Machine Learning Applications in Medical Diagnosis & its Adversarial Impacts

Shubham Agarwal
Saarland University
Saarbrücken

Awantee Deshpande
Saarland University
Saarbrücken

Abstract

Diabetic Retinopathy (DR) is an eye disease associated with long-standing diabetes. With the intention of automating and improving the current state of DR detection, several datasets have been created for analysis by different ML models and techniques. In this project, we attempt to train and test a Convolutional Neural Network on the Kaggle DR dataset, and then subject this model to some adversarial attacks to assess the robustness of the model along with the dataset efficacy. In lieu of the sparsity and non-uniformity of the data samples across all the classes, we use a Generative Adversarial Network to generate images using this dataset. In the end, we report the performance of the classification tasks performed, the effectiveness of the attacks as well as the quality of the synthetic dataset generated and also the discuss the short-comings, issues and future prospects.

1. Introduction

1.1. Machine Learning

Machine Learning has made it possible to automate the task of identifying rules, patterns, and trends within complex data in order to gain information from it. This information can then be used for further prediction and decision making. There are four generally defined approaches to machine learning, namely - supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. They find a huge application in almost every field of industry today.

1.2. Adversarial Attacks on ML Models

Adversarial examples are inputs to a machine learning model that are intentionally crafted to force the model to make a mistake. Typically, adversarial examples are engineered by taking real data and making intentional changes to it in order to fool the algorithm that will process it.

Adversarial attacks can be of two types:-

1. **Black box attacks** - The attacker does not have any access or knowledge about the underlying ML model, and relies on a different model (or no model) to generate adversarial data to transfer to the target model.
   e.g. Spatial attack, Inversion attack, Gaussian Blur attack
2. **White box attacks** - The attacker has complete access to the ML Model and its underlying parameters.
   e.g. Fast Gradient Signed Method attack

Thus, it becomes necessary to have a thorough understanding of how different attacks, training techniques, and dataset properties affect the accuracy of the ML models.

1.3. ML in Medical Analysis and Diagnosis

Nowadays, with modern medical imagery techniques, it has become possible to harness large amounts of medical data. Past case histories make valuable training data that provides a better probability of correctly diagnosing the next case based on the symptoms, which are used as features for training. It has been found that deep learning techniques achieved by the implementation of Neural Networks have given good results over various medical repositories.

Machine learning is especially conducive in healthcare because it will help lower the cost of healthcare, make diagnosis a faster process, and improve the patient's overall experience.

2. Background

2.1. Diabetic Retinopathy: An Overview

Diabetic retinopathy (DR) is a complication of diabetes that severely affects the retina of the eye due to increased blood sugar levels. If not diagnosed in time, it can gradually become more serious and can affect the patient's vision and lead to blindness. Detecting DR manually by a trained clinician is a time-consuming process. Machine learning can play a valuable role here by automating the process of analysing the colour fundus photographs of the retina. Detecting and learning the stages of DR makes this a multi-class or binary classification problem.

Based on protocol, the dataset is divided into 5 classes - Class 0 being the normal eye, and the further classes progressing with each stage as mild, moderate, severe, non-proliferative DR, and proliferative DR.

2.2. Convolutional Neural Network

Deep learning primarily relies on unsupervised methods to train neural networks and then fine tune them through supervised learning methods. CNNs have been dominant in such deep learning tasks for some time now. They are designed to automatically and adaptively learn spatial hierarchies of features through back-propagation by using multiple building blocks, such as convolution layers, pooling layers, and fully connected layers. They work especially well for any kind of data that has a grid based pattern - which means that CNNs find a large application in imaging techniques. Because of the rich feature set of medical images, CNNs are a preferred learning method over medical databases for diagnosis and classification.

