Explaining Deep Neural Networks

The Good, the Bad and the Ugly

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

Knowledge Graph Conference
Workshop on Application of Reasoning on Complex and Evolving Data: Methods and Use-Cases

THALES

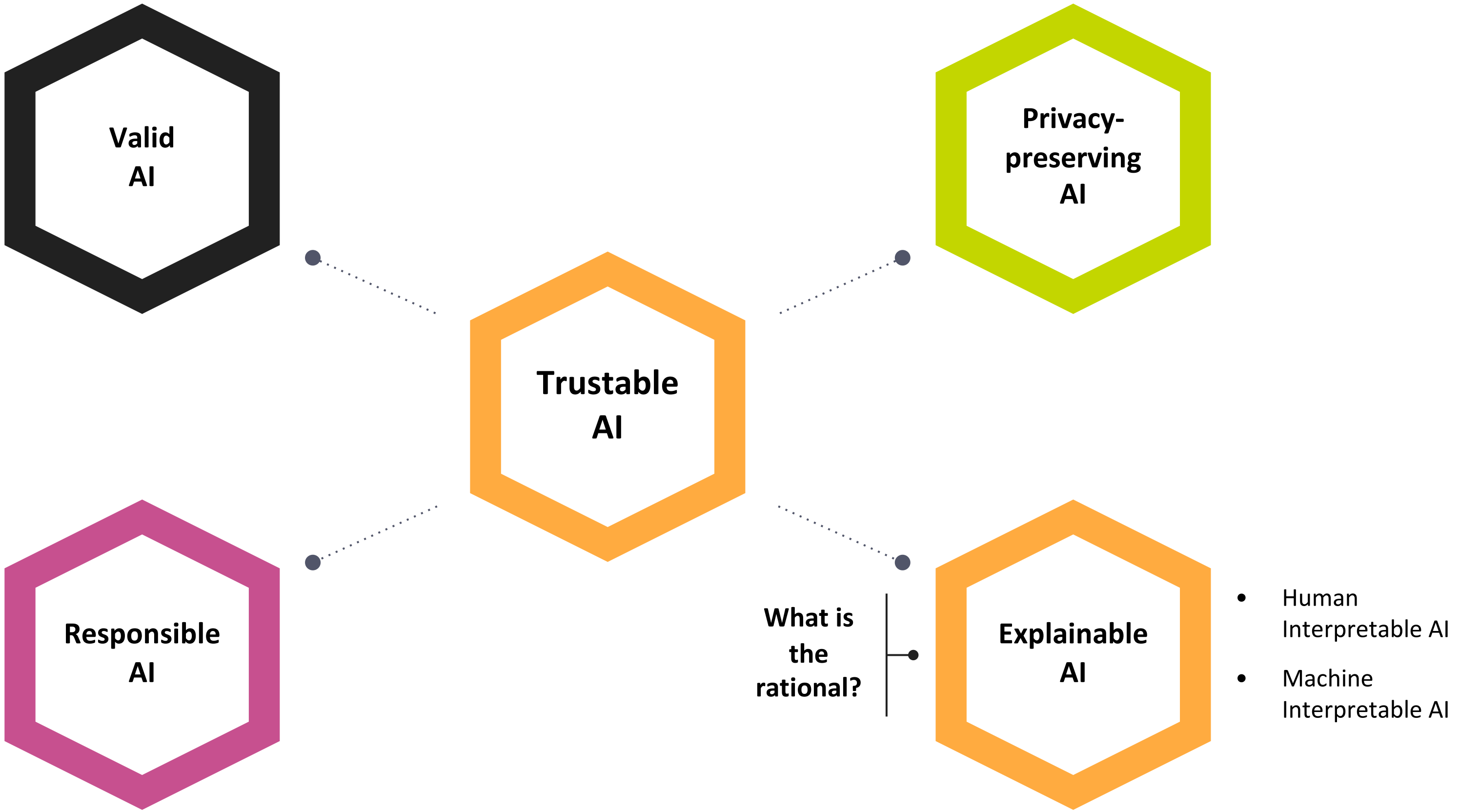
INVENTEURS DU MONDE NUMÉRIQUE

May 2, 2022

https://tinyurl.com/hs73b88u

Scope

AI Adoption: Requirements



Part I

Introduction and Motivation

Explanation - From a Business Perspective

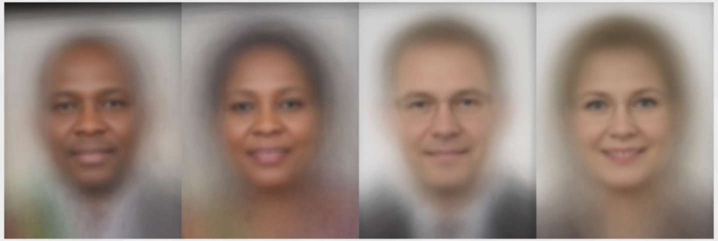
Business to Customer AI

Critical Systems (1)

Critical Systems (2)

… and even More

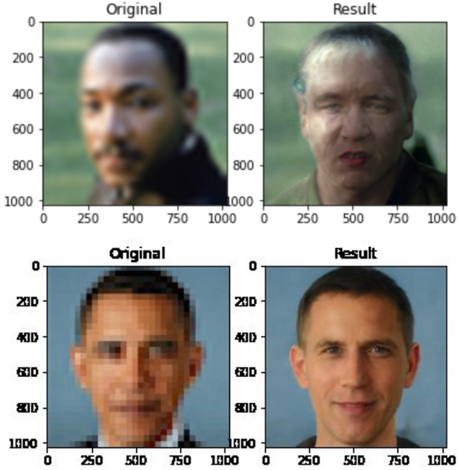
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



