Exploring deep neural network explanation technologies potential for research using scientific image datasets

Elena Ranguelova, Netherlands eScience centre, Amsterdam

Many scientific challenges can only be tackled with big datasets and black box (deep) neural network (DNN) models. Despite their high predictive accuracy, DNNs lack inherent explainability necessary for system verification, improvement, learning and social acceptance [1], [2]. Therefore, means to interpret a black-box model are very important especially to different domain scientists to not only trust the DNNs, but to get to the knowledge captured by the model as a source of scientific insights. Many post-hoc methods interpret the DNN reasoning by quantification and visualization of the "importance" of individual pixels with respect to the classification decision, producing sensitivity [3], [4], deconvolution [5] or relevance [6] heatmaps. While the usefulness of a heatmap can be judged subjectively by a human, an objective quality measure is needed for systematic interpretation of such maps [7]. Finally, to be adopted by the wide scientific community, well-documented and reliable open-source software and suitable datasets demonstrating the usefulness of the heatmaps on an intuitive level are needed. The heatmap evaluation technique is relevant in any scientific domain where DNN models are used for predictions on images. For example, social scientists studying informal settlements (slums) [8], could benefit greatly from insights on what are the visual characteristics of the hard to define notion of a slum.

For evaluating the potential of explainability technologies being adoption by the scientific community, a simple dataset and simple classification task was needed. The simplest benchmark dataset used widely is MNIST hand-written digits [9] which has 10 classes. Since this is too complex for the desired purpose, the simple 2-class Triangles and Squares (T&S) datasets were proposed [10]. They contain each 100k gray images of randomly oriented (32 x 32) and scaled (64 x 64) triangles and squares with uniform object and background. Applying the Layer-wise Relevance Propagation (LRP) technology to explain the decision 'triangle' or 'square' had shown not intuitive heatmaps [10]. The explanation of an NN model decision for this simple classification task was different than the human reasoning of counting corners. At the same time, it is not a surprise than a data driven model will not necessarily apply the same classification strategy as a human would.

Therefore, it is important to investigate the requirements and design better task and benchmark dataset to decrease the chance of a human strategy being very different than the one employed by an NN. For example, one might consider image dataset containing *k* number of similar objects (e.g. circles) per image with label- the number of objects, i.e. the task for the network will be to learn counting image objects.

The claims that the quality of the LRP is better compared to sensitivity and deconvolution heatmaps [7] have not been verified on simple image datasets with the domain scientists as users in mind, while many alternatives exist. The project will investigate many existing technologies offered as Open Source software (see below for XAI software list) in order to verify the capabilities for explainability and potential for adoption by the wide scientific community. If the time permits, the possibility of converting the created NN model(s) to and implementing the support by the chosen XAI technology of some emerging NN standardised formats (see the list below) will be also pursued.

Pre-requisites

Linear algebra; programming (e.g. Python, Jupyter notebooks); Git; familiarity with or interest for (deep) neural networks, specifically CNNs for image classification and eXplainable Artificial Intelligence (XAI) technologies

Technologies

XAI: DeepExplain, iNNvestigate, SHAP, Eli5, Skater, Yellowbrick, Lucid

NN standard formats: NNEF, OpenVX, ONNX, MMdnn

References

[1] G. Montavon, W. Samek, and K. M¨uller, "Methods for Interpreting and Understanding Deep Neural Networks," CoRR, vol. abs/1706.07979, 2017. http://arxiv.org/abs/1706.07979

[2] S. Chakraborty, R. J. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. D. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of Deep Learning Models: A Survey of Results," 2017.

[3] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. M¨uller, "How to Explain Individual Classification Decisions," J. Mach. Learn. Res., vol. 11, pp. 1803–1831, Aug. 2010. http://dl.acm.org/citation.cfm?id=1756006.1859912

[4] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," CoRR, vol. abs/1312.6034, 2013. http://arxiv.org/abs/1312.6034

[5] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.

[6] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. M¨uller,W. Samek, and O´ . D. Sua´rez, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," in PloS one, 2015.

[7] W. Samek, A. Binder, G. Montavon, S. Bach, and K. M¨uller, "Evaluating the visualization of what a Deep Neural Network has learned," CoRR, vol. abs/1509.06321, 2015. http://arxiv.org/abs/1509.06321

[8] M. Kuffer, K. Pfeffer, and R. Sliuzas, "Slums from Space - 15 Years of Slum Mapping using Remote Sensing," Remote Sensing, vol. 8, no. 6, p. 455, 2016.

[9] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[10] Ranguelova, E., Pauwels, E. J., & Berkhout, J. (2018). Evaluating layer-wise relevance propagation explainability maps for artificial neural networks. In *Proceedings - IEEE 14th International Conference on eScience, e-Science 2018* (pp. 377-378). [8588726] (Proceedings - IEEE 14th International Conference on eScience, e-Science 2018). Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/eScience.2018.00107