Local Interpretable Model-Agnostic Explanations for Multitarget Image Regression

Kira Vinogradova[1,2,*], Gene Myers[1,2]

[1]Center for Systems Biology Dresden
[2]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

Abstract

Convolutional neural networks are state-of-the-art for the majority of computer vision tasks, including estimation of optical aberrations in microscopy 3D data defined as a multitarget image regression problem. A novel approach to making multitarget 3D image regression explainable, Image-Reg-LIME, based on the local interpretable model-agnostic explanations (LIME) method, is presented in this study. The explanations are provided as heat maps showing which parts of the input influence the output positively and negatively. We modify LIME to explain the predictions of the image regression model for estimation of the amplitudes of optical aberrations. Additionally, we propose a modification that allows explaining why the ground truth value was not predicted. This research shows that Image-Reg-LIME is a valid method for explaining the estimation of optical aberrations in 3D images.

Keywords

Model-agnostic XAI, Post-hoc, CNN, 3D Image regression

1. Introduction

Methods for making artificial intelligence (AI) explainable are becoming more available as the field develops. Numerous methods have been developed during the past decade, such as CAM [1], Grad-CAM [2], Grad-CAM++ [3], Smoothgrad [4], LIME [5], SHAP [6], RISE [7], LRP [8]. However, explainability in computer vision mainly focuses on image classification and remains underexplored in image regression.

Image regression is a task of predicting a finite rational number from image data. Examples of such a task are: estimation of the human age [9], counting of tumor cells [10]. The output of the *multitarget (multi-output) image regression* is an array of rational numbers. Estimation of the human head pose [11] and estimation of optical aberrations are examples of multitarget image regression.

An explainable AI method Seg-Grad-CAM [12] has been applied to the segmentation network involved in the object pose estimation [13] (multitarget regression in 6D). The decisions of the convolutional long short-term memory model trained on the daily temperature and precipitation mapsto predict the river streamflow (single regression target) were explained by visualizing important regions in these maps using a technique [14] based on Grad-CAM [2]. Another

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.
vinograd@mpi-cbg.de (K. Vinogradova); myers@mpi-cbg.de (G. Myers)

gradient-based method was applied to a U-Net-based CNN with a parallel path with GAP for the task of determining the spatial position of the crack tip [15]. U-Noise [16], originally designed to explain segmentation, has been adapted to brain age estimation [17] (single-target regression in 3D). A method called interpretable classification and regression with feature attribution mapping (ICAM-reg) [18] (modification of (ICAM) [19], a technique previously developed for image classification) was applied to the brain age estimation task.

A monochromatic *optical aberration* can be defined as a deviation of a monochromatic (i.e., with a single wavelength [20]) or quasi-monochromatic [21] light beam from the trajectory proposed by geometrical optics. Monochromatic aberrations lead to image deterioration, with the shape of the distortion depending on the type of the aberration. Aberrations of lower orders influence image quality more than those of higher orders.

The Zernike polynomials [22] are used to describe aberrations in an optical system. Each Zernike polynomial corresponds to a specific type of aberration and is orthogonal to other polynomials. The wavefront ϕ is the sum of the Zernike polynomials Z_i multiplied by their amplitudes a_i, where the index i [23, 24] corresponds to the aberration type:

$$\phi = \sum_i a_i Z_i \tag{1}$$

2. Methods

2.1. Estimation of Aberrations

In a previous study, a deep convolutional neural network PhaseNet [25] was trained for multi-target 3D image regression under supervision. PhaseNet was trained on a simulated data set to find the amplitudes a_i (Eq.1) and tested on experimentally acquired 3D images of fluorescent beads. Later, the method was proved to be applicable to an object of a more complex shape [26]. In both studies, 11 aberrations (the aberrations of the second order, excluding *defocus*, third, and fourth order) contributing to the image quality the most were considered. The 3D microscopic images of fluorescent beads containing these 11 aberration types, which were made publicly available by the authors [25], were used in this research.

