Explaining Deep Neural Networks

The Good, the Bad and the Ugly

... and Where Every Little Knowledge Helps

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

Canadian Consortium on Responsible AI

THALES

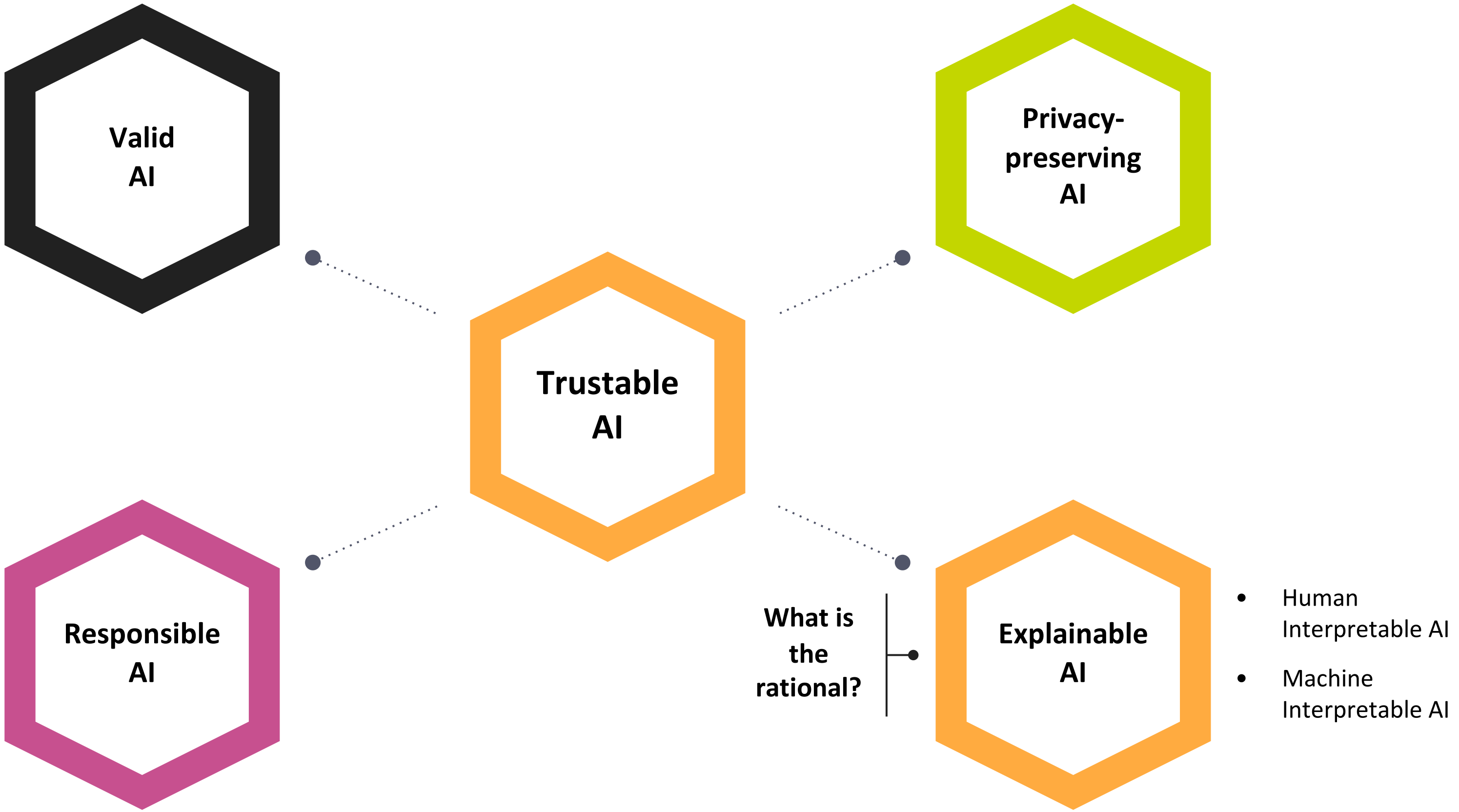
Inria
INVENTEURS DU MONDE NUMÉRIQUE

February 24, 2022

https://tinyurl.com/2p9cpdvj

Scope

AI Adoption: Requirements



Part I

Introduction and Motivation

Explanation - From a Business Perspective

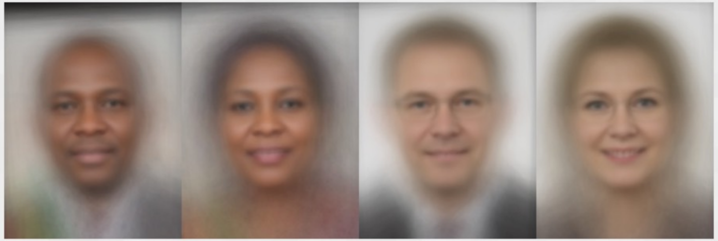
Business to Customer AI

Critical Systems (1)

Critical Systems (2)

… and even More

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Joy Buolamwini, Timnit Gebru: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT 2018: 77-91

Colin, but at home. @colinmadland · Sep 18, 2020
any guesses?
60 1.1K 7.2K

Colin, but at home. @colinmadland · Sep 18, 2020
29 659 6K

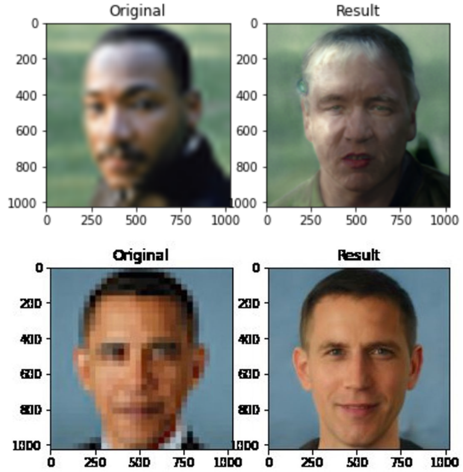
Colin, but at home. @colinmadland · Sep 18, 2020
Turns out @zoom_us has a crappy face-detection algorithm that erases black faces...and determines that a nice pale globe in the background must be a better face than what should be obvious.
96 4.3K 21.1K

Colin, but at home. @colinmadland · Sep 18, 2020
Images were taken and shared with permission.
4 174 6K

Colin, but at home. @colinmadland · Sep 18, 2020
Geez...any guesses why @Twitter defaulted to show only the right side of the picture on mobile?
ALT
83 2K 19.2K

https://techcrunch.com/2020/10/02/twitter-may-let-users-choose-how-to-crop-image-previews-after-bias-scrutiny/

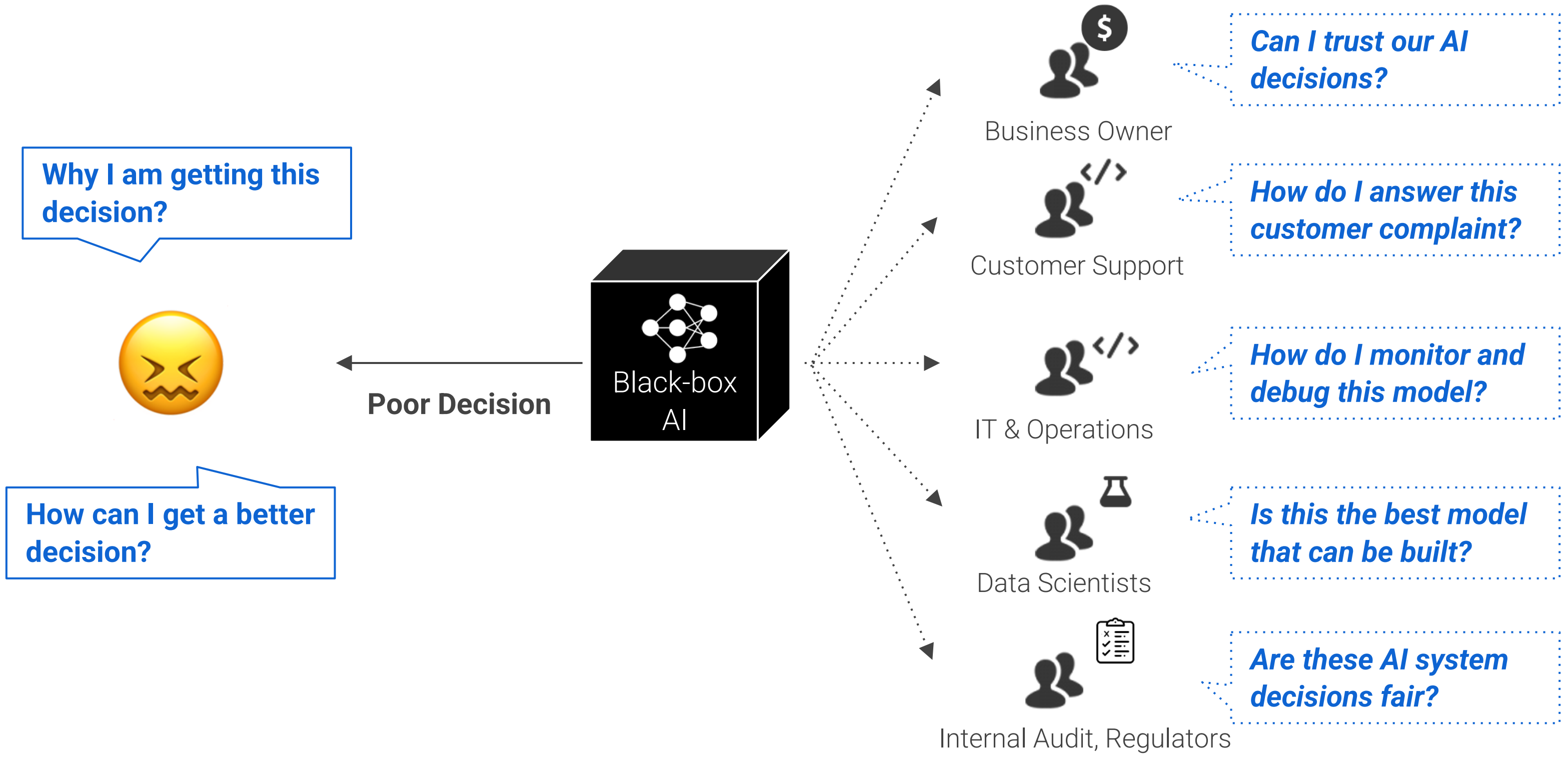
https://www.cbsnews.com/news/apple-credit-card-goldman-sachs-disputes-claims-that-apple-card-is-sexist/



https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

Explanation - In a Nutshell

AI as a Black-box: Source of Confusion and Doubt



Credit: Lecue et al., Tutorial on XAI. AAAI 2020. https://xaitutorial2020.github.io/

Evaluation - XAI: One Objective, Many Metrics

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?

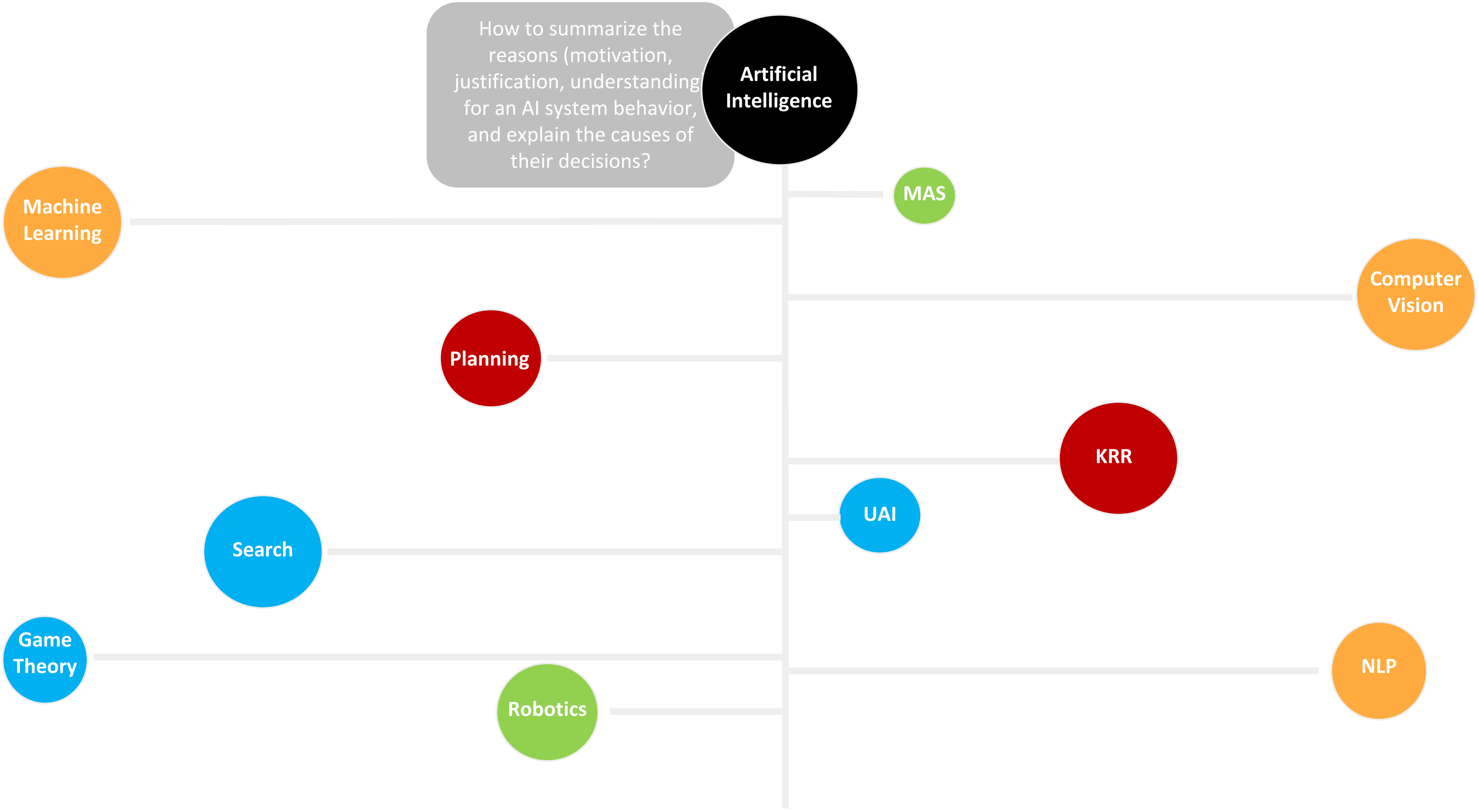
Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan

Part II

Explanation in AI (Focus Deep Neural Networks)

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

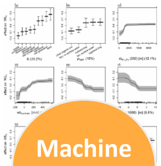


XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

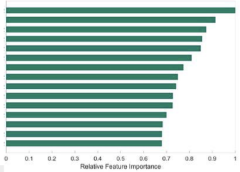
Artificial Intelligence

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

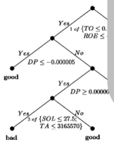
Dependency Plot



Feature Importance



Surrogate Model



Machine Learning

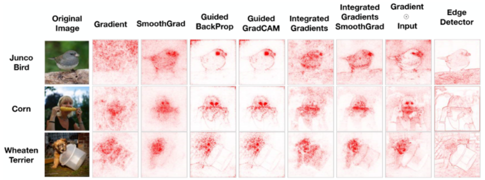
Which features are responsible of classification?

