# Fine-Grained Visual Classification

**Aniketh Garikaparthi**
aniketh20360@iiitd.ac.in

**Deeptanshu**
deeptanshu20371@iiitd.ac.in

**Rahul Sehgal**
rahul20232@iiitd.ac.in

## 1 Problem Statement

Canines are an important part of our lives. With a large number of dog breeds, identification of dog breeds can be a cumbersome task for the average person. Classification of dog breeds is considered a problematic fine-grained categorization problem, with many similarities and dissimilarities between classes. Our goal is to train models that can perform this task of dog-breed classification reliably and efficiently. Dog-breed classification offers an interesting opportunity to discover and broaden our knowledge related to fine-grained classification, which can further be applied to other classification tasks.

## 2 Related work and Baselines

The paper which published the dataset provides us with the baseline results as a function of accuracy with respect to the number of samples in each class (Khosla et al., 2011). They achieve an accuracy of 22% with 100 training images per class. A graph was plotted to demonstrate how much data was required for training. But there were no details regarding what classifier was used so it could not be reproduced.

(Luo et al., 2020) uses ResNet-50 as a backbone for their algorithm and also as a baseline for evaluation for the task of fine-grained image classification. The reported accuracy for this model was 88.1%. They use an approach that involves a semantic groping module, which arranges feature channels with different properties into different groups. Then a feature enhancement module was used, increasing the performance by improving the sub-features. These additions to the model resulted in a boost in performance in all the backbones using all the datasets. With ResNet-50 as a backbone trained on stanford dogs dataset, the accuracy improved to 88.8%.

(Kim et al., 2022) implements a neural tree decoder with a vision transformer as the backbone. The vision transformer provides attention-based contextual image patches, while the neural tree decoder is used to solve the limitations faced by the vision transformers. This also aids in the interpretability of the results. The proposed method with DeiT-B as the backbone results in an impressive 93.6% accuracy, which is the current state of the art for this dataset.

(James, 2023) use ResNet-50 and MobileNetV2, achieving accuracies of 83.3% and 79.5% respectively. Global Average pooling and a dense output layer with softmax activation were used in the implementation of MobileNet.

(Darvish et al., 2018) mentions that vision transformers require a lot of data, which is not generally available. Thus, they make the use of general adversarial networks (GANs) to generate synthetic data which can then be used to boost performance.

## 3 Data-set Details

The dataset was first published in 2011 (Khosla et al., 2011) as the Stanford Dogs dataset. This dataset contains 120 classes with a total of 20,580 images (approximately 150 images per class). The annotations include bounding-box and class labels. The dataset is made for the task of fine-grain image classification, as there are a lot of similarities between the different classes. The backgrounds of the images consist of many humans and man-made environments, resulting in a lot of variation. The images present in this dataset were taken from ImageNet.

## 4 Experiment setup and results

We employed the models, MobileNetV2 and ResNet-50 as the two baselines for our approach to fine-grained image classification on the Stanford Dogs

| Model | Training Accuracy | Validation Accuracy | Testing Accuracy |
|---|---|---|---|
| ResNet-50 - Reported | - | - | 88.1 |
| MobileNetV2 - Reported | - | - | 79.5 |
| ResNet-50 - Reproduced | 99.77 | 79.54 | 80.07 |
| MobileNetV2 - Reproduced | 98.52 | 70.84 | 74.14 |

Table 1: Comparision of baseline accuracies on the Stanford Dogs dataset between previously reported and the reproduced.

dataset (Khosla et al., 2011).

### 4.1 Data Pre-Processing

Each image from the training, validation, and testing set is resized to a 224x224 image by centre cropping the images for each set and selecting a portion of size 224x224 pixels. The pixels of all the images are then scaled by dividing them by 255, along with normalizing the values of pixels.

### 4.2 Implementation Details

We initialized both models with weights pre-trained on the ImageNet dataset. Further, we employed the Adam Optimizer with learning rate, *lr* set to 0.0001, and Cross Entropy Loss as the loss criterion. The model's accuracy is validated on 20% of the original dataset and finally tested on the test dataset containing 10% of the original dataset.

