Evaluating Augmentation Techniques: CutMix and MixUp for ResNet Model on the Stanford Dogs Dataset - Not Always Guaranteed Performance Improvements

KyoungGyu Lee
qwaszx3801@gmail.com

Abstract

In this study, we aimed to compare the effects of augmentation techniques, CutMix and MixUp, using the Stanford Dogs Dataset. We evaluated the performance under various augmentation settings on the ResNet50 model, including basic augmentations such as horizontal flipping and brightness adjustment. The experimental results showed that while CutMix and MixUp can improve performance, they do not always guarantee better results. In the experiments, CutMix and MixUp demonstrated similar performance, with some differences: CutMix tended to converge faster during the initial training phase compared to MixUp, which exhibited a more stable learning process. This study analyzes the impact of data augmentation techniques on model performance and discusses scenarios where these techniques may not be beneficial.

1. Introduction

Deep learning models require a large amount of data to achieve high performance. However, in practice, it is often challenging to secure sufficient data. To address this issue, data augmentation techniques are widely used. Data augmentation enhances the generalization ability of models by creating new data through various transformations of existing data. Common data augmentation techniques include rotation, translation, horizontal flipping, and brightness adjustment.

Recently, in addition to these basic augmentation techniques, more complex methods such as CutMix and MixUp have been proposed. CutMix generates new images by cutting out a portion of one image and replacing it with a part of another image, allowing the model to learn from diverse backgrounds and objects. MixUp creates new images by weightedly combining two images, enabling the model to learn more generalized features.

In this study, we aim to compare the effects of CutMix and MixUp using the Stanford Dogs dataset. The Stanford Dogs dataset includes 120 dog breeds and provides images for each breed. We will use the ResNet50 model to evaluate the performance of various augmentation settings and analyze the impact of each augmentation technique on the model's performance. Through this analysis, we intend to highlight that CutMix and MixUp do not always guarantee performance improvements.

2. Related works

In recent studies, advancements in data augmentation techniques have significantly contributed to the improvement of deep learning model performance. Both CutMix and MixUp are recent data augmentation methods that aim to enhance model generalization by synthesizing new images from two existing images.

Cutmix: technique that involves cutting out a portion of one image and combining it with another image, allowing the model to learn features from both the original and the other image simultaneously. Proposed by Naver CLOVA, this method is implemented by cutting a rectangular patch from a random location in one image and pasting it onto the same location in another image. This technique helps the model to recognize objects in various situations by randomly altering important parts of the image.

Mixup: proposed by Zhang et al (2018), is a technique that generates new images by weightedly combining two images. This method creates synthetic images by taking the weighted average of the pixel values of two images and similarly mixing their labels with the same weights. Through this approach, the model learns from various image combinations, which helps prevent overfitting and enhances generalization performance.

Each of these techniques has its own advantages and disadvantages. This study aims to compare and analyze these methods to evaluate their impact on model performance.

3. Method

Used 6 models of ResNet50 for each train dataset.

3.1. Dataset

The Stanford Dogs dataset was utilized. The Stanford Dogs dataset consists of a total of 120 dog breeds, with an average of 150 images per breed. Comprising 20,580 images in total, this dataset is well-suited for image classification tasks such as breed classification. Each image is labeled according to the dog breed, and in this study, these labeled images were used for training and validation.

3.2. Model

In the experiments, the ResNet50 model was employed. ResNet50, a Residual Network, was proposed by Kaiming He et al. (2016). This model is composed of 50 layers and utilizes residual blocks to increase the depth of learning while maintaining performance. In this study, the ResNet50 model was created and trained under various augmentation settings.

Data Augmentation:

- **Basic Augmentation:** Horizontal flipping, brightness adjustment
- **CutMix:** Cutting out a portion of one image and combining it with another image
- **MixUp:** Generating new images by weightedly combining two images

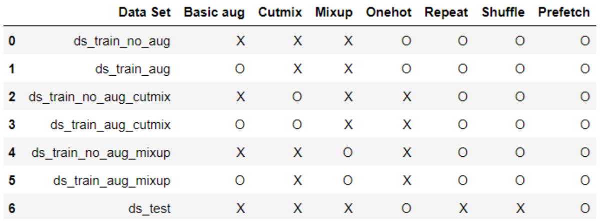
Experimental Setup:

- Generation of datasets with and without augmentation
- Generation of datasets with and without the application of CutMix and MixUp
- Training and performance evaluation of the ResNet50 model for each setting

Through this setup, we aimed to analyze the impact of each augmentation technique on the model's performance.