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

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

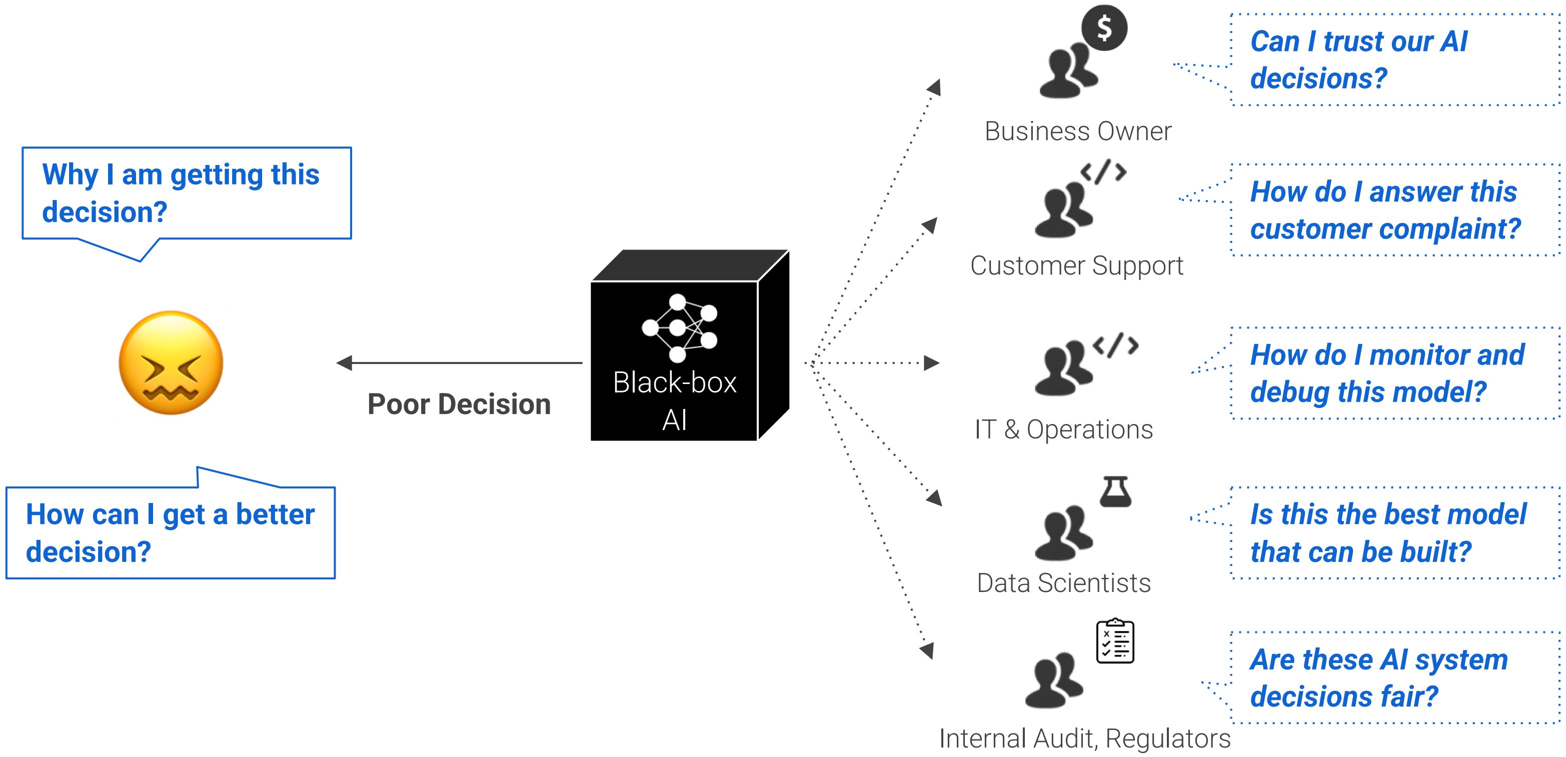
https://www.cbsnews.com/news/apple-credit-card-goldman-sachs-disputes-claims-that-apple-card-is-sexist/



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

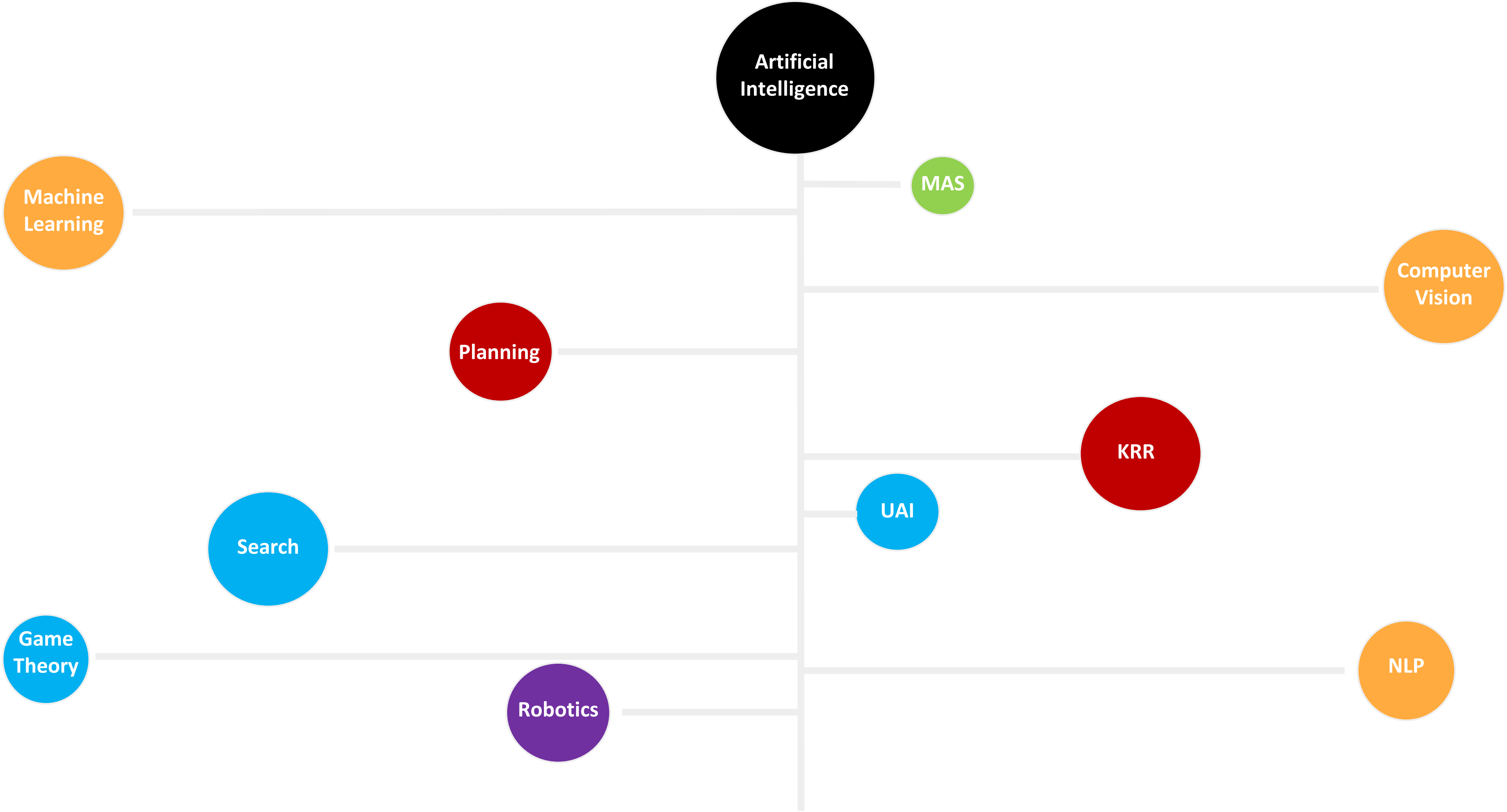
Is the explanation complete, partial, restricted?

