Interpretability (XAI) Part 1

Presenter: Zhengrui Guo
Contact: zguobc@connect.ust.hk

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

What is Interpretability

And

Why it matters

Interpretability (XAI): Introduction

The transparency and ability to explain is useful at three different stages of Artificial Intelligence (AI) evolution :

- First, when AI is significantly weaker than humans and not yet reliably deployable
- Second, when AI is on par with humans and reliably deployable
- Third, when AI is significantly stronger than humans

Interpretability (XAI): Introduction

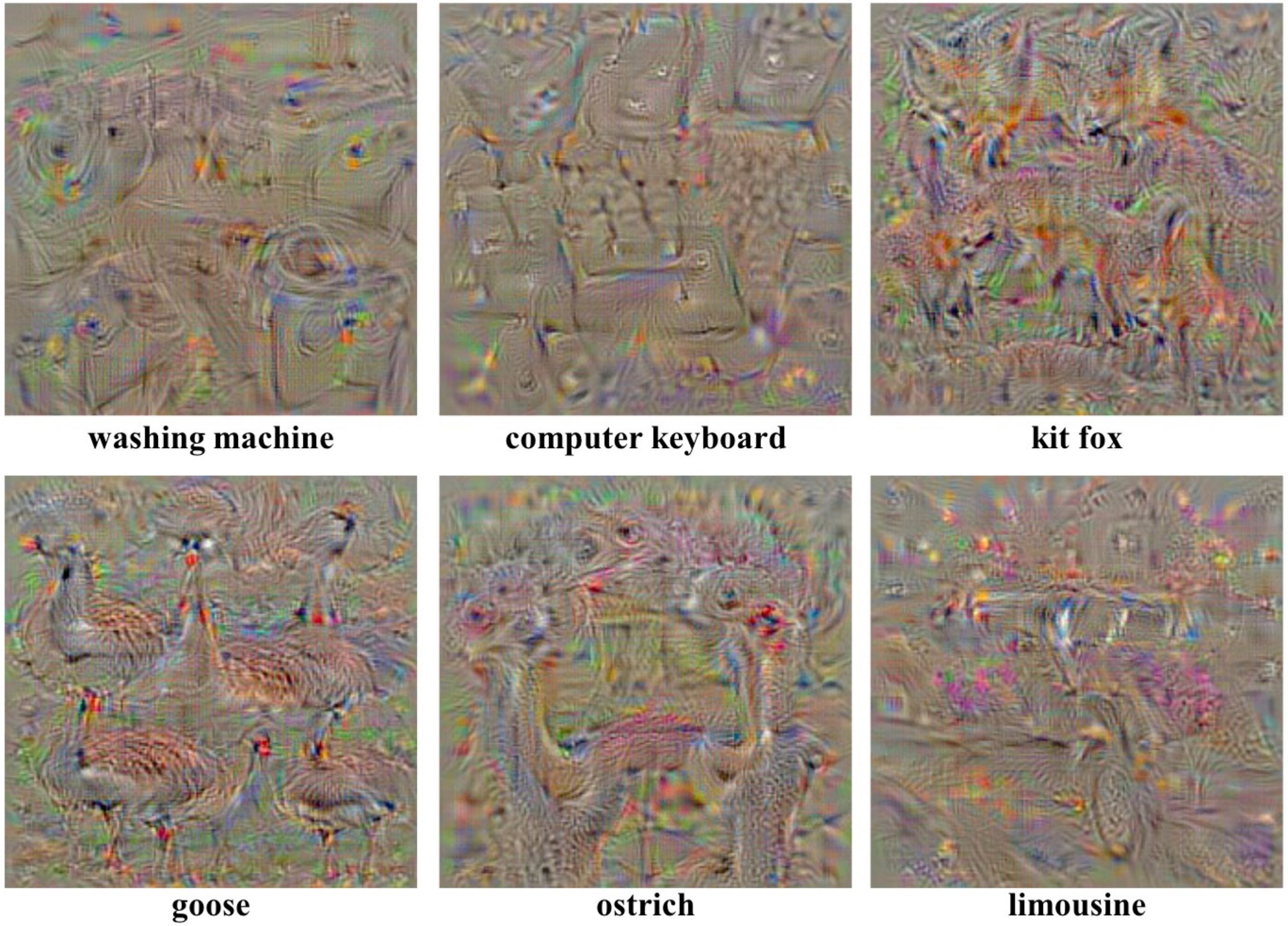
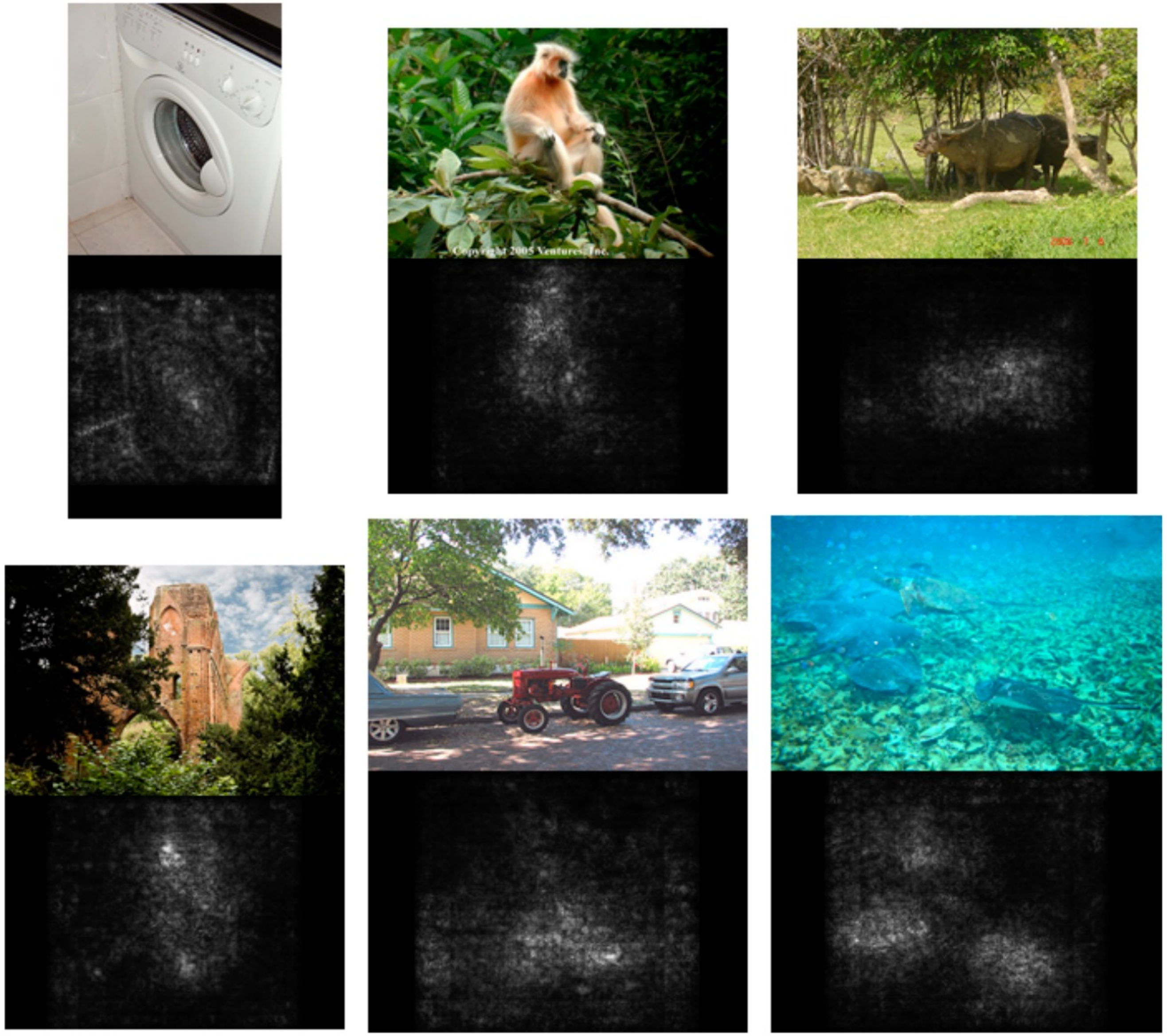


Figure 1: **Numerically computed images, illustrating the class appearance models, learnt by a ConvNet, trained on ILSVRC-2013.** Note how different aspects of class appearance are captured in a single image. Better viewed in colour.

Interpretability (XAI): Introduction



Interpretability (XAI): Introduction

(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

Saliency Maps

Versus

Class Activation Maps

Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

Contributions:

- Use the numerical optimization of the input image to obtain the understandable visualizations of CNN classification models;
- They propose a method for computing the spatial support of a given class in a given image (image-specific class saliency map) using a single back-propagation pass through a classification CNN;
- They apply the generated saliency maps to weakly supervised object localization.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

CNN implementation details:

- Conv64 - Conv256 - Conv256 - Conv256 - Conv256 - Full4096 - Full4096 - Full1000
- Trained on ImageNet with 1.2M training images, labelled into 1000 classes.
- On ImageNet validation set, the network achieves the top-1/top-5 classification error of 39.7%/17.7%.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

Method 1: Class Model Visualization:

- Let $S_c(I)$ be the score of the class c, computed by the classification layer of the CNN for an image I.
- Find an L_2-regularized image such that the score S_c is high:

$$\mathrm{argmax}_I S_c(I) - \lambda\left||I|\right|_2^2$$

- Fixing the parameters of CNN, a local-optimal image I can be found by back propagation. (The optimization is performed w.r.t the input image)

