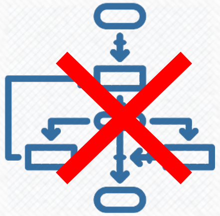
Quantifying the Importance of Latent Features in Neural Networks

Amany Alshareef, Nicolas Berthier, Sven Schewe, Xiaowei Huang

Department of Computer Science
University of Liverpool, UK

UNIVERSITY OF LIVERPOOL

Challenges in Deep Neural Network Testing



→ Deep neural networks learn by example
→ Do **not** have a specific control-flow structure

→ Most testing techniques propose structural coverage
→ Tend at transforming in the input data

The factor causing the adversarial vulnerability is the distortion in the **latent feature space**.

→ **Need to explore the internal logic of the learning models**

UNIVERSITY OF LIVERPOOL

Aim and Contribution

Understand the DNN underlying decision processes

Analyse the latent features learnt by the model

Generate additional test cases based on that

Estimate the **importance** of a neural network's **latent features** by analysing an associated **Bayesian network's sensitivity** to **distributional shifts**

UNIVERSITY OF LIVERPOOL

Bayesian Network

A dimensionality reduction technique using feature extraction algorithms to abstracts the behaviour of a DNN.

Constructing a Bayesian Network:

1. Hidden features extraction

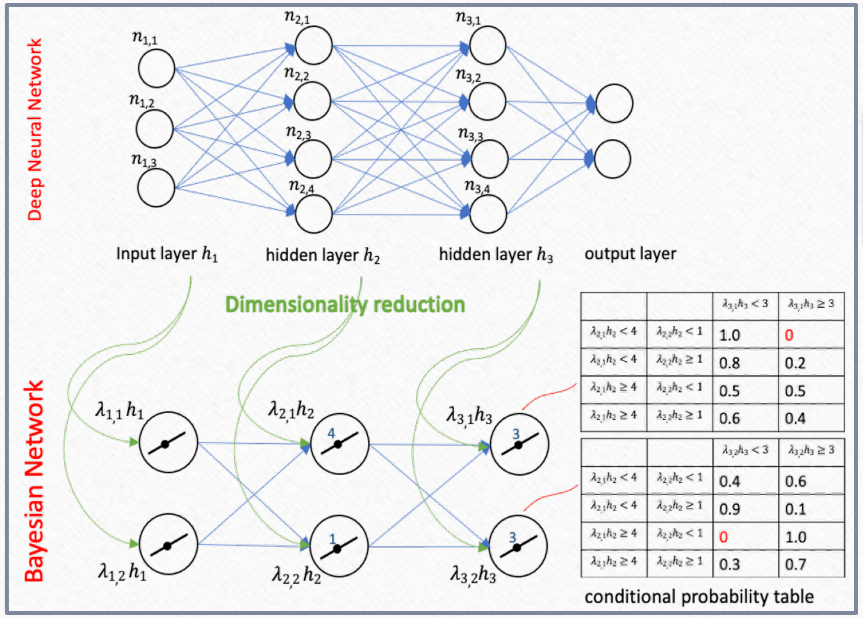
   Map from a high-dimensional space into a *feature space*.

2. Feature space discretisation

   Discretise each feature component into finite feature intervals.

3. Probability tables construction

   Associate each feature with a marginal or a conditional probability table.



		$\lambda_{3,1}h_3 < 3$	$\lambda_{3,1}h_3 \geq 3$
$\lambda_{2,1}h_2 < 4$	$\lambda_{2,2}h_2 < 1$	1.0	0
$\lambda_{2,1}h_2 < 4$	$\lambda_{2,2}h_2 \geq 1$	0.8	0.2
$\lambda_{2,1}h_2 \geq 4$	$\lambda_{2,2}h_2 < 1$	0.5	0.5
$\lambda_{2,1}h_2 \geq 4$	$\lambda_{2,2}h_2 \geq 1$	0.6	0.4

		$\lambda_{3,2}h_3 < 3$	$\lambda_{3,2}h_3 \geq 3$
$\lambda_{2,1}h_2 < 4$	$\lambda_{2,2}h_2 < 1$	0.4	0.6
$\lambda_{2,1}h_2 < 4$	$\lambda_{2,2}h_2 \geq 1$	0.9	0.1
$\lambda_{2,1}h_2 \geq 4$	$\lambda_{2,2}h_2 < 1$	0	1.0
$\lambda_{2,1}h_2 \geq 4$	$\lambda_{2,2}h_2 \geq 1$	0.3	0.7

conditional probability table

BN-based Latent Feature Analysis

Leverage the Bayesian Network to estimate the sensitivity of an individual feature to a controlled distribution shift.