2.2. LIME

The Local Interpretable Model-Agnostic Explanations (LIME) method [5] was chosen for this research because it has received high recognition by the community, is model-agnostic, and outputs both positively and negatively contributing features. The latter means that the method can answer a twofold question, "Which parts of image X support prediction Y, and which vote against it?" LIME is an algorithm designed for explaining the predictions of any black-box classifier (including image classifiers) and of black-box regressors trained on tabular data.

It works by approximating the behavior of the complex model locally by learning a white box model (such as linear or logistic regression) around the prediction made for a specific instance. For image classification, this is done by dividing the input image into superpixels, randomly perturbing the input by occlusion of multiple superpixel segments with the mean or a predefined value, and observing the changes in the predictions. The impact of the segments on

the prediction is weighted according to a user-defined distance metric (e.g., cosine similarity). To explain a classification result, LIME requires the prediction function that outputs continuous values (probabilities).

2.3. Proposed: Image-Reg-LIME

First, to make LIME work in 3D, we replaced the default superpixel segmentation algorithm Quick Shift [27] with Simple Linear Iterative Clustering (SLIC) [28] from Scikit-image library [29]. The size of the input images was 32^3, therefore the following parameters were chosen to ensure that the segments are large enough to be meaningful and small enough to demonstrate precise explanations: number of segments $= 8^3 = 512$, compactness $= 0.01$. The channel axis was set to $None$ since the images were grayscale. The rest of the parameters' values remained the default.

Our second modification is in setting the occlusion value to zero (black pixels) because of the nature of the aberrations data with the black background, instead of the default mean pixel value across the input. To explain the target regression prediction ("Why was the value a_i predicted for the target class i?"), we propose to use the original prediction function f of the regression model as is because the network outputs continuous values.

Our key contribution is instructing Local Interpretable Model-Agnostic Explanations for Multitarget Image Regression (Image-Reg-LIME) to answer the question with a reference value: "Why the value a_i^* (e.g., ground truth) was not predicted instead of a_i for the target i?" This is achieved by instructing the method to select the perturbed data set close to the desired value a_i^*. The perturbed examples, which receive predictions close to a_i^*, are weighted with greater values in the output explanation, according to the cosine similarity distance metric.

3. Results

Figure 1A shows a single 2D plane (plane number 27) of an example 3D input image with experimentally introduced $oblique\ astigmatism$ aberration with an amplitude of $0.093\mu m$. The sample image and the network are from the PhaseNet publication [25] and the associated GitHub page [30]. The network predicted the amplitude of $0.088\mu m$ for the target aberration. We replaced the last linear activation function of PhaseNet with $softmax$ and retrained the network to classify the aberration types.

Figure 1B demonstrates the explanation of classification. This experiment was used as a sanity check to test SLIC parameters and the applicability of LIME to these data. Figure 1C shows the explanation of the predicted regression value and Figure 1D answers, why the ground truth was not predicted.

The result of asking Image-Reg-LIME, "Why did the model **not** predict the amplitude of 0.093 μm?" is shown in Figure 2 (2D plane 24 from the z stack). In other words, it helps to understand, "Why was amplitude $0.068\mu m$ predicted for vertical astigmatism instead of amplitude $0.093\mu m$?" The positive impact (in the yellow-red spectrum) stands for the decision "**not 0.093** μm", the segments with a negative impact (in blue) are **against** this decision, meaning that they are actually supporting the opposite decision of predicting 0.093 μm. The segments supporting the decision "**not 0.093** μm" outweighed those against it, therefore, the final decision,

