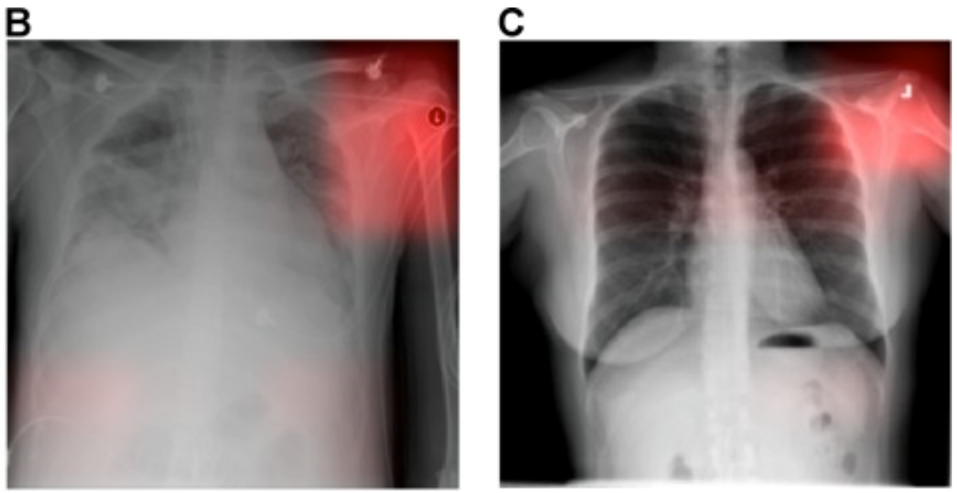
Explainable models and why we need them

Kate Saenko
Boston University

Why do we need explainable models?

CNNs learn to predict pneumonia by detecting hospital which took the image

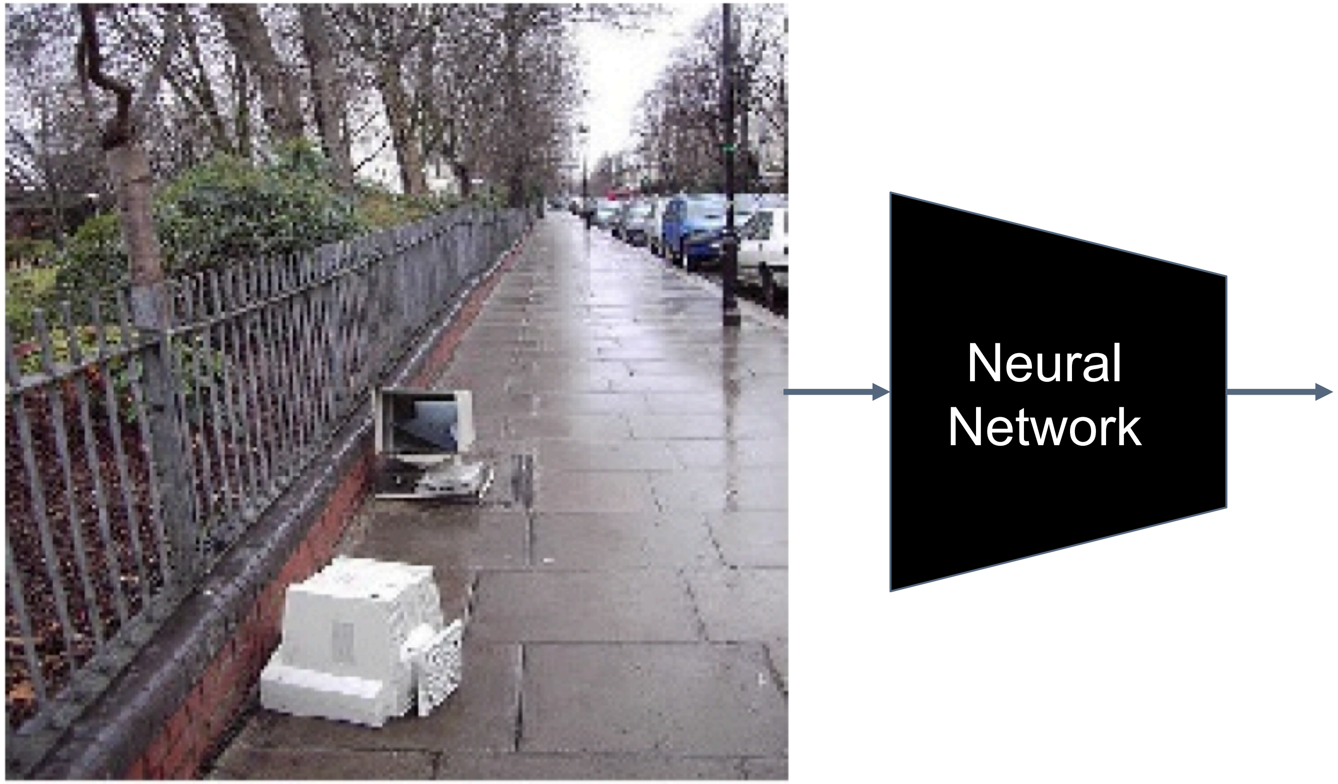
- Study on detecting pneumonia using 158,323 chest radiographs
- CNNs robustly identified hospital system and department within a hospital
- CNN has learned to detect a metal token that radiology technicians place on the patient in the corner of the image



Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study.
[Zech JR]1, [Badgeley MA]2, [Liu M]2, [Costa AB]3, [Titano JJ]4, [Oermann EK]3. https://www.ncbi.nlm.nih.gov/pubmed/30399157

Can we explain the network’s decision?

Why did the classifier predict “car”?



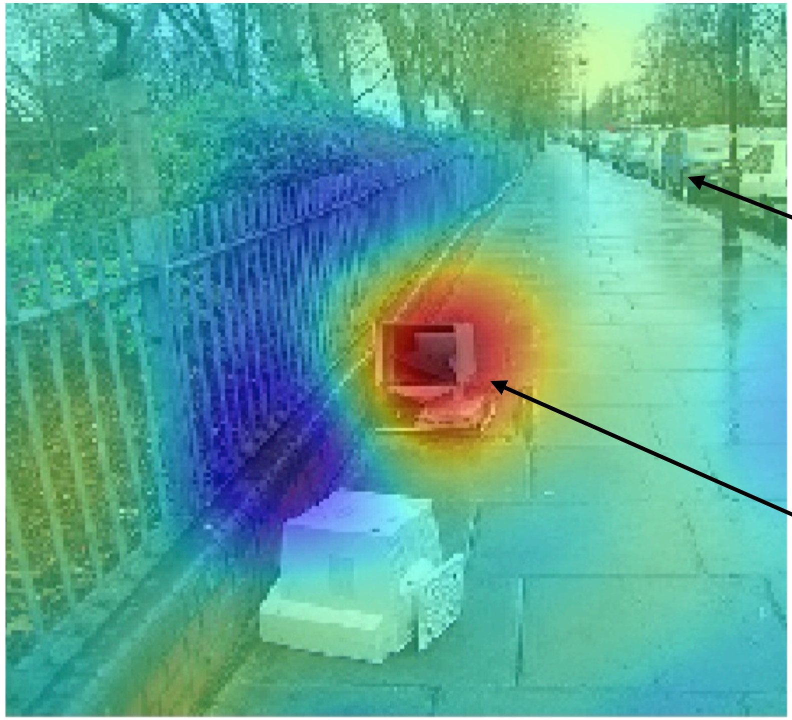
Prediction: “car” 64%

Saliency “explanation”

Why did the classifier predict “car”? –correct but for wrong reasons!

Prediction: “car” 64%

Explanation for “car”



Cars parked

actual reason is a monitor and street context

Saliency “explanation”

Why did the classifier predict “cow” (incorrect)?

