Augmenting Convolutional Networks with Attention-based Aggregation for Breast Cancer Detection

Filippo Betello*
Sapienza University of Rome
Rome, Italy
betello.1835108@studenti.uniroma1.it

Federico Carmignani*
Sapienza University of Rome
Rome, Italy
carmignani.1845479@studenti.uniroma1.it

Eleonora Lopez
Sapienza University of Rome
Rome, Italy
eleonora.lopez@uniroma1.it

Danilo Comminiello
Sapienza University of Rome
Rome, Italy
danilo.comminiello@uniroma1.it

Aurelio Uncini
Sapienza University of Rome
Rome, Italy
aurelio.uncini@uniroma1.it

***Abstract*—Vision transformers based on self-attention between image patches have shown great potential in image classification. Recently, they have impacted the area of medical image analysis with astonishing results. In this paper we present how a novel architecture, PatchConvNet S60, can be applied to this task using one of the latest dataset of breast cancer images: CBIS-DDSM. The results are presented with two metrics: classification accuracy and Area Under the ROC Curve (AUC). In addition to that, we generate the attention map that point out the areas having a greater probability to be affected by cancer and we compare them with the Region of Interest (ROI), both for positive and negative images.**

I. Introduction

In 2021 in Italy breast cancer is the most diagnosed cancer in women: every year 55,000 new diagnosis and 12,500 death are reported [**1**]. Among them, 87% survive after the 5 years from the prognosis. In 2020 globally, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths [**2**]: at the year, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer.
Often the first recognizable symptom is a lump or thickened area in the breast. It can be achieved a survival probability of 90% or higher, if the disease is identified early. Only 5-10% of cancers are inherited from parents; on the other hand, 85-90% of breast cancers are due to genetic abnormalities that happen as a result of the aging process [**3**]. Evidently, there is the necessity of setting a routine breast cancer screening for all the population after the puberty.

Dosovitskiy et al. [**4**] introduced the Vision Transformer (ViT) architecture, stimulated by the progress of the self-attention based neural networks of transformer models in NLP (Natural Language Processing), for the image classification application: the complete training process in these models is based on splitting the input image into many different patches of same size (word in case of NLP). In recent times, transformers have also contributed on the field of medical image analysis for disease diagnosis: He et al. [**5**] gather more than 170 existing Transformer-based methods in the field. The works [**6**], [**7**], [**8**], [**9**] and [**10**] all present different architectures that use transformers in the area of medical images. Chen et al. [**11**] discovered that a four-image transformer-based model significantly outperforms state-of-the-art multi-view CNNs.
In our work, we use an innovative architecture [**12**] on the CBIS-DDSM breast cancer dataset [**13**], which contains more than 10,000 images and the number of participants is 1,566; then we compare the ground truth of the Region of Interest (ROI) with the one coming from the attention mechanism of the architecture. The whole process can be seen in Fig. 2.

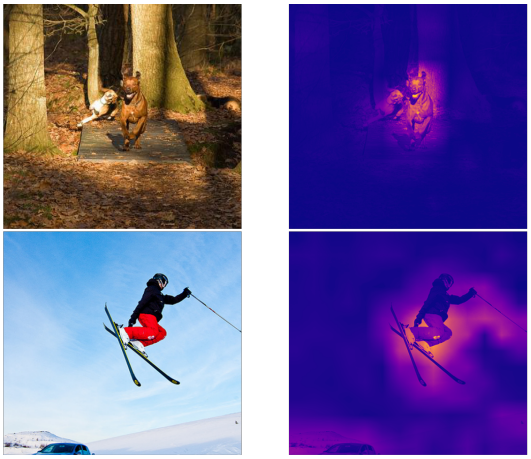

Fig. 1. Visualization the behavior of the attention mechanism in common images.

*equal contribution

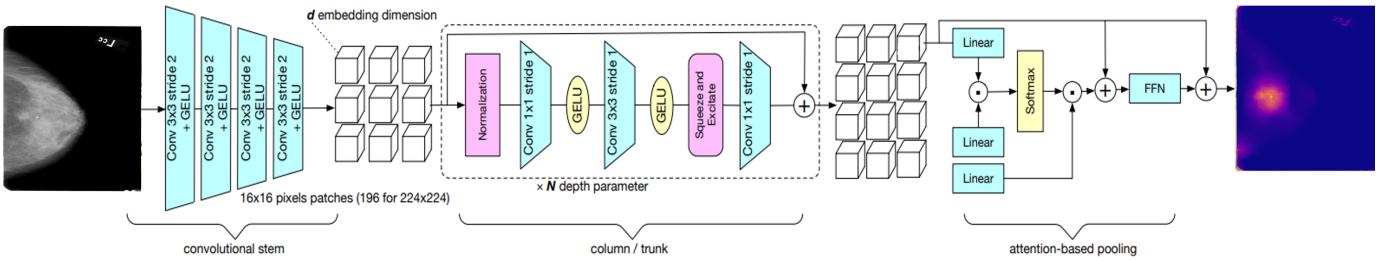

Fig. 2. The full architecture, with the convolutional stem on the left, the convolutional main block in the middle and the attention-based pooling on the right.

II. Background

The model used is the **PatchConvNet S60**: they replace the final average pooling by an attention-based aggregation layer that weights how each patch is involved in the classification decision. Moreover, using a patch-based convolutional network parametrized by only 2 parameters (width and depth), it is possible to maintain the input patch resolution across all the layers. It is very important in our context because we must avoid loss of resolution. Furthermore, they show that the S60 model is less complex than the popular ResNet-50 [**14**], using four different metrics to measure the performance: throughput, FLOPs, number of parameters and peak memory usage. In addition to that, they show that S60 model has higher classification accuracy on ImageNet [**15**] than ResNet-50. Moreover, they test this architecture on segmentation and detection obtaining outstanding result using ADE20k [**16**] and COCO [**17**] datasets.
Jetley et al. [**18**] discovered that attention improves performance on multiclass classification by 7% on CIFAR-100: they use an internal metric that is intended to have a high value when the image patch contains parts of the dominant image category. Heat-maps are useful to visualize results because it helps human to understand if the neural network is looking at the correct portion of the image, or if it is totally wrong. The paper [**19**] explains how the attention mechanism works: instead of paying attention to the last state as is usually done with RNNs, it extracts information from the whole sequence, doing a weighted sum of all the past states. This allows assign greater weight or importance to a certain element of the input for each element of the output. Eventually the authors go further using the input vector in three different ways: the Query, the Key and the Value in order to obtain even better results. This attention mechanism have a key role in the generation of heat-maps.
In Fig. 1 it is possible to see an application of this mechanism on common images: it is clear that the network is highlighting the correct patch of the images.

Our goal is to use this powerful approach to the CBIS-DDSM dataset. The images are divided into *positive*, if there is a cancer, and *negative* otherwise. With this in mind, we generate the heat-maps corresponding to each image: the patch which is responsible for the classification will be highlighted and this will be also the one where the cancer is located in the positive images. This could give a huge support to the community, avoiding the problem of finding the portion of the breast affected by cancer.

III. Methods

As mentioned before, the dataset used is CBIS-DDSM,which is an updated and standardized version of the Digital Database for Screening Mammography (DDSM): it contains 6,775 case studies, but only 1,566 partecipants. This is because the dataset is structured such that each participant has multiple patient IDs. Moreover, the ROI annotations for the abnormalities in the DDSM were provided to indicate a general position of lesions, but not a precise segmentation for them. The CBIS-DDSM collection addresses that problem by releasing a standardized version of the DDSM. Inside the dataset all images have different dimension. Because of that, first of all, we decide to resize the images to 224x224 dimension also to be consistent with the results obtained by the architectures in [**12**]. In addition to that, we opt for patches of dimension 16x16: in this way we split the image and generate 196 different patches which will contribute in the creation of the final heat-map.
We performed some preliminary experiments, in order to understand the best combination between opimizers and learning rate (lr). In the end, we decide to use Lamb [**20**] optimizer, a variant of AdamW [**21**]. The principle is the layerwise adaptation strategy, used in order to accelerate training of deep neural networks using large mini-batches. We train for 25 epochs fixing the value of the lr to $5x10^{-4}$. In addition to that, we use the pre-trained model located in [**22**], obtained by the authors in [**12**] with ImageNet 1k, to achieve better result.

IV. Results

In this section we present our main experimental results. Moreover, we include also a visual comparison between original image, ROI (which we overlap to the original image in order to understand better the corresponding portion of the image) and the heat-map generated by the architecture.
Before visualizing the images, we wanted to know if our approach was good or not. In Fig. 3 is possible to see the behavior of the loss and accuracy across the epochs. From

the picture is possible to figure out that the validation loss has a small value for all the epochs and stops around 0,7. On the other hand, the validation accuracy is more or less stable around 70. It is a nice result for this task.

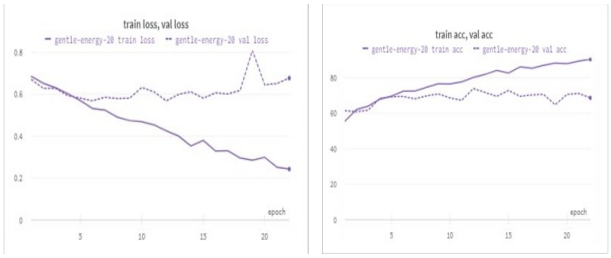

Fig. 3. On the left there is the plot of the loss; on the other side, there is the accuracy plot.

Moreover, we also used another performance metric to validate our results: the Area Under the ROC Curve (AUC). A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a model. This curve plots two parameters: True Positive Rate (TPR) and False Positive Rate (FPR). A ROC curve plots TPR vs. FPR at different classification thresholds. As the name suggest, AUC measures the entire two-dimensional area underneath the whole ROC curve and it ranges in value from 0 to 1. AUC is desirable for the following two reasons: it is scale invariant and it is classification-threshold-invariant. In our training, we obtain **0,73** of AUC.
In addition to that, we generate two different types of heat-maps: one for positive and one for negative carcinogenic images. In Fig. 4 is possible to see three different non carcinogenic images: on the left there is the original image, in the middle there is the ROI which is overlapped in the original image and on the right the corresponding heat-map.

The results are interesting: as we expect in the heat-maps there is no big change in colors. This is motivated by the fact that there is no cancer in this images.
On the other hand, in Fig. 5 are shown three different positive images: as in the previous case, on the left there is the original image, centered there is the ROI overlapped in the original image, and on the right the generated heat-map.

The results are astonishing: in all three pictures the colors are not uniformly distributed. In fact the corresponding ROI for each image, which is unknown for the model and it is an item to validate our results, is highlighted. Moreover, only the patch that is similar with the ROI is highlighted, meaning that the cancer is probably located in that place.

V. Conclusion

In this paper we introduce transformers in medical image analysis. We use a patch based model [**12**] applied to CBIS-DDSM dataset. Using a lr of $5x10^{-4}$, we trained for 25 epochs. We obtained a value of 0,73 for the AUC, which is

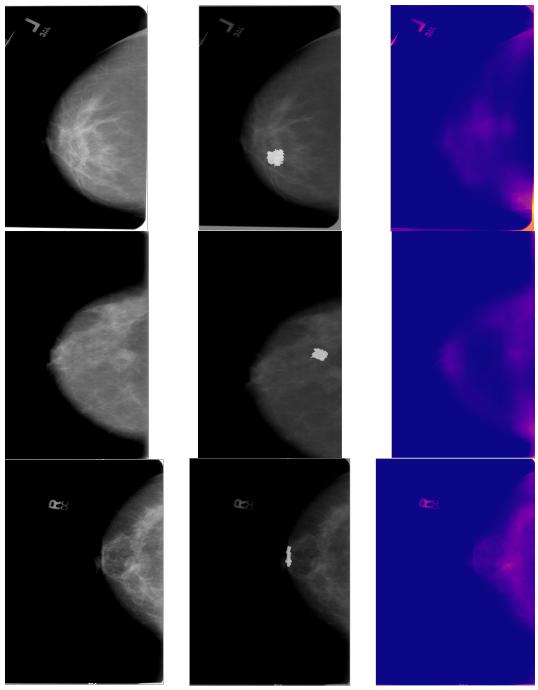

Fig. 4. Non-carcinogenic images: on the left there is the original image, in the middle there is the ROI and on the right the corresponding heat-map