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

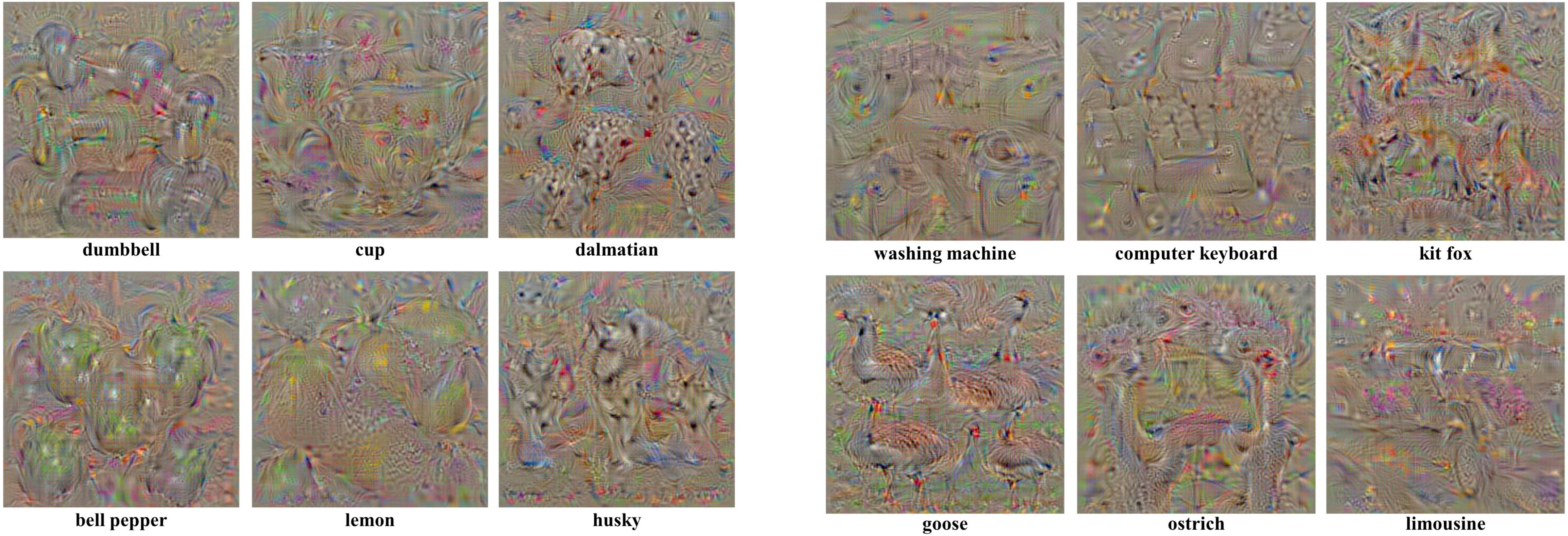


Figure 1: **Numerically computed images, illustrating the class appearance models, learnt by a ConvNet, trained on ILSVRC-2013.** Note how different aspects of class appearance are captured in a single image. Better viewed in colour.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

Method 1: Class Model Visualization:

- Note that the unnormalized class scores $S_c(I)$ is used, rather than the class posteriors returned by the soft-max layer:

$$P_c = \frac{\exp S_c}{\sum_c \exp S_c}$$

- They argue that the maximization of the class posterior can be achieved by minimizing the scores of other classes.
- Therefore, they optimize $S_c(I)$ to ensure that the optimization concentrates only on the class c.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

Method 2: Image-specific class saliency visualization:

- Given an image I_0, a class c, and a classification CNN with the class score function $S_c(I)$, the goal is to rank the pixels of I_0 based on their influence on the score $S_c(I_0)$.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

Method 2: Image-specific class saliency visualization:

- Given an image I_0, a class c, and a classification CNN with the class score function $S_c(I)$, the goal is to rank the pixels of I_0 based on their influence on the score $S_c(I_0)$.

A motivational Example:

- Consider the linear score model for the class c:

$$S_c(I) = w_c^T I + b_c$$

- In this case, it is easy to see that the magnitude of elements of w defines the importance of the corresponding pixels of I for the class c.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Method 2: Image-specific class saliency visualization:

- Given an image I_0, a class c, and a classification CNN with the class score function $S_c(I)$, the goal is to rank the pixels of I_0 based on their influence on the score $S_c(I_0)$.

A motivational Example:

- While for CNN, $S_c(I)$ is highly non-linear.
- Given an image I_0, one can approximate $S_c(I)$ with a linear function in the neighborhood of I_0 by computing the first-order Taylor expansion:

$$S_c(I) \approx w^T I + b$$

with $w = \frac{\partial S_c}{\partial I}\Big|_{I_0}$

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

Method 2: Image-specific class saliency visualization:

- Given an image I_0 with m rows and n columns, and a class c, the class saliency map $M \in R^{m \times n}$ is computed as follows:
  1. Obtain the derivative $w = \frac{\partial S_c}{\partial I}|_{I_0}$ by backpropagation
  2. Rearrange the elements of the vector w to obtain the saliency map:
     - For grey-scale image, the map is computed as $M_{ij} = |w_{h(i,j)}|$, in which $h(i,j)$ is the index of the element w, corresponding to the image pixel in the i-th row and j-th column.
     - For multi-channel image, the map is computed as $M_{ij} = \max_c |w_{h(i,j,c)}|$, in which $h(i,j,c)$ is the index of the element w, corresponding to the image pixel in the i-th row, j-th column, and c-th channel.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

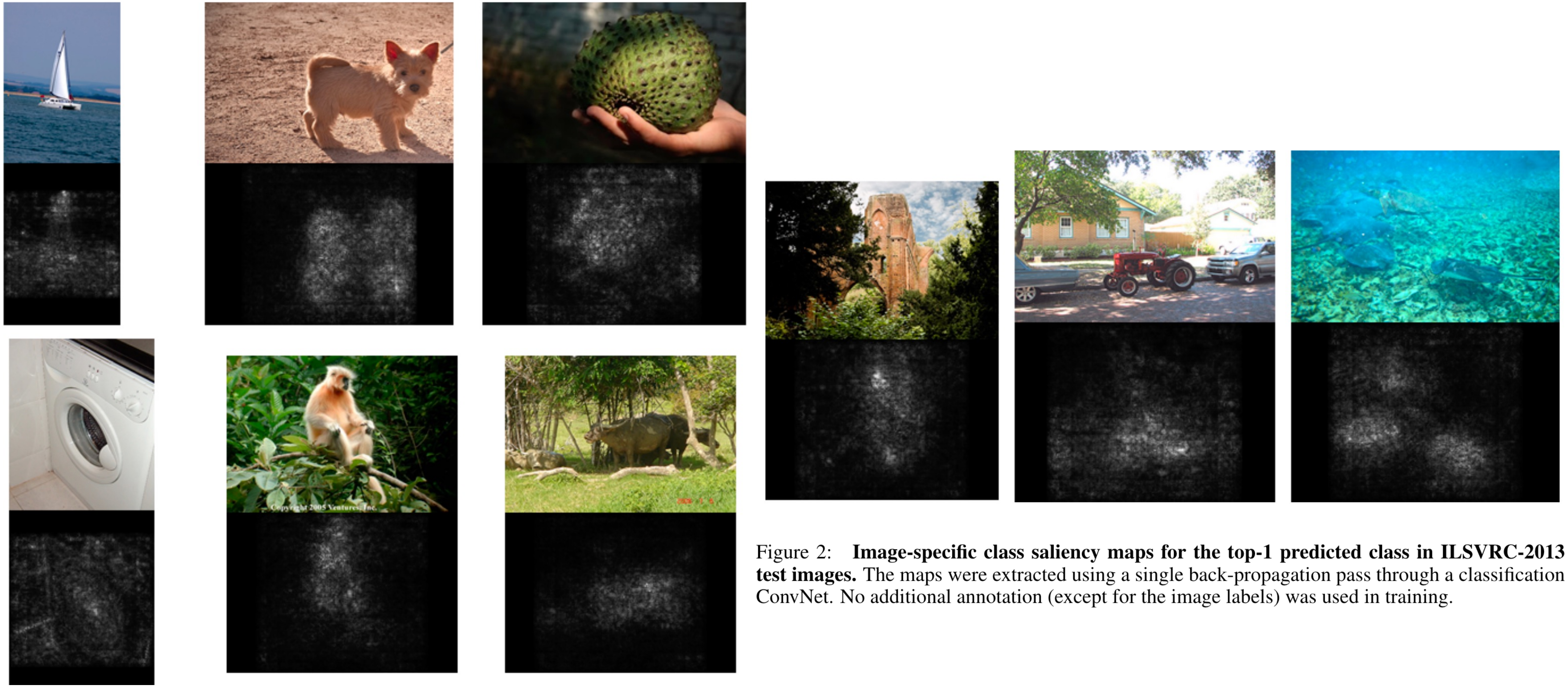


Figure 2: **Image-specific class saliency maps for the top-1 predicted class in ILSVRC-2013 test images.** The maps were extracted using a single back-propagation pass through a classification ConvNet. No additional annotation (except for the image labels) was used in training.

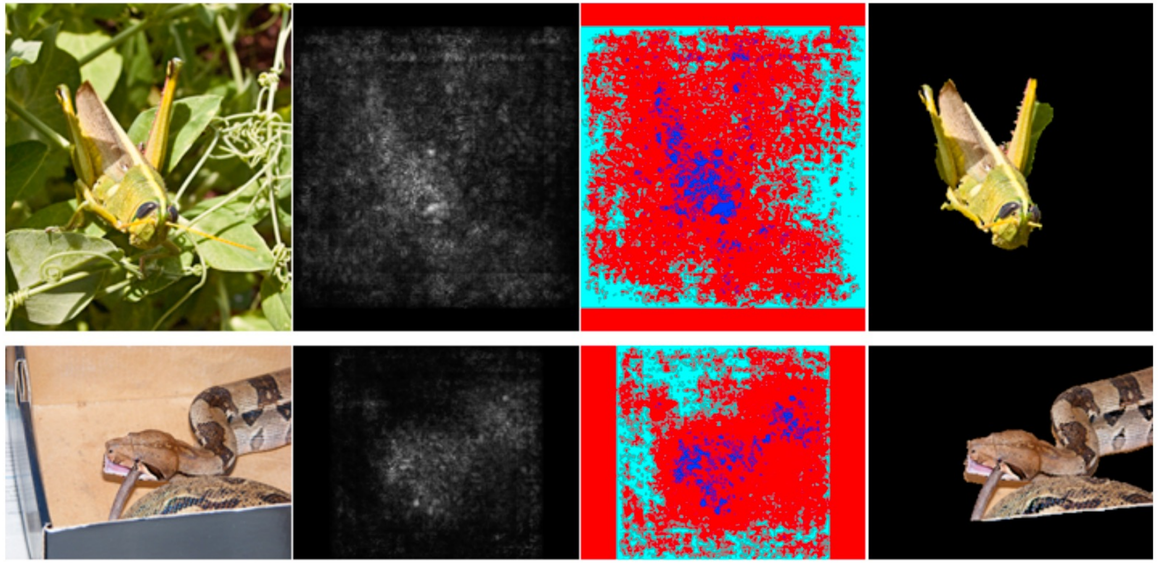
Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

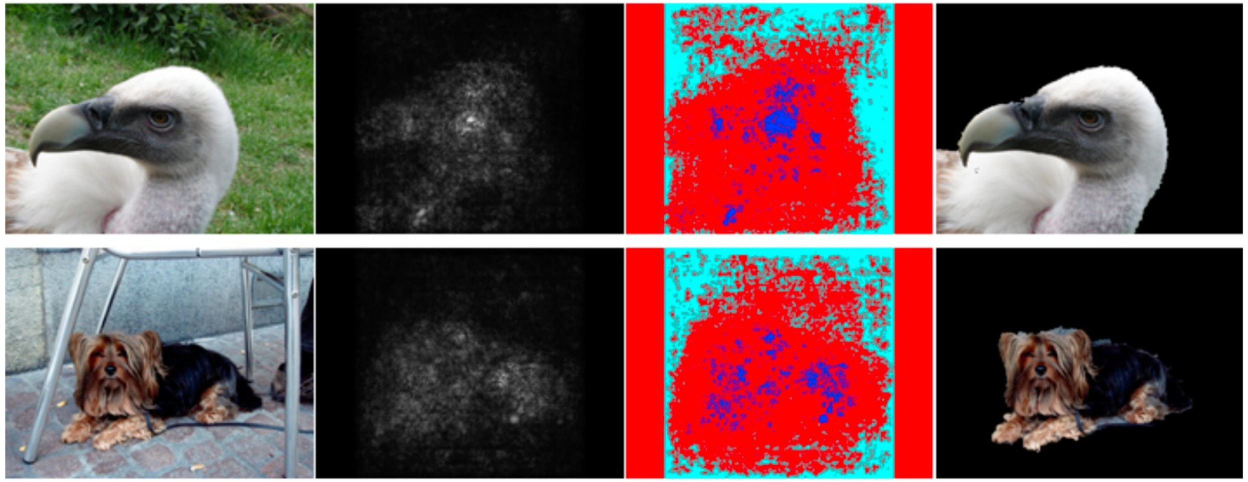
Interpretability (XAI): Saliency Maps-based method

