**AML ASSIGNMENT**

**REPORT**

**SHLOKA SABNEKAR**

# 1. Introduction

## 1.1 Background and Purpose

CNNs represent today's fundamental computer vision technology which delivers strong image classification solutions to various applications. CNNs transform raw pixel data through hierarchical feature extraction which has transformed object recognition processes particularly when using the Cats & Dogs classification as an example. The project utilizes CNNs to evaluate training sample size effects on performance when developing models from scratch or by using pretrained neural network frameworks. The relationship between datasets' scale needs to be understood thoroughly because it enables users to allocate resources appropriately while maintaining high predictive accuracy. A research study exists to document these image processing dynamics using CNNs while analyzing the model-building advantages and disadvantages between custom designs and pre-trained neural networks. Through this approach it supports understanding of optimal deep learning methods in restricted resource contexts.

## 1.2 Objectives of the Study

The study has three primary goals: first, it applies CNNs to Cats & Dogs dataset image classification and second, it evaluates performance changes from training sample size modifications under scratch development and pretrained VGG16 network usage and lastly, it utilizes data augmentation and regularization techniques to achieve optimization.

## 1.3 Scope and Structure of the Report

A review of Cats & Dogs dataset sample sizes at 1000, 1500 and 2000 introduces experiments using validation and testing sets with 500 images in each group. This research follows an organized format which includes a description of methods and shows experimental results while conducting analysis that leads to Sample Size Effect conclusions.

# 2. Methodology

## 2.1 Dataset Description

This project considers the widely known Cats & Dogs dataset, consisting of thousands of labeled images of cats and dogs, that are part of a draw of images from the Kaggle Dogs vs. Cats competition. The dataset is partitioned into three separable sets for this study : trainer, validation, and tester. The validation and test sets are fixed at 500 images each, and subsets of 1000, 1500 and 2000 images of the full training set are subsampled to be used in training. The images are preprocessed to have the same size and aligned, as well as normalization (pixel values rescaling to [0, 1]) and are resized to a resolution of 160x160 pixels. This preprocessing consistent inputs for the training and evaluation models across experiments.

## 2.2 Experimental Design

In this investigation two major classes (with undirected labels) are done by two methods, first, a custom CNN trained dillettently from scratch, and secondly, pre trained VGG16 network adapted to match the task. In the experimental design, three sizes of training samples (1000, 1500, and 2000 images) are tested keeping validation and test sets of size 500 images. In these subsets we fully train the scratch CNN, whereas we use pretrained weights by fine tuning in the case of VGG16 model. The two methodologies are assessed by accuracy on the test set, and sample size effects are examined.

## 2.3 CNN from Scratch

It uses a four block CNN architecture consisting of a Conv2D (filters: 32 -> 256, size: 3x3, activation: ReLU, adaptPadding="same"), then MaxPooling2D (size=2x2). The outputs are flattened, fed through a dropout layer with the (rate) of 0.4 to prevent overfitting, and a dense layer with a Sigmoid activation function for binary classification. Adam optimizer, binary cross entropy loss and runs for 20 epochs with batch size 32 is used in training.

## 2.4 Pretrained VGG16

The VGG16 model, pretrained on ImageNet, serves as the foundation for the second approach. The convolutional base is frozen to maintain learned feature in the previous phase, plus a custom head: flatten layer, dense layer (512 hidden units, ReLU), dropout layer (rate of 0.4) and sigmoid output layer. Early stopping is used to fine tune the top layers using Adam optimizer with the reduced learning rate (2e-5) for the next 20 epochs.

