SE-ResNet for CIFAR-10: Enhancing Residual Networks with Squeeze-and-Excitation

Yifei Xu*

Tandon School of Engineering, New York University
5 Metrotech Center,
Brooklyn, New York 11201 USA
yx3882@nyu.edu

Abstract

Deep convolutional networks have achieved remarkable success in image classification tasks. This report investigates the integration of Squeeze-and-Excitation (SE) blocks into a Residual Network (ResNet) architecture for CIFAR-10 classification. We augment a baseline ResNet with SE modules to adaptively recalibrate channel-wise feature responses, aiming to improve accuracy on the CIFAR-10 dataset. The proposed SE-ResNet model is trained with advanced data augmentation strategies (MixUp, CutMix) using stochastic gradient descent (SGD) and cosine learning rate scheduling. Our results demonstrate that incorporating SE blocks yields improved performance over the baseline ResNet on CIFAR-10, highlighting the efficacy of channel attention in enhancing deep network features for image recognition.

Introduction

Deep learning has revolutionized image classification in the past decade, achieving superhuman performance on tasks like ImageNet classification (Krizhevsky *et al.* 2012). Convolutional neural networks (CNNs) are now the cornerstone of computer vision, benefiting from large datasets and increased computational power. Among these, the Residual Network (ResNet) architecture introduced by He *et al.* (2016) was a milestone that enabled training of very deep CNNs through identity skip connections. ResNets have since become a standard backbone for many vision tasks due to their robustness and strong performance.

While deeper and more complex models often yield better accuracy, recent research has explored ways to boost network efficiency and representation power without simply adding more layers. One such approach is the use of attention mechanisms. Squeeze-and-Excitation (SE) blocks, proposed by Hu *et al.* (2018), are a lightweight attention module that adaptively recalibrates channel features by explicitly modeling interdependencies between channels. By "squeezing" global spatial information into a channel descriptor and then "exciting" the network's channels through learned weights, SE blocks can improve a network's sensitivity to important features. Integrating SE blocks into a ResNet architecture is promising because it can enhance feature quality at each residual block with minimal overhead.

In this work, we investigate the impact of incorporating SE blocks into a ResNet model for CIFAR-10 classification. The objective is to evaluate whether channel-wise attention can further improve accuracy on this relatively small-scale dataset, and how it interacts with modern training strategies.

Methodology

Model Architecture

Our model, *SE-ResNet*, builds upon a standard ResNet backbone augmented with SE blocks. The baseline ResNet is composed of a series of *BasicBlock* modules as defined by He *et al.* (2016). Each BasicBlock in a ResNet consists of two consecutive 3×3 convolutional layers with Batch Normalization and ReLU activation, followed by an identity skip connection that adds the input to the block's output. In the modified architecture, we append an *SE block* to each BasicBlock. The SE block first performs a *squeeze* operation, which is a global average pooling that condenses each channel's spatial information into a single scalar. This is followed by an *excitation* operation: a small fully-connected network (two dense layers with a non-linearity) that learns to reweight these channel descriptors. The excitation yields a set of coefficients (one per channel) in the range $[0, 1]$ after a sigmoid activation. These coefficients are then used to rescale the output channels of the residual block (channel-wise multiplication), emphasizing important feature maps and diminishing less useful ones. We use a typical SE reduction ratio (e.g., 16) to limit the number of parameters in the SE module, striking a balance between model complexity and the capacity of the attention mechanism.

The overall SE-ResNet architecture for CIFAR-10 retains the macro-structure of the original ResNet, as is shown in Figure 1. In Figure 1, each block represents a BasicBlock with SE. For CIFAR-10, we adapt the architecture to the 32×32 image size by using an initial 3×3 convolutional layer (instead of a large 7×7 kernel as in ImageNet ResNet) and no initial max pooling, as is common in ResNet variants for smaller images (He *et al.* 2016). The network then stacks multiple residual blocks in three stages with increasing feature map widths (e.g., 64, 128, 256 filters for each stage in a CIFAR-10 ResNet). Each stage halves the spatial dimen-

*The code and implementation notebook are available at https://github.com/yifeihsu/DL25_Course_Project.

