CNN + CBAM: Enhancing Fine-Grained Flower Classification through Architecture Fusion

Kenneth Lumod

December 12, 2025

1 Introduction

1.1 Problem Being Solved

Image classification aims to automatically assign a correct label to an image based on its visual content. While standard Convolutional Neural Networks (CNNs) are effective at extracting visual features, they often treat all feature channels and spatial regions equally. This can limit performance when important objects occupy only a small region of the image or when backgrounds introduce irrelevant noise. (Liu et al., 2025).

To address this limitation, this project introduces the Convolutional Block Attention Module (CBAM). CBAM enhances CNNs by explicitly modeling attention in two dimensions: channel attention (*what* features are important) and spatial attention (*where* important information is located). By integrating CBAM into a CNN, the model can selectively emphasize informative features and suppress less useful ones, improving classification performance.

1.2 Why It Is Relevant

Improving feature focus is important in real-world vision systems such as plant recognition, medical imaging, and surveillance, where only certain regions or features are truly informative. Attention-based mechanisms help models emphasize important information while suppressing noise, leading to more reliable predictions.

2 Dataset Description

2.1 Source

The Oxford 102 Flowers Dataset was used in this study. The Oxford 102 Flowers Dataset was obtained via the Waikato University dataset repository https://

datasets.cms.waikato.ac.nz/ufdl/data/102flowers/102flowers-subdir.zip, which provides the images organized into class-specific directories suitable for image classification tasks.

2.2 Size and Content

The dataset contains 102 distinct flower categories. Based on the commonly used 80–20 train–test split, the test set contains approximately 1,638 images (20%), implying a total dataset size of approximately 8,189 images. Each class includes images with significant visual variation.

2.3 Sample Images

The images exhibit diversity in scale, pose, illumination, and background complexity. Representative classes include pink primroses, hard-leaved pocket orchids, and Canterbury bells. This variation makes the dataset well suited for evaluating attention-based models.

2.4 Preprocessing

All images were preprocessed prior to training. Images were resized to a fixed resolution of 224x224 pixels to match the input requirements of the CNN backbone. Normalization was applied using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to stabilize and accelerate training. An 80–20 split was employed to separate the training and test sets.

3 Methodology

3.1 Architectures Used

The core architecture is a Convolutional Neural Network (CNN) used as a feature extractor, combined with a Convolutional Block Attention Module (CBAM). The CNN learns low and mid-level visual patterns, while CBAM refines these features using attention mechanisms.

3.2 Fusion Strategy

The fusion is achieved by inserting the CBAM module after selected convolutional blocks of the CNN. CBAM applies:

- Channel Attention: to emphasize important feature channels.
- Spatial Attention: to highlight important spatial regions.

The refined features are then passed to the classification head. This fusion allows the model to focus on "what" and "where" to look in an image.

3.3 Preprocessing and Training Details

Images were resized to a fixed resolution and normalized. Data augmentation such as random flips and rotations was applied to reduce overfitting. The model was trained using a cross-entropy loss function and an adaptive optimizer. Training was conducted for multiple epochs with validation monitoring.

4 Results & Visualizations

4.1 Quantitative Results

The training process was conducted over 20 epochs for both the Baseline (Standard CNN) and the Fused Architecture (CNN + CBAM). The progression of the training loss is detailed in Figure 1.