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

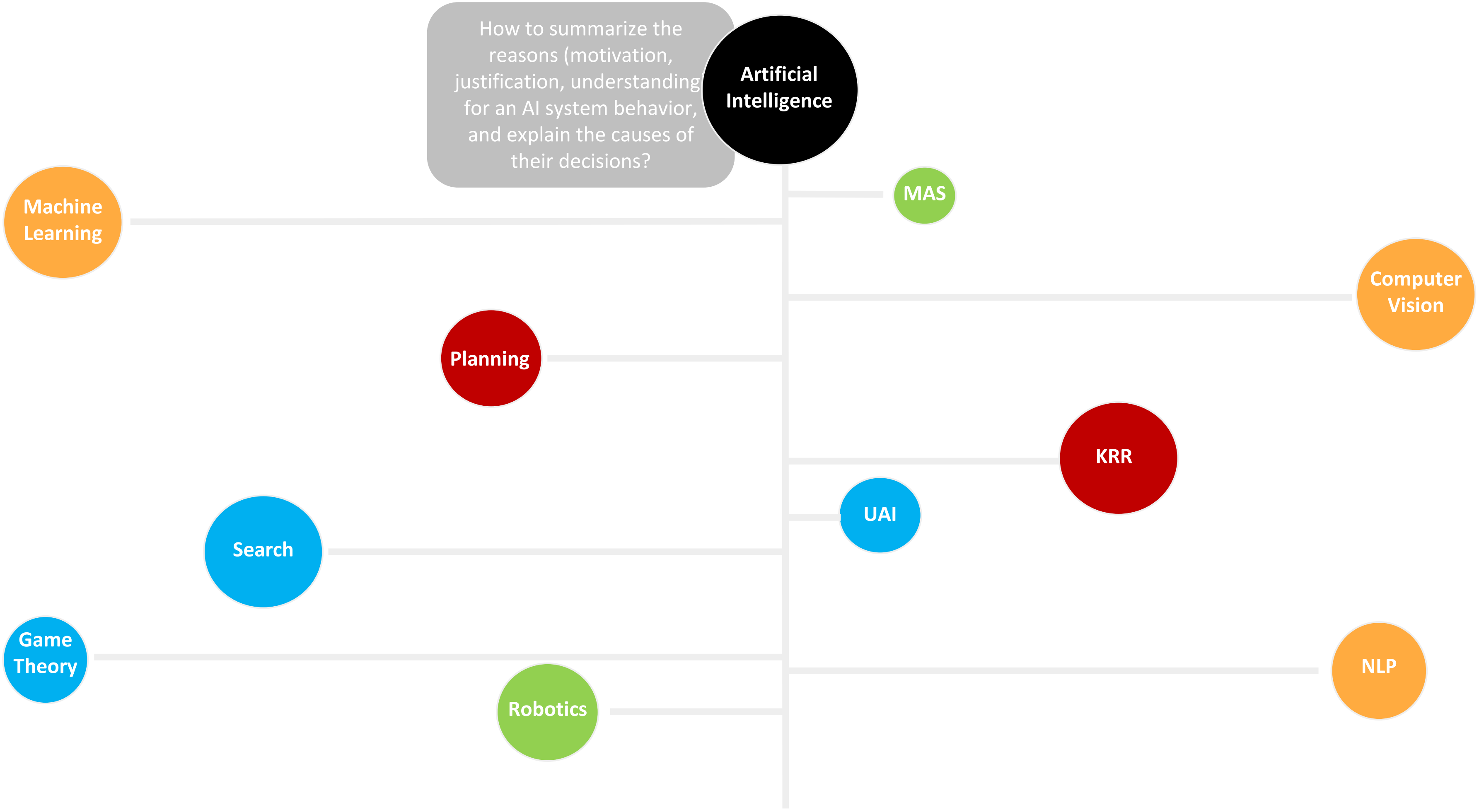
Part II

Explanation in AI (Focus Deep Neural Networks)

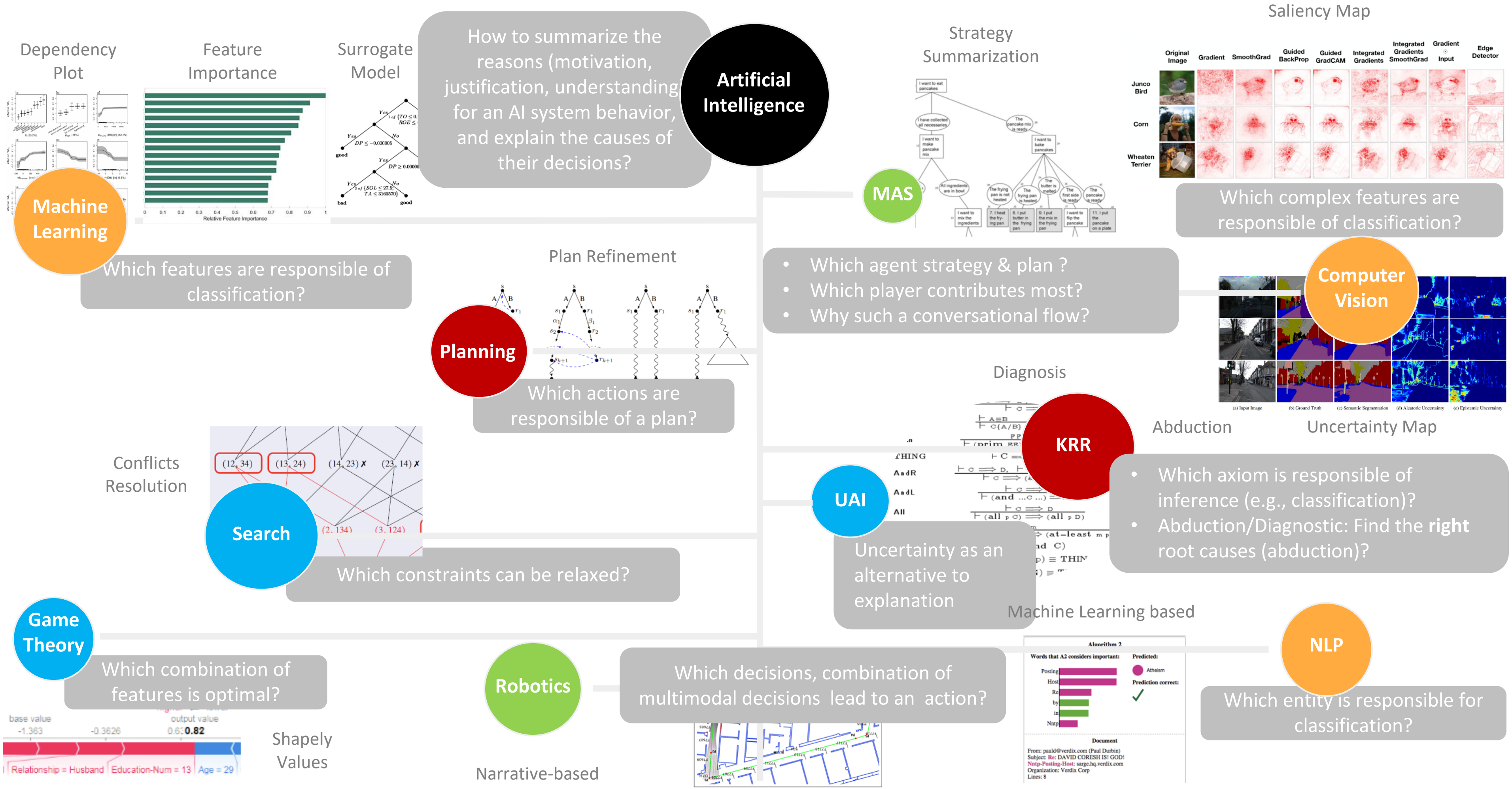
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