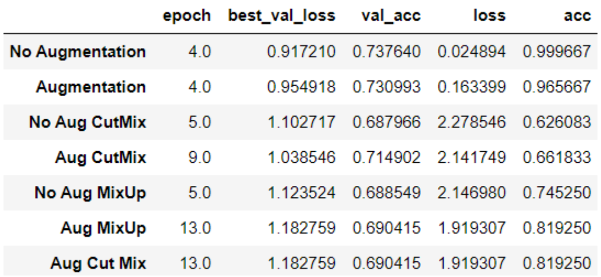
3.3. Performance Evaluation

The model performance was evaluated based on validation loss and accuracy. For each model, the optimal validation loss and the corresponding epoch's loss, accuracy, and validation accuracy were compared. This allowed for a quantitative assessment of the effectiveness of each augmentation technique.



	Data Set	Basic aug	Cutmix	Mixup	Onehot	Repeat	Shuffle	Prefetch
0	ds_train_no_aug	X	X	X	O	O	O	O
1	ds_train_aug	O	X	X	O	O	O	O
2	ds_train_no_aug_cutmix	X	O	X	X	O	O	O
3	ds_train_aug_cutmix	O	O	X	X	O	O	O
4	ds_train_no_aug_mixup	X	X	O	X	O	O	O
5	ds_train_aug_mixup	O	X	O	X	O	O	O
6	ds_test	X	X	X	O	X	X	O

Figure 1. Train Data Set



	epoch	best_val_loss	val_acc	loss	acc
No Augmentation	4.0	0.917210	0.737640	0.024894	0.999667
Augmentation	4.0	0.954918	0.730993	0.163399	0.965667
No Aug CutMix	5.0	1.102717	0.687966	2.278546	0.626083
Aug CutMix	9.0	1.038546	0.714902	2.141749	0.661833
No Aug MixUp	5.0	1.123524	0.688549	2.146980	0.745250
Aug MixUp	13.0	1.182759	0.690415	1.919307	0.819250
Aug Cut Mix	13.0	1.182759	0.690415	1.919307	0.819250

Figure 2. Comparison of Model Performance with Various Data Augmentation

4. Result

For the model without mixed augmentations like CutMix and MixUp, the optimal validation loss was achieved at epoch 4. In contrast, models with CutMix and MixUp augmentations reached optimal validation loss at epoch 5 without basic augmentation. When basic augmentation was applied, these models converged at epochs 9 and 15, indicating the need for more training to reach convergence.

Additionally, the model without mixed augmentations showed higher validation accuracy. Among the models with mixed augmentations, the one with CutMix exhibited lower accuracy on the training data compared to the model with MixUp. This suggests that the CutMix model receives more noise, which can be interpreted as CutMix introducing more variability in the data.

5. Discussion

First, we will visualize and compare the entire training process for each of the six training datasets corresponding to the different augmentation methods mentioned earlier. Then, we will compare the results based on the presence or absence of basic augmentation in three scenarios (baseline dataset, CutMix-applied dataset, MixUp-applied dataset). Finally, we will compare CutMix and MixUp with and without basic augmentation.

