

Exploring the Impact of Dataset Characteristics on CutMix Augmentation Effectiveness

Sung-Hoon Kim

AIFEL Research 13[±]

Abstract

Data augmentation is a crucial technique that helps models learn more generalized representations from limited data. To this end, various methods have been proposed, including region-based (also known as patch-based) augmentation strategies. CutMix is widely considered one of the most effective region-based augmentation methods. While CutMix has shown remarkable performance improvements across various benchmarks, it is not universally beneficial. In this study, we investigate the conditions under which CutMix leads to performance gains, as well as scenarios where it may hinder learning. Our findings suggest that the effectiveness of CutMix is highly dependent on the characteristics of dataset.

1. Introduction

1.1. The role of Data Augmentation

Convolutional neural network (CNN) classifiers generally exhibit better performance when trained on large-scale datasets. However, when only limited data is available, models tend to overfit to specific features of the classes in the training set, resulting in poor generalization to unseen data. Data augmentation serves as a key technique to mitigate this overfitting issue and enhance the model's generalization capability. Accordingly, this study aims to investigate whether the augmentation method CutMix can effectively fulfill its intended role of improving generalization across different dataset characteristics.

1.2. CutMix Method

Among various data augmentation techniques, region-based augmentation methods manipulate specific regions of an image—such as by removing, mixing, or replacing them. These methods help reduce the model's reliance on partial features and encourage learning a broader set of discriminative cues, thereby providing a regularization effect. Among such approaches, **CutMix** has been widely recognized for its effectiveness.

CutMix generates augmented samples by cutting a patch from one image and pasting it onto another. The corresponding labels are also mixed in proportion to the area of the patch, encouraging the model to learn from more diverse and informative regions.





	ResNet-50	Mixup [48]	Cutout [3]	CutMix
Image				
Label	Dog 1.0	Dog 0.5 Cat 0.5	Dog 1.0	Dog 0.6 Cat 0.4
ImageNet Cls (%)	76.3 (+0.0)	77.4 (+1.1)	77.1 (+0.8)	78.6 (+2.3)
ImageNet Loc (%)	46.3 (+0.0)	45.8 (-0.5)	46.7 (+0.4)	47.3 (+1.0)
Pascal VOC Det (mAP)	75.6 (+0.0)	73.9 (-1.7)	75.1 (-0.5)	76.7 (+1.1)

Table 1: Overview of the results of Mixup, Cutout, and CutMix on ImageNet classification, ImageNet localization, and Pascal VOC 07 detection (transfer learning with SSD finetuning) tasks. Note that CutMix significantly improves the performance on various tasks.

1.3. Motivation and Research Objectives

While the effectiveness of CutMix is evident—as demonstrated in Table 1—it does not always yield positive results. In fact, there have been multiple cases, particularly observed through the Aiffel research classes, where CutMix hinders training and even degrades model performance.

This study aims to investigate the conditions under which CutMix enhances performance, as well as the conditions that lead to negative outcomes. In particular, we focus on how the characteristics of the dataset influence the effectiveness of CutMix.

By doing so, we hope to prevent counterproductive and inefficient outcomes that may arise from the inappropriate use of CutMix. Furthermore, our findings are expected to contribute to the development of more informed and effective strategies for applying augmentation techniques.

2. Method

2.1. Datasets

In this study, the dataset itself serves as the primary experimental variable. We conducted our experiments on two datasets: Stanford Dogs and Food-101. Datasets such as CIFAR and ImageNet, where the effectiveness of CutMix has already been widely demonstrated, were excluded from this study.

Stanford Dogs requires distinguishing between classes based on fine-grained visual features. Since all classes represent dog breeds, they share many common attributes, and the discriminative cues are often limited to very small regions such as the face or even smaller parts. We aim to investigate whether CutMix, which overlays patches between images, disrupts learning by obscuring these key features, or instead encourages the model to learn from a broader set of features, thereby improving performance.

Food-101, like Stanford Dogs, also has many classes that look similar to each other. However, the features that help distinguish between classes are not limited to small areas like the face or eyes. We chose these two datasets because they are both fine-grained, but they differ in how the important features for classification are spread across the images.

2.2. Architecture

We applied the ResNet-50 architecture for our experiments. Since the performance of CutMix reported in Table 1 was based on ResNet-50, we used the same architecture to maintain consistency. ResNet is widely recognized as a standard benchmark model for image classification, and given that our experiments also focus on classification tasks, ResNet-50 is an appropriate choice to validate the effectiveness of CutMix. By using the same model architecture as the original paper, we ensure that any differences in results are not due to architectural variations.

2.3. Experimental Setup

In this experiment, the patch size used in CutMix was determined according to the following procedure:

$$\begin{aligned} r_w &\sim \mathcal{U}(0, 1), \quad r_h \sim \mathcal{U}(0, 1) \\ W &= \left\lfloor W_{\text{img}} \times \sqrt{1 - r_w} \right\rfloor \\ H &= \left\lfloor H_{\text{img}} \times \sqrt{1 - r_h} \right\rfloor \end{aligned}$$

r is a random real number sampled from a uniform distribution between 0 and 1. W_{img} and H_{img} represent the width and height of the original image, respectively. The width and height of the patch are calculated by multiplying the original dimensions by $\sqrt{1-r}$, and then taking the integer part.

The location of the patch is also randomly determined. A random center point (x, y) is selected within the image dimensions, and the patch is positioned around this point. To ensure the patch does not exceed the image boundaries, its coordinates are clipped as follows:

$$\begin{aligned} x_{\min} &= \max(0, x - \frac{W}{2}) \\ y_{\min} &= \max(0, y - \frac{H}{2}) \\ x_{\max} &= \min(W_{\text{img}}, x + \frac{W}{2}) \\ y_{\max} &= \min(H_{\text{img}}, y + \frac{H}{2}) \end{aligned}$$

This ensures that the entire patch remains within the bounds of the image, regardless of the randomly chosen size and position.

The learning rate was set to 0.01, and the optimizer used was stochastic gradient descent (SGD), following the original CutMix paper. No learning rate decay was applied. All experiments were conducted using TensorFlow version 2.10.1.

2.4. Experiment

We trained and compared models on datasets augmented with four different strategies:

No Augmentation: This serves as the baseline model trained on data without any augmentation.

Basic Augmentation: To provide a reference for augmentation effects, basic data augmentations including random left right flipping and random brightness adjustment were applied.

Mixup: This augmentation technique blends images and labels similarly to CutMix. Previous studies have reported that CutMix overcomes some limitations of Mixup. In this study, Mixup was included to observe under which conditions such similar augmentation methods might negatively impact performance. Basic Augmentation was also applied alongside Mixup.

CutMix: Models were trained on data augmented using the CutMix method described in Section 2.3. Basic Augmentation was applied in conjunction.

