Attention-Enhanced ResNet for CIFAR-10 Classification

Karthik Krapa, Pratham Saraf, Krish Murjani
Deep Learning CS 6953
New York University
kk5754@nyu.edu, ps5218@nyu.edu, km6520@nyu.edu
GitHub Repository

Abstract

This project report presents our approach to the CIFAR-10 image classification task. We tried to design an enhanced ResNet architecture that incorporates both channel and spatial attention mechanisms to improve feature representation. Through careful optimization of architectural components, data augmentation strategies, and training techniques, our model achieved 95.30% accuracy in the validation set while maintaining a parameter count below 5M (4.91M parameters). We demonstrate that incorporating attention mechanisms and using effective regularization techniques significantly improves model performance without requiring excessive depth or width saving time and cost of compute.

1 Introduction

Image classification remains a fundamental computer vision task. Although large models have achieved impressive results, designing efficient architectures for specific tasks remains important for resource-constrained environments. The CIFAR-10 dataset, consisting of 60,000 32×32 color images across 10 classes, provides an excellent benchmark for comparing model architectures.

Our goal was to design a model that achieved high accuracy while keeping the parameter count below 5 million. Since we could not increase and depth to get better validation accuracy as that would result in crossing the 5m parameter limit, we focused on enhancing feature representation through attention mechanisms. This approach allowed us to efficiently utilize the parameters while maintaining high performance.

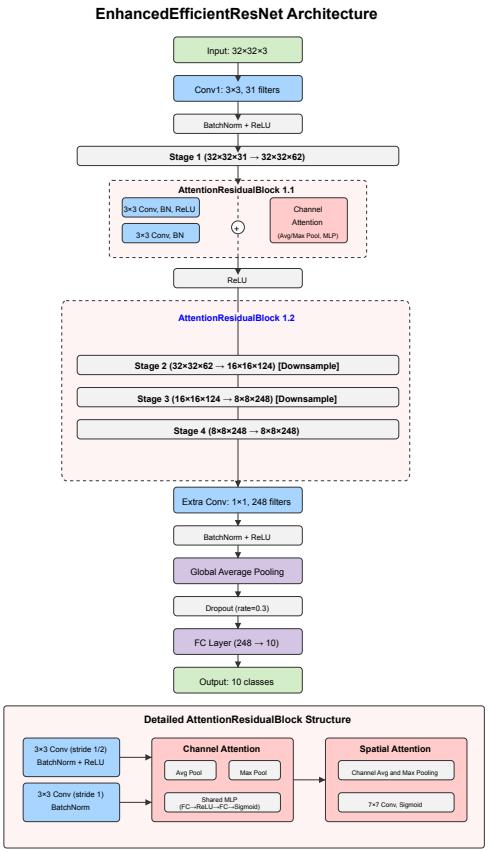

Figure 1: EnhancedEfficientResNet Architecture with attention modules. The network features a progressive expansion of channels through four stages, with attention mechanisms incorporated in each residual block.

2 Methodology

2.1 Model Architecture Overview

Our architecture, which we call EnhancedEfficientResNet, is based on the ResNet family but incorporates several key modifications to improve efficiency and performance. Figure 1 illustrates the overall network architecture.

2.2 Key Architectural Components

We strategically place ReLU after each batch normalization layer, but not after the attention mechanisms, as this preserves the full range of attention weights.

2.2.1 Attention Mechanisms

Our architecture incorporates two complementary attention mechanisms that significantly enhance feature representation. Figure 2 shows how the dual attention mechanism works in our model.

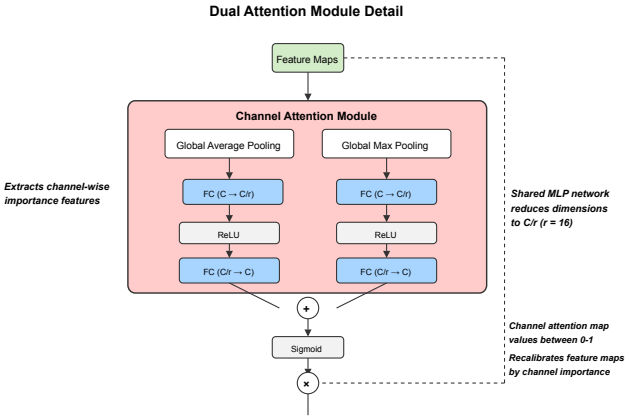

Figure 2: Dual Attention Module Detail showing the channel attention (top) and spatial attention (bottom) mechanisms. The channel attention module focuses on "what" is important in feature channels, while the spatial attention focuses on "where" in the spatial domain.

