ζ-*mixup*: Richer, More Realistic Mixing of Multiple Images

Kumar Abhishek[1] KABHISHE@SFU.CA
Colin J. Brown[2] COLIN.BROWN@HINGEHEALTH.COM
Ghassan Hamarneh[1] HAMARNEH@SFU.CA
[1] *School of Computing Science, Simon Fraser University, Canada*
[2] *Hinge Health, Canada*

Editors: Accepted for publication at MIDL 2023

Abstract

Data augmentation (DA), an effective regularization technique, generates training samples to enhance the diversity of data and the richness of label information for training modern deep learning models. *mixup*, a popular recent DA method, augments training datasets with convex combinations of original samples pairs, but can generate undesirable samples, with data being sampled off the manifold and with incorrect labels. In this work, we propose ζ-*mixup*, a generalization of *mixup* with provably and demonstrably desirable properties that allows for convex combinations of $N \geq 2$ samples, thus leading to more realistic and diverse outputs that incorporate information from N original samples using a p-series interpolant. We show that, compared to *mixup*, ζ-*mixup* better preserves the intrinsic dimensionality of the original datasets, a desirable property for training generalizable models, and is at least as fast as *mixup*. Evaluation on several natural and medical image datasets shows that ζ-*mixup* outperforms *mixup*, CutMix, and traditional DA methods.

Keywords: data augmentation, mixup, intrinsic dimensionality, data manifold

1. Introduction

Given the large parameter space of deep learning models, training on small datasets tends to cause the models to overfit to the training samples, which is especially a problem when training with data from high dimensional input spaces such as images, and consequently, benefits from data augmentation (DA) techniques for improved generalization performance. *mixup* (Zhang et al., 2018), a popular DA method, generates convex combinations of pairs of original training samples and linear interpolations of corresponding labels with a hyperparameter $\lambda \sim [0, 1]$. The primary hypothesis of *mixup* and many derivatives is that a model should behave linearly between any two training samples, even if the distance between samples is large. This implies that we may train the model with synthetic samples that have very low confidence of realism; in effect, over-regularizing. We instead argue that we should only synthesize examples with high confidence of realism, and that a model should only behave linearly nearby training samples, supported by research in cognitive sciences showing that human perception between object category boundaries is warped and not as linear as *mixup* seems to suggest (Beale and Keil, 1995; Newell and Bülthoff, 2002).

Consider the $\mathcal{K}$-class classification task, where we are provided with a dataset of m points $\{x_i\}_{i=1}^m$ in a $\mathcal{D}$-dimensional ambient space $\mathbb{R}^{\mathcal{D}}$ with the corresponding labels $\{y_i\}_{i=1}^m$ in a label space $\mathcal{L} = \{l_1, \cdots, l_{\mathcal{K}}\} \in \mathbb{R}^{\mathcal{K}}$. Keeping in line with the manifold hypothesis (Cayton, 2005; Fefferman et al., 2016), which states that complex data manifolds in high dimensional

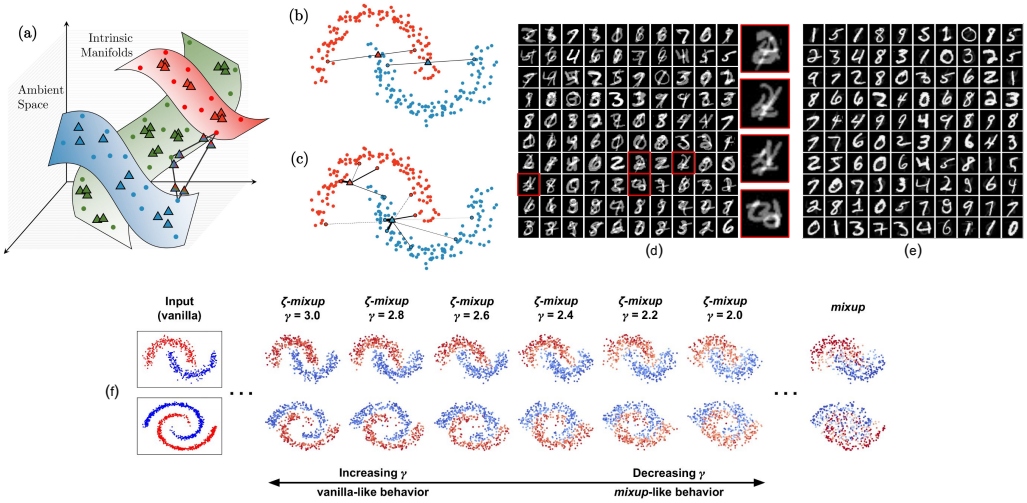

Figure 1: (a) An overview of ζ-*mixup* with original ($\circ$) and synthetic ($\triangle$) samples. Note how *mixup* ((b), (d)) does not respect individual class boundaries and can generate incorrect samples, that lie off the data manifold, with incorrect labels. ζ-*mixup* ((a), (c), (e)) can mix any number of samples (e.g., 3 in (a), 4 or 8 in (c), and 25 in (e)) and the generated samples remain close to the original distribution while incorporating rich information from several samples. (f) The hyperparameter γ in ζ-*mixup* formulation can control the diversity of the synthetic samples.

