Synthetic Data Augmentation using GAN for Improved Breast Cancer Classification

Calvin Chi
UC Berkeley
calvin.chi@berkeley.edu

Edward Fang
UC Berkeley
edward.fang@berkeley.edu

Eric Wang
UC Berkeley
ericwang1996@berkeley.edu

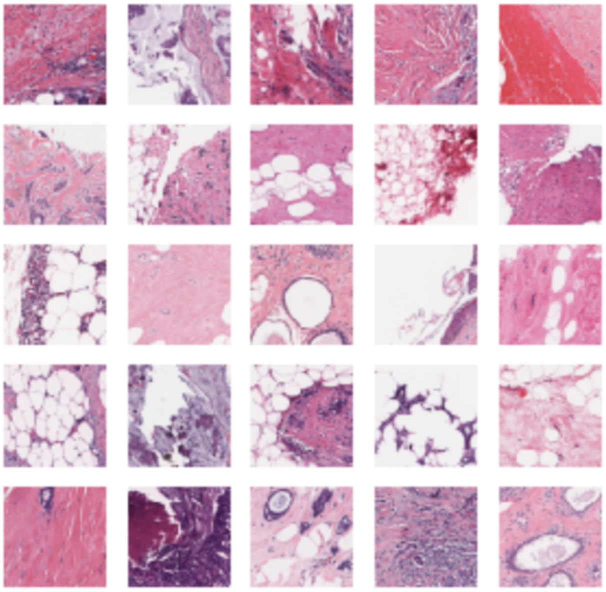


Figure 1: Visualization of randomly sampled histology images.

Abstract

The application of computer vision to biomedical imaging holds many promises, from biomedical research to clinical diagnosis. With the appropriate dataset, deep learning can in principle detect disease at rates comparable to a physician's diagnosis if not better. However, large quantities of high quality data are often challenging to obtain in the biomedical space. In this project, we explore whether image synthesis with generative adversarial networks (GANs) can address this challenge. Specifically, we evaluate the effectiveness of deep convolutional GAN (DCGAN) and cycle adversarial network (CCAN) in augmenting training data to improve breast cancer classification. Additionally, we also report heuristics that improve visual similarity between generated and real images. Overall, we found DCGAN to be the most effective method to improve breast cancer classification accuracy. Accuracy and precision improved by around 5% *and* 12% *respectively as a result of data augmentation. However, recall decreased by nearly* 15%. *Our work shows that it is possible to use image synthesis with GANs to improve certain performance metrics, but also highlights the many challenges associated with this task.*

1. Introduction

Medical imaging is the technique and process of creating visual representations of the interior of a body for clinical analysis and medical intervention, as well as visual representations of the function of organs or tissues. Examples include X-ray radiography, magnetic resonance imaging, and microscopy [6].

Since the analysis of biomedical images involves pattern recognition, efforts have already been made in using machine learning as a way to advance computer-aided diagnosis [14]. However, challenges in this application include availability and cost of data. For example, it costs time for pathologists to manually label large quantities of images. Conflicts of interest prevent clinics and hospitals from sharing data with each other. In this paper, we investigate the utility of data augmentation techniques for disease diagnosis in the context of breast histology images.

We evaluate the effectiveness of data augmentation methods to improve cancer classification performance. We explore whether the generative adversarial network (GAN) can synthesize images to improve classification performance [2]. The two types of GANs we will work with are Cycle Consistent Adversarial Networks (CCAN) and Deep Convolutional Generative Adversarial Networks (DCGAN) [10, 16].

2. Related Works

The applications of computer vision to biomedical imaging is an active area of interdisciplinary research [11]. Successful deep learning applications often require large datasets that capture population variation. However, there exist challenges in collecting large quantities of well-annotated, high-quality biomedical images. These challenges include human labor by the medical professional community in data collection and data sharing due to legal concerns and/or conflicts of interest. Thus, successful image synthesis that retains or captures biomedical signal could greatly address these challenges. GANs have been lightly explored as a method for medical image synthesis. We discuss two applications in computed tomography (CT) synthesis from magnetic resonance imaging and retinal vessel segmentation synthesis.

CT is an non-invasive imaging technique that enables visualization of cross sections of a human body. It has applications in preventive medicine and disease screening. However, CT exposes patients to radiation which can cause harmful side effects. Magnetic resonance imaging (MRI) is the preferred alternative, as patients are not exposed to any radiation. Context-aware GANs have shown compelling qualitative performance on synthesizing CT scans based on MRIs [7], notably outperforming all other state-of-the-art methods such as Atlas, SR and SRF+.

