Improving mammogram classification by generating artificial images

Ana Peterka†
University of Ljubljana,
Faculty of Computer and
Information Science,
Ljubljana, Slovenia
anapeterka1151@gmail.com

Zoran Bosnić
University of Ljubljana,
Faculty of Computer and
Information Science,
Ljubljana, Slovenia
zoran.bosnic@fri.uni-lj.si

Evgeny Osipov
Luleå University of Technology,
Department of Computer Science,
Electrical and Space Engineering,
Luleå, Sweden
evgeny.osipov@ltu.se

ABSTRACT

Training a deep convolutional neural network (DCNN) from the scratch is difficult, because it requires large amounts of labeled training data. This is a big problem especially in the medical domain, since datasets are scarce and data is often imbalanced. This can result in overfitting the model. Fine-tuning a model that has been pre-trained on a large dataset shows promising results. Another approach is to augment the dataset with artificially generated learning examples. In this paper, we augment the learning set with artificially generated images that are produced by conditional infilling GAN. The results that we obtained show that we can relatively easily generate realistically looking mammograms that improve the classification of benign and malignant mammograms.

KEYWORDS

data augmentation, transfer learning, CNN, ResNet-50, GAN, ciGAN

1 Introduction

Breast cancer is a cancer that is found in the tissue of the breast, when abnormal cells grow in an uncontrolled way. It can affect both women and men, though it is prevalent in women. Statistics show that it has the highest mortality rate of any cancer in women worldwide and that 1 in 8 women in the EU will develop breast cancer before the age of 85[1]. Screening mammography helps diagnose cancer at an early stage, which significantly increases the survival rates. However, the evaluation of mammograms performed by doctors and radiologists is tedious, lengthy and error prone, as it results in a high number of false positives.

New approaches in deep learning (DL), in particular convolutional neural networks (CNNs), have proven their potential for medical imaging classification tasks. This could relieve radiologists and give patients quicker and more accurate diagnosis. However, the performance of CNNs are dependent on large labeled datasets, which are hard to obtain in the medical imaging field due privacy concerns of the patients and the time consuming expert annotations. Furthermore, the data is often imbalanced, meaning that pathologic findings are relatively very rare. This can result in overfitting the model and bad generalization ability.

So far, this problem has been addressed with transfer learning and data augmentation techniques. In this paper, we evaluate these techniques on the CBIS-DDSM dataset, which is a publicly available dataset that contains benign and malignant mammograms. We propose a novel approach of generating new images with Generative Adversarial Networks (GANs) combined with traditional data augmentation, such as horizontal flipping, rotations etc., and evaluate if increasing the dataset helped to achieve better classification. We also test if fine tuning a ResNet-50 model helps improve the results.

The paper is structured as follows. Section 2 presents the related work, Section 3 describes the data augmentation techniques used, Section 4 the training process, Section 5 the evaluation metrics used and the results, and in Section 6 we state our conclusions and discuss the prospective future work.

2 Related Work

This section provides a brief review of past work that falls down to three categories:

1. improved classification with traditional data augmentation,
2. improved classification with generating synthetic images using generative adversarial network,
3. transfer learning and fine tuning.

The problem with small datasets, especially in the medical domain, is that models that are trained on them tend to overfit the data. There are a lot of approaches to reduce it, like batch normalization, dropout, data augmentation and also transfer learning. Traditional data augmentation based on affine transformations, such as translation, rotation, shearing, flipping and scaling, is the most widely used and very easy to implement. They are ubiquitous in computer vision tasks and show very promising results [1]. However, they do not bring any new visual features that could additionally improve the generalization of the CNN.

Synthetic image generation with GANs enables more variability to the dataset and further improves robustness of the

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

[1] https://www.europadonna.org/breast-cancer-facs/

classification network. GANs were inspired by game theory, where two neural networks are pitted against each other using a minmax strategy. They were first introduced in [2], and they have recently been applied to many different medical imaging applications, mostly for image to image translation and image inpainting. In [3], the authors used conditional infilling GAN to synthesize lesions on mammograms.

Transfer learning and fine tuning for mammography medical images was the main topic in [4] and [5]. In [4], they demonstrated that a whole image model trained on DDSM can be easily transferred to INbreast without using its lesion annotations and using only a small amount of training data. In [5], the authors showed that fine tuning ResNet-50 model pre-trained on ImageNet can be used to perform tumor classification in CBIS-DDSM dataset.

