THALES

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

Freddy Lecue, Chief AI Scientist
@freddylecue

Context


BNSF
BNSF

Markets we serve where XAI is crucial

Aerospace

Space

Ground Transportation

Defence

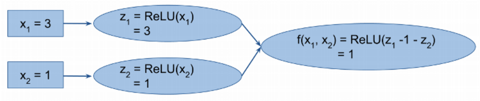
Security

Trusted Partner For A Safer World

THALES

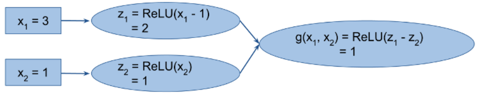
Approach

Some XAI Approaches (1)



Network $f(x_1, x_2)$
Attributions 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	$x_1 = 1.5,\ x_2 = -0.5$



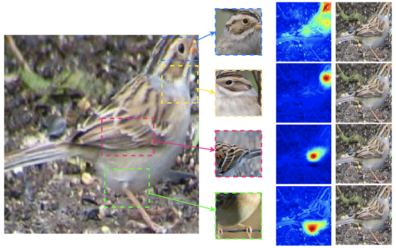
Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

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

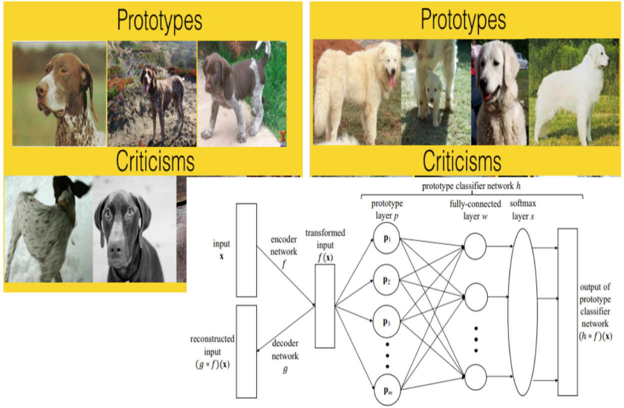
Attribution for Deep Network (Integrated gradient-based)

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje: Learning Important Features Through Propagating Activation Differences. ICML 2017: 3145-3153



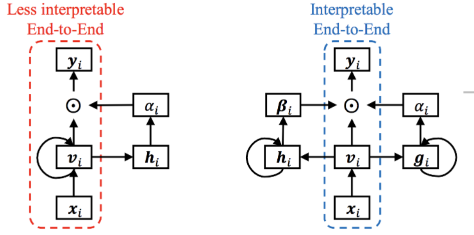
Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin: This looks like that: deep learning for interpretable image recognition. CoRR abs/1806.10574 (2018)



Example-based / Prototype

Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin: Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. AAAI 2018: 3530-3537

Been Kim, Oluwasanmi Koyejo, Rajiv Khanna:Examples are not enough, learn to criticize! Criticism for Interpretability. NIPS 2016: 2280-2288

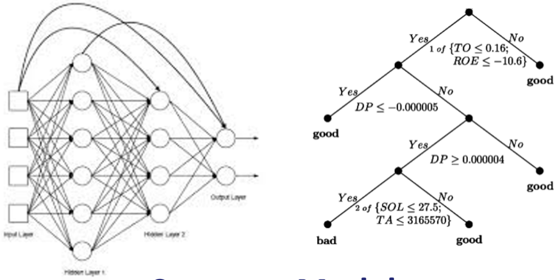


(a) Standard attention model (b) RETAIN model

Attention Mechanism

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, Walter F. Stewart: RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. NIPS 2016: 3504-3512