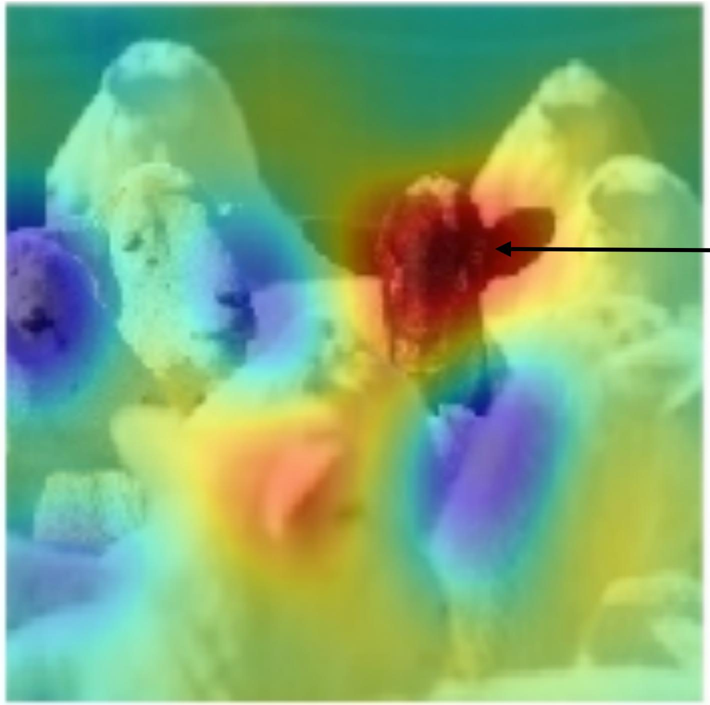
Prediction: “cow” 76%

Saliency "explanation"

Why did the classifier predict "cow" (incorrect)?

Prediction: "cow" 76%

Explanation for "cow"



MODEL BIAS:
most sheep are white, so model mistakes black sheep for cows

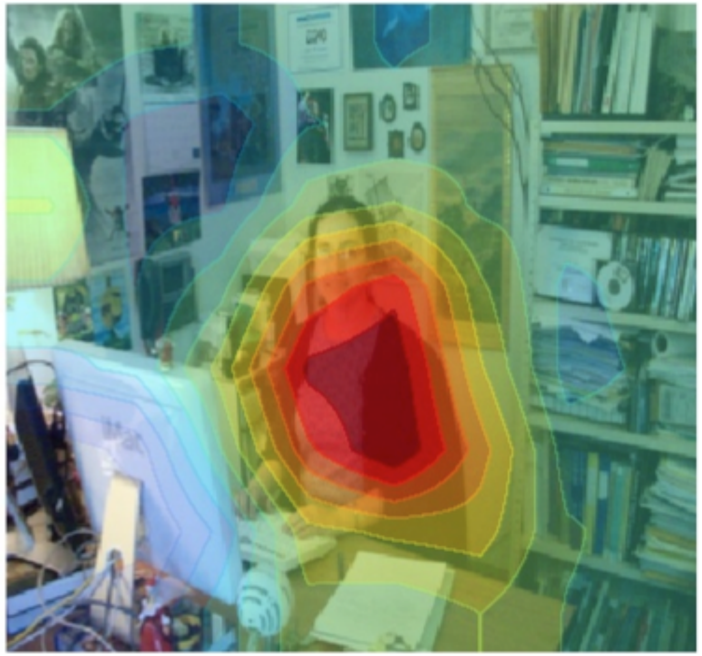
explainable AI

- explain prediction
- improve the model
- discover bias

Wrong

Baseline: *A **man** sitting at a desk with a laptop computer.*

Right for the Right Reasons

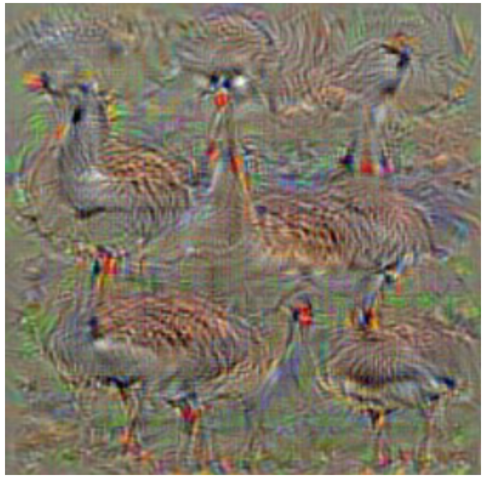


Our Model: *A **woman** sitting in front of a laptop computer.*

Burns, Hendricks, et. al, Women Also Snowboard: Overcoming Bias in Captioning Models, ECCV 2018

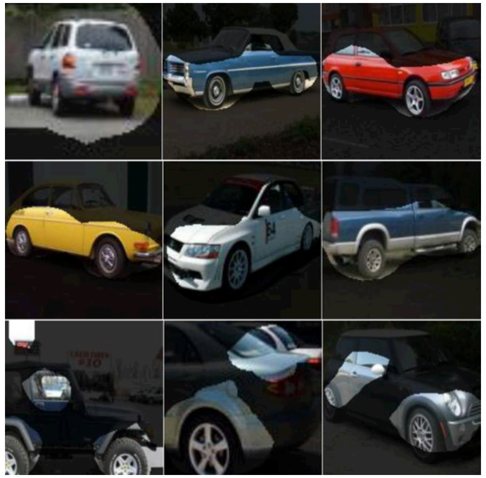
eXplainable Artificial Intelligence (XAI)

Activation Maximization	Generate iconic images	Network dissection	Interpretable models

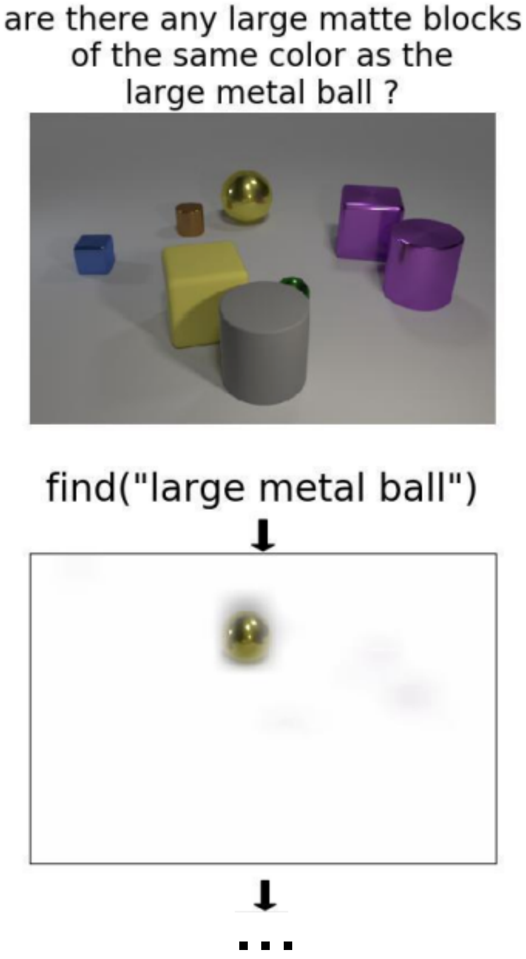


ostrich

volcano



car neuron



Simonyan et al. **Deep inside convolutional networks: Visualising image classification models and saliency maps.** ICLR Workshop 2014.

