Explaining Pretrained Deep Neural Network MRI Cancer Classification Outcomes

Kinshu Gupta

Data Science Program

Quinnipiac University

Introduction

Deep learning techniques have demonstrated remarkable performance in medical imaging analysis and diagnosis, surpassing traditional computer vision techniques involving manual feature extraction and classification. In many cases, these techniques have even outperformed medical practitioners in diagnosing diseases from medical imaging [1].

However, this class of machine learning techniques faces fundamental flaws due to their black-box nature and lack of transparency in their decision outcomes, making their use in real-world medical settings problematic [2], [3].

While extensive research focuses on improving the performance of these models, the lack of interpretability associated with deep learning techniques in medical imaging analysis remains an issue to be addressed [4], [5], [6].

Recognizing the critical importance of enhancing interpretability, this study examines the use of input gradients to provide insights into the decision-making process of a pretrained VGG-16 Convolutional Neural Network model trained on prostate lesion MR imaging data for malignancy classification. By overlaying these input gradients on the original images, we aim to elucidate the model's reasoning, enabling the utilization of deep learning's powerful pattern recognition capabilities while maintaining explainability and transparency.

Approach

Dataset

The dataset comprised MR imaging data from 346 participants which was obtained from the ProstateX international challenge focused on the classification of clinically significant prostate cancers. T2-Weighted axial plane image series were used for the analysis [7]. Images were standardized and cropped to lesion centers. An 80-20 split was used for training and testing, respectively. Flip transformations were employed to augment the data and address class imbalance.

Model

Pretrained VGG-16 CNN model weights [8] were loaded and used as a feature extractor for the cropped prostate lesion MR images. The extracted features were then fed into a custom shallow neural network classifier. The classifier architecture consisted of fully connected layers and was trained to predict whether a lesion's malignancy was clinically significant (Gleason score > 7). An Adam optimizer and binary cross-entropy loss were utilized during the training process.

Interpretation

Cropped malignant prostate lesion MR images with accurate predictions were selected for analysis. Logits corresponding to the positive class (malignancy) were chosen as the tensor for which backpropagation was performed through PyTorch's autograd [9] module. Using matplotlib [10], the original MR images were overlaid with heatmaps generated from the corresponding input gradients, with an alpha value of 0.5 for transparency.

Results

After preprocessing, 321 MR images of cropped prostate lesions remained for analysis. The feature extraction layers of the VGG-16 [11] model were kept fixed with the pretrained weights obtained using PyTorch [9].

The classifier head was trained for 100 epochs with checkpoints to evaluate the model's performance on the test set at the end of every epoch. The parameters of the models with the highest test set evaluation metrics for accuracy, sensitivity, and minimum of accuracy and sensitivity were saved at the end of each checkpoint as a form of early stopping.

This training technique consistently resulted in a model with the highest minimum of sensitivity and specificity consistently around epochs 40 to 50. The best model was chosen as the model with the highest minimum of sensitivity and specificity. This model achieved test set evaluation metrics of 63.08% accuracy, 66.67% sensitivity, and 62.26% specificity.

For the analysis of the input gradients, the resulting image with the overlay of the input gradients on the original cropped prostate lesion image visualizes a heatmap of regions in the image that most contributed to the malignant classification outcome of the model. Regions in red indicate a positive contribution towards a malignant classification, whereas blue regions indicate a lower or negative contribution towards a malignant classification of the lesion MR image.

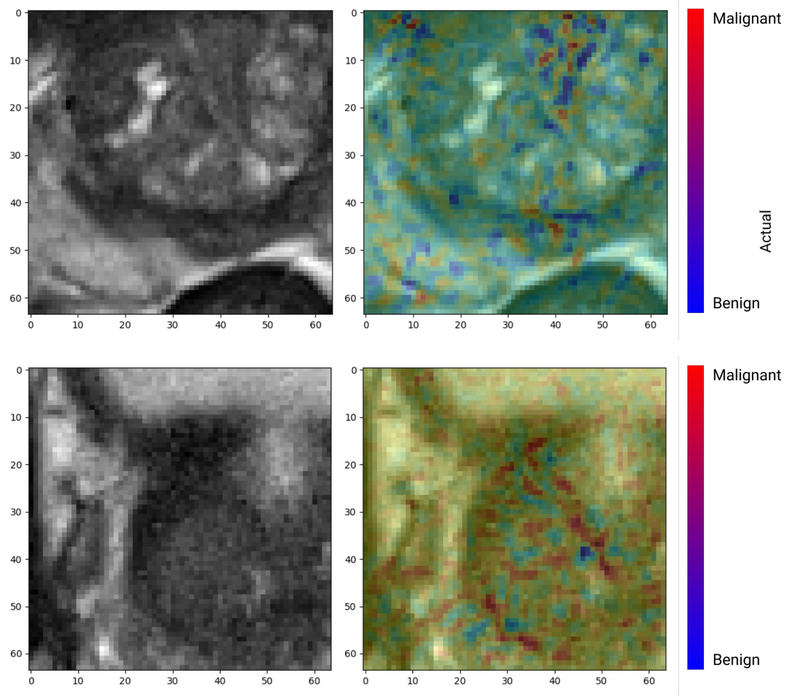


Figure 1. Input Gradients of Prostate Lesion Centered MRI slice

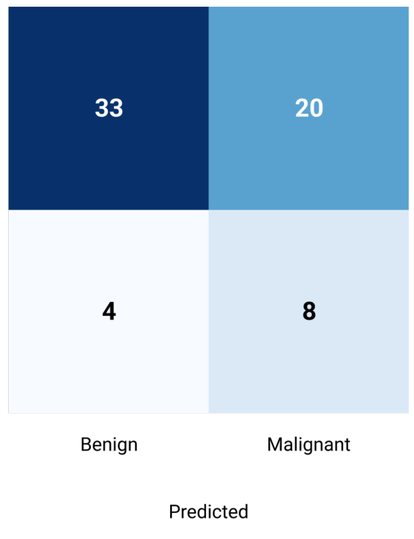


Figure 2. Model Predictions and Ground Truth for Clinical Significance of Lesions

Discussion

This study comprised two main components: building a classification model to categorize cropped prostate lesions as cancerous or non-cancerous, and enhancing the interpretability of the model's decision-making process.