2.3. Security & Sensitivity of Medical Data

Diagnosing medical data is an error-sensitive task - one cannot afford to misdiagnose an ill patient as healthy, or vice versa (with different levels of severity). This makes it imperative for the underlying machine learning algorithm to be trained as precisely as possible without a large margin of error. Medical images especially are highly sensitive to ML attacks[7] because

1. They have very few classes
2. The complex image structure makes them more vulnerable
3. Overparameterising medical imaging tasks can make them more susceptible to ML attacks

2.4. Generative Adversarial Network

Generative Adversarial Networks use a dual NN architecture that follows the principles of game theory. Two neural networks are pitted against each other with the goal of generating synthetic, artificial data that can be passed off as real data. The generator neural network creates new data instances from noise. The discriminator decides if the generated instance belongs to the actual training dataset or not. The entire architecture is composed of a dual feedback loop, where the discriminator is in a feedback loop with the true labelled data, and the generator is in a feedback loop with the discriminator.

2.4.1 GANs in Medical Imaging

Image generation is a valuable contribution in the field of medicine due to the lack of available data for medical analysis. Each aspect of a GAN can be used in medical imaging. The generator can be used where medical data is sparse and patient privacy is a concern. The discriminator can be used as an identifier for abnormal images.

A lot of research is being done in formulating applications of GANs for the processing of medical data. GANs have been implemented for the synthesis of skin lesions, retinal scans, MR images etc.[**?**] [**?**] [**?**]

3. Dataset

The largest open-source dataset available for Diabetic Retinopathy is hosted by Kaggle and has actively been used in numerous other projects by researchers. The dataset contains approximately 35 thousand training images and as many test images altogether. A sample image from the dataset is shown in Figure 1:

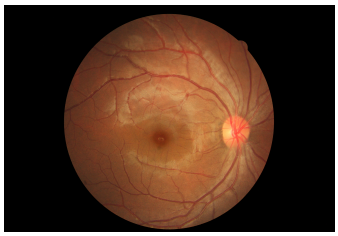

Figure 1: Example of an image with type 0 (no symptoms)

However, the dataset is largely imbalanced across different classes, as approximately 70% of the dataset attributes to class 0 (i.e., no disease). On the contrary, there are less than 800 images for class 4 (i.e. the last stage of the disease). Due to this huge imbalance, it is difficult to train the model equivalently on all the classes. For this reason, we randomly picked 700 images from each class in the training dataset for our entire project, i.e., a dataset of size 3500 images for both, training and testing.

4. Experiments

We perform two relatively different experiments on the same dataset to achieve the common goal of accurately predicting the class of the data points among multiple classes:

1. Deep Neural Network (CNN) for accurate multi-class classification.
2. Generative Adversarial Network to create synthetic dataset for further training.

There are a few common steps that we performed in both the experiments, which are described below. We describe each of them separately further.

Data Preprocessing & Augmentation

We implemented simple image pre-processing steps such as reshaping and scaling down the images to 200 x 200 and

further flattening them. We also performed additional data transformations for the GAN during data loading.

We used different image augmentation techniques while training such as shifting pixels. horizontal and vertical flip, image rotation and transformation & random brightness. The augmentation techniques helped to generate artificial images on the fly train the model better for generalization and robustness.

Model Parameters

We performed multiple iterations of training over different parameters to achieve optimal accuracy over the test dataset. The batch size ranged from 8 to 128 over different no. of epochs ranging from 10 to 300. Moreover, we used different no. of pooling layers and convolution filters for the same.

4.1. Experiment I - CNN

In initial steps, we implemented a standard CNN model for classifying the data points to their respective classes. The architecture of the model has three convolution layers with ReLU and Soft-max as Activation functions, followed by Max-Pooling. We also added dense and dropout between the two layers. We used the same model for all classification tasks in the project.

We describe the classification task in steps as follows:

1. Firstly, we performed training over all the data points and their classes and then tested the performance over different parameters to accurately identify the optimal parameters for further classification.

2. Then, we decided to perform a binary classification between class 0 and all other classes (i.e., no disease vs different stages of the disease) to understand the performance of the classifier when asked to predict between the normal vs diseased symptoms. We also performed one-to-one classification between each classes and class 0 (i.e. 0 vs 1, 0 vs 2, 0 vs 3, 0 vs 4) to narrow down the classes the classifier fails to perform for accurate prediction.