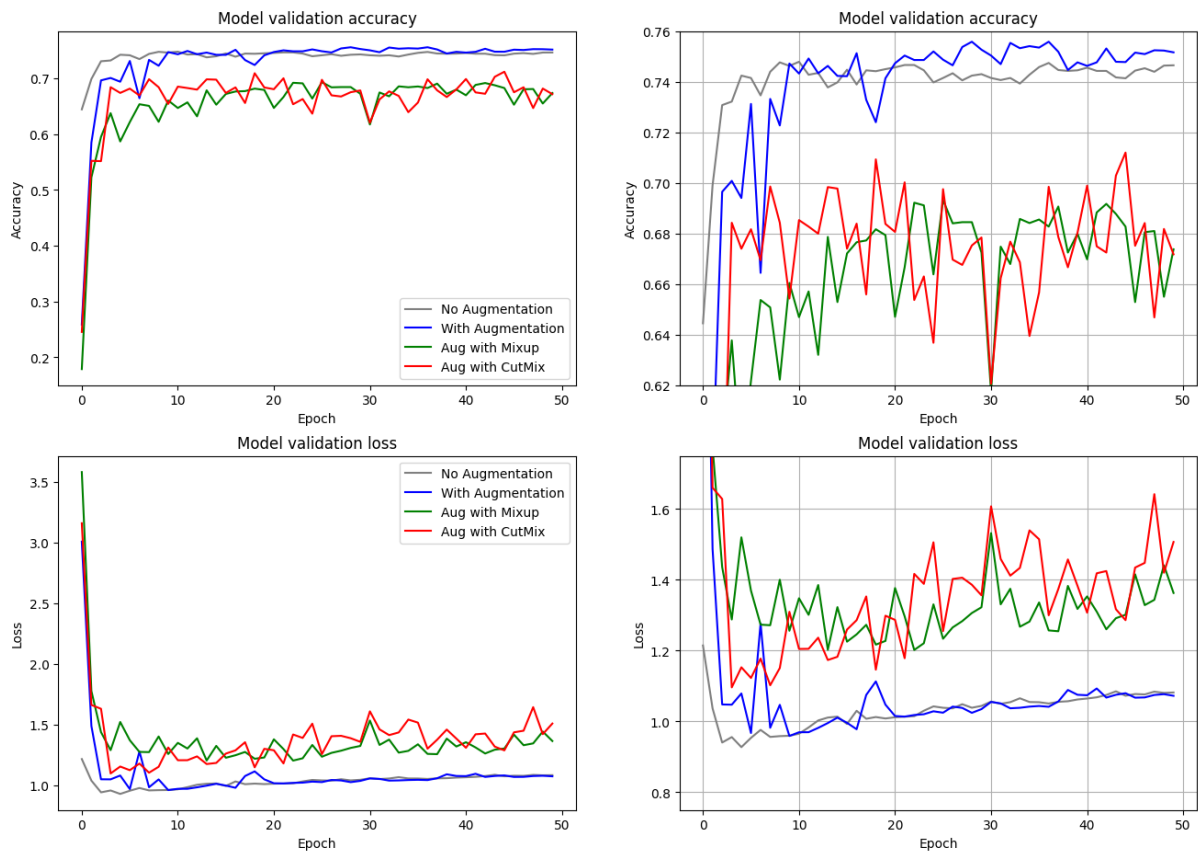
2.5. Metric

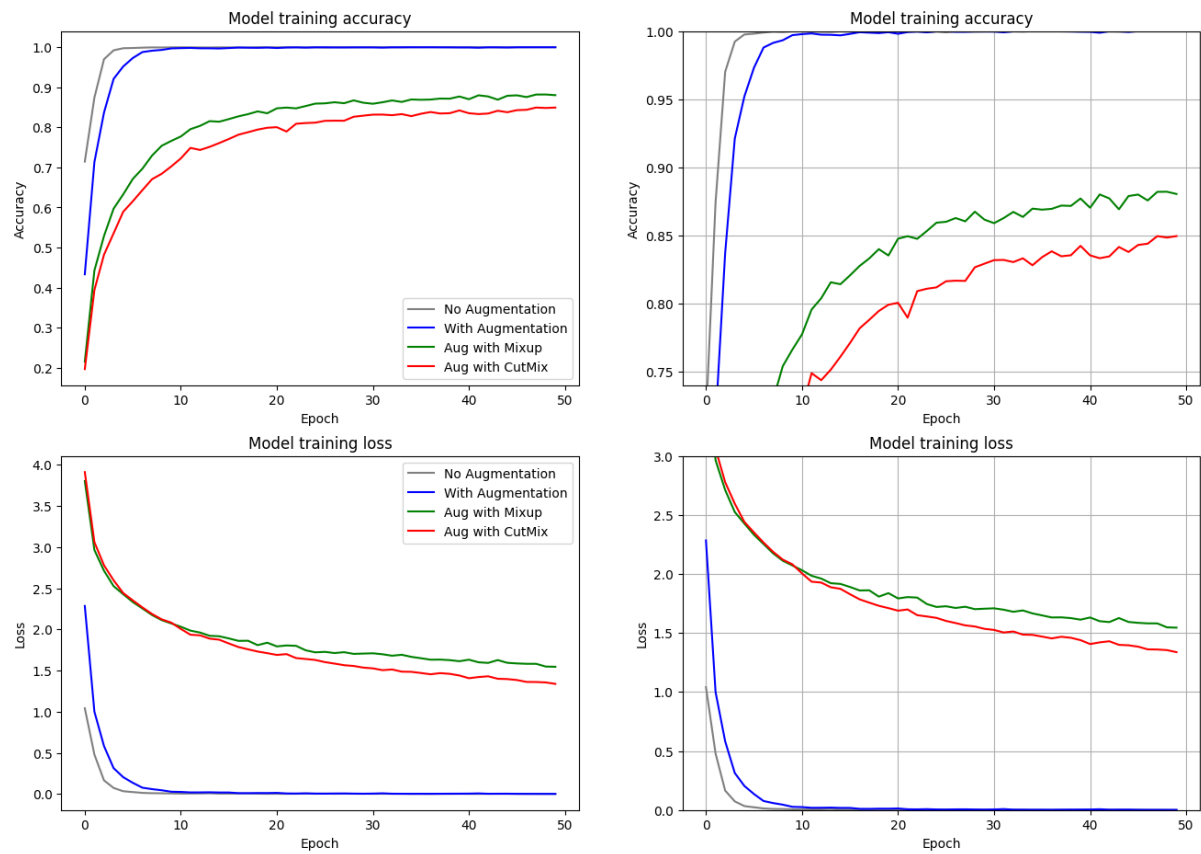
In classification tasks, common evaluation metrics include Accuracy and F1 Score. The F1 Score is particularly useful to address the issue of skewed class distributions, where Accuracy can be misleading. However, since the datasets used in this study have balanced class distributions, we adopt Accuracy as the sole evaluation metric for its simplicity and interpretability.

