Optimization-based Uncertainty Attribution Via Learning Informative Perturbations

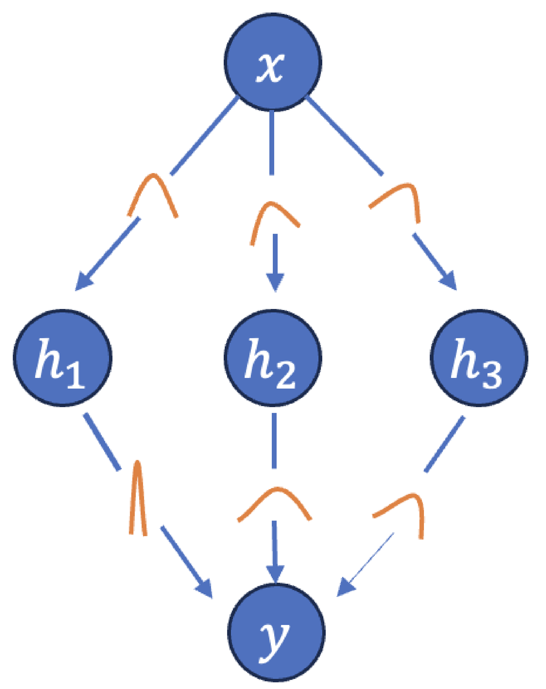
Hanjing Wang (RPI) , Bashirul Azam Biswas (RPI), and Qiang Ji (RPI)

Introduction

- Bayesian deep learning (BDL) model
  - Treat parameters θ as random variables

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

  - Well-founded framework for uncertainty quantification (UQ).



- Uncertainty quantification
  - Seek to determine the confidence in the predictions, given the imperfect inputs
  - UQ can make the model say "No" to the predictions.
  - Two types of uncertainty:
    - Epistemic uncertainty U_e captures the insufficient knowledge of the modeling process.
    - Aleatoric uncertainty U_a occurs due to the data noise.
    - Total uncertainty $U_t = U_a + U_e$
  - The measure of uncertainty: Denote x as the input and y as the output. Give a classification model that outputs $p(y|x,\theta)$, we have

$$\underbrace{\mathcal{H}[p(y|x,D)]}_{\text{Total}} = \underbrace{\mathrm{I}[y,\theta|x,D]}_{\text{Epistemic}} + \underbrace{\mathbb{E}_{p(\theta|D)}[\mathcal{H}[p(y|x,\theta)]]}_{\text{Aleatoric}}$$

- Uncertainty attribution (UA)
  - Focus on understanding and explaining the sources and causes of uncertainty.
  - The proposed method localizes the high uncertain regions to determine "**where is wrong**"?
  - Challenges
    - Not well-explored area
    - Current explainable AI methods focus on attribution of the classification score for deterministic neural networks
    - Gradient-based UA is often noisy and hard to interpret.

Proposed Method: Optimization-based Uncertainty Attribution

- Basic Formulation:

$$M^* = \arg\min_M U\big((1-M)\odot x + M \odot \hat{x}\big) + \lambda||M||$$

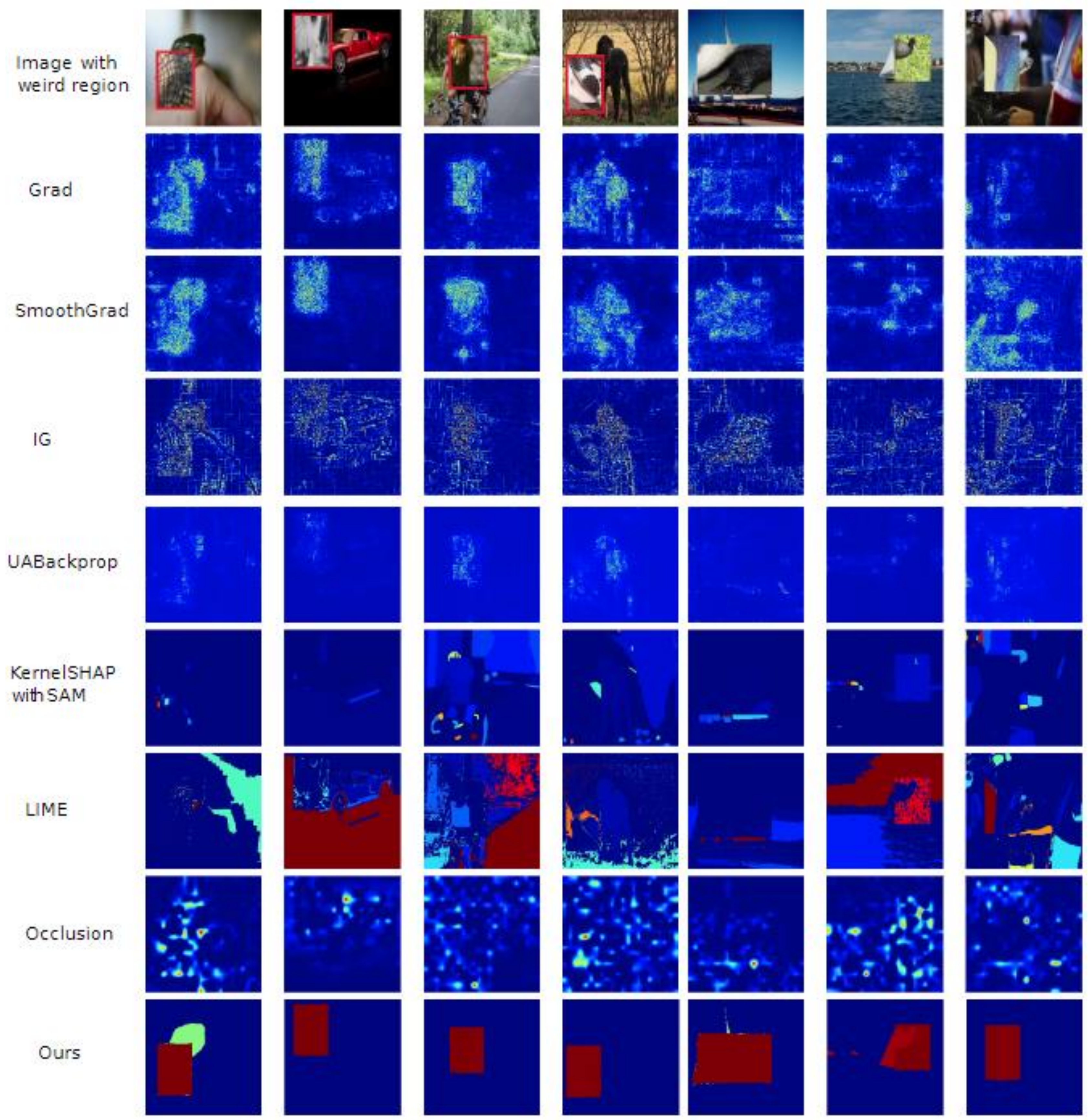
M: the binary mask that highlights areas in the inputs significantly contributing to uncertainty.
U: the function of uncertainty.
$\hat{x}$: the perturbed input with reduced uncertainty.

- Three improvements:
  - SAM-guided Mask Parameterization
    - We parameterize M by a linear combination of segments derived from the pre-trained Segment Anything model (SAM), i.e., $M = \sum_i w_i M_i$
    - Each segment M_i is inherently binary and delineate areas corresponding to semantically meaningful and human-understandable concepts.
  - Learnable Perturbation
    - $\hat{x}$ is learned by $g_\phi(x)$ where $g_\phi(\cdot)$ is a blurring function parameterized by ϕ, allowing for the precise and dynamic adjustment of perturbations.
  - Gumbel-sigmoid Reparameterization For Binary Weights
    - The Binary weight w_i is parameterized using Gumbel-sigmoid function to keep its binary nature under continuous optimization.

Quantitative Experiments

- Detection of Problematic Regions
  - Evaluate methods for detecting image anomalies using semi-synthetic data with known problematic areas.
  - Quantitative metrics include Intersection over Union (IoU) and anomaly detection accuracy (ADA).
- Faithfulness Test
  - Faithfulness in uncertainty attribution quantifies the accuracy with which a method's explanations reflect the actual influence of input features on the model's uncertainty.
  - We refine the most problematic pixels of the input, for example, by updating 2% of the pixels with the highest UA scores, and then observe the reduction in uncertainty after the alteration

Uncertainty Attribution Maps Examples



Method	C10	
	IoU	ADA
Grad	0.109	0.054
SmoothGrad	0.297	0.226
IG	0.208	0.152
UA-Backprop	0.271	0.178
KernelSHAP	0.324	0.306
KernelSHAP + SAM	0.676	0.670
LIME	0.281	0.258
Occlusion	0.218	0.160
Ours	**0.713**	**0.700**

Method	Refining using Gaussian Blurring						
	C10		C100		SVHN		Avg. Performance
	2%	5%	2%	5%	2%	5%	2% + 5%
Grad	0.343	0.440	0.099	0.136	0.053	0.031	0.184
SmoothGrad	0.331	0.418	0.079	0.126	0.109	0.105	0.195
IG	0.392	0.481	0.096	0.141	0.032	0.066	0.201
UA-Backprop	0.374	0.453	0.088	0.118	0.076	0.151	0.210
KernelSHAP	0.823	0.902	0.319	0.448	0.122	0.098	0.452
LIME	0.378	0.355	**0.415**	**0.547**	0.131	0.222	0.341
Occlusion	0.730	0.786	0.297	0.390	0.423	0.539	0.528
Ours	**0.860**	**0.918**	0.386	0.487	**0.869**	**0.939**	**0.743**