Explainability of Deep Neural Networks with Comparative Neuron Activations and Gradients

2023.06.19

경북대학교 IT대학 컴퓨터학부

남우정

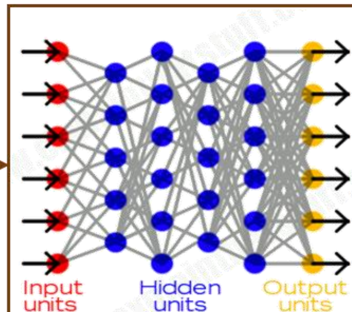
Overview of eXplainable Artificial Intelligence (XAI)

Explaining the decision of Deep Neural Networks beyond the complex and nonlinear internal structures

Current Deep Learning	Explainable Deep Learning
▪ Transparency of Deep Neural Networks (DNNs) is hampered by complex and nonlinear internal structures ▪ Despite the tremendous performance of deep learning, **"Clever hans" phenomenon** could be occurred	▪ Providing the rationale of the decision accessible to humans, leading to higher confidence in the ability of the model ▪ Reducing the potential risk of an unpredictable phenomenon and helping debug the model

Training Data

Learning Process



Learned Function

This is a cat
(p = .93)

Output

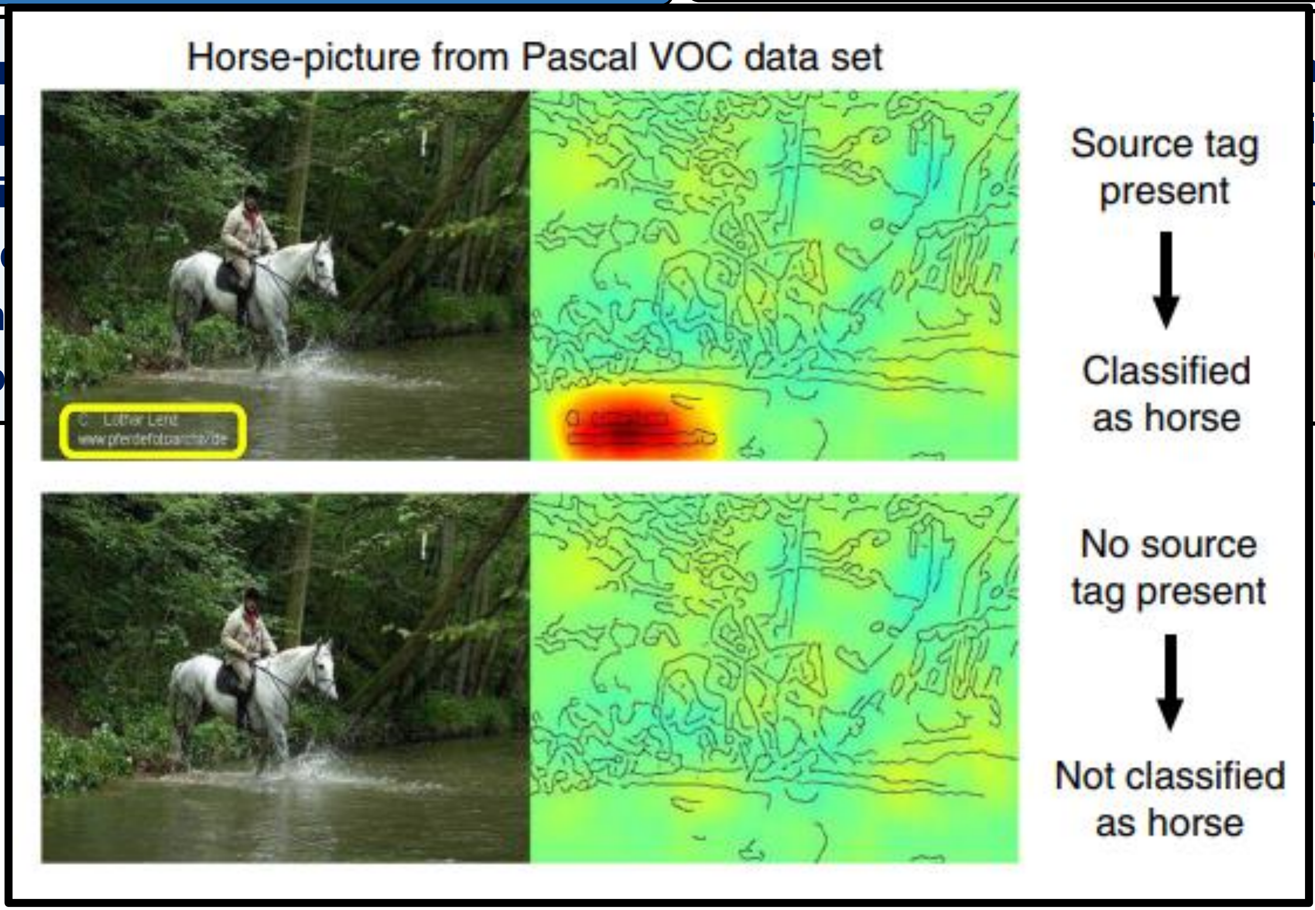
User

"Black box" inner structure of deep learning

Overview of eXplainable Artificial Intelligence (XAI)

Explaining the decision of Deep Neural Networks beyond the complex and nonlinear internal structures

Current Deep Learning	Explainable Deep Learning
▪ Transpare (DNNs) is l nonlinear i ▪ Despite th deep learn could be o	e decision ing to higher the model of an and helping

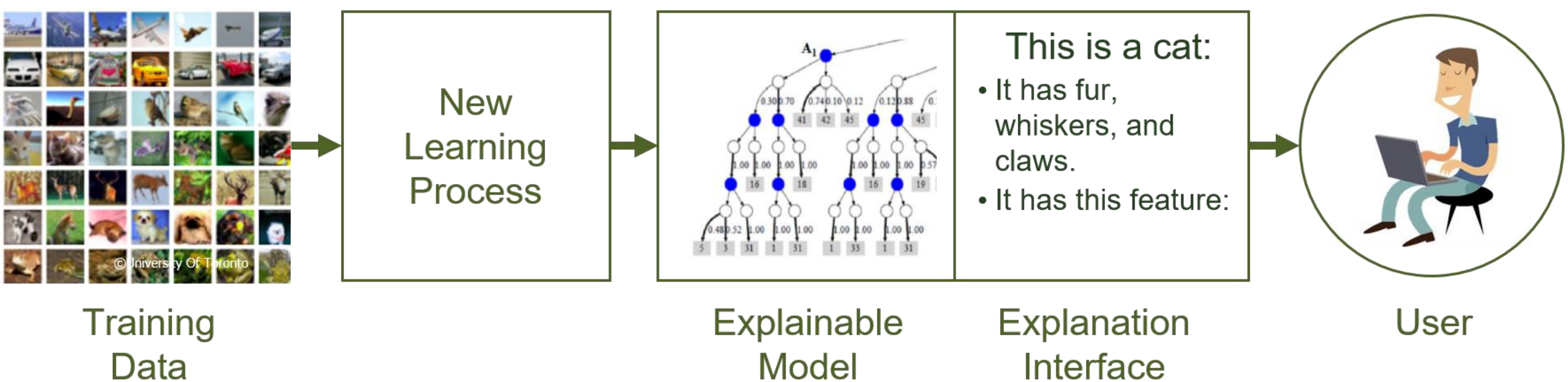


"Clever hans" phenomenon of the Fisher vector classifier [S. Lapuschkin, 2019]

Overview of eXplainable Artificial Intelligence (XAI)

Explaining the decision of Deep Neural Networks beyond the complex and nonlinear internal structures

Current Deep Learning	Explainable Deep Learning
▪ Transparency of Deep Neural Networks (DNNs) is hampered by complex and nonlinear internal structures ▪ Despite the tremendous performance of deep learning, **“Clever hans” phenomenon** could be occurred	▪ Providing the rationale of the decision accessible to humans, leading to **higher confidence** in the ability of the model ▪ Reducing the **potential risk of an unpredictable phenomenon** and helping debug the model

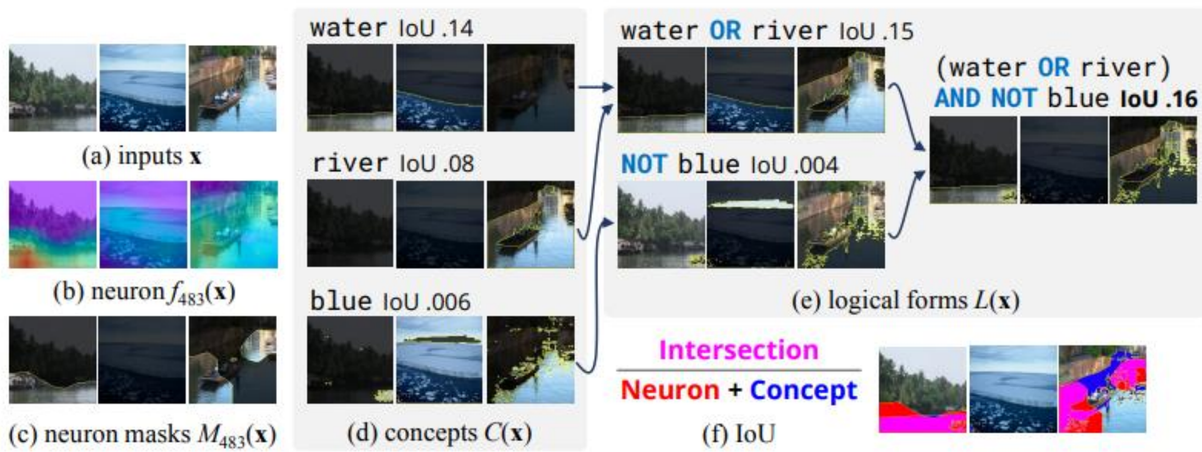


Overview of explainable machine learning

Related Works

- **Revealing the evidence of model decision**
  - **Visualizations of Intermediate Features**
    - ✓ **Visualizing intermediate features by maximizing the activated neurons**
  - **Concept-based explanation**
    - ✓ **Visualizing how a model learned a class in terms of concepts**

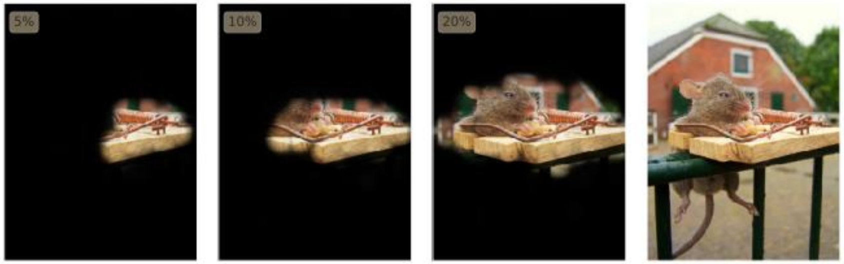
Visualizing Intermediate Features
[Mahendran, A. et.al., 2016]



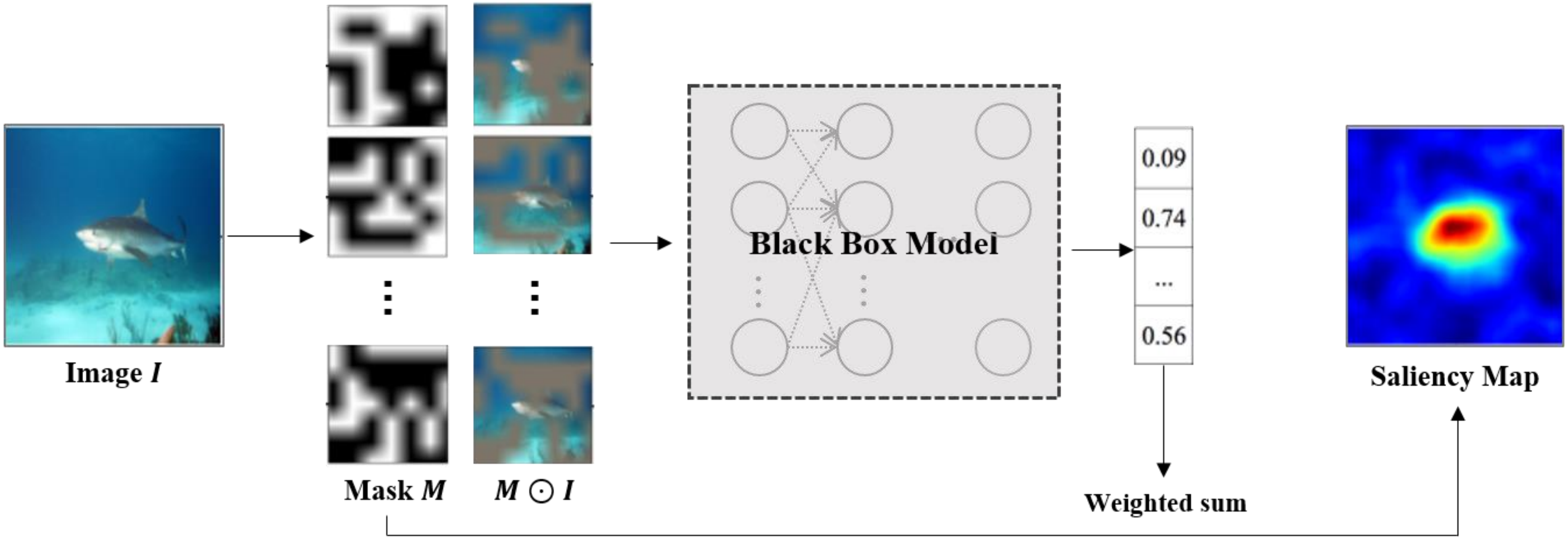
Concept based explanations of neuron
[Jess, M. et.al., 2020]

Related Works (Cont.)

