Style Transfer as Data Augmentation Technique for Breast Cancer Imaging Datasets

Filip Ryzner
MIT
ryznerf@mit.edu

Bharat Khurana
MIT
bkhurana@mit.edu

Abstract

We have investigated the use of neural network based style transfer algorithm as a data augmentation tool for breast cancer images dataset. In particular, our goal was to generate synthetic malignant breast cancer images by combining content of benign breast cancer mammography images and style of malignant breast cancer mammography images. The quality of generated images for this algorithm was found to depend on three important hyperparameters – the convolutional layer used to calculate content loss, the number of steps for the optimization process and the ratio of content weight to style weight. The label for some generated images was shown to change to "malignant" while the input images to the algorithm (copies of respective content images) had the label "benign". It was shown that the set of hyperparameters that generate realistic image can vary for different combinations of content image and style image. An autoencoder was trained to function as an anomaly detector to separate the realistic generated images from the non-realistic generated images. We trained a binary classifier model with output classes "benign" and "malignant". This classifier was used to predict the labels for the images generated by the style transfer algorithm. Certain generated images do not meet the threshold set for being conclusively classified as truly malignant. This is most likely due to the sensitivity to the set of hyperparameters, which need to be optimized for each combination of content image and style image. This is a major obstacle to upscaling the algorithm to generate large amounts of synthetic data.

1. Introduction

Deep learning algorithms for computer-aided diagnosis have attracted a great deal of attention during the last few years. Human experts use images collected via various methods to diagnose a variety of diseases. However, the unavailability of trained experts in many parts of the world leads to increased mortality from these treatable diseases due to late diagnosis. Advancements in computer-aided diagnosis present a promising opportunity to improve the delivery of healthcare to parts of the world where the availability of human expertise is limited. For example, [30] have demonstrated an algorithm consisting of a 121-layer convolutional neural network able to detect pneumonia from chest X-rays with performance better than that of practising radiologists. The algorithm was extended to the detection of 13 other diseases using chest X-rays. [13] has demonstrated a deep learning algorithm for automated detection of brain haemorrhage from computed tomography scans with accuracy comparable to radiologists. [14] have demonstrated a deep learning algorithm with high sensitivity and specificity for detection of referable diabetic retinopathy using retinal fundus photographs collected from adults having diabetes. [20] have demonstrated a deep learning approach for abnormality detection and localization in chest X-rays with high accuracy.

A commonly encountered problem during the training of classifiers based on deep neural networks is that the number of examples of some classes is much more than that of other classes in the training dataset. This problem is particularly severe for medical diagnosis as the frequency of one class (healthy person) can be 1000 times higher than the frequency of another class (patient with cancer). This difference is called class imbalance [4]. A model can achieve very high accuracy on this kind of dataset by classifying every person as healthy (the negative case). However, this model will not be informative in a real-world scenario, where a precise diagnosis is required. Several approaches addressing the problem of class imbalance have been reported, ranging from the data augmentation using transformations like changing brightness/contrast and rotating images, oversampling from the minority class, reweighting the examples from different classes to generating new images for minority classes using generative adversarial networks (GANs) or style transfer algorithms [28], [4].

The goal of our project is to investigate the use of the style transfer algorithm [11] for generating synthetic examples of an underrepresented class. Conventional approaches

mostly rely on using image transformations to augment the dataset which include rotations of images by random angles, addition of random noise to the images and changing the brightness/contrast of images. Moreover, Generative Adversarial Networks (GANs) have been extensively used for data augmentation to alleviate the problem of class imbalance in computer aided diagnosis of diseases [33]. However, GANs are known to be computationally expensive to train and have to be trained over extensive time-periods to produce good results [22].

However, [28] have reported a use the neural style transfer algorithm to ameliorate the issue of imbalance in datasets. In their study the neural style transfer algorithm, is used to generate a new example mimicking the malignant class by using an example from benign class as the content image and an example from the malignant class as the style image. The paper states that the style was successfully transferred and the generated example is claimed to have style similar to the malignant class example. Moreover, they also mention that an application of this approach is possible for the breast mammography images and provide the output of the algorithm, in which they demonstrate a change in the inner structure of the breast after the style transfer.

Therefore, our main goal is to investigate the possibilities of data augmentation for dataset balancing using the style transfer as outlined in the [28]. Specifically, we are interested in investigating the reproducibility in terms of hyperparameters and scalability of this approach as the base algorithm is much less computationally demanding than the state-of-the-art GANs. Furthermore, we investigate the role of different hyperparameters for application of this approach to augmentation of breast cancer datasets. Next, we attempt to train an autoencoder model for anomaly detection to segregate the realistic generated images from those which do not look realistic. Contrary to [28], we aim to evaluate the effectiveness of the style transfer algorithm by training a binary classifier discriminating between benign and malignant examples of mammography screenings.

2. Related Work

In this section, we conduct a comprehensive overview of literature relevant for our work. We begin by reviewing literature demonstrating the traditional approaches to adressing dataset imbalance, then move to cover style transfer and finally end with the description of applications of GANs to this problem.

