CLUE: A Method for Explaining Uncertainty Estimates

Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, José Miguel Hernández-Lobato

{ja666, usb20, tah47, aw665, jmh233}@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Intersecting Explainability and Uncertainty in ML

Can we interrogate probabilistic models (e.g. BNNs) about their reasoning when they are uncertain?

- Help ML practitioners identify input regions where training data is sparse; e.g. identify under-represented groups (by age, gender, etc.)
- Direct end-users' (e.g. medical professionals) attention to anomalous characteristics in uncertain inputs.

Applying Explainability Techniques to Uncertainty

- **Uncertainty Sensitivity Analysis [1]:** Gradient-based approach. In high dimensions can take points off data manifold.
- **Feature Importance Approaches:** highlight features which provide evidence for a given class. What if there is conflicting evidence for multiple classes or a lack of evidence for any class?

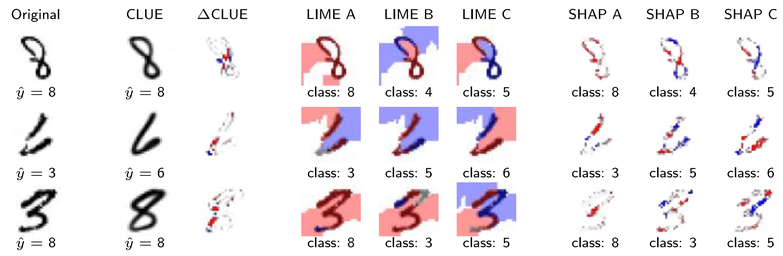


Figure: MNIST test digits for which our BNN's predictive entropy is large. Feature importance techniques, do not reveal the features responsible for uncertainty.

A Proposed Workflow

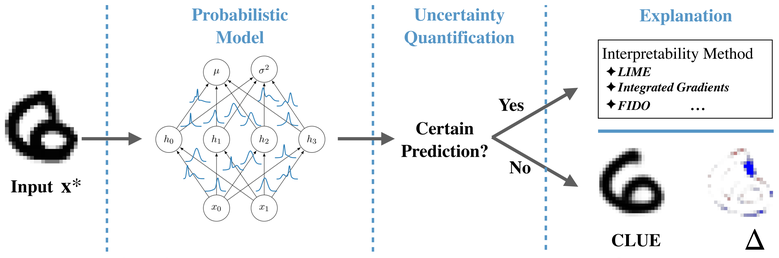


Figure: Workflow for automated decision making with transparency. Our BNN classifier produces a distribution over outputs. In an example high uncertainty case, CLUE allows us to identify features which are responsible for class ambiguity in the input (denoted by Δ and highlighted in dark blue). For certain decisions we use feature importance.

References

[1] S. Depeweg, J. M. Hernández-Lobato, S. Udluft, and T. A. Runkler. Sensitivity analysis for predictive uncertainty. In *ESANN*, 2017.

Counterfactual Latent Uncertainty Explanations

We ensure our counterfactuals represent plausible inputs by introducing an auxiliary latent variable DGM. **CLUE finds points in latent space which generate inputs similar to an original observation $\mathbf{x}_0$ but are assigned low uncertainty $\mathcal{H}$:**

$$\mathcal{L}(\mathbf{z}) = \mathcal{H}(\mathbf{y}|\mu_\theta(\mathbf{x}|\mathbf{z})) + d(\mu_\theta(\mathbf{x}|\mathbf{z}), \mathbf{x}_0), \quad (1)$$

$$\mathbf{x}_{\text{CLUE}} = \mu_\theta(\mathbf{x}|\mathbf{z}_{\text{CLUE}}) \text{ where } \mathbf{z}_{\text{CLUE}} = \arg\min_{\mathbf{z}} \mathcal{L}(\mathbf{z}). \quad (2)$$

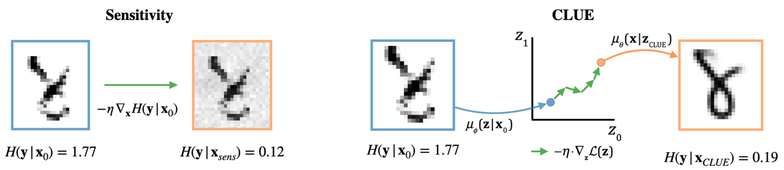


Figure: Left: A step in the direction of maximum sensitivity leads to a noisy input configuration for which H is small. Right:CLUE's uncertainty-based objective in terms of a DGM's latent variable $\mathbf{z}$ produces a plausible digit with a corrected lower portion.

