Uncovering the Semantics of Deep Neural Networks with Knowledge Graphs

Freddy Lecue (@freddylecue)
http://www-sop.inria.fr/members/Freddy.Lecue/

Global Software Technology Summit 2022
Edinburgh

THALES

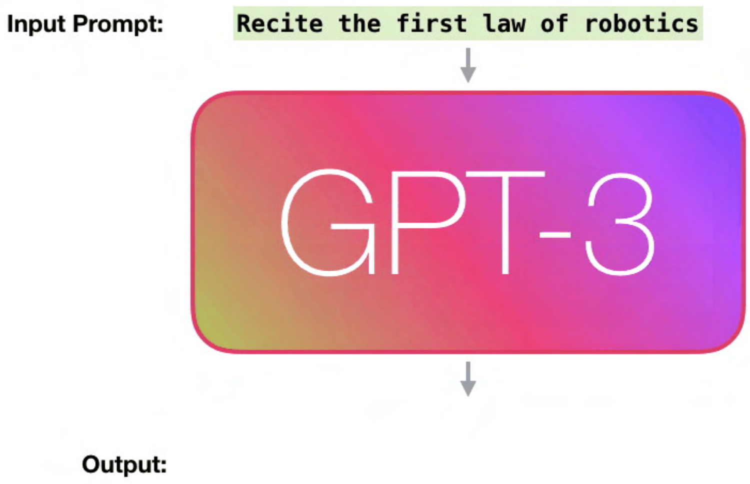
Inria
INVENTEURS DU MONDE NUMÉRIQUE

July 7, 2022

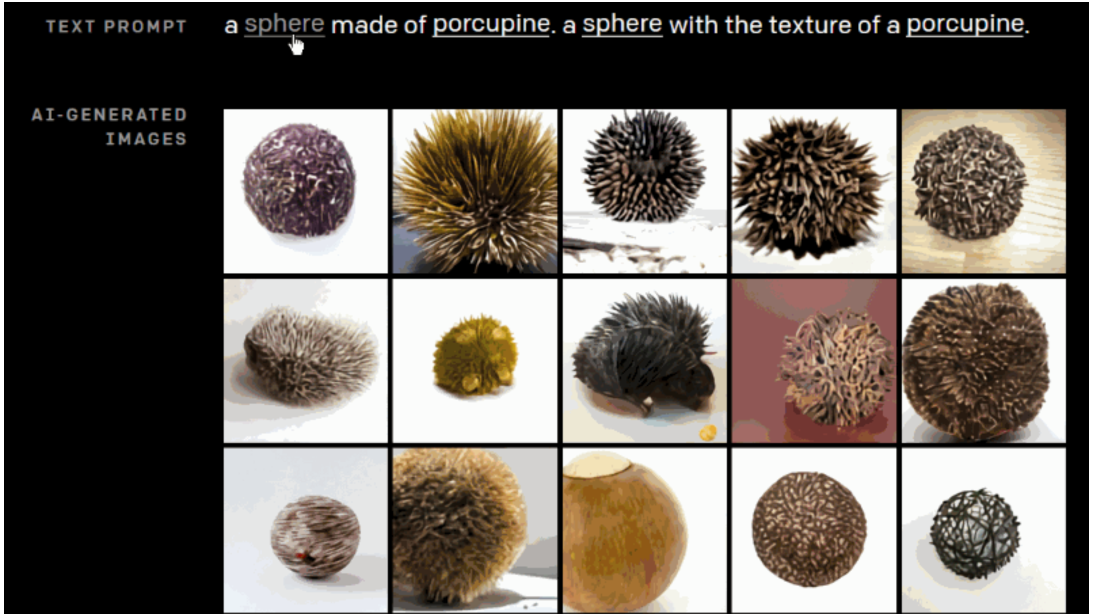
https://tinyurl.com/2p83nt59

Large-Scale Models

From Language Models such as Open AI GPT3



https://jalammar.github.io/how-gpt3-works-visualizations-animations/



Created by: Mckay Wrigley

with Applications: (1) Programming with OpenAI Codex

sentiments.ts write_sql.go parse_expenses.py addresses.rb

```
#!/usr/bin/env ts-node

import { fetch } from "fetch-h2";

// Determine whether the sentiment of text is positive
// Use a web service
async function isPositive(text: string): Promise<boolean> {
  const response = await fetch(`http://text-processing.com/api/sentiment/`, {
    method: "POST",
    body: `text=${text}`,
    headers: {
      "Content-Type": "application/x-www-form-urlencoded",
    },
  });
  const json = await response.json();
  return json.label === "pos";
}
```

Copilot

Replay

https://github.com/features/copilot/

with Applications: (2) Solving Mathematical Questions with Google Minerva

Question: Assume that the variance of the first n natural numbers is 10, and the variance of the first m even natural numbers is 16. Compute $m + n$.

Model output: Let σ^2 be the variance of the first m even natural numbers, and τ^2 be the variance of the first n natural numbers. Then, we have $\tau^2 = \frac{n^2-1}{12}$ and $\sigma^2 = \frac{m^2-1}{3}$. We solve the system of equations:

$$\frac{n^2-1}{12} = 10$$
$$\frac{m^2-1}{3} = 16.$$

This translates into the system

$$n^2 = 121$$
$$m^2 = 48 + 1 = 49.$$

Therefore, $n = \boxed{11}$ and $m = \boxed{7}$, so $n + m = \boxed{18}$.

Question: For every $a, b, b \neq a$ prove that

$$\frac{a^2+b^2}{2} > \left(\frac{a+b}{2}\right)^2.$$

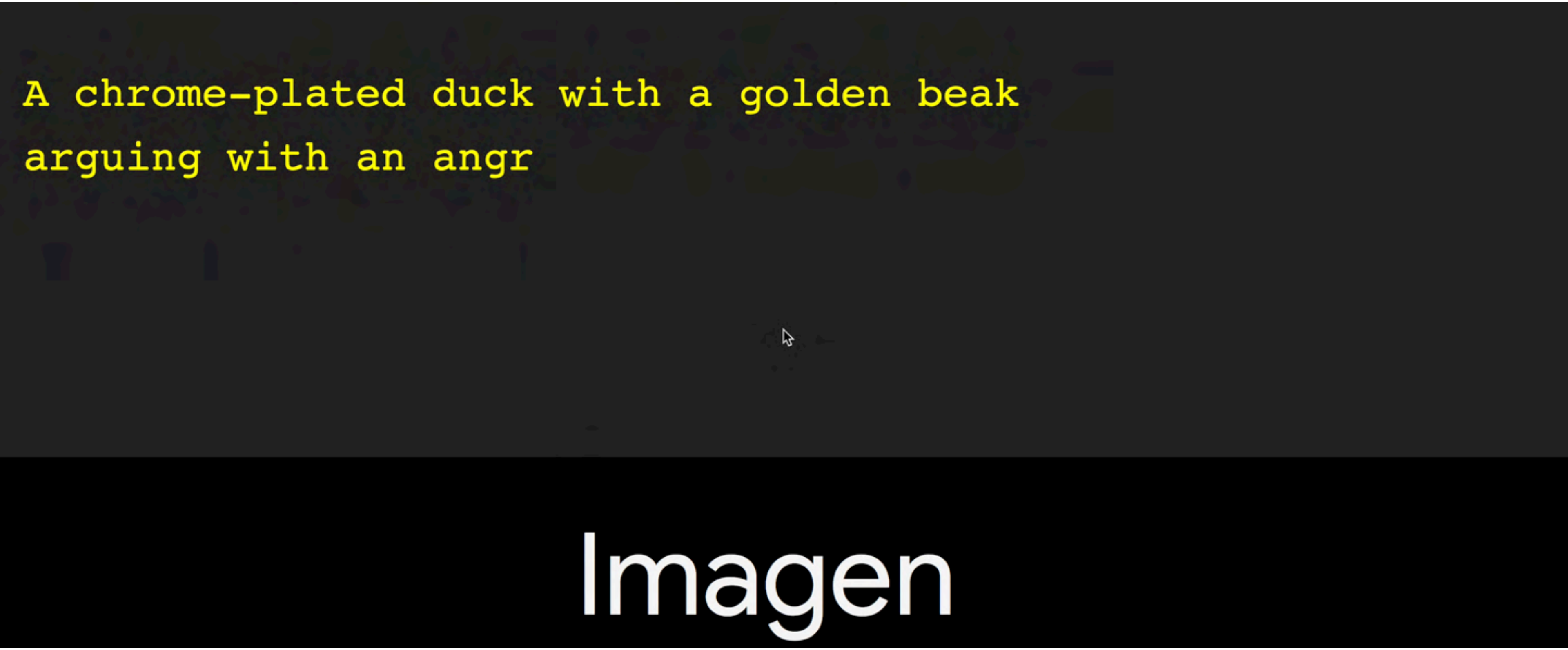
Model output:

$$\frac{a^2+b^2}{2} > \left(\frac{a+b}{2}\right)^2$$
$$\iff \frac{a^2+b^2}{2} > \frac{a^2+b^2+2ab}{4}$$
$$\iff a^2+b^2 > \frac{a^2+b^2+2ab}{2}$$
$$\iff 2a^2+2b^2 > a^2+b^2+2ab$$
$$\iff a^2+b^2 > 2ab$$
$$\iff a^2+b^2-2ab > 0$$
$$\iff (a-b)^2 > 0$$

which is true, because the square of a real number is positive.

https://ai.googleblog.com/2022/06/minerva-solving-quantitative-reasoning.html

To Text-to-Image Models (1) such as Google Imagen



To Text-to-Image Models (2) such as Open AI DALL-E-2

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

riding a horse lounging in a tropical resort in space playing basketball with cats in space

in a photorealistic style in the style of Andy Warhol as a pencil drawing

→

DALL·E 2

https://openai.com/dall-e-2/

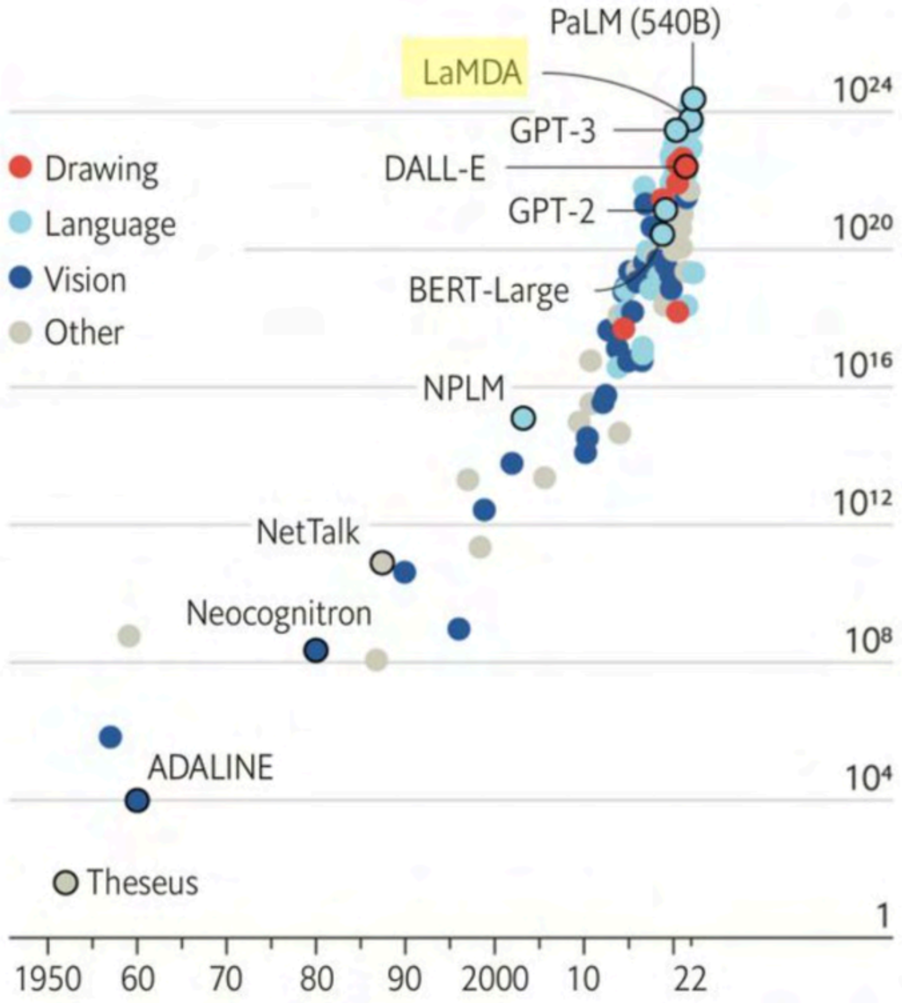
And Towards Generalist Models such as DeepMind Gato