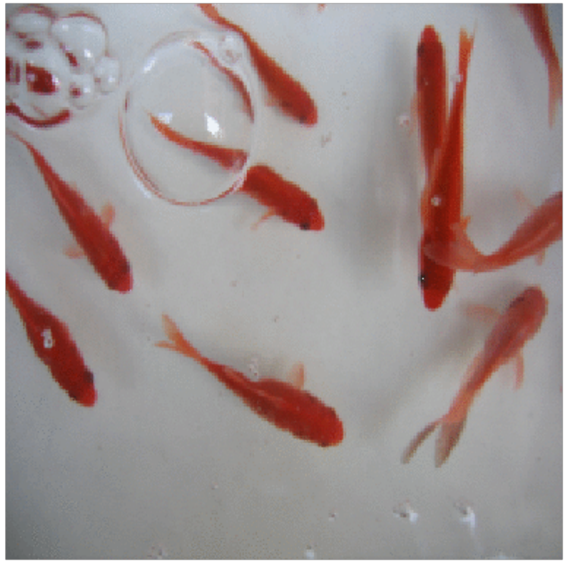
Nguyen et al. **Plug & play generative networks: Conditional iterative generation of images in latent space.** CVPR 2017

Bau et al. **Network Dissection: Quantifying Interpretability of Deep Visual Representations.** CVPR 2017

Hu et al. **Explainable Neural Computation via Stack Neural Module Networks.** ECCV 2018

Background: XAI via saliency detection

Image classification

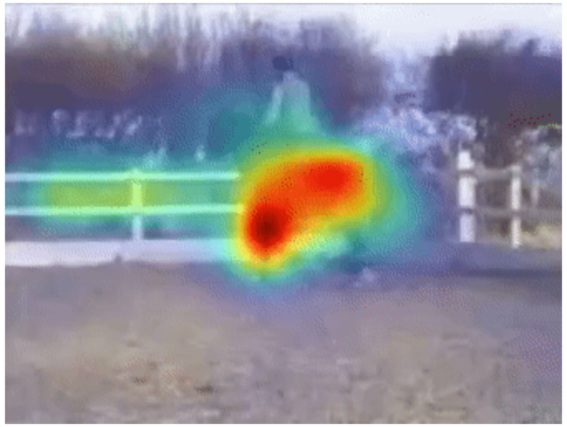


Goldfish

Image captioning

"A horse and a **carriage** on a city street."

Action recognition



Horse riding

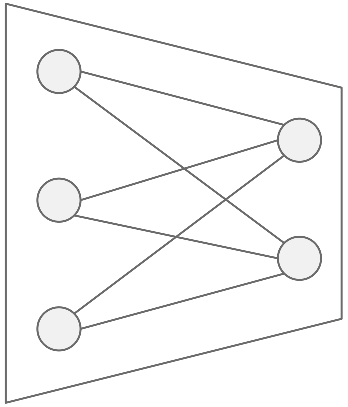
Sentiment analysis

“Despite its flaws, this is still a fascinating story.”

positive

Background: White box vs. Black box

White-box model



Excitation backprop

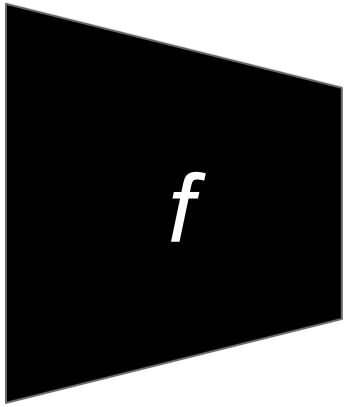
Grad-CAM

Meaningful Perturbations

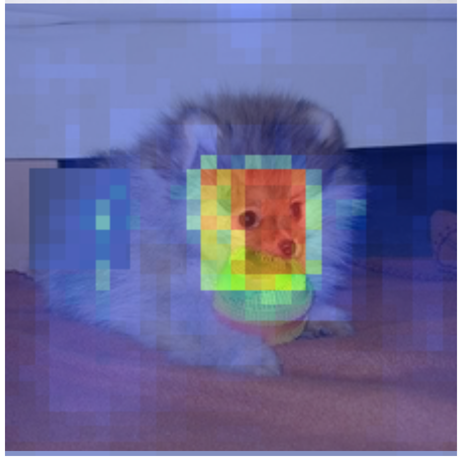
...

Many many works on this!

Black-box model



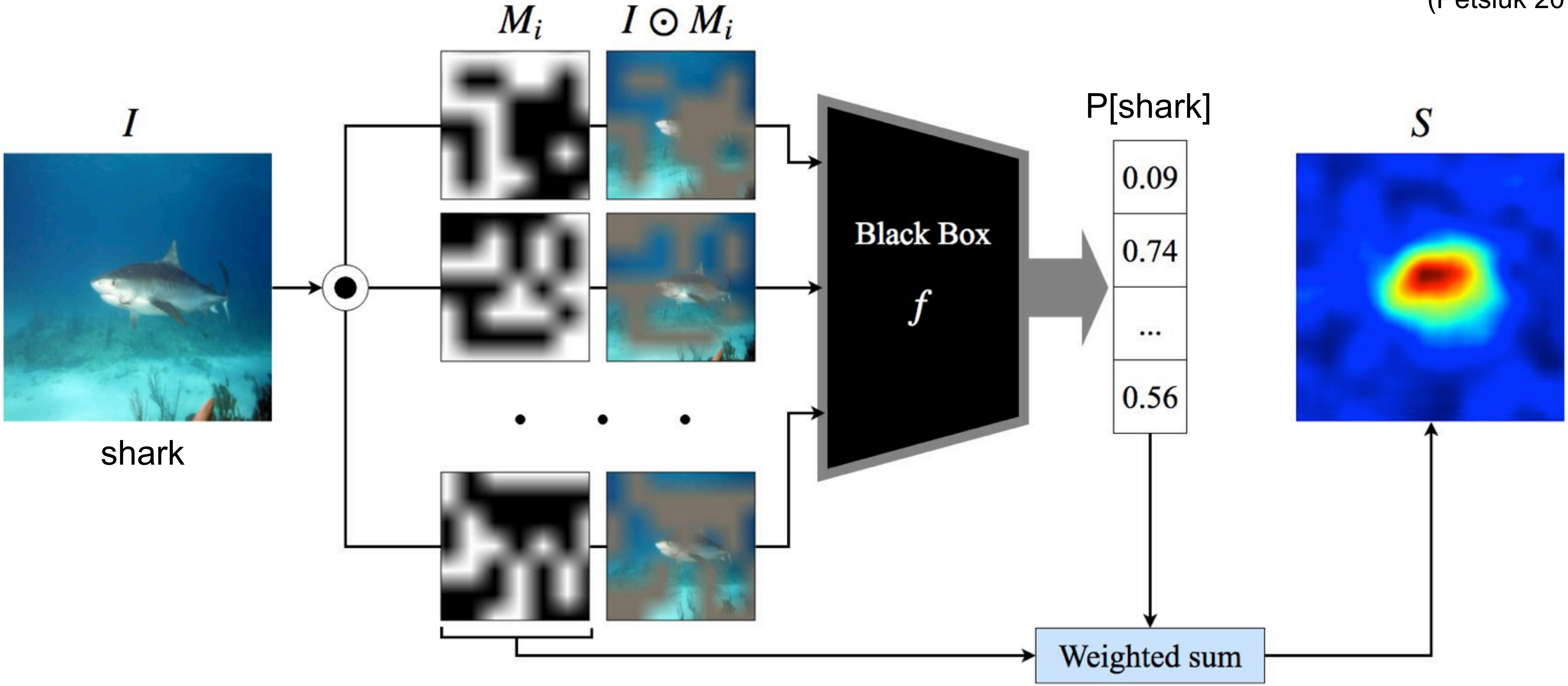
Sliding occlusion



LIME

RISE: randomly mask input, measure output

(Petsiuk 2018)



RISE: Randomized Input Sampling for Explanation of Black-box Models, Petsiuk, Das, Saenko, BMVC 2018

