A CRITICAL REVIEW ON MACHINE LEARNING APPLICATIONS IN SCIENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

The application of machine learning in sciences has seen exciting advances in recent years. As a widely-applicable technique, anomaly detection has been long studied in the machine learning community. Especially, deep neural nets-based out-of-distribution detection has made great progress for high-dimensional data. Recently, these techniques have been showing their potential in scientific disciplines. We take a critical look at their applicative prospects including data universality, experimental protocols, model robustness, etc. We discuss examples that display transferable practices and domain-specific challenges simultaneously, providing a starting point for establishing a novel interdisciplinary research paradigm in the near future.

1 INTRODUCTION

The advances in the deep learning revolution have been expanding their influence in many domains and accelerating research in a generalized interdisciplinary manner. Neural nets serving as general function approximators have been employed in scientific applications including object identification/classification, anomaly/novelty detection, autonomous control, and neural net-based simulation, etc. Great successes have been made in multiple scientific disciplines. A typical example is AlphaFold (Jumper et al., 2021) for accurate protein structure prediction. At the same time, many progresses have been made in physical sciences (Baldi et al., 2014; Ribli et al., 2019), biology (Ching et al., 2018), molecule generation / drug discovery (Gottipati et al., 2020), medical imaging (Zhang et al., 2020), etc.

Despite the successes, there are many challenges or unrecognized pitfalls in transporting machine learning techniques into more traditional science domains. Not being aware of these possible pitfalls could result in vain efforts and sometimes catastrophic consequences in real-world model deployment. In the following, we take a holistic look at the current collaborative scheme in machine learning applications for sciences in this new cross-disciplinary research era. The differences between general machine learning (mainly focused on computer vision (CV) and natural language processing (NLP)) and tailored scientific applications reside in all parts of the pipeline. The following aspects build an intertwined picture in the modern machine learning-assisted scientific discovery: **Nature of the data**; **Inference process**; **Benchmarks**; **Uncertainty quantification**; **Generalization and Robustness**. Keeping the differences in mind helps shape the research guidelines toward a well-focused and suited technology transfer. Adapting the workflow according to the needs promotes and secures scientific applications and transforms the research paradigm in a universal manner. Finally, the interplay between machine learning and scientific discovery benefits from a communal understanding of the field vocabulary, the publishing traditions, the collaboration schemes, and the academic setups (Wagstaff, 2012). This new regime solicits novel community infrastructures for more impactful research works in the next few years.

2 AN EXAMPLE: OUT-OF-DISTRIBUTION DETECTION

Thanks to the capacity of processing high-dimensional data, deep neural networks-based out-of-distribution (OOD) detection in computer vision and natural language processing has shown great potential and seen much progress in the past few years (Hendrycks & Gimpel, 2017; Vaze et al.,

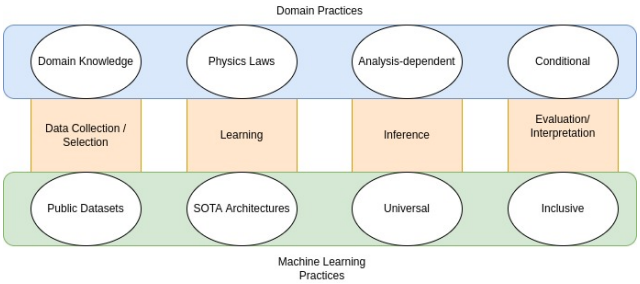

Figure 1: Connection and differences in the pipelines for domain science practices and machine learning practices.

2021; Ahmed & Courville, 2020; Ren et al., 2019; Pang et al., 2021). Models trained on in-distribution (ID) data are expected to "know" what they don't know. Thus they are used to detect unseen patterns as anomalies, by the associated uncertainties or likelihood. On the other hand, OOD detection also serves as a check for model calibration and robustness against dataset shifts.

At the same time, as an important emerging research area, scientific applications of machine learning are seeing great success in advancing the discovery of novel natural phenomena. Anomaly detection techniques have been used to search for (rare) novel particles for high energy physics , galactic activities for astrophysics , and novel molecular mechanisms for molecular physics . Despite these successes, OOD in the machine learning community often follows common workflow and conventions, which might result in unexpected failures in real-world applications. The components of the typical workflow (Fig. 1) reflect fundamental differences and focuses in these two streams of research. The motives and protocols share some common aspects, yet display essential differences and focuses. Being aware of these differences in research protocols facilitates effective application and forges innovative and impactful collaboration. We discuss these aspects respectively in the following.

