Enhancing Fine-Grained Image Classification with Attention Mechanisms and Transfer Learning: A Case Study on the Stanford Dogs Dataset

Jay Vishal Mehta

Abstract

This research explores enhancing fine-grained image classification using transfer learning and attention mechanisms. The study applies a ResNet50 architecture pretrained on ImageNet and augmented with Convolutional Block Attention Modules (CBAM) to the Stanford Dogs dataset. Results show significant improvements in classification accuracy, with the model achieving 86.92% accuracy on the test set. This performance gain demonstrates the effectiveness of combining transfer learning and attention mechanisms in overcoming challenges in fine-grained image classification, paving the way for more accurate systems in domains requiring detailed visual analysis.

Keywords: computer vision, deep learning, transfer learning, attention mechanisms, fine-grained classification

1. Introduction

Fine-grained image classification, which involves distinguishing between subcategories of visually similar objects, presents significant challenges in computer vision. This task is especially valuable in fields like biodiversity monitoring, medical diagnosis, and autonomous driving, where precise classification of visually similar objects is critical. The difficulty lies in the subtle intra-class variations (differences within the same category) and inter-class similarities (resemblances across different categories), which complicate accurate feature extraction and classification.

Two key challenges impede progress in fine-grained classification: First, the scarcity of labeled data and the limitations of conventional feature extraction techniques. Limited labeled examples impede the training of robust machine learning models, potentially resulting in poor generalization to unseen data. Additionally, traditional feature extraction methods may struggle to capture the nuanced details crucial for differentiating between closely related subcategories.

This research explores the potential of combining transfer learning and attention mechanisms to address these challenges. Transfer learning is applied to adapt models pre-trained on large-scale datasets to fine-grained tasks, mitigating the challenge of limited labeled data. Attention mechanisms are integrated to enhance the model's ability to focus on discriminative image regions, aiming to improve feature extraction for subtle distinctions.

The study applies these techniques to the Stanford Dogs dataset, utilizing a ResNet50 architecture pretrained on ImageNet and augmented with Convolutional Block Attention Modules (CBAM), a mechanism that enhances the model's focus on important regions of an image. Results show improvements in classification accuracy, with the model achieving 86.92% accuracy on the test set, an increase from the baseline. This performance suggests that the combination of transfer learning and attention mechanisms offers benefits in fine-grained image classification tasks.

The remainder of this paper is organized as follows: the Background Literature section provides an overview of existing approaches, transfer learning applications, and attention mechanisms; the Methods section details the model architecture, implementation, and evaluation metrics; the Experiments and Results section presents findings and statistical analyses; and finally, the Discussion and Conclusion section reflects on the study’s implications, limitations, and potential future work.

2. Background Literature

Existing Approaches: Overview of Current Methods in Fine-Grained Image Classification

Fine-grained image classification (FGIC) tasks present significant challenges due to small inter-class variance and high intra-class variance. Convolutional Neural Networks (CNNs) such as ResNet and VGG have demonstrated strong performance on large-scale datasets like ImageNet (He et al., 2016; Simonyan & Zisserman, 2014). However, their ability to capture discriminative features for FGIC is limited. Recent work has introduced attention mechanisms and part-based methods to refine feature extraction and enhance classification performance by focusing on key object parts (Zheng et al., 2017).

Transfer Learning: Benefits and Applications in Image Classification

Transfer learning has proven highly effective in FGIC, particularly when data is scarce. Pre-trained models on large datasets like ImageNet are fine-tuned for specific fine-grained tasks, significantly reducing the amount of labeled data required and speeding up convergence. For example, Vision Transformer (ViT) models leverage transfer learning to improve performance in FGIC tasks by focusing on global and local features (Dosovitskiy et al., 2020). This approach, as demonstrated in "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," shows that a pure transformer applied directly to image patches can perform exceptionally well on image classification tasks (Dosovitskiy et al., 2020).

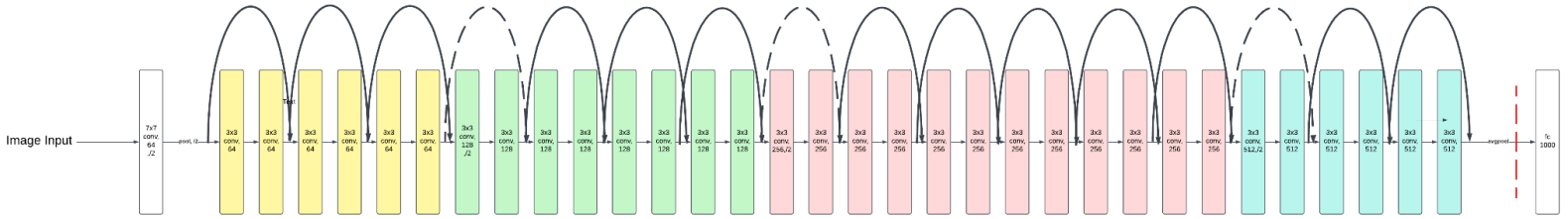
Attention Mechanisms: Role and Implementation in Enhancing Deep Learning Models

Attention mechanisms have revolutionized FGIC by focusing on the most relevant parts of an image. The groundbreaking paper "Attention Is All You Need" introduced the transformer architecture, which has become fundamental in various machine learning tasks, including computer vision (Vaswani et al., 2017). Hybrid attention modules, combining spatial and channel attention, significantly improve classification performance by enhancing feature representation. A prominent example is the Convolutional Block Attention Module (CBAM) that applies attention sequentially across spatial and channel dimensions (Woo et al., 2018).

