TUE-PM-333

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

On the Pitfall of Mixup for Uncertainty Calibration

Deng-Bao Wang, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, Min-Ling Zhang

Southeast University, Tencent, CUHK

Outline

- Background
- Mixup for Calibration: Does It Really Work?
- How to Fix Mixup's Issue?
- Experiments

Outline

- Background   Uncertainty Calibration
- Mixup for Calibration: Does It Really Work?
- How to Fix Mixup's Issue?
- Experiments

Outline

- Background Uncertainty Calibration
- Mixup for Calibration: Does It Really Work? ✕
- How to Fix Mixup's Issue?
- Experiments

Outline

- ☐ Background Uncertainty Calibration
- ☐ Mixup for Calibration: Does It Really Work? ✗
- ☐ How to Fix Mixup's Issue? The MIT Approach
- ☐ Experiments

Outline

- ☐ Background   Uncertainty Calibration
- ☐ Mixup for Calibration: Does It Really Work?
- ☐ How to Fix Mixup's Issue?   The MIT Approach
- ☐ Experiments

Background

Real-world Applications of Calibration

Autonomous driving

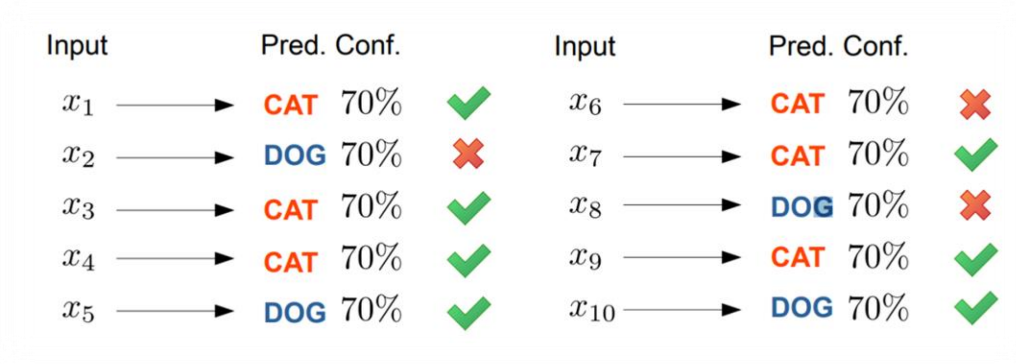
Smart healthcare

Uncertainty Calibration is important for Safety-aware Scenarios

Background

Intuitive explanation:

The average confidence reflects models's accuracy



Formally:

A perfect classifier satisfies: $\mathbb{P}(\hat{y} = y | \hat{p} = p) = p$

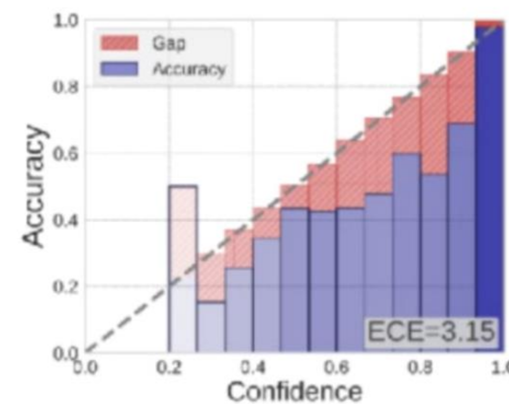
Background

Intuitive explanation:

The *average confidence* *reflects models's* *accuracy*

Input	Pred.	Conf.		Input	Pred.	Conf.	
x_1 →	CAT	70%	✓	x_6 →	CAT	70%	✗
x_2 →	DOG	70%	✗	x_7 →	CAT	70%	✓
x_3 →	CAT	70%	✓	x_8 →	DOG	70%	✗
x_4 →	CAT	70%	✓	x_9 →	CAT	70%	✓
x_5 →	DOG	70%	✓	x_{10} →	DOG	70%	✓

In a real case



Formally:

A *perfect classifier* *satisfys:* $\mathbb{P}(\hat{y} = y | \hat{p} = p) = p$

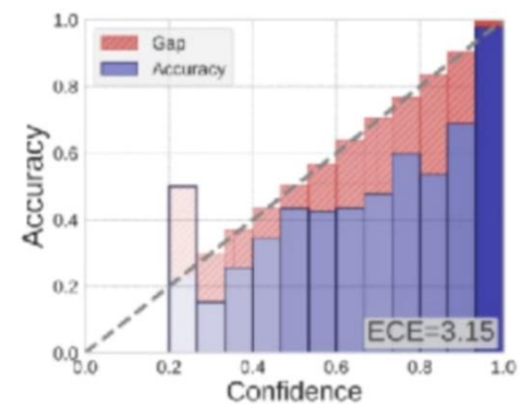
Background

Intuitive explanation:

The average confidence reflects models's accuracy

Input		Pred.	Conf.		Input		Pred.	Conf.	
x_1	$\longrightarrow$	CAT	70%	✓	x_6	$\longrightarrow$	CAT	70%	✗
x_2	$\longrightarrow$	DOG	70%	✗	x_7	$\longrightarrow$	CAT	70%	✓
x_3	$\longrightarrow$	CAT	70%	✓	x_8	$\longrightarrow$	DOG	70%	✗
x_4	$\longrightarrow$	CAT	70%	✓	x_9	$\longrightarrow$	CAT	70%	✓
x_5	$\longrightarrow$	DOG	70%	✓	x_{10}	$\longrightarrow$	DOG	70%	✓

In a real case



Formally:

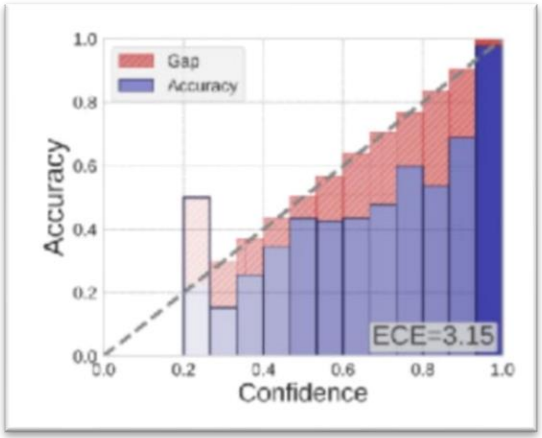
A perfect classifier satisfys: $\mathbb{P}(\hat{y} = y | \hat{p} = p) = p$

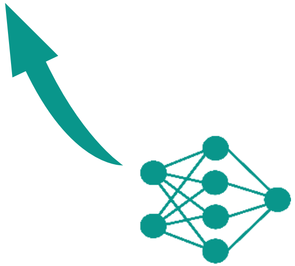
Evaluation Metric: $ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$

