How Thales relies on Explainable AI to accelerate adoption of AI in critical systems

Freddy Lecue, Chief AI Scientist
@freddylecue
Context
Markets we serve where XAI is crucial

Aerospace  Space  Ground Transportation  Defence  Security

Trusted Partner For A Safer World
Approach
Some XAI Approaches (1)

Network $f(x_1, x_2)$
Attributes at $x_1 = 3, x_2 = 1$

- **Integrated gradients**: $x_1 = 1.5, x_2 = -0.5$
- **DeepLift**: $x_1 = 1.5, x_2 = -0.5$
- **LRP**: $x_1 = 1.5, x_2 = -0.5$

Network $g(x_1, x_2)$
Attributes at $x_1 = 3, x_2 = 1$

- **Integrated gradients**: $x_1 = 1.5, x_2 = -0.5$
- **DeepLift**: $x_1 = 2, x_2 = -1$
- **LRP**: $x_1 = 2, x_2 = -1$

**Attribution for Deep Network** (Integrated gradient-based)


**Example-based / Prototype**

**Surogate Model**

**Attention Mechanism**

Some XAI Approaches (2)

Interpretable Units

Uncertainty Map
Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

Visual Explanation
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19

Saliency Map / Features Attribution-based
Some XAI Approaches – Towards Semantics

ConceptSHAP

Chih-Kuan Yeh, Been Kim, Sercan Ömer Arik, Chun-Liang Li, Tomas Pfister, Pradeep Ravikumar: On Completeness-aware Concept-Based Explanations in Deep Neural Networks. NeurIPS 2020

ACE


Circuits in CNNs

https://distill.pub/2020/circuits/zoom-in/

Compositional Explanations

Jesse Mu, Jacob Andreas: Compositional Explanations of Neurons. NeurIPS 2020
Example of an End-to-End XAI System

- Humans may have follow-up questions
- Human–Machine interactions are required
- Explanations cannot answer all users' concerns in one shot
  - Many different stakeholders
  - Many different objectives
  - Many different expertise

XAI: Let’s Add some Semantics

What is the impact of semantic representation on units in Neural Networks?

Jesse Mu, Jacob Andreas: Compositional Explanations of Neurons. NeurIPS 2020
Evaluation
XAI Evaluation

- **Comprehensibility**: How much effort for correct human interpretation?
- **Succinctness**: How concise and compact is the explanation?
- **Actionability**: What can one action, do with the explanation?
- **Reusability**: Could the explanation be personalized?
- **Accuracy**: How accurate and precise is the explanation?
- **Completeness**: Is the explanation complete, partial, restricted?

Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan
XAI Evaluation – more Human (Role)-based Evaluation Needed

<table>
<thead>
<tr>
<th>Task</th>
<th>Image Recognition</th>
<th>Sentiment Analysis</th>
<th>Key Word Detection</th>
<th>Heartbeat Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Image</td>
<td>Text</td>
<td>Audio</td>
<td>Sensory data (ECG)</td>
</tr>
<tr>
<td>Dataset</td>
<td>Cifar-10</td>
<td>Sentiment140</td>
<td>Speech Commands</td>
<td>MIT-BIH Arrhythmia</td>
</tr>
<tr>
<td>Classes</td>
<td>10</td>
<td>2</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 2: An overview of the application tasks and datasets used in our study

<table>
<thead>
<tr>
<th>Explanation Method</th>
<th>Image Study</th>
<th>Text Study</th>
<th>Audio Study</th>
<th>Sensor Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIME</td>
<td>47.7 ± 4.5%</td>
<td>70.4 ± 3.6%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Anchor</td>
<td>38.9 ± 4.3%</td>
<td>25.8 ± 3.5%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SHAP</td>
<td>33.7 ± 4.3%</td>
<td>59.9 ± 3.8%</td>
<td>34.7 ± 4.8%</td>
<td>32.8 ± 3.3%</td>
</tr>
<tr>
<td>Saliency Maps</td>
<td>39.4 ± 4.3%</td>
<td>-</td>
<td>46.1 ± 5.1%</td>
<td>40.4 ± 3.5%</td>
</tr>
<tr>
<td>Grad-CAM++</td>
<td>50.8 ± 4.5%</td>
<td>-</td>
<td>48.1 ± 5.3%</td>
<td>42.0 ± 3.5%</td>
</tr>
<tr>
<td>ExMatchina</td>
<td>89.6 ± 2.6%</td>
<td>43.7 ± 3.9%</td>
<td>70.9 ± 4.7%</td>
<td>84.8 ± 2.5%</td>
</tr>
</tbody>
</table>