Method 2: Image-specific class saliency visualization:

Weakly Supervised Object Localization:

- These class saliency maps can be used for object localization (in spite of being trained on image labels only)





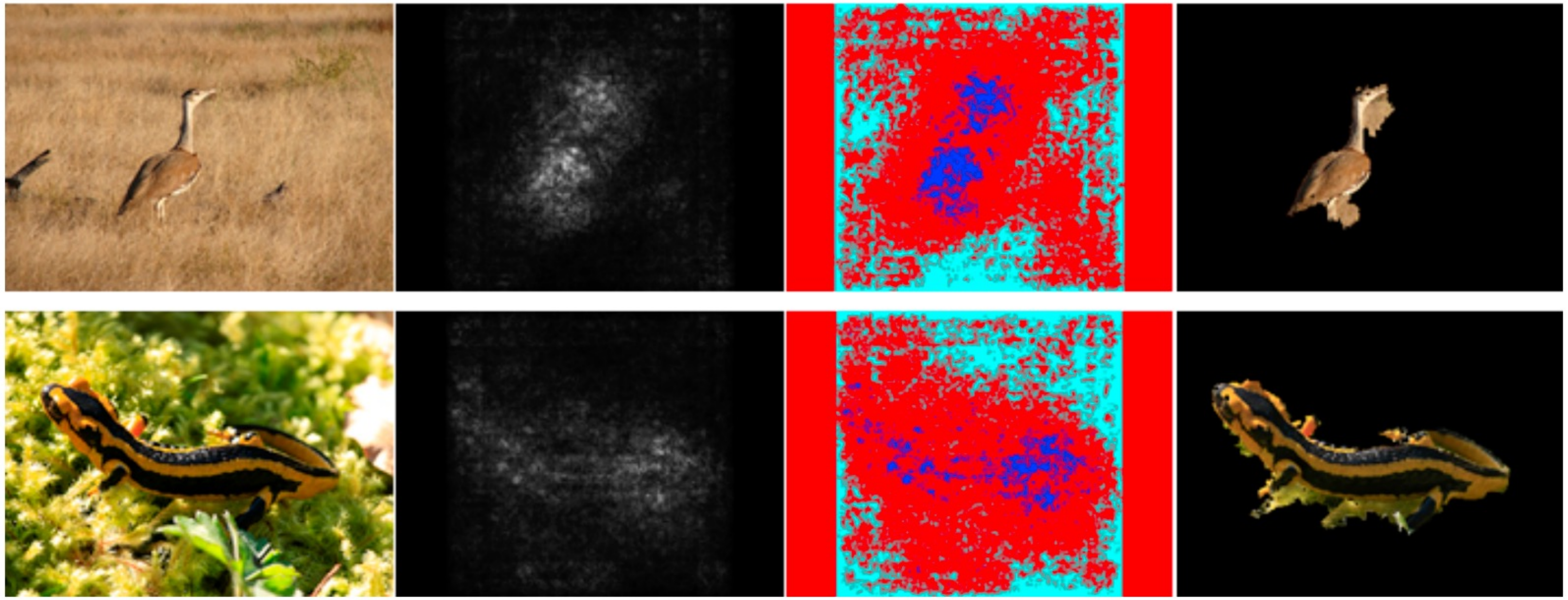
Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Interpretability (XAI): Saliency Maps-based method

Method 2: Image-specific class saliency visualization:

Weakly Supervised Object Localization:

- These class saliency maps can be used for object localization (in spite of being trained on image labels only)



Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Class Activation Maps (CAM)

Interpretability (XAI): CAM-based method

Introduction to CAM:

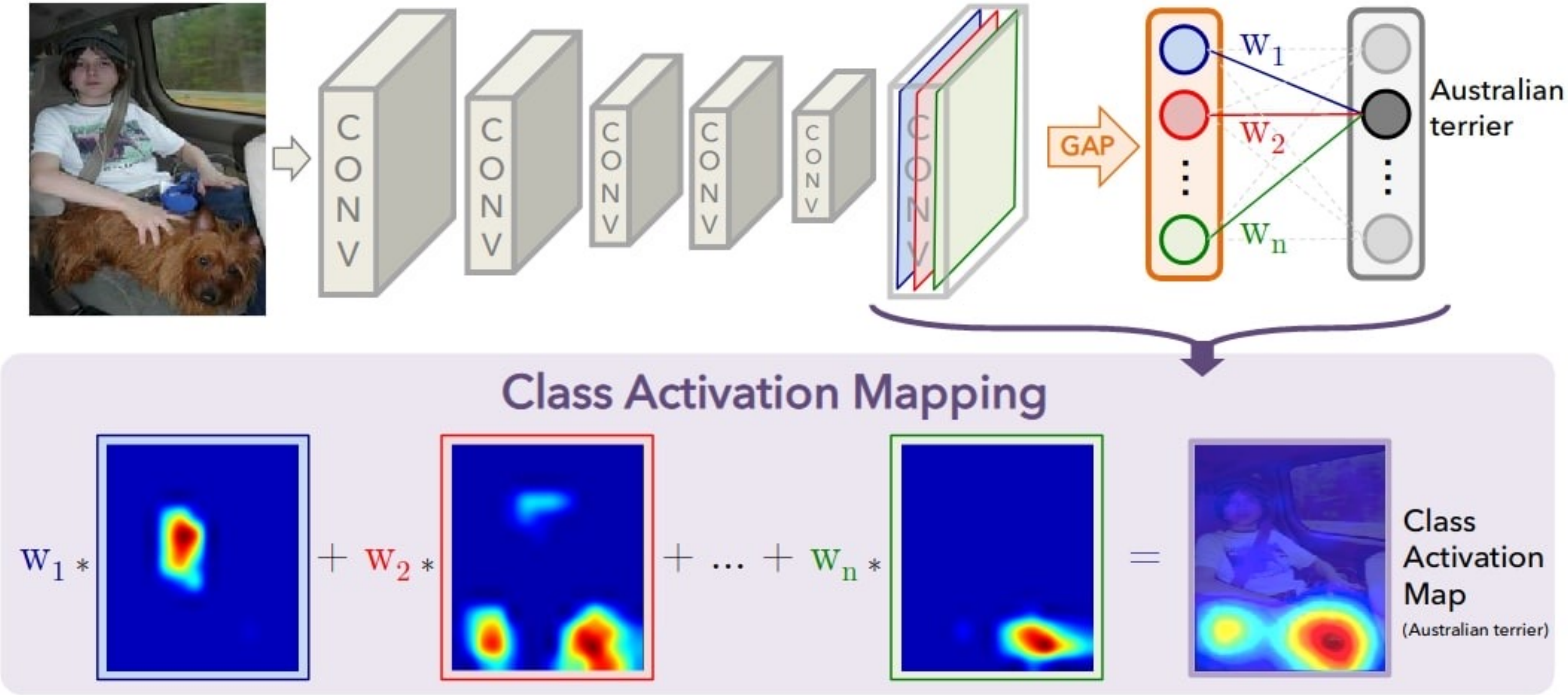
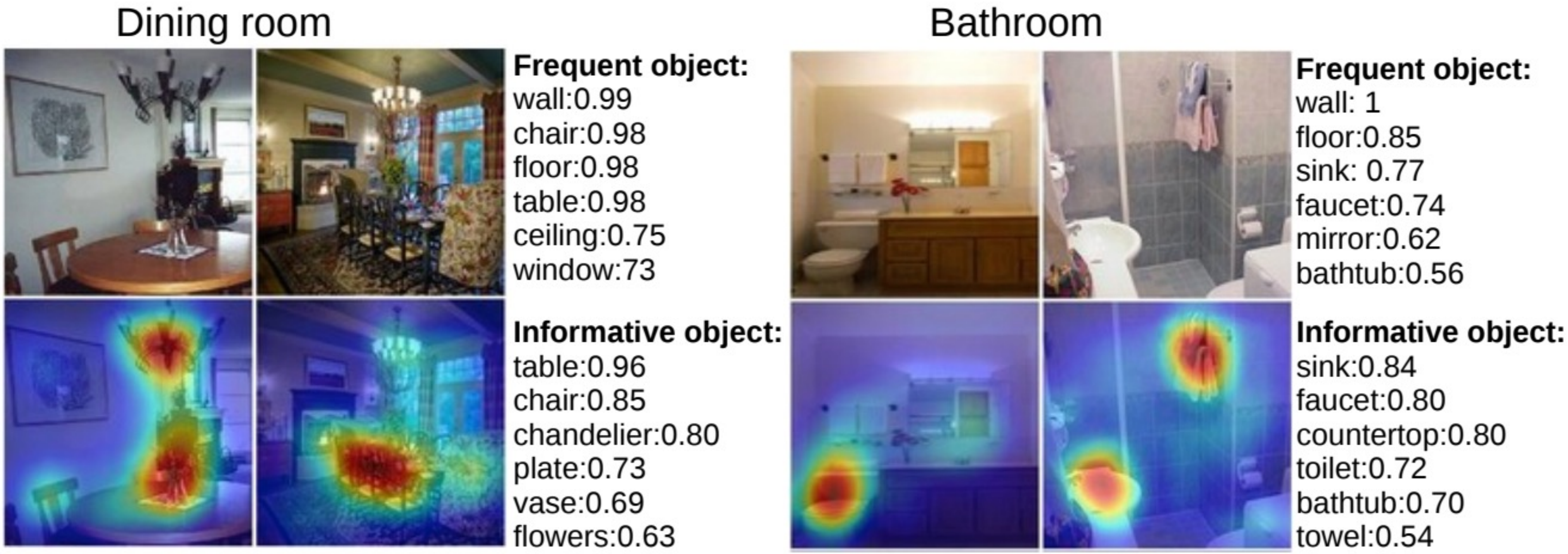