Background

Solve the calibration issue: Post-hoc Calibration

raw outputs

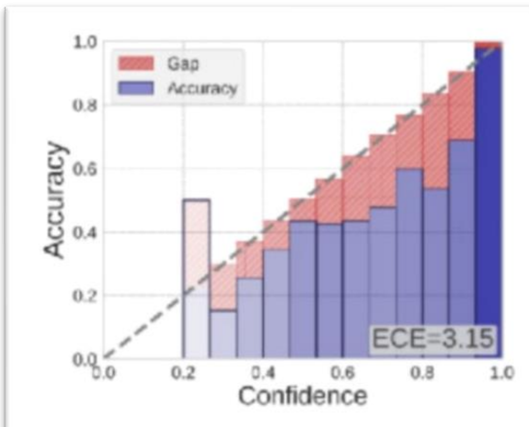


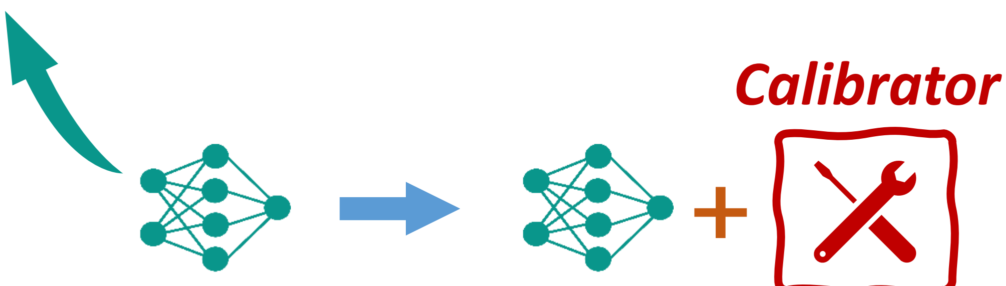


Background

Solve the calibration issue: Post-hoc Calibration

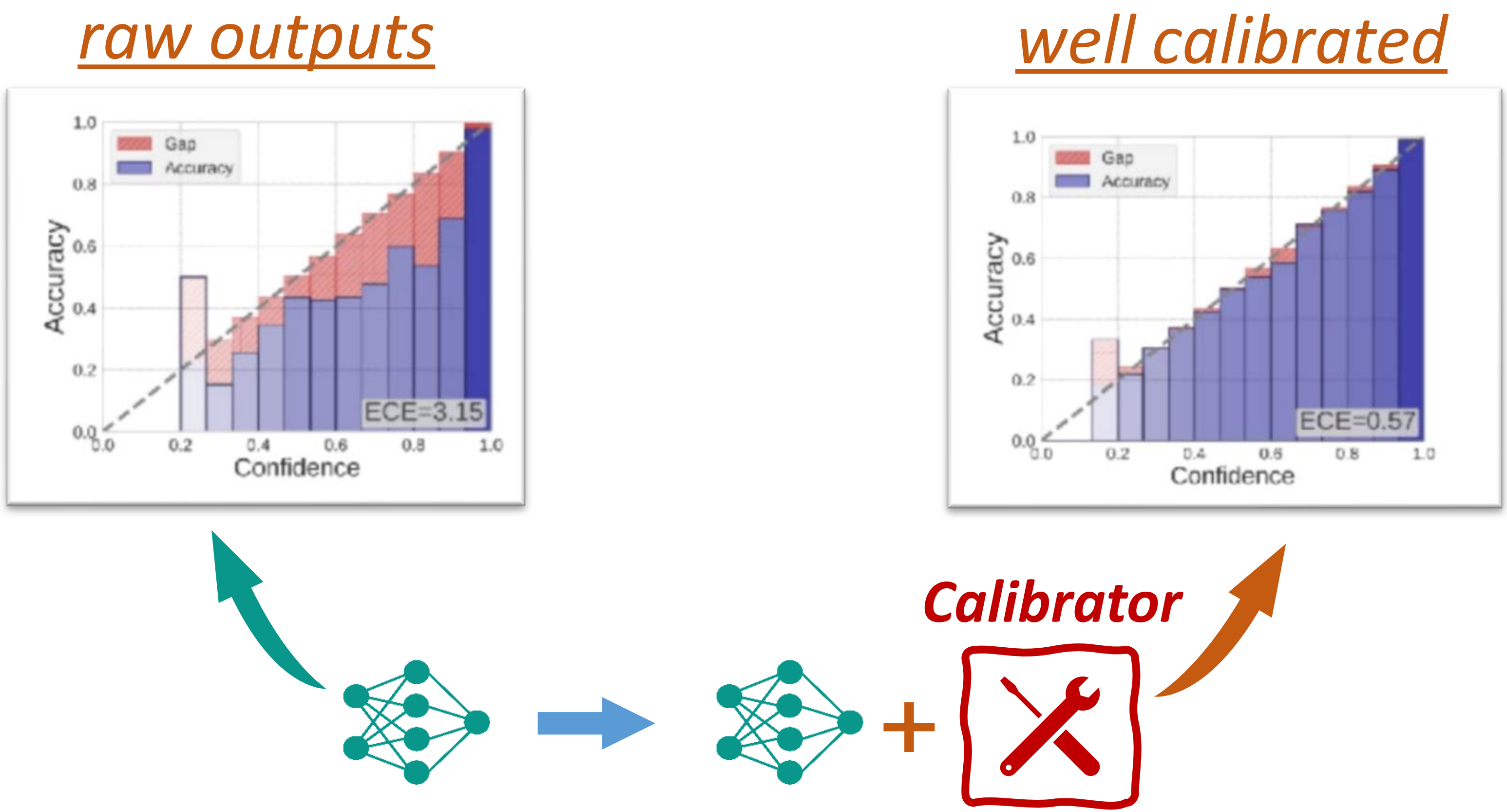
raw outputs





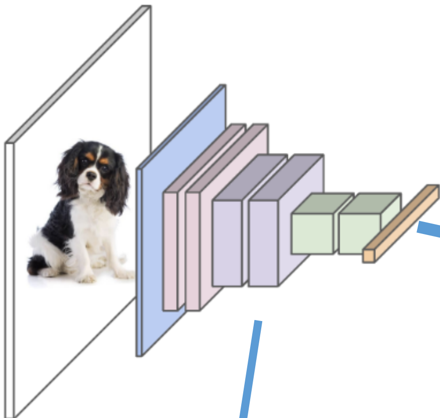
Background

Solve the calibration issue: Post-hoc Calibration



Background

Temperature Scaling



$$P_i = \frac{e^{\boxed{\frac{y_i}{T}}}}{\sum_{k=1}^{n} e^{\boxed{\frac{y_k}{T}}}}$$

Model trained on training set

Find the best T in validation set

Temperature Scaling (*TS*) achieves surprising performance

Outline

- ◻ Background
- ◻ **Mixup for Calibration: Does It Really Work?**
- ◻ How to Fix Mixup's Issue?
- ◻ Experiments

Mixup for Calibration: Does It Really Work?

[1] On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks, *NeurIPS 2020*

They empirically find that DNNs trained with mixup are significantly better calibrated.

[2] When and How Mixup Improves Calibration, *ICML 2022*

They theoretically prove that Mixup improves calibration in high-dimensional settings.

[3] Combining Ensembles and Data Augmentation Can Harm Your Calibration, *ICLR 2021*

Mixup may hurt calibration in some cases!

Mixup for Calibration: Does It Really Work?

[1] On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks, *NeurIPS 2020*

They empirically find that DNNs trained with mixup are significantly better calibrated.

[2] When and How Mixup Improves Calibration, *ICML 2022*

They theoretically prove that Mixup improves calibration in high-dimensional settings.

[3] Combining Ensembles and Data Augmentation Can Harm Your Calibration, *ICLR 2021*

Mixup may hurt calibration in some cases!

[4] Pitfalls of in-domain uncertainty estimation and ensembling in deep learning, *ICLR 2019*

Comparison without post-hoc calibration might be not fair.

 +  **Does mixup really improve Calibration?**

Mixup for Calibration: Does It Really Work?

$$P_i = \frac{e^{\boxed{\frac{y_i}{T}}}}{\sum_{k=1}^{n} e^{\boxed{\frac{y_k}{T}}}}$$

Temperature Scaling (TS)

Datasets	Metrics	ERM				mixup ($\alpha=0.1$)				mixup ($\alpha=0.5$)				mixup ($\alpha=1.0$)			
SVHN	ECE	2.15	2.67	2.43	2.56	3.96	1.89▲	2.46	1.38▲	11.2	9.48	8.04	9.37	14.8	13.7	13.8	12.9
	Calibrated ECE	0.50	0.87	0.75	0.90	0.99▼	1.03▼	1.08▼	1.05▼	1.23▼	1.21▼	1.28▼	1.21▼	1.12▼	1.18▼	1.14▼	1.04▼
	Optimal ECE	0.24	0.56	0.45	0.58	0.75▼	0.74▼	0.85▼	0.68▼	1.12▼	0.95▼	0.95▼	0.88▼	1.04▼	0.98▼	0.88▼	0.78▼
CIFAR-10	ECE	3.33	3.99	3.78	3.47	2.57▲	2.22▲	2.55▲	2.53▲	6.87	6.25	6.55	6.20	12.1	11.5	10.5	11.2
	Calibrated ECE	0.65	0.79	0.83	0.65	1.04▼	1.07▼	1.08▼	1.12▼	1.15▼	1.15▼	0.95▼	1.05▼	0.94▼	0.91▼	0.83	0.76▼
	Optimal ECE	0.59	0.63	0.61	0.52	0.97▼	0.98▼	1.01▼	1.01▼	0.97▼	1.03▼	0.88▼	0.88▼	0.85▼	0.80▼	0.71▼	0.65▼
CIFAR-100	ECE	10.9	12.5	11.9	11.7	2.43▲	6.63▲	5.95▲	5.59▲	10.8▲	3.89▲	3.91▲	3.85▲	13.0	7.44▲	7.50▲	7.55▲
	Calibrated ECE	2.56	2.41	2.64	2.42	1.76	1.87	1.37	1.67	1.22	2.63▼	3.21▼	2.57▼	1.25	2.66▼	3.02▼	3.52▼
	Optimal ECE	2.45	2.29	2.44	2.31	1.60	1.59	1.23	1.45	0.98	2.46▼	3.04▼	2.39▼	1.09	2.54▼	2.85▼	3.38▼
Tiny-ImageNet	ECE	23.2	20.5	20.7	21.6	8.57▲	7.51▲	9.76▲	10.4▲	3.98▲	3.92▲	2.16▲	3.29▲	6.85▲	7.44▲	4.98▲	5.93▲
	Calibrated ECE	1.33	1.23	1.36	1.33	1.32	1.28▼	1.55▼	2.08▼	1.33	1.46▼	1.52▼	1.82▼	1.49▼	1.65▼	2.26▼	2.00▼
	Optimal ECE	1.14	1.00	1.16	1.16	1.02	1.05▼	1.40▼	1.93▼	1.08	1.21▼	1.23▼	1.60▼	1.20▼	1.30▼	1.91▼	1.69▼

