Diverse and Amortised Counterfactual Explanations for Uncertainty Estimates

Dan Ley, Umang Bhatt, and Adrian Weller

University of Cambridge

Abstract

To interpret uncertainty estimates from differentiable probabilistic models, [1] proposed generating a single Counterfactual Latent Uncertainty Explanation (CLUE) for a given data point where the model is uncertain. [2] formulated δ-CLUE, the set of CLUEs within a δ ball of the original input in latent space – however, we find that many CLUEs generated by this method are very similar, hence redundant. We propose DIVerse CLUEs (∇-CLUEs), a set of CLUEs which each provide a distinct explanation. We further introduce GLobal AMortised CLUEs (GLAM-CLUEs), which represent amortised mappings that apply to specific groups of uncertain inputs, taking them and efficiently transforming them in a single function call into inputs that a model will be certain about. Our experiments show that ∇-CLUEs and GLAM-CLUEs both address shortcomings of CLUE and provide beneficial explanations of uncertainty estimates to practitioners.

Introduction

[1] proposes a method for finding an explanation of a model's predictive uncertainty of a given input by searching in the latent space of an auxiliary deep generative model (DGM), identifying a single possible change to the input such that the model becomes more certain in its prediction. This is termed CLUE (Counterfactual Latent Uncertainty Explanation). However, there are limitations to CLUE, including the lack of a framework to deal with a potential diverse set of plausible explanations, despite proposing methods to generate them. CLUE introduces a latent variable DGM with decoder $\mu_\theta(\mathbf{x}|\mathbf{z})$ and encoder $\mu_\phi(\mathbf{z}|\mathbf{x})$. $\mathcal{H}$ refers to any differentiable uncertainty estimate of a prediction $\mathbf{y}$. CLUE minimises: $\mathcal{L}(\mathbf{z}) = \mathcal{H}\left(\mathbf{y}|\mu_\theta(\mathbf{x}|\mathbf{z})\right) + d\left(\mu_\theta(\mathbf{x}|\mathbf{z}), \mathbf{x}_0\right)$ to yield $\mathbf{x}_{\text{CLUE}} = \mu_\theta\left(\mathbf{x}|\mathbf{z}_{\text{CLUE}}\right)$ where $\mathbf{z}_{\text{CLUE}} = \text{argmin}_{\mathbf{z}} \mathcal{L}(\mathbf{z})$. We propose ∇-CLUE and GLAM-CLUE, full details of which can be found in the paper.

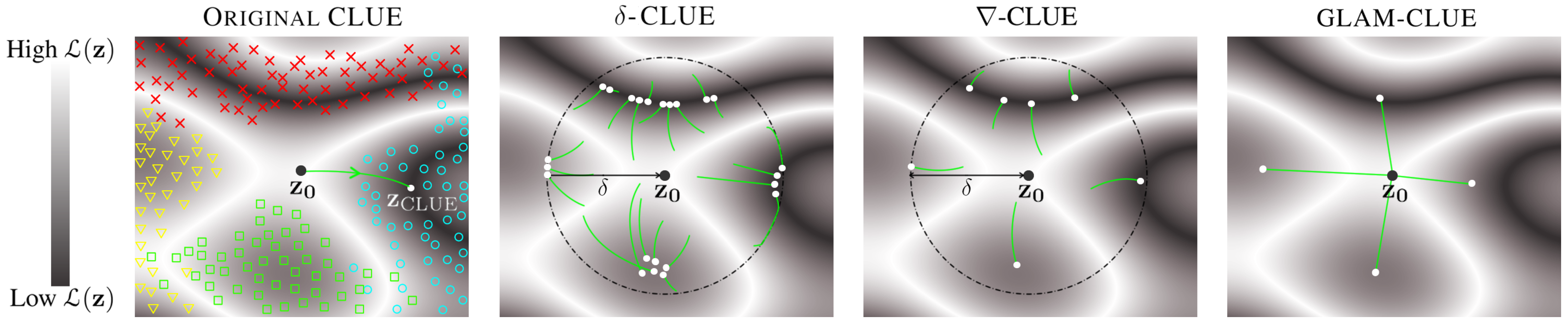

Figure 1: Conceptual colour map of objective function $\mathcal{L}(z)$ with $\mathbf{z}_0$ located in high cost region. White circles indicate explanations found. Left: Gradient descent to region of low cost [1]. Training points in colour. Left Centre: Gradient descent constrained to δ-ball [2]. Diverse starting points yield diverse local minima, albeit with many redundant solutions. Right Centre: Direct optimisation for diversity (∇-CLUE). Right: Efficient mappings without gradient descent (GLAM-CLUE).