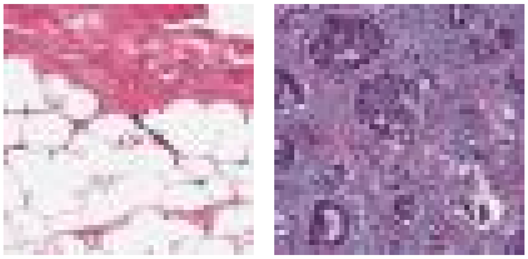
Another study attempted to generate synthetic retinal vessel segmentation data through a dual generative adversarial approach. Large amounts of clinical data remain private and thus restricted from public access. The motivation behind this study was to create synthetic images generated from real, private datasets in order to increase the amount of publicly available research data. Using retinal vessel segmentation data from the DRIVE database, the dual generative adversarial approach is able to produce realistic, synthetic retinal vessel segmentation data [3]. Artifacts in synthesized images have been reported to be a challenge in the dual generative adversarial approach. We encountered similar challenges with artifacts in the application of DCGANs, and in this project we investigate ways to address these challenges [10].

In this project, we also explore image-to-image mapping with cycle consistent GANs. Cycle consistent GANs have demonstrated impressive results on image-to-image translation, and we explore whether cycle consistent GANs can successfully translate a non-cancerous image to a cancerous image and vice versa for the purpose of data augmentation [16].

3. Dataset

Our dataset from Kaggle consists of 5,547 breast histology images of size $50 \times 50 \times 3$, of which 2,788 images are labeled as invasive ductal carcinoma (IDC) and 2,759 images are labeled as non-IDC [5]. IDC is the most common form of breast cancer.

We use a 67-33 split rule on the 5,547 images to form training and validation sets. A fixed seed was used to generate these training and validation sets. After splitting, the training set contained 1,819 non-IDC images and 1,897 IDC images, while the validation set contained 940 non-IDC images and 891 IDC images. The same validation set was used across all evaluations to establish a baseline for comparison.



(a) Non-Invasive Ducctal Carcinoma (b) Invasive Ducctal Carcinoma

Figure 2: Classification of Breast Cancer Histology

4. Methods

We generate synthetic images by feeding the training dataset as an input to GANs. In the process, we address challenges encountered in training GANs and different approaches we investigated to resolve these challenges. Afterwards, for each set of synthetic images generated by an image synthesis method, we will evaluate its effectiveness in increasing test classification performance on a neural network trained for binary classification of cancer vs non-cancer. Computation was provided by EC2 of Amazon Web Services.

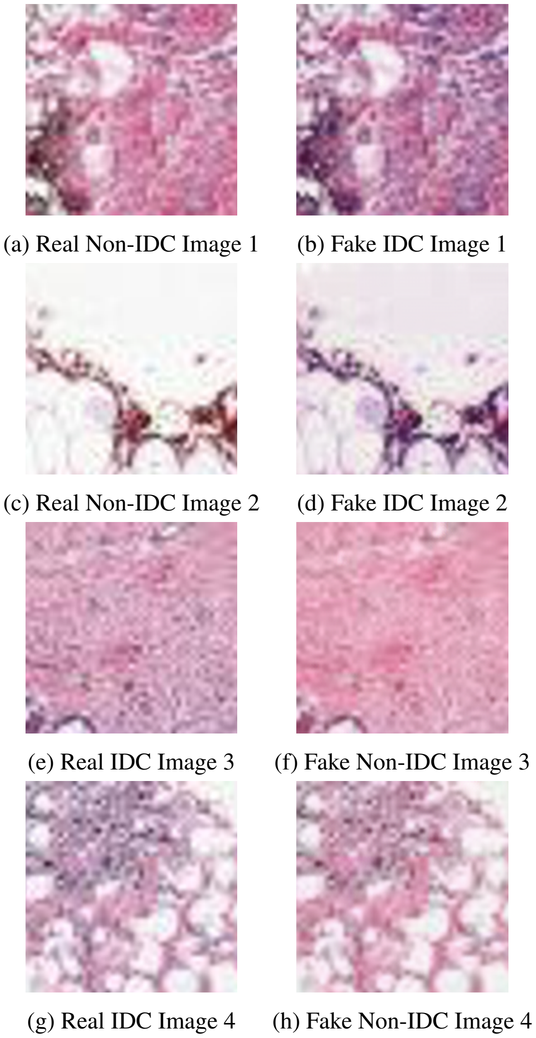
4.1. Cancer Classification with Transfer Learning

Training a deep convolutional neural network from scratch would require significantly more images than available in our dataset, which would be extremely costly in the context of histology as experts would need to label millions of images for this method to be effective. Alternatively, transfer learning is a method that utilizes the pre-trained weights of convolutional layers and then retrains the final fully connected layer to create a class specific classifier.

