Utilizing Generative Modeling for Data Augmentation to Develop a Low-Cost Revolutionary Machine Learning Solution to Diagnose Skin Lesions

OVERVIEW

- Skin lesions are defined as areas of the skin anywhere on the body that are abnormal from the skin around it
- These lesions can occur in benign and often harmless forms such as birthmarks, moles, and acne
- Skin lesions may also be malignant in forms such as basal cell carcinoma, squamous cell carcinoma, and melanoma
- Skin cancers must be treated as soon possible by surgical removal or other medication in order to prevent the spread of the disease across the skin and other body parts
- Various types of skin cancers plague numerous individuals on a global scale
  - According to the American Academy of Dermatology, cancerous melanoma affects around 1 million Americans
  - 3 million Americans are affected by non-melanoma skin cancers such as basal cell carcinoma
- 1 in 5 people will develop skin cancer in their lifetime, with around 9500 people diagnosed daily in the United States
- In remote areas of the world and third-world countries, Skin cancer leads to around 7000 annual deaths
- Benign skin diseases are often harmless, but still may require treatment in order to recover from the symptoms.
- A machine learning algorithm that can diagnose skin lesions with the correct skin disease can greatly improve chances of recovery for victims.

PROJECT OBJECTIVE

The object of this project is to develop a Convolutional Neural Network(CNN) model that can accurately diagnose skin lesions to one of seven classes using the MNIST Skin Cancer dataset and data augmentation methods from a Deep Convolutional Generative Adversarial Network(DCGAN).

TERMINOLOGY

Skin Lesions - areas of the skin that abnormal from the skin around it and occur in benign and malignant forms

Melanocytic Nevus – a noncancerous disorder of pigment-producing skin cells(moles); usually does not require treatment, but some cases require removal of the mole

Seborrheic Keratosis – a benign skin condition common in older adults that appears as a waxy brown, black, or tan growth

Melanoma - a serious disease in which pigment-producing cells that give color to the skin become cancerous; treatment may require surgery, radiation, medications, or chemotherapy

Basal Cell Carcinoma – a common skin cancer that begins in the basal cells and appears as a white waxy lump or a brown scaly patch on sun-exposed areas such as the face and neck; treatment include prescription creams or surgery to remove the cancer

Actinic Keratosis - A non-cancerous disease in which rough, scaly patches form skin caused by years of sun exposure; treatment involves removal by surgery

Vascular Lesions – represent benign common abnormalities of the skin and underlying tissues, such as birthmarks

Dermatofibroma - noncancerous skin growth that may look like a small, firm bump, like a mole; does not require treatment but can be removed

Convolutional Neural Network(CNN) - a type of artificial neural network used primarily for image recognition and processing, due to its ability to recognize patterns in images.

Generative Adversarial Network(GAN) - a machine learning neural network model in which two neural networks known as the discriminator and generator train each other in an unsupervised manner to become more accurate in their predictions.

Deep Convolutional Generative Adversarial Network(DCGAN) – a special type of generative adversarial network for images that incorporates convolutional and convolutional-transpose layers in the generator and discriminator, respectively.

Data Augmentation - technique of artificially increasing the training set by creating modified copies of a dataset using existing data.

Dense Neural Network – a traditional neural network model in which neurons from each layer are connected to one another. Each neuron contains a **bias** which is added to the input and each connection contains a **weight** which is multiplied by the input followed by a non-linear activation function during forward propagation.

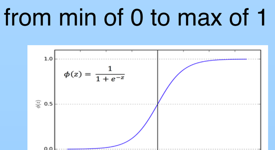
Loss Function - a function that computes the error between the current output of the algorithm and the expected output

Gradient Descent - an optimization algorithm that minimizes the cost function in iterative steps

Epoch - one pass of the entire training set across the network

F1-Score - a metric that measures accuracy by combining the recall and precision of a binary classifier; it goes from min of 0 to max of 1

Sigmoid Activation Function - function takes any real value as input and outputs values in the range 0 to 1; ideal for classification



Softmax Activation Function – used in situations in machine learning that deal with multi-class predictions