- **Revealing the evidence of model decision**
  - **Perturbation-based approach**
    - ✓ **Analyzing the variations of decision when distorting the input image**



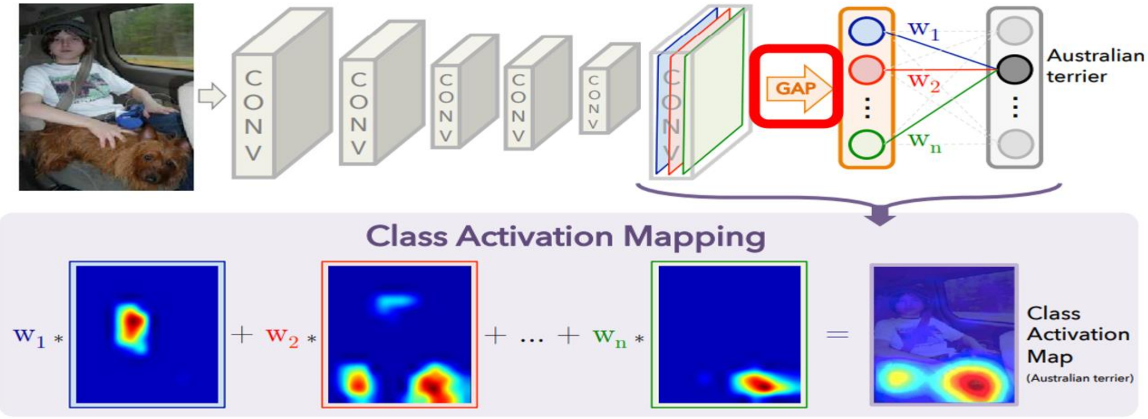
Extremal perturbation [Fong, R. et.al., 2015]



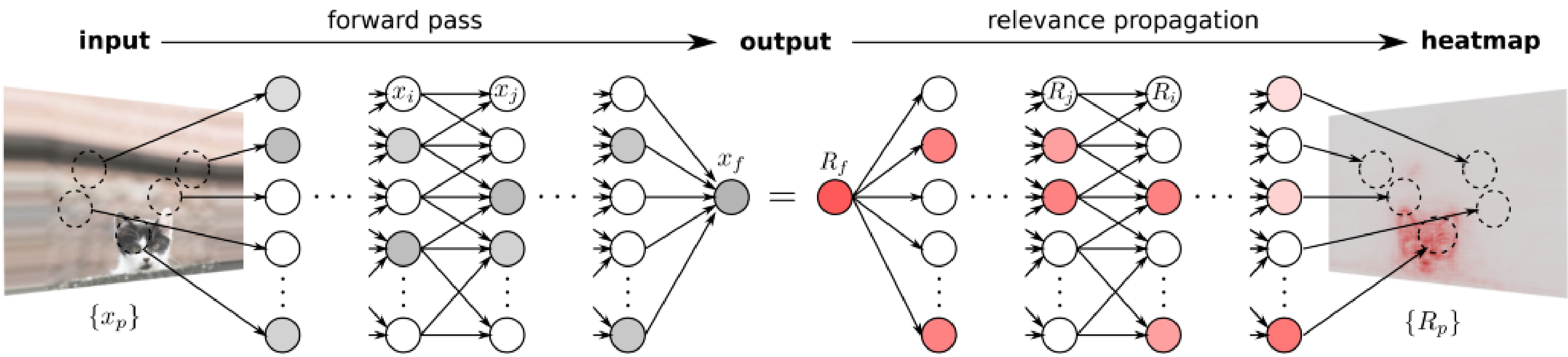
Overall framework of Perturbation-based approach [V. Petsiuk, 2018]

Related Works (Cont.)

- **Revealing the evidence of model decision**
  - **Decomposing the network decision**
    - ✓ **Aim to seek the relevant parts of the input image by following the backward propagation rule that preserve the evidence of decision**



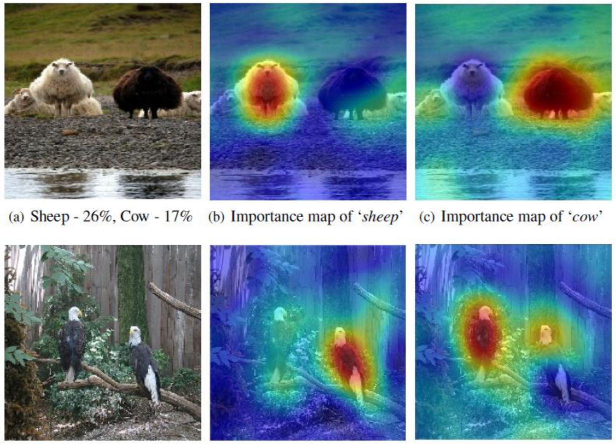
Overview of Class Activation Mapping [Zhou, B. et.al., 2016]



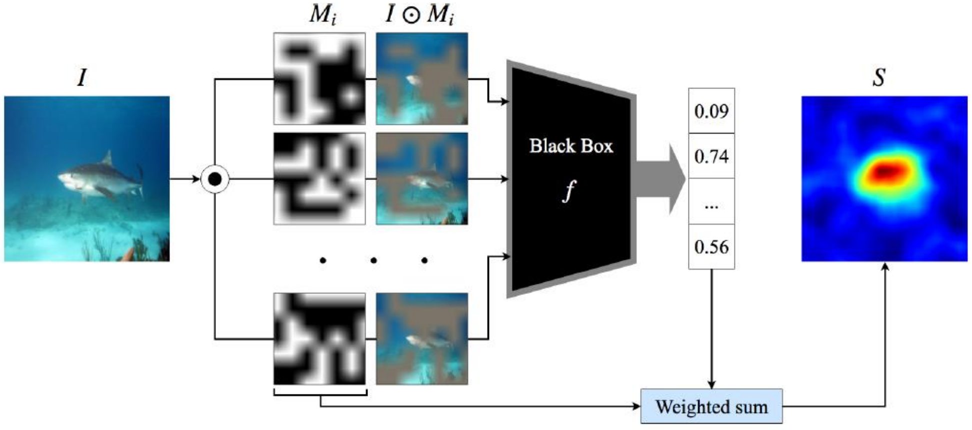
Overview of Layerwise Relevance Propagation [Montavon, G. et.al., 2017]

Perturbation-based Approaches

- **RISE [BMVC, 2018]**
  - Generating an importance map indicating how salient each pixel is for the model's prediction
    - ✓ Generating N binary masks smaller than the input image and upsample to size with the input image
    - ✓ Running the model using the masked image to get confidence scores
    - ✓ A saliency map for each pixel is obtained as a weighted sum of confidence scores and masks



Generated a pixel importance map for each decision (redder is more important)

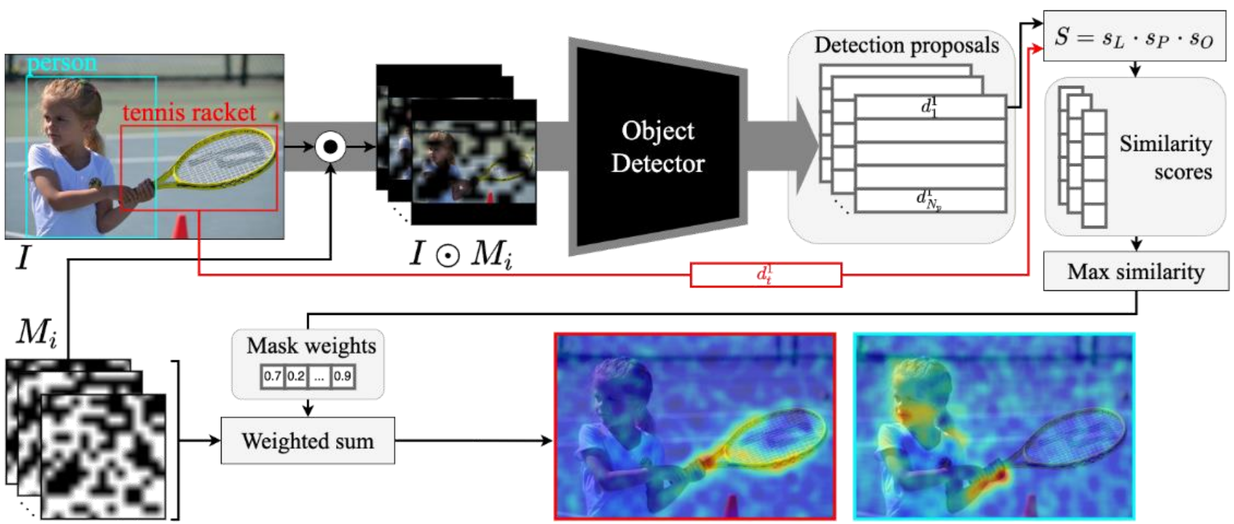


Overview of the method proposed in this paper

Vitali, Abir, and Kate. "RISE: Randomized Input Sampling for Explanation of Black-box Models." *BMVC* 2018

Perturbation-based Approaches

- **D-RISE [CVPR, 2018]**
  - Generating saliency maps that show image areas that most affect the prediction in both localization and classification tasks
    - ✓ Producing a masked image using randomly generated masks
    - ✓ Run the detector to produce several proposals for each masked image
    - ✓ Computing pairwise similarities between ground truth and predicted vectors and get the maximum score to generate the weight
    - ✓ Computing a weighted sum of masks with respect to get saliency maps

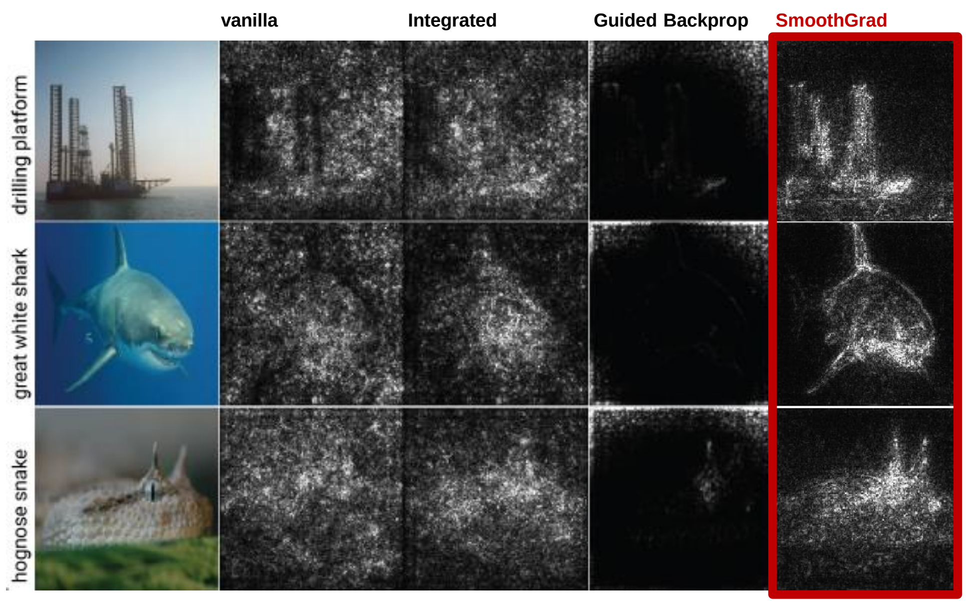


An overview of D-RISE framework

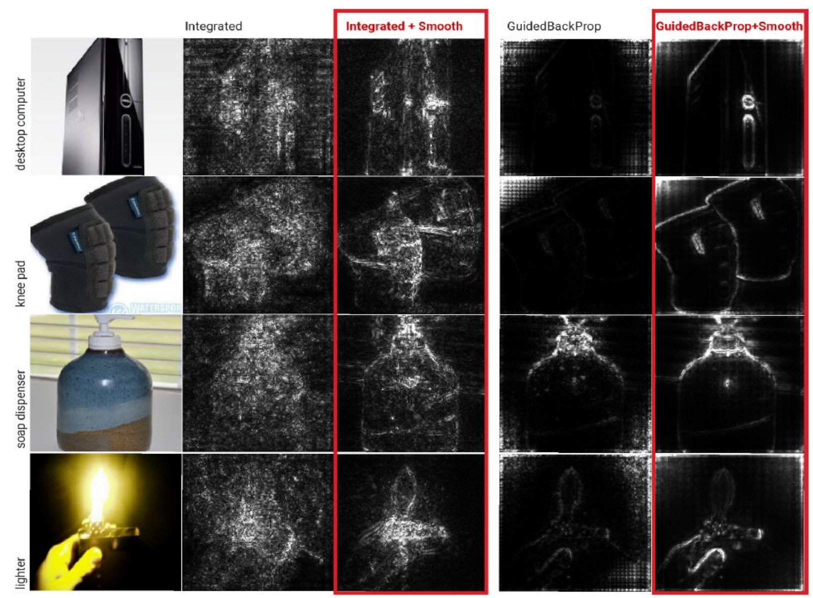
Vitali, Rajiv, Varun, Vlad, Ashutosh, Vicente, and Kate. “D-RISE: Black-box Explanation of Object Detectors via Saliency Maps.” CVPR 2021