Classical approaches for addressing the problem of class imbalance usually aim to make the number of examples in the two classes appear to be equal [4]. This can be done by random oversampling from the minority class, which has been found to be effective but may also lead to overfitting [12]. Another approach that has been found to be effective in some cases is undersampling from majority class in which examples from majority class are randomly removed until all classes have the same number of examples. However, a drawback of this approach is that a part of the available data is being discarded [5]. Another approach to address the problem of class imbalance is to assign weights to examples of different classes [4]. Weight assigned to examples of a class is inversely proportional to frequency of that class, resulting in equal importance to correctly classifying examples of both majority class and minority class. Additionally, certain traditional transformations can be used to augment data which include rotation, reflection, shearing, zooming-in or zooming-out, changing the brightness or contrast, sharpening, blurring or addition of Gaussian noise [28].

Moreover, GANs are currently a popular tool for medical image synthesis to alleviate the problems of class imbalance due to data scarcity and overfitting [33]. [10] have demonstrated the use of GAN to generate synthetic samples for three classes of liver lesions which when combined with real training data led to improvement of both sensitivity and specificity for lesion classification task. [3] have demonstrated generation of brain MR images using GAN which neuroradiologists found to be comparable to real ones in quality. [15] has used the GANs to synthesis and balance lung nodule dataset. Other interesting examples are covered in [33]. However training of GANs is known to be computationally expensive and somewhat unstable [22].

Finally, neural network based style transfer algorithm has been proposed augment training datasets with class imbalance as discussed by [28]. This family of algorithms aims to combine the semantic high-level information from one image with style or texture of another image. Prior to advent of neural network based style transfer algorithms a number of other approaches were used for this purpose. [8] introduced a correspondence map that included features of target image like image intensity to constrain texture synthesis. [18] used image analogies to transfer texture from an already stylized image onto a target image. [1] used the high frequency texture information from one image along with the coarse scale of the target image to achieve style transfer. Their algorithm was further improved by [25] who additionally used edge orientation information to inform texture transfer. However, all these algorithms suffer from a common limitation – they use low level image features of target image to inform texture transfer.

Ideally the algorithm should extract high level semantic information from the target image and render it in the style of source image. [11] have demonstrated that convolutional neural networks (CNNs) trained for object detection can learn representations of images that enable separation of their content information from their style information thus making it possible to combine style of one image with

content of another image. [21] demonstrated the training of feed-forward transformation networks with perceptual loss function which was able to increase the speed of style transfer algorithm drastically. However, their network was tied to a fixed set of styles and does not work for arbitrary new styles. [19] introduced an adaptive instance normalization layer which enabled high speed style transfer without being restricted to a pre-defined set of styles.

3. Methodology

In this section we will first describe the individual methods and components that are relevant for formulating our approach and then describe our approach for generating new malignant class images using the style transfer algorithm introduced by [11].

3.1. Style Transfer Fundamentals

In this section we describe the fundamentals of the neural style transfer algorithm introduced in [11], which serves as the foundation of our approach. The algorithm introduced by them uses CNNs optimised for object recognition to explicitly infer high level information from the image. The algorithm is able to infer style from one input image, mix it with a content extracted from another input image, and produce a result of high perceptual quality. The algorithm relies on the CNN architecture trained using two loss functions, namely the content loss and the style loss.

Assuming p and x are the input image (copy of the content image) and the content image respectively, and P_l and X_l are their feature representations for convolutional layer l, then the content loss is defined as the squared-error loss between the two feature representations and is given by:

$$L_{content}(p, x, l) = \frac{1}{2} \sum (F^l_{i,j} - P^l_{i,j})^2 \tag{1}$$

Along the processing hierarchy of the network, the input image is converted into representations that are increasingly more sensitive to actual content of the image and less sensitive to exact details. Thus, one of the higher layers of the network is used for calculation of content loss. Feature space to represent the style of an image can be built as a correlation matrix for different filter responses in a layer. This correlation matrix is called Gram matrix (G^l) and is given by:

$$G^l = (F^l)(F^l)^T \tag{2}$$

Here F^l is a matrix such that its rows are vectorized versions of the different filter responses in the layer l. If p and y are the original image and the style image and G^l and Y^l are their Gram matrices for layer l, then the contribution of layer l to style loss is given by:

$$E_l = \frac{1}{4N_l^2 M_l^2} \Sigma \left(G^l_{ij} - Y^l_{ij}\right)^2 \tag{3}$$

Here N_l is the number of filters in layer l and M_l is the number of pixels in each feature map of layer l. The total style loss is given by:

$$L_{style}(p, y) = \sum w_l E_l \tag{4}$$

where w_l are the weighting factors for contribution of each layer to style loss. Finally, the overall loss function we minimize during the style transfer algorithm defined as:

$$L_{\text{total}}(p, x, y) = \alpha L_{\text{content}}(p, x) + \beta L_{\text{style}}(p, y) \tag{5}$$

where α and β are the weighting factors for contribution of content loss and style loss to the total loss.

3.1.1 Trade-off Between Content and Style Matching

We note that there exists a trade-off between matching the content and the style in the algorithm. Specifically, the overall loss $L_{\text{total}}(p, x, y)$ that we minimize is a linear combination of the content loss $L_{content}(p, x, l)$ and the style loss $L_{style}(p, x, y)$, thus enabling us to control their relative importance. If the emphasis on minimizing the style loss is high, we are able to obtain an appearance like the style image, essentially giving a texturized version of it, but the output image will lack content of the content image. On the other hand, if the emphasis on minimizing content loss is high, we can obtain the objects and their distribution in space from the content image correctly in the output image, however the appearance will not match that of the style image.