2.1 Datasets and Input Representation

The first and foremost element in the pipeline is the data under consideration. Computer vision and natural language processing, as the main application areas of machine learning, have restricted input formats. In contrast, data in sciences come in diverse formats, including raw device response records in particle detectors, light spectra received by telescopes, X-ray images, etc.

These differences motivate intriguing research topics specifically for scientific applications. On the one hand, the input formats of scientific data request novel tailored neural architectures that are aware of the underlying inductive biases. For instance, many scientific data have inherent symmetries or are constrained by physics laws or associated invariances. In the same vein, equivariant neural networks (Cohen & Welling, 2016) and geometric deep learning (Bronstein et al., 2021) have pioneered neural architectures and mechanisms for incorporating the concepts of symmetries. They have been gaining attention in particle physics (Bogatskiy et al., 2020), computational chemistry (Anderson et al., 2019; Batzner et al., 2022), etc. At the same time, these approaches have boosted research in the usual machine learning applications (computer vision, natural language processing, etc.). Similar motives can broaden the research horizon of "classical" machine learning, forge multidisciplinary collaborations, and facilitate innovative research advances. On the other hand, attention-based transformers (Vaswani et al., 2017) have reigned the NLP community and more recently intruded computer vision (Dosovitskiy et al., 2020). They have been successful in processing multiple data formats (Reed et al., 2022). These achievements indicate potential directions for cross-disciplinary research protocols with a uniform framework and reusable and shareable modules that can be interfaced with scientific tasks.

Regarding the data collection process, while image labeling is expensive for machine learning researchers, simulation data with labels in sciences are relatively cheaper to generate (although it could take much computing resources in some cases). That means we can have as much data as we need. As a practical result, some strategies (e.g., those dedicated to small datasets) are not necessarily

appropriate for scientific applications. On the other hand, models trained on simulated data are confronted with performance degradation-associated distribution shifts when applied to real data. This will be further discussed in section 2.4.

2.2 Inference, Decision, and Evaluation

In most scientific applications, model inference [1] is closely related to the underlying hypotheses to be tested and thus the evaluation metrics should depend on the chosen test sets which are supposed to resemble the real-world deployment circumstances. In other words, the components of the test sets define the inference strategies. At the same time, the test sets are determined by the science under investigation.

The well-adopted practice in the ML community is testing the model on all the classes for a benchmark dataset and reporting class-inclusive metrics (Hendrycks & Gimpel, 2017). In most of the research literature, the model evaluation metrics are reported for classifying two large benchmark datasets separately collected (e.g., models trained on CIFAR are tested against other datasets such as SVHN, or even tested on uniform noise as the OOD). This convention is, of course, convenient and effective for a preliminary assessment.

However, in scientific applications, depending on the context, due to the common imbalance of in-distribution classes, it is more suitable to test against class-conditional in-distribution data. A generic example is that background events have different class weights, sometimes in extreme imbalance. If the phenomenon of interest is detected under the circumstance of a specific set of background events, we will need to adjust our evaluation metrics to be calculated for the distribution of focus.

Test datasets One problem of current scientific applications is the lack of consensus on benchmark test datasets. Researchers might work under different selected subsets, different data qualities, and different simulation settings. Even in more focused sub-fields, researchers usually report model metrics on their own datasets. This brings variations and concerns in model comparison. From another practical point of view, robustness examination and stress tests also benefit from a "complete" set of test sets. Research schemes might bias toward specific scenarios if they are only tested on a limited set of data.

Evaluation metrics For anomaly detection, a widely-reported metric is the Area Under the ROC curve (AUC) score based on the binary ID/OOD classification. OOD classes are usually from other datasets different from the training datasets. The selection of ID and OOD datasets is rotating around common benchmarks MNIST/Fasion-MNIST or CIFAR-10/CIFAR-100/ImageNet/SVHN. However, this kind of inclusive measure doesn't give much information on context-sensitive use cases. Meanwhile, different datasets have intrinsic dataset shifts which affect the validity of this approach. Though there are recent proposals (Ahmed & Courville, 2020) on transforming the research scheme to a more specific within-benchmark evaluation, we haven't seen large-scale changes in the ML community.