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, 2015



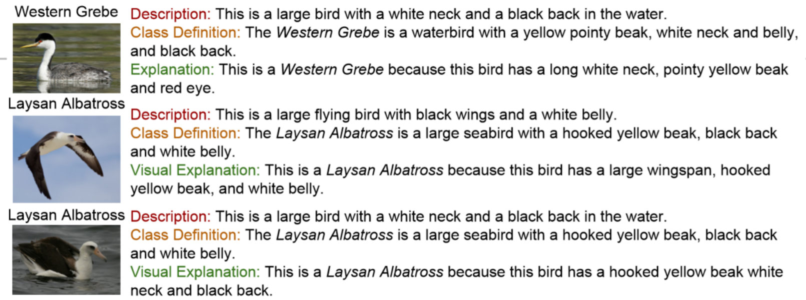
Surogate Model

Mark Craven, Jude W. Shavlik: Extracting Tree-Structured Representations of Trained Networks. NIPS 1995: 24-30

Some XAI Approaches (2)

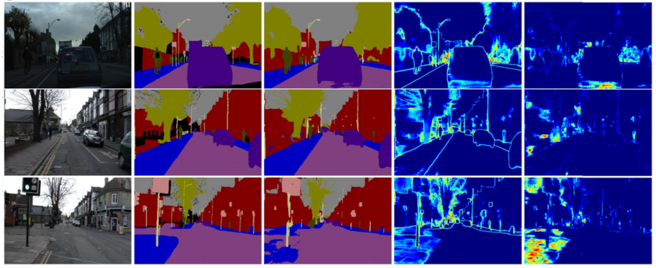
Interpretable Units

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327



Visual Explanation

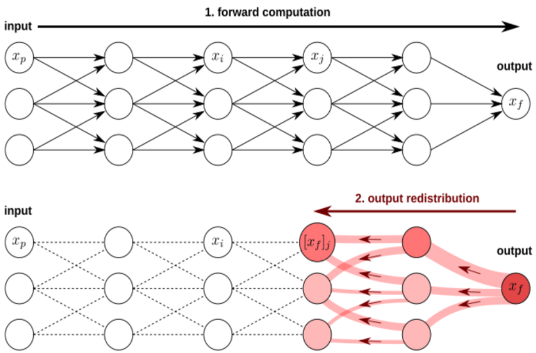
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19

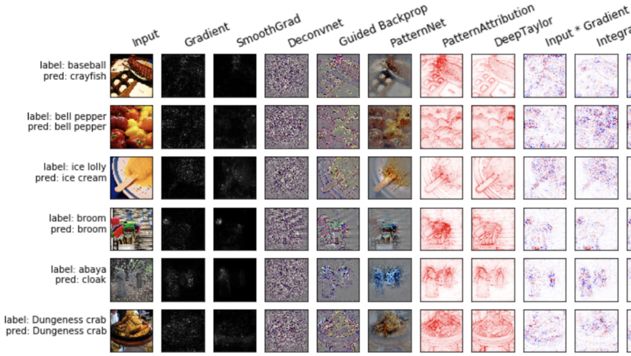


(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Uncertainty Map

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





Saliency Map / Features Attributiin-based

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps. NeurIPS 2018: 9525-9536

Some XAI Approaches – Towards Semantics

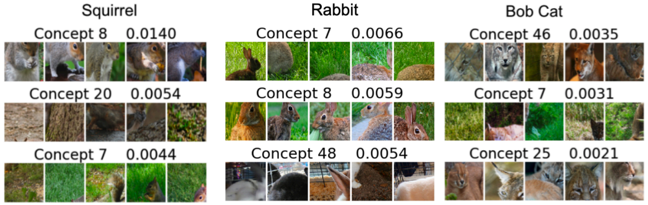
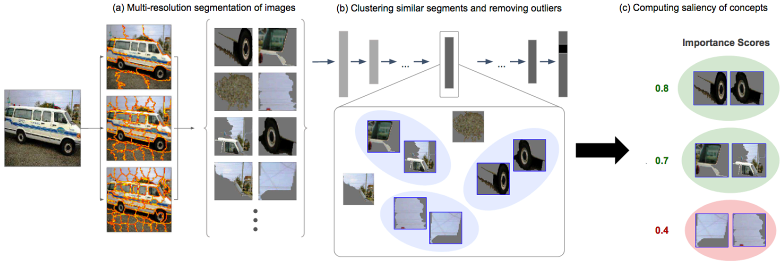


Figure 3: Concept examples with the samples that are the nearest to concept vectors in the activation space in AwA. The per-class ConceptSHAP score is listed above the images.

