Resilience of Bayesian Layer-Wise Explanations under Adversarial Attacks

Kevin Lee, Shrijesh Siwakoti, Ananth Chillarige

Motivation

- DNN’s have a black box nature, in that there is not much intuition behind the decision making process
- DNN’s are also vulnerable to Adversarial Attacks (hence this class)

Paper argues that these two challenges are **linked**

Background: LRP

- Layer-Wise Relevance Propagation is a method by which we can analyze each the relevance of each pixel to an image's classification
- This Paper introduces the notion of "LRP robustness", and realizes that this robustness is very low under adversarial attacks

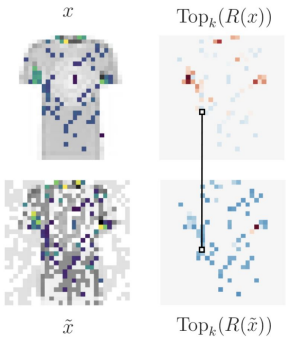


Figure 2: Top_{200} pixels in an image x from Fashion MNIST dataset and an FGSM adversarial perturbation $\tilde{x}$.

Definition 3.1. *Let x be an image with relevance heatmap $R(x,w)$ and $\tilde{x}$ an adversarial perturbation with relevance heatmap $R(\tilde{x},w)$. Let $Top_k(R)$ denote the set of k most relevant pixel indexes in a heatmap R, where $P \in \mathbb{N}$ is the total number of pixels in x and $k \leq P$. The k-*LRP robustness *of x w.r.t. the attack $\tilde{x}$ is*

$$k\text{-}LRP_\rho(x,\tilde{x},w) := \frac{|Top_k(R(x,w)) \cap Top_k(R(\tilde{x},w))|}{k}. \quad (5)$$

Background: Bayesian NN's

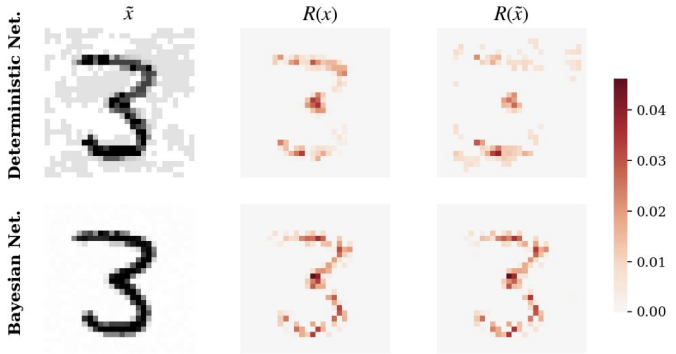
- Captures uncertainty in models by forming an ensemble of DNN models that produce predictions through the posterior predictive distribution.
- The posterior is obtained by combining the prior and likelihood terms
  - Likelihood term takes into account how probable the data is given the network's weights
  - **P(W|D) = P(D|W) P(W)**, where P(W) is the prior, P(D|W) is the likelihood, P(W|D) is the posterior
  - Weights modeled as a distribution

$$p(f(x)|D) = \int dw p(f(x)|w)p(w|D)$$
$$\simeq \sum_{w_j \sim p(w|D)} p(f(x)|w_j)$$

Background: Bayesian NN's (cont.)

- Recent research shows that BNN's are adversarially robust to attacks, which raises the question of how this extends to LRP Robustness
- Bayesian LRP is naturally generalized to the average of all deterministic heat maps from the ensembles

$$\mathbb{E}_{p(w|D)}\left[k\text{-LRP}_{\rho}\left(x, \tilde{x}, w, l\right)\right].$$

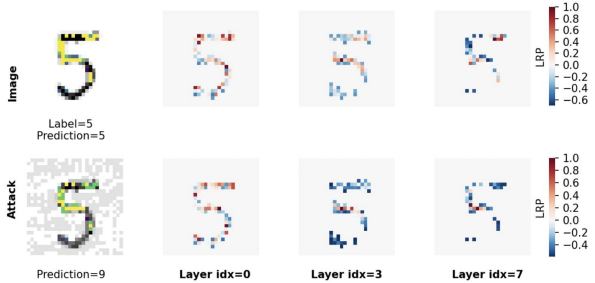


Methods - LRP Robustness

- Uses k-LRP robustness of relevance heatmaps for adversarial attacks to assess how adversarial perturbations affect the explanations
- $k\text{-}LRP_p\ (x, \tilde{x}, w)$ is the fraction of common k most relevant pixels for x and $\tilde{x}$
- Analyze the behavior of LRP representations in internal layers of the network
  - Extends the computation of LRP heatmaps to any feature representation of input x at any learnable layer
- LRP heatmaps do not refer to explanations in the classification phase but rather in the learning phases

Methods - LRP Robustness + Bayes

- LRP heatmaps for image x and FGSM adv. attack x̃. Each layer shows the 100 most relevant pixels.