Mixup for Calibration: Does It Really Work?

$$P_i = \frac{e^{\boxed{\frac{y_i}{T}}}}{\sum_{k=1}^{n} e^{\boxed{\frac{y_k}{T}}}}$$

Temperature Scaling (TS)

Datasets	Metrics	ERM				mixup ($\alpha = 0.1$)				mixup ($\alpha = 0.5$)				mixup ($\alpha = 1.0$)			
SVHN	ECE	2.15	2.67	2.43	2.56	3.96	1.89▲	2.46	1.38▲	11.2	9.48	8.04	9.37	14.8	13.7	13.8	12.9
	Calibrated ECE	0.50	0.87	0.75	0.90	0.99▼	1.03▼	1.08▼	1.05▼	1.23▼	1.21▼	1.28▼	1.21▼	1.12▼	1.18▼	1.14▼	1.04▼
	Optimal ECE	0.24	0.56	0.45	0.58	0.75▼	0.74▼	0.85▼	0.68▼	1.12▼	0.95▼	0.95▼	0.88▼	1.04▼	0.98▼	0.88▼	0.78▼
CIFAR-10	ECE	3.33	3.99	3.78	3.47	2.57▲	2.22▲	2.55▲	2.53▲	6.87	6.25	6.55	6.20	12.1	11.5	10.5	11.2
	Calibrated ECE	0.65	0.79	0.83	0.65	1.04▼	1.07▼	1.08▼	1.12▼	1.15▼	1.15▼	0.95▼	1.05▼	0.94▼	0.91▼	0.83	0.76▼
	Optimal ECE	0.59	0.63	0.61	0.52	0.97▼	0.98▼	1.01▼	1.01▼	0.97▼	1.03▼	0.88▼	0.88▼	0.85▼	0.80▼	0.71▼	0.65▼
CIFAR-100	ECE	10.9	12.5	11.9	11.7	2.43▲	6.63▲	5.95▲	5.59▲	10.8▲	3.89▲	3.91▲	3.85▲	13.0	7.44▲	7.50▲	7.55▲
	Calibrated ECE	2.56	2.41	2.64	2.42	1.76	1.87	1.37	1.67	1.22	2.63▼	3.21▼	2.57▼	1.25	2.66▼	3.02▼	3.52▼
	Optimal ECE	2.45	2.29	2.44	2.31	1.60	1.59	1.23	1.45	0.98	2.46▼	3.04▼	2.39▼	1.09	2.54▼	2.85▼	3.38▼
Tiny-ImageNet	ECE	23.2	20.5	20.7	21.6	8.57▲	7.51▲	9.76▲	10.4▲	3.98▲	3.92▲	2.16▲	3.29▲	6.85▲	7.44▲	4.98▲	5.93▲
	Calibrated ECE	1.33	1.23	1.36	1.33	1.32	1.28▼	1.55▼	2.08▼	1.33	1.46▼	1.52▼	1.82▼	1.49▼	1.65▼	2.26▼	2.00▼
	Optimal ECE	1.14	1.00	1.16	1.16	1.02	1.05▼	1.40▼	1.93▼	1.08	1.21▼	1.23▼	1.60▼	1.20▼	1.30▼	1.91▼	1.69▼

Without TS

Mixup for Calibration: Does It Really Work?

$$P_i = \frac{e^{\boxed{\frac{y_i}{T}}}}{\sum_{k=1}^{n} e^{\boxed{\frac{y_k}{T}}}}$$

Temperature Scaling (TS)

Datasets	Metrics	ERM				mixup ($\alpha=0.1$)				mixup ($\alpha=0.5$)				mixup ($\alpha=1.0$)			
SVHN	ECE	2.15	2.67	2.43	2.56	3.96	1.89▲	2.46	1.38▲	11.2	9.48	8.04	9.37	14.8	13.7	13.8	12.9
	Calibrated ECE	0.50	0.87	0.75	0.90	0.99▼	1.03▼	1.08▼	1.05▼	1.23▼	1.21▼	1.28▼	1.21▼	1.12▼	1.18▼	1.14▼	1.04▼
	Optimal ECE	0.24	0.56	0.45	0.58	0.75▼	0.74▼	0.85▼	0.68▼	1.12▼	0.95▼	0.95▼	0.88▼	1.04▼	0.98▼	0.88▼	0.78▼
CIFAR-10	ECE	3.33	3.99	3.78	3.47	2.57▲	2.22▲	2.55▲	2.53▲	6.87	6.25	6.55	6.20	12.1	11.5	10.5	11.2
	Calibrated ECE	0.65	0.79	0.83	0.65	1.04▼	1.07▼	1.08▼	1.12▼	1.15▼	1.15▼	0.95▼	1.05▼	0.94▼	0.91▼	0.83	0.76▼
	Optimal ECE	0.59	0.63	0.61	0.52	0.97▼	0.98▼	1.01▼	1.01▼	0.97▼	1.03▼	0.88▼	0.88▼	0.85▼	0.80▼	0.71▼	0.65▼
CIFAR-100	ECE	10.9	12.5	11.9	11.7	2.43▲	6.63▲	5.95▲	5.59▲	10.8▲	3.89▲	3.91▲	3.85▲	13.0	7.44▲	7.50▲	7.55▲
	Calibrated ECE	2.56	2.41	2.64	2.42	1.76	1.87	1.37	1.67	1.22	2.63▼	3.21▼	2.57▼	1.25	2.66▼	3.02▼	3.52▼
	Optimal ECE	2.45	2.29	2.44	2.31	1.60	1.59	1.23	1.45	0.98	2.46▼	3.04▼	2.39▼	1.09	2.54▼	2.85▼	3.38▼
Tiny-ImageNet	ECE	23.2	20.5	20.7	21.6	8.57▲	7.51▲	9.76▲	10.4▲	3.98▲	3.92▲	2.16▲	3.29▲	6.85▲	7.44▲	4.98▲	5.93▲
	Calibrated ECE	1.33	1.23	1.36	1.33	1.32	1.28▼	1.55▼	2.08▼	1.33	1.46▼	1.52▼	1.82▼	1.49▼	1.65▼	2.26▼	2.00▼
	Optimal ECE	1.14	1.00	1.16	1.16	1.02	1.05▼	1.40▼	1.93▼	1.08	1.21▼	1.23▼	1.60▼	1.20▼	1.30▼	1.91▼	1.69▼

With TS

Outline

- Background
- Mixup for Calibration: Does It Really Work?
- **How to Fix Mixup's Issue?**
- Experiments

Mixup for Calibration: Does It Really Work?

Remark 1. [3] *Let* $\lambda \sim \mathrm{Beta}_{[\frac{1}{2},1]}(\alpha, \alpha)$ *and* $j \sim$ *Uniform* $([n])$ *be two random variables with* $\alpha > 0$, $n > 0$ *and let* $\bar{\lambda} = \mathbb{E}_{\lambda}\lambda$. *The mixed sample* $(\widetilde{x}_i, \widetilde{y}_i)$ *as in Equaton (1) for any* $i \in [n]$ *can be reformulated as:*

$$\begin{aligned} \widetilde{x}_i &= \bar{x} + \bar{\lambda}(x_i - \bar{x}) + (\lambda - \bar{\lambda})x_i + (1-\lambda)x_j - (1-\bar{\lambda})\bar{x}, \\ \widetilde{y}_i &= \underbrace{\bar{y} + \bar{\lambda}(y_i - \bar{y})}_{\text{Data Transformation } x'_i, y'_i} + \underbrace{(\lambda - \bar{\lambda})y_i + (1-\lambda)y_j - (1-\bar{\lambda})\bar{y}}_{\text{Random Perturbation } \epsilon_i^x, \epsilon_i^y}, \end{aligned} \quad (2)$$

where $\bar{x}, \bar{y}$ *are the mean of inputs and labels of all training samples, and the perturbation terms satisfy* $\mathbb{E}_{\lambda,j}\epsilon_i^x = \mathbb{E}_{\lambda,j}\epsilon_i^y = 0$.

