XAI in practice: medical case study using DIANNA

Bryan Cardenas Guevara, Damian Podareanu, Matthieu Laneuville

SURF, Amsterdam, The Netherlands

February 28, 2022

For this work, we explored the benefits and limitations of using DIANNA[1] as an explainable AI tool during a research project in the medical sector. The EU-funded Examode project aims in particular at improving decision-making adoption of extreme-scale analysis and tools[2]. This project includes both image and text modalities and the subject matter makes the robustness of the AI model critical.

1 Problem description.

Biomedical data often comes in different modalities: images, text or feature data that attempt to describe the problem as well as possible. However, the models that are usually deployed for classification or image segmentation consider only one modality. In this research project, the objective is to develop and investigate methods based on deep neural networks to use as many modalities as possible to learn compact yet highly descriptive vector-representations of histopathology multi-modal data. The data size and network design are complex and done at a large scale, meaning that the training process can become increasingly expensive. Algorithmic and data parallelism together with the computational requirements is SURF's key motivation to explore petaflop scaling behavior.

1.1 Dataset overview.

The unit of work in this project are so-called whole-slide images (WSI) of biopsies from the colon, lung or other tissues. Each image is in gigapixel size: images with a dimension of 80000 by 20000 pixels are not uncommon. At the time of writing, the dataset comprises about 1500 WSIs collected from five Dutch medical centers for a total size of about 3 TB. A fraction of those images are linked to text diagnostic reports describing the level of malignancy of the tissue (on a 1 to 5 scale). The raw size of the images and the technical nature and diversity of writing in the pathology diagnoses (multilingual) are central challenges to the task.

[1] https://github.com/dianna-ai/dianna
[2] https://www.examode.eu

1.2 Task description.

In this task, the goal is to learn the link between the WSIs and the diagnostic reports. The neural network learning process provides us with vector-representations of the multi-modal data, which could in turn be used to classify the degree of malignancy indicated in the diagnosis. For this purpose we use OpenAI's CLIP [3] that allows us to to train visual models with natural language supervision or vice-versa. CLIP can be used for a wide variety of tasks without directly optimizing it for those tasks: no hard-labels are required. Additionally, CLIP is easily adaptable for other modalities, which makes the method lucrative for multi-modal representation learning. As a consequence, we can perform zero-shot classification and large-scale image retrieval using natural language. That is, a text description is provided as a substitute for a hard-label to categorize an image.

2 Explainability.

2.1 Possible applications.

While the task is classification, the goal is to help clinicians with decision making. The transparency of the classification algorithm or its capacity to explain decisions post-hoc is therefore important. Saliency maps showing the image features that lead to a specific categorization could help understand if predictions are robust. This is the approach we focused on during this work. Further, the saliency maps could potentially be used as a rough segmentation proxy for classification models. That is, the enormous costs of annotating the gigapixel images in a pixel-wise manner by pathologists can be ameliorated by using simple classification methods, as opposed to deploying expensive segmentation methods. Additionally, the saliency maps permit machine learning engineers or researchers to discover any biases that a classification method may pick up due to data imbalances.

Here is an illustration of this particular problem: in this histopathology project, the gigapixel whole slide data often comes from several labs. The tissue samples are photographed using scanners that can differ from each other in terms of staining, brightness or sharpness. An imbalanced class that is correlated with one of these scanner artifacts could be picked up by the neural network, leading to feature extraction and classification of these artifacts rather than the local morphology of the tissue. Having tools like DIANNA embeded in the workflow could really act here as a debugging tool for non-trivial problems introduced by (biased) data. It serves as a focussed spotlight to highlight and discover these bias-points and test its robustness against unseen data.

2.2 Explainability in practice.

Using DIANNA, we were able to easily test several model-agnostic, local explanation methods, which perform well under well-known data sets. LIME learns an interpretable model locally around the prediction by first using inputs cast in an interpretable rep-

resentation (bag-of-word for text and super-pixels following meaningful features for images) [4]. RISE on the other hand generates an importance map indicating how salient each pixel is for the prediction [2]. It works on black-box models by masking parts of the input and measuring the variations in predictions. Using vanilla implementations of those algorithms did not produce a good visual explanation for the results. Exploring hyperparameters for those methods on top of the architecture and hyper-parameter explorations for the base model proved too time consuming. Part of this is due to the sheer size of the data which makes systematic search prohibitive, but part is also due to the complexity of systematically searching that space. The 'auto-tune' feature available in DIANNA helped us explore the space more efficiently.

