DeepRings: A Concentric-Ring Based Visualization to Understand Deep Learning Models

João Alves
DETI, IEETA
University of Aveiro
Aveiro, Portugal

Tiago Araújo
PPGCC
Federal University of Pará
Belém, Brazil

Bernardo Marques
DETI, IEETA
University of Aveiro
Aveiro, Portugal

Paulo Dias
DETI, IEETA
University of Aveiro
Aveiro, Portugal

Beatriz Sousa Santos
DETI, IEETA
University of Aveiro
Aveiro, Portugal

***Abstract*—Artificial Intelligent (AI) techniques, such as machine learning (ML), have been making significant progress over the past decade. Many systems have been applied in sensitive tasks involving critical infrastructures which affect human well-being or health. Before deploying an AI system, it is necessary to validate its behavior and guarantee that it will continue to perform as expected when deployed in a real-world environment. For this reason, it is important to comprehend specific aspects of such systems. For example, understanding how neural networks produce final predictions remains a fundamental challenge. Existing work on interpreting neural network predictions for images via feature visualization often focuses on explaining predictions for neurons of one single convolutional layer. Not presenting a global perspective over the features learned by the model leads the user to miss the bigger picture. In this work we focus on providing a representation based on the structure of deep neural networks. It presents a visualization able to give the user a global perspective over the feature maps of a convolutional neural network (CNN) in a single image, revealing potential problems of the learning representations present in the network feature maps.**

***Index Terms*—Deep Learning Interpretability, Convolutional Neural Networks Feature Visualization, Concentric Ring Design**

I. Introduction

A great amount of progress in AI techniques, such as machine learning (ML), has been done in the last decade. Areas like personal assistants, logistics, surveillance systems, high-frequency trading, health care, and scientific research have been impacted by the development of these techniques. Transferring decision processes to an AI system might lead to faster and more consistent decisions, freeing human resources for more creative tasks [1].

While some AI systems have already been deployed, what remains a truly limiting factor for a broader adoption of AI technology is the inherent and undeniable risks that come with giving up human control and oversight to 'intelligent' machines [2]. Clearly, for sensitive tasks involving critical infrastructures and affecting human well-being or health, it is crucial to limit the possibility of improper, non-robust, and unsafe decisions and actions [3]. As such, it is of uttermost importance to validate the behavior of an AI system, before deploying it. Hence, it is possible to establish guarantees that it will continue to perform as expected when deployed in a real-world environment. With this objective in mind, several ways for humans to verify the agreement between the AI decision structure and their own ground-truth knowledge have been explored [1], [4]–[6].

Simple models such as shallow decision trees or response curves are readily interpretable, but their predicting capability is limited [7]. More recent deep learning based neural networks provide far superior predictive power, but at the price of behaving as a 'black-box' where the underlying reasoning is much more difficult to extract. Moreover, deep learning is increasingly used in decision-making tasks, due to its high performance on previously-thought hard problems and a low barrier to entry for building, training, and deploying neural networks [8].

Explainable AI (XAI) has developed as a subfield of AI, focused on exposing complex AI models to humans in a systematic and interpretable manner. Interpretability as a way to explain these AI models does not have a clear definition centering around human understanding, varying according to the aspect of the model to be understood: its internals [9], operations [10], mapping of data [11], or representation [12]. Some XAI techniques have already proven useful by revealing to the user unsuspected flaws or strategies in commonly used ML models [8], [13]. However, many questions remain on whether these explanations are robust, reliable, and sufficiently comprehensive to fully assess the quality of the AI system. The case for transparency has been made in many settings, including government policy, business, charity, and algorithms [14]. This topic is also given keen interest in laws such as the General Data Protection Regulation[1] (introduced in the EU in 2018) which seek to provide users with meaningful information about algorithmic decisions.

Image classification tasks have been heavily influenced by the deep learning approach. A type of recent approaches in this domain attempt to identify the parts of a given image which are most salient, i.e. those parts which in a sense were most responsible for leading to the system's prediction [12], [15]. Another type of approach on interpreting neural network predictions for images is via feature visualization. This technique studies what each neuron codes for, or what

Identify applicable funding agency here. If none, delete this.

[1]GDPR Legal Text

information its firing represents. The intuition behind these approaches is that inspecting the preferred stimuli of a unit can shed light into what the neuron is doing [6], [16]. They often focus on explaining predictions showing feature maps from one single convolutional layer at a time [6], [17]. As large-scale model predictions are often computed from a consecutive number of layers that learn hierarchical representations [6], the limitation of not presenting all feature maps at once can lead the user to miss the bigger picture. One of the first works to explore visualization of complete deep networks is Yosinski's work [16]. It uses tabs to keep track of activation of each layer. It is possible to observe gradient ascent on input, and the images on the dataset that are most activated by the selected channels. Recent works [8] use feature maps from layer to layer to build semantic graphs.

The works of [18] and [19] show that common visualization solutions on literature do not comprise the features of a Deep Neural Netork (DNN) representation. A complete state of the art report [20] on multilayer networks presents a graph based approach with implicit hierarchies, and even on this work there is no direction for the DNNs visualization. A new representation based on the structure of DNN graphs may be needed as none is found on the literature. Aiming at providing a representation that considers this structure, a visualization able to present feature maps from several convolutional layers at once and able to convey their hierarchical structure is proposed.

The remaining of this paper is structured as follows: Section 2 presents the visualization proposed, describes the architecture containing the visualization design and the machine learning system. In Section 3, we discuss the impact of our approach. Finally, in Section 4 we draw some conclusions and present ideas for future work.

II. Concentric-Ring Visualization

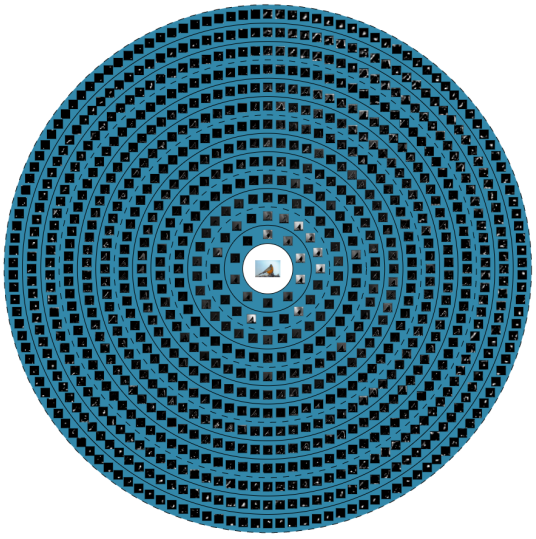

Fig. 1. DeepRings- Concentric-ring based visualization aimed to display the feature maps of VGG16 [21] layers.

In this section, we propose and describe a visualization platform and the architecture linking it to a machine learning engine. The visualization is based on a concentric-ring design, where the number of rings depend on the number of layers of a specific model architecture.

Our visualization, depicted in Figure 1, is a concentric-ring design and each ring has several image placeholders embedded near its outer border. After computing the feature maps, these placeholders are replaced by them using the following criteria: Each ring contains the feature map of one convolutional layer - inner rings contain the feature maps from the first layers and as we move away from the center the feature maps correspond to those computed in deeper layers. The number of placeholders is static for each layer, but in each transition it increases as we move towards the final layers. This decision was made because convolutional neural networks (CNN) have more feature maps as we get close to the final layers.

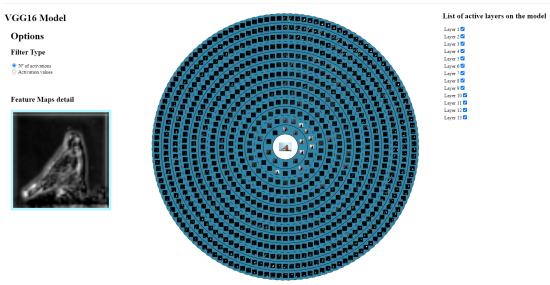

Fig. 2. Visualization platform- Concentric-ring based visualization generated using feature maps ordered by a specific metric is displayed in the center; on the left- details of the feature map selected and the possible sorting metrics; on the right- selection of the convolutional layers to visualize.

In this visualization, the user is able to define how many and which layers s/he wants to visualize and specify what activation metric is considered more relevant. It is possible to choose between two metrics defined by the number of neurons activated or the intensity value of the activated neurons, respectively. The most relevant filter per layer based on the user-defined metric is show directly above the visualization center, and as we rotate clockwise we find with a decreasing importance degree the remaining filters. To create the visualization shown in Figure 1 the metric used was the intensity value of the activated neurons and all the layers from the VGG16 architecture [21] are displayed.

Besides displaying the concentric-ring based visualization, our platform (Figure 2) is also interactive, allowing the user to hover over the feature maps presented in the rings to visualize (on the left) a detailed version of them. Furthermore, the user has the possibility of selecting the convolutional layers s/he wishes to visualize by selecting and deselecting each one of the check box associate with each layer. In Figure 3 the visualization presents only eight out of the thirteen convolutional layers with the collapsed ones represented by thin red semi-rings.

A ML engine is required to obtain the feature maps from a specific image using a prebuilt or an user-defined model.

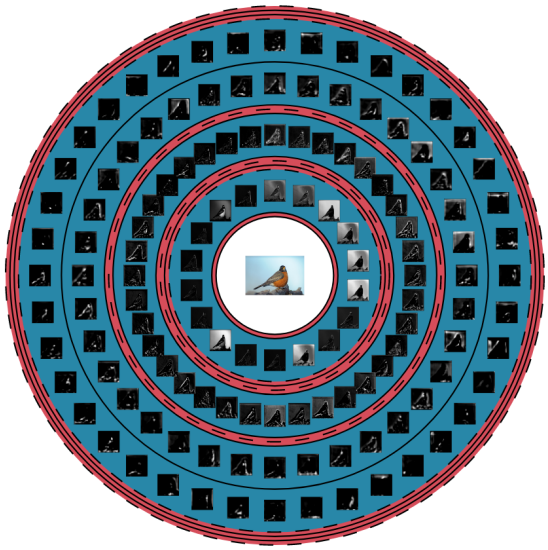

Fig. 3. DeepRings displaying only the feature maps of the layers selected by the user.

The ML engine receives an image from the client and using a preloaded model computes the feature maps. After this operation, the server sends back this information to be displayed using the proposed visualization.

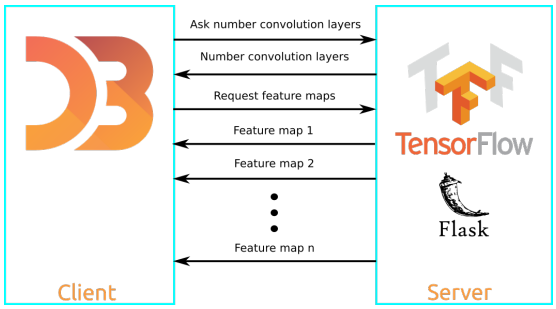

Fig. 4. System Architecture and Information Flow. Front-end visualization uses D3.js to create the visualization. The information required to be displayed is requested to a Flask server which computes the feature maps using TensorFlow.

In this project the main focus was on the visualization part, and the tools were selected as a consequence of this decision. We used D3.js[2] to develop the visualization as is a robust and well-established framework, robust and flexible library, and TensorFlow was used as the machine learning platform. To establish the communication between these two technologies we used a Flask web server[3]. The system architecture together with an example of a possible information exchange between client and server is depicted in Figure 4.

III. Discussion

The layered hierarchy of DNNs mapped to the visualization layers with its respective feature maps allows the observation of patterns between them. In Figure 3 we can see parts of the bird body and its contours activated in several layers of the network, indicating these features are used throughout the network. In Figure 2 the initial features maps activate almost in the whole image, but it gets more specific as we move forward. This shows that general purpose filters are used at the beginning and get more refined as the image advance.

The proposed visualization design does not scale to very deep networks, when it is necessary to display every single layer because the area of each ring tends to become smaller when the number of layers increase as illustrated in Figure 1. However, this issue may be alleviated since the user has the freedom to select what layers are displayed having also the option of visualizing a detailed version of a specific feature map. This can be used to further analyze specific layers within the same representation.

A problem with deploying CNNs in critical domains that require low latency is the system response time. From Figure 3 we can notice that the activation of several feature maps in the last layer is very similar between them. This suggests the existence of redundant filters in the CNN, which leads to unnecessary computation. Removing these filters would potentially alleviate the computational burden leading to a reduction of the system response time.

The visualization layout is also able to show the user that the learned features are hierarchical. The outer rings (final layers) are composed by activation patterns from previous layers. The behavior of learning hierarchical features representations is present in most modern CNN architectures, and it is easily visible using this layout.

A. Exploratory Study

We performed an exploratory study with three domain experts to understand if our system has potential to help end users and researchers to better understand the ”black box” underlying model. The domain experts have background on Computer Vision and use Deep Learning on their work for more than two years. We asked the participants to explore the visualization and performed a simplified Thinking Aloud [22] observation protocol, with no direct tasks to perform, followed by some questions. The sessions took between one and two hours. The following questions were asked:

- What do you observe as positive and negative aspects?
- What do you think of the interactions?
- What features do the visualization lack?
- What can be discovered using the visualization?
- What can be better on the representation?
- Is there any conceptual weakness?

While all participants perceived minor bugs and display errors, they also highlighted some important aspects of lack of information. The class label and certainty are missing aspects of the image input. The visualization also needed an indicator of where the order of feature maps start on the rings. It was not clear that only a specific number of feature maps per layer were presented leading the domain experts to assume that they visualizing all of them.

[2]D3 Website
[3]Flask Website

Even without the information of missing feature maps, the domain experts highlighted that filtering by the best on a specific criteria helps the user find patterns, not overwhelming the user when it shows only the best ones. The domain experts praised the overview presentation of the network in a circular shape, as it also shows the hierarchical structure. Quoting one of the domain experts: "With the visualization, a user starts from the input and observes the abstractions, observing that it goes from the shape, to background removal and class abstraction". They also suggested new features to be included, and the ones aligned with the application roadmap are highlighted in the next section on the future works. Another remark done by one of the domain experts was the fact that this representation type allows to spot errors during training as it allows the user to quickly perceive possible erroneous feature maps while the network is learning. With a pre-trained network, he also pointed out that the visualization suggests that pruning some networks layers could be helpful to reduce inference time as the final layers seem to have less relevant information.

IV. Conclusion and Future Work

Deep learning as an AI technique is increasingly used in decision-making tasks, and for this reason it is important to understand how neural networks learn their internal representations. In this work, we presented an interactive visualization platform able to present a global perspective over the feature maps of a CNN in a single image, showing what features a deep learning model has considered to make predictions. This representation crystallizes the knowledge regarding learning of hierarchical features, while revealing the existence of redundant filters in CNN models.

As future work, we plan to allow the user to change dynamically the model architecture as well as the number of feature maps to be visualized per layer. In addition, we also aim to let the user define their own metrics for the feature maps, allowing them to obtain new insights about the model. Besides displaying feature maps, incorporating other type of representations like saliency maps in this visualization could also be a viable path to improve user understanding. A larger user study can be also conducted to evaluate the potential of the proposed visualization.

Acknowledgment

To all participants involved in the study, thanks for collaborating and providing relevant feedback. This study was supported by IEETA - Institute of Electronics and Informatics Engineering of Aveiro, funded by National Funds through the FCT - Foundation for Science and Technology, in the context of the project [UID/CEC/00127/2019]. This study was also supported by PPGCC - UFPA - Computer Science Graduate Program of Federal University of Pará, funded by National Funds through the CAPES Edital nº47/2017.

References

[1] W. Samek and K.-R. Müller, *Towards Explainable Artificial Intelligence*. Cham: Springer International Publishing, 2019, pp. 5–22.

[2] D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *CoRR*, vol. abs/1606.06565, 2016.

[3] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 2015.

[5] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[6] 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.

[7] T. M. Hehn, J. F. P. Kooij, and F. A. Hamprecht, "End-to-End Learning of Decision Trees and Forests," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 997–1011, 2020.

[8] F. Hohman, H. Park, C. Robinson, and D. H. Polo Chau, "Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1096–1106, jan 2020.

[9] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, pp. 80–89.

[10] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *IJCAI-17 workshop on explainable AI (XAI)*, vol. 8, no. 1, 2017.

[11] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[14] R. Mortier, H. Haddadi, T. Henderson, D. McAuley, and J. Crowcroft, "Human-data interaction: The human face of the data-driven society," *SSRN Electronic Journal*, 10 2014.

[15] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," in *5th International Conference on Learning Representations 2017*, 2017.

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

[17] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017.

[18] A. Komarek, J. Pavlik, and V. Sobeslav, "Network visualization survey," in *Computational Collective Intelligence*, M. Núñez, N. T. Nguyen, D. Camacho, and B. Trawiński, Eds. Cham: Springer International Publishing, 2015, pp. 275–284.

[19] Y. Shaobo and W. Lingda, "A key technology survey and summary of dynamic network visualization," in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 2017, pp. 474–478.

[20] F. McGee, M. Ghoniem, G. Melançon, B. Otjacques, and B. Pinaud, "The state of the art in multilayer network visualization," *Computer Graphics Forum*, vol. 38, no. 6, pp. 125–149, 2019.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] R. Jääskeläinen, "Think-aloud protocol," *Handbook of translation studies*, vol. 1, pp. 371–374, 2010.