Inspired by progress in large-scale language modelling, DeepMind applied a similar approach towards building a single generalist agent beyond the realm of text outputs. Gato works as a multi-modal, multi-task, multi-embodiment generalist policy. The same network with the same weights can play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens.

https://www.deepmind.com/publications/a-generalist-agent

Gato

All are Great Examples of what can be achieved by larger and larger models



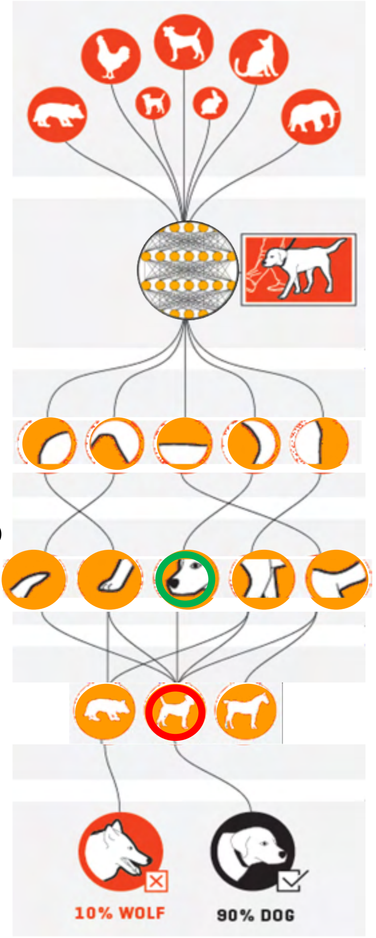
AI Training runs, estimated computed resources used,
Floating-point operations, selected systems, by type, log scale

Compute Trends Across Three Eras of Machine Learning. Jaime Sevilla.
https://ar5iv.labs.arxiv.org/html/2202.05924

They are Artificial Neural Network Powered-Models with different flavors

Input Layer

Training Data



Input (unlabeled image)

Hidden Layer

Neurons respond to simple shapes

1st Layer

Neurons respond to more complex structures

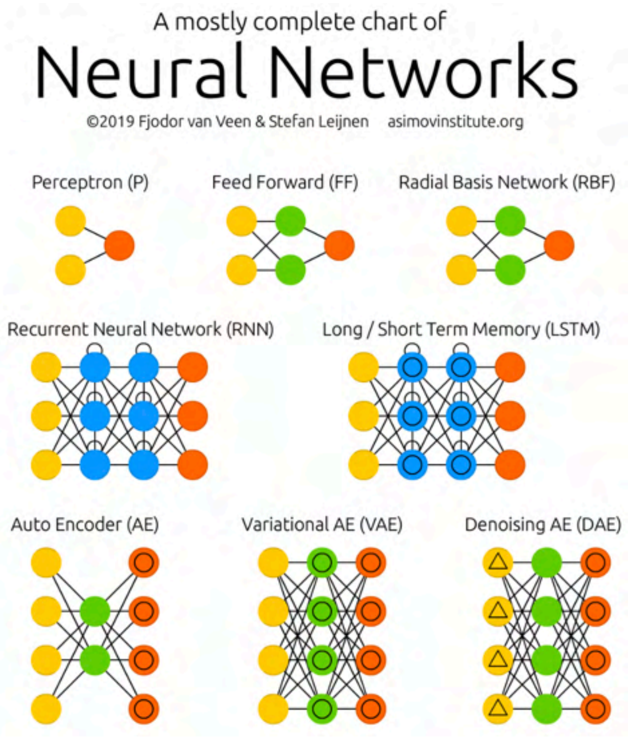
2nd Layer

Neurons respond to highly complex, abstract concepts

nth Layer

Output Layer

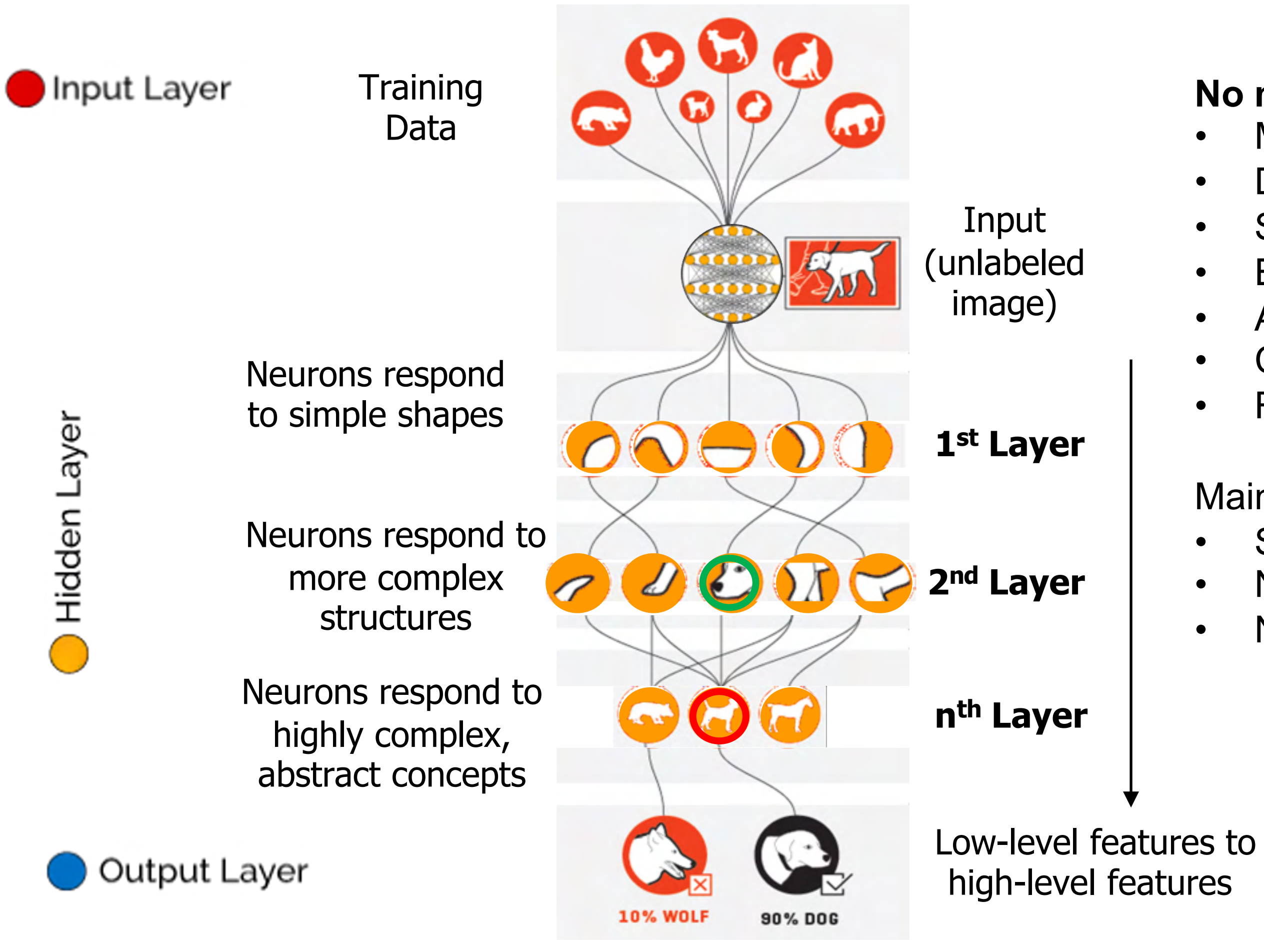
Low-level features to high-level features



https://www.asimovinstitute.org/neural-network-zoo/

Limitations

They are Artificial Neural Network Powered-Models with:



No mechanisms beyond fine-tuning for:

- Modularization / Atomization
- Discovery
- Search
- Extraction
- Adaptation
- Composition
- Re-purpose of the models parts

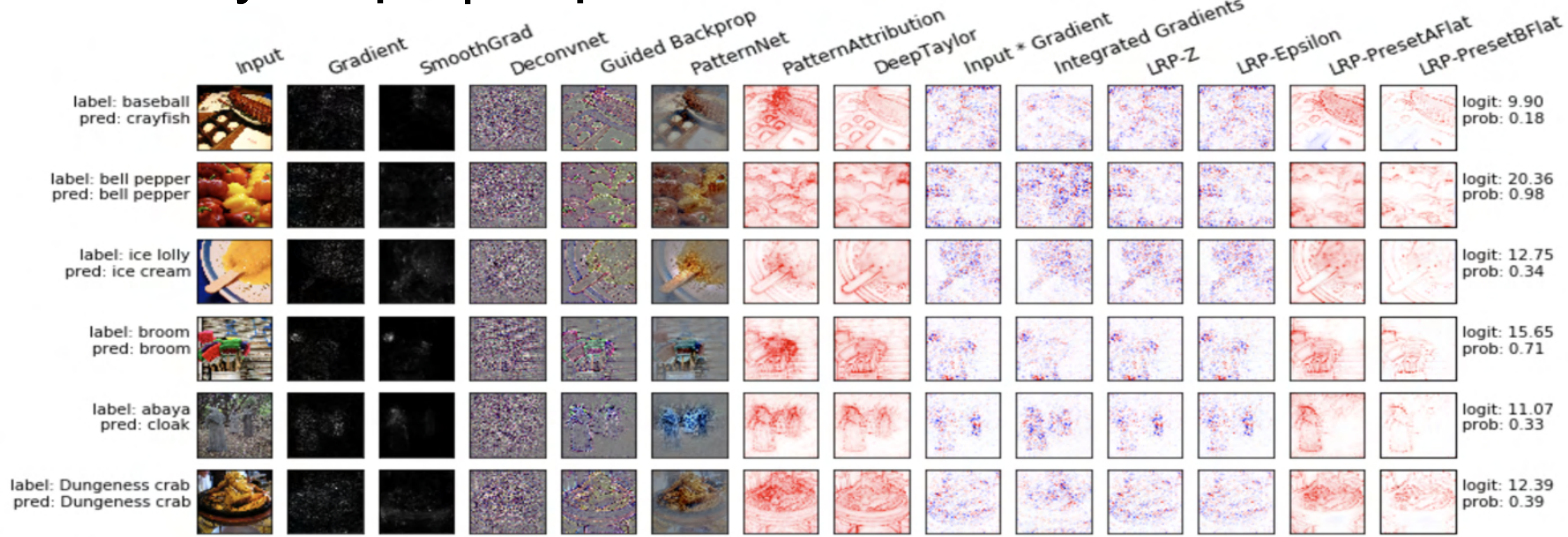
Mainly because such models have:

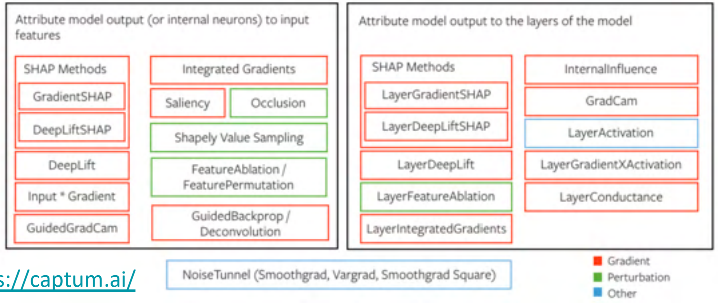
- Spurious correlation
- No explicit semantics encoded
- No causal relations enforced

Strong need for causality- / explanation-first models

Towards Knowledge Extraction and Search

Saliency Maps | Super-Pixels





https://captum.ai/

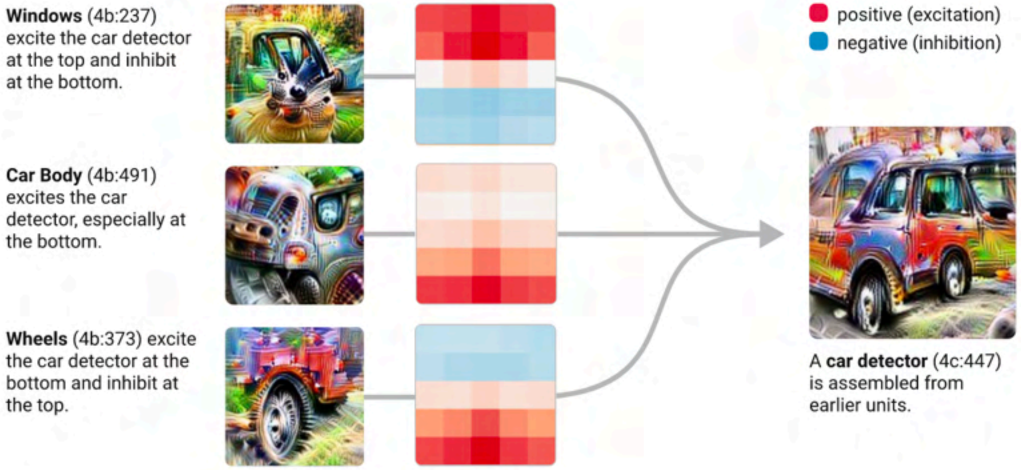
- [Interaction] No human interaction
- [Construction] Purely architecture / gradient based
- [Validation] Qualitative | Highly subjective
- [Knowledge] None is required

