Eurographics Workshop on Visual Computing for Biology and Medicine (2019)
J. Byška and G. Mistelbauer (Poster Chairs)

Recalibrating Deep Learning Models Attention with Human Attention in Medical Image Classification

A. Adhikari, M. Kruithof, A. Opbroek, S. Raaijmakers, I. Tolios

TNO, The Hague, The Netherlands

Artificial Intelligence (AI) is currently on the rise, as Machine Learning models are increasingly being adopted to provide solutions to real-world problems. These models are becoming substantially more complex than in the past (e.g., Deep Learning), in turn making them progressively less interpretable and transparent [DBH18]. This constitutes a problem in sensitive domains such as in Life and Health Sciences, where large-scale adoption of AI in medical decision support is still lacking. In addition, the often noisy, unstructured character of medical data can affect the effectiveness of such systems [Hol18]. In this context, leveraging on the knowledge and experience of human experts, by allowing them to understand the reasoning of the AI and to provide feedback, would help in improving the quality and usefulness of AI systems. In turn, this would also increase interpretability, transparency and therefore trust on the system.

In this work, we propose an interactive graphical approach for exploiting human experts' knowledge to improve the accuracy of an AI-based image classification system in medical imaging. The user gets an explanation of the reasoning of the AI in the form of an attention map showing to which parts of the image the AI is directing its attention [FHYF19]. The user can then provide feedback on this explanation, which in turn is used to retrain the AI.

In our tool, the user is able to view and inspect (zoom, probe, pan) the original image and the outcome of classification, and to read and show the attention map in the form of a color map (either as an overlay on top of the original image, or on another panel). He can then provide feedback by drawing one or more Regions of Interest (ROI) on the attention map. The user can also check the differences between the initial and the calibrated attention masks. A mock-up of the tool is shown in Figure 1.

We test our proposed solution on a classification task, where a Deep Learning model is trained to classify blood smear slide images as *infected* or *non-infected* by malaria. An expert user is able to access the outcome of the prediction, and a corresponding attention map. By means of our tool, the expert can then graphically adjust the map. In turn, the model can recalibrate its attention by leveraging on the user's feedback.

More specifically, the deep learning model is trained on the NIH malaria data set[†], which contains labeled images of Giemsa-stained thin blood smear slides acquired by a standard light microscope. An attention layer is added after the input layer of the model, and is trained end-to-end. This layer produces the attention map. If the user decides to correct it, the modified map is incorporated in the loss function of the model. The model is then retrained and optimized to take both the original training data and the human feedback into account through the modified loss function. Finally, the user is shown the effect of the feedback on the model, and can accept the new model or roll back to the previous version.

Preliminary tests of the proposed system allowed us to identify some interesting research questions for future work on this direction: (1) how to properly balance knowledge-driven information (user feedback) and data-driven information (e.g. a labeled training set) in one AI model; (2) how to effectively visualize and explain the effect of feedback to the user; (3) what are other types of visual explanations and interactions that allow for an efficient communication about the inner workings of the ML model; (4) at which representational level in the network should attention and feedback be addressed? Higher levels tend to encode more conceptual information.

Acknowledgements

This research was carried out within the Early Research Programme (ERP) Hybrid AI of TNO. We would like to thank Riccardo Satta for his helpful comments.

References

[DBH18] DOŠILOVIĆ F. K., BRČIĆ M., HLUPIĆ N.: Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (2018), IEEE, pp. 0210–0215. 1

[FHYF19] FUKUI H., HIRAKAWA T., YAMASHITA T., FUJIYOSHI H.: Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 10705–10714. 1

† https://ceb.nlm.nih.gov/repositories/malaria-datasets/

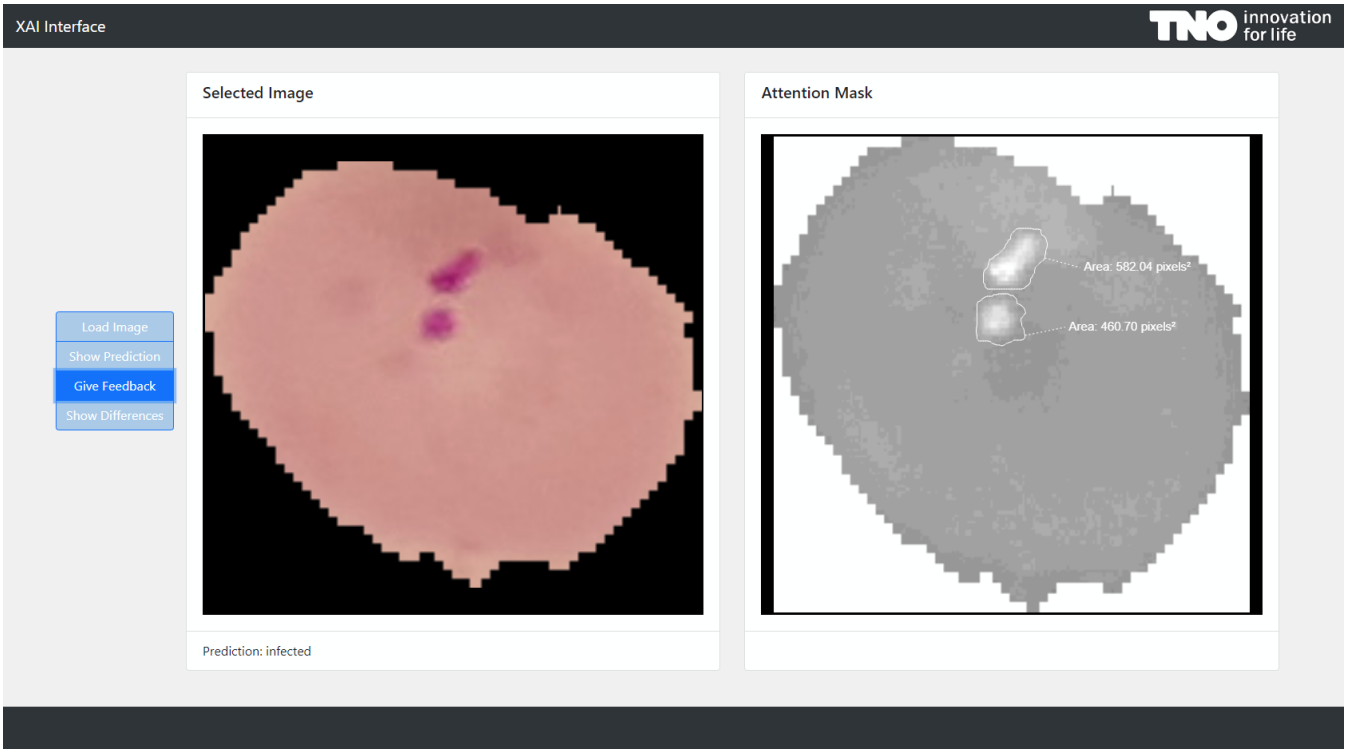

Figure 1: *Mock-up of our tool. An image is loaded on the left panel. The user can get the prediction and the attention of the model by pressing the "Show Prediction" button. The attention map can be seen on the right panel. The user is able to zoom, pan and probe both images. If the user decides to correct the attention of the model, he can draw one or multiple Regions of Interest on the mask, and the modified map is incorporated in the loss function of the model. By pressing the "Show Differences" button, the user can view the differences between the model's attention and the adjusted attention map.*

[Hol18] HOLZINGER A.: From machine learning to explainable ai. *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)* (2018), 55–66. 1