CLEVR-XAI: A Benchmark Dataset for the Ground Truth Evaluation of Neural Network Explanations

Leila Arras, Ahmed Osman, Wojciech Samek

MOTIVATION — Goal of e**X**plainable **A**rtificial **I**ntelligence (XAI)

Now: How to evaluate XAI **quantitatively**?

Post-hoc XAI methods:

- **LRP** (Bach et al. PLOS ONE 2015)
- **Integrated Gradients** (Sundararajan et al. ICML 2017)
- **Guided Backpropagation** (Springenberg et al. ICLR 2015)
- **Deconvnet** (Zeiler et al. ECCV 2014)
- **Grad-CAM** (Selvaraju et al. ICCV 2017)

amongst others

METHOD — Previous Works

Task: **Single** object classification on **real-world** images

Evaluation Methods:

Evaluation of XAI	Task (dataset)	Properties			
		synthetic	selective	same model	same data
Pixel Perturbation	object classification (ImageNet etc.)	✗	✗	✓	✗
Data Randomization Test		✗	✗	✗	✗
Model Randomization Test		✗	✗	✗	✓
Remove and Retrain		✗	✗	✗	✗
Object Localization/Segmentation		✗	✗	✓	✓
Pointing Game		✗	✗	✓	✓

Perturb input images according to pixels´ relevances
→ Track impact on the prediction

Pb: out-of-manifold input data

(i.a. Bach et al. PLOS ONE 2015, Samek et al. TNNLS 2017)

Modify training set with input perturbed according to pixels' relevances
→ Retrain model and eval performance

Pb: local minima, model fidelity

(Hooker et al. NeurIPS 2019)

Localization/Segmentation as a proxy task
→ Intersection over union, per pixel accuracy, pointing game accuracy

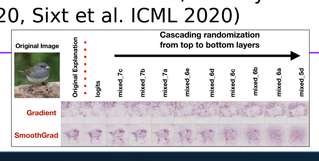
Pb: object context & biases ignored

(i.a. Simonyan et al. ICLR 2014)

Randomize data or model's layers
→ Track change in heatmap

Pb: out-of-manifold input data, "prior" of unmodified layers is ignored, metric for heatmap comparison is unclear

(i.a. Adebayo et al. NeurIPS 2018, Adebayo et al. NeurIPS 2020, Sixt et al. ICML 2020)



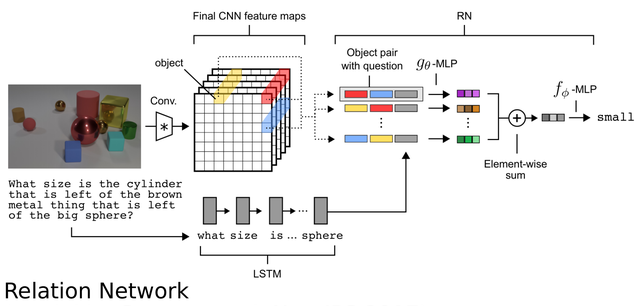
METHOD — Our Approach

Our Task: **Visual question answering** (VQA) on **synthetic** images

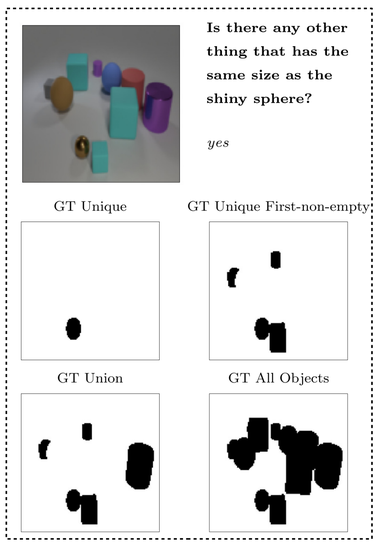
Our Dataset: We introduce **CLEVR-XAI** composed of

- **40k simple questions** with a single target object
- **100k complex questions** with multiple target objects **annotated with Ground Truth masks**

(derived from the CLEVR data generator by Johnson et al. CVPR 2017)



Relation Network
(model by Santoro et al. NeurIPS 2017)



Advantages:

- ✔ **selective** recognition task (multiple target objects per image, modulated by the question)
- ✔ **controlled** setup (synthetic images, unbiased and uninformative image background)
- ✔ **realistic** images (complex questions, objects occluded, shadow, various object attributes 2 sizes x 3 shapes x 2 materials x 8 colors, random locations, up to 10 objects per image)

Note: In our benchmarking we use a **simple** Relation Network to first **validate** XAI methods, but future work could use any other computer vision model that can solve the VQA task.