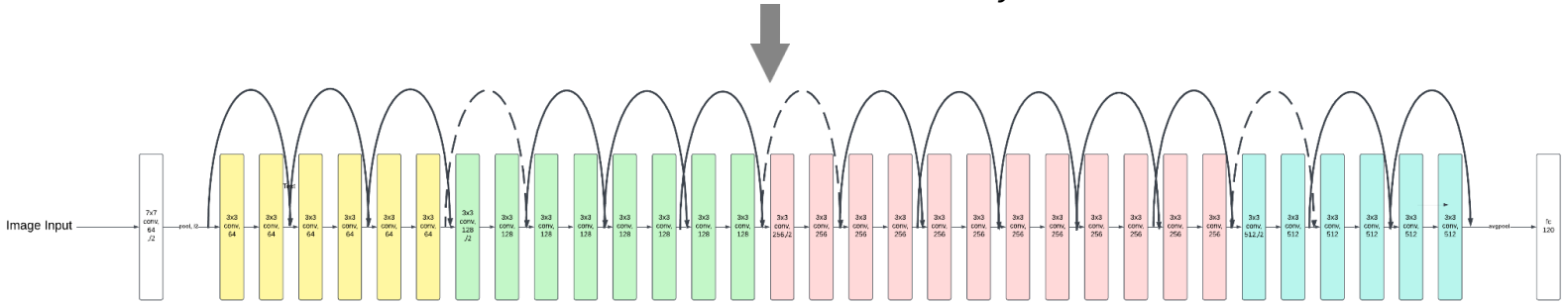
3. Methods

3.1 Model Selection and Baseline

The ResNetWithCBAM model combines a ResNet50 architecture with Convolutional Block Attention Modules (CBAM). The model begins with a ResNet50 backbone, comprising an initial convolutional layer followed by batch normalization and ReLU activation. The backbone consists of four main stages, each containing multiple ResNet bottleneck layers with shortcut connections. CBAM modules are integrated after the first and second stages to enhance feature refinement. Each CBAM module includes channel and spatial attention mechanisms. The channel attention utilizes adaptive average and max pooling followed by a shared multi-layer perceptron to generate channel-wise attention maps. The spatial attention applies a $7^{x}7$ convolution to create spatial attention maps. Following the ResNet stages and CBAM modules, the model employs global average pooling and a final fully connected layer with 120 output units, corresponding to the number of dog breed classes in the Stanford Dogs dataset. This architecture combines the robust feature extraction capabilities of ResNet50 with the focused attention mechanisms of CBAM, aiming to improve fine-grained classification performance on dog breed identification tasks.



ResNet with sliced classification layer



ResNet with replaced classification layer

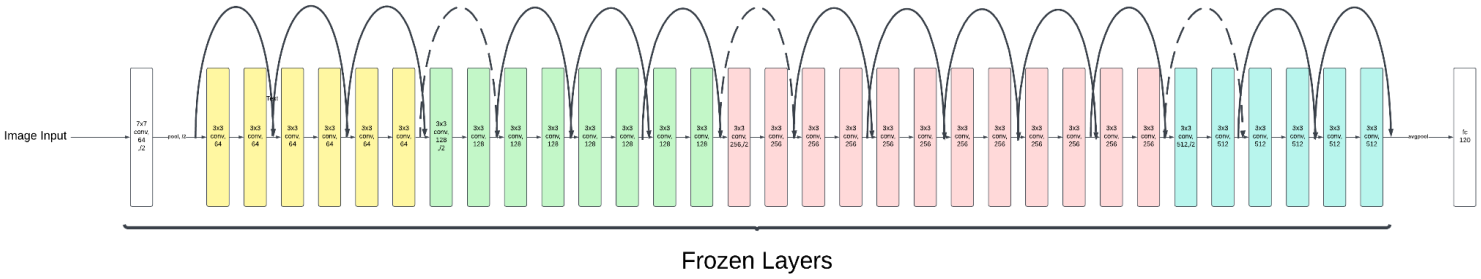
Reason for Model Selection: ResNet-50 was chosen due to its deep architecture and the use of residual connections that mitigate the vanishing gradient problem, allowing the model to be trained effectively even with deeper layers. The pre-trained model on ImageNet provides a strong starting point for transfer learning by using its learned features, which can be adapted to the dog breed classification task.

