école normale supérieure paris—saclay

ENS Paris-Saclay

Mathématiques Vision Apprentissage

Introduction to Medical Imaging Analysis

Authors:
de LAVALLADE Pauline
MAHE Sylvanus
MAYARD Hippolyte

Advisor:
DELINGETTE Hervé
PENNEC Xavier

December 2020

Contents

1 Abstract

With the exponentional growth of deep learning methods and range of aplications, we also see a considerable increase in cyber attacks against those methods. Healthcare data is now almost always stored on computers and keeping such sensitive data safe is becoming more challenging. Hence, Becker et al. studied the efficiency of these attacks on mammography images using Cycle Generative Adversarial Networks and if they are detectable by the experienced human eye of radiologists. Here we will critically analyse how the authors simulated attacks on medical images and if these attacks were detectable. We then show how we replicated their experiment and we propose further work.

2 Introduction

The emergence of Deep Learning has opened a large range of solutions for Artificial Intelligence in Healthcare. However, as the Deep Learning scientific community grows, another field is also expanding which goal is to identify the weaknesses of Deep Learning models' weaknesses and attack them. This new research interest is called Adversarial Attacks. In [1], Goodfellow and al. highlighted the weaknesses of neural network architectures and challenged those weaknesses by attacking a Deep Convolutional Neural Network using Fast Gradient Signed Method (FGSM). This adversarial attack method alters specific pixels of an image which results in the misclassification of the image by the algorithm while it still appears the same to the human eye. Figure 1 below displays an example of a FGSM attack. We can see that the CNN classifies the original image as a panda with 57.7% confidence. But after the attack, the network misclassifies the image as a gibbon while we still see a panda.

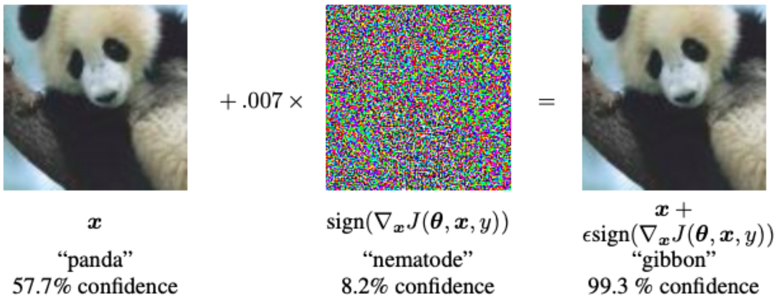
$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure 1: Fast Gradient Signed Method (FGSM), Goodfellow and al. [1]

In this perspective of adversarial attacks, we propose here to study the following paper: *Injecting and removing malignant features in mammography with CycleGAN: Investigation of an automated adversarial attack using neural networks*, Anton S. Becker, Lukas Jendele, Ondrej Skopek, Nicole Berger, Soleen Ghafoor, Magda Marcon, Ender Konukoglu [2]. This paper shows how attacks such as inserting or removing malignant features in mammography medical images, can be detected by radiologists.

The main goal behind this paper [2] is to train two Generative Adversarial Networks that inject or remove malignant features on breast mammography images in order to determine if radiologists are able to detect these attacks on the modified images. In other words, the authors aim to quantify the efficiency of these attacks at fooling radiologists. For these image alteration, the authors use a specific type of Generative Adversarial Networks (GANs) [3] that is called cycle-consitant GANs (CycleGANs) [4]. Under the hypothesis that GANs can learn an implicit representation of malignant features on mammography images, the networks will either insert or remove malignant features from a benign or malignant mammography image respectively. Further details on architecture and training of CycleGANs are given in Section 3.

