Infinite Class Mixup

Thomas Mensink Pascal Mettes

BMVC 2023

What is Mixup?

Generate additional samples by interpolating samples:

$$\tilde{\boldsymbol{x}} = \lambda\, \boldsymbol{x}_a + (1-\lambda)\, \boldsymbol{x}_b \qquad \tilde{\boldsymbol{y}} = \lambda\, \boldsymbol{y}_a + (1-\lambda)\, \boldsymbol{y}_b$$

Mixing of two images — Mixing of one-hot labels

- Easy and effective way to enrich any training set
- Improves accuracy, calibration & out-of-distribution generalization
- Assumption: mixing should be done at output probability level
- Mixup wants linearity, but its not enforced in classifier space!

Motivation

Image: *40% dog and 60% cat*

Mixup: *Classify as 40% dog 60% cat.*
Infinite Class Mixup: *Classify this image 100% to the new class: 40% dog and 60% cat.*

Mixup with Infinite Classes

Code available

Main idea: for each mixed sample, we construct a unique interpolated classifier:

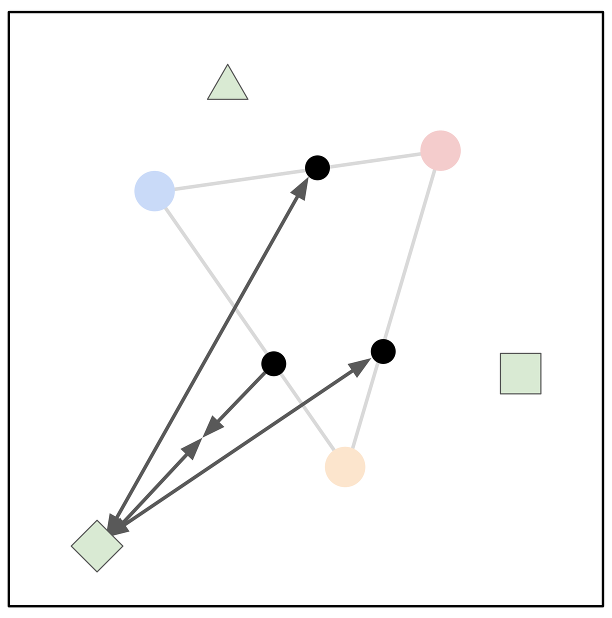
$$\boldsymbol{w}_{\tilde{c}} = \lambda\, \boldsymbol{w}_a + (1-\lambda)\, \boldsymbol{w}_b = W\tilde{\boldsymbol{y}}$$

Optimize likelihood of interpolated classes:

$$p(C_{\tilde{\boldsymbol{y}}}|\tilde{\boldsymbol{x}}) \propto \exp(f_\theta(\tilde{\boldsymbol{x}})^\top W\tilde{\boldsymbol{y}})$$

via contrastive learning in two ways:

Class-axis



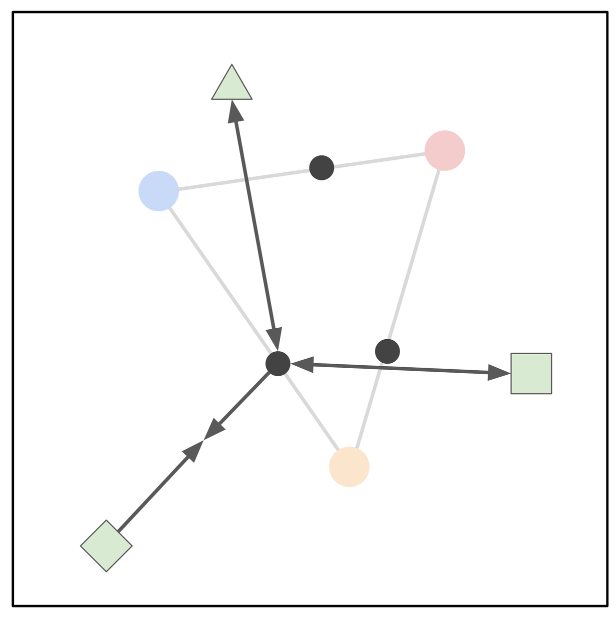
Contrast example against all interpolated classifiers in the batch.

In the spirit of conventional contrastive learning.

Pair-axis

Contrast classifier against all interpolated samples in the batch.

Exists because of the unique construction of each classifier. Not (directly) applicable in Mixup.



Optimization: $f_\theta(\tilde{\boldsymbol{x}})^\top W\tilde{\boldsymbol{y}}$ is a B x B matrix for batch size B, we perform cross-entropy loss across both axes.