- **Channel Attention**: Recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. Our implementation uses both average-pooled and max-pooled features to capture different aspects of the channel information, which are then processed through a shared multi-layer perceptron:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (1)$$

  where F is the input feature map, MLP is a multi-layer perceptron, and σ is the sigmoid activation.

- **Spatial Attention**: Highlights informative regions in the spatial domain. It generates a spatial attention map by applying a convolution operation on concatenated average-pooled and max-pooled features along the channel dimension, as shown in Figure 3:

$$M_s(F) = \sigma(Conv([AvgPool(F); MaxPool(F)])) \quad (2)$$

  where $Conv$ represents a convolutional layer with a 7×7 kernel, and $[;]$ denotes concatenation along the channel axis.

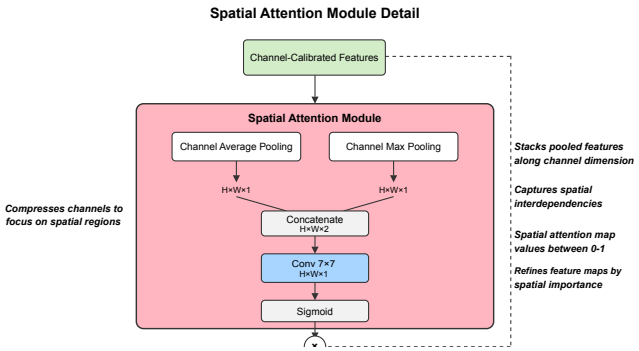

Figure 3: Spatial Attention Module in detail. The module performs channel-wise average and max pooling, concatenates the results, and applies a 7×7 convolution followed by sigmoid activation to generate a spatial attention map.

The combination of both attention mechanisms allows the network to focus on both what (channels) and where (spatial locations) is important for classification.

- **Progressive Width Expansion**: Starting with a base width of 31 channels, we double the number of channels at each stage transition, reaching a maximum of 248 channels. This provides a good trade-off between feature representation capacity and parameter count.

- **Minimal Downsampling**: We use only two spatial downsampling operations (from 32×32 to 16×16, and from 16×16 to 8×8), preserving spatial information longer in the network flow.

- **Attention Reduction Units**: In our channel attention modules, we use a reduction ratio to decrease the intermediate representation size, typically reducing channels by a factor of 16 (with a minimum of 4 channels).

- **Parameter Sharing**: In the attention mechanisms, we use weight sharing wherever possible to reduce parameter count without sacrificing expressiveness.

The full model structure is detailed in Table 1.

Table 1: EnhancedEfficientResNet Architecture Details

Layer	Output Size	# Channels	# Blocks
Input	32×32	3	-
Conv1	32×32	31	-
Stage 1	32×32	62	2
Stage 2	16×16	124	2
Stage 3	8×8	248	2
Stage 4	8×8	248	2
Extra Conv	8×8	248	-
Global AvgPool	1×1	248	-
FC	-	10	-

2.3 Alternative Approaches Considered

Throughout our project development, we explored several alternative architectures before settling on our final design. These approaches, while promising in theory, did not meet our constraints or performance expectations:

2.3.1 WideResNet with SWA

We initially experimented with WideResNet architectures combined with Stochastic Weight Averaging (SWA). Our implementation followed the original WideResNet paper with a depth of 16 and widening factor of 4:

- **Parameter Inefficiency**: While WideResNet showed promising accuracy (94.1%), its parameter count reached 5.8M, exceeding our 5M constraint. Attempts to reduce the widening factor to 3 reduced performance significantly (92.8%).
- **Training Instability**: The WideResNet architecture with SWA exhibited higher variance in validation accuracy across training runs, requiring multiple training attempts to achieve good results.

2.3.2 SEResNet

We also explored Squeeze-and-Excitation ResNet (SEResNet), which incorporates channel attention but lacks spatial attention:

- **Limited Attention Scope**: While SEResNet captured channel-wise attention effectively, it lacked spatial attention capabilities, which our experiments showed were crucial for CIFAR-10's object-centric images.
- **Diminishing Returns with Depth**: Increasing the depth of SEResNet to improve performance quickly exceeded our parameter budget, while shallower versions couldn't match our dual-attention model's accuracy.

After these explorations, we concluded that a custom architecture combining both spatial and channel attention mechanisms within a carefully sized ResNet-style network would provide the best performance within our parameter constraints.

