Multi-task learning to improve performance consistency in mammogram classification

Mickael Tardy[1,2] MICKAEL.TARDY@EC-NANTES.FR

Diana Mateus[1] DIANA.MATEUS@EC-NANTES.FR

[1] *Ecole Centrale de Nantes, LS2N, UMR CNRS 6004, Nantes, France*
[2] *Hera-MI, SAS, Nantes, France*

Editors: Under Review for MIDL 2022

Abstract

Breast cancer is the most prevalent cancer amongst women. Its regular screening, often based on mammograms, significantly reduces the mortality. Deep learning has shown good performances in coping with screening-generated imaging data, however there are still open questions related to the imbalance, noisiness, and heterogeneity of the data. We propose to address these challenges with Multi-Task Learning, combining tasks such as classification, regression, segmentation, and reconstruction. Our approach allows to obtain consistent performances of AUC $\approx$ 0.80 across different vendors (including those unknown during training) on the primary breast cancer classification task, while fulfilling well secondary tasks including an F_1 score of 0.96 on 4-class vendor classification, and F_1 score of 0.64 on 4-class density classification.

Keywords: Breast cancer, classification, multi-task learning, mammography

1. Introduction

Breast cancer remains one of the most prevalent types of cancer creating a serious public health concern (Siegel et al., 2021). Luckily, the mortality rate is comparatively low, given its prevalence, thanks to the advances in screening and diagnostic tools. Indeed, when detected early, appropriate treatment can be advised to allow for recovery. The screening often starts with a clinical exam followed by mammography (X-ray imaging exam). Mammograms are interpreted by radiologists, who look for early signs of cancer in the breast. As cancer awareness raises, and screening programs become a usual practice in developed countries, a substantial amount of mammography imaging is generated every year. Large image datasets are an ideal testbed for developing and training deep-learning-based algorithms. However, large amounts of mammography data from screening bring other challenges, such as high-resolution images, class imbalance, label noise, and an overall lack of precise labels, making the brute force deployment of Convolutional Neural Networks unsuitable for the task.

2. Method and Experiments

Some customized techniques have been proposed to deal with the aforementioned challenges, including weakly supervised training (Shen et al., 2020), data synthesis (Wu et al., 2018), or multi-view approaches (Yang et al., 2021), etc. In our case, we propose to address

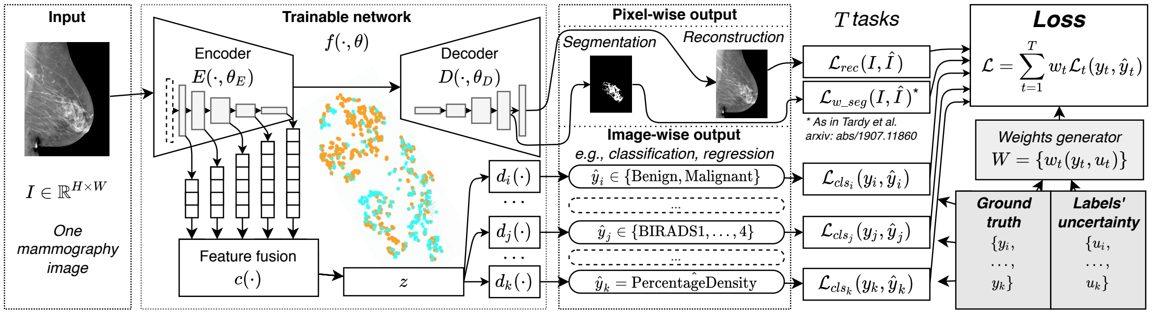
Input
$I \in \mathbb{R}^{H \times W}$
One mammography image
Trainable network
$f(\cdot, \theta)$
Encoder
$E(\cdot, \theta_E)$
Decoder
$D(\cdot, \theta_D)$
Feature fusion
$c(\cdot)$
z
$d_i(\cdot)$
$d_j(\cdot)$
$d_k(\cdot)$
Pixel-wise output
Segmentation
Reconstruction
Image-wise output
e.g., classification, regression
$\hat{y}_i \in \{\text{Benign, Malignant}\}$
$\hat{y}_j \in \{\text{BIRADS1}, \dots, 4\}$
$\hat{y}_k = \text{Percentage}\hat{\ }\text{Density}$
T tasks
$\mathcal{L}_{rec}(I, \hat{I})$
$\mathcal{L}_{w\_seg}(I, \hat{I})^*$
* As in Tardy et al. arxiv: abs/1907.11860
$\mathcal{L}_{cls_i}(y_i, \hat{y}_i)$
$\mathcal{L}_{cls_j}(y_j, \hat{y}_j)$
$\mathcal{L}_{cls_k}(y_k, \hat{y}_k)$
Loss
$\mathcal{L} = \sum_{t=1}^{T} w_t \mathcal{L}_t(y_t, \hat{y}_t)$
Weights generator
$W = \{w_t(y_t, u_t)\}$
Ground truth
$\{y_i, \dots, y_k\}$
Labels' uncertainty
$\{u_i, \dots, u_k\}$

Figure 1: Overview of the proposed method: for a given mammography image, an auto-encoder-like neural network is trained with multiple objectives.

the mammogram classification problem through multi-task learning, coping with several of the above challenges at once (Tardy et al., 2022). The proposed method is illustrated in Figure 1. We aim at higher and more consistent performances by training a neural network simultaneously with multiple objectives, and thereby allowing the network to learn more informative features from a given sample.

Our multi-task learning approach is inspired by clinical screening practices. When radiologists interpret a mammography exam, for each breast, they analyze two views acquired from different angles, looking for similarities amongst them. In addition, they evaluate the overall distribution of the parenchymal tissue and its symmetry in the two breasts, with dense breasts being more exposed to a risk of developing cancer. In this sense, the incidence and density provide additional information for the interpretation, and presumably a trainable algorithm can also benefit from this information to enhance the overall knowledge. Fortunately, these labels (density, view angle) are generally available from clinical reports, allowing to avoid the expensive annotation burden. Finally, it has been shown, that learning relevant features is also possible from an unsupervised reconstruction task with an auto-encoder-like setup (without skip connections between the encoder and decoder). Combining several objectives also allows reducing the influence of the noise of each isolated label, with cancer classification being usually the noisiest.

We argue that learning several objectives from the input samples yields higher and more consistent performances on the main task of cancer classification. Moreover, multi-task learning allows completing secondary clinically relevant objectives, such as, for example, breast density estimation, which is also beneficial from a computational speed standpoint, as the algorithm is capable of providing several predictions in one forward pass.

Compared to our original work (Tardy et al., 2022), we further experimented with complementary outputs, such as percentage density regression (based on BI-RADS 4th edition), breast normality classification (based on ACR classification), dense tissue distribution segmentation as in (Tardy et al., 2019), and vendor classification. We observe that these objectives do not decrease the performance of the primary breast cancer classification task, while the network also performs well on the additional tasks, achieving an F1 score of 0.95 on vendor classification and an MAE of 14% on percentage density estimation.

3. Results and Discussion

On the breast cancer classification task we obtain a consistent AUC of ≈ 0.80 with several datasets, including in total (6) different vendors. We underline that the IMS Giotto images[1] were under-represented in training, and that the Siemens images (i.e., INB) were not included in training at all. Combining all datasets, we achieve F_1 score of 0.60 for 4-class density classification, and a Mean Absolute Error (MAE) of 14% for percentage density regression. Our method has an advantage of being trainable with only some of the objectives, allowing efficiently using available data without the need for the expert annotations.

Our work has some limitations, opening paths for future research. We see only a slight performance increase when introducing noisy and imbalanced datasets such as VTB. More advanced losses and sample balancing schemes might further improve the feature extractor network. We used prior knowledge about the training dataset to scale the losses, which could be both, restrictive and unreliable. A different approach, including, for example, pseudo-labeling, could be adopted. From the engineering standpoint, we may envision more advanced networks designs. We explored an extension of the bottleneck, replacing one-dense-layer networks with larger ones (adding 64-neurons-layers in our case), and achieved comparable performances, when training on the same dataset which illustrates the potential of reduced overfitting even with more complex networks.

4. Disclaimer

This short paper discusses essentials of the work published in (Tardy et al., 2022).

References

Yiqiu Shen et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *arXiv preprint arXiv:2002.07613*, feb 2020. URL http://arxiv.org/abs/2002.07613.

Rebecca L. Siegel et al. Cancer Statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71 (1):7–33, jan 2021. ISSN 0007-9235. doi: 10.3322/caac.21654.

Mickael Tardy et al. A closer look onto breast density with weakly supervised dense-tissue masks. *arXiv*, jul 2019. ISSN 23318422. URL http://arxiv.org/abs/1907.11860.

Mickael Tardy et al. Leveraging Multi-Task Learning to Cope With Poor and Missing Labels of Mammograms. *Frontiers in Radiology*, 1:19, jan 2022. ISSN 2673-8740. doi: 10.3389/fradi.2021.796078.

Eric Wu et al. Conditional infilling GANs for data augmentation in mammogram classification. Technical report, 2018. URL https://arxiv.org/pdf/1807.08093.pdf.

Zhicheng Yang et al. MommiNet-v2: Mammographic multi-view mass identification networks. *Medical Image Analysis*, 73, oct 2021. ISSN 13618423. doi: 10.1016/j.media.2021.102204. URL https://pubmed.ncbi.nlm.nih.gov/34399154/.

[1] IMS Giotto images were not part of our original work (Tardy et al., 2022)