Robust and Predictable Out-of-Distribution (OOD) Data Detection and Reasoning for Safety-Critical Systems

November 26, 2021

Deep Neural Network (DNN) based https://www.overleaf.com/project/6180a985ca2f4286974ced09intelligent systems have been rapidly deployed in the real worhttps://www.overleaf.com/project/6180a985ca2f4286974ced09ld, both as standalone software as well as components in cyber-physical systems. Many of them, such as autonomous driving and energy management, are safety and security critical. Among DNN perception systems, object/data classification is a very important, if not the central, task. Inevitably, it will face unknown or novel objects/data in the real world, namely Out-of-Distribution (OOD) samples, hence it is very likely to give wrong outputs in such situations. Detecting, reasoning about and rejecting OOD samples is fundamental to increasing a DNN's robustness in the open world. Over the years, many detection methods have been proposed by the OOD community. Some of them utilize the object/data classifier outputs as OOD features, while others look into generative learning techniques.

So far, many OOD detection methodologies have been proposed and tested over benchmark data sets. Broadly, these methods can be grouped into two categories. One group works through a supervised learning framework. By tapping into the statistics of neural weights in well-trained deep neural classifiers, novel metric scores that estimate the trustworthiness of a classifiers' outputs were proposed (e.g., [15, 16, 18]). OOD samples would receive trust scores lower than In-Distribution (ID) samples. More recently, an unsupervised generative model based approach has been explored by utilizing the information bottleneck property of Variational Autoencoder (VAE) to learn a posterior distribution approximation $q(z \mid x)$ to infer the unknown true latent distribution $p(z)$ of a training data set. Equation 1 is the current widely adopted evidence lower bound (ELBO) objective formulated by [13], where $\beta = 1$. The first term on the right hand side, likelihood, measures the reconstruction accuracy of input. The second term measures the distribution's discrepancy in latent space. The learned latent space is better disentangled when adding an adjustable hyperparameter β to balance the two terms [12].

$$\log p(x) \geqslant \mathbb{E}_{q(z|x)}\big[\log p(x \mid z)\big] - \beta D_{KL}(q(z \mid x) \parallel p(z)) \tag{1}$$

Typically, the posterior approximation resides in a low dimensional latent space $\mathbb{Z}$, hundreds to thousands of times lower than the corresponding input space $\mathbb{X}$. Given test samples x, the VAE decoder reconstructs inputs with high likelihoods if inputs are ID, and otherwise with low likelihoods. The encoder of the VAE produces data-driven latent variables. The discrepancy of these latent variables' distribution to the latent prior $p(z)$ is commonly measured by Kullback–Leibler (KL) divergence. Detection methods utilized either the distribution discrepancy in latent space (e.g, [21, 19, 7, 23]) or the likelihood in input space as an OOD measure [2, 4]. For OOD reasoning, methodologies to map key input features (also called generative factors) to specific dimensions of the latent space have also been developed [19]. The increasing deployment of DNNs into safety-critical systems in the open world have driven OOD detection into an active research field. More related literature can be found in a comprehensive survey [1].

Given the complex nature of the OOD detection and reasoning task, partly due to the high dimensionality of inputs, most solutions in the literature deploy DNNs as presented above. Analyzing and quantifying the uncertainty in such DNN based OOD decisions is then fundamental to the design and formal analysis of safety-critical systems. For example, the VerifAI toolkit supports a variety of formal modelling of autonomous systems as well as dynamic input environments for automatic falsification of design specifications at a system level as well as runtime monitoring of simulation-based verification [22]. Along the same lines, [11] proposed a method for screening and classifying simulation-based test data with responsibility sensitive safety specifications. In the context of formal design and analysis, these works focused on verifying system-level conformity or violation against certain safety specifications. On the other hand, many statistical learning-based techniques have also been proposed to characterize the uncertainty in DNNs directly, such as variational inference [6], Monte Carlo dropout [8], and deep ensembles [14].

PhD Objectives:

1. **Hybrid-AI based OOD detection and reasoning**: Although some of the above OOD detection methods reported impressive performance in detecting certain key OoD factors (e.g., rain intensity and illumination levels for autonomous driving), their generalizability to other more complex OOD factors and robustness to noisy inputs have not yet been investigated. Besides, approaches based on generative models such as VAEs, require the latent space dimension to be much higher than the number of expected OOD factors, thus introducing challenges in the selection of hyperparameters (e.g., value of β or the number of latent dimensions). In summary, effective OOD detection and reasoning methods that will perform robustly across multiple high dimensional data sets for real-time safety-critical applications are still lacking.

   To address the above issues, in this task we plan to explore multi-modal inputs (e.g., radar and lidar together with video streams) and physics-based system models to augment the OOD detection and reasoning capabilities of DNN based solutions. In particular, focusing on the generative model based approaches (e.g., VAEs and their variants), we aim to develop an integrated Hybrid-AI technique that uses the physics-based models parameterized with low dimensional data (e.g., expected trajectory for objects obtained with radar/lidar data) to improve the generalizability and robustness of the OOD decisions.

