Deep Generative Breast Cancer Screening and Diagnosis

Shayan Shams, Richard Platania, Jian Zhang, Joohyun Kim, Kisung Lee, and Seung-Jong Park(✉)

Louisiana State University, Baton Rouge, USA
sjpark@cct.lsu.edu

Abstract. Mammography is the primary modality for breast cancer screening, attempting to reduce breast cancer mortality risk with early detection. However, robust screening less hampered by misdiagnoses remains a challenge. Deep Learning methods have shown strong applicability to various medical image datasets, primarily thanks to their powerful feature learning capability. Such successful applications are, however, often overshadowed with limitations in real medical settings, dependency of lesion annotations, and discrepancy of data types between training and other datasets. To address such critical challenges, we developed DiaGRAM (Deep GeneRAtive Multi-task), which is built upon the combination of Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN). The enhanced feature learning with GAN, and its incorporation with the hybrid training with the region of interest (ROI) and the whole images results in higher classification performance and an effective end-to-end scheme. DiaGRAM is capable of robust prediction, even for a small dataset, without lesion annotation, via transfer learning capacity. DiaGRAM achieves an AUC of 88.4% for DDSM and even 92.5% for the challenging INbreast with its small data size.

1 Introduction

Breast cancer is the most common and fatal cancer among adult women [12]. According to the National Cancer Institute, approximately one in eight women will develop an invasive form of this cancer at some point in their lives [11]. Frequent screenings through mammograms can help detect early signs of breast cancer. However, certain challenges, such as false negatives, unnecessary biopsies, and low screening rate in some rural areas, overshadow the effectiveness of mammogram screening [8,9]. We believe deep learning aided software is a promising direction to achieve highly accurate screening, reducing the number of false negatives and unnecessary biopsies, while at the same time expanding screening capacity and coverage. Deep learning makes this possible by learning hidden features and correlations that might not be visible to humans [5]. Towards this goal, our work aims to provide an end-to-end deep learning system. There are several challenges that we need to overcome.

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 859–867, 2018.
https://doi.org/10.1007/978-3-030-00934-2_95

Firstly, limited training data makes it difficult to achieve highly accurate diagnosis. Secondly, not all data have lesion annotations because making the annotations is a very expensive and time consuming task. Therefore, developing an accurate model that can conduct inference on whole images without annotation is very important. Lastly, it is desirable that models should be robust and adaptable to heterogeneous datasets.

To address these challenges, we propose DiaGRAM (Deep GeneRAtive Multi-task), an end-to-end system that combines a Generative Adversarial Networks (GANs) [4] with discriminative learning using a multi-task learning strategy, to enhance classification performance when training data is limited. We also employ transfer learning to adapt a model trained with one type of data to another.

Generative Adversarial Networks (GANs) are often used to produce data when the analytic form of the data distribution is hard to obtain. Instead of using GAN as a data augmenting device, we use GAN to enhance feature learning. Insights from deep learning show us that features that capture the characteristics of the data, that are learned without label information by unsupervised methods, can still be helpful for discriminative tasks such as classification. For example, stacked autoencoders or deep belief network (DBN) can be used to pre-train the weights of a discriminative model in an unsupervised fashion, then fine-tune the model using the label information. DiaGRAM's design follows this insight with some modification. Rather than taking a two-stage process, DiaGRAM is end-to-end. It extracts features that are good both for the discriminative tasks (i.e., patch and image classification) and for the GAN's generative task (i.e., differentiate the real patches from the generated ones). The latter task ensures that the learned features capture the data characteristics, and thus can help classification, in a way similar to pre-training by autoencoders or DBNs.

Previously, there have been several works related to applying deep learning towards mammogram classification [1–3,10,13,14]. Most of these works focus on either mass segmentation, detection, or classification. A recent survey regarding deep learning in medical imaging analysis mentioned the lack of GAN-based approaches, pointing out the absence of any peer-reviewed papers regarding this subject [7]. Our proposed framework, DiaGRAM, is capable of both mass and whole image classification and inherently agonistic for the mentioned above challenges and thus allows an end-to-end solution for breast cancer screening and diagnosis purposes.

