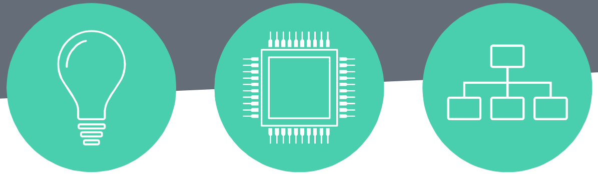
Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates



Overview

01 Counterfactual Latent Uncertainty Explanation (CLUE)

02 Why this paper?

03 δ-CLUE

04 Diversity Metrics

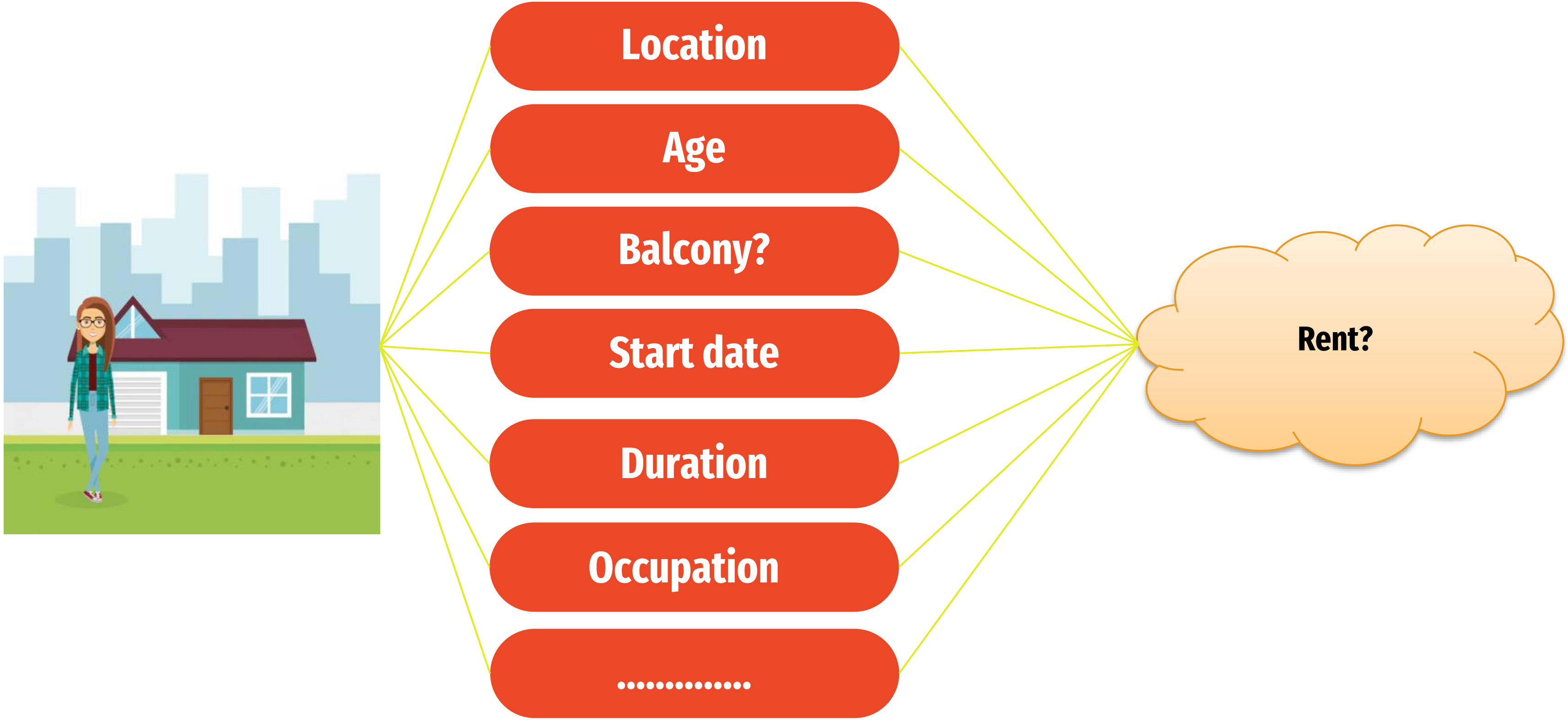
05 ∇-CLUE

07 GLAM-CLUE

08 Performance Test

09 Future Work

Counterfactual Explanations



Counterfactual Latent Uncertainty Explanation (CLUE)

What is the smallest on-manifold change that can be done to an input so that our model becomes more certain

