ENVISIONING DEEP LEARNING NETWORKS

P. Archana Menon[1] , R. Gunasundari[2]*

Abstract

Deep learning techniques are referred to as ‘black box models’ because of its opaque nature. Interpretability techniques and explainable methods help to make the complex network transparent and explain the user why a particular decision is made by the model. Different visualization techniques used for the interpretation of deep learning networks for both image and non-image classification are investigated in this paper.

Keywords— Deep Learning, Black Box, Interpretability, Explain ability, Visualization Techniques, Activation maps, XAI

I. INTRODUCTION

Deep Learning techniques are applied in various domains. There are areas where transparency of the model is not important rather the performance matters. It is crucial in some of the safety critical domains such as medical treatment, judicial, financial etc. to understand why the network makes a particular decision. Explainable AI and Interpretability techniques help to reveal the mystery behind the networks by making the network more transparent. These techniques translate the happenings inside a network into an output form which a person can interpret.

The opaqueness of ML and DL models make the adoption of such models in crucial domains more crucial. Recent developments in XAI show the importance of transparency in decision making. The remainder of the paper is organized as follows. Section II comments on the different visualization methods available for both image and non-image data classification. A few of the techniques used by some of the researchers are also discussed here. The paper concludes in Section III.

II. VISUALIZATION METHODS

In situations where AI model takes a decision or it has a supportive role, there should be some way to verify these decisions. These verification or explanation helps one to build trust in the AI system. Explaining and interpreting the results is becoming an inevitable component in AI architecture.

Interpretability methods can be applied either after the training of the network or can be build along with the network. Applying interpretability techniques after training the model helps to save time. These techniques explain the predictions of a model using visual representations. Network behavior can be visualized using with the help of saliency maps, heat maps, feature importance maps etc.

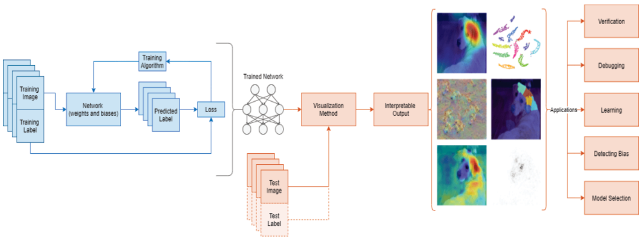

Fig. 1.Visualization Methods

A. Interpretability Methods for Image Classification

Among the different interpretability techniques available, which one to choose purely depends upon the network and the interpretation we want? The interpretations could be global or local [1]. The general behavior of the network is explained by the global explanations. Local explanations are responsible for explaining specific input/output

[1] Department of Computer Science
[2] Department of Computer Applicaiton
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

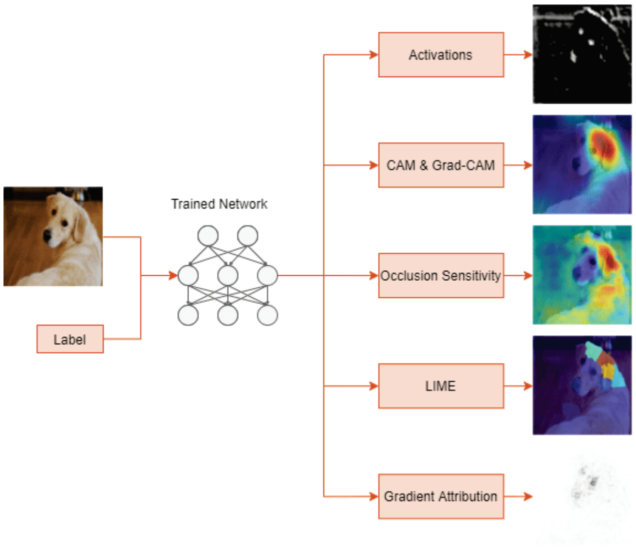

Fig. 2. Visualization Methods for Image Classification

Different visualization methods available for classifying images are shown in Fig. 2. Researchers use different interpretability techniques for visualizing their model. Major interpretability techniques used for image classification in deep learning are explained here.

1) Activations: Activation maps are visual representation of activation numbers at various layers of the network [2]. Here, we examine the output activations of each layer. Low level features of an image will be learnt in the first convolution layers. In the deeper layers, the network learns more complicated features.

2) CAM: CAM (Class Activation Mapping) exposes the strongly influenced region of an image by the predictions of a network. CAM provides GUI based toolkit for classification mosels.It allows us to view feature-activation and class activation maps [3].