Strategy Summarization

MAS

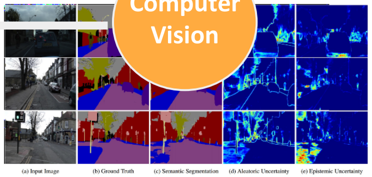
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



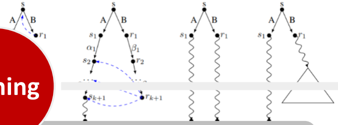
Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

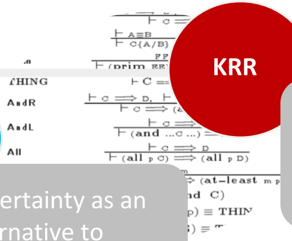
Plan Refinement



Planning

Which actions are responsible of a plan?

Diagnosis



KRR

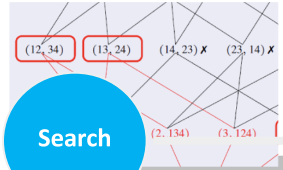
Abduction

- Which axiom is responsible of inference (e.g., classification)?
- Abduction/Diagnostic: Find the **right** root causes (abduction)?

UAI

Uncertainty as an alternative to explanation

Conflicts Resolution



Search

Which constraints can be relaxed?

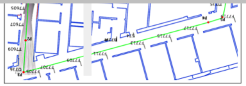
Game Theory

Which combination of features is optimal?

Shapely Values

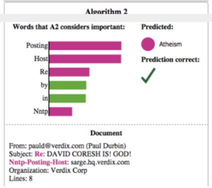
Robotics

Which decisions, combination of multimodal decisions lead to an action?



Narrative-based

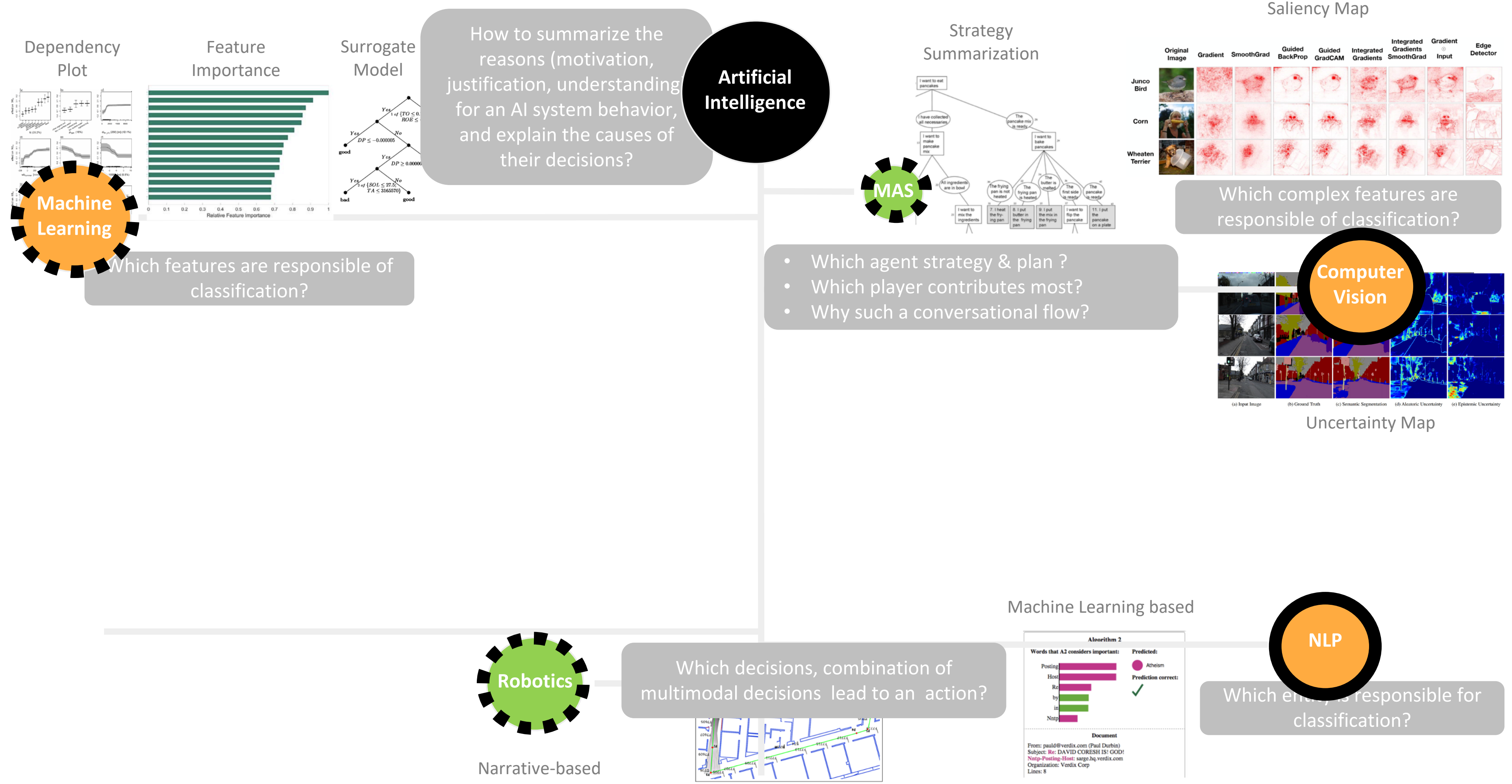
Machine Learning based



NLP

Which entity is responsible for classification?

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



Part III

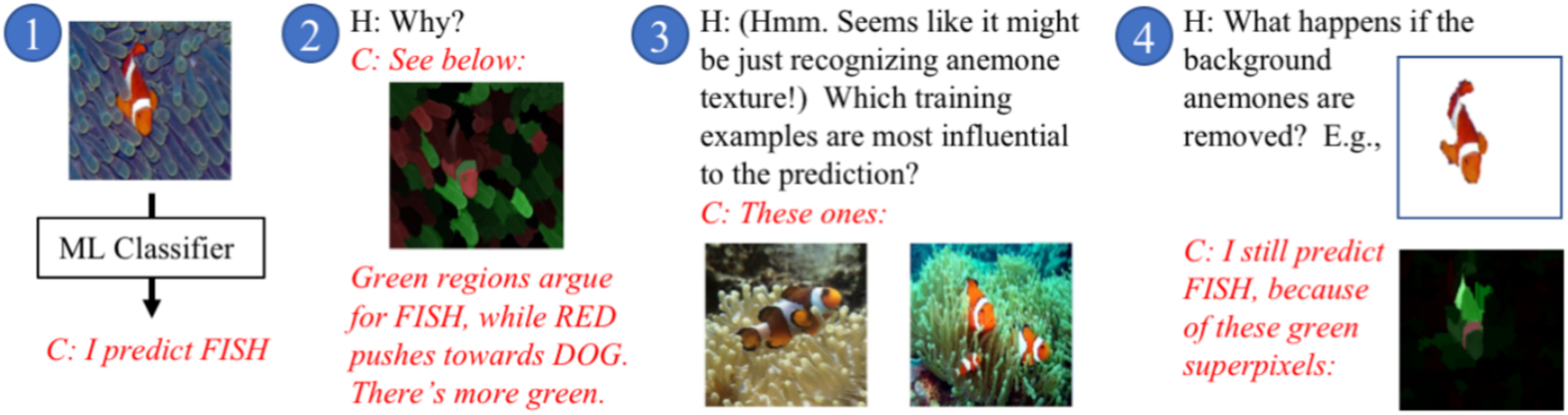
XAI:

The Good,

The Bad, and

The Ugly

The Good: Multimodal End-to-End XAI System



- Systems do handle **humans follow-up questions**
- Human – Machine interactions ARE at **FOUNDATIONAL**
- **Examples / prototypes DO help**
- Explanations **DO NOT answer all users' concerns in one shot**
  - Many different stakeholders
  - Many different objectives
  - Many different experiise

The Good

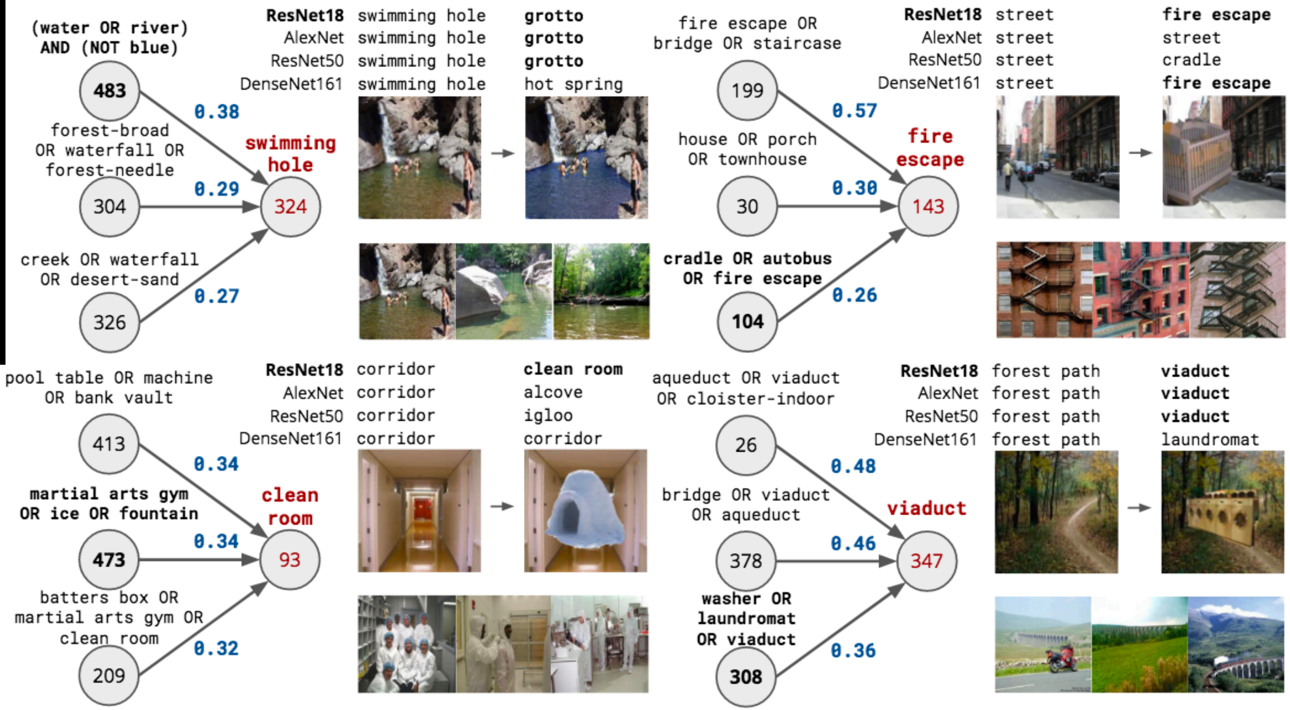
- [Interaction] Human are in the loop (What-if / counterfactual)
- [Construction] Iterative explanation search
- [Validation] Operator as opposed to developer driven
- [Knowledge] Domain knowledge is required

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

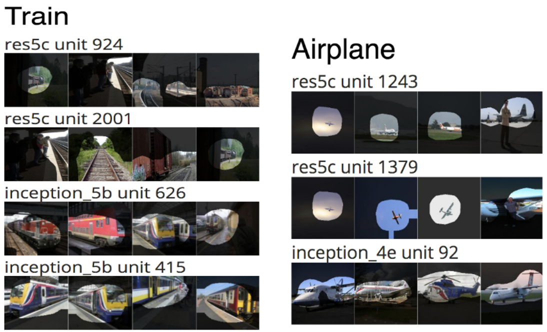
The (not so) **Bad**: Network Dissection | Neurons Composition

