BMC Digital Health

COMMENT

Open Access

Towards improving the visual explainability of artificial intelligence in the clinical setting

Adrit Rao[1] and Oliver Aalami[1*]

Abstract

Improving the visual explainability of medical artificial intelligence (AI) is fundamental to enabling reliable and transparent clinical decision-making. Medical image analysis systems are becoming increasingly prominent in the clinical setting as algorithms are learning to accurately classify diseases in various imaging modalities. Saliency heat-maps are commonly leveraged in the clinical setting and allow clinicians to visually interpret regions of an image that the model is focusing on. However, studies have shown that in certain scenarios, models do not attend to clinically significant regions of an image and perform inference using insignificant visual features. Here, we discuss the importance of focusing on visual explainability and an effective strategy that has the potential to improve a model's ability to focus more on clinically relevant regions of a given medical image using attention mechanisms.

Keywords Computer vision, Deep learning, Explainability, Medical image analysis

Introduction

Artificial intelligence (AI) and deep learning have the potential to significantly impact clinical decision-making and healthcare workflows [1]. As more AI systems are being translated into the clinical setting, we are uncovering both the strengths and limitations present in human-centered design and the interactions between algorithms and healthcare providers [2]. A major impact of AI-based advancements is occurring in *medical image analysis* [3]. Medical image analysis refers to the study of applying deep learning techniques to classify or detect disease in a wide range of medical imaging modalities (e.g. x-ray, MRI, CT). The main objective of these systems is to hasten the time to diagnosis, democratize access to imaging specialists, and lower the overall workload of physicians by reducing manual time-consuming processes. Research in the area of medical image analysis commonly leverages state-of-the-art developments in the general field of AI and involves applying these techniques to specific clinical problems. In recent years, these systems have become increasingly capable leading to the wider adoption of AI-based imaging diagnostic tools at the point-of-care and the rise in approval of AI-based medical devices by the US Food and Drug Administration (FDA) [4]. When deploying algorithms to provide diagnostic assistance to clinicians, it is important to enable explainable insights in terms of *how* the algorithm is performing inference. When AI-based medical image analysis systems are deployed, they commonly provide the following to the user: a diagnostic prediction, percentage of confidence (or level of certainty), and a heat-map or visualization of which parts of the image the model used to perform inference. Typically, increasing the accuracy of AI is the highest priority during development stages. However, understanding the visual explainability of algorithms through a clinical perspective is important in ensuring

*Correspondence:
Oliver Aalami
aalami@stanford.edu
[1] Stanford University School of Medicine, Stanford, CA, USA

usability and reliability. Here, we discuss the importance of visual explainability in healthcare, current limitations, and potential strategies which can be employed to improve the reliability of AI when deployed in the clinical setting.

Visual explainability of medical AI

Saliency methods that generate heat-maps are commonly leveraged in medical image analysis, allowing clinicians to visualize regions of importance that a model is focusing on in a given image [5]. When these models are deployed in a clinical setting, the ability for a clinician to fully interpret results in a transparent manner is of high importance. Solely focusing on a model's predicted diagnosis limits the confidence in the model for clinical decision-making and lacks visual interpretation. When a clinician traditionally interprets a medical image, they are able to apply cognitive attention to specific regions of the image which are relevant to their diagnosis. The ability to assess the importance of features in a medical image is essential as it allows the clinician to make a final diagnosis based only on relevant information. This cognitive process of *attention* allows the clinician to disregard unnecessary features using prior knowledge ensuring that the diagnosis is based on regions of an image that are important to the specific clinical task and target imaging modality. Studies have shown that medical image analysis models tend to evaluate features more generally when compared to clinicians which can lead to predictions based on irrelevant regions of an image [6]. In one study, it was discovered that a model was using insignificant pieces of text in an x-ray to make predictions regarding COVID-19 rather than clinically significant regions [7]. This is a major limitation of medical AI which can be a significant risk in the clinical setting as it can lead to false interpretations. Replicating the cognitive capability of attention within AI systems is of high importance as it has the potential to improve not only performance but visual explainability in the clinical setting. Finding, developing, and validating strategies that can tune algorithms to perform inference similar to the cognitive process of clinicians can enhance AI interpretation in the clinical setting and lead to more reliable decision-making.