DIVerse CLUE

δ-CLUE [2] introduces a method for generating a set of CLUEs by restricting the search in latent space to a ball of radius δ. However, many CLUEs found therein are redundant. We introduce metrics D (Table 1) to measure the diversity in sets of CLUEs such that we can optimise for it directly: we term this DIVerse CLUE (∇-CLUE). By optimising simultaneously over k counterfactuals, we minimise (note that we apply the function D in latent space here):

Diversity Metric	Function (D)
Determinantal Point Processes	$\det(\mathbf{K})$ where $\mathbf{K}_{i,j} = \dfrac{1}{1 + d(\mathbf{x}_i, \mathbf{x}_j)}$
Average Pairwise Distance	$\frac{1}{\binom{k}{2}} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(\mathbf{x}_i, \mathbf{x}_j)$
Coverage	$\frac{1}{d'} \sum_{i=1}^{d'} \left(\max_j (\mathbf{x}_j - \mathbf{x}_0)_i + \max_j (\mathbf{x}_0 - \mathbf{x}_j)_i \right)$
Prediction Coverage	$\dfrac{1}{c'} \sum_{i=1}^{c'} \max_j [(\mathbf{y}_j)_i]$
Distinct Labels	$\dfrac{1}{c'} \sum_{j=1}^{c'} \mathbf{1}_{[\exists i \,:\, y_i = j]}$
Entropy of Labels	$-\dfrac{1}{\log c'} \sum_{j=1}^{c'} p_j(k) \log p_j(k)$

Table 1: Diversity metrics D with arbitrary distance metric d.

$$\mathcal{L}(\mathbf{z}_1, \ldots, \mathbf{z}_k) = -\lambda_D D(\mathbf{z}_1, \ldots, \mathbf{z}_k) + \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}(\mathbf{z}_i)$$

$$\text{where} \quad \mathcal{L}(\mathbf{z}_i) = \mathcal{H}\left(\mathbf{y}|\mu_\theta(\mathbf{x}|\mathbf{z}_i)\right) + d\left(\mu_\theta(\mathbf{x}|\mathbf{z}_i), \mathbf{x}_0\right),$$

$$\text{to yield} \quad X_{\text{CLUE}} = \mu_\theta\left(X|Z_{\text{CLUE}}\right)$$

$$\text{where} \quad Z_{\text{CLUE}} = \underset{\mathbf{z}_1, \ldots, \mathbf{z}_k}{\text{argmin}} = \mathcal{L}(\mathbf{z}_1, \ldots, \mathbf{z}_k).$$

