Skin Lesion Classification Using GAN based Data Augmentation

Conference Paper *in* Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference · July 2019

DOI: 10.1109/EMBC.2019.8857905

3 authors:

Haroon Rashid
National University of Sciences and Technology
2 PUBLICATIONS **13** CITATIONS
SEE PROFILE

M. Asjid Tanveer
National University of Sciences and Technology
7 PUBLICATIONS **38** CITATIONS
SEE PROFILE

Hassan Aqeel Khan
National University of Sciences and Technology
10 PUBLICATIONS **43** CITATIONS
SEE PROFILE

Skin Lesion Classification Using GAN based Data Augmentation

Haroon Rashid[1], M. Asjid Tanveer[1] and Hassan Aqeel Khan[1]

***Abstract*— Early detection and frequent monitoring are critical for survival of skin cancer patients. Unfortunately, in practice a significant number of cases remain undetected until advanced stages, reducing the chances of survival. An appealing approach for early detection is to employ automated classification of dermoscopic images acquired via low-cost, smartphone-based hardware. By far, the most successful classification approaches on this task are based on deep learning. Unfortunately, most medical image classification tasks are unable to leverage the true potential of deep learning due to limited sizes of training datasets. Investigation of novel data generation techniques is thus an appealing option since it can enable us to augment our training data by a large number of synthetically generated examples. In this work, we investigate the possibility of obtaining realistic looking dermoscopic images via generative adversarial networks (GANs). These images are then employed to augment our existing training set in an effort to enhance the performance of a deep convolutional neural network on the skin lesion classification task. Results are compared with conventional data augmentation strategies and demonstrate that GAN based augmentation delivers significant performance gains.**

I. INTRODUCTION

Deep learning has enabled the resolution of complex learning problems that conventional, rule-based approaches have struggled with [8]. Algorithms based on deep learning have approached human-level performance on a number of complex computer vision and image classification tasks. For example deep learning algorithms were shown to match the performance of 21 board certified dermatologists on a skin cancer classification task in [3]. Consequently, deep learning techniques are currently being widely used in the medical imaging domain for various applications like disease identification. However, deep architectures require a large number of training examples to learn useful representations. Unfortunately, unlike other applications, generation of large-scale medical imaging datasets (for supervised learning) is an expensive and time-consuming process since it requires specialized equipment and trained medical practitioners for acquisition and labelling. Consequently, the size of the training set becomes a bottleneck that prevents us from tapping the true potential of deep learning in medical imaging applications. One such example of limited data is the ISIC 2018 Lesion Diagnosis Challenge [11] where only 10,000 total samples are available, belonging to seven different classes. This number is quite small in comparison to the millions of images used to train deep neural networks for natural image classification tasks. In limited data settings the most popular option is to fine-tune an existing deep learning architecture. If however, the data generation process is easy to replicate then it can be used to generate a substantial amount of synthetic data to train a network from scratch. For example in [7], synthetically generated text data is used for classification of text in natural scenes. Unfortunately, generation of synthetic medical images is not easy. In this context, the emergence of Generative Adversarial Networks (GANs) [5] can prove to be very helpful due to their ability to learn the underlying data distribution and generate realistic looking images. If the GAN generated images are high quality then we should be able to use them for training deep learning models and obtain substantial performance gains over transfer learning based approaches. Here we test this hypothesis using the ISIC skin lesion classification task as an example.

As with any image classification task use of large volumes of training delivers significant performance boosts on the skin lesion classification task. This is evident from the top four entries to the ISIC 2018 lesion classification challenge [6] all of which augmented the training data using images from external sources. For the learning algorithm these approaches applied transfer learning to DenseNet and Resnet architectures. We therefore, use fine-tuned DenseNet and Resnet architectures as our baseline models and compare their performance with a CNN trained, from scratch, on a dataset augmented by synthetic images generated via a GAN trained to learn the underlying data distribution of skin lesion images. There have been some prior efforts to apply GANs to skin lesion. For example, in [1, 2] GANs were used to generate highly realistic images of skin lesions. However, these works are focused on generation and synthesis of realistic images and have not demonstrated how much (if any) performance gain can be achieved by augmenting the training data with these synthetic images. GAN generated data has been employed to achieve performance gains on a liver lesion classification task in [4]. This work is therefore, one of the first efforts to formally gauge performance gains obtained via GAN based augmentations on a skin lesion classification task.