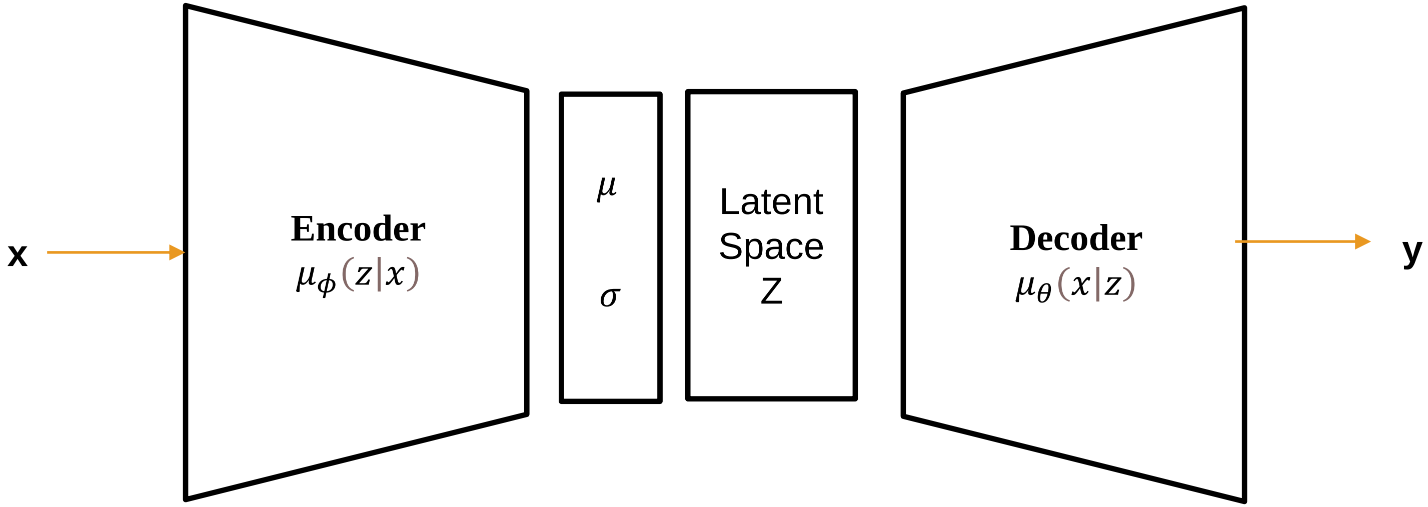
High Uncertainty

Incorrect prediction

More counterfactuals

Uncertainty explanations are a precedent for model explanation

CLUE Generation with Auxiliary Deep Generative Model with VAE



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

yields

$$\mathbf{x}_{\mathrm{CLUE}} = \mu_\theta(\mathbf{x}|\mathbf{z}_{\mathrm{CLUE}})$$

where

$$\mathbf{z}_{\mathrm{CLUE}} = \arg\min_{\mathbf{z}} \mathcal{L}(\mathbf{z})$$

$\mathcal{H}$ = Uncertainty estimate of a prediction y

$d(\mathbf{x}, \mathbf{x}_0) = \lambda_x d_x(\mathbf{x}, \mathbf{x}_0) +$
$\lambda_y d_y(f(\mathbf{x}), f(\mathbf{x}_0))$, where $f(\mathbf{x}) = \mathbf{y}$

Why this paper?

Limitations of CLUE

- **Multipicity Issue** — δ-CLUE
- **Redundancy** — ∇-CLUE (DIVerse CLUE)
- **Computational Inefficiency** — GLAM-CLUE

δ-CLUE vs CLUE

δ-CLUE

Multiplicity is achieved by searching randomly in different areas of latent space

- Sampling around an input in latent space
- Gradient descent

Vs

CLUE also does this, but:

- Finds minima in a limited region of space
- Might strays far away from Counterfactuals

$$\mathbf{x}_{\delta-\mathrm{CLUE}} = \mu_\theta\left(\mathbf{x}\middle|\mathbf{z}_{\delta-\mathrm{CLUE}}\right) \quad \text{where} \quad \mathbf{z}_{\delta-\mathrm{CLUE}} = \arg\min_{\mathbf{z}:\ \rho(\mathbf{z},\mathbf{z}_0)\leq\delta} \mathcal{L}(\mathbf{z})$$

$$\mathbf{z_0} = \mu_\phi(\mathbf{z}|\mathbf{x_0})$$

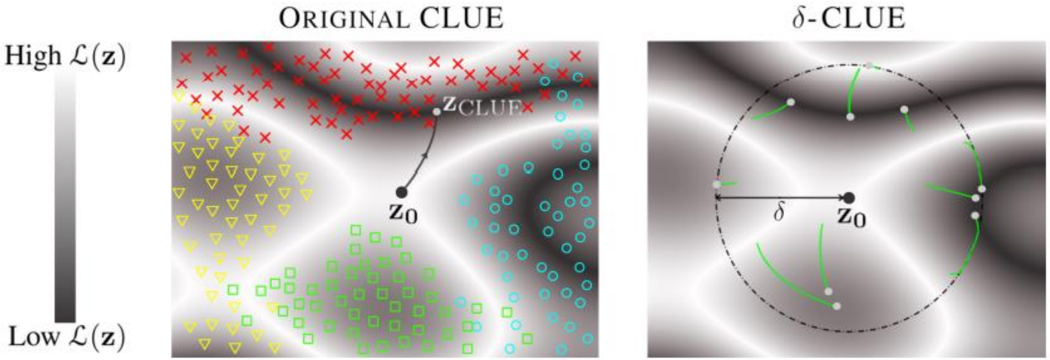
$$\rho(\mathbf{z},\mathbf{z}_0) = ||\mathbf{z}-\mathbf{z}_0||_2$$