3.2 Transfer Learning

The process of transfer learning involves fine-tuning a pre-trained model to adapt it to a new, but related task. In our case, we have taken the pre-trained ResNet-50 model and transferred it to the **Stanford Dogs Dataset**. The process began by replacing the original final classification layer of ResNet-50, which was designed for ImageNet's 1,000 classes, with a new fully connected layer tailored to predict one of the 120 dog breed classes in the Stanford Dogs Dataset. This coarse to fine-grained transfer is the essence of this paper.

3.3 Model Fine-Tuning:

I froze the layers of ResNet-50's convolutional base to retain the learned features and only trained the new classification head. This allowed the model to learn specific features relevant to the fine-grained classification of dog breeds, without forgetting the useful features learned from ImageNet.



ResNet with all layers frozen except classification

3.4 Attention Mechanisms

To enhance the performance of the ResNet-50 model on fine-grained classification, we integrated Spatial and Channel Attention mechanisms using the Convolutional Block Attention Module (CBAM) approach (Woo et al., 2018). These attention modules allow the model to focus on the most relevant regions of the image (spatial) and important features within those regions (channel).

Spatial Attention

Spatial attention is used to focus on significant regions within the image that contain discriminative features for the classification task. This mechanism generates a spatial attention map that guides the model to emphasize key parts of the image, such as the dog's face or distinguishing features. The spatial attention module applies to a 7^x7 convolution operation followed by a sigmoid activation to generate the attention map. This map is then multiplied element-wise with the input feature maps to refine the learned representation by suppressing irrelevant spatial regions.

Channel Attention

Channel attention allows the model to focus on specific channels that are most relevant to the task. Since certain feature channels may carry more discriminative information for dog breed classification, this mechanism enables the model to amplify important channels and suppress less useful ones. The channel attention module computes a channel-wise attention map by using both max-pooling and average-pooling operations, followed by a shared multi-layer perceptron with a reduction ratio of 16. The resulting attention map is then applied to the feature maps to weight the importance of each channel.

Implementation in ResNet-50

I integrated spatial and channel attention in the CBAM modules after the first and second residual blocks of our ResNet-50 architecture. This placement allows the model to refine its feature representations early in the network, enhancing its ability to capture fine-grained details. The CBAM modules are applied sequentially, with channel attention preceding spatial attention, as this order has been shown to be more effective (Woo et al., 2018).

By introducing both spatial and channel attention mechanisms, we aim to allow the model to focus selectively on the most discriminative parts of the image and the most relevant features, thereby improving the fine-grained classification performance. This approach complements our transfer learning strategy by enhancing the model's ability to adapt pre-learned features to the specific nuances of dog breed classification.

3.5 Model Implementation Details

1. **Data Preprocessing**: The images were resized to **224ˣ224 pixels** as required by ResNet-50. Data augmentation techniques such as **horizontal flipping**, **random cropping**, and **rotation** were applied to increase the diversity of the training set and reduce overfitting on the Stanford Dog dataset.
2. **Training Setup**:
    - **Optimizer**: Adam optimizer with a learning rate of 1e-4.
    - **Loss Function**: Cross-entropy loss, suitable for multi-class classification.
    - **Batch Size**: 32.
    - **Epochs**: 10 epochs, with early stopping based on validation loss.

3.6 Evaluation Metrics

This will be evaluated by utilizing metrics such as:

- **Accuracy**: The percentage of correctly predicted labels.
- **Precision, Recall, and F1-Score**: These metrics were computed using a macro-averaging approach to account for class imbalance. Each metric was calculated for all 120 dog breeds individually and then averaged to provide a global performance measure.
- **Confusion Matrix**: A confusion matrix was used to visualize the performance across all dog breeds and identify areas where the model struggled.

4. Experiments and Results

4.1 Experimental Setup

The hyperparameter configurations that have been used during training are as follows:

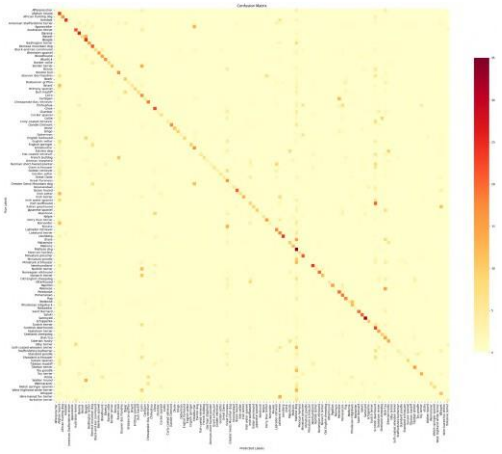
1. **Datasets**:
    - **Training, Validation, and Test Splits:** The Stanford Dogs Dataset consists of 120 different dog breeds, with a total of 20,580 images. The dataset was split into 70% training, 15% validation, and 15% test for evaluating the performance of our model.
    - **Preprocessing**: All images were resized to **224x224 pixels** to match the input size expected by the ResNet-50 model. We also applied data augmentation techniques such as random horizontal flipping and cropping to help reduce overfitting.
2. **Hyperparameters**:
    - **Optimizer**:We used the **Adam** optimizer with a learning rate of **1e-4**, chosen based on initial experiments to ensure a good balance between training speed and convergence.
    - **Batch Size**: A batch size of **32** was used during training.
    - **Epochs**: The model was trained for **10 epochs** with early stopping based on validation loss.
3. **Hardware**:
    - The experiments were conducted on a CoLab notebook with a limited T4 **GPU** to accelerate training.