In this paper, we will first use traditional data augmentation techniques and later additionally augment the dataset with applying the ciGAN (conditional infilling GAN). We will evaluate the improvements with a fine tuned ResNet-50 model.

3 Augmenting the dataset

In this section, we first describe the dataset, then we explain the traditional data augmentation methods used and a GAN method for synthesizing new images.

3.1 The CBIS-DDSM dataset

CBIS-DDSM [6] is a publicly available dataset that contains digitized images from scanned films of mammogram images and it is a subset of the DDSM dataset that consists of only benign and malign cases. The data was acquired from 1566 patients and it contains both mediolateral oblique (MLO) and craniocaudal (CC) views of each breast. Images are grayscale, and they have corresponding binary masks that indicate mass and ROI images of that mass. Images are in DICOM format, which is the standard for medical imaging information. The data is already split in the training and testing set. We used a part of the testing set as a validation set for the classification network.

3.2 Traditional data augmentation

To compensate for the lack of training images, we used classical data augmentation techniques, in particular horizontal flipping, rotations of up to 30°, and zoom range from 0.75 to 1.25 and test if this improved the performance of the CNN.

3.3 Data augmentation with GANs

To further augment and balance the dataset, we use a GAN variant, called conditional infilling GAN (ciGAN) [3]. GANs are a type of generative models, which means they are able to produce novel examples, based on the training data. They consist of two neural networks, a generator and a discriminator, which are pitted against each other. Generator tries to capture the data's distribution while the discriminator tries to distinguish real and generated examples. By training them simultaneously, the generator will get better at generating realistic data, while the discriminator gets better at distinguishing real and fake data. In the case of ciGAN, the generator is based on a cascaded refinement network (CRN) [8], where features are generated at multiple scales before being concatenated, which yields a more realistic image synthesis.

In our approach, we apply the ciGAN to sample a location on a healthy mammogram and then synthesize a lesion in its location, as shown in Figure 1. The input is a concatenated stack of:

- a corrupted image (one channel grayscale image with lesion replaced by uniform distribution of values between 0 and 1),
- a binary mask that marks lesion (1 representing the location of the lesion, and the zeros elsewhere), and
- the class label ([1,0] representing the non-malignant class, and [0,1] representing the malignant class).

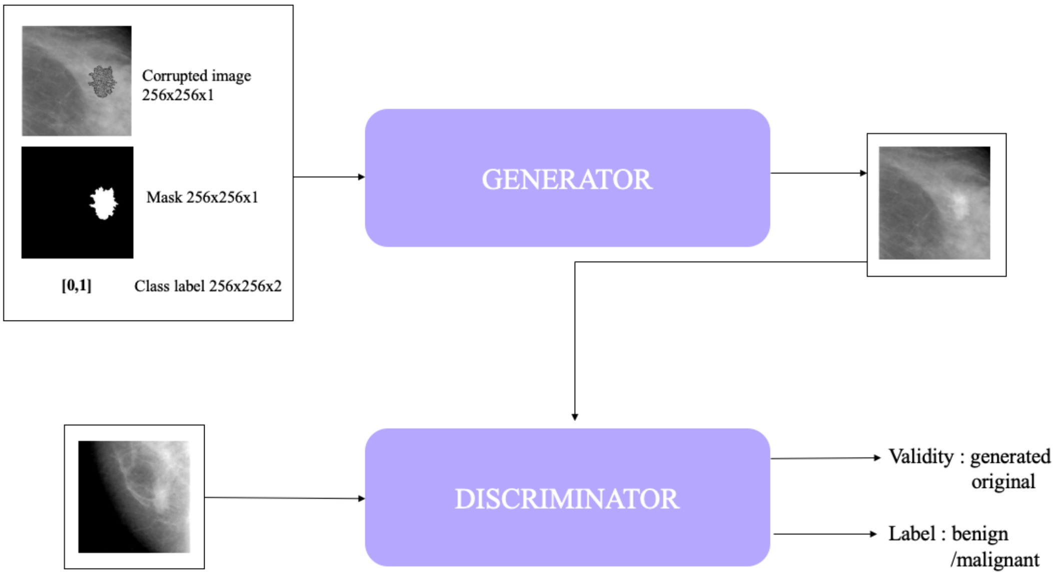

Figure 1: The ciGAN architecture. The input consists of two one channel images, and 2 class channels for indicating malignant/benign label. Output of the generator is, together with the real image fed into the discriminator, which predicts whether each image is either generated or original and also whether the image contains benign or malignant lesions.

