Gradient-based Uncertainty Attribution For Explainable Bayesian Deep Learning

Authors: Hanjing Wang, Dhiraj Joshi, Shiqiang Wang, Qiang Ji

Session and Poster ID: WED-AM-364

Rensselaer IBM

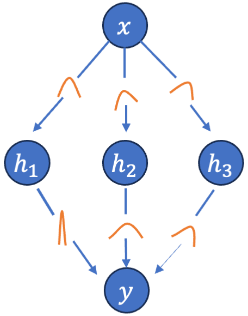
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Introduction to Uncertainty Attribution (UA)

Uncertainty Quantification 1

Explainable

Uncertainty Attribution 2



Bayesian neural networks

- Treat parameters as random variables

Posterior Likelihood Prior

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

evidence

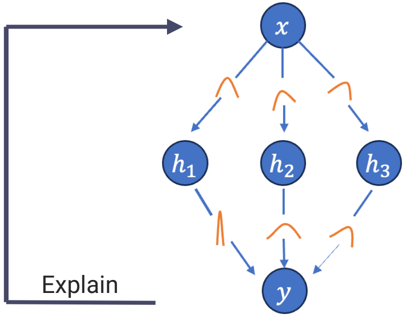
- Well-founded framework for uncertainty quantification

The input has high uncertainty

Something is wrong of the input

Where is wrong with the input

?

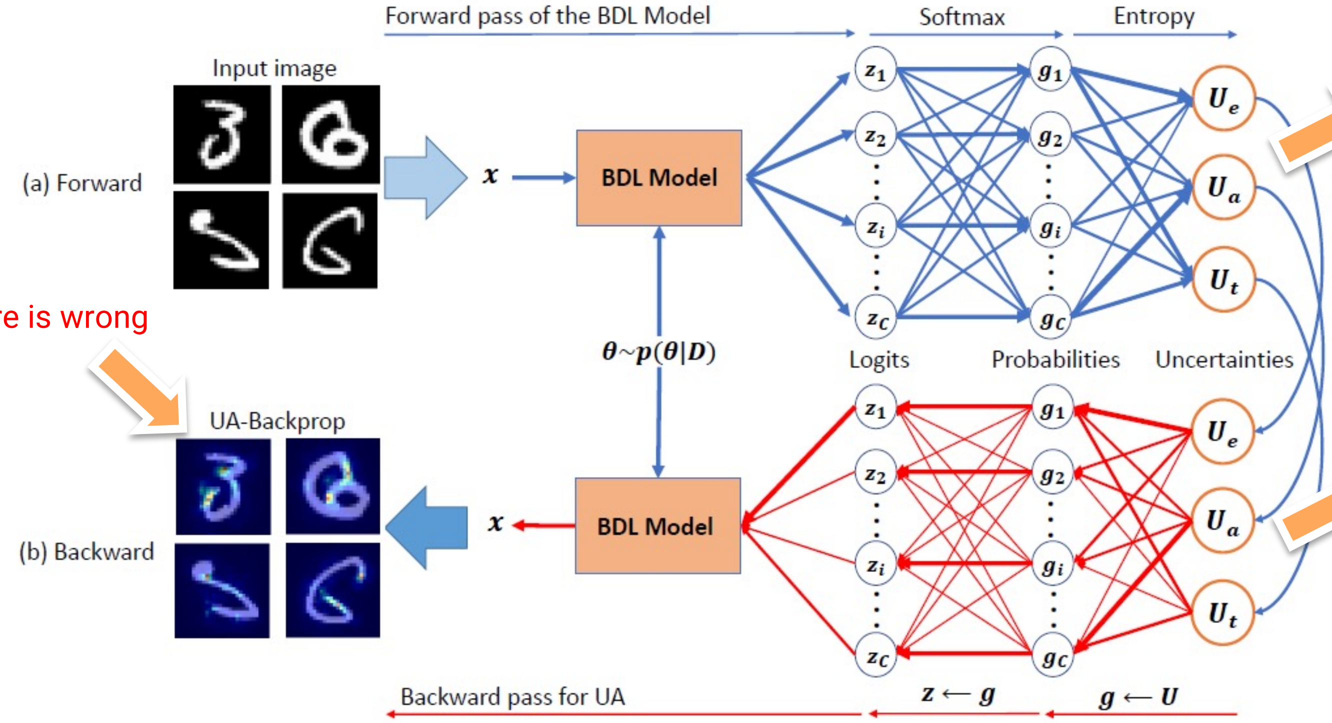


Uncertainty attribution (UA) focuses on understanding and explaining the sources of uncertainty. Localize the high uncertain regions to determine “where is wrong”?

