BNL-203606-2018-FORE

Scoping Study of Machine Learning Techniques for Visualization and Analysis of Multi-source Data in Nuclear Safeguards

Y. Cui,

May 2018

Nonproliferation and National Security Department
Brookhaven National Laboratory

U.S. Department of Energy
USDOE National Nuclear Security Administration (NNSA), Office of Nonproliferation and Verification Research and Development (NA-22)

DISCLAIMER

Scoping Study of Machine Learning Techniques for Visualization and Analysis of Multi-source Data in Nuclear Safeguards

U.S. DEPARTMENT OF ENERGY

NNSA

OFFICE OF NONPROLIFERATION AND ARMS CONTROL (NPAC)

INTERNATIONAL NUCLEAR SAFEGUARDS

Access values of machine learning techniques in processing of multi-source data in nuclear safeguards

Background/State of the Art

- Nuclear safeguards deploy many different technologies: radiation detection, video surveillance, seals, various NDA measurements, inspection information, satellite imagers, etc.
- Each technology measures a specific property of nuclear materials or activity related to operation of nuclear facility.
- Analysis of multi-source data in an integrated way could improve the effectiveness and efficiency of nuclear safeguards.
- Machine learning is a cutting-edge technology in multi-source data analysis that has been proven in social media and other scientific research, and has great potentials in nuclear safeguards.

Innovation

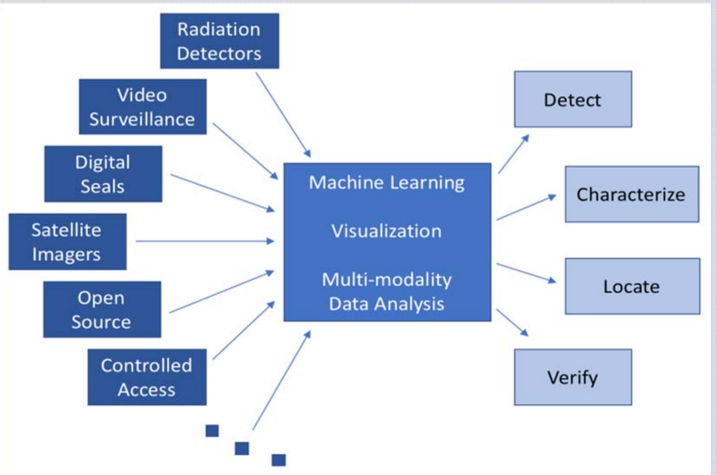
- We will survey the machine learning techniques, specifically in visualization and analysis of multi-source data and assess their values to nuclear safeguards.
- This work will provide full pictures of visualization and analysis techniques for multi-source data in nuclear safeguards, which hasn't been done yet.
- This work will also provide a pathway of technology development in terms of machine learning for multi-source data in nuclear safeguards.

R&R #: BNL-203606-2018

Approach, Metrics and Outcomes

MAIN GOAL

- A survey of machine learning techniques for visualization and analysis of multi-source data
- A guidance for adapting and developing machine learning techniques for nuclear safeguards



HOW IT WORKS

- Work with NA-241 and the IAEA to examine data sets from each modality.
- Examine all the techniques that we have studied and/or developed in the past multi-source data projects.
- Conduct literature search to study the efforts and achievements done by other research teams.
- Develop a technical report on the survey results and provide guidance for technology development.

Impact

- More effective and efficient nuclear safeguards
  - Detection, characterization, location of nuclear materials
  - Characterization and verification of operation of nuclear facilitates
- STR 385, T.5.R1
- STR 386, SGIM-008
- **Start of FY TRL = 2**
- **End of FY TRL (Planned) = 2**

Goals/Action Plan

Current FY

- Survey of machine learning techniques for multi-source data
- A comprehensive report of this study

Future FY

- We will propose to develop machine learning techniques dedicated to multi-source data analysis in nuclear safeguards

Team

BNL (Nonproliferation and National Security Department and Computational Science Initiatives)

Principal Investigator: Yonggang Cui, ycui@bnl.gov, (631)344-5351

Scoping Study of Machine Learning Techniques for Visualization and Analysis of Multi-source Data in Nuclear Safeguards

Project Categories

24.1.3.2 Containment and Surveillance
☐ or 2.3 Other C&S
☒
☐ or ☐

PROPOSED STATEMENT OF WORK

Abstract

In implementation of nuclear safeguards, many different techniques are being used to monitor operation of nuclear facilities and safeguard nuclear materials, ranging from radiation detectors, flow monitors, video surveillance, satellite imagers, digital seals to open source search and reports of onsite inspections/verifications. Each technique measures one or more unique properties related to nuclear materials or operation processes. Because these data sets have no or loose correlations, it could be beneficial to analyze the data sets together to improve the effectiveness and efficiency of safeguards processes. Advanced visualization techniques and machine-learning based multi-modality analysis could be effective tools in such integrated analysis. In this project, we will conduct a survey of existing visualization and analysis techniques for multi-source data and assess their potential values in nuclear safeguards.