The (not so) **Bad**

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



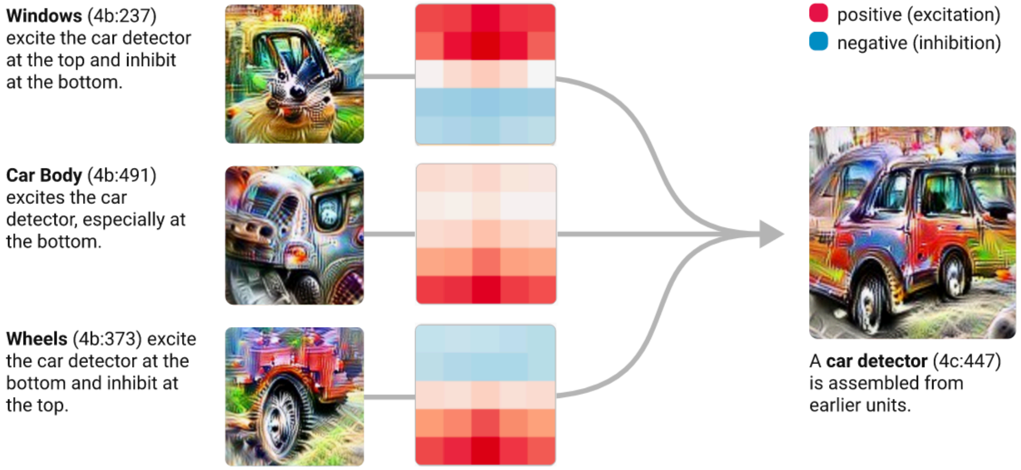
Jesse Mu, Jacob Andreas: Compositional Explanations of Neurons. NeurIPS 2020



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

The Bad: Feature Visualization

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



The Bad

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

CLIP
Resent 50
Layer 4

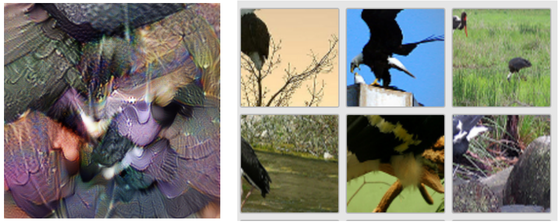
https://microscope.openai.com/models

Unit 118

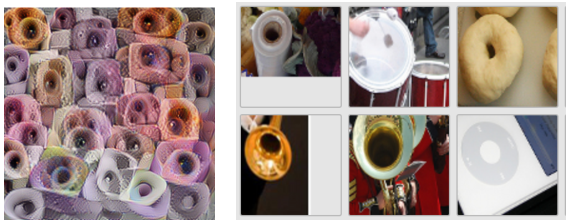
Unit 55

Resnet 50 v2
Block4/unit_3/add

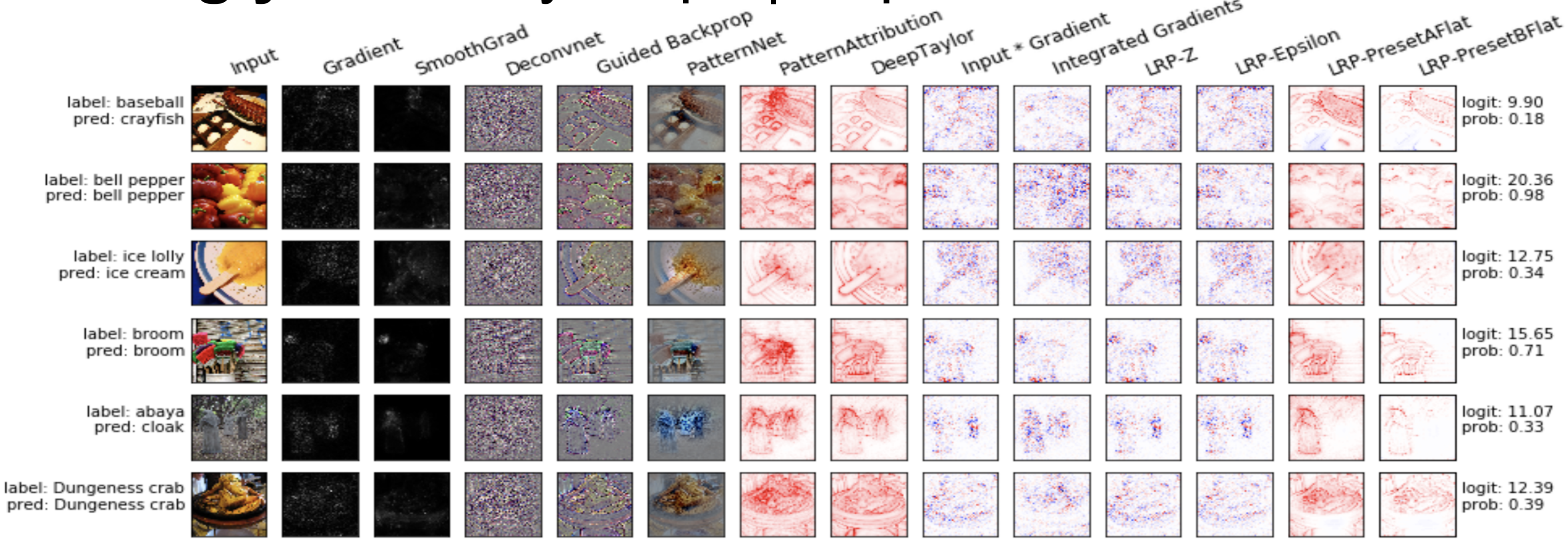
Unit 546

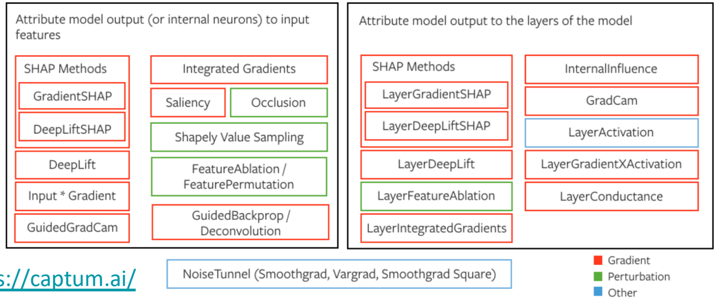


Unit 562



The Ugly: Saliency Maps | Super-Pixels





https://captum.ai/

The Ugly

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

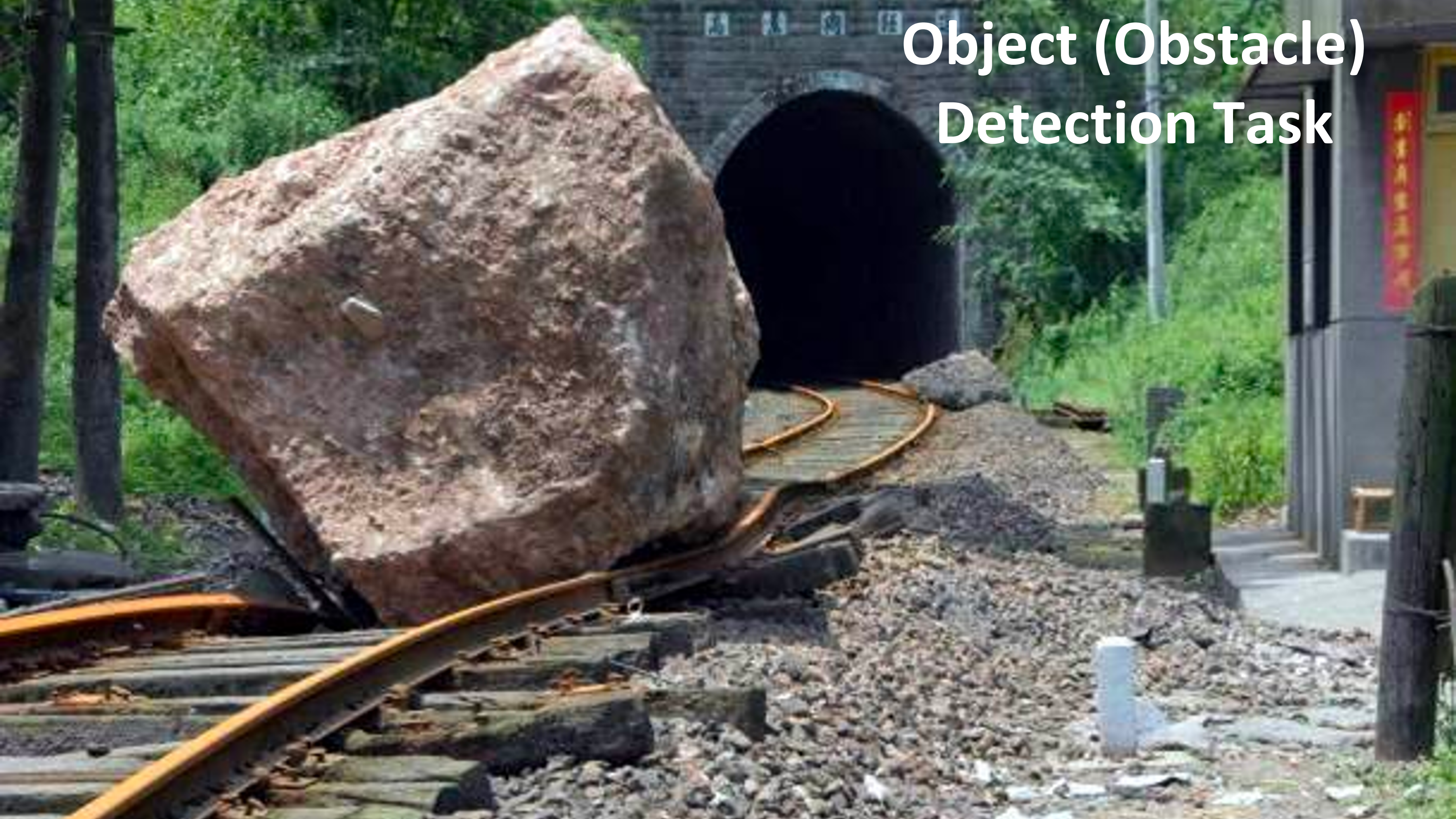
Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps. NeurIPS 2018: 9525-9536

Part IV

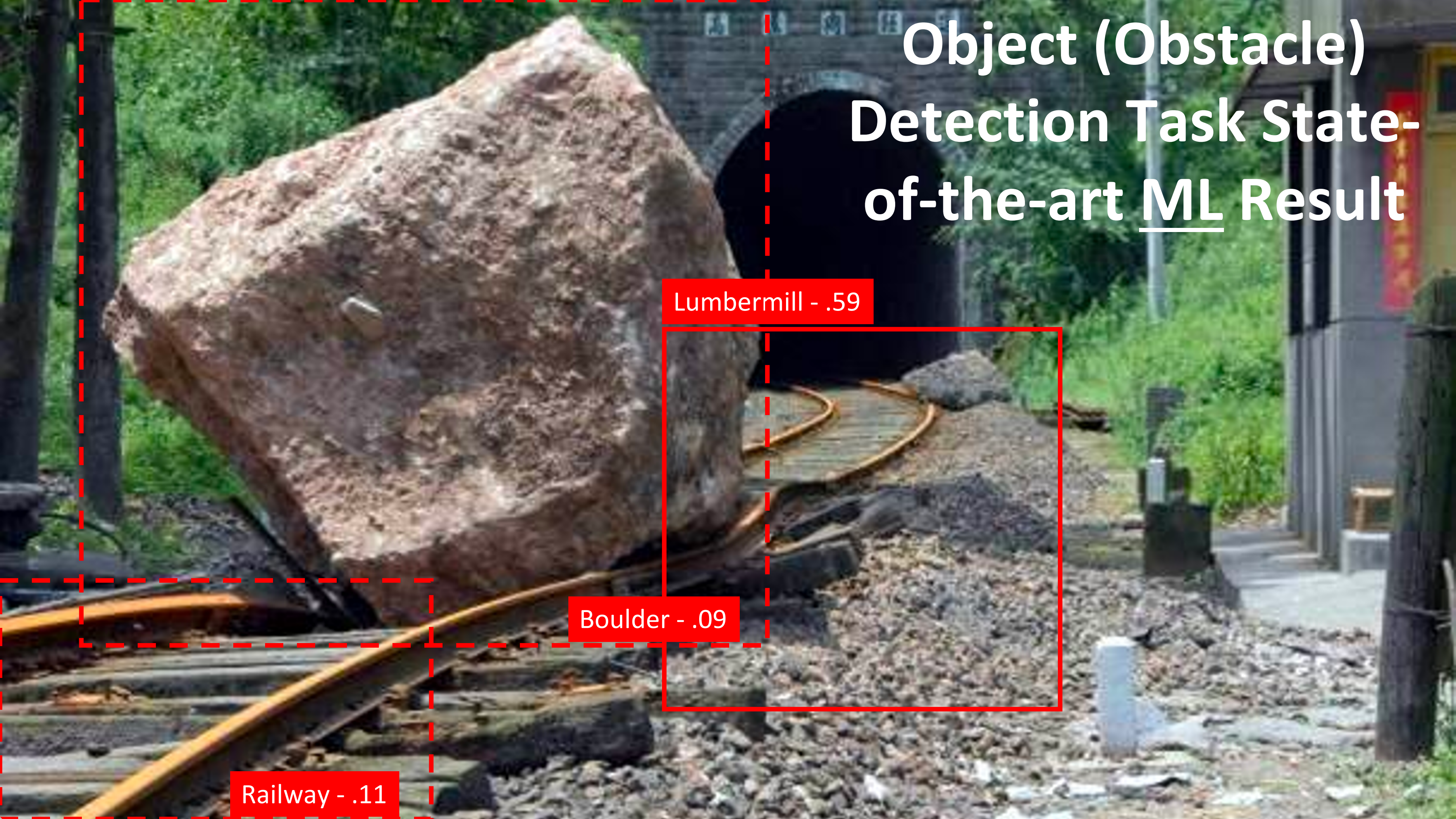
On Boosting Neural Networks Interpretation with Graphs

How Does
it
Work
in Practice?