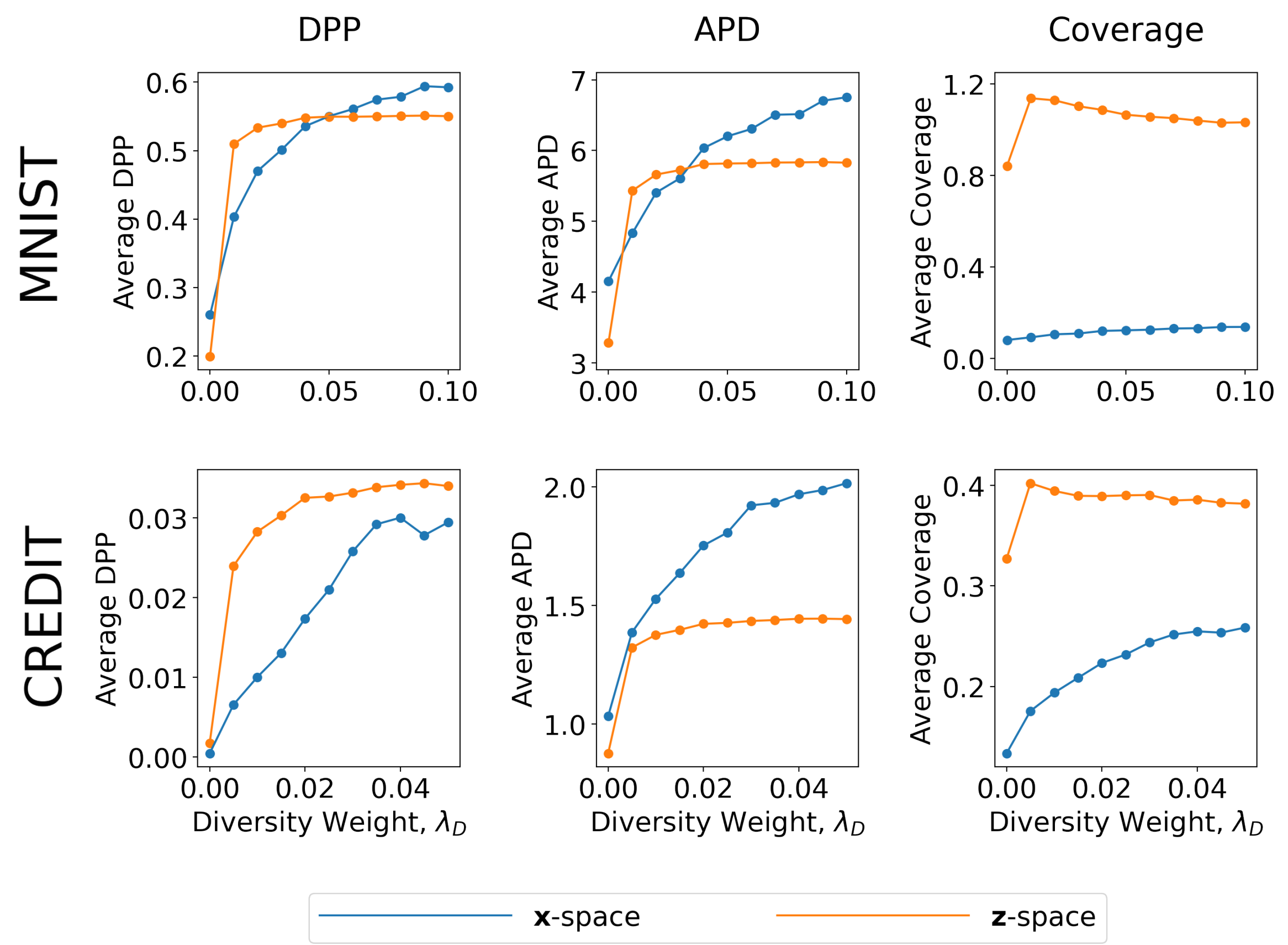

Figure 2: Effect of λ_D on diversity. DPP, APD and Coverage metrics applied to the set of $k = 10$ ∇-CLUEs.

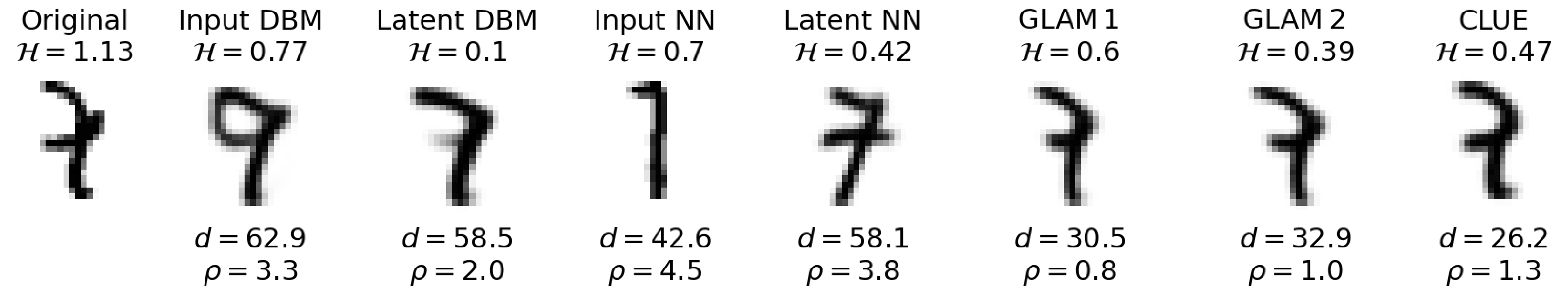

Figure 3: Comparison of explanations for an uncertain input (left) by the baselines, GLAM-CLUE, and CLUE. $\mathcal{H}$ is uncertainty, d is input space distance, ρ is latent space distance. Low $\mathcal{H}$ in baselines have unrealistically high d from the original.

GLobal AMortised CLUE

We desire a computationally efficient method that only requires a finite portion of the dataset from which we learn global properties of uncertainty; we propose GLobal AMortised CLUE (GLAM-CLUE), a method that achieves this with considerable speedups. High certainty points are taken from the training data to learn such mappings (we call this GLAM 1), but we demonstrate improvements by instead using CLUEs generated from uncertain points in the training data (GLAM 2). At inference time, GLAM-CLUE performs significantly faster than CLUE by **average CPU time** (Table 2). We show that ∇-CLUE and GLAM-CLUE address the shortcomings of CLUE.