2 Methods

2.1 Model Overview

Figure 1(a) shows our model architecture which consists of four components: generator network, feature extraction network, discriminator network, and extended classification network. The feature extraction network and the extended classification network form a path for mammogram classification. The generator network, the feature extraction network, and the discriminator network form a GAN. (Note that the "discriminator" of the original GAN paper [4] corresponds

to the combination of both our feature extraction network and our discriminator). The main novel feature of our model is that it fuses, using a multi-task learning strategy, part of the image classification path with part of the GAN path to extract features that can help both tasks.

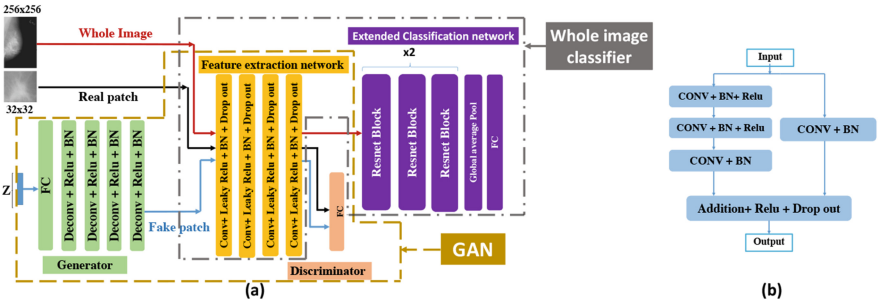

Fig. 1. (a) DiaGRAM architecture (b) Residual block in DiaGRAM

2.2 GAN-Enhanced Deep Classification

Two types of images are considered in our model. One is the whole mammogram images and the other is patches from mammograms. Let $\{(\mathbf{I}_i, \mathbf{t}_i)\}_{i=1}^{N}$ be a collection of N mammogram images ($\mathbf{I}_i$) and their labels ($\mathbf{t}_i$). Some mammogram datasets (such as DDSM) include regions of interest (ROI) on the image. These regions of interest serve as image patches in our learning. Since ROIs may differ in size, we resize them to the same size, $s \times s$. We denote by $\{(\mathbf{C}_j, \mathbf{t}_j)\}_{j=1}^{M}$ a set of M patch images and their labels. In both cases, the label $\mathbf{t}_i$ is an indicator vector (i.e., if the i-th image belongs to class k, the k-th entry of the corresponding label vector has value 1 ($\mathbf{t}_i^{(k)} = 1$) and all other entries have value 0). We describe the components of our model in the following:

Generator: The generator is a deep neural network that takes as input a random vector and produces an image patch. It comprises of one fully connected and four deconvolution layers. We denote by $\mathcal{G}$ the generator network and $\boldsymbol{\theta}_g$ its parameters. Let $\mathbf{z} \in \mathbb{R}^n$ be a random vector whose entries are drawn uniformly in the range $[-1, 1]$. Also, let $\mathcal{G}(\mathbf{z}; \boldsymbol{\theta}_g) \in \mathbb{R}^{s \times s}$ be the size $(s \times s)$ image patch generated. For a set of random vectors $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M\}$, the generator can produce a set of patches $\{\mathcal{G}(\mathbf{z}_1; \boldsymbol{\theta}_g), \mathcal{G}(\mathbf{z}_2; \boldsymbol{\theta}_g), \ldots, \mathcal{G}(\mathbf{z}_M; \boldsymbol{\theta}_g)\}$.

Feature Extraction Network: The purpose of the feature extraction network is to discover features that may be present in both a patch and a whole mammogram image and that can be useful in the classification of both. This is the common component between the GAN and the image classifiers. We employ a four-layered CNN as the feature extraction network. We denote by $\mathcal{F}$ the feature extraction network and $\boldsymbol{\theta}_f$ its parameters. Given an input x, we denote by $\mathcal{F}(x; \boldsymbol{\theta}_f)$ the output (features maps) from the network. The feature extraction

network may take an image $\mathbf{I}$ as input and give output $\mathcal{F}(\mathbf{I};\boldsymbol{\theta}_f)$, or it may take a patch $\mathbf{C}$ (or generated patch $\mathcal{G}(\mathbf{z})$) as input and give output $\mathcal{F}(\mathbf{C};\boldsymbol{\theta}_f)$ (or $\mathcal{F}(\mathcal{G}(\mathbf{z});\boldsymbol{\theta}_f)$). Note that since $\mathbf{C}$ and $\mathcal{G}(\mathbf{z})$ are of the same size, the feature maps of $\mathcal{F}(\mathbf{C};\boldsymbol{\theta}_f)$ and $\mathcal{F}(\mathcal{G}(\mathbf{z});\boldsymbol{\theta}_f)$ have the same size, whereas the feature maps of $\mathcal{F}(\mathbf{I};\boldsymbol{\theta}_f)$ have a size different from them.