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

Feature Visualization

- [Interaction] No human interaction
- [Construction] Neuron activation | Content-based
- [Validation] Qualitative | ML Developer focus
- [Knowledge] Implicitly

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



CLIP
Resent 50
Layer 4

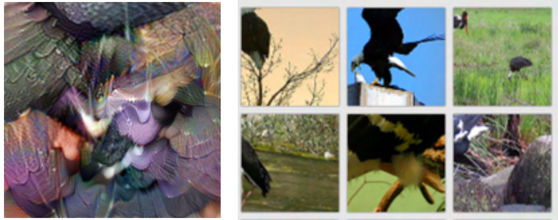
https://microscope.openai.com/models

Unit 118

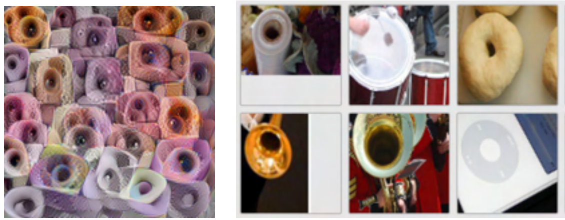
Unit 55

Resnet 50 v2
Block4/unit_3/add

Unit 546

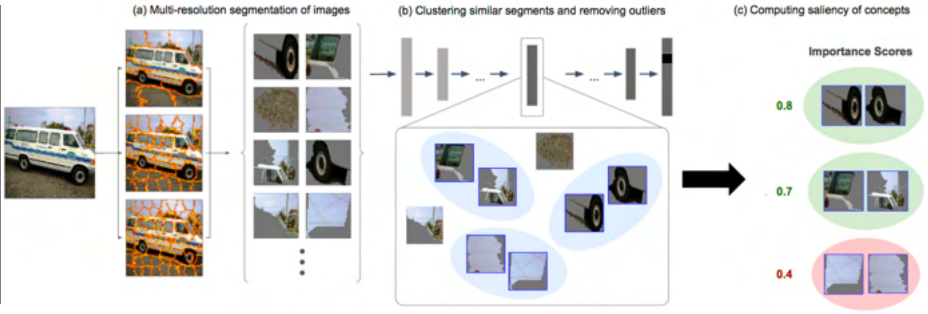


Unit 562



Concept Annotation | Towards Semantics

- [Interaction] No human interaction
- [Construction] Ablation
- [Validation] Quantitative (wrt Precision)
- [Knowledge] Implicitly



ACE

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

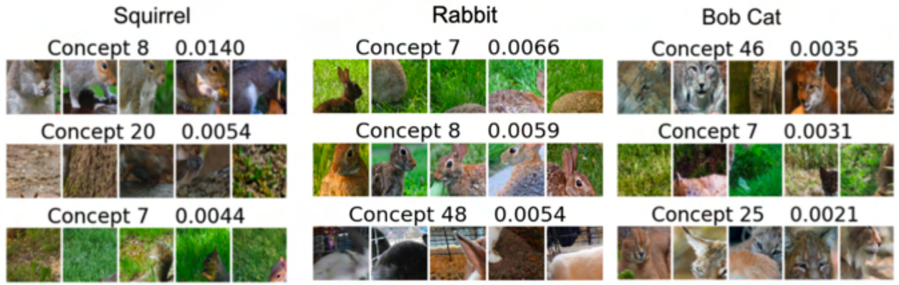
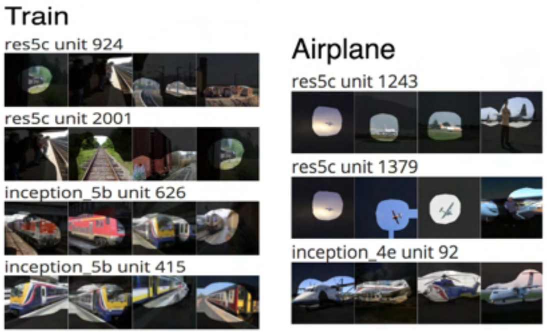


Figure 3: Concept examples with the samples that are the nearest to concept vectors in the activation space in AwA. The per-class ConceptSHAP score is listed above the images.

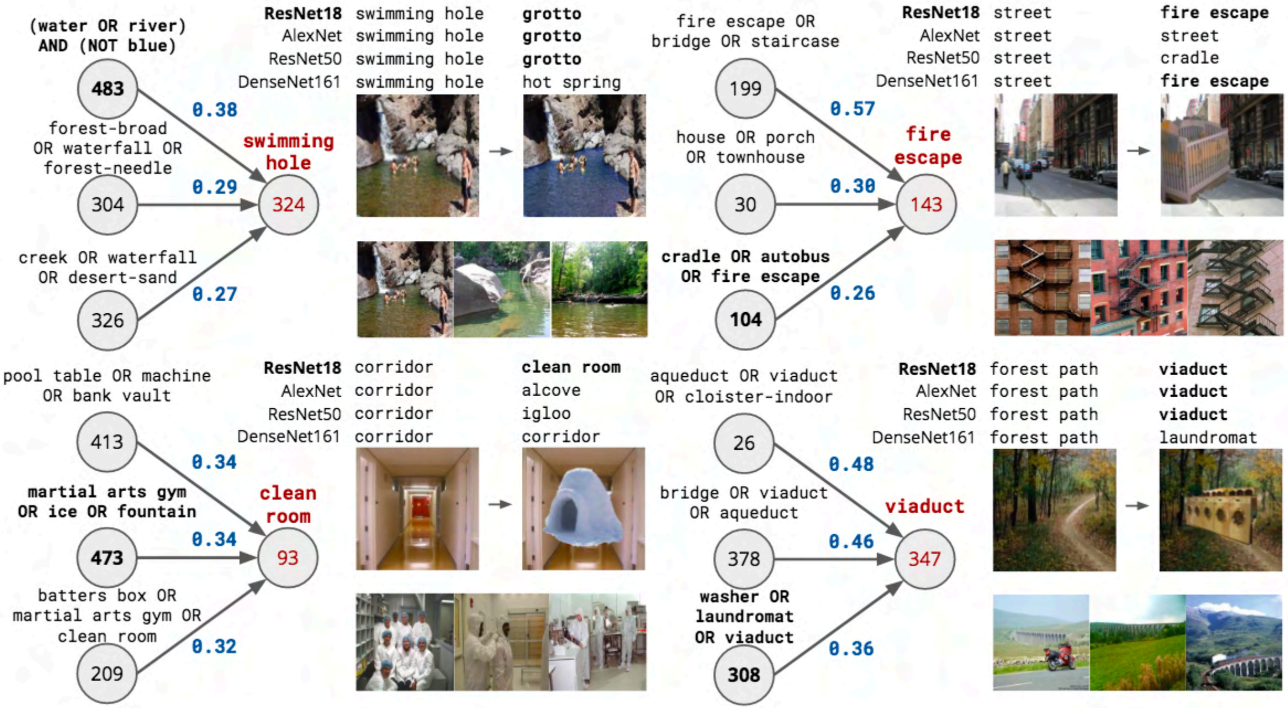
ConceptSHAP Chih-Kuan Yeh, Been Kim, Sercan Ömer Arik, Chun-Liang Li, Tomas Pfister, Pradeep Ravikumar:On Completeness-aware Concept-Based Explanations in Deep Neural Networks. NeurIPS 2020

Network Dissection | Neurons Composition

- [Interaction] No human interaction
- [Construction] Concept-firing
- [Validation] Qualitative and quantitative (wrt IoU)
- [Knowledge] Implicitly



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



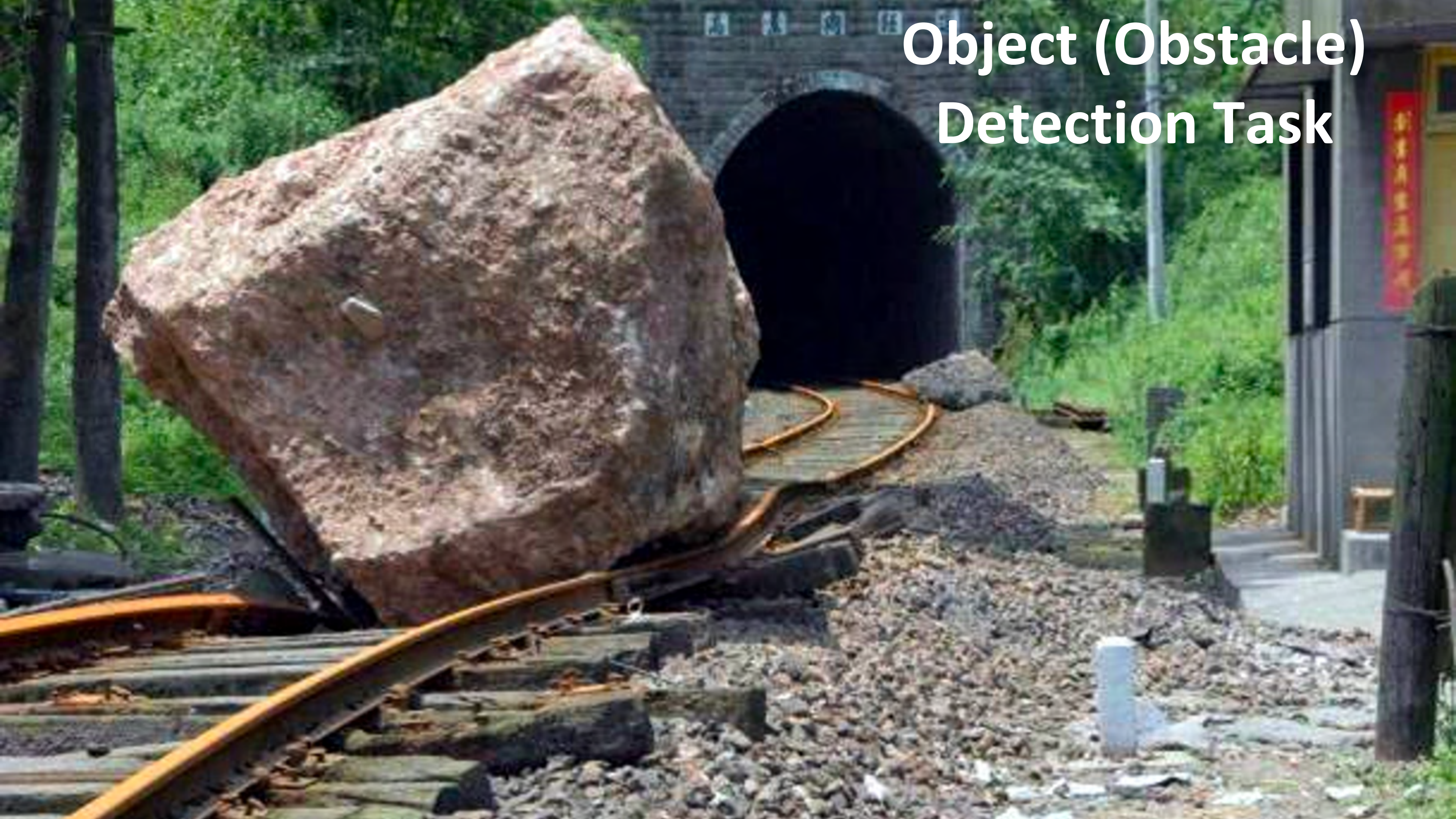
Jesse Mu, Jacob Andreas: Compositional Explanations of Neurons. NeurIPS 2020