METHOD — Our Metrics

Mass Accuracy: $\frac{R_{within}}{R_{total}}$, where $R_{within} = \sum_{k=1,\, s.t.\ p_k \in Ground\ Truth}^{|Ground\ Truth|} R_{p_k}$ and $R_{total} = \sum_{k=1}^{N} R_{p_k}$

Rank Accuracy: $\frac{|P_{top\,K} \cap Ground\ Truth|}{|Ground\ Truth|}$, where $P_{top\,K} = \{p_1, p_2, ..., p_K | R_{p_1} > R_{p_2} > ... > R_{p_K}\}$

$|Ground\ Truth| = K$: size of the ground truth mask $\quad N$: size of the image

R_{p_k}: relevance of pixel p_k **after pooling** along channels (we use **10 pooling techniques**)

RESULTS — Comparison of XAI methods

Results using CLEVR-XAI simple:

Best mean performance: LRP and Integrated Gradients (+ LRP lowest variance)!

Increased XAI accuracy with model confidence

Increased XAI accuracy with target object size

Relevance Mass Accuracy	GT Single Object all		GT Single Object s.t. proba > 0.99999		GT Single Object s.t. nb pixels > 1000	
	Mean (std)	Median	Mean (std)	Median	Mean (std)	Median
LRP [20]	**0.85** (0.17)	**0.91**	**0.90** (0.09)	**0.93**	**0.91** (0.09)	**0.94**
Excitation Backprop [37]	0.80 (0.20)	0.87	0.85 (0.14)	0.90	0.88 (0.13)	0.92
IG [19]	0.71 (0.27)	0.81	0.75 (0.25)	0.85	0.80 (0.24)	0.90
Guided Backprop [22]	0.58 (0.20)	0.62	0.63 (0.16)	0.66	0.76 (0.13)	0.78
Guided Grad-CAM [39]	0.58 (0.24)	0.63	0.64 (0.21)	0.68	0.83 (0.13)	0.86
SmoothGrad [46]	0.60 (0.33)	0.69	0.61 (0.33)	0.72	0.80 (0.25)	0.91
VarGrad [34]	0.58 (0.34)	0.68	0.60 (0.34)	0.71	0.80 (0.25)	0.91
Gradient [16]	0.49 (0.35)	0.49	0.51 (0.34)	0.53	0.82 (0.25)	0.93
Gradient × Input [18]	0.43 (0.34)	0.37	0.44 (0.34)	0.39	0.77 (0.27)	0.89
Deconvnet [42]	0.18 (0.16)	0.13	0.18 (0.16)	0.13	0.42 (0.21)	0.41
Grad-CAM [39]	0.09 (0.10)	0.05	0.10 (0.10)	0.07	0.29 (0.10)	0.28

Poorest performance: Deconvnet and Grad-CAM!

Highest variance: all gradient-based XAI methods!

More results and discussion in our paper

SUMMARY — Outlook

- First objective ground truth and VQA based evaluation framework for XAI in computer vision
- Systematic comparison of 10+ popular XAI methods using novel quantitative metrics
- Surprising findings regarding the strengths and limitations of current XAI methods
- Dataset, code and framework are publicly available for future XAI research
- Open questions: Can an XAI method reach the 1.0 relevance accuracy upper bound? Do we need further model constraints to achieve this?

Leila Arras, Ahmed Osman, Wojciech Samek
CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations
In: Information Fusion, Vol. 81, Pages 14-40, 2022

Dataset & Code: github.com/ahmedmagdiosman/clevr-xai

Further information: www.heatmapping.org