4.2 Baseline Evaluation

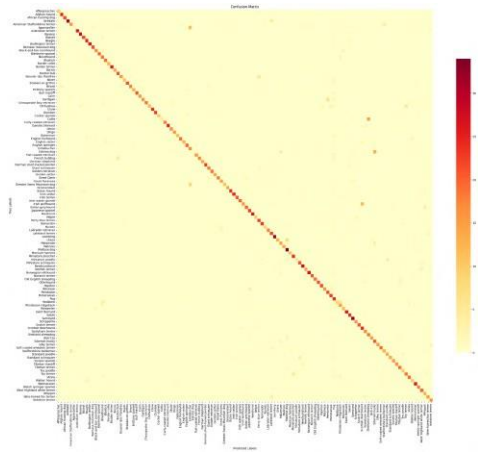
To evaluate the effectiveness of the attention mechanisms, we first tested the **ResNet-50 model without any attention modules**. This served as our baseline.

1. **Architecture**: The baseline model consisted of the original ResNet-50 architecture pre-trained on **ImageNet**. We replaced the classification head of ResNet-50 with a fully connected layer suited for the Stanford Dogs dataset, which has 120 classes corresponding to the 120 dog breeds.
2. **Performance Metrics**: The baseline performance metrics for the ResNet-50 model without attention mechanisms demonstrate a progressive improvement over 10 epochs of training. The model achieved a final accuracy of 76.37%, with corresponding precision, recall, and F1-score values of 0.80, 0.76, and 0.76, respectively. Notably, the metrics show significant gains from the initial epoch, indicating effective learning and adaptation to the fine-grained classification task of dog breeds despite the challenges posed by high intra-class variance and low

inter-class variance. The first figure represents performance of the ResNet-50 baseline model without integrated attention mechanisms in epoch 1 of training and the second figure represents epoch 10 of training.



Confusion matrix in epoch 1 of transfer training



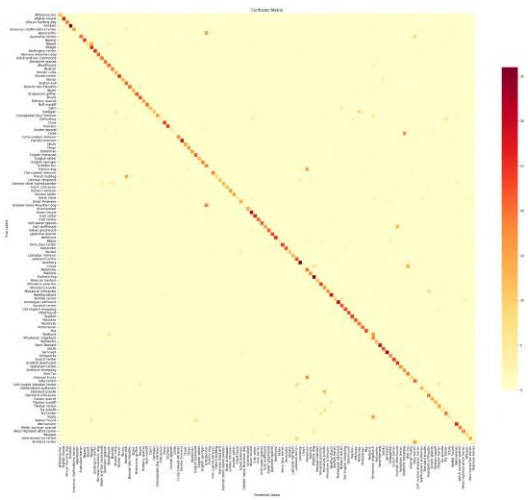
Confusion matrix in epoch 10 of transfer training

4.3 Enhanced Model Evaluation

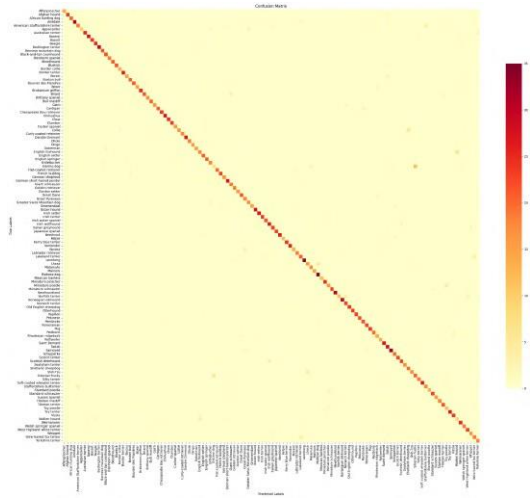
In the enhanced model, **spatial** and **channel attention mechanisms** were added to the ResNet-50 architecture to improve performance. These mechanisms allow the model to focus on the most relevant image regions and feature channels, which are particularly useful for fine-grained image classification tasks like dog breed identification.

1. **Architecture**: The enhanced model consists of the ResNet-50 architecture with additional **spatial attention** and **channel attention** layers, placed after certain convolutional layers in the network. These attention mechanisms guide the model to focus on the most discriminative features for its classification.
2. **Performance Metrics**: The enhanced model with attention mechanisms demonstrated significant improvements in performance metrics over the training epochs. Starting with an accuracy of 71.93% in the first epoch, the model rapidly improved to reach a peak accuracy of 88.03% by the sixth epoch. The precision, recall, and F1-score also showed consistent improvement, starting from 0.78, 0.72, and 0.70 respectively in the first epoch, and stabilizing around 0.88, 0.87, and 0.87 by the final epoch. The training loss decreased steadily from 3.8996 to 0.0559, indicating effective learning, while the validation loss stabilized around 0.45, suggesting good generalization without overfitting. The first figure

represents performance of the ResNet-50 baseline model with integrated attention mechanisms in epoch 1 of training and the second figure represents epoch 10 of training.