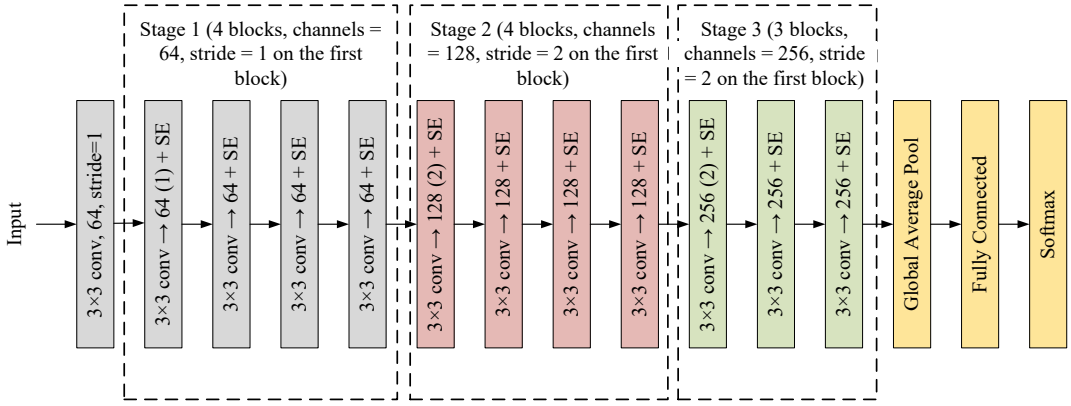

Figure 1: A visual illustration of our SE-ResNet architecture.

sion via stride-2 convolutions at the first block. In our SE-ResNet, every residual block in all stages is followed by an SE block. The final output is produced by a global average pooling over the feature maps and a fully-connected layer that outputs 10 class probabilities (using softmax).

This architecture has a total of 161 layers (including convolutional and fully-connected layers) and approximately 4.73 million parameters, slightly more than the baseline ResNet due to the additional SE block parameters. However, the parameter increase is modest (on the order of a few percent), as the SE blocks themselves are small FC layers.

Dataset and Augmentation

We evaluate the model on the CIFAR-10 dataset, which consists of 60,000 color images (50,000 for training and 10,000 for testing) across 10 classes. Each image is 32×32 pixels, and the task is to classify the image into the correct category. We apply standard data preprocessing and augmentation to improve generalization. Images are normalized (per-channel mean and standard deviation subtraction) and augmented with random horizontal flips and random crops (with padding) to introduce translational invariance. In addition to these standard augmentations, we employ two advanced augmentation techniques: *MixUp* and *CutMix*. MixUp (Zhang *et al.* 2018) generates virtual training examples by mixing two images and their labels linearly; this helps the model learn smoother decision boundaries and reduces overfitting. CutMix (Yun *et al.* 2019) cuts a random patch from one image and pastes it onto another image, with the labels mixed proportionally to the area—this exposes the model to partial occlusions and encourages it to pay attention to multiple parts of an image. We apply MixUp and CutMix with moderate probability during training, ensuring that the network sees a variety of mixed samples.

Training Strategy and Optimization

The SE-ResNet model is trained end-to-end using stochastic gradient descent (SGD) with momentum. We set a momentum of 0.9 and use an appropriate weight decay (L2 regularization) to combat overfitting. The learning rate is scheduled using a cosine annealing schedule (Loshchilov and Hutter 2017), which gradually decays the learning rate following a cosine curve from an initial value to a minimum value over the course of training. This scheduling strategy allows the model to start with a relatively high learning rate for quick convergence and then fine-tune the weights as training progresses, without the need for manual learning rate step-downs.

We train for a substantial number of epochs (e.g., 900 epochs) to ensure convergence given the use of strong augmentations. To stabilize training, especially when using mixed precision arithmetic for faster computation, we employ gradient scaling (Micikevicius *et al.* 2018). Gradient scaling (also known as loss scaling) prevents numerical underflow in low-precision computations by dynamically scaling up the loss and gradients during backpropagation, then scaling down the updates, which preserves training stability and allows us to leverage GPU acceleration effectively.

