Differential Data Augmentation Techniques for Medical Imaging Classification

Ashish Vinodkumar, Pranav Manjunath, Tommy Tseng
Duke University

Abstract

In this paper we explore various data augmentation techniques such as Gaussian Blur, Gaussian Noise, Random Rotation, and Color Jitters, on the Digital Database for Screening Mammography (DDSM) using a VGG-16 model architecture. We learn that data augmentation in general boost model performance and reduce overfitting. Specifically, Gaussian Blur showcased to have the highest testing accuracy, and Random Rotation followed by Gaussian Blur has the highest Recall score. When plotting the ROC curves, Gaussian Blur tends to have the highest AUC score when compared to the other models. As a result, our analysis showed that given we are dealing with Medical Imaging DDSM data, Gaussian Blur showed to be the superior data augmentation technique across both the model validation metrics.

Objectives

1. To build a Deep Learning Model that can accurately predict Breast Cancer through Mammogram images
2. Understand if adding data augmentation techniques to train a Deep Learning Model will help improve model performance
3. Compare the model results for various data augmentation techniques and conclude which technique performs the best when predicting Breast Cancer.

Data Source

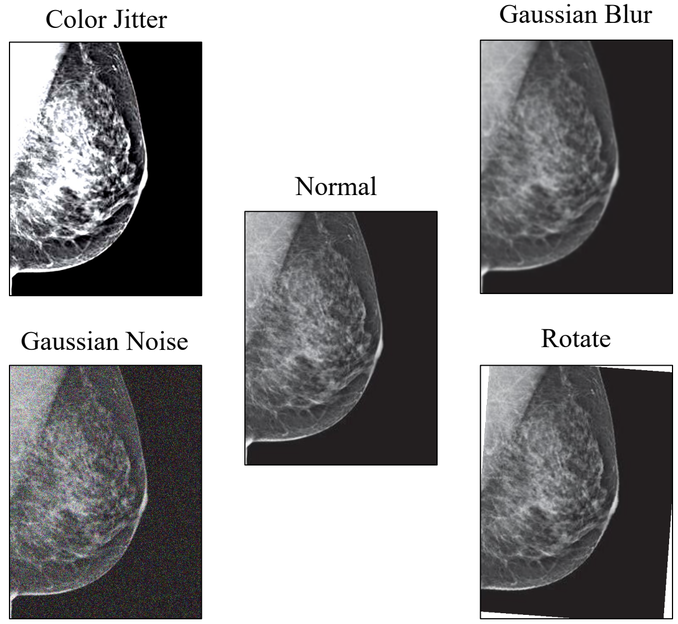
We extracted 2000 images from the Digital Database for Screening Mammography (DDSM). The images were classified into 2 classes, positive (malignant) and negative (benign) with each class having equal proportions of images. The dataset was split by a 80-20 percent ratio, which gave us 1600 training and 400 testing images.

Methods

We decided to use a **VGG-16 Model** with learning rate set to 1e-4, L2 regularization set to 1e-7 and dropout parameter p set to 0.5. Trained the model on 3200 images (1600 original + 1600 augmented) stratified on class and tested the model performance on 400 images.

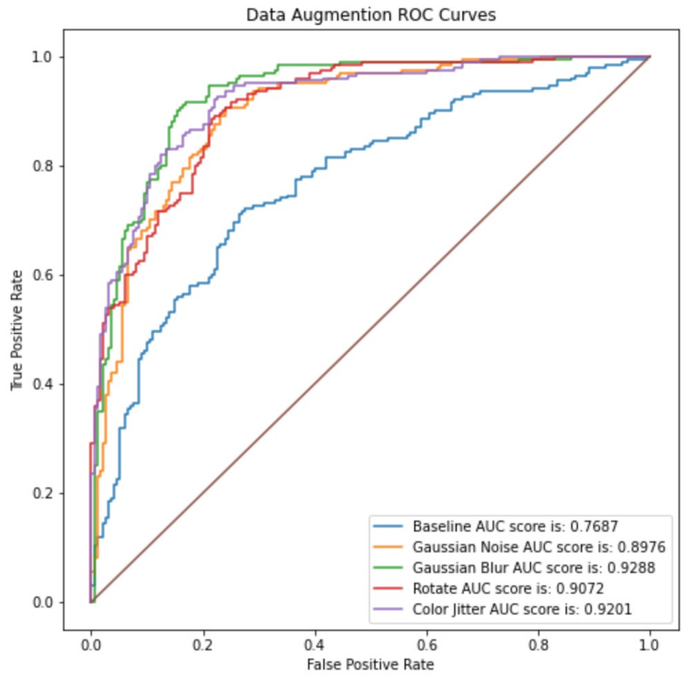
For Data Augmentation Techniques, we implemented