Discriminator: The discriminator network takes features produced by the feature extraction network and performs patch classification. It consists of a single fully connected layer that has $m+1$ neurons, where m is the number of classes in the patch images. We denote by $\mathcal{D}$ the network and $\boldsymbol{\theta}_d$ its parameters. The first m neurons of $\mathcal{D}$ are softmax units. Given a patch $\mathbf{C}$, the output from the i-th neuron ($\mathcal{D}^{(i)}(\mathcal{F}(\mathbf{C};\boldsymbol{\theta}_f);\boldsymbol{\theta}_d)$) computes the probability that the patch belongs to class i. Let y be the variable for the patch's label. We have:

$$P(y=i|\mathbf{C}) = \mathcal{D}^{(i)}(\mathcal{F}(\mathbf{C};\boldsymbol{\theta}_f);\boldsymbol{\theta}_d). \tag{1}$$

The $(m+1)$-th neuron is a sigmoid neuron and computes the probability that a patch is from a real image (not generated). We denote its output by $\mathcal{D}^{(m+1)}(\mathcal{F}(x;\boldsymbol{\theta}_f);\boldsymbol{\theta}_d)$ and have:

$$P(r=1|x) = \mathcal{D}^{(m+1)}(\mathcal{F}(x;\boldsymbol{\theta}_f);\boldsymbol{\theta}_d) \tag{2}$$

where x is a patch (real or generated) and r is the variable that takes value 1 if the patch is from a real image and 0 otherwise.

Extended Classification Network: Features produced by the feature extraction network are local features from a small region. Deep CNNs often contain many layers and neurons in higher layers that respond to larger-size features that are constructed from small-size features reacted to by lower layer neurons. We follow the same idea, taking the feature maps produced by the feature extraction network and passing them through more layers of the CNN before the final classification. We call the additional layers the extended classification network. It consists of six Residual network blocks [6] and an output layer that gives the class probability. We denote by $\mathcal{E}$ the extended classification network and by $\boldsymbol{\theta}_e$ its parameters. For a whole image $\mathbf{I}$, the i-th output of $\mathcal{E}$ is the probability that the image belongs to the i-th class:

$$P(y=i|\mathbf{I}) = \mathcal{E}^{(i)}(\mathcal{F}(\mathbf{I};\boldsymbol{\theta}_f);\boldsymbol{\theta}_e) \tag{3}$$

2.3 Training

Our model combines multiple network components together for better feature extraction and classification. To train the model, we employ multiple loss functions. Given a random vector $\mathbf{z}$, the generator loss is:

$$\mathcal{L}_g(\mathbf{z}) = -logP(r=1|\mathcal{G}(\mathbf{z};\boldsymbol{\theta}_g)) \tag{4}$$

Our discriminator performs two tasks and thus involves two losses: the loss for distinguishing the real patches from the generated ones and the loss for patch

classification. Given a patch $\mathbf{C}$ and a random vector $\mathbf{z}$, the loss for distinguishing the real from the generated is:

$$\mathcal{L}_d(\mathbf{C}, \mathbf{z}) = -\big[\log P(r = 1|\mathbf{C}) + \log P(r = 0|\mathcal{G}(\mathbf{z}; \boldsymbol{\theta}_g))\big]. \tag{5}$$

For patch classification, we use the cross-entropy loss. Given a patch $\mathbf{C}$ and its label indicator vector $\mathbf{t}$, the loss is as follows:

$$\mathcal{L}_c(\mathbf{C}, \mathbf{t}) = -\sum_k \mathbf{t}^{(k)} \log P(y = k|\mathbf{C}) \tag{6}$$

Finally the cross-entropy loss for whole image classification, given an image $\mathbf{I}$ and its label indicator vector $\mathbf{t}$, is:

$$\mathcal{L}_i(\mathbf{I}, \mathbf{t}) = -\sum_k \mathbf{t}^{(k)} \log P(y = k|\mathbf{I}) \tag{7}$$

The overall training process is presented in Algorithm 1. During a training iteration, we update the parameters of the model components using stochastic gradient descending on the related losses.