Mission Relevance

It is evident that the IAEA has showed interest in artificial intelligence/machine learning technology in applications of nuclear safeguards. The interests are identified in the following areas in the IAEA Research and Development Plan and the IAEA Development and Implementation Support Program:

STR 385, T.5.R1 – Identify, evaluate and test promising applications of robotics and machine learning/artificial intelligence to improve the effectiveness and efficiency of safeguards.

STR 386, SGIM-008 – Surveillance Techniques. Outcome 9.) Investigated feasibility of intelligent systems for analysing non-quantitative data, eliciting analyst conclusions, and aggregating analyst conclusion across multiple disparate data sources in order to assist analysts in drawing broad State-level conclusions with a measured degree of confidence.

Scope of Work

1. **Overview of proposed project including a detailed description of the technology or process to be developed.**

   Artificial intelligence/machine learning has been applied to many fields ranging from social media, consumer electronics, medical health to cybersecurity and scientific research, and has shown its unique strength in processing large scale data sets from a variety of sources. Nuclear safeguards is one of the applications that can potentially benefit from this technology.

In implementation of nuclear safeguards, many different techniques are being used to monitor operation of nuclear facilities and safeguard nuclear materials, ranging from radiation detectors, flow monitors, video surveillance, satellite imagers, and digital seals to open source search and reports of onsite inspections/verifications. Each technique measures one or more specific properties related to nuclear materials or operation processes. Our experience in multi-modality data analysis indicates that, by integrating data sets that are not or loosely correlated, machine learning methods can give much better performance (or accuracy) over traditional data analysis methods. We believe that applying artificial intelligence/machine learning technology to the multi-source data collected in nuclear safeguards, we can improve the effectiveness and efficiency of safeguards process.

In this project, we will conduct a survey of machine learning techniques and identify potential techniques with great values for nuclear safeguards applications. Specifically, we will focus on two areas: advanced visualization techniques that can help visualize and interpret data and multi-modality analysis techniques that can find the hidden features underneath the integrated multi-source data sets.

Previously BNL has been funded to conduct research in both areas and has applied our techniques to other fields of scientific research. A high-level summary of our relevant work is reported below.

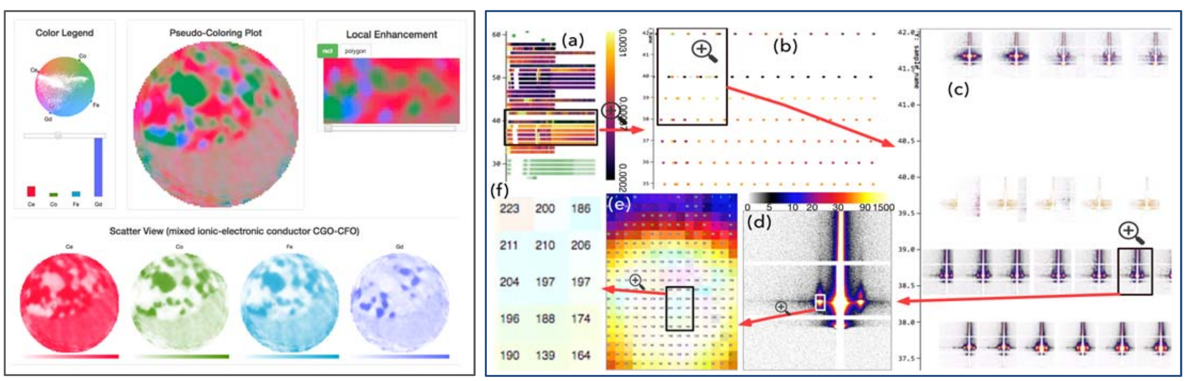


Figure 1: Two visualization methods: (left) ColorMapND to map multivariate data to color, (right) MultiSciView to visualize images and their attributes in level-of-details.

