Understanding the Role of CutMix in ETF During Neural Collapse

Nazmus Saqib, Joon Min-Gil*
Dept. of Computer Engineering, Jeju National University, South Korea.

Abstract

CutMix enhances deep neural network performance by generating convex combinations of training samples and labels. Despite its success, its effect on the data manifold during training remains underexplored. We investigate this through Neural Collapse, where last-layer features converge to a simplex equiangular tight frame (ETF). Using a controlled setup, we compare CutMix with MixUp in promoting ETF formation. Our results show that CutMix forms ETF structures more gradually than MixUp. However, it offers improved generalization and localization performance. We analyze geometric configurations to uncover mechanisms behind this behavior. This study sheds light on how CutMix influences representation learning dynamics.

Ⅰ. Introduction

Consider a classification problem characterized by an input and an output space, which are expressed as $\chi \epsilon \mathbb{R}^D$ and $\gamma := \{0,1\}^C$, respectively. Given a training set of samples, $\{(x_i, y_i)\}_{i=1}^N$, with $x_i \epsilon \chi$ denoting the i^{th} input data point and $y_i \epsilon \gamma$ representing the corresponding label, the goal is to train a model $f_\theta : \chi \rightarrow \gamma$ by finding parameters θ that minimizes the cross-entropy loss $CE(f_\theta(x_i), y_i)$ incurred by the model prediction $f_\theta(x_i)$ relative to the true target y_i, averaged over the training set, $\frac{1}{N}\sum_{i=1}^N CE(f_\theta(x_i), y_i)$.

Papyan et al. [1] observed that during loss function optimization, neural networks exhibit a phenomenon known as *Neural Collapse*, where the last-layer activations and classifier weights converge to the geometric structure of a simplex equiangular tight frame (ETF). This configuration reflects the network's inherent tendency to arrange class representations such that they are aligned with their corresponding classifiers, possess equal norms, and are equally spaced in angle—thereby achieving optimal class separation in the feature space. Understanding Neural Collapse is challenging due to the complex architecture and intrinsic non-linearity of neural networks.

A. CutMix

CutMix [2] is a popular data augmentation strategy, which generates new training examples through convex combinations of existing data points and corresponding labels.

$$x_{ii'}^{\lambda} = M \odot x_i + (1 - M) \odot x_{i'}, y_{ii'}^{\lambda} = \lambda y_i + (1 - \lambda) y_{i'} \quad (1)$$

where $M \epsilon \{0,1\}^{W \times H}$ denotes a binary mask indicating the dropout and fill which happened by exchanging information between two images. 1 is a binary mask filled with ones, and $\odot$ is the elementwise multiplication. Like Mixup [3], another popular data augmentation, CutMix also uses λ which is a symmetric Beta $\beta(\alpha, \alpha)$ distribution, where α is set to 1. The loss associated with CutMix can be mathematically represented as:

$$E_{\lambda \sim D_\lambda} \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N CE(f_\theta(x_{ii'}^{\lambda}), y_{ii'}^{\lambda}) \quad (2)$$

B. Problem Statement and Contribution

Despite the widespread use and demonstrated efficacy of the CutMix data augmentation in enhancing generalization of deep neural networks, the underlying mechanism demands Neural Collapse for the following primary question:

"Does CutMix induce either similar or different geometrical configurations from Neural Collapse? Whatever the pattern it induces, how rapid or gradual the process to form the configurations is compared to other data augmentation strategies, and what is the theoretical reason behind this?"

This study aims to uncover the potential geometric configurations in the last layer activations resulting from CutMix.

Ⅱ. Result Summary

The results of our primary empirical investigation are presented in Fig. 1. This figure presents a comparative study between CutMix and Mixup to illustrate the delayed formulation of the ETF in CutMix. Our study incorporates the WideResNet-40-10 architecture on the dataset CIFAR10 using Adam as the optimizer.