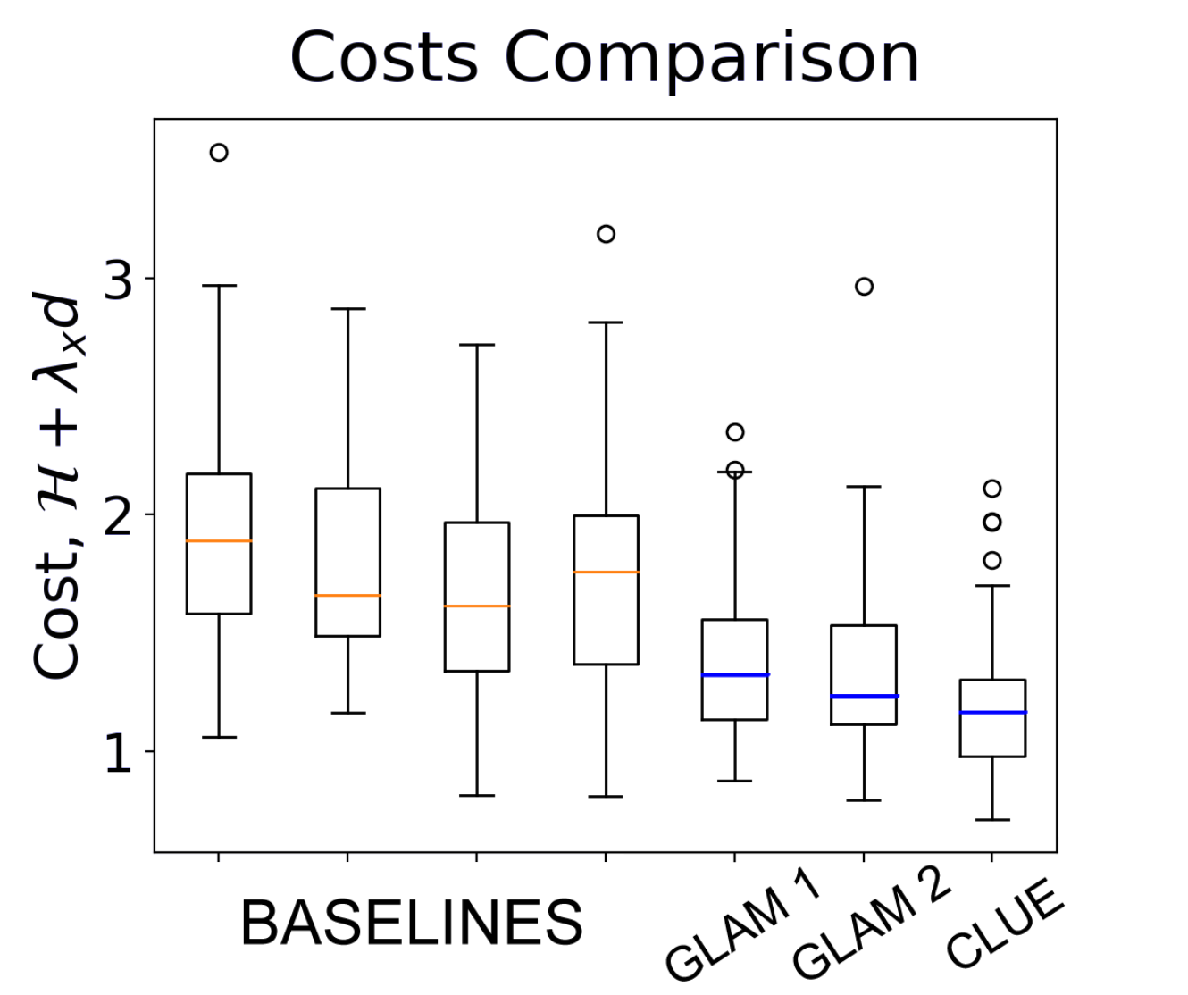

Figure 4: GLAM-CLUE vs baselines when mapping uncertain 7s to certain 7s in MNIST. Total costs $\mathcal{H} + \lambda_x d$.

Input DBM	Latent DBM	Input NN
0.0306	0.0262	0.0236
Latent NN	GLAM-CLUE	CLUE
0.0245	0.0238	4.68

Table 2: Average CPU time in **seconds** to compute **one** MNIST counterfactual explanation.

References

[1] Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a CLUE: A method for explaining uncertainty estimates. In *International Conference on Learning Representations*, 2021.

[2] Dan Ley, Umang Bhatt, and Adrian Weller. δ-CLUE: Diverse sets of explanations for uncertainty estimates. In *ICLR Workshop on Security and Safety in Machine Learning Systems*, 2021.

Acknowledgements

UB acknowledges support from DeepMind and the Leverhulme Trust via the Leverhulme Centre for the Future of Intelligence (CFI) and from the Mozilla Foundation. AW acknowledges support from a Turing AI Fellowship under grant EP/V025379/1, The Alan Turing Institute under EPSRC grant EP/N510129/1 and TU/B/000074, and the Leverhulme Trust via CFI. The authors thank Javier Antorán for his comments and pointers.