Visualization techniques have been known to excel at knowledge communication and insight discovery by using encodings to transform abstract data into meaningful representations [1]. It can be adopted to visualize and evaluate the raw and analyzed data [2], but also be integrated with complex analysis models for better understanding and diagnosis [3]. For multi-source data, the capabilities of data integration, multivariate data exploration, correlation analysis and machine learning model visualization are indispensable. Our team has been funded by DOE and the BNL LDRD program to develop tools for implementation of all these functionalities. Specifically, we developed ColorMapND [4] to map multivariate data to color (Fig. 1 left) and solved the challenge of integration of images of multiple features into a single pseudo-color one. A feature similarity-based approach was provided to reduce the data redundancy and present meaningful mapping to reflect the data correlation. We also developed MulSciView [4], a cross-data spaces visualization and exploration method to understand datasets from a number of experiments. We devised a multilevel scatterplot to allow smooth zooming across data and attribute spaces (Fig. 1 right). Attribute correlation analysis and filtering were also linked to the scatterplot for further analysis. Moreover, we developed a visual analytics approach to learn and understand heterogeneous datasets using hypergraph [4]. This approach can show high-order relations in structural data from multiple sources and enables the users to validate learning model such as classification interactively. Finally, we developed a visualization framework to show the evolution of training in multiple levels of details of a neural network [4] so that multifaceted features learned in the training process can be visualized in overview, layer and neuron details.

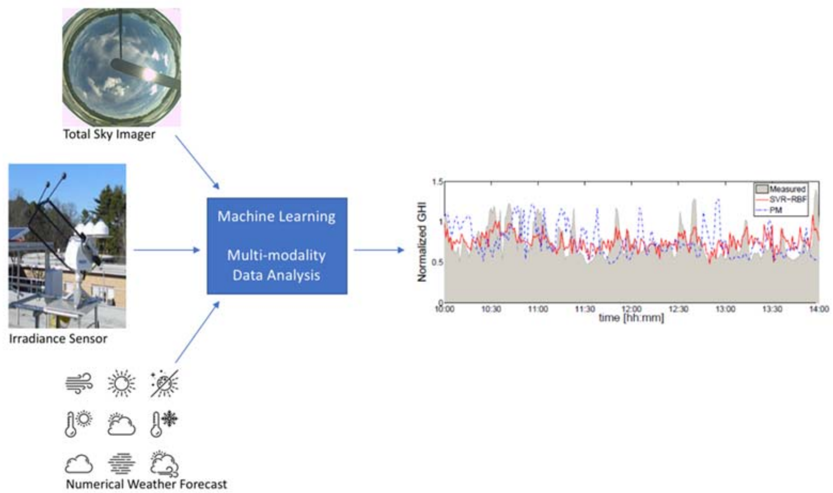


Figure 2: Machine learning algorithm for multi-modality data analysis that integrate cloud images, irradiance sensor readings and weather forecast information to forecast generation of solar energy at a grid-tied solar farm.

In terms of multi-modal data analysis, we are able to fuse various sources of information into prediction, classification and causal inference tasks. Multi-modal data analysis can be done in various ways such as feature extension [11, 12, 13, 14], Bayesian approach [9], multi-kernel learning [10], multi-task learning and representation learning [8], etc. Our earlier work includes combining text mining with social network analysis for email prioritization [11], so that we can infer the social importance from social network analysis and email message importance from text analysis. By combining social feature with text analytics, our work significantly boosted email prioritization accuracy with very few numbers of training examples. In another use case, we integrated Numerical Weather Forecasting, solar irradiance sensors, and ground-based sky images to improve the predictability of confined view of ground-based sky imagery system [12,13], which significantly boosted prediction accuracy (Figure 2). In causal analysis, to suppress noises in causal inference from biological systems, we adopted a Bayesian approach to integrate two-modality information and got significantly better causal discovery [9]. In this case, one modality mainly provides the crucial causal inference information and the other modality complement uncertain cases of causal inference, so that the additional modality information can be used for clarification as a Bayesian prior. In addition, we have done multi-modal brain image analysis using functional MRI (Magnetic Resonance Imaging) where we combined brain morphometry, brain connectome and clinical data to predict dementia [14]. Our method proved that multi-source data analysis was able to improve the classification tasks better than sole brain connectome study.

In this project, we will extend our research to nuclear safeguards. We will use our experience and knowledge in this field to survey the machine learning techniques that could potentially be applied to nuclear safeguards and assess their values in this specific application.

2. **Description of project trajectory events, including overview of tasks, milestones, deliverables, and decision points**

Task 1: Identify the IAEA's priorities in data analysis for nuclear safeguards

We will consult with NA-241 and the IAEA to learn the data sources for each nuclear safeguards scenario including data format, volume and features of interests. We will also work with the IAEA to identify the use cases of high priorities. The information will help us define the scope of our survey and identify visualization and intelligent data analysis techniques that may benefit this application.