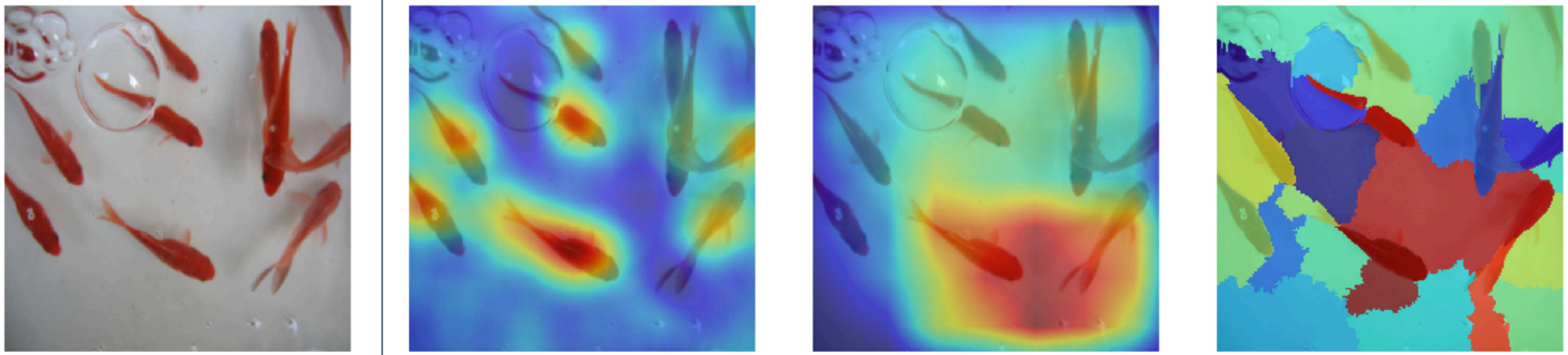
RISE: Qualitative Examples

(Petsiuk 2018)

- What the network actually sees, not what a human sees: “high-fidelity” explanation

pixels important for prediction

goldfish



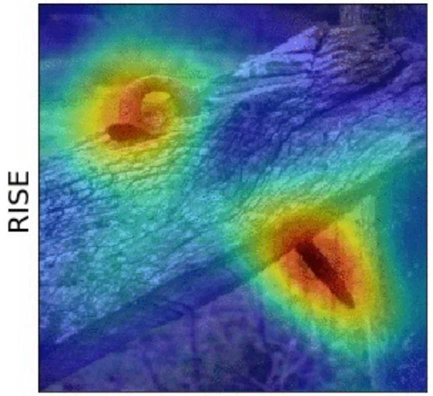
Ours | GradCAM | LIME

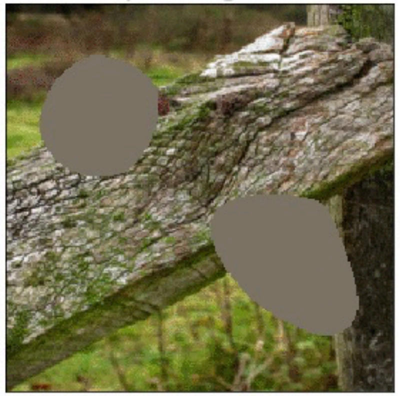
RISE: Randomized Input Sampling for Explanation of Black-box Models, Petsiuk, Das, Saenko, BMVC 2018

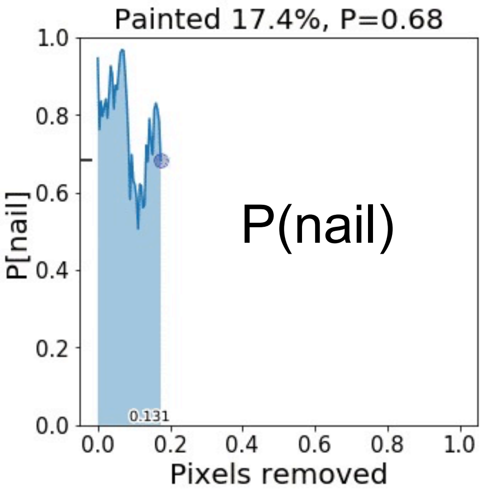
RISE: Evaluation

(Petsiuk 2018)

Explaining: nail





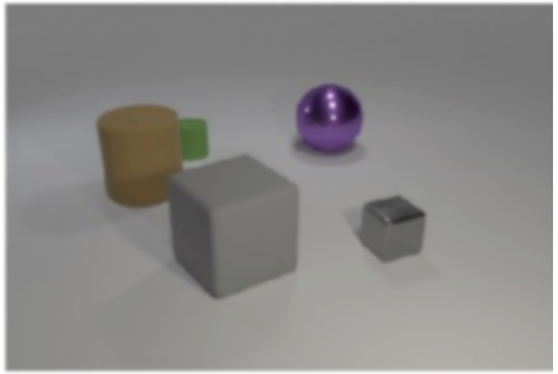


Method	ResNet50		VGG16	
	Deletion	Insertion	Deletion	Insertion
Grad-CAM [23]	0.1232	0.6766	0.1087	0.6149
Sliding window [31]	0.1421	0.6618	0.1158	0.5917
LIME [21]	0.1217	0.6940	0.1014	0.6167
RISE (ours)	**0.1076±0.0005**	**0.7267±0.0006**	**0.0980±0.0025**	**0.6663±0.0014**

Causal metrics on ImageNet dataset

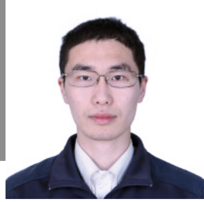
RISE: Randomized Input Sampling for Explanation of Black-box Models, Petsiuk, Das, Saenko, BMVC 2018

Can we go beyond a single heatmap?



Q: There is a small gray block; are there any spheres to the left of it?

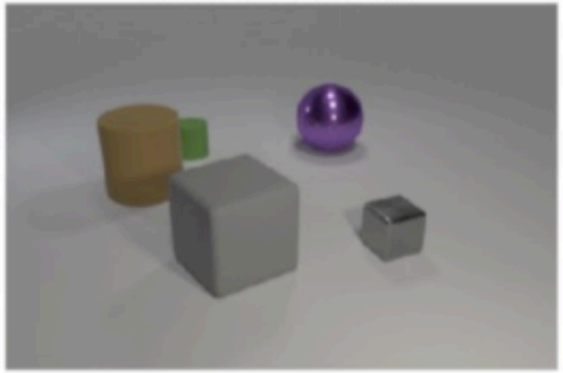
Neural modules learn a "program"



(Hu 2018)

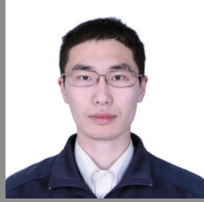
input: There is a small gray block; are there any spheres to the left of it?

input image



Hu, Andreas, Darrell, Saenko, Explainable Neural Computation via Stack Neural Module Networks, ECCV'18

Stack neural module networks (Hu et al. 2018)