Algorithm 3: δ-CLUE

Inputs: δ, k, $\mathcal{S}$, r, $\mathbf{x}_0$, d, ρ, $\mathcal{H}$, μ_θ, μ_ϕ

Initialise $\varnothing$ of CLUEs: $X_{\text{CLUE}} = \{\}$;
Set δ-ball centre of $\mathbf{z}_0 = \mu_\phi(\mathbf{z}|\mathbf{x}_0)$;
for $1 \leq i \leq k$ **do**
    Set initial value of $\mathbf{z}_i = \mathcal{S}(\mathbf{z}_0, r, i, k)$;
    while *loss $\mathcal{L}$ has not converged* **do**
        Decode: $\mathbf{x} = \mu_\theta(\mathbf{x}|\mathbf{z}_i)$;
        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 $\mathbf{z}_i$ with $\nabla_{\mathbf{z}}\mathcal{L}$;
        if $\rho(\mathbf{z}_i, \mathbf{z}_0) > \delta$ **then**
            Project $\mathbf{z}_i$ onto the surface of the δ-ball as $\mathbf{z}_i = \delta \times \frac{\mathbf{z}_i - \mathbf{z}_0}{\rho(\mathbf{z}_i, \mathbf{z}_0)}$;
        end if
    end while
    Decode explanation: $\mathbf{x}_{\delta-\text{CLUE}} = \mu_\theta(\mathbf{x}|\mathbf{z}_i)$;
    if $\mathcal{H}(\mathbf{y}|\mathbf{x}_{\delta_{\text{CLUE}}}) < \mathcal{H}_{\text{threshold}}$ **then**
        $X_{\text{CLUE}} \leftarrow X_{\text{CLUE}} \cup \mathbf{x}_{\delta_{\text{CLUE}}}$;
    end if
end for

Outputs: X_{CLUE}, a set of $n \leq k$ CLUEs



Source: Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 7. 2022.

Different trials on δ-CLUE Algorithm

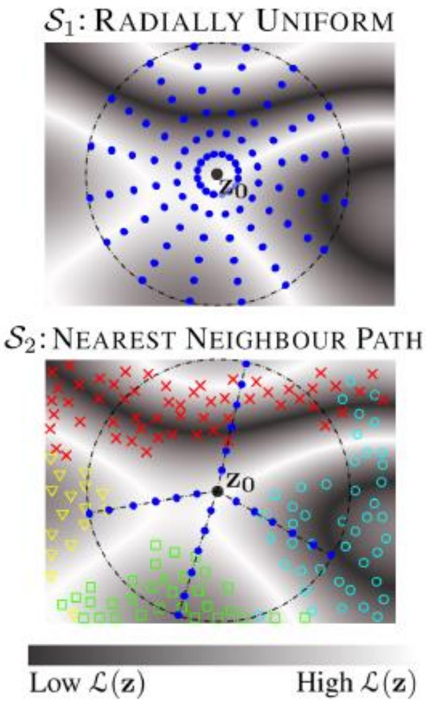
01 Range of δ values from 0.5 to 3.5

02 Two latent space loss functions:

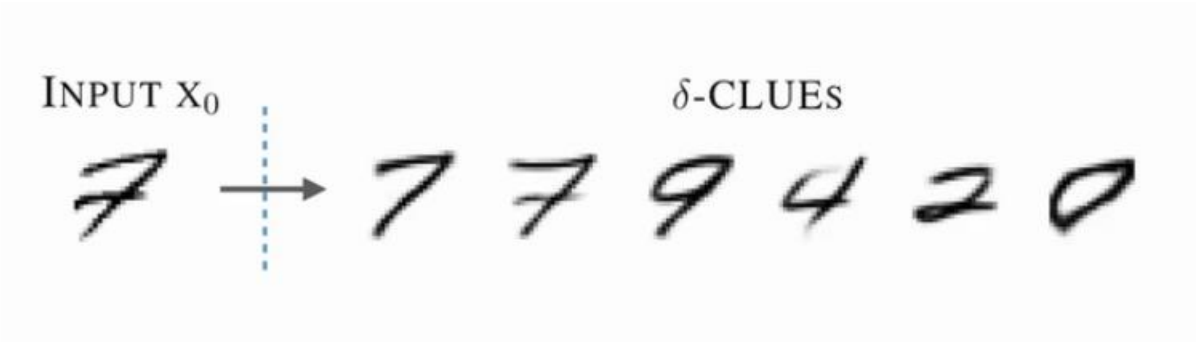
- Uncertainty : $\mathcal{L}_{\mathcal{H}} = \mathcal{H}$
- Distance : $\mathcal{L}_{\mathcal{H}+d} = \mathcal{H} + d$

