Engineer/Master 2 level internship

Statistical Method For Analyzing DNN Classifier Decisions

Research unit: LIST3N
University of Technology of Troyes
12 rue Marie Curie - CS 42060 - 10004 TROYES CEDEX
Duration: 6 months
Start date: February-March 2026
Contact: A. Baussard (alexandre.baussard@utt.fr)

CONTEXT

Anomaly detection is an important topic in data analysis. The definition of an anomaly may vary depending on the scope or purpose, but generally it is an observation that deviates significantly from the rest of the other observations as if it were generated by a different process. Thus, in some cases, anomaly detection can improve data quality by deleting or replacing anomalous data. In other cases, anomalies reflect an event and provide useful new knowledge. For example, anomaly detection can prevent material damage and thus encourage preventive maintenance in the industrial sector. They can also be a sign of the presence of a new phenomenon that will have to be taken into account, such as the detection of a new class linked, for example, to the appearance of a new model of vehicle, aircraft or other in surveillance applications.
In recent years, methods based on deep neural networks have shown their potential in many areas. However, the deployment of these models in operational contexts requires that they not only perform well but also be robust in the face of any new situation that varies more or less strongly from those encountered during the learning process. It has been shown that unfortunately these networks can provide a very high level of confidence to a result when it is not correct. It is therefore necessary to develop methods to reduce these risks of error. To do this, it is necessary to understand and explain the decisions made by these networks in order to be able to detect any decision errors potentially related to the presence of an anomaly. Thus, there are several advantages in the development of these methods, which will be the subject of this project, such as: detecting and processing anomalies or non-distribution data, having a better understanding of decision-making, providing an operator with the level of confidence in the decision, etc.

SUBJECT

In this project we will focus on the detection of out-of-distribution (OOD) data in Depp Neural Network (DNN) for image classification applications. The methods of interest are based on the definition of statistical decisions from DNN outputs. The detector is then independent of the DNN classification models considered, which ensures the conservation of the classifier's performance. There are other approaches that consist of modifying DNN but they require a new learning phase that can lead to performance losses.

An approach has already been developed in the laboratory and will therefore constitute the basic building block of the work. The objectives of the proposed work are as follows:

- Test and evaluate two PyTorch packages including several OOD detection methods for image classification tasks and to compare them on different criteria (ease of use, performance, robustness, coding quality, etc.). One of these libraries will be used to compare the performance of our approach with the state-of-the-art.
- Contribute to finalizing the statistical approach developed in the laboratory, in particular by proposing analyses of the detector behavior. Propose evolutions and improvements of the proposed pipeline.
- Depending on the progress of the work, an initial solution for novelty detecting (i.e. detecting the appearance of a new class among the detected data as OOD) may be implemented.

REFERENCES

[1] T.M. Carvalho, M. Vellasco, J.S. Amaral, Out-of-Distribution Detection in Deep Learning Models: A Feature Space-Based Approach, IJCNN, 2023
[2] R. Chalapathy, S. Chawla, Deep Learning for anomaly detection: A survey, arXiv:1901.03407v2, 2019.
[3] M. Haroush, T. Frostig, D. Soudry, A Statistical Framework for Efficient Out of Distribution Detection in Deep Neural Networks, ICLR 2022.
[4] Y-C. Hsu, Y. Shen, H. Jin, Z. Kira, Generalized ODIN: Detecting Out-of-distribution Image without Learning from Out-of-distribution Data, CVPR 2020.
[5] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, NeurIPS 2018.

DESIRED PROFILE

Master's or engineering students with a background in applied mathematics (particularly statistics) and machine learning (deep learning).
It is necessary to know Python, PyTorch and scikit-learn.