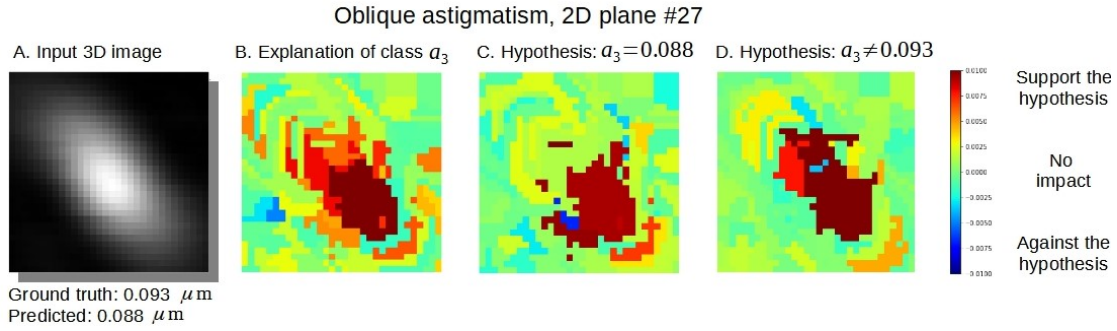

Figure 1: Comparisson of LIME for classification and Image-Reg-LIME for multitarget image regression.

dictated by the features with positive weights, was "**not 0.093** μm". This visualization helps to understand, "Why was amplitude $0.068\mu m$ predicted for vertical astigmatism instead of amplitude $0.093\mu m$?"

The areas with the largest positive weights in the explanation (orange and dark red segments) overlap with the locations of the largest difference in Figure 2C (bright spots in Figure 2D) between the real input image and the image that received the desired prediction. This suggests that the model did not predict the reference value because of this difference.

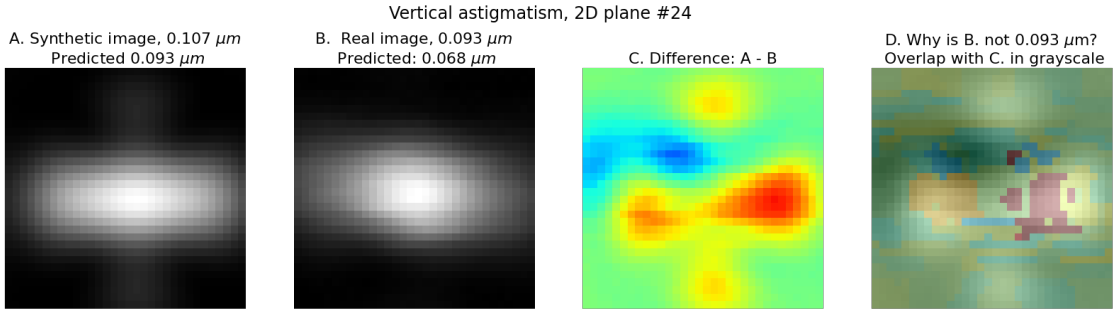

Figure 2: Example result for the question with a reference value.

4. Conclusions

First, the results prove the applicability of LIME to the classification of optical aberrations in experimental 3D microscopy data. Second, the results suggest that the explanations of the proposed method for explainable 3D image regression Image-Reg-LIME highlight input features that are responsible for the prediction of the output regression value. Third, Image-Reg-LIME is shown to point out to input features responsible for an incorrect regression prediction. Moreover, it can be used to explain why one regression value was predicted instead of another used-defined value. The method has the potential to be used for image regression in other application domains, which could be a scope of future work.

References

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[3] A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (2018). doi:10.1109/wacv.2018.00097.

[4] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, arXiv preprint arXiv:1706.03825 (2017). Presented at Workshop on Visualization for Deep Learning, ICML.

[5] M. T. Ribeiro, S. Singh, 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, San Francisco, CA, USA, August 13-17, 2016, pp. 1135–1144.

[6] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.

[7] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, ArXiv abs/1806.07421 (2018).

[8] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.

[9] R. Angulu, J. R. Tapamo, A. O. Adewumi, Age estimation via face images: a survey, EURASIP Journal on Image and Video Processing 2018 (2018). doi:10.1186/s13640-018-0278-6.

[10] Y. Xue, N. Ray, J. Hugh, G. Bigras, Cell counting by regression using convolutional neural network, in: G. Hua, H. Jégou (Eds.), Computer Vision – ECCV 2016 Workshops, Springer International Publishing, Cham, 2016, pp. 274–290.