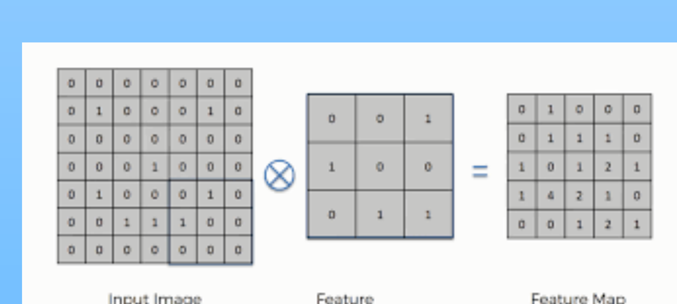
PROCESSES: Feature Detection

•Convolutional Learning employs the convolution operation in which multiple filter matrices called feature detectors which are multiplied in an element-wise manner across each corresponding pixel value in image to detect crucial features

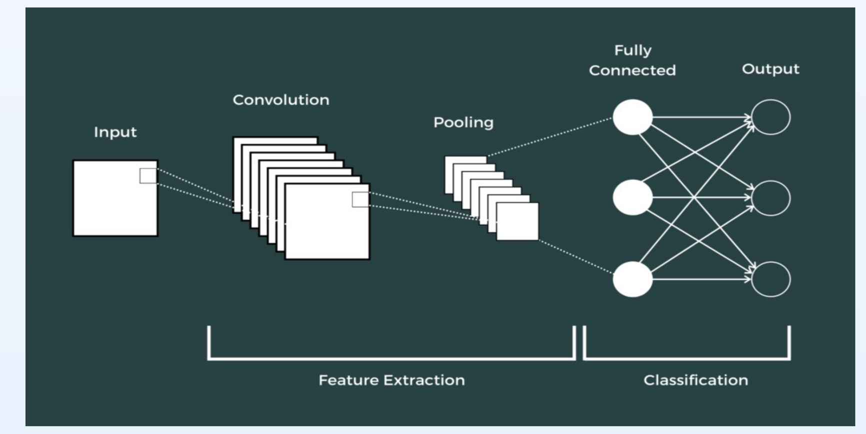
•This operation generates a feature map which is slightly reduced in dimensionality when compared to the original image

•The resulting feature map is put into a non-linear activation function in order to detect the presence of a feature learned by the model

•Allows for meaningful information to be extracted from the special structure of image



PROCESSES: Convolutional Learning



Visualization of the Forward Propagation of a Convolutional Neural Network

The image is preprocessed and fed into the model

The convolution operation with non-linear activation followed by max pooling is performed on the image a given amount of times

The resulting data is flattened if necessary and inputted into the dense neural network for classification

PROCESSES: Generative Modeling for Images

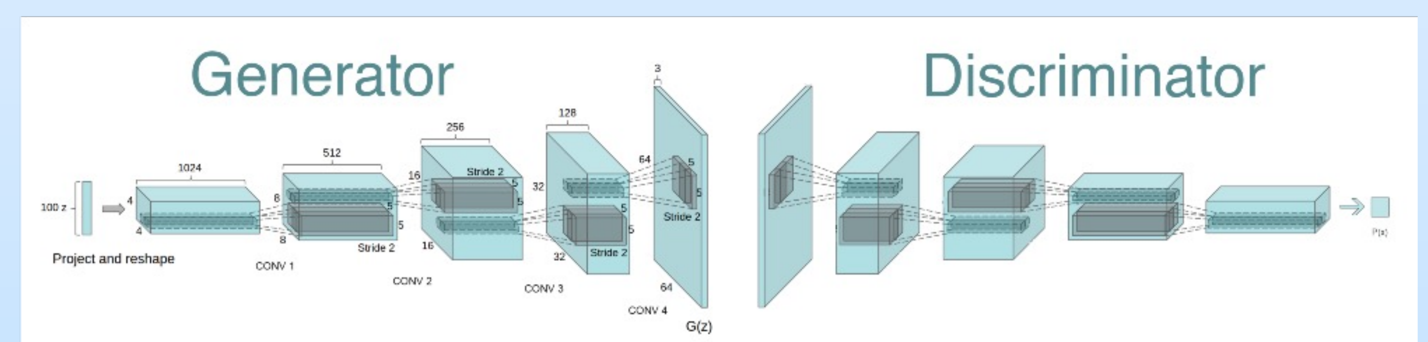


Diagram of Deep Convolutional Generative Adversarial Network

•The Generative Adversarial Network consists of two parts:

•The Generative model attempts to generate realistic data from a noise vector that is indistinguishable to the discriminator

•The Discriminator model attempts to discern between generated and real data

•The DCGAN Model learns in unsupervised manner

•The generator of the DCGAN model contains transpose convolutional layers with non-linear activations which continuously generate images from a noise vector

•The discriminator likewise contains convolutional layers with non-linear activations which feed to an output neuron with a final non-linear activation

•Both parts alternate turns in training of DCGAN in order to decrease the loss in the corresponding other model.