2.3 Model Examination and Failure Analysis

Failure mode analysis can serve as a portal for understanding the methods proposed. It's not a "bonus" compared with the AUC. Rather, it helps reveal crucial aspects of the model learning mechanisms and provides insights into the robustness under variable deployment circumstances. And these inspections often lead to further research ideas and model improvement directions. For instance, OOD detection in computer vision might fail because of the pixel correlation (Ren et al., 2019). This inspection leads to the remedies proposed there.

On the other hand, scientific data may have complex nature that makes model examination difficult compared with the cases in CV or NLP. Usually, they come in formats not "human-readable". To interpret model performance and examine model failures would require extra interfaces or tools. Sometimes, involving theoretical and expert-designed features becomes necessary.

[1] We follow the definition in deep learning for the terminology "inference", which mostly refers to the decision-making process at test time.

2.4 GENERALIZATION AND ROBUSTNESS

Building generalizable and robust models is the goal of every machine learning practitioner. There are different understandings of what "robustness" (Basart, 2021) stands for, especially across different domains. In the ML community, adversarial attacks (Goodfellow et al., 2015) caused by perturbations in the input space are under frequent scrutiny. Models robust to malicious attacks have been developed and investigated. In scientific applications, we may have a different definition and focus regarding model "robustness". In anomalous signal detection, we would like the model to 1) be effective across an extensive range of unseen signals, 2) have less distortion under the simulation-to-data shift, and 3) be invariant to spurious correlations.

Uncertainty Quantification and Model Calibration In counting experiments, which is the main probing method in particle physics, statistical and systematic uncertainties (Barlow, 1989; Cranmer, 2014; Demortier, 2008) are the common uncertainties we consider for interpreting the results. Inserting neural nets into the pipeline could result in complicated uncertainty quantification. Calibration of the outputs and estimation of the effects of nuisance parameters need extra effort. Meanwhile, the terminologies are different in the machine learning community. Uncertainties in deep neural models are usually categorized into *epistemic uncertainty* (model uncertainty) and *aleatoric uncertainty* (data uncertainty) (Gal, 2016). At first sight, the *aleatoric uncertainty* seems identical to the *statistical uncertainty*, while the *epistemic uncertainty* shares a similar meaning with the *systematic uncertainty*.

Quantifying what the neural networks don't know, or the uncertainty of the outputs, is important to ensure we have a trustworthy deployment of the models (Ovadia et al., 2019; Malinin & Gales, 2018). Calibration is one method to quantify the quality of model uncertainty. (A calibrated model should have a decent alignment between the model confidence and the actual likelihood.) Especially for OoD detection, out-of-distribution uncertainty ("know what is unknown by the model") is also an important indicator. One typical method which has a long history is ensemble models (Lakshminarayanan et al., 2017). At the same time, in the framework of Bayesian Neural Networks, the Monte Carlo drop-out (Gal & Ghahramani, 2016) technique can serve as a surrogate for estimating model uncertainty.

Meanwhile, taking epistemic uncertainty into account can increase model robustness and improve OOD detection performance (Lakshminarayanan et al., 2017). Correspondingly in scientific applications, systematic uncertainty has been associated with nuisance parameters (Dorigo & De Castro Manzano, 2020; d'Agnolo et al., 2022; Ghosh et al., 2021). And incorporating this uncertainty in the training strategy can result in more robust and powerful models.

Distribution Shift Due to the large dimensionality under consideration in deep learning models, the effects under distribution shift are hard to quantify. Robustness under distribution shift can be realized through adversarial training strategies (Ganin et al., 2016; Li et al., 2018). Other approaches (Magliacane et al., 2018) leverage domain invariances to achieve robustness. Similar approaches have been taken in High Energy Physics (Louppe et al., 2017), though used in another context.

In the same vein, for many scientific domains, simulation plays an important role in modeling the background. Especially for supervised learning, data with label information are mostly dependent on the simulator. Models trained on simulation data will have degraded performance when directly applied to real data. Though there are workarounds for background estimation, achieving domain invariance brings more opportunities and makes even complex search strategies possible.