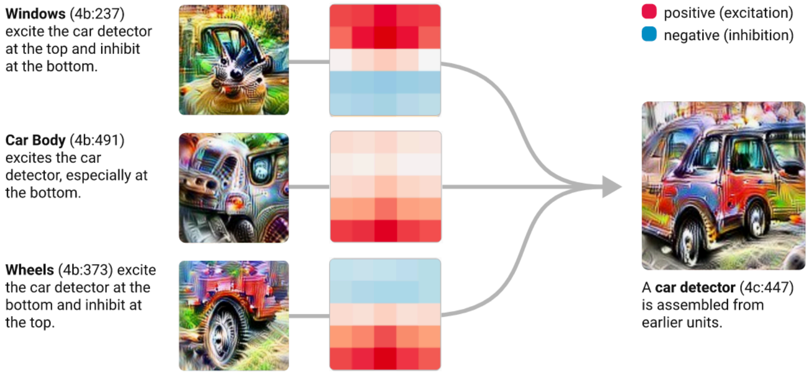
ConceptSHAP

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

Amirata Ghorbani, James Wexler, James Y. Zou, Been Kim:Towards Automatic Concept-based Explanations. NeurIPS 2019: 9273-9282



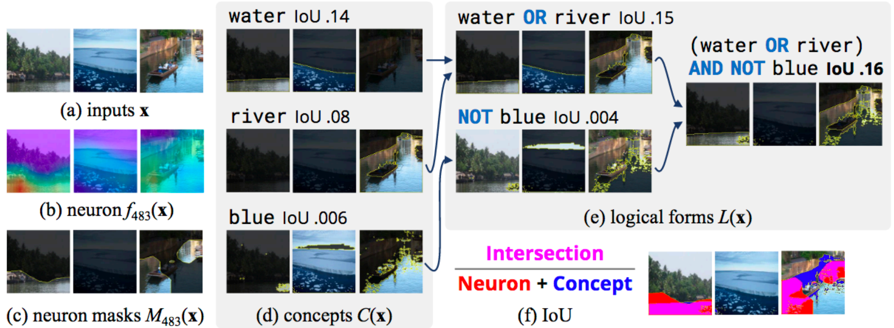


Figure 1: Given a set of inputs (a) and scalar neuron activations (b) converted into binary masks (c), we generate an explanation via beam search, starting with an inventory of primitive concepts (d), then incrementally building up more complex logical forms (e). We attempt to maximize the IoU score of an explanation (f); depicted is the IoU of $M_{483}(\mathbf{x})$ and `(water OR river) AND NOT blue`.