Gradient-based Approaches

- **SmoothGrad [ICML, 2019]**
  - ➢ Introducing a simple approach that improves the quality of saliency maps by iteratively injecting noises into the input image
    - ✓ Computing the average of sensitive maps that are generated from noise-injected images to generate final saliency maps



Qualitive evaluation of different methods

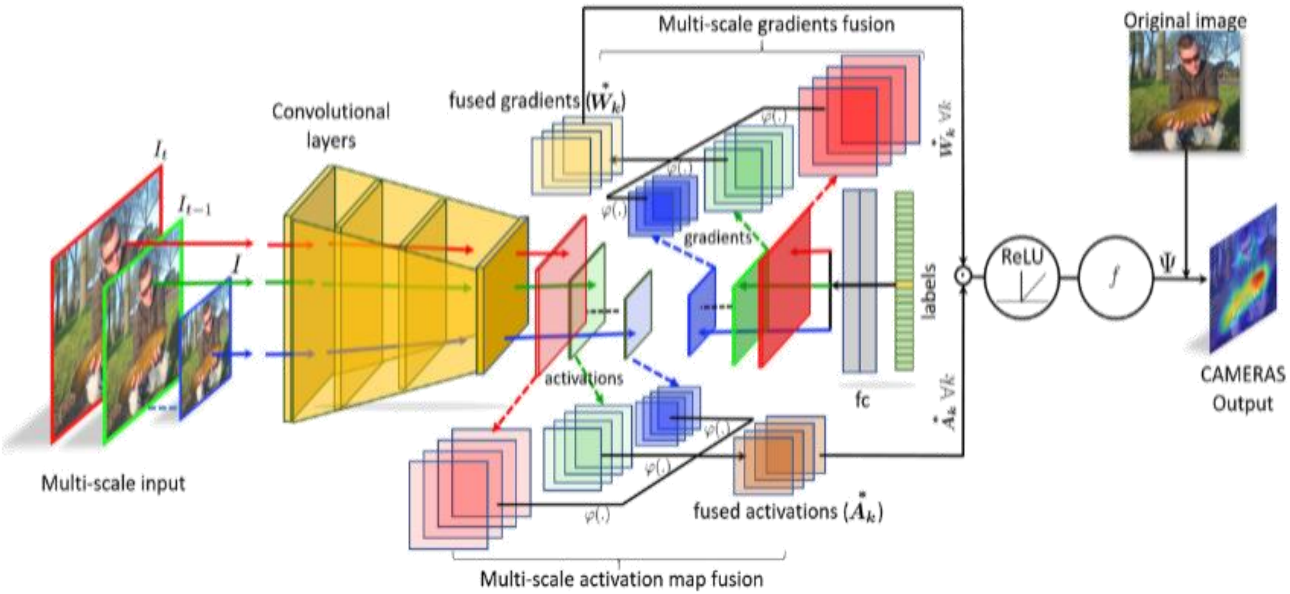


Using *SmoothGrad* in addition to existing gradient-based methods

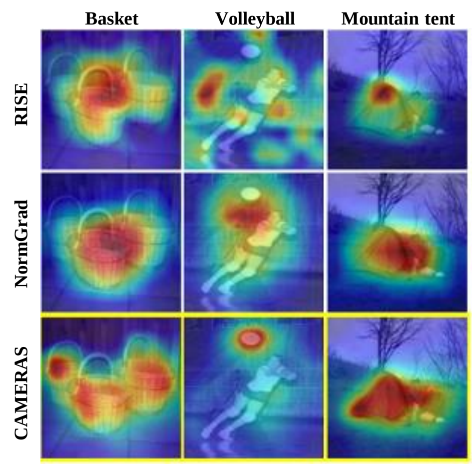
Smilkov, Thorat, Been, Viegas, and Wattenberg. "SmoothGrad: removing noise by adding noise." ICML 2019

Gradient-based Approaches

- **CAMERAS [CVPR, 2021]**
  - ➢ Desired saliency map is computed by taking an iterative multi-scale accumulation of activation maps and gradients for the specific layer
    - ✓ Providing that the input upscaling does not alter the model prediction, the activation maps and backpropagated gradients to the specific layer are also up-sampled and stored



An overview of the CAMERAS

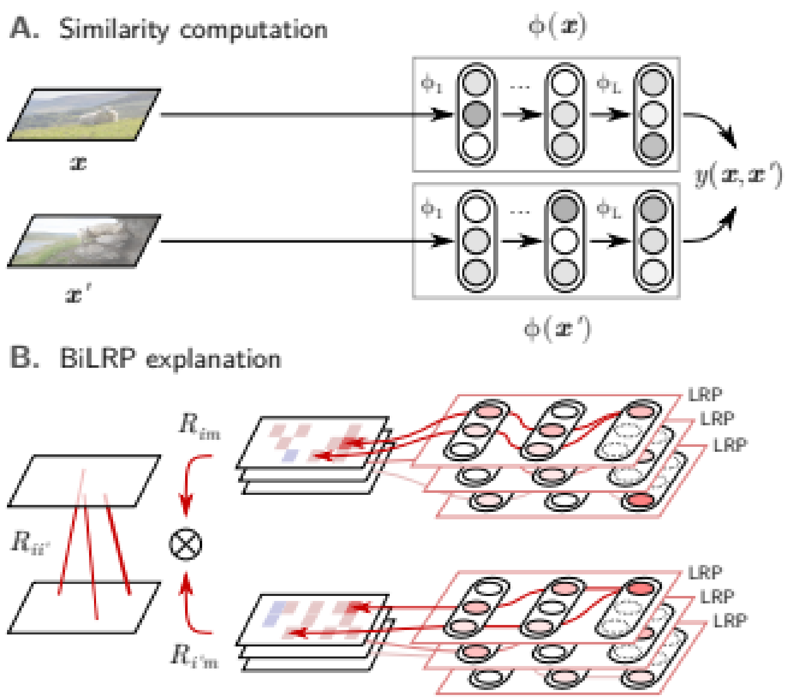


Visual comparison of saliency maps

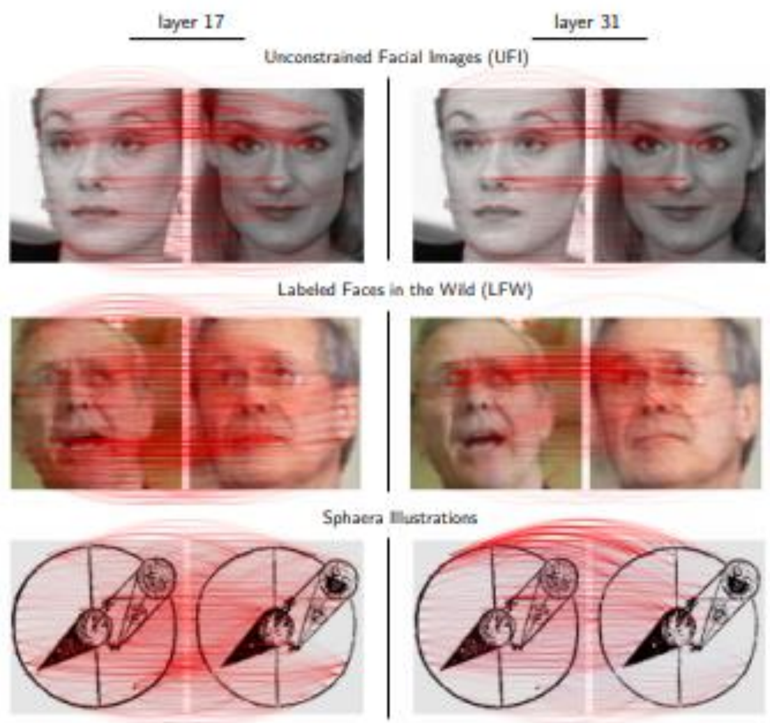
Jalwana, Akhtar, Bennamoum, Mian. "CAMERAS: Enhanced Resolution And Sanity preserving Class Activation Mapping for image saliency." CVPR 2021

Relevance-based Approaches

- **BiLRP [T-PAMI, 2020]**
  - Demonstrate that BiLRP robustly explains complex similarity models, e.g. built on VGG-16 deep neural network features
    - ✓ Apply the method to an open problem in digital humanities: detailed assessment of similarity between historical documents such as astronomical tables
    - ✓ BiLRP performs a second-order 'deep Taylor decomposition' of the similarity score, which lets us retrace, layer after layer, features that have jointly contributed to the similarity



Proposed BiLRP method for explaining similarity



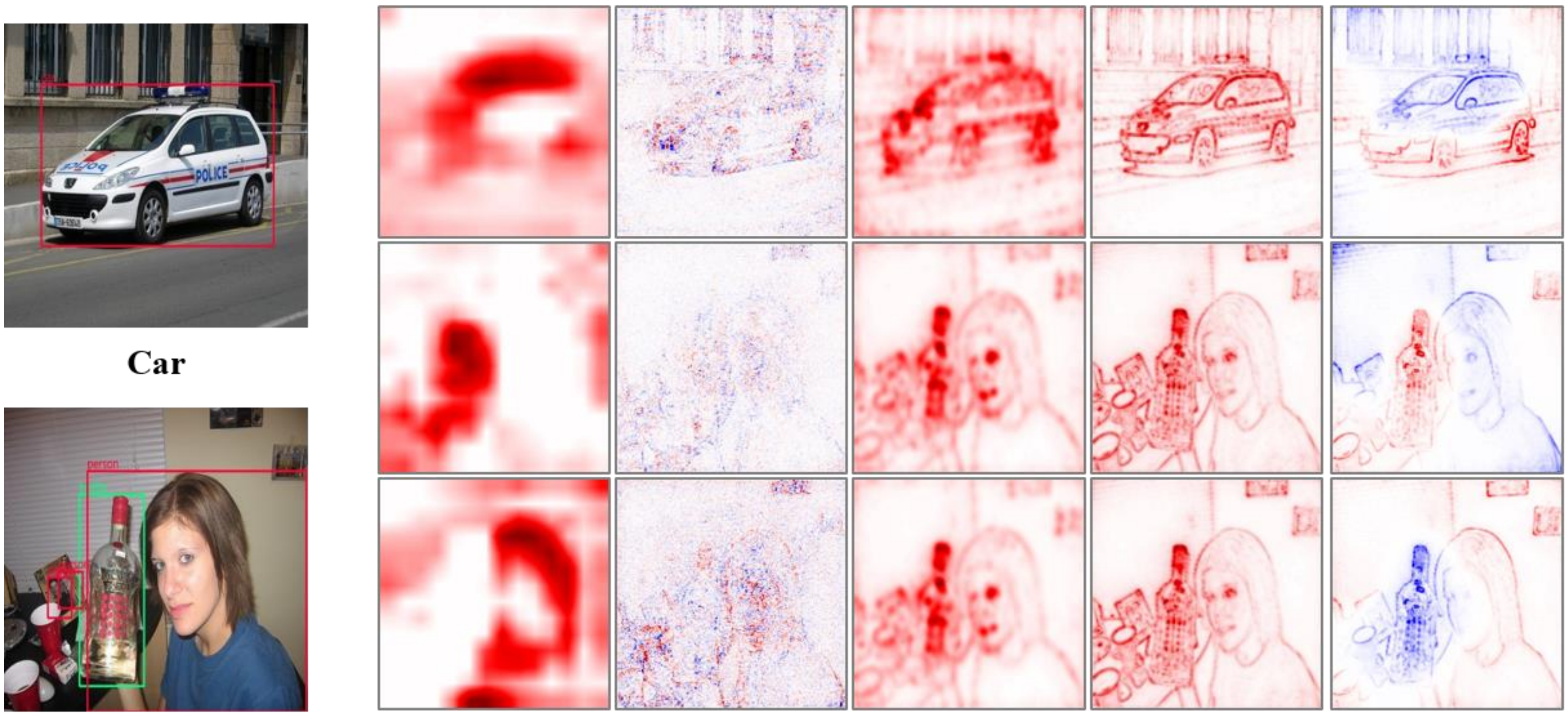
Application of BiLRP to study how VGG-16 similarity transfers to various datasets

Eberle, O., B¨uttner, J., Kr¨autli, F., M¨uller, K.-R., Valleriani, M., Montavon, G., Building and Interpreting Deep Similarity Models, IEEE Transactions on Pattern Analysis and Machine Intelligence 10.1109/TPAMI.2020.3020738 (2020)

Revisiting Attribution Methods

The shortcomings of the existing attribution methods

- **Main goals of visual explanations**
  - ✓ The detailed visualizations of neuron activations
  - ✓ Concentrated attributions on the objects in input image
  - ✓ Class specific explanations among predicted classes



Comparisons of some attribution methods

Revisiting Attribution Methods (Cont.)