3. Lastly, we performed adversarial attacks on both the above classification models by leveraging the standardized open-source library - *Foolbox*. We perform a total of four different image-based adversarial attacks such as FGSM, Spatial attack, Gaussian Blur and so on. They try to force the model to misclassify the data point by creating external perturbations on the given image.

   By doing so, we intended to understand the performance of the classifier to handle noisy and adversarial inputs.

4.2. Experiment II - GAN

The initially obtained dataset is largely imbalanced. As this is mostly the case with any other medical and health based tasks, we attempted to implement a GAN which would generate images as close as possible to the real dataset, so as to create a synthetic dataset which could be used by the classifier to further optimize and enhance its accuracy in prediction. Moreover, training the classifier with the synthetic dataset generated by the GAN along with the original dataset would also help to defend against adversarial attacks and at the same time handle the noisy inputs well.

We describe this experiment in steps as follows:

1. We implement DCGAN and train the network (over different range of batches and epochs) to generate synthetic dataset for each of the class separately.

2. We then perform multi-class classification on the original dataset by training additionally on the synthetic dataset.

3. Similarly, we perform binary classification over the original dataset in conjunction with the synthetic dataset.

5. Results

We evaluated the performance of each of the classification task individually from both the experiments respectively.

5.1. Result I

The first experiment mostly focused on the classification of the dataset in their respective classes under different configurations (i.e. multi-class and different binary settings). We report the evaluation statistics obtained from each of the tasks in Table 1.

We observe that the accuracy obtained from the multi-class task is 53.4% - well above the random baseline (20%) but doesn't prove to be significantly useful for real world applications. With the intuition of predicting only the existence of the disease (not the stage), we obtained an accuracy of about 87% with binary classification, which infers that the model is at least able to distinguish between DR and normal features. When performing classification between two classes, we observe that it is equivalently difficult for the classifier to distinguish between non-diseased and initial stages of the disease. We suspect that the features in this case have huge overlap and thus makes it difficult to correctly classify them.

Then we implemented a few adversarial attacks on the first two tasks listed in Table 1. We report the evaluation statistics obtained in Table 2. The attacks performed are as follows:

Task	Accuracy	Precision	Recall	F1 Score
Multi-Class	53.4%	55%	53%	53%
Binary (0vs*)	88.5%	87%	88%	87%
Binary (0vs1)	52.4%	97%	53%	67%
Binary (0vs2)	55.3%	62%	55%	56%
Binary (0vs3)	50.3%	50%	50%	50%
Binary (0vs4)	70.0%	73%	70%	71%

Table 1: Evaluation Report for Experiment 1 - CNN

1. *FGSM Attack* - We started with the basic white-box gradient-based attack to estimate the effects of gradient-based perturbation on the classifier. We observe that this attack is fairly able to force the classifier to mis-predict in case of multi-class prediction. However, although the mean accuracy is reduced in case of binary classification, it is not drastically decreased as seen in the former task.
2. *Spatial Attack* - This attack tries to create random rotations and transformations in the input image to force it to be mis-classified without knowing much about the ML model or its architecture. We observe that as above, the spatial attack seems pretty ineffective for binary classification, while it's the other way round for the multi-class task. We can verify that the spatiality of the input is important, as the images are either left or right aligned as per the eye.
3. *Inversion Attack* - In this attack, the pixels of the images are randomly inverted to distort the view of the model for prediction. This is also a blackbox attack with similar results. Inverting the features doesn't really affect the binary classification unless the feature-related pixels are directly affected.
4. *Gaussian Blur Attack* - This form of attack blurs the image so that the features are not clearly available to the classifier for identification. This attack is also comparatively effective on the multi-class task.

From the attacks above, we infer that it is relatively easy to fool the model to mis-classify when multiple classes are involved and the features overlap. This is not the case when just predicting the existence of the disease, that is, when there are few features against many features.

Attacks	Multi-Class	Binary
FGSM	0.43%	60.14%
Spatial	5.57%	72.86%
Inversion	9.71%	74.86%
Gaussian Blur	6.0%	70.43%

Table 2: Adversarial Attacks on Experiment 1

5.2. Result II

We evaluated the performance of the DCGAN to generate images, which is expected to be as feature-rich as the original dataset for training purposes, implemented by running the original CNN classification task in Experiment 1 again by training over the synthetic dataset in addition to the original dataset and then testing its accuracy. The sample images generated by the GAN for class 2 is shown in Figure 2.