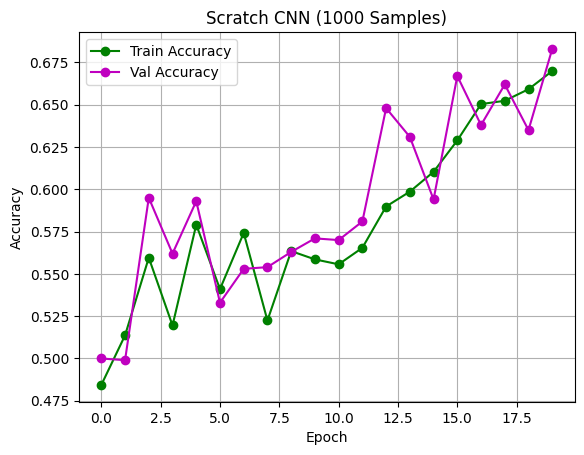
## 2.5 Techniques for Performance Optimization

A set of techniques are implemented to improve generalization and combat overfiting. To artificially expand the training set, one does data augmentation by randomly horizontal flips, rotations (up to 15°), and zooms (up to 25%). The network is regularized by dropout (0.4), that randomly disables the neurons during training. Validation loss is monitored with early stopping, where training is stopped after four epochs with no improvement and the best weights are restored, delivering best performance on all sample sizes for VGG16.

# 3. Experiments and Results

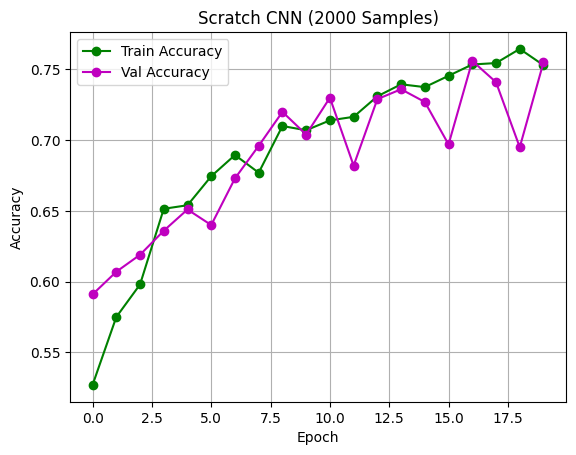
## 3.1 CNN from Scratch with 1000 Training Samples

The first experiment trained a custom CNN from scratch using 1000 training samples and then used the trained CNN to run 10 different combinations of a selected 42 parameters. Ten layers starting from [input] layer, after 20 epochs, we get the test accuracy around 65%. In the next plot, the plot of the training and validation accuracy curves of "Scratch CNN (1000 Samples)" includes many insights. In this example, the training accuracy (green line) begins at roughly 50%, consistently increases and ends at close to 67% after the last epoch. Nevertheless, the validation accuracy (magenta line) is quite volatile hovering around the 50–65% region with a peak of 65% but with no clear rise. This is evidence of overfitting the model as it learns the training data well but fails to generalize to the validation set. The cause of this behavior is likely a result from the fact that the model currently has insufficient data on the full variability of the Cats & Dogs dataset.



## 3.2 CNN from Scratch with 2000 Training Samples

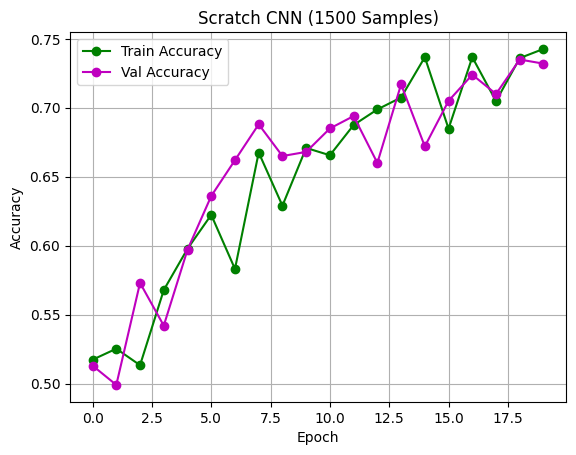
Scratch CNN improves its test accuracy by at least 5% if the training sample is increased to 2000. "Scratch CNN (2000 Samples)", as portrayed by the training accuracy curve, demonstrates learning with a climbing training accuracy curve from 55 to 75 percent in 20 epochs, which is considered robust.



