

Applications, Challenges & Adversarial Influence of ML in Medicine

Machine Learning in Cybersecurity, 2019-20

{Saarland University, CISPA}

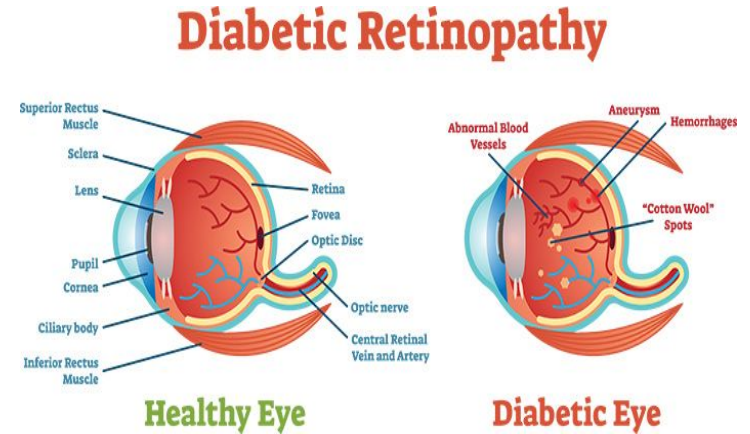
Shubham Agarwal, Awantee Deshpande



Motivation

Diabetic Retinopathy : A Background

- Diabetic retinopathy (DR) is a complication of diabetes that severely affects eyes.
- Its diagnosis is divided into 5 different stages - type 0 denotes no symptoms, type 1 to 4 denotes gradual increase in the severity of the disease.



Machine Learning: Scope and Application

Article | [Open Access](#) | Published: 20 September 2019

Deep learning algorithm predicts diabetic retinopathy progression in individual patients

Filippo Arcadu, Federico
& Marco Prunotto

npj Digital Medicine

10k Accesses | 1

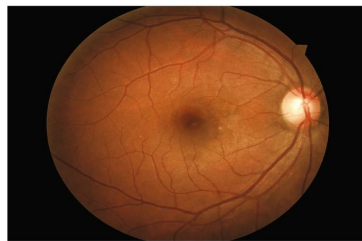
DeepMind's AI can detect over 50 eye diseases as accurately as a doctor

The system analyzes 3D scans of the retina and could help speed up diagnoses in hospitals

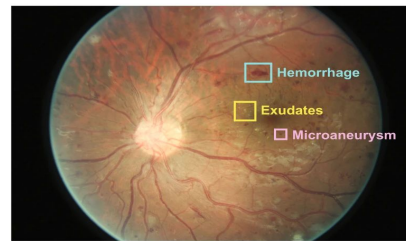
By [James Vincent](#) | Aug 13, 2018, 11:01am EDT