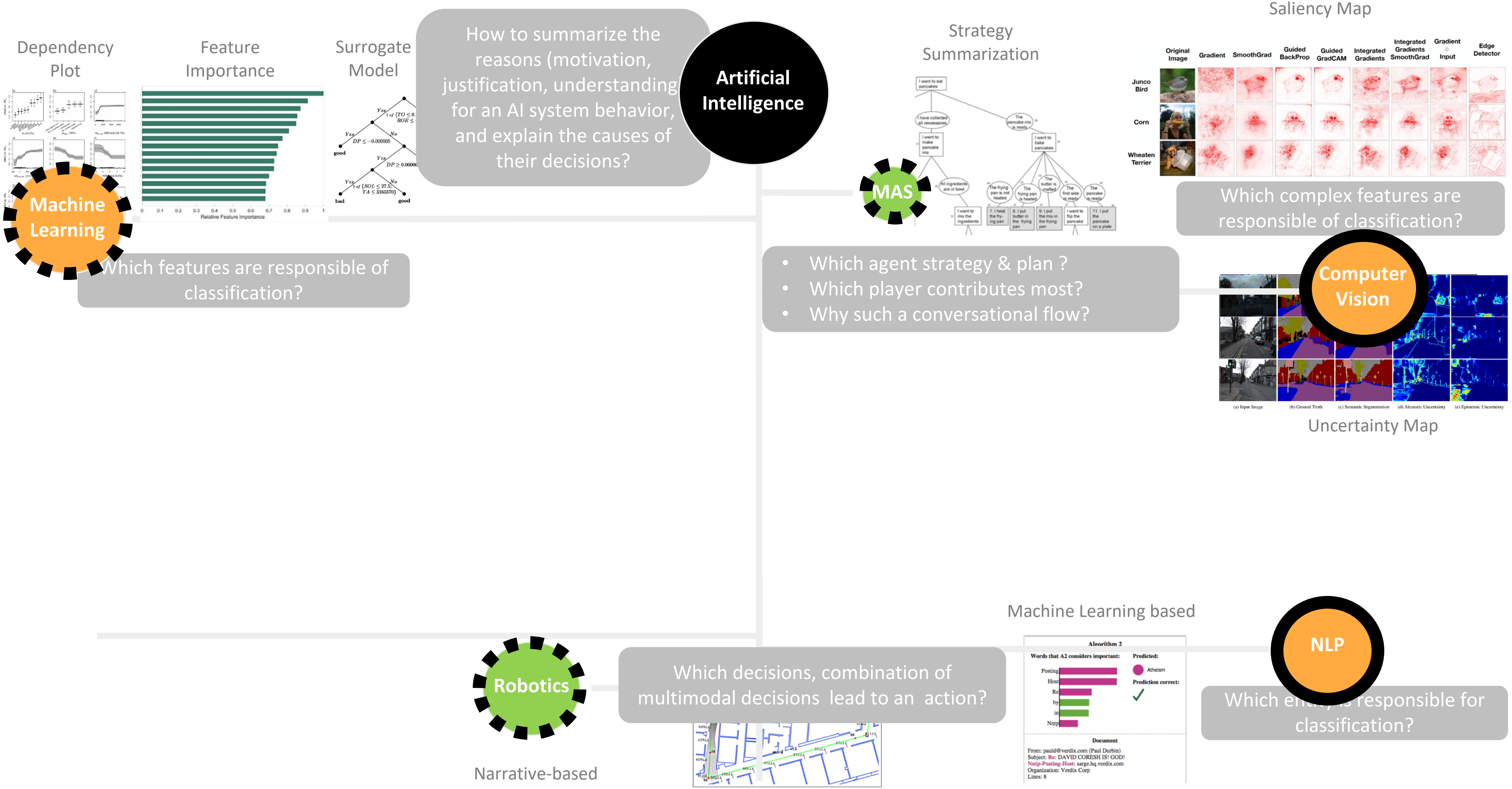
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



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



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



Part III

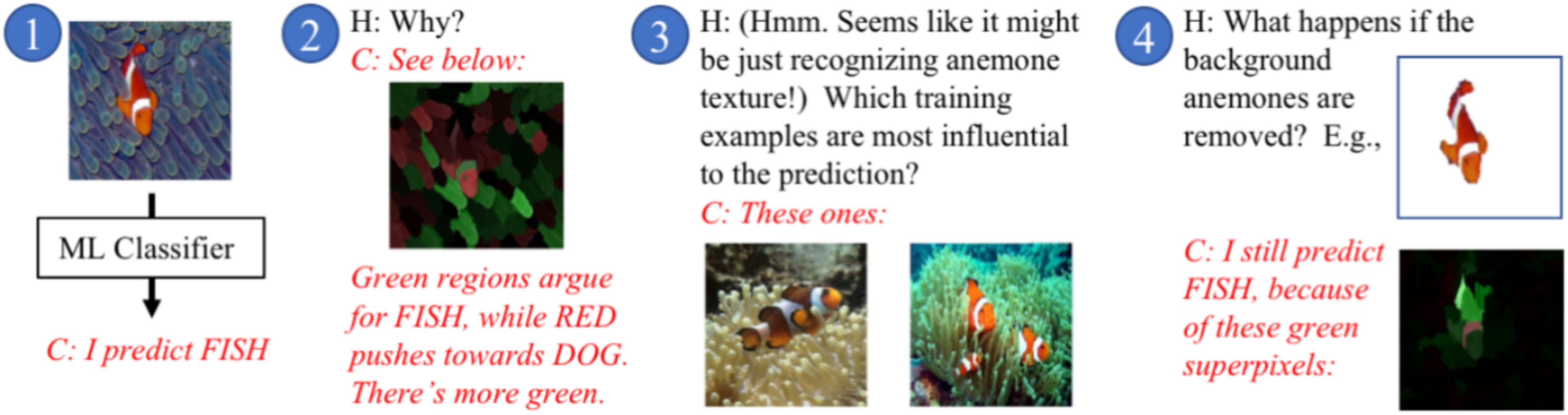
XAI:

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

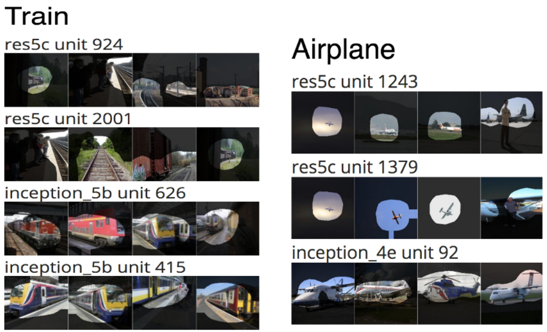
- [Interaction] Human are in the loop (What-if / counterfactual)
- [Construction] Iterative explanation search
- [Validation] Operator as opposed to developer driven
- [Knowledge] Domain knowledge is required

