Preliminary Study on the Impact of Attention Mechanisms for Medical Image Classification

Tiago Gonçalves[1,2]
tiago.f.goncalves@inesctec.pt

Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1]Faculdade de Engenharia
Universidade do Porto
Porto, Portugal

[2]INESC TEC
Porto, Portugal

Abstract

Despite their high performance, deep learning algorithms still work as black boxes and are not capable of explaining their predictions in a human-understandable manner, thus leading to a lack of transparency which may jeopardise the acceptance of these technologies by the healthcare community. Therefore, the topic of explainable artificial intelligence (xAI) appeared to address this issue. There are three main approaches to xAI: pre-, in- and post-model methods. In medical images, important information is generally spatially constricted. Hence, to ensure that models focus on the important parts of the images and learn relevant features, several attention mechanisms have been proposed and demonstrated increased performances. This work proposes a comparative study of the application of different attention mechanisms in deep neural networks and the evaluation of their impact on the performance of the models and the quality of the learned features.

1 Introduction

The democratised access to data and the increase of the availability of computational power allowed deep learning (DL) methodologies to achieve nearly-human performances in several areas of science, business and government. The popularity and success of DL in computer vision is mainly due to the introduction of convolutional neural networks (CNNs), which are designed to process unstructured data (*e.g.*, images) [6]. In medical image classification, the main task is to output a diagnosis (*e.g.*, presence or absence of a disease) based on one or more input images. Given the high predictive performance rates of CNNs in other computer vision tasks (*e.g.*, natural image recognition), the application of DL algorithms in medical image classification occurred almost naturally.

2 Explainable Artificial Intelligence

Despite the high performances achieved by DL-based algorithms, their transition into real-world applications is not trivial, due to their complexity (*i.e.*, high-number of parameters) and their black-box behaviour, which may jeopardise their acceptance by the clinical community. Therefore, the topic of explainable artificial intelligence (xAI) appeared intending to contribute to a more transparent AI. Although there is no clear distinction between explainability and interpretability, one may think of these as a three-stage process [3]: pre-model methods focus on understanding the data distribution before building the model, through exploratory data analysis; in-model methods seek to integrate interpretability inside the model (*e.g.*, models based in rules, models based in cases, the use of regularisation techniques during training to obtain sparser or monotonic models); post-model methods are related to a posterior analysis of the model predictions (*e.g.*, using the gradient information to identify the areas of the image that mostly contribute to the final decision, inserting a perturbation and observing the prediction, inverting the representations back to the input pixel space or connecting the representations to semantic concepts). In healthcare applications, it is fundamental to assess the quality of these explanations, for the sake of transparency, ethics and fairness [8].

3 Attention Mechanisms

The intuition behind the application of attention mechanisms in DL algorithms is inspired by the field of psychology, according to which humans tend to selectively concentrate on a part of the information. For instance, the human visual system tends to selectively focus on specific parts of an image while ignoring others. Following this rationale, it is recognised that in AI systems, some parts of the inputs may be more relevant than others (*e.g.*, in automatic translation systems, only a subset of words is relevant). The use of attention was initially proposed in [1] for the task of neural machine translation. Recently, a CNN with a multi-level dual-attention mechanism (MLDAM) has been proposed for macular optical coherence tomography classification [7]. The main novelty of this work in the context of medical image classification is the joint application of a *self-attention* and a *multi-level attention* mechanisms that allow the network to learn relevant features in coarser as well as finer sub-spaces. Regarding the impact of the application of attention mechanisms in the interpretability of the DL algorithms, we point to the work proposed by [2], which approaches the field of interpretability through an analysis of the saliency maps produced by the gradient-weighted class activation mapping (Grad-CAM) [9].

4 Data

We decided to perform experiments on two different use-cases using medical images: breast cancer detection in mammography (CBIS-DDSM data set) and pathology detection in chest X-ray (MIMIC-CXR data set). Each data set contains images of two different classes (binary classification): normal (*i.e.*, without lesion or pathology) and abnormal (*i.e.*, with lesion or pathology).

5 Implementation

We performed a comparative study using three state-of-the-art pre-trained deep learning models as backbones: VGG-16 [10], ResNet-50 [4] and DenseNet-121 [5]. To assess the influence of the use of attention mechanisms, we adapted the MLDAM architecture described in [7] for each of the backbones. We performed experiments with four use-cases: **baseline** (*i.e.*, only the backbone is trained), **baseline with data augmentation** (*i.e.*, the backbone is trained with data augmentation strategies), **baseline and MLDAM** (*i.e.*, the backbone with MLDAM is trained), **baseline and MLDAM with data augmentation** (*i.e.*, the backbone with MLDAM is trained with data augmentation strategies). All the images are resized to the final size of 224×224 and a z-normalisation is applied to each RGB channel. The data augmentation strategy employed in this work is composed of several random rotations, random translations, random scaling, and random horizontal flips. Each model is trained for a maximum of 300 epochs, with binary cross-entropy as the loss function and Adaptive Moment Estimation (Adam) with learning rate 1×10^{-4} as the optimisation algorithm. The batch size varied from 1 to 4, depending on the available GPU memory. We save the best model's parameters in both training and validation sets according to the value of the loss. We tested all the trained models (using the best weights in both training and validation sets) in the test set of each database and computed the accuracy, precision, recall and F1-score. We generated saliency maps [11] for the positive and negative samples of the test set that were correctly predicted by all the use-cases related to all backbones, to assure a fair intra- and inter-comparison. It is important to note that these saliency maps were generated using the models loaded with the best weights in the validation set of each database.