Mixup

↓

Data Augmentation

+

Label Smoothing

Label smoothing hurts post-hoc calibration! ***[4]***

[3] On mixup regularization, JMLR 2022

[4] Rethinking Calibration of Deep Neural Networks: Do Not Be Afraid of Overconfidence, NeurIPS 2021

Mixup for Calibration: Does It Really Work?

Remark 1. [3] *Let* $\lambda \sim \mathrm{Beta}_{[\frac{1}{2},1]}(\alpha,\alpha)$ *and* $j \sim$ *Uniform* $([n])$ *be two random variables with* $\alpha > 0$, $n > 0$ *and let* $\bar{\lambda} = \mathbb{E}_\lambda \lambda$. *The mixed sample* $(\widetilde{x}_i, \widetilde{y}_i)$ *as in Equaton (1) for any* $i \in [n]$ *can be reformulated as:*

$$\begin{aligned}\widetilde{x}_i &= \bar{x}+\bar{\lambda}(x_i-\bar{x}) + (\lambda-\bar{\lambda})x_i+(1-\lambda)x_j-(1-\bar{\lambda})\bar{x},\\ \widetilde{y}_i &= \underbrace{\bar{y}+\bar{\lambda}(y_i-\bar{y})}_{\text{Data Transformation } x'_i, y'_i} + \underbrace{(\lambda-\bar{\lambda})y_i+(1-\lambda)y_j-(1-\bar{\lambda})\bar{y}}_{\text{Random Perturbation } \epsilon_i^x, \epsilon_i^y},\end{aligned} \tag{2}$$

where $\bar{x}, \bar{y}$ *are the mean of inputs and labels of all training samples, and the perturbation terms satisfy* $\mathbb{E}_{\lambda,j}\epsilon_i^x = \mathbb{E}_{\lambda,j}\epsilon_i^y = 0$.

Mixup

Data Augmentation

+

Label Smoothing

Label smoothing hurts post-hoc calibration! ***[4]***

[3] On mixup regularization, JMLR 2022

[4] Rethinking Calibration of Deep Neural Networks: Do Not Be Afraid of Overconfidence, NeurIPS 2021

Mixup for Calibration: Does It Really Work?

Remark 1. [3] *Let* $\lambda \sim \mathrm{Beta}_{[\frac{1}{2},1]}(\alpha,\alpha)$ *and* $j \sim$ *Uniform* $([n])$ *be two random variables with* $\alpha > 0$, $n > 0$ *and let* $\bar{\lambda} = \mathbb{E}_\lambda \lambda$. *The mixed sample* $(\widetilde{x}_i, \widetilde{y}_i)$ *as in Equaton (1) for any* $i \in [n]$ *can be reformulated as:*

$$\begin{aligned}\widetilde{x}_i &= \underbrace{\bar{x}+\bar{\lambda}(x_i-\bar{x})}_{\text{Data Transformation } x'_i,\,y'_i} + \underbrace{(\lambda-\bar{\lambda})x_i+(1-\lambda)x_j-(1-\bar{\lambda})\bar{x}}_{\text{Random Perturbation } \epsilon^x_i,\,\epsilon^y_i}, \\ \widetilde{y}_i &= \bar{y}+\bar{\lambda}(y_i-\bar{y}) + (\lambda-\bar{\lambda})y_i+(1-\lambda)y_j-(1-\bar{\lambda})\bar{y},\end{aligned} \tag{2}$$

where $\bar{x}, \bar{y}$ *are the mean of inputs and labels of all training samples, and the perturbation terms satisfy* $\mathbb{E}_{\lambda,j}\epsilon^x_i = \mathbb{E}_{\lambda,j}\epsilon^y_i = 0$.

Mixup

↓

Data Augmentation

+

Label Smoothing

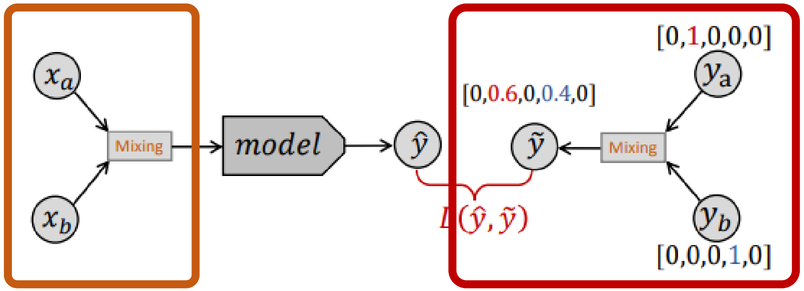
Label smoothing hurts post-hoc calibration! ***[4]***

[3] On mixup regularization, JMLR 2022

[4] Rethinking Calibration of Deep Neural Networks: Do Not Be Afraid of Overconfidence, NeurIPS 2021

How to Fix Mixup's Issue?

Avoiding label smoothing, but remaining data augmentation



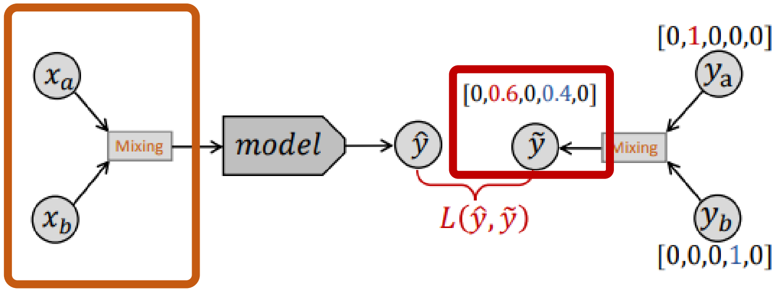
Data augmentation

(a) Mixup

Label smoothing

How to Fix Mixup's Issue?

Avoiding label smoothing, but remaining data augmentation



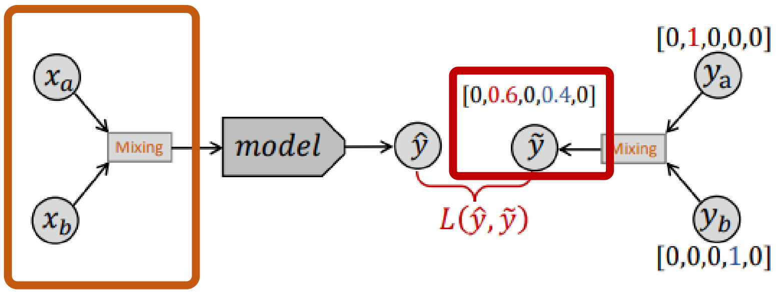
(a) Mixup

Data augmentation

Label smoothing

How to Fix Mixup’s Issue?

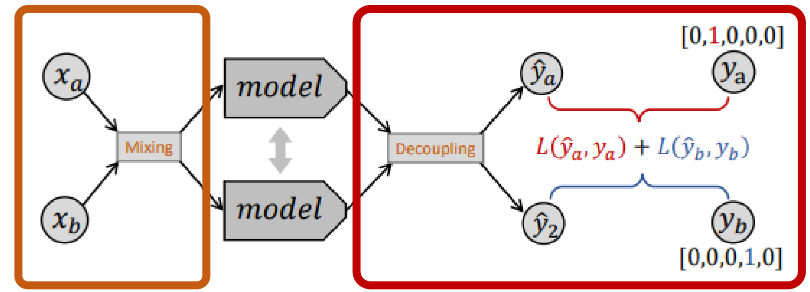
Avoiding label smoothing, but remaining data augmentation



(a) Mixup

Data augmentation

Label smoothing



(b) MIT-L

~~Label smoothing~~

Decoupling

How to Fix Mixup's Issue?