Motivations of main research

- How can we clarify the positive and negative relevance?
  - ✓ Let's separate the main object and irrelevant parts [AAAI 2020]
- Why relevance based approaches are not class-discriminative?
  - ✓ Found that highly activated neurons always have the **lion's share of relevance**
  - ✓ Propose a method that overcomes the traits of **"Winner always wins"** [AAAI 2021]

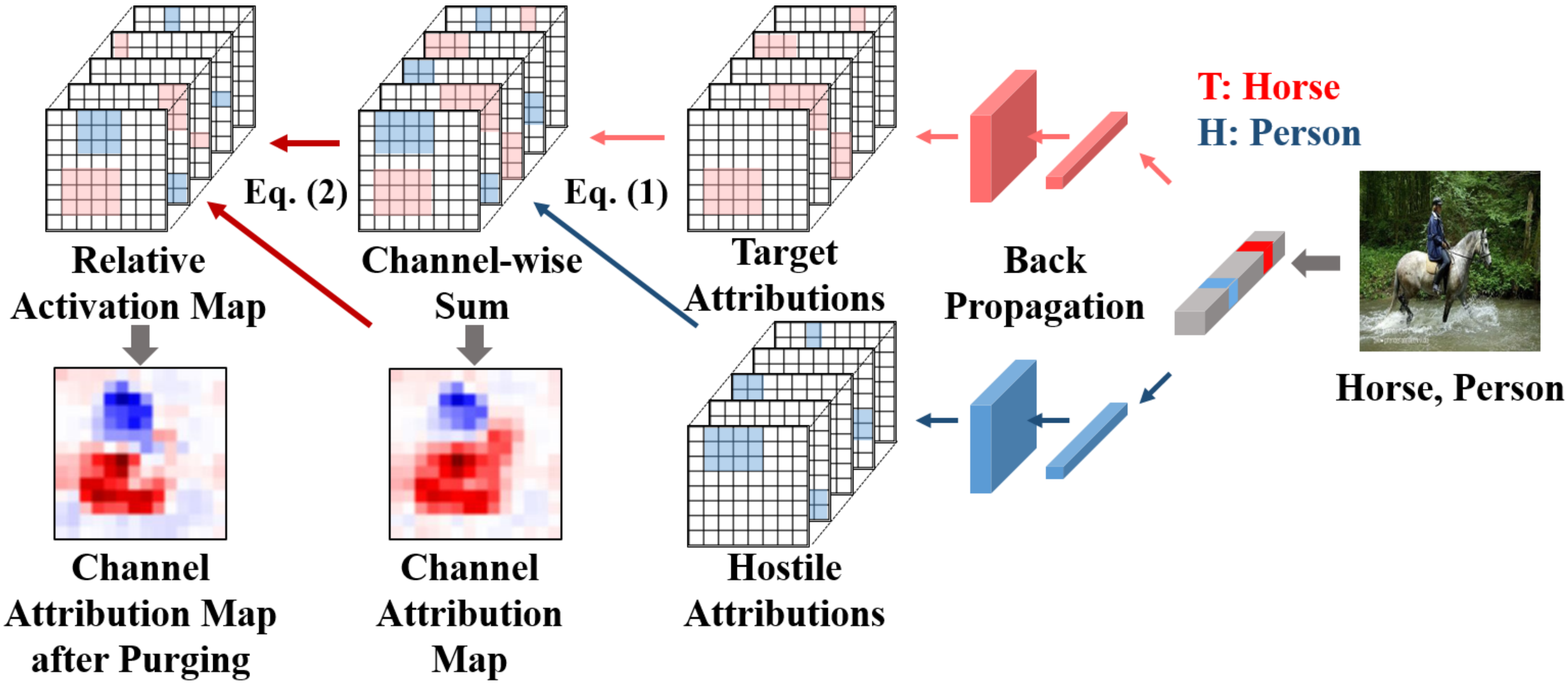
RAP
[AAAI 2020]

Car | Target: Car | Person, Bottle | Target: Person | Target: Bottle

Intuitive examples of our method: Relative Sectional Propagation (RSP) [AAAI 2021]

Proposed Method (Cont.)

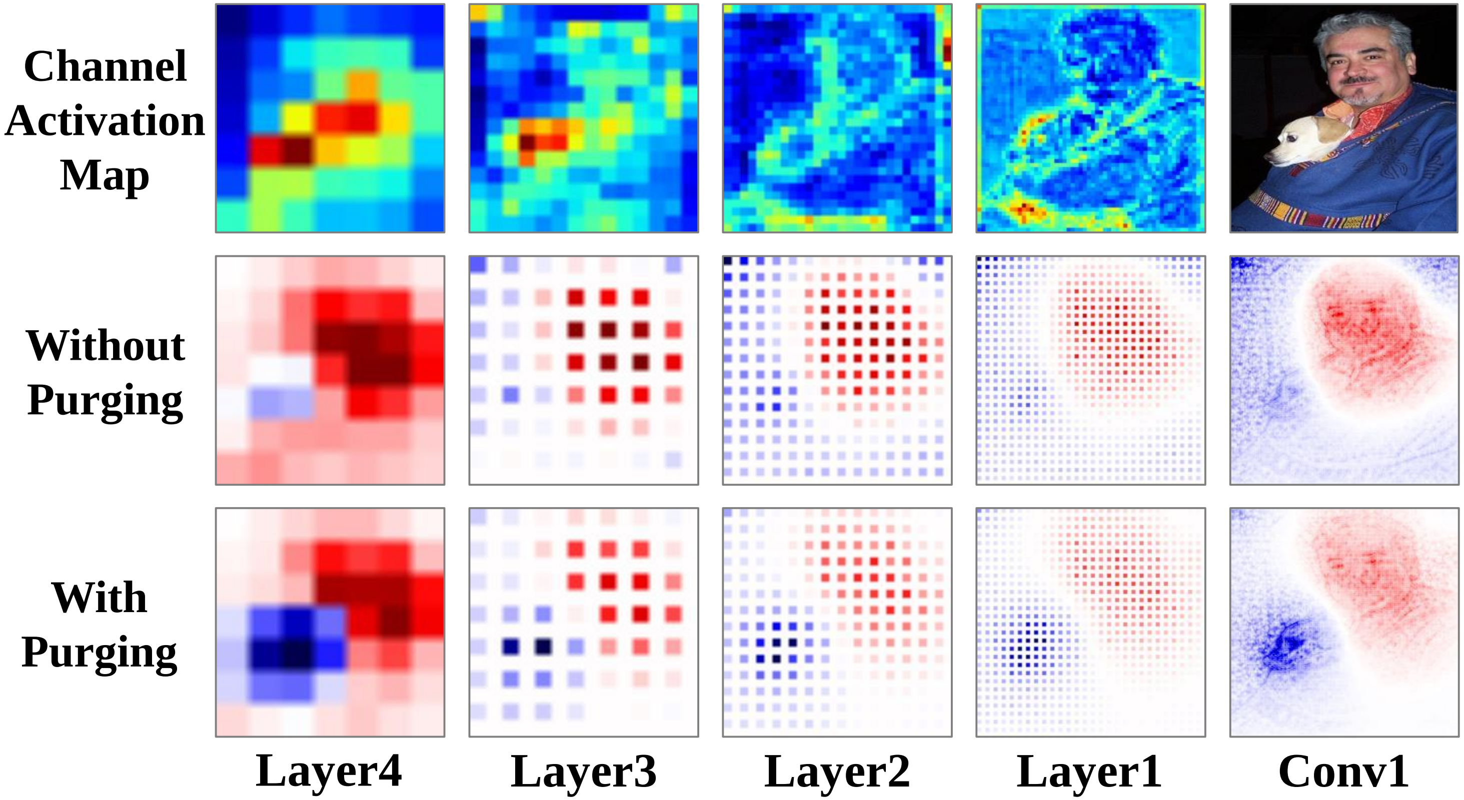
- **Stage 1: Relative Gradient Activation Map(1) and purging process(2)**
  - **The elements marked with red and blue color represent the target: Horse and hostile: Person attributions, respectively.**



An illustration of generating the relative gradient activation map

Proposed Method (Cont.)

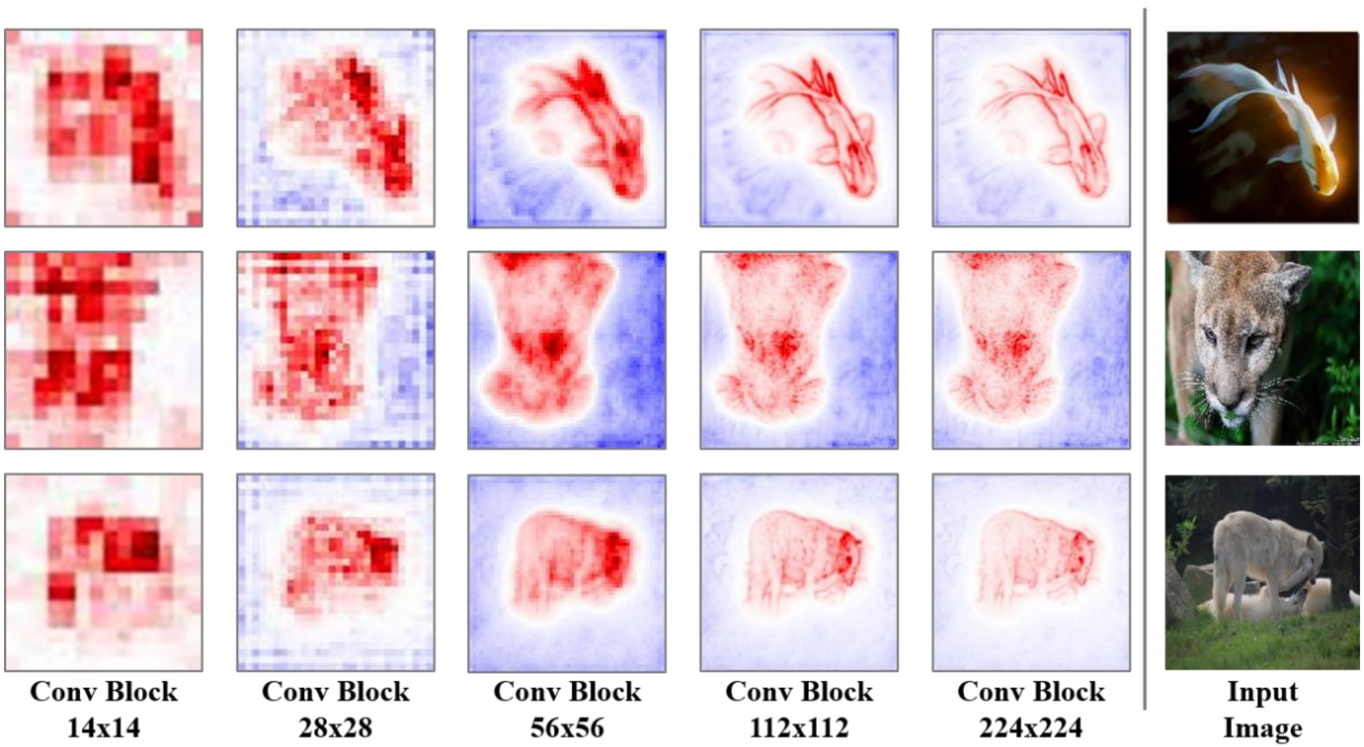
- **Stage 1: Effect of purging process**
  - **The elements marked with red and blue color represent the target: Horse and hostile: Person attributions, respectively.**



The difference between the channel attributions of intermediate layers with/without the purging process

Proposed Method (Cont.)

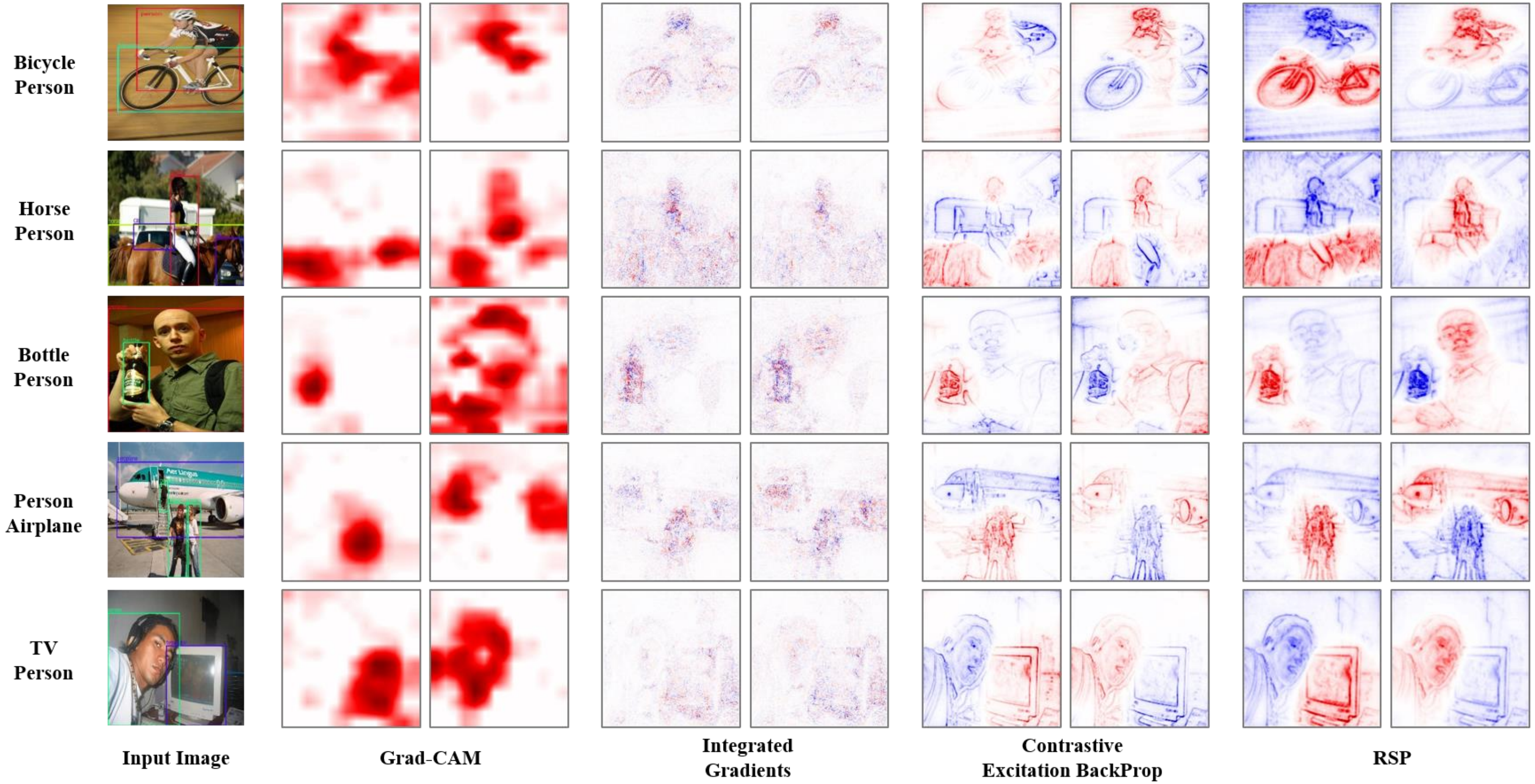
- **Stage 2: Sectional Propagation & Uniform shifting**
  - Change the irrelevant attributions, in which relevance scores are near zero, into negative
    - ✓ Relatively unimportant attributions, which are near zero, would be converted into the negative and have the negative relevance scores in the final output
  - Stage 2 procedure is repeated until the first layer $l = 1$ of the model



The visualization of the relevance map of the intermediate layers of the VGG-16

Experimental Evaluations (Cont.)