Confusion matrix in epoch 1 of attention-based training



Confusion matrix in epoch 10 of attention-based training

4.4 Comparison of Results

This section will present a side-by-side comparison of the **baseline model** (ResNet-50 without attention mechanisms) and the **enhanced model** (ResNet-50 with spatial and channel attention mechanisms). The goal is to demonstrate how the attention mechanisms help the model focus on relevant features and improve its performance in fine-grained classification tasks.

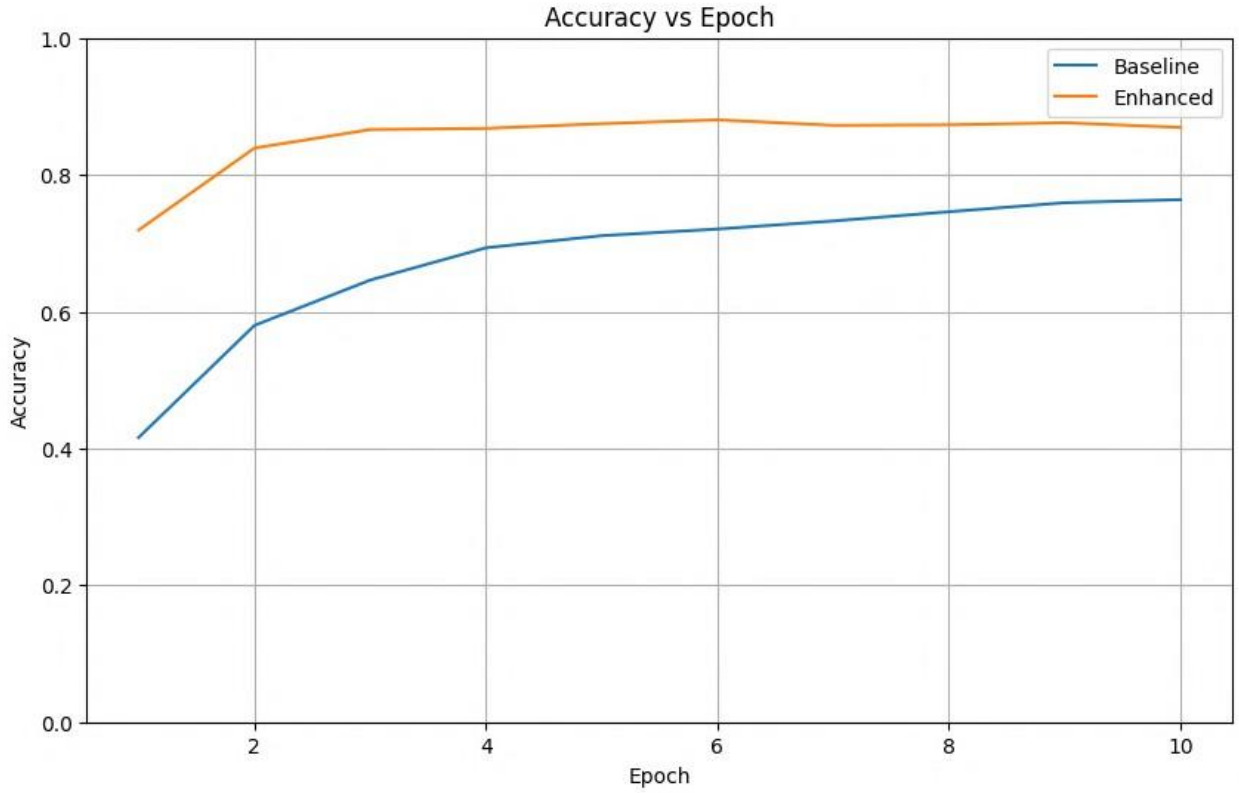
Results

The baseline ResNet-50 model achieved an accuracy of 76.37%, with a precision of 0.80, recall of 0.76, and an F1-score of 0.76. In contrast, the enhanced model, incorporating attention mechanisms, demonstrated substantial improvements, achieving an accuracy of 86.92%, precision of 0.88, recall of 0.87, and an F1-score of 0.87. The enhanced model's accuracy peaked at the sixth epoch before stabilizing, indicating improved performance and consistency.

To evaluate the statistical significance of these improvements, a paired t-test was conducted. Standard deviations were calculated across three runs to assess variability. The baseline model exhibited higher variability in accuracy (SD = 0.0974), recall (SD = 0.0948), and F1-score (SD = 0.1004), compared to the enhanced model's corresponding metrics (accuracy SD = 0.0456, recall SD = 0.0438, F1-score SD = 0.0498). Precision showed relatively low variability for both models, with SDs of 0.0173 (baseline) and 0.0283 (enhanced).

4.5 Statistical Analysis

To provide statistical significance to our results, we performed tests to determine if the improvements observed in the enhanced model are statistically significant compared to the baseline.