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

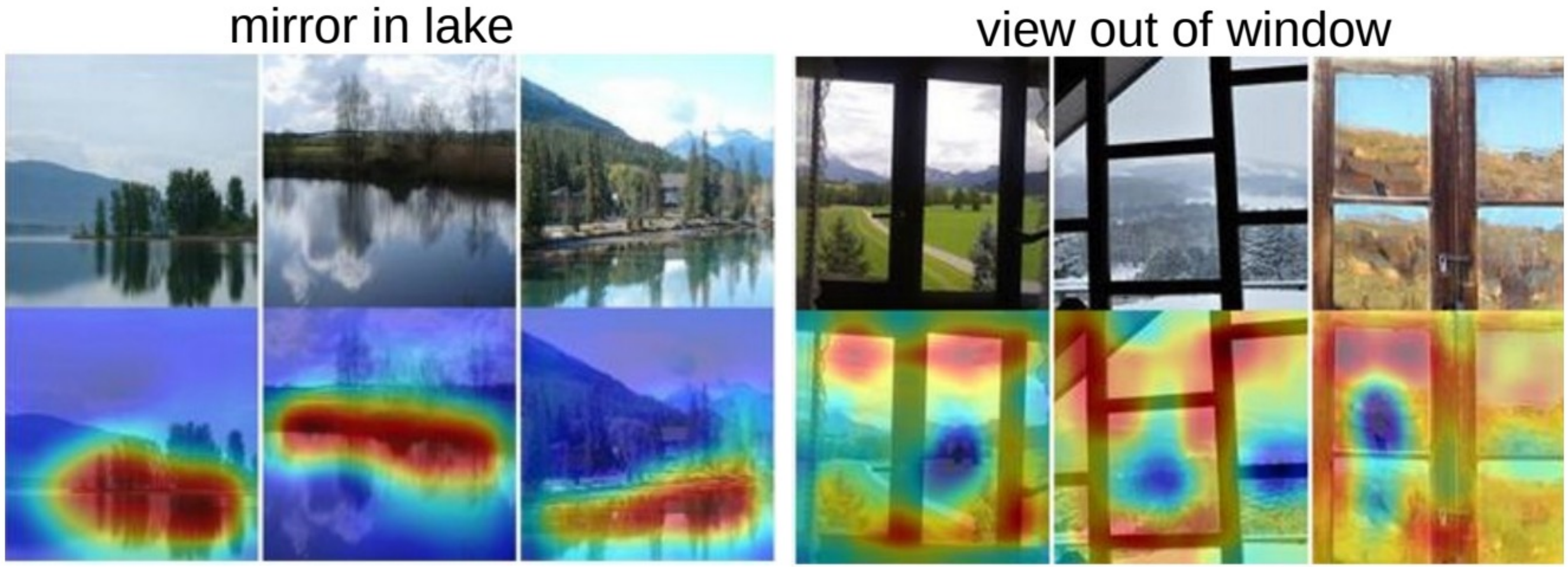
Interpretability (XAI): CAM-based method

Application of CAM: informative objects detection



Interpretability (XAI): CAM-based method

Application of CAM: informative regions for the concept learned from weakly labelled images



Interpretability (XAI): CAM-based method

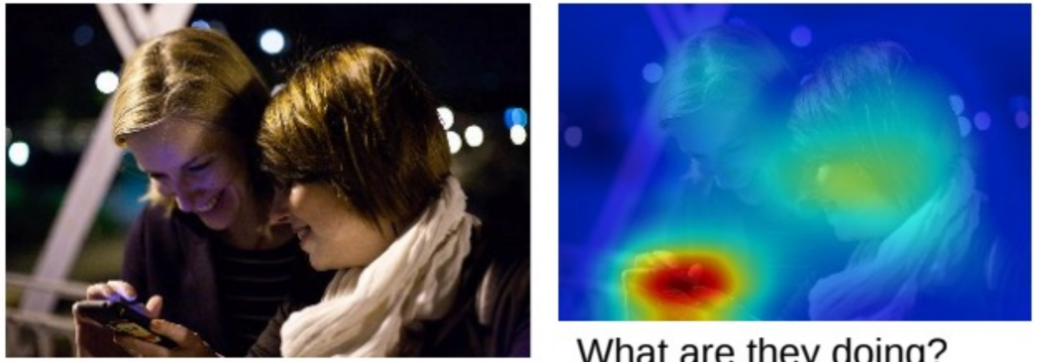
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Application of CAM: weakly supervised text detector

Interpretability (XAI): CAM-based method

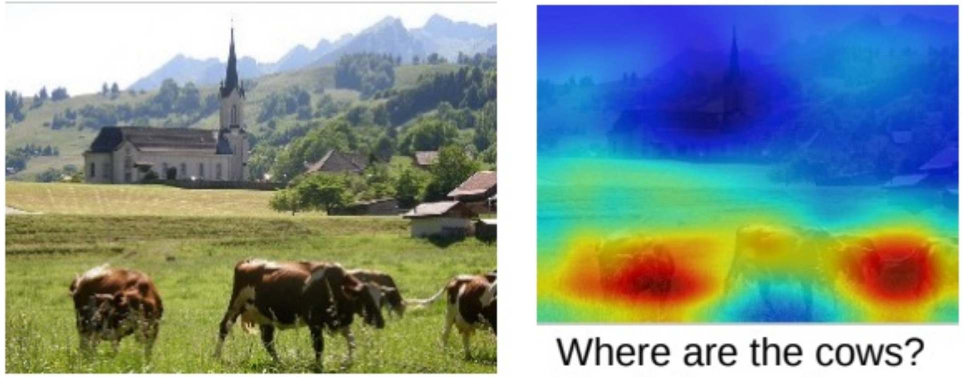
Application of CAM: Visualization for the predicted answer in VQA

What is the color of the horse?
Prediction: brown



What are they doing?
Prediction: texting

What is the sport?
Prediction: skateboarding



Where are the cows?
Prediction: on the grass

Interpretability (XAI): CAM-based method

Disadvantages of CAM:

- Specific design of network architecture: FCN layer -> GAP layer
- Only focused on the classification problem

Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Interpretability (XAI): CAM-based method

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

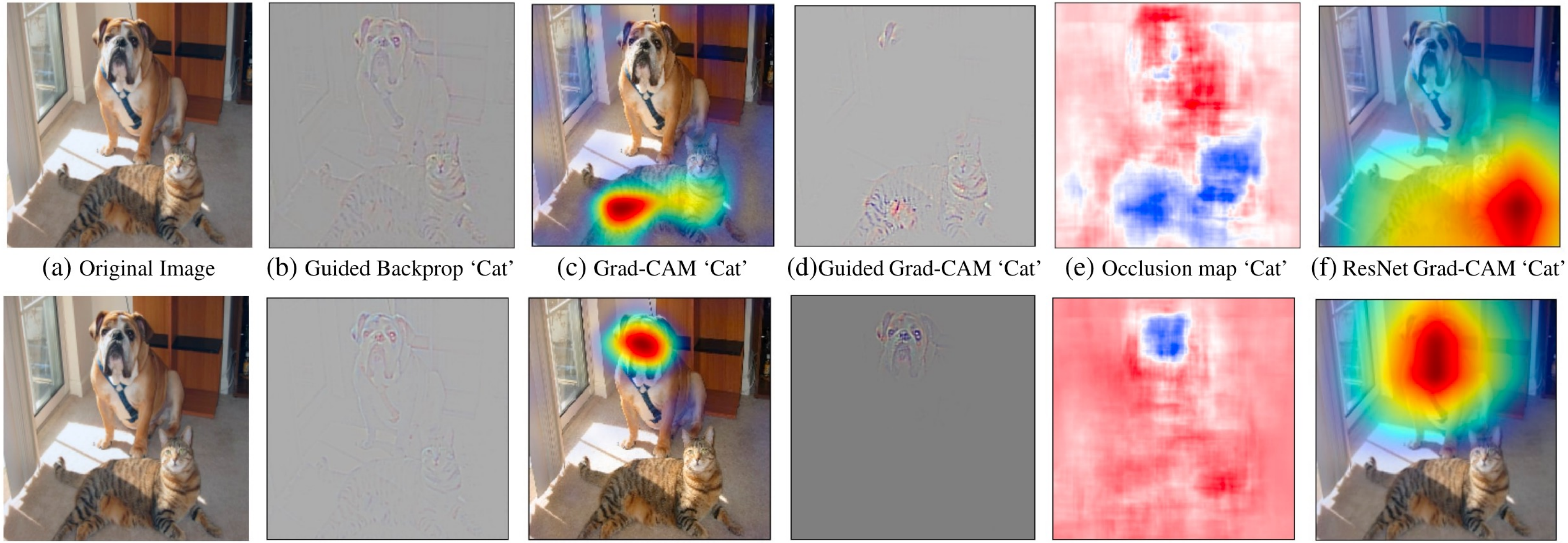
Motivation of Grad-CAM:

- CNNs' lack of **decomposability into individually intuitive components** makes them hard to interpret;
- **Trade-off** between interpretability and accuracy;
- **Shortage of CAM**: trades off model complexity and performance for more transparency into the working of the model
- **What makes a good visual explanation**:
  - Class discriminative
  - High-resolution

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Interpretability (XAI): CAM-based method

Visualization of a number of methods:



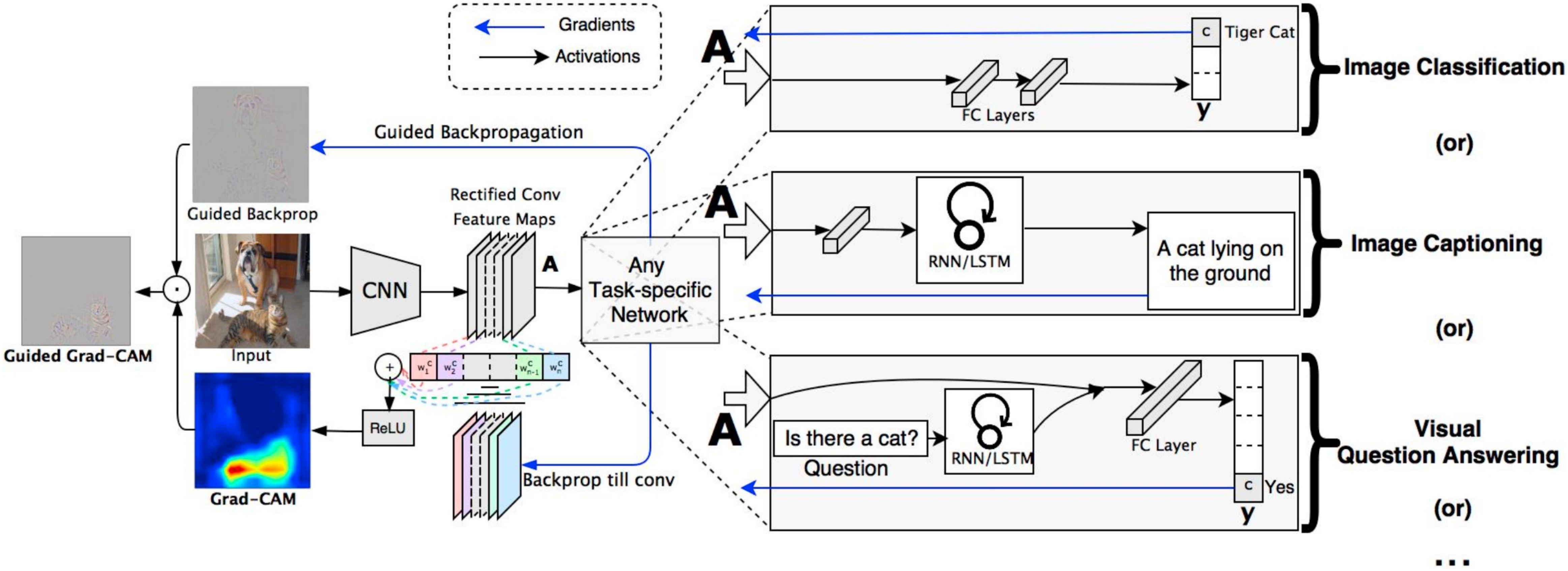
(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d)Guided Grad-CAM 'Cat' (e) Occlusion map 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j)Guided Grad-CAM 'Dog' (k) Occlusion map 'Dog' (l)ResNet Grad-CAM 'Dog'