3.1.2 Effect of Different Layers of the Convolutional Neural Network

Using several layers for calculation of style loss helps to preserve local image structures at a large scale leading to a more continuous and smoother visual experience. Thus, using more layers for calculation of style loss helps to generate visually appealing images. Using a lower layer in the hierarchy for calculation of content loss causes the algorithm to match the detailed pixel information between the content image and the generated image. Moreover, it causes the texture of the style image to be merely blended over the content image. However, using a higher layer in the hierarchy for calculation of content loss causes the algorithm to extract the actual semantic content from the content image which is properly merged with the texture of the style image.

3.1.3 Limitations of the Style Transfer Algorithm

The biggest limiting factor for this algorithm is the resolution of the generated image. Both the dimensionality of the

optimization problem and the number of units in convolutional neural network increase linearly with the number of pixels. Thus, the speed of the algorithm depends on the resolution of the generated image. Another limitation is the fact that the problem of style transfer is not a very well defined one. Style of an image can refer to a variety of things – the brush strokes in a painting, the color map or the dominant shapes in an image. We consider style transfer to be successful if the generated image looks like the style image but displays objects from the content image. This criterion is hard to quantify mathematically. Presence of low-level noise in the generated image can be another issue.

3.2. Anomaly Detection Model

The anomaly detector, in this case based on the deep neural network architecture, is responsible for identifying datapoints, which are deemed to be outliers to such extent that there arise suspicion that they may be coming from a different distribution or may be generated by a different underlying process [16]. We utilize the anomaly detector to automatically evaluate whether an image generated by the style transfer algorithm appears close enough to the real world images, which we already have in our dataset.

The anomaly detector we use in our work is based on the autoencoder architecture inspired by [9]. Specifically, an autoencoder is a special type of a neural network architecture designed to encode the input into a lower dimensional compressed and meaningful representation, and then decode it back such that the reconstructed input is similar as possible to the original one [2]. We measure the reconstruction error using the commonly utilized mean squared error (MSE), which can be defined as: $\frac{1}{n}\Sigma(x-\widehat{x})^2$, where x is the original point and $\widehat{x}$ corresponds to its reconstruction.

We have opted for the convolutional autoencoder architecture as the it has a smaller number of parameters and generally requires less training time compared to the conventional autoencoders, which are based solely on fully connected layers [6].

3.3. Classifier Model

We use binary classifier based on the Resnet-50 network architecture introduced by [17] and trained via transfer learning from a model pre-trained on the ImageNet 1000 dataset [24]. Moreover, we have adjusted the first layer of the Resnet-50 model from requiring an input with 3 channels to expecting a single channel input as we are using grey-scale images.

We used the Binary Cross Entropy loss function for the training of the classifier. The Binary Cross Entropy can be formalized as:

$$\mathcal{L} = -\frac{1}{N_b}\sum_{i=1}^{N_b} y_i \cdot \log \hat{y}_i + (1-y_i) \cdot \log\left(1-\hat{y}_i\right) \quad (6)$$

Layer	Input	Output
$3\times3\times64$	x $(1\times H\times W)$	x_{0-1} $(64\times H\times W)$
$3\times3\times64$	x_{0-1}	x_{0-2} $(64\times H\times W)$
MaxPool	x_{0-2}	x_{1-1} $(64\times 1/2H\times 1/2W)$
$3\times3\times128$	x_{1-1}	x_{1-2} $(128\times 1/2H\times 1/2W)$
$3\times3\times128$	x_{1-2}	x_{1-3} $(128\times 1/2H\times 1/2W)$
MaxPool	x_{1-3}	x_{2-1} $(128\times 1/4H\times 1/4W)$
$3\times3\times256$	x_{2-1}	x_{2-2} $(256\times 1/4H\times 1/4W)$
$3\times3\times256$	x_{2-2}	x_{2-3} $(256\times 1/4H\times 1/4W)$
MaxPool	x_{2-3}	x_{3-1} $(256\times 1/8H\times 1/8W)$
$3\times3\times512$	x_{3-1}	x_{3-2} $(256\times 1/8H\times 1/8W)$
$3\times3\times512$	x_{3-2}	x_{3-3} $(256\times 1/8H\times 1/8W)$
MaxPool	x_{3-3}	x_{4-1} $(256\times 1/8H\times 1/16W)$
$3\times3\times512$	x_{4-1}	x_{4-2} $(512\times 1/16H\times 1/16W)$
$3\times3\times512$	x_{4-2}	x_{4-3} $(512\times 1/16H\times 1/16W)$
UpSample	x_{4-3}	up_{3-1} $(512\times 1/8H\times 1/8W)$
$3\times3\times256$	$[up_{3-1}, x_{3-3}]$	up_{3-2} $(256\times 1/8H\times 1/8W)$
$3\times3\times256$	up_{3-2}	up_{3-3} $(256\times 1/8H\times 1/8W)$
UpSample	up_{3-3}	up_{2-1} $(256\times 1/4H\times 1/4W)$
$3\times3\times128$	$[up_{2-1}, x_{2-3}]$	up_{2-2} $(128\times 1/4H\times 1/4W)$
$3\times3\times128$	up_{2-2}	up_{2-3} $(128\times 1/4H\times 1/4W)$
UpSample	up_{2-3}	up_{1-1} $(128\times 1/2H\times 1/2W)$
$3\times3\times64$	$[up_{1-1}, x_{1-3}]$	up_{1-2} $(64\times 1/2H\times 1/2W)$
$3\times3\times64$	up_{1-2}	up_{1-3} $(64\times 1/2H\times 1/2W)$
UpSample	x_{1-3}	up_{0-1} $(64\times H\times W)$
$3\times3\times64$	$[up_{0-1}, x_{0-2}]$	up_{0-2} $(64\times H\times W)$
$3\times3\times64$	up_{0-2}	up_{0-3} $(64\times H\times W)$
$3\times3\times3$	up_{0-3}	$output$ $(3\times H\times W)$

Figure 1. Anomaly detector architecture

where N_b is the training batch size, y_i is the true classification of the example i (0 - benign, 1 - malignant), and $\hat{y}_i$ is the prediction produced by the model, which is a continuous value corresponding to the class prediction of the model $\hat{y}_i \in [0,1]$.

We used the AdamW optimizer introduced in [27] to optimize the above defined loss function. We have generally obtained better results on this problem when using the Adam optimizer [23] compared to the stochastic gradient descent optimizer (SGD) [31]. The main benefit of AdamW over the Adam is the decoupling of weight decay from the gradient udpate. The weight decay is used for regularization, specifically the L_2 regularization, and the aim of using regularization is to ensure better model generalization. Besides the learning rate, the optimizer several parameters that need to be set, specifically, the exponential decay rate for the first moment estimates β_1, the exponential decay rate for the second-moment estimates β_2 and the ϵ, which is a very small number to prevent division by zero.

The Adam optimizer can be described by the following equations:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\, g_t \quad (7)$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\, g_t^2 \quad (8)$$

where g_t is the gradient on the current batch and m_t, v_t are moving averages. The obtained variables m_t, v_t are then substituted into the following equation to obtain the weight update:

$$w_t = w_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad (9)$$

where η represents the learning rate.

3.4. Automated pipeline

Style transfer algorithm for different pairs of images appears to be very sensitive to the choice of the hyperparameters, therefore to inspect the ability of the style transfer algorithm across a large number of samples and to inspect the quality of these new samples, we have created an automated pipeline for generating and evaluating the quality of generated images on larger scale.

The automated pipeline, depicted in the Figure 2, starts with a random sampler, randomly drawing a pairing of the malignant and benign images for the style and content image respectively. Next the style transfer algorithm with baseline hyperparameters (iterations = 250, style weight = 100000, content weight = 1, convolutional layer = 5) generates a new image. This image is then passed to the anomaly detector, which decides whether it looks consistent enough with the real data. If the image fails, the style transfer is re-run with different parameters, if it passes the test, then it is evaluated by the pre-trained classifier. We evaluates the strength of the believe about the class membership of the new image via a softmax function. If the image is classified as benign or as malignant, but with a softmax value below the required treshold level, the parameters of style transfer are changed and the cycle repeats. Otherwise, the image is saved and a new image pairing drawn. The hyperparameter combinations for style transfer were chosen based on exploration described in the results section. Moreover, we set the maximum number of repeated trials per image to 7.

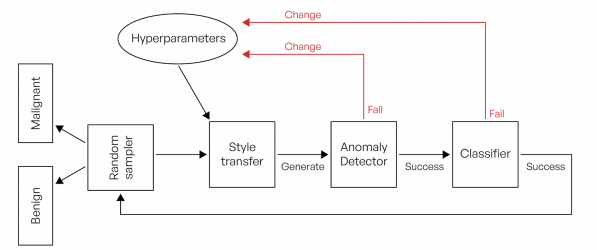


Figure 2. The automated pipeline outline

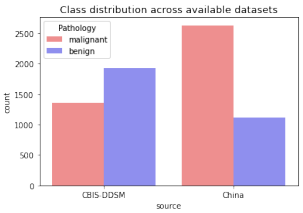
4. Data

4.1. Datasets

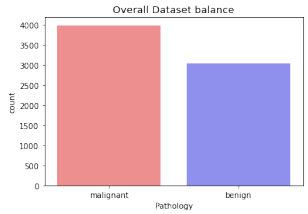
To increase the amount of data available, we have merged two certified datasets containing mammography screenings evaluated by professional radiologists.

The first datasets is the well known Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) released in 2017, [26], which we complement by the recently released Chinese Mammography Database (CMMD), [7]. Both datasets contain examples of malignant and benign mammography findings.

We provide a visual description of the class distribution across the two datasets (Fig. 3a) as well as the class distribution in the merged dataset (Fig. 3b). From these we observe that while the individual datasets, especially the CBIS-DDSM, suffer from class imbalance, the merged dataset is actually relatively balanced.



(a) Class distribution by source



(b) Class distribution in the dataset

4.2. Data Usage

As outlined before, the goal of our project is to investigate how reliably the neural style transfer algorithm [11] can be applied to the dataset imbalance problem. However, both the anomaly detector and classifier, forming the automated pipeline, have to be trained on some data, and ideally later shown generated images, based on previously unobserved samples to increase the robustness of this approach.

Therefore, we have opted to split and divide our data in the following way: 80% - training set for anomaly detector and classifier; 10% - validation set for the training of the anomaly detector and the classifier; and 10% - hold-out set for generating new images using the style transfer. The data-set was split into these parts randomly, but care was taken to maintain the ratio of malignant and benign cases as in the full data-set.

5. Results

5.1. Style transfer hyperparameter exploration

The possibility of generating synthetic malignant breast cancer images using a style transfer algorithm with malignant breast cancer image as style image and benign breast cancer image as content image has been studied. Specifically, the style transfer algorithm consists of three important hyperparameters – the layer of convolutional network used for calculation of content loss, number of steps for the optimization process and the ratio of weights for content loss and style loss. The role of these hyperparameters in the style transfer process has been investigated.

