Directions in Interpretability

Ruth Fong

CVPR 2022, Human-Centered AI Tutorial
June 20, 2022
Slides and links available at ruthfong.com
What is interpretability?

Research focused on explaining complex AI systems in a human-interpretable way.
Why interpretability?

- 🕵️‍♂️ Science
- 👫 Trust
-🤖 Learning
An incomplete retrospective: the first decade of deep learning

**CNNs (2012-2016)**
- AlexNet, VGG16,
- GoogLeNet, ResNet50

**GANs (2014-2018)**
- GAN, ProGAN, CycleGAN

**Transformers (2017-now)**
- Transformer, BERT, ViT

**Self-supervised learning (2016-now)**
- Colorization, MOCO, SWaV

**Diffusion models (2020-now)**
- DDPM, DALL-E 2, Imagen

[Krizhevsky et al., NeurIPS 2012; Zhu* & Park* et al., ICCV 2017; Zhang et al., ECCV 2016; Dosovitskiy* et al., ICLR 2021; Ramesh et al., arXiv 2022]
An incomplete retrospective: the first decade of interpretability

Feature visualization (2013-2018)
Activation Max., Feature Inversion, Net Dissect, Feature Vis.

Attribution heatmaps (2013-2019)
Gradient, Grad-CAM, Occlusion, Perturbations, RISE

Interpretable-by-design (2020-now)
Concept Bottleneck, ProtoPNet, ProtoTree

[Selvaraju et al., ICCV 2017; Fong* & Patrick* et al., ICCV 2019; Bau* & Zhou* et al., CVPR 2017; Olah et al., Distill 2017; Koh*, Nguyen*, Tang* et al., ICML 2020]
An incomplete retrospective: the first decade of interpretability

Attribution heatmaps (2013-2019)
- Gradient, Grad-CAM, Occlusion, Perturbations, RISE

Primarily focused on understanding and approximating **CNNs**
- Exceptions:
  - GANPaint [Bau et al., ICLR 2019]
  - Transformer Circuits [Elhage et al., 2021]

Interpretable-by-design (2020-now)
- Concept Bottleneck, ProtoPNet, ProtoTree

[Selvaraju et al., ICCV 2017; Fong* & Patrick* et al., ICCV 2019; Bau* & Zhou* et al., CVPR 2017; Olah et al., Distill 2017; Koh*, Nguyen*, Tang* et al., ICML 2020]
Directions for the next decade of interpretability

1. Develop interpretability methods for *diverse domains*
   - Beyond CNN classifiers: self-supervised learning, generative models, etc.

2. Center *humans* throughout the development process
   - In design, co-develop methods with real-world stakeholders.
   - In evaluation, measure human interpretability and utility of methods.
   - In deployment, package interpretability tools for the wider community.
Roadmap

1. **Automated** evaluation of interpretability → **human-centered** evaluation
   HIVE: Evaluating the Human Interpretability of Visual Explanations.

2. Explanations via **labelled attributes** → explanations via **labelled attributes and unlabelled features**
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.

3. Interpretability of **supervised** models → interpretability of **self-supervised** models
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.

4. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.
Roadmap

1. **Automated** evaluation of interpretability $\rightarrow$ **human-centered** evaluation
   HIVE: Evaluating the Human Interpretability of Visual Explanations.

2. Explanations via **labelled attributes** $\rightarrow$ explanations via **labelled attributes and unlabelled features**
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.

3. Interpretability of **supervised** models $\rightarrow$ interpretability of **self-supervised** models
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.

4. **Static** visualizations $\rightarrow$ **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.
Explanation form factors: Why did the model predict Y?

**Heatmap** explanations (e.g. Grad-CAM)

**Prototype** explanations (e.g. ProtoPNet)

**Counterfactual** explanations (e.g. SCOUT)

Concept-based explanations (e.g. Concept Bottleneck)

[Selvaraju et al., ICCV 2017; Koh*, Nguyen*, Tang* et al., ICML 2020; Chen* & Li* et al., NeurIPS 2019; Wang & Vasconcelos, CVPR 2020]
Post-hoc explanations

Explanation
(not part of model design)
Interpretable-by-design models
Current metrics focus on heatmap evaluation

- Weak localization performance [Zhang et al., ECCV 2016]
- Perturbation analysis
  - Deletion game [Samek et al., TNNLS 2017]
  - Retrain classifiers with removed features [Hooker et al., NeurIPS 2019]
- Sensitivity to...
  - output neuron [Rebuffi*, Fong*, Ji* et al., CVPR 2020]
  - model parameters [Adebayo et al., NeurIPS 2018]
  - ...
Selectivity to output class

[Mahendran & Vedaldi, ECCV 2016; Rebuffi et al., CVPR 2020]
Sensitivity to model parameters (a.k.a. sanity checks)

[Adebayo et al., NeurIPS 2018]
Current metrics focus on heatmap evaluation

- Sheng & Huang, HCOMP 2020
  Guess the incorrectly predicted label

- **Nguyen et al., NeurIPS 2021**
  Is this prediction correct?