Algorithm 1: CLUE

Inputs: datapoint $\mathbf{x}_0$, distance $d(\cdot)$, Uncertainty estimator $\mathcal{H}$, DGM decoder $\mu_\theta(\cdot)$, DGM encoder $\mu_\phi(\cdot)$

Set initial value $\mathbf{z} = \mu_\phi(\mathbf{z}|\mathbf{x}_0)$;
while *loss $\mathcal{L}$ is not converged* **do**
  Decode: $\mathbf{x} = \mu_\theta(\mathbf{x}|\mathbf{z})$;
  Use predictor to obtain $\mathcal{H}(\mathbf{y}|\mathbf{x})$;
  $\mathcal{L} = \mathcal{H}(\mathbf{y}|\mathbf{x}) + d(\mathbf{x}, \mathbf{x}_0)$;
  Update $\mathbf{z}$ with $\nabla_z \mathcal{L}$;
end

Output: $\mathbf{x}_{\text{CLUE}} = \mu_\theta(\mathbf{x}|\mathbf{z})$

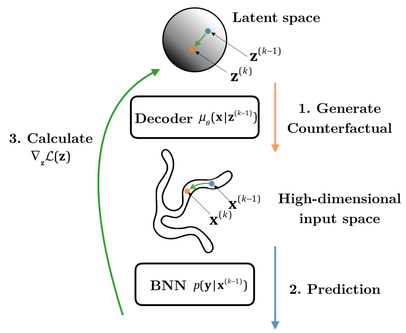
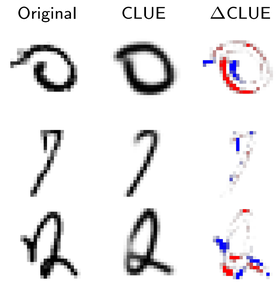


Figure: Latent codes are decoded into inputs for which a BNN generates uncertainty estimates; Our objective's gradients are back-propagated to latent space.

Explaining Uncertainty with Counterfactuals

CLUEs can be generated for classification and regression tasks on both tabular and image data. We highlight changes $\Delta_{CLUE}=(\mathbf{x}_{CLUE} - \mathbf{x}_0)^2$.



(a) MNIST

	Original
LSAT	41.0
UGPA	3.7
Race	Asian
Sex	Female

	CLUE
LSAT	41.0
UGPA	3.0
Race	White
Sex	Female

(b) LSAT

Figure: Example image and tabular CLUEs together with their corresponding Δ_{CLUE}.

Diversity in Plausible Counterfactuals

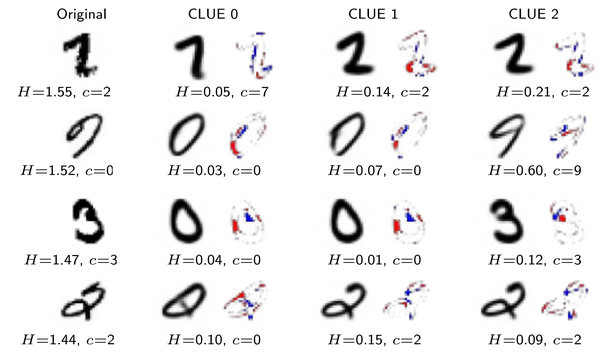


Figure: CLUEs for MNIST digits found from different random initialisations of $\mathbf{z}_0$.

	Original	CLUE	CLUE	CLUE
Age	Greater than 45	Greater than 45	25-45	25-45
Race	African-American	African-American	African-American	African-American
Sex	Female	Female	Male	Male
Current Charge	Felony	Misdemeanour	Misdemeanour	Felony
Reoffended Before	No	No	No	No
Prior Convictions	1	0	0	1
Days Served	0	0	0	0

Figure: CLUEs for COMPAS points found from different random initialisations of $\mathbf{z}_0$.

User Study

We run a forward simulation task. Users are told to predict if a BNN will be certain or uncertain about a new point after having been shown context points together with corresponding counterfactual explanations. We employ the COMPAS and LSAT tabular datasets. 40 ML graduate students take our survey, answering 18 questions, 9 per dataset.

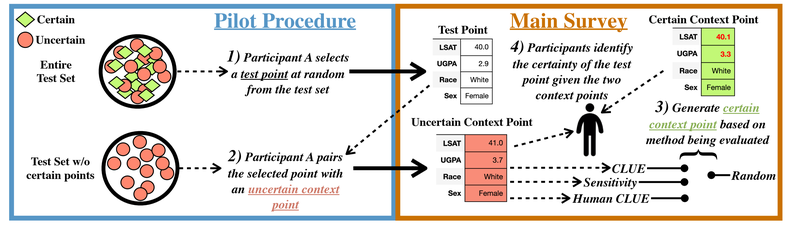


Figure: Experimental workflow for our tabular user study.

Table: Average participant accuracy by variant. Human selected refers to pilot users selecting counterfactuals points manually. We include random certain points as a baseline. We compare CLUE vs baselines using unpaired Wilcoxon signed-rank tests.

	CLUE	Human Choice	Uncertainty Sensitivity	Random Explanation
Accuracy (%)	**82.22**	62.22	52.78	61.67
p-value	-	2.34e-5	2.60e-9	1.47e-5