First, we took a combination of content image and style image and generated images for different choices of number of steps – $150, 300, 500, 650, 1000$ and 1200. The values of content weight = 1, style weight = $100,000$ were kept fixed and 'conv_5' convolutional layer was used for calculation of content loss. The combination of content image and style

image has been shown in Fig. 4 and the generated images have been shown in Fig. 5.

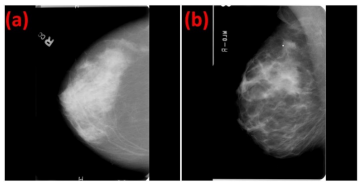


Figure 4. (a) Content and (b) Style images for analysis, case 1

As we increase the number of steps from 150 to 500, the style transferred from the style image to the content image becomes progressively more prominent as shown in Fig.5. However, for higher number of steps ($>= 650$) the generated images no longer look realistic. This trend was observed for several other combinations of content image and style image (not shown here). On the other hand, using too few steps causes the generated image to be almost same as input image which is the content image in the implementation being used by us.

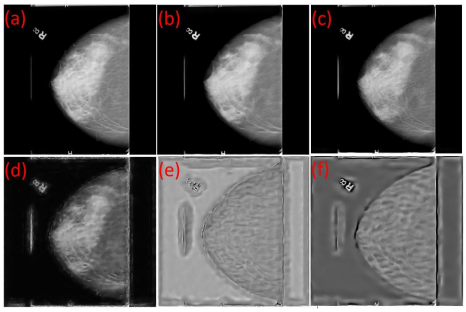


Figure 5. Images generated for content weight = 1m style weight = 100,000, and 'conv_5' convolutional layer for content loss. Number of steps were chosen to be: (a) 150 (b) 300 (c) 500 (d) 650 (e) 1000 (f) 1200

Next, we examined the effect of the ratio of content weight and style weight by using a fixed style weight = 100,000, number of steps was kept fixed at 500, and 'conv_5' convolutional layer used for calculation of content loss. We tested content weights - 0.1, 1, 10, 100 or 200. The generated images are shown in Fig. 6. Clearly the generated image looks very similar to the content image for content weight = 100 or 200 while significant changes can be seen for lower content weights of 0.1, 1 or 10.

Lastly, we examined the effect of choice of the convolutional layer used to calculate content loss while fixing content weight = 1, style weight = 100,000 and number of steps = 500, the convolutional layer for calculation of content loss was chosen as either 'conv_1', 'conv_2', 'conv_3', 'conv_4' or 'conv_5'. The combination of content image and style image chosen for this are shown in Fig. 7, and the generated images are shown in Fig. 8.

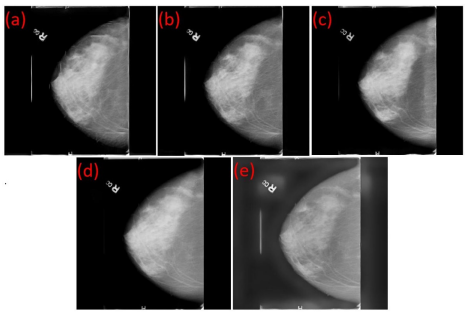


Figure 6. Images generated with number of steps = 500, style weight = 100,000 and 'conv_5' convolutional layer for content loss. Content weight was chosen to be (a) 0.1 (b) 1 (c) 10 (d) 100 (e) 200.

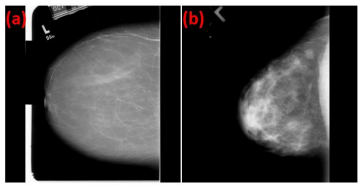


Figure 7. (a) Content and (b) Style images for analysis, case 2

For layers lower in hierarchy, i.e. 'conv_1', the algorithm tries to match the exact pixel values, thus merging of the texture from the style image and the actual contents from the content image is not proper and the generated image does not look realistic (i.e. granular texture). However, for layers higher in the network hierarchy, i.e. 'conv_5', the algorithm can extract the actual semantic content from the content image, thus merging of the style and content is better, resulting in image which looks realistic. These results can be observed in the Figure 8

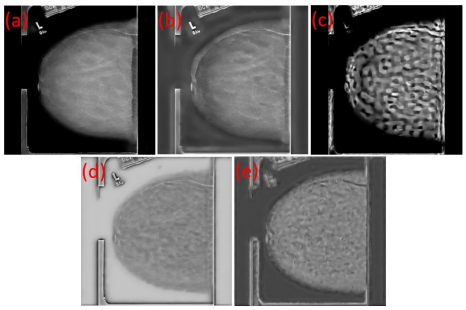


Figure 8. Images generated with Number of steps = 500,styl e weight = 100,000 were kept fixed and 'conv_5' convolutional layer for content loss. Content weight was chosen to be (a) 0.1 (b) 1 (c) 10 (d) 100 (e) 200

The set of hyperparameters that generates realistic images can vary significantly for different combinations of

content image and style image. An example of content image, style image and generated image for number of steps = 300, content weight = 1, style weight = 100,000 and using 'conv_5' convolutional layer for content loss has been shown in Fig. 9. Clearly the image does not look realistic. However, the same set of hyperparameters has given a realistic image for a previous example discussed here.