II. DATASET

The dataset used for our experiments has been provided by International Skin Imaging Collaboration as a part of ISIC 2018 challenge [2]. The goal of the challenge is

[1]H. Rashid, M. Tanveer & H. Khan are with the Department of Electrical Engineering, School of Electrical Engineering & Computer Science, National University of Sciences & Technology, Sector H-12, Islamabad, Pakistan. 14beehrashid@seecs.edu.pk, 14beemtanveer@seecs.edu.pk, hassan.aqeel@seecs.edu.pk

This work was generously supported by NVIDIA Corporation's GPU grant program. The authors are grateful to NVIDIA for donating a TitanX GPU to support our research.

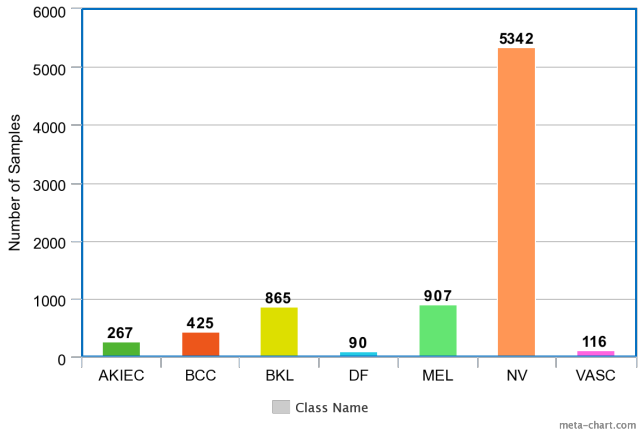


Fig. 1. Histogram of all classes in the ISIC skin lesion classification dataset.

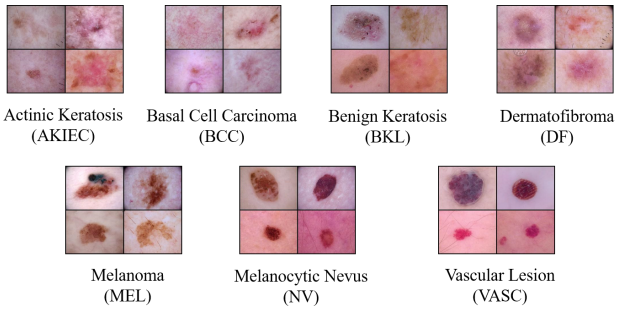


Fig. 2. Samples images from the seven skin lesion classes.

to perform the automatic diagnosis of melanoma from dermoscopic images. The challenge is composed of three parts: (a) Lesion Segmentation (b) Detection & Localization of Visual Dermoscopic Features/Patterns and (c) Disease Classification. We focus on the third task where the objective is to classify Lesion images into one of the seven categories: Melanoma (MEL); Melanocytic Nevus (NV); Basal Cell Carcinoma (BCC); Actinic Keratosis (AKIEC); Benign Keratosis (BKL); Dermatofibroma (DF); Vascular Lesion (VASC). Our training dataset contains a total of 8,000 training images from all seven categories. A histogram indicating the distribution of training images is shown in Fig. 1. Some representative images from all seven classes are shown in Fig. 2. An additional set of 2,000 images with similar class distribution is employed as a test set for performance evaluation. The images in the dataset consists of images at multiple resolutions. Resolution heterogeneity and class imbalance make this a challenging dataset to classify correctly. The test images are selected randomly however, we ensure that the class-distribution remains preserved and the proportion of images from each class is similar in both the train and test sets.

III. Model Architecture and Training

The classification model used in this work is based on the Generative Adversarial Networks (GANs) [5] which have become popular recently due to their ability to generate excellent synthetic data. GANs belong to the generative class of models and are typically used to either (explicitly) learn, or sample from, complex probability distributions of image data. A GAN works by training two networks simultaneously. The first, Generator network, produces synthetic images (typically by learning the underlying data distribution). The second network is referred to as the Discriminator and its objective is to estimate the probability that an input sample is synthetic (i.e., came from the generator) or real. The intuition behind choosing GANs as the basis of our proposed solution is to model the underlying distribution of data and then use it to discriminate between seven target categories with very small inter-class variation. The generator is a deconvolutional network that takes in noise and produces an image of a specific resolution. The discriminator, on the other hand, is a Convolutional Neural Network whose job is to perform binary classification of an input image as fake or real. In our setting, the discriminator also acts as a classifier that learns to distinguish between seven target classes. This results in a total of $n+1(=8)$ units in the output layer of discriminator where n is the number of target classes. The objective during the training is for the generator to maximize the probability of the discriminator network making a mistake. This turns the optimization problem into a two-player mini-max game where the generator is trying to fool the discriminator and discriminator is getting better and better to discriminate between real and fake data. For such an optimization problem, it has been shown theoretically [5] that a unique solution exists with generator recovering the data distribution whereas discriminator outputs a probability 0.5 for both real and fake data.

