DATA AUGMENTATION TECHNIQUES FOR SMALL SIZE MAMMOGRAM IMAGES

By

GANPAT KUMAR
18BCE069

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
Ahmedabad 382481

[DATA AUGMENTATION TECHNIQUES FOR SMALL SIZE MAMMOGRAM IMAGES]

Minor Project Report

Submitted in partial fulfillment of the requirements

For the degree of

Bachelor of Technology in Computer Science & Engineering

By

GANPAT KUMAR
18BCE069

Guided By
PARITA OZA
[DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING]

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
Ahmedabad 382481

CERTIFICATE

This is to certify that the minor project entitled **"[**DATA AUGMENTATION TECHNIQUES FOR SMALL SIZE MAMMOGRAM IMAGES]" submitted by [GANPAT KUMAR (18BCE069)], towards the partial fulfillment of the requirements for the degree of Bachelor of Technology in Computer Science and Engineering of Nirma University is the record of work carried out by him/her under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination.

Parita Oza
Associate Professor,
Computer Science and Engineering Dept.,
Institute of Technology,
Nirma University,
Ahmedabad

Dr. Madhuri Bhavsar,
Professor and HOD,
Computer Science and Engineering Dept.,
Institute of Technology,
Nirma University,
Ahmedabad

ACKNOWLEDGEMENT

I would like to express my gratitude toward the staff of the computer science and engineering department as well as the faculties of Nirma University for providing me with a great opportunity to complete a project on "Data Augmentation Techniques For Small Size Mammogram Images". My sincere thanks go to Parita Oza without his support and guidance for the completion of this project. I am ensuring that this project was done by me and not copied from anywhere.

ABSTRACT/ Outline

On a variety of computer vision tasks, deep convolutional neural networks have excelled. However, in order to avoid overfitting, these networks mainly rely on massive data. Overfitting is when a network learns a function with a lot of variation in order to properly model the training data. Many application domains, such as medical image analysis, unfortunately, do not have access to huge data. Data Augmentation, a data-space solution to the problem of limited data, is the subject of this Project. Data Augmentation is a set of approaches for increasing the size and quality of training datasets so that stronger Deep Learning models may be generated. Geometric transformations(Rotation , Flipping , Translation, Clipping), colour space augmentations, and random erasing are some of the image enhancement algorithms described in this review. We show how Data Augmentation may help companies improve the performance of their models and expand constrained datasets to take advantage of big data's possibilities in this report.

LIST OF FIGURES

CONTENTS

1. Introduction

The invention of convolutional neural networks, deep neural networks have been effectively applied to computer vision applications such as picture classification, object recognition, and image segmentation (CNNs). Many fields of research are attempting to enhance current benchmarks by applying deep convolutional networks to Computer Vision problems. One of the most difficult issues is to improve these models' generalisation capacity. The term "generalizability" refers to the ability to apply a concept to the difference in performance between a model's performance when tested on previously observed data (training data) and data it has never seen before (testing data). The training data has been overfed by models with weak generalizability. Plotting the data is one approach to find out if you're overfitting. During training, accuracy in training and validation is measured at each epoch.

The validation error must decrease in tandem with the training error in order to develop usable Deep Learning models. Data augmentation is a very effective way to accomplish this. The enriched data will reflect a broader range of data points, reducing the distance between the training and validation sets, as well as any future testing sets. Data Augmentation is one of the most used techniques to overcome the over-fitting problem in small size data sets like in paper we have an MIAS data set(322 images). There are other Technique also available to overcome this problem those are[1]

1. Dropout: During training, a dropout is a regularisation approach that zeros out the activity levels of randomly selected neurons. Instead of relying on the prediction abilities of a limited subset of neurons in the network, this constraint drives the network to learn more robust properties.
2. Batch Normalization: Another regularisation technique that normalises the collection of activations in a layer is batch normalisation. Each activation is normalised by subtracting the batch mean and dividing by the batch standard deviation. This normalisation approach, along with standardisation, is a common preprocessing technique for pixel values.
3. Transfer Learning: Another intriguing approach for preventing overfitting is Transfer Learning. Transfer Learning involves training a network on a large dataset like ImageNet and then using those weights as the beginning weights in a new classification assignment. Rather than copying the complete network, including

fully-connected layers, only the weights in convolutional layers are usually replicated. Because many image datasets have low-level spatial properties that are better learned with huge data, this is a very effective method. It's still a work in progress to figure out how transferred data domains interact.