a

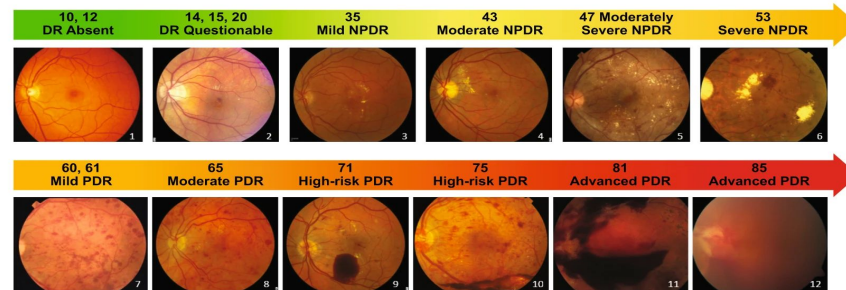
Patient without DR



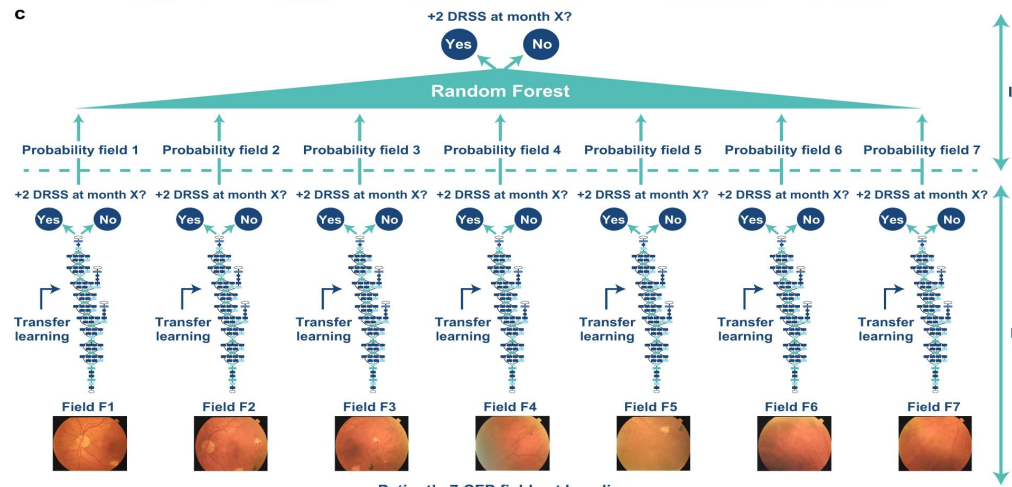
Patient with DR



b

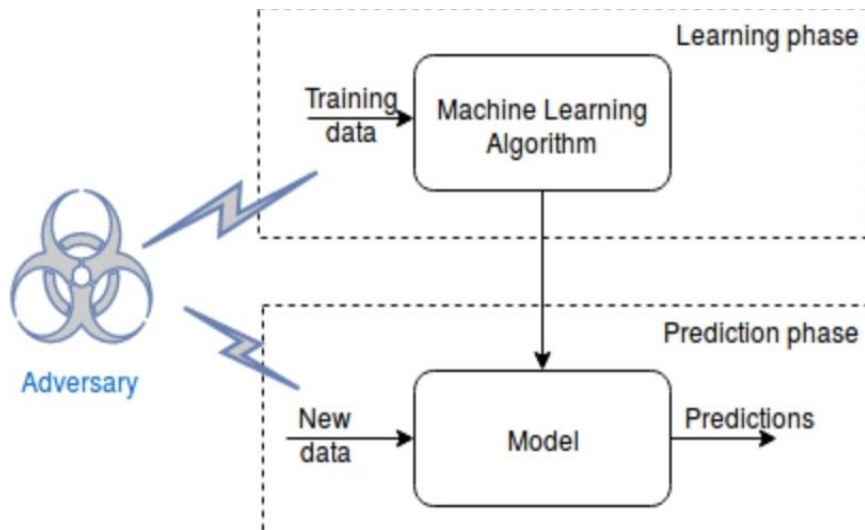


c



However ...

Adversarial Machine Learning comes into picture - a technique employed in the field of machine learning which attempts to fool models through malicious attempts.





Poisoning Attacks

- An ANN attacker poisons the training data by injecting carefully designed samples to eventually compromise the whole learning process.

Evasion Attacks

- An ANN attacker modifies samples at test time to evade detection; that is, to be misclassified as legitimate.

KDnuggets™ | [Subscribe to KDnuggets News](#) |  
[SOFTWARE](#) | [News/Blog](#) | [Top stories](#) | [Opinions](#) | [Tutorials](#) | [JOE](#)

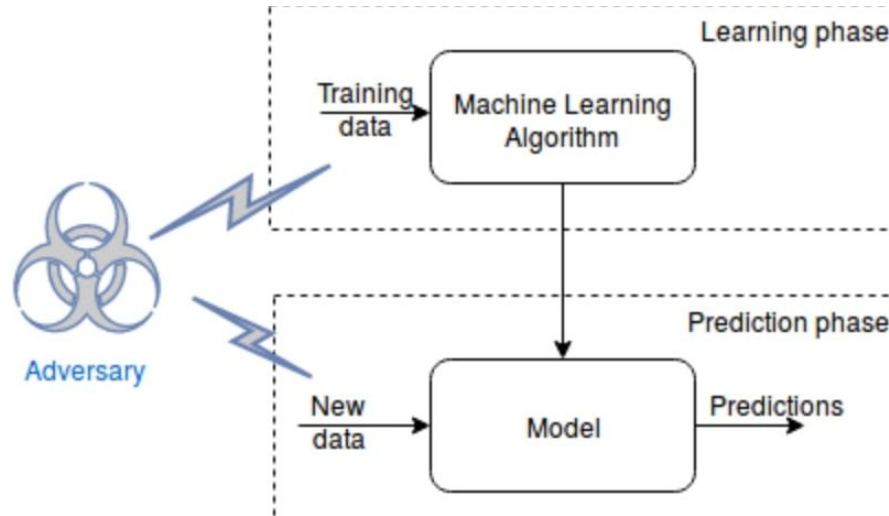
Sources: <https://mc.ai/adversarial-robustness>
<https://www.kdnuggets.com/2019/06/machine-learning-adversarial-attacks.html>
https://www.researchgate.net/figure/Adversarial-Machine-Learning_fig1_318227376