- Bayesian LRP robustness can be generalized to the bayesian setting. The average of all the deterministic heatmaps from the ensemble:

$$\mathbb{E}_{p(w|D)}\left[k\text{-LRP}_{\rho}\left(x, \tilde{x}, w, l\right)\right].$$

Methods - Geometric Meaning

To better conceptualize, discussing thermodynamic limit of infinite data and infinite expressivity of the network will help understand the Bayesian impact on LRP robustness

Main ingredients:

- data manifold $\mathcal{M}_D$
- a piecewise smooth submanifold of the input space where the data lie
- the true input/output function g(x)

Methods - Geometric Meaning

- Du et al. [2019], Mei et al. [2018], Rotskoff and Vanden-Eijnden [2018] proved that the DNN f(x,w) trained via SGD will converge to the true underlying function g(x) over the whole data manifold $\mathscr{M}_D$
- Because the data manifold is assumed piecewise smooth, it is possible to define a tangent space to the data manifold almost everywhere
- It also defines two operators $\nabla_x^{\perp}$ and $\nabla_x^{\parallel}$
  - define the gradient along the normal and tangent directions to the data manifold $\mathscr{M}_D$ at a point x of a function defined over the whole input space

Methods - Geometric Meaning

Adversarial attacks evaluates the gradient of the loss function

$$\nabla_x L(f,g) = \frac{\delta L(f,g)}{\delta f}\frac{\partial f}{\partial x}$$

- In regards to the thermodynamic limit, the DNN function f(x,w) coincides with the true function everywhere on the data manifold, and the tangent gradient of the loss function is identically zero
- The normal gradient of the loss is unconstrained by the data, and in a high dimensional setting, might achieve very high values along certain directions, creating weaknesses that may be exploited by an adversarial attacker
- BNNs are robust against adversarial attacks

Methods - Geometric Means

- The normal gradient of the loss function is proportional to the normal gradient of the prediction function
- So, the zero-averaging property under the posterior will be inherited by the gradient of the prediction function making

$$E_{p(w|D)}[\nabla_x^{\perp} f(x, w)] = 0.$$

- BNNs in the thermodynamic limit will only retain relevant directions along the data manifold, which correspond to genuine directions of high relevance

Experimental Results

- Compared LRP Robustness using MNIST and Fashion MNIST datasets
- Adversarial attacks include Fast Gradient Sign Method (FGSM) and Projected Gradient Method (PGD)

- Bayesian explanations are more robust under adversarial attacks than deterministic architectures.
- LRP robustness scores for both the VI and HMC training were significantly higher than their deterministic counterparts.

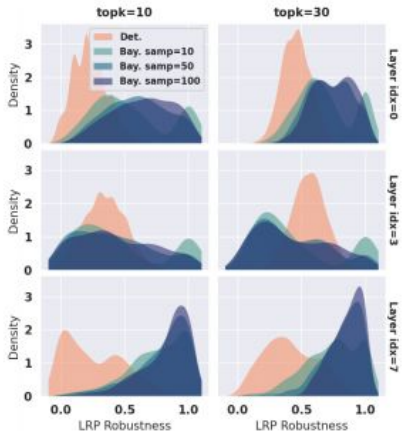


Figure 4: LRP robustness distributions computed on 500 test points from MNIST dataset. Bayesian networks are trained with VI and tested on an increasing number of samples (10, 50, 100). Layer indexes refer to the learnable layers in the architecture (Tab. 1 in the Appendix), which is shared across models.

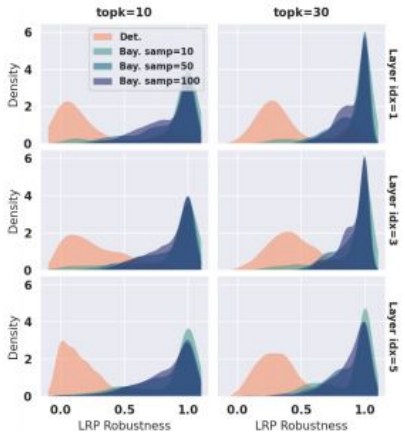


Figure 6: LRP robustness distributions computed on 500 test points from MNIST dataset. Bayesian networks are trained with HMC and tested on an increasing number of samples (10, 50, 100). Layer indexes refer to the learnable layers in the architecture (Tab. 2 in the Appendix), which is shared across models.

