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. Our code with inference pipeline is available at: DL Project Final

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.

We have also artificially generated data by making the use of Generative Adversarial Networks (GANs). We used a pre-trained GAN, Big GAN 256, to gener-

ate 30 images for each class, which were appended to the training data. Our truncation variable was set to 0.5 to ensure a balance between accuracy and variations of the generated data.

Figure 1: Example of bounding box on image

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 dataset (Khosla et al., 2011). We have taken 3 different setups of training data. First is taking the training data as it is, second is applying the bounding box annotations and third is mixing the artificially generated data with our original data.

4.1 Data Pre-Processing

We are taking a split of 70-20-10 for the training-validation-testing sets from the whole dataset.

- **Bounding-Box** - We are annotating the training data bounding box annotations available with Stanford dogs dataset by overlaying each training data image with the bounding box rectangle. The resulting image included the original image with the bounding box drawn over it.
- **GAN generated Data** - For using the data generated using GANs, we had to merge it with our existing training data. For this, we created the dataset classes for both the datasets and then concatenated them together. Further, we shuffled this concatenated dataset to ensure our model receives the artificial data evenly and not all at once while training.

(a) Examples of GAN generated images of different dog classes

(b) Variations of GAN generated images for one dog class

Figure 2: Examples of GAN generated images

Further, Each image from the training, validation, and testing set is resized to a 224x224 image by center 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.

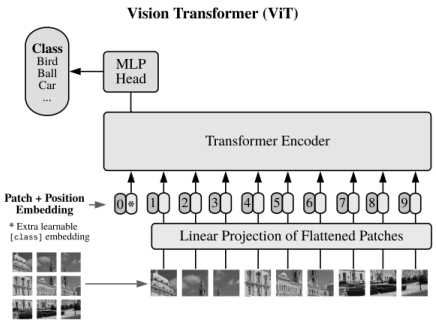


Figure 3: Diagram of the ViT model-(Best Performance),

	Testing set Accuracies for different cases		
Model	**Unbounded**	**Bounded**	**GAN+ Unbounded**
Previous Baseline-ResNet50	80.07	-	-
Previous Baseline-MobileNetV2	72.02	-	-
ResNet-50	81.29	81.97	81.63
MobileNetV2	72.4	74.1	75.36
ViT	86.97	**87.85**	87.026
Ensemble	-	84.45	-

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

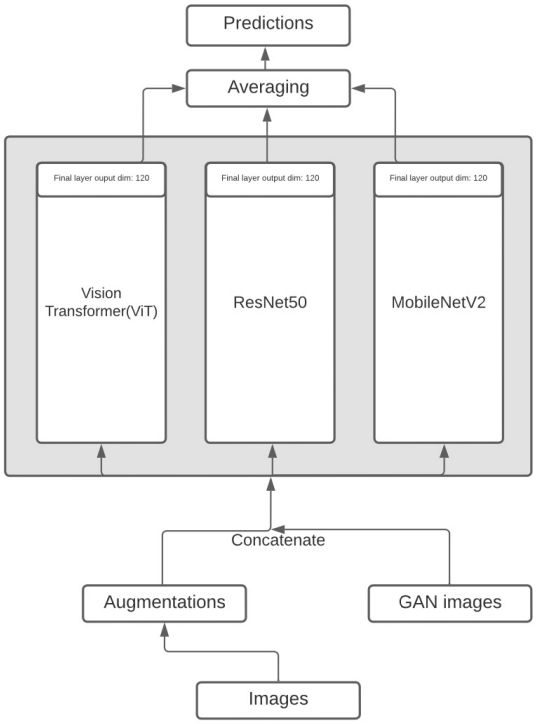


Figure 4: Diagram of the Ensemble model

4.2 Implementation Details

We have employed four models for our fine-grain classification task, which are RestNet50, MobileNetV2, Vision-Transformer(ViT) and an Ensemble model which includes all three previously stated models. We initialized the models with weights pre-trained on the ImageNet dataset. Further, we employed the Adam Optimizer with learning rate, *lr* set to 1e-5 for RestNet50, Vision-Transformer(ViT) and Ensemble model, and 1e-4 for MobileNetV2 and Cross Entropy Loss as the loss criterion. The model's accuracy is validated on the validation set which is about 20% of the original dataset and finally tested on the test dataset containing 10% of the original dataset.

The Resnet50 and MobileNetV2 models had earlier been used as our baseline models and are now used to compare with new models as well as compare their results with augmentations on training data.

- **ResNet-50, MobileNetV2, ViT** - The stated models have been pre-trained on the ImageNet dataset, hence having different output dimensions of the final classification layer. We have changed the output dimensions of the final classification layer of the models to be 120 for our use case of 120 classes.
- **Ensemble model** - The Ensemble model from Figure 4 also incorporates the dimensionality changes of ResNet50, MobileNetV2 and ViT. While doing the forward pass of this model, we are averaging the outputs of the three models and using them for prediction.

$$preds_{ensemble} = \frac{preds_{ViT} + preds_{RN50} + preds_{MNV2}}{3} \quad (1)$$

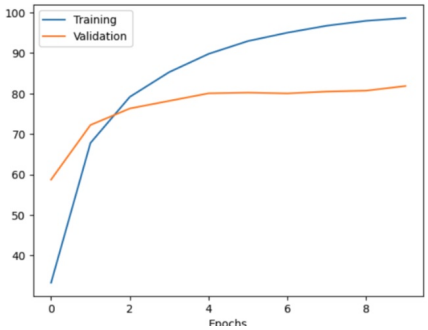
We have tested our models for different cases such as unbounded-training images, bounded-training images, and unbounded-training images concatenated with GAN generated images to give us an accurate representation of developments with data-augmentations.