Algorithm 1. Training algorithm

for number of training iterations **do**
  for k *steps* **do**
    $S_C \leftarrow$ *Sample a minibatch of m patches*
    $S_I \leftarrow$ *Sample a minibatch of n images*
    $S_z \leftarrow$ *Sample a minibatch of m random vectors*
    Update the feature extract network and the discriminator by descending on their parameter gradients:

$$\nabla_{(\boldsymbol{\theta}_f, \boldsymbol{\theta}_d)} \frac{1}{m} \Big(\sum_{\mathbf{C} \in S_C, \mathbf{z} \in S_z} \mathcal{L}_d(\mathbf{C}, \mathbf{z}) + \sum_{(\mathbf{C}, \mathbf{t}) \in S_C} \mathcal{L}_c(\mathbf{C}, \mathbf{t}) \Big)$$

    Update the feature extract network and the extended classifier by descending on their parameter gradients:

$$\nabla_{(\boldsymbol{\theta}_f, \boldsymbol{\theta}_e)} \frac{1}{n} \sum_{(\mathbf{I}, \mathbf{t}) \in S_I} \mathcal{L}_i(\mathbf{I}, \mathbf{t})$$

  end for
  $S_z \leftarrow$ *Sample a minibatch of m random vectors*
  Update the generator by descending on its parameter gradient:

$$\nabla_{\boldsymbol{\theta}_g} \frac{1}{m} \sum_{\mathbf{z} \in S_z} \mathcal{L}_g(\mathbf{z})$$

end for

2.4 Transfer Learning

Digital mammography has been widely adopted in modern hospitals, providing a clearer image in comparison with the film mammography of the past. For example, INbreast is a digital mammography dataset. To build an accurate model for small-size datasets such as INbreast, we utilize transfer learning. We train a DiaGRAM model using a larger dataset with region annotations (DDSM). Then, we take out the classification path (the feature extraction and the extended classification networks) from the model, fine-tune it in a supervised mode with INbreast training data, and use it as a classifier for INBreast data.

3 Experiments and Results

In this section, we present the experimental results of DiaGRAM for the DDSM and INbreast datasets and discuss the benefit of combining the GAN with discriminative learning using a multi-task learning strategy. For fair comparisons, we use 5-fold cross validation to evaluate DiaGRAM. The reported AUC is the result from 5-fold cross validation.

Since the DDSM dataset is used for multi-task learning, we use annotated lesion and whole mammogram images, which are 3,500 images in total, divided into cancer and benign. We utilize several common data augmentation methods to reduce over-fitting and improve overall accuracy. For instance, we rotate and mirror images across the y-axis randomly. We use the overlay files to extract the region of interests, which have various shapes. We crop the smallest possible square that can fully contain a ROI and resize it to 32×32. Thus, we generate 25,000 cropped images of ROIs. For the INbreast dataset, we convert BI-RADS 4, 5, and 6 to cancerous samples and 1 and 2 to negative samples. Since it is not clear that BI-RADS 3 samples are benign or cancerous, we exclude 23 mammograms, which were labeled as BI-RADS 3.

Since the INbreast dataset is not large enough to train a model from scratch, we use transfer learning, which is explained in Sect. 2.4, and fine-tune DiaGRAM

Table 1. Comparison with other works for whole image classification.

Paper	End-to-end	Dataset	Accuracy	AUC
Ball and Bruce [1]	✗	DDSM	87%	N/A
Varela et al. [13]	✗	DDSM	81%	N/A
Domingues et al. [3]	✗	INbreast	89%	N/A
Dhungel et al. [2]	✗	INbreast	$(95 \pm 5)\%$	(91 ± 12)
Dhungel et al. [2]	✓	INbreast	$(91 \pm 2)\%$	(76 ± 23)
Zhu et al. [14]	✓	INbreast	$(90 \pm 2)\%$	$(89 \pm 4)\%$
DiaGRAM	✓	DDSM	$89 \pm 3.4\%$	$88.4 \pm 2.9\%$
	✓	INbreast	$93.5 \pm 2.9\%$	$92.5 \pm 2.4\%$

for 20 epochs using the dataset. In Table 1, the best results of previous works using DDSM or INbreast are reported. DiaGRAM achieves a mean AUC of 92.5% and 88.4% for INbreast and DDSM datasets, respectively, and provides superior AUC and accuracy over other previous works for both datasets. ROC curves for both datasets are plotted in Fig. 2.

