INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 06, june 2023, pp : 1408-1411

SUPPORTING DNN SAFETY ANALYSIS AND RETRAINING USING UNSUPERVISED LEARNING

Prof. Sneha A. Khaire[1], Ketan Dhole[2], Sarika Bhingane[3], Ruthik Wankhede[4], Kunal Kshirsagar[5]

[1]Assistant Professor, Department of Information Technology, Sandip Institute of Technology and Research Centre, College of Engineering, Nashik - 422213

[2,3,4,5]BE Students, Department of Information Technology, Sandip Institute of Technology and Research Centre, College of Engineering, Nashik - 422213

DOI: https://www.doi.org/10.58257/IJPREMS31567

ABSTRACT

The development of deep neural networks (DNNs) has greatly advanced the field of artificial intelligence, but their use in safety-critical applications such as autonomous driving, medical diagnosis, and financial forecasting requires rigorous analysis and verification to ensure their reliability and trustworthiness. In this context, unsupervised learning has emerged as a promising technique for supporting DNN safety analysis and retraining, by enabling the detection of anomalies, errors, and biases in the input data, as well as the identification of data-driven features and representations that can enhance the generalization and robustness of the model. This paper presents an overview of the recent research on unsupervised learning methods for DNN safety, including autoencoders, generative models, clustering, and outlier detection, and their applications in detecting adversarial attacks, handling missing data, improving fault tolerance, and mitigating dataset bias. We also discuss the challenges and opportunities of incorporating unsupervised learning into the DNN development pipeline and highlight the need for further research and standardization to ensure the scalability, interpretability, and reproducibility of these methods.

Keywords: DNN, Pipeline, Research, Unsupervised Learning.

1. INTRODUCTION

Deep neural networks (DNNs) have made significant advancements in various domains, including image recognition, natural language processing, and decision-making systems. However, the widespread adoption of DNNs in safety-critical applications necessitates careful analysis and retraining to ensure their reliability and trustworthiness. Safety concerns arise due to potential vulnerabilities, biases, and adversarial attacks that can compromise the performance and robustness of DNN models. Unsupervised learning techniques have emerged as a promising approach to address these concerns and enhance DNN safety analysis and retraining. Traditional supervised learning relies on labeled data, where each input sample is associated with a corresponding output label. While supervised learning has proven successful, it requires large amounts of labeled data, which can be expensive and time-consuming to acquire, especially in safety-critical domains. Unsupervised learning, on the other hand, offers an alternative by leveraging unlabeled data to identify underlying patterns, structures, and anomalies within the data. The main objective of using unsupervised learning in the context of DNN safety analysis and retraining is twofold. Firstly, it aims to enhance the understanding of the input data distribution and identify potential outliers or abnormal patterns that may adversely affect the DNN's performance. Secondly, it facilitates the identification of data-driven features and representations that can improve the DNN's generalization and robustness. This paper provides an overview of the research and developments in using unsupervised learning techniques for supporting DNN safety analysis and retraining. We discuss various methods such as autoencoders, generative models, clustering, and outlier detection, and their applications in detecting adversarial attacks, handling missing data, improving fault tolerance, and mitigating dataset bias. By incorporating unsupervised learning into the DNN development pipeline, it is possible to identify and address potential safety issues proactively. Moreover, unsupervised learning methods can aid in identifying biases and enhancing fairness in DNN models, making them more reliable and trustworthy across diverse user populations. However, there are challenges and opportunities associated with the integration of unsupervised learning into DNN safety analysis. These include interpretability of unsupervised models, scalability to large-scale datasets, and the need for standardization to ensure reproducibility and comparability of results. Addressing these challenges will be crucial to promote the adoption of unsupervised learning techniques in the development of safe and reliable DNNs. In summary, this paper aims to highlight the significance of unsupervised learning in supporting DNN safety analysis and retraining. We provide insights into the various methods and applications, discuss the challenges, and emphasize the need for further research and standardization in this evolving field. By harnessing the power of unsupervised learning, we can enhance the safety and reliability of DNNs in critical applications, paving the way for their widespread adoption.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 06, june 2023, pp : 1408-1411

2. METHODOLOGY

The methodology for supporting DNN safety analysis and retraining using unsupervised learning typically involves several steps. Here is a generalized methodology that can be followed:

- **Data Preprocessing:** Prepare the dataset for unsupervised learning by cleaning, normalizing, and transforming the data as required. This step ensures that the data is in a suitable format for the unsupervised learning algorithms.
- **Feature Extraction:** Extract meaningful features from the dataset to capture the essential characteristics of the input data. This step helps in representing the data in a more compact and informative manner, facilitating subsequent analysis.
- **Unsupervised Learning:** Apply unsupervised learning algorithms to the preprocessed dataset. Common unsupervised learning techniques include clustering algorithms (e.g., k-means, hierarchical clustering) and dimensionality reduction methods (e.g., PCA, t-SNE). These algorithms can reveal patterns, structures, and relationships in the data without relying on labeled examples.
- **Anomaly Detection:** Utilize unsupervised learning for anomaly detection to identify unexpected or abnormal patterns in the data. Anomalies can indicate potential safety risks or vulnerabilities in the DNN model. Techniques such as density-based clustering, one-class SVM, or autoencoders can be employed for anomaly detection.
- **Bias Detection and Mitigation:** Analyze the learned representations and latent features to identify biases in the data or model behavior. Unsupervised learning can help in uncovering unintended biases that might affect the fairness and reliability of the DNN model. Once biases are detected, appropriate mitigation techniques can be applied, such as debiasing algorithms or data augmentation strategies.
- **Adversarial Defense:** Use unsupervised learning to detect and understand adversarial perturbations. Adversarial attacks aim to manipulate the model's decision-making process by introducing carefully crafted inputs. Unsupervised techniques can help in identifying abnormal patterns or perturbations in the input space, enabling the development of robust defenses against adversarial attacks.
- **Retraining and Improvement:** Incorporate the insights gained from the unsupervised analysis into the retraining process. The identified anomalies, biases, and adversarial vulnerabilities can guide the selection and generation of new training samples. This process aims to improve the model's robustness, generalization, and safety by retraining with the enhanced dataset.
- **Evaluation and Validation:** Evaluate the performance of the retrained model using appropriate metrics and validation techniques. This step ensures that the model's safety, reliability, and generalization capabilities have improved after the application of unsupervised learning techniques.

It is worth noting that the specific implementation details and choice of algorithms may vary depending on the specific application and problem domain. Additionally, the methodology should be complemented with other techniques, such as supervised learning, formal verification, and extensive testing, to ensure a comprehensive approach to DNN safety analysis and retraining.

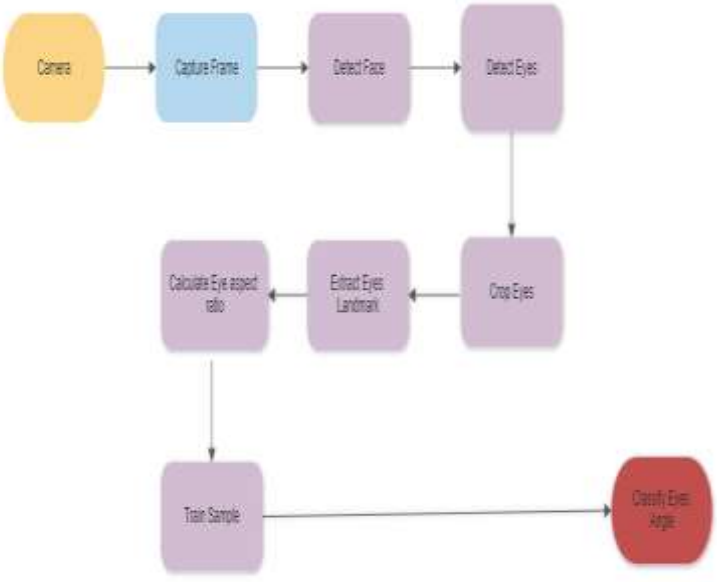

Figure 1: Workflow of proposed system

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 06, june 2023, pp : 1408-1411

Impact Factor : 5.725

3. MODELING AND ANALYSIS

There are various important factors to consider when modelling and analysing data to facilitate DNN safety analysis and retraining utilizing unsupervised learning. The DNN error root causes can be found using the HUDD tool, which also supports safety analysis. Additionally, it can retrain DNNs more successfully than current methods. Another method to enable DNN safety assessments and retraining is heatmap-based unsupervised learning. There has been little advancement in automated functional safety analysis assistance for DNN-based systems. HUDD has various restrictions, such as the fact that it can only evaluate DNN implementations that have been extended to compute LRP. In order to enable DNN safety analysis and retraining, heatmap-based unsupervised learning techniques look for scenarios that are underrepresented in the test set and may pose substantial risks that could result in DNN errors. These techniques can also create fictitious data to enhance the performance of the DNN and pinpoint the most crucial input qualities that cause problems. DNNs may be retrained more successfully utilizing these techniques, which will increase their precision and security in safety-critical systems.