State of the Art Machine Learning Applied to Critical Systems


Object (Obstacle)
Detection Task

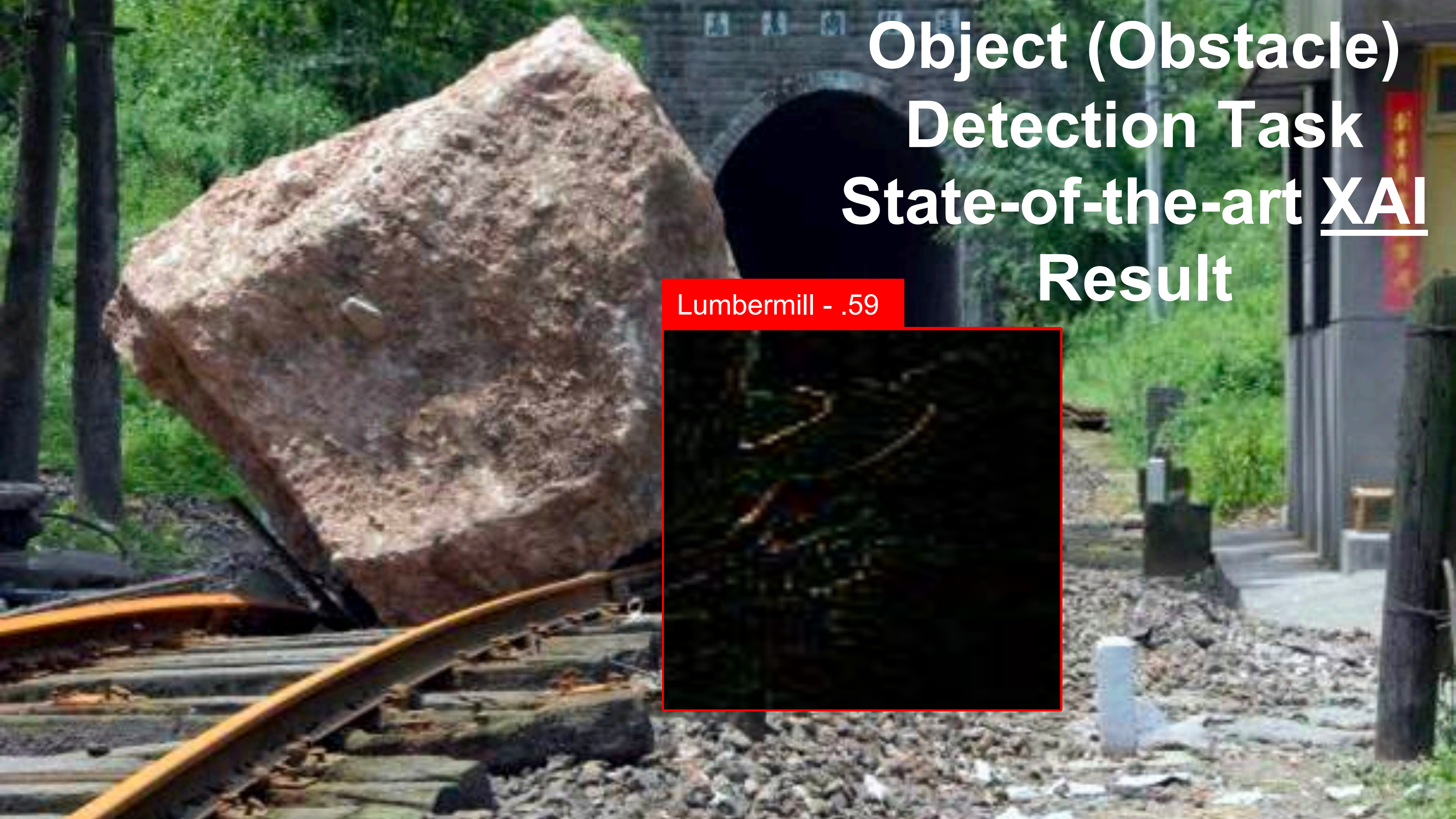
Object (Obstacle) Detection Task State-of-the-art ML Result


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 <u>XAI</u> Result



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

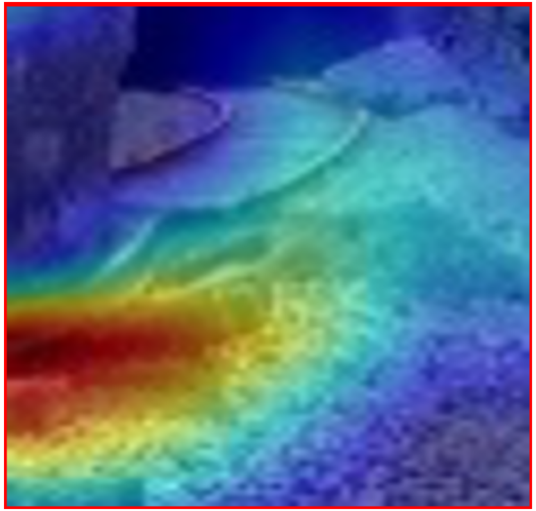
Unfortunately, this is of NO use for a human behind the system

Let’s stay back

Why this Explanation?
(meta explanation)

After Human Reasoning…

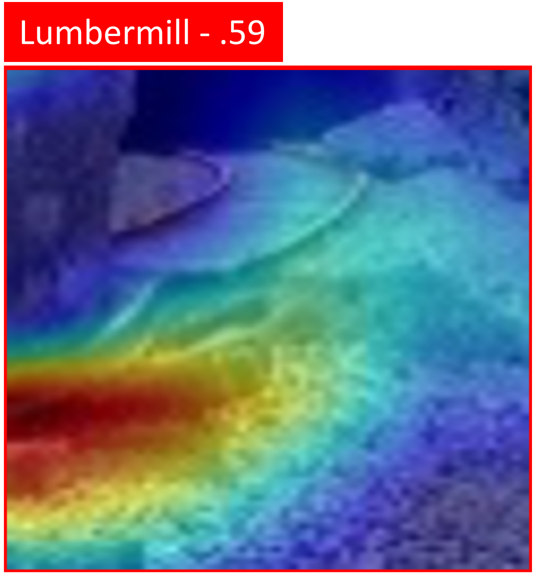
Lumbermill - .59

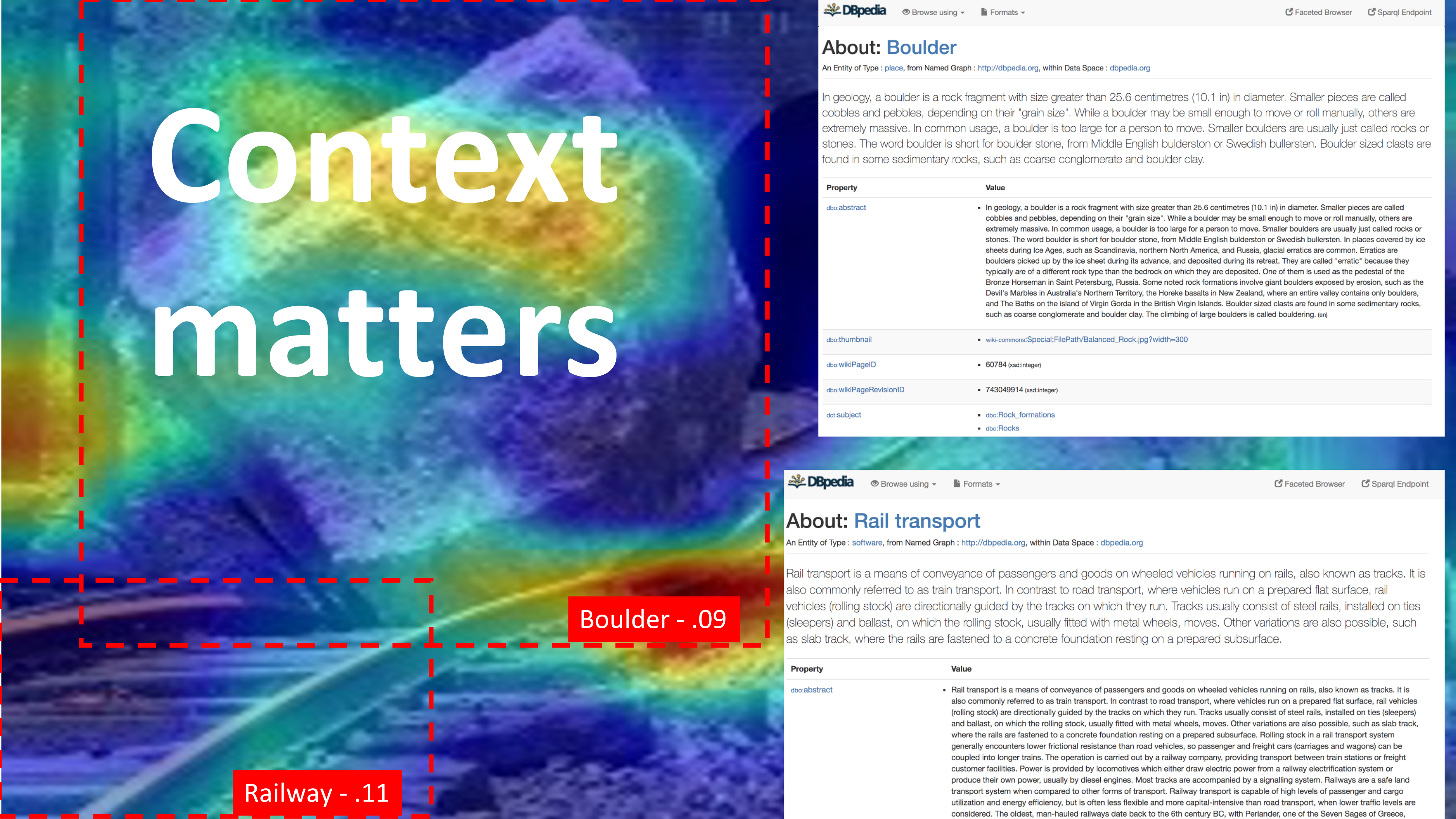


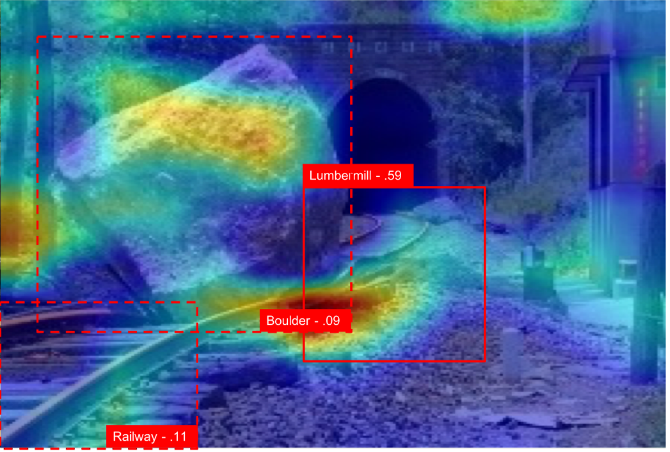
DBpedia | Browse using | Formats | Faceted Browser | Sparql Endpoint

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?