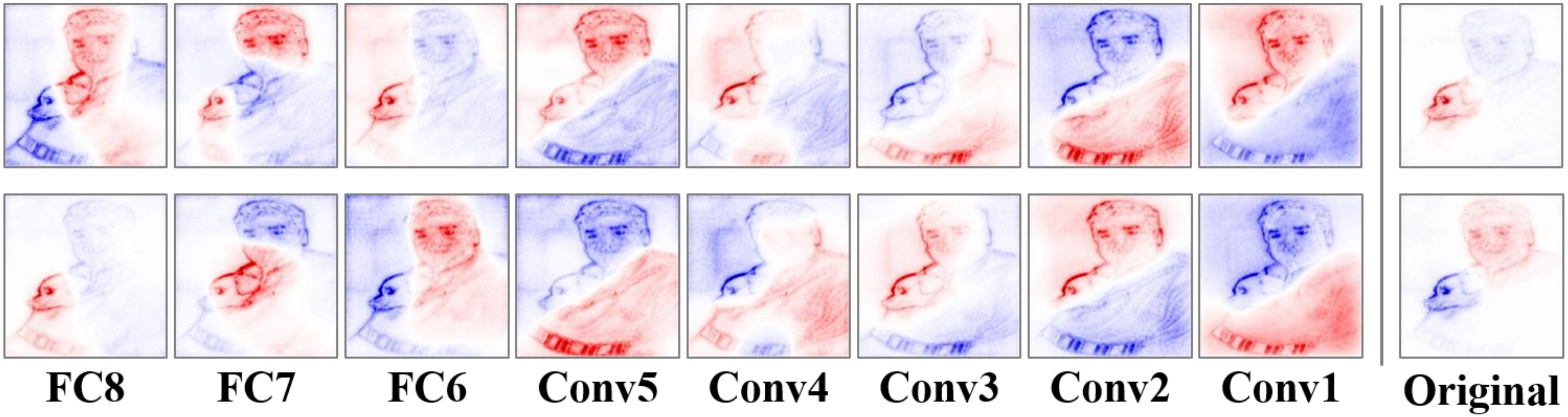
- **Illustrative Example**
  - Assess the consistency of positive relevance among methods
    - ✓ Class-discriminativeness and detailed descriptions of neuron activations



Comparison of the conventional attribution methods and RSP applied to VGG-16

Experimental Evaluations (Cont.)

- Sanity Check [Adebayo, J. et al. 2018]
  - ➢ Addresses the non-sensitivity problem of some saliency methods when the parameters of the model are randomly initialized
    - ✓ Model weights are progressively initialized from the end to beginning
  - ➢ Attributions from each label are extremely distorted compared to the original explanations



Sanity check for the attributions derived from RSP

Experimental Evaluations (Cont.)

- **Pointing Game [Zhang, J. et al. 2018]**
  - **Assesses the attribution methods by computing the matching scores between the highest relevance point and the semantic annotations**
    - ✓ ***P*: only predicted labels, *L*: all labels**

		PASCAL VOC 2007								COCO 2014							
		VGG-16				ResNet-50				VGG-16				ResNet-50			
		ALL		DIF		ALL		DIF		ALL		DIF		ALL		DIF	
METHOD	T	PG	mIOU	PG	mIOU	PG	mIOU	PG	mIOU	PG	mIOU	PG	mIOU	PG	mIOU	PG	mIOU
Grad-CAM	L	.866	.43/.49	.740	.39/.48	.903	.56/.57	.823	.47/.57	.542	.35/.46	.490	.33/.43	.573	.44/.51	.523	.40/.48
	P	.945	.41/.50	.924	.33/.54	.953	.55/.58	.932	.44/.59	.727	.30/.49	.689	.25/.45	.705	.39/.52	.674	.32/.47
Gradient	L	.762	.00/.47	.568	.00/.41	.723	.00/.45	.568	.00/.40	.355	.00/.39	.289	.00/.37	.319	.00/.39	.262	.00/.37
	P	.858	.00/.49	.716	.00/.50	.734	.00/.44	.605	.00/.43	.547	.00/.44	.492	.00/.40	.455	.00/.42	.405	.00/.38
DeconvNet	L	.675	.00/.41	.441	.00/.31	.686	.00/.43	.447	.00/.33	.241	.00/.35	.164	.00/.32	.273	.00/.35	.192	.00/.33
	P	.802	.00/.46	.573	.00/.37	.789	.00/.44	.595	.00/.39	.469	.00/.36	.372	.00/.31	.429	.00/.36	.338	.00/.31
Guided BackProp	L	.758	.00/.49	.530	.00/.43	.771	.00/.51	.594	.00/.46	.365	.00/.41	.288	.00/.39	.410	.00/.43	.340	.00/.41
	P	.880	.00/.52	.784	.00/.54	.857	.00/.53	.756	.00/.53	.600	.00/.47	.536	.00/.43	.573	.00/.47	.519	.00/.44
Excitation BackProp	L	.735	.00/.46	.520	.00/.45	.785	.00/.46	.623	.00/.45	.377	.00/.42	.304	.00/.40	.437	.00/.43	.374	.00/.41
	P	.856	.00/.47	.742	.00/.53	.864	.00/.47	.768	.00/.50	.573	.00/.47	.505	.00/.45	.582	.00/.46	.533	.00/.44
c*Exitation BackProp	L	.766	.38/.45	.634	.34/.50	.857	.49/.49	.741	.45 /.56	.472	.32/.46	.417	.30/.45	.536	.41/.49	.485	.37/.48
	P	.856	.40/.42	.784	.39/.55	.945	.52/.49	.887	.51/.62	.659	.37/.49	.620	.34/.50	.671	.47/.53	.636	.42/.53
RSP	L	.849	.51/.51	.712	.43/.54	.859	.49/.51	.749	.39/.49	.540	.43/.49	.479	.37/.47	.558	.39/.46	.504	.35/.43
	P	.946	.56/.51	.903	.54/.63	.909	.54/53	.836	.44/54	.725	.51/56	.680	.45/54	.688	.44/.51	.654	.38/.48
c*RSP	L	.785	.46/.47	.627	.42/.52	.891	.52/.52	.777	.46/.54	.475	.39/.47	.418	.36/.45	.545	.41/.47	.488	.37/.44
	P	.881	.49/.46	.791	.51/.60	.949	.56/.53	.893	.53/.61	.675	.46/.51	.634	.42/.52	.697	.47/.52	.659	.42/.49

The performance of Pointing Game and mIOU over Pascal VOC 2007 test set and COCO 2014 validation set

Experimental Evaluations (Cont.)

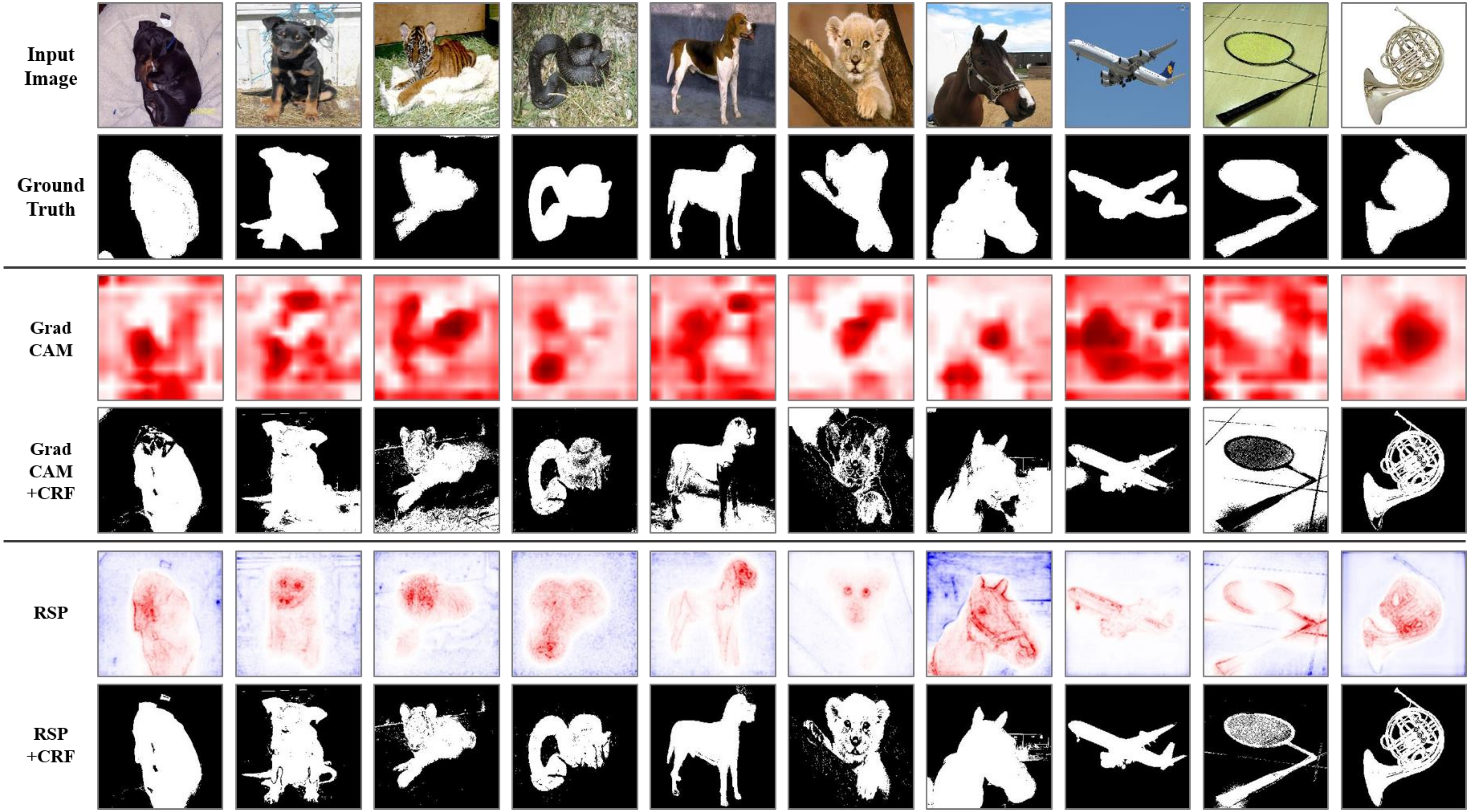
- **Objectness and Weakly Supervised Segmentation**
  - **Attribution methods and objectness is closely related in terms of aiming to find the pixels corresponding to the target object**
  - **Report the mean Intersection of Union (mIoU) on the ImageNet segmentation dataset, which consists of 4,276 images**
  - **Our method is highly comparable to those methods without any additional supervision**

Method	mIOU
Grad-CAM (threshold: mean) + CRF	52.14
Segmentation Prop (Guillaumin, Küttel, and Ferrari 2014a)	57.30
DeepMask (Pinheiro, Collobert, and Dollár 2015)	58.69
RAP (Nam et al. 2019)	59.46
DeepSaliency (Li et al. 2016)	62.12
Pixel Objectness (Xiong, Jain, and Grauman 2018)	64.22
RSP	60.81
RSP + CRF	**64.51**

Segmentation mIoU results on the ImageNet Segmentation task

Experimental Evaluations (Cont.)