Challenges

- Not well-explored area
- Previous methods focus on attribution of the classification score of the deterministic model

Overview of the Method: UA-Backprop



Where is wrong

Explain three types of uncertainty:

- epistemic uncertainty U_e
- aleatoric uncertainty U_a
- total uncertainty U_t

Specially designed:

- Gradient-based: finish attribution within a backward pass
- The attribution map achieves the completeness property
- Gradients are smoothed

UA-Backprop Details

Forward Pass

$\theta^s \sim p(\theta|D) \Rightarrow \{z^s\}_{s=1}^S \; \{g^s\}_{s=1}^S$

UQ

$$U_t = \sum_{i=1}^{C} -(\frac{1}{S}\sum_{s=1}^{S} g_i^s)\log(\frac{1}{S}\sum_{s=1}^{S} g_i^s)$$

$$U_a = \sum_{i=1}^{C} \frac{1}{S}\sum_{s=1}^{S} -g_i^s \log g_i^s$$

$$U_e = U_t - U_a$$

Uncertainty → Softmax Probability

$U \to g$

Find the contribution of each g_i to U

$$U_{t,g_i} = -(\frac{1}{S}\sum_{s=1}^{S} g_i^s)\log(\frac{1}{S}\sum_{s=1}^{S} g_i^s)$$

$$U_{a,g_i} = \frac{1}{S}\sum_{s=1}^{S} -g_i^s \log g_i^s$$

$$U_{e,g_i} = U_{t,g_i} - U_{a,g_i}$$

Softmax Probabilties → Logits

$g \to z$

Find the contribution of each z_i to U by exploring all possible path $g_j \to z_i$.

$$U_{z_i^s} = \sum_{j=1}^{C} c_{g_j^s \to z_i^s} U_{g_j}$$

$$c_{g_j^s \to z_i^s} = \phi_i\left(\left\{\frac{\partial g_j^s}{\partial z_k^s}\right\}_{k=1}^{C}, \tau_1\right) = \frac{\exp\left(\frac{\partial g_j^s}{\partial z_i^s}/(g_j^s \cdot \tau_1)\right)}{\sum_{k=1}^{C}\exp\left(\frac{\partial g_j^s}{\partial z_k^s}/(g_j^s \cdot \tau_1)\right)}$$