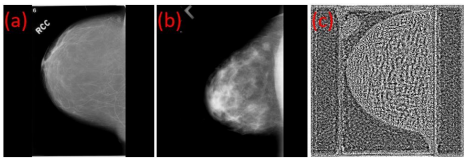


Figure 9. (a) Content image (b) Style image (c) Generated image for number of steps = 300, content weight = 1, style weight = 100,000 and 'conv_5' layer was used to calculate content loss. The generated image clearly does not look realistic while the same set of hyperparameters gave a realistic image for a different combination of content image and style image shown in a previous figure.

5.2. Anomaly Detector and Classifier Training

5.2.1 Anomaly Detector

We trained the anomaly detector for 200 epochs, using the MSE loss function optimized by the Stochastic Gradient Descent with the momentum parameter set to 0.9, [32], and learning rate guided by a scheduler, changing gradually from 0.1 to 0.0001.

After training the anomaly detector, we have fed it both all the training examples and all the test examples to separately observe the L_1 reconstruction error ($L_1 = \sum_{i=1}^{n} |y_{\text{true}} - y_{\text{predicted}}|$, as recommended in [9]), which we use to derive a threshold for the anomalous examples of the generated images.

The L1 reconstruction losses observed across the training and testing set are visualised in the Figure 10. The mean L1 error was around 0.2 on both sets and the 95% percentile was around 0.5. Considering anything above the 95% to be a rarely occurring outlier and and after an additional manual evaluation of visibly incoherent images generated by the style transfer, we have decided to set the decision threshold for anomaly to $L_1 = 0.5$.

5.2.2 Classifier

We have trained the model for 75-epochs with a batch size of 32, moreover the learning rate for the AdamW optimizer was set to $lr = 0.0001$, the weight decay was kept at the suggested 0.01, $\epsilon = 1e-08$ and the beta parameters were set to $\beta_1 = 0.9, \beta_2 = 0.99$. We have achieved a 97.32% and 78% accuracy on the training set and testing set respectively. While the accuracy metric is not perfect, it is sensible to use in terms of the binary classification. Moreover,

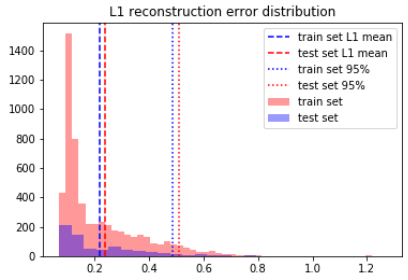


Figure 10. L1 loss distribution in the anomaly detector

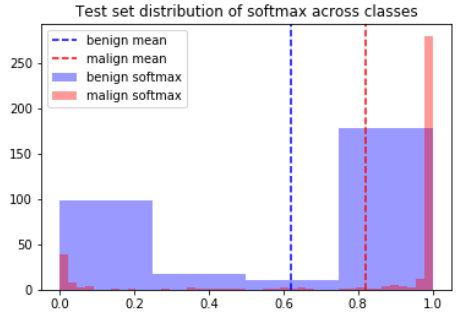


Figure 11. Softmax distribution across the all test cases

we have checked the confusion matrix and found that most of the accuracy error comes from mis-classification of the benign cases, which we actually do not study in our case as we care about evaluating whether a case is malignant or not and we have the ground truth for the benign case in the style transfer algorithm. However, to address this classifier issue in our automated pipeline, we have only considered as valid examples cases, where the benign image was initially classified as benign and then switched to malignant.

For the malignant case we found that across all test cases the average softmax value was 0.81 across all malignant cases. Based on this we set the softmax cutoff in our pipeline to 0.7 in order to require a stronger response from the classifier to only keep examples about which we believe have high chance of actually being malignant. While this is a decision that significantly impacts the outcome of our method, we believe that especially in the case of healthcare related datasets, we should be very certain that the produced results are of a the highest quality and can be comfortably used in further medical related procedures. The softmax distribution for the test cases is visible in the Figures 11.

5.3. Style Transfer Generation Results

Our testing using our automated pipeline has lead to several findings. First of all, we have found that the anomaly detector with the treshold of 0.5 generally work well and is

able to detect cases generated images, which are visibly not close to the real ones. Examples of these cases are presented in Figures 12 and 13. However, we have also detected cases where the anomaly detector fails and they get subsequently classified as malignant by the Classifier, thus resulting in invalid example potentially entering the dataset and harming the effect of our technique. We present such example in the Figure 14. It appears that our anomaly detector generally struggles with cases, where the structure of the breast is filled with a fine grained pattern.

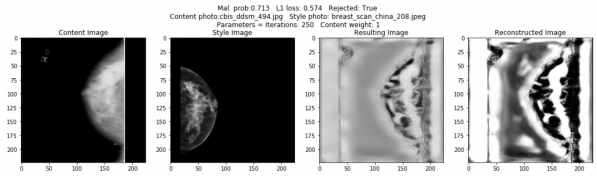


Figure 12. Anomaly detector success example 1

Secondly, we have found that most of the example generated by the Style transfer, do not meet the required 0.7 probability value treshold we required from in the classifier, thus result in a failed case. Therefore the amount of the cases that do not pass the quality control in our testing pipeline is very high, resulting in extremely high number of attempts to obtain a valid case. Examples of two successful cases are presented in the Figures 15 and 16. We observe that in both images there is a slight change in the internal pattern style of the breast, which we believe is the change from benign to malignant structure.