The generator is comprised of multiple convolutional blocks. The first convolutional block receives input stack, downsampled to the 4x4 resolution. Resolution is doubled between consecutive blocks. So the next convolutional block is fed with concatenation of the output from the first layer, upsampled to the 8x8 and an input stack resized to 8x8. This is repeated until resolution of 256x256 is obtained. The discriminator has similar, but inverse structure.

4 Generating artificial images

4.1 Preprocessing

To extract patches of 256x256 pixels that are fed into ciGAN, we used a sliding window technique. The program loops through the whole mammogram image with the stride of 128 and checks if the rectangular region overlaps the majority of the breast. It also checks whether the patch contains lesion or it shows only normal breast tissue, and labels it accordingly. This is done by comparing the same region of the corresponding binary mask. At the end the patch dataset contains 5466 images, 1743 of them are normal, 2198 benign and 1525 malignant.

After acquiring a dataset of patches, the program loops through all the patches containing only normal tissue. For each normal patch, it randomly chooses one patch that contains a lesion. The patch with lesion is then randomly zoomed in/out by a small factor, to obtain more diverse masses. Next, we check whether on the same location as is lesion, on the normal patch, is only breast tissue and not background. If not, the next random lesion patch is chosen and the whole process is repeated until a suitable match is found.

Once there is a suitable pair obtained, the normal image is corrupted, by replacing the area defined by the mask of the lesion with uniform distribution.

4.2 Loss functions

The ciGAN model is trained by utilizing three loss functions [3]:

- Perceptual loss: is a loss calculated between the ground truth and the output image. But unlike a per-pixel loss, which is based on differences between pixels, it measures the discrepancy between high-level perceptual features extracted from pretrained networks [10]. It encourages the generator to output images with similar high level features as the original image. In this case, the VGG-19 [11] convolutional neural network is used, pretrained on the ImageNet dataset. It is defined as

$$L(R,S) = \sum_l \parallel \phi(L)_l - \phi(S)_l \parallel_1$$

  where R denotes a real image, S a synthetic image and ϕ a feature function;

- Boundary Loss: is used to encourage smoothing between infilled components and the context of the generated image. It is a L1 difference between the real and generated images at the boundary and defined as

$$B(R,S) = \parallel w \odot (R - S) \parallel_1$$

  where *w* denotes the mask with Gaussian filter of standard deviation 10 applied, and $\odot$ is the element wise product;

- Adversarial Loss: is the general GAN loss. It is defined as a distance between the true and the generated distribution at the current iteration. Its goal is to converge to the equilibrium in the minmax game between generator G and discriminator D, as follows:

$$\min_G \max_D L(D,G)$$
$$L(G,D) = E_{c,R}[\log D(c,R)] + E_R[\log(1 - D(c,S)]$$

  where c denotes the class label.

4.3 Training

The ciGAN is first pretrained on VGG-19 loss for 300 epochs. Then the training of discriminator and generator are alternating, when loss for either drops below 0.3 for additional 2000 epochs. The ciGAN produces realistic images as shown in Figure 2.

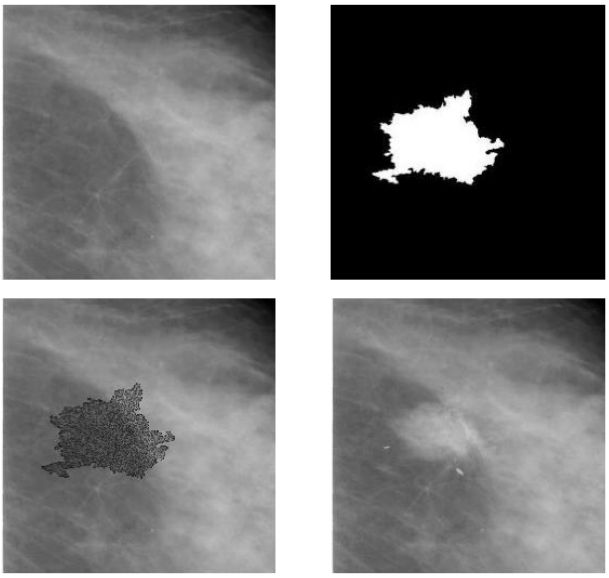

Figure 2: A generated sample from ciGAN. The first image is the normal image without a lesion, the second one is the binary mask representing the random malignant lesion, the third one is the corrupted image and the last one is the synthesized image with malignant lesion.