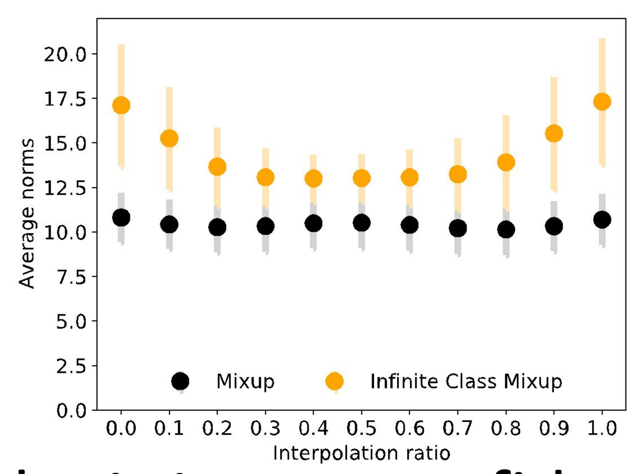
Experiments

contrastive axis		CIFAR-100 batch size		
class-axis	pair-axis	64	128	512
✓		74.90	76.75	76.17
	✓	75.38	77.62	76.09
✓	✓	**76.20**	**77.90**	**77.08**

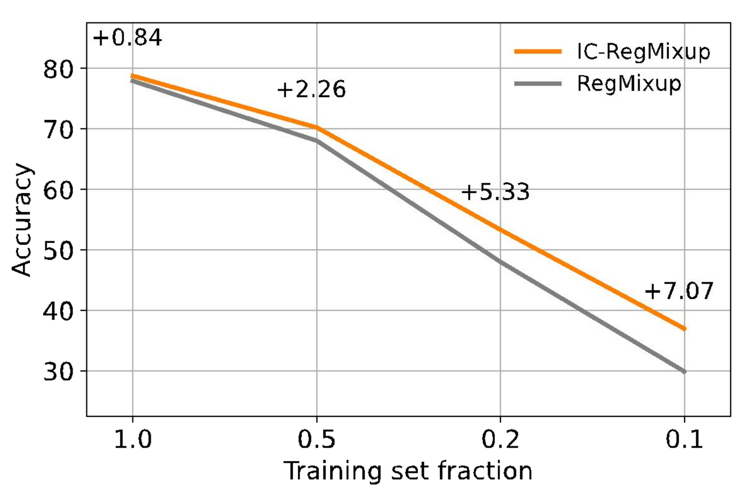
The class- and pair-axes are complementary.

	LT-CIFAR100 0.1	LT-CIFAR100 0.01	LT-CIFAR10 0.1	LT-CIFAR10 0.01
ERM	58.54	37.44	88.63	71.87
Mixup	62.68	39.21	89.63	72.82
IC-Mixup	**64.30**	**43.31**	**89.89**	**76.81**
	+1.62	+4.10	+0.26	+3.99
Remix	61.36	38.04	89.57	72.65
IC-Remix	**64.56**	**46.01**	**90.26**	**79.28**
	+3.20	+5.97	+0.67	+6.63

Improves long-tailed recognition.



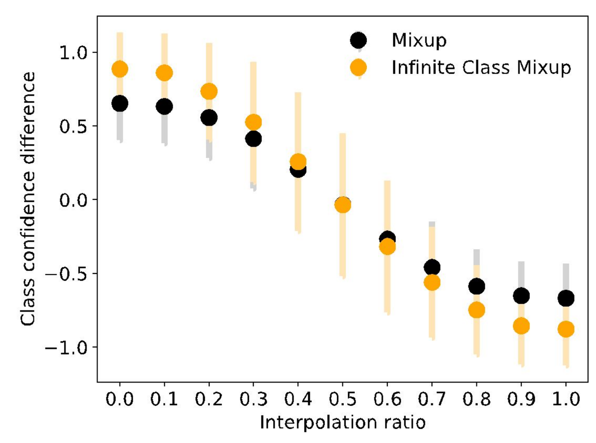
Insight 1: Lower confidence for ambiguous samples.



Enhances recent RegMixup (NeurIPS 2022).

	ciFAIR-10
Bietti et al. [4]	51.03
Oyallon et al. [40]	54.21
Kayhan and Gemert [22]	55.00
Ulicny et al. [49]	56.50
Kobayashi [27]	57.50
Brigato et al. [6]	58.22
+ IC-RegMixup	**61.84**

State-of-the-art in data-constrained learning.



Insight 2: Better differentiation between interpolations.