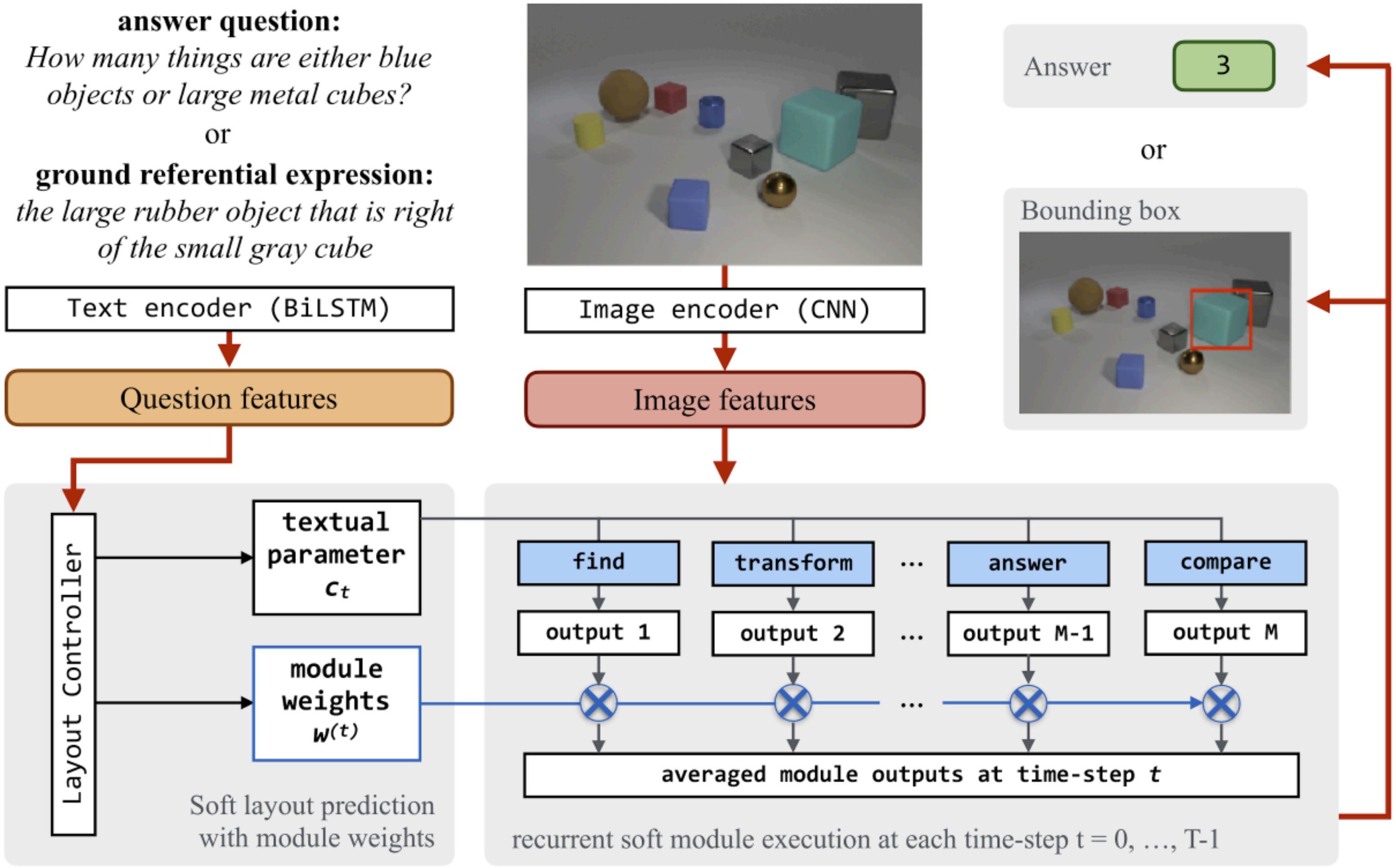


(Hu 2018)

Differentiable: replacing previous *discrete* execution graph with *continuous* soft layout (via module weights), not requiring "expert layout" supervision or RL.

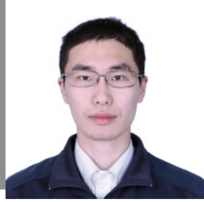
Interpretable as humans can understand its reasoning steps and detect its failure.

Multi-task by sharing a common set of sub-tasks (modules).



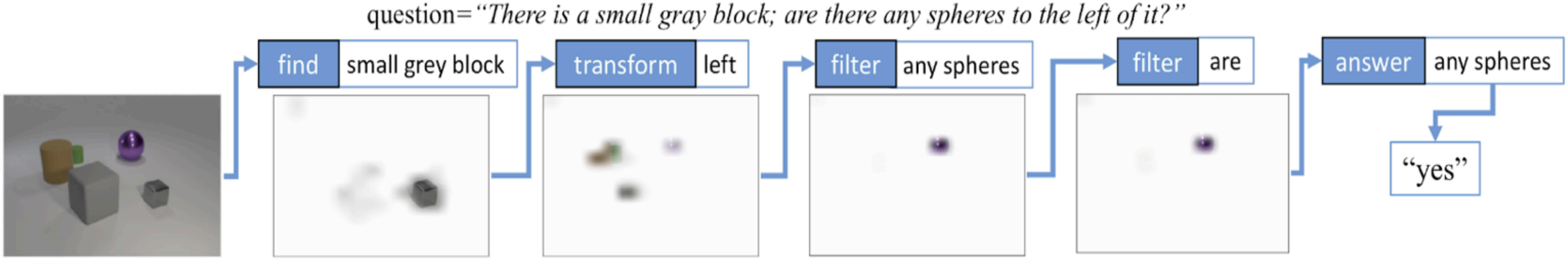
Hu, Andreas, Darrell, Saenko, Explainable Neural Computation via Stack Neural Module Networks, ECCV'18

Interpretability evaluation of NMNs (Hu et al. 2018)

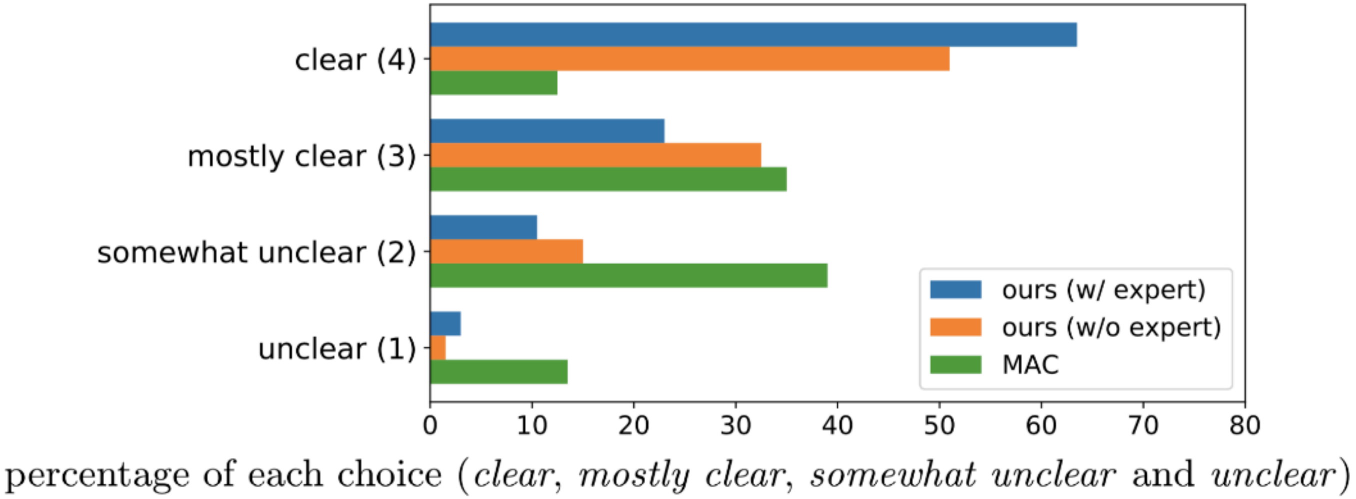


(Hu 2018)

We let human users judge (from the image and text attentions) whether the internal computation is clear to them. **Our model is much more often rated as "clear"**.



Question:
Are the internal reasoning steps above clear and understandable to you?



Can we explain similarity?

Why are these similar?

Both outdoors? Both are animals? Household pets?
Near/in a forest?