[KDnuggets Home](#) » [News](#) » [2019](#) » [Jun](#) » [Opinions](#) » Why Machine

Why Machine Learning is vulnerable to adversarial attacks and how to fix it

However...

ML models may be subject to various adversarial attacks which could significantly disrupt its intended behavior.

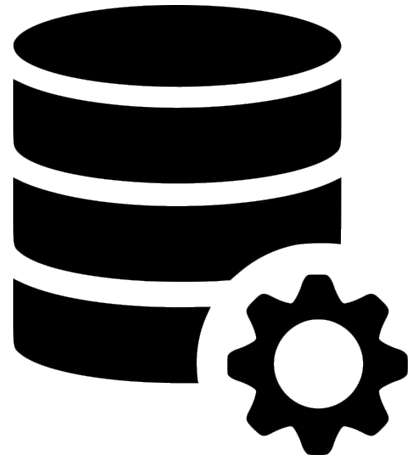


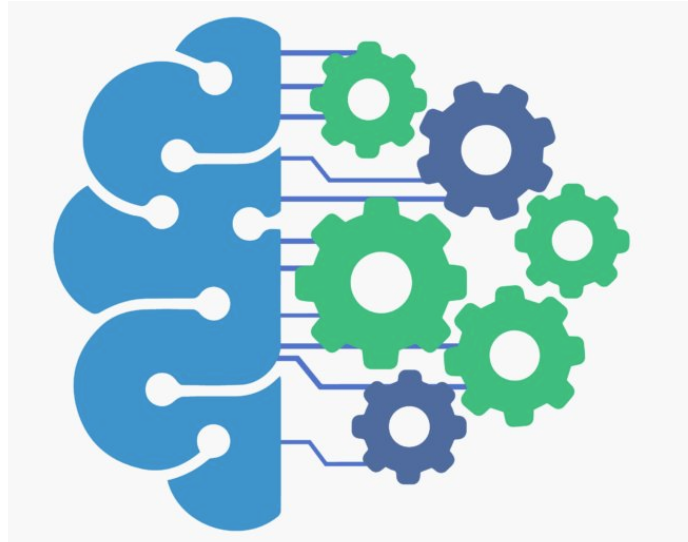
Goal...

- Analysis of the performance of DNN on classification of the diabetic retinopathy disease.
- Performance of attacks on the NN model and analysis of their respective effectiveness.
- GAN implementation to generate synthetic data and overcome imbalanced dataset issue for future training tasks.
- Evaluation of the quality of the synthetic dataset over original dataset.

Training Dataset

- **Kaggle Diabetic Retinopathy Detection Database** - High Resolution Images of fundus (interior portion) of an eye for each class/type.
- **Training Data Points Available** - 35,216.
- **Training Data Used** - 3500 (700 images from each class).