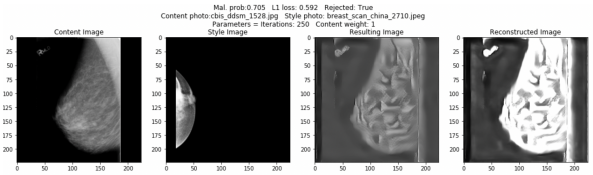


Figure 13. Anomaly detector success example 2

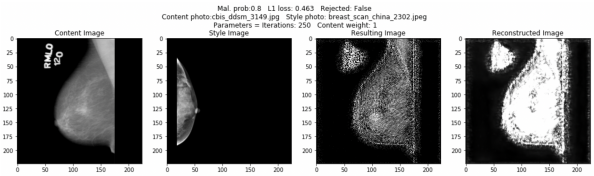


Figure 14. Anomaly detector failure example

6. Discussion

In this section we provide a discussion of our findings, potential causes that may have led to them.

In overall, we have found that synthesizing images for malignant breast cancer using the neural style transfer algorithm by [11] is extremely sensitive to the three hyperparameters of the style transfer model, namely the convolutional layer used for calculation of content loss, the number of steps and the ratio of content weight and style weight. However, we were able to demonstrate for several combinations of a content image (a mammography screening image containing a benign finding) and a style image (a mammography screening image containing a malignant finding) that it is possible to adjust the above mentioned hyperparameters to produce realistic looking images of the malignant class. We have demonstrated that the content image which was previously labelled as benign case, by the classifier trained on real data, was classified as malignant after the style transfer algorithm was used to apply the style of another malignant example from the dataset. However, manually optimizing the hyperparameters for each combination of content image and style image makes it difficult to upscale this algorithm for generating synthetic malignant breast cancer images. Even, the automated pipeline introduced in our work does not fully fix this problem as it often requires iterating over different sets of hyperparameters to eventually produce a new valid malignant example.

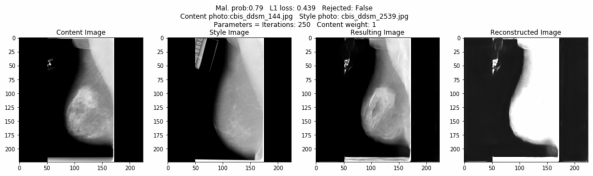


Figure 15. Accepted generated malignant example 1

While the style transfer algorithm has been applied to generate malignant skin cancer images in a past study using dataset with colored images ([28]), the dataset for our project has grey scale images which may make it harder for the algorithm to extract the relevant style characteristics from the malignant breast cancer images. The color map is an important characteristic, which is considered to be style by this algorithm. However, grey scale color map not contain the information equivalent to the RGB color map.

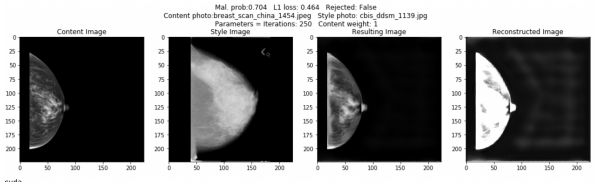


Figure 16. Accepted generated malignant example 2

Moreover, the lack of radiological skills make it difficult to evaluate the performance of style transfer algorithm

as even when we observe an image that seems structurally correct and is classified as malignant after the style transfer, there remains uncertainty about the quality of the image as there is no guarantee that the classifier has actually classified the image as malignant due to the fact that the benign finding has been changed to a malignant finding via the style transfer. This provides a significant barrier to the application of this method even if it was finetuned as there is a high risk of introducing an unwanted bias into the dataset that would be balanced using this data augmentation method.

Finally, we have developed an anomaly detection model in order to try to classify the generated images as realistic or non-realistic. This anomaly detector is based on an autoencoder as described in the methodology section. However, we have observed that while the autoencoder learns to reliably reconstruct shapes from the image embeddeding, thus the shape of the breast pictured in the mammography image, it fails to recover the inner structure/texture of the breast. This failure of inner structure recovery is the same for real data as well as images generated by the style transfer, which may not be truly realistic looking even to an amateur observant. We believe that the failure to capture the inner structure of the breast is a result of the shortcomings of the MSE loss, which are magnified by the fact that the images used are only gray-scale. We suspect that during the training the autoencoder quickly gets stuck in a local optima, where the MSE is minimised recovering the shape of the breast and uniformly painting the inner part of the breast with grey color (color between black and white). This way the autoencoder reliably detects the black regions of the image, which do not contain the breast, thus perfectly minimises the loss over that region. Moreover, as the inner parts of the breast are usually covered with extremely complex thin white structures, uniformly expecting the content of the breast to have homogeneous color is a reliable way to minimise the loss over different cases, thus the results. Therefore, in this case we believe that the anomaly detector appears difficult to train and may not present a suitable and reliable approach due to the difficulty of reconstructing the inner part of the breast in the mammography images. These shortcomings of the MSE loss are also discussed in [29].

Moreover, while it was not an initial goal, we realize that this approach basically mimics the algorithm of the Generative Adversarial Networks as the anomaly detector basically poses as the discriminator and the style transfer model is a Generator of a sense. However, there are significant differences as first of all the models are not trained together as in GANs, secondly the generator does not learn a latent posterior distribution, but rather just combines two inputs. Lastly the style transfer based generator is extremely sensitive to the setting of the hyperparameters, whose available combinations were chosen manually for the automated generating pipeline.