Transfer learning on Resnet-18 was used as a baseline for comparison [4] [9]. The convolutional layers in Resnet are able to learn comprehensive and informative features that are effective in image classification. We chose to use these convolutional layers as a feature extractor for our histology dataset and use transfer learning to retrain the last fully connected layer to create an IDC classifier. Pretrained weights from PyTorch were used and the final layer was retrained to predict whether a given image was IDC or non-IDC. A variety of image perturbation methods were used, including rotation, horizontal and vertical flips, and random noise to introduce robustness to the model. Relative to not using image perturbations, including image perturbations had little impact on accuracy, precision and recall.

4.2. Cycle Consistent Adversarial Networks

CCANs have achieved compelling results for many image-to-image translation tasks, such as zebras-to-horses, summer-to-winter, and painting and photo styles [16]. An important characteristic of CCANs is that they can be trained on two sets of images that aren't directly related. For images of horses and zebras, a ”pair” of training samples need not have the same background or pose. For our histology images, a ”pair” of training samples from IDC and non-IDC need not share the same histology. However, CCANs perform best when the unpaired images share similar visual content [16], suggesting that not all pairings will produce adequate results.


(a) Real Non-IDC Image 1 (b) Fake IDC Image 1

(c) Real Non-IDC Image 2 (d) Fake IDC Image 2

(e) Real IDC Image 3 (f) Fake Non-IDC Image 3

(g) Real IDC Image 4 (h) Fake Non-IDC Image 4

Figure 3: Results from CCAN

In the scenario of binary IDC classification, our assumption is that there is a mapping from non-IDC images to and from IDC images. In other words, we expect that given a real, non-IDC histology we are able to predict what that same histology would look like if it were IDC. Similarly, given a real, IDC histology we should be be able to predict the synthetic non-IDC histology.

Using CCANs in this context allows us to augment the dataset with IDC versions of non-IDC images and vice versa. Alternatively, if we do not reverse the labels when applying CCAN's, then this augmentation tests the robustness of our model as healthy looking synthetic IDC histologies are injected into the mix. We trained a CCAN on unpaired images of non-IDC and IDC images from the training set and then applied the trained CCAN to images in the training set. The CCAN was trained for 25 epochs and the cycle consistency loss converged at 0.002 after only a few epochs. This produced a set of synthetic training images that were the fake opposites of the real training images. During transfer learning, we tested combinations of real data only, synthetic data only and combined data.

In Figure 3, we visualize the results of applying the trained CCAN to the training set. Figures 3*a*, 3*c* are real, non-IDC images and Figures 3*b*, 3*d* are the synthesized, IDC images corresponding to the non-IDC images. Similarly, Figures 3*e*, 3*g* are real, IDC images and Figures 3*f*, 3*h* are the synthesized, non-IDC images corresponding to the non-IDC images. These results seem promising to non-experts.

We see that the CCAN is able to learn the features of non-IDC and IDC histologies and generate synthetic histologies given a real histology. Notably, the fake IDC images are darker and more purple in color than the real non-IDC images, and the fake non-IDC images are lighter in color and more pink than the real IDC images. However, Figures 3*g*, 3*h* also demonstrates that the CCAN does not always significantly alter the image–the synthesized image is fairly similar to the original image. This highlights a potential issue: if the labels of the synthesized images are opposite to the source image and if the synthesized image is similar to the source image, our combined dataset may contain conflicting labels for similar images. We will further analyze this caveat in the results section. Still, the majority of results are interesting. Of course, histology is much more complicated than a simple shift in hue, but to a non-expert eye the synthesized IDC images seem ”more cancerous” than the real non-IDC images and the synthesized non-IDC images seem ”less cancerous” than the real IDC images.

4.3. Clustering of Histology Images

GANs can produce conflated images that do not belong to a particular class when trained on complex and diverse datasets [1]. As Figure 1 shows, our histology images are diverse in hue, texture and morphology.