03 Two initialisation schemes like:

- Radially Uniform
- Nearest Neighbour

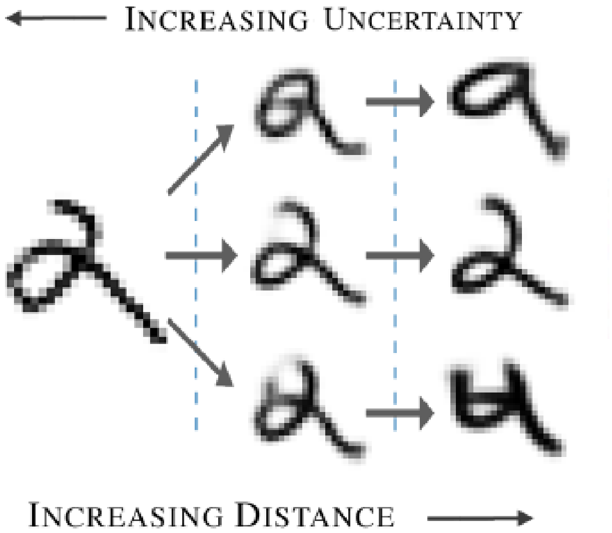


Source: Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 7. 2022.



Source: Ley, Dan, Umang Bhatt, and Adrian Weller. "{\delta}-CLUE: Diverse Sets of Explanations for Uncertainty Estimates." *arXiv preprint arXiv:2104.06323* (2021).

Uncertainty vs Distance Trade-off



The hyperparameters (λ_x, λ_y') controls this trade-off

Source: Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 7. 2022.

Diversity Metrics (D)

01

Determinantal Point Process

$$\det(\mathbf{K}) \text{ where } \mathbf{K}_{i,j} = \frac{1}{1 + d(\mathbf{x}_i, \mathbf{x}_j)}$$

02

Average Pairwise Distance

$$\frac{1}{\binom{k}{2}} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(\mathbf{x}_i, \mathbf{x}_j)$$

03

Coverage

$$\frac{1}{d'} \sum_{i=1}^{d'} (\max_j (\mathbf{x}_j - \mathbf{x}_0)_i + \max_j (\mathbf{x}_0 - \mathbf{x}_j)_i)$$

04

Prediction Coverage

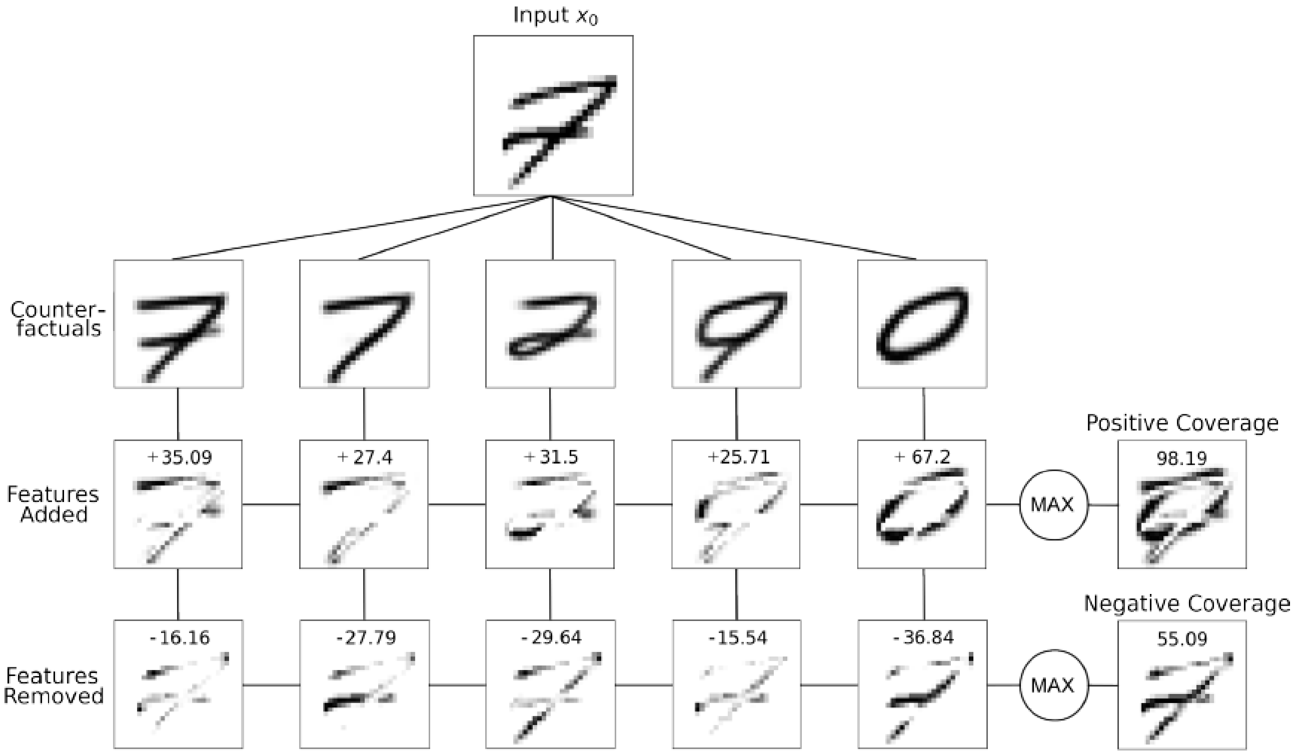
$$\frac{1}{c'} \sum_{i=1}^{c'} \max_j [(\mathbf{y}_j)_i]$$