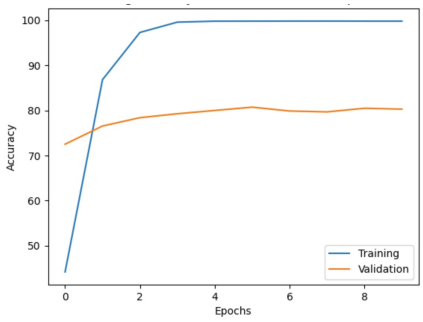
### 4.3 Hyperparameter Tuning

We experimented with varying combinations of learning rates, batch sizes, decaying learning rates and weight decay parameter in the Adam Optimizer to prevent overfitting as much as possible. It was noted that MobileNetV2 performed significantly better and converged faster with a smaller batch size of 32, whereas ResNet-50 provided a higher accuracy when trained on a larger batch size of 128.

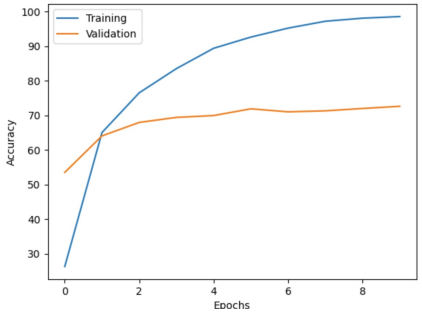
We experimented with decaying learning rates and weight decay to prevent overfitting on the training split. Still, it was noted that these methods only slowed the training process without offering any better generalization and hence, were not used in the final run.

### 4.4 Results

Table 1 shows the reported and reproduced accuracies of the ResNet-50 model and the MobileNetV2 model on the Stanford Dogs dataset. In both cases, overfitting is observed as we can see from Figure 1 showing the training and validation accuracies for both the models.




(a) ResNet-50




(b) MobileNetV2

Figure 1: Training and Validation accuracies vs epochs on both models

- **ResNet-50** - The ResNet-50 model is able to achieve high training accuracy, greater than 95% in just 3 epochs. From Figure 1 (a), we can see the difference in the accuracies between the training and the validation sets, suggesting an overfit on the training data. However, the accuracies obtained on the validation(79.54%) and testing sets(80.07%) are at par with the previously obtained baselines having different splits.

- **MobileNetV2** - The MobileNetV2 model is able to achieve high training accuracies, greater than 95% in 8-9 epochs. As MobileNetV2 is a smaller model than the ResNet-50, it takes slightly less time for training and more epochs to obtain optimum accuracy. From Figure 1 (b), we can see the difference in the accuracies between the training and the validation sets, suggesting an overfit on the training data. However, the accuracies obtained on the validation(70.84%) and testing(74.14%) sets are at par with the previously obtained baselines having different splits.

## 5 Observations and Future work

From our results, we observe that our baseline models are close to the existing baselines available for the task in terms of accuracy metrics. This gives us an idea regarding the scope of improvement which can involve utilizing bounding boxes to analyze the regions of interest in images which can improve the performance on the task of fine-grained classification for dog breeds. This approach provides insight into the scope of potential improvements and enables the model to better learn the fine features of the images.

We can also make use of Attention-based models like the Vision Transformer, which has been able to achieve higher accuracies for fine-grained image classification(Kim et al., 2022). More data augmentations can also be added along with the use of Generative Adversarial Networks, which can synthetically produce new data based on the original training data. These methods have been found to improve the performance of fine-grained image classification.(Darvish et al., 2018)

## References


Mahdi Darvish, Mahsa Pouramini, and Hamid Bahador. 2018. Towards fine-grained image classification with generative adversarial networks and facial landmark detection. In *2018 26th Iranian Conference on Electrical Engineering (ICEE)*, pages 1597–1602. IEEE.

Hailey James. 2023. Cs109 final project: Dog superbreed classification.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO.

Sangwon Kim, Jaeyeal Nam, and Byoung Chul Ko. 2022. ViT-NeT: Interpretable vision transformers with neural tree decoder. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11162–11172. PMLR.

Wei Luo, Hengmin Zhang, Jun Li, and Xiu-Shen Wei. 2020. Learning semantically enhanced feature for fine-grained image classification. *CoRR*, abs/2006.13457.