SAND2021-1560R

Breakouts 4 & 5: Robustness and Validation of Model and Digital Twins Deployment

Chairs: David Stracuzzi, Jenifer Shafer

Co-chairs: Svitlana Volkova, Jaideep Ray

Participants:

Matt Barone	Sandia National Laboratory
Sharlotte Kramer	Sandia National Laboratory
Dan Ratner	SLAC National Accelerator Laboratory
Maria Glenski	Pacific Northwest National Laboratory
Andy Huang	Sandia National Laboratory
Jaideep Ray	Sandia National Laboratory
Laura Pullum	Oak Ridge National Laboratory

Introduction:

For digital twins (DTs) to become a central fixture in mission critical systems, a better understanding is required of potential modes of failure, quantification of uncertainty, and the ability to explain a model's behavior. These aspects are particularly important as the performance of a digital twin will evolve during model development and deployment for real-world operations. The requirements for safety-critical DT systems, robustness, accountability, transparency and fairness, are defined below:

- **Robust** DTs must be resilient to variations in data inputs.
- **Accountable** DTs must demonstrate reliability when applied in key circumstances and the ability to review historical predictions and inferences (closely related to transparency).
- **Fair** DTs must be equitable across representative subsets e.g., across subpopulations of users impacted by DT outputs.
- **Transparent** DTs must enable high quality and correct interpretation of model behavior to identify points of failure through data inputs and model predictions. This can be accomplished either by interactive explanations of model behavior or by quantifying the predictive performance of the model. An example of DTs could include a combination of uncertainty quantification (UQ), out-of-distribution analysis, and traditional performance analysis (cross-validation, etc). Below we define requirements for safety-critical digital twin systems

Also required is a rigorous approach to DT validation and verification (V&V) for three reasons: (1) to provide a basis for trust in digital twin adoption, (2) to reduce risk of digital twin backsliding, and (3) to decrease likelihood of accidents with a digital twin. DT V&V is challenging because testing standards depend on the area of application and the specifics of machine learning approach used in the development of the DT.

Guiding questions:

- How do we define robustness?
- How do we define validation?
- How are current definitions of robustness and validation insufficient, and how could they be improved?
- What does it mean to do V&V on a digital twin model when some parts of the model are represented by ML (statistical) models while other parts are represented by mathematical (theoretical) models?
- Given that ML models extrapolate statistically rather than via an explicit encoding of underlying physics, what are the implications for DT model development, verification, and validation?
- What terms and methods that are typically well-defined with respect to V&V for (equation-based) physics models need to be changed?
- How does the problem change when we talk about DTs that are updated over time (i.e. DT models that are incrementally revised as we collect data from real systems)?
- What are the near-term research needs to develop appropriate V&V methods?

Key challenges:

- Many fundamental questions about DT V&V and robustness still remain open. For example, broadly accepted definitions of verification, validation, and robustness for DTs as well as for ML do not exist. Definition of V&V (safety, fairness, interpretability, etc) varies between communities and provides constraints on solution space.
- V&V is a multidimensional process that needs to account for robustness (including extrapolation), fairness, interpretability, safety, and so on.
- Currently, it is not clear what metrics are important for DT V&V. Moreover, quantitative metrics might be insufficient; they have to be paired with qualitative evaluation. What is testing cadence and stopping criteria for DT V&V?

State of the art:

- Focus of DOE effort is on DT for high-consequence (aka safety-critical) systems -- without effective V&V the adoption will not happen.
- There is extensive literature in classical V&V, UQ, etc. There is growing literature in ML evaluation including recent work in adversarial attacks, input perturbation, UQ for some ML models, OOD, training optimization, etc. as described in detail below.

  Related work that extends traditional evaluation of ML models to address the issues of robustness, accountability, fairness, or transparency of model performance consider these issues independently as summarized in Table 2. DT evaluation will require the integration of multiple packages or standalone tools with distinct requirements, environments, or interfaces in order to perform a multi-faceted evaluation with tools often facing similar limitations. For example, existing tools that address model robustness are largely focused on adversarial attacks (ART, Advertorch, Foolbox, Advbox, OpenAttack, TextFooler) or gradient attacks (Foolbox). Others are also tied to a specific architecture framework, such as Pytorch (Advertorch), which limits the flexibility when applied to previously or independently developed models.

Table 2. Summary of existing tools for ML evaluation, highlighting related work across each of the four dimensions of interest.