Our objective is to learn the generator's distribution p_g over the data $\mathbf{x}$. The generator, $G(\mathbf{z};\mathbf{W}_g)$, is a neural network with parameters $\mathbf{W}_g$ and input noise vector $\mathbf{z}$ with prior distribution $p_{\mathbf{z}}(\mathbf{z})$. The discriminator $D(\mathbf{x};\mathbf{W}_d)$ is a convolutional neural network whose output is an $(n+1)$ dimensional vector indicating the probability whether its input vector, $\mathbf{x}$, comes from one of the $n(=7)$, skin lesion classes, or from the generator distribution p_g. During training the discriminator, D, attempts to maximize the probability of assigning the correct label to its input $\mathbf{x}$ by trying to differentiate between images synthetically generated by sampling p_g and those taken from the actual data samples. At the same time, the generator, G, is trained to mislead the discriminator by attempting to minimize $log(1-D(G(\mathbf{z})))$. Both these objectives can be achieved by optimizing the following value function [5]:

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

The block diagram of the GAN based system employed for classification is shown in Fig. 3. It is highlighted that the above function is designed only to differentiate between synthetic and real images whereas, our ultimate objective is to be able to differentiate between the different classes within the images. Therefore, the above loss function is employed

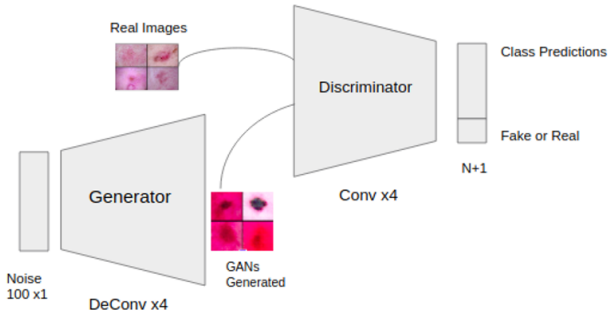
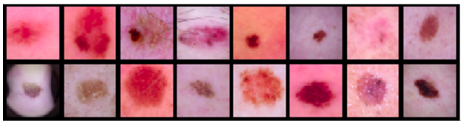
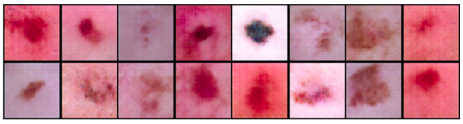


Fig. 3. Model Architecture for Generative Adversarial Network based classifier. The generator takes in noise and produces an image that it learns from data distribution. While discriminator takes in real as well as generated data and performs softmax classification over target classes as well as discriminates fake or real.



(a) Real images sampled from dataset



(b) GAN generated, synthetic images

Fig. 4. Real and GAN generated synthetic images.

for the $(n+1)-th$ unit in the final layer of D which is designated for predicting the probability of a an image being real or fake. The remaining n units in this layer are trained using a standard cross-entropy loss function as below:

$$L_{supervised} = -\sum_{D(\mathbf{x})} p(D(\mathbf{x}))log(q(D(\mathbf{x}))) \qquad (2)$$

Here, $p(D(\mathbf{x}))$ is the correct class label corresponding to the input $\mathbf{x}$ and $p(D(\mathbf{x}))$ is the class probability predicted by the discriminator D. Addition of an extra label for the fake class enables us to simultaneously train the discriminator network over two loss functions; one for detecting fake and real images and the other for classification within real images.

IV. Experiments and Results

Each sample of training data is resized to 224x224 before feeding it into the discriminator network. The generator and discriminator networks both have 5 layers each. The generator consists of 4 deconvolutional layers with `ReLU` activations and a single deconvolutional layer, at the end, with `tanh` activation. The discriminator consists of 4 convolutional layers with `Leaky-ReLU` activations and a single fully-connected softmax layer at the end. During the training process, the discriminator is fed with real data as well as the data from the generator network. In addition to GAN generated samples; conventional data augmentation strategies such as: random cropping, Gaussian blurring, addition of salt & pepper noise; are also employed to ensure that the model can learn robust representations. After a few epochs of training, the generator produces very realistic images that are also used for further training. Training is performed for a total of 24 epochs with a learning rate of 0.0001. Synthetic images generated after training of the GAN are shown in Fig. 4 (b); real images from the training data are also shown for comparison in Fig. 4 (a). Notice that the GAN generated images are very realistic and difficult to differentiate from the actual images. This is encouraging and demonstrates that the generator is able to learn a good representation of the image data distribution.