05

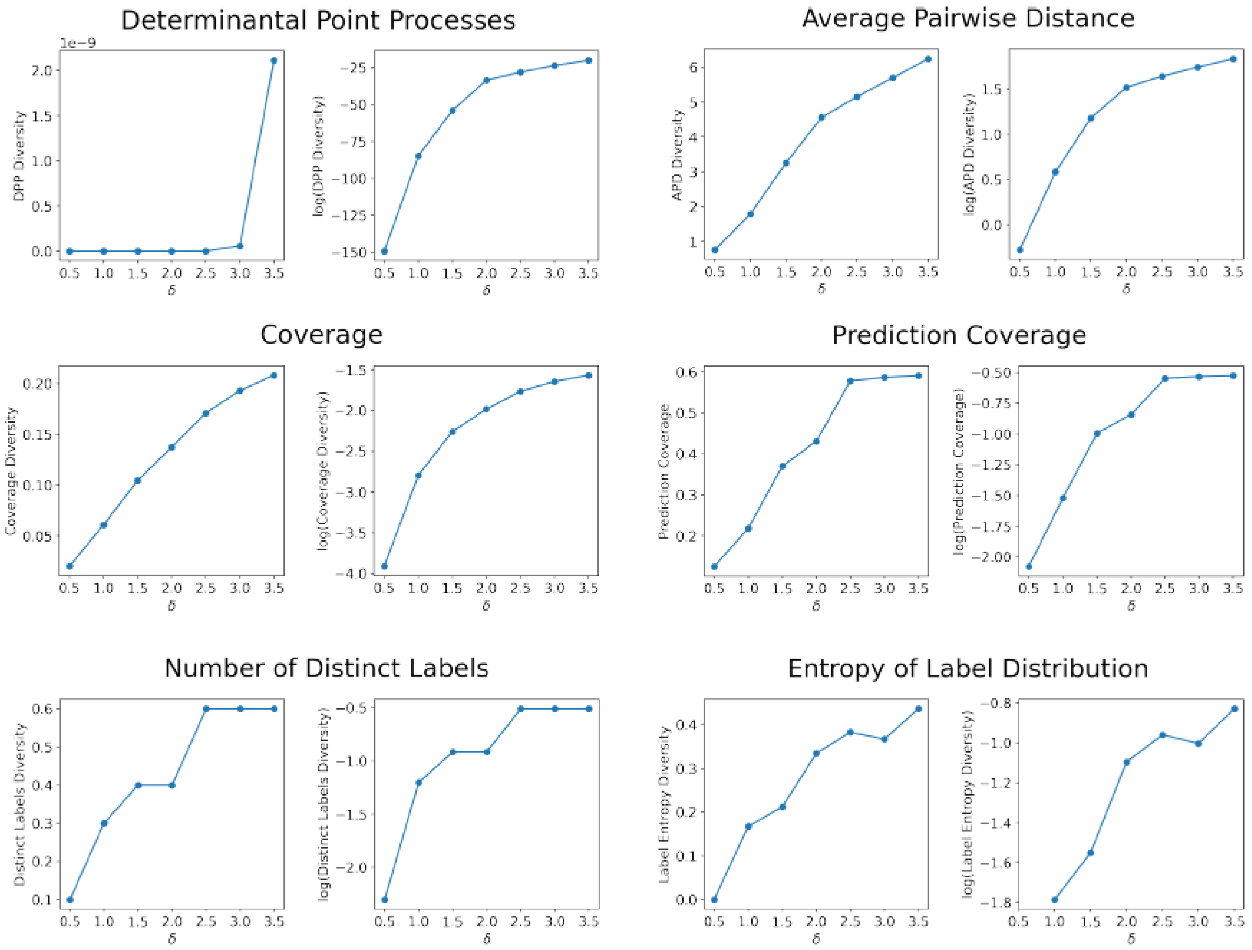
Distinct Labels/ Entropy of Labels

$$-\frac{1}{\log c'} \sum_{j=1}^{c'} p_j(k) \log p_j(k)$$

Coverage as a Metric



Source: Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 7. 2022.



