Topic: FovEx - Foveation based Explanations for deep neural networks

In contemporary society, the pervasive integration of Deep Learning (DL) and Machine Learning (ML) techniques has assumed a pivotal role across diverse domains such as science, technology, and public life [1]. The reliance on intelligent systems is continuously escalating, evident in applications ranging from autonomous vehicles and adaptive email filters to cutting-edge proactive law maintenance models. With the growing prevalence of intricate DL models, there arises an imperative need to comprehensively comprehend their internal mechanisms and glean insights into their outcomes. This imperative forms the foundational motivation driving the pursuit of Explainable Artificial Intelligence (XAI) [2]. Within the domain of Explainable Artificial Intelligence (XAI), diverse methodologies have emerged to elucidate the inner workings of Deep Learning (DL) models when presented with specific inputs, particularly within the context of computer vision tasks.

XAI methods broadly fall into two categories: gradient-based approaches, exemplified by techniques such as GradCAM [3] and GradCAM++ [4], and attribution methods like LRP [5]. Additionally, a distinct paradigm within XAI is perturbation-based methods, which assess how small alterations to the input impact the network's decision-making process. These perturbation-based methods offer an intuitive means of explanation and are applicable even to any kind of black-box models, obviating the need to scrutinize either the activations or gradients [1].

In the context of neural network models, Neural Visual Attention (NeVA) [6], is devised to generate human-like visual scanpaths in a top-down and unsupervised manner. NeVA is constructed around three core elements: 1) a differentiable foveation mechanism that simulates the human foveated vision with a central region of high visual acuity and peripheral coarse resolution, 2) a task model pretrained on a visual downstream task, and 3) an attention mechanism responsible for selecting the next point of interest based on the current stimulus. By its very definition, NeVA highlights the locations in the input image that are the most relevant to solving the corresponding downstream task.

Our study aims to leverage NeVA to generate explanations to interpret black-box models effectively. An explanation is a rule that specifies how a black box model behaves given specific inputs. In our case, the explanations represent attribution maps that highlight the regions of the image crucial for model prediction. NeVA possesses the unique advantage of being model agnostic, i.e., applicable to various architectural paradigms, including Convolutional Neural Networks (CNNs) and Transformers, without requiring any modifications. We intend to adopt NeVA to generate explanations for widely-used models like ResNet50 [7] and Vision Transformer (ViT) [8] on benchmark datasets such as ImageNet [9] and PASCAL VOC 2012 [10]. We plan to conduct a thorough evaluation of these generated explanations, employing both qualitative and quantitative assessments. NeVA, with its human-like exploration of input data, offers the potential for deeper insights into the predictive behavior of deep learning models.

The proposed work comprises the following key components:

- Apply NeVA to ResNet50 [7] and ViT [8] models to generate scanpths on ImageNet [9] and PASCAL VOC [10] datasets.
- Creation of explanations from the generated scanpaths. More specifically, we convert scanpaths to attribution maps (i.e., saliency maps). We analyze the importance of different receptive field sizes.
- Qualitative assessment of the generated explanations through visualizations and class-specific explanation generation.
- Quantitative evaluation of explanation faithfulness through metrics including % Drop [4], % Increase [4], Insertion [11], and Deletion [11].
- Quantitative assessment of explanation localization using the pointing game methodology [12].
- Quantitative evaluation of explanation plausibility in terms of similarity to human gaze using MIT1003 [13] human attention dataset
- Comparison of NeVA-generated explanations with popular methods like gradCAM [3], gradCAM++ [14] (gradient-based method), and Meaningful Perturbations [15] (perturbation-based method) on ImageNet [9] and PASCAL VOC [10] datasets, both qualitatively and quantitatively.

This comprehensive analysis will shed light on the effectiveness of NeVA-generated explanations in comparison to established explanation techniques, facilitating a deeper understanding of model behavior.

Supervisors:	Dr. Dario Zanca, and Prof. Dr. Björn Eskofier
External advisors:	Dr. Matteo Tiezzi (University of Siena)
Student:	Mahadev Prasad Panda
Start – End:	October 20th 2023 - April 19th 2024

References

[1] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021.

[2] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

[3] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.

[5] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In Alessandro E.P. Villa, Paolo Masulli, and Antonio Javier Pons Rivero, editors, *Artificial Neural Networks and Machine Learning – ICANN 2016*, pages 63–71, Cham, 2016. Springer International Publishing.

[6] Leo Schwinn, Doina Precup, Bjoern Eskofier, and Dario Zanca. Behind the machine's gaze: Neural networks with biologically-inspired constraints exhibit human-like visual attention. *Transactions on Machine Learning Research*, 2022.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[11] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[12] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 111–119, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[13] Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113, 2009.

[14] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, 2021.

[15] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.