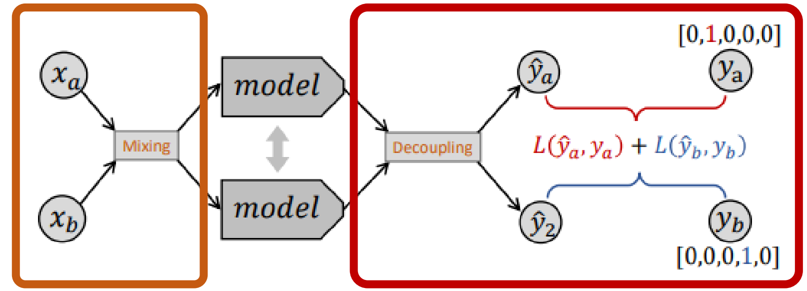
Avoiding label smoothing, but remaining data augmentation

Remark 2. *Recall the basic idea of mixup: linear interpolations of feature vectors should lead to linear interpolations of the output space. Based on this assumption, by mixing two samples twice with $\lambda_1 \neq \lambda_2 \in (0,1)$, as is*

$$\begin{aligned} \widetilde{x}_1 &= \lambda_1 x_a + (1-\lambda_1) x_b, \\ \widetilde{x}_2 &= \lambda_2 x_a + (1-\lambda_2) x_b, \end{aligned} \tag{3}$$

we can decouple these two samples in outputs space:

$$\begin{aligned} \widehat{y}_a &= \frac{f(\widetilde{x}_1) - f(\widetilde{x}_2)(1-\lambda_1)/(1-\lambda_2)}{\lambda_1 - \lambda_2(1-\lambda_1)/(1-\lambda_2)}, \\ \widehat{y}_b &= \frac{f(\widetilde{x}_1) - f(\widetilde{x}_2)\lambda_2/\lambda_1}{1-\lambda_2-(1-\lambda_1)\lambda_2/\lambda_1}. \end{aligned} \tag{4}$$



(b) MIT-L

~~Label smoothing~~

Decoupling

Outline

- Background
- Mixup for Calibration: Does It Really Work?
- How to Fix Mixup's Issue?
- **Experiments**

Experiments

ECE ↓

	Backbones	ERM	Mixup (0.1)	Mixup (0.5)	Mixup (1.0)	Mixup (DT)	Mixup (TO)	Mixup (SC)	Mixup (IO)	**MIT-A**	**MIT-L** ($\Delta\lambda > \frac{1}{2}$)	**MIT-A** ($\Delta\lambda > \frac{1}{2}$)
SVHN	ResNet18	0.50 (2)	0.99 (7)	1.23 (11)	1.12 (10)	1.01 (8)	1.05 (9)	0.50 (3)	0.61 (5)	**0.47 (1)**	0.65 (6)	0.53 (4)
	ResNet50	0.87 (6)	1.03 (7)	1.21 (9)	1.18 (8)	1.55 (11)	1.39 (10)	0.59 (4)	0.50 (2)	**0.49 (1)**	0.52 (3)	0.66 (5)
	ResNet110	0.75 (6)	1.08 (7)	1.28 (9)	1.14 (8)	1.39 (10)	1.43 (11)	0.50 (2)	0.60 (4)	**0.48 (1)**	0.53 (3)	0.70 (5)
	ResNet152	0.90 (6)	1.05 (8)	1.21 (9)	1.04 (7)	1.28 (10)	1.37 (11)	0.57 (2)	0.65 (4)	0.61 (3)	**0.53 (1)**	0.67 (5)
	Avg. gain	—	+ 0.28	+ 0.47	+ 0.36	+ 0.55	+ 0.55	– 0.21	– 0.16	– 0.24	– 0.19	– 0.11
CIFAR-10	ResNet18	0.65 (6)	1.04 (8)	1.15 (9)	0.94 (7)	1.70 (11)	1.45 (10)	0.62 (4)	0.61 (3)	**0.56 (1)**	0.59 (2)	0.62 (5)
	ResNet50	0.79 (6)	1.07 (8)	1.15 (9)	0.91 (7)	1.81 (11)	1.64 (10)	0.65 (4)	**0.46 (1)**	0.63 (3)	0.59 (2)	0.68 (5)
	ResNet110	0.83 (7)	1.08 (9)	0.95 (8)	0.83 (6)	1.52 (10)	1.56 (11)	0.54 (3)	**0.50 (1)**	0.52 (2)	0.54 (4)	0.78 (5)
	ResNet152	0.65 (4)	1.12 (9)	1.05 (8)	0.76 (7)	1.55 (11)	1.42 (10)	0.67 (5)	**0.48 (1)**	0.57 (3)	0.50 (2)	0.67 (6)
	Avg. gain	—	+ 0.34	+ 0.34	+ 0.13	+ 0.91	+ 0.78	– 0.11	– 0.21	– 0.15	– 0.17	– 0.04
CIFAR-100	ResNet18	2.56 (9)	1.76 (5)	**1.22 (1)**	1.25 (2)	5.24 (11)	3.33 (10)	2.00 (7)	1.87 (6)	1.44 (3)	2.18 (8)	1.75 (4)
	ResNet50	2.41 (7)	1.87 (2)	2.63 (8)	2.66 (9)	4.86 (11)	4.55 (10)	**1.82 (1)**	2.10 (5)	1.90 (3)	2.15 (6)	1.97 (4)
	ResNet110	2.64 (7)	**1.37 (1)**	3.21 (9)	3.02 (8)	4.70 (11)	4.45 (10)	1.76 (2)	1.93 (3)	1.98 (4)	2.25 (6)	2.00 (5)
	ResNet152	2.42 (6)	1.67 (2)	2.57 (8)	3.52 (9)	4.19 (11)	3.97 (10)	**1.65 (1)**	1.98 (4)	1.71 (3)	2.47 (7)	2.17 (5)
	Avg. gain	—	– 0.84	– 0.10	+ 0.10	+ 2.24	+ 1.56	– 0.69	– 0.53	– 0.74	– 0.24	– 0.53
Tiny-ImageNet	ResNet18	1.33 (4)	1.32 (2)	1.33 (3)	1.49 (8)	2.22 (11)	1.55 (10)	1.38 (5)	**1.30 (1)**	1.47 (7)	1.54 (9)	1.41 (6)
	ResNet50	1.23 (3)	1.28 (4)	1.46 (6)	1.65 (9)	2.08 (11)	1.83 (10)	1.59 (8)	1.58 (7)	**1.18 (1)**	1.23 (2)	1.36 (5)
	ResNet110	1.36 (4)	1.55 (8)	1.52 (7)	2.26 (11)	2.14 (10)	**1.28 (1)**	1.92 (9)	1.49 (6)	1.35 (3)	1.29 (2)	1.39 (5)
	ResNet152	1.33 (3)	2.08 (10)	1.82 (7)	2.00 (8)	1.81 (6)	2.06 (9)	1.46 (5)	2.19 (11)	1.43 (4)	**1.17 (1)**	1.30 (2)
	ResNet18†	1.12 (2)	1.43 (5)	1.22 (3)	1.31 (4)	2.83 (11)	1.90 (10)	1.58 (7)	1.72 (8)	1.56 (6)	1.79 (9)	**1.11 (1)**
	ResNet152†	1.96 (6)	1.57 (4)	2.75 (9)	2.74 (8)	4.83 (10)	6.60 (11)	**1.19 (1)**	1.68 (5)	1.37 (3)	2.58 (7)	1.26 (2)
	Avg. gain	—	+ 0.14	+ 0.29	+ 0.51	+ 1.26	+ 1.14	+ 0.13	+ 0.27	0.00	+ 0.21	– 0.08