On Boosting Neural Networks Interpretation with Graphs

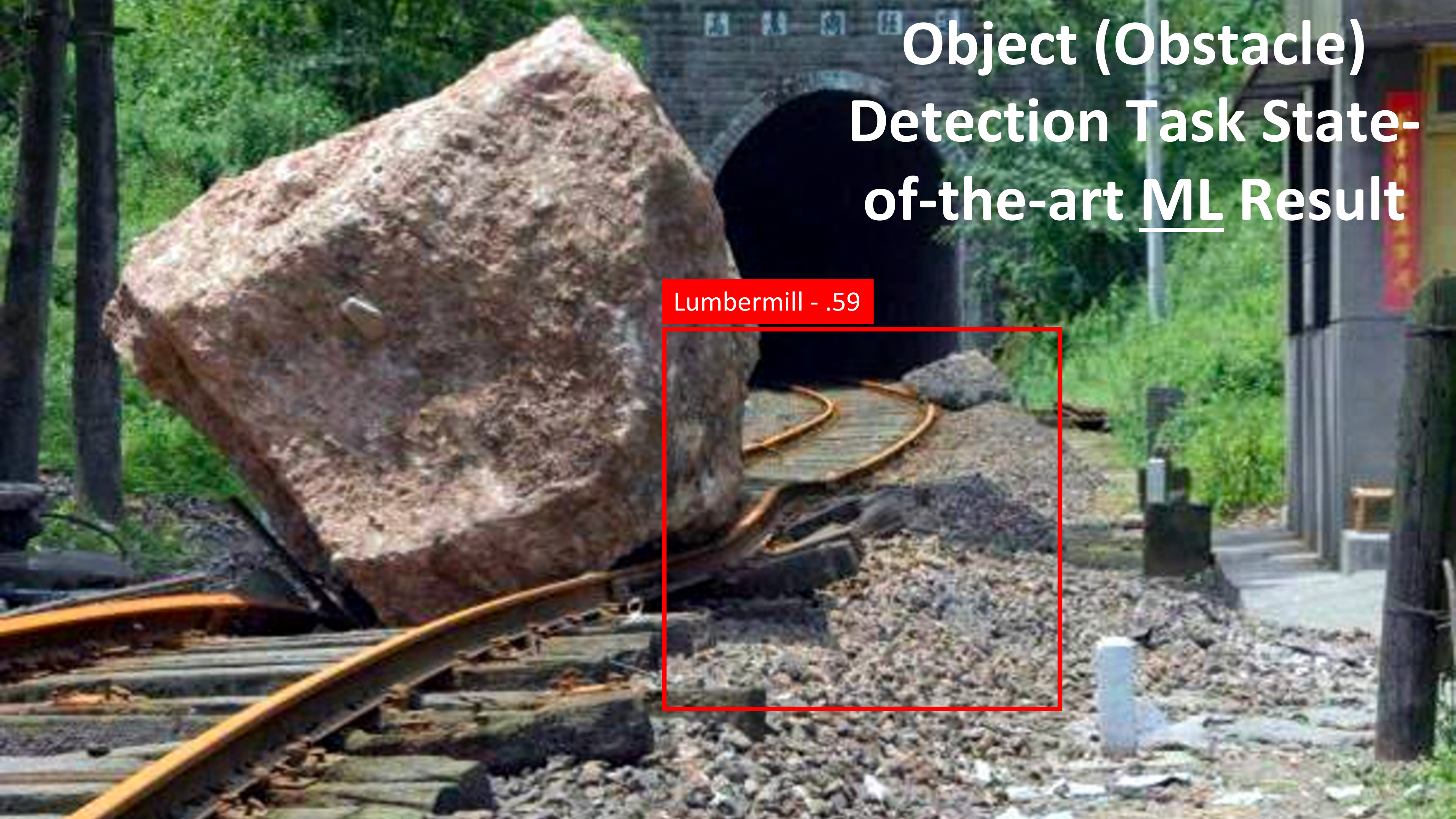
How Does it Work in Practice?

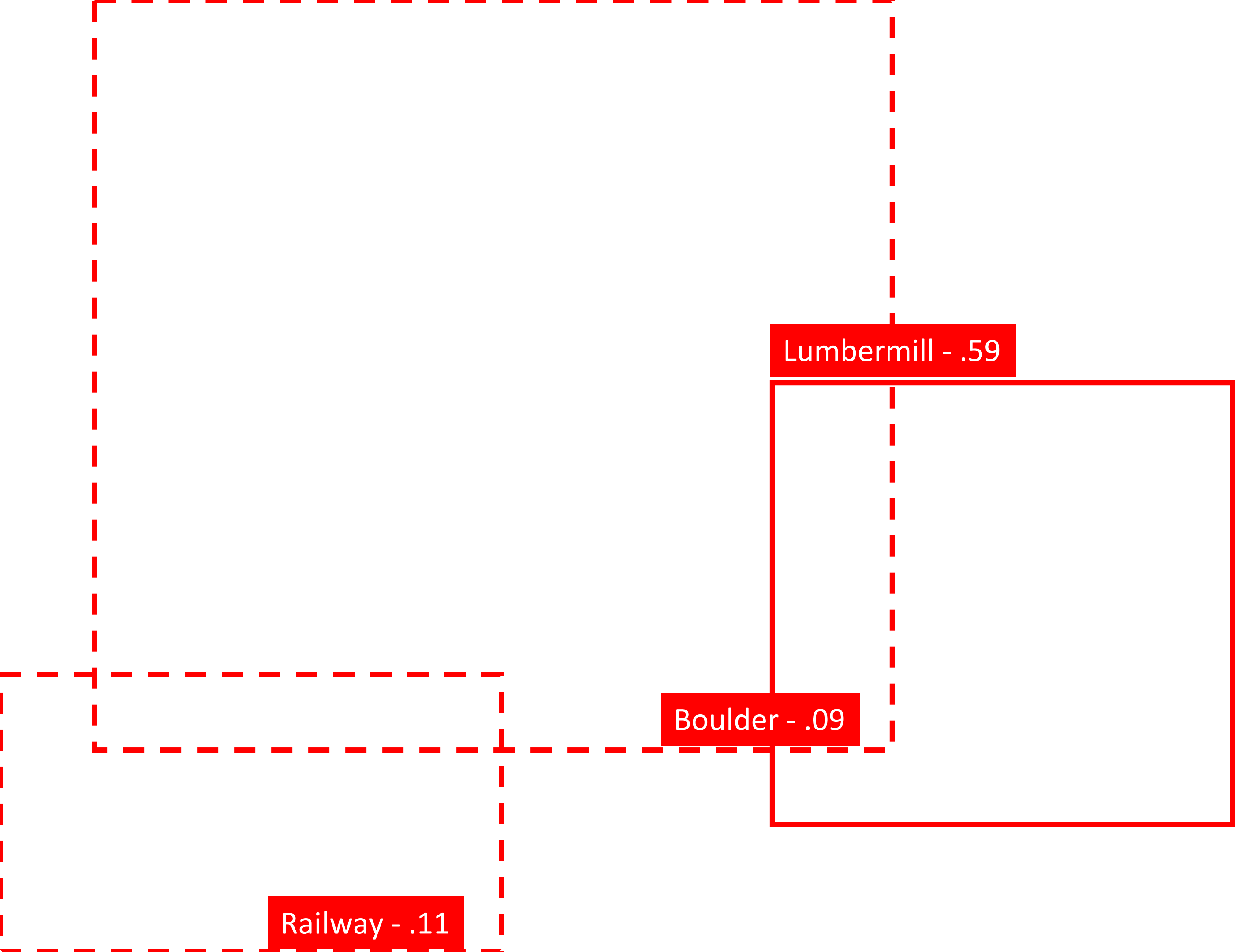
State of the Art Machine Learning Applied to Critical Systems

Object (Obstacle) Detection Task



Object (Obstacle) Detection Task State-of-the-art ML Result

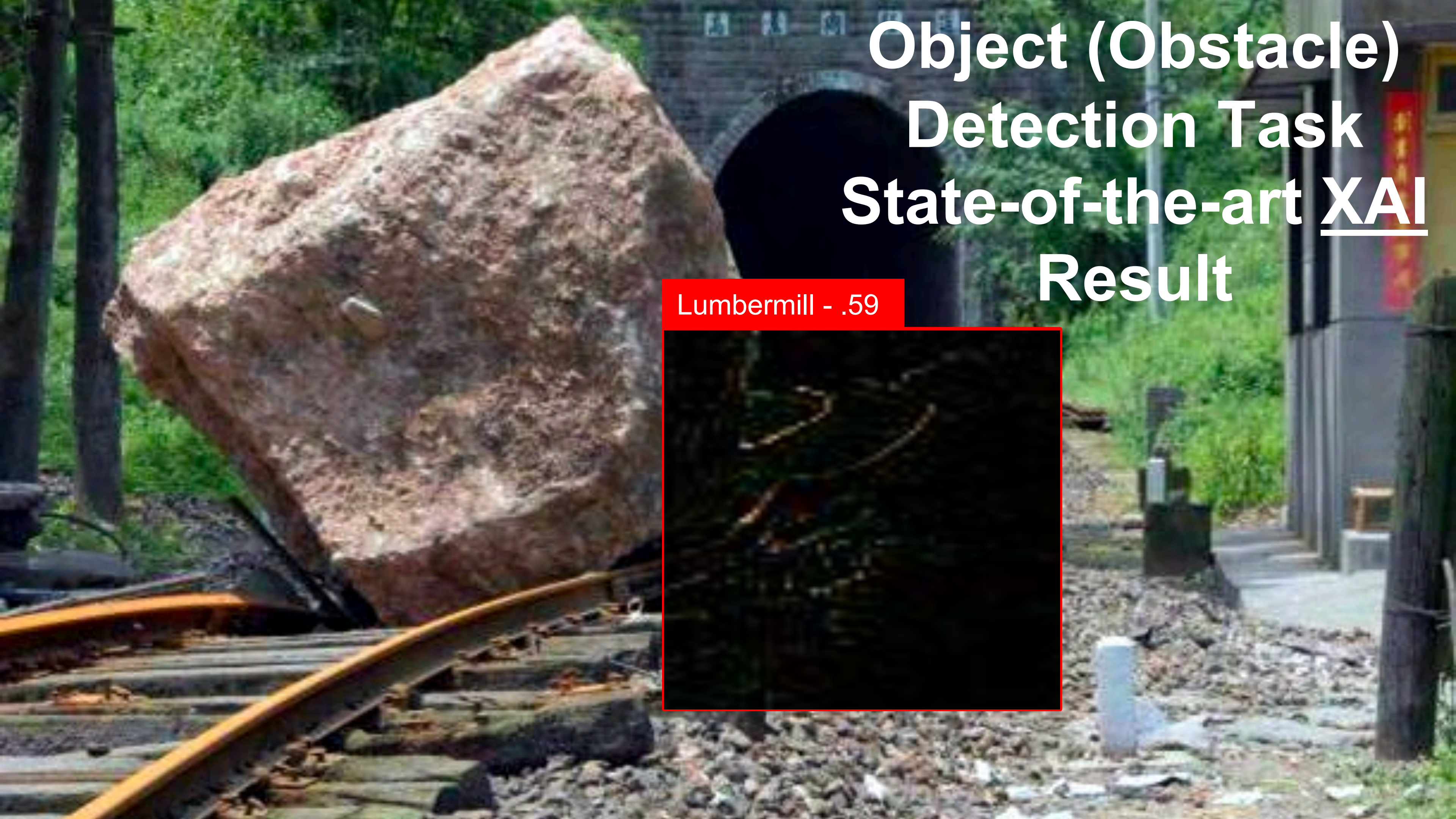



Lumbermill - .59
Boulder - .09
Railway - .11

State of the Art XAI Applied to Critical Systems


Object (Obstacle)
Detection Task
State-of-the-art XAI
Result
Lumbermill - .59

Object (Obstacle) Detection Task State-of-the-art XAI Result



Object (Obstacle) Detection Task State-of-the-art XAI Result

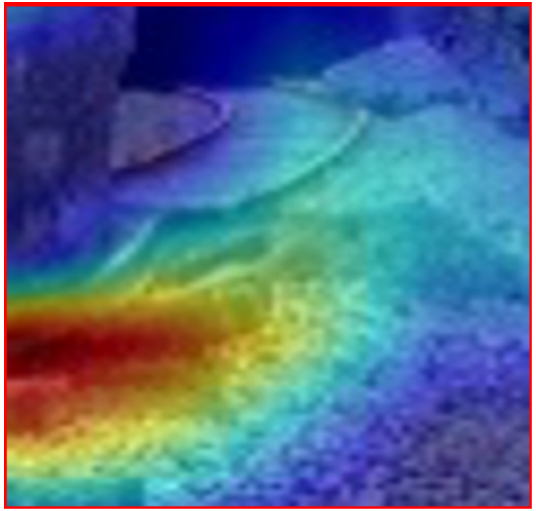
Unfortunately, this is of NO use for a human behind the system