Source: Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 7. 2022.

Diversity Optimization : ∇-CLUE

01

Select the Diversity Metric

02

Optimize k counterfactuals

Simultaneous Diversity Optimization

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

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

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

$$Z_{\text{CLUE}} = \arg\min_{\mathbf{z}_1, \ldots, \mathbf{z}_k} = \mathcal{L}(\mathbf{z}_1, \ldots, \mathbf{z}_k)$$

Sequential Diversity Optimization

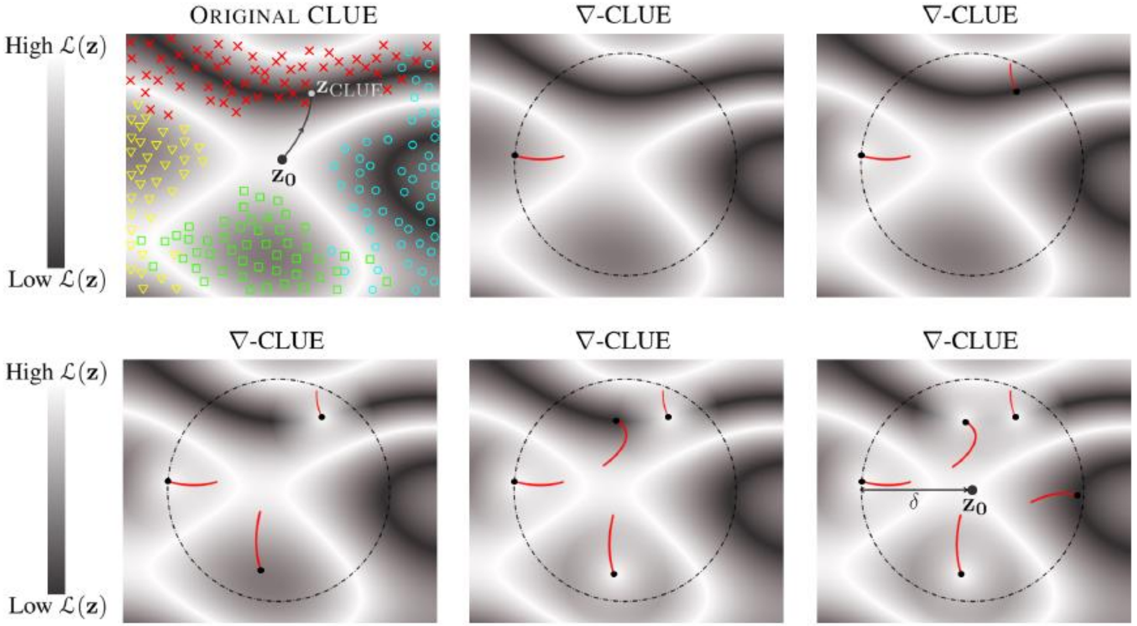
$$\mathcal{L}(\mathbf{z}) = \lambda_D D(Z_{\text{CLUE}} \cup \mathbf{z}) + \mathcal{H}\left(\mathbf{y} | \mu_\theta(\mathbf{x}|\mathbf{z})\right) + d\left(\mu_\theta(\mathbf{x}|\mathbf{z}), \mathbf{x}_0\right)$$

Algorithm 1: ∇-CLUE (simultaneous)

Inputs: $\delta, k, \mathcal{S}, r, \mathbf{x}_0, d, \rho, \mathcal{H}, \mu_\theta, \mu_\phi, D, \lambda_D$

Initialise $\varnothing$ of CLUEs: $X_{\text{CLUE}} = \{\}$;
Set δ-ball centre of $\mathbf{z}_0 = \mu_\phi(\mathbf{z}|\mathbf{x}_0)$;
for $1 \leq i \leq k$ **do**
    Set initial value of $\mathbf{z}_i = \mathcal{S}(\mathbf{z}_0, r, i, k)$;
end for
while *loss* $\mathcal{L}$ *has not converged* **do**
    for $1 \leq i \leq k$ **do**
        Decode: $\mathbf{x}_i = \mu_\theta(\mathbf{x}|\mathbf{z}_i)$;
        Use predictor to obtain $\mathcal{H}(\mathbf{y}|\mathbf{x}_i)$;
        $\mathcal{L}(\mathbf{z}_i) = \mathcal{H}(\mathbf{y}|\mathbf{x}_i) + d(\mathbf{x}_i, \mathbf{x}_0)$;
    end for
    $\mathcal{L}(\mathbf{z}_1, ..., \mathbf{z}_k) = -\lambda_D D(\mathbf{z}_1, ..., \mathbf{z}_k) + \frac{1}{k}\sum_{i=1}^{k} \mathcal{L}(\mathbf{z}_i)$;
    Update $\mathbf{z}_1, ..., \mathbf{z}_k$ with $\nabla_{\mathbf{z}_1,...,\mathbf{z}_k} \mathcal{L}(\mathbf{z}_1, ..., \mathbf{z}_k)$;
    for $1 \leq i \leq k$ **do**
        Constrain $\mathbf{z}_i$ to δ ball using $\rho(\mathbf{z}_i, \mathbf{z}_0)$;
    end for
end while
for $1 \leq i \leq k$ **do**
    Decode explanation: $\mathbf{x}_i = \mu_\theta(\mathbf{x}|\mathbf{z}_i)$;
    if $\mathcal{H}(\mathbf{y}|\mathbf{x}_i) < \mathcal{H}_{\text{threshold}}$ **then**
        $X_{\text{CLUE}} \leftarrow X_{\text{CLUE}} \cup \mathbf{x}_i$;
    end if
end for