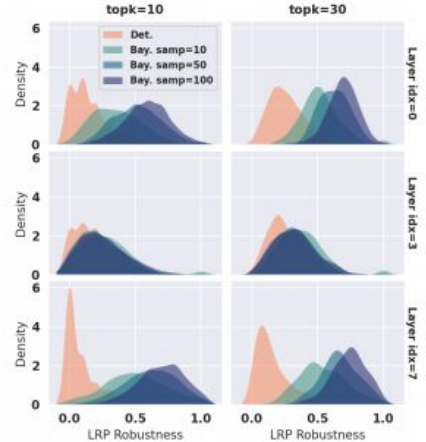


Figure 5: LRP robustness distributions computed on 500 test points from Fashion MNIST dataset. Bayesian networks are trained with VI and tested on an increasing number of samples (10, 50, 100). Layer indexes refer to the learnable layers in the architecture (Tab. 1 in the Appendix), which is shared across models.

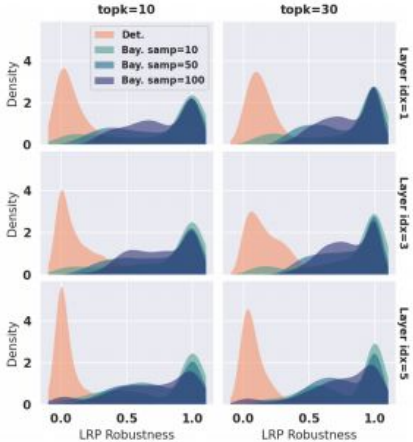


Figure 7: LRP robustness distributions computed on 500 test points from Fashion MNIST dataset. Bayesian networks are trained with HMC and tested on an increasing number of samples (10, 50, 100). Layer indexes refer to the learnable layers in the architecture (Tab. 2 in the Appendix), which is shared across models.

- LRP robustness is correlated with Softmax robustness with Bayesian NNs.
- The attacks that are more successful (lower softmax robustness) alter the interpretation of the classification more substantially.
- Therefore, BNNs are likely to represent genuine directions of change of the true underlying decision function along the data manifold.

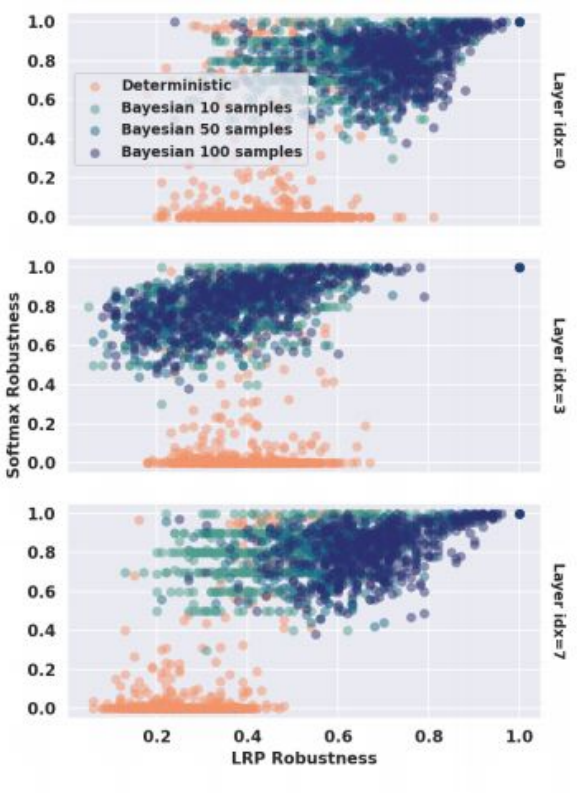


Figure 8: LRP vs Softmax Robustness of deterministic and Bayesian NNs trained on Fashion MNIST dataset. LRP Robustness is computed on the 100 most relevant pixels, i.e. using Top_{100}. Bayesian Networks are trained with VI and tested on an increasing number of samples $(10, 50, 100)$. Layer indexes refer to the learnable layers in the architecture (Tab. 1 in the Appendix), which is shared across all models.

Conclusion

- Bayesian networks, in the limit of infinite data, remedy the weaknesses of Deterministic NNs where the gradients of the loss function and the prediction functions are orthogonal to the data manifold, by averaging out irrelevant gradient directions.
- BNN models also have higher LRP robustness values than DNN networks.
- With BNN models, the LRP robustness correlates with softmax robustness, meaning that they capture relevant parameterizations of the data manifold.