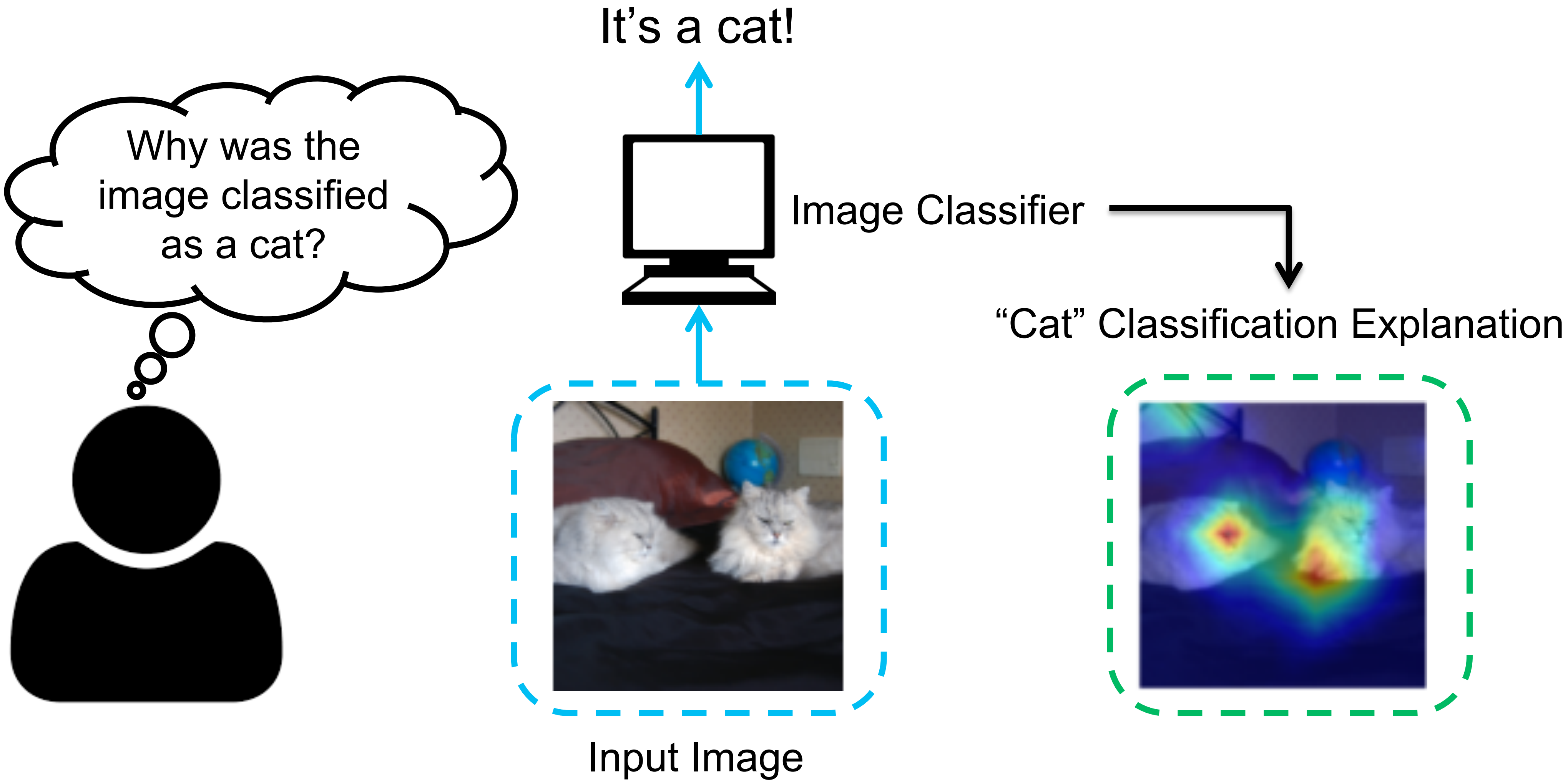
Explaining Decisions of Image Similarity Models, Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, David Forsyth, 2019

Why do these match?

Both are jewelry? Both gold? Both shiny or sparkly?

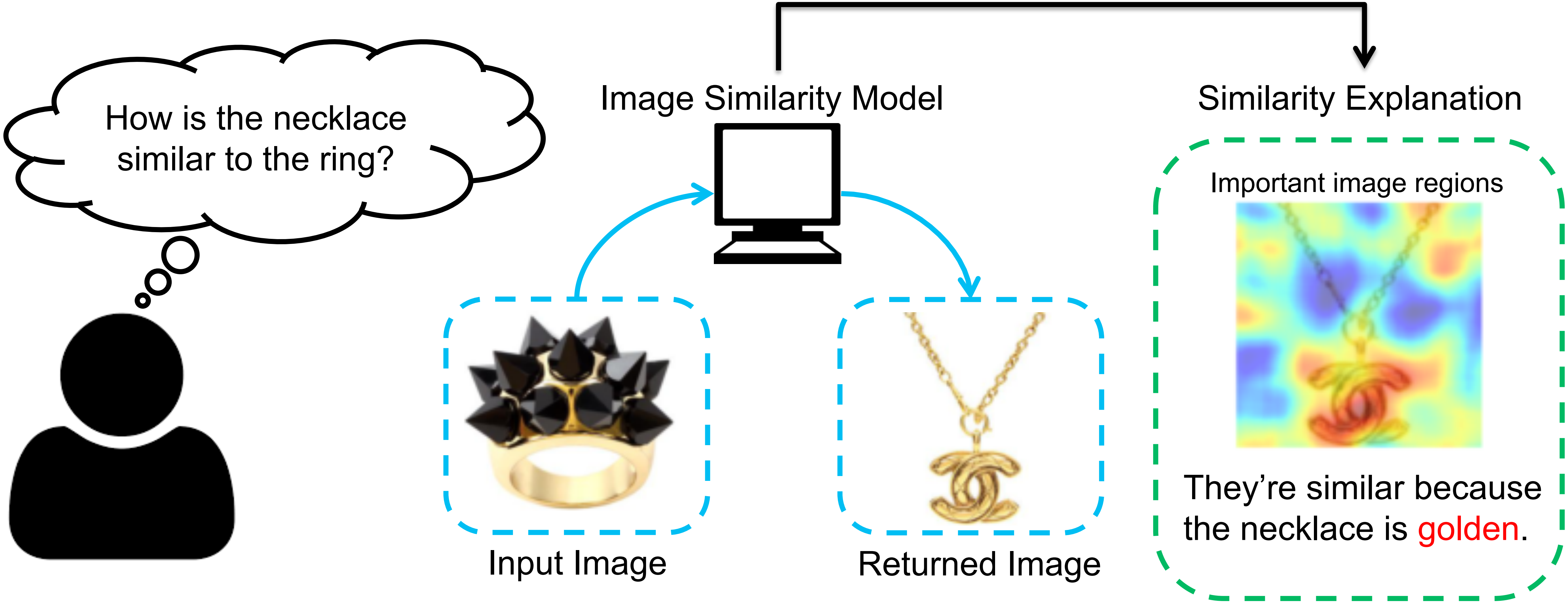
Explaining Decisions of Image Similarity Models, Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, David Forsyth, 2019