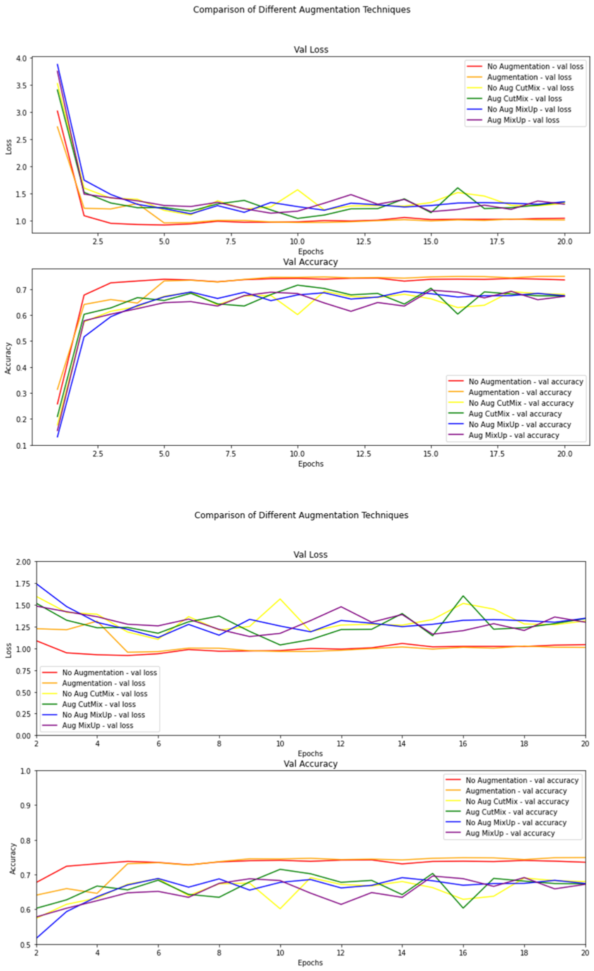


Figure 3. Overall Model

Overall Model Graph Interpretation

Contrary to the expectation that applying CutMix or MixUp would enhance performance, the results showed otherwise. The model without any augmentation (red) exhibited the fastest decrease in loss and convergence, which was somewhat anticipated. However, the models with CutMix or MixUp did not demonstrate the best performance. The model with basic augmentation (left-right flip, brightness adjustment, orange) showed a trend similar to the model without augmentation (red) and appeared to yield the best performance as training continued.

Comparison in each Model with or without Basic Augmentation

Comparison in each Model with or without Basic Augmentation As mentioned in the Overall Model Graph Interpretation, the first graph shows that the model without basic augmentation converges faster and the model with basic augmentation appears to yield better performance. This suggests that basic augmentation effectively increases data diversity, enabling the model to learn better.

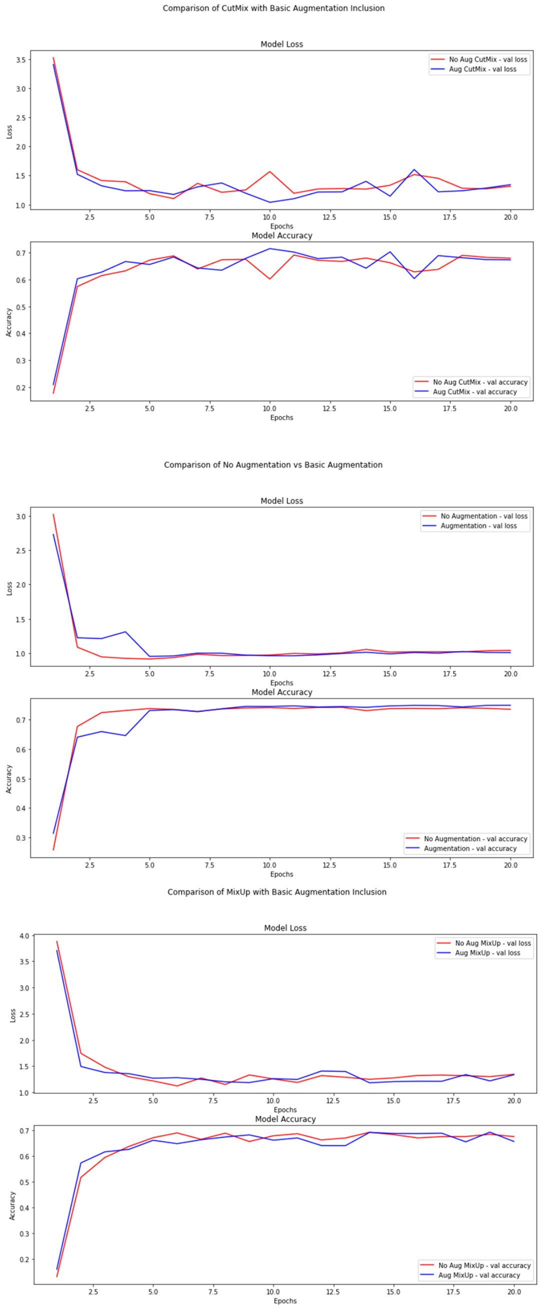


Figure 4. Comparison in each Model with or without Basic Augmentation

Similarly, the models with the mixed augmentations, CutMix and MixUp, also seem to perform better when basic augmentation is applied. This indicates that even with the application of mixed augmentations, the positive impact of basic augmentation is evident.

However, in the case of mixed augmentations, it appears that models with basic augmentation converge faster than those without it. This could be because mixed augmentations increase data complexity, and the additional data diversity provided by basic augmentation helps create a more uniform and regularized dataset.