Attention for medical AI

In the general field of AI and deep learning, enabling models to focus on more relevant regions of an image has been studied. A notable strategy used to improve the "focus" of models is *attention mechanisms*. Attention mechanisms aim to replicate the cognitive capability of attention in terms of deep learning. For general image classification purposes, popular implementations include squeeze-and-excitation (SE) [8], convolutional block attention module (CBAM) [9], global-context blocks (GC) [10], among others. Such mechanisms use various operations to tune a model to focus on features in an image that are semantically important in a fully data-driven manner. Across various general AI benchmarks, these mechanisms have significantly improved the relevance of saliency heat-maps to a wide range of specific contexts. Additionally, these mechanisms can be simply integrated within the common image classification model, the convolutional neural network (CNN). Studies have applied attention techniques to CNNs for medical image analysis tasks. One study showed that by applying attention to skin cancer classification, the performance of standard algorithms can be improved [11]. Another study showed that by using attention mechanisms, models can learn to suppress irrelevant regions of medical images from saliency maps [12]. Additionally, in the area of medical image segmentation, attention-based approaches are being leveraged to improve the focus of models toward relevant fields of view when segmenting skin lesions and more [13]. In our study, we aimed to understand the clinical implications of attention and to answer the question "Does attention improve the visual explainability of medical AI?" from the perspective of practicing clinicians [14]. We trained a common image classification model (ResNet-18 [15]) across the x-ray and dermatological image modalities, with and without the use of various types of attention (SE, CBAM, and GC). Then, we visualized saliency heat-maps from the baseline model and subsequently each attention-augmented variant across medical image samples. We then anonymously consulted clinicians in each respective medical domain and presented the samples as shown in Fig. 1 in a randomized manner without specifying which samples were derived using attention [14]. The goal of this observational study was to understand if attention improves the "focus" of models in terms of using clinically relevant regions to perform inference and if this improves visual explainability.

Evaluation of attention for medical AI

In our observational study, we concluded that in the majority of images across these medical image modalities, attention mechanisms improved the visual explainability of the AI in terms of clinical relevance based on subjective feedback from clinicians. The clinicians surveyed, in all but one situation, selected an attention-augmented model over the baseline [14]. The reasoning behind these decisions was based on the ability of the models with attention to focus on more clinically significant regions of the medical image and the low amount of focus that the baseline model had in comparison. Additionally, all attention-augmented models showed an