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 our models performed better and converged faster with a batch size of 32, with the exception of the ResNet-50 model, which performed best with a batch size of 128. The Vision Transformer worked optimally with a learning rate of 1e-5, whereas the other two gave optimal results with a learning rate of 1e-4.

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 were not used in the final run.



(a) ViT-(bounding box case)

Figure 5: Training and Validation accuracies vs epochs on the best model

4.4 Results

Table 1 shows all our models' reported and reproduced accuracies on the Stanford Dogs dataset. In all cases, overfitting is observed to distinctive levels, as we can see from Figure 5 showing the training and validation accuracies on our best model.

For the bounding box annotations, we had also tried cropping the training data, but that resulted in severe drops in performance. Our inference from this experiment was that cropping the background inhibits our models' ability to learn to ignore the background on its own.

- **ResNet-50** - The ResNet-50 model is able to achieve high training accuracies, greater than 95% in about 5-6 epochs for all different cases. The model is able to achieve about 81.29% accuracy on the test set for the case of unbounded images, while the case of bounded images gets an accuracy of about 81.97%. The model attains a test accuracy of 81.63% on un-bounded images with GAN-generated images, suggesting bounded-box improving the overall accuracy for the model.

- **MobileNetV2** - The MobileNetV2 model is able to achieve high training accuracies, greater than 95% in about 5-6 epochs for all different cases. The model can acheive about 72.4% accuracy on the test set for the case of unbounded images, while the case of bounded images gets an accuracy of about 74.10%. The model attains a test accuracy of 75.36% on un-bounded images with GAN-generated images, suggesting GAN images improve the overall accuracy of the model.

- **ViT** - The Vision Transformer(ViT) model can achieve high training accuracies, greater than 95% in about 3-4 epochs for all different cases. The model is able to achieve about 86.97% accuracy on the test set for the case of unbounded images, while the case of bounded images gets an accuracy of about 87.51%. The model attains a test accuracy of 87.02% on un-bounded images with GAN-generated images, suggesting bounded-box improving the overall accuracy for the model.

- **Ensemble** - The Ensemble model, which comprises all three models stated, is able to achieve high training accuracies, greater than 95% in about seven epochs. The model can achieve about 84.45% accuracy on the test set for the case of bounded images. This shows us that by ensembling the outputs of our models, we are not getting better results than ViT, which is performing best in all distinctive cases.

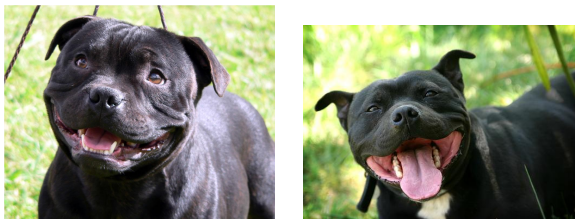
Our results have shown improvements from previously reported baseline results. These results depict how different models learn unique features with different data augmentations.

5 Observations and Error Analysis

We observed that the ViT model outperformed the previous baseline models we implemented. A key difference was that the ViT could better generalize

the given data and showed higher validation and test accuracies than other models. A general trend observed between all three models was high training accuracies nearing 99%, and there was significant overfitting.

A deeper look at the class-wise F-1 scores revealed that a few classes with low scores were interrelated and had high inter-class similarities and low inter-class variance. A good example of such an occurrence is shown below.



(a) Staffordshire (b) American Staffordshire

Figure 6: Training and Validation accuracies vs epochs on the best model

As noticed from the Figures 6 (a) and (b), belonging to Staffordshire Bullterrier and American Staffordshire Terrier, there are little to no noticeable differences between the breeds. The differences between the two breeds lie in the size of the dog; our proposed model implementation fails to distinguish between such instances where the differences are subtle.

6 Contribution

Aniketh was involved in finding existing literature on the topic of Fine-Grained Image Classification and taking inspiration from methodologies and experimentations done in published works on the same topic. In Experimental Setup, Aniketh worked on implementing bounding boxes and hyperparameter tuning.

Deeptanshu was responsible for augmenting the dataset by using GANs Model to generate 30 images per class and then training of the models that used this data. Deeptanshu's contribution also involved ideation and finding out potential solutions from existing work, along with evaluation of the results generated.

Rahul was involved in implementing and tuning the model architectures for our task, i.e. Resnet50, MobileNetV2, ViT and our Ensemble model for different cases discussed, while taking inspiration from previous works on Fine-grained image classification along with ideation of augmentation tasks.

While we all focused on a particular task, we reviewed each other's work and helped each other out along the way.

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.