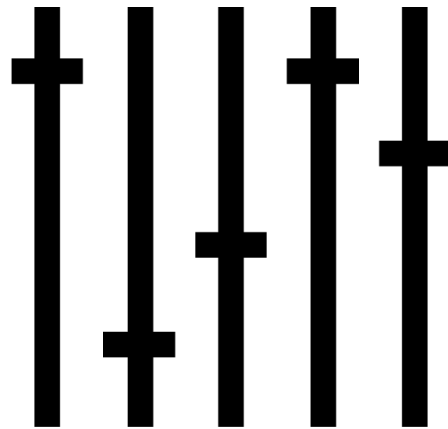


Experiments

Experiment I - CNN

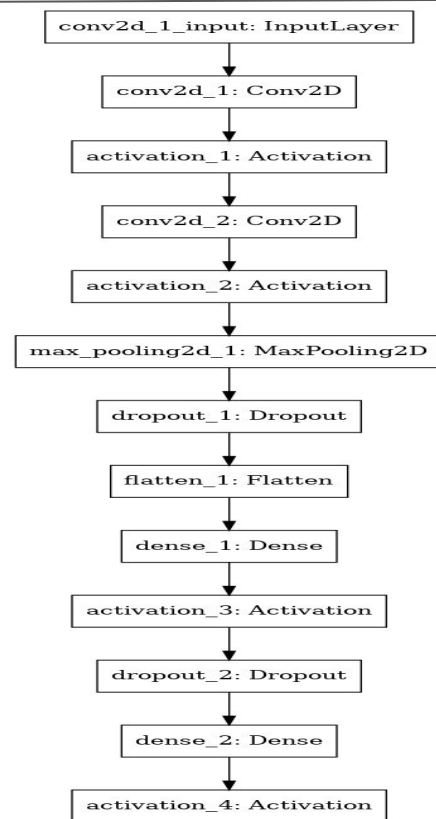
Standard CNN Model to perform multiple classification tasks:

1. Multi-class classification.
2. Binary Classification (type 0 vs Rest).
3. Binary Classification (0vs1, 0vs2, 0vs3, 0vs4).
4. Evaluation of adversarial attacks on tasks 1 & 2.



Experiment I - CNN (continued)

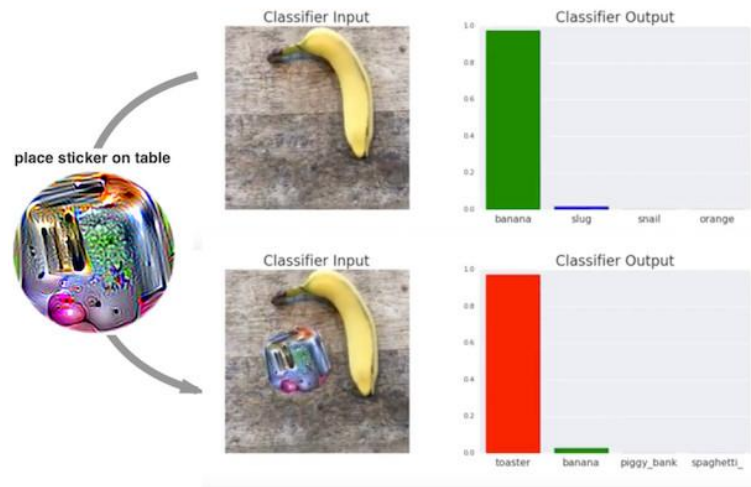
- Architecture Summary:
 1. 3 Convolution Layers.
 2. Activation Functions (ReLU & Softmax).
 3. Dropout & Dense Layers.
- Hyperparameters:
 1. Batch Size : 8 - 128.
 2. Epochs : 10 - 300.
 3. Kernel Size : 3.



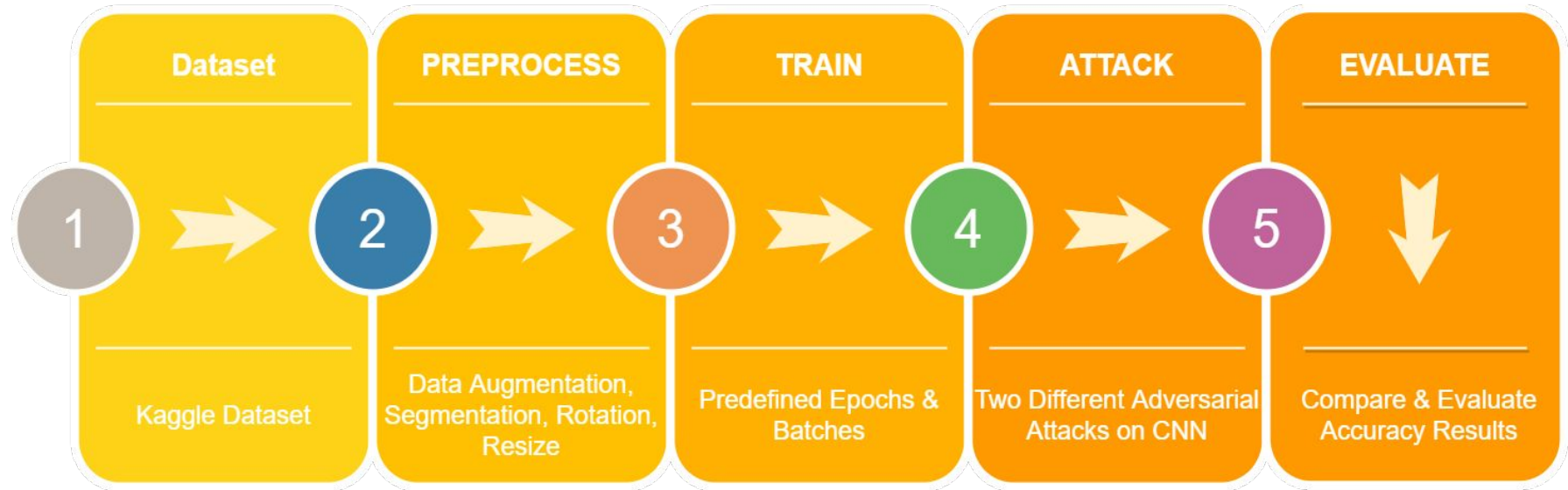
Experiment I - CNN (continued)