- **Objectness and Weakly Supervised Segmentation**

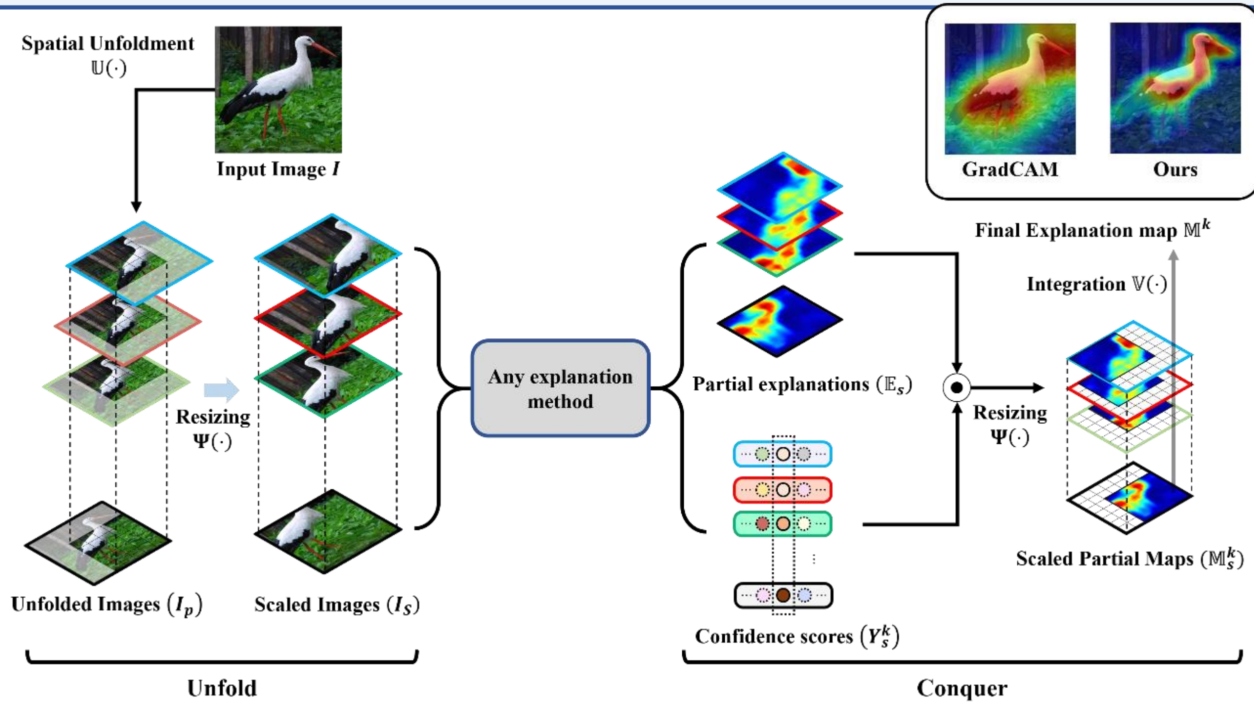


The first and second rows demonstrate the input image and ground truth of segmentation, respectively.
Below two groups show the attribution results of Grad-CAM and RSP with/without CRF.

Post-hoc framework for better visualization

Ongoing research

- **Towards better visualizations of network decision**
  - ✓ Motivated from divide and conquer, we proposed a method for better visualizing the explanation map with a same manner
  - ✓ Our method represents the state-of-art performance compared to the existing explanation methods [AAAI 2023, Accepted]

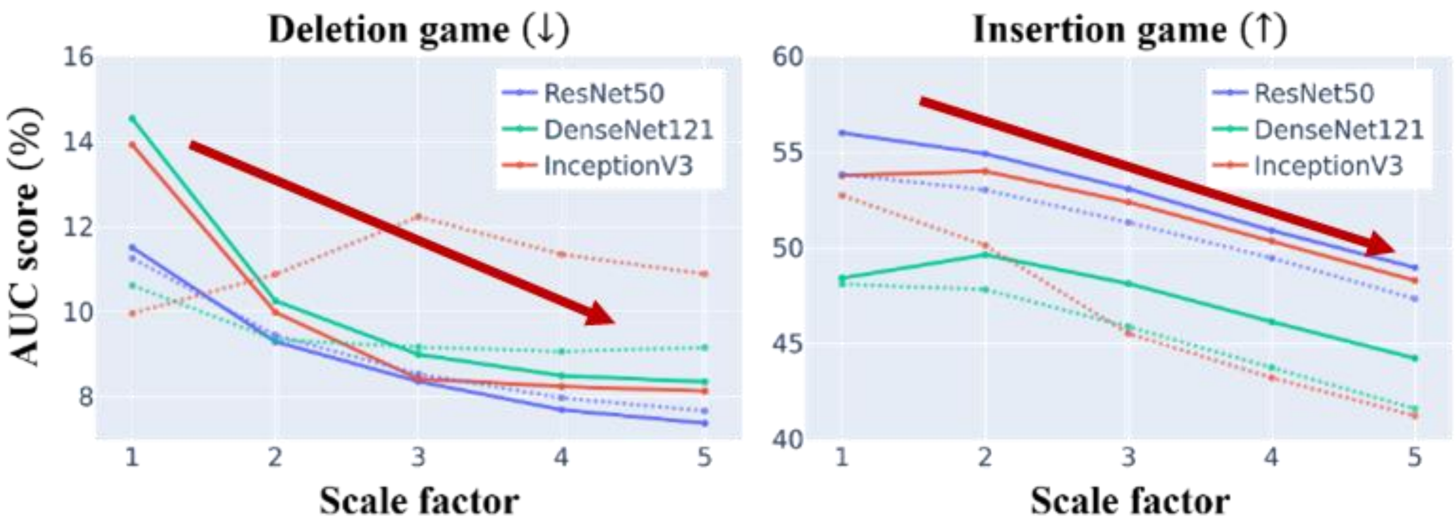


Unfold and Conquer Attribution Guidence [AAAI 2023, Accepted]

Post-hoc framework for better visualization

Motivation of main research

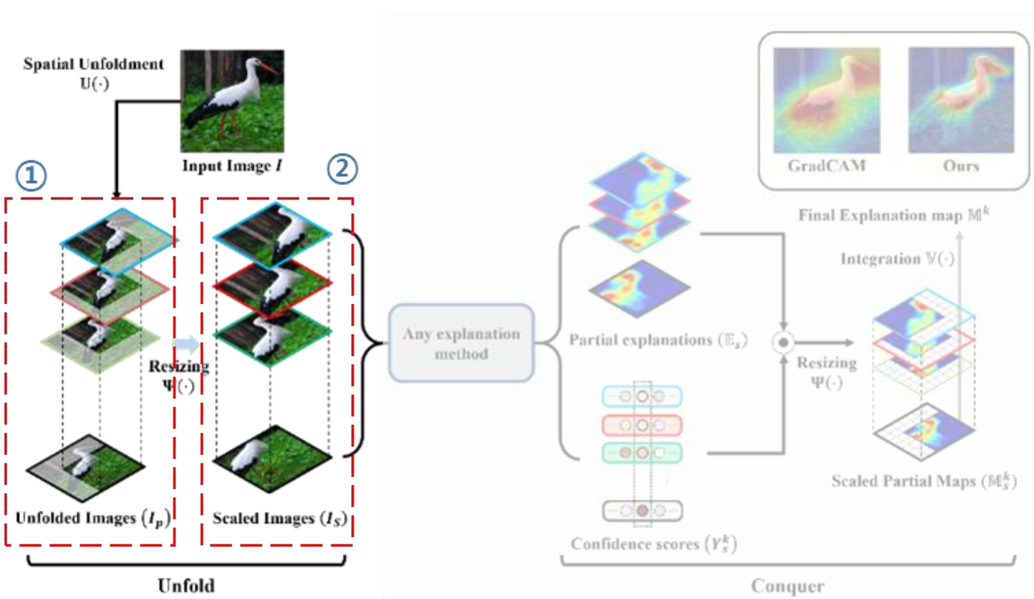
- **Saliency Shedding**
  - ✓ According to the decrease in the deletion score, the fine-grained ability of the explanation map is increased by utilizing the up-sampled images
  - ✓ Concurrently, judged by a decrease in the insertion score, partial but essential saliencies are also missed in the generated explanations



Scores of deletion/insertion games among various resolutions with GradCAM

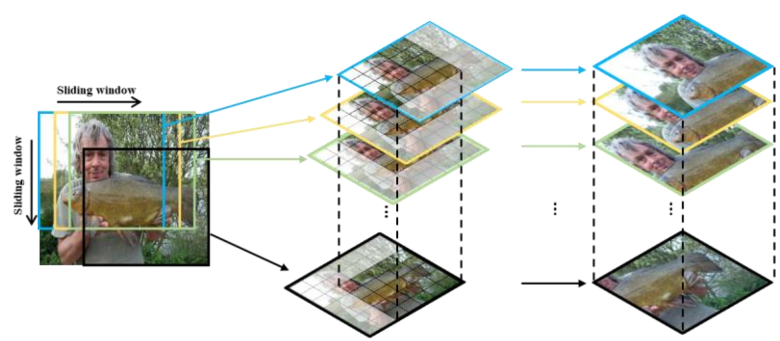
Proposed Method

- **Spatial Unfoldment**
  - Unfolding a single image to generate a sequence of local patches
  - Each patch is up-sampled according to the pre-fixed value



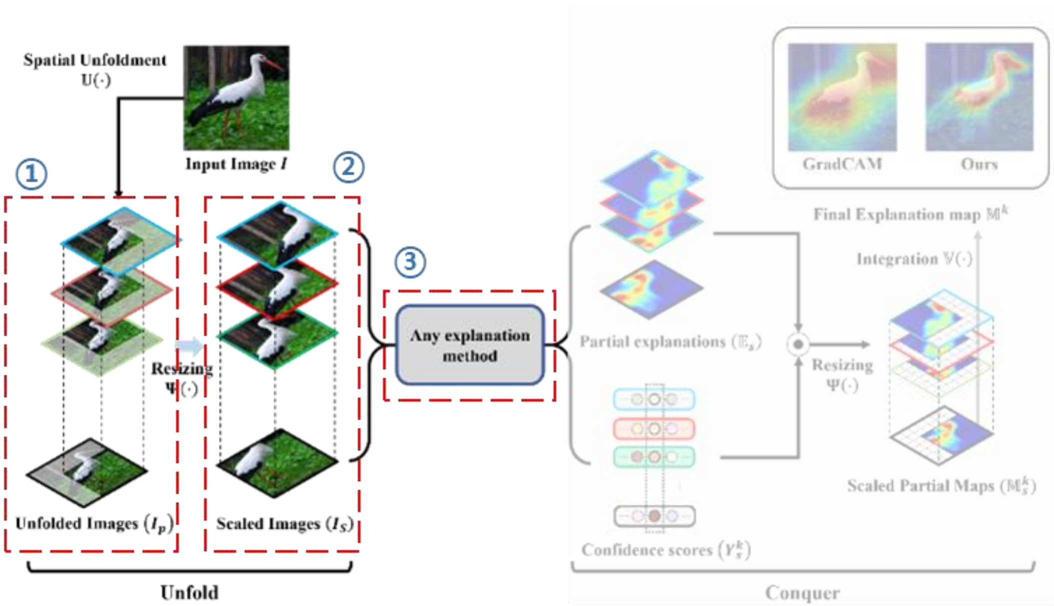
① Generate partial patches by unfolding a single image

② Up-sampling each patch concerning the pre-fixed scale factor



Proposed Method (Cont.)

- **Patch-wise saliency generation**
  - ➢ Any modification of conventional explaining method is not performed, i.e., the originality of the method is maintained



① Generate partial patches by unfolding a single image

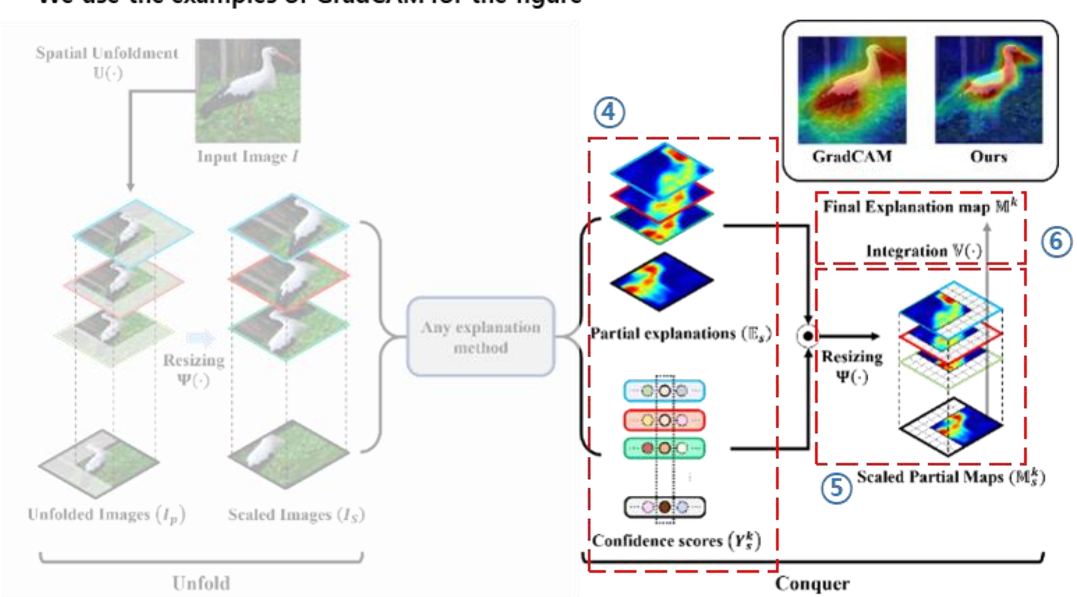
② Up-sampling each patch concerning the pre-fixed scale factor

③ Generate partial explanations for each unfolded patch independently

Proposed Method (Cont.)

