**Report: Impact of Training Sample Size and Network Selection on Model Performance**

**INTRODUCTION:**

This study investigates the effects of different neural network designs and training sample sizes on model performance in a Cats vs. Dogs image categorization challenge. Both models built from the ground up and those that make use of pre-trained networks are examined in the study. Furthermore, data augmentation and other optimization approaches are used to improve the model's generalization to new data and prevent overfitting**.**

Additionally, the diversity of the training data is increased by the use of data augmentation techniques including image flipping, rotation, and scaling. By preventing overfitting, these strategies make sure the model works effectively on both training and unseen data.

The report's main goal is to investigate how the efficacy and versatility of machine learning models for image classification are impacted by variables such as sample size, neural architecture, and optimization techniques like data augmentation.In a Cats vs. Dogs image classification challenge, this study investigates the effects of variations in neural network architectures and training sample sizes on model performance. Both scratch-built models and models using pre-trained networks are examined in the study.

The study aims to understand how sample size impacts the model's accuracy and performance by altering the number of training samples. The performance of several neural network architectures, whether built from the ground up or improved using pre-trained models, is compared.

**1. Initial Configuration for Model Training from Scratch:**

• Training collection: 1000 pictures

• Validation collection: 500 pictures

Five hundred pictures make up the test set.

**Outcomes of Performance:**

Accuracy of training: 78.2%; accuracy of validation: 73.6%

• 70.1% test accuracy

These findings show that while the model could learn from a limited dataset, overfitting was a problem. In contrast to training accuracy, test accuracy is lower, which emphasizes this problem. The gap implied that methods such as expanding the training set size or augmenting data would be required for improved generalization.

**2. Methods of Optimization:**

**Implementation of Data Augmentation:**

• To lessen overfitting and enhance model performance, data augmentation was used to produce a variety of permutations of the training photos.

**Outcomes following data augmentation**:

69.6% is the training accuracy.

• 65.4% is the validation accuracy.

• 60.87% is the test accuracy.

Data augmentation in this instance did not result in better performance, indicating that a bigger training set was necessary for the model to properly benefit from the augmentation strategies.

**3. Expanding the Training Sample Dimensions**

**Fresh Configuration:**

• Training collection: 1500 pictures

• Validation collection: 500 pictures

Five hundred pictures make up the test set.

**Outcomes of Performance:**

Accuracy of Training: 77.8%

Accuracy of Validation: 75.0%

Accuracy of the test: 77.8%

With a larger training sample size, the model demonstrated a discernible increase in test and validation accuracy. As evidenced by the better and more reliable test accuracy, this showed that the model's capacity to generalize had improved and that overfitting had decreased.

**4. Establishing the Optimal Training Sample Dimension**

In order to determine the ideal training sample size, tests were carried out using a collection of

• 1500 training photos

• Validation set: one thousand pictures

Test collection: 500 pictures

**Performance Outcomes:**

• Accuracy of Training: 72.11%

72.9% is the validation accuracy.

• 72.1% is the test accuracy.

Across training, validation, and test datasets, this configuration produced consistent results by striking a solid compromise between performance and overfitting.

**5. Making Use of a VGG16 Pre-trained Network**

Feature Extraction for VGG16:

• To avoid starting the model from scratch, feature extraction was carried out using the VGG16 network that had already been trained.

**Performance Outcomes:**

99.78% training accuracy and 97.65% validation accuracy

Across all datasets, the pre-trained model performed noticeably better than the model that was created from scratch. This suggests that employing a pre-trained network is a practical strategy, particularly in situations when there is a shortage of data.

Using a pre-trained network to augment data:

**Accuracy of training:** 97.06%; accuracy of validation: 97.85%

• 97.1% is the test accuracy.

Performance was further improved by introducing data augmentation using the pre-trained VGG16, highlighting the importance of integrating augmentation with a pre-trained network to increase generalization.Final Thoughts :

• **Training Sample Size:** The greater accuracy seen when the sample size was increased indicates that larger training datasets result in better generalization and enhanced model performance.

• **Pre-trained Networks**: When working with sparse data, models built using pre-trained architectures, such as VGG16, performed noticeably better than those trained from scratch. Transfer learning challenges are best suited for pre-trained models since they may utilize the prior knowledge that is already ingrained in them.

The efficacy of data augmentation, a helpful method for enhancing generalization, is contingent upon the network architecture and the size of the training set. Using it in conjunction with a pre-trained network produced the greatest results.

The greatest results were obtained in this investigation using a pre-trained network (VGG16), data augmentation, and a higher training sample size.