2. **Probabilistic quantification of OOD uncertainty (Probably Approximately Correct - PAC)** Learning-based techniques for characterizing the uncertainty in DNNs, such as those presented above, do not provide mathematically provable probabilistic guarantees over the probability of performance of tasks learned with available data when dealing with concrete DNN models. Furthermore, some of the ensemble approaches require multiple DNN inferences to produce one output, and hence their execution time places a challenge to deployments in real-time applications.

   In contrast, PAC learning [10] offers a mathematical framework for seeking a probability bound over the learning task of our interest, namely, OOD detection. Over the past years, PAC learning has been advanced by both machine learning and optimization fields. In the former area, many works contributed to the theory along the PAC-Bayesian learning direction [9]. For example, when tapping into data-dependent oracle prior [5], non-vacuous classification performance bounds were obtained over benchmark data sets MNIST and Fashion-MNIST. Recently, the scenario approach [3], a general probabilistic framework for robust control design, has been applied for robustness analysis of Support Vector Machines (SVM) [20]. Orthogonally, [17] transformed the analysis of the robustness of DNNs to adversarial perturbation into a linear programming optimization problem. However, their approach does not entail a probabilistic guarantee as in PAC learning.

   So far, the scenario optimization approach has not been used to characterize the probabilistic uncertainty in DNN based OOD detection and reasoning tasks. In this research activity, we plan to explore this direction using black-box (deriving guarantees based only on DNN inputs/outputs) and grey-box (additionally using DNN intermediate layer statistics to derive guarantees) techniques. The objective would be to derive a relatively simple surrogate model using scenario optimization to characterize the uncertainty in the DNN based OOD detection and reasoning task.

References

[1] Saikiran Bulusu et al. “Anomalous example detection in deep learning: A survey”. In: *IEEE Access* 8 (2020), pp. 132330–132347.

[2] F. Cai and X. Koutsoukos. “Real-time Out-of-distribution Detection in Learning-Enabled Cyber-Physical Systems”. In: *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)* (2020), pp. 174–183.

[3] Marco C Campi, Simone Garatti, and Maria Prandini. “The scenario approach for systems and control design”. In: *Annual Reviews in Control* 33.2 (2009), pp. 149–157.

[4] Erik Daxberger and José Miguel Hernández-Lobato. “Bayesian variational autoencoders for unsupervised out-of-distribution detection”. In: *arXiv preprint arXiv:1912.05651* (2019).

[5] Gintare Karolina Dziugaite et al. “On the role of data in PAC-Bayes”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 604–612.

[6] Sebastian Farquhar, Michael A Osborne, and Yarin Gal. “Radial Bayesian neural networks: beyond discrete support in large-scale Bayesian deep learning”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1352–1362.

[7] Yeli Feng, Daniel Jun Xian Ng, and Arvind Easwaran. “Improving Variational Autoencoder Based Out-of-Distribution Detection for Embedded Real-Time Applications”. In: *ACM Transactions Embedded Computer Systems* 20.5s (2021).

[8] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[9] Benjamin Guedj and John Shawe-Taylor. “A Primer on PAC-Bayesian Learning”. In: *ICML 2019-Thirty-sixth International Conference on Machine Learning*. 2019.

[10] David Haussler. *Probably approximately correct learning*. University of California, Santa Cruz, Computer Research Laboratory, 1990.

[11] Mohammad Hekmatnejad, Bardh Hoxha, and Georgios Fainekos. “Search-based test-case generation by monitoring responsibility safety rules”. In: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2020, pp. 1–8.

[12] Irina Higgins et al. “beta-vae: Learning basic visual concepts with a constrained variational framework”. In: (2016).

[13] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: *arXiv preprint arXiv:1612.01474* (2016).

[16] Kimin Lee et al. “A simple unified framework for detecting out-of-distribution samples and adversarial attacks”. In: *Advances in neural information processing systems* 31 (2018).

[17] Wang Lin et al. “Robustness verification of classification deep neural networks via linear programming”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 11418–11427.

[18] Philipp Oberdiek, Matthias Rottmann, and Hanno Gottschalk. “Classification uncertainty of deep neural networks based on gradient information”. In: *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer. 2018, pp. 113–125.

[19] Shreyas Ramakrishna et al. "Efficient Out-of-Distribution Detection Using Latent Space of β-VAE for Cyber-Physical Systems". In: *CoRR* abs/2108.11800 (2021). URL: https://arxiv.org/abs/2108.11800.

[20] Roberto Rocchetta, Milan Petkovic, and Qi Gao. "Scenario-based Generalization bound for Anomaly Detection Support Vector Machine Ensembles". In: *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*. 2020.

[21] Vijaya Kumar Sundar et al. "Out-of-Distribution Detection in Multi-Label Datasets using Latent Space of beta-VAE". In: *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE. 2020, pp. 250–255.

[22] Hazem Torfah et al. "Formal Analysis of AI-Based Autonomy: From Modeling to Runtime Assurance". In: *International Conference on Runtime Verification*. Springer. 2021, pp. 311–330.

[23] Michael Yuhas et al. "Embedded Out-of-Distribution Detection on an Autonomous Robot Platform". In: Destion '21. Nashville, Tennessee: Association for Computing Machinery, 2021, pp. 13–18. ISBN: 9781450383165. DOI: 10.1145/3445034.3460509.