Logits to Input

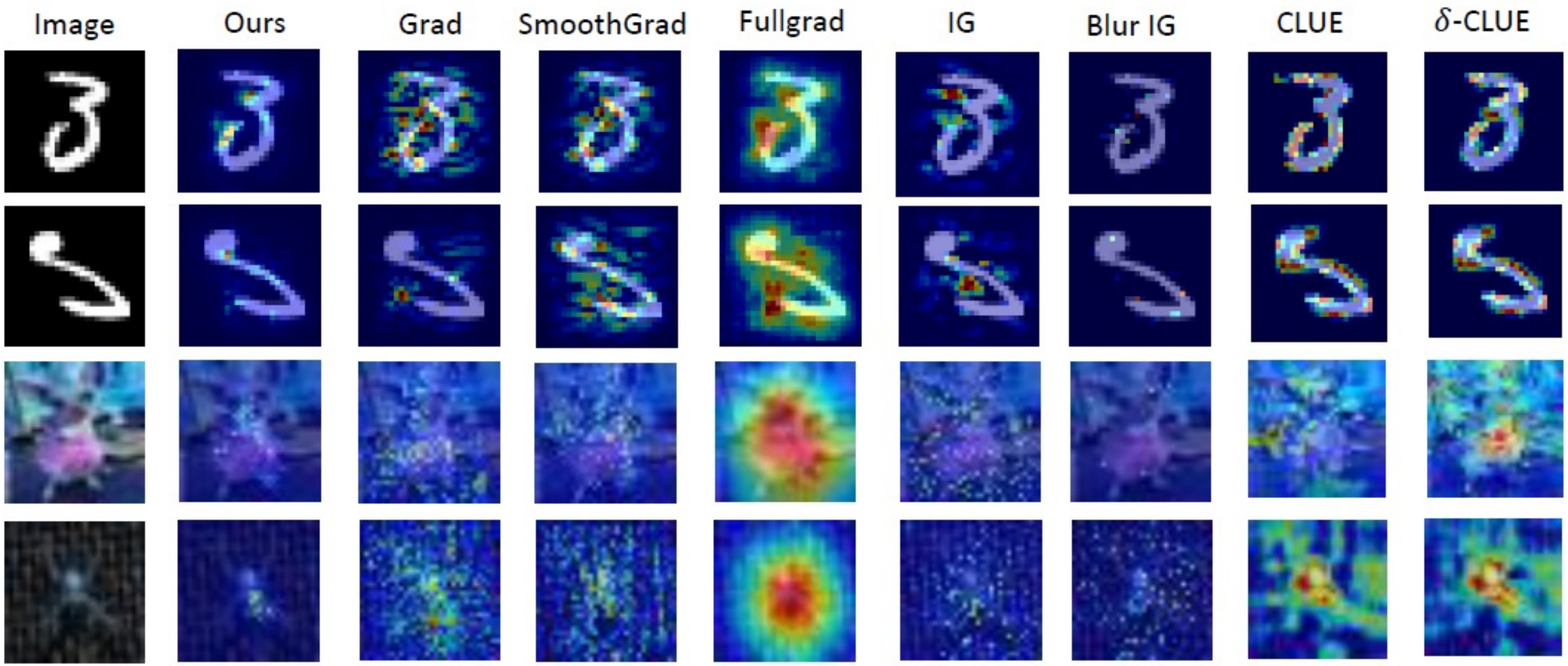
$z \to x$

Find the contribution of each x to U by exploring the path $U \to g \to z \to x$.

$$M(\boldsymbol{x}) = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{C} U_{z_i}^s M_i^s(\boldsymbol{x})$$

$$M_i^s(\boldsymbol{x}) = \psi\left(\left|\frac{\partial z_i^s}{\partial \boldsymbol{x}} \odot \boldsymbol{x}\right| + \sum_l \left|\frac{\partial z_i^s}{\partial \boldsymbol{b}_l^s} \odot \boldsymbol{b}_l^s\right|, \tau_2\right)$$

Qualitative Evaluation: Where Is Wrong?

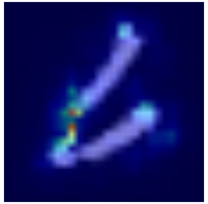


Quantitative Evaluation: Blurring Test

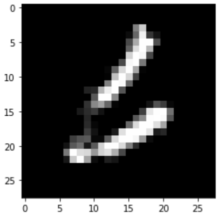
Iteratively blur up to 2%/5% of the image pixels, following the order of high UA scores (blur the problematic regions)