Table 3: Results of the Mechanical Turk study evaluating user preference for DNN explanation methods across image, text, audio, and sensory input domains. Survey questions individually compare two methods at a time, with each explanation compared to all other available methods equally. Results indicate the rate by which users selected a particular method when it is an available explanation, with 95% bootstrap confidence intervals.

Figure 2: Depiction of surveyed explanation methods for image, text, and ECG input.
XAI Evaluation – more Human (Role)-based Evaluation Needed

<table>
<thead>
<tr>
<th>Domain</th>
<th>Model Purpose</th>
<th>Explainability Technique</th>
<th>Stakeholders</th>
<th>Evaluation Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finance</td>
<td>Loan Repayment</td>
<td>Feature Importance</td>
<td>Loan Officers</td>
<td>Completeness [34]</td>
</tr>
<tr>
<td>Insurance</td>
<td>Risk Assessment</td>
<td>Feature Importance</td>
<td>Risk Analysts</td>
<td>Completeness [34]</td>
</tr>
<tr>
<td>Content Moderation</td>
<td>Malicious Reviews</td>
<td>Feature Importance</td>
<td>Content Moderators</td>
<td>Completeness [34]</td>
</tr>
<tr>
<td>Finance</td>
<td>Cash Distribution</td>
<td>Feature Importance</td>
<td>ML Engineers</td>
<td>Sensitivity [69]</td>
</tr>
<tr>
<td>Facial Recognition</td>
<td>Smile Detection</td>
<td>Feature Importance</td>
<td>ML Engineers</td>
<td>Faithfulness [7]</td>
</tr>
<tr>
<td>Content Moderation</td>
<td>Sentiment Analysis</td>
<td>Feature Importance</td>
<td>QA ML Engineers</td>
<td>$\ell_2$ norm</td>
</tr>
<tr>
<td>Healthcare</td>
<td>Medicare access</td>
<td>Counterfactual Explanations</td>
<td>ML Engineers</td>
<td>Normalized $\ell_1$ norm</td>
</tr>
<tr>
<td>Content Moderation</td>
<td>Object Detection</td>
<td>Adversarial Perturbation</td>
<td>QA ML Engineers</td>
<td>$\ell_2$ norm</td>
</tr>
</tbody>
</table>

Table 1: Summary of select deployed local explainability use cases

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, Peter Eckersley: Explainable machine learning in deployment.

FAT* 2020: 648-657

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, Finale Doshi-Velez: An Evaluation of the Human-Interpretability of Explanation. CoRR abs/1902.00006 (2019)

Observations: patient, wearing glasses, lazy
Recommendation: milk, gauze

The alien’s preferences:
- lazy or nervous → nodding
- nodding and wearing glasses → clumsy
- helpful or clumsy → smile
- helpful and cold or brave and passive → candy or dairy and fruit
- sleepy or patient and obedient → spinach and grains or dairy
- brown or red or patient or laughing → dairy and fruit or grains
- crying or sleepy and faithful → grains and option or fruit

Ingredients:
- Vegetables: corn, carrot, spinach
- Spinach: broccoli, chilies, cheese
- Dairy: milk, butter, yogurt
- Fruits: orange, banana, grapes
- Candy: chocolates, taffy, caramel
- Graham logs, rice, pasta

Is the alien happy with the recommended meal?
- Yes
- No

Through Amazon Mechanical Turk (900 subjects all together)