2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

SEMI-SUPERVISED LEARNING WITH GENERATIVE ADVERSARIAL NETWORKS FOR CHEST X-RAY CLASSIFICATION WITH ABILITY OF DATA DOMAIN ADAPTATION

Ali Madani, Mehdi Moradi, Alexandros Karargyris, Tanveer Syeda-Mahmood*

IBM Research, Almaden Research Center - San Jose, CA
mmoradi@us.ibm.com

ABSTRACT

Deep learning algorithms require large amounts of labeled data which is difficult to attain for medical imaging. Even if a particular dataset is accessible, a learned classifier struggles to maintain the same level of performance on a different medical imaging dataset from a new or never-seen data source domain. Utilizing generative adversarial networks in a semi-supervised learning architecture, we address both problems of labeled data scarcity and data domain overfitting. For cardiac abnormality classification in chest X-rays, we demonstrate that an order of magnitude less data is required with semi-supervised learning generative adversarial networks than with conventional supervised learning convolutional neural networks. In addition, we demonstrate its robustness across different datasets for similar classification tasks.

***Index Terms*—** semi-supervised learning, chest x-ray.

1. INTRODUCTION

Medical imaging presents a unique set of difficulties when attempting to develop deep learning algorithms. Due to privacy laws, healthcare industry standards, and the lack of integration of medical information systems, data is not as abundant as other fields of computer vision. While efforts to alleviate these issues are ongoing [1, 2], as of now they hamper the speed of innovation of deep learning algorithms as they inherently require large amounts of data for tasks such as image classification [3] or semantic segmentation [4].

In many cases even if data is available, it is unstructured or lacks proper labeling. To address this, one would annotate the medical images which proves to be an expensive and time-consuming process. Often, one can only feasibly label a small portion of the images while having a much larger portion of unlabeled images. For supervised learners such as a convolutional neural network, only the small labeled portions can be utilized for training. In the past, we have approached the problem of data scarcity, particularly the lack of samples in disease categories at the time of training, by using the available samples from the normal class to train a segmentation network. The features produced by this segmentation model are used along with the whole image in training a disease classification network. This is a way of learning the distribution of data in one class, and taking advantage of it in distinguishing that class from others [5]. This concept of "learning normal" as a way to improve disease classification is recently also applied using a generative model [6].

In addition, many deep learned classifiers tend to overfit to a particular data domain source. For any given image classification task in medical imaging, one strives to train a classifier that separates images based on the structural or physiological variations that define the target classes. However, there are other sources of variance such as scanner type and imaging protocol that can differentiate images. As a result, when deep learned classifiers are trained on a particular training dataset and then tested in production on data from a different domain source, there is usually a drop in performance.

In this study, we address both issues of labeled data scarcity and data domain variance with generative adversarial networks (GANs) [7]. GANs utilize two networks, a discriminator and a generator, involved in a minimax game to find the Nash equilibrium of these two networks. In short, the generator seeks to create as realistic images as possible and the discriminator seeks to distinguish between data that is real vs generated as shown in Fig 1.

Fig. 1. Generative Adversarial Network Framework

Generative adversarial networks have been developed for a variety of computer vision tasks in the field of medical imaging. Most of the applications revolve around image synthesis [8, 9] and segmentation [10, 11] in addition to tasks in image de-noising [12]. Before GANs, there were some results utilizing deep learning frameworks for semi-supervised learning [13]. In this study, we are particularly interested in the development and application of GANs for semi-supervised

978-1-5386-3636-7/18/$31.00 ©2018 IEEE

learning in line with advances for digit recognition [14, 15]. We also show the use of the semi-supervised framework to handle domain variability in medical imaging with GANs in contrast to other techniques that use non-generative adversarial training algorithms [16, 17].

The contribution of this paper is in proposing an architecture and learning algorithm that converts a GAN into a semi-supervised classifier particularly for disease detection in chest X-ray images, trained on a fairly small size dataset of annotated images. We show a massive reduction of effort in annotation using this architecture as compared with a traditional convolutional neural network (CNN). We also show that the resulting model, is more tolerant when data from a new source is presented for classification.

2. MATERIALS AND METHODS

Our goal here is building a discriminator that separates images depicting disease instances from normal chest X-rays. For this purpose, we propose an architecture where we utilize GANs in semi-supervised training. An unsupervised GAN with converged generator produces new images from its implicitly learned manifold. Here, we propose a semi-supervised GAN-based architecture that is developed by adapting the generator to take advantage of both labeled and unlabeled data, as shown in Fig. 2 [18]. As the model converges, the discriminator separates generated from real disease, or real normal images. As a result, both labeled and unlabeled data can contribute to the convergence of the model. This is useful for scenarios where there is a small amount of labeled data and large amount of unlabeled data.

