δ-CLUE: Diverse Sets of Explanations for Uncertainty Estimates

Dan Ley, Umang Bhatt, and Adrian Weller

University of Cambridge

Abstract

To interpret uncertainty estimates from differentiable probabilistic models, recent work has proposed generating Counterfactual Latent Uncertainty Explanations (CLUEs). However, for a single input, such approaches could output a variety of explanations due to the lack of constraints placed on the explanation. Here we augment the original CLUE approach, to provide what we call δ-CLUE. CLUE indicates *one* way to change an input, while remaining on the data manifold, such that the model becomes more confident about its prediction. We instead return a *set* of plausible CLUEs: multiple, diverse inputs that are within a δ ball of the original input in latent space, all yielding confident predictions.

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. Termed CLUE (Counterfactual Latent Uncertainty Explanation), this method is effective for generating plausible changes to an input that reduce uncertainty. 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: $p_\theta(\mathbf{x}) = \int (p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$, with encoder $q_\phi(\mathbf{z}|\mathbf{x})$. The predictive mean of the DGM is $\mathbb{E}_{p_\theta(\mathbf{x}|\mathbf{z})}[\mathbf{x}] = \mu_\theta(\mathbf{x}|\mathbf{z})$ and of the encoder is $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\mathbf{z}] = \mu_\phi(\mathbf{z}|\mathbf{x})$ respectively. $\mathcal{H}$ refers to any differentiable uncertainty estimate of a prediction $\mathbf{y}$. CLUE minimises:

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

$$\text{to yield } \mathbf{x}_{\text{CLUE}} = \mu_\theta\left(\mathbf{x}|\mathbf{z}_{\text{CLUE}}\right) \quad (2)$$

$$\text{where } \mathbf{z}_{\text{CLUE}} = \operatorname*{argmin}_{\mathbf{z}} \mathcal{L}(\mathbf{z}). \quad (3)$$

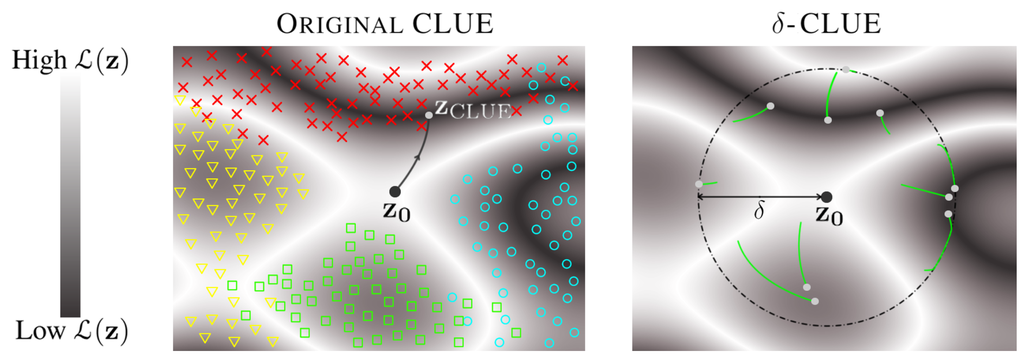

Figure 1: Conceptual colour map of objective function $\mathcal{L}(z)$ with $\mathbf{z}_0$ located in high cost region. Left: Gradient descent to region of low cost (original CLUE algorithm). Training points are shown in colour. Right: Gradient descent constrained to δ-ball at every step. Diverse starting points yield diverse local minima. White circles indicate CLUEs found.

Approach

We propose to modify the original method to generate a set of solutions that are all within a specified distance δ of $\mathbf{z}_0 = \mu_\phi(\mathbf{z}|\mathbf{x_0})$ in latent space: $\mathbf{z}_0$ is the latent space representation of the uncertain input $\mathbf{x}_0$ being explained. We achieve multiplicity by initialising the search in different areas of latent space using varied initialisation methods; some may randomly initialise within the δ-ball, while others could use training data or class boundaries to determine starting points. Figure 1 contrasts the original and proposed objectives.