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

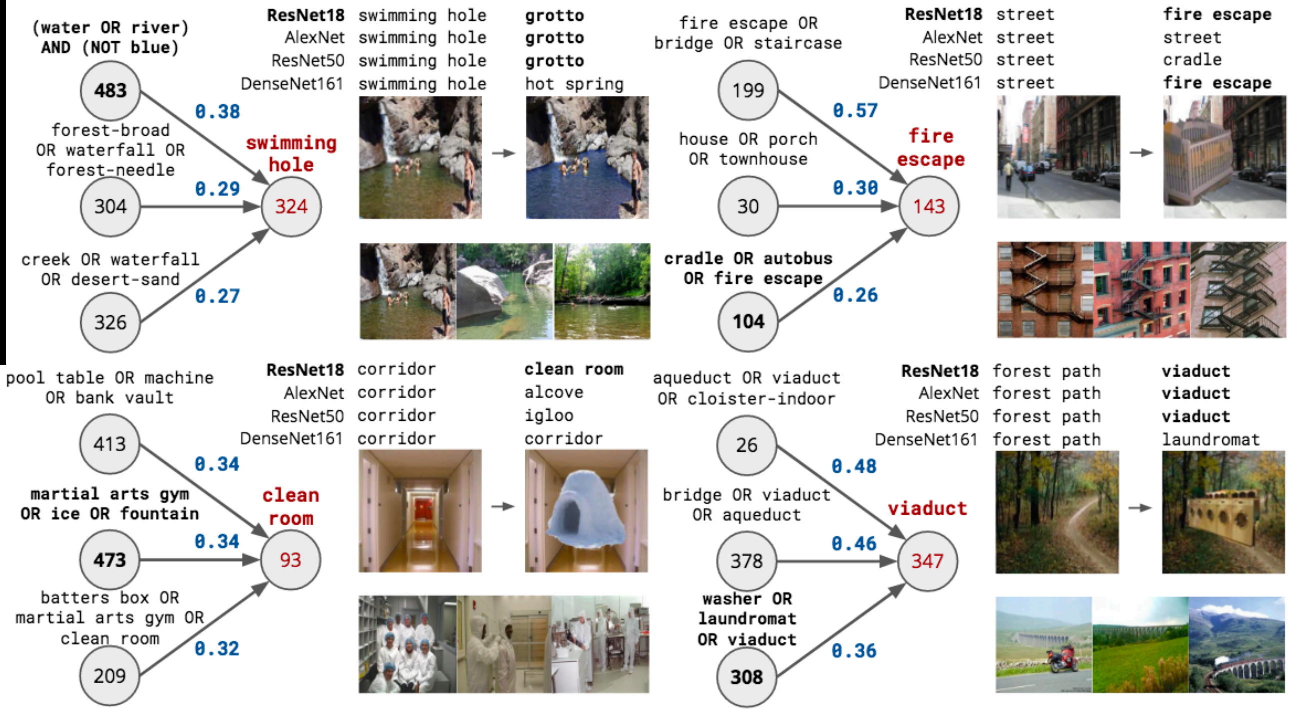
The (not so) **Bad**: Network Dissection | Neurons Composition

The (not so) **Bad**

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



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



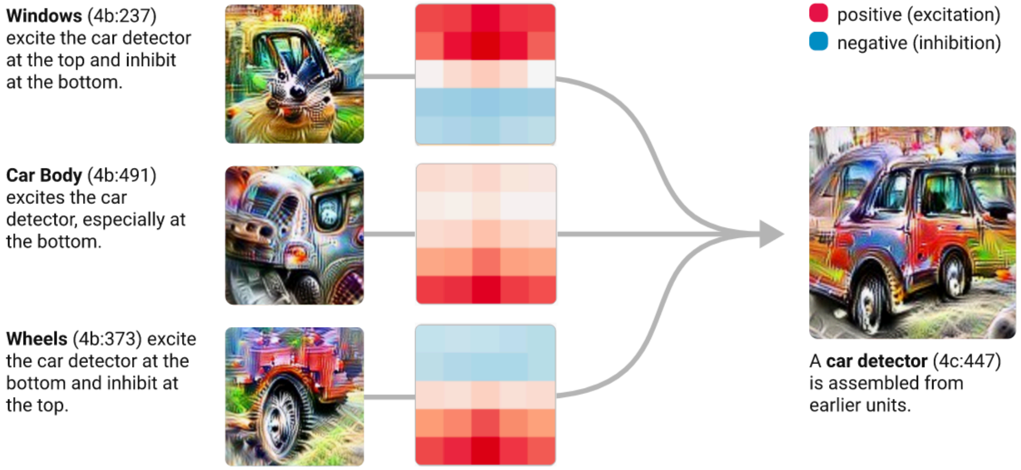
Jesse Mu, Jacob Andreas: Compositional Explanations of Neurons. NeurIPS 2020

The Bad: Feature Visualization

The Bad

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

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

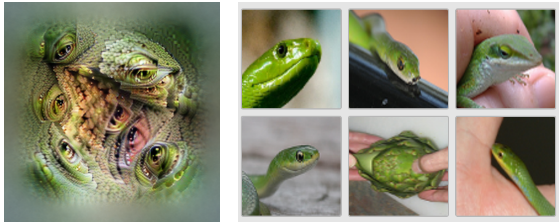


CLIP
Resent 50
Layer 4

https://microscope.openai.com/models

Unit 118

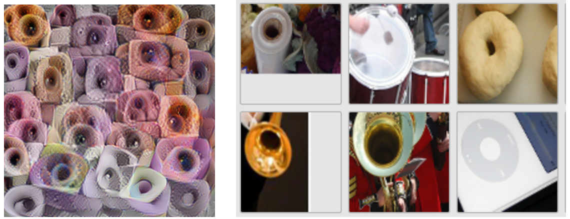
Unit 55



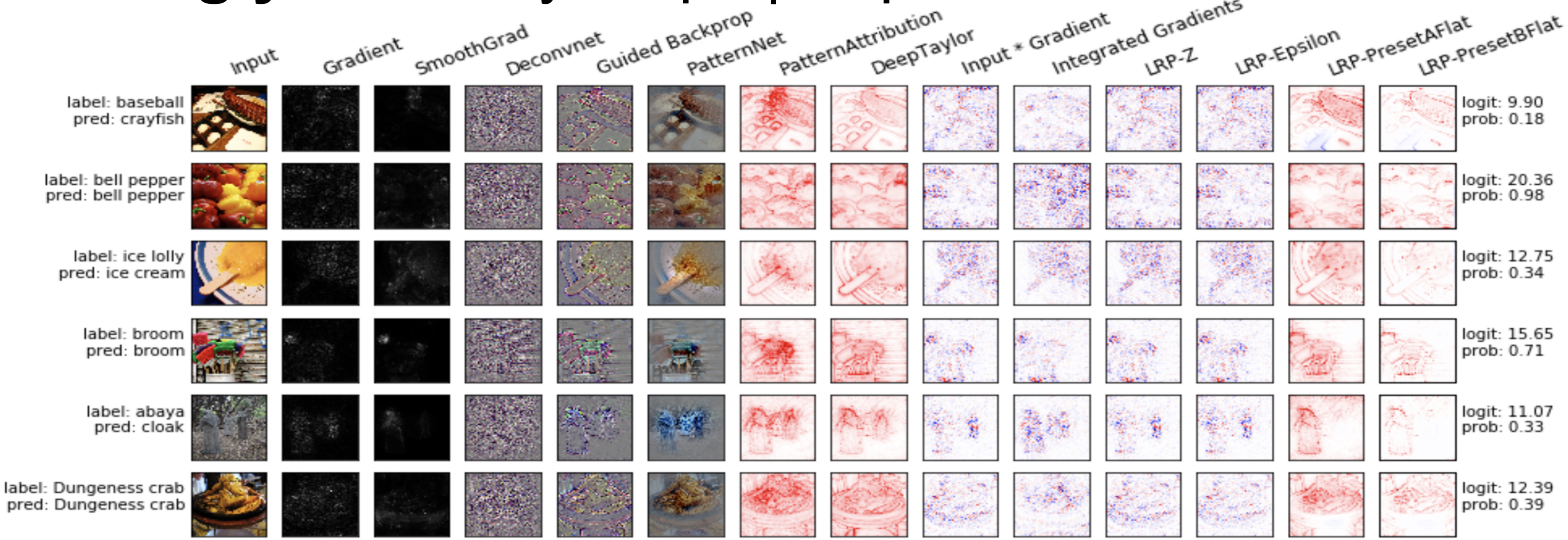
Resnet 50 v2
Block4/unit_3/add

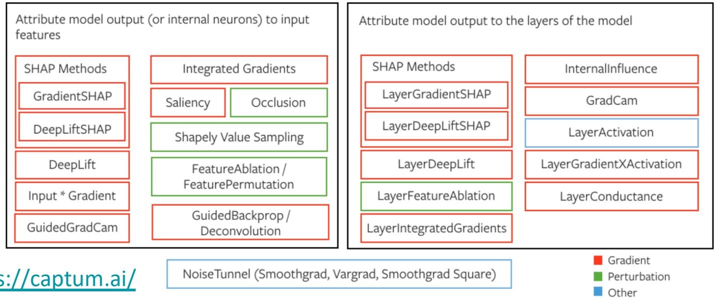
Unit 546

Unit 562



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 Interpretating Visual Question Answering Results with Graphs

What is Visual Question Answering (VQA)?