3.1 Performance Enhanced by GAN

To investigate whether the GAN is effective in enhancing classification performance, we created a model variant that does not include GAN and compare the performance of DiaGRAM to that of the variant. The variant without GAN contains the feature extraction network, the discriminator (without the neuron that outputs the probability whether a patch is real or generated), and the extended classification network. It performs two tasks: patch classification (combining the feature extraction network and the discriminator) and whole image classification (combining the feature extraction network and the extended classification network). The variant was trained in a multi-task learning fashion using the losses in Eqs. 6 and 7.

As shown in Fig. 3, the model variant without GAN suffered a drop of 2.9% on AUC (85.5% compared to DiaGRAM's 88.4%) for the DDSM dataset. This indicates that having the GAN in the model indeed contributes to the model's high performance. It demonstrates that the task of discriminating fake data from real data can be leveraged to learn latent and hidden features that will improve classification performance.

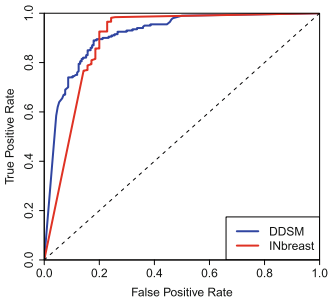

Fig. 2. ROC curves for DDSM and INbreast.

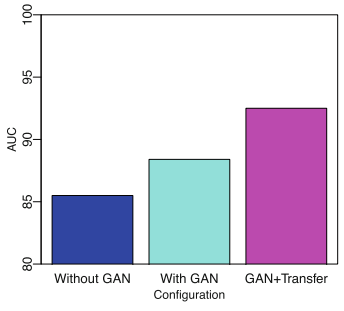

Fig. 3. AUC for different configurations.

4 Conclusion

In this work, we introduced DiaGRAM (Deep GeneRAtive Multi-task), an end-to-end deep learning solution for breast cancer screening and diagnosis purposes. DiaGRAM employs two main approaches to achieve highly accurate mammogram diagnosis: (1) it combines a GAN with a deep classifier to learn features

that benefit both, (2) and transfer learning is used to adapt the model trained with one type of data to another. We conducted a set of experiments using the DDSM and the INbreast datasets. The results showed better performance of DiaGRAM on both the accuracy and the AUC measures when compared to prior works. DiaGRAM also demonstrated transfer learning capacity as the model trained on DDSM dataset and adapted to the INbreast dataset showed good performance. In future works, we plan to extend the techniques used in this paper for real medical settings, focusing on usabilities for screening and diagnosis procedure.

Acknowledgments. This work was partially funded by NIH grants (P20GM103458-10, P30GM110760-03, P20GM103424), NSF grants (MRI-1338051, IBSS-L-1620451, SCC-1737557, RAPID-1762600), LA Board of Regents grants (LEQSF(2016-19)-RD-A-08 and ITRS), and IBM faculty awards.

References

1. Ball, J.E., Bruce, L.M.: Digital mammographic computer aided diagnosis (CAD) using adaptive level set segmentation. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2007, pp. 4973–4978. IEEE (2007)
2. Dhungel, N., Carneiro, G., Bradley, A.P.: The automated learning of deep features for breast mass classification from mammograms. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 106–114. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_13
3. Domingues, I., et al.: Inbreast-database masses characterization. XXIII CBEB (2012)
4. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
5. He, K., et al.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
6. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Litjens, G.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
8. Ong, M.S., Mandl, K.D.: National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at $4 billion a year. Health Aff. **34**(4), 576–583 (2015)
9. Orwat, J.: Comparing rural and urban cervical and breast cancer screening rates in a privately insured population. Soc. Work Publ. Health **32**(5), 311–323 (2017)
10. Platania, R., et al.: Automated breast cancer diagnosis using deep learning and region of interest detection (BC-DROID). In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 536–543. ACM (2017)
11. Siegel, R.: Cancer statistics, 2014. CA Cancer J. Clin. **64**(1), 9–29 (2014)
12. Teh, Y.C.: Opportunistic mammography screening provides effective detection rates in a limited resource healthcare system. BMC Cancer **15**(1), 405 (2015)

13. Varela, C.: Use of border information in the classification of mammographic masses. Physics Med. Biol. **51**(2), 425 (2006)
14. Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 603–611. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_69