Robustness	Accountability	Fairness	Transparency
ART (Nicolae, et al., 2018) **Advertorch** (Ding, Wang, & Jin, 2019) **Foolbox** (Rauber, Zimmermann, Bethge, & Brendel, 2020) **Advbox** (Goodman, et al., 2020) **OpenAttack** (Zeng, et al., 2020) **TEAPOT** (Michel, Neubig, Li, & Pino, 2019) **TextAttack** (Morris, et al., 2020) **TextFooler** (Jin, Jin, Zhou, & Szolovits, 2019)	**GLUE** (Wang, et al., 2018) **SuperGLUE** (Wang, et al., 2019) **SQuAD 2.0** (Rajpurkar, Jia, & Liang, 2018) **ROAR** (Hooker, Erhan, Kindermans, & Kim, 2019)	**Aequitas** (Saleiro, 2018) **AI Fairness 360** (Bellamy, et al., 2019) **Fairlearn** (Bird, et al., 2020) **Fairness Indicators** [1] **FairSight** (Ahn & Lin, 2019) **ML-fairness-gym** [2] **Scikit-Fairness** [3]	**interpretML** (Nori, Jenkins, Koch, & Caruana, 2019) **SHAP** (Lundberg & Lee, 2017) **Captum** (Kokhlikyan, et al., 2020) **LIT** (Tenney, et al., 2020) **TreeExplainer** (Lundberg, et al., 2019)

[1] https://github.com/tensorflow/fairness-indicators

[2] https://github.com/google/ml-fairness-gym

[3] https://github.com/koaning/scikit-fairness

Roadblocks: coordinated focus of effort and funding for fundamental research.

Research opportunities:

- Develop broadly accepted, operational definitions of DT robustness, verification, validation, fairness, etc. The community has to converge on fundamental definitions for validation, verification, robustness, fairness in context of DT in addition to specific applications and users. For example, in robust DT, outputs do not change drastically if the inputs are changed a little. DTs fail gracefully if fed non-physical inputs. Validation has no change in goal as the DT still needs to reproduce physics, though the associated techniques may change. Verification may entail decoupling of the DT into the PDE-part and ML-part. Test the PDE- part as we do today - compare with exact solutions, discretization error convergence etc. The ML-part could be attacked with novel methods that go beyond LIME and other one-dimensional evaluation approaches presented in the related work.

- Develop qualitative approaches and quantitative metrics for evaluating DT robustness, fairness, interpretability, and so on. For example, robustness can be tackled with sensitivity analysis, and validation with forward UQ nominally, but may need to solve Bayesian inverse problems to infer input uncertainties from other experiments. For verification, the ML piece needs to have gone through at least cross-validation, which is standard practice, but it does not speak to predictive uncertainty, extrapolation ability, or safety in the field. If the ML piece is a representation of training data from high-fidelity simulations, explanations need to be generated regarding the ML model e.g., via Deep Taylor Series or Generalized Additive Models [collectively called Locally Interpretable

Model-agnostic Explanations] and make sure it behaves something like the high-fidelity simulations. This appears to be a brand new topic with no publications in this area.

- Fundamental research in UQ mathematics is required for support DT V&V. Major research questions include but are not limited to:
    - What requirements are placed on UQ for ML?
    - What constitutes a sufficient UQ evaluation of an ML model or an ensemble? (There are many sources of uncertainty in an ML model; a complete evaluation is at least as intractable as it is for equational models).
    - How do we combine/propagate UQ for ML components with physics equations? (Forward UQ nominally, but may need to solve Bayesian inverse problems to infer input uncertainties from other experiments).

- Fundamental research in formal methods for evaluating critical model parts and scaling these formal methods is needed.

- Redesign of "real world" systems to facilitate ML V&V and model development (e.g. data instrumentation).

Impact if research opportunities are addressed?

- What new scientific capabilities will be enabled?
    - Data-driven insights into physics/engineering theory.
    - Precise analysis/quantification of model applicability.
    - Model-driven automation of critical applications.
- New methods and techniques will allow to perform:
    - Certification, validation, robustness, etc. methods and evaluation techniques
    - Data-driven models that admit a physics-based chain of reasoning for SciML models.
    - Improve the accuracy of scientific models without a significant increase in computational cost.

Summary:

Targeted continuous evaluation of DT is required to estimate the tradeoff between the risks and benefits of deploying DT models. In order to achieve that we need to admit that our current ML evaluation process is broken[1], and since the ML component is a critical part of DT, fundamentally novel approaches for evaluating mission critical DT systems are needed.

The ubiquitous method of multi-fold train and validate followed by evaluation on a final holdout test set assumes the data is sampled from a distribution representing the data that the model will see after deployment during the operational phase. For a number of reasons, this assumption often does not hold. The objective of the model developer is often to evaluate the

[1]https://www.technologyreview.com/2020/11/18/1012234/training-machine-learning-broken-real-world-heath-nlp-computer-vision/

model performance, optimize hyperparameters, and give the developer confidence in the model's performance. The focus is not on *identifying as many modes of model failure as possible and determining how best to correct or mitigate them*, but it should be.