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Interpretability (XAI): CAM-based method

Method: Grad-CAM



Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Interpretability (XAI): CAM-based method

Method: Grad-CAM

- To obtain the class-discriminative localization map Grad-CAM $L^c_{Grad-CAM} \in R^{u \times v}$ of width u and height v for any class c:
  - ❖ Compute the gradient of the score for class c, y^c (before the softmax), w.r.t feature map activations A^k of a convolutional layer, i.e.,

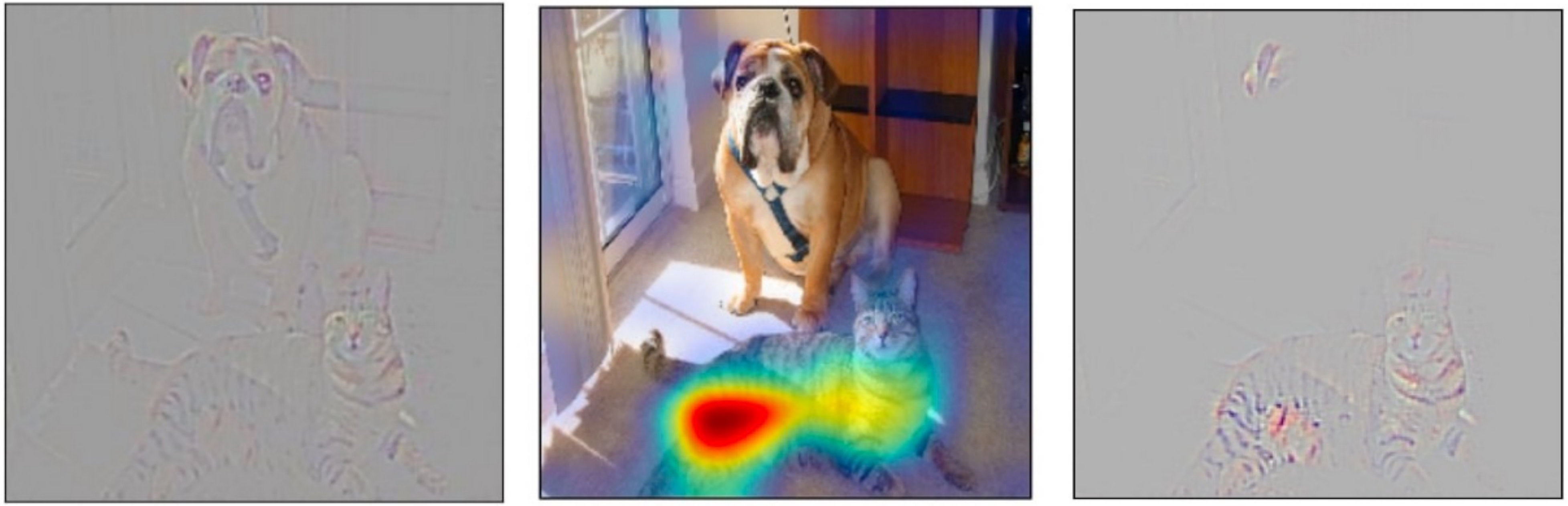
$$\frac{\partial y^c}{\partial A^k}$$

  - ❖ Obtain the neuron importance weights α_k^c: $\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A^k_{ij}}}_{\text{gradients via backprop}}$
  - ❖ Obtain Grad-CAM: $L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$

Interpretability (XAI): CAM-based method

Method: Guided Grad-CAM

- Fuse Grad-CAM with Guided Backpropagation via element-wise multiplication

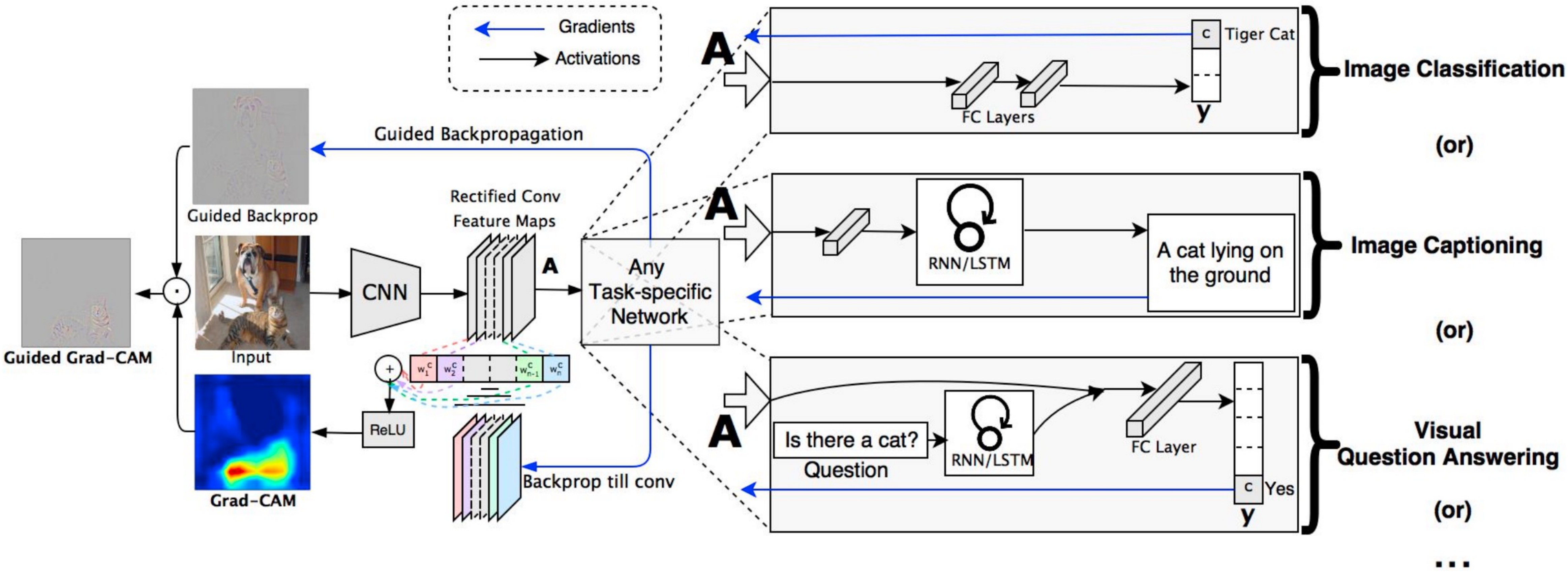


(b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d)Guided Grad-CAM 'Cat'

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Interpretability (XAI): CAM-based method

Method: Grad-CAM

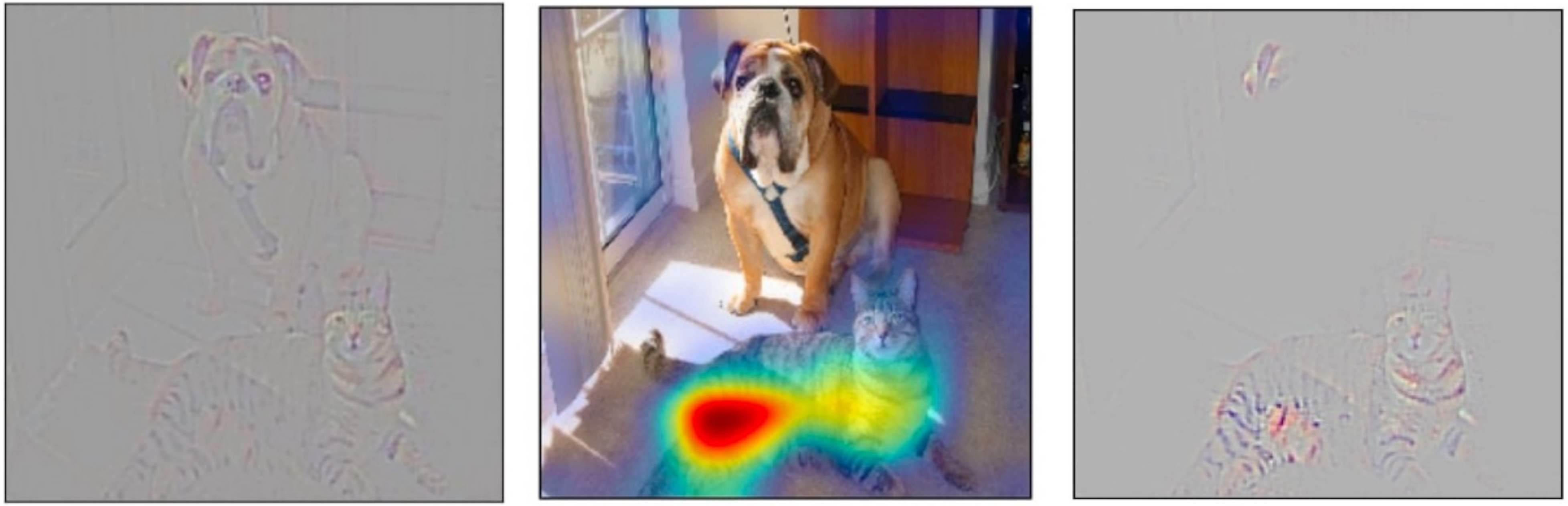


Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Interpretability (XAI): CAM-based method

Method: Guided Grad-CAM

- Fuse Grad-CAM with Guided Backpropagation via element-wise multiplication



(b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d)Guided Grad-CAM 'Cat'

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Interpretability (XAI): CAM-based method

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Experiment: 1. Weakly-supervised Localization

		Classification		Localization	
		Top-1	**Top-5**	**Top-1**	**Top-5**
VGG-16	Backprop [51]	30.38	10.89	61.12	51.46
	c-MWP [58]	30.38	10.89	70.92	63.04
	Grad-CAM (ours)	30.38	10.89	**56.51**	46.41
	CAM [59]	33.40	12.20	57.20	**45.14**
AlexNet	c-MWP [58]	44.2	20.8	92.6	89.2
	Grad-CAM (ours)	44.2	20.8	68.3	56.6
GoogleNet	Grad-CAM (ours)	31.9	11.3	60.09	49.34
	CAM [59]	31.9	11.3	60.09	49.34

Table 1: Classification and localization error % on ILSVRC-15 val (lower is better) for VGG-16, AlexNet and GoogleNet. We see that Grad-CAM achieves superior localization errors without compromising on classification performance.

Interpretability (XAI): CAM-based method

Experiment: 2. Weakly-supervised Segmentation

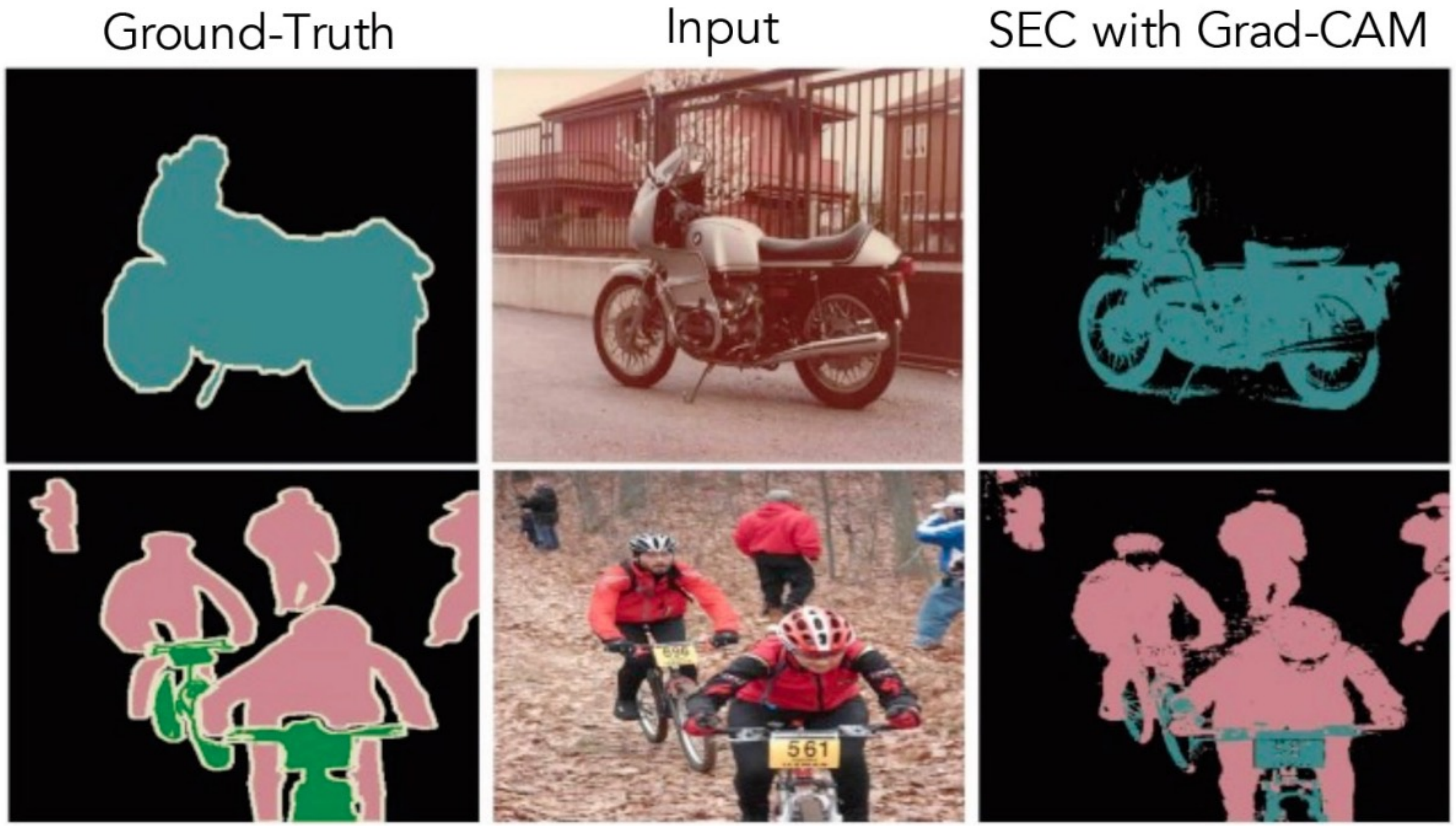
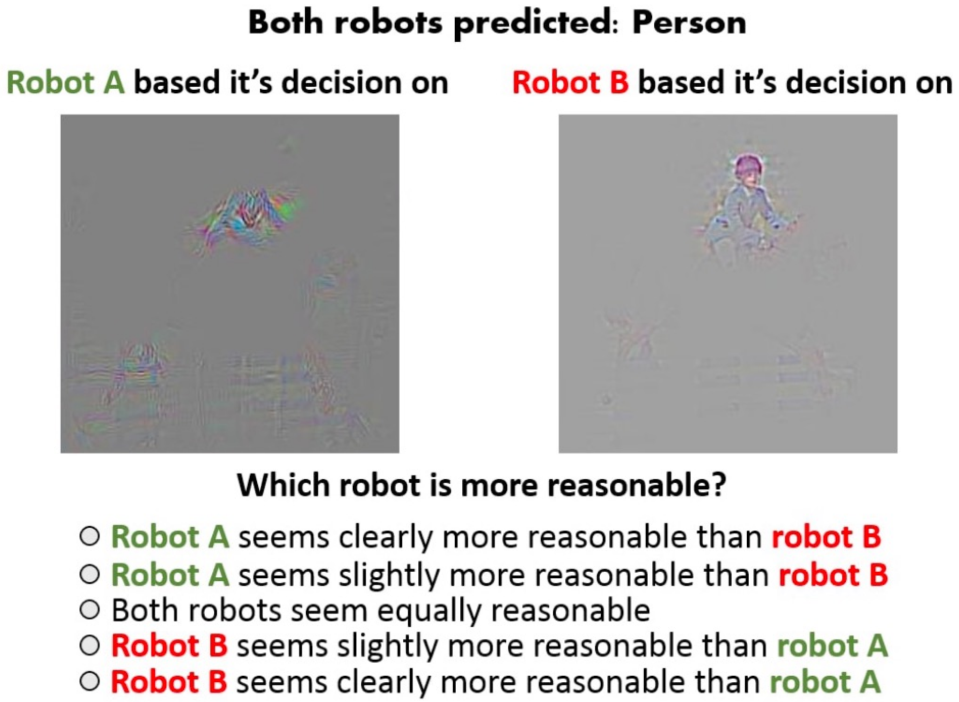


Fig. 4: PASCAL VOC 2012 Segmentation results with Grad-CAM as seed for SEC [32].

Human study: Evaluating Class Discrimination

(a) Raw input image. Note that this is not a part of the tasks (b) and (c)

(b) AMT interface for evaluating the class-discriminative property



(c) AMT interface for evaluating if our visualizations instill trust in an end user

Fig. 5: AMT interfaces for evaluating different visualizations for class discrimination (b) and trustworthiness (c). Guided Grad-CAM outperforms baseline approaches (Guided-backprop and Deconvolution) showing that our visualizations are more class-discriminative and help humans place trust in a more accurate classifier.

Interpretability (XAI): CAM-based method

Human study: Evaluating Class Discrimination

Method	Human Classification Accuracy	Relative Reliability	Rank Correlation w/ Occlusion
Guided Backpropagation	44.44	+1.00	0.168
Guided Grad-CAM	61.23	+1.27	0.261

Table 2: Quantitative Visualization Evaluation. Guided Grad-CAM enables humans to differentiate between visualizations of different classes (Human Classification Accuracy) and pick more reliable models (Relative Reliability). It also accurately reflects the behavior of the model (Rank Correlation w/ Occlusion).

Interpretability (XAI): CAM-based method

Diagnosis CNN with Grad-CAM: Analyzing failure modes for VGG-16

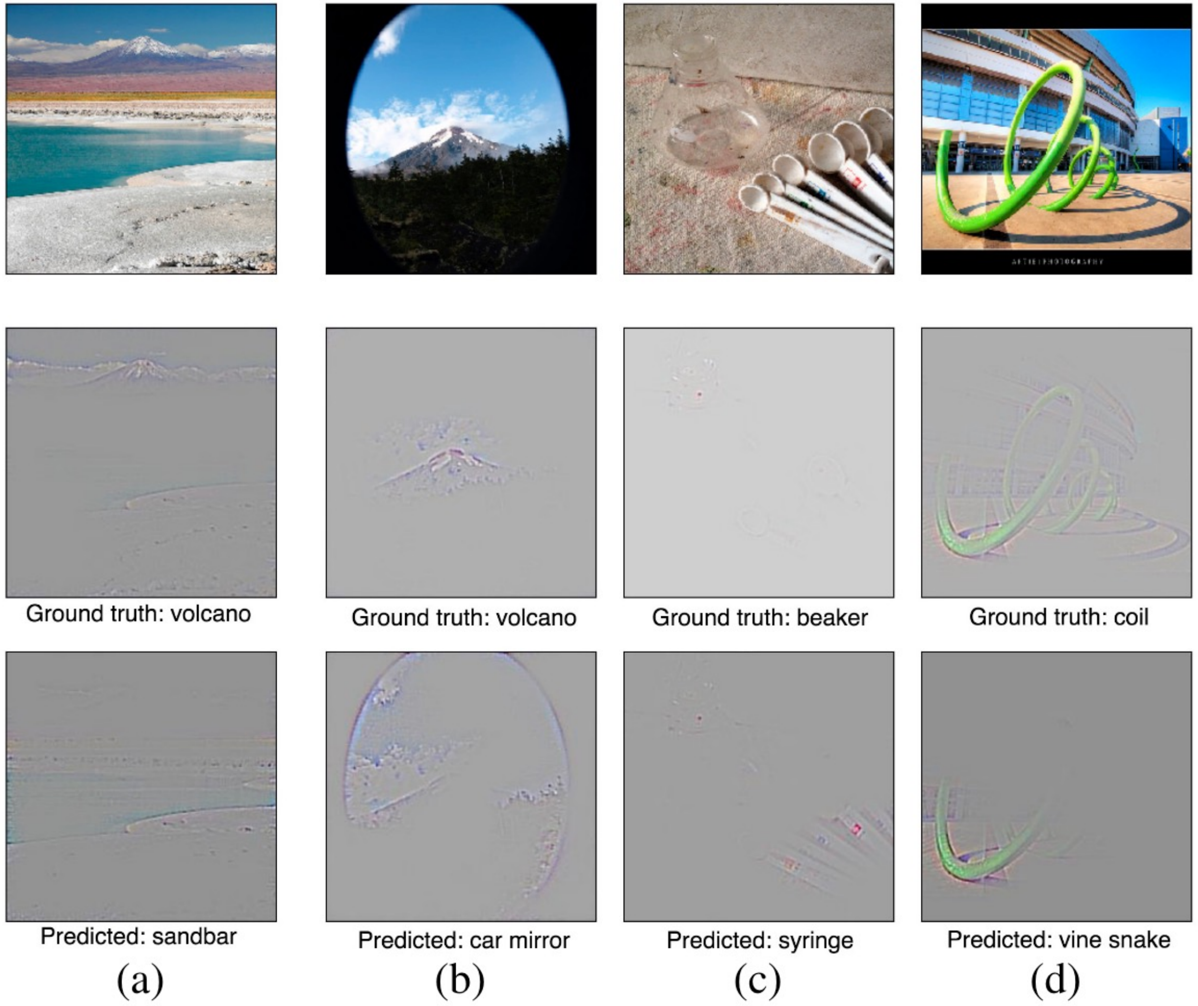


Fig. 6: In these cases the model (VGG-16) failed to predict the correct class in its top 1 (a and d) and top 5 (b and c) predictions. Humans would find it hard to explain some of these predictions without looking at the visualization for the predicted class. But with Grad-CAM, these mistakes seem justifiable.