Task 2. Survey of visualization and multi-modality data analysis techniques

In this task, we will conduct the actual survey. Given the information from Task 1, we will examine our past and ongoing projects in the fields of visualization and multi-modality analysis. We will evaluate the similarities of data sets and features between these projects and the nuclear safeguards application and assess the values of visualization and machine learning techniques in this specific application. In addition to looking at our internal projects, we will expand the scope of the survey to cover techniques developed by other research teams by conducting literature search.

Task 3. Generate a technical report

In this task, we will generate a final report for the project. In the report, we will summarize all the techniques under the survey and generate a matrix to compare their potential utilities in nuclear safeguards. Based on the comparison results, we will recommend the best candidate techniques that may fulfill the requirements of nuclear safeguards and provide suggestions to improve the nuclear safeguards implementation so that it can take advantages of the innovative techniques to the most extent.

Participating Laboratories

Principal Investigator Table

Lab	Lab Program Manager	Principal Investigator	PI's Email	PI's Phone
BNL	Peter Vanier	Yonggang Cui	ycui@bnl.gov	(631)344-5351

Summary Table

Task No.	Event Type	Event Title	Responsible Lab	Event Date
1	Milestone	Meeting with the IAEA to identify data sources and priority use cases	BNL	12/15/2018
	Deliverable	A report to summarize the meeting outputs	BNL	12/31/2018
2	Milestone	Complete survey of visualization techniques for multi-source data	BNL	07/31/2019
	Milestone	Complete survey of multi-modal data analysis techniques	BNL	07/31/2019
3	Milestone	Complete the final technical report	BNL	09/30/2019
	Deliverable	Final technical report	BNL	09/30/2019

Pertinent references

[1] F. Hohman, M. Kahng, R. Pienta, D. H. Chau, Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. arXiv:1801.06889. Feb 2018.

[2] S. Liu, D. Maljovec, B. Wang, P. T. Bremer, V. Pascucci, Visualizing high- dimensional data: Advances in the past decade, IEEE Transactions on Visualization and Computer Graphics 23 (3) (2017) 1249–1268.

[3] Zeiler, Matthew D and Fergus, Rob. Visualizing and under- standing convolutional networks. In European Confer- ence on Computer Vision, pp. 818–833. Springer, 2014.

[4] S. Cheng, W. Xu, K. Mueller, “ColorMapND: A Data-Driven Approach for Mapping Multivariate Data to Color,” IEEE Trans. on Visualization and Computer Graphics (TVCG), 2018.

[5] W. Zhong, W. Xu, K. G. Yager, G. S. Doerk, J. Zhao, Y. Tian, S. Ha, C. Xie, Y. Zhong, K. Mueller, K. K. Van Dam, “MultiSciView: Multivariate Scientific X-ray Image Visual Exploration with Cross-Data Space Views,” Visual Informatics Journal; IEEE PacificVAST workshop co-held with IEEE PacificVIS conference, Kobe, Japan, Apr. 2018.

[6] C. Xie, W. Zhong, W. Xu, K. Mueller, “Visual Analytics of Heterogeneous Data using Hypergraph Learning,” ACM Transactions on Intelligent Systems and Technology, Special Issue on Visual Analytics, 2018

[7] W. Zhong, C. Xie, Y. Zhong, Y. Wang, W. Xu, S. Cheng and K. Mueller, “Evolutionary Visual Analysis of Deep Neural Networks,” Int’l Conf. on Machine Learning (ICML’17) Workshop on Visualization for Deep Learning, Sydney, Australia, Aug. 2017.

[8] Ngiam, Jiquan, et al. "Multimodal deep learning." *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011.

[9] Shun Yao, Shinjae Yoo, Dantong Yu, "Prior knowledge driven Granger causality analysis on gene regulatory network discovery". BMC Bioinformatics, 16:273, August 2015.

[10] Guillaumin, Matthieu, Jakob Verbeek, and Cordelia Schmid. "Multimodal semi-supervised learning for image classification." *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010.

[11] Yoo, Shinjae, et al. "Mining social networks for personalized email prioritization." *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.

[12] Xu, Jin, et al. "Solar irradiance forecasting using multi-layer cloud tracking and numerical weather prediction." *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. ACM, 2015.

[13] Xu, Jin, et al. "Sensor network based solar forecasting using a local vector autoregressive ridge framework." *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. ACM, 2016.

[14] Wang, Yun, et al. "Accurate Prediction of Alzheimer's Disease Using Multi-Modal MRI and High-Throughput Brain Phenotyping." *bioRxiv* (2018): 255141.