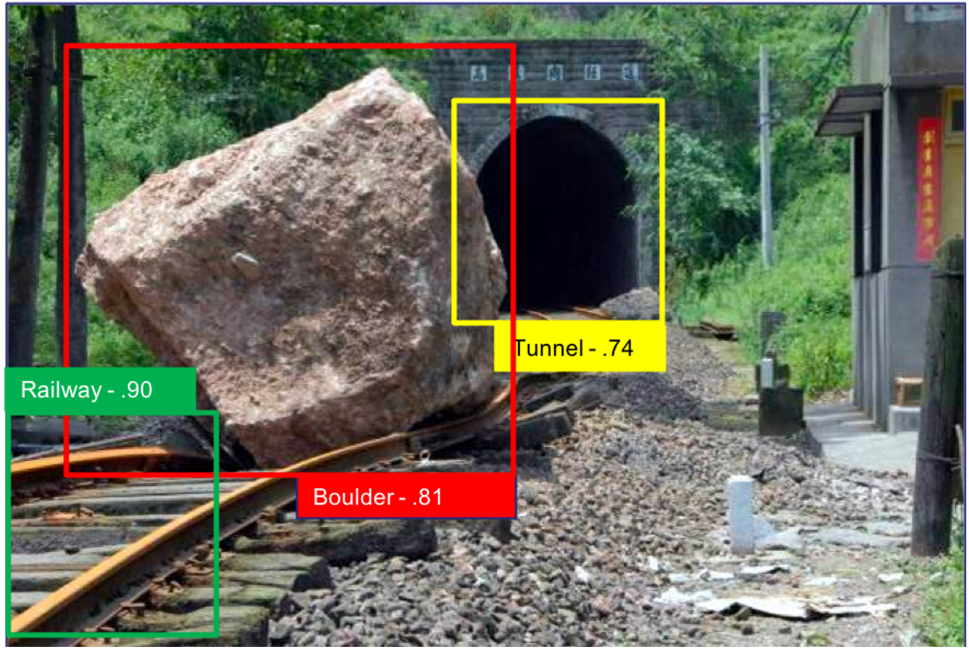



Context matters
Boulder - .09
Railway - .11
DBpedia
Browse using
Formats
Faceted Browser
Sparql Endpoint
About: Boulder
An Entity of Type : place, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org
In geology, a boulder is a rock fragment with size greater than 25.6 centimetres (10.1 in) in diameter. Smaller pieces are called cobbles and pebbles, depending on their "grain size". While a boulder may be small enough to move or roll manually, others are extremely massive. In common usage, a boulder is too large for a person to move. Smaller boulders are usually just called rocks or stones. The word boulder is short for boulder stone, from Middle English bulderston or Swedish bullersten. Boulder sized clasts are found in some sedimentary rocks, such as coarse conglomerate and boulder clay.
Property
Value
dbo:abstract
In geology, a boulder is a rock fragment with size greater than 25.6 centimetres (10.1 in) in diameter. Smaller pieces are called cobbles and pebbles, depending on their "grain size". While a boulder may be small enough to move or roll manually, others are extremely massive. In common usage, a boulder is too large for a person to move. Smaller boulders are usually just called rocks or stones. The word boulder is short for boulder stone, from Middle English bulderston or Swedish bullersten. In places covered by ice sheets during Ice Ages, such as Scandinavia, northern North America, and Russia, glacial erratics are common. Erratics are boulders picked up by the ice sheet during its advance, and deposited during its retreat. They are called "erratic" because they typically are of a different rock type than the bedrock on which they are deposited. One of them is used as the pedestal of the Bronze Horseman in Saint Petersburg, Russia. Some noted rock formations involve giant boulders exposed by erosion, such as the Devil's Marbles in Australia's Northern Territory, the Horeke basalts in New Zealand, where an entire valley contains only boulders, and The Baths on the island of Virgin Gorda in the British Virgin Islands. Boulder sized clasts are found in some sedimentary rocks, such as coarse conglomerate and boulder clay. The climbing of large boulders is called bouldering. (en)
dbo:thumbnail
wiki-commons:Special:FilePath/Balanced_Rock.jpg?width=300
dbo:wikiPageID
60784 (xsd:integer)
dbo:wikiPageRevisionID
743049914 (xsd:integer)
dct:subject
dbc:Rock_formations
dbc:Rocks
DBpedia
Browse using
Formats
Faceted Browser
Sparql Endpoint
About: Rail transport
An Entity of Type : software, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org
Rail transport is a means of conveyance of passengers and goods on wheeled vehicles running on rails, also known as tracks. It is also commonly referred to as train transport. In contrast to road transport, where vehicles run on a prepared flat surface, rail vehicles (rolling stock) are directionally guided by the tracks on which they run. Tracks usually consist of steel rails, installed on ties (sleepers) and ballast, on which the rolling stock, usually fitted with metal wheels, moves. Other variations are also possible, such as slab track, where the rails are fastened to a concrete foundation resting on a prepared subsurface.
Property
Value
dbo:abstract
Rail transport is a means of conveyance of passengers and goods on wheeled vehicles running on rails, also known as tracks. It is also commonly referred to as train transport. In contrast to road transport, where vehicles run on a prepared flat surface, rail vehicles (rolling stock) are directionally guided by the tracks on which they run. Tracks usually consist of steel rails, installed on ties (sleepers) and ballast, on which the rolling stock, usually fitted with metal wheels, moves. Other variations are also possible, such as slab track, where the rails are fastened to a concrete foundation resting on a prepared subsurface. Rolling stock in a rail transport system generally encounters lower frictional resistance than road vehicles, so passenger and freight cars (carriages and wagons) can be coupled into longer trains. The operation is carried out by a railway company, providing transport between train stations or freight customer facilities. Power is provided by locomotives which either draw electric power from a railway electrification system or produce their own power, usually by diesel engines. Most tracks are accompanied by a signalling system. Railways are a safe land transport system when compared to other forms of transport. Railway transport is capable of high levels of passenger and cargo utilization and energy efficiency, but is often less flexible and more capital-intensive than road transport, when lower traffic levels are considered. The oldest, man-hauled railways date back to the 6th century BC, with Periander, one of the Seven Sages of Greece,



- **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

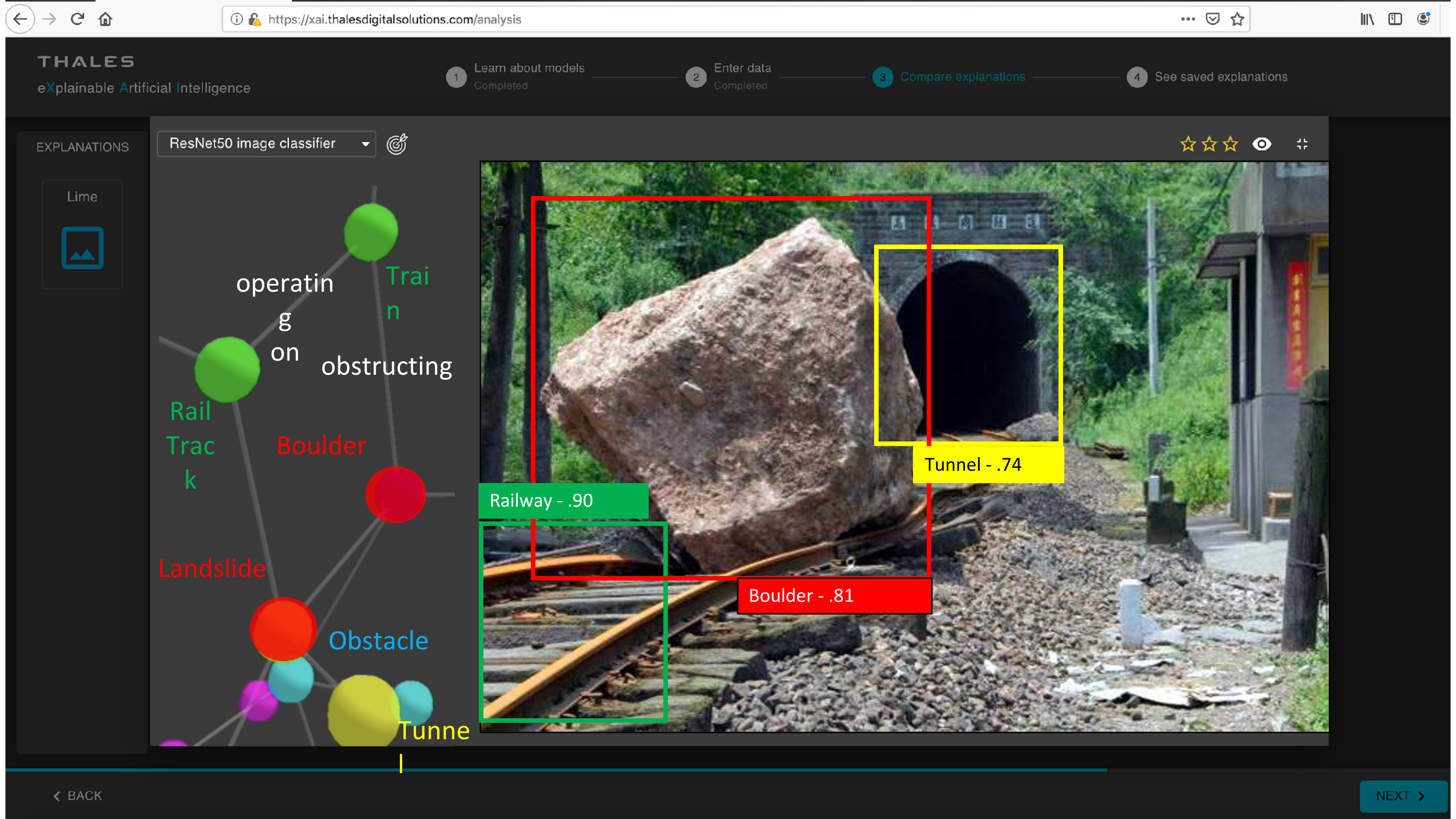
Transitionin g



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**

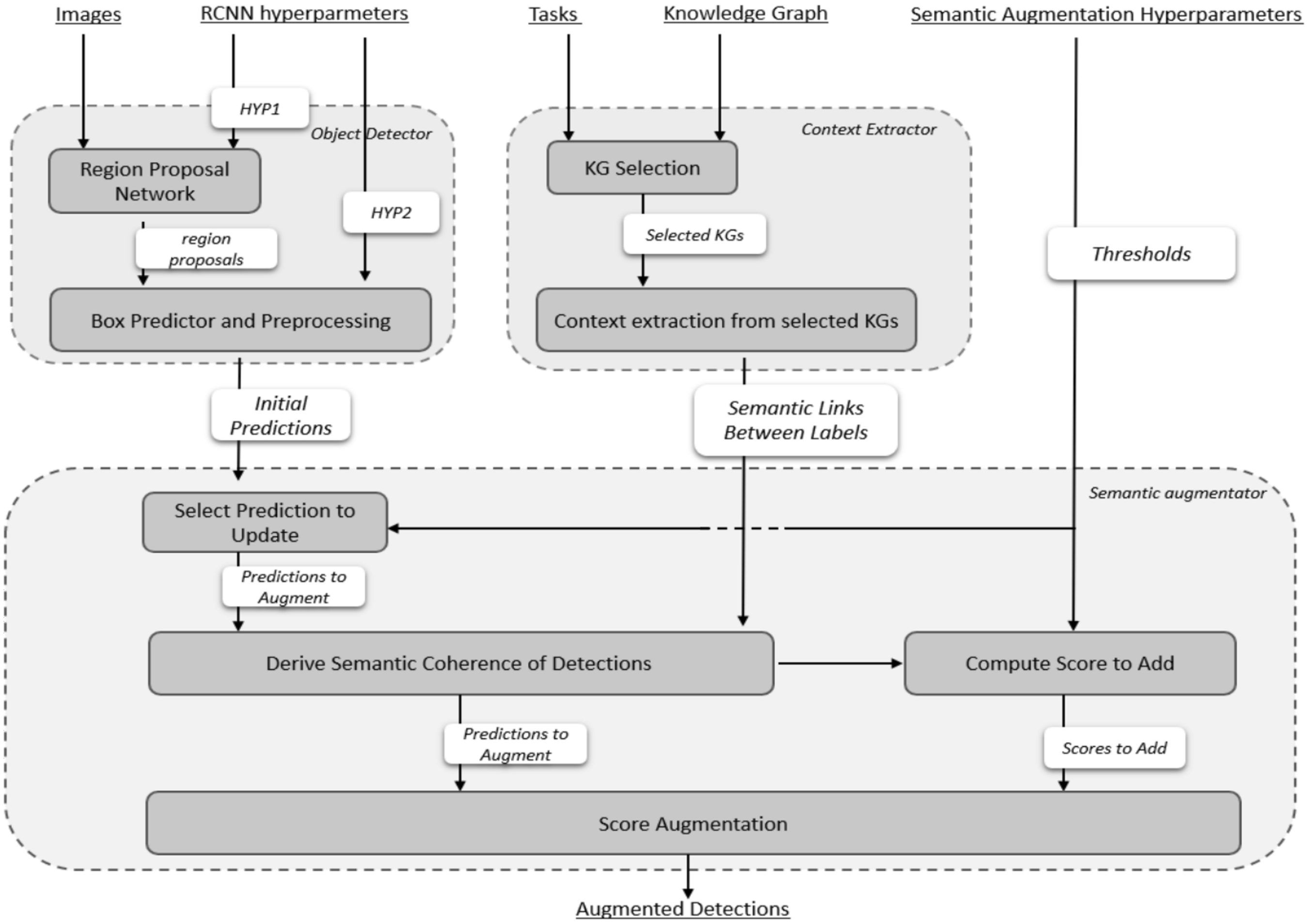
XAI Thales Platform

- 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

Part IV

On Interpretating Visual Question Answering Results with Graphs

What is Visual Question Answering (VQA)?