1. Probability calculation for input sample under BN

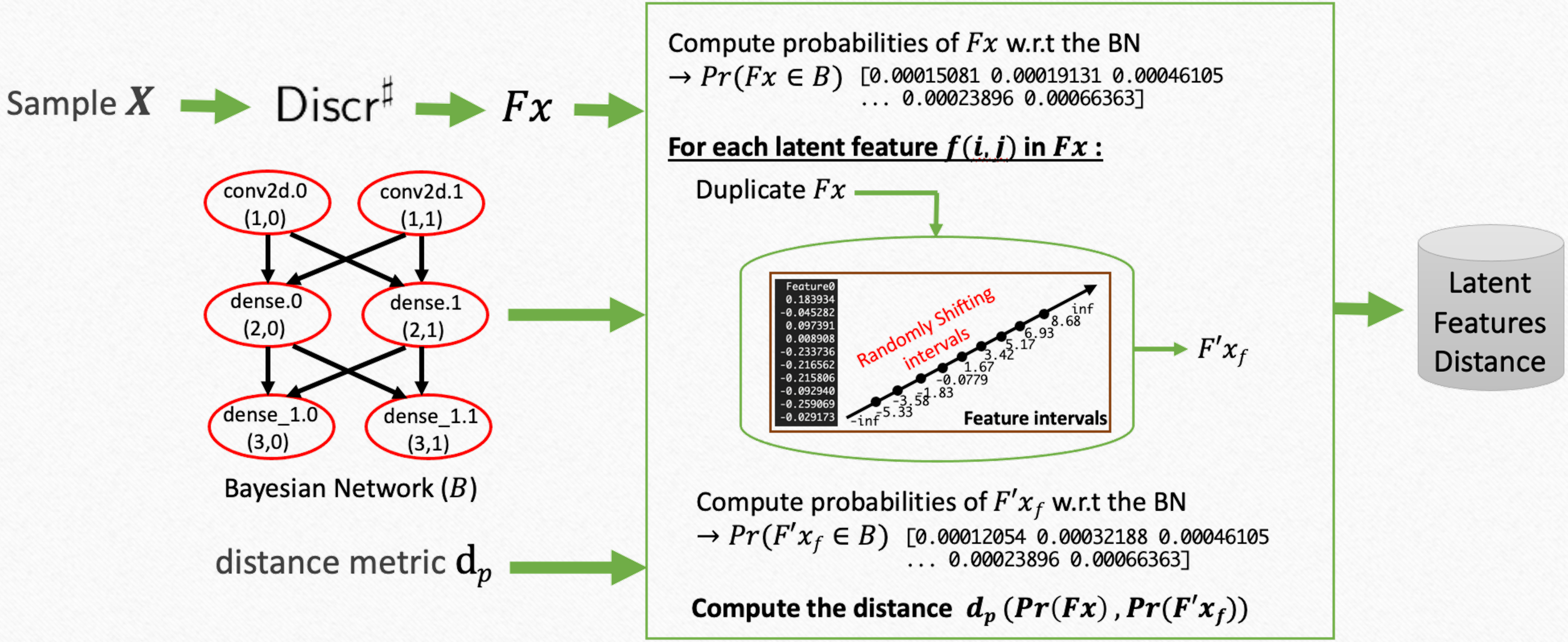
   Perform the feature projection and discretisation step to input sample to obtain the associated feature intervals, and then calculate their probability belonging to the BN distribution.

2. Latent features perturbation

   For each latent feature, randomly shifting its intervals in a selected feature space.

3. Distance computation

   Compute the distance between the original probability vector and the probability vector obtained from the perturbed features.



BN analysis technique to compute the sensitivity of extracted latent features

UNIVERSITY OF LIVERPOOL

Experiments

1) Feature Importance

Associate each extracted feature f with a weight w_f based on the measured sensitivity distance.

Higher distribution change → Higher importance score

distance / perturbed feature	d_{L_1}	d_{L_2}	d_{L_∞}	d_{JS}	$\boxed{d_{corr}}$	d_{cos}	d_{MSE}	d_{RMSE}	d_{MAE}	d_{AF}
(1, 0)	150	0.726	0.009 56	0.224	0.142	0.114	0.000 000 879	0.000 937	0.000 249	0.278
(1, 1)	340	1.18	0.009 89	0.353	0.448	0.361	0.000 002 32	0.001 52	0.000 567	0.735
(2, 0)	325	1.09	0.009 46	0.365	0.332	0.267	0.000 001 98	0.001 41	0.000 541	0.625
(2, 1)	360	1.16	0.0103	0.393	0.395	0.323	0.000 002 24	0.001 50	0.000 600	0.710
(3, 0)	276	0.880	0.008 89	0.258	0.170	0.137	0.000 001 29	0.001 14	0.000 460	0.408
(3, 1)	315	1.07	0.009 60	0.324	0.318	0.264	0.000 001 92	0.001 39	0.000 525	0.608

Example distance measures of the MNIST model

$$w_f = \frac{e^{d_f}}{\sum_{f \in T} e^{d_f}}$$

$w_{(1,\,1)} = 0.192$

$w_{(2,\,1)} = 0.182$

Experiments

2) Sensitivity to Adversarial Distribution Shift

Assess whether the BN abstraction can detect the shift in the distribution of inputs that occurs when the NN is subject to adversarial examples.