Supporting Material

1. Ahn, Y., & Lin, Y.-R. (2019). Fairsight: Visual analytics for fairness in decision making. *IEEE transactions on visualization and computer graphics.* IEEE.
2. Arendt, D. L. (2020). Parallel embeddings: a visualization technique for contrasting learned representations. *Proceedings of the 25th International Conference on Intelligent User Interfaces.*
3. Arendt, D., Huang, Z., Shrestha, P., Ayton, E., Glenski, M., & Volkova, S. (2020). CrossCheck: Rapid, Reproducible, and Interpretable Model Evaluation. *arXiv preprint.*
4. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., . . . others. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development.*
5. Bird, S., Dud{\'\i}k, M., Edgar, R., Horn, B., Lutz, R., Milan, V., . . . Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Technical Report MSR-TR-2020-32.* Microsoft.
6. Chouldechova, A. R. (2020). *A snapshot of the frontiers of fairness in machine learning.* Communications of the ACM, 63(5).
7. Ding, G. W., Wang, L., & Jin, X. (2019). {AdverTorch} v0.1: An Adversarial Robustness Toolbox based on PyTorch. *arXiv preprint arXiv:1902.07623.*
8. Goodman, D., Xin, H., Yang, W., Yuesheng, W., Junfeng, X., & Huan, Z. (2020). Advbox: a toolbox to generate adversarial examples that fool neural networks. *arXiv preprint.*
9. Haibe-Kains, B. A. (2020). Transparency and reproducibility in artificial intelligence. *Nature 586,* , E14–E16.
10. Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems.*
11. Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2019). Is BERT Really Robust? Natural Language Attack on Text Classification and Entailment. *arXiv preprint.*
12. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., . . . others. (2020). Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint.*
13. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30.*
14. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., . . . Lee, S.-I. (2019). Explainable AI for Trees: From Local Explanations to Global Understanding. *arXiv preprint.*
15. Mehrabi, N. M. (2019). *A survey on bias and fairness in machine learning.* arXiv preprint arXiv:1908.09635.
16. Michel, P., Neubig, G., Li, X., & Pino, J. M. (2019). On Evaluation of Adversarial Perturbations for Sequence-to-Sequence Models. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.*
17. Michel, P., Neubig, G., Li, X., & Pino, J. M. (2019). On Evaluation of Adversarial Perturbations for Sequence-to-Sequence Models. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).*
18. Modas, A. S.-M.-D. (2019). *Sparsefool: a few pixels make a big difference.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
19. Molnar, C. (2019). *Interpretable Machine Learning.*
20. Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. *arXiv preprint.*

21. Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., . . . others. (2018). Adversarial Robustness Toolbox v1. 0.0. *arXiv preprint.*
22. Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint.*
23. Pruthi, D. a. (2019). Combating Adversarial Misspellings with Robust Word Recognition. *The 57th Annual Meeting of the Association for Computational Linguistics (ACL).*
24. Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).*
25. Rauber, J., Zimmermann, R., Bethge, M., & Brendel, W. (2020). Foolbox Native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX. *Journal of Open Source Software.*
26. Saldanha, E., Praggastis, B., Billow, T., & Arendt, D. (2019). ReLVis: Visual Analytics for Situational Awareness During Reinforcement Learning Experimentation. *EuroVis (Short Papers).*
27. Saleiro, P. a. (2018). Aequitas: A Bias and Fairness Audit Toolkit. *arXiv preprint.*
28. Silva, S. H. (2020). *Opportunities and challenges in deep learning adversarial robustness: A survey.* arXiv preprint arXiv:2007.00753.
29. Stracuzzi, D.J., Chen, M.G., Darling, M.C., Peterson, M.G., & Vollmer, C. (2017). *Uncertainty quantification for machine learning.* SAND2017-6776., Sandia National Laboratories.
30. Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., . . . Yuan, A. (2020). The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for {NLP} Models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics.
31. Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., . . . Yuan, A. (2020). The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for {NLP} Models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Association for Computational Linguistics.
32. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., . . . Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems.*
33. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.*
34. Yang, F. a. (2020). How do visual explanations foster end users' appropriate trust in machine learning? *Proceedings of the 25th International Conference on Intelligent User Interfaces.*
35. Zeng, G., Qi, F., Zhou, Q., Zhang, T., Hou, B., Zang, Y., . . . Sun, M. (2020). OpenAttack: An Open-source Textual Adversarial Attack Toolkit. *arXiv preprint.*

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under Contract DE-NA0003525.