[11] X. Liu, W. Liang, Y. Wang, S. Li, M. Pei, 3d head pose estimation with convolutional neural network trained on synthetic images, 2016 IEEE International Conference on Image Processing (ICIP) (2016) 1289–1293.

[12] K. Vinogradova, A. Dibrov, G. Myers, Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract), in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 13943–13944.

[13] J. Zhang, W. Li, S. Liang, H. Wang, J. Zhu, Adversarial samples for deep monocular 6d object pose estimation, 2022. arXiv:2203.00302.

[14] L. Schmidt, E. Gusho, W. de Back, K. Vinogradova, R. Kumar, O. Rakovec, S. Attinger, J. Bumberger, Spatially-distributed Deep Learning for rainfall-runoff modelling and system understanding, Technical Report, Copernicus Meetings, 2020. doi:10.5194/

egusphere-egu2020-20736.
[15] D. Melching, T. Strohmann, G. Requena, E. Breitbarth, Explainable machine learning for precise fatigue crack tip detection, Scientific Reports 12 (2022) 9513.
[16] T. Koker, F. Mireshghallah, T. Titcombe, G. Kaissis, U-noise: Learnable noise masks for interpretable image segmentation, arXiv preprint arXiv:2101.05791v3 (2021).
[17] K.-M. Bintsi, V. Baltatzis, A. Hammers, D. Rueckert, Voxel-level importance maps for interpretable brain age estimation, in: Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data: 4th International Workshop, IMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings, Springer-Verlag, 2021, p. 65–74. doi:10.1007/978-3-030-87444-5_7.
[18] C. Bass, M. Da Silva, C. Sudre, L. Z. Williams, H. S. Sousa, P.-D. Tudosiu, F. Alfaro-Almagro, S. P. Fitzgibbon, M. F. Glasser, S. M. Smith, et al., Icam-reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans, IEEE Transactions on Medical Imaging (2022).
[19] C. Bass, M. da Silva, C. Sudre, P.-D. Tudosiu, S. Smith, E. Robinson, Icam: Interpretable classification via disentangled representations and feature attribution mapping, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 7697–7709.
[20] S. H. Schwartz, Geometrical and Visual Optics, McGraw-Hill, 2013.
[21] E. Hecht, Optics, 2017.
[22] F. Zernike, Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode, physica 1 (1934) 689–704.
[23] R. J. Noll, Zernike polynomials and atmospheric turbulence∗, J. Opt. Soc. Am. 66 (1976) 207–211. doi:10.1364/JOSA.66.000207.
[24] L. N. Thibos, R. A. Applegate, J. T. Schwiegerling, R. Webb, Standards for reporting the optical aberrations of eyes, Journal of Refractive Surgery 18 (2002) S652–S660. doi:10.3928/1081-597X-20020901-30.
[25] D. Saha, U. Schmidt, Q. Zhang, A. Barbotin, Q. Hu, N. Ji, M. J. Booth, M. Weigert, E. W. Myers, Practical sensorless aberration estimation for 3d microscopy with deep learning, Optics express 28 (2020) 29044–29053.
[26] K. Vinogradova, E. W. Myers, Estimation of optical aberrations in 3d microscopic bioimages, in: 2022 7th International Conference on Frontiers of Signal Processing (ICFSP), IEEE, 2022, pp. 97–103.
[27] A. Vedaldi, S. Soatto, Quick shift and kernel methods for mode seeking, in: D. Forsyth, P. Torr, A. Zisserman (Eds.), Computer Vision – ECCV 2008, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 705–718.
[28] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (2012) 2274–2282. doi:10.1109/TPAMI.2012.120.
[29] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, the scikit-image contributors, scikit-image: image processing in Python, PeerJ 2 (2014) e453. doi:10.7717/peerj.453.
[30] D. Saha, M. Weigert, U. Schmidt, Phasenet, https://github.com/mpicbg-csbd/phasenet/, 2020.