Towards Human-centric Evaluation of Machine Learning Models

Yara Rizk, David Piorkowski, Vatche Isahagian, Vinod Muthusamy
IBM Research

ABSTRACT

Evaluating machine learning (ML) models is an established standard in the research community. Metrics, benchmarks and evaluation setups are adopted to determine whether one approach is "better" than another. However, many of these evaluation metrics do not always accurately represent how well the model will actually perform when deployed in the real world within a complex software system. In this talk, we discuss the need for more human-centric evaluation metrics that assess the performance of ML models in context of the users' tasks; i.e., can we measure how well people are getting their work done (with and without ML models).

ACM Reference Format:
Yara Rizk, David Piorkowski, Vatche Isahagian, Vinod Muthusamy. 2022. Towards Human-centric Evaluation of Machine Learning Models. In *Proceedings of The Workshop on Applied Machine Learning Management at The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD WAMLM '22).* ACM, New York, NY, USA, 1 page. https://doi.org/XXXXXXX.XXXXXXX

1 TRADITIONAL ML EVALUATION METRICS

ML models have been adopted in many real-world products with varying degrees of success; for example, autonomous robots have used deep learning models to perform scene understanding (e.g., pedestrian classification and traffic sign recognition).

They are generally evaluated using publicly available benchmarks using setups like cross-validation where a subset of the data is hidden from the model and reserved for testing. Performance metrics measure error rates (e.g., classification accuracy, F1-score, false positives, mean squared error, ROC curves, etc.), computational efficiency (e.g., training/prediction time on CPUs vs. GPUs vs. TPUs), number of training samples, and model size (e.g., number of hyper-parameters in a neural network).

For some applications, ML state-of-the-art performance did not immediately translate to good "real-world" performance [1]. For others, the models did not achieve wide adoption despite performing well on these traditional metrics [2].

2 HUMAN-CENTRIC MINDSET TO ML EVALUATION METRICS

When ML is part of an end-user facing system, measures of performance should extend to include how well that system is helping end users accomplish their tasks. Consider a conversational agent that helps business users automate some of their tasks. Traditional performance measures would focus on the underlying ML components such as the intent classifier as the main source of improvement. Such a tool's performance would more appropriately be measured by how effectively it helps a user complete their tasks. Improvements to the intent classifier may *incidentally* help the end user, but without evaluating if the improvement translates to an end user's ability to complete tasks it is unclear if the overall tool's performance has improved. In the worst case, it may be possible to improve the underlying ML, but decrease the end user's ability to do their tasks. Extending this idea of performance further, measures of how effectively end users onboard a tool or recover from errors can likewise be part of the larger performance evaluation.

As human-AI collaboration solutions become more commonplace, performance measures should consider both the perspectives of the AI and the human when evaluating "goodness" of the overall collaboration. Yet, determining what the appropriate measures are and how to aggregate the ML- and human-centric measures together remain an open research question. Future work must investigate various collaborations and determine possible generalizations.

REFERENCES

[1] Thilo Stadelmann et al. 2018. Deep learning in the wild. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, 17–38.
[2] Julian Varghese. 2020. Artificial intelligence in medicine: Chances and challenges for wide clinical adoption. *Visceral medicine* 36, 6 (2020), 443–449.

A APPENDIX

Relevance. Related to making a successful product from ML tech and measuring its value (by focusing on the end user instead of the model's accuracy or its reliance on state-of-the-art approaches).

Discussion Points. Metrics that can better evaluate ML-based products; shortcomings of existing metrics.

Company and Project Portrait. IBM Research is an industrial research lab developing advancements in AI, quantum and many other fields. The talk draws on the authors' experience working on the Watson Orchestrate product that relies on ML models.

Main Presenter Bio. Yara Rizk is a researcher at IBM Research. She received her doctorate in Electrical and Computer Engineering from the American University of Beirut (AUB) in 2018. Prior, she obtained a Bachelor of Engineering in Computer and Communication Engineering from AUB, Lebanon, in 2012. Her research interests span, artificial intelligence, machine learning, multi-agent systems, business automation, and robotic process automation. Her work has led to multiple peer-reviewed publications in leading academic journals and international conferences.

KDD WAMLM '22, Aug. 14–18, 2022, Washington D.C.

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX