The Journal of Open Source Software

DIANNA: Deep Insight And Neural Network Analysis

Elena Ranguelova [1], **Christiaan Meijer** [1], **Leon Oostrum** [1], **Yang Liu** [1], **Patrick Bos** [1], **Giulia Crocioni** [1], **Matthieu Laneuville** [2], **Bryan Cardenas Guevara** [2], **Rena Bakhshi** [1], **and Damian Podareanu** [2]

1 Netherlands eScience Center, Amsterdam, the Netherlands **2** SURF, Amsterdam, the Netherlands

DOI: 10.21105/joss.04493

Software

- Review
- Repository
- Archive

Editor: Patrick Diehl

Reviewers:

- @Athene-ai
- @sara-02

Submitted: 22 March 2022
Published: 01 December 2022

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

The growing demand from science and industry inspired rapid advances in Artificial Intelligence (AI). The increased use of AI and the reliability and trust required by automated decision making and standards of scientific rigour, led to the boom of eXplainable Artificial Intelligence (XAI). DIANNA (Deep Insight And Neural Network Analysis) is the first Python package of systematically selected XAI methods supporting the Open Neural Networks Exchange (ONNX) format. DIANNA is built by and designed for all research software engineers and researchers, also non-AI experts.

Statement of need

AI systems have been increasingly used in a wide variety of fields, including such data-sensitive areas as healthcare (Alshehri & Muhammad, 2021), renewable energy (Kuzlu et al., 2020), supply chain (Toorajipour et al., 2021) and finance. Automated decision-making and scientific research standards require reliability and trust of the AI technology (Xu, 2019). Especially in AI-enhanced research, a scientist need to be able to trust a high-performant, but opaque AI model used for automation of their data processing pipeline. In addition, XAI has the potential for helping any scientist to "find new scientific discoveries in the analysis of their data" (Hey et al., 2020). Furthermore, tools for supporting repeatable science are of high demand (Feger, 2020).

DIANNA addresses these needs of researchers in various scientific domains who make use of AI models, especially supporting non-AI experts. DIANNA provides a Python-based, user-friendly, and uniform interface to several XAI methods. To the best of our knowledge, it is the only library using Open Neural Network Exchange (ONNX) (Bai et al., 2019), the open-source, framework-agnostic standard for AI models, which supports repeatability of scientific research.

State of the field

There are numerous Python XAI libraries, many are listed in the Awesome explainable AI (Wang & others, 2022) repository. Popular and widely used packages are Pytorch (Paszke et al., 2019), LIME (Ribeiro et al., 2016), Captum (Kokhlikyan et al., 2020), Lucid (Schubert & contributors, 2021), SHAP (Lundberg & Lee, 2017), InterpretML (Nori et al., 2019), PyTorch CNN visualizations (Ozbulak, 2019) Pytorch GradCAM (Gildenblat & contributors, 2021), Deep Visualization Toolbox (Yosinski et al., 2015), ELI5 (Korobov et al., 2022). However, these libraries have limitations that complicate adoption by scientific communities:

- **Single XAI method** or **single data modality.** While libraries such as SHAP, LIME, Pytorch GradCAM. have gained great popularity, their methods are not always suitable for the research task and/or data modality. For example, GradCAM is applicable only to images.

Most importantly, each library in that class addresses AI explainability with a different method, complicating comparison between methods.

- **Single Deep Neural Network (DNN) format/framework/architecture.** Many XAI libraries support a single DNN format: Lucid supports only TensorFlow (Abadi et al., 2015), Captum - PyTorch (Paszke et al., 2019) and iNNvestigate (Alber et al., 2019) is aimed at Keras users exclusively. Pytorch GradCAM supports a single method for a single format and Convolutional Neural Network Visualizations even limits the choice to a single DNN type. While this is not an issue for the current most popular framework communities, not all mature libraries support a spectrum of XAI methods. Most importantly, tools that support a single framework are not "future-proof". For instance, Caffe (Jia et al., 2014) was the most popular framework in the computer vision (CV) community in 2018, but it has since been abandoned.
- **Unclear choice of supported XAI methods.** ELI5 supports multiple frameworks/formats and XAI methods, but it is unclear how the selection of these methods was made. Furthermore, the library has not been maintained since 2020, so any methods in the rapidly changing XAI field proposed since then are missing.
- **AI expertise is necessary.** The Deep Visualization Toolbox requires DNN knowledge and is only used by AI experts mostly within the CV community.

In addition, on more fundamental level, the results of XAI research does not help to make the technology understandable and trustworthy for non (X)AI experts:

- **Properties of the output of the explainer.** There is no commonly accepted methodology to systematically study XAI methods and their output.
- **Human interpretation intertwined with the one of the explainer.** This is a major problem in the current XAI literature, and there has been limited research to define what constitutes a meaningful explanation in the context of AI systems (Lu et al., 2019).
- **Lack of suitable (scientific) datasets.** The most popular and simplest dataset used as "toy-example" is the MNIST dataset of handwritten digits (LeCun et al., 2010), composed of 10 classes and with no structural variation in the content. Such a dataset is too complex for non-AI experts to intuitively understand the XAI output and simultaneously too far from scientific research data.
- **Plethora of current AI model formats.** The amount and the speed with which they become obsolete is another important show-stopper for reproducibility of AI-enabled science.
- **Lack of funding.** Some libraries, such as iNNvestigate, are becoming obsolete due to the lack of support for research projects that sponsored their creation.

Key Features

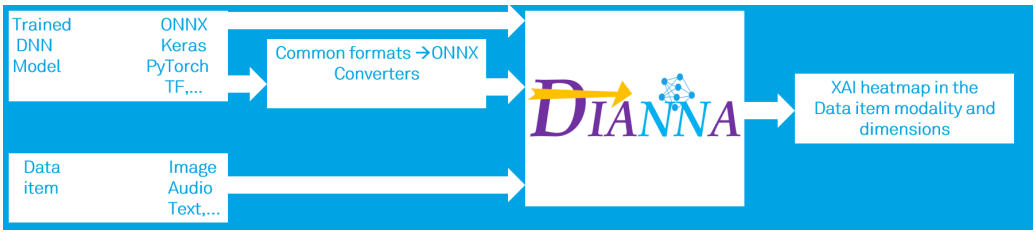

Figure 1: High level architecture of DIANNA

DIANNA is an open source XAI Python package with the following key characteristics:

- **Systematically chosen diverse set of XAI methods.** We have used a relevant subset of the thorough objective and systematic evaluation criteria defined in (Sokol & Flach,

2019). Several complementary and model-architecture agnostic state-of-the-art XAI methods have been chosen and included in DIANNA (Ranguelova & Liu, 2022).

- **Multiple data modalities.** DIANNA supports images and text, we will extend the input data modalities to embeddings, time-series, tabular data and graphs. This is particularly important to scientific researchers, whose data are in domains different than the classical examples from CV and natural language processing communities.
- **Open Neural Network Exchange (ONNX) format.** ONNX is the de-facto standard format for neural network models. Not only is the use of ONNX very beneficial for interoperability, enabling reproducible science, but it is also compatible with runtimes and libraries designed to maximize performance across hardware. To the best of our knowledge, DIANNA is the first and only XAI library supporting ONNX.
- **Simple, intuitive benchmark datasets.** We have proposed two new datasets which enable systematic research of the properties of the XAI methods' output and understanding on an intuitive level: Simple Geometric Shapes (Oostrum et al., 2021) and LeafSnap30 (Ranguelova et al., 2021). The classification of tree species on LeafSnap data is a great example of a simple scientific problem tackled with both classical CV and a deep learning method, where the latter outperforms, but needs explanations. DIANNA also uses well-established benchmarks: a simplified MNIST with 2 distinctive classes only and the Stanford Sentiment Treebank (Socher et al., 2013).
- **User-friendly interface.** DIANNA wraps all XAI methods with a common API.
- **Modular architecture, extensive testing and compliance with modern software engineering practices.** It is very easy for new XAI methods which do not need to access the ONNX model internals to be added to DIANNA. For relevance-propagation type of methods, more work is needed within the ONNX standard (Levitan, 2020) and we hope our work will boost the development growth of ONNX (scientific) models. We welcome the XAI research community to contribute to these developments via DIANNA.
- **Thorough documentation.** The package includes user and developer documentation. It also provides instructions for conversion between ONNX and Tensorflow, Pytorch, Keras or Scikit-learn.

Used by

DIANNA is currently being used in the project "Recognizing symbolism in Turkish television drama" (Verhaar, 2022). An important task is the development of an effective model for detecting and recognizing symbols in videos. DIANNA is being used to increase understanding of AI models in order to explore how to improve them.

DIANNA is also currently being used in the "Visually grounded models of spoken language" project, which builds on previous work by (Alishahi et al., 2017; Chrupala, 2018; Chrupała et al., 2017, 2019). The goal is a multimodal model that projects image and sound data into a common embedded space. Within DIANNA, we are developing XAI methods to visualize and explain these embedded spaces in their complex multimodal network contexts.

Finally, DIANNA has also been used in the EU-funded Examode medical research project (Guevara et al., 2022). It deals with very large datasets and, because it aims to support physicians in their decision making, needs transparent and trustworthy models.

