Network Inversion for Uncertainty-Aware Out-of-Distribution Detection (Student Abstract)

Pirzada Suhail, Rehna Afroz Shaik, Gouranga Bala, Amit Sethi

IIT Bombay
psuhail@iitb.ac.in

Abstract

Out-of-distribution (OOD) detection and uncertainty estimation (UE) are critical components for building safe machine learning systems, however the two problems have, until recently, separately been addressed. In this work, we propose a novel framework that combines network inversion with classifier training to simultaneously address both OOD detection and uncertainty estimation. For a standard n-class classification task, we extend the classifier to an (n+1)-class model by introducing a "garbage" class, initially populated with random gaussian noise to represent outlier inputs. After each training epoch, we use network inversion to reconstruct input images corresponding to all output classes that initially appear as noisy and incoherent and are therefore excluded to the garbage class for retraining the classifier. This cycle of training, inversion, and exclusion continues iteratively till the inverted samples begin to resemble the in-distribution data more closely, with a significant drop in the uncertainty, suggesting that the classifier has learned to carve out meaningful decision boundaries while sanitising the class manifolds by pushing OOD content into the garbage class. During inference, this training scheme enables the model to effectively detect and reject OOD samples by classifying them into the garbage class. Furthermore, the confidence scores associated with each prediction can be used to estimate uncertainty for both in-distribution and OOD inputs. Unlike prior approaches, our method requires no external OOD datasets or post-hoc calibration, offering a simple and interpretable solution to ensure robustness in classification under distributional shift while providing a unified solution to the dual challenges of OOD detection and uncertainty estimation.

Introduction

The increasing deployment of machine learning models in high-stakes, real-world applications—such as autonomous driving, medical diagnosis, and financial decision-making—has underscored the importance of model reliability and robustness. A key limitation of modern neural networks is their tendency to produce overconfident predictions (Suhail and Sethi 2025) even on inputs that lie far outside the training distribution. This makes it crucial to develop models capable of both out-of-distribution (OOD) detection—the ability to identify inputs that fall outside the training distribution—and uncertainty estimation (UE)—the ability to quantify confidence in predictions to ensure safe decision-making under distributional shift.

Both capabilities are vital for trustworthiness in deployment scenarios where the data encountered during inference may deviate from the training distribution in subtle or unexpected ways. Although these two problems are inherently linked, most existing approaches treat them separately, often relying on post-hoc calibration techniques or auxiliary OOD datasets, which may not always be available.

Recent work in (Ansari et al. 2022) proposed Autoinverse, a framework for neural network inversion that prioritizes solutions near reliable training samples, using embedded regularization and predictive uncertainty minimization to improve robustness. Later (Lu et al. 2023) introduced a semantically coherent OOD detection (SCOOD) approach by combining uncertainty-aware optimal transport with dynamic cost modeling and inter-cluster enhancements. While (Chen et al. 2024) developed a Gaussian process-based model that operates solely on in-distribution data, defining predictive uncertainty scores without requiring OOD examples during training. Similarly, (Charpentier, Zügner, and Günnemann 2020) presents PostNet, which employs normalizing flows to model posterior distributions over predicted probabilities, allowing reliable uncertainty estimation and effective OOD discrimination—even without OOD supervision.

In this work, we propose a novel framework that leverages network inversion(Suhail 2024; Suhail and Sethi 2024), not only to detect OOD inputs but also to estimate prediction uncertainty, unifying the two objectives in a single training procedure. By extending a standard (n+1)-class model with an auxiliary garbage class, and iteratively refining the model using inverted reconstructions, we encourage the network to carve out clean decision boundaries while isolating ambiguous or anomalous regions.

Methodology

Our unified training approach integrates out-of-distribution (OOD) detection and uncertainty estimation (UE) into a single framework using network inversion and an auxilary garbage class. For an n-class classification task, we extend the classifier to an (n+1)-class model by introducing an additional "garbage" class designed to absorb anomalous inputs.

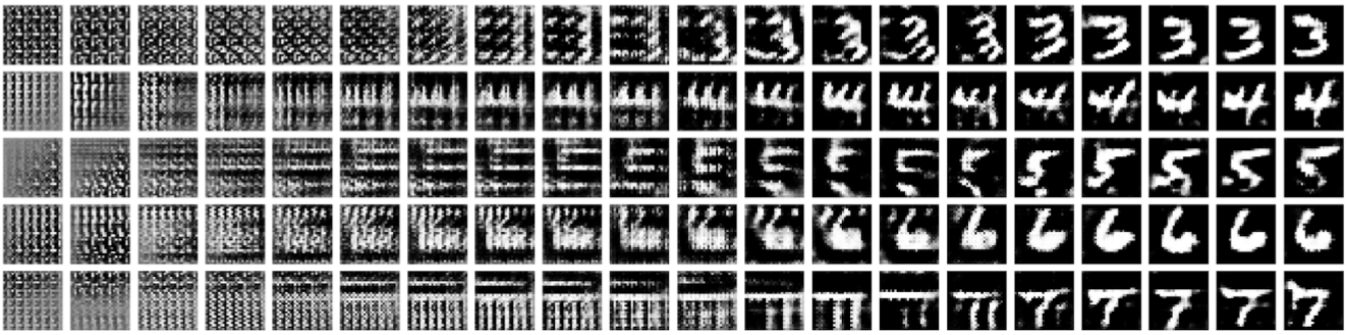

Figure 1: Inverted Samples across epochs, for 5 different classes beginning to resemble the training data.