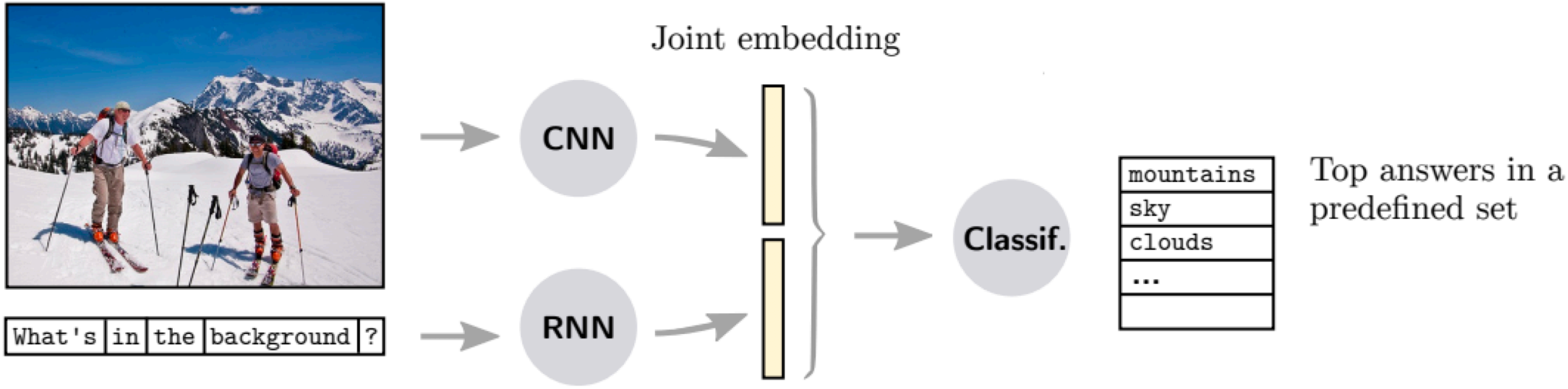
The objective of a VQA model combines **visual** and **textual** features in order to **answer questions** grounded in an **image**.

What's in the background?

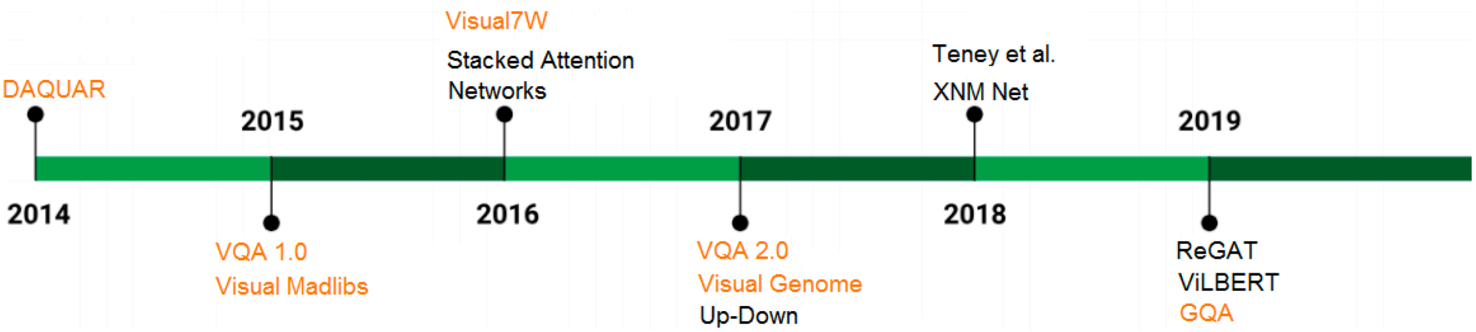
Where is the child sitting?

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 %)

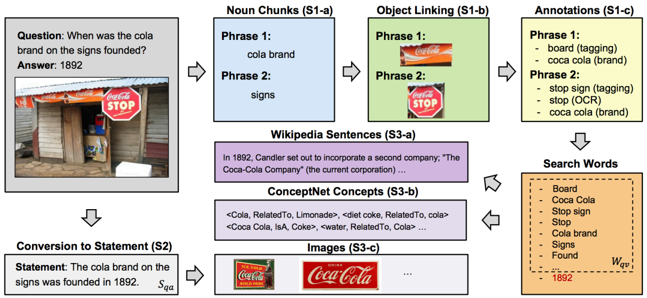
But they have limitations:

- 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.

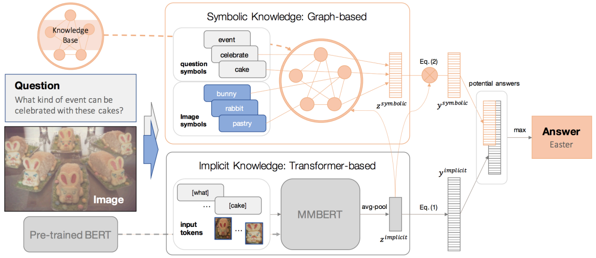
State of the Art in Visual Question Answering + Graph

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



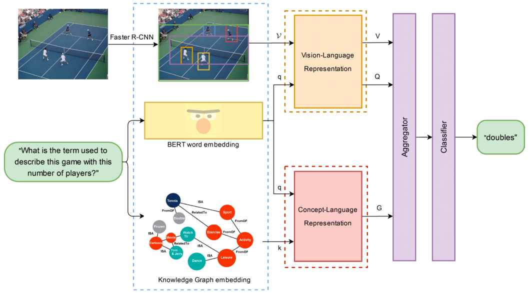
Search-based (MAVEx)

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



Graph-Embedding-based (KRISP)

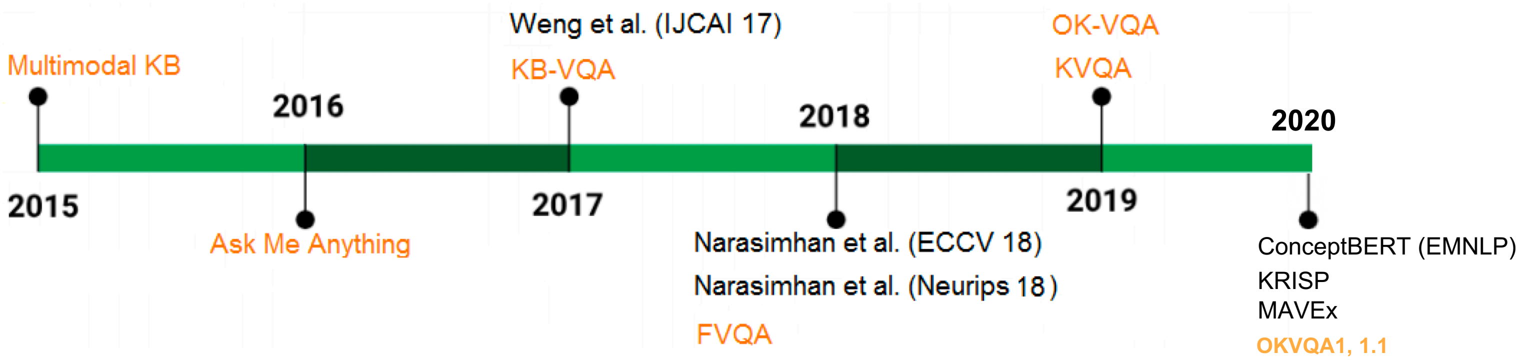
https://arxiv.org/pdf/2012.11014.pdf



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 **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?