Prior work: explain classifier



This work: explain similarity model

(Plummer 19

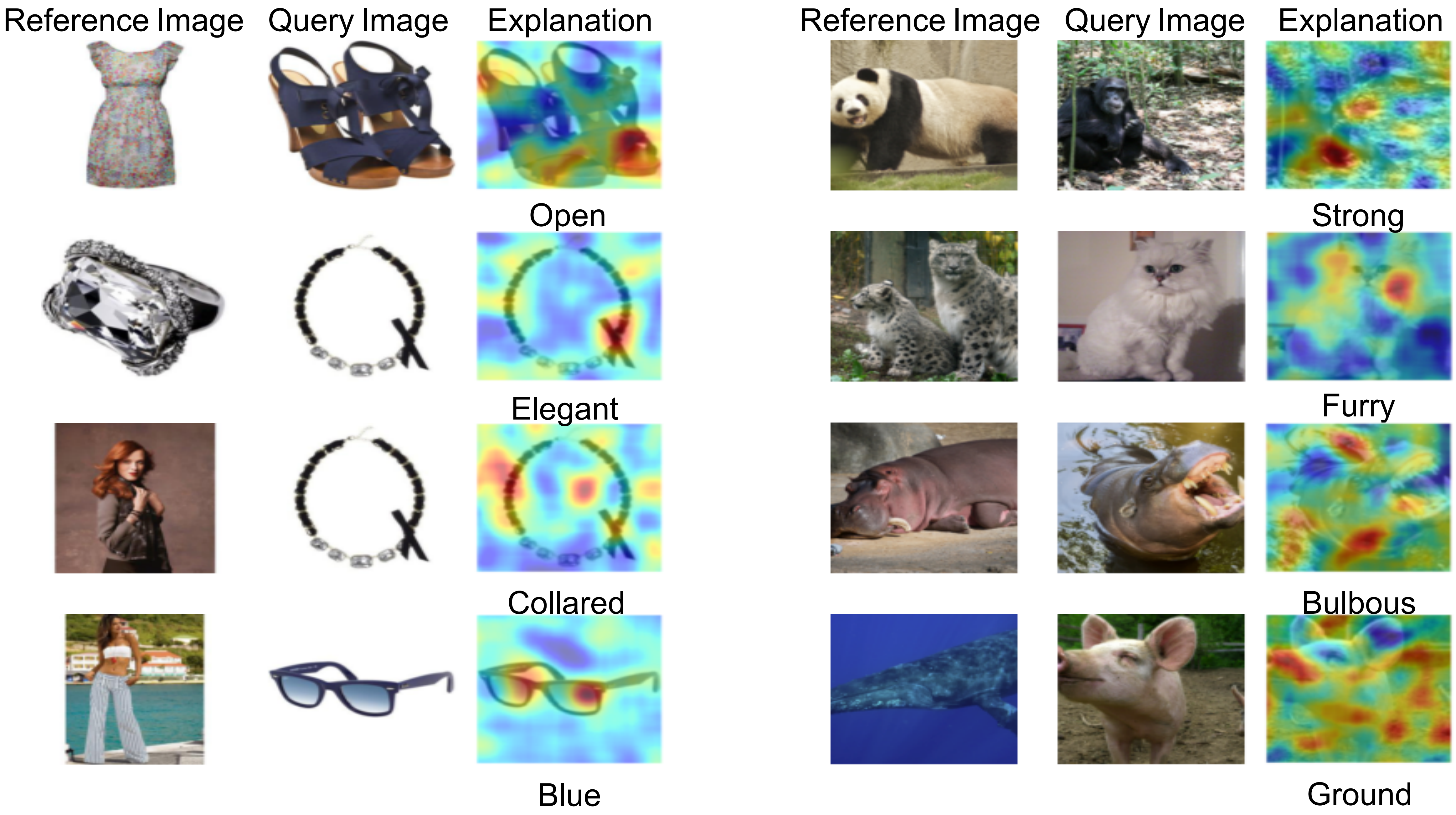


Explaining Decisions of Image Similarity Models, Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, David Forsyth, 2019

Desirable Qualities of Explanations

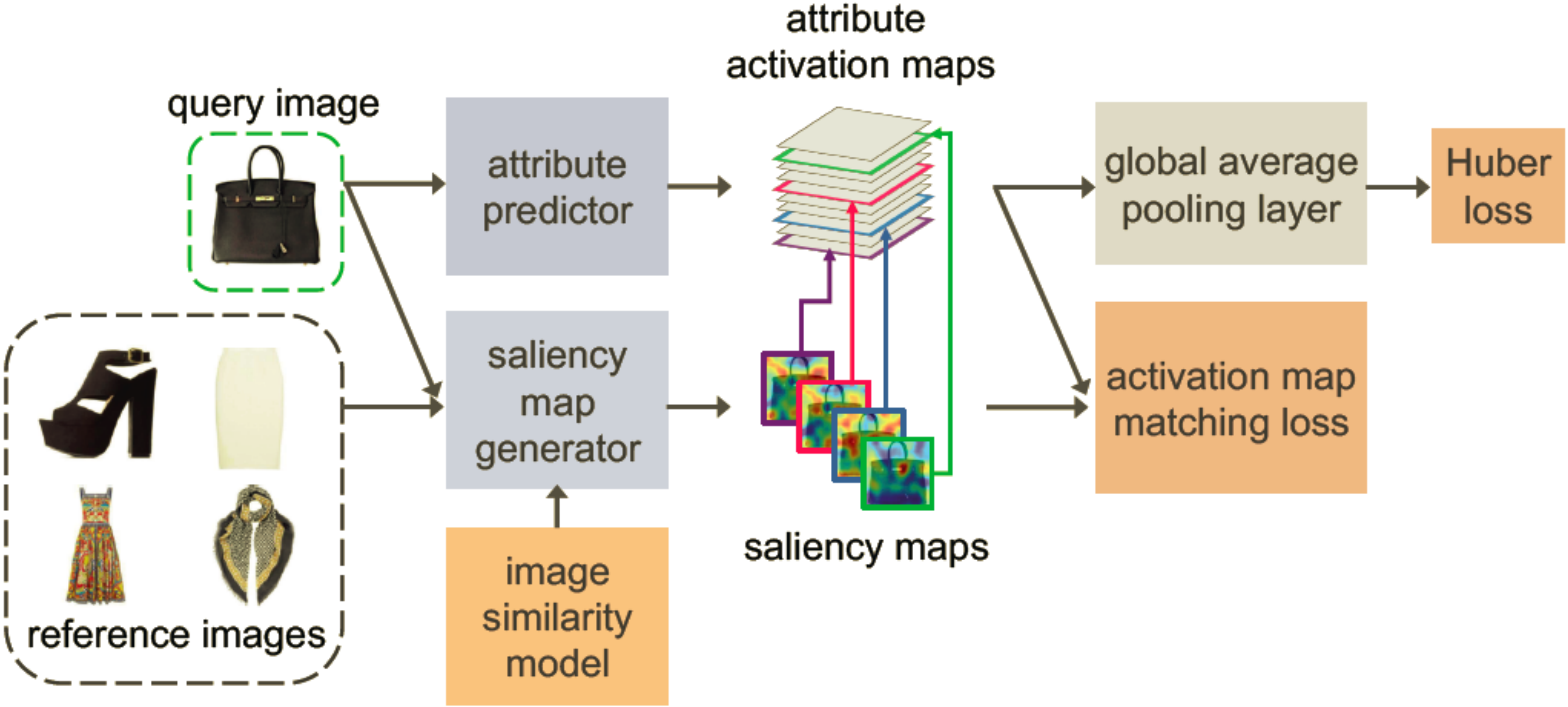
- Human interpretable
- Considers both images (i.e. changing one image affects the explanation of the other)
- Explains model behavior

Explaining Decisions of Image Similarity Models, Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, David Forsyth, 2019



Explaining Decisions of Image Similarity Models, Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, David Forsyth, 2019