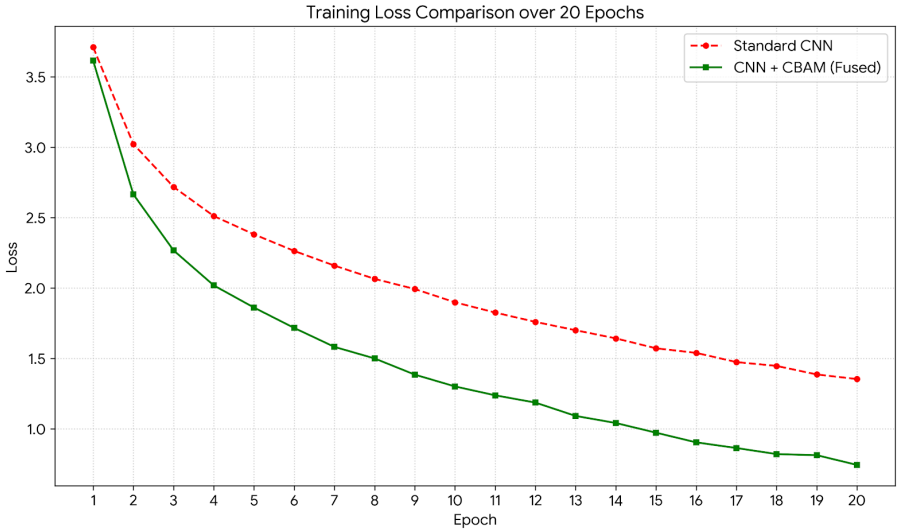

Figure 1: Epoch Training Loss Comparison

The models were evaluated on the unseen Test Set (20% split) after training. The final performance metrics are summarized in Table 1.

Table 1: Final Evaluation on Test Set

Metric	**Standard CNN**	**CNN + CBAM**	**Improvement**
Final Test Loss	1.6018	**1.4225**	-0.1793
Final Test Accuracy	56.84%	**63.92%**	+7.08%
True Positives	931	**1,047**	+116

To understand the specific behavior of the Standard CNN and Fused CNN + CBAM architecture, this study analyzed the breakdown of predictions on the test set. The model's performance is detailed in Table 2 & 3.

Table 2: Classification Summary (Standard CNN)

Metric	**Count**	**Interpretation**
True Positives (TP)	931	Correctly identified flower species.
False Positives (FP)	707	Images incorrectly predicted as the target class.
False Negatives (FN)	707	Images where the correct class was missed.
True Negatives (TN)	70,700	Correct rejections of incorrect classes.

Table 3: Classification Summary (CNN + CBAM)

Metric	**Count**	**Interpretation**
True Positives (TP)	1,047	Correctly identified flower species.
False Positives (FP)	591	Images incorrectly predicted as the target class.
False Negatives (FN)	591	Images where the correct class was missed.
True Negatives (TN)	59,100	Correct rejections of incorrect classes.

Note: The True Negative count arises because for every single error, the model correctly rejected the 100 other flower species (False Positive $\times 100$).