RESOURCES

1. The Google Collaboratory Programming Environment powered by Google Cloud
2. The MNIST Skin Cancer Dataset
3. The Python 3 Programming Language
4. The Tensorflow Python Library
5. The Pandas Python Library
6. The Numpy Python Library
7. The Sklearn Python Library

MNIST SKIN CANCER DATASET

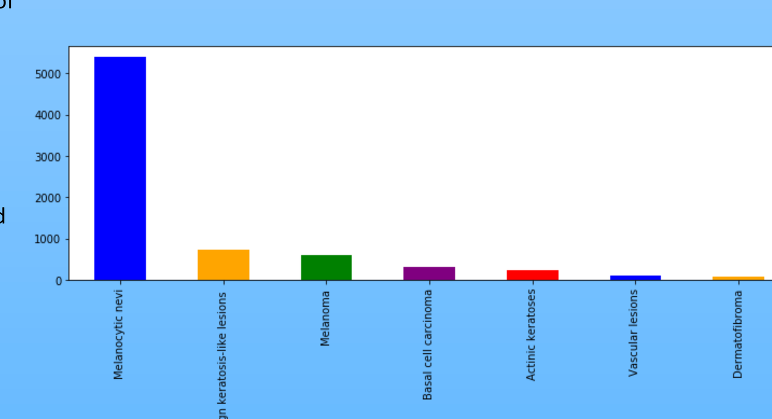
•The MNIST Skin Cancer Dataset contains multiple dermatoscopic images the of pigmented skin lesions in the size of 600 x 450 pixels with associated annotations

•These annotations/labels for each image were independently confirmed by professional dermatologists

•Contained 7470 images of pigmented lesions along with proper metadata and labels

•Showed an uneven distribution of data between the defined classes of skin lesions

•This dataset along with the data augmentation methods will be employed for training the classification algorithm



METHODOLOGY

Data Augmentation

1. Download and pre-process the images and labels using the Pandas and Numpy python libraries and resize images by a factor of 0.25
2. Construct the discriminator of the DCGAN model using Tensorflow with three convolutional layers with 64 filters each along with the leaky RELU activation function which can be expressed as:

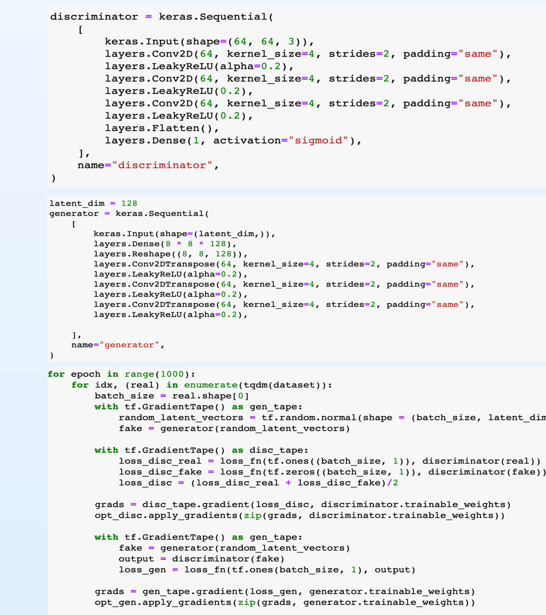
   $$R(z) = max(0.2z, z)$$

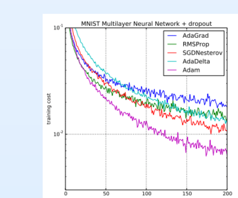
3. Finish the model with an output neuron with the sigmoid activation function
4. Construct the generator using Tensorflow with three transpose convolutional layers with 64 filters each along with the leaky RELU activation function
5. Create a latent noise vector of size 128 for the input of the generator
6. Compile the both the generator and discriminator with the binary cross-entropy loss function and the ADAM gradient descent algorithm

   $$L = -\frac{1}{N}\sum_{i=1}^{N} y_i log(\hat{y}_i) + (1 - y_i)log(1 - y_i)$$

7. Train the DCGAN model for 1000 epochs on images of Melanocytic Nevus and use the generator model to generate 500 images
8. Repeat the process to generate 500 images of the other 6 categories of skin lesions.