In our proposed method, the loss function is the same as in Eq 1, with the δ requirement as:

$$\mathbf{x}_{\delta-\text{CLUE}} = \mu_\theta\left(\mathbf{x}|\mathbf{z}_{\delta-\text{CLUE}}\right) \quad (4)$$

$$\text{where } \mathbf{z}_{\delta-\text{CLUE}} = \operatorname*{argmin}_{\mathbf{z}:\ \rho(\mathbf{z},\mathbf{z}_0)\leq\delta} \mathcal{L}(\mathbf{z}) \quad (5)$$

$$\text{and } \mathbf{z_0} = \mu_\phi(\mathbf{z}|\mathbf{x_0}) \quad (6)$$

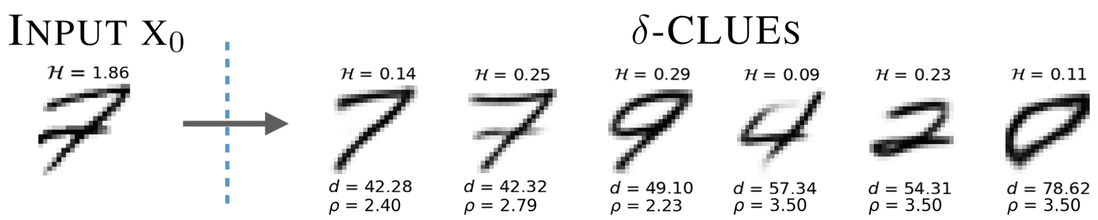

Figure 3: We produce a **diverse set** of candidate explanations that show how to reduce predictive uncertainty while still remaining close to x_0 in both input and latent space ($\mathcal{H}$ is uncertainty, d is input space distance, ρ is latent space distance).

Results

The value of δ is shown to control the performance of CLUEs found:

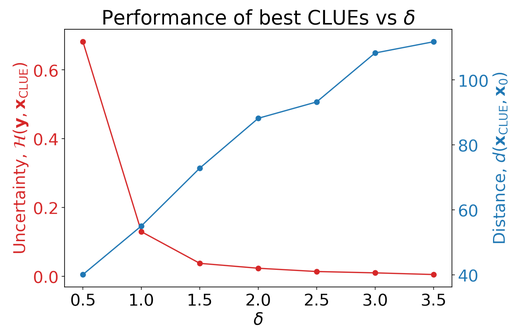

Figure 2: Increasing the size of the δ ball yields lower uncertainty CLUEs at higher distances from the original input.

We demonstrate diversity in the CLUEs found on MNIST. Diversity arises via convergence to multiple class labels or to different modes within these labels.

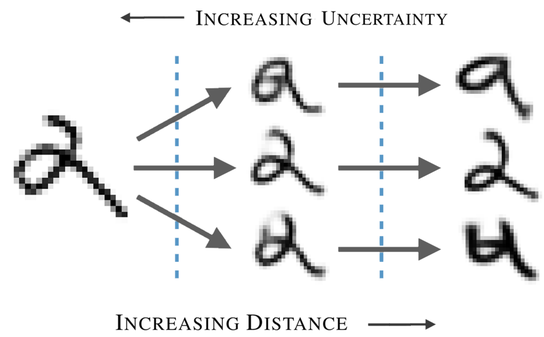

Figure 4: Visualisation of the trade off between uncertainty $\mathcal{H}$ and distance d (diverse examples discovered by δ-CLUE).

Future Work

As recent work considered specifying the exact level of uncertainty desired in a sample [2] and has considered using DGMs to find counterfactual explanations though not for uncertainty [3], we posit that leveraging DGMs to study the *diversity* of plausible explanations is a promising direction to pursue. δ-CLUE is just one step towards realising this goal.

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] Serena Booth, Yilun Zhou, Ankit Shah, and Julie Shah.
Bayes-TrEx: Model transparency by example.
arXiv e-prints, pages arXiv–2002, 2020.

[3] Shalmali Joshi, Oluwasanmi Koyejo, Been Kim, and Joydeep Ghosh.
xGEMs: Generating examplars to explain black-box models.
arXiv preprint arXiv:1806.08867, 2018.

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 helpful comments and pointers.