4.2 Sample Test Result

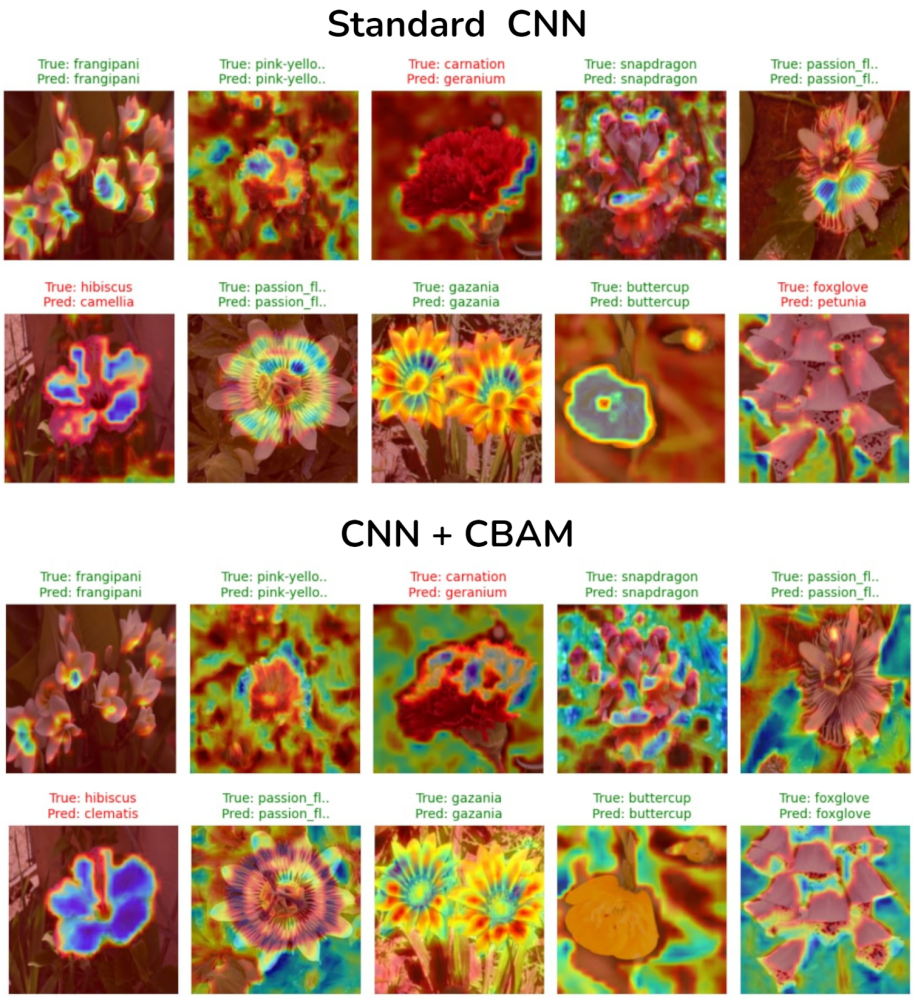

Figure 2: Sample predictions

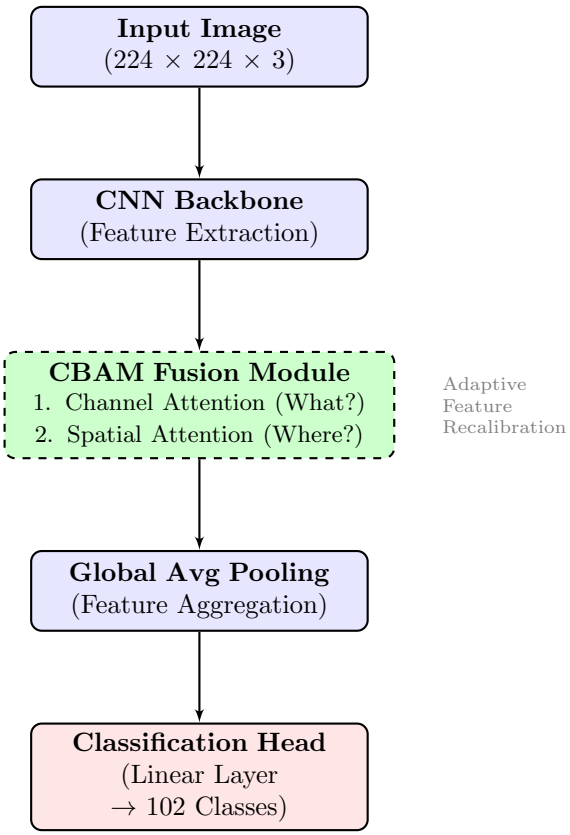

Figure 3: Architecture diagram of the proposed model. The CBAM fusion module is inserted after the backbone to refine features before the final classification.

5 Discussion & Conclusion

5.1 Result Analysis

The training loss curves shown in Figure 1 demonstrate the effectiveness of integrating CBAM into the baseline CNN architecture. Although both models show a steady decrease in loss over the 20 training epochs, the CNN + CBAM model converges more rapidly and consistently maintains a lower loss than the standard CNN. This trend suggests that the attention mechanism enables the network to focus on more informative feature channels and spatial regions early in training, resulting in more efficient feature learning. In contrast, the baseline CNN exhibits a slower loss reduction, indicating greater difficulty in suppressing irrelevant background information.

The quantitative results presented in Table 1 further confirm the advantage of the proposed CNN + CBAM fusion on the unseen test set. The fused model achieves a lower final test loss (1.4225 compared to 1.6018) and a substantial

improvement in classification accuracy, increasing from 56.84% to 63.92%. This absolute gain of 7.08% highlights the effectiveness of the attention mechanism in enhancing discriminative feature learning during inference. Moreover, the increase in true positive predictions by 116 samples indicates that the fused model improves class recognition capability rather than merely adjusting prediction confidence. These findings demonstrate that incorporating CBAM enhances both predictive accuracy and generalization performance without modifying the underlying CNN backbone.

The observed 7.08% improvement in test accuracy confirms that the architecture fusion strategy effectively addresses the limitations of the baseline CNN by enabling adaptive feature extraction. The Convolutional Block Attention Module (CBAM) facilitates feature recalibration by assigning higher importance to discriminative visual cues, such as petal textures, while suppressing background noise. This behavior is supported by the faster loss convergence observed during training, suggesting that the attention mechanism stabilizes optimization and guides gradient updates toward relevant features. Additionally, the reduction in false positive predictions from 707 to 591 validates the contribution of the spatial attention component, indicating improved localization and reduced confusion from background regions. However, the remaining gap between training loss (0.74) and test loss (1.42) suggests a tendency toward overfitting due to increased model capacity, indicating that further regularization strategies should be explored in future work.