3) Grad-CAM: Gradient-weighted Class Activation Mapping makes use of the classification score of the convolutional features. It produces a heat map by using the gradients of the target image from the final convolution layer. The map highlights important regions of the image. A high gradient represents the region with high influence [4]

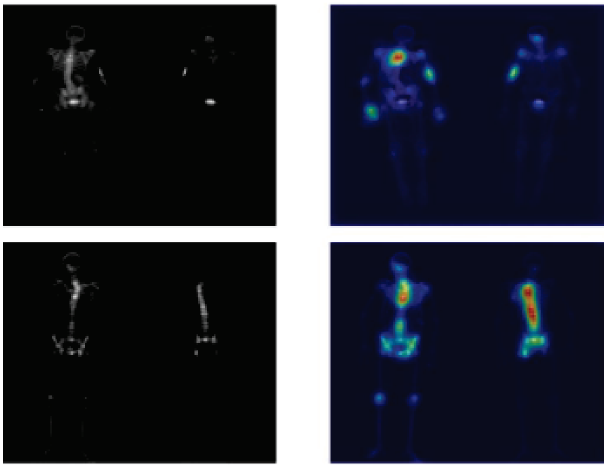

Fig. 3.Correctly classified BS images with activation maps

Identification of bone scintigraphies containing metastatic bone disease was done by [13]. Fig. 3 [13] shows correctly classified BS images with activation maps.

4) Occlusion sensitivity: This method measures network sensitivity to small perturbations in the input data. This technique disturbs small areas of input data with an occluding mask and measures the change in probability score [5].

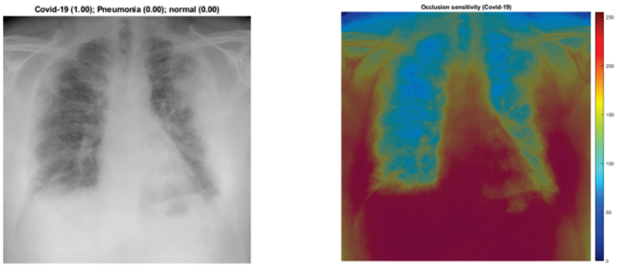
Covid-19 (1.00); Pneumonia (0.00); normal (0.00)
Occlusion sensitivity (Covid-19)
250
200
150
100
50
0

Fig. 4.Occlusion Sensitivity map on Covid-19 class

CovidNet, a Deep Neural Network proposed by [14] used limited training parameters for the classification of Covid-19 patients. The fig. 4 [14] shows Occlusion Sensitivity map on Covid-19 class on a chest X-ray image.

5) LIME: Local Interpretable Model-Agnostic Explanations approximates the deep network behaviour in a simpler manner. LIME focuses on training local surrogate models for explaining each prediction. It identifies the regions in an image which are strongly associated with the prediction [6]. This method initially segment an image into features and generate many synthetic images which inturn classify these images with a DNN. Asimple linear regression model is used to fit the fetures in each synthetic images and then computes the importance of each feature [1].

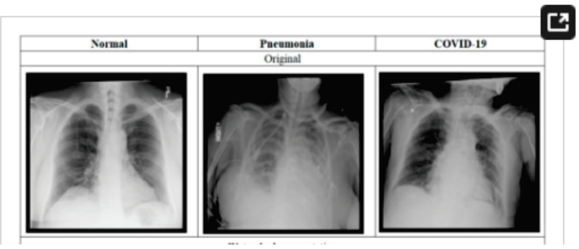

Fig. 5. Chest X-ray images

Fig 5 [15] shows chest X-ray images of persons who are healthy, affaected by pneumonia and affected by Covid-19 disease.

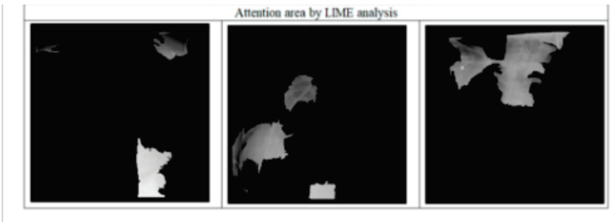

Fig. 6.Chest X-ray images after LIME analysis

Fig. 6 [15] shows chest X-ray images of different categories of persons after the nalysis by LIME algorithm.

6) Gradient Attribution: This shows the user which pixels are responsible for making the clasification decision by bulding pixel resolution maps [10]. It computes the gradient of the class score with respect to the input pixels [11]. The maps produced by this method and the input image have the same size.

7) Deep Dream: It is a feature visualization technique which visualizes the patterns learned by a network. It synthesizes images which are responsible for activating the network layers [7].

Fig. 7 [18] is an example image which is generated using Deep Dream. Deep Dream, uses a CNN to enhance the patterns in images and creates a dream-like hallucinogenic appearance in the over-processed images.