- Colin* & Fel* et al., arXiv 2021
  What did the model predict (choose one of two)?
Is this prediction correct?

AI’s top-1 predicted label: **lorikeet**

A confidence 20%

B heatmaps
- GradCAM
- EP
- SOD

C 3 nearest neighbors in **lorikeet**

user

**Yes** vs **No**

groundtruth label: “bee eater”

[Nguyen et al., NeurIPS 2021]
HIVE: Evaluating the Human Interpretability of Visual Explanations

1. Within method $\rightarrow$ **Cross-method comparison**

2. Automated evaluation $\rightarrow$ **Human-centered evaluation**

3. Intuition-based reasoning $\rightarrow$ **Falsifiable hypothesis testing**
Our contributions

- Novel human study design for evaluating 4 diverse interpretability methods
- **First human study** for interpretable-by-design and prototype methods
- Quantify the utility of explanations in distinguishing between *correct and incorrect predictions*
- Quantify how users would trade off between *interpretability and accuracy*
- **Open-source** HIVE studies to encourage reproducible research

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.]
1. Cross-method comparison

- **Grads-CAM**
  - heatmap
  - post-hoc

- **BagNet**
  - prototype
  - interpretably-by-design

- **ProtoPNet**
  - prototype

- **ProtoTree**

[Selvaraji et al., ICCV 2017; Brendel & Bethge, ICLR 2019; Chen* & Li* et al., NeurIPS 2019, Nauta et al., CVPR 2021]
2. Human-centered evaluation

**Agreement task**
How confident are you in the model’s prediction?

**Distinction task**
Which class do you think is correct?

*Experimental set-up: AMT studies with N=50 participants each*

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.; Chen* & Li* et al., NeurIPS 2019]
2. Human-centered evaluation

**Agreement task**

How confident are you in the model’s prediction?

**Finding #1:** Prototype similarities often do **not** align with human notions of similarity.

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.; Chen* & Li* et al., NeurIPS 2019]
2. Human-centered evaluation

**Agreement task**
How confident are you in the model’s prediction?

**Finding #1:** Prototype similarities often do not align with human notions of similarity.

**Finding #2:** Agreement task reveals confirmation bias.

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.; Chen* & Li* et al., NeurIPS 2019]
2. Human-centered evaluation

**Agreement task**

How confident are you in the model’s prediction?

**Finding #1:** Prototype similarities often **do not align** with human notions of similarity.

**Finding #2:** Agreement task reveals **confirmation bias**.

---

**Q. What do you think about the model’s prediction?**

- Fairly confident that prediction is correct
- Somewhat confident that prediction is correct
- Somewhat confident that prediction is incorrect
- Fairly confident that prediction is incorrect

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.]
2. Human-centered evaluation

**Distinction task**
Which class do you think is correct?

**Finding #3:** Participants struggle to identify the **correct class**, esp. for incorrect predictions.

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.; Selvaraju et al., ICCV 2017]
2. Human-centered evaluation

**Distinction task**
Which class do you think is correct?

Finding #3: Participants struggle to identify the **correct class**, esp. for incorrect predictions.

Goal: Interpretability should help humans identify and explain model errors.

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.]
3. Falsifiable hypothesis testing

**Finding #1:** Prototype similarities often do not align with human notions of similarity.

**Finding #2:** Agreement task reveals confirmation bias.

**Finding #3:** Participants struggle to identify the correct class, esp. for incorrect predictions.

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.]
3. Falsifiable hypothesis testing

**Finding #1:** Prototype similarities often do not align with human notions of similarity.

**Finding #2:** Agreement task reveals confirmation bias.

**Finding #3:** Participants struggle to identify the correct class, esp. for incorrect predictions.

**Finding #4:** Participants prefer interpretability over accuracy, esp. in high-risk settings.

---

**Interpretability-accuracy tradeoff**

Q: What is the minimum accuracy of a baseline model that would convince you to use it over a model with explanations?