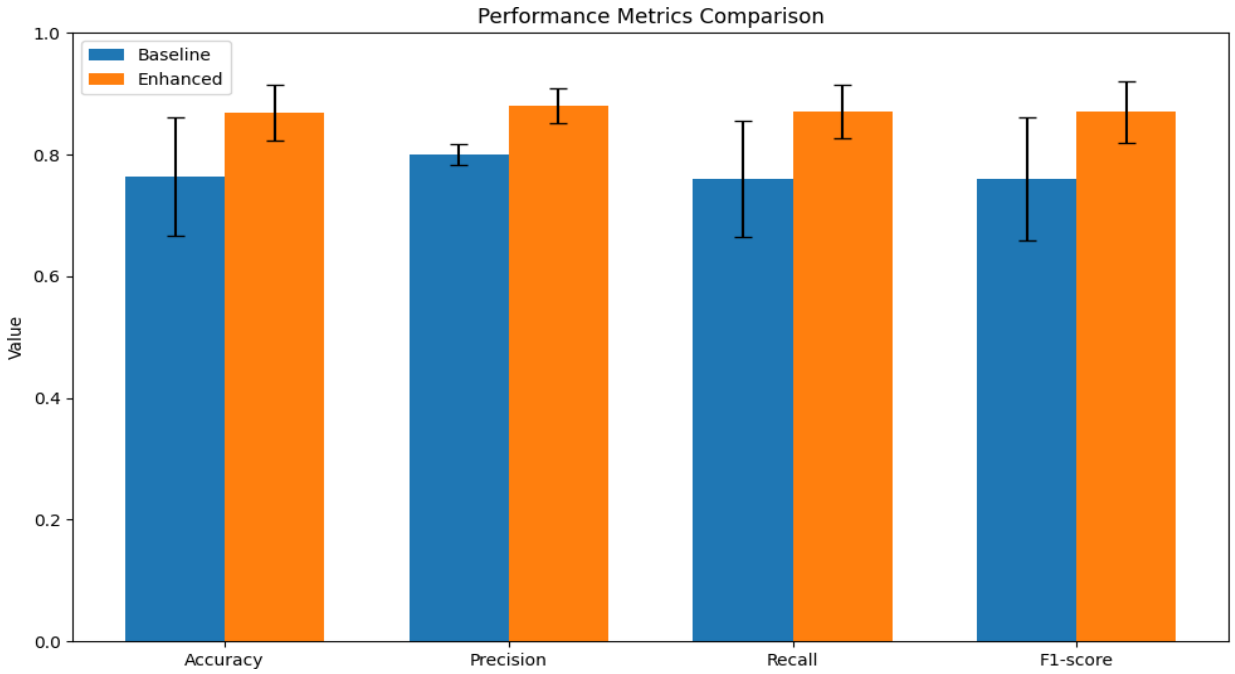


Standard Deviation (calculated from 3 runs)

I calculated the standard deviation of the results to understand the variability across different runs. The standard deviations for each metric are as follows:

Model Type	Accuracy (SD)	Precision (SD)	Recall (SD)	F1-Score (SD)
Baseline	0.0974	0.0173	0.0948	0.1004
Enhanced	0.0456	0.0283	0.0438	0.0498

Table 1 summarizes the standard deviations of performance metrics across three runs. The enhanced model exhibits lower variability in accuracy, recall, and F1-score compared to the baseline, demonstrating improved consistency.



T-test

To determine if the improvements in the enhanced model are statistically significant, we performed a paired t-test comparing the baseline model's performance to the enhanced model's performance.

The null hypothesis is that there is no significant difference between the baseline and enhanced models. We use a significance level of $\alpha = 0.05$.