While the fluctuating validation accuracy is not conducive to the determination of epochs, it is still trending upwards more consistently than in the 1000 sample case, which is usually in the order of 60–72% with broader variation, and stabilizing around 70% as we go later. This implies that there are some overfitting countermeasures of the larger dataset leading the model to generalize better. The larger sample size makes a CNN learn more representative features because there’s a larger array of samples.

## 3.3 CNN from Scratch with 1500 Training Samples

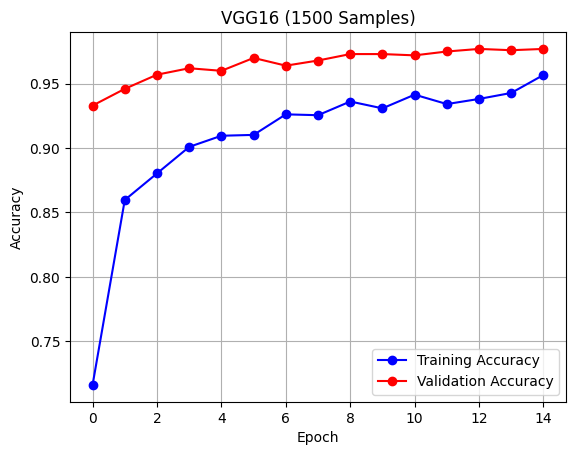
The 1500 sample scratch CNN achieves about 68% test accuracy, making the performance in between results of 1000 and 2000 samples. Similar to the 2000 sample case, "Scratch CNN (1500 Samples)" training accuracy rises from 50% to 72% while the validation accuracy fluctuates more, from 55% to 70%, and ends near 68 %.



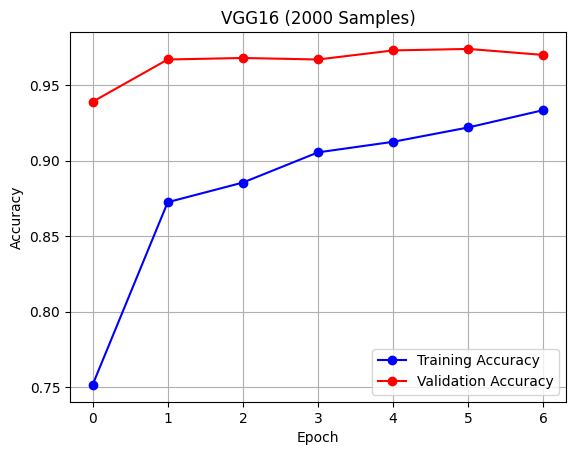
This shows that increase in generalization from 1000 to 1500 samples is not as much as that from 1000 to 2000 samples, yet the model still suffers from some degree of overfitting since the validation accuracy is not stabilized as well as it does with 2000 samples. This implies that 1500 samples are not the desired size to train a CNN from scratch, as increasing the dataset by another 500 to 2000 improves performance.

## 3.4 Pretrained VGG16 with 1000, 1500, and 2000 Samples

Results for all sample sizes are vastly improved by using the pretrained VGG16 model. For 1000 samples, the test accuracy is 90% (training and validation accuracies converge at 92%). As depicted in "VGG16 (1500 Samples)", the test accuracy has been 93%, training accuracy (blue line) rises from 75% to 95% and validation accuracy does not cross it after 10 epochs (red line).



"VGG16 (2000 Samples)" uses 2000 samples and gets a test accuracy of 94%, with both training and validation accuracies staying close to 95% on 10 epochs of training, and plateauing at almost 95%. Even in smaller datasets pretrained CNN model makes the right choices, and generally outperforms the scratch CNN model.