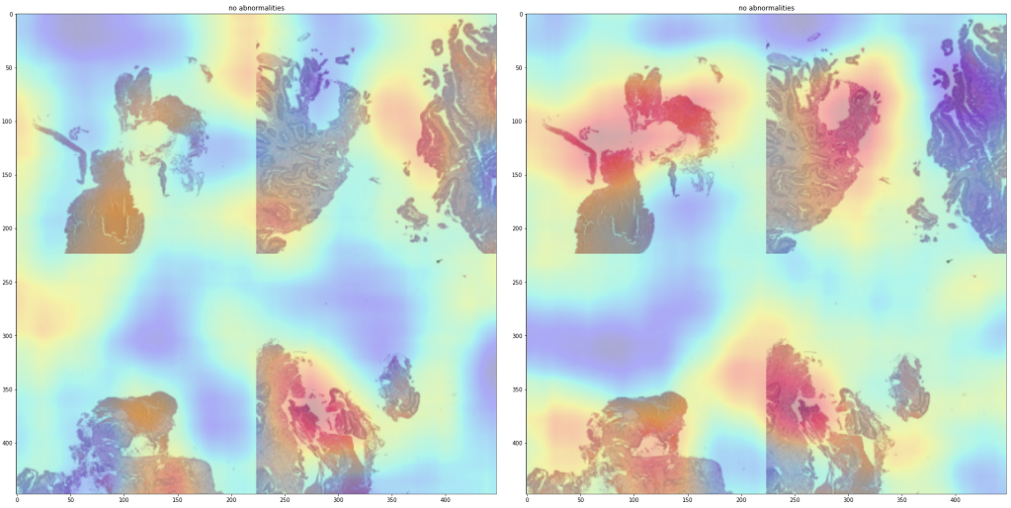

Figure 1: Example saliency map obtained for four patches using the "no abnormalities" query and the RISE method. We use a number of features of 8 and increase the number of masks from left (100) to right (1000). Increasing the number of masks even more still shows no convergence.

The size of the WSI represents another challenge for those algorithms: meaningfully masking a 20 gigapixel image is not effective and super-pixel segmentation algorithms were not always designed with medical images in mind. The images were pre-processed specifically for CLIP and the same preprocessing method had to be applied in the explainability pipeline. However, the LIME and RISE outputs proved to be difficult to interpret when dealing with the preprocessing of multiple images. To illustrate this see, the Figure above with the saliency map of the patches. We chose four patches as toy-example where one of the classes was visibly apparent. Note that the collection of patches should be seen as just one image. We can observe that the saliency map is highlighting regions in an inconsistent manner given different RISE parameter choices, whereas the neural network is correctly classifying these images with a high probability.

This highlight an issue we had with saliency-based explainability methods: what constitutes a good explanation? We found it very easy to over-interpret those maps

and ended up looking for ways to understand the calculations behind that saliency map, creating the need for an explanation of the explanation. The cause of this inconsistency could be present in either the classification method employed or the ability of the explainability tool to deal with gigapixel preprocessing approaches. To diagnose this, LIME and RISE should be tested with a simpler model or dataset, suggesting that researchers using XAI methods should become expert in explainability on top of their domain of expertise. DIANNA mitigates this by providing easy access to benchmark data to help researchers develop a stronger understanding of the methods faster. An interesting overview of what constitutes a good explanation from the social sciences perspective can be found in Miller [1].

Finally, while LIME and RISE are known to perform well when the machine learning algorithm itself performs well (high accuracy), during a research project the accuracy only slowly rises and the state of the art may be well below 90%. In those cases, what should be expected from XAI methods? Is failure of the explainability layer an indication that the model isn't good enough yet (irrespective of accuracy targets), or can the model be good enough for its purpose but the XAI method simply not appropriate? Can those methods be used to decide when the accuracy of a model has reached a reasonable threshold (i.e., when it can be explained reasonably well)?

3 Conclusion.

Using explainability methods consistently in existing research pipelines requires a non-negligible effort at this stage. Tools like DIANNA lower the barrier to entry and allow easy comparison of existing models, which is bound to help. However, the state of the field suggests that, for research applications, a good understanding of XAI methods is still needed for explainability to shine, despite the lower technical barrier to entry.

Bibliography

[1] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[2] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[4] M. T. Ribeiro, S. Singh, and 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*, pages 1135–1144, 2016.