Data Augmentation and Its Impact on Model Performance on CIFAR-10

Mohamad Nabih Alkhateeb
Department of Computer Science
University of Texas at Arlington
Arlington, TX, USA
mxa0786@mavs.uta.edu

***Abstract*—This paper investigates the effects of various data augmentation strategies on convolutional neural network (CNN) performance using the CIFAR-10 image classification dataset. We explore how basic transformations such as cropping, flipping, rotation, and color jittering impact accuracy and robustness, offering insights for enhancing models trained on small to mid-sized datasets.**

***Index Terms*—data augmentation, CIFAR-10, convolutional neural networks, image classification, deep learning**

I. Introduction

In the realm of deep learning, particularly for image classification tasks, data augmentation has proven to be a powerful strategy to improve model generalization. This project focuses on analyzing the impact of various augmentation techniques on the performance of convolutional neural networks (CNNs) trained on the CIFAR-10 dataset.

CIFAR-10 consists of 60,000 32x32 color images across 10 object classes and is widely used to benchmark machine learning algorithms. Due to its relatively small size, models trained on it are susceptible to overfitting, making it ideal for studying data augmentation methods.

Our objective is to apply a set of image transformations—such as random rotation, cropping, flipping, and color jittering—to the training data and evaluate their effects on model performance metrics like accuracy and robustness. By comparing baseline models with augmented ones, we aim to identify augmentation strategies that yield the highest performance gains. This will help build best practices for using augmentations in small to mid-sized image datasets.

Beyond academic interest, the broader implication of this work lies in its real-world applicability. Improving the robustness and generalization of image classification models has direct impacts on critical fields such as autonomous driving, healthcare diagnostics, and security systems, where high-accuracy visual recognition is essential. Data augmentation techniques make it possible to train more resilient models even with limited datasets, thereby reducing the need for extensive data collection and labeling. This can accelerate innovation in industries that rely heavily on computer vision and help make AI technologies more accessible and reliable across a wide range of societal applications.

II. Literature Review

Shorten and Khoshgoftaar (2019) conducted a comparative analysis of augmentation techniques, emphasizing how geometric transformations like warping, rotation, and elastic distortions impact model performance across various datasets. The study finds that augmentation in data-space, particularly using elastic distortions model, consistently improves classification performance and reduces overfitting more effectively than feature-space techniques like SMOTE and DBSMOTE. Their results suggest that overly simplistic augmentations may fail to add meaningful diversity to training data, while more complex distortions can significantly enhance generalization. This insight influenced our choice to include more nuanced transformations such as random affine changes, rather than relying solely on traditional flipping or cropping.

Cubuk et al. (2019) proposed AutoAugment, a method that uses reinforcement learning to discover augmentation policies that optimize validation accuracy. This method employed a controller RNN that explores a discrete search space of augmentation operations with defined types, probabilities, and magnitudes for each to optimize validation accuracy on a given dataset. AutoAugment's success on CIFAR-10 and other datasets demonstrated that well-designed augmentation policies can rival even architectural changes in boosting performance. While we do not replicate AutoAugment's search process, their results serve as an upper-bound reference for evaluating the effectiveness of our handcrafted augmentations. They also highlighted the importance of combining multiple augmentation operations rather than applying them in isolation.

In their 2021 survey, Shorten and Khoshgoftaar explored Image Data Augmentation (IDA) as a crucial technique to improve deep learning models in computer vision, particularly when training data is limited. IDA helps mitigate overfitting by artificially expanding datasets through transformations or synthetic image generation. The researchers also presented a detailed taxonomy of augmentation techniques, classifying them into geometric, photometric, learned, and adversarial categories. This broader perspective helped contextualize where our work fits within the field. Specifically, we are currently focused on geometric and photometric augmentations due

to their computational efficiency, but may expand to more advanced techniques like GAN-based or learned augmentations in future work. Their discussion also cautions about diminishing returns when stacking too many augmentations, which we considered when designing our experimental setup.

Overall, these sources provide a theoretical and practical foundation for our investigation. They guided our selection of transformations, informed the structure of our experiments, and clarified the trade-offs between augmentation complexity and computational cost. Importantly, they emphasized that augmentation should be tailored to the dataset and task, which aligns with our goal of systematically evaluating their impact on CIFAR-10 classification.

III. Implementation

The implementation of this project was developed in Python using the PyTorch framework. Supplementary libraries included torchvision for dataset handling and augmentations, NumPy for data operations, Matplotlib and Seaborn for visualization, and scikit-learn for evaluation metrics. These tools were chosen due to their popularity, robustness, and strong community support within the machine learning and deep learning ecosystems.

We worked with the CIFAR-10 dataset, which consists of 60,000 color images of size 32×32 across 10 distinct classes. The dataset is split into 50,000 training images and 10,000 test images. Each image contains 3 RGB channels, and the model takes input in the format $3 \times 32 \times 32$. The 10 classes represent common object categories, making CIFAR-10 a well-balanced and widely adopted benchmark for image classification.

To prepare the data, we normalized pixel values to fall within the range $[-1, 1]$ using the mean and standard deviation of the CIFAR-10 dataset. Label values were kept as integers from 0 to 9. This normalization ensures consistent model behavior across different input scales and accelerates convergence during training.

Data augmentation was applied only to the training dataset. The transformations used were:

- **RandomHorizontalFlip:** with probability 0.5
- **RandomCrop:** with padding of 4 pixels
- **RandomRotation:** limited to 15 degrees
- **ColorJitter:** applied with slight variations in brightness and contrast

These augmentations were selected based on existing literature and are known to improve generalization by introducing spatial and photometric diversity in the training data.

The convolutional neural network model consists of three convolutional blocks followed by ReLU activations and max-pooling layers. The flattened output is fed into two fully connected layers with dropout regularization and a final Softmax layer for classification. This architecture is intentionally lightweight and interpretable, allowing for fast iteration while being expressive enough to learn from the CIFAR-10 dataset. The model was optimized using the Adam optimizer with a learning rate of 0.001, and trained using CrossEntropyLoss as the objective function. Batch size was set to 64 and the model was trained for 10 epochs.

Training progress was tracked by recording accuracy and loss over each epoch. As shown in Figures 1 and 2, the baseline model reached 91.09% training accuracy by epoch 10, while the augmented model plateaued at 68.88%. The loss plot reveals that although the baseline model minimizes training loss faster, the augmented model maintains a steadier and more gradual descent, suggesting stronger generalization.

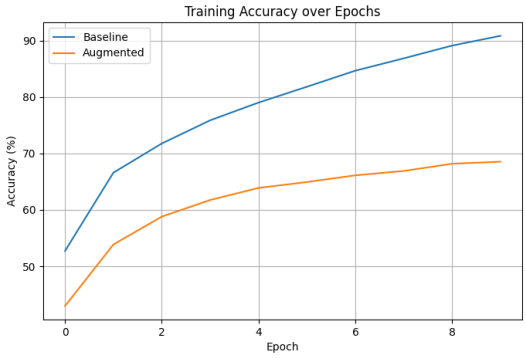

Fig. 1. Training accuracy over 10 epochs (baseline vs. augmented).

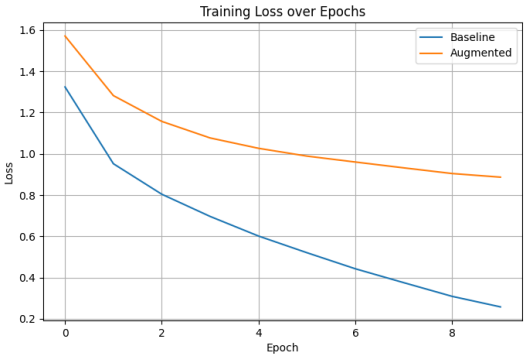

Fig. 2. Training loss over 10 epochs (baseline vs. augmented).

After training, the models were evaluated on the CIFAR-10 test set using classification accuracy, confusion matrices, classification reports, and per-class performance metrics. The baseline model achieved a final test accuracy of 70.93%, while the augmented model achieved 72.64%, confirming a modest but consistent improvement from data augmentation.

Figures 3 and 4 show the confusion matrices, which visualize model errors and accuracy across each class. The augmented model exhibits reduced confusion between visually similar classes, reflecting the positive influence of diverse training data.

The classification report includes precision, recall, and F1-score for each class, helping evaluate whether a model is biased toward specific categories. The augmented model shows balanced improvements across several classes.

Figure 5 presents a bar chart comparing final test accuracy of both models. This offers a quick visual summary of overall performance improvements resulting from augmentation.

Figure 6 shows a per-class test accuracy comparison, revealing which classes benefited the most from augmentation. Classes like ”cat” and ”ship” saw notable gains, possibly due to the added variability aiding the network in learning more generalized features.

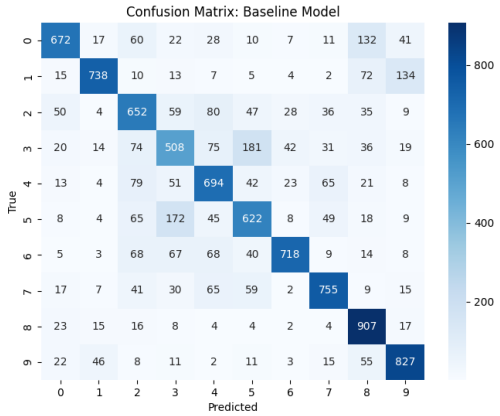

Fig. 3. Confusion matrix for baseline model.

