information related to the concepts are represented along the axes.
Most activated

Concept information now lies along the axes.
Because this image has warm colors, it lies mainly along the bed axis at layer 1.

See how an image travels through the layers.
See how an image travels through the layers
See how an image travels through the layers
See how an image travels through the layers
See how an image travels through the layers
See how an image travels through the layers
See how an image travels through the layers
Advantages of CW over BatchNorm

➢ No sacrifice in accuracy
  - accuracy is on par with standard CNNs

➢ Easy to use
  - warm-start from pretrained model requires only one additional epoch of further training
  - Note: requires training data for the concepts to define the axes

➢ Disentangles the latent space
Interpretable deep CNNs for computer vision:

- Prototype Network
- Case-based reasoning
- Fine Annotation
- Concept Whitening
- Neural Disentanglement

Strictly better than saliency

Strictly better than concept vectors
Take-aways

• There is no scientific evidence supporting a tradeoff between interpretability and accuracy in deep learning.
  • Interpretability helps troubleshoot and helps accuracy

• It is a matter of time until companies try to use black box models for biopsy decisions...
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Computer Science > Machine Learning

[Submitted on 5 Feb 2020 v1, last revised 10 Oct 2020 (this version, v4)]

Concept Whitening for Interpretable Image Recognition

Zhi Chen, Yiğit Bey, Cynthia Rudin

What does a neural network encode about a concept as we traverse through the layers? Interpretability in machine learning models is crucial for high stakes decisions and troubleshooting. This work introduces a framework for understanding the representations of neural networks and provides a tool for interpreting their success and failures.

Computer Science > Machine Learning

[Submitted on 20 Mar 2021]

Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, Chudi Zheng

Interpretability in machine learning (ML) is crucial for high stakes decisions and troubleshooting. This work provides fundamental principles for interpretable ML, and dispel common misunderstandings that dilute the importance of this crucial topic. We also identify 10 technical challenge areas in interpretable machine learning and provide history and background on each problem. Some of these problems are classically important, and some are recent problems that have arisen in the last few years. These problems are: (1) Optimization of scoring systems; (2) Place constraints into generalized additive models to encourage sparsity and interpretability; (4) Modern case-based reasoning, including neural networks and matching for causal inference; (5) Complete supervised disentanglement of neural networks; (6) Complete or partial unsupervised disentanglement of neural networks; (7) Dimensionality reduction for data visualization; (8) Machine learning models that can incorporate physics and other generative or causal constraints; (9) Characterization of the "Rashomon set" of good models; and (10) Interpretable reinforcement learning. This survey is suitable as a starting point for statisticians and computer scientists interested in working in interpretable machine learning.

All papers are here: https://users.cs.duke.edu/~cynthia/