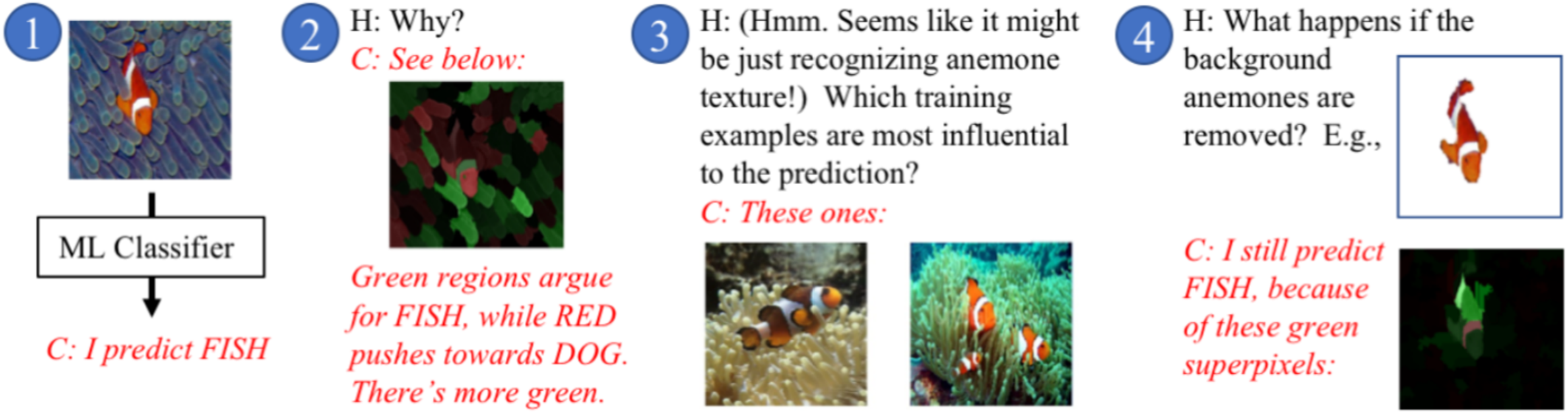
Circuits in CNNs

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

Compositional Explanations

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

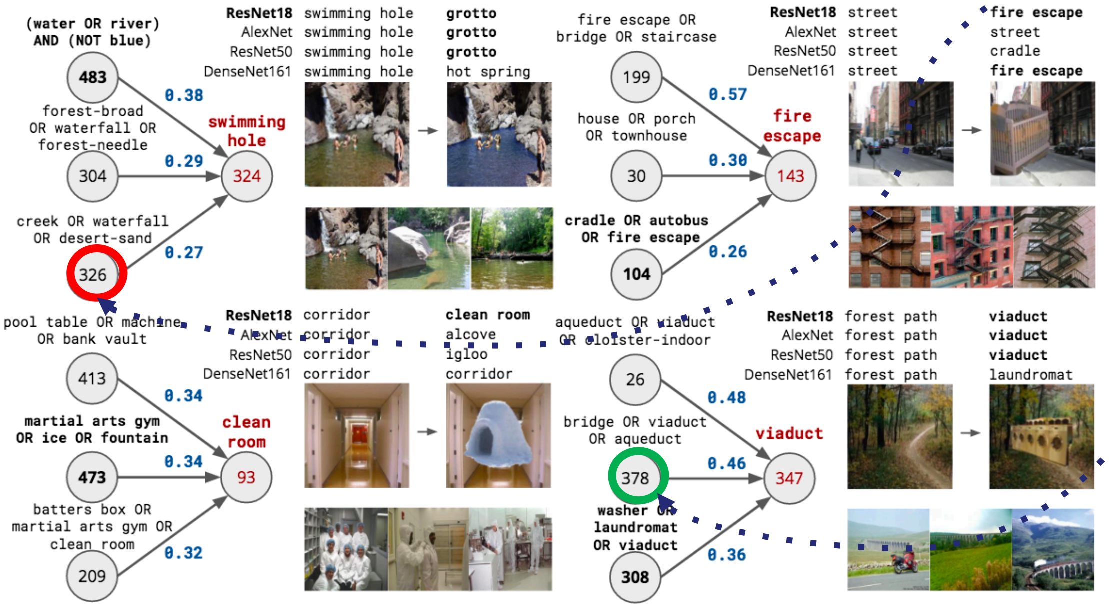
Example of an End-to-End XAI System



- **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

Weld, D., and Gagan Bansal. "The challenge of crafting intelligible intelligence." Communications of ACM (2018).

XAI: Let's Add some Semantics



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

Low-level features to high-level features

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

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

Task	Image Recognition	Sentiment Analysis	Key Word Detection	Heartbeat Classification
Domain	Image	Text	Audio	Sensory data (ECG)
Dataset	Cifar-10	Sentiment140	Speech Commands	MIT-BIH Arrhythmia
Classes	10	2	10	5