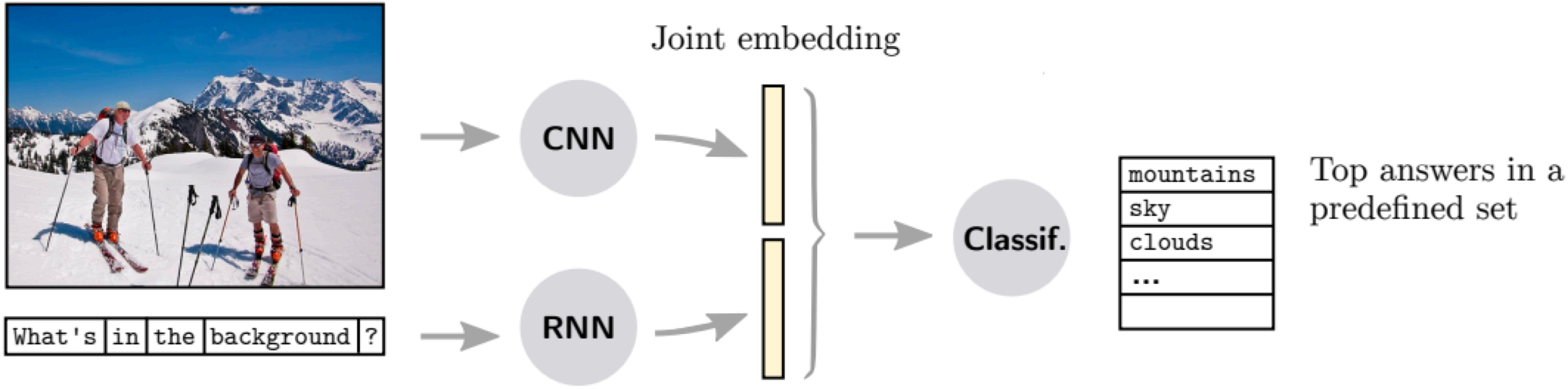
The objective of a VQA model combines **visual** and **textual** features in order to **answer questions** grounded in an **image**.

What's in the background?

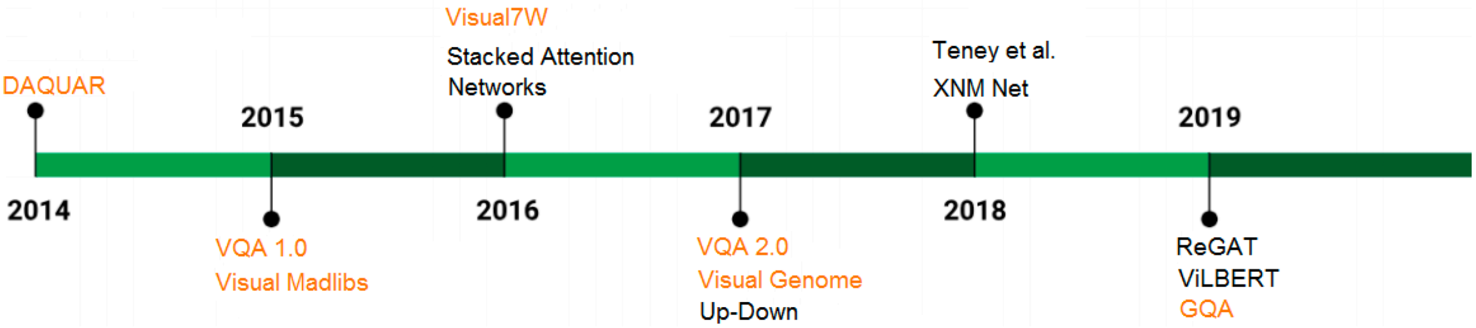
Where is the child sitting?

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 %)

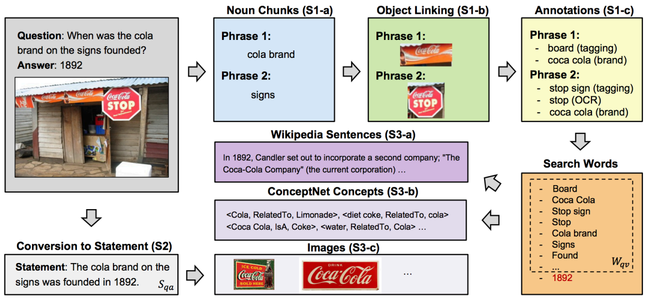
But they have limitations:

- 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.

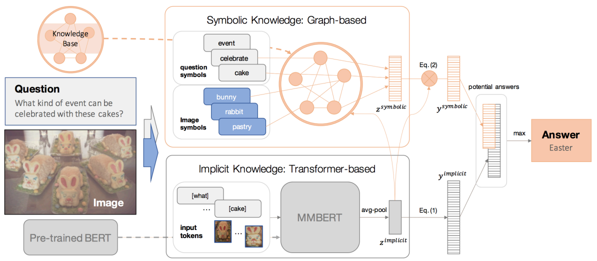
State of the Art in Visual Question Answering + Graph

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



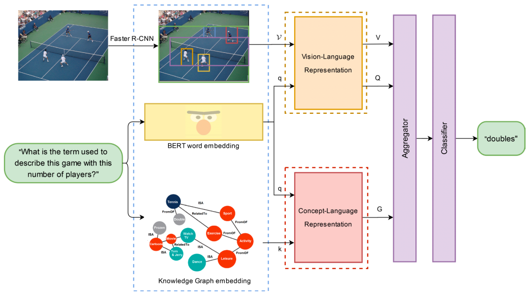
Search-based (MAVEx)

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



Graph-Embedding-based (KRISP)

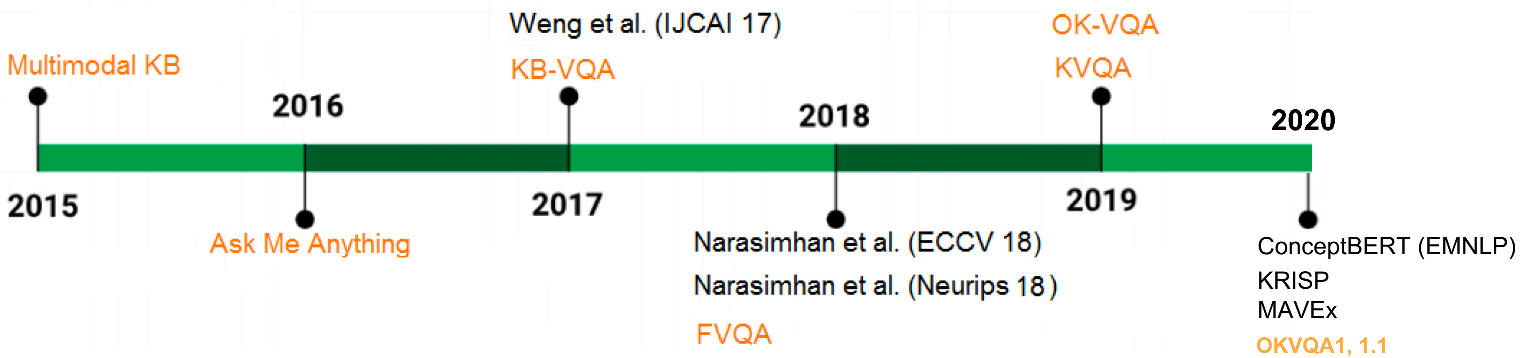
https://arxiv.org/pdf/2012.11014.pdf



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 **<u>retrieve explanations</u>** of a VQA model during inference?
- How to expose articulated knowledge (i.e., **<u>composition of knowledge graph triples</u>**) to explain how an answer is related to the question, objects of the images and concepts?

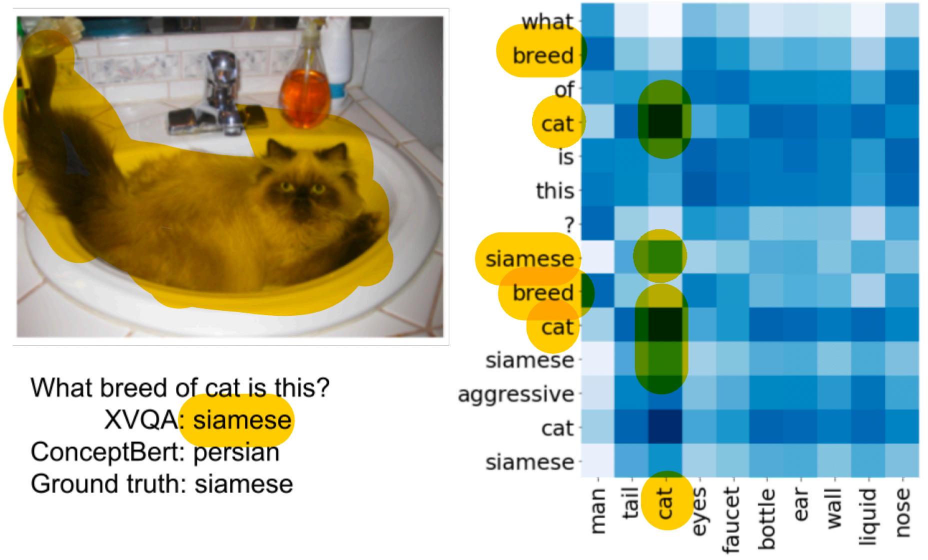
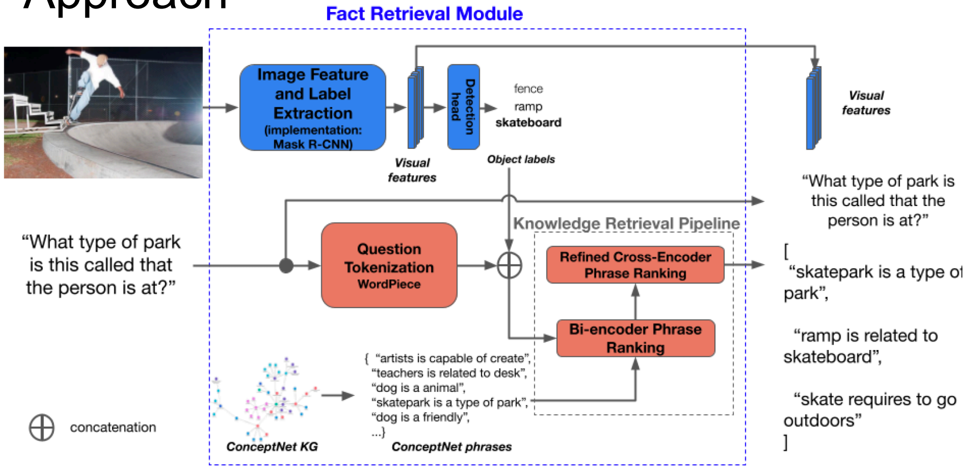


Figure 1: An example of VQA task with question: *What breed of cat is it?* on the left image, and our XVQA Answer: *Siamese*. XVQA also exhibits explanations from the optimal transfer map between (i) question tokens (vertical tokens on the right image: cat, breed), graph entities (vertical tokens after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

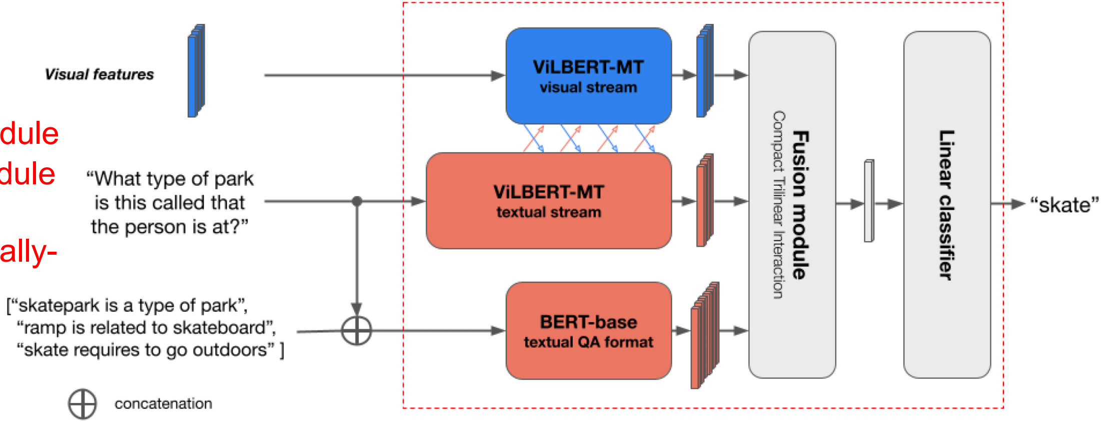
1) Bi-Encoder Phrase Ranking to compute query agnostic fact phrase embeddings
2) Refined Cross-Encoder Phrase Ranking for each model

[redacted]

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

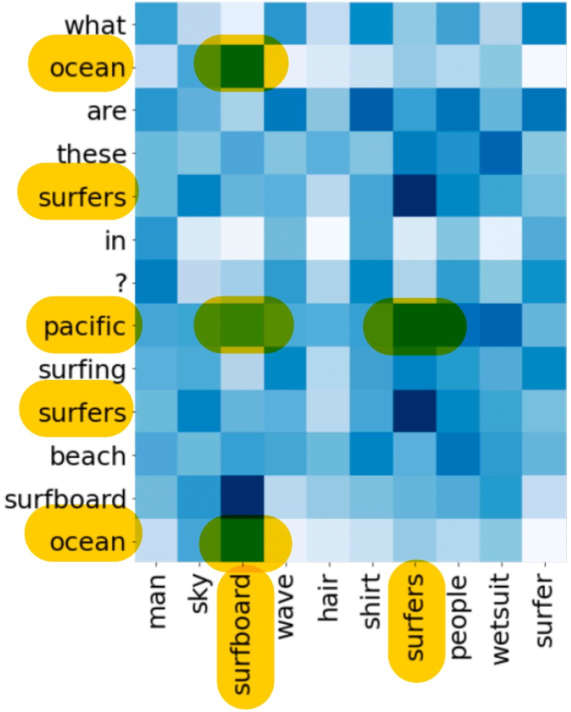
Model / Data type	OK-VQAv1	OK-VQAv1.1
XVQA	33.2%	39.7%
XVQA (without facts)	32.6%	38.9%
XVQA (oracle case)	46.3%	54.7%
ConceptBERT	33,0%	–
ViLBERT	35.2%	41.6%
KRISP	38.35%	38.9%
MAVEx	–	40.5%
MAVEx (oracle case)	–	43.5%

eXplainable Visual Question Answering using Knowledge Graphs (4)

Qualitative Results

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

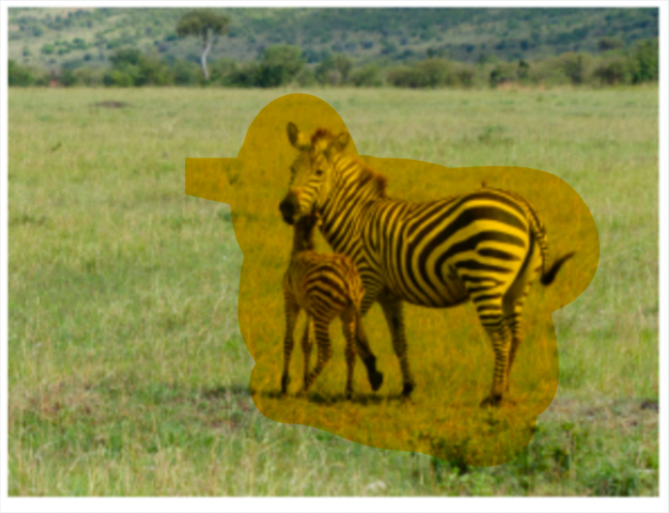
XVQA: pacific
ConceptBert: surf
Ground truth: pacific



(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*.