It appears that there is no single prototypical example of an IDC histology image, just as there is no single prototypical example of a non-IDC histology image. The differences between members of IDC histology images or non-IDC histology images could be significant enough to produce conflation in generated histology images.

To minimize the probability that poorly generated images are due to over-diversification of our dataset, we decided to run DCGAN on images that cluster together. We performed feature extraction with Scale Invariant Feature Transform (SIFT), an algorithm that allows us to generate a list of meaningful features from raw pixel data. We then used a Bag-of-Words model to create a histogram of features for each image and used the histogram as a feature vector representing the image. Running k-means on these feature vectors allowed us to generate clusters of each images which became the datasets we used to test out different GANs. The parameter k in k-means was selected via manual inspection of images in clusters. Empirically, we found that choosing $k = 4$ provided a good balance between having large clusters and having high degrees of similarity within clusters.

4.4. DCGAN

Deep Convolutional GAN (DCGAN) is an extension of the original GAN that provides improved training stability and ability to learn a hierarchy of representations from object parts to scenes in both the generator and discriminator. Changes to the original GAN include

- Introducing transpose convolution to the generator
- Using batch normalization in both the generator and discriminator
- Use ReLU activation in generator for all layers except for the output, which uses Tanh
- Use LeakyReLU activation in the discriminator for all layers.

Further details on DCGANs are described in the original paper [10]. Our DCGAN software is based off the TensorFlow DCGAN implementation by Taehoon Kim, which extends the architecture in the original DCGAN paper by updating the generator network twice for each discriminator update [13].

We first tested out this implementation on a randomly chosen 10% of bedroom images from the LSUN dataset, totaling 303,125 images [15]. Figure 4 visualizes the generated images and the implementation was deemed a success before moving forward with the histology dataset.

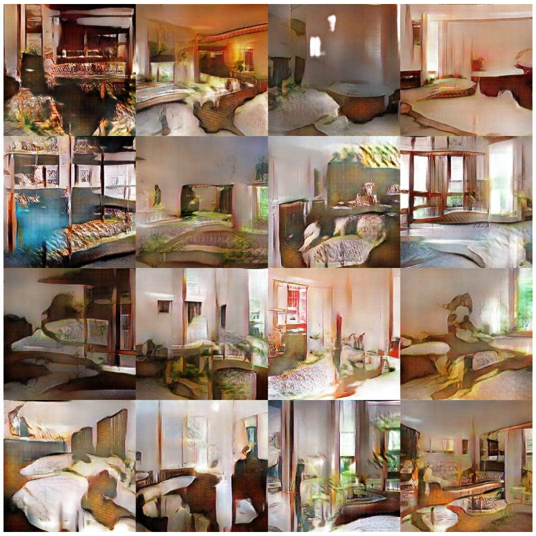
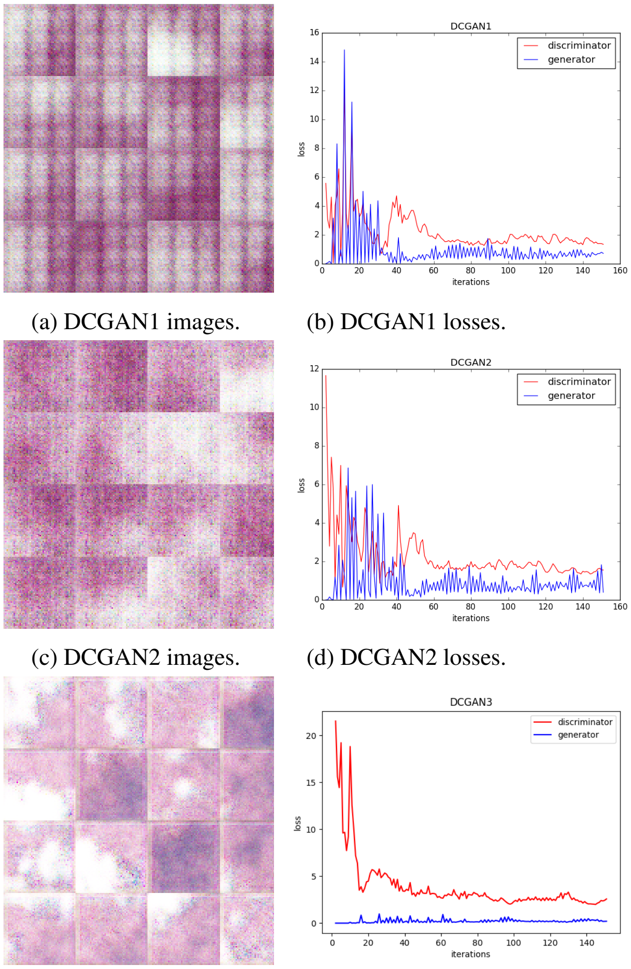


Figure 4: Visualization of generated images from DCGAN trained on LSUN bedroom dataset.