1. **Data Augmentation Techniques For Small Size Mammogram Images**

   When We have a small size of data, that time prediction on the train data set is very difficult. and overfitting is always that time. To overcome this problem we use data Augmentation. Our Project is to Perform some data augmentation on a small size of Mammogram images set.

2. **Objectives**

   Objectives for the project included, identifying the recent research in the field of Data Augmentation. Also, find which techniques of data augmentation are suitable for our project implementation and also apply those techniques in implementation to get Better Result.

3. **Problem Statement**

   Read Different-Different data augmentation techniques and identify which data augmentation technique is best suitable for our project. And create implementations related to this.

Literature Survey

Data augmentation is a technique for increasing the size of a training dataset artificially by producing changed versions of the images in the dataset. More data can lead to more skilled deep learning neural network models, and augmentation approaches can provide variations of the images that can increase the fit models' ability to generalise what they've learned to new images. Many researchers are working on the Data Augment Methodology to Overcome the Over-fitting Problem and present their solution. One of them is Auto-Augment. AutoAugment[3] is a programme that searches for better data augmentation policies automatically. In Their implementation, They created a search space in which a policy is made up of multiple sub-policies, one of which is picked at random for each image in each mini-batch. Each operation is an image processing function such as translation, rotation, or shearing, and the probabilities and magnitudes with which the functions are performed make

up a sub-policy. They utilise a search technique to discover the best policy that will provide the neural network with the best validation accuracy on a given dataset. Some Data augmentation technique[1] is Discuss in all researchs is

1. Rotation : The image is rotated right or left on an axis between 1° and 359° for rotation augmentations. The rotation degree parameter has a big impact on the rotation augmentations' safety. On digit identification tasks like MNIST, slight rotations like 1 to 20 or 1 to 20 could be useful, but when the rotation degree grows, the data's label is no longer preserved post-transformation. This Part Also we try in the implementation part of the MIAS Data set.
2. Flipping: Horizontal axis flipping is far more prevalent than vertical axis flipping. This is one of the simplest augmentations to use, and it's worked well on datasets like CIFAR-10 and ImageNet. This is not a label-preserving transformation for datasets involving text recognition, such as MNIST or SVHN.
3. Cropping: By cropping a central patch of each image, cropping photographs can be utilised as an useful processing step for image data with mixed height and width dimensions. Furthermore, random cropping can be employed to create a similar effect to translations. Random cropping differs from translations in that cropping reduces the size of the input, such as (256,256)–>(224, 224), whereas translations keep the image's spatial dimensions. This may or may not be a label-preserving transformation, depending on the cropping reduction threshold. In our Project, we Try from this Data augmentation.
4. Translation: To prevent positional bias in data, shifting images left, right, up, or down can be a highly effective adjustment. For example, if all of the images in a dataset are precisely centred, as is usual in face recognition datasets, the model must also be validated on perfectly centred images. The leftover space can be filled with a constant value such as 0 s or 255 s, or it can be filled with random or Gaussian noise as the original image is translated in a direction. This padding preserves the image's spatial dimensions after augmentation

Methodology

The workflow starts by importing the MIAS dataset, which is obtained from the Kaggle dataset . Then we Set the Image Label from Benign, Malignant and Normal to not Normal and Normal. Then we Apply Data Augmentation on the Image set and then using the CNN model we Train and test the model and get the result.