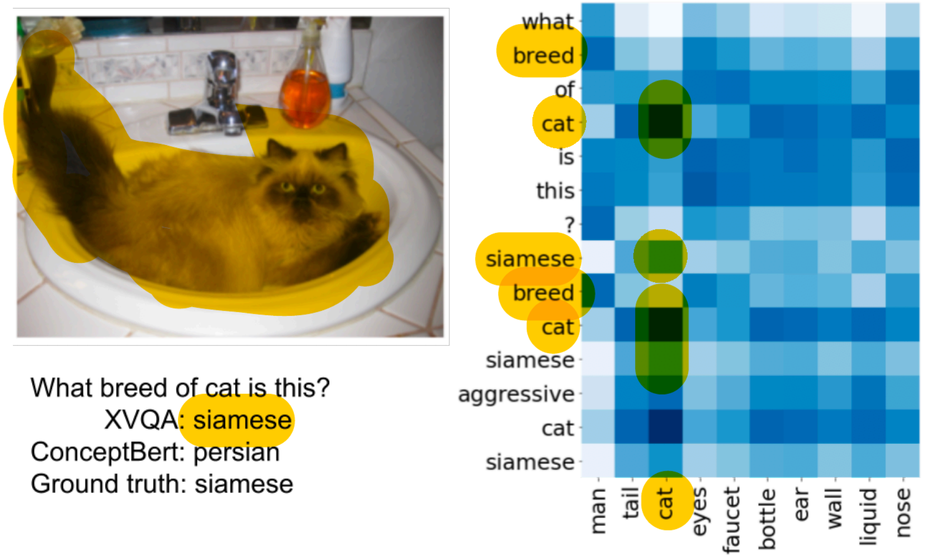
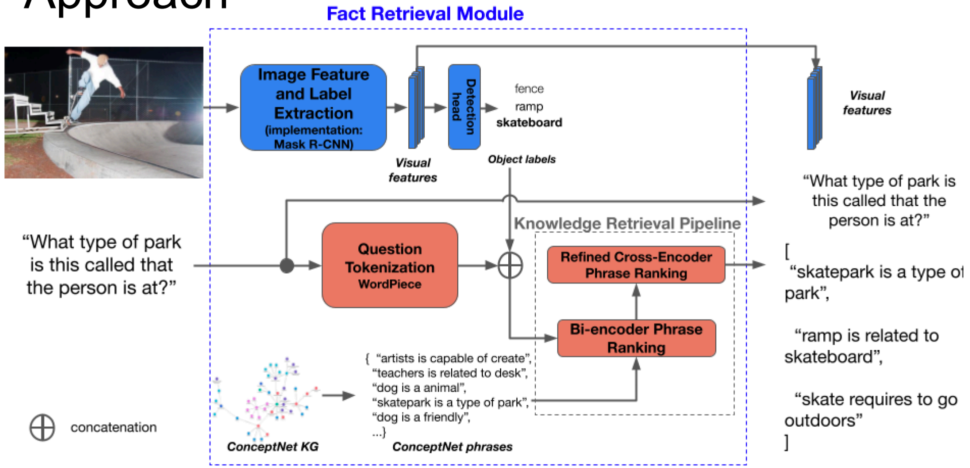


Figure 1: An example of VQA task with question: *What breed of cat is it?* on the left image, and our XVQA Answer: *Siamese*. XVQA also exhibits explanations from the optimal transfer map between (i) question tokens (vertical tokens on the right image: cat, breed), graph entities (vertical tokens after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

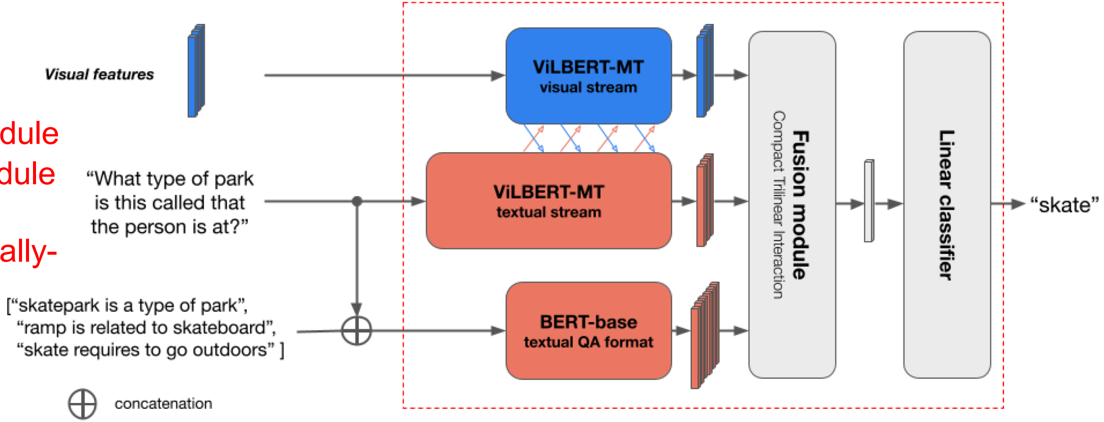
1) Bi-Encoder Phrase Ranking to compute query agnostic fact phrase embeddings
2) Refined Cross-Encoder Phrase Ranking for each model

[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)