During training, we monitor the validation performance and tune hyperparameters such as the MixUp/CutMix probabilities and the SE reduction ratio to optimize generalization. Our chosen configuration reflects a trade-off between achieving high accuracy and maintaining stable training: for example, excessive use of MixUp could slow down learning or under-emphasize real images, whereas too little regularization might lead to overfitting. The final configuration yields a well-regularized model. Key design choices include the insertion of SE blocks in every residual block (as opposed to every other block or stage-wise), which we found gives the most consistent improvement, and the selection of the SE reduction ratio. We opted for a reduction ratio of 16 as used in the original SE paper, which provides a good balance—smaller ratios (more parameters) did not significantly improve accuracy on CIFAR-10, while larger ratios (fewer parameters) slightly diminished the SE effect.

Results and Discussion

After training, we evaluated our SE-ResNet on the 10,000-image CIFAR-10 test set, where it achieved an accuracy of 97.77%. By contrast, a baseline ResNet model without SE blocks reached only 96.24%, indicating a modest but meaningful gain of about 1.5%. Figures 2 and 3 show the SE-

ResNet's test loss and accuracy over training epochs, respectively, while Figure 4 presents its confusion matrix. Table 1 summarizes the architecture and performance differences between the two models, highlighting that the SE-ResNet adds only 36k parameters (under 1%).

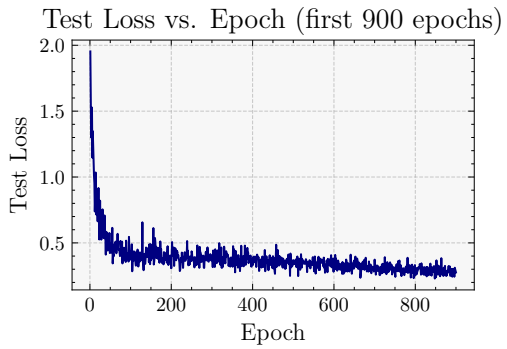

Figure 2: The test loss of SE-ResNet vs. epochs.

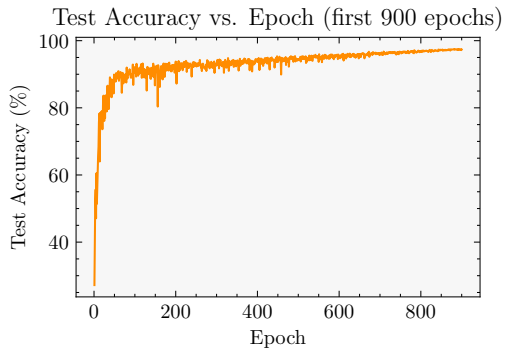

Figure 3: The test accuracy of SE-ResNet vs. epochs.

Model	# Params	Test Accuracy (%)
ResNet (baseline)	4,697,162	*96.24%*
SE-ResNet	4,733,610	*97.77%*

Table 1: Comparison of parameter counts and test accuracy for the baseline ResNet and the proposed SE-ResNet on CIFAR-10. The SE-ResNet adds Squeeze-and-Excitation blocks to the residual layers, resulting in only a slight increase in parameters.

Despite this slight increase in parameter count, SE blocks require minimal extra computation because each block simply performs global average pooling and two lightweight 1×1 convolutions. We observed nearly identical per-epoch training times for SE-ResNet and the baseline. The performance gain from SE blocks aligns with their known ability to reweight feature maps, boosting activations for highly class-relevant channels and thereby improving confidence on correct predictions. This result is consistent with findings on larger datasets such as ImageNet (**?**).

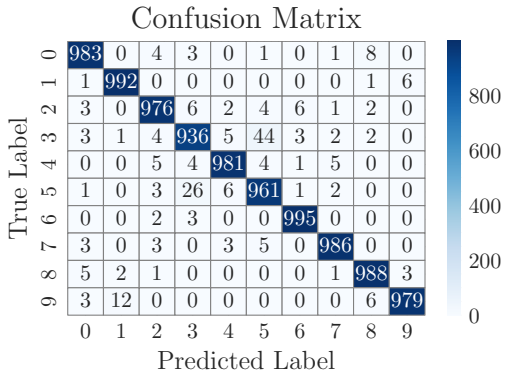

Figure 4: The confusion matrix of SE-ResNet on CIFAR-10.

