C-Mixup: Improving Generalization in Regression

Huaxiu Yao*[1], Yiping Wang*[2], Linjun Zhang[3], James Zou[1], Chelsea Finn[1]

[1]Stanford University, [2]Zhejiang University, [3]Rutgers University

tl;dr: a simple interpolation-based method (C-Mixup) to improve generalization on regression tasks by interpolating examples with closer labels

Background

Mixup in Deep Learning

A learning model

$$\mathcal{D}_{tr} = \{x_i, y_i\}_{i=1}^N \rightarrow \text{Classifier},$$

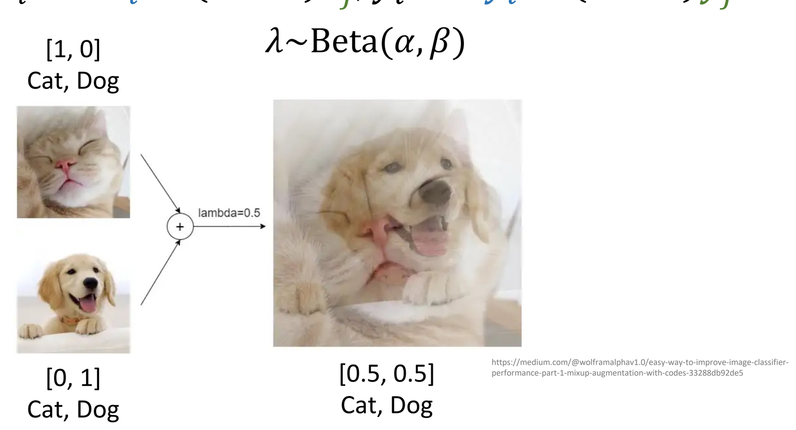
mixup[1]

$$\tilde{\mathcal{D}}_{tr} = \{\tilde{x}_i, \tilde{y}_i\}_{i=1}^N \rightarrow \text{Classifier},$$

where

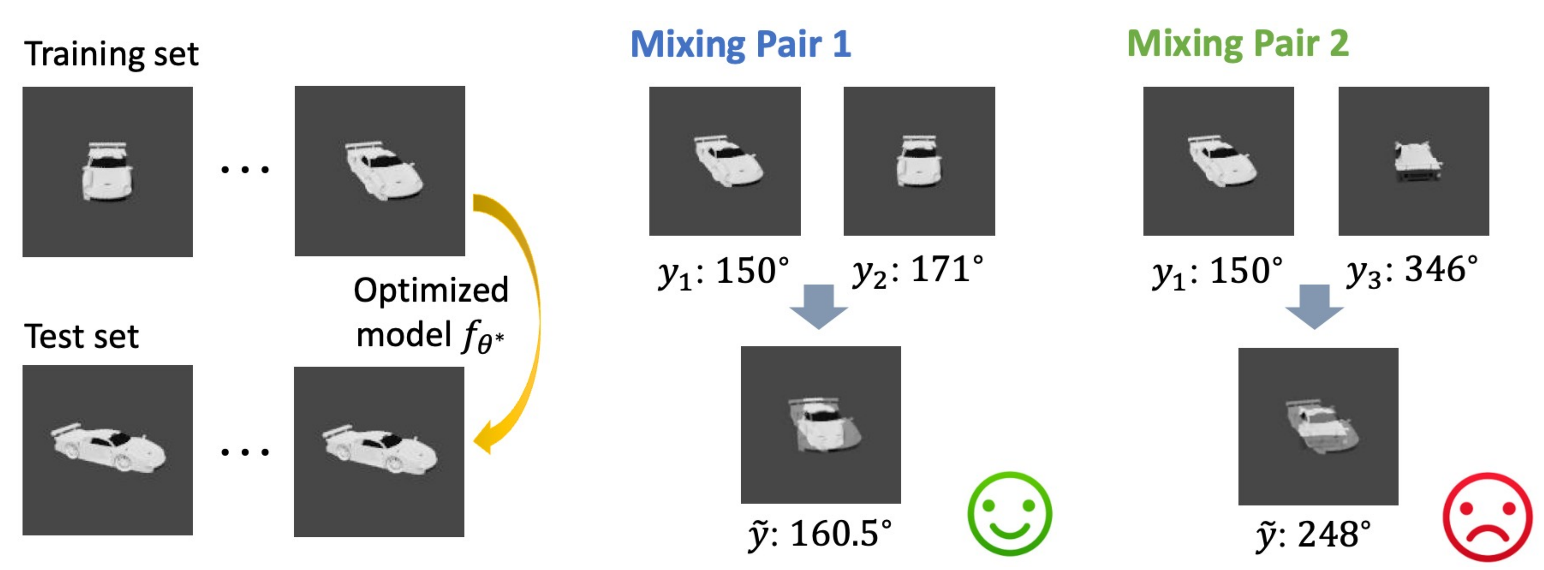
$$\tilde{x}_i = \lambda x_i + (1-\lambda) x_j, \tilde{y}_i = \lambda y_i + (1-\lambda) y_j$$

$$\lambda \sim \text{Beta}(\alpha, \beta)$$



Why mixup may Fail in Regression?