Adversarial Attacks Performed on first two classification tasks:

1. FGSM Attack
2. Spatial Attack
3. Inversion Attack
4. Gaussian Blur Attack



Experiment I - System Flow

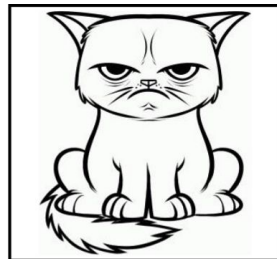


Experiment II - GAN

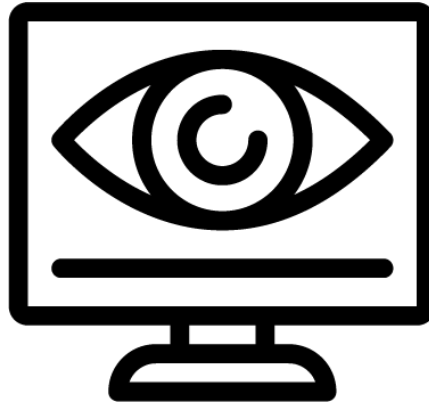
- DC-GAN to generate synthetic dataset for all the classes over original database.
- The images for each class are separately generated to retain the labels for further use.
- Different classes were trained on different hyperparameters based on original dataset.



\mathcal{X}



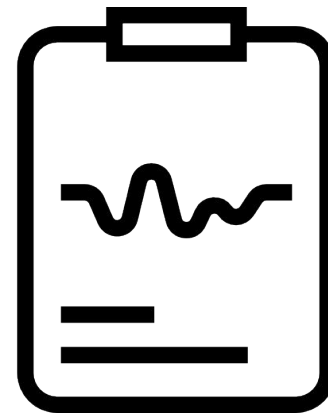
$G(z)$



Evaluation

Results from Experiment I

- Multiple classification tasks to obtain optimal accuracy.
- Classified between each pair of classes to better understand the ambiguities faced by the classifier.



Results from Experiment I

Task	Accuracy	Precision	Recall	F1 Score
Multi-Class	53.4%	55%	53%	53%
Type 0 vs Rest	88.5%	87%	88%	87%
Type 0 vs Type 1	52.4%	97%	53%	67%
Type 0 vs Type 2	55.3%	62%	55%	56%
Type 0 vs Type 3	50.3%	50%	50%	50%
Type 0 vs Type 4	70%	73%	70%	71%

Results from Experiment I

Task	Accuracy	Precision	Recall	F1 Score
Multi-Class	53.28%	55%	53%	53%
Type 0 vs Rest	88.5%	87%	88%	87%
Type 0 vs Type 1	52.4%	97%	53%	67%
Type 0 vs Type 2	55.3%	62%	55%	56%
Type 0 vs Type 3	50.3%	50%	50%	50%
Type 0 vs Type 4	70%	73%	70%	71%

Results from Experiment I

Task	Accuracy	Precision	Recall	F1 Score
Multi-Class	53.28%	55%	53%	53%
Type 0 vs Rest	88.5%	87%	88%	87%
Type 0 vs Type 1	52.4%	97%	53%	67%
Type 0 vs Type 2	55.3%	62%	55%	56%
Type 0 vs Type 3	50.3%	50%	50%	50%
Type 0 vs Type 4	70%	73%	70%	71%