Category	Description
0	Needs additional imaging evaluation and/or prior mammograms for comparison
1	Negative
2	Benign finding(s)
3	Probably benign finding(s). Short-interval follow-up is suggested.
4	Suspicious anomaly. Biopsy should be considered.
5	Highly suggestive of malignancy. Appropriate action should be taken.
6	Biopsy proven malignancy

Table 1: Breast Imaging Reporting and Data System Assessment Categories

3 Method

As mentioned above, the main objective of the paper [2] is to train a CycleGAN neural network to inject and remove cancerous structures in mammography images and see if such manipulation can be detected by experienced radiologists. Two separate experiments were carried out: one on low resolution images and the other on higher resolution images to test the generalization capacity of the architecture developed in large dimensions. The goal of this section is to describe the main components of the method used in [2]

3.1 An introduction to Generative Adversarial Networks

The emergence of Generative Adversarial Networks (GANs), introduced by Goodfellow and al. [3], has been a powerful innovation in image synthesis. They established a new approach for training an image synthesis model, based on a two–player minimax game. The Generator aims at generating realistic images while the Discriminator discriminates the fake images from the real ones. If we formalize the problem, G transforms a noise $z \sim p_x(z)$ from a prior distribution $p_x(z)$ to a realistic image synthesis. On the other hand, D tries to discriminate the images, that is to distinguish which images are from the real sample distribution $x \sim p_{data}(x)$ or the fake sample. This optimisation problem is summarised in the equation below:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_x(z)}[\log(1 - D(G(z)))] \tag{1}$$

3.2 CycleGANs

CycleGANs (Cycle-Consistent Adversarial Networks) were introduced by [4] to address the lack of twinned data in image translation problems. i.e. switching from one data distribution to another or here how to convert cancerous mammograms into healthy mammograms

and vice versa. CycleGANs enable the translation of images from a domain X to a domain Y in the absence of twinned data, e.g. they enable the mapping G of X and Y. To make this possible [4] introduced an inverse mapping F from Y to X as well as a loss of coherence so that $F(G(X)) \approx X$.
For each of the data sets X , Y [4] introduces two adversarial discriminators D_X and D_Y to complete the GANs setting (explained above). The objective function contains two types of losses, the adversarial losses of the learned mappings (e.g (2) for G), and the cycle consistency loss(3) to ensure that the Mappings F and G are indeed inverse to each other.

$$\begin{aligned}\mathcal{L}_{\text{GAN}}\left(G, D_Y, X, Y\right) = & \mathbb{E}_{y \sim p_{\text{data}}(y)}\left[\log D_Y(y)\right] \\ & + \mathbb{E}_{x \sim p_{\text{data}}(x)}\left[\log \left(1 - D_Y(G(x))\right]\right.\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{\text{cyc}}\left(G, F\right) = & \mathbb{E}_{x \sim p_{\text{data}}(x)}\left[\|F(G(x)) - x\|_1\right] \\ & + \mathbb{E}_{y \sim p_{\text{data}}(y)}\left[\|G(F(y)) - y\|_1\right]\end{aligned} \tag{3}$$

The total objective function is written:

$$\begin{aligned}\mathcal{L}\left(G, F, D_X, D_Y\right) = & \mathcal{L}_{\text{GAN}}\left(G, D_Y, X, Y\right) \\ & + \mathcal{L}_{\text{GAN}}\left(F, D_X, Y, X\right) \\ & + \lambda \mathcal{L}_{\text{cyc}}(G, F)\end{aligned} \tag{4}$$

and the corresponding maximisation problem:

$$G^*, F^* = \arg \min_{G, F} \max_{D_x, D_Y} \mathcal{L}\left(G, F, D_X, D_Y\right) \tag{5}$$

3.3 Modelling

The feature injection / removing described in [2] is part of the image translation problem. More specifically, it is part of the translation problem with unpaired data, as we do not have both cancerous and non-cancerous mammograms for the same subject.Thus [2] uses a modified version of CycleGANs [4] with two pairs of generator and discriminator to translate images of healthy subjects into cancerous images and vice versa.

3.4 Data selection and training

Experiment 1: Two public datasets were used for this first experiment: BCDR [5] and INbreast [6]. A total of 680 mammograms corresponding to 334 subjects (including 362 healthy controls). These images were then rescaled to 256×256 px and then augmented using rotation, contrast shifting and scaling to increase the size of the dataset before being sent as input to the cycleGAN architecture. The training was done on a computer with a GTX 1070 graphics processor using Adam (Adaptive moment estimation) optimizer with $\beta = 0.5$ and a learning rate $lr = 0.0002$
Experiment 2: The goal of this experiment is to access the generalization of the CycleGAN architecture the larger dimension input images. Here the size of the images increases to 512×408 px and private data (302 cancer / 590 healthy) is also added. They performed data augmentation and the training was done with GeForce GTX TITAN X/Xp GPUs cluster. The model parameters remained the same.

