Comparison of MixUp and CutMix Data Augmentation on CIFAR-10

Mihir Ranjan

December 17, 2025

Contents

1 Introduction

For this assignment, I decided to compare the differences in validation accuracy between MixUp and CutMix on the CIFAR-10 dataset. For context, both MixUp and CutMix are popular augmentation strategies with the goal of improving validation accuracy for image classification. MixUp creates new training examples by taking a weighted linear interpolation of two existing samples and their labels, in other words, it blends the images. CutMix creates new training images by cutting a patch from one image and pasting it onto another image.

2 Methodology

2.1 Dataset and Model Setup

These data augmentation strategies are applied to the CIFAR-10 dataset, which consists of 60,000 small, 32 by 32 pixel color images across 10 classes. These classes are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Furthermore, this project will use the AdamW optimizer and the ResNet-18 model where $f(x;\theta)$ is the network with parameters θ, producing class logits $z \in \mathbb{R}^{10}$

2.2 MixUp

For two random samples (x_i, y_i) and (x_j, y_j)

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

y_i, y_j are one-hot label vectors where the model is trained to predict $\tilde{y}$ given $\tilde{x}$. The loss is defined as

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

This encourage linearly interpolated behavior between samples, so decision boundaries become smoother and reduces overfitting.

2.3 CutMix Formulation

Let the binary mask $M \in \{0,1\}^{H\times W}$ denote cut region

$$\tilde{x} = Mx_i + (1-M)x_j,\ \tilde{y} = y_i + (1-)y_j$$

where $\lambda = \frac{\text{area}(M)}{HW}$. 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 forces the model to reason about multiple spatial regions and associate them with the correct classes, which improves localization and robustness

2.4 Training Configuration

The dataset is CIFAR-10, the model is ResNet-18, the opitimizer is AdamW with a learning rate 5×10^{-4} and weight decay of 0.02. The scheduler is a warm-up with 1 epoch, turning to cosine annealing. We ran the model for 100 epochs each with a batch size of 128 and augmentation of random crops, flip and $\alpha = 0.4$ for mix-up and $\alpha = 1.0$ for CutMix. Lastly AMP was the precision

3 Experimental Results

3.1 Training and Validation Performance

Method	Final Accuracy	Best Accuracy	Epoch
MixUp	94.47%	94.47%	100
CutMix	94.76%	94.86%	80

Table 1: Validation accuracy comparison between MixUp and CutMix on CIFAR-10.

4 Discussion

From the results, the mix-up improves generalization by encouraging linearity in the data-label space. CutMix yields stronger gains when spatial features are informative

5 Conclusion

MixUp and CutMix are simple, parameter-free augmentations that enhance robustness and generalization in small image datasets, such as CIFAR-10. AdamW's decoupled regularization works well with these augmentations by maintaining stable optimization. Empirically, CutMix slightly outperforms MixUp on CIFAR-10, suggesting that spatially mixed features provide a stronger regularization signal for convolutional networks

References

[1] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, *mixup: Beyond Empirical Risk Minimization*, Proceedings of the International Conference on Learning Representations (ICLR), 2018. Available at: https://arxiv.org/abs/1710.09412

[2] S. Yun, D. Han, S. Oh, S. Chun, J. Choe, and Y. Yoo, *CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019. Available at: https://arxiv.org/abs/1905.04899