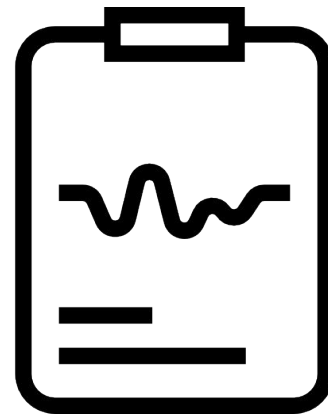
Results from Experiment I

Task	Accuracy	Precision	Recall	F1 Score
Multi-Class	53.28%	55%	53%	53%
Type 0 vs Rest	88.5%	87%	88%	87%
Type 0 vs Type 1	52.4%	97%	53%	67%
Type 0 vs Type 2	55.3%	62%	55%	56%
Type 0 vs Type 3	50.3%	50%	50%	50%
Type 0 vs Type 4	70%	73%	70%	71%

Results from Experiment I

- Adversarial perturbations on input to force the model to misclassify data point.

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%

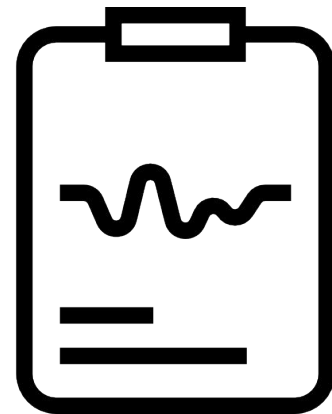


Results from Experiment II

- In addition to the original dataset, trained on the synthetic dataset generated from GAN to evaluate its performance.

Task	Accuracy in Experiment 1	Accuracy with GAN
Multi-Class	53.2%	23.4%
Binary	87%	74.1%

- We infer from the results that GAN doesn't perform well to imitate the original features.