Let’s stay back

Why this Explanation?
(meta explanation)

After Human Reasoning…

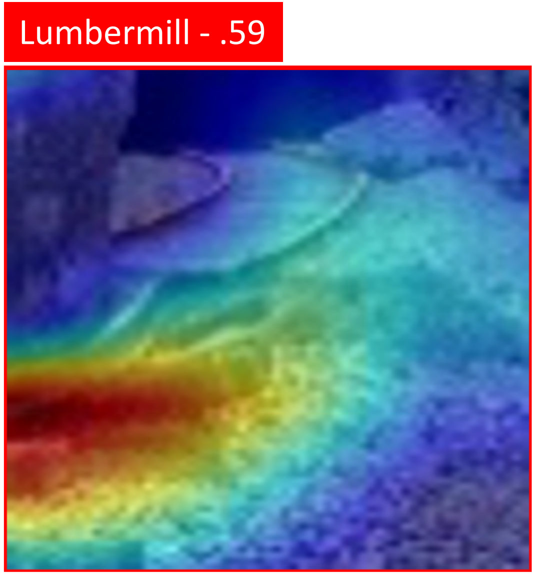
Lumbermill - .59

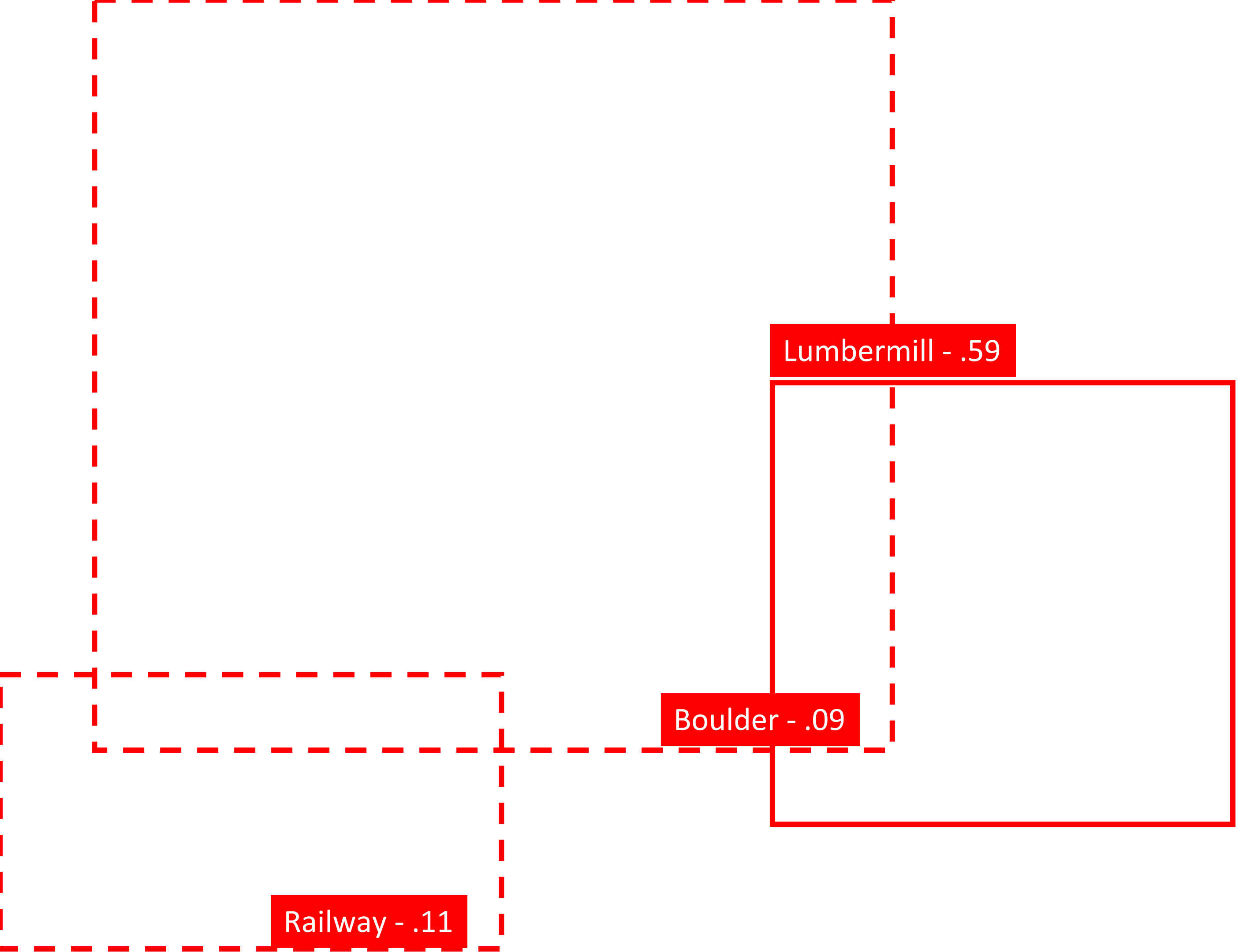


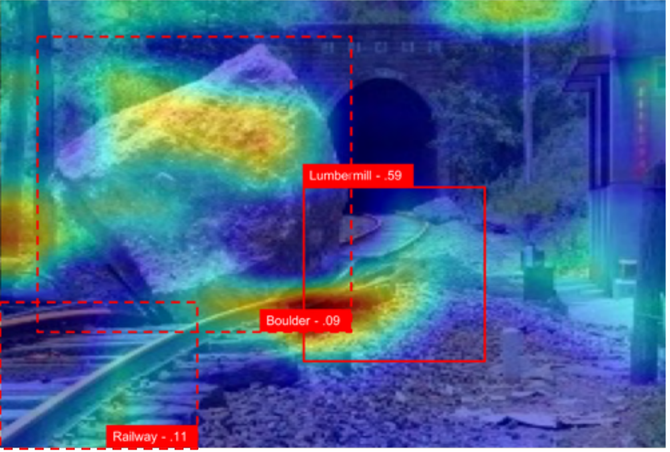
DBpedia | Browse using | Formats | Faceted Browser | Sparql Endpoint

dbo:wikiPageID	352327 (xsd:integer)
dbo:wikiPageRevisionID	734430894 (xsd:integer)
dct:subject	dbc:Sawmills dbc:Saws dbc:Ancient_Roman_technology dbc:Timber_preparation dbc:Timber_industry
http://purl.org/linguistics/gold/hypernym	dbr:Facility
rdf:type	owl:Thing dbo:ArchitecturalStructure
rdfs:comment	A sawmill or lumber mill is a facility where logs are cut into lumber. Prior to the invention of the sawmill, boards were rived (split) and planed, or more often sawn by two men with a whipsaw, one above and another in a saw pit below. The earliest known mechanical mill is the Hierapolis sawmill, a Roman water-powered stone mill at Hierapolis, Asia Minor dating back to the 3rd century AD. Other water-powered mills followed and by the 11th century they were widespread in Spain and North Africa, the Middle East and Central Asia, and in the next few centuries, spread across Europe. The circular motion of the wheel was converted to a reciprocating motion at the saw blade. Generally, only the saw was powered, and the logs had to be loaded and moved by hand. An early improvement was the developm (en)
rdfs:label	Sawmill (en)
owl:sameAs	wikidata:Sawmill dbpedia-cs:Sawmill dbpedia-de:Sawmill dbpedia-es:Sawmill

What is missing?

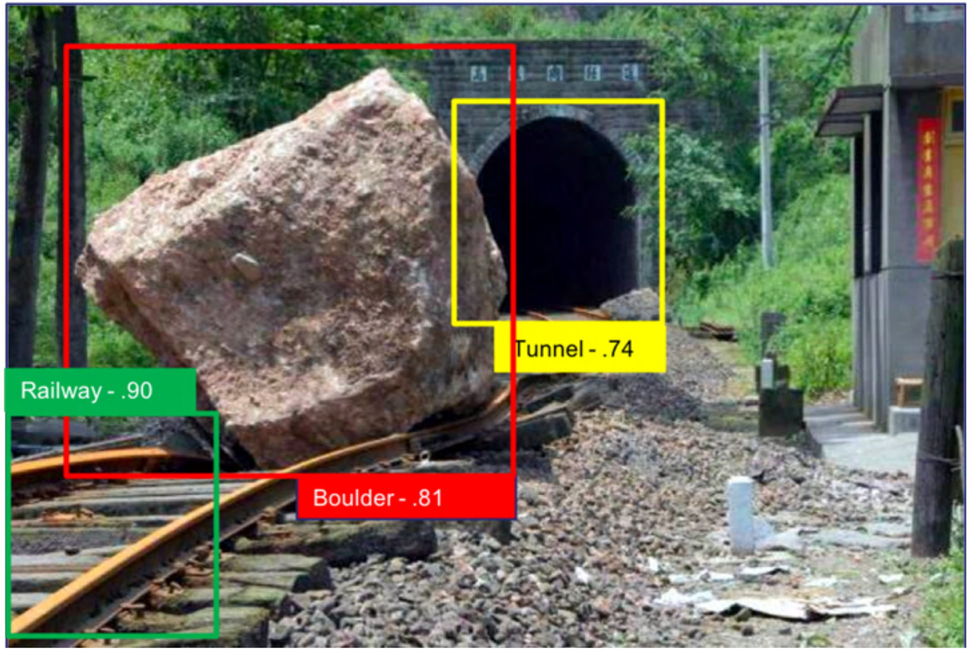



Lumbermill - .59
Boulder - .09
Railway - .11



- **Hardware**: **High performance, scalable, generic** (to different FGPA family) **& portable** CNN dedicated **programmable** processor implemented on an FPGA for **real-time embedded inference**
- **Software**: Knowledge graph extension of object detection