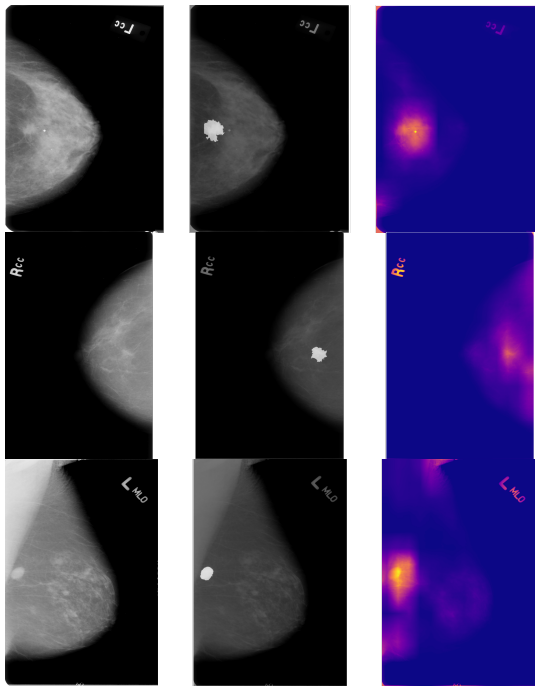

Fig. 5. Three different positive images: on the left there is the original image, centered there is the ROI and on the right the generated heat-map.

good for this task. In addition to that, our main goal was to generate good heat-maps for each image, highlighting which part is affected by cancer in the positive images. The results were remarkable: in fact for positive images the heat-maps highlight correctly the patches corresponding to the ROI. On the other hand, in negative images, colors were uniformly distributed meaning that no cancer was found.

We believe that this approach could be the next big thing in medical imaging, but a more sophisticated approach is requested: it is necessary to compare the AUC of different patch-based models with state of the art models in medical imaging, in order to understand if these models are robust enough compared to them. We hope that our work will inspire other researcher to continue with these approach, because it could help a lot the doctors to read the images in an appropriate way.

REFERENCES

[1] https://www.salute.gov.it/portale/tumori/dettaglioContenutiTumori.jsp?lingua=italiano&id=5538&area=tumori&menu=vuoto. Accessed: 2022-05-27.

[2] https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Accessed: 2022-05-27.

[3] https://www.breastcancer.org/facts-statistics. Accessed: 2022-05-27.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[5] Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen, "Transformers in Medical Image Analysis: A Review", 2022.

[6] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis", 2022.

[7] Ali Hatamizadeh, Yucheng Tang et al. "UNETR: Transformers for 3D Medical Image Segmentation", 2021.

[8] Behnaz Gheflati and Hassan Rivaz, "Vision transformers for classification of breast ultrasound images", 2022.

[9] Yuhao Mo, Chu Han et al., "HoVer-Trans: Anatomy-aware HoVer-Transformer for ROI-free Breast Cancer Diagnosis in Ultrasound Images", 2022.

[10] Zhu He, Mingwei Lin, Zeshui Xu, Zhiqiang Yao, Hong Chen, Adi Alhudhaif, Fayadh Alenezi, "Deconv-transformer (DecT): A histopathological image classification model for breast cancer based on color deconvolution and transformer architecture", 2022.

[11] Xuxin Chen, Ke Zhang, Neman Abdoli, Patrik W. Gilley, Ximin Wang, Hong Liu, Bin Zheng and Yuchen Qiu, "Transformers Improve Breast Cancer Diagnosis from Unregistered Multi-View Mammograms", 2022.

[12] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Piotr Bojanowski, Armand Joulin, Gabriel Synnaeve, Hervé Jégou. Augmenting Convolutional networks with attention-based aggregation, 2021.

[13] https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM. Accessed: 22-06-06.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition, 2015.

[15] Jia Deng, Wei Dong, Richard Socher, Lia-Jia Li, "ImageNet: A large-scale hierarchical image database", 2009.

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. Conference on Computer Vision and Pattern Recognition, 2017.

[17] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan et al. Microsoft coco: Common objects in context. In European Conference on Computer Vision, 2014.

[18] Saumya Jetley, Nicholas A. Lord, Namhoon Lee Philip H. S. Torr, "Learn to pay attention", 2018.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention is all you need", 2017.

[20] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh, "Large batch optimization for deep learning: Training BERT in 76 minutes", 2020.

[21] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam", 2017.

[22] https://github.com/facebookresearch/deit/blob/main/README_patchconvnet.md. Accessed: 2022-06-13