In overall, some improvements in the autoencoder model may be required to upscale this algorithm for generating synthetic malignant breast cancer images by automatically rejecting any generated images that do not look realistic. Therefore, the current algorithm should only be used to generate synthetic malignant breast cancer images by manually optimizing the relevant hyperparameters, thus has limited scalability. Moreover, we recommend that the selected generated examples are presented to radiology expert for further evaluation before using them to balance a real world dataset used in classifier training.

7. Conclusion

In this project we investigated the possibility of using style transfer as a data augmentation tool to generate synthetic malignant breast cancer images. A benign breast cancer image was used as the content image while a malignant breast cancer image is used as the style image in this algorithm. The impact of three important hyperparameters, the choice of number of steps, the convolutional layer used to calculate content loss and the ratio of content weight to style weight, on the quality of generated images was examined. For some individual cases we show that the label of an input image (a copy of the content image) changes from benign to malignant on application of the style transfer using a malignant breast cancer image as style image. However, the set of hyperparameters which generates realistic image can be different for different combinations of content image and style image. We developed an automated image generating pipeline containing an autoencoder functioning as an anomaly detector and segregating realistic generated images from non-realistic ones as well as a classifier used to evaluate whether a produced image appears to be truly malignant. Moreover, we found that the autoencoder can reconstruct the shape of the breasts in an image but not the texture. Effectiveness of the style transfer algorithm was evaluated by classifying the generated images using a binary classifier with two output classes – malignant and benign. A certain fraction of the generated images is classified as benign, but the intended outcome was to generate synthetic malignant breast cancer images. There are several possible reasons for this. One of the important characteristics that the algorithm considers "style" is the color map of an image. However, images in the dataset used by us are greyscale images. Moreover, the set of hyperparameters that generate realistic image need to be optimized for each combination of content image and style image. This is a major obstacle to upscaling this algorithm to generate large amounts of synthetic data images.

Individual contribution - Filip Ryzner

- Collected and pre-processed all the datasets for the project
- Coming up with the core ideas and methodology for the project
- Made significant contribution to writing the report - responsible for relevant parts of methodology and results as well as the overall review
- Programming and fine tuning of the automated generating pipeline and all related components (style transfer part reused from PSet 7)
- Converted partners contribution into latex

References

[1] N Ashikhmin. Fast texture transfer. *IEEE computer Graphics and Applications*, 23(4):38–43, 2003. 2

[2] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv preprint arXiv:2003.05991*, 2020. 4

[3] Camilo Bermudez, Andrew J Plassard, Larry T Davis, Allen T Newton, Susan M Resnick, and Bennett A Landman. Learning implicit brain mri manifolds with deep learning. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105741L. International Society for Optics and Photonics, 2018. 2

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 1, 2

[5] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16, 2015. 2

[6] Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau. Autoencoder-based network anomaly detection. In *2018 Wireless Telecommunications Symposium (WTS)*, pages 1–5. IEEE, 2018. 4

[7] C Cui, L Li, H Cai, Z Fan, L Zhang, T Dan, J Li, and J Wang. The chinese mammography database (cmmd): An online mammography database with biopsy confirmed types for machine diagnosis of breast. *The Cancer Imaging Archive*, 1, 2021. 5

[8] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001. 2

[9] Ye Fei, Chaoqin Huang, Cao Jinkun, Maosen Li, Ya Zhang, and Cewu Lu. Attribute restoration framework for anomaly detection. *IEEE Transactions on Multimedia*, 2020. 4, 7

[10] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. 2

[11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 1, 2, 3, 5, 8

[12] Anjana Gosain and Saanchi Sardana. Handling class imbalance problem using oversampling techniques: A review. In *2017 international conference on advances in computing, communications and informatics (ICACCI)*, pages 79–85. IEEE, 2017. 2

[13] Monika Grewal, Muktabh Mayank Srivastava, Pulkit Kumar, and Srikrishna Varadarajan. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 281–284. IEEE, 2018. 1

[14] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016. 1

[15] Changhee Han, Yoshiro Kitamura, Akira Kudo, Akimichi Ichinose, Leonardo Rundo, Yujiro Furukawa, Kazuki Umemoto, Yuanzhong Li, and Hideki Nakayama. Synthesizing diverse lung nodules wherever massively: 3d multi-conditional gan-based ct image augmentation for object detection. In *2019 International Conference on 3D Vision (3DV)*, pages 729–737. IEEE, 2019. 2

[16] Douglas M Hawkins. *Identification of outliers*, volume 11. Springer, 1980. 4

[17] 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. 4

[18] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001. 2

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 3

[20] Mohammad Tariqul Islam, Md Abdul Aowal, Ahmed Tahseen Minhaz, and Khalid Ashraf. Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850*, 2017. 1

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[22] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 4

[25] Hochang Lee, Sanghyun Seo, Seungtaek Ryoo, and Kyunghyun Yoon. Directional texture transfer. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, pages 43–48, 2010. 2

[26] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017. 5

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[28] Agnieszka Mikolajczyk and Michal Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE, 2018. 1, 2, 8

[29] PyTorch Lightning. Deep autoencoders tutorial, 2022 [Online]. 9

[30] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1

[31] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 4

[32] Ersan Yazan and M Fatih Talu. Comparison of the stochastic gradient descent based optimization techniques. In *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–5. IEEE, 2017. 7

[33] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019. 2