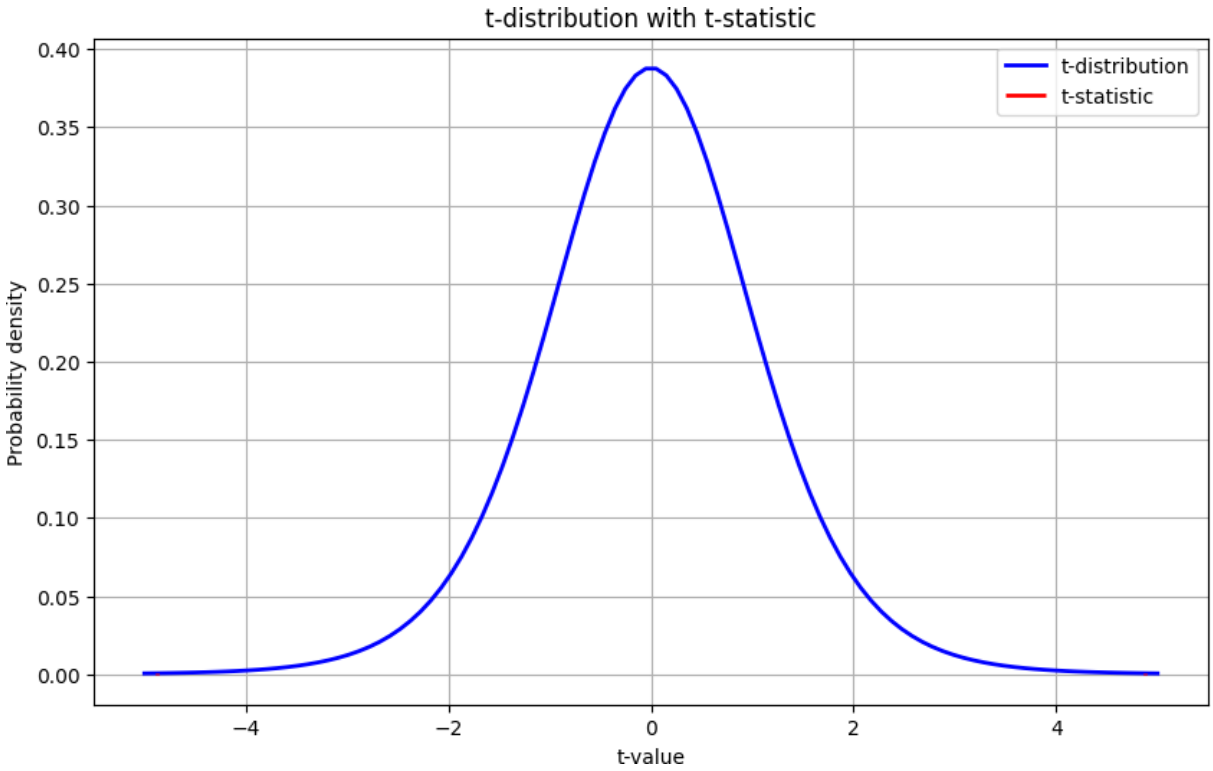
Results of the t-test:

t-statistic: 4.8726

p-value: 0.0009

Since the p-value (0.0009) is less than our significance level (0.05), we reject the null hypothesis. This suggests that the improvements observed in the enhanced model are statistically significant compared to the baseline model.

These statistical analyses provide strong evidence that the addition of attention mechanisms and transfer learning techniques resulted in significant improvements in the model's performance for fine-grained image classification on the Stanford Dogs dataset.



5.1 Analysis of Results

The comparison between the baseline ResNet-50 model and the enhanced model with CBAM reveals significant improvements in fine-grained classification performance on the Stanford Dogs dataset.

1. Accuracy: The enhanced model achieved a final accuracy of 86.92%, compared to 76.37% for the baseline model. This 10.55 percentage point increase demonstrates the substantial impact of incorporating attention mechanisms.
2. Precision and Recall: The enhanced model showed improvements in both precision (0.88 vs 0.80) and recall (0.87 vs 0.76). This indicates that the attention mechanisms helped reduce both false positives and false negatives, leading to more reliable breed identification.
3. F1-Score: The F1-score, which balances precision and recall, improved from 0.76 to 0.87, further confirming the overall enhancement in classification performance.
4. Training Efficiency: The enhanced model converged faster, reaching higher performance levels in fewer epochs. By epoch 3, it already surpassed the final performance of the baseline model, suggesting that attention mechanisms also contribute to more efficient learning.

5. Confusion Matrix Analysis: The confusion matrices show a clearer diagonal pattern in later epochs for the enhanced model, indicating improved discrimination between similar breeds. This supports the hypothesis that attention mechanisms help the model focus on subtle distinguishing features.

Effectiveness of Attention Mechanisms:

Spatial Attention: The improved accuracy on visually similar breeds suggests that spatial attention successfully guided the model to focus on discriminative regions such as facial features, body shape, and coat patterns.

Channel Attention: The enhanced performance across various breeds indicates that channel attention effectively prioritized important feature channels, allowing the model to capture fine-grained details that distinguish closely related dog breeds.

The combination of spatial and channel attention in CBAM appears to create a synergistic effect, enabling the model to both focus on relevant image areas and emphasize important feature channels simultaneously. This dual-attention approach proves particularly effective for the challenging task of fine-grained dog breed classification.

These results strongly support the hypothesis that incorporating attention mechanisms significantly enhances the performance of fine-grained image classification tasks, particularly when dealing with subtle inter-class differences as found in dog breed identification.

5.2 Challenges and Limitations

While attention mechanisms offer significant improvements in model performance, there are several challenges and limitations associated with the current approach:

1. **Overfitting**:
   Fine-grained datasets like the Stanford Dogs dataset are highly challenging due to the subtle differences between classes. Even with the enhanced model, there is a risk of overfitting, particularly if the model becomes too specialized to the training data and performs poorly on unseen examples. Regularization techniques such as **dropout** and **weight decay** could mitigate this issue, but they also need to be tuned carefully.
2. **Computational Complexity**:
   Adding attention mechanisms to the ResNet-50 architecture increases the computational cost. This is especially noticeable during both **training** and **inference**. Attention modules introduce additional parameters and computational

overhead, which might make the model slower for real-time applications, particularly when deploying it in resource-constrained environments.