3.5 Readout experiment

To measure the radiologists' performances, they used metrics such as the receiver operating characteristic (ROC) curve and the area under the curve (AUC). ROC curve is a probability curve and AUC represents the degree or the measure of separability. It quantifies how accurately the radiologists can classify the images as original or modified but also how good they are at diagnosing malignant features. A high AUC means that the radiologist is good at identifying modified images and malignant features. To compare the AUC and ROC results obtained [2], they also used Delong's non parametric test [7].

4 Results

4.1 First Experiment

Becker et al., conducted 2 separate experiments to test the accuracy of their Cycle GANs algorithms on small mammography images. The experiments differed in the assessment method and in the input images' resolution.

In the first experiment, the researchers observed that the CycleGANs first started to adjust global features such as contrast and brightness and learned to remove or add glandular tissue which increased the breasts density. It then went on to detect skin-thickening features as malignant observations. Finally the CycleGANs learned to add or remove features on more local areas. This included the alterations of mass-like lesions or benign calcifications into fat and soft tissue masses. However, when the algorithm tried adding malignant looking masses with poorly defined limits, it usually did so by adding that feature on top on preexisting smaller benign features. This rendered the altered image less real to the human eye. Furthermore, Becker et al. noticed that on this first experiment the images with added features looked more real than the ones with removed features and that grid-like patterns started to appear on the images after 160k steps. These problems caused the altered images to be very identifiable to the human eye. So for this experiment, the best results were observed on images generated by a network trained less than 160k steps. Becker et al. randomly chose 30 altered images and 30 original images and presented them to 3 radiologists who only knew that some images had been altered within this set. 40 images of the set were presented in pairs, i.e. the original one and its modified version, and the last 20 images were presented alone. The 3 radiologists were asked to try and identify which images were original and which ones were altered as well as the level of malignancy of cancer features. They found that only the most experienced radiologist could identify which images contained CycleGAN modifications but his diagnostic performance decreased during the experiment compared to his control performance (Table 2). The last two radiologists did not manage to identify the CycleGAN modifications better than chance while their diagnostic performance seemed roughly unaffected.

4.2 Second Experiment

The researchers then further investigated the occurrences of artifacts in images with a higher resolution and a higher number of training steps. A higher resolution means there will be finer details in the image and a finer texture as well. This is important as the CycleGANs must now adapt to finer features when modifying images. Becker et al. hypothesised the artefacts would be easier to spot with a higher resolution. During training they observed more pronounced grid-like patterns than in the first experiment emerging at around 45-50k training iterations even thought he learning pattern seemed similar. Testing was conducted on 6 image pairs and 6 single images that were selected after 35k training steps. They wanted to analyse the occurence of different artefacts in these images. At step 70k, they selected 12 healthy images and 12 cancerous ones and like in the first experiment half of those images were modified by CycleGANs and half were presented with the original image alongside. So the 3 radiologists were presented with a total of 72 images that they had to diagnose and identify as original or modified. Because of the higher image resolutions, the radiologists could now all identify which images were modified which confirmed the authors hypothesis that artefacts were easier to identify at higher resolution (Table 3). However, the radiologists diagnostic performances dropped significantly. Finally, they concluded that the CycleGANs could not produce more artifacts at later training stages as the identification of modifications remained high for both early and late training iterations stops.

Readers	AUC originals	AUC modified	AUC CycleGANs
1	0.85	0.63	0.66
2	0.75	0.67	0.48
3	0.77	0.69	0.50

Table 2: Results of the ROC analysis for the first experiment

Readers	AUC originals	AUC modified	AUC CycleGANs early	AUC CycleGANs late
1	0.78	0.69	0.93	0.93
2	0.77	0.59	0.93	1.0
3	0.84	0.60	0.95	0.90

Table 3: Results of the ROC analysis for the second experiment

5 Experiment replication

We replicated experiment 1 from the paper but we only had access to one dataset: InBreast. This dataset contained 401 images of both mediolateral oblique (MLO) and craniocaudal (CC) views of healthy and cancerous mammography images.

5.1 Implementation of Experiment 1

We chose to try and replicate Experiment 1 from the paper to critically assess how the CycleGANs learn cancer features and can modify healthy and cancerous images. We implemented the experiment on the InBreast dataset using all images (CC and MLO views) and we classified the images with a BiRads score of 3 or more as cancerous (see Table 1) and the ones with a score of less than 3 were labeled as healthy. We could not replicate the exact experiment as we did not have access to the other dataset they use and we could not have our results analysed by radiologists. However, we could still clearly see that the network could alter images pretty well and seemed to correctly learn what kind of features are cancerous (e.g. mass, lesions). Moreover, just like in the original experiment, we observed grid like-patterns on some generated images towards the end of training (after 200 epochs). So we decided to stop the training at 200 epochs, you can see results of that experiment in Fig. 2.

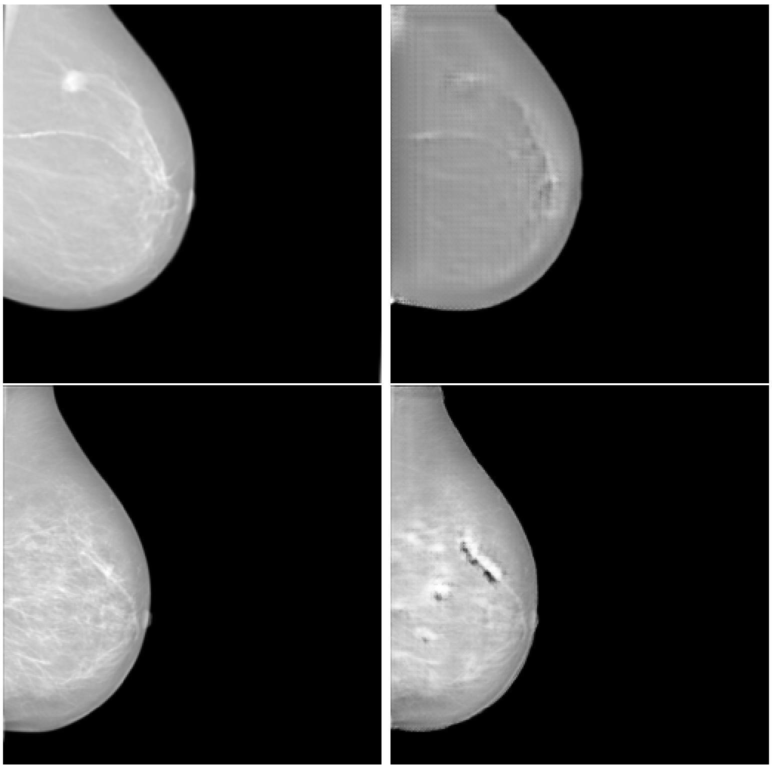

Figure 2: Original Images on the left and modified ones on the right from replication of experiment 1

5.2 Further experiment

As a further work, we chose to implement the Experiment 1 again but this time we chose a different classification argument: we decided to label all the images presenting a mass as cancerous and the ones that do not as healthy. The mass feature was stored in the same metadata as the BiRads. We chose to do so to limit the amount of malignant features the algorithm needs to learn to masses only as we hoped these images would present better results. We ran this experiment on the InBreast dataset again on both MLO and CC views. As far as we could tell, the images looked realistic and seemed to show better results than in the first experiment we ran (Fig. 3).