1) Gaussian Blur – Kernel Size: 17
2) Rotate – 10 degrees
3) Color Jitter – Contrast = 10
4) Gaussian Noise - Mean = 0.2, Std Dev. = 0.09



Results

	BASELINE	GAUSSIAN NOISE	GAUSSIAN BLUR	ROTATE	COLOR JITTER
Accuracy Score	0.705	0.82	0.86	0.83	0.845
Precision Score	0.766234	0.82	0.86	0.794643	0.870968
Recall Score	0.59	0.82	0.86	0.89	0.81
F1 Score	0.666667	0.82	0.86	0.839623	0.839378
Balanced Accuracy Score	0.705	0.82	0.86	0.83	0.845
Area Under the Curve	0.768675	0.897625	0.928825	0.9072	0.920075
Average Precision Score	0.768464	0.881193	0.9066	0.90808	0.922229



When comparing all classification metrics, **all data augmentations have performed better than baseline model**, indicating that adding data augmentation to the training dataset helps improve model performance. Specifically, **Gaussian Blur has the highest testing accuracy** (86%)

Furthermore, in the context of healthcare application, we would also want to maximize the recall and minimize the type-2 error. In such case, the augmentation technique **with highest testing recall score is Random Rotation** (89.0%).

Looking at the ROC curve with AUC values, we naturally observe that the **Baseline model has the lowest AUC score** of 0.7687 and that **Gaussian Blur has the highest AUC score** of 0.9288.

In conclusion, we found **Gaussian Blur to be the most effective augmentation technique** that has the highest classification metrics.

Conclusion

Having compared the VGG-16 model architecture driven models with 4 data augmentation techniques: Gaussian Blur, Gaussian Noise, Random Rotation, and Color Jitters; we observe that Gaussian Blur is the most effective data augmentation technique as it has the best scores for majority of classification metrics such as testing accuracy, AUC, and F1 Score. It also has the 2nd best recall score which is critical given the medical domain and the high price for false-negatives. As a result, Gaussian Blur is the most effective augmentation technique from our analysis with the DDSM dataset.

References

1. Hussain,Zeshan,et al.“(PDF)Differential Data Augmentation Techniques for Medical Imaging Classification Tasks.” ResearchGate, Apr. 2018, https://www.researchgate.net/publication/325532618_Differential_Data_Augmentation_Techniques_for_Medical_Imaging_Classification_Tasks.

2.Scuccimarra, Eric A. “DDSM Mammography.” Kaggle, 3 July 2018, https://www.kaggle.com/skooch/ddsm-mammography.

3.Raj, Ananta. “DDSM Breast Cancer VGG-19 Features Extraction.” Kaggle, Kaggle, 15 Apr.2021, https://www.kaggle.com/vortexkol/ddsm-breast-cancer-vgg-19-features-extraction.

4.“Gaussian Blur.” Wikipedia, Wikimedia Foundation, 15 Oct. 2021, https://en.wikipedia.org/wiki/Gaussian blur.

5.Swain, Anisha.“Noise in Digital Image Processing.” Medium,Image Vision,9Aug.2020, https://medium.com/image-vision/noise-in-digital-image-processing-55357c9fab71.

6.“Torchvision.transforms.”Torchvision.transforms-Torchvision 0.11.0 Documentation, https://pytorch.org/vision/stable/transforms.html.

7. How to Add Noise to MNIST Dataset When Using Pytorch." PyTorch Forums, 1 Nov. 2019, https://discuss.pytorch.org/t/how-to-add-noise-to-mnist-dataset-when-using-pytorch/59745.