ambient spaces are actually made up of samples from manifolds with low intrinsic dimensionalities ($\mathcal{D}_{\text{int}}$), we assume that the m points are samples from $\mathcal{K}$ manifolds $\{\mathcal{M}_i\}_{i=1}^{\mathcal{K}}$ with $\mathcal{D}_{\text{int}}$ as $\{d_i\}_{i=1}^{\mathcal{K}}$, where $d_i << \mathcal{D}\ \forall i \in [1, \mathcal{K}]$ (Fig. 1 (a)). We seek an augmentation method that facilitates a denser sampling of each intrinsic manifold $\mathcal{M}_i$, thus generating more real and more diverse samples with richer labels. Following Wood et al. (2021); Wood (2021), we consider three criteria for evaluating the quality of synthetic data: **(i) realism:** allowing the generation of correctly labeled synthetic samples close to the original samples, ensuring the realism of the synthetic samples, **(ii) diversity:** facilitating the synthesis of more diverse samples by allowing exploration of the input space, and **(iii) label richness** when generating synthetic samples while still staying on the manifold of realistic samples. Additionally, we aim for: **(iv) valid probabilistic labels** along with **(v) computationally efficient** augmentation of training batches (e.g., avoiding inter-sample distance calculations).

To this end, we propose to synthesize a new sample $(\hat{x}_k, \hat{y}_k)$ as $\hat{x}_k = \sum_{i=1}^{N} w_i x_i$; $\hat{y}_k = \sum_{i=1}^{N} w_i y_i$, where w_is are the weights assigned to the N samples being mixed. One such suitable weighting scheme is to sample weights from the terms of a p-series, i.e., $w_i = i^{-p}$, which is a convergent series for $p \geq 1$. Extending the idea of local synthetic instances for connectome augmentation (Brown et al., 2015), we adopt the following formulation: given N samples (where $2 \leq N \leq m$ and thus, theoretically, the entire dataset), a $N \times N$ random permutation matrix π, and the resulting randomized ordering of samples $s = \pi[1, 2, \ldots, N]^T$, the weights are defined as $w_i = \frac{s_i^{-\gamma}}{C}$, $i \in [1, N]$, where the hyperparameter γ allows us to control how far the synthetic samples can stray away from the original samples. C is the normalization constant to ensure that w_i must satisfy $w_i \geq 0\ \forall i$ and $\sum_{i=1}^{N} w_i = 1$, such that $\hat{y}_k$ is a valid probabilistic label, where $C = \sum_{j=1}^{N} j^{-\gamma}$ is the N-truncated Riemann

Table 1: Classification error on CIFAR datasets averaged over 3 runs ($\gamma \in \mathrm{U}[\gamma_{\min}, 4.0]$).

Method	CIFAR-10 ResNet-18	CIFAR-100 ResNet-18	Method	CIFAR-10 ResNet-18	CIFAR-10 ResNet-50	CIFAR-100 ResNet-18	CIFAR-100 ResNet-50
ERM	5.48	23.33	CutMix	4.13	4.08	19.97	18.99
mixup	4.68	21.85	+ ζ-*mixup*	**3.84**	**3.61**	**19.54**	**18.86**
ζ-*mixup*	**4.42**	**21.35**					

Table 2: Micro-averaged F1 score on skin lesion image datasets ($\gamma = 2.8$).

Method	ISIC 2016 ResNet-18	ISIC 2016 ResNet-50	ISIC 2017 ResNet-18	ISIC 2017 ResNet-50	ISIC 2018 ResNet-18	ISIC 2018 ResNet-50	DermoFit ResNet-18	DermoFit ResNet-50
ERM	0.7836	0.8127	0.7383	0.6867	**0.8756**	0.8653	0.8269	0.8500
mixup	0.7968	0.8179	0.7333	0.7433	0.8394	0.8601	0.8577	0.8500
ζ-*mixup*	**0.8654**	**0.8602**	**0.7633**	**0.7733**	**0.8756**	**0.9016**	**0.8731**	**0.8962**

zeta function (Riemann, 1859) $\zeta(z)$ evaluated at $z = \gamma$, and thus we call our method ζ-*mixup*. Since there exist $N!$ possible $N \times N$ random permutation matrices, given N original samples, ζ-*mixup* can synthesize $N!$ new samples for a single γ, unlike *mixup* which can only synthesize 1 new sample per sample pair for a single λ. Moreover, as a result of its formulation, ζ-*mixup* presents two desirable properties: **(1)** for all values of $\gamma \geq \gamma_{\min} = 1.72865$, the weight assigned to one sample is greater than the sum of weights assigned to all other samples, implicitly introducing the desired notion of linearity in only the locality of original samples; and **(2)** for $N = 2$ and $\gamma = \log_2\left(\frac{\lambda}{1-\lambda}\right)$, ζ-*mixup* simplifies to *mixup*.