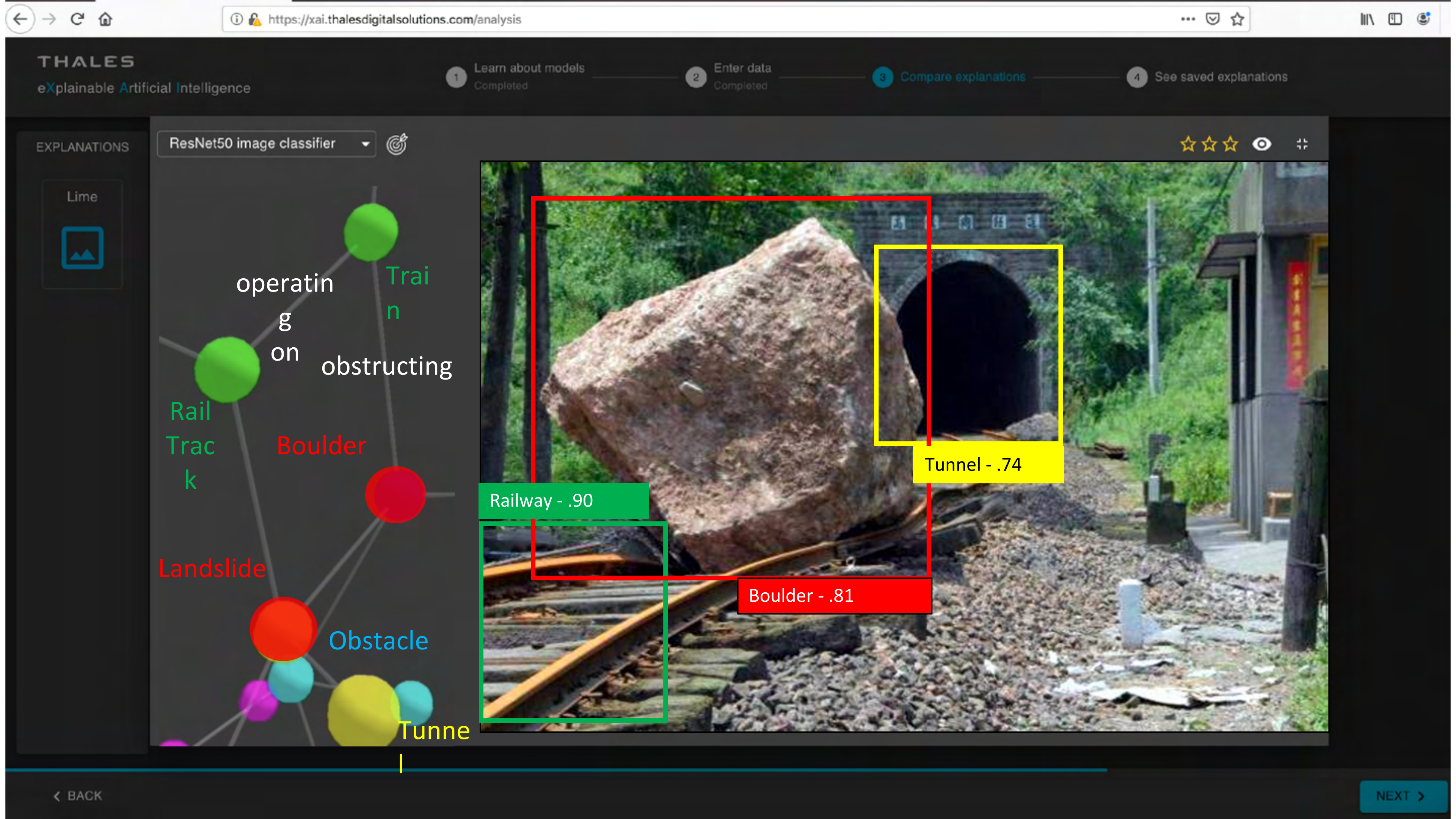
Transitionin g



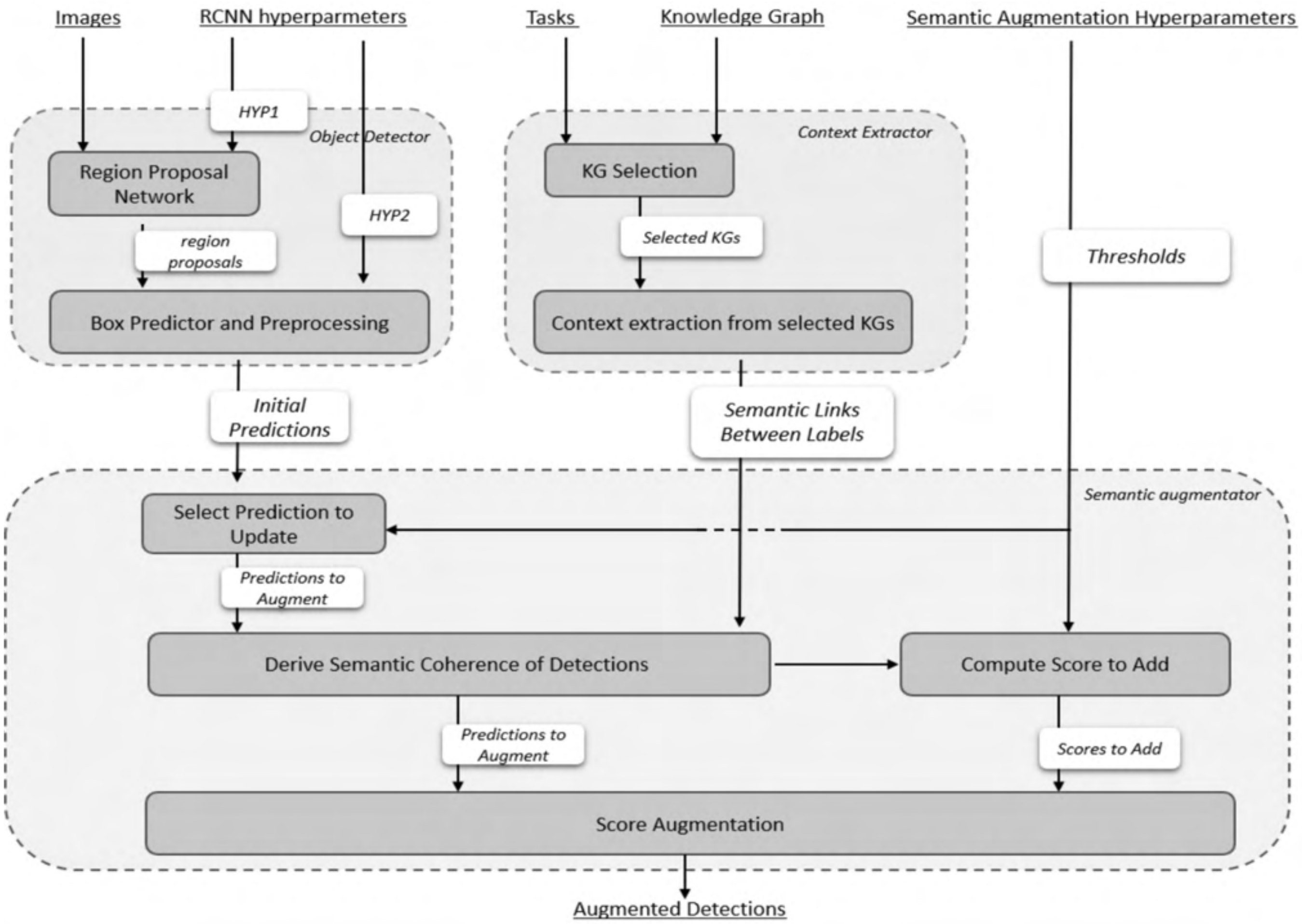
This is an **Obstacle: Boulder** obstructing the train: XG142-R on **Rail_Track** from City: Cannes to City: Marseille at **Location: Tunnel VIX** due to **Landslide**

XAI Thales Platform

- Higher accuracy with no intensive fine-tuning
- Human interpretable explanation
- Running on the edge at inference time


https://xai.thalesdigitalsolutions.com/analysis
THALES
eXplainable Artificial Intelligence
1 Learn about models
Completed
2 Enter data
Completed
3 Compare explanations
4 See saved explanations
EXPLANATIONS
Lime
ResNet50 image classifier
operating
on
obstructing
Train
Rail Track
Boulder
Landslide
Obstacle
Tunnel
Railway - .90
Boulder - .81
Tunnel - .74
BACK
NEXT

Knowledge Graph in Machine Learning - An Implementation



Freddy Lécué, Jiaoyan Chen, Jeff Z. Pan, Huajun Chen: Augmenting Transfer Learning with Semantic Reasoning. IJCAI 2019: 1779-1785

Freddy Lécué, Tanguy Pommellet: Feeding Machine Learning with Knowledge Graphs for Explainable Object Detection. ISWC Satellites 2019: 277-280

Freddy Lécué, Baptiste Abeloos, Jonathan Anctil, Manuel Bergeron, Damien Dalla-Rosa, Simon Corbeil-Letourneau, Florian Martet, Tanguy Pommellet, Laura Salvan, Simon Veilleux, Maryam Ziaeefard: Thales XAI Platform: Adaptable Explanation of Machine Learning Systems - A Knowledge Graphs Perspective. ISWC Satellites 2019: 315-316

Jiaoyan Chen, Freddy Lécué, Jeff Z. Pan, Ian Horrocks, Huajun Chen: Knowledge-Based Transfer Learning Explanation. KR 2018: 349-358

On Interpretating Visual Question Answering Results with Graphs

What is Visual Question Answering (VQA)?

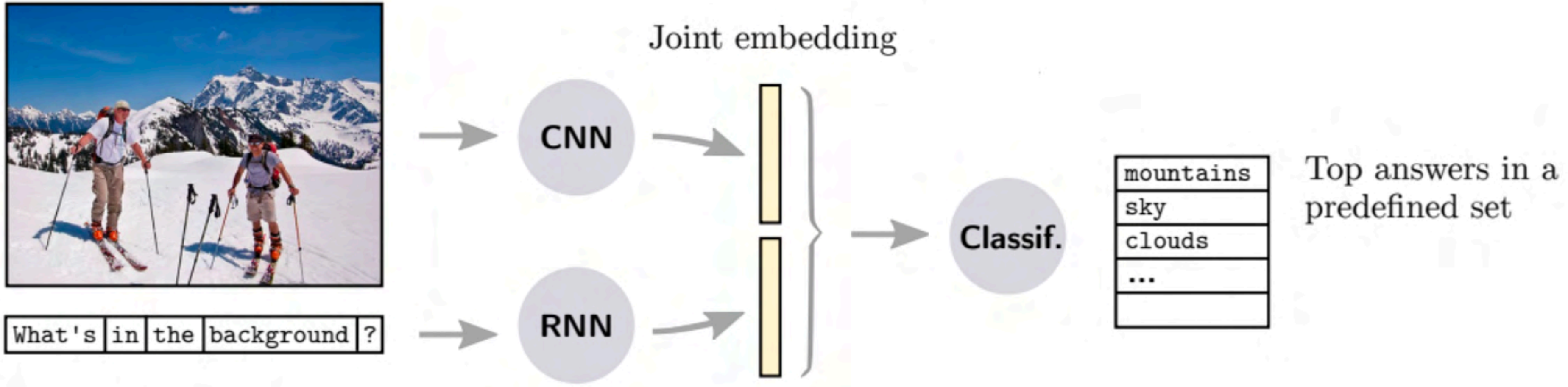
The objective of a VQA model combines **visual** and **textual** features in order to **answer questions** grounded in an **image**.

What's in the background?

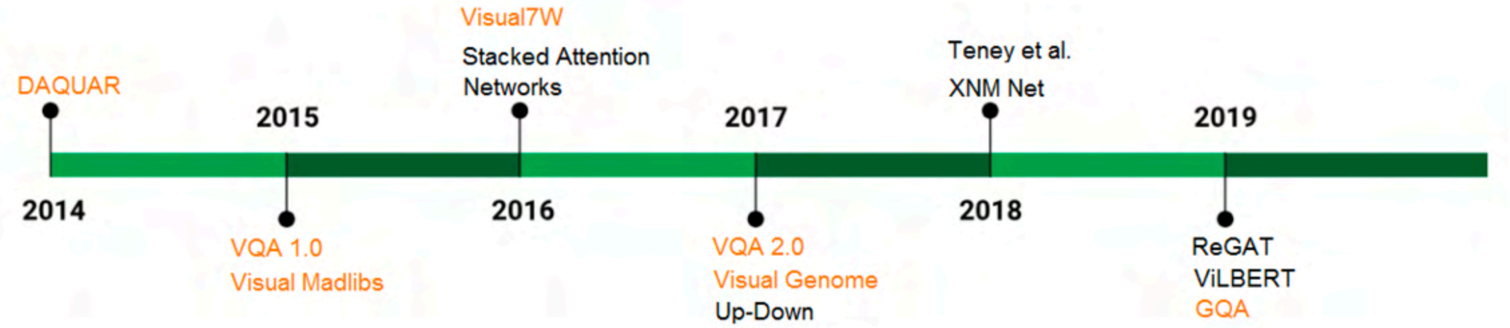
Where is the child sitting?

State of the Art in Visual Question Answering

Most approaches combine **Convolutional Neural Networks** (CNN) with **Recurrent Neural Networks** (RNN) to learn a mapping directly from input images (vision) and questions to answers (language)



Major breakthrough in VQA (models and real-image dataset)



Accuracy Results:

DAQUAR [2] (13.75 %), VQA 1.0 [1] (54.06 %), Visual Madlibs [3] (47.9 %), Visual7W [4] (55.6 %), Stacked Attention Networks [5] (VQA 2.0: 58.9 %, DAQAUR: 46.2 %), VQA 2.0 [6] (62.1 %), Visual Genome [7] (41.1 %), Up-down [8] (VQA 2.0: 63.2 %), Teney et al. (VQA 2.0: 63.15 %), XNM Net [9] (VQA 2.0: 64.7 %), ReGAT [10] (VQA 2.0: 67.18 %), ViLBERT [11] (VQA 2.0: 70.55 %), GQA [12] (54.06 %)

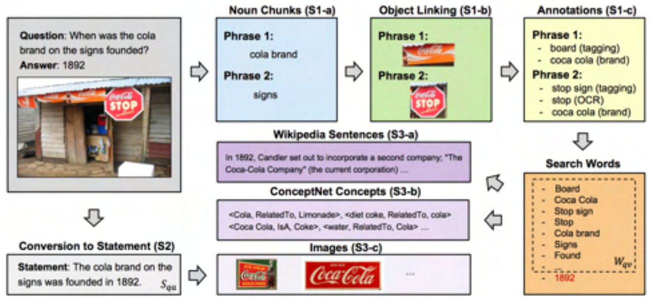
But they have limitations:

- Answers are required to be in the image
- Knowledge is limited

Therefore some questions cannot be correctly answered as some level of (basic) reasoning is required.