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

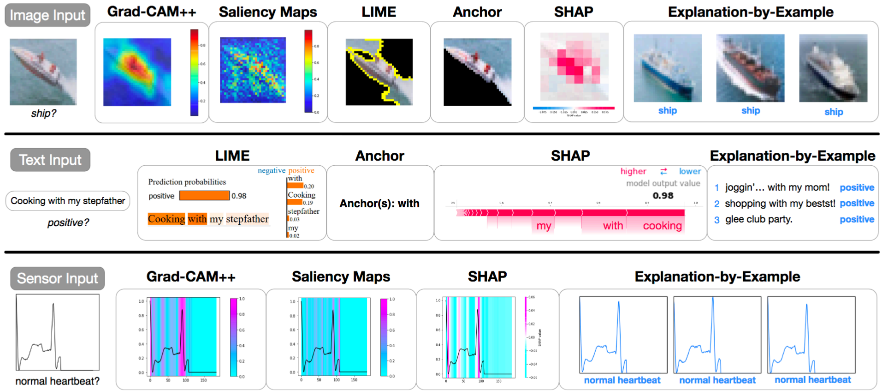


Figure 2: Depiction of surveyed explanation methods for image, text, and ECG input.

Explanation Method	Image Study	Text Study	Audio Study	Sensor Study
LIME	$47.7 \pm 4.5\%$	$\mathbf{70.4 \pm 3.6\%}$	-	-
Anchor	$38.9 \pm 4.3\%$	$25.8 \pm 3.5\%$	-	-
SHAP	$33.7 \pm 4.3\%$	$59.9 \pm 3.8\%$	$34.7 \pm 4.8\%$	$32.8 \pm 3.3\%$
Saliency Maps	$39.4 \pm 4.3\%$	-	$46.1 \pm 5.1\%$	$40.4 \pm 3.5\%$
Grad-CAM++	$50.8 \pm 4.5\%$	-	$48.1 \pm 5.3\%$	$42.0 \pm 3.5\%$
ExMatchina	$\mathbf{89.6 \pm 2.6\%}$	$43.7 \pm 3.9\%$	$\mathbf{70.9 \pm 4.7\%}$	$\mathbf{84.8 \pm 2.5\%}$

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.

Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, Mani B. Srivastava:How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods. NeurIPS 2020

XAI Evaluation – more Human (Role)-based Evaluation Needed

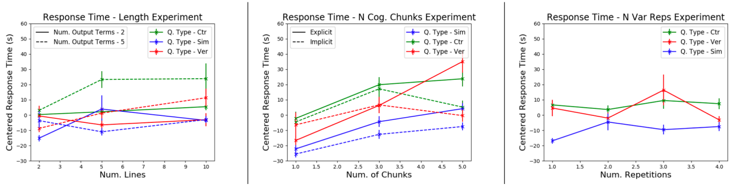
Domain	Model Purpose	Explainability Technique	Stakeholders	Evaluation Criteria
Finance	Loan Repayment	Feature Importance	Loan Officers	Completeness [34]
Insurance	Risk Assessment	Feature Importance	Risk Analysts	Completeness [34]
Content Moderation	Malicious Reviews	Feature Importance	Content Moderators	Completeness [34]
Finance	Cash Distribution	Feature Importance	ML Engineers	Sensitivity [69]
Facial Recognition	Smile Detection	Feature Importance	ML Engineers	Faithfulness [7]
Content Moderation	Sentiment Analysis	Feature Importance	QA ML Engineers	ℓ_2 norm
Healthcare	Medicare Access	Counterfactual Explanations	ML Engineers	normalized ℓ_1 norm
Content Moderation	Object Detection	Adversarial Perturbation	QA ML Engineers	ℓ_2 norm

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

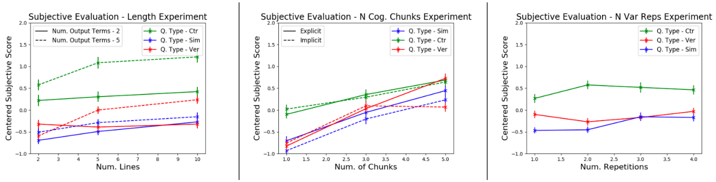
Through Amazon Mechanical Turk (900 subjects all together)

Response Time



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)

Subjective Satisfaction



Accuracy