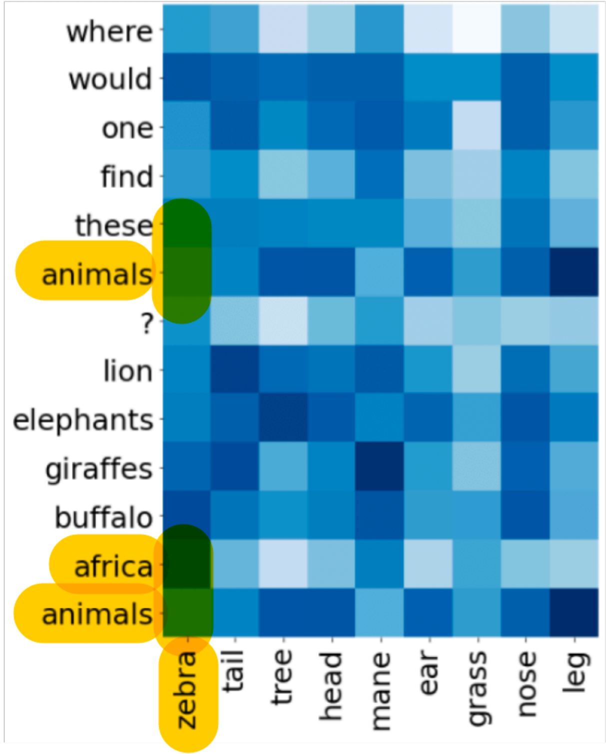
eXplainable Visual Question Answering using Knowledge Graphs (5)

Qualitative Results



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

XVQA: africa
ConceptBert: africa
Ground truth: africa



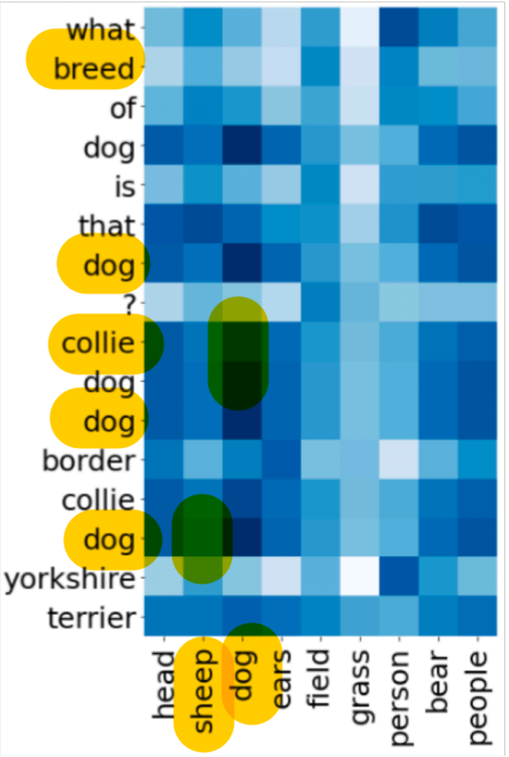
(2) Here the optimal transfer map is between (i) question tokens (vertical tokens: animals), graph entities (vertical tokens: africa, animals) and (ii) detected object (horizontal tokens: zebra) embeddings i.e., *africa has animals*.

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

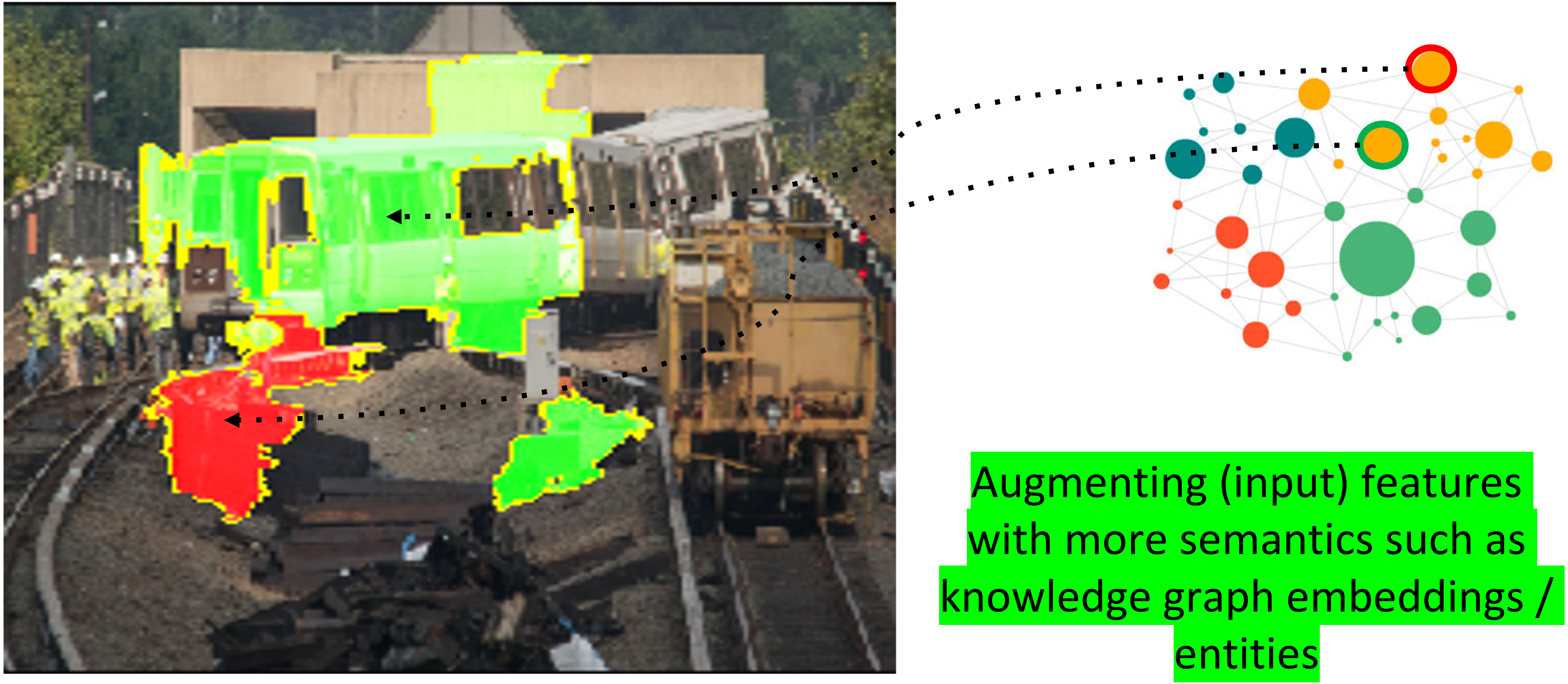
- **Retrieving explanations** of a VQA model during inference is a complex task
- Exposing articulated knowledge (i.e., **composition of knowledge graph triples**) to explain how an answer is related to the question, objects of the images and concepts is highly depending **on relevant retrieved knowledge**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

Part V

Even More Opportunities for Knowledge Graphs in Deep Neural Networks

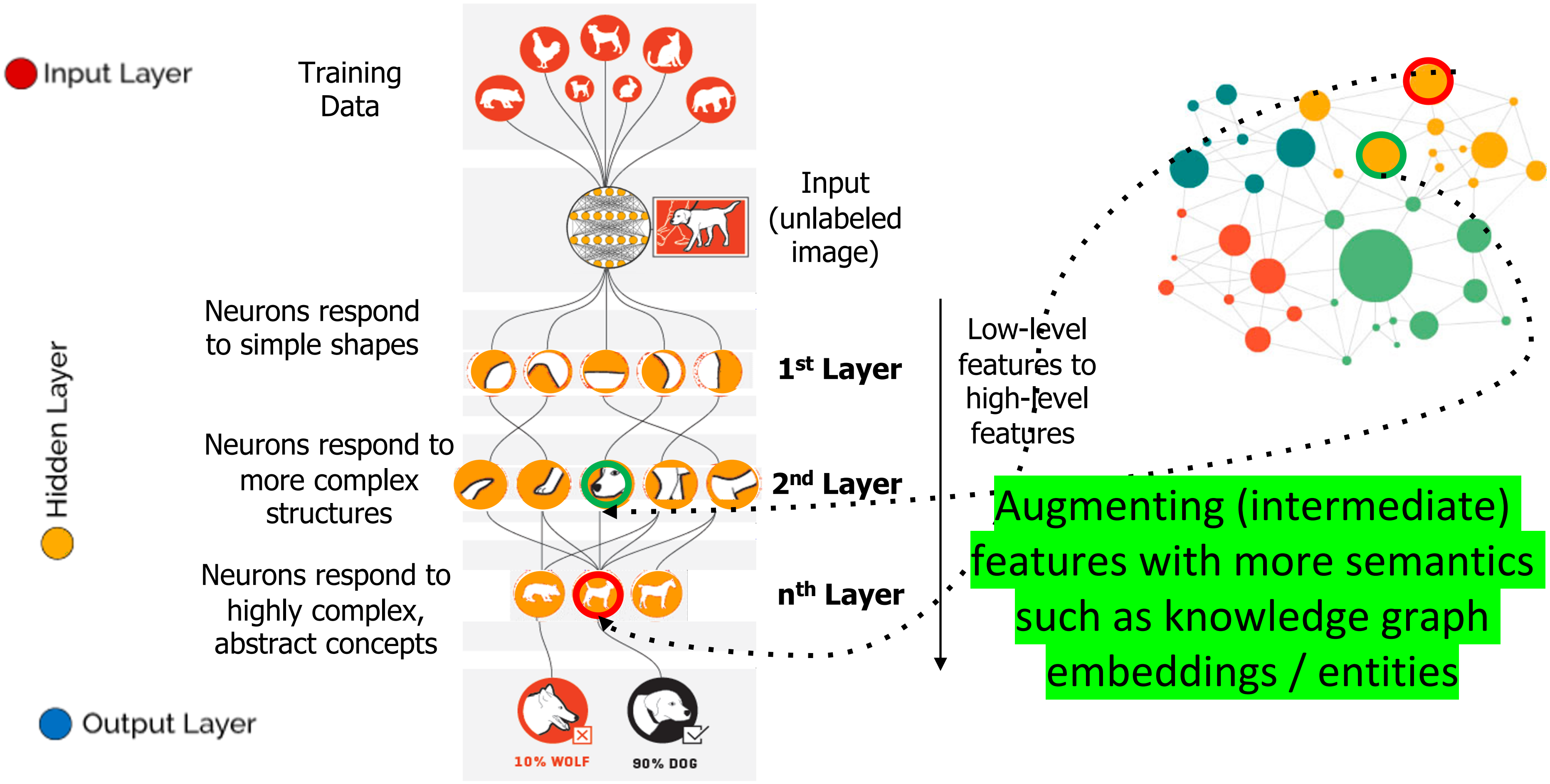
Knowledge Graph in Machine Learning (1)