This garbage class is initially populated with random Gaussian noise, representing OOD samples.

Between successive training epochs, we perform network inversion as in (Suhail and Sethi 2024) to reconstruct samples from the input space of the classifier for all output classes. Given the vastness of the input space, during early training stages, these reconstructions tend to be visually incoherent and do not resemble real data, reflecting the model's incomplete or uncertain understanding of the class manifolds. These reconstructions are assigned to the garbage class and added to the training set for the subsequent epochs. In subsequent epochs the classifier is trained using a weighted cross-entropy loss to account for the class imbalance introduced by addition of garbage samples.

By iteratively repeating this cycle of training, inversion, and exclusion, the model gradually learns to refine the decision boundaries while pushing anomalous content into the garbage class. As the training progresses, inverted samples in Fig 1 begin to look like training data, indicating that the classifier has effectively carved out the in-distribution manifold while isolating outliers into the garbage class.

During inference, this training procedure equips the classifier to identify and reject out-of-distribution (OOD) inputs by assigning them to the garbage class. Additionally, the softmax confidence scores corresponding to class predictions can be used to assess the model's uncertainty. Low softmax confidence on in-distribution predictions indicates ambiguous or uncertain inputs, while high confidence in the garbage class suggests a strong belief that the input is OOD. We quantify uncertainty using the softmax confidence values across all $n+1$ output classes by capturing how sharply peaked or spread out the model's predictive distribution is. The uncertainty estimate for a prediction $\mathbf{p}$ is given by:

$$\mathrm{UE}(\mathbf{p}) = 1 - \frac{\sum_{i=1}^{n+1} \left(p_i - \frac{1}{n+1} \right)^2}{\sum_{i=1}^{n+1} \left(\delta_{i,k} - \frac{1}{n+1} \right)^2} \tag{1}$$

where $k = \arg\max_i p_i$ and $\delta_{i,k}$ is the Kronecker delta. The resulting score ranges from 0 to 1, providing an interpretable measure of confidence by computing the squared distance between the predicted vector $\mathbf{p}$ and the uniform distribution, normalized by the maximum possible distance under a one-hot prediction.

Quantitative Results

We evaluate the effectiveness of our approach to uncertainty-aware out-of-distribution detection across four benchmark image classification datasets: MNIST (Deng 2012), FashionMNIST (Xiao, Rasul, and Vollgraf 2017), SVHN, and CIFAR-10. To assess OOD detection performance, we follow a one-vs-rest evaluation strategy: the model is trained exclusively on one dataset and evaluated on the remaining three as OOD sources.

Train \ Test	**MNIST**	**FMNIST**	**SVHN**	**CIFAR-10**
MNIST	99.1	89.5	99.1	99.4
FMNIST	85.2	92.6	96.3	95.7
SVHN	93.6	94.9	89.4	87.6
CIFAR-10	97.8	95.7	88.2	85.5

Table 1: Accuracy for both in- and out-of-distribution datasets.

Table 1 presents the accuracy for uncertainty-aware OOD detection across all pairs of datasets. Each row corresponds to a model trained on one of the datasets and diagonal entries represent the in-distribution (ID) performance measured on the standard test set of the training dataset. Off-diagonal entries indicate OOD detection performance, where the accuracy represents how well the model distinguishes out-of-distribution samples by correctly classifying them into the garbage class. High values across both diagonal and off-diagonal entries demonstrate that the model maintains strong classification performance on ID data while reliably identifying OOD inputs. We also observe that while the majority of OOD samples are correctly assigned to the garbage class, a small percentage of the samples can still be misclassified into in-distribution classes. However, a significant finding is that on average, the least confidently classified in-distribution sample is still more confidently classified compared to the most confidently misclassified out-of-distribution sample, making them easily separable.

Future work can also consider the use of n separate garbage classes—one for each of the in-distribution classes—for fine-grained separation of OOD samples and weighted individual OOD sample contribution to the loss while retraining the classifier based on uncertainty.

References

Ansari, N.; Seidel, H.-P.; Ferdowsi, N. V.; and Babaei, V. 2022. Autoinverse: Uncertainty Aware Inversion of Neural Networks. In *Advances in Neural Information Processing Systems(NeurIPS)*.

Charpentier, B.; Zügner, D.; and Günnemann, S. 2020. Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, Y.; Sung, C.-L.; Kusari, A.; Song, X.; and Sun, W. 2024. Uncertainty-Aware Out-of-Distribution Detection with Gaussian Processes. *arXiv:2412.20918*.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Lu, F.; Zhu, K.; Zhai, W.; Zheng, K.; and Cao, Y. 2023. Uncertainty-Aware Optimal Transport for Semantically Coherent Out-of-Distribution Detection. In *Computer Vision and Pattern Recognition (CVPR)*.

Suhail, P. 2024. Network Inversion of Binarised Neural Nets. In *The Second Tiny Papers Track at ICLR 2024*.

Suhail, P.; and Sethi, A. 2024. Network Inversion of Convolutional Neural Nets. In *Muslims in ML Workshop co-located with NeurIPS 2024*.

Suhail, P.; and Sethi, A. 2025. Network Inversion for Generating Confidently Classified Counterfeits. arXiv:2503.20187.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:1708.07747.