4. RESULTS AND DISCUSSION

Screenshots of the system:

Figure 2: Opening Window

Figure 3: Registration

Figure 4: Start

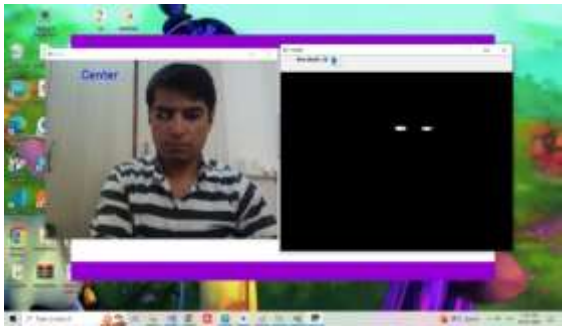

Figure 5: Center

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 06, june 2023, pp : 1408-1411

e-ISSN : 2583-1062

Impact Factor : 5.725

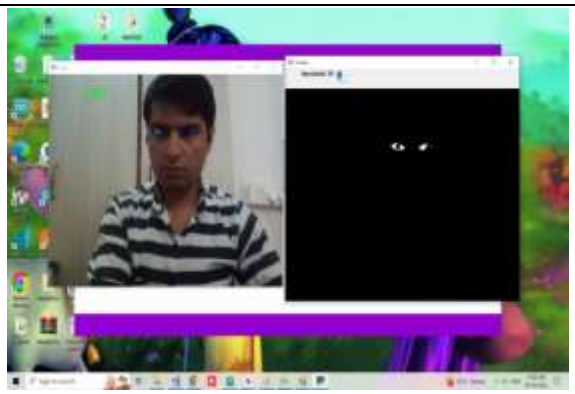

Figure 6: Left

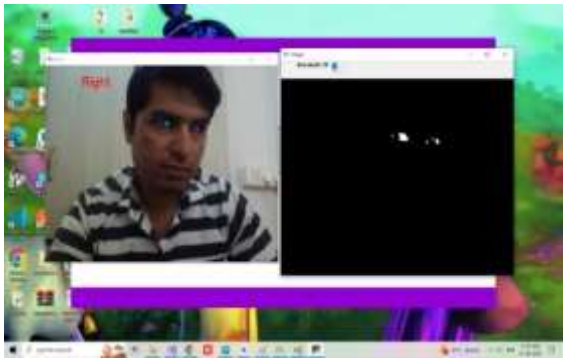

Figure 7: Right

5. CONCLUSION

In conclusion, a viable strategy to increase the dependability and robustness of artificial intelligence systems is to facilitate deep neural network (DNN) safety analysis and retraining utilizing unsupervised learning techniques.

Machine learning models are trained on unlabeled data through the process of "unsupervised learning," which enables them to find patterns and structures in the data without having to be explicitly supervised. This strategy has various advantages for DNN safety analysis and retraining. It's crucial to keep in mind, though, that unsupervised learning might not be enough on its own for thorough DNN safety analysis and retraining. To ensure a comprehensive approach to AI safety, it should be used in conjunction with other strategies and approaches including supervised learning, formal verification, and thorough testing. The reliability, robustness, and security of AI systems can be greatly improved by providing DNN safety analysis and retraining utilizing unsupervised learning approaches. We may strengthen generalization abilities, get deeper understanding of model behavior, and create stronger defenses against adversarial attacks by utilizing the potential of unsupervised learning. Unsupervised learning must, however, be used in conjunction with other methods to ensure a thorough and accurate assessment of the security of AI systems.

6. REFERENCES

[1] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." arXiv preprint arXiv:1412.6572 (2014).

[2] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. "Robustness May Be at Odds with Accuracy." arXiv preprint arXiv:1805.12152 (2018).

[3] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. "Adversarial Examples Are Not Bugs, They Are Features." arXiv preprint arXiv:1905.02175 (2019).

[4] Anish Athalye, Nicholas Carlini, and David Wagner. "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples." arXiv preprint arXiv:1802.00420 (2018).

[5] Youssef Mroueh, Tom Sercu, Aniruddha Kembhavi, and Samuel Rota Bulò. "Unsupervised Learning for Physical Interaction through Video Prediction." arXiv preprint arXiv:1704.06888 (2017).

[6] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. "High-Dimensional Continuous Control Using Generalized Advantage Estimation." arXiv preprint arXiv:1506.02438 (2015).

[7] David Held, Xinyang Geng, Carlos Florensa, and Pieter Abbeel. "Automatic Goal Generation for Reinforcement Learning Agents." arXiv preprint arXiv:1705.06366 (2017).

[8] Saurabh Kumar, Eriq Augustine, Lerrel Pinto, and Abhinav Gupta. "Visual Foresight: Model-Based Deep Reinforcement Learning for Vision-Based Robotic Control." arXiv preprint arXiv:1610.00696 (2016).