1. Dataset Used: In this Paper we use MIAS dataset that contains 322 images of Benign, Malignant and normal patients. We Have the CSV file that contain Seven feature "REFNUM(Indicate the patient Id)", "BG", "CLASS(Indicate that patient is normal or have the disease)", "SEVERITY(Indicate Benign, Malignant and normal patient)", "X(Indicate the Location of the Disease symbol in image)", "Y", " RADIUS(Radius of disease part)".Base on SEVERITY Benign, Malignant and normal form we convert them into normal and no normal patient contain 209 and 123 images.
2. Data Augmentation: Now We apply two parts of the data Augmentation Technique one is the rotation of the image and the second one is the flipping of the image. and Assign to the label of each image. and get 44640 images with each pixel size of 224 X 224.

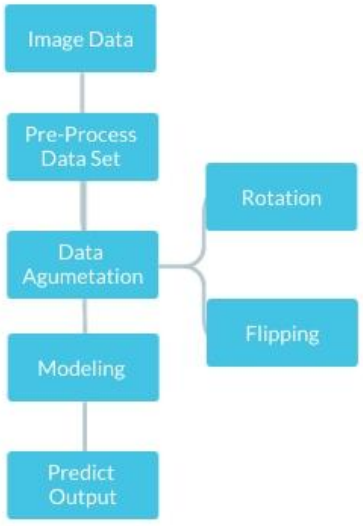

Fig. 1. Block Diagram of Proposed Methodology

Figure 3.1: Block Diagram For Proposed Methodology

3. Training and Testing
    a. We split the data into a training set and testing set. Training part contains 35712 images , 8928 images for testing the model we have built.
    b. Now We use the Sequential model with 2 hidden layers . Also, we use 0.25 dropout to overcome the over-fitting. We use kernel size is (3X3) and relu or sigmoid activation function for the model.

Result Analysis

we get 0.69314 loss and approx 50% accuracy . We Also Show How Accuracy and Loses vary with Each Epoch(In our Project we Consider Only for 3 Epoch).