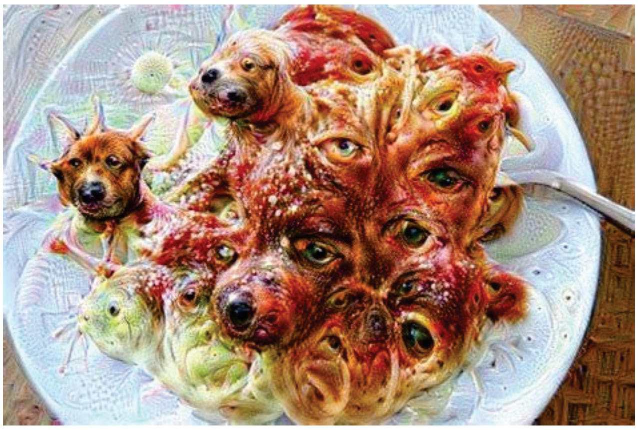

Fig. 7.An image generated using Deep Dream

8) T-SNE: T-distributed Stochastic Neighbourhood Embedding (tSNE) is an algorithm mainly used for data visualization and exploration. It calculates similarity measure between pairs of instances in the high dimensional space and in the low dimensional space and using cost function, it optimizes the similarity measures [8]. It is generally used to visualise the high dimensional data in low (2 or 3) dimensions.

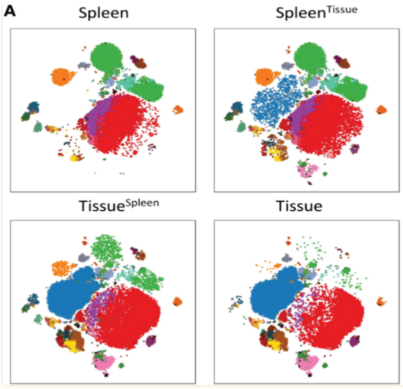

Fig. 8.Samples of spleen and tissue generated using t-SNE

[17] used t- SNE to display high- parameter data which is shown in Fig. 8 [17]. They generated t-SNE plots using parental samples and built samples for constructing artificial dataset.

9) Maximal and minimal activating images: If we find out the images which are strongly or weakly activating the neural network, it helps us to understand why the neural network is making correct or incorrect classifications [9].

Based on the application and requirement user can choose any of the visualization techniques available to plot the output of the model. Saliency map on X-ray images using various methods are shown in the Fig. 9 [16].

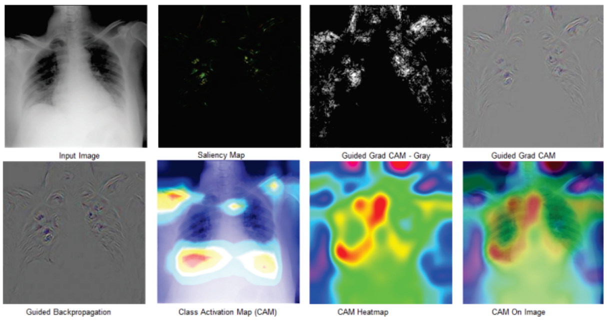

Fig. 9.Saliency map on X-ray Image using different methods

B. Interpretability Methods for Non-image Data Interpretability techniques are mainly employed in image classification problems because of the fact that non-image data interpretation is challenging as they are non-visual in nature.

1) Grad-CAM: Grad- CAM can also be used to visualize the classification decisions of a 1-D CNN trained on time series data. For time series data, it computes the important time steps for the classification decision of the network [12].

2) LIME: LIME can be used for explaining tabular as well as text data. In tabular data, samples for LIME are taken from the training data's mass center [1]. In th ecase of text data, new texts are generated from the original ones by removing words from the original text randomly. Dataset would contain only 1 and 0 as values where the feature value 1 represents word is included and 0 if it is removed [1].

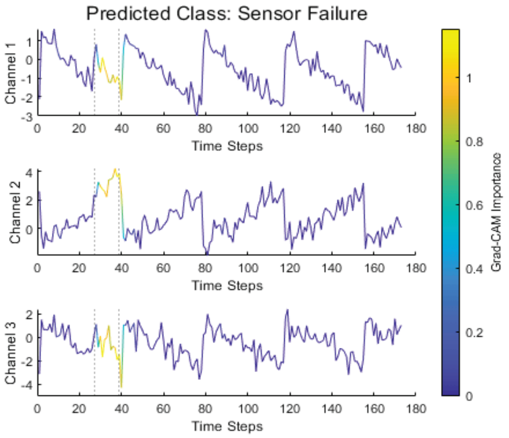

Fig. 10. Classifiying time-series data as Normal or Sensor Failure

In fig. 10, the map generated using LIME highlights the region s used by the network for the classification of time-series data as "Normal" or "Sensor Failure"

.