We tested our DCGAN on cancer images from only one cluster of histology images to minimize the probability that DCGAN may perform poorly due to large variation among input images. The cluster that was randomly chosen contained 404 images. We tested a total of three DCGAN versions, with one of the three being the original implementation, with the goal of generating believable histology images. In addition, we monitored discriminator and generator loss to ensure training stability and convergence. All results were compared after exactly 25 epochs. Figure 5 shows the results from all 3 DCGANs.

The first version, which we call DCGAN1, is the original DCGAN implementation. We can see in Figure 5a that the generated images were both low resolution and have ”checkerboard artifacts”, the square tiling pattern repeated across the image. The second version, which we call DCGAN2, implemented tips suggested and aggregated online [12]. Specifically, we implemented label smoothing and added Gaussian noise from $N(0, 0.2)$ to images fed into the discriminator. The label smoothing we implemented involves multiplying labels of real images by 0.9 in training the discriminator. From Figure 5c, although the checkerboard artifacts were dampened, finer, visible artifacts remained. Furthermore, the resolution of the generated images remained low.

The final version, which we call DCGAN3, changed stride lengths in transpose deconvolution to 1. This version came from the observation that checkerboard artifacts could arise when the kernel size is not divisible by the stride length. This could cause regions of the feature maps to systemically receive output from a kernel multiple times, creating checkerboard artifacts [8]. We observe from Figure 5e that DCGAN3 was the most effective in minimizing artifacts and maximizing resolution.



(a) DCGAN1 images. (b) DCGAN1 losses.

(c) DCGAN2 images. (d) DCGAN2 losses.

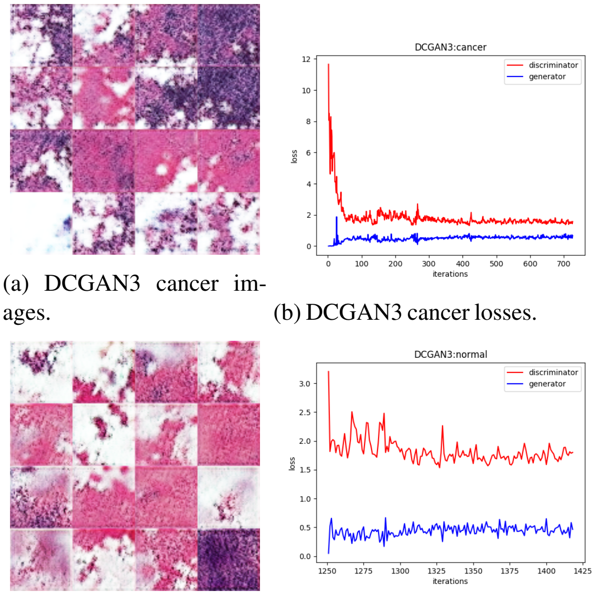
(e) DCGAN3 images. (f) DCGAN3 losses.

Figure 5: Results from three DCGAN versions.

We observe from the losses shown in Figures 5b, 5d, 5f that DCGAN does exhibit training stability because the generator and discriminator losses appear to be converging. Of the three DCGANs, DCGAN3 appear to have the least variance in training loss, providing further support that setting strides to 1 is the correct setting for this breast histology dataset.

The set of cancer and non-cancerous histology images that comprised the final DCGAN-generated synthetic dataset was generated from DCGAN3 trained for 50 epochs on the training set for the cancer prediction classifier. Visualization results are shown in Figure 6.

Compared to Figure 5e, it is observed that additional epochs improved the resolution of the generated images. As with before, the discriminator and generator losses remained stable throughout training and appeared to be converging.



(a) DCGAN3 cancer images. (b) DCGAN3 cancer losses.

(c) DCGAN3 normal images. (d) DCGAN3 normal losses.