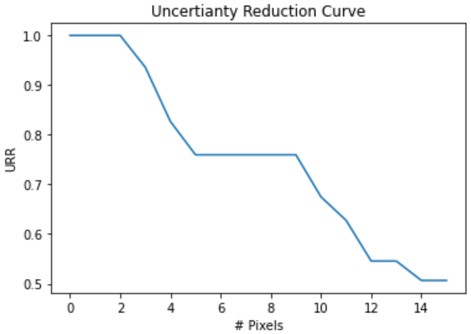
Image

Uncertainty Map



Blurred Image



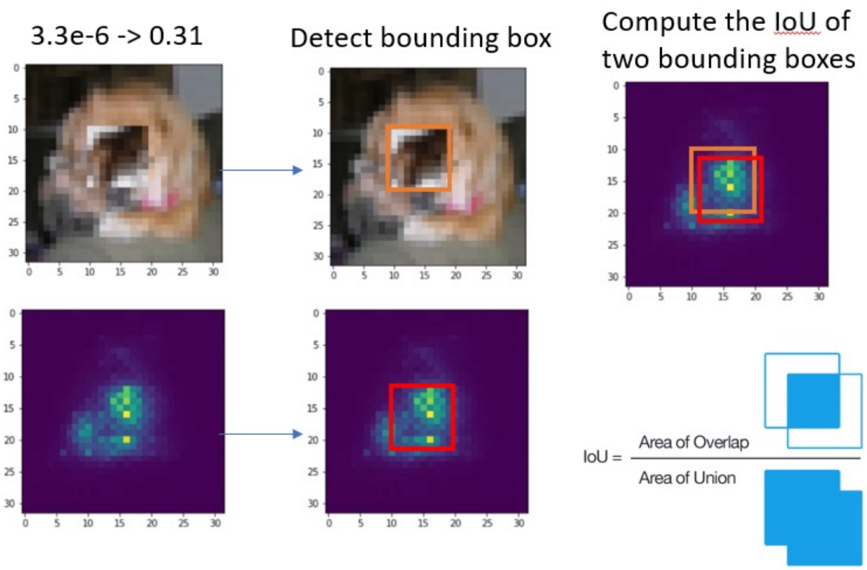


Method	Maximum Uncertainty Reduction Rate (MURR) ↑								
	MNIST		C10		C100		SVHN		Avg. Performance
	%2	%5	%2	%5	%2	%5	%2	%5	%2 + %5
Ours	0.648	0.850	0.629	0.848	**0.195**	0.302	0.625	0.758	**0.607**
Grad	0.506	0.741	0.578	0.798	0.165	0.276	0.555	0.705	0.541
SmoothGrad	0.601	0.779	0.566	0.800	0.154	0.255	0.575	0.735	0.558
FullGrad	**0.691**	0.869	0.555	0.772	0.156	0.274	0.565	0.709	0.574
IG	0.434	0.725	0.632	0.827	0.159	0.270	0.649	0.773	0.559
Blur IG	0.305	0.515	**0.693**	**0.971**	0.184	**0.318**	**0.762**	**0.896**	0.581
CLUE	0.614	0.874	0.291	0.628	0.074	0.148	0.171	0.352	0.394
δ-CLUE	0.625	**0.901**	0.415	0.577	0.073	0.150	0.146	0.295	0.398

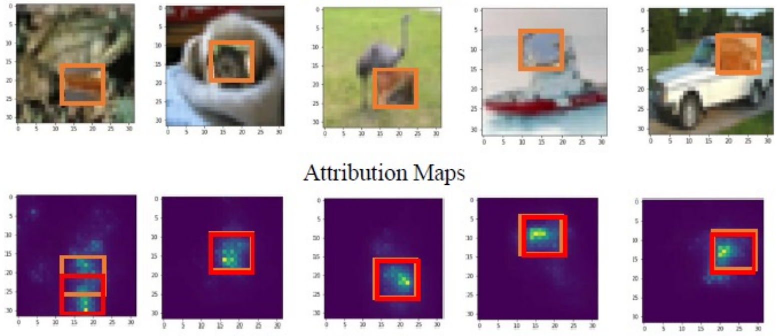
Method	Area under the Uncertainty Reduction Curve (AUC-URR) ↓								
	MNIST		C10		C100		SVHN		Avg. Performance
	%2	%5	%2	%5	%2	%5	%2	%5	%2 + %5
Ours	0.667	0.445	0.664	0.484	**0.901**	**0.821**	0.526	0.407	**0.614**
Grad	0.709	0.534	0.701	0.538	0.912	0.843	0.613	0.448	0.662
SmoothGrad	0.675	0.461	0.730	0.551	0.919	0.860	0.584	0.424	0.651
FullGrad	**0.603**	0.429	0.696	0.543	0.924	0.859	0.596	0.455	0.638
Blur IG	0.816	0.667	**0.638**	0.466	0.914	0.851	0.541	0.402	0.662
IG	0.752	0.529	0.731	**0.444**	0.905	0.824	**0.523**	**0.298**	0.626
CLUE	0.709	0.397	0.861	0.624	0.966	0.926	0.919	0.815	0.777
δ-CLUE	0.665	**0.395**	0.793	0.710	0.968	0.924	0.932	0.848	0.779

Quantitative Evaluation: Anomaly Detection

The process of anomaly detection



Visualizations



Quantitative Results

Method	C10		C100		SVHN		Avg. Performance	
	IoU	ADA	IoU	ADA	IoU	ADA	IoU	ADA
Ours	**0.353**	**0.285**	**0.363**	**0.375**	**0.217**	**0.124**	**0.311**	**0.261**
Grad	0.141	0.090	0.167	0.135	0.198	0.096	0.169	0.107
SmoothGrad	0.321	0.260	0.316	0.245	0.212	0.114	0.283	0.206
FullGrad	0.341	**0.285**	0.320	0.295	0.206	0.114	0.289	0.231
IG	0.171	0.090	0.170	0.105	0.139	0.052	0.160	0.082
Blur IG	0.182	0.125	0.318	0.290	0.150	0.078	0.217	0.164
CLUE	0.253	0.210	0.208	0.180	0.115	0.042	0.192	0.114
$\delta-$CLUE	0.248	0.240	0.229	0.220	0.105	0.044	0.194	0.168