To facilitate this investigation, a relatively simple VGG-16 architecture [11] was chosen over more complex models. Additionally, the study only incorporated axial image series from the MR image data, rather than utilizing a more comprehensive set of MR image series as employed in other research.

These design choices enabled a thorough examination of the model's interpretability but may have contributed to the suboptimal classification performance.

While the benefits of using pretrained weights have been demonstrated in various works, it is possible that a set of parameters more relevant to the prostate lesion classification task could yield better results for the feature extraction layers.

Simple augmentation techniques were applied to address the class imbalance. However, these techniques may have been overly simplistic compared to more advanced methods involving the addition of noise, blurs, shears, and other transformations [12].

Previous works have shown instances where models have learned to focus on irrelevant image artifacts or regions as primary determinants of classification outcomes, which would render the model unusable in real-world scenarios [13], [14], [15].

Despite these challenges, the resulting visualizations demonstrate great potential as tools to validate the decision-making process of deep neural networks to ensure that it is not relying on spurious features as indicators of malignancy.

By examining the input gradients, researchers can verify that the model is attending to clinically relevant regions and features when making its predictions.

Future Directions

Leveraging Comprehensive Data

Preliminary results suggest incorporating sagittal, coronal, and Ktrans MR image series, in addition to axial data used here, could improve performance and provide a more holistic representation of prostate lesions. With robust interpretability techniques established, we can extend our analysis to leverage comprehensive MR data.

Exploring Advanced Architectures

Architectures like U-Net [16], specifically designed for medical imaging analysis, have demonstrated success and may better capture lesion intricacies compared to general models used in this study. Applying our interpretability methods to such advanced architectures could enable high-performing yet transparent models.

Comparative Interpretability Analysis

Beyond input gradients, techniques like Class Activation Maps [17] and Grad-CAM [18] are worth investigating to provide complementary perspectives on model reasoning. A comparative analysis combining multiple interpretability approaches could uncover deeper insights into decision-making nuances.

References

[1] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep Learning Applications in Medical Image Analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018, doi: 10.1109/ACCESS.2017.2788044.
[2] A. J. London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability," *Hastings Cent. Rep.*, vol. 49, no. 1, pp. 15–21, 2019, doi: 10.1002/hast.973.
[3] C. J. Hoofnagle, B. van der Sloot, and F. Z. Borgesius, "The European Union general data protection regulation: what it is and what it means*," *Inf. Commun. Technol. Law*, vol. 28, no. 1, pp. 65–98, Jan. 2019, doi: 10.1080/13600834.2019.1573501.
[4] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.
[5] M. A. Gulum, C. M. Trombley, and M. Kantardzic, "A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging," *Appl. Sci.*, vol. 11, no. 10, Art. no. 10, Jan. 2021, doi: 10.3390/app11104573.
[6] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 4, p. e1312, 2019, doi: 10.1002/widm.1312.
[7] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "SPIE-AAPM PROSTATEx Challenge Data." The Cancer Imaging Archive, 2017. doi: 10.7937/K9TCIA.2017.MURS5CL.
[8] T. maintainers and contributors, "TorchVision: PyTorch's Computer Vision library," *GitHub repository*. GitHub, 2016. [Online]. Available: https://github.com/pytorch/vision
[9] A. Paszke *et al.*, "PyTorch: an imperative style, high-performance deep learning library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 8026–8037.
[10] The Matplotlib Development Team, "Matplotlib: Visualization with Python." [object Object], Apr. 09, 2024. doi: 10.5281/ZENODO.592536.
[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015. doi: 10.48550/arXiv.1409.1556.
[12] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning Augmentation Policies from Data." arXiv, Apr. 11, 2019. Accessed: Apr. 22, 2024. [Online]. Available: http://arxiv.org/abs/1805.09501
[13] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nat. Mach. Intell.*, vol. 3, no. 7, pp. 610–619, Jul. 2021, doi: 10.1038/s42256-021-00338-7.
[14] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging," *Proc. ACM Conf. Health Inference Learn.*, vol. 2020, pp. 151–159, Apr. 2020, doi: 10.1145/3368555.3384468.
[15] K. Saab, S. Hooper, M. Chen, M. Zhang, D. Rubin, and C. Re, "Reducing Reliance on Spurious Features in Medical Image Classification with Spatial Specificity," in *Proceedings of the 7th Machine Learning for Healthcare Conference*, PMLR, Dec. 2022, pp. 760–784. Accessed: Apr. 22, 2024. [Online]. Available: https://proceedings.mlr.press/v182/saab22a.html
[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.
[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Jun. 2016, pp. 2921–2929. doi: 10.1109/CVPR.2016.319.
[18] 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," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

Acknowledgments

Professor James Soda
Professor Mehmet Gulum
Professor Jesse Kalinowski
Professor Michael Ben-Avie
Data Science Program
College of Arts and Sciences
Quinnipiac University