Acknowledgements

This work was supported by the Netherlands eScience Center and SURF.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jozefowicz, R., Jia, Y., Kaiser, L., Kudlur, M., ... Zheng, X. (2015). *TensorFlow, Large-scale machine learning on heterogeneous systems*. https://doi.org/10.5281/zenodo.4724125

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., & Kindermans, P.-J. (2019). iNNvestigate neural networks! *Journal of Machine Learning Research*, *20*(93), 1–8. http://jmlr.org/papers/v20/18-540.html

Alishahi, A., Barking, M., & Chrupala, G. (2017). Encoding of phonology in a recurrent neural model of grounded speech. *CoRR*, *abs/1706.03815*. https://doi.org/10.18653/v1/k17-1037

Alshehri, F., & Muhammad, G. (2021). A comprehensive survey of the internet of things (IoT) and AI-based smart healthcare. *IEEE Access*, *9*, 3660–3678. https://doi.org/10.1109/ACCESS.2020.3047960

Bai, J., Lu, F., Zhang, K., & others. (2019). *ONNX: Open neural network exchange*. https://github.com/onnx/onnx; GitHub.

Chrupala, G. (2018). Symbolic inductive bias for visually grounded learning of spoken language. *CoRR*, *abs/1812.09244*. https://doi.org/10.18653/v1/p19-1647

Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 613–622. https://doi.org/10.18653/v1/P17-1057

Chrupała, G., Gelderloos, L., Kádár, Ákos, & Alishahi, A. (2019). On the difficulty of a distributional semantics of spoken language. *Proceedings of the Society for Computation in Linguistics*, *2*. https://doi.org/10.7275/extq-7546

Feger, S. S. (2020). Interactive tools for reproducible science - understanding, supporting, and motivating reproducible science practices. *ArXiv*, *abs/2012.02570*.

Gildenblat, J., & contributors. (2021). *PyTorch library for CAM methods*. https://github.com/jacobgil/pytorch-grad-cam; GitHub.

Guevara, B. C., Podareanu, D., & Laneuville, M. (2022). *XAI in practice: Medical case study using DIANNA*. Zenodo. https://doi.org/10.5281/zenodo.6303282

Hey, T., Butler, K., Jackson, S., & Thiyagalingam, J. (2020). Machine learning and big scientific data. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *378*(2166), 20190054. https://doi.org/10.1098/rsta.2019.0054

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv Preprint arXiv:1408.5093*.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-Richardson, O. (2020). *Captum: A unified and generic model interpretability library for PyTorch*. https://arxiv.org/abs/2009.07896

Korobov, M., Lopuhin, K., & others. (2022). *ELI5*. https://github.com/TeamHG-Memex/eli5; GitHub.

Kuzlu, M., Cali, U., Sharma, V., & Güler, Ö. (2020). Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. *IEEE Access*, *8*, 187814–187823. https://doi.org/10.1109/ACCESS.2020.3031477

LeCun, Y., Cortes, C., & Burges, C. J. C. (2010). *THE MNIST DATABASE of handwritten digits*. http://yann.lecun.com/exdb/mnist/.

Levitan, S. (2020). *Contribute to the open neural network eXchange (ONNX)* [Medium]. https://medium.com/codait/contribute-to-the-open-neural-network-exchange-onnx-5cfff6889761

Lu, J., Lee, D., Kim, T. W., & Danks, D. (2019). Good explanation for algorithmic transparency. *SSRN*. https://doi.org/10.2139/ssrn.3503603

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *CoRR*, *abs/1705.07874*. http://arxiv.org/abs/1705.07874

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. *arXiv Preprint arXiv:1909.09223*.

Oostrum, L., Liu, Y., Meijer, C., Ranguelova, E., & Bos, P. (2021). *Simple geometric shapes* (Version 1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5012825

Ozbulak, U. (2019). PyTorch CNN visualizations. In *GitHub repository*. https://github.com/utkuozbulak/pytorch-cnn-visualizations; GitHub.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Ranguelova, E., & Liu, Y. (2022). *How to find your Artificial Intelligence explainer* [Medium]. https://blog.esciencecenter.nl/how-to-find-your-artificial-intelligence-explainer-dbb1ac608009

Ranguelova, E., Meijer, C., Oostrum, L., Liu, Y., & Bos, P. (2021). *LeafSnap30* (Version v.1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5061353

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *CoRR*, *abs/1602.04938*. https://doi.org/10.18653/v1/n16-3020

Schubert, L., & contributors. (2021). Lucid. In *GitHub repository*. GitHub. https://github.com/tensorflow/lucid

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://aclanthology.org/D13-1170

Sokol, K., & Flach, P. A. (2019). Explainability fact sheets: A framework for systematic assessment of explainable approaches. *CoRR*, *abs/1912.05100*. http://arxiv.org/abs/1912.05100

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research*, *122*, 502–517. https://doi.org/10.1016/j.jbusres.2020.09.009

Verhaar, P. (2022). Mediating islam in the digital age. In *Digital Scholarship@Leiden*. University of Leiden. https://www.digitalscholarshipleiden.nl/articles/mediating-islam-in-the-digital-age

Wang, Y., & others. (2022). *Awesome-explainable-AI*. https://github.com/wangyongjie-ntu/Awesome-explainable-AI#python-librariessort-in-alphabeta-order; GitHub.

Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions*, *26*(4), 42–46. https://doi.org/10.1145/3328485

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *Deep Learning Workshop, International Conference on Machine Learning (ICML)*.