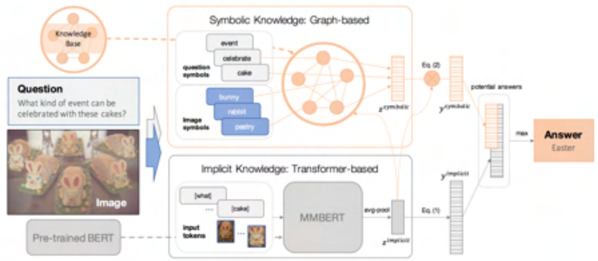
State of the Art in Visual Question Answering + Graph

Most approaches aims at extending VQA Neural Network architectures with **<u>knowledge graphs</u>** in different ways



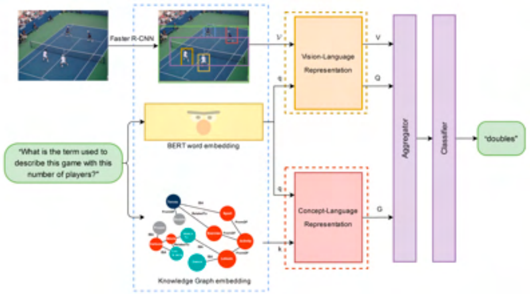
Search-based (MAVEx)

https://arxiv.org/pdf/2103.12248.pdf



Graph-Embedding-based (KRISP)

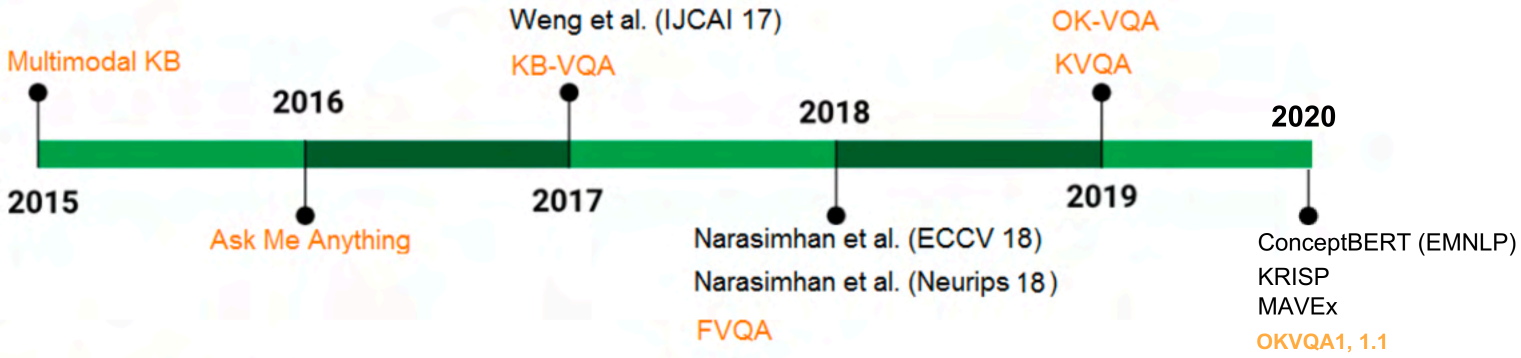
https://arxiv.org/pdf/2012.11014.pdf



Graph-Fusion-based (ConceptBERT)

https://aclanthology.org/2020.findings-emnlp.44/

Major breakthrough in OKVQA (models and real-image dataset)



Accuracy Results:

Multimodal KB [17] (NA), Ask me Anything [18] (59.44 %), Weng et al (VQA 2.0: 59.50 %), KB-VQA [19] (71 %), FVQA [20] (56.91 %), Narasimhan et al. (ECCV 2018) (FVQA: 62.2 %) , Narasimhan et al. (Neurips 2018) (FVQA: 69.35 %), OK-VQA [21] (27.84 %), KVQA [22] (59.2 %)

But they **ALSO** have limitations:

- No explanation

Therefore no insight on how the solutions have any semantic relations to the questions and image

eXplainable Visual Question Answering using Knowledge Graphs (1)

Core Question:

- How to **retrieve explanations** of a VQA model during inference?

- How to expose articulated knowledge (i.e., **composition of knowledge graph triples**) to explain how an answer is related to the question, objects of the images and concepts?

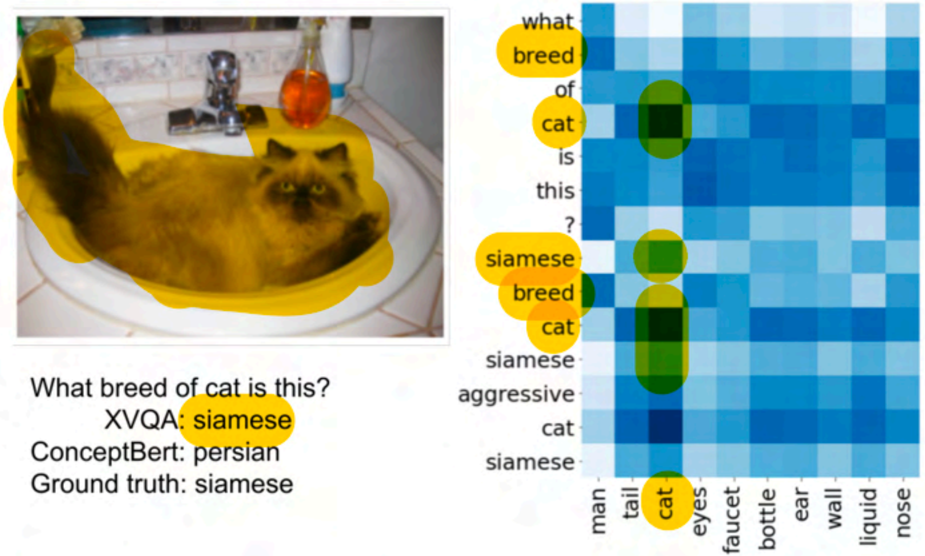
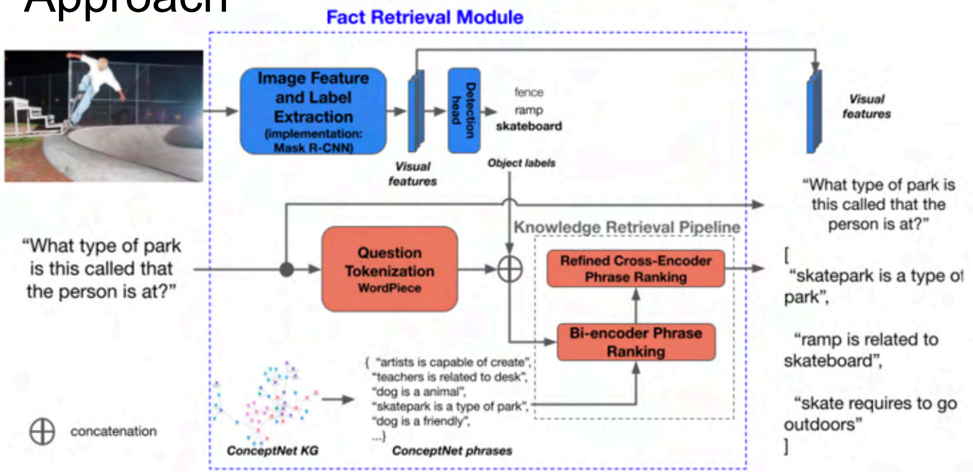


Figure 1: An example of VQA task with question: *What breed of cat is it*? on the left image, and our XVQA Answer: *Siamese*. XVQA also exhibits explanations from the optimal transfer map between (i) question tokens (vertical tokens on the right image: cat, breed), graph entities (vertical tokens after question on the right image: siamese, cat, breed) and (ii) detected object embeddings (horizontal tokens on the right image: cat) i.e., *siamese is a cat breed*.

eXplainable Visual Question Answering using Knowledge Graphs (2)

Approach



Fact Retrieval Module

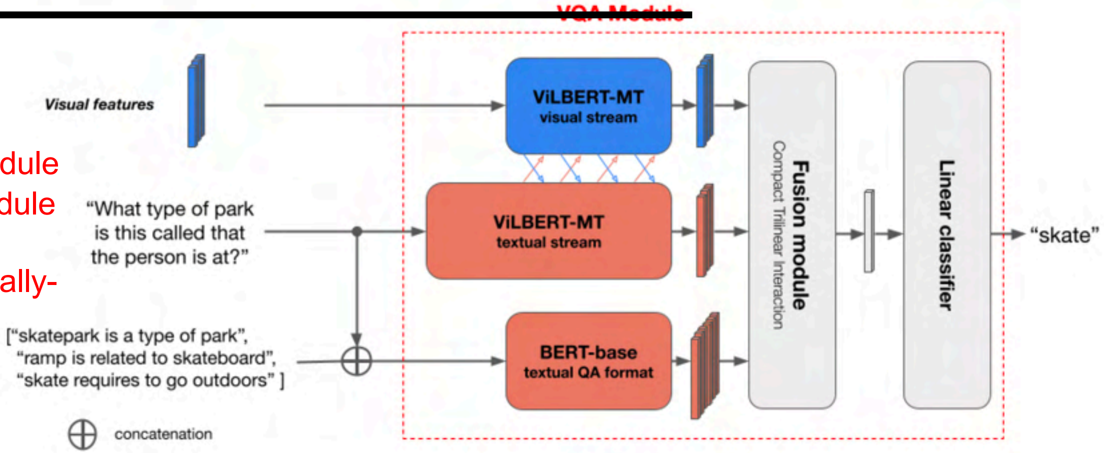
We perform text retrieval on facts from ConceptNet to collect relevant OK related to each question-image pair

1) Bi-Encoder Phrase Ranking to compute query agnostic fact phrase embeddings
2) Refined Cross-Encoder Phrase Ranking for each model

VQA Module

A parallel stream architecture with a vision language module along with a BERT-base textual question answering module

1) Capturing image and text data into dense semantically-rich representations,
2) Aligning these representations from different modalities,
3) Enriching them with outside knowledge



eXplainable Visual Question Answering using Knowledge Graphs (3)

Quantitative Results

Model / Data type	OK-VQAv1	OK-VQAv1.1
XVQA	33.2%	39.7%
XVQA (without facts)	32.6%	38.9%
XVQA (oracle case)	46.3%	54.7%
ConceptBERT	33,0%	–
ViLBERT	35.2%	41.6%
KRISP	38.35%	38.9%
MAVEx	–	40.5%
MAVEx (oracle case)	–	43.5%

eXplainable Visual Question Answering using Knowledge Graphs (4)

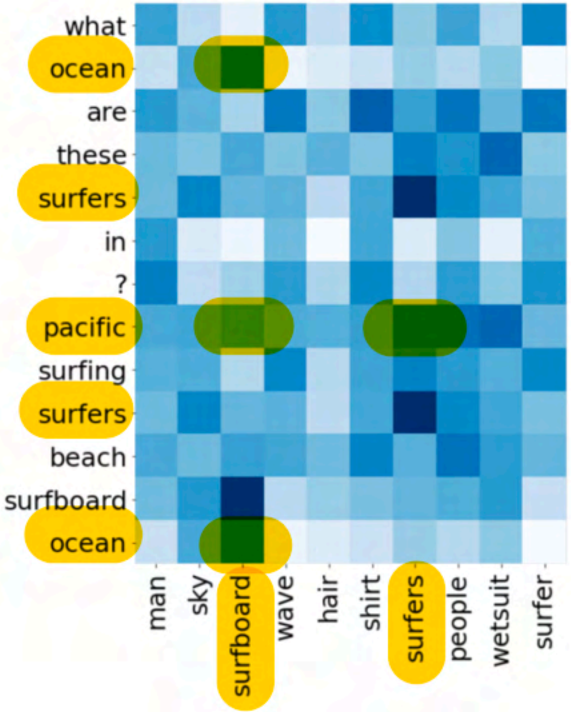
Qualitative Results

(1) Question: What ocean are these surfers in?