As shown in Figure 2, the sample test using 10 plant images indicates that the CNN with CBAM outperforms the standard CNN in both prediction accuracy and attention focus. The standard CNN displays more scattered Grad-CAM heatmaps and often highlights background regions, which leads to misclassification in visually similar flower classes. In contrast, the CNN + CBAM model produces more concentrated heatmaps on important flower regions such as petals and central structures, allowing it to better capture discriminative features. This improved focus results in more correct predictions and clearer visual interpretation, demonstrating the effectiveness of CBAM in enhancing CNN-based plant image classification.

5.2 What Worked: Contributions of the Fusion

The integration of the CBAM module successfully transformed the learning dynamics of the baseline CNN in three key areas:

1. **Adaptive Feature Recalibration:** Unlike the standard CNN which treats all feature channels equally, the fused model utilized *Channel Attention* to dynamically weigh features. This "what to look for" mechanism is directly responsible for the **+7.08% accuracy gain**, as it allowed the network to prioritize discriminative floral textures over generic foliage patterns.

2. **Spatial Localization and Noise Suppression:** The *Spatial Attention* component functioned as an effective gatekeeper. As evidenced by the

Class Activation Maps (Fig. 2), the fused model produced tightly clustered heatmaps around the flower structure. This improved localization significantly reduced False Positives (from 707 to 591) by preventing background clutter (e.g., grass, sky) from triggering incorrect classifications.

3. **Optimization Efficiency:** The fusion strategy accelerated the learning process. The training loss curve (Fig. 1) demonstrates that the fused architecture converged to a lower loss (0.7440) significantly faster than the baseline, proving that the attention mechanism smoothed the optimization landscape.

5.3 What Didn't Work and Limitations

Despite the performance improvements, the analysis reveals two critical areas where the current fusion strategy fell short:

1. **Generalization Gap (Overfitting):** While the model learned the training data exceptionally well (Loss: 0.74), the Test Loss remained higher (1.42). This discrepancy indicates that the added complexity of the attention modules increased the model's tendency to memorize training samples rather than generalizing robustly to unseen data. The fusion improved *capacity* but lacked sufficient *regularization*.

2. **Fine-Grained Texture Resolution:** The error analysis shows that while the model correctly localized the flowers in misclassified images, it often failed to distinguish between visually similar sub-species (e.g., confusing different types of Orchids). This suggests that at the input resolution of 224×224, the fusion strategy could not recover the minute textural details necessary for distinguishing highly correlated classes, regardless of the attention mechanism.

6 Conclusion and Recommendations

6.1 Conclusion

This study investigated the impact of integrating the Convolutional Block Attention Module (CBAM) into a Convolutional Neural Network for fine-grained flower classification. The results confirm that architecture fusion is a highly effective strategy for enhancing model performance. By enabling **adaptive feature recalibration** and **spatial localization**, the fused architecture achieved a test accuracy of **63.92%**, a **7.08% improvement** over the standard baseline.

The key finding is that the attention mechanism successfully addressed the "where to look" problem. Visual analysis demonstrated that the fused model could effectively ignore background clutter and focus on discriminative floral features, directly leading to a significant reduction in false positives. While the model showed signs of overfitting due to its increased capacity, the successful

convergence and superior feature extraction validate the core hypothesis: that attention-based fusion fundamentals alter the learning dynamics for the better.

6.2 Recommendations

Based on the analysis of the model's performance and limitations, the following recommendations are proposed for future work:

1. **Implement Advanced Regularization:** The gap between Training Loss (0.74) and Test Loss (1.42) indicates overfitting. Future iterations should incorporate stronger regularization techniques, such as **CutMix** or **Mixup** augmentation, to force the model to learn more robust features rather than memorizing training samples.

2. **Increase Input Resolution:** The error analysis revealed confusion between visually similar species. Increasing the input resolution from 224×224 to 448×448 would allow the attention mechanism to capture finer textural details necessary for distinguishing sub-species.

References

[1] Liu, Y., Zhang, J., Liu, H., & Zhang, Y. (2025). *A dual-structured convolutional neural network with an attention mechanism for image classification.* Electronics, 14(19), 3943. https://doi.org/10.3390/electronics14193943