3. Result

3.1. Stanford Dogs dataset

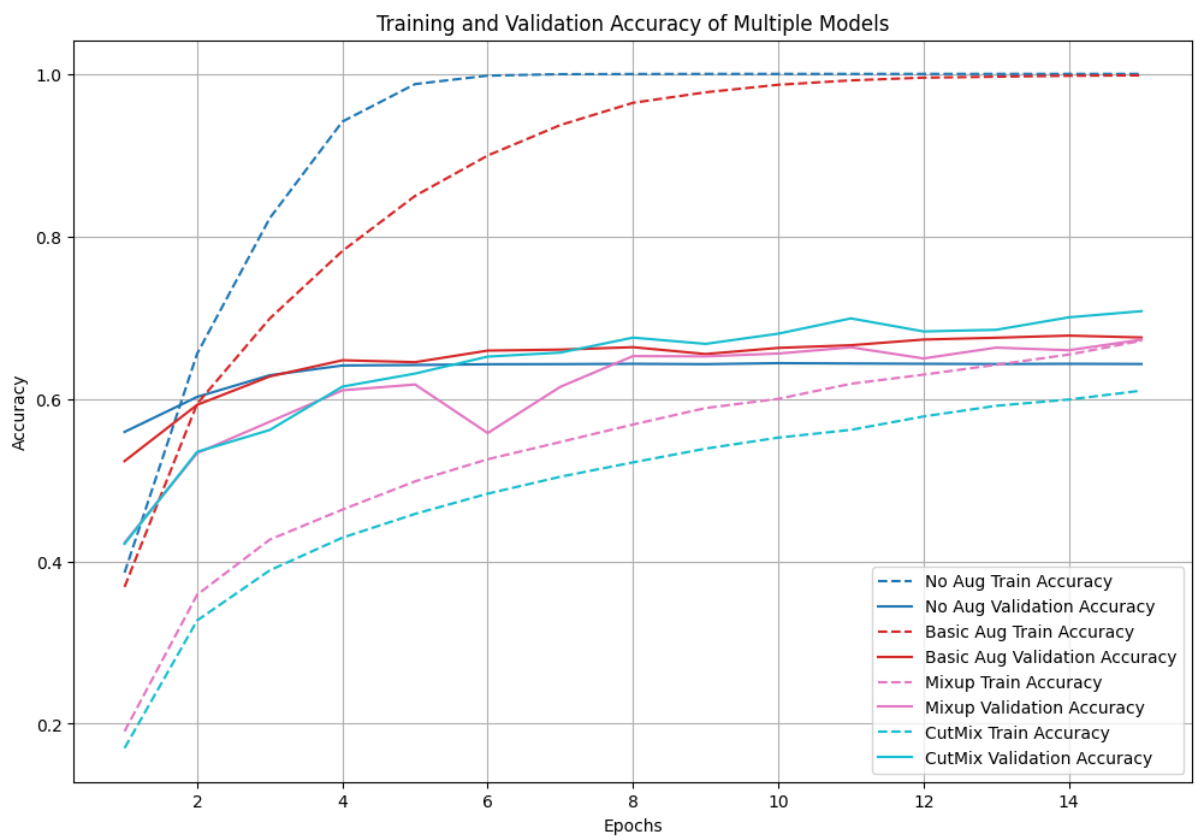
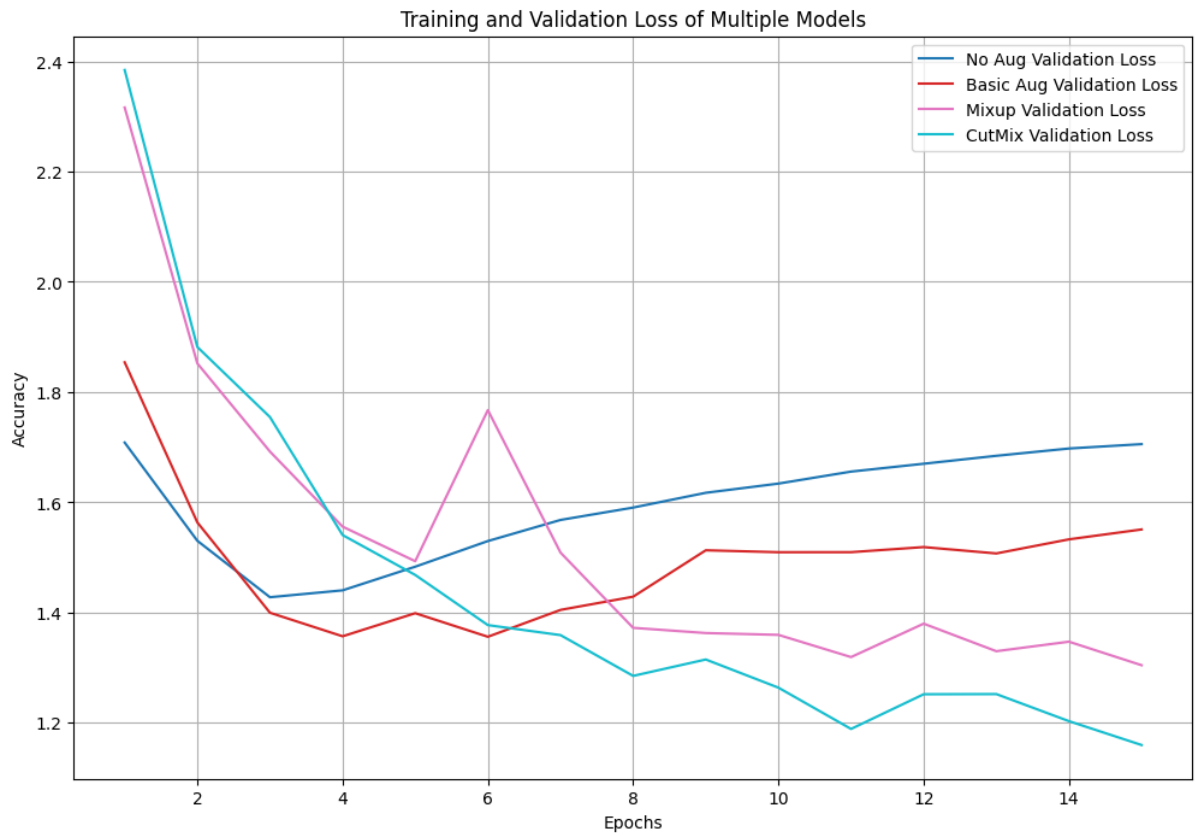
The results on the Stanford Dogs dataset are as follows:



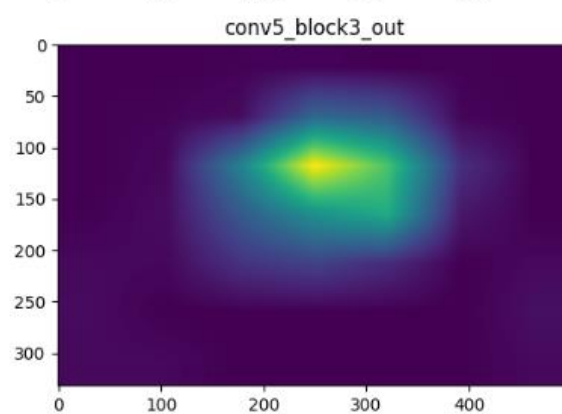
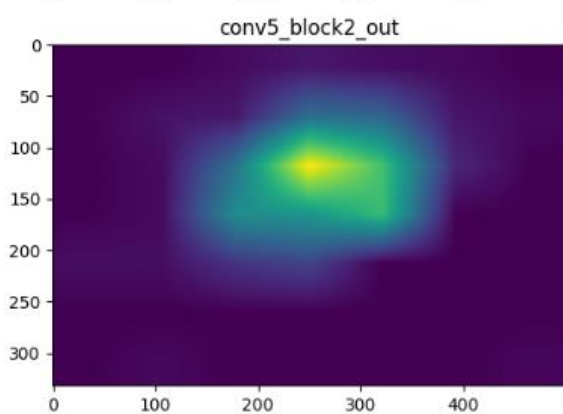
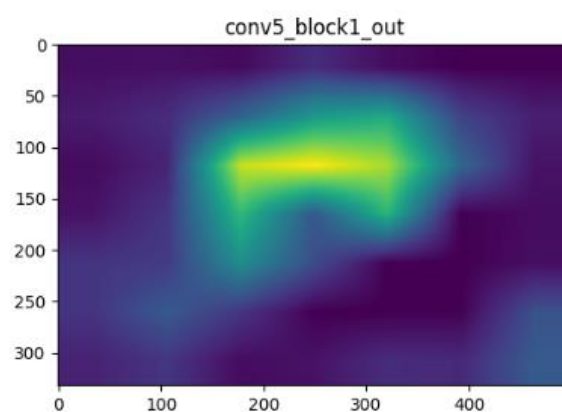
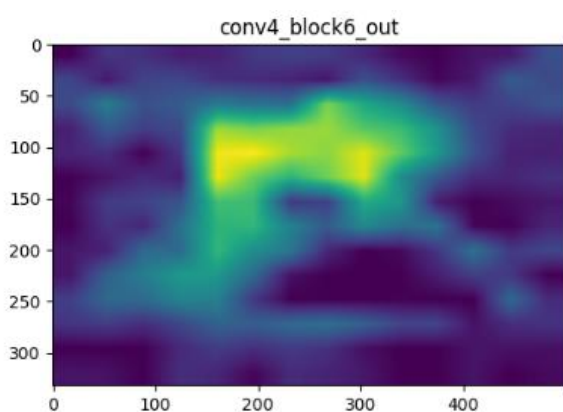


3.2. Food-101 dataset





4. Discussion



5. Conclusion

References