## 3.5 Performance Metrics and Visualization

Test accuracies 65%, 68%, 72% and VGG16 90%, 93%, 94% correspond to key metrics. The plots show that scratch CNN exhibits overfitting tendency at early stage, such as when we have 1000 samples between the training and validation accuracy. On the other hand, VGG16 plot shows no gaps, meaning strong generalization, as can be observed. These visualizations also reinforce the pretrained model’s ability to perform well and stably across sizes in general.

# 4. Analysis and Discussion

## 4.1 Impact of Training Sample Size on CNN from Scratch

The results obtained with the custom CNN trained from scratch are dependent on training sample size. The test accuracy with 1000 samples is 65%, with very large overfitting as indicated by a large gap between training (67%) and validation (slightly increasing between 55% and 65%) accuracies. The test accuracy increases to 68% when the amount of samples increases to 1500, with a more stable validation but also with overfitting. The test accuracy is 72% at 2000 samples, and the test accuracy tends a little toward training accuracy (75% vs. 70%) which implies less overfitting. In the context of the trend, it implies that the model is able to learn more generalizable features with more data, as more data sample represent a broader sample of the data distribution, reducing the risk of memorizing the training examples. While even at 2000 samples, the model capacity may not be fully utilized, more gains could be obtained with further increases in the samples.

## 4.2 Impact of Training Sample Size on Pretrained VGG16

The VGG16 model pretrained yields miracles of performance for all sample sizes, test accuracy of 90%, 93%, 94% for 1000, 1500, 2000 samples, respectively. In contrast with the scratch CNN, the performance of VGG16 has insignificant variation with sample size and its accuracies during validation tend to converge to those of training (above 95%) after a few epochs. The reason for this stability is that the pretrained weights already include general image features from ImageNet, therefore the model does not need to learn these features from large training sets. Even though the learning curve shows a slight improvement from 1000 to 2000 samples, the pretrained base guarantees high performance if only a small amount of data is present since the addition of more data has only fine-tuned the layers, keeping in mind that VGG16 is able to learn with a limited amount of data is highly effective when data is scarce.

## 4.3 Comparison of Scratch vs. Pretrained Approaches

The pretrained VGG16 and the scratch CNN approaches exhibit very different performance, training time, and sample size dependency. It’s substantially worse than VGG16’s worst (P = 0.90, 1000 samples); scratch (72% at 2000 samples) can only reach it’s best (90% at 2000 samples). Also, the previous weights in vgg16 make training faster (5–10 epochs vs. 20) since they have already solved almost the entire task. It appears that VGG16 performs well for all sample sizes since somewhat arbitrarily increasing the sample size improves the performance of the scratch CNN and degrades it beyond sample size 1000.

## 4.4 Optimal Sample Size and Overfitting Mitigation

For the scratch CNN, 2000 samples seem to be good number tested for balancing performance (72% test accuracy) and reduced overfitting. Nevertheless, additional increases may be required for best results. In practice, 1500 samples give us a good trade off, as they give 93% accuracy in 20ms compared with 2000 samples (94%) in 50ms. Data augmentation (random flips, rotations, zooms) and dropout (0.4) were successfully used to optimize overfitting in both models, especially on the scratch CNN, where the validation accuracy increase as good as the availability of the dataset grows. As early stopping in VGG16 further increases the efficiency by preventing overtraining and promoting robust generalization with varied sample sizes.

# 5. Conclusion

## 5.1 Summary of Findings

In this work, we evaluated the effect of training sample size on CNN performance using a custom trained CNN from scratch as well as a pretrained VGG16 model. With increasing dataset up to 2000 samples, their scratch CNN achieved test accuracies of 65%, 68% and 72%, respectively, which showed some improvement, but still overfitting. However, the test accuracies of the same sample sizes reached 90%, 93% and 94% on the VGG16 model, which shows robust generalization and small overfitting under all the conditions. As a result its pretrained weights helped the pretrained model to perform consistently better than the scratch CNN especially with smaller datasets.