3 SUMMARY

The coming decades call for a novel collaboration scheme of machine learning and sciences, which will foster great opportunities for machine learning researchers and domain scientists. By clarifying the current research protocols, identifying obstacles and pitfalls, discussing possible solutions, and conceiving future directions, we can advance this career further. In order to make the discussion more concrete, we utilize anomaly detection – a popular topic in machine learning and an emerging research direction in sciences – to take a detailed look at the current gaps. We discussed different components in the pipeline, and analyzed the communal practices and the differences in the machine

learning community and the science. We also discussed opportunities and challenges, and proposed potential solutions for the new generation's collaborative innovation in this new era.

REFERENCES

Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3154–3162, Apr. 2020. doi: 10.1609/aaai.v34i04.5712. URL https://ojs.aaai.org/index.php/AAAI/article/view/5712.

Brandon M. Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. In *NeurIPS*, 2019.

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for Exotic Particles in High-Energy Physics with Deep Learning. *Nature Commun.*, 5:4308, 2014. doi: 10.1038/ncomms5308.

Roger Barlow. *Statistics. A guide to the use of statistical methods in the physical sciences*. 1989.

Steven Basart. Towards Robustness of Neural Networks. *arXiv e-prints*, art. arXiv:2112.15188, December 2021.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13, 2022.

Alexander Bogatskiy, Brandon M. Anderson, Jan T. Offermann, Marwah Roussi, David W. Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics. *ArXiv*, abs/2006.04780, 2020.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velivckovi'c. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *ArXiv*, abs/2104.13478, 2021.

Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, 2016.

Kyle Cranmer. Practical Statistics for the LHC. In *2011 European School of High-Energy Physics*, pp. 267–308, 2014. doi: 10.5170/CERN-2014-003.267.

Raffaele Tito d'Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Learning new physics from an imperfect machine. *Eur. Phys. J. C*, 82(3):275, 2022. doi: 10.1140/epjc/s10052-022-10226-y.

Luv Demortier. P Values and Nuisance Parameters. 2008. doi: 10.5170/CERN-2008-001.23. URL https://cds.cern.ch/record/1099967.

T. Dorigo and P. De Castro Manzano. Dealing with Nuisance Parameters using Machine Learning in High Energy Physics: a Review. 7 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yarin Gal. Uncertainty in deep learning. 2016.

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

Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. In *J. Mach. Learn. Res.*, 2016.

Aishik Ghosh, Benjamin Nachman, and Daniel Whiteson. Uncertainty-aware machine learning for high energy physics. *Phys. Rev. D*, 104(5):056026, 2021. doi: 10.1103/PhysRevD.104.056026.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Karam M. J. Thomas, Simon Blackburn, Connor W. Coley, Jian Tang, Sarath Chandar, and Yoshua Bengio. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *ICML*, 2020.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ArXiv*, abs/1610.02136, 2017.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A. Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI*, 2018.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In *NIPS*, 2017.

Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

Andrey Malinin and Mark John Francis Gales. Predictive uncertainty estimation via prior networks. In *NeurIPS*, 2018.

Yaniv Ovadia, Emily Fertig, J. Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *ArXiv*, abs/1906.02530, 2019.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection. *ACM Computing Surveys (CSUR)*, 54:1 – 38, 2021.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent, 2022.

J. Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *ArXiv*, abs/1906.02845, 2019.

Dezső Ribli, Bálint Ármin Pataki, and István Csabai. An improved cosmological parameter inference scheme motivated by deep learning. *Nature Astron.*, 3(1):93–98, 2019. doi: 10.1038/s41550-018-0596-8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *CoRR*, abs/2110.06207, 2021. URL https://arxiv.org/abs/2110.06207.

Kiri Wagstaff. Machine learning that matters. *CoRR*, abs/1206.4656, 2012. URL http://arxiv.org/abs/1206.4656.

Jianpeng Zhang, Yutong Xie, Yi Li, Chunhua Shen, and Yong Xia. Covid-19 screening on chest x-ray images using deep learning based anomaly detection. *ArXiv*, abs/2003.12338, 2020.