Ⅲ. Experiments

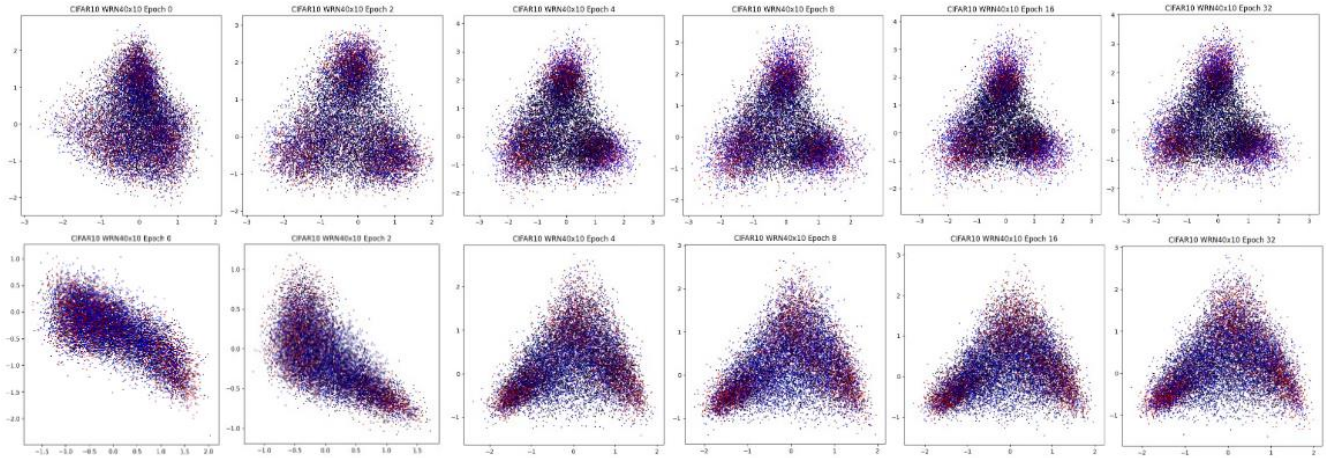

Fig 1. Evolution of three class-conditional feature activations under **Mixup** (top) and **CutMix** (bottom) training. Columns show epochs 0, 2, 4, 8, 16, and 32 (left → right). Mixup rapidly organizes the penultimate-layer features into a simplex ETF, whereas CutMix exhibits a slower, more gradual emergence of the ETF geometry.

For the above dataset and network pair, we visualize the last-layer activations for a subset of the training dataset consisting of three randomly selected classes. After obtaining the last-layer activations, they undergo a two-step projection: first onto the classifier for the subset of three classes, then onto a two-dimensional representation of a three-dimensional simplex ETF.

IV. Theoretical Justification:

CutMix tends to reduce regional dropouts, which aims to improve utilization for better generalization and localization performance. To do so, CutMix pastes inactive patches, so large portions of the image remain purely from one class while others come from another class. This creates piecewise-linear effects in feature space, where subregions belong to different classes. Thus, CutMix enjoys the property that there is no uninformative pixel during training to make it more efficient by retaining the advantages of regional dropout to attend to non-discriminative parts of objects. Our investigation reveals that while maintaining better generalization in training, CutMix fails to provide an early illustration of ETF. The theoretical reason behind this is CutMix's creation of piecewise-linear effects in feature space, while subregions belong to different classes. As a result, early in training, class features are more scattered and less aligned to an ETF structure. On the other hand, MixUp blends every pixel between two images, producing a linear combination of entire feature maps. This introduces strong linearity in feature space from the start, enforcing class-mean vectors to move quickly toward the symmetric, equidistant arrangement of ETF. We assume that, at both ends, the label mixing ratio λ directly influences the effect. MixUp's earlier formation of ETF directly influenced by λ which directly matches the pixel-mixing ratio, ensuring a consistent and immediate label-feature relationship. On the other hand, CutMix's λ depends on the patch area, and feature activations are influenced by contextual cues (background, object parts), which slows down the uniform convergence of class features toward the ETF vertices. The reduction of intra-class variance for CutMix is more gradual than MixUp, as CutMix preserves spatial structures, maintains higher-class variance longer, which slows down the ETF process. MixUp faces fast variance reduction by smoothing decision boundaries fast.

A. From data manifold perspective

As a large portion of local features are unchanged in the original image rather than the changed patches, CutMix creates piecewise manifold transitions instead of piecewise manifold transitions where local subregions belong to different class manifolds.

Ⅲ. Conclusion

We provide an investigative study about how gradual CutMix reforms the ETF in Neural Collapse. Though CutMix reforms it gradually, CutMix presents better generalization and localization performance compared to other data augmentation strategies, which present an earlier formation of ETF. In future works, we will carry out additional experiments for different datasets and networks, and theoretically characterize the optimal last-layer features.

ACKNOWLEDGMENT

This research was supported by the Regional Innovation System & Education(RISE) program through the Jeju RISE center, funded by the Ministry of Education(MOE) and the Jeju Special Self-Governing Province, Republic of Korea(2025-RISE-17-001).

References

[1] Papyan, Vardan, X. Y. Han, and David L. Donoho. "Prevalence of neural collapse during the terminal phase of deep learning training." *Proc. of the National Academy of Sciences*, vol. 117, no. 40, pp. 24652-24663, 2020.

[2] Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." *arXiv preprint arXiv:1710.09412* (2017).

[3] Yun, Sangdoo, et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features." *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 6022-6031, 2019.