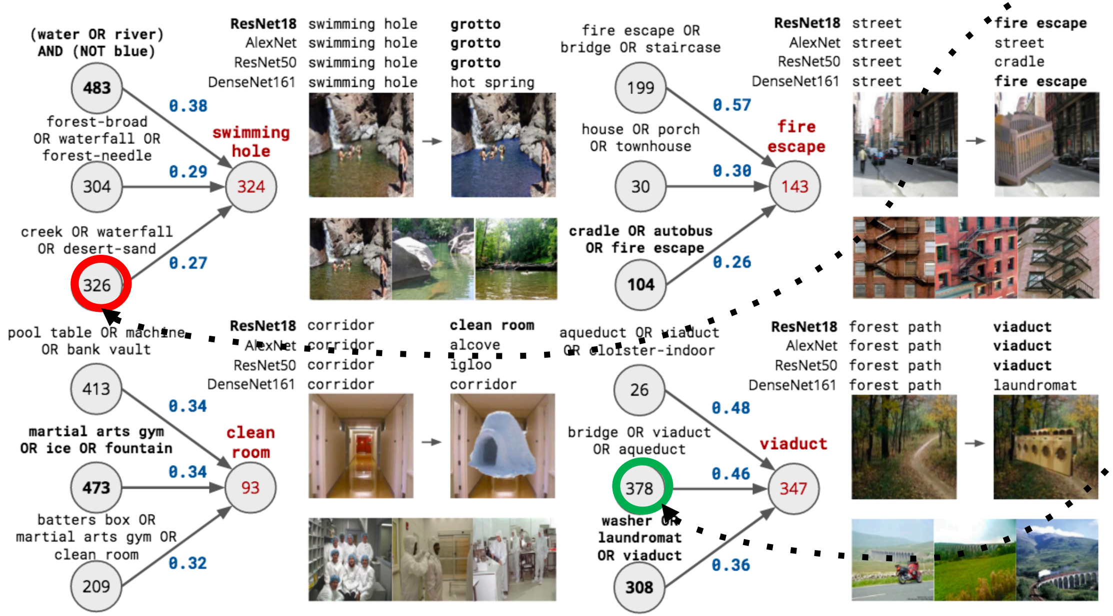
Augmenting (input) features with more semantics such as knowledge graph embeddings / entities

https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret

Knowledge Graph in Machine Learning (2)



Knowledge Graph in Machine Learning (3)



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

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

Knowledge Graph in Machine Learning (4)

Input Layer

Training Data

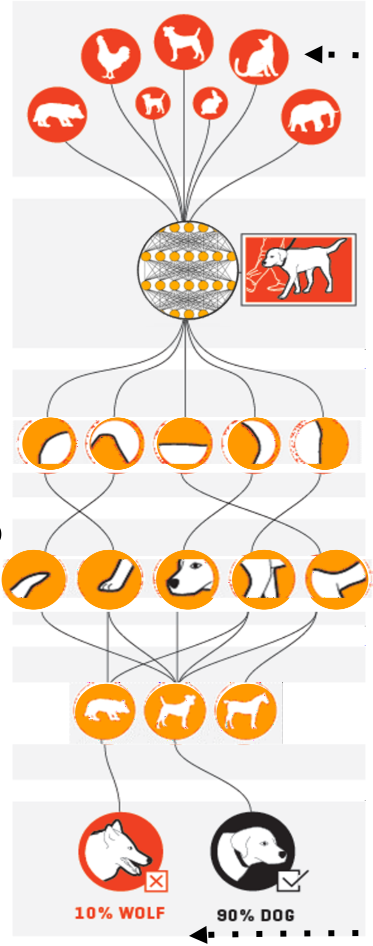
Hidden Layer

Neurons respond to simple shapes

Neurons respond to more complex structures

Neurons respond to highly complex, abstract concepts

Output Layer



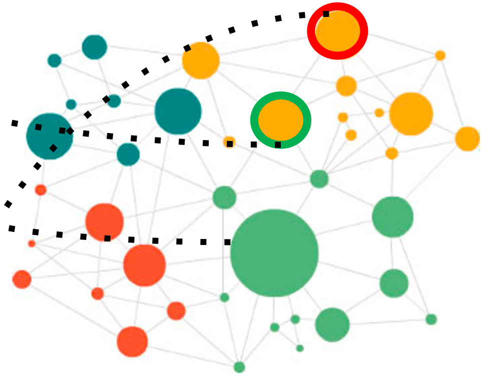
Input (unlabeled image)

1st Layer

2nd Layer

nth Layer

Low-level features to high-level features



Augmenting (input, intermediate) features – output relationship with more semantics to capture causal relationship

Knowledge Graph in Machine Learning (5)

Description 1: This is an orange train accident

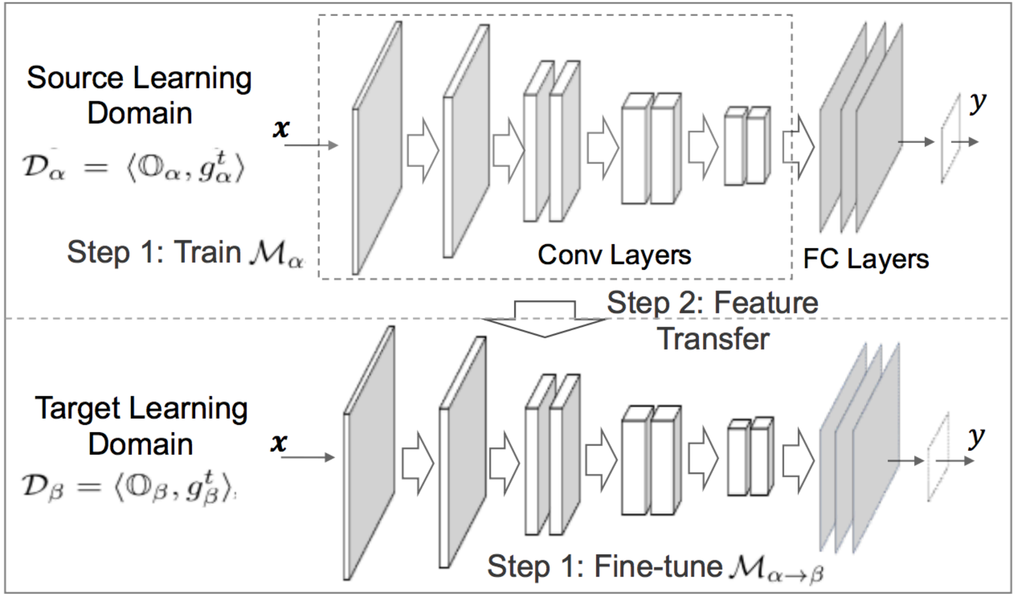
Description 2: This is a train accident between two speed merchant trains of characteristics X43-B and Y33-C in a dry environment

Description 3: This is a public transportation accident

Augmenting models with semantics to support personalized explanation

Knowledge Graph in Machine Learning (6)

"How to explain transfer learning with appropriate knowledge representation?



Augmenting input features and domains with semantics to support interpretable transfer learning

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

Knowledge Graph in Machine Learning (7)

"How to explain concept drift in Machine Learning?

Augmenting input features and domains with semantics to interpret concept drift in Machine Learning

Without semantics augmentation

With semantics augmentation

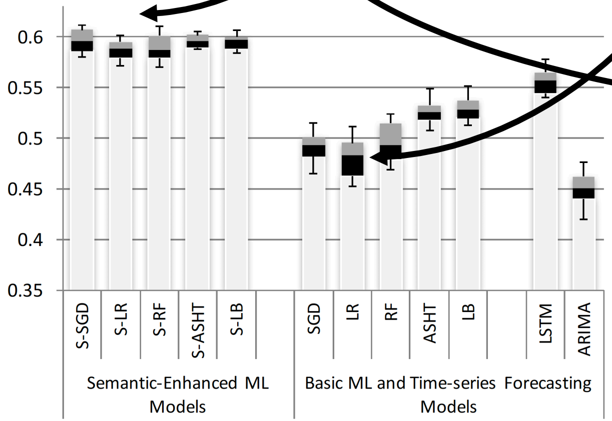


Figure 6: [Beijing Context] Baseline Comparison of Forecasting Macro-F1 Score (Evaluation of Algorithm 1-3), where $\Delta = 6$.

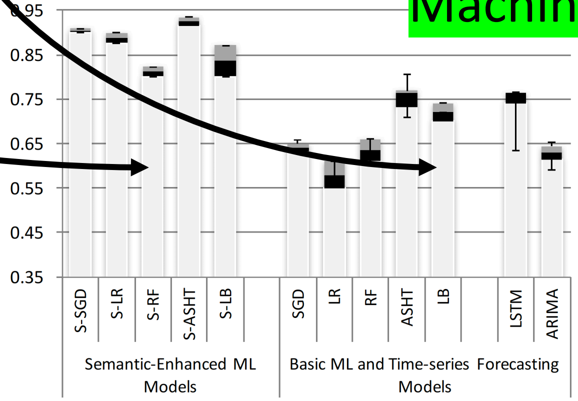
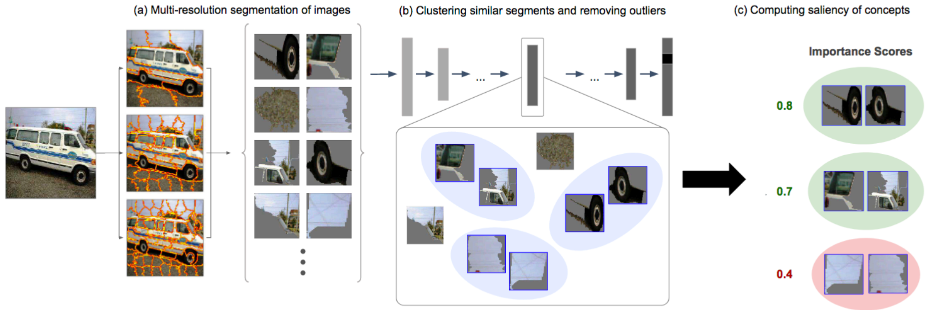


Figure 7: [Dublin Context] Baseline Comparison of Forecasting Macro-F1 Score (Evaluation of Algorithm 1-3), where $\Delta = 6$.

Jiaoyan Chen and Freddy Lécué and Jeff Z. Pan and Shumin Deng and Huajun Chen. Knowledge graph embeddings for dealing with concept drift in machine learning. Journal of Web Semantics. (2021) http://www.sciencedirect.com/science/article/pii/S1570826820300585

Knowledge Graph in Machine Learning (8)

- Towards more semantic interpretation



ACE

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

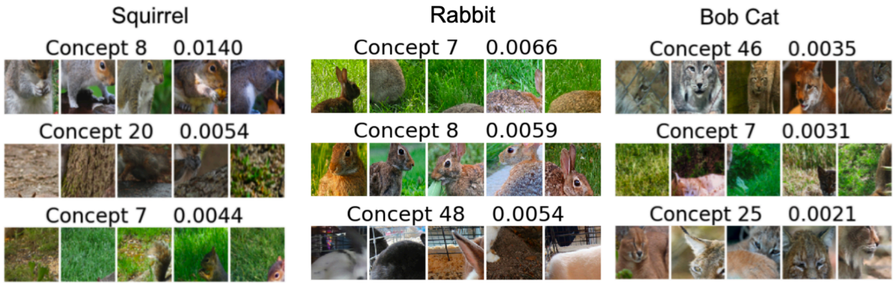


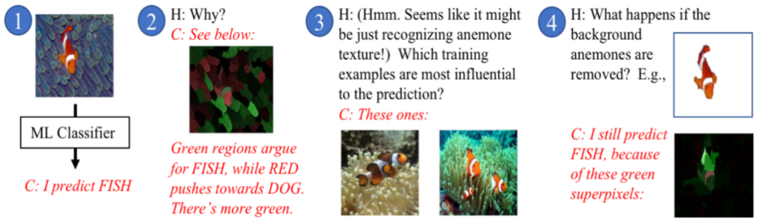
Figure 3: Concept examples with the samples that are the nearest to concept vectors in the activation space in AwA. The per-class ConceptSHAP score is listed above the images.

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

Part VI

Conclusion

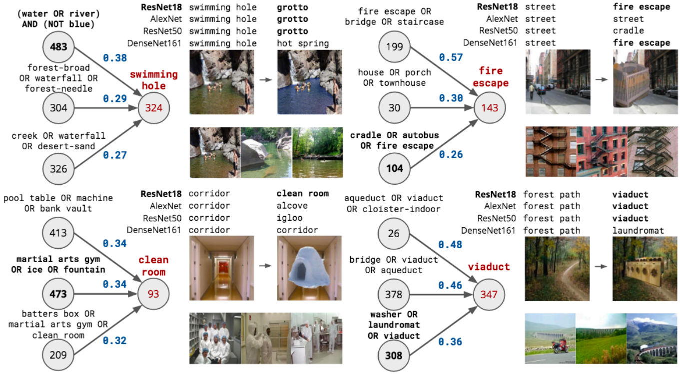
The Good: Multimodal End-to-End XAI System



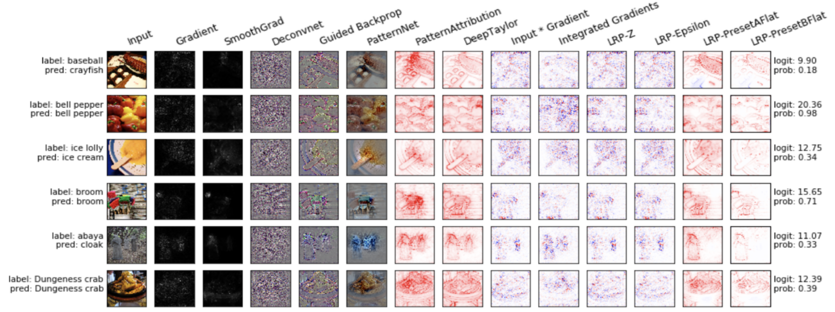
The Bad: Feature Visualization

Knowledge Graph as Semantic Glue for XAI in Deep Neural Networks

The (not so) **Bad**: Network Dissection Neurons Composition



The Ugly: Saliency Maps Super-Pixels



Thanks! Questions?

- Feedback most welcome :-)
  - freddy.lecue@inria.fr (@freddylecue)
  - freddy.lecue@thalesgroup.com
- Slides: https://tinyurl.com/hs73b88u

INVENTEURS DU MONDE NUMÉRIQUE