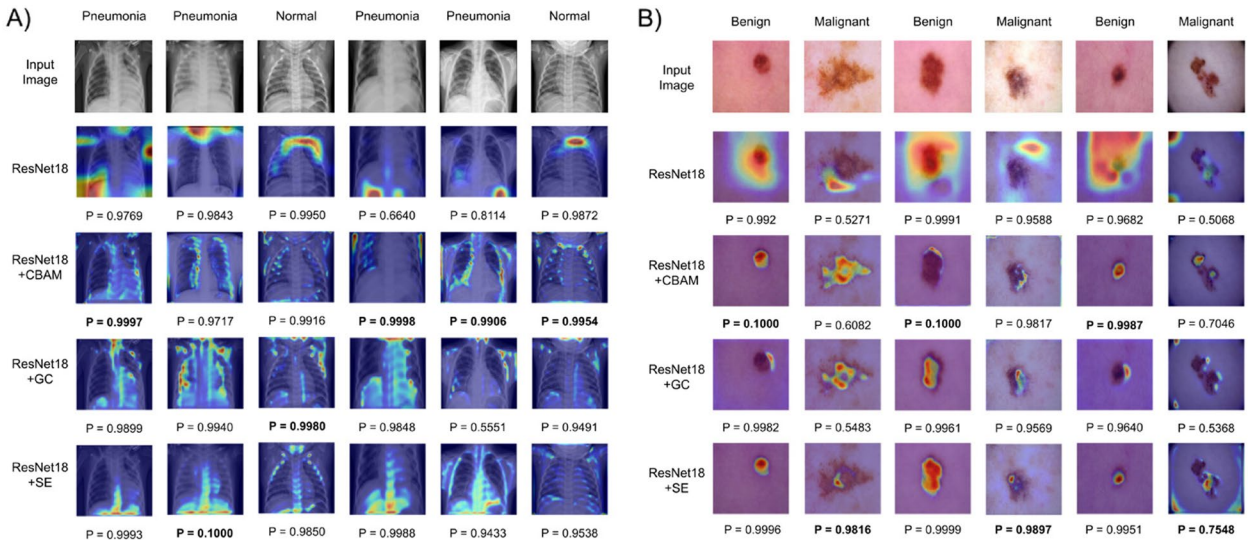

Fig. 1 **A** Chest x-ray (CXR) pneumonia dataset saliency visualizations, **B** Skin (dermatological) cancer dataset saliency visualizations. ResNet-18 is the baseline model and each successive model (downwards) is attention-augmented with CBAM, GC, and SE attention mechanisms. *P* represents the level of confidence that the model has in its diagnostic prediction

increase in accuracy over the baseline. It is important to note that this is only a single study, and further experimentation must be done across various institutions, specialists, and imaging modalities to provide more comprehensive results. Attention mechanisms are only a single strategy that can be used to potentially improve the "focus" of models towards clinically significant regions of an image. Other approaches have enabled models to focus on more clinically significant regions by analyzing gaze data using observational supervision from clinicians [16].

Conclusions

In order to increase the clinical confidence in medical AI systems, more effort needs to be made to find ways to integrate mechanisms and capabilities which can enable models to become more explainable, not just more "accurate". Attention mechanisms, as described, is an approach that can be used to potentially help models learn to focus on clinically relevant regions in medical images. There are various other tools developed in the general field of computer vision to increase the attention of models. By doing so, AI models can potentially provide higher utility in the clinical setting and lead to more robust and clinically explainable decision-making. Making models more explainable can also provide a means to implement quality assurance systems before deploying in any given population.

Acknowledgements
Not applicable

Authors' contributions
Both authors drafted and revised the manuscript. A.R. added the technical aspects of the manuscript and O.A. added clinical content and applicability data. Both author(s) edited and approved the final manuscript.

Funding
No funding received.

Availability of data and materials
The datasets generated and/or analysed during the current study are available here: chest x-ray dataset (https://data.mendeley.com/datasets/rscbjbr9sj/2) and dermatological dataset (https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 23 February 2023 Accepted: 25 May 2023
Published online: 11 July 2023

References

1. Ting DS, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. Nat Med. 2018;24(5):539–40.
2. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med. 2022;5(1):1–15.
3. Suzuki K. Overview of deep learning in medical imaging. Radiol Phys Technol. 2017;10(3):257–73.

4. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3(1):1–8.
5. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl. 2020;32(24):18069–83.
6. Saporta A, Gui X, Agrawal A, Pareek A, Truong S Q, Nguyen CD, Rajpurkar P, et al. Benchmarking saliency methods for chest X-ray interpretation. medRxiv. 2021.02.28.21252634.
7. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell. 2021;3:610–19.
8. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 7132–41.
9. Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). 2018. p. 3–19.
10. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 7794–803.
11. Datta SK, Shaikh MA, Srihari SN, Gao M. Soft attention improves skin cancer classification performance. In: Interpretability of machine intelligence in medical image computing, and topological data analysis and its applications for medical data. Cham: Springer; 2021. p. 13–23.
12. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D. Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal. 2019;53:197–207.
13. Azad R, Asadi-Aghbolaghi M, Fathy M, Escalera S. Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. In: European conference on computer vision. Cham: Springer; 2020. p. 251–66.
14. Rao A, Park J, Woo S, Lee JY, Aalami O. Studying the effects of self-attention for medical image analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. p. 3416–25.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.
16. Saab K, Hooper SM, Sohoni NS, Parmar J, Pogatchnik B, Wu S, Ré C. Observational supervision for medical image classification using gaze data. In: International conference on medical image computing and computer-assisted Intervention. Cham: Springer; 2021. p. 603–14.

Publisher's Note