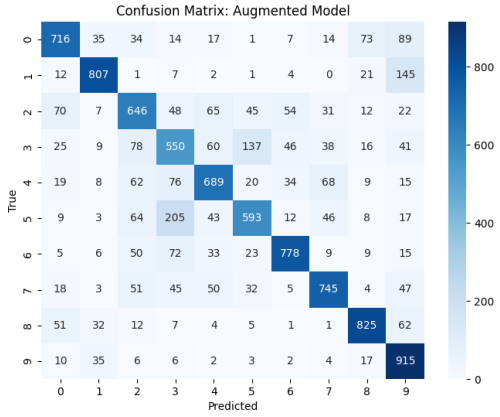

Fig. 4. Confusion matrix for augmented model.

IV. Limitations

While our findings demonstrate the benefits of data augmentation on model generalization, there are several limitations to consider. First, we only experimented with a single CNN architecture, which may not fully capture how augmentations interact with more complex or deeper models. Second, the augmentations applied were limited to common geometric and photometric transformations; we did not explore learned or adversarial augmentation techniques that could further enhance performance. Lastly, CIFAR-10, while a popular benchmark, is a relatively small and clean dataset compared to real-world

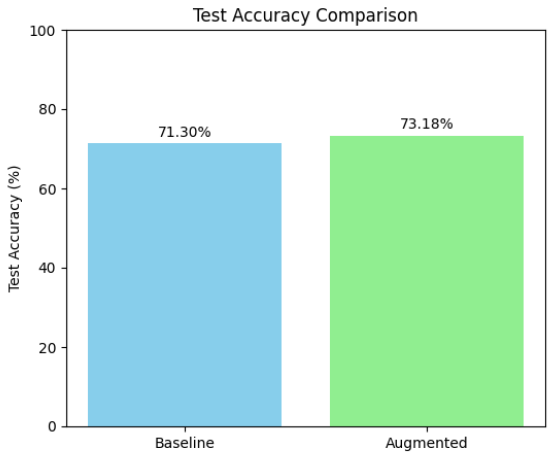

Fig. 5. Test accuracy comparison: baseline vs. augmented.

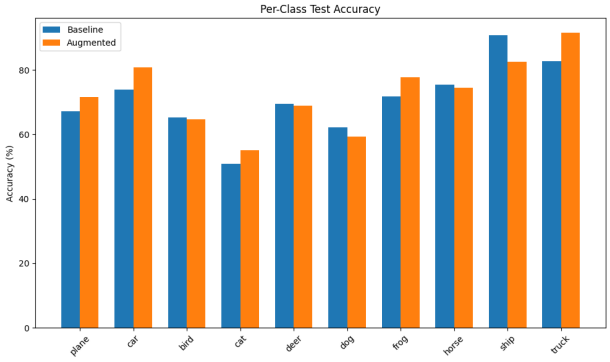

Fig. 6. Per-class test accuracy for each model.

image data, which may limit the generalizability of our results to more challenging domains.

V. Conclusion

This project explored the impact of common data augmentation techniques on the performance of convolutional neural networks trained on the CIFAR-10 dataset. Through careful experimentation and visualization, we demonstrated that data augmentation, even with relatively simple transformations, provides a measurable benefit in improving model generalization.

While the baseline model achieved higher training accuracy, it plateaued in generalization with a final test accuracy of 70.93%. The augmented model, though slower to converge during training, reached a higher test accuracy of 72.64%, validating that diversity introduced through augmentation prevents overfitting and helps the model perform better on unseen data.

Additional evaluation through confusion matrices and classification reports revealed that the augmented model made fewer severe misclassifications and achieved better balance across class predictions. Particularly, performance gains were notable in harder-to-classify categories such as “cat” and

"ship." A per-class accuracy breakdown confirmed that augmentation strategies provided enhanced feature robustness for several visually similar classes.

These findings affirm the practical importance of data augmentation in real-world applications where dataset size may be limited or costly to scale. Incorporating augmentation enables the development of more reliable and adaptable models without architectural changes or added labeling overhead. This can be particularly impactful in fields such as medical imaging, autonomous systems, and edge AI, where collecting diverse datasets is often infeasible.

Future work may involve exploring automated augmentation techniques (e.g., AutoAugment), adversarial augmentations, or applying the same strategies to more complex datasets. Additionally, testing across different architectures would help validate the generality of these results. Overall, this work underscores that even simple augmentation pipelines can lead to meaningful improvements in classification performance, reinforcing their value as a fundamental tool in deep learning workflows.

ACKNOWLEDGMENT

The author would like to express his sincere gratitude to Professor Nadra Guizani for her valuable guidance and feedback throughout the course of this project. Her insights in the field of machine learning greatly contributed to the development and direction of this work.

REFERENCES

[1] C. Shorten and T. M. Khoshgoftaar, "Understanding data augmentation for classification: When to warp?," *Journal of Big Data*, vol. 6, no. 1, pp. 1–21, 2019.

[2] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 113–123.

[3] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 8, no. 1, pp. 1–54, 2021.