TABLE I

Precision and Recall for all the seven classes. Results are obtained as a result of proposed GANs based Classification

Class	Precision	Recall	F1-score
ACKIEC	0.73	0.69	0.710
BCC	0.85	0.90	0.870
BKL	0.80	0.81	0.804
DF	0.89	0.74	0.808
MEL	0.81	0.79	0.800
NV	0.94	0.95	0.945
VASC	0.89	0.92	0.904

Lesion classification results on the test set are shown in Table I. The highest F1-score (0.945) is obtained for Melanocytic Nevus (NV) which is not suprising due to the large number of training samples of this class. The worst F1-score (0.710) is obtained for Actinic Keratosis (AKIEC). The overall average F1-score is 0.834.

In order to gauge the performance gains obtained by employing GAN based augmentation we benchmark our algorithm with two popular existing network architectures, ResNet and DenseNet, which are widely employed in a number of medical image classification tasks. ResNet and DenseNet were employed by some of the top scoring entries [10, 9] to the ISIC 2018 challenge [6] that did not employ data augmentation from external sources. Both the benchmark networks were fine-tuned using the same training data as the GAN based model. Furthermore, conventional augmentation techniques (rotation, flipping, salt & pepper noise etc.) were also applied during training. However, no synthetic images obtained from the generator were employed for augmentation purposes when training the benchmark networks. The metric employed for performance comparison is the balance accuracy score. This metric is selected to cater for clas imbalance in the data. Results are shown in Tabel II. It can be observed that GAN based augmentation outperforms both DenseNet and ResNet-50 even though these two architectures are much more deeper than the GAN based classifier. For instance, the DenseNet architecture consists of 103 convolutional layers whereas, the ResNet50 architecture consists of 50 residual layers. We would like to emphasize here that we are talking about fine-tuned deep networks

TABLE II
PERFORMANCE COMPARISON OF VARIOUS ALGORITHMS ON ISIC 2018 DATASET

Approach	Balance Accuracy Score
Semi-Supervised GANs	**0.861**
DenseNet	0.815
ResNet-50	0.792

here. It is indeed highly likely that deeper networks may outperform our GAN based classifier if trained from scratch however; medical image data is almost always in short supply and in such scenarios fine-tuning is the only viable option when using deep neural network architectures.

V. CONCLUSIONS

We have investigated a semi-supervised approach based on synthetic data augmentation produced from generative adversarial networks (GANs). Using this strategy we are able to classify skin lesion images with high accuracy. Furthermore, results also demonstrate that augmenting the training data using GAN generated image samples does result in substantial performance improvement when compared to the conventional approach of fine-tuning existing deep neural network architectures. These kind of novel in data generation and augmentation strategies can be very helpful application areas, such as medical imaging, where large sized training datasets are generally not available. We aim to further explore customized architectures for data generation and classification as part of our future work.

ACKNOWLEDGMENT

The authors are grateful to the International Skin Imaging Collaboration (ISIC) for giving us access to their image data online for research purposes.

REFERENCES

[1] Christoph Baur, Shadi Albarqouni, and Nassir Navab. "Generating highly realistic images of skin lesions with GANs". In: *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018, pp. 260–267.

[2] Alceu Bissoto et al. "Skin lesion synthesis with generative adversarial networks". In: *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018, pp. 294–302.

[3] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), p. 115.

[4] Maayan Frid-Adar et al. "Synthetic data augmentation using GAN for improved liver lesion classification". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 289–293.

[5] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[6] *ISIC Challenge 2018 Leaderboard*. `https://challenge2018.isic-archive.com/leaderboards/`. Accessed: 2019-02-17.

[7] Max Jaderberg et al. "Synthetic data and artificial neural networks for natural scene text recognition". In: *arXiv preprint arXiv:1406.2227* (2014).

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[9] Katherine M Li and Evelyn C Li. "Skin Lesion Analysis Towards Melanoma Detection via End-to-end Deep Learning of Convolutional Neural Networks". In: *arXiv preprint arXiv:1807.08332* (2018).

[10] Yongsheng Pan and Yong Xia. "Residual Network based Aggregation Model for Skin Lesion Classification". In: *arXiv preprint arXiv:1807.09150* (2018).

[11] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific data* 5 (2018), p. 180161.