Sample Images of Pigmented Lesions from the MNIST Skin Cancer Database





Graph showing performance of ADAM stochastic gradient descent algorithm in comparison to other algorithms

Development of Classification Algorithm

1. Construct the CNN classifier using Tensorflow with 3 convolutional layers with 64 filters and corresponding max pooling and RELU activation functions
2. Append the dense layer of size 128 and an output neuron to the model with the Softmax activation function for proper classification:

   $$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

3. Combine the generated and original data and use the Sci-Kit Learn library to allocate around 20% of the original to testing and 10% to validation
4. Compile the CNN model with the Categorical Cross entropy function and ADAM gradient descent algorithm and train it for 50 epochs
5. Evaluate the model's performance on the test set

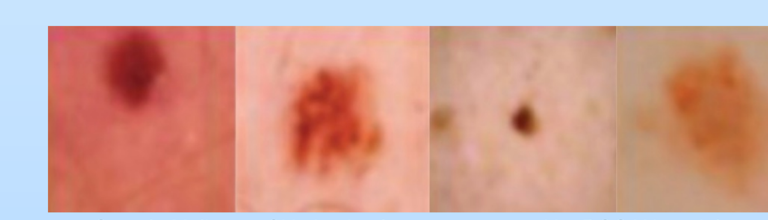
RESULTS AND DATA ANALYSIS

•The Convolutional Neural Network achieved an accuracy of around **91%** on the test set.

•The classifier achieved a recall of **0.89** and precision of **0.83**

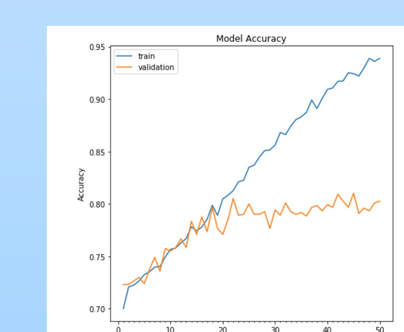
•The F1-score of this model is **0.86**

•When the data augmentation methods using the DCGAN were not employed, the model achieved an accuracy of around **73%** and a F1-score of around **0.64**

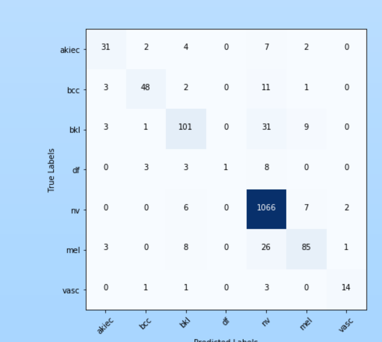


Sample Malignant Skin Lesion Images generated by DCGAN model

$$F_1 = \frac{2(Precision \cdot Recall)}{Precision + Recall}$$



Plot of Accuracy of Model on Training and Validation Set versus Training Iterations during Gradient Descent



Confusion Matrix of the Model on the Test set

DISCUSSIONS/CONCLUSIONS

•The goal of this project was to create an accurate classification machine learning algorithm that can diagnose images of skin lesions

•The CNN model was able to generally detect cancerous skin lesions with a relatively high accuracy of 91% and F1-score of 0.86

•This algorithm could be improved upon in multiple different ways:

•Compiling skin lesion data from different datasets for both the classifier and data augmentation in order to disregard any potential inconsistences in the primary dataset that was utilized

•Incorporating early-stopping in order to prevent the classifier from overfitting to the dataset and determine the number of epochs to best minimize the error in the diagnosing the skin lesions.

•Utilizing more convolutional layers in the DCGAN in order to generate more realistic images of skin lesions

This CNN model and its approach in development have tremendous potential:

•The researcher is in the process of making the model accessible to the public through a mobile application called **SkinScan** that can be integrated in health care facilities

•The trained model is also in the process of being uploaded to a low-cost miniature computer such as a Raspberry Pi($35) connected to a USB web camera($10), and LCD Display(30$) and with a proper graphical user interface which will give access to possible patients in remote areas of the world

•Total cost of all components of prototype device comes to around $100 which enable widespread usability and accessibility for patients and healthcare infrastructure with inadequate funds

•Data augmentation using the DCGAN architecture can be vital to solving many other problems with machine learning that deal with a lack of consistent data.

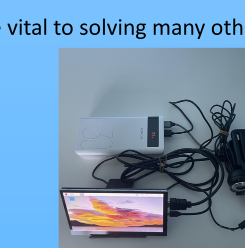


Image of Raspberry Pi Prototype for Low-Cost Solution to Diagnose Skin Lesions using the Trained Model