Active Monitoring of Neural Networks

Lukina, A.; Schilling, Christian ; Henzinger, Thomas A.

Publication date
2021

Document Version
Final published version

Published in
BNAIC/BeneLearn 2021

Citation (APA)
Lukina, A., Schilling, C., & Henzinger, T. A. (2021). Active Monitoring of Neural Networks. In E. L. A. Leiva, C. Pruski, R. Markovich, A. Najjar, & C. Schommer (Eds.), *BNAIC/BeneLearn 2021: 33rd Benelux Conference on Artificial Intelligence and 30th Belgian-Dutch Conference on Machine Learning* (pp. 685-687)

Active Monitoring of Neural Networks

Anna Lukina[1 [0000−0001−9525−0333]], Christian Schilling[2 [0000−0003−3658−1065]], and Thomas A. Henzinger[3 [0000−0002−2985−7724]]

[1] Delft University of Technology, Delft, The Netherlands
a.lukina@tudelft.nl
[2] University of Konstanz, Konstanz, Germany
christian.schilling@uni-konstanz.de
[3] IST Austria, Klosterneuburg, Austria
thomas.henzinger@ist.ac.at

Abstract. Neural-network classifiers are trained to achieve high prediction accuracy. However, their performance still suffers from frequently appearing inputs of unknown classes. As a component of a cyber-physical system, the classifier in this case can no longer be reliable and is typically retrained. We propose an algorithmic framework for monitoring reliability of a neural network. In contrast to static detection, a monitor wrapped in our framework operates in parallel with the classifier, communicates interpretable labeling queries to the human user, and incrementally adapts to their feedback.

Keywords: monitoring · neural networks · novelty detection.

Automated classification is an essential part of numerous modern technologies and one of the most popular applications of deep neural networks [4]. Neural-network image classifiers have fast-forwarded technological development in many research areas, e.g., automated object localization as a stepping stone to successful real-world robotic applications [9]. Such applications require a high level of reliability from the neural networks.

However, when deployed in the real world, neural networks face a common problem of novel input classes appearing at prediction time, leading to possible misclassifications and system failures. For example, consider a scenario of a neural network used for labeling inputs and making decisions about the next actions for an automated system with limited human supervision: a robot assistant learning to recognize objects in a new home. Assume the neural network is trained well on a dataset containing examples of a finite set of classes. However, after this robot is deployed in the real home, novel classes of objects can appear and confuse the neural network. The inherent misclassifications can stay undetected and accumulate over time, eventually reducing overall accuracy.

The likelihood of severe system damage increases with the frequency and diversity of novel input classes. Typically, this risk is addressed by detecting novel inputs, augmenting the training dataset, and retraining the classifier from scratch [5]. This procedure is not only inefficient, but also leaves the system

vulnerable until such a dataset has been collected. Techniques to incrementally adapt classifiers at prediction time are beneficial for improving accuracy in real-world applications [8, 7]. They, however, do not provide desired interpretability for the human. Approaches to run-time monitoring of neural networks were therefore introduced [6]. In particular, approaches based on abstractions [2, 3, 1, 10] proved to be effective at detecting novel input classes. In addition, they provide transparency of neural-network monitoring.

Crucially, these monitors are constructed offline and remain static at prediction time. Functionalities they are still lacking are distinguishing between "known" and "unknown" novelties and selectively adapting at prediction time.

We propose an active monitoring framework for neural networks that detects novel input classes, obtains the correct labels from a human authority, and adapts the neural network and the monitor to the novel classes, all at prediction time. The framework contains a mechanism for automatic switching between monitoring and adaptation based on run-time statistics. Adaptation consists of either learning new classes (when enough data has been collected) or retraining with more up-to-date information (when the run-time performance is unsatisfactory), where retraining is applied to the network and the monitor independently. A trained neural-network model accompanied by our framework, as an external observer and mediator between the neural network and the human, achieves improved transparency of operation through informative interaction.

Furthermore, we propose a new monitor designed for the adaptive setting. Introducing a quantitative metric at the hidden layers of the neural network, the monitor timely warns about inputs of novel classes and reports its own confidence to the authority. This allows for assessing the need of model adaptation. The quantitative metric allows for easy adaptation at prediction time to newly introduced labels and successfully maintains overall classification accuracy on inputs of known and previously novel classes combined. As such, our framework is an interactive and interpretable tool for informed decision making in neural-network based applications.

Our framework is independent of the choice of the dataset and the neural-network architecture. The only requirements for applicability of our approach are access to the output of the feature layer(s). We plan to extend our procedure toward real-world applications with particular need of active monitoring, e.g., in robotics for the trained controller to gradually adapt to the behavior of the authority. Other interesting directions are time-critical applications where the adaptation of the monitor or the neural network need to be delayed to uncritical phases, and scenarios where novel inputs occur rarely. In addition, the underlying method of our framework can serve as a suitable tool for designing an algorithmic approach to explainability of a neural network's predictions.

References

1. Chen, Y., Cheng, C., Yan, J., Yan, R.: Monitoring object detection abnormalities via data-label and post-algorithm abstractions. CoRR **abs/2103.15456** (2021), https://arxiv.org/abs/2103.15456

2. Cheng, C., Nührenberg, G., Yasuoka, H.: Runtime monitoring neuron activation patterns. In: DATE. pp. 300–303. IEEE (2019), https://doi.org/10.23919/DATE.2019.8714971
3. Henzinger, T.A., Lukina, A., Schilling, C.: Outside the box: Abstraction-based monitoring of neural networks. In: ECAI. Frontiers in Artificial Intelligence and Applications, vol. 325, p. 2433–2440. IOS Press (2020), http://doi.org/10.3233/FAIA200375
4. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing **234**, 11–26 (2017), https://doi.org/10.1016/j.neucom.2016.12.038
5. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113**, 54–71 (2019), https://doi.org/10.1016/j.neunet.2019.01.012
6. Rahman, Q.M., Corke, P., Dayoub, F.: Run-time monitoring of machine learning for robotic perception: A survey of emerging trends. IEEE Access **9**, 20067–20075 (2021), https://doi.org/10.1109/ACCESS.2021.3055015
7. Rebuffi, S., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: Incremental classifier and representation learning. In: CVPR. pp. 5533–5542. IEEE Computer Society (2017), https://doi.org/10.1109/CVPR.2017.587
8. Royer, A., Lampert, C.H.: Classifier adaptation at prediction time. In: CVPR. pp. 1401–1409. IEEE Computer Society (2015), https://doi.org/10.1109/CVPR.2015.7298746
9. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: IROS. pp. 23–30. IEEE (2017), https://doi.org/10.1109/IROS.2017.8202133
10. Wu, C., Falcone, Y., Bensalem, S.: Customizable reference runtime monitoring of neural networks using resolution boxes. CoRR **abs/2104.14435** (2021), https://arxiv.org/abs/2104.14435