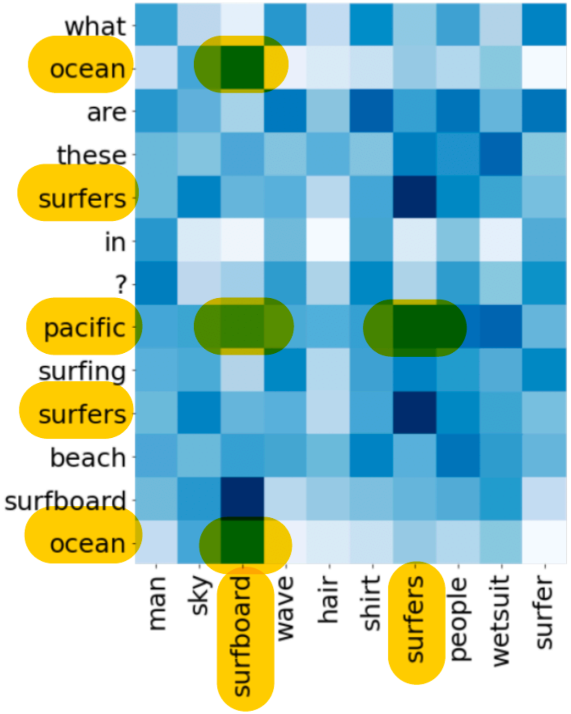
Qualitative Results

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

XVQA: pacific

ConceptBert: surf

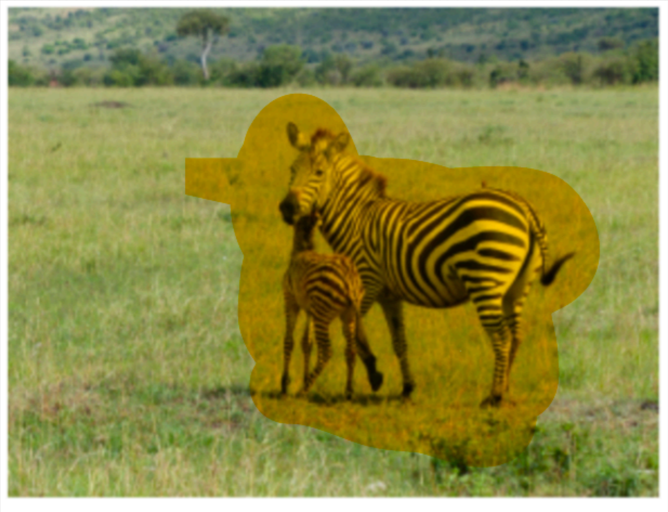
Ground truth: pacific



(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*.

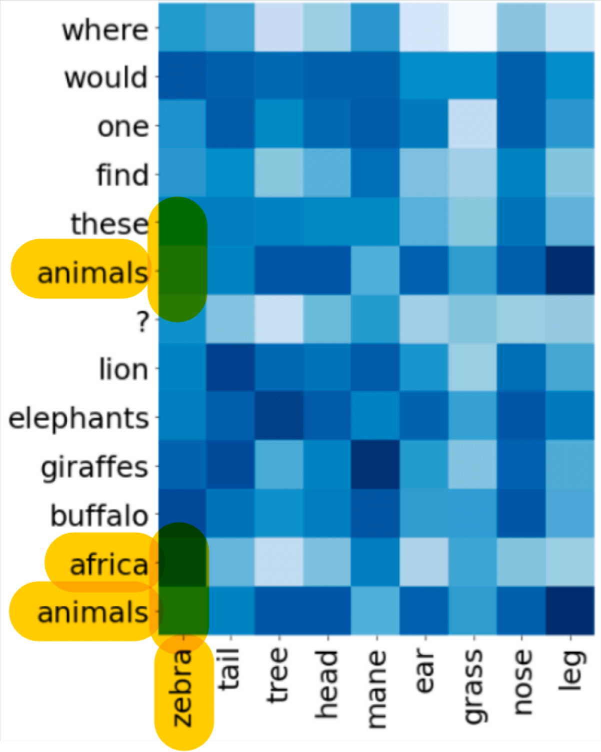
eXplainable Visual Question Answering using Knowledge Graphs (5)

Qualitative Results



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

XVQA: africa
ConceptBert: africa
Ground truth: africa



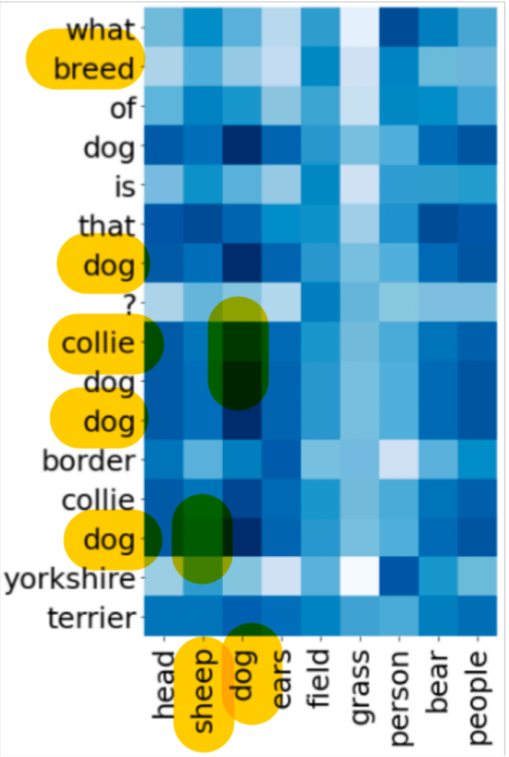
(2) Here the optimal transfer map is between (i) question tokens (vertical tokens: animals), graph entities (vertical tokens: africa, animals) and (ii) detected object (horizontal tokens: zebra) embeddings i.e., *africa has animals*.

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

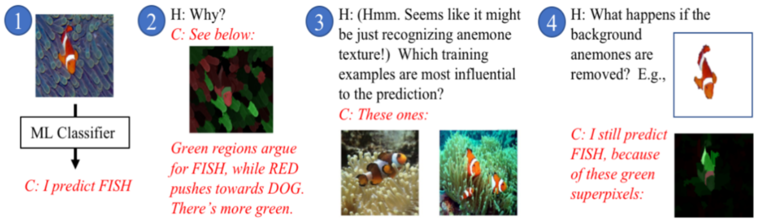
- **Retrieving explanations** of a VQA model during inference is a complex task
- Exposing articulated knowledge (i.e., **composition of knowledge graph triples**) to explain how an answer is related to the question, objects of the images and concepts is highly depending **on relevant retrieved knowledge**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

Conclusion

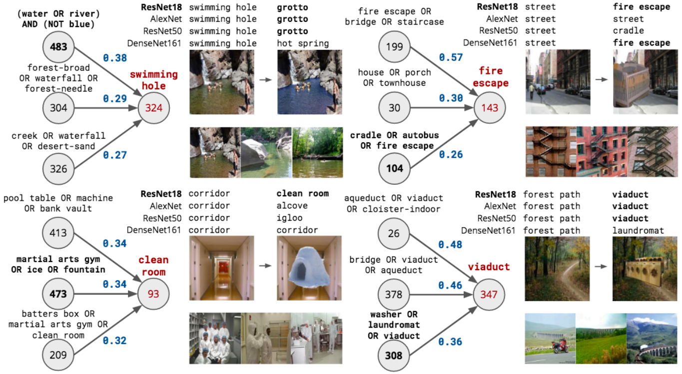
The Good: Multimodal End-to-End XAI System



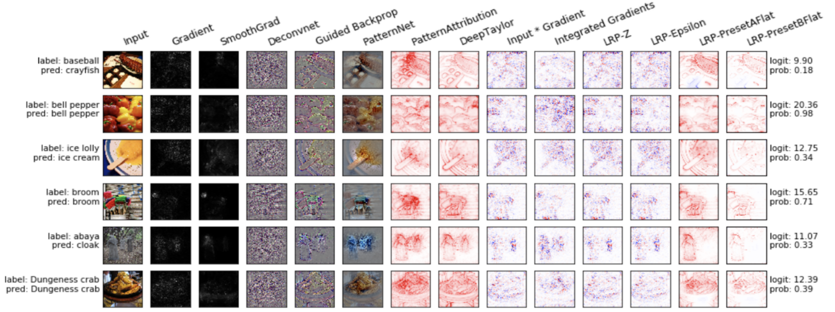
The Bad: Feature Visualization

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/y4wc2xj9

THALES

Inria
INVENTEURS DU MONDE NUMÉRIQUE