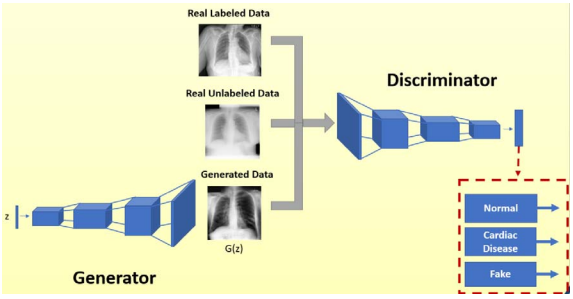

Fig. 2. Semi-supervised GAN-based Architecture.

2.1. Data

We focused on identifying cardiac abnormalties in chest X-ray. Two datasets of chest x-ray were used in this study: one from the National Institute of Health (NIH) prostate, lung, colorectal, and ovarian (PLCO) cancer dataset [19]; the other from the NIH Chest X-Ray collection from Indiana University [20]. In the NIH PLCO dataset, there are around 196,000 X-ray digital images of which we had a subset of around 36,000 frontal chest X-rays. A subset of 4500 images were used with labels of normal or abnormal which were subsequently rescaled to 128x128 pixels and histogram equalized as shown in Fig 3. The abnormal class samples were originally tagged for any patient with cardiomegaly, congestive heart failure, or cardiac abnormality. We took a subsample of 100 images to confirm correct ground-truthing with a trained radiologist. The second dataset, from Indiana University, was used to examine performance of deep learned classifiers on different data source domains. We utilized 400 cases of normal chest x-rays and 313 cases of cardiomegaly (as the abnormal class). Worth noting, the full NIH PLCO dataset includes more disease variance than exclusively cardiomegaly. The datasets were split into 3 groups: training, validation, and testing in a 80:10:10 ratio.

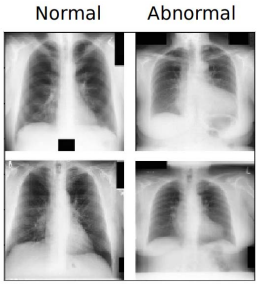

Fig. 3. Sample chest X-ray images from NIH PLCO dataset. Left column corresponds to normal chest X-rays; right column are abnormal chest x-rays due to cardiovascular abnormality.

2.2. Unsupervised GAN for generating chest X-ray image samples

In the basic GAN architecture (Fig. 1), the generator, G, takes a vector $\mathbf{z}$, sampled from random Gaussian noise or conditioned with structured input, and transforms it to $\mathbf{p}_G = G(\mathbf{z})$ to mimic the data distribution, p_{data}. Batches of the generated images and real images are sent to the discriminator, D, where it assigns a label 0 for real or a label 1 for generated. The cost of the discriminator, $J^{(D)}$, and the generator, $J^{(G)}$, are as follows:

$$J^{(D)} = -\frac{1}{2}\mathbf{E}_{x\sim p_{data}} \log D(x) - \frac{1}{2}\mathbf{E}_z \log(1 - D(G(x)))$$

$$J^{(G)} = -\frac{1}{2}\mathbf{E}_z \log D(G(x))$$

Our GAN implementation is a modified version of [21]. In this architecture, the generator is fed a 100x1 input vector which is projected and rescaled. There are then four convolutional layers with 2D-upsampling layers interlaced in between to scale to the appropriate 128x128 image size. To avoid sparse gradients, most non-linear activations were

applied with the leaky rectified linear unit (ReLU) function which has a small negative slope for the negative domain. The discriminator network is a similar network with a series of convolutions with stride of 2 to replace the need for max-pooling. Dropout is used for regularization and leaky ReLU is again used except for the final one node activation with a sigmoid function. A set of normal images were used to train a GAN to produce samples of normal images. A second GAN was trained using only abnormal training data to produce abnormal examples. Each GAN was trained for 500 epochs. This network was used only to show that a GAN can be used to produce realistic chest X-ray images.