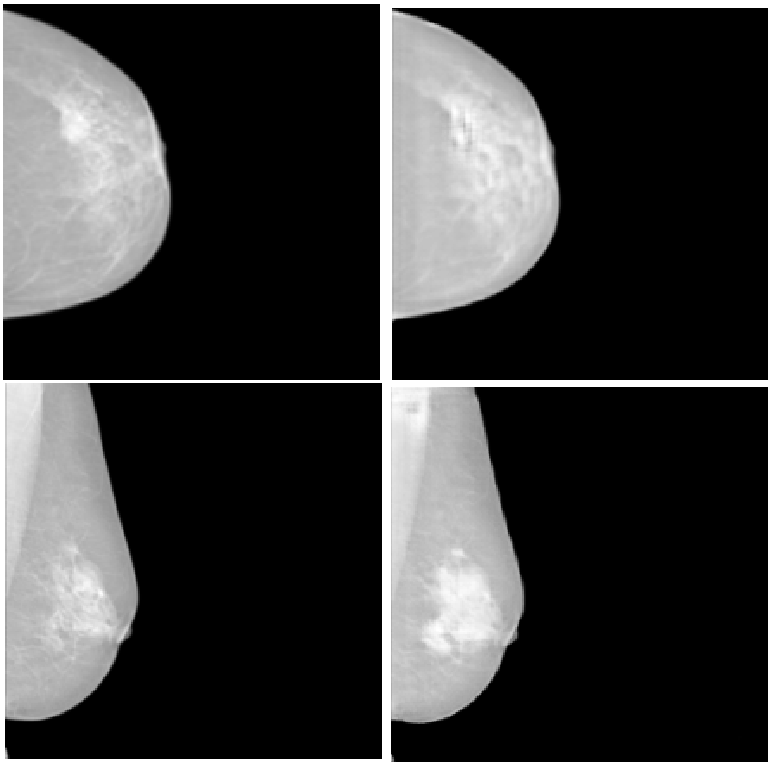

Figure 3: Original Images on the left and modified ones on the right from our further experiment

The results obtained from this experiment show that CycleGANs can learn malignant features from mammoygraohy images and modify images to change their classification from healthy to cancerous and vice verse. However, Becker et al. made assumptions at the beginning of their paper that CycleGANs can learn implicit representations of what cancer looks like and can alter images so that they are misdiagnosed by radiologists. Furthermore, in our second experiment we hypothesised that the images presenting a mass were cancerous when that is not necessarily the case. Some masses can be benign.
The code for both experiments is available at:
https://github.com/hippolytelrm/Breast_ CycleGAN_ reproduction

6 Conclusion

To conclude, Becker et al., showed that CycleGANs can efficiently inject or remove malignant features in mammographic images and with a low resolution and fool radiologists to think modified images are original. Our experiments showed similar results even though our dataset was smaller. Finally, our second experiment seemed ot exhibit better results when cancer features are limited to the presence of a mass, hence we believe that focusing algorithms to learn one malignant feature at a time would yield better results at modifying images. However, the main motivation of the study was to understand how we can detect adversarial attacks on medical imaging data and the main solution here seems to be to keep images at high resolution which can then lead to memory problems on computers. Hence further studies focusing on the detection of adversarial attacks are needed to protect such sensitive data. Future work might also include generalising this method to other medical imaging techniques such as echography, MRI or scanners.

References

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples.* 2015. arXiv: 1412.6572 [stat.ML].

[2] Anton S. Becker et al. "Injecting and removing malignant features in mammography with CycleGAN: Investigation of an automated adversarial attack using neural networks". In: *CoRR* abs/1811.07767 (2018). arXiv: 1811.07767. URL: http://arxiv.org/abs/1811.07767.

[3] Ian J. Goodfellow et al. *Generative Adversarial Networks.* 2014. arXiv: 1406.2661 [stat.ML].

[4] Jun-Yan Zhu et al. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *CoRR* abs/1703.10593 (2017). arXiv: 1703.10593. URL: http://arxiv.org/abs/1703.10593.

[5] D. Moura and Miguel Ángel Guevara-López. "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis". In: *International Journal of Computer Assisted Radiology and Surgery* 8 (2013), pp. 561–574.

[6] Inês C. Moreira et al. "INbreast: Toward a Full-field Digital Mammographic Database". In: *Academic Radiology* 19.2 (2012), pp. 236–248.

[7] David M. DeLong Elizabeth R. DeLong and Daniel L. Clarke-Pearson. "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach". In: *Biometrics* 44 (1988), pp. 837–845.