We obtain accuracy in the range of 23% - 30% over different iterations and batches when run on multi-classification task. which is worse than the results obtained before with the original dataset. Thus, we infer that the synthetic images generated. though visually similar, does not contain the features learnt by the model to correctly classify them and thus is almost comparable to the random selection performance (i.e., 20

However, when training on the same synthetic dataset for binary classification gives an accuracy of over 70% which is fairly comparable to the original result of 87%. Thus, it makes sense to say that GAN could be approximately as useful as the original dataset in just predicting the existence of the dataset but doesn't help in predicting the exact class of the input.

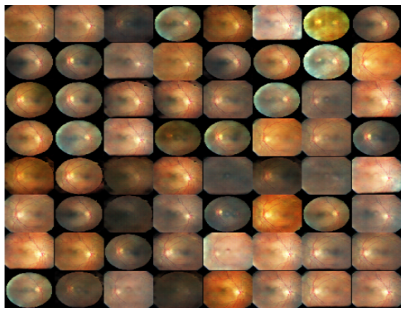

Figure 2: Class 2 Images Generated by DCGAN

6. Discussion

The application of ML has been on the rise in the medical domain in the past few years. However, the fact that these applications are from being tested in adversarial conditions has been discovered recently which makes it even more difficult to estimate the practicality of a certain model to perform as-is. Also, keeping into mind the sensitivity of the domain and its consequences, it's fair to say that true-negatives would be far worse than false-positives in these kinds of applications.

We observe from both the experiments that the features overlap and are sparse across the classes making it difficult to for the classifier to learn. Further, we understand that the model architecture used in both CNN & DCGAN is quite

simple and improving the CNN may help improve the prediction accuracy. Similarly, deploying CGAN or WGAN might capture better details from the input instead of DCGAN. Moreover, using weighted loss would help to try and capture details from the input with fewer instances available. From the results obtained by attacking the CNN, we claim that it's relatively easy to attack models with multiple classes when the dataset is imbalanced or small in size.

Conclusion

In this project, we started with the classification of different stages of Diabetic Retinopathy disease with initial input dataset obtained from Kaggle. After performing multiple iterations of classification to identify the set of classes where features largely overlap, we reported the accuracy of each of the task, In general, binary classification is comparatively better than multi-class and could be further tubed to perfection. However, the second experiment focused on creating synthetic dataset by leveraging the GAN architecture did not yield significant results as training over the adversarial data did not help better the performance of the multi-class classification further.

We believe that by using a weighted loss function would make sense and parallelly, architecture CGAN and WGAN might capture the details in the input missed by DCGAN and is worth exploring.

References

[1] Alceu Bissoto, Fábio Perez, Eduardo Valle, and Sandra Avila. 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*, pages 294–302. Springer, 2018.

[2] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

[3] Pedro Costa, Adrian Galdran, Maria Inês Meyer, Michael David Abràmoff, Meindert Niemeijer, Ana Maria Mendonça, and Aurélio Campilho. Towards adversarial retinal image synthesis. *arXiv preprint arXiv:1701.08974*, 2017.

[4] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. *arXiv preprint arXiv:1712.02779*, 2017.

[5] Samuel G Finlayson, Hyung Won Chung, Isaac S Kohane, and Andrew L Beam. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*, 2018.

[6] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[7] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *arXiv preprint arXiv:1907.10456*, 2019.

[8] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[9] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[10] Per Welander, Simon Karlsson, and Anders Eklund. Generative adversarial networks for image-to-image translation on multi-contrast mr images-a comparison of cyclegan and unit. *arXiv preprint arXiv:1806.07777*, 2018.

[11] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.

[12] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 39–49, 2017.

[13] Yi Zhou, Boyang Wang, Xiaodong He, Shanshan Cui, Fan Zhu, Li Liu, and Ling Shao. Dr-gan: Conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. *arXiv preprint arXiv:1912.04670*, 2019.