![Graph showing required accuracy gains for low-risk, medium-risk, and high-risk settings.]

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.]
Challenges for human evaluation

- Skill cost: web development skills
- Financial cost: budget for AMT experiments
- Time cost: human study design and iteration (e.g. task feasibility, IRB approval, quality control)

**Takeaway:** As a research community, invest in and reward human evaluation studies (like dataset development).
Roadmap

1. **Automated** evaluation of interpretability → **human-centered** evaluation
   HIVE: Evaluating the Human Interpretability of Visual Explanations.

2. Explanations via **labelled attributes** → explanations via **labelled attributes and unlabelled features**
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.

3. Interpretability of **supervised** models → interpretability of **self-supervised** models
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.

4. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.
Concept-based explanations

Why did the model predict **sheepdog**?

**Pro:** Labelled concepts are interpretable to humans
**Concept Bottleneck: Linear Combination of Labelled Attributes**

**Predict present or absence of attribute**

**Linearly combine with attribute weights**

\[ \sum \]

**sheepdog**

**Con:** Problems with predicting fractional values
- hard to interpret
- can encode hidden information

[Koh*, Nguyen*, Tang* et al., ICML 2020]
Concept Bottleneck: Linear Combination of Labelled Attributes

Predict present or absence of attribute

Linearly combine with attribute weights

Con: Problems with predicting fractional values
- hard to interpret
- can encode hidden information

[Koh*, Nguyen*, Tang* et al., ICML 2020]
ELUDE: Explanation via a Labelled and Unlabelled DEcomposition of features

**Goal:** Approximate behavior of original CNN

[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]
**ELUDE: Decomposition of labelled and unlabelled features**

**Goal:** Approximate behavior of original CNN

1. Linearly combine **ground-truth, labelled attributes**

2. Learn remaining **unlabelled features as low-rank space**

---

**Feature activations for sheepdog**

- \( f_1 \) = 8.2
- \( f_2 \) = 4.5
- \( f_3 \) = -7.6

**Feature weights for sheepdog**

- \( +1.1 \) for \( f_1 \)
- \( -0.3 \) for \( f_2 \)
- \( -0.7 \) for \( f_3 \)

**Ground-truth presence/absence of attributes**

- fur: 1, paw: 1, tree: 0

**Attribute weights for sheepdog**

- fur: \( +1.2 \)
- paw: \( +0.7 \)
- tree: \( -0.6 \)

---

[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]
**Attributes only**: % of model explained via labelled attributes decreases as task complexity increases

<table>
<thead>
<tr>
<th>Task</th>
<th>% Explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-way scene classification (indoor vs. outdoor)</td>
<td>95.7</td>
</tr>
<tr>
<td>16-way scene classification (home/hotel, workplace, etc.)</td>
<td>46.2</td>
</tr>
<tr>
<td>365-way scene classification (airfield, bowling alley, etc.)</td>
<td>28.8</td>
</tr>
</tbody>
</table>

Without fractional values encoding hidden information, attribute-only approaches are limited.

[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]
**Attributes only:** % of model explained via labelled attributes decreases as task complexity increases

<table>
<thead>
<tr>
<th>Scene group</th>
<th>TPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>home/hotel</td>
<td>99.0</td>
</tr>
<tr>
<td>comm-buildings/towns</td>
<td>93.5</td>
</tr>
<tr>
<td>water/ice/snow</td>
<td>60.6</td>
</tr>
<tr>
<td>forest/field/jungle</td>
<td>40.2</td>
</tr>
<tr>
<td>workplace</td>
<td>14.2</td>
</tr>
<tr>
<td>shopping-dining</td>
<td>12.4</td>
</tr>
<tr>
<td>cultural/historical</td>
<td>6.5</td>
</tr>
<tr>
<td>cabins/gardens/farms</td>
<td>4.7</td>
</tr>
<tr>
<td>outdoor-transport</td>
<td>3.2</td>
</tr>
<tr>
<td>indoor-transport</td>
<td>0.0</td>
</tr>
<tr>
<td>indoor-sports/leisure</td>
<td>0.0</td>
</tr>
<tr>
<td>indoor-cultural</td>
<td>0.0</td>
</tr>
<tr>
<td>mountains/desert/sky</td>
<td>0.0</td>
</tr>
<tr>
<td>outdoor-manmade</td>
<td>0.0</td>
</tr>
<tr>
<td>outdoor-fields/parks</td>
<td>0.0</td>
</tr>
<tr>
<td>industrial-construction</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Without fractional values encoding hidden information, attribute-only approaches are limited.