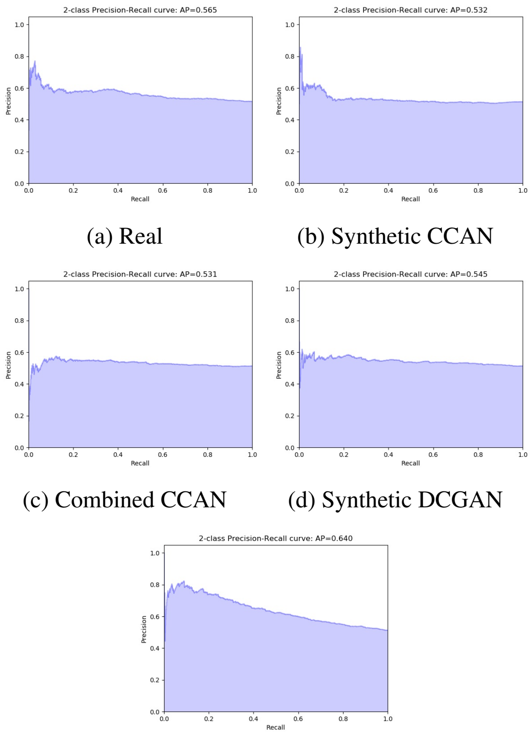
Figure 6: DCGAN3 synthetic dataset.

5. Results

Table 1 compares breast cancer classification performance before and after data augmentation with synthesized images. Classification accuracy was most significantly improved after augmenting the training set with synthesized images from DCGAN3. Augmenting the training set with synthesized images from CCAN only resulted in a minor gain in classification accuracy. Data augmentation with DCGAN improved the precision by $\sim$ 12% but decreased the recall by $\sim$ 15%. With CCAN, the precision improved by $\sim$ 3% and recall dropped by $\sim$ 7%. Accuracy improved by $\sim$ 5% and $\sim$ 1%, respectively. To summarize, data augmentation with DCGAN or CCAN appears to improve the detection rate of non-IDC images but hurt the detection rate of IDC images. Significant challenges remain because recall is often the most important metric to optimize in biomedical applications.

Method	Precision	Recall	Accuracy
Real	0.642	0.856	0.682
CCAN	0.513	1.000	0.513
CCAN and Real	0.675	0.783	0.695
DCGAN	0.495	0.051	0.471
DCGAN and Real	0.760	0.695	0.731

Table 1: Effect of data augmentation on breast cancer classification performance. Metrics are reported based on the validation dataset.



(a) Real

(b) Synthetic CCAN

(c) Combined CCAN

(d) Synthetic DCGAN

(e) Combined DCGAN

Figure 7: Precision recall for each model

Additionally, Table 1 shows that training on synthetic images generated from either CCANs or DCGANs resulted in classifiers with poor classification accuracy. Based on precision and recall metrics, the classifier trained on CCAN images mostly predicted IDC while the classifier trained on DCGAN images mostly predicted non-IDC.

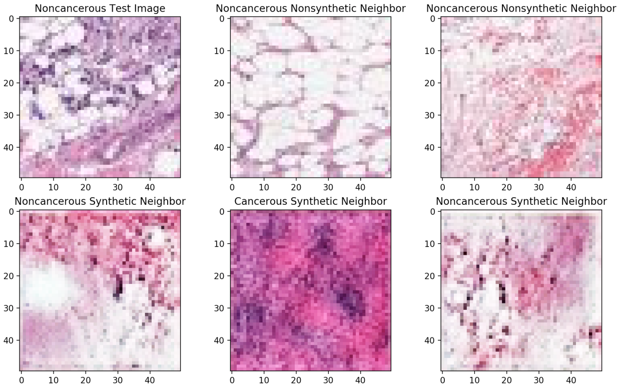

Noncancerous Test Image
Noncancerous Nonsynthetic Neighbor
Noncancerous Nonsynthetic Neighbor
Noncancerous Synthetic Neighbor
Cancerous Synthetic Neighbor
Noncancerous Synthetic Neighbor
0
10
20
30
40

Figure 8: Nearest neighbors of test image correctly classified after data augmentation.

After data augmentation with DCGAN3, precision and accuracy improved compared to the baseline model. In Figure 8 we visualize the nearest neighbors of an image that was incorrectly classified by the baseline classifier but correctly classified after data augmentation. An image's nearest neighbors in the training set should have great influence on its classification. Indeed, the two synthetic cancer images are part of the five nearest neighbors of the test image in Figure 8 and probably influenced its classification.