Interpretability (XAI): CAM-based method

Diagnosis CNN with Grad-CAM: Effect of adversarial noise on VGG-16

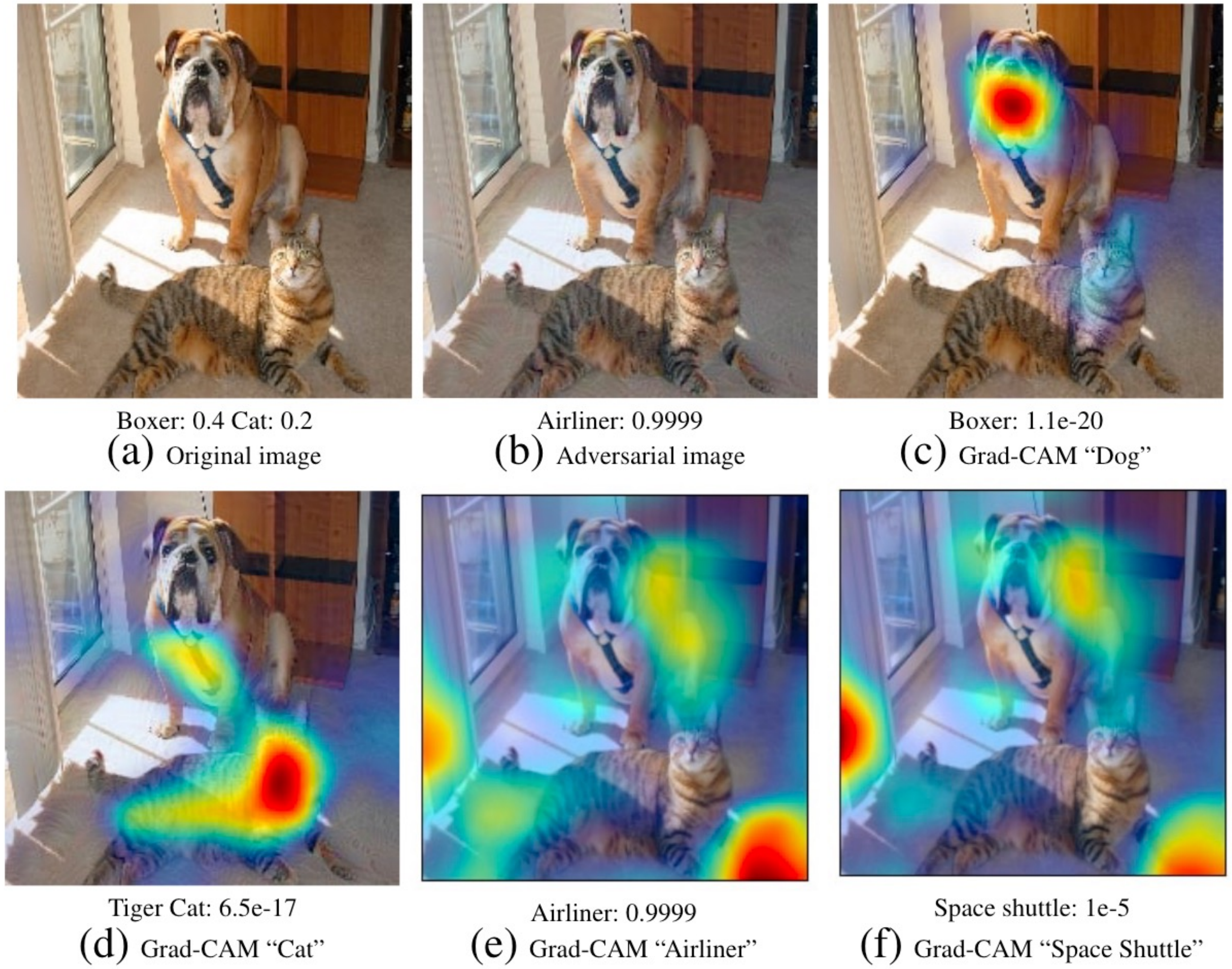


Fig. 7: (a-b) Original image and the generated adversarial image for category "airliner". (c-d) Grad-CAM visualizations for the original categories "tiger cat" and "boxer (dog)" along with their confidence. Despite the network being completely fooled into predicting the dominant category label of "airliner" with high confidence (>0.9999), Grad-CAM can localize the original categories accurately. (e-f) Grad-CAM for the top-2 predicted classes "airliner" and "space shuttle" seems to highlight the background.

Interpretability (XAI): CAM-based method

Diagnosis CNN with Grad-CAM: Identifying bias in dataset

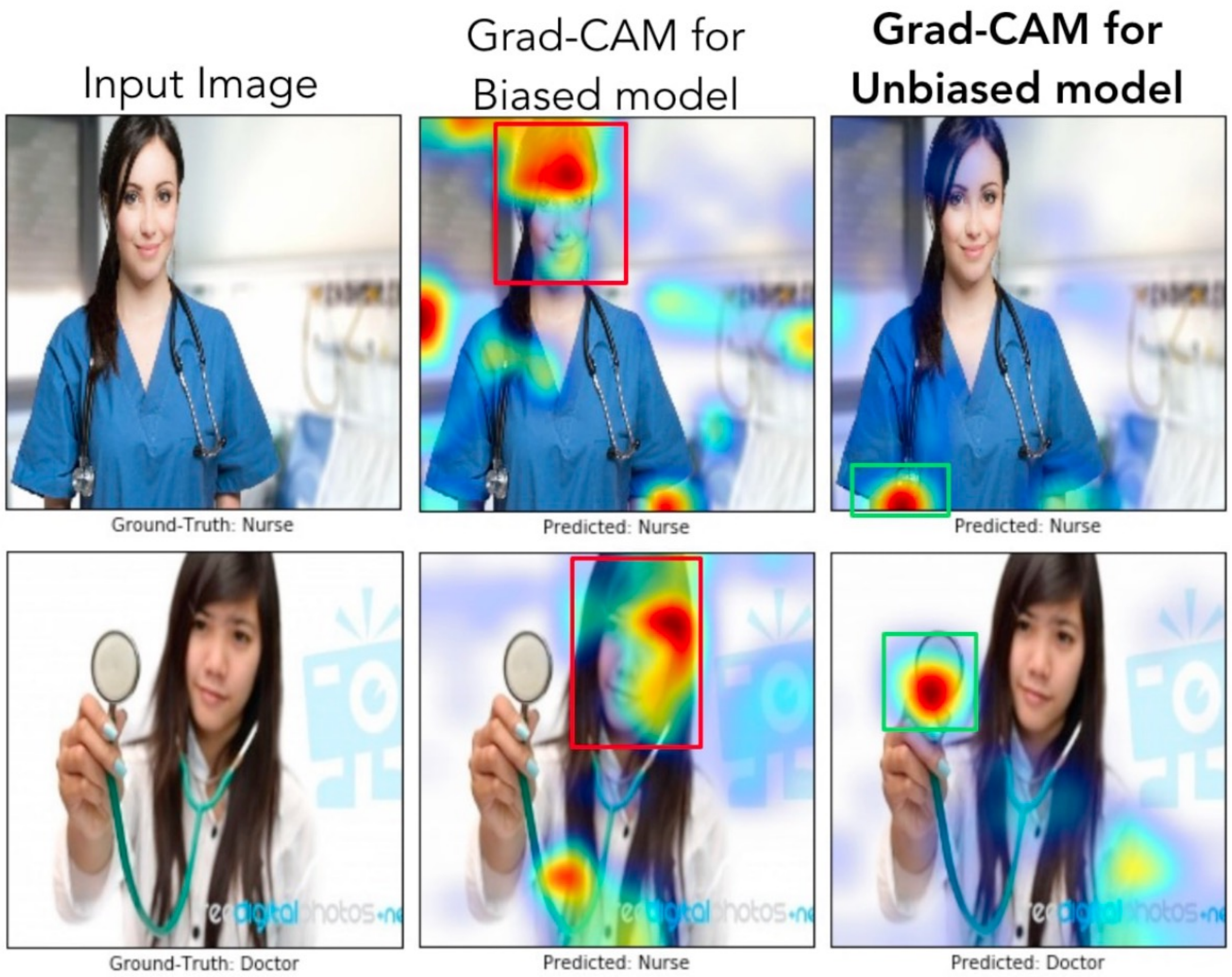
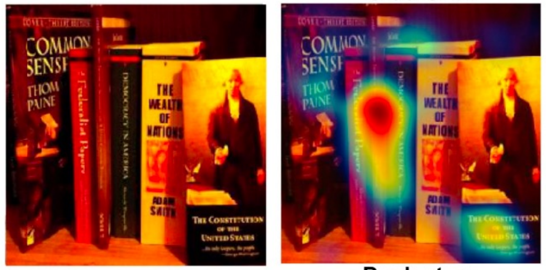


Fig. 8: In the first row, we can see that even though both models made the right decision, the biased model (model1) was looking at the face of the person to decide if the person was a nurse, whereas the unbiased model was looking at the short sleeves to make the decision. For the example image in the second row, the biased model made the wrong prediction (misclassifying a doctor as a nurse) by looking at the face and the hairstyle, whereas the unbiased model made the right prediction looking at the white coat, and the stethoscope.