III. CONCLUSION

ML and DL algorithms are able to achieve excellent performance in various domains. It could make decision and optimize the results of even sophisticated problems. Due to the non-linear structure of the models, they are called as black boxes. The black box nature of the deep learning models can be explained with the help of supporting interpretability techniques. These techniques, when applied on the trained model, reveal the mystery of the neural network and it explains how the model came up with a particular decision. Different visualization techniques for interpreting the image and non-image data are discussed in this paper. A few of the techniques used by some of the researchers are also mentioned here.

REFERENCES

[1] Molnar, C. (2023, March 2). Interpretable machine learning. christophm.github.io. https://christophm.github.io/interpretable-ml-book/ (accessed March 22, 2023)

[2] Sarkar, T. (2019, October 14). Activation maps for deep learning models in a few lines of code. Medium, https://towardsdatascience.com/activation-maps-for-deep-learning-models-in-a-few-lines-of-code-ed9ced1e8d21, (accessed March 22, 2023).

[3] Cam-visualizer: Class activation map visualization toolkit. Intel. (n.d.), https://www.intel.com/content/www/us/en/developer/articles/reference-implementation/class-activation-map-visualizer.html (accessed March 22, 2023).

[4] Reiff, D. (2022, May 12). Understand your algorithm with grad-cam. Medium. https://towardsdatascience.com/understand-your-algorithm-with-grad-cam-d3b62fce353 (accessed March 22, 2023)

[5] Googlenet. MATLAB & Simulink. (n.d.). https://www.mathworks.com/help/deeplearning/ug/understand-network-predictions-using-occlusion.html (accessed March 22, 2023)

[6] Arteaga, C. (2019, October 22). Interpretable machine learning for image classification with lime. Medium. https://towardsdatascience.com/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13 (accessed March 22, 2023)

[7] Tutorials: Tensorflow Core. TensorFlow. (n.d.). https://www.tensorflow.org/tutorials (accessed March 22, 2023)

[8] Violante, A. (2018, August 30). An introduction to T-Sne with python example. Medium. https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1 (accessed March 22, 2023)

[9] Googlenet. MATLAB & Simulink. (n.d.). https://www.mathworks.com/help/deeplearning/ug/visualize-image-classifications-using-maximal-and-minimal-activating-images.html (accessed March 22, 2023)

[10] Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2019). Gradient-based attribution methods. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 169–191. https://doi.org/10.1007/978-3-030-28954-6_9

[11] Googlenet. MATLAB & Simulink. (n.d.). https://www.mathworks.com/help/deeplearning/ug/investigate-classification-decisions-using-gradient-attribution-techniques.html (accessed March 22, 2023)

[12] GradCAM. MATLAB & Simulink. (n.d.). https://www.mathworks.com/help/deeplearning/ug/interpret-time-series-classifications-with-grad-cam.html (accessed March 22, 2023)

[13] Ibrahim, A., Vaidyanathan, A., Primakov, S. et al. Deep learning based identification of bone scintigraphies containing metastatic bone disease foci. Cancer Imaging 23, 12 (2023). https://doi.org/10.1186/s40644-023-00524-3

[14] Muhammad Aminu, Noor Atinah Ahmad, Mohd Halim Mohd Noor, Covid-19 detection via deep neural network and occlusion sensitivity maps, Alexandria Engineering Journal, Volume 60, Issue 5, 2021, Pages 4829-4855, ISSN 1110-0168, https://doi.org/10.1016/j.aej.2021.03.052.

[15] Chen, K.-Y., Lee, H.-C., Lin, T.-C., Lee, C.-Y., & Ho, Z.-P. (2023). Deep learning algorithms with lime and similarity distance analysis on COVID-19 chest X-ray dataset. International Journal of Environmental Research and Public Health, 20(5), 4330. https://doi.org/10.3390/ijerph20054330

[16] Ghoshal, Biraja & Tucker, Allan. (2020). Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection.

[17] Carlos P. Roca, Oliver T. Burton, Julika Neumann, Samar Tareen, Carly E. Whyte, Vaclav Gergelits, Rafael V. Veiga, Stéphanie Humblet-Baron, Adrian Liston, A cross entropy test allows quantitative statistical comparison of t-SNE and UMAP representations, Cell Reports Methods, Volume 3, Issue 1, 2023, 100390, ISSN 2667-2375, https://doi.org/10.1016/j.crmeth.2022.100390.

[18] Prakash, A. (2020, April 29). Exploring deep dream using tensorflow 2.0. Medium. https://pub.towardsai.net/exploring-deep-dream-using-tensorflow-2-0-93ecd1091fa3 (accessed March 22, 2023)

.