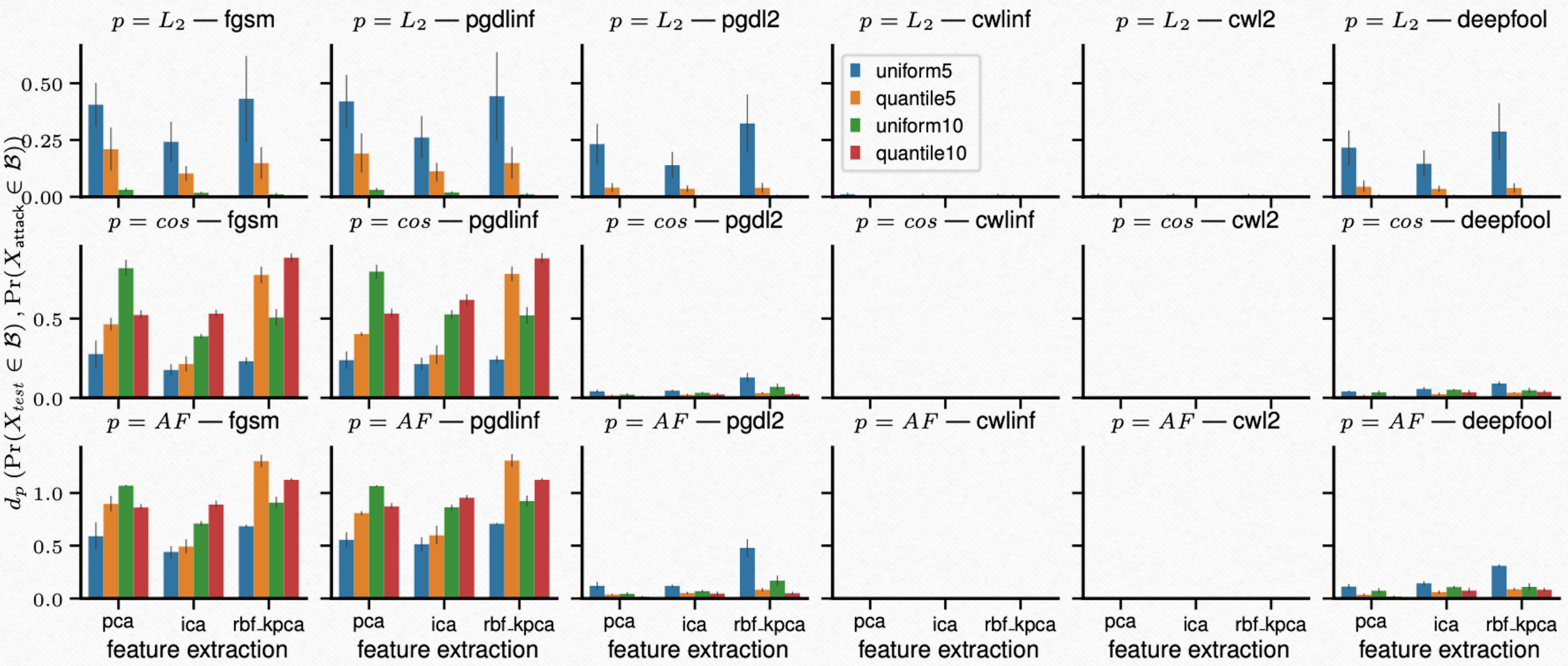
- 👽 **fgsm** is the Fast Gradient Sign Method;
- 👽 **pgdlinf** and **pgdl2** are the Projected Gradient Descent approach;
- 👽 **cwlinf** and **cwl2** ;
- 👽 **Deepfool.**

→ Generate an adversarial dataset X_{attack} from the validation dataset X_{test}

→ Calculate the distance between their probability vectors

UNIVERSITY OF LIVERPOOL

Sensitivity to Adversarial Distribution Shift (MNIST)



UNIVERSITY OF LIVERPOOL

Results

Computing distances between two BN probability distributions clean and perturbed by intervals-shift or adversarial attacks
→ **Detect distribution shift**
→ **Reveal important features**

Conclusion

Advanced a novel technique that employs a BN abstraction to investigate how to measure the importance of high level features when they are used by the neural network to make classification decisions

UNIVERSITY OF LIVERPOOL

Utility of the Feature Weights

First, **visualising** the most **important features** provides insight into the model's internal decisions by highlighting dominating regions in the feature space.

Second, using the **importance measurement** to design high-level **testing** metrics that evaluate the robustness of the DNN.

Third, utilising the obtained **importance** in the **training** process and force the DNN to adjust its parameters according to the most relevant features to the prediction.

UNIVERSITY OF LIVERPOOL

THANK YOU!

UNIVERSITY OF LIVERPOOL