Comparison of Model in each Basic Augmentation

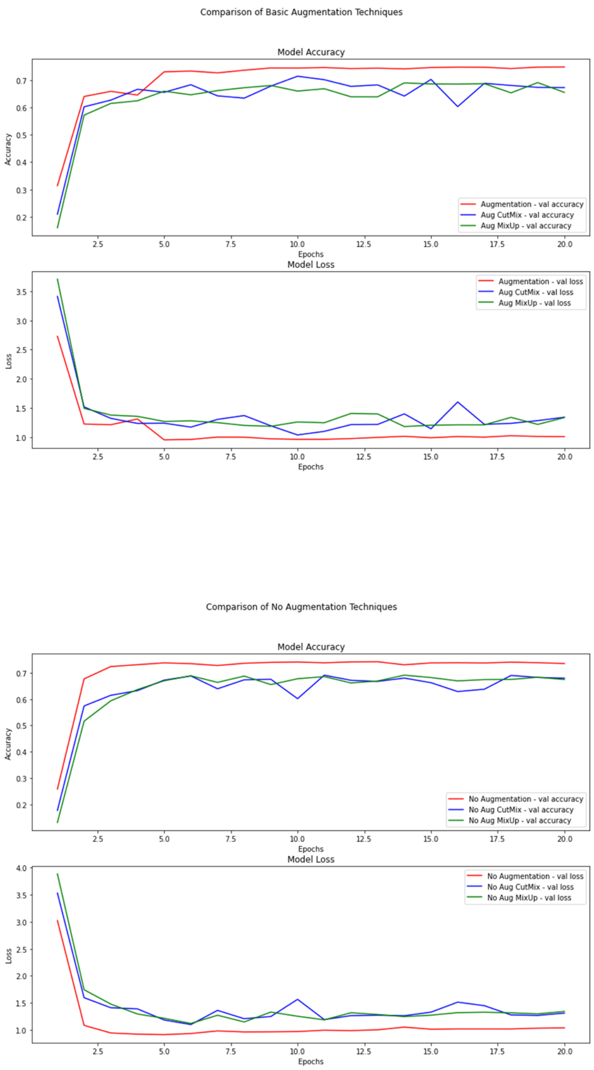


Figure 5. Comparison of Model in each Basic Augmentation

This graph similarly shows that the model without basic augmentation performs better. When comparing CutMix and MixUp, we observe that CutMix leads to a faster decrease in validation loss, but MixUp provides more stable learning compared to CutMix. Consequently, as training progresses, both techniques exhibit similar performance.

6. Conclusion

In experiments using the Stanford Dogs dataset, the application of CutMix and MixUp augmentation techniques did not always yield the expected performance improvements. These results are consistent with several key studies. Here, we analyze the potential reasons for these outcomes by referencing relevant literature.

6.1. Class Homogeneity

The Stanford Dogs dataset consists of high-resolution images of various dog breeds, which are structurally very similar. This homogeneity can cause issues when applying augmentation techniques like CutMix and MixUp. Yun et al. (2019) emphasized that the CutMix technique, which involves mixing image regions, can hinder the model's ability to learn distinct class features. In our results, the model struggled to distinguish between similar breeds, leading to decreased accuracy.

6.2. Application Methods of Augmentation Techniques

Techniques like MixUp and CutMix, which mix pixels or cut and paste parts of images, can blur important visual features. Zhang et al. (2018) noted that MixUp can cause models to learn ambiguous features. Similarly, Yun et al. (2019) discussed that applying CutMix can replace significant parts of an image (e.g., a dog's face or ears) with parts of another image. This issue was particularly problematic for the Stanford Dogs dataset, where critical parts for accurate classification were mixed.

6.3. Dataset Characteristics

This study found that the model's performance on the Stanford Dogs dataset was suboptimal compared to expectations. This could be due to the characteristics of the dataset. According to Misra et al. (2015) and He et al. (2016), mixing high-resolution, fine-grained images like those in the Stanford Dogs dataset can disrupt the critical details necessary for the model to recognize specific classes accurately.

6.4. Model Complexity

The Stanford Dogs dataset consists of high-resolution and detailed images, making it challenging for simple augmentation techniques to effectively enhance model generalization. He et al. (2016) demonstrated that while deep residual networks can handle complex datasets, inappropriate augmentation can still confuse the learning process. Our results showed that models trained with CutMix and MixUp did not always outperform the baseline.

7. Analysis of Unexpected Outcomes

The unexpected performance outcomes in experiments with CutMix and MixUp can be attributed to the homogeneity of the dataset, the disruptive nature of the augmentation techniques, insufficient hyperparameter tuning, and the inherent complexity of the dataset. Future research should consider these factors and explore more sophisticated augmentation methods or improved tuning strategies to achieve better performance.

References

1. Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., Yoo, Y. (2019). CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 6023-6032).

2. Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D. (2018). MixUp: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR).*

3. He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M. (2019). Bag of Tricks for Image Classification with Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 558-567).

4. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778).

5. Misra, I., Zitnick, C. L., Hebert, M. (2015). Exploring the Limits of Weakly Supervised Pretraining. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*