SANE: Attribute-based explanation model

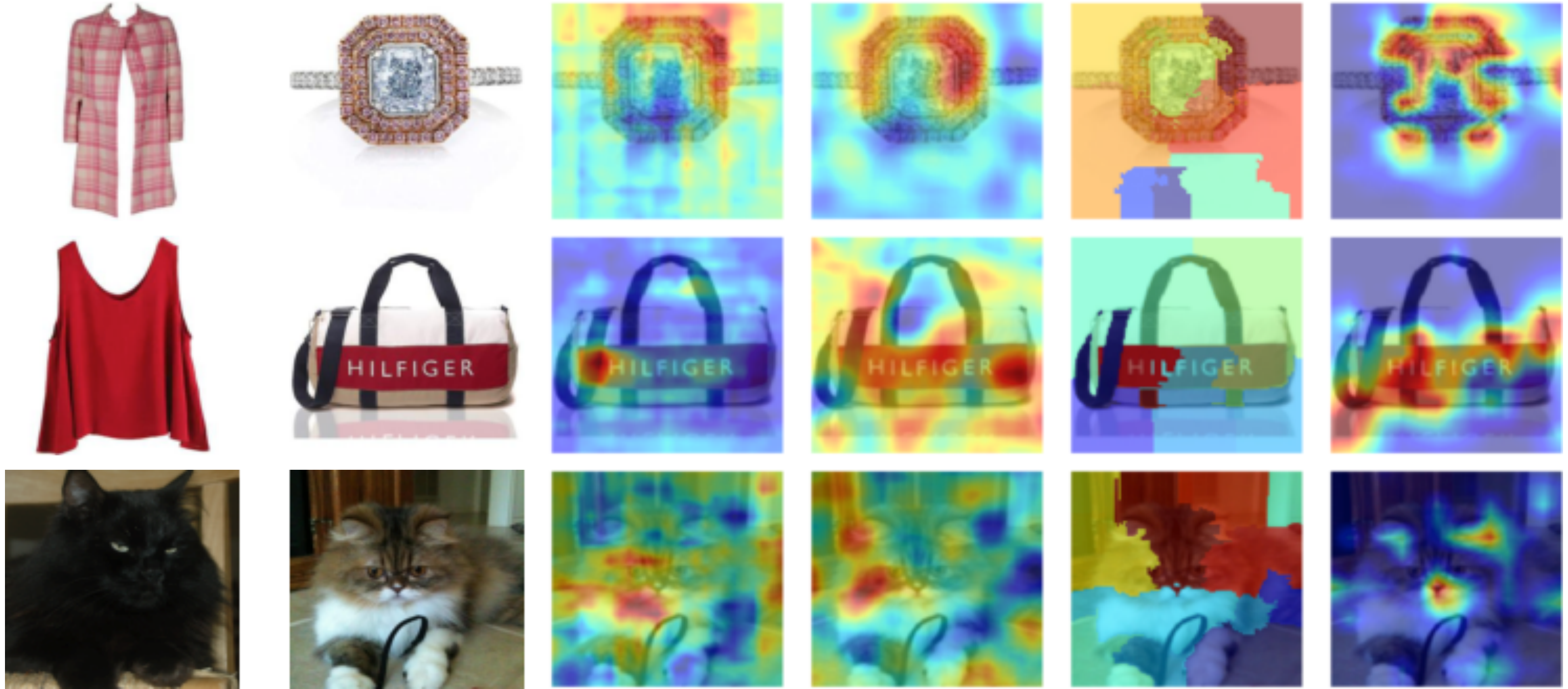


Explaining Decisions of Image Similarity Models, Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, David Forsyth, 2019

Saliency Map Performance

(Plummer 19)

		Polyvore Outfits		Animals with Attributes 2	
Method	Fixed Reference?	Insertion (↑)	Deletion (↓)	Insertion (↑)	Deletion (↓)
Sliding Window	Y	60.2	53.6	76.9	76.8
RISE	Y	**62.0**	**52.0**	76.5	77.1
LIME	Y	58.4	55.4	77.0	**71.2**
Mask	Y	59.4	53.3	74.5	77.3



Attribute "removal" metric

(Plummer 19)

Input Image	Attribute to Remove	Returned Image	
	"studded"		measure drop in similarity
	"lace"		measure drop in similarity

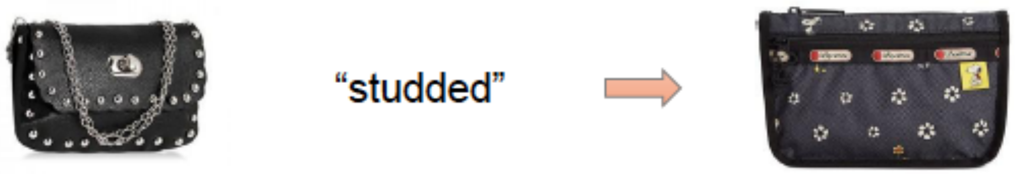
Explaining Decisions of Image Similarity Models, Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, David Forsyth, 2019

Attribute “removal” evaluation

(Plummer 19)

Method	Polyvore Outfits mAP	Polyvore Outfits Top1 Accuracy	Polyvore Outfits Attr Removal	Animals with Attributes 2 mAP	Animals with Attributes 2 Top1 Accuracy	Animals with Attributes 2 Attr Removal
Random	–	1.3	0.2	–	38.1	0.4
Attribute Classifier	24.2	49.1	0.5	66.5	73.9	0.9
FashionSearchNet [1]	24.5	49.1	0.4	66.7	75.2	1.1
FashionSearchNet + Map Matching	–	49.8	1.5	–	77.8	1.4
SANE	**25.7**	50.0	2.2	**67.1**	77.1	1.8
SANE + Map Matching	–	51.7	2.9	–	**85.5**	2.3
SANE + Map Matching + Prior (Full)	–	**52.2**	**3.5**	–	85.1	**2.7**

Input Image Attribute to Remove Returned Image



measure drop
in similarity

Explaining Decisions of Image Similarity Models, Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, David Forsyth, 2019

Summary

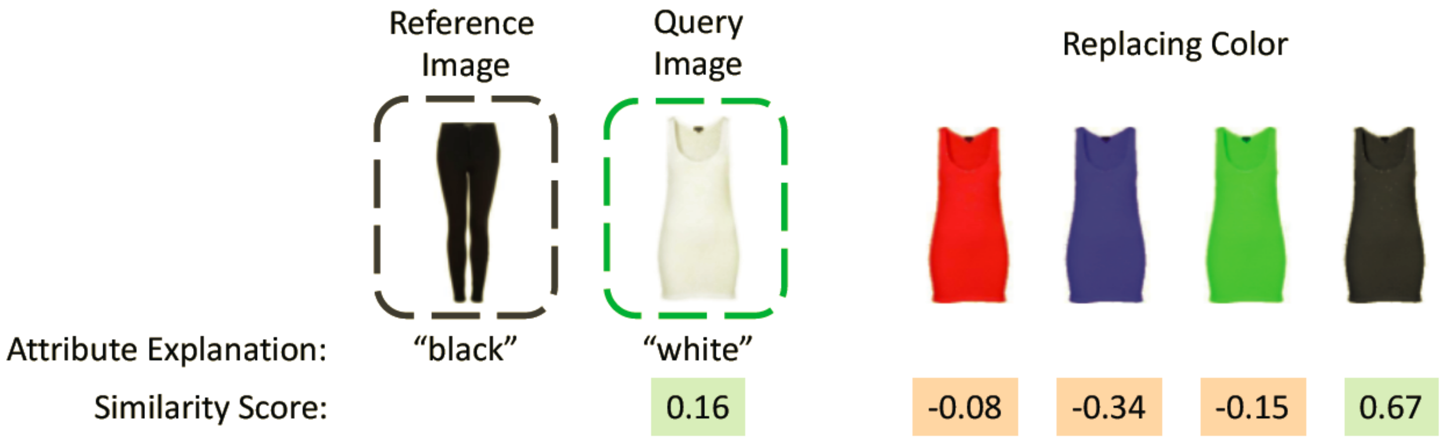
- A causal saliency explanation model (RISE)
- Naturally explainable modular networks
- Explaining a similarity model with attributes

Where to go next?

- need evaluation metrics for XAI
- disentangled representations

Datasets

- Polyvore Outfits - 365,054 images, 205 attributes
- Animals with Attributes 2 - 37,322 images, 50 animal classes, 85 attributes



Explaining Decisions of Image Similarity Models, Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, David Forsyth, 2019