Outputs: X_{CLUE}, a set of $n \leq k$ diverse CLUEs



Sequential ∇-CLUE Optimization

Source: Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 7. 2022.

GLobal AMortised CLUE (GLAM-CLUE)

Input data

Finite set of Counterfactuals

Training Step

A mapper which learn global properties of uncertainty in latent space

$$\mathbf{z}_{\text{certain}} = G(\mathbf{z}_{\text{uncertain}})$$

Inference Step

Apply this mapper to test data to see the effect on uncertain data

Output data

$$\mathcal{L}(\boldsymbol{\theta}|Z_{\text{uncertain}}, X_{\text{certain}}) =$$

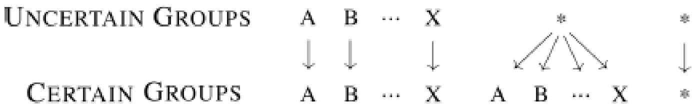
$$\lambda_\theta \|\boldsymbol{\theta}\|_1 + \frac{1}{|Z_{\text{uncertain}}|} \sum_{\mathbf{z} \in Z_{\text{uncertain}}} \min_{\mathbf{x} \in X_{\text{certain}}} \|\mu_\theta(\mathbf{z} + \boldsymbol{\theta}) - \mathbf{x}\|_2^2 \quad (2)$$

Algorithm 2: GLAM-CLUE (Training Step)

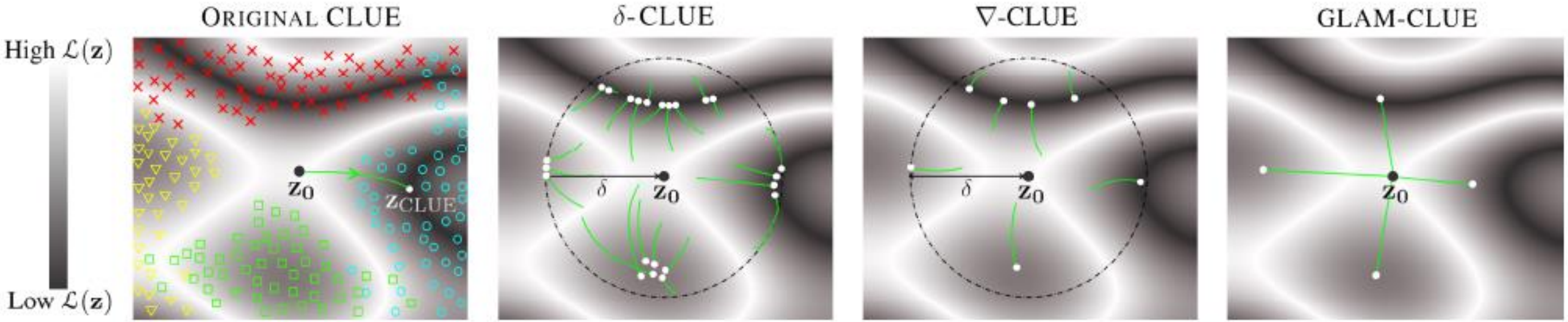
Inputs: Inputs $X_{\text{uncertain}}, X_{\text{certain}}$, groups $Y_{\text{uncertain}}, Y_{\text{certain}}$, DGM encoder μ_ϕ, loss $\mathcal{L}$, trainable parameters $\boldsymbol{\theta}$

for all groups $(i \to j)$ in $(Y_{\text{uncertain}}, Y_{\text{certain}})$ **do**
  Select X_i from $X_{\text{uncertain}}, Y_{\text{uncertain}}$;
  Select X_j from $X_{\text{certain}}, Y_{\text{certain}}$;
  Encode: $Z_i = \mu_\phi(Z|X_i)$;
  while *loss $\mathcal{L}$ has not converged* **do**
    Update $\boldsymbol{\theta}_{i\to j}$ with $\nabla_{\boldsymbol{\theta}_{i\to j}} \mathcal{L}(\boldsymbol{\theta}_{i\to j}|Z_i, X_j)$;
  end while
end for

Outputs: A collection of mapping parameters $\boldsymbol{\theta}_{i\to j}$ for given mappers $G_{i\to j}$ that take uncertain inputs from group i and produce nearby certain outputs in group j



$$\mathbf{z}_j = G_{i\to j}(\mathbf{z}_i) = \mathbf{z}_i + \boldsymbol{\theta}_{i\to j}$$



Source: Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 7. 2022.