Directly applying mixup in Regression may produce arbitrary labels



Goal: building **interpolation-based models** to improve the **generalization in regression**

C-Mixup: Mixup for Regression

Key idea: interpolating examples with similar labels

Changing the sampling probability of mixing pairs

$$P((x_i, y_j) | (x_i, y_j)) \propto \exp(-\frac{d(i,j)}{2\sigma^2})$$

d: distance between examples i and j

Similar examples

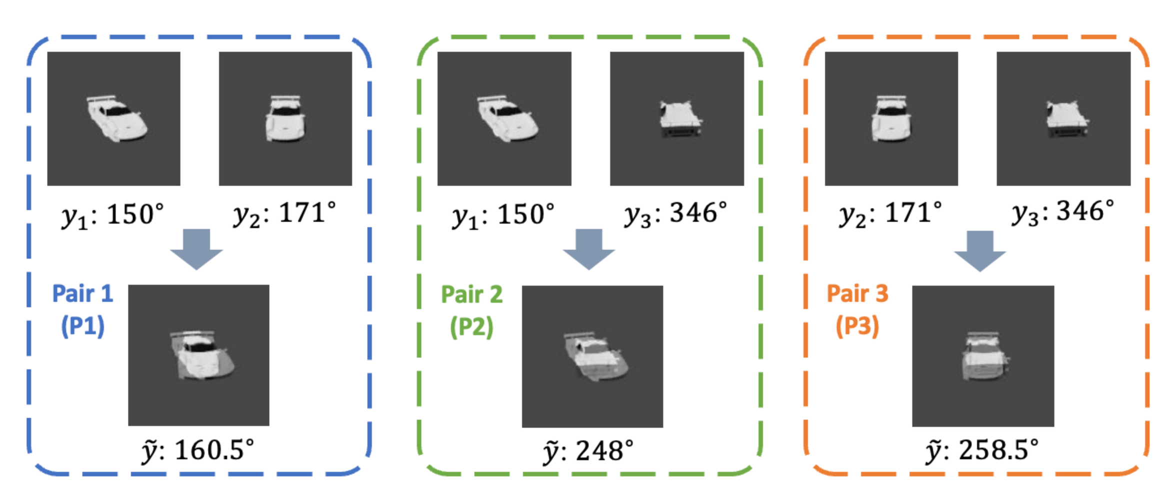
Higher probability to be mixed

Natural way: compute the distance using the input feature x

$$d(i,j) = d(x_i, x_j)$$

Drawbacks:

- Lacking good distance metrics to capture structured feature information
- Distance between features can be easily influenced by feature noise



Mixing Pairs ($\lambda = 0.5$)

Vanilla mixup
Prob(P1) = Prob(P2) = Prob(P3)

$d(i,j) = d(x_i, x_j)$
Prob(P1) ≈ Prob(P3) >> Prob(P2)

C-Mixup: $d(i,j) = d(y_i, y_j)$
Prob(P1) >> Prob(P2) > Prob(P3)

C-Mixup

Examples with closer labels ➡ Higher probability to be mixed

$$d(i,j) = d(y_i, y_j)$$

+ Benefit both in-distribution and out-of-distribution generalization
+ Calculating label distance is computationally efficient

Looking for **theoretical** support? ➡ mean square error: C-Mixup < min(mixup, $d(x_i, x_j)$)

(linear or monotonic non-linear model)