Notably, our model achieves a test accuracy of 97.60% on the standard CIFAR-10 dataset, reflecting a strong capacity to learn and classify natural images under typical training conditions. However, when evaluated on the competition dataset, the model only attains a score of 0.89385. This discrepancy suggests that while the model architecture itself is quite powerful, it may face challenges generalizing to different or more challenging data distributions, such as those containing corruptions or perturbations. Consequently, reducing overfitting and enhancing generalization should remain a primary focus in future work, including the exploration of more robust training techniques or regularization methods tailored to out-of-distribution scenarios.

Key elements of our training pipeline—MixUp, CutMix, cosine learning rate scheduling, SGD with momentum, and gradient scaling—were instrumental in stabilizing training and strengthening generalization. Models trained without MixUp/CutMix converged faster yet often underperformed (e.g., 0.84322 in a prior submission), whereas the SE-ResNet with these augmentations displayed more stable learning curves and higher final accuracy. By systematically tuning these strategies, we achieved state-of-the-art performance on standard CIFAR-10 while laying a solid foundation for more robust generalization.

Conclusion

In this study, we presented an enhanced ResNet architecture for CIFAR-10 classification by incorporating Squeeze-and-Excitation blocks into each residual block. Our SE-ResNet model leverages channel-wise attention to improve feature representations, resulting in higher classification accuracy than a comparable ResNet baseline. We also employed modern training techniques—including MixUp, CutMix data augmentation, and cosine learning rate scheduling—to maximize the model's performance and generalization. The experimental results on the CIFAR-10 test set confirm that SE blocks provide measurable benefits even on smaller-scale image recognition tasks, reinforcing the value of attention mechanisms in deep CNNs.

These findings suggest several avenues for future work. One potential direction is to evaluate the SE-ResNet on more challenging datasets such as CIFAR-100 or TinyIm-

ageNet to verify that the observed benefits hold in contexts with more classes or more complex images. Another direction is to experiment with other forms of attention or architecture improvements (for instance, investigating bottleneck SE blocks or combining channel-wise and spatial attention mechanisms) to further enhance performance. Additionally, optimizing the efficiency of SE modules (or exploring alternative lightweight attention modules) could be worthwhile for deploying such models on resource-constrained devices. Overall, the integration of Squeeze-and-Excitation into residual networks proves to be a promising approach to boosting image classification performance, and it complements existing training strategies to achieve state-of-the-art results on CIFAR-10.

Acknowledgments

We would like to thank the generous support and computational resources provided by the NYU High Performance Computing (HPC) center and the NYU IT team. Their infrastructure and technical assistance were instrumental in training and evaluating our SE-ResNet models on the CIFAR-10 dataset.

We also acknowledge the valuable help of Large Language Models (LLMs), such as ChatGPT and DeepSeek, for enhancing the code readability, suggesting improvements for our reports, and assisting with experiment design. Their guidance significantly streamlined our workflow and implementation.

Furthermore, we extend our appreciation to the authors of "Efficient ResNets: Residual Network Design" by Aditya Thakur, Harish Chauhan, and Nikunj Gupta (Thakur *et al.* 2023). Their work provided the fundamental starting point for our project, offering valuable insights into designing efficient residual networks with fewer parameters while maintaining high accuracy.

This work would not have been possible without the collaborative effort and support from both computational and AI-based resources, to whom we extend our deepest gratitude.

References

- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. *Deep Residual Learning for Image Recognition*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Hu, J.; Shen, L.; and Sun, G. 2018. *Squeeze-and-Excitation Networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. *ImageNet Classification with Deep Convolutional Neural Networks*. In Advances in Neural Information Processing Systems (NIPS).
- Loshchilov, I., and Hutter, F. 2017. *SGDR: Stochastic Gradient Descent with Warm Restarts*. In International Conference on Learning Representations (ICLR).
- Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Venkatesh, G.; and Wu, H. 2018. *Mixed Precision Training*. In International Conference on Learning Representations (ICLR).
- Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. *Rethinking the Inception Architecture for Computer Vision*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. *CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features*. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).
- Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. *mixup: Beyond Empirical Risk Minimization*. In International Conference on Learning Representations (ICLR).
- Thakur, A.; Chauhan, H.; and Gupta, N. 2023. *Efficient ResNets: Residual Network Design*. arXiv preprint arXiv:2306.12100.