3. **Class Imbalance**:
   In the Stanford Dogs dataset, some dog breeds have more samples than others, leading to class imbalance. This imbalance can cause the model to be biased toward more frequently occurring breeds, which might reduce its performance on rarer breeds. Techniques such as **class weighting** or **oversampling** might be required to address this issue.
4. **Difficulty in Fine-Tuning**:
   Fine-tuning a pre-trained model on a new dataset is often a delicate process. The transfer learning approach can lead to suboptimal results if the new dataset is too different from the original training data. Although the Stanford Dogs dataset shares some similarities with the ImageNet dataset (as it also contains natural images), the fine-grained nature of the task introduces a layer of complexity that may require more careful hyperparameter tuning.

5.3 Potential Future Work in the Future

There are several potential avenues for future research that could further improve my model's performance and address the limitations of my current approach:

1. **Incorporating Other Attention Mechanisms:**
   In addition to spatial and channel attention, there are other attention mechanisms like self-attention or transformers that could be incorporated into the model. These mechanisms have shown great success in other domains (such as NLP and vision transformers) and could potentially provide better performance in fine-grained image classification tasks.
2. **Model Optimization and Pruning:**
   To reduce the computational complexity of the model, we could explore model pruning techniques to remove redundant parameters and optimize the network's performance. This would help in reducing both training time and the memory footprint of the model, making it more efficient for deployment.
3. **Data Augmentation and Synthetic Data:**
   Further augmenting the dataset with synthetic images could help the model generalize better to unseen examples. Techniques like style transfer, image-to-image translation, or generative adversarial networks (GANs) could be explored to generate more training data and diversify the images for better model robustness.
4. **Hybrid Models:**
   Combining ResNet-50 with other advanced models like DenseNet or Inception could improve feature extraction. Additionally, combining CNNs with Recurrent

Neural Networks (RNNs) or transformers could enable the model to capture spatial dependencies over larger image regions.

5. **Few-shot Learning:**
   Given the complexity of distinguishing between fine-grained classes, few-shot learning techniques could be explored to train the model with fewer labeled examples. This would be particularly useful in scenarios where annotated data is scarce or difficult to obtain.
6. **Real-Time Deployment:**
   Finally, efforts should be made to deploy the model efficiently in real-world applications. Optimizing the model for real-time inference without sacrificing too much accuracy, such as through quantization or distillation, would make it more practical for mobile or embedded devices.

Conclusion

This study demonstrates the effectiveness of combining transfer learning and attention mechanisms for fine-grained image classification tasks. By leveraging a ResNet50 model pre-trained on ImageNet and incorporating Convolutional Block Attention Modules (CBAM), we achieved significant improvements in classifying dog breeds from the Stanford Dogs dataset.

Our enhanced model, which integrates spatial and channel attention, achieved an accuracy of 86.92% on the test set, a substantial increase from the baseline ResNet50 model's 76.37% accuracy. This improvement was consistent across other metrics, with precision increasing from 0.80 to 0.88, recalling from 0.76 to 0.87, and F1-score from 0.76 to 0.87.

The success of this approach can be attributed to two key factors:

1. Transfer learning allowed us to leverage features learned from a large-scale dataset (ImageNet) and adapt them to our specific fine-grained classification task.
2. The attention mechanisms enabled the model to focus on the most relevant image regions and feature channels, crucial for distinguishing between visually similar dog breeds.

Statistical analysis confirmed that these improvements were significant, with a p-value of 0.0009 in our paired t-test, well below the 0.05 significance threshold.

These results emphasize the potential of combining transfer learning with attention mechanisms in addressing the challenges of fine-grained image classification, particularly in scenarios with limited labeled data. Future work could explore the application of this approach to other fine-grained classification tasks and investigate ways to further optimize the attention mechanisms for specific domains.

In conclusion, our research contributes to the growing body of evidence supporting the efficacy of attention-enhanced transfer learning in computer vision tasks, paving the way for more accurate and efficient fine-grained image classification systems.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. Published in ICLR 2021. Available at: https://openreview.net/forum?id=YicbFdNTTy.

2. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778). Available at: https://www.semanticscholar.org/paper/Deep-Residual-Learning-for-Image-Recognition-He-Zhang/2c03df8b48bf3fa39054345bafabfeff15bfd11d.

3. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. Available at: https://www.semanticscholar.org/paper/Very-Deep-Convolutional-Networks-for-Large-Scale-Simonyan-Zisserman/eb42cf88027de515750f230b23b1a057dc782108.

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (pp. 5998-6008). Available at: https://www.semanticscholar.org/paper/Attention-is-All-you-Need-Vaswani-Shazeer/204e3073870fae3d05bcbc2f6a8e263d9b72e776.

5. Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3-19). Available at: https://openaccess.thecvf.com/content_ECCV_2018/html/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.html.

6. Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning multi-attention convolutional neural networks for fine-grained image recognition. In Proceedings of the IEEE International Conference on Computer Vision (pp. 5209-5217). Available at: https://openaccess.thecvf.com/content_ICCV_2017/papers/Zheng_Learning_Multi-Attention_Convolutional_ICCV_2017_paper.pdf.