6 Results and Discussion

An extended version of these results is publicly available in a GitHub repository[1]. Table 1 and Table 2 present the accuracy results obtained for the test set of the CBIS-DDSM and MIMIC-CXR, respectively. In both cases, we can observe that the models' predictive performance does not suffer abrupt changes. Figure 1 and Figure 2 present examples of saliency maps obtained for images with label "0" of the CBIS-DDSM

[1] https://github.com/TiagoFilipeSousaGoncalves/attention-mechanisms-healthcare/blob/main/reports/Report.pdf

Table 1: Accuracy results obtained for the test set of the CBIS-DDSM data set: (a) - Baseline, (b) - Baseline with Data Augmentation, (c) - Baseline and MLDAM, (d) - Baseline and MLDAM with Data Augmentation.

Model	Weights	(a)	(b)	(c)	(d)
DenseNet-121	Training	0.6650	0.6142	0.6210	0.5871
	Validation	0.6108	0.5584	0.6396	0.6125
ResNet-50	Training	0.6514	0.6396	-	0.5973
	Validation	0.5956	0.6176	0.5939	0.5854
VGG-16	Training	0.6650	0.6074	0.5939	0.5854
	Validation	0.6244	0.6514	0.6210	0.6041

Table 2: Accuracy results obtained for the test set of the MIMIC-CXR data set: (a) - Baseline, (b) - Baseline with Data Augmentation, (c) - Baseline and MLDAM, (d) - Baseline and MLDAM with Data Augmentation.

Model	Weights	(a)	(b)	(c)	(d)
DenseNet-121	Training	0.8451	0.8312	0.8349	0.8386
	Validation	0.8629	0.8498	0.8535	0.8666
ResNet-50	Training	0.8340	0.8424	0.8470	0.8386
	Validation	0.8535	0.8694	0.8563	0.8414
VGG-16	Training	0.8507	0.8330	0.8293	0.8461
	Validation	0.8629	0.8731	0.8535	0.8647

and MIMIC-CXR, respectively. Taking into account the reported effects of the use of attention mechanisms in the quality of the features learned during training, we were expecting the saliency maps to highlight clear differences between the baseline (with or without data augmentation) and the baseline with an attention mechanism (with or without data augmentation). However, the saliency maps obtained suggest that the behaviour of the models is either similar or completely disparate, without apparent meaning. This lack of consistency makes it difficult to relate direct benefits to the use of attention mechanisms on the properties of post-model explanation methods.

7 Conclusions and Future Work

Our experiments did not present conclusive results on the impact of attention mechanisms in two healthcare use-cases, using three different state-of-the-art backbones. Hence, further work should be devoted to: 1) the development of new experiences with different data processing and augmentation strategies, since it is not clear if these steps are harming the performance of the models; 2) the design of different attention mechanisms that capture features from different scales or levels, since, to the authors' knowledge, there is not a clear pipeline on which are the best scales or levels that should be incorporated in an MLDAM module; 3) generate saliency maps with other methods to see if the results that we obtained are dependent of the post-model interpretability method or not; 4) experiment different state-of-the-art backbones to see if their behaviour differs from the ones we used in this work; 5) try different data sets and different tasks to assess if results are data or task-dependent.

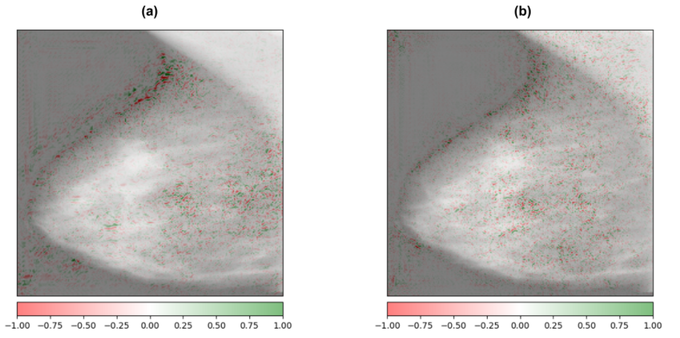

Figure 1: Examples of saliency maps obtained for an image with label "0" of the CBIS-DDSM data set, using the DenseNet-121 backbone model: (a) - Baseline, (b) - Baseline and MLDAM.

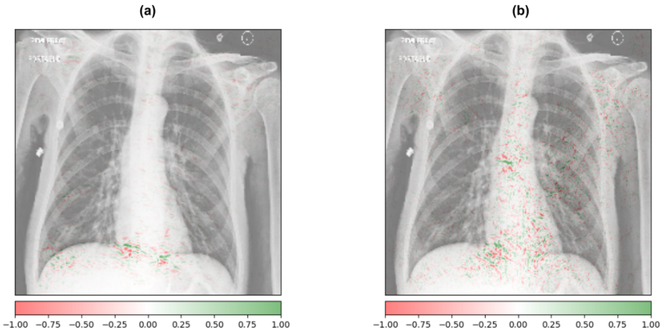

Figure 2: Examples of saliency maps obtained for an image with label "0" of the MIMIC-CXR data set, using the DenseNet-121 backbone model: (a) - Baseline, (b) - Baseline and MLDAM.

Acknowledgements

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership and the PhD grant "2020.06434.BD".

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Cunjian Chen and Arun Ross. An explainable attention-guided iris presentation attack detector. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 97–106.

[3] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*, March 2017. arXiv: 1702.08608.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14539.

[7] Sapna S Mishra, Bappaditya Mandal, and Niladri B Puhan. Multi-level dual-attention based cnn for macular optical coherence tomography classification. *IEEE Signal Processing Letters*, 26(12):1793–1797, 2019.

[8] Cynthia Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv:1811.10154 [cs, stat]*, September 2019. arXiv: 1811.10154.

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.