Interpretability (XAI): CAM-based method

Textual explanations with Grad-CAM:

Grad-CAM



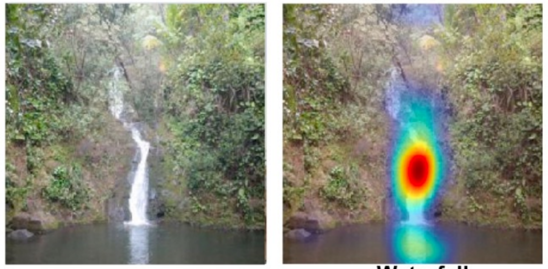
Book-store

Important concepts for `Book-store`

Positive Neuron ID	Positive Concept	Negative Neuron ID	Negative Concept
78	Book	237	Sky
318	Book	357	road
502	Striped	148	Water
311	Shelf	404	Car
156	Swirly	71	Flower

(a)

Grad-CAM



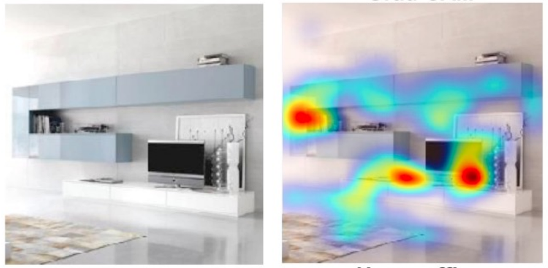
Waterfall

Important concepts for `Waterfall`

Positive Neuron ID	Positive Concept	Negative Neuron ID	Negative Concept
117	Waterfall	115	Corridor
106	Closet	166	Road
148	Water	494	Bus
143	Water	106	Laundromat
216	Stratified	412	Grid

(b)

Grad-CAM



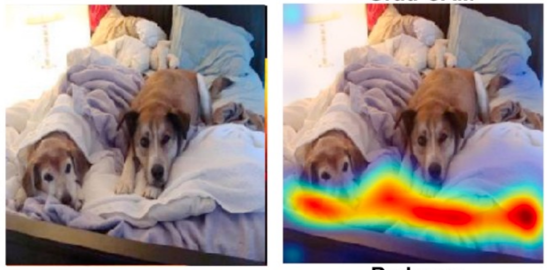
Home-office

Important concepts for `Home-office`

Positive Neuron ID	Positive Concept	Negative Neuron ID	Negative Concept
78	Book	186	Chequered
312	Desk	237	Sky
75	Office	494	Swimming pool
492	Stove	334	Sidewalk
305	Screen	498	Crosswalk

(c)

Grad-CAM



Bedroom

Important concepts for `Bedroom`

Positive Neuron ID	Positive Concept	Negative Neuron ID	Negative Concept
317	Bed	187	Spiralled
290	Bed	294	Pantry
226	Painting	26	Toiled
175	Cushion	9	Shoe shop
117	Waterfall	182	Amusement park

(d)

Grad-CAM

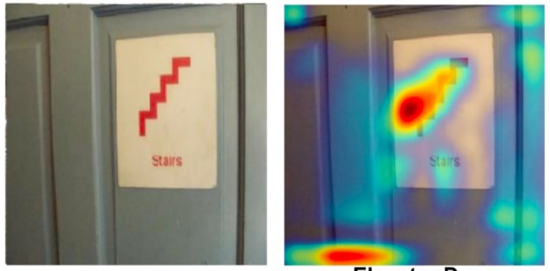
Rope-bridge

Important concepts for `Rope-bridge`

Positive Neuron ID	Positive Concept	Negative Neuron ID	Negative Concept
148	Water	242	House
166	Water	101	House
266	Bridge	351	House
106	Closet	490	Dog
143	Water	477	Tree

(e)

Grad-CAM



Elevator Door

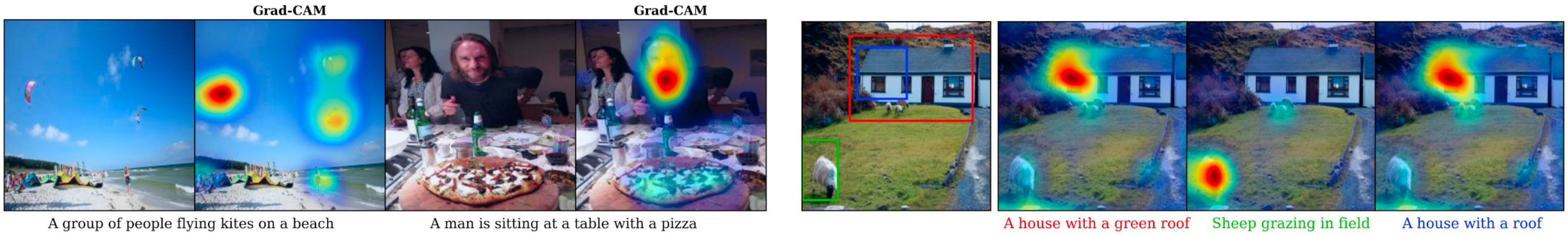
Important concepts for `Elevator Door`

Positive Neuron ID	Positive Concept	Negative Neuron ID	Negative Concept
323	Cabinet	78	Book
479	Crosswalk	61	Classroom
431	Staircase	294	Pantry
194	Meshed	485	Lacelike
20	Track	384	Toilet

(f)

Interpretability (XAI): CAM-based method

Grad-CAM for Image Captioning:

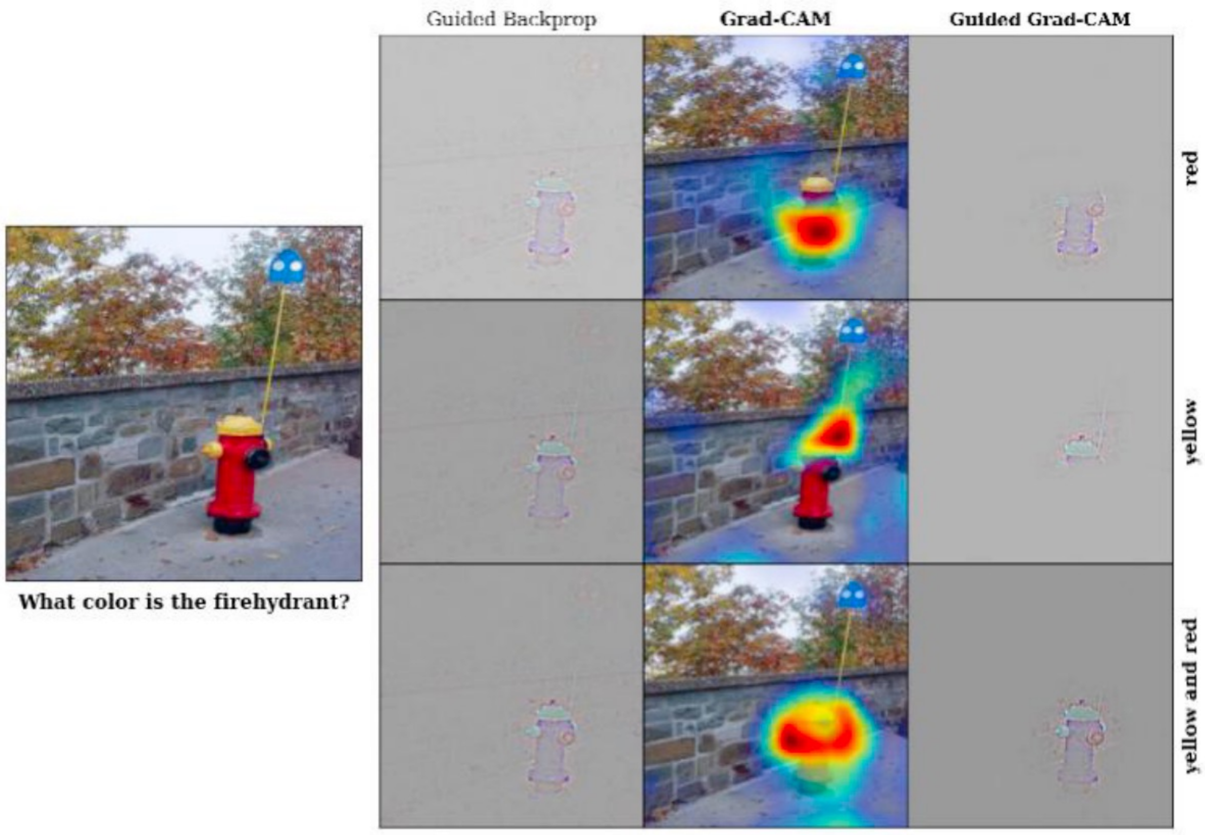


(a) Image captioning explanations (b) Comparison to DenseCap

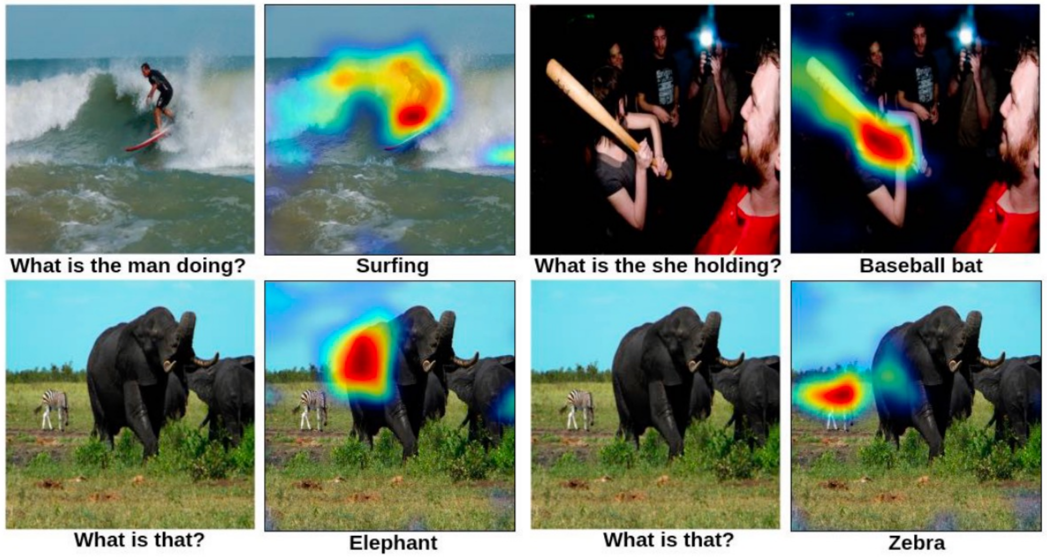
Fig. 10: Interpreting image captioning models: We use our class-discriminative localization technique, Grad-CAM to find spatial support regions for captions in images. Fig. 10a Visual explanations from image captioning model [31] highlighting image regions considered to be important for producing the captions. Fig. 10b Grad-CAM localizations of a *global* or *holistic* captioning model for captions generated by a dense captioning model [29] for the three bounding box proposals marked on the left. We can see that we get back Grad-CAM localizations (right) that agree with those bounding boxes – even though the captioning model and Grad-CAM techniques do not use any bounding box annotations.

Interpretability (XAI): CAM-based method

Grad-CAM for Visual Question Answering (VQA):



(a) Visualizing VQA model from [38]



(b) Visualizing ResNet based Hierarchical co-attention VQA model from [39]

Fig. 12: Qualitative Results for our VQA experiments: (a) Given the image on the left and the question "What color is the firehydrant?", we visualize Grad-CAMs and Guided Grad-CAMs for the answers "red", "yellow" and "yellow and red". Grad-CAM visualizations are highly interpretable and help explain any target prediction – for "red", the model focuses on the bottom red part of the firehydrant; when forced to answer "yellow", the model concentrates on it's top yellow cap, and when forced to answer "yellow and red", it looks at the whole firehydrant! (b) Our approach is capable of providing interpretable explanations even for complex models.

Thank you