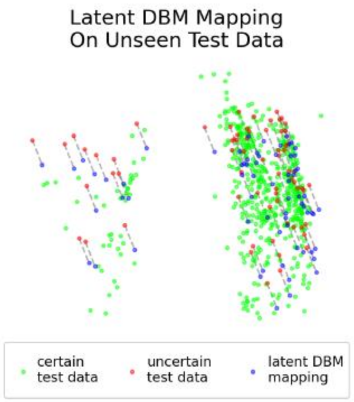
Performance Test

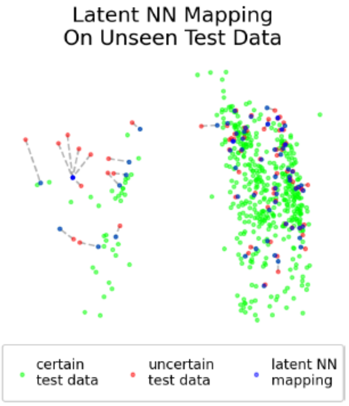
01 **Difference Between Means (DBM)**

Uncertain data to certain data in input or latent space

02 **Nearest Neighbours (NN)**

Used in high certainty training data in input or latent space





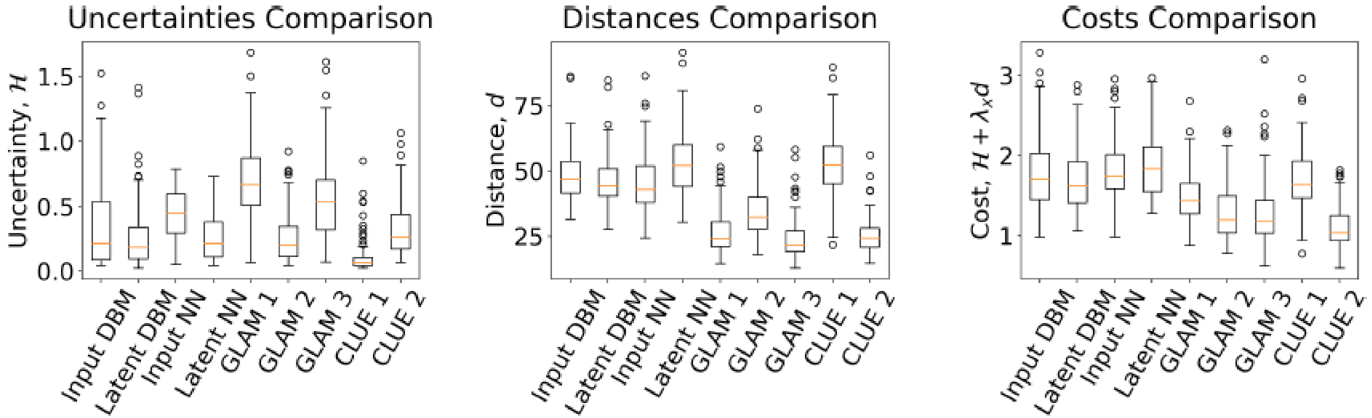
Source: Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 7. 2022.

Performance Comparison

Input DBM	Latent DBM	Input NN
0.0306	0.0262	0.0236

Latent NN	GLAM-CLUE	CLUE
0.0245	0.0238	4.68

GLAM-CLUE outperforms these baselines
……*almost* **200** *times faster*



Source: Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 7. 2022.

Future Work

01

Data set dimensions

Using higher dimensional data set

02

Introduce different metric

Use FID scores to replace simple distance metric in evaluation and optimisation

03

Use different DGMs

Use DGM alternative like GANs instead of VAEs

Conclusion

01 **Making CLUE more useful in practice**

02 **Proposed δ-CLUE and ∇-CLUE to tackle the multiplicity and diversity issues**

03 **Introduced GLAM-CLUE which tackles the computational inefficiency caused on large data sets with δ-CLUE and ∇-CLUE**

References

- https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73
- Antorán, Javier, et al. "Getting a clue: A method for explaining uncertainty estimates." *arXiv preprint arXiv:2006.06848* (2020).
- Ley, Dan, Umang Bhatt, and Adrian Weller. "{\delta}-CLUE: Diverse Sets of Explanations for Uncertainty Estimates." *arXiv preprint arXiv:2104.06323* (2021).
- https://slideslive.com/38955757/deltaclue-diverse-sets-of-explanations-for-uncertainty-estimates?ref=recommended

THANK YOU

Sruthi Aikkara

Matriculation number: **229386**

Supervisor: **Jelle Hüntelmann**