5 Evaluation and results

For evaluation of results three metrics were used. The first one is accuracy, which tells us how many examples were correctly classified. The second one is recall/sensitivity, which is the fraction between true positives and the sum of true positives and false positives. It is the most important metric in this case, due to the risk of overlooking cancer. The third one is Area Under Curve (AUC), which measures area under the ROC curve.

We evaluate the results by performing 4 experiments:

1. **Shallow CNN** [12]: we implement it as the baseline. The network is fed a patch and classifies it as either malignant or benign. It consists of three convolutional blocks, composed of 3x3 Convolutions, Batch Normalization, ReLU activation function and Max Pooling, followed by

three Dense layers, and softmax function for binary classification.

2. **ResNet-50**: we classify the data using a fine tuned ResNet-50 [13] model to see if transfer learning improves the results.
3. **ResNet-50 + Traditional** data augmentation,
4. **ResNet-50 + Traditional** data augmentation and generated **artificial** images.

As mentioned in [5], we fine tuned the Resnet-50 [12] model with ImageNet weights. It is an extremely deep neural network with 150+ layers and consists of convolutional layers, pooling layers and multiple residual blocks. In the residual blocks, the layers are fed into the next layer and also directly into the layers about two to three hops away. The input to the ResNet-50 model is a patch of a size 224x224x3. Since mammograms have only grayscale channels, the color information is copied over all three channels. We used the Adam optimizer with an initial learning rate of 10^{-5}, $\beta1 = 0.9, \beta2 = 0.999, e = 10^{-8}$ and ImageNet weight initialization. We trained it for 50 epochs with batch size of 32 and a 0.9 learning rate decay every 30 epochs.

Table 1 shows the obtained results. We can see that already using only fine tuning using ResNet-50 improved the results. After combining ResNet-50 with traditional data augmentation, we obtained even better performance metrics. Nevertheless, by increasing the dataset with relatively small amounts of synthetic images while simultaneously balancing it, we improved accuracy and AUC even more, but obtaining a slight decrease in the recall.

Table 1: The obtained accuracy, recall and AUC scores

	accuracy	recall	AUC
Shallow CNN	0.57267	0.44810	0.54943
ResNet-50	0.60155	0.55769	0.59443
ResNet-50 + traditional	0.67132	0.64231	0.66666
ResNet-50 + traditional + artificial	0.76145	0.61538	0.71638

6 Conclusion

In this paper we discussed overcoming the obstacle of small and imbalanced mammography dataset. We proposed an approach for artificial generation of images that are produced by a conditional infilling GAN (ciGAN). The results showed that we can relatively easily generate realistically looking mammograms that improve the classification of benign and malignant mammograms. Further, we evaluated the learning performance when using fine-tuning, classical data augmentation and synthetic examples. The results showed that each of these techniques improved classification, yielding the best results using all three together.

Testing these methods on different medical datasets shall be the subject of future work. As well, one may consider using these methods on bigger data sets and improve the current state of the art algorithms. Since the ciGAN's discriminator was also conditioned on class, we intend on extracting its features and using it for classification on other mammography dataset, for example on the INBreast dataset. We also plan on adding more synthetic images to the dataset, to see if we can further improve the classification.

Currently, the mammogram classification is performed by the doctors and radiologists, but we hope that improving the classification with the use of machine learning combined with these and similar techniques could relieve them of such tasks in the near future.

REFERENCES

[1] Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit, 11*.

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).

[3] Wu, E., Wu, K., Cox, D., & Lotter, W. (2018). Conditional infilling GANs for data augmentation in mammogram classification. In *Image Analysis for Moving Organ, Breast, and Thoracic Images* (pp. 98-106). Springer, Cham.

[4] Shen, L. (2017). End-to-end training for whole image breast cancer diagnosis using an all convolutional design. *arXiv preprint arXiv:1711.05775*.

[5] Agarwal, R., Diaz, O., Lladó, X., & Martí, R. (2018, July). Mass detection in mammograms using pre-trained deep learning models. In *14th International Workshop on Breast Imaging (IWBI 2018)* (Vol. 10718, p. 107181F). International Society for Optics and Photonics.

[6] Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, *4*, 170177, https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM.

[7] Odena, A., Olah, C., & Shlens, J. (2017, July). Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning* (pp. 2642-2651).

[8] Chen, Q., & Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 1511-1520).

[9] Johnson, J., Alahi, A., & Fei-Fei, L. (2016, October). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694-711). Springer, Cham.

[10] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[11] Lévy, D., & Jain, A. (2016). Breast mass classification from mammograms using deep convolutional neural networks. *arXiv preprint arXiv:1612.00542*.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).