Experiments

ECE ↓

	Backbones	ERM	Mixup (0.1)	Mixup (0.5)	Mixup (1.0)	Mixup (DT)	Mixup (TO)	Mixup (SC)	Mixup (IO)	**MIT-A**	**MIT-L** ($\Delta\lambda > \frac{1}{2}$)	**MIT-A** ($\Delta\lambda > \frac{1}{2}$)
SVHN	ResNet18	0.50 (2)	0.99 (7)	1.23 (11)	1.12 (10)	1.01 (8)	1.05 (9)	0.50 (3)	0.61 (5)	**0.47 (1)**	0.65 (6)	0.53 (4)
	ResNet50	0.87 (6)	1.03 (7)	1.21 (9)	1.18 (8)	1.55 (11)	1.39 (10)	0.59 (4)	0.50 (2)	**0.49 (1)**	0.52 (3)	0.66 (5)
	ResNet110	0.75 (6)	1.08 (7)	1.28 (9)	1.14 (8)	1.39 (10)	1.43 (11)	0.50 (2)	0.60 (4)	**0.48 (1)**	0.53 (3)	0.70 (5)
	ResNet152	0.90 (6)	1.05 (8)	1.21 (9)	1.04 (7)	1.28 (10)	1.37 (11)	0.57 (2)	0.65 (4)	0.61 (3)	**0.53 (1)**	0.67 (5)
	Avg. gain	—	+ 0.28	+ 0.47	+ 0.36	+ 0.55	+ 0.55	– 0.21	– 0.16	– 0.24	– 0.19	– 0.11
CIFAR-10	ResNet18	0.65 (6)	1.04 (8)	1.15 (9)	0.94 (7)	1.70 (11)	1.45 (10)	0.62 (4)	0.61 (3)	**0.56 (1)**	0.59 (2)	0.62 (5)
	ResNet50	0.79 (6)	1.07 (8)	1.15 (9)	0.91 (7)	1.81 (11)	1.64 (10)	0.65 (4)	**0.46 (1)**	0.63 (3)	0.59 (2)	0.68 (5)
	ResNet110	0.83 (7)	1.08 (9)	0.95 (8)	0.83 (6)	1.52 (10)	1.56 (11)	0.54 (3)	**0.50 (1)**	0.52 (2)	0.54 (4)	0.78 (5)
	ResNet152	0.65 (4)	1.12 (9)	1.05 (8)	0.76 (7)	1.55 (11)	1.42 (10)	0.67 (5)	**0.48 (1)**	0.57 (3)	0.50 (2)	0.67 (6)
	Avg. gain	—	+ 0.34	+ 0.34	+ 0.13	+ 0.91	+ 0.78	– 0.11	– 0.21	– 0.15	– 0.17	– 0.04
CIFAR-100	ResNet18	2.56 (9)	1.76 (5)	**1.22 (1)**	1.25 (2)	5.24 (11)	3.33 (10)	2.00 (7)	1.87 (6)	1.44 (3)	2.18 (8)	1.75 (4)
	ResNet50	2.41 (7)	1.87 (2)	2.63 (8)	2.66 (9)	4.86 (11)	4.55 (10)	**1.82 (1)**	2.10 (5)	1.90 (3)	2.15 (6)	1.97 (4)
	ResNet110	2.64 (7)	**1.37 (1)**	3.21 (9)	3.02 (8)	4.70 (11)	4.45 (10)	1.76 (2)	1.93 (3)	1.98 (4)	2.25 (6)	2.00 (5)
	ResNet152	2.42 (6)	1.67 (2)	2.57 (8)	3.52 (9)	4.19 (11)	3.97 (10)	**1.65 (1)**	1.98 (4)	1.71 (3)	2.47 (7)	2.17 (5)
	Avg. gain	—	– 0.84	– 0.10	+ 0.10	+ 2.24	+ 1.56	– 0.69	– 0.53	– 0.74	– 0.24	– 0.53
Tiny-ImageNet	ResNet18	1.33 (4)	1.32 (2)	1.33 (3)	1.49 (8)	2.22 (11)	1.55 (10)	1.38 (5)	**1.30 (1)**	1.47 (7)	1.54 (9)	1.41 (6)
	ResNet50	1.23 (3)	1.28 (4)	1.46 (6)	1.65 (9)	2.08 (11)	1.83 (10)	1.59 (8)	1.58 (7)	**1.18 (1)**	1.23 (2)	1.36 (5)
	ResNet110	1.36 (4)	1.55 (8)	1.52 (7)	2.26 (11)	2.14 (10)	**1.28 (1)**	1.92 (9)	1.49 (6)	1.35 (3)	1.29 (2)	1.39 (5)
	ResNet152	1.33 (3)	2.08 (10)	1.82 (7)	2.00 (8)	1.81 (6)	2.06 (9)	1.46 (5)	2.19 (11)	1.43 (4)	**1.17 (1)**	1.30 (2)
	ResNet18†	1.12 (2)	1.43 (5)	1.22 (3)	1.31 (4)	2.83 (11)	1.90 (10)	1.58 (7)	1.72 (8)	1.56 (6)	1.79 (9)	**1.11 (1)**
	ResNet152†	1.96 (6)	1.57 (4)	2.75 (9)	2.74 (8)	4.83 (10)	6.60 (11)	**1.19 (1)**	1.68 (5)	1.37 (3)	2.58 (7)	1.26 (2)
	Avg. gain	—	+ 0.14	+ 0.29	+ 0.51	+ 1.26	+ 1.14	+ 0.13	+ 0.27	0.00	+ 0.21	– 0.08

Experiments

ECE ↓

						With LS		Without LS				
	Backbones	ERM	Mixup (0.1)	Mixup (0.5)	Mixup (1.0)	Mixup (DT)	Mixup (TO)	Mixup (SC)	Mixup (IO)	**MIT-A**	**MIT-L ($\Delta\lambda>\frac{1}{2}$)**	**MIT-A ($\Delta\lambda>\frac{1}{2}$)**
SVHN	ResNet18	0.50 (2)	0.99 (7)	1.23 (11)	1.12 (10)	1.01 (8)	1.05 (9)	0.50 (3)	0.61 (5)	**0.47 (1)**	0.65 (6)	0.53 (4)
	ResNet50	0.87 (6)	1.03 (7)	1.21 (9)	1.18 (8)	1.55 (11)	1.39 (10)	0.59 (4)	0.50 (2)	**0.49 (1)**	0.52 (3)	0.66 (5)
	ResNet110	0.75 (6)	1.08 (7)	1.28 (9)	1.14 (8)	1.39 (10)	1.43 (11)	0.50 (2)	0.60 (4)	**0.48 (1)**	0.53 (3)	0.70 (5)
	ResNet152	0.90 (6)	1.05 (8)	1.21 (9)	1.04 (7)	1.28 (10)	1.37 (11)	0.57 (2)	0.65 (4)	0.61 (3)	**0.53 (1)**	0.67 (5)
	Avg. gain	—	+ 0.28	+ 0.47	+ 0.36	+ 0.55	+ 0.55	– 0.21	– 0.16	– 0.24	– 0.19	– 0.11
CIFAR-10	ResNet18	0.65 (6)	1.04 (8)	1.15 (9)	0.94 (7)	1.70 (11)	1.45 (10)	0.62 (4)	0.61 (3)	**0.56 (1)**	0.59 (2)	0.62 (5)
	ResNet50	0.79 (6)	1.07 (8)	1.15 (9)	0.91 (7)	1.81 (11)	1.64 (10)	0.65 (4)	**0.46 (1)**	0.63 (3)	0.59 (2)	0.68 (5)
	ResNet110	0.83 (7)	1.08 (9)	0.95 (8)	0.83 (6)	1.52 (10)	1.56 (11)	0.54 (3)	**0.50 (1)**	0.52 (2)	0.54 (4)	0.78 (5)
	ResNet152	0.65 (4)	1.12 (9)	1.05 (8)	0.76 (7)	1.55 (11)	1.42 (10)	0.67 (5)	**0.48 (1)**	0.57 (3)	0.50 (2)	0.67 (6)
	Avg. gain	—	+ 0.34	+ 0.34	+ 0.13	+ 0.91	+ 0.78	– 0.11	– 0.21	– 0.15	– 0.17	– 0.04
CIFAR-100	ResNet18	2.56 (9)	1.76 (5)	**1.22 (1)**	1.25 (2)	5.24 (11)	3.33 (10)	2.00 (7)	1.87 (6)	1.44 (3)	2.18 (8)	1.75 (4)
	ResNet50	2.41 (7)	1.87 (2)	2.63 (8)	2.66 (9)	4.86 (11)	4.55 (10)	**1.82 (1)**	2.10 (5)	1.90 (3)	2.15 (6)	1.97 (4)
	ResNet110	2.64 (7)	**1.37 (1)**	3.21 (9)	3.02 (8)	4.70 (11)	4.45 (10)	1.76 (2)	1.93 (3)	1.98 (4)	2.25 (6)	2.00 (5)
	ResNet152	2.42 (6)	1.67 (2)	2.57 (8)	3.52 (9)	4.19 (11)	3.97 (10)	**1.65 (1)**	1.98 (4)	1.71 (3)	2.47 (7)	2.17 (5)
	Avg. gain	—	– 0.84	– 0.10	+ 0.10	+ 2.24	+ 1.56	– 0.69	– 0.53	– 0.74	– 0.24	– 0.53
Tiny-ImageNet	ResNet18	1.33 (4)	1.32 (2)	1.33 (3)	1.49 (8)	2.22 (11)	1.55 (10)	1.38 (5)	**1.30 (1)**	1.47 (7)	1.54 (9)	1.41 (6)
	ResNet50	1.23 (3)	1.28 (4)	1.46 (6)	1.65 (9)	2.08 (11)	1.83 (10)	1.59 (8)	1.58 (7)	**1.18 (1)**	1.23 (2)	1.36 (5)
	ResNet110	1.36 (4)	1.55 (8)	1.52 (7)	2.26 (11)	2.14 (10)	**1.28 (1)**	1.92 (9)	1.49 (6)	1.35 (3)	1.29 (2)	1.39 (5)
	ResNet152	1.33 (3)	2.08 (10)	1.82 (7)	2.00 (8)	1.81 (6)	2.06 (9)	1.46 (5)	2.19 (11)	1.43 (4)	**1.17 (1)**	1.30 (2)
	ResNet18†	1.12 (2)	1.43 (5)	1.22 (3)	1.31 (4)	2.83 (11)	1.90 (10)	1.58 (7)	1.72 (8)	1.56 (6)	1.79 (9)	**1.11 (1)**
	ResNet152†	1.96 (6)	1.57 (4)	2.75 (9)	2.74 (8)	4.83 (10)	6.60 (11)	**1.19 (1)**	1.68 (5)	1.37 (3)	2.58 (7)	1.26 (2)
	Avg. gain	—	+ 0.14	+ 0.29	+ 0.51	+ 1.26	+ 1.14	+ 0.13	+ 0.27	0.00	+ 0.21	– 0.08