6. Conclusion

This project demonstrated that it is possible to increase breast histology classification accuracy by augmenting the training set with synthesized images. However, with the admittedly limited number of methods we attempted, most of the improved accuracy was due to improving the true negative detection rate instead of the true positive detection rate. In biomedical settings, it is often more important to have high recall.

Nevertheless, we learned that using DCGANs to generate histology images may be a promising direction for data augmentation to improve breast cancer detection. Additionally, we learned how to improve the existing DCGAN implementation to produce images that more closely resembled histology images, namely choosing a stride length that could divide the kernel size perfectly to reduce checkerboard artifacts. There were several limitations we had to work with in this project, namely lower resolution histology images, limited time to fully explore different GAN architectures, and training time. Although our generator and discriminator errors appear to be converging, in general we found increasing the number of epochs to improve the resemblance between generated images and real images. As Figure 8 shows, our synthetic images were among the closest neighbors to some breast histology images in our validation set. We believe it is possible that with a better dataset, a more comprehensive and principled exploration of different GAN architectures, and more training time, that recall could be improved.

We also learned that our usage of CCANs was not well-suited to augmenting breast histology images. We believe the primary reason for this is that a significant portion of the generated synthetic images were too similar to the original source image. Thus our assumption that the CCAN will map real, non-IDC images to synthetic IDC images and vice versa is partially incorrect. Instead, we believe our training set was augmented with similar images with opposite labels, which is expected to decrease classifier performance. This hypothesis is confirmed empirically - the recall for the model trained on both synthetic CCAN data and real data are slightly worse than the baseline model, as indicated in Table 1. A future direction is to investigate whether retaining significantly different image pairs generated by CCANs could prove to be an effective data augmentation method. This could be implemented using sum of squared distance between the CCAN output image and the source image as a metric for evaluation.

The state-of-the-art GANs are capable of generating images that fool the untrained human eye. However, we believe challenges remain in applying GANs as a way to augment biomedical images. In principle, the generator network needs only to produce images that fool the discriminator, without the requirement to generate images that capture true biological signal. At the same time, the discriminator need not use features specific to biology to distinguish between real and fake images. Even if the generator is capable of capturing true biological signal, there is no guarantee that the generated images contain new signal that could improve the performance of a trained classifier on the entire population of images. Despite these challenges and lack of theory that guarantees success for similar applications, our project demonstrates empirically that synthesizing biomedical images with GANs could improve breast cancer classification accuracy.

References

[1] Towards effective gans for data distributions with diverse modes. *ICLR conference paper*, 2018. 3

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. 2014. arXiv:1406.2661v1 [stat.ML]. 1

[3] J. T. Guibas, T. S. Virdi, and P. S. Li. Synthetic medical images from dual generative adversarial networks. *CoRR*, abs/1709.01872, 2017. 2

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 2

[5] A. Janowczyk. Breast histology images. https://www.kaggle.com/simjeg/lymphoma-subtype-classification-fl-vs-cll, 2014. Classify IDC vs non IDC images. 2

[6] Medical Imaging. Medical imaging — Wikipedia, the free encyclopedia, 2018. [Online; accessed 29-April-2018]. 1

[7] D. Nie, R. Trullo, C. Petitjean, S. Ruan, and D. Shen. Medical image synthesis with context-aware generative adversarial networks. *CoRR*, abs/1612.05362, 2016. 2

[8] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill 1*, 2016. no 10, e3. 4

[9] PyTorch. Torchvision models. https://pytorch.org/docs/master/torchvision/models.html. 2

[10] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. 2016. arXiv:1511.06434v2 [cs.LG]. 1, 2, 4

[11] Shadi Albarqouni. Deep learning papers on medical image analysis. https://github.com/albarqouni/Deep-Learning-for-Medical-Applications#classification. 1

[12] Soumith Chintala. How to train a gan? tips and tricks to make gans work. https://github.com/soumith/ganhacks, 2016. 4

[13] Taehoon Kim. Dcgan in tensorflow. https://github.com/carpedm20/DCGAN-tensorflow, 2018. 4

[14] M. Wernick, Y. Yang, J. Brankov, G. Yourganov, and S. Strother. Machine learning in medical imaging. *IEEE Signal Process Mag*, 27:25–38, 2010. 1

[15] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. 2015. arXiv:1506.03365v3 [cs.CV]. 4

[16] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. 1, 2, 3