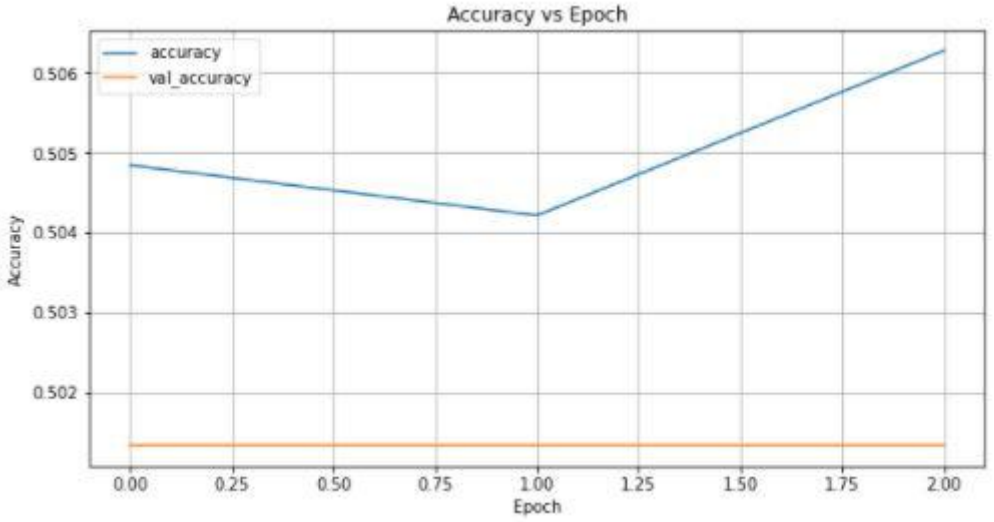

Figure 4.1: Accuracy vs Epoch

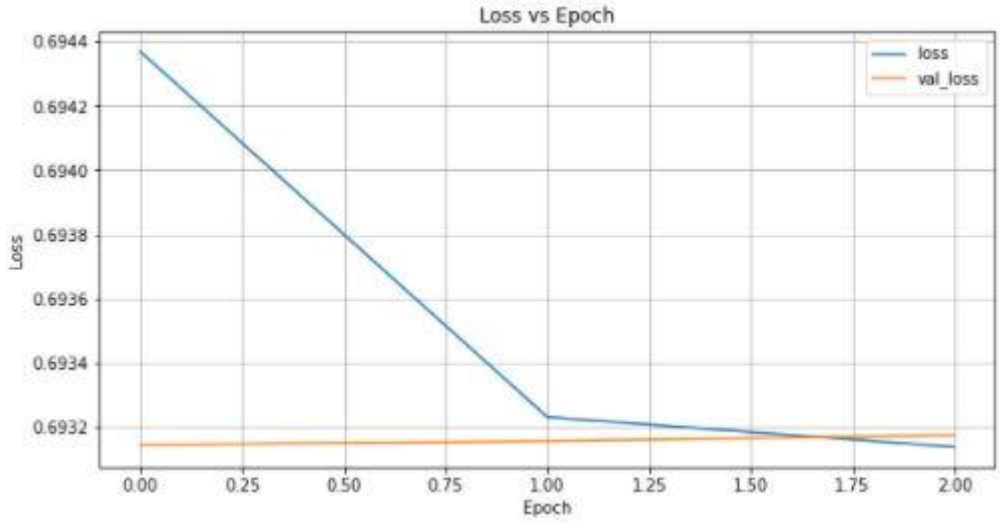

Figure 4.2: Losses vs Epoch

Future work and Conclusion

Due to a lack of data, this study provides a series of Data Augmentation solutions to the problem of overfitting in Deep Learning models. To avoid overfitting, deep learning models rely on large amounts of data. Using the approaches mentioned in this work, artificially inflating datasets achieves the benefits of big data in the limited data domain. Data augmentation is a powerful tool for improving dataset quality. Many augmentations have been proposed, all of which may be classified as either data warping or oversampling. Data Augmentation has a bright future ahead of it. The potential for using search algorithms that combine data warping and oversampling methods is immense. Deep neural networks' layered architecture opens up a lot of possibilities for Data Augmentation. The input layer is where the majority of the augmentations surveyed function. Some, like DisturbLabel, are derived from hidden layer representations, and one is even manifested in the output layer. The label space and the space of intermediate representations are two under-explored areas of Data Augmentation with promising results. Although many of these approaches and principles can be applied to other data domains, this paper concentrates on applications for images data. and Also we Discuss Some methodology proposed by some research paper authors. We Also Show Implementation on MIAS data set using the Sequential model. And get accuracy Approx 50%

References

1. Connor Shorten* and Taghi M. Khoshgoftaar *A survey on Image Data Augmentation for Deep Learning* Shorten and Khoshgoftaar J Big Data (2019) 6:60
2. Hang Min, Devin Wilson, Yinhuang Huang, Siyu Liu, Stuart Crozier, Andrew P Bradley, Shekhar S.Chandra, *Fully Automatic Computer-Aided Mass Detection And Segmentation Via Pseudo-Color Mammograms And Mask R-Cnn June 2019.*
3. Ekin D. Cubuk ∗, Barret Zoph∗, Dandelion Mane, Vijay Vasudevan, Quoc V. Le ´Google Brain, *AutoAugment:Learning Augmentation Strategies from Data* Apr 2019.
4. Zeshan Hussain1, Francisco Gimenez2, PhD, Darvin Yi2, Daniel Rubin2, MD, MS 1 Stanford University, Department of Computer Science, Stanford, CA; Stanford University, Department of Radiology, Stanford, CA *Differential Data Augmentation Techniques for Medical Imaging ClassificationTasks* AMIA Annu Symp Proc. 2017; 2017: 979–984. Published online 2018 Apr 16.