Experiments

ECE ↓

	Backbones	ERM	Mixup (0.1)	Mixup (0.5)	Mixup (1.0)	Mixup (DT)	Mixup (TO)	Mixup (SC)	Mixup (IO)	MIT-A	MIT-L ($\Delta\lambda>\frac{1}{2}$)	MIT-A ($\Delta\lambda>\frac{1}{2}$)
SVHN	ResNet18	0.50 (2)	0.99 (7)	1.23 (11)	1.12 (10)	1.01 (8)	1.05 (9)	0.50 (3)	0.61 (5)	**0.47 (1)**	0.65 (6)	0.53 (4)
	ResNet50	0.87 (6)	1.03 (7)	1.21 (9)	1.18 (8)	1.55 (11)	1.39 (10)	0.59 (4)	0.50 (2)	**0.49 (1)**	0.52 (3)	0.66 (5)
	ResNet110	0.75 (6)	1.08 (7)	1.28 (9)	1.14 (8)	1.39 (10)	1.43 (11)	0.50 (2)	0.60 (4)	**0.48 (1)**	0.53 (3)	0.70 (5)
	ResNet152	0.90 (6)	1.05 (8)	1.21 (9)	1.04 (7)	1.28 (10)	1.37 (11)	0.57 (2)	0.65 (4)	0.61 (3)	**0.53 (1)**	0.67 (5)
	Avg. gain	—	+ 0.28	+ 0.47	+ 0.36	+ 0.55	+ 0.55	– 0.21	– 0.16	– 0.24	– 0.19	– 0.11
CIFAR-10	ResNet18	0.65 (6)	1.04 (8)	1.15 (9)	0.94 (7)	1.70 (11)	1.45 (10)	0.62 (4)	0.61 (3)	**0.56 (1)**	0.59 (2)	0.62 (5)
	ResNet50	0.79 (6)	1.07 (8)	1.15 (9)	0.91 (7)	1.81 (11)	1.64 (10)	0.65 (4)	**0.46 (1)**	0.63 (3)	0.59 (2)	0.68 (5)
	ResNet110	0.83 (7)	1.08 (9)	0.95 (8)	0.83 (6)	1.52 (10)	1.56 (11)	0.54 (3)	**0.50 (1)**	0.52 (2)	0.54 (4)	0.78 (5)
	ResNet152	0.65 (4)	1.12 (9)	1.05 (8)	0.76 (7)	1.55 (11)	1.42 (10)	0.67 (5)	**0.48 (1)**	0.57 (3)	0.50 (2)	0.67 (6)
	Avg. gain	—	+ 0.34	+ 0.34	+ 0.13	+ 0.91	+ 0.78	– 0.11	– 0.21	– 0.15	– 0.17	– 0.04
CIFAR-100	ResNet18	2.56 (9)	1.76 (5)	**1.22 (1)**	1.25 (2)	5.24 (11)	3.33 (10)	2.00 (7)	1.87 (6)	1.44 (3)	2.18 (8)	1.75 (4)
	ResNet50	2.41 (7)	1.87 (2)	2.63 (8)	2.66 (9)	4.86 (11)	4.55 (10)	**1.82 (1)**	2.10 (5)	1.90 (3)	2.15 (6)	1.97 (4)
	ResNet110	2.64 (7)	**1.37 (1)**	3.21 (9)	3.02 (8)	4.70 (11)	4.45 (10)	1.76 (2)	1.93 (3)	1.98 (4)	2.25 (6)	2.00 (5)
	ResNet152	2.42 (6)	1.67 (2)	2.57 (8)	3.52 (9)	4.19 (11)	3.97 (10)	**1.65 (1)**	1.98 (4)	1.71 (3)	2.47 (7)	2.17 (5)
	Avg. gain	—	– 0.84	– 0.10	+ 0.10	+ 2.24	+ 1.56	– 0.69	– 0.53	– 0.74	– 0.24	– 0.53
Tiny-ImageNet	ResNet18	1.33 (4)	1.32 (2)	1.33 (3)	1.49 (8)	2.22 (11)	1.55 (10)	1.38 (5)	**1.30 (1)**	1.47 (7)	1.54 (9)	1.41 (6)
	ResNet50	1.23 (3)	1.28 (4)	1.46 (6)	1.65 (9)	2.08 (11)	1.83 (10)	1.59 (8)	1.58 (7)	**1.18 (1)**	1.23 (2)	1.36 (5)
	ResNet110	1.36 (4)	1.55 (8)	1.52 (7)	2.26 (11)	2.14 (10)	**1.28 (1)**	1.92 (9)	1.49 (6)	1.35 (3)	1.29 (2)	1.39 (5)
	ResNet152	1.33 (3)	2.08 (10)	1.82 (7)	2.00 (8)	1.81 (6)	2.06 (9)	1.46 (5)	2.19 (11)	1.43 (4)	**1.17 (1)**	1.30 (2)
	ResNet18†	1.12 (2)	1.43 (5)	1.22 (3)	1.31 (4)	2.83 (11)	1.90 (10)	1.58 (7)	1.72 (8)	1.56 (6)	1.79 (9)	**1.11 (1)**
	ResNet152†	1.96 (6)	1.57 (4)	2.75 (9)	2.74 (8)	4.83 (10)	6.60 (11)	**1.19 (1)**	1.68 (5)	1.37 (3)	2.58 (7)	1.26 (2)
	Avg. gain	—	+ 0.14	+ 0.29	+ 0.51	+ 1.26	+ 1.14	+ 0.13	+ 0.27	0.00	+ 0.21	– 0.08