[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]
**Features + attributes:** Unlabelled features correspond to human-interpretable concepts

- Bowling alleys?
- People eating?
- Outdoor sports fields?
- Castle-like buildings?

<table>
<thead>
<tr>
<th>Scene group</th>
<th>TPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>home/hotel</td>
<td>99.0</td>
</tr>
<tr>
<td>comm-buildings/towns</td>
<td>93.5</td>
</tr>
<tr>
<td>water/ice/snow</td>
<td>60.6</td>
</tr>
<tr>
<td>forest/field/jungle</td>
<td>40.2</td>
</tr>
<tr>
<td>workplace</td>
<td>14.2</td>
</tr>
<tr>
<td>shopping-dining</td>
<td>12.4</td>
</tr>
<tr>
<td>cultural/historical</td>
<td>6.5</td>
</tr>
<tr>
<td>cabins/gardens/farms</td>
<td>4.7</td>
</tr>
<tr>
<td>outdoor-transport</td>
<td>3.2</td>
</tr>
<tr>
<td>indoor-transport</td>
<td>0.0</td>
</tr>
<tr>
<td>indoor-sports/leisure</td>
<td>0.0</td>
</tr>
<tr>
<td>indoor-cultural</td>
<td>0.0</td>
</tr>
<tr>
<td>mountains/desert/sky</td>
<td>0.0</td>
</tr>
<tr>
<td>outdoor-manmade</td>
<td>0.0</td>
</tr>
<tr>
<td>outdoor-fields/parks</td>
<td>0.0</td>
</tr>
<tr>
<td>industrial-construction</td>
<td>0.0</td>
</tr>
</tbody>
</table>

[attributes only]
[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]
Challenges for concept-based methods

- Attributes-only approaches are incomplete
- Develop more methods to explain the “remainder”
  - Interpretable Basis Decomposition (IBD) [Zhou et al., ECCV 2018]
  - Automatic Concept-based Explanations (ACE) [Ghorbani et al., NeurIPS 2019]
  - ConceptSHAP [Yeh et al., NeurIPS 2020]
- Ensure that concept-based explanations are truly human-interpretable

**Takeaway:** Be realistic about the benefits and limitations of an interpretability method and work towards addressing the limitations.
Roadmap

1. **Automated** evaluation of interpretability → **human-centered** evaluation
   HIVE: Evaluating the Human Interpretability of Visual Explanations.

2. Explanations via **labelled attributes** → explanations via **labelled attributes and unlabelled features**
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.

3. Interpretability of **supervised** models → interpretability of **self-supervised** models
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.

4. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.
Supervised Learning

(x, y, sheepdog)
Self-Supervised Learning
Visual Concept
Self-Supervised Learning
Self-Supervised Learning

Unlabelled data

Learn clusters
(e.g. DeepCluster, SeLa, SwaV)

Learn features
(e.g. SimCLR, MoCo, ...)

k-means

cluster 1

cluster 2

cluster K
Learnability

Learnability

Describability

"dessert with chocolate sauce"

Describability

dessert with chocolate sauce

Manual

Describability

dessert with chocolate sauce

Manual

OR

Automatic

(A)

(B)

Evaluation

Learnability

ImageNet cluster purity:
how correlated is a cluster’s contents to a single ImageNet label?

$purity = 1 \rightarrow$ cluster only contains images from a single ImageNet label

[Asano et al., ICLR 2020; He et al., CVPR 2020]
Evaluation

Learnability

Describability

[Asano et al., ICLR 2020; He et al., CVPR 2020]
Findings

**ImageNet cluster purity**

**SeLa: cluster 393 (0.668)**
a newborn baby lying on a bed

**SeLa: cluster 332 (0.542)**
a snake on a hand

**MoCo: cluster 2335 (0.459)**
view of the mountains from the lake

Follow up: Laina et al., ICLR 2022.
Measuring the Interpretability of Unsupervised Representations via Quantized Reverse Probing.