Experiments

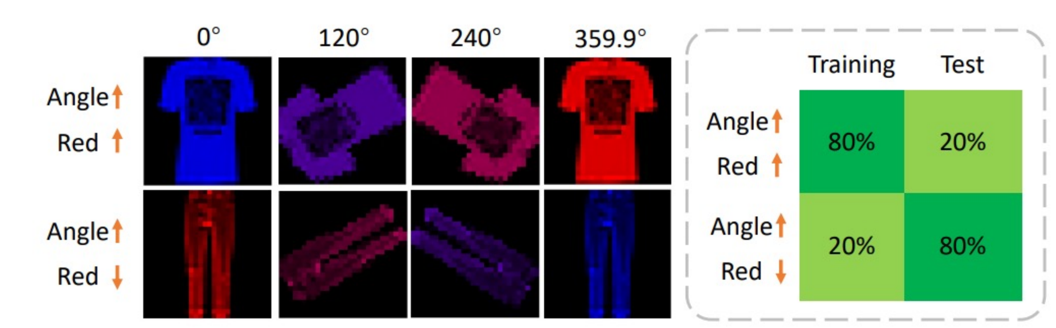
In-distribution Results

	Tabular				Time-series				Video	
	Airfoil		NO2		Exchange-Rate		Electricity		Echo	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
ERM	2.901	1.753%	0.537	13.615%	0.0236	2.423%	0.0581	13.861%	5.402	8.700%
mixup	3.730	2.327%	0.528	13.534%	0.0239	2.441%	0.0585	14.306%	5.393	8.838%
Mani mixup [2]	3.063	1.842%	0.522	13.382%	0.0242	2.475%	0.0583	14.556%	5.482	8.955%
k-Mixup	2.938	1.769%	0.519	13.173%	0.0236	2.403%	0.0575	14.134%	5.518	9.206%
Local Mixup	3.703	2.290%	0.517	13.202%	0.0236	2.341%	0.0582	14.245%	5.652	9.313%
MixRL	3.614	2.163%	0.527	13.298%	0.0238	2.397%	0.0585	14.417%	5.618	9.165%
C-Mixup (Ours)	2.717	1.610%	0.509	12.998%	0.0203	2.041%	0.0570	13.372%	5.177	8.435%

Out-of-Distribution Results

Subpopulation shift: mitigate spurious correlation

RCF-MNIST



Domain shift: generalize to new domains

PovertyMap [3]

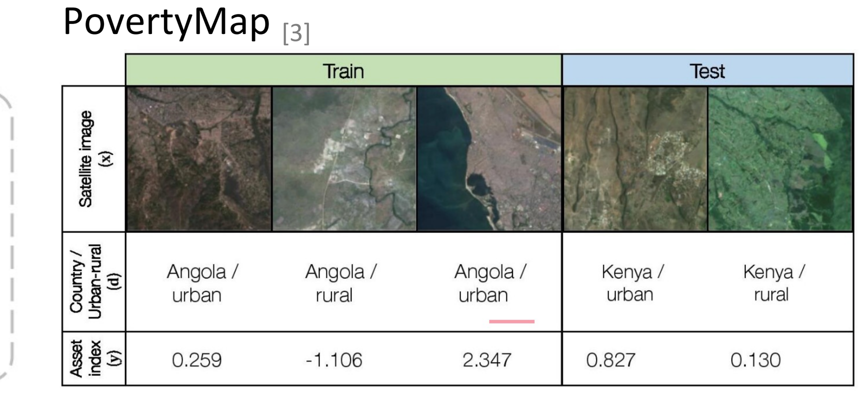


	Image	Tabular							Drug	
	Sub. Shift	Domain Shift								
	RCF-MNIST	PovertyMap (R)		Crime (RMSE)		SkillCraft (RMSE)		DTI (R)		
	Avg. (RMSE) ↓	Avg. ↑	Worst ↑	Avg. ↓	Worst ↓	Avg. ↓	Worst ↓	Avg. ↑	Worst ↑	
ERM	0.162	0.80	0.50	0.134	0.173	5.887	10.182	0.464	0.429	
IRM	0.153	0.77	0.43	0.127	0.155	5.937	7.849	0.478	0.432	
IB-IRM	0.167	0.78	0.40	0.127	0.153	6.055	7.650	0.479	0.435	
V-REx	0.154	**0.83**	0.48	0.129	0.157	6.059	7.444	0.485	0.435	
CORAL	0.163	0.78	0.44	0.133	0.166	6.353	8.272	0.483	0.432	
GroupDRO	0.232	0.75	0.39	0.138	0.168	6.155	8.131	0.442	0.407	
Fish	0.263	0.80	0.30	0.128	0.152	6.356	8.676	0.470	0.443	
mixup	0.176	0.81	0.46	0.128	0.154	5.764	9.206	0.465	0.437	
C-Mixup (Ours)	**0.146**	0.81	**0.53**	**0.123**	**0.146**	**5.201**	**7.362**	**0.498**	**0.458**	

Analysis

I. Different distance metrics

	DTI
Feature/Repr. Distance	0.477
C-Mixup	**0.498 ↑**

II. Scalability: batch-wise C-Mixup

Apply C-Mixup on every batch

	Dataset	Airfoil	NO2	Exchange-Rate	Electricity
RMSE ↓	C-Mixup-batch	2.792 ± 0.135	0.510 ± 0.007	0.0205 ± 0.0017	0.0576 ± 0.0002
	C-Mixup	2.717 ± 0.067	0.509 ± 0.006	0.0203 ± 0.0011	0.0570 ± 0.0006
MAPE ↓	C-Mixup-batch	$1.616 \pm 0.053\%$	$12.894 \pm 0.180\%$	$2.064 \pm 0.218\%$	$13.697 \pm 0.155\%$
	C-Mixup	$1.610 \pm 0.085\%$	$12.998 \pm 0.271\%$	$2.041 \pm 0.134\%$	$13.372 \pm 0.106\%$

References

[1] Zhang, Hongyi, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).

[2] Verma, Vikas, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. "Manifold mixup: Better representations by interpolating hidden states." In International Conference on Machine Learning, PMLR, 2019.

[3] Koh, Pang Wei, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu et al. "Wilds: A benchmark of in-the-wild distribution shifts." *ICML 2021.*