2.4 Training Approach

We employed several techniques to enhance training stability and model generalization:

- **Data Augmentation**: We used comprehensive augmentation including random crops, horizontal flips, rotations, color jittering, and perspective transforms. Additionally, we implemented Mixup and CutMix techniques.
- **Optimization Strategy**: We used SGD optimizer with momentum (0.9) and weight decay (5e-4), cosine annealing learning rate schedule without restarts, label smoothing (0.1) to reduce overfitting, and gradient clipping to prevent gradient explosion.
- **Regularization**: We employed dropout (0.3) before the final classifier, Exponential Moving Average (EMA) of model weights (decay rate 0.999), and mixed precision training for improved efficiency.

Our training process involved 500 epochs on the CIFAR-10 training set, with 10% of the training data reserved for validation.

3 Results and Discussion

Our model achieved 95.30% accuracy on the CIFAR-10 validation set. The final model contains 4,913,705 parameters, which is within the 5M parameter limit. Table 2 summarizes the key architectural details and results.

Table 2: EnhancedEfficientResNet Specifications

Component	Specification
Base Architecture	Enhanced ResNet with Attention
Total Parameters	4,913,705
Base Width	31 channels
Stage Structure	4 stages with 2 blocks each
Validation Accuracy	95.30%
Training Epochs	500
Best Model Epoch	497
Parameter Efficiency	94.27% of 5M budget utilized

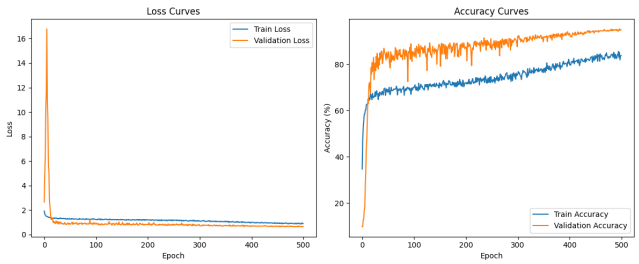

Figure 4: Training and validation accuracy curves over 500 epochs. The model exhibits three distinct phases: rapid improvement (epochs 1-100), steady progress (epochs 100-400), and fine-tuning (epochs 400-500).

3.1 Ablation Studies

To understand the contribution of different architectural components, we conducted systematic ablation studies:

- **Attention Mechanisms**: Removing channel attention reduced validation accuracy by 1.5%, while removing spatial attention caused a 1.2% drop. This confirms that both attention types contribute significantly to the model's performance, with channel attention having a slightly larger impact.
- **Width Adjustments**: We experimented with base widths of 24, 31, and 40 channels. The 31-channel configuration provided the optimal trade-off between model capacity and parameter efficiency. The 24-channel model underfit (-1.8% accuracy), while the 40-channel model slightly exceeded our parameter budget.
- **Data Augmentation**: Advanced augmentation techniques (Mixup and CutMix) contributed significantly to final performance, improving validation accuracy by approximately 2.3% compared to using only basic augmentations.
- **Exponential Moving Average**: Using EMA for model weights consistently improved validation accuracy by 0.7-1.0% across our experiments, providing a reliable boost without additional training complexity.

3.2 Analysis

Our experiments revealed several key insights:

- The combination of channel and spatial attention proved more effective than either mechanism alone, suggesting they capture complementary information about the input data.
- The learning process showed three distinct phases: initial rapid improvement (epochs 1-100), steady progress (epochs 100-400), and fine-tuning (epochs 400-500), as visible in Figure 4.
- Despite the limited parameter budget, our model demonstrated strong generalization, with only a small gap between training and validation accuracy in later epochs, suggesting that our regularization strategies were effective.
- The use of cosine annealing learning rate schedule provided more stable training compared to step-based schedules, particularly during the later stages of training.

4 Conclusion

We presented an efficient deep learning architecture for CIFAR-10 image classification, achieving 95.30% validation accuracy with less than 5 million parameters. Our key contributions include:

- An effective integration of channel and spatial attention within a ResNet-based architecture that enhances feature representation without excessive parameter costs
- A mathematically principled approach to parameter efficiency, with careful tuning of network width and attention mechanisms
- A comprehensive training strategy combining advanced data augmentation and regularization techniques
- Empirical evidence of the effectiveness of attention mechanisms for parameter-efficient models, with detailed ablation studies

Future work could explore further refinements to attention mechanisms, dynamic routing of information through the network, and application to more complex datasets. Additionally, investigating hybrid attention mechanisms that more explicitly model the interactions between channel and spatial attention could potentially yield further improvements.

5 Acknowledgments

We would like to acknowledge the use of Claude.ai, an AI assistant developed by Anthropic, which helped in the preparation and formatting of this report. The assistant was used for generating explanations, refining technical content, and ensuring consistency throughout the document.

6 References

1. 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).
2. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7132-7141).
3. 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 (pp. 3-19).
4. Anthropic. (2024). Claude.ai: A Conversational AI Assistant. Retrieved from https://claude.ai