Comparison of MixUp and CutMix Data Augmentation on CIFAR-10

Mihir Ranjan

November 7, 2025

Contents

1 Introduction

Data augmentation is a key regularization technique used to improve the generalization of deep learning models, especially on small datasets such as CIFAR-10. In this report, I compare two advanced augmentation methods, **MixUp** and **CutMix**, to evaluate their effects on validation accuracy when training a ResNet-18 model with the AdamW optimizer.

MixUp and CutMix are designed to prevent overfitting and encourage smoother decision boundaries. **MixUp** generates synthetic samples by blending two random images and their corresponding labels, while **CutMix** creates composite images by replacing a random rectangular region from one image with a patch from another image. Both strategies aim to improve robustness and model calibration by forcing the network to learn from mixed inputs.

2 Methodology

2.1 Dataset and Model Setup

The experiments were conducted using the CIFAR-10 dataset, which contains 60,000 color images of size 32×32 pixels, categorized into 10 distinct classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The model used is a ResNet-18 architecture, denoted as $f(x;\theta)$ with parameters θ, producing class logits $z \in \mathbb{R}^{10}$.

Training utilized the AdamW optimizer with a learning rate of 5×10^{-4} and weight decay of 0.02. A cosine annealing learning rate schedule was applied following one warm-up epoch. Each model was trained for 100 epochs with a batch size of 128 using random cropping, horizontal flipping, and mixed-precision training (AMP).

2.2 MixUp Formulation

Given two random samples (x_i, y_i) and (x_j, y_j), the MixUp transformation is defined as:

$$\lambda \sim \text{Beta}(\alpha, \alpha), \quad \tilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \tilde{y} = \lambda y_i + (1-\lambda)y_j,$$

where y_i and y_j are one-hot label vectors. The corresponding loss function is:

$$\mathcal{L}_{\text{MixUp}} = \lambda\, CE(f(\tilde{x}), y_i) + (1-\lambda)\, CE(f(\tilde{x}), y_j).$$

This approach encourages linear behavior between training examples, leading to smoother decision boundaries and improved regularization.

2.3 CutMix Formulation

CutMix modifies the input by cutting and pasting image patches rather than blending pixels. Let $M \in \{0,1\}^{H \times W}$ be a binary mask representing the cut region:

$$\tilde{x} = M \odot x_i + (1-M) \odot x_j, \quad \tilde{y} = \lambda y_i + (1-\lambda)y_j,$$

where $\lambda = \frac{\text{area}(M)}{HW}$ and $\odot$ denotes element-wise multiplication. The loss is defined as:

$$\mathcal{L}_{\text{CutMix}} = \lambda\, CE(f(\tilde{x}), y_i) + (1-\lambda)\, CE(f(\tilde{x}), y_j).$$

This method compels the model to associate specific spatial regions with the correct labels, improving localization and robustness.

2.4 Training Configuration Summary

- **Dataset:** CIFAR-10
- **Model:** ResNet-18
- **Optimizer:** AdamW (lr $= 5 \times 10^{-4}$, weight decay $= 0.02$)
- **Scheduler:** Warm-up (1 epoch) $\rightarrow$ Cosine Annealing
- **Epochs:** 100
- **Batch size:** 128
- **Augmentation:** Random crop, flip, MixUp ($\alpha = 0.4$), CutMix ($\alpha = 1.0$)
- **Precision:** Automatic Mixed Precision (AMP)

3 Experimental Results

3.1 Training and Validation Performance

The validation accuracies for MixUp and CutMix are shown in Table 1. Both methods significantly improved performance over standard training, with CutMix achieving slightly higher accuracy.

Method	Final Accuracy	Best Accuracy	Epoch
MixUp	94.6%	94.6%	99
CutMix	95.0%	95.1%	91

Table 1: Validation accuracy comparison between MixUp and CutMix.

3.2 Observations

CutMix not only achieved a marginally higher validation accuracy but also converged faster, stabilizing by epoch 90 compared to MixUp's later convergence. The differences, though small, were consistent across runs.

4 Discussion

Both MixUp and CutMix effectively regularize training by introducing interpolated or composite samples that force the model to generalize beyond memorized examples.

MixUp improves generalization by enforcing linearity between input-label pairs, promoting smooth decision boundaries in high-dimensional space. However, it often produces visually unrealistic samples, which can make training less effective when spatial consistency is important.

CutMix, in contrast, preserves the semantic structure of images by mixing spatially meaningful regions. This helps the model learn to associate local features with the correct classes, which often leads to better calibration and robustness. The results suggest that CutMix's spatial reasoning advantage yields a stronger regularization signal, leading to better generalization and slightly faster convergence.

It is worth noting that both methods benefit significantly from the AdamW optimizer's decoupled weight decay, which provides stable convergence under aggressive augmentation.

5 Conclusion

In this experiment, both MixUp and CutMix achieved strong validation performance on CIFAR-10, demonstrating their effectiveness as augmentation strategies for small-scale image classification tasks.

CutMix slightly outperformed MixUp in both accuracy and convergence rate, suggesting that augmentations preserving spatial structure provide a more effective learning signal for convolutional models. The observed improvement, though modest, aligns with existing literature and indicates that hybrid approaches—combining MixUp and CutMix—may yield further gains.

Future work should include multiple randomized runs with statistical significance testing to better quantify the differences, as well as evaluations on larger datasets such as ImageNet.

References

1. Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. *MixUp: Beyond Empirical Risk Minimization*. International Conference on Learning Representations (ICLR), 2018.

2. Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. *CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features*. IEEE International Conference on Computer Vision (ICCV), 2019.

3. Loshchilov, Ilya, and Frank Hutter. *Decoupled Weight Decay Regularization*. International Conference on Learning Representations (ICLR), 2019.

4. Krizhevsky, Alex. *Learning Multiple Layers of Features from Tiny Images*. Technical Report, University of Toronto, 2009.