Results from Experiment II - Sample I

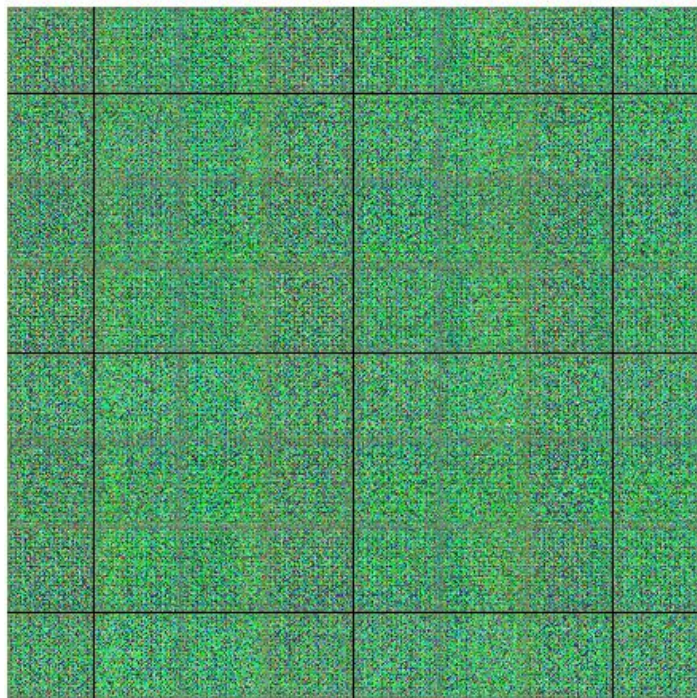


Sample Training Image

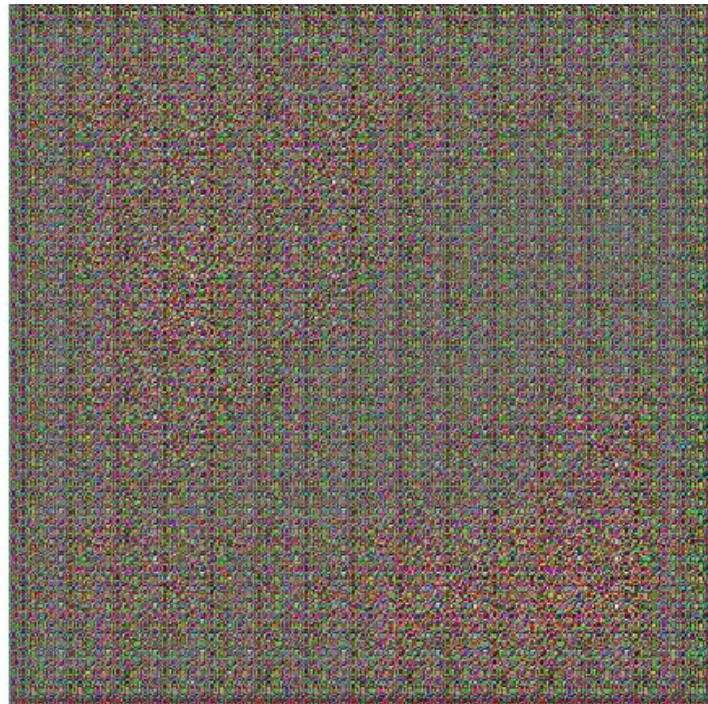


GAN Generated Image

Results from Experiment II - Sample II



Overall Training Across Epochs



Training Across Epoch for Single Image

Discussion

- Standard CNN model used for classification, can improve on the model architecture further.
- Imbalanced original dataset an issue.
- Features overlap across certain classes, weighted loss might help in this case.
- DCGAN is ineffective in capturing the necessary features, doesn't really help improve the initial multi-class classification.



Summary

- Binary Classification performs significantly well when compared to pinpoint classification to identify exact class of each input.
- The adversarial attacks and their results from experiment 1 strongly suggests to consider the adversarial conditions for the ML models used in sensitive applications such as in medicine.
- Much like other medical diagnosis problems, imbalanced dataset remains an issue & other variant of GANs such as CGAN & WGAN is worth exploring for generating synthetic dataset.

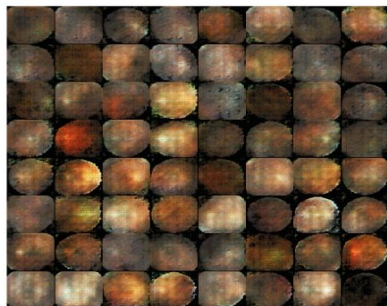
Related Work

- 1) DR-GAN: Conditional Generative Adversarial Network for Fine-Grained Lesion Synthesis on Diabetic Retinopathy Images [Yi Zhou, Boyang Wang, Xiaodong He, Shanshan Cui, Fan Zhu, Li Liu, Ling Shao]
- 2) Pathological Evidence Exploration in Deep Retinal Image Diagnosis [Yuhao Niu, Lin Gu, Feng Lu, Feifan Lv, Zongji Wang, Imari Sato, Zijian Zhang, Yangyan Xiao, Xunzhang Dai, Tingting Cheng]
- 3) Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems [Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, Feng Lu]
- 4) Synthesizing retinal and neuronal images with generative adversarial nets [He Zhaoab, Huiqi Li, Sebastian Maurer-Stroh, Li Cheng]
- 5) Practical Black-Box Attacks against Machine Learning [Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami]

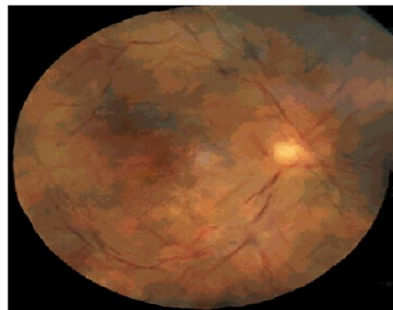
Questions?

Goal...

- Analysis of the performance of DNN on classification of the diabetic retinopathy disease.
- Performance of attacks on the NN model and analysis of their respective effectiveness.
- GAN implementation to generate synthetic data and overcome imbalanced dataset issue for future training tasks.
- Evaluation of the quality of the synthetic dataset over original dataset.



Overall Training Across Epochs



Training Across Epoch for Single Image

Summary

- Binary Classification performs significantly well when compared to pinpoint classification to identify exact class of each input.
- The adversarial attacks and their results from experiment 1 strongly suggests to consider the adversarial conditions for the ML models used in sensitive applications such as in medicine.
- Much like other medical diagnosis problems, imbalanced dataset remains an issue & other variant of GANs such as CGAN & WGAN is worth exploring for generating synthetic dataset.