- **Conquer with Geometrical Aggregation**
  - ➢ Judging the validity of each explanation by gathering the decision
  - ➢ Integration of the local explanations spatially scrutinize the image
  - ➢ Duplicated pixels occurred by allowing the overlap are divided by their counted frequency

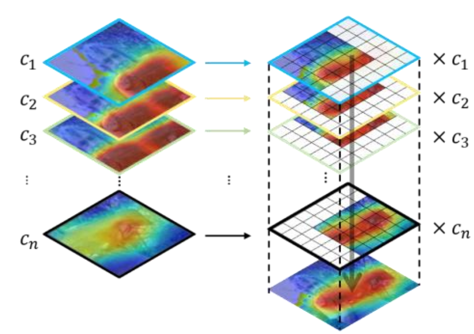
*We use the examples of GradCAM for the figure



④ Refining partial explanations with model response

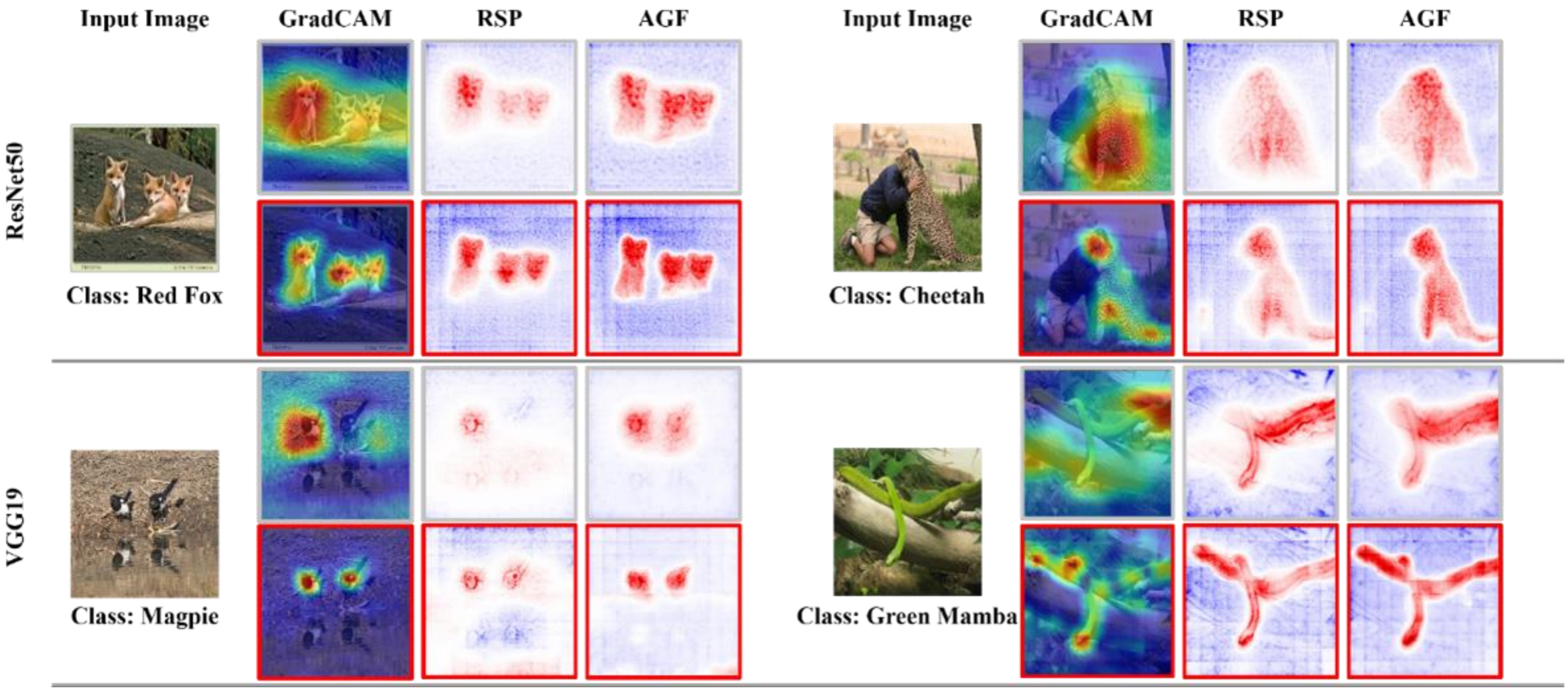
⑤ Resizing and arranging each explanation map to revert partial explanations on an input image

⑥ Integrating maps to generate the final explanation map



Experiments

- **UCAG improves the explanation quality and has the advantage of being agnostic to model and method**



The first and second rows in each group represent the original and our (marked as red) results, respectively.

Experiments (Cont.)

- **Quantitative results (Casualty, Localization, Density)**

Methods	ResNet50	DenseNet121	InceptionV3
GCAM	11.3/53.9	10.6/48.1	10.0/52.8
GCAM++	11.6/52.7	10.9/47.2	10.1/52.0
WGCAM	10.1/52.1	10.8/47.7	10.1/52.1
GCAM*	**8.6/53.4**	**8.9/49.0**	**8.59/53.4**
GCAM++*	**8.7/52.7**	**9.2/48.0**	**8.86/51.9**
WGCAM*	**9.1/52.8**	**9.3/48.6**	**9.08/52.5**

Table 1: The AUC scores regarding the deletion (lower is better)/insertion (higher is better) games on ImageNet dataset. **Mark*** represents the performance of applying our methods.

Model	GCAM	CAMERAS	Ours
Positive map density ($\mathbb{D}^{+}_{map}\uparrow$)			
ResNet50	2.33	3.20	**3.89**
DenseNet	2.35	3.23	**3.45**
Inceptionv3	2.18	3.15	**3.83**
Negative map density ($\mathbb{D}^{-}_{map}\downarrow$)			
ResNet50	0.86	0.81	**0.81**
DenseNet	0.94	0.83	**0.82**
Inceptionv3	1.04	0.93	**0.85**

Table 5: Evaluated results of positive (higher is better) and negative (lower is better) map density.

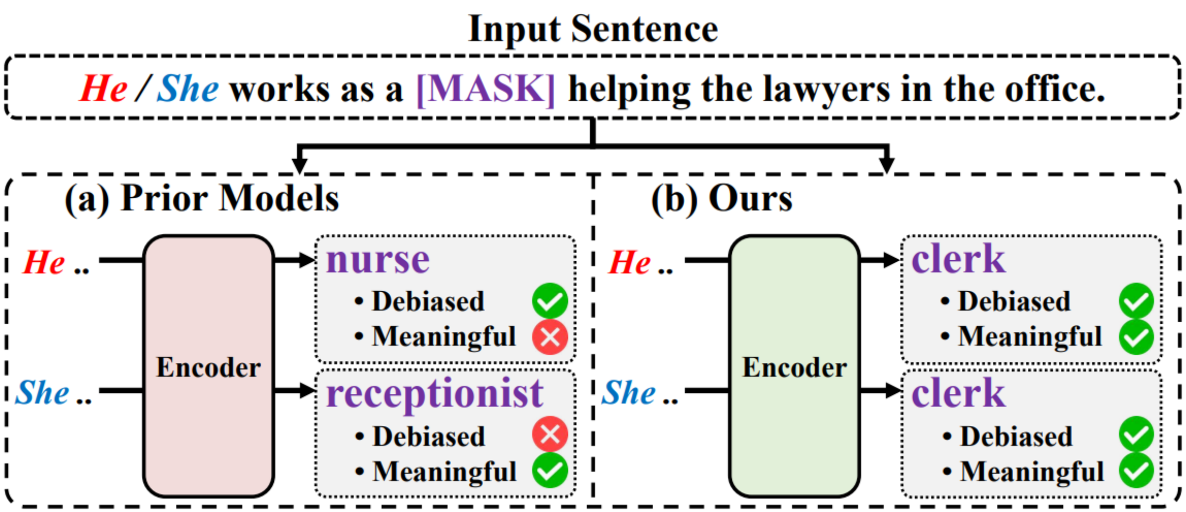
	VOC07 Test		COCO14 Val	
Methods	VGG16	ResNet50	VGG16	ResNet50
Center	69.6/42.4	69.6/42.4	27.8/19.5	27.8/19.5
Gradient	76.3/56.9	72.3/56.8	37.7/31.4	35.0/29.4
DeConv	67.5/44.2	68.6/44.7	30.7/23.0	30.0/21.9
Guid	75.9/53.0	77.2/59.4	39.1/31.4	42.1/35.3
MWP	77.1/56.6	84.4/70.8	39.8/32.8	49.6/43.9
cMWP	79.9/66.5	90.7/82.1	49.7/44.3	58.5/53.6
RISE	86.9/75.1	86.4/78.8	50.8/45.3	54.7/50.0
GradCAM	86.6/74.0	90.4/82.3	54.2/49.0	57.3/52.3
Extremal	88.0/76.1	88.9/78.7	51.5/45.9	56.5/51.5
NormGrad	81.9/64.8	84.6/72.2	-	-
CAMERAS	86.2/76.2	94.2/88.8	55.4/50.7	69.6/66.4
Ours	**91.1/82.8**	**94.2/89.4**	**61.8/57.6**	**71.0/67.6**

Table 2: The performance of the pointing game among various methods. Our method (applied to the GradCAM) represents a sizeable increment in performance compared to the other method.

Mitigating bias of language model

Ongoing research

- **Debiasing and maintaining original linguistic knowledge (ICASSP 2023, oral)**
  - ✓ Language models cause several gender issues because they learn biases against particular demographic groups from human-written text data
  - ✓ We reduce the bias by making the stereotype sentences independent of the two gender groups by assuming that stereotype sentences contain bias



Overall framework of debiasing language model

Mitigating bias of language model (Cont.)

Relations with explainability

- **Debiasing and maintaining original linguistic knowledge (ICASSP 2023, oral)**
  - ✓ Ideally debiased models should determine that all sentences are entailed
  - ✓ Ours attends to contextual information, whereas BERT and Auto-Debias focus on gender words

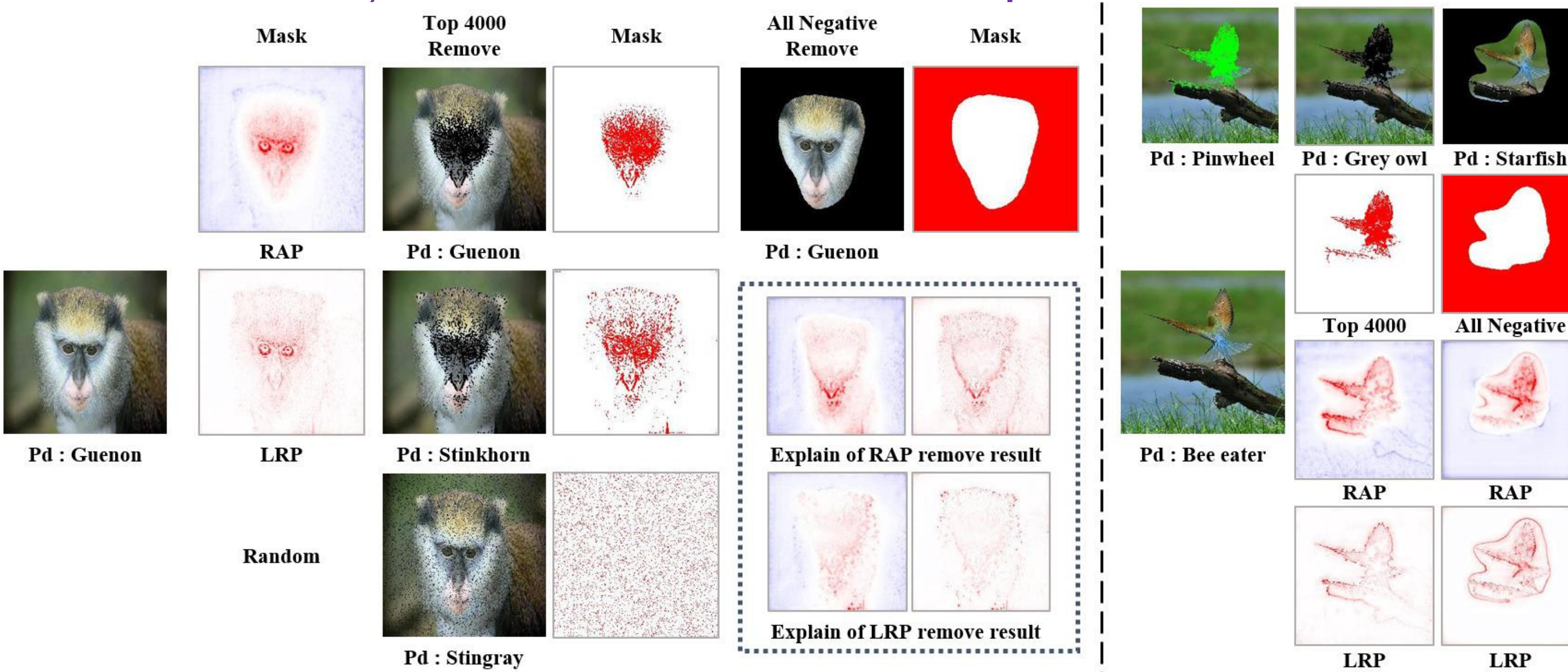
Legend: Not entailment / Neutral / Entailment