2.3. Proposed Semi-supervised GAN-based Classifier

A GAN architecture was developed for semi-supervised training as shown in Fig 2. The main difference of the semi-supervised GAN with the unsupervised one is the structure of the loss function to incorporate both labeled and unlabeled real data. The loss function can be divided into three parts. The output layer of the discriminator has K+1 classes, where K=2 for normal and abnormal, and the (K+1) class is for generated images. The loss function (L) is defined for each type of data ($L_{labeled}$, $L_{unlabeled}$, $L_{generated}$) separately and the total loss is used in optimization. In the notation below, x corresponds to an image, y corresponds to the label, p_{data} is the real data distribution, G is the generator, and $p_{model}(.|.)$ is the predicted class probability. These are defined as below:

$$
\begin{aligned}
L &= L_{labeled} + L_{unlabeled} + L_{generated} \\
L_{labeled} &= -\mathbf{E}_{x,y\sim p_{data}} \log p_{model}(y|x, y < K+1) \\
L_{unlabeled} &= -\mathbf{E}_{x\sim p_{data}} \log(1 - p_{model}(y = K+1|x) \\
L_{generated} &= -\mathbf{E}_{x\sim G} \log p_{model}(y = K+1|x)
\end{aligned}
$$

As the loss function for unlabeled data shows, these samples can be classified as any of the K classes of interest (K=2 here) and contribute to loss when they are classified as generated class K+1. As a result, this architecture allows the un-labeled real data to contribute to learning, reducing the amount of labeling effort required to achieve a certain level of accuracy.

For our experiments, we used rescaled images to 32x32 pixels and incorporated virtual batch normalization. Virtual batch normalization designates a portion of the data as a reference batch to use in conjunction with the current example in its calculation of normalization statistics [15].

2.3.1. Cardiac Abnormality Classification

We utilized the proposed semi-supervised architecture for cardiac abnormality classification task in chest X-ray. The goal was to examine the performance of the GAN with only a portion of the data as labeled. In parallel, a CNN was trained on the labeled portion of the data to compare accuracy. Although the architectures were different, we attempted to keep the number of the layers and parameters similar for both CNN and GAN. The amount of labeled data used for the experiments were 10, 25, 100, 250, 500, 1000, and 2000 images of each class.

2.3.2. Integration Across Different Data Domains

The semi-supervised GAN-based classifier was also used for testing the performance across different data domains. For these experiments, we designate the NIH PLCO data as "Dataset 1" and Indiana University data as "Dataset 2". In these experiments, a CNN was first trained on Dataset 1, then tested on a never-seen Dataset 2. The accuracy was recorded. Then, the GAN was trained only on the Dataset 1, then tested on the Dataset 2. Lastly, the GAN was trained on both datasets by treating the Dataset 1 as labeled and the Dataset 2 as unlabeled. After training, the architecture was tested on the Dataset 2.

3. RESULTS

3.1. Generated Chest X-Ray Images

We are able to produce images that resemble chest X-rays from a qualitative perspective. Afterwards, we randomly sampled vectors from a normal distribution to be fed-forward through the generator network. As shown in Fig 4, the sampled images capture the global structural elements such as the lungs, spine, and heart along with local visual signatures such as the ribs, aortic arch, and the unique curvature of the lower lungs.

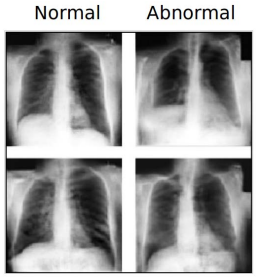

Fig. 4. Sample generated chest X-rays by generator of GAN

3.2. Performance of the Proposed Semi-supervised GAN-based Method for Varying amounts of Unlabeled Data

For the semi-supervised GAN-based model, we examine the performance for image classification tasks when labeled data is scarce. In Fig. 5 and Table 1, we can see that the GAN requires an order of magnitude fewer labeled samples to achieve comparable results. For example, the proposed semi-supervised model only needs 10 labeled images for each class to achieve an accuracy of 73.08%– an accuracy

Labeled Data #	CNN Accuracy	GAN Accuracy
10	51.27%	73.08%
25	51.27%	75.52%
100	57.81%	79.58%
250	71.23%	81.51%
500	78.69%	84.25%
1000	80.80%	84.57%
2000	83.12%	85.10%

Table 1. When labeled data is limited, the semi-supervised GAN-based network requires an order of magnitude less labeled training data to achieve comparable performance to a supervised CNN classifier.

that requires somewhere between 250 to 500 labeled images for a conventional CNN.

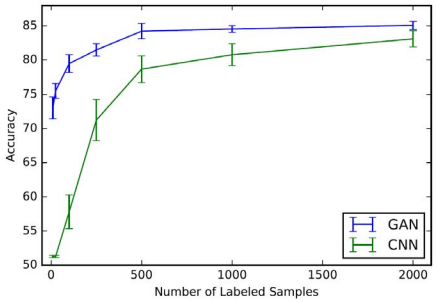

Fig. 5. Accuracy performance of semi-supervised GAN-based method vs supervised CNN for varying amounts of labeled samples for training.