98.3% 100.0% 93.3% 95.0%

[Asano et al., ICLR 2020; He et al., CVPR 2020]
Challenges for novel frontiers in deep learning

- Need to contextualize interpretability to the novel frontiers
- Lack of access to standardized implementations

**Takeaway:** Collaboration and buy-in from novel research areas is crucial for interpretability in those frontiers.
**Roadmap**

1. **Automated** evaluation of interpretability $\rightarrow$ **human-centered** evaluation  
   HIVE: Evaluating the Human Interpretability of Visual Explanations.

2. Explanations via **labelled attributes** $\rightarrow$ explanations via **labelled attributes and unlabelled features**  
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.

3. Interpretability of **supervised** models $\rightarrow$ interpretability of **self-supervised** models  
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.

4. **Static** visualizations $\rightarrow$ **interactive** visualizations  
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.  
   Interactive Similarity Overlays.
Interpretability Tools

Current tools render **static images.**

Future tools should be **interactive!**

[Fong et al., ICCV 2019; Selvaraju et al., ICCV 2017; Bau et al., CVPR 2017; Mahendran & Vedaldi, IJCV 2016; Olah et al., Distill 2018; Fong et al., VISxAI 2021]
Interpretability: Interactive, Exploratory, Easy-to-use

How can we easily explore hypotheses about the model?

Acknowledgement: Chris Olah
Interactive Similarity Overlays

Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
Spatial Activations

\[ f_a \rightarrow \text{golden retriever} \]
Interactive Similarity Overlays

\[ a_{6,5} = [17.7, 0, 103.4, 6.81, 0, 0, 0, 0, 32.0, 0, 0, 0, \ldots] \]

[Olah et al., Distill 2018] 62
Interactive Similarity Overlays

[Fong et al., VISxAI 2021. Interactive Similarity Overlays.]
Live Demo: Interactive Similarity Overlays

Interactive visualizations empower practitioners to easily explore model behavior.

[Fong et al., VISxAI 2021. Interactive Similarity Overlays.]
Preview: Interactive Visual Feature Search

bit.ly/interactive_search

Devon Ulrich

Acknowledgement: David Bau
Challenges for interactive visualizations

- Skills cost: web development skills
  - HuggingFace Spaces, Gradio, Streamlit
- Potential misuse: Intuition-based insights should be validated via quantitative experiments
- Poor incentives: software tooling for research is often not rewarded
- Inadequate publishing structures: Sparse publishing venues for interactive articles and/or visualizations
  - Distill journal hiatus
  - CVPR demo track
- Lack of cross-talk: HCI and AI communities are developing interpretability tools fairly independently

**Takeaway:** Relevant research communities should collectively invest in and reward software tooling for research, particularly interactive tools.
Takeaways from challenges in interpretability

- **Human studies:** As a research community, invest in and reward human evaluation studies (like dataset development).

- **(Concept-based) interpretability:** Be realistic about the benefits and limitations of an interpretability method and work towards addressing the limitations.

- **New frontiers:** Collaboration and buy-in from novel research areas is crucial for interpretability in those frontiers.

- **Interactive visualizations:** Relevant research communities should collectively invest in and reward software tooling for research, particularly interactive tools.
Directions for the next decade of interpretability

1. Develop interpretability methods for **diverse domains**
   - Beyond CNN classifiers: self-supervised learning, generative models, etc.

2. Center **humans** throughout the development process
   - In design, co-develop methods with real-world stakeholders.
   - In evaluation, measure human interpretability and utility of methods.
   - In deployment, package interpretability tools for the wider community.
An incomplete retrospective: the first decade of interpretability

**Feature visualization (2013-2018)**
- Activation Max., Feature Inversion,
- Net Dissect, Feature Vis.

**Attribution heatmaps (2013-2019)**
- Gradient, Grad-CAM,
- Occlusion, Perturbations, RISE

**Interpretable-by-design (2020-now)**
- Concept Bottleneck, ProtoPNet,
- ProtoTree

- Primarily focused on understanding and approximating CNNs

---

2012

- Orig Img
- Mask
- Grad CAM

- Cabbage butterfly

2022

- Concept Bottleneck
- ProtoTree

---

[Selvaraju et al., ICCV 2017; Fong* & Patrick* et al., ICCV 2019; Bau* & Zhou* et al., CVPR 2017; Olah et al., Distill 2017; Koh*, Nguyen*, Tang* et al., ICML 2020]
Into the future: the next decade of interpretability
We’re hiring postdocs!

bit.ly/vai-lg-postdoc

Talk acknowledgements: Brian Zhang, Sunnie S. Y. Kim, Vikram V. Ramaswamy, Olga Russakovsky
Thank You