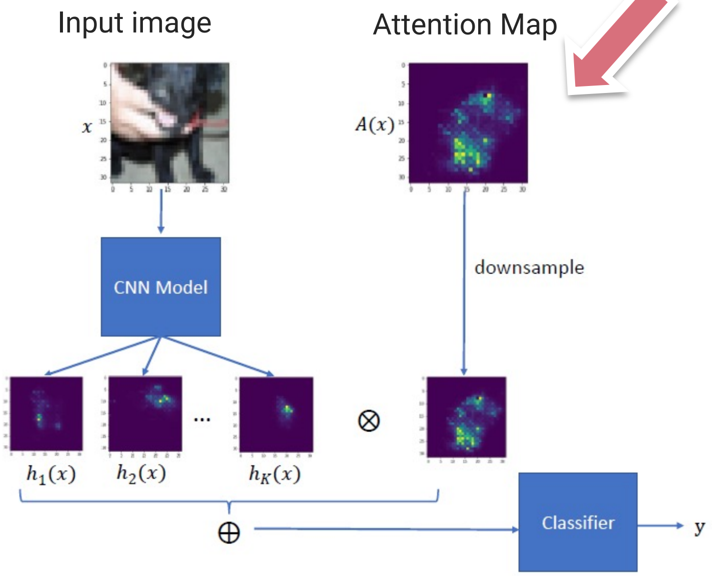
Uncertainty Attribution Maps As Attention

Given the uncertainty attribution map $M(x)$

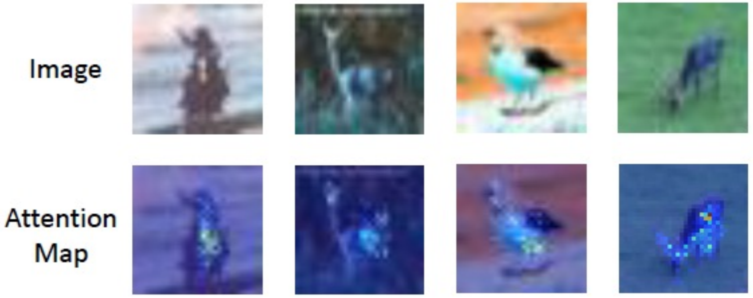
$$A(x) = \big(1 - M(x)\big) \odot M(x)$$

$A(x)$ strengthens more informative regions by

- Ignore the unimportant background information
- Ignore the problematic regions



Overall Framework



Examples of Attention Maps

Acc (%) ↑ and NLL ↓ for uncertainty mitigation evaluation

Method	MNIST		C10		C100		SVHN		Avg. Performance	
	ACC	NLL	ACC	NLL	ACC	NLL	ACC	NLL	ACC	NLL
Ours	91.95	**0.287**	**36.48**	**1.768**	12.12	4.326	**65.13**	**1.489**	**51.42**	**1.968**
Grad	91.35	0.302	31.60	1.938	12.13	4.422	63.74	1.578	49.71	2.060
SmoothGrad	90.68	0.324	32.05	1.942	**12.57**	4.508	62.35	1.628	49.41	2.100
FullGrad	91.39	0.300	32.85	1.920	12.06	4.574	62.38	1.568	49.67	2.091
IG	**91.98**	0.350	34.43	1.829	11.89	**4.265**	64.31	1.511	50.65	1.989
Blur IG	91.57	0.288	32.20	1.935	12.34	4.630	65.04	1.526	50.29	2.095
CLUE	91.64	0.348	33.34	1.846	12.15	4.299	60.01	1.572	49.29	2.016
δ-CLUE	91.76	0.350	35.02	1.809	12.22	4.362	62.71	1.612	50.43	2.033
No attention	90.78	0.358	31.62	1.921	12.02	4.536	60.64	1.569	48.77	2.096

Thank You