3.3. Addressing Data Source Domain Overfitting

The last set of experiments involve testing the performance of learned classifiers on different data domains. Assuming each dataset has its associated biases from data collection, we usually observe a drop in performance when tested on a new dataset. When trained on 80% of Dataset 1, a CNN is able to achieve 81.93% on a held-out 10% test set from Dataset 1. But when the same model is tested on all of Dataset 2, the accuracy drops to 57.8%, a hallmark of overfitting. Our proposed GAN-bases semi-supervised model under the same training scenario is more robust as it only drops to 76.4% in accuracy on Dataset 2. The most impressive was the last training scenario where the semi-supervised model was trained on 100% of Dataset 1 with labels and 80% of Dataset 2 without labels. In this context, we were able to achieve an accuracy of 93.7% when tested on a 20% held-out from Dataset 2. Without any labeling of Dataset 2, the classifier is able to achieve high accuracy. *This provides a low cost solution for model adaptation to a new data source as it removes the need for labeling of the data from the new source prior to re-training.*

Training Scenario	Dataset 2 Accuracy
CNN model trained on dataset 1	57.8%
GAN model trained on dataset 1	76.4%
Semi-supervised GAN-based net trained on labeled dataset 1 and unlabeled dataset 2	93.7%

Table 2. GAN training is more robust to overfitting to domain source artifacts than vanilla CNNs.

4. DISCUSSION AND CONCLUSIONS

Generative adversarial network training provides unique advantages for common problems in medical imaging. In this study, we validate that deep generative adversarial networks are able to learn the visual structure in medical imaging domains (particularly chest X-rays). Generated samples from the generator network present both the global and local structure that define particular classes of chest X-rays. Next, we proposed a semi-supervised architecture of GANs that is capable of learning form both labeled and unlabeled images. As shown in the results, annotation effort is reduced considerably to achieve similar performance through supervised training techniques. We attribute this to the GAN being able to learn structure in the unlabeled data in an unsupervised learning fashion – which significantly offsets the low number of labeled data samples. Lastly, we show that the semi-supervised GAN-based is robust to data source domain issues. When models are trained on one dataset and tested on a different dataset, a supervised CNN shows a larger drop in accuracy compared to a semi-supervised GAN-based classifier. If re-training of the models in the new domain is feasible, one can use unlabeled data from the second domain with the semi-supervised model; whereas for a CNN one needs to go through the costly process of labeling.

A limitation of the current study is the small number of abnormalities within our data. The more challenging problem of tackling multi-label disease classification in the context of chest X-ray, using the semi-supervised approach is among our current research goals.

5. REFERENCES

[1] Yaniv Gur et al., "Towards an efficient way of building annotated medical image collections for big data studies," in *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS Workshop)*, 2017, pp. 87–95.

[2] Mehdi Moradi, Yufan Guo, Yaniv Gur, M Negahdar,

and Tanveer Syeda-Mahmood, "A cross-modality neural network transform for semi-automatic medical image annotation," in *International Conference on Medical Image Computing and Computer-Assisted Interventions*, 2016, pp. 300–307.

[3] Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *npj Digital Medicine*, 2018.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[5] Ken C L Wong, Alexandros Karargyris, Tanveer Syeda-Mahmood, and Mehdi Moradi, "Building disease detection algorithms with very small numbers of positive samples," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2017, pp. 471–479.

[6] Thomas Schlegl, Philipp Seebck, Sebastian Waldstein, and Georg Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging (IPMI)*, 2017, pp. 146–157.

[7] Ian Goodfellow and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[8] Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum, "Deep mr to ct synthesis using unpaired data," in *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2017, pp. 14–23.

[9] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood, "Chest x-ray generation and data augmentation for cardiovascular abnormality classification," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2018.

[10] Pim Moeskops, Mitko Veta, Maxime W Lafarge, Koen AJ Eppenhof, and Josien PW Pluim, "Adversarial training and dilated convolutions for brain mri segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 56–64. Springer, 2017.

[11] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 408–416.

[12] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Isgum, "Generative adversarial networks for noise reduction in low-dose ct," *IEEE Transactions on Medical Imaging*, 2017.

[13] Christoph Baur, Shadi Albarqouni, and Nassir Navab, "Semi-supervised deep learning for fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 311–319.

[14] Augustus Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[15] Tim Salimans et al., "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[16] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al., "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 597–609.

[17] Maxime W Lafarge, Josien PW Pluim, Koen AJ Eppenhof, Pim Moeskops, and Mitko Veta, "Domain-adversarial neural networks to address the appearance variability of histopathology images," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 83–91. Springer, 2017.

[18] Diederik P Kingma et al., "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems*, 2014, pp. 3581–3589.

[19] Martin M Oken et al., "Screening by chest radiograph and lung cancer mortality: the prostate, lung, colorectal, and ovarian (plco) randomized trial," *Jama*, vol. 306, no. 17, pp. 1865–1873, 2011.

[20] Dina Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2015.

[21] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.