XVQA: pacific

ConceptBert: surf

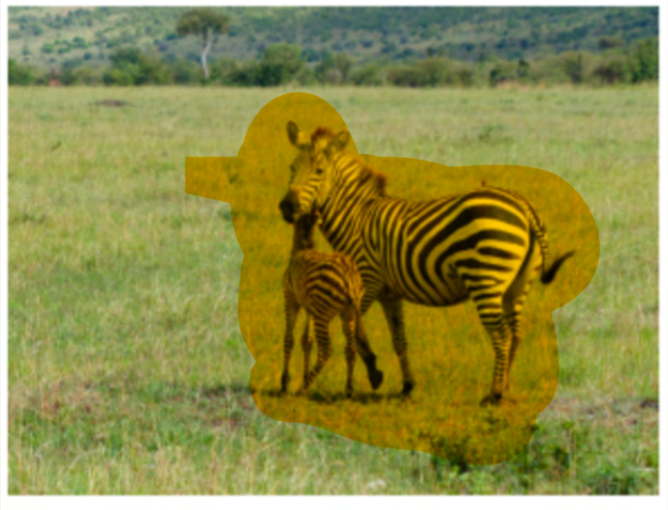
Ground truth: pacific

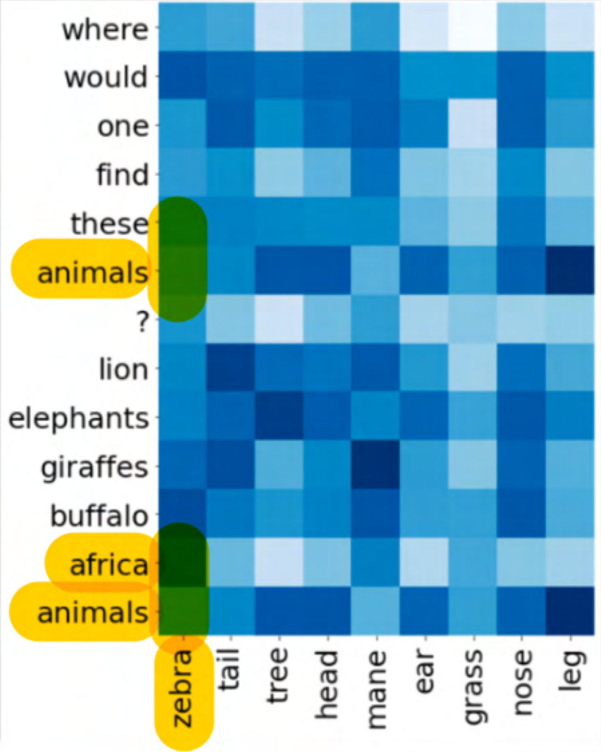


(1) XVQA exhibits explanations from the optimal transfer map between (i) question tokens (vertical tokens: ocean, surfers), graph entities (vertical tokens: surfers, ocean, pacific) and (ii) detected object (horizontal tokens: surfers, surfboard) embeddings i.e., *surfing isAnActivityIn pacific*, *surfboard isRelatedTo ocean*.

eXplainable Visual Question Answering using Knowledge Graphs (5)

Qualitative Results





(2) Question: Where would one find these animals?

XVQA: africa
ConceptBert: africa
Ground truth: africa

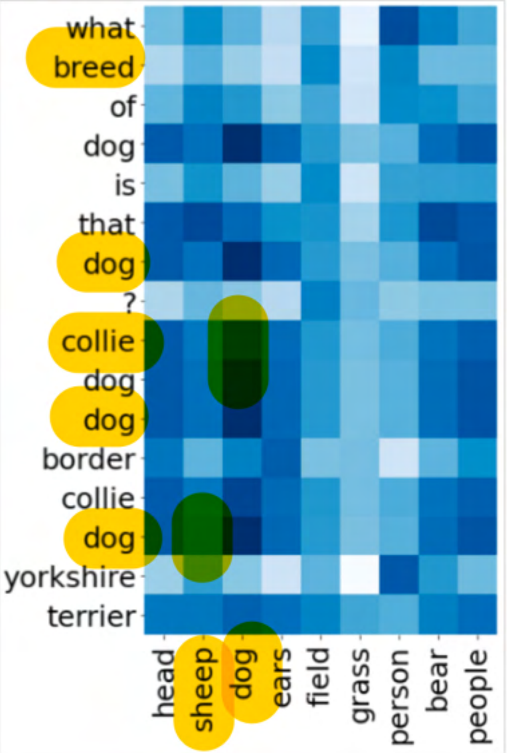
(2) Here the optimal transfer map is between (i) question tokens (vertical tokens: animals), graph entities (vertical tokens: africa, animals) and (ii) detected object (horizontal tokens: zebra) embeddings i.e., *africa has animals*.

eXplainable Visual Question Answering using Knowledge Graphs (6)

Qualitative Results

(3) Question: What breed of dog is that dog?

XVQA: collie
ConceptBert: shepherd
Ground truth: collie



(3) Here the optimal transfer map is between (i) question tokens (vertical tokens: dog, breed), graph entities (vertical tokens: collie, dog) and (ii) detected object (horizontal tokens: sheep, dog) embeddings i.e., *collie isA dog*.

eXplainable Visual Question Answering using Knowledge Graphs (7)

Lessons Learnt

- **Retrieving explanations** of a VQA model during inference is a complex task
- Exposing articulated knowledge (i.e., **composition of knowledge graph triples**) to explain how an answer is related to the question, objects of the images and concepts is highly depending **on relevant retrieved knowledge**
- **High potential for improvement**

Model / Data type	OK-VQAv1	OK-VQAv1.1
XVQA	33.2%	39.7%
XVQA (without facts)	32.6%	38.9%
XVQA (oracle case)	46.3%	54.7%
ConceptBERT	33,0%	–
ViLBERT	35.2%	41.6%
KRISP	38.35%	38.9%
MAVEx	–	40.5%
MAVEx (oracle case)	–	43.5%

Towards Causal Explanations

Context

- Explainability and interpretability play an important role for adopting deep neural networks in critical systems
- State-of-the-art systems focus on correlation rather than causal mechanisms

Goal

- Computing causal graph to extract causal explanations

Approach

- Hybrid AI:
  - Deep Neural Networks
  - Abduction mechanisms

Thales DIS

Results

Causal Analysis of LeNet on MNIST

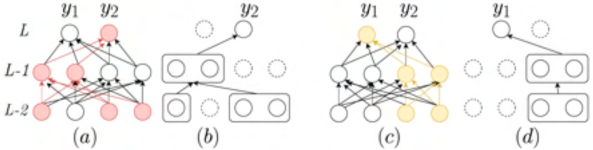


Figure 2: Causal connections within the last three layers of a neural network. (a) and (c) Coloured paths (red/yellow) transpose signals between layers to labels y_1/y_2, respectively. (b) and (d) Two abstract graphs are obtained by causal inference. In each graph, neutral neurons (marked in dots) hold variant information which don't influence models' behaviour for the corresponding label.

Table 1: Quantitative comparison between explanation methods using 1000 samples of ImageNet data including 10 classes. The results are the median values.

Method	Lipschitz Estimate ↓		IROF ↑	
	ResNet18	ConvNext	ResNet18	ConvNext
Occlusions [36]	-	-	14.45	**30.50**
RISE [23]	146.31	115.33	16.36	29.44
DeconvNet [37]	63.19	6.67	18.60	29.38
Saliency [31]	21.66	4.77	12.52	29.77
IG [33]	25.20	21.06	14.53	28.67
InputXGrad [29]	30.74	6.07	14.36	28.27
GradShape [2]	38.4	43.58	14.41	29.39
GuidedBackProb [32]	6.29	6.65	14.76	29.38
MWP [38]	1.08	2.68	14.20	27.66
Causal (ours)	**0.36**	**0.02**	**21.7**	**30.22**

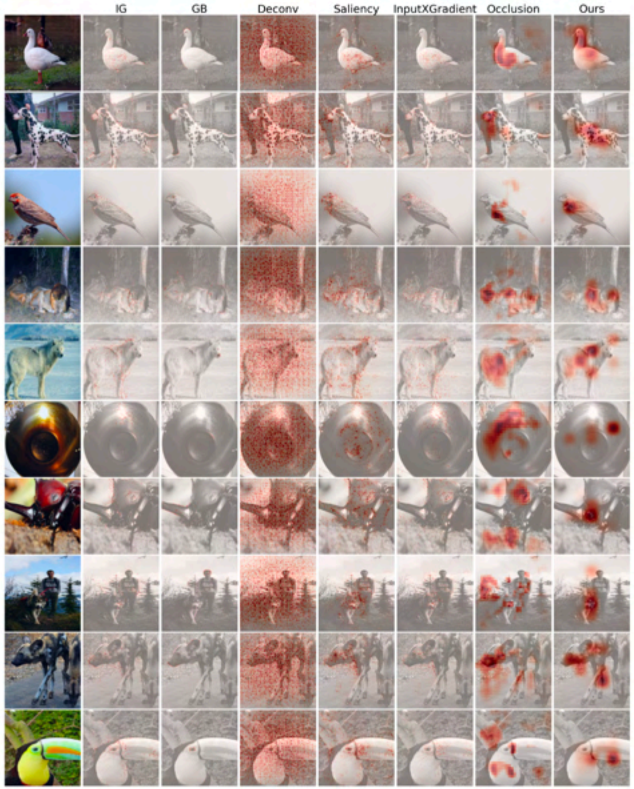


Figure 15: Explanations Comparison between several methods on ResNet18 and ImageNet

Conclusion

Concept Annotation | Semantics

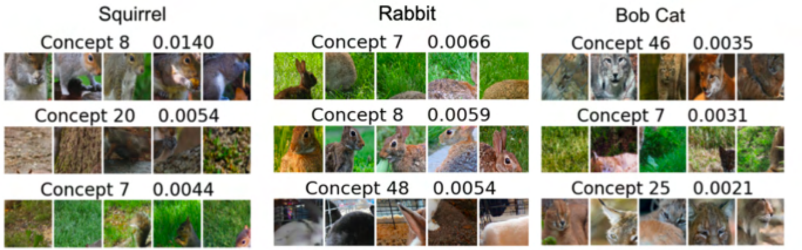
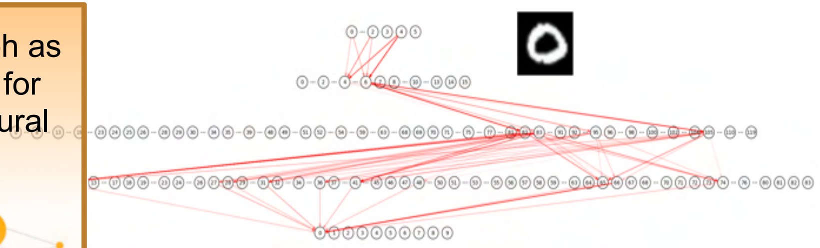
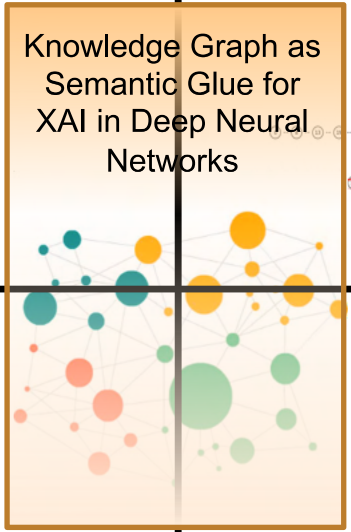


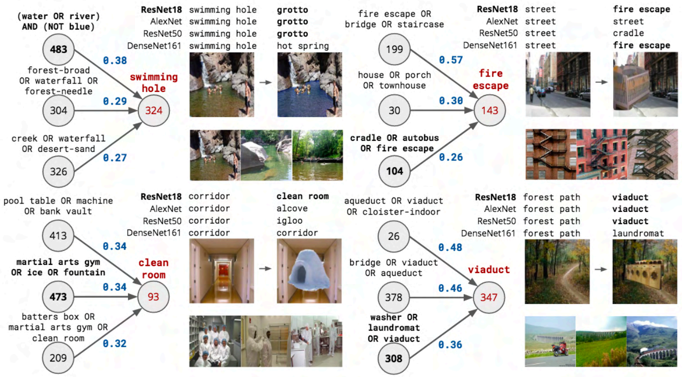
Figure 3: Concept examples with the samples that are the nearest to concept vectors in the activation space in AwA. The per-class ConceptSHAP score is listed above the images.

Causal Relations

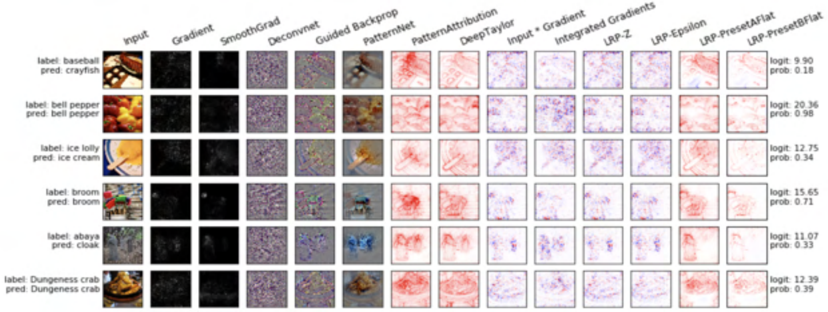




Network Dissection Neurons Composition



Saliency Maps Super-Pixels



Thanks! Questions?

- Feedback most welcome :-)
  - **freddy.lecue@inria.fr (@freddylecue)**
  - **freddy.lecue@thalesgroup.com**
- Slides: https://tinyurl.com/2p83nt59