2. Results and Discussion

Using a PCA-based local $\mathcal{D}_{\text{int}}$ estimator calculated using a k-nearest neighborhood around each sample, with $k = 128$ (Fukunaga and Olsen, 1971), we find that $\mathcal{D}_{\text{int}}$ for CIFAR-10 and CIFAR-100 using ζ-*mixup* are lower than using *mixup*: 26.83 ± 6.53 (versus 35.43 ± 9.47) and 24.76 ± 6.22 (versus 32.41 ± 8.65), respectively, thus showing that ζ-*mixup* indeed preserves the low $\mathcal{D}_{\text{int}}$ that natural image datasets lie in (Ruderman, 1994; Pope et al., 2021), while *mixup*'s off-manifold sampling leads to an inflated estimate of local $\mathcal{D}_{\text{int}}$. Tables 1 and 2 show the classification performance using traditional DA techniques, e.g., rotation, flipping, and cropping ("ERM"), against those trained with *mixup* and ζ-*mixup* outputs as well as compare the benefit of applying ζ-*mixup* to an orthogonal DA method, CutMix (Yun et al., 2019), as evaluated on natural: CIFAR-10 and CIFAR-100 and medical (skin lesion): ISIC 2016 (Gutman et al., 2016), 2017 (Codella et al., 2018), and 2018 (Codella et al., 2019), and DermoFit (Ballerini et al., 2013) image datasets. We report the error rate and the micro-averaged F1-score for natural and medical image datasets, respectively, since the latter are class-imbalanced. We observe that ζ-*mixup* improves performance across the board. Our optimized ζ-*mixup* implementation is $2.1\times$ faster than the original *mixup* implementation, while similar training time is recorded for both of them for CIFAR-10/100 ($\sim$ 1h 20m).

Conclusion: We proposed ζ-*mixup*, a parameter-free multi-sample generalization of the popular *mixup* technique for data augmentation that combines $N \geq 2$ samples without significant computational overhead. The ζ-*mixup* formulation allows for the weight assigned to one sample to dominate all the others, thus ensuring the synthesized samples are on or close to the original data manifold. This leads to generating samples that are more realistic and, along with allowing $N > 2$, generates more diverse samples with richer labels compared to *mixup*. Future work will include exploring ζ-*mixup* in the learned feature space.

Acknowledgments

The authors are grateful to StackOverflow user `obchardon` and Ashish Sinha for code optimization suggestions and to Saeid Asgari Taghanaki for initial discussions. The authors are also grateful for the computational resources provided by NVIDIA Corporation and Digital Research Alliance of Canada (formerly Compute Canada). Partial funding for this project was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. Springer, 2013.

James M Beale and Frank C Keil. Categorical effects in the perception of faces. *Cognition*, 57(3):217–239, 1995.

Colin J Brown, Steven P Miller, Brian G Booth, Kenneth J Poskitt, Vann Chau, Anne R Synnes, Jill G Zwicker, Ruth E Grunau, and Ghassan Hamarneh. Prediction of motor function in very preterm infants using connectome features and local synthetic instances. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 69–76. Springer, 2015.

Lawrence Cayton. Algorithms for manifold learning. *University of California at San Diego Technical Report*, 12(1-17):1, 2005.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Keinosuke Fukunaga and David R Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 100(2):176–183, 1971.

David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1605.01397*, 2016.

Fiona N Newell and Heinrich H Bülthoff. Categorical perception of familiar objects. *Cognition*, 85(2):113–143, 2002.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

Bernhard Riemann. Ueber die anzahl der primzahlen unter einer gegebenen grosse. *Ges. Math. Werke und Wissenschaftlicher Nachlaß*, 2(145-155):2, 1859.

Daniel L Ruderman. The statistics of natural images. *Network: computation in neural systems*, 5(4):517, 1994.

Erroll Wood. Synthetic data with digital humans. Microsoft Sponsor Session, CVPR 2021, 2021. URL https://www.microsoft.com/en-us/research/uploads/prod/2019/09/2019-10-01-Synthetic-Data-with-Digital-Humans.pdf.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3681–3691, 2021.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.