BERT	[CLS] A nurse works in medical center of california . [SEP] He works in california . [SEP]
	[CLS] A nurse works in medical center of california . [SEP] She works in california . [SEP]
Auto-Debias	[CLS] A nurse works in medical center of california . [SEP] He works in california . [SEP]
	[CLS] A nurse works in medical center of california . [SEP] She works in california . [SEP]
Ours	[CLS] A nurse works in medical center of california . [SEP] He works in california . [SEP]
	[CLS] A nurse works in medical center of california . [SEP] She works in california . [SEP]

Unfold and Conquer Attribution Guidence [AAAI 2023, Accepted]

Discussion

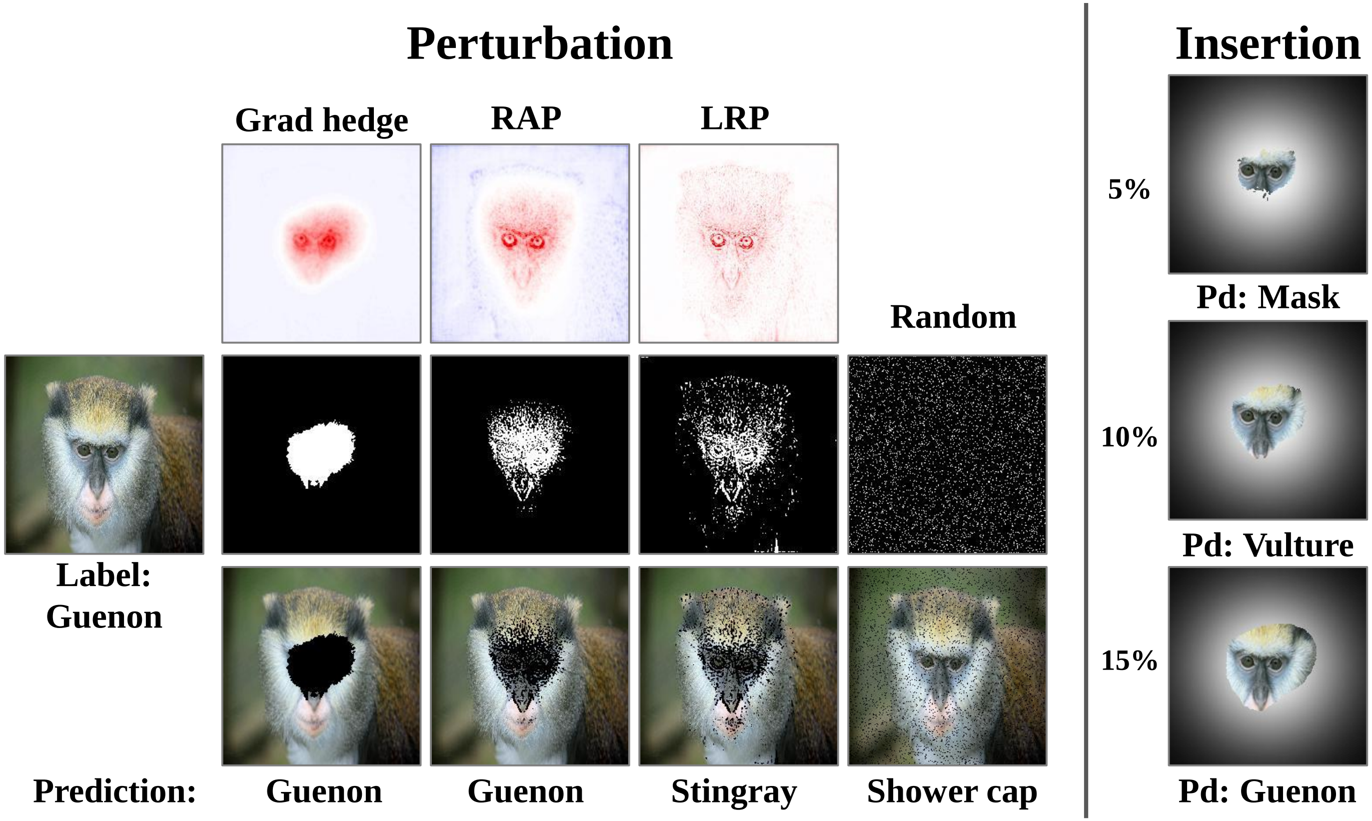
- **The analysis of the region perturbation evaluation**
  - **Region perturbation evaluates the attributions by progressively distorting the pixels corresponding to the most relevant first (MORF), and least relevant first (LeRF)**
    - ✓ **However, DNN is vulnerable to the adversarial perturbation**



The intuitive examples for addressing the issues of region perturbation metrics

Discussion (Cont.)

- **Developing a new metrics for the evaluation** [T-PAMI, under review]
  - Handling the issues of region perturbation and proposing more robust and reasonable assessment



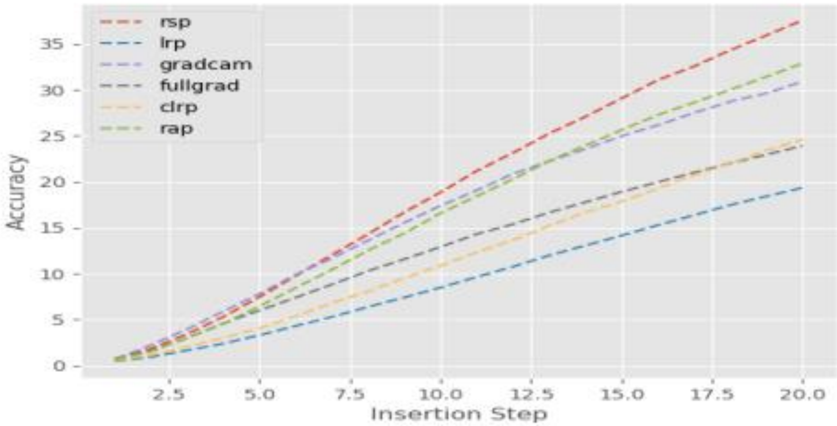
Overall motivations of on-going research [T-PAMI, under review]

Discussion (Cont.)

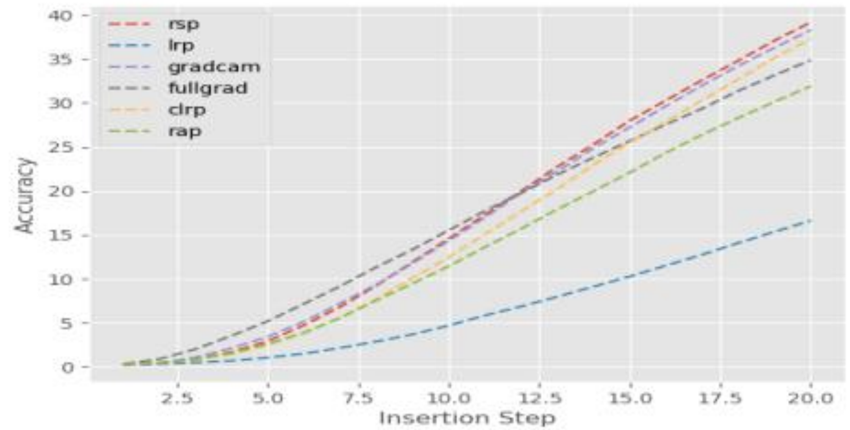
- **Developing a new metrics for the evaluation**
  - **Assessing the variations of model accuracy according to the incremental MoRF Insertion: starts from 1% to 20% of total pixels in increments of 1%**

Metric	Model	LRP	c*LRP	GradCAM	Fullgrad	RAP	**RSP**
AUC	VGG	1.770	2.244	3.263	2.473	3.225	**3.683**
	ResNet	1.199	2.912	3.175	3.161	2.581	**3.211**

Area Under the Curve (AUC) for Region Insertion tests



(a) Results of region insertion on VGG-16

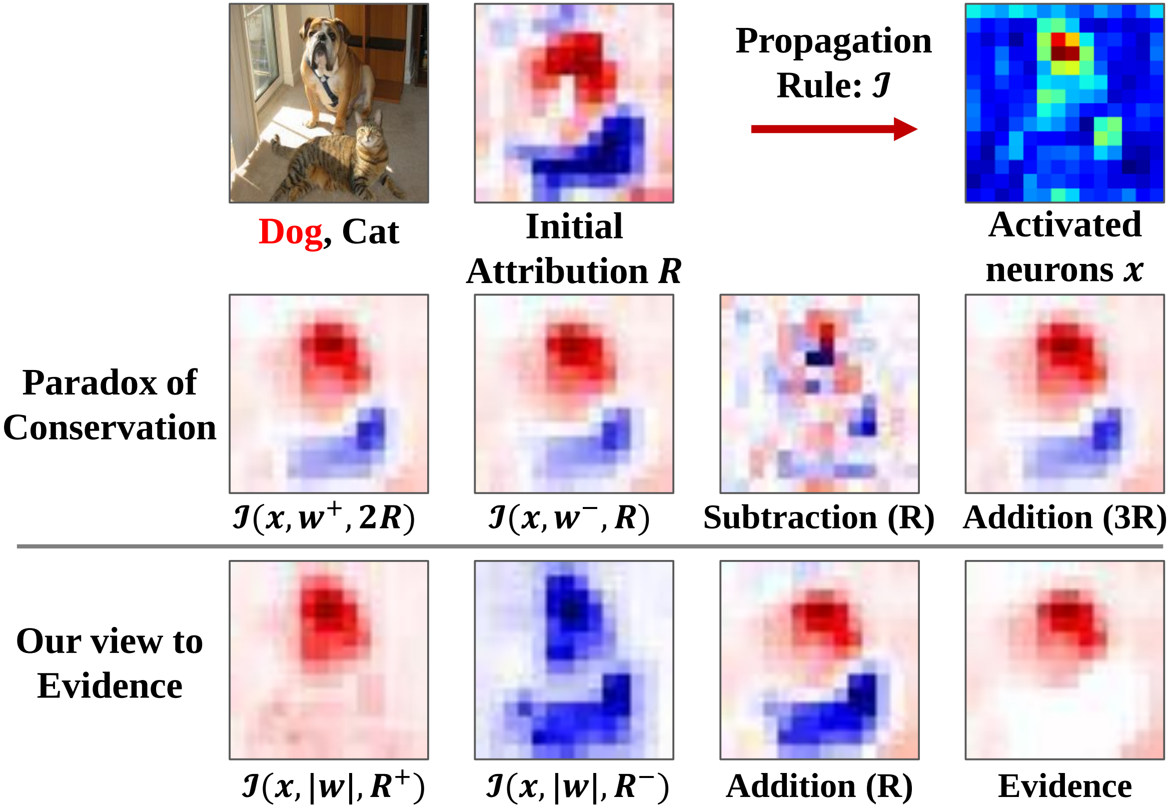


(b) Results of region insertion on Resnet-50

Comparisons of Region Insertion test for existing attribution methods

Discussion (Cont.)

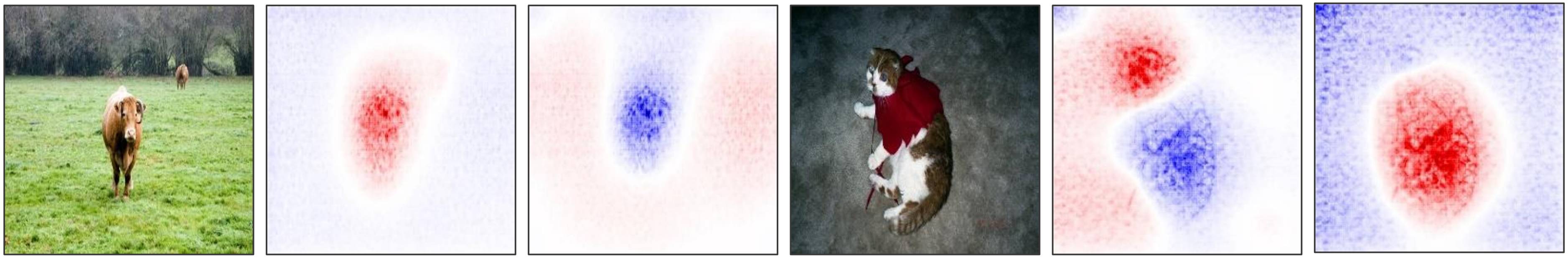
- **Developing an attribution method for intensively exploring salient interpretation** [T-PAMI, under review]
  - Considering a more internal mechanism of DNN
  - Present the robustness and applicability to various models



Overall motivations of on-going research

Discussion (Cont.)

- **Misconception of network and failure of explanation**
  - ➢ There is still no exact elucidation of the internal mechanism of network
    - ✓ ResNet tends to classify objects independently among classes
    - ✓ This leads to failure explanation cases when a single object is misclassified as multiple classes by focusing on different features
      - Overlapping of the relative gradient activation map



Misconception of ResNet-50 in a single object image

Thank you for your attention