Experiments

Without LS

Accuracy ↑

	Backbones	ERM	Mixup (0.1)	Mixup (0.5)	Mixup (1.0)	Mixup (DT)	Mixup (TO)	Mixup (SC)	Mixup (IO)	**MIT-A**	**MIT-L** ($\Delta\lambda>\frac{1}{2}$)	**MIT-A** ($\Delta\lambda>\frac{1}{2}$)
SVHN	ResNet18	95.4 (5)	95.5 (4)	94.8 (7)	94.5 (8)	95.6 (3)	**96.0 (1)**	94.3 (9)	93.5 (10	95.0 (6)	93.2 (11)	95.7 (2)
	ResNet50	96.0 (4)	96.0 (3)	95.8 (5)	95.5 (8)	95.5 (7)	95.7 (6)	95.3 (9)	94.9 (10	96.2 (2)	94.3 (11)	**96.2 (1)**
	ResNet110	96.0 (5)	96.1 (4)	96.3 (3)	95.8 (7)	95.6 (8)	95.9 (6)	95.4 (9)	95.3 (10	96.5 (2)	95.0 (11)	**96.7 (1)**
	ResNet152	96.2 (5)	96.6 (2)	96.4 (4)	96.2 (6)	95.6 (8)	95.9 (7)	95.5 (9)	95.5 (10	96.5 (3)	94.9 (11)	**96.7 (1)**
	Avg. gain	—	+ 0.11	− 0.10	− 0.40	− 0.32	− 0.06	− 0.78	− 1.12	+ 0.14	− 1.55	+ 0.41
CIFAR-10	ResNet18	94.5 (9)	95.1 (6)	95.7 (3)	95.8 (2)	93.9 (11)	94.5 (8)	94.4 (10)	94.7 (7)	95.5 (4)	95.2 (5)	**95.9 (1)**
	ResNet50	94.4 (9)	95.3 (7)	95.8 (3)	**96.0 (1)**	93.1 (11)	94.2 (10)	94.5 (8)	95.3 (6)	95.8 (4)	95.7 (5)	96.0 (2)
	ResNet110	94.7 (9)	95.7 (6)	**96.3 (1)**	96.2 (2)	93.7 (11)	94.3 (10)	95.1 (8)	95.4 (7)	96.1 (4)	96.0 (5)	96.1 (3)
	ResNet152	95.1 (8)	95.8 (7)	96.4 (2)	**96.7 (1)**	93.9 (11)	94.8 (10)	95.0 (9)	95.8 (6)	96.3 (4)	96.2 (5)	96.4 (3)
	Avg. gain	—	+ 0.78	+ 1.36	+ 1.53	− 1.01	− 0.21	+ 0.08	+ 0.64	+ 1.27	+ 1.12	+ 1.41
CIFAR-100	ResNet18	74.4 (8)	75.3 (7)	76.8 (2)	**77.2 (1)**	72.4 (11)	76.4 (4)	72.6 (9)	72.5 (10	76.2 (5)	75.9 (6)	76.6 (3)
	ResNet50	73.9 (9)	76.4 (6)	78.3 (2)	77.8 (3)	68.2 (11)	75.1 (7)	72.9 (10)	74.5 (8)	**78.3 (1)**	76.6 (5)	77.7 (4)
	ResNet110	76.1 (9)	77.9 (6)	**80.1 (1)**	79.3 (2)	70.9 (11)	77.3 (7)	74.6 (10)	76.7 (8)	78.7 (4)	77.9 (5)	79.1 (3)
	ResNet152	75.3 (9)	78.2 (6)	79.7 (2)	79.6 (3)	72.5 (11)	76.9 (7)	75.1 (10)	76.7 (8)	79.1 (4)	78.2 (5)	**79.8 (1)**
	Avg. gain	—	+ 2.01	+ 3.79	+ 3.55	− 3.92	+ 1.50	− 1.12	+ 0.18	+ 3.14	+ 2.24	+ 3.38
Tiny-ImageNet	ResNet18	46.1 (9)	46.6 (7)	47.4 (5)	47.8 (4)	36.5 (11)	47.1 (6)	43.0 (10)	46.6 (8)	**49.5 (1)**	48.5 (3)	49.3 (2)
	ResNet50	49.3 (7)	49.5 (6)	50.0 (5)	50.4 (4)	37.5 (11)	49.0 (8)	46.4 (10)	48.8 (9)	51.4 (2)	51.0 (3)	**51.8 (1)**
	ResNet110	48.5 (3)	43.6 (7)	42.7 (9)	42.6 (10)	35.6 (11)	44.6 (4)	43.9 (6)	43.5 (8)	48.6 (2)	44.4 (5)	**50.8 (1)**
	ResNet152	47.3 (2)	44.7 (5)	42.3 (9)	44.6 (6)	34.5 (11)	45.5 (4)	43.0 (8)	39.7 (10	46.1 (3)	43.8 (7)	**50.0 (1)**
	ResNet18†	53.6 (6)	53.5 (8)	54.0 (5)	53.5 (7)	44.1 (11)	**54.7 (1)**	49.7 (10)	50.5 (9)	54.5 (3)	54.4 (4)	54.7 (2)
	ResNet152†	62.4 (6)	63.2 (2)	**63.7 (1)**	63.0 (3)	49.6 (11)	62.6 (4)	58.8 (10)	59.9 (9)	61.9 (7)	62.5 (5)	61.6 (8)
	Avg. gain	—	− 1.01	− 1.18	− 0.87	− 11.5	− 0.60	− 3.70	− 3.04	+ 0.81	− 0.41	+ 1.83

Experiments

Accuracy ↑

	Backbones	ERM	Mixup (0.1)	Mixup (0.5)	Mixup (1.0)	Mixup (DT)	Mixup (TO)	Mixup (SC)	Mixup (IO)	MIT-A	MIT-L ($\Delta\lambda > \frac{1}{2}$)	MIT-A ($\Delta\lambda > \frac{1}{2}$)
SVHN	ResNet18	95.4 (5)	95.5 (4)	94.8 (7)	94.5 (8)	95.6 (3)	**96.0 (1)**	94.3 (9)	93.5 (10	95.0 (6)	93.2 (11)	95.7 (2)
	ResNet50	96.0 (4)	96.0 (3)	95.8 (5)	95.5 (8)	95.5 (7)	95.7 (6)	95.3 (9)	94.9 (10	96.2 (2)	94.3 (11)	**96.2 (1)**
	ResNet110	96.0 (5)	96.1 (4)	96.3 (3)	95.8 (7)	95.6 (8)	95.9 (6)	95.4 (9)	95.3 (10	96.5 (2)	95.0 (11)	**96.7 (1)**
	ResNet152	96.2 (5)	96.6 (2)	96.4 (4)	96.2 (6)	95.6 (8)	95.9 (7)	95.5 (9)	95.5 (10	96.5 (3)	94.9 (11)	**96.7 (1)**
	Avg. gain	—	+ 0.11	− 0.10	− 0.40	− 0.32	− 0.06	− 0.78	− 1.12	+ 0.14	− 1.55	+ 0.41
CIFAR-10	ResNet18	94.5 (9)	95.1 (6)	95.7 (3)	95.8 (2)	93.9 (11)	94.5 (8)	94.4 (10)	94.7 (7)	95.5 (4)	95.2 (5)	**95.9 (1)**
	ResNet50	94.4 (9)	95.3 (7)	95.8 (3)	**96.0 (1)**	93.1 (11)	94.2 (10)	94.5 (8)	95.3 (6)	95.8 (4)	95.7 (5)	96.0 (2)
	ResNet110	94.7 (9)	95.7 (6)	**96.3 (1)**	96.2 (2)	93.7 (11)	94.3 (10)	95.1 (8)	95.4 (7)	96.1 (4)	96.0 (5)	96.1 (3)
	ResNet152	95.1 (8)	95.8 (7)	96.4 (2)	**96.7 (1)**	93.9 (11)	94.8 (10)	95.0 (9)	95.8 (6)	96.3 (4)	96.2 (5)	96.4 (3)
	Avg. gain	—	+ 0.78	+ 1.36	+ 1.53	− 1.01	− 0.21	+ 0.08	+ 0.64	+ 1.27	+ 1.12	+ 1.41
CIFAR-100	ResNet18	74.4 (8)	75.3 (7)	76.8 (2)	**77.2 (1)**	72.4 (11)	76.4 (4)	72.6 (9)	72.5 (10	76.2 (5)	75.9 (6)	76.6 (3)
	ResNet50	73.9 (9)	76.4 (6)	78.3 (2)	77.8 (3)	68.2 (11)	75.1 (7)	72.9 (10)	74.5 (8)	**78.3 (1)**	76.6 (5)	77.7 (4)
	ResNet110	76.1 (9)	77.9 (6)	**80.1 (1)**	79.3 (2)	70.9 (11)	77.3 (7)	74.6 (10)	76.7 (8)	78.7 (4)	77.9 (5)	79.1 (3)
	ResNet152	75.3 (9)	78.2 (6)	79.7 (2)	79.6 (3)	72.5 (11)	76.9 (7)	75.1 (10)	76.7 (8)	79.1 (4)	78.2 (5)	**79.8 (1)**
	Avg. gain	—	+ 2.01	+ 3.79	+ 3.55	− 3.92	+ 1.50	− 1.12	+ 0.18	+ 3.14	+ 2.24	+ 3.38
Tiny-ImageNet	ResNet18	46.1 (9)	46.6 (7)	47.4 (5)	47.8 (4)	36.5 (11)	47.1 (6)	43.0 (10)	46.6 (8)	**49.5 (1)**	48.5 (3)	49.3 (2)
	ResNet50	49.3 (7)	49.5 (6)	50.0 (5)	50.4 (4)	37.5 (11)	49.0 (8)	46.4 (10)	48.8 (9)	51.4 (2)	51.0 (3)	**51.8 (1)**
	ResNet110	48.5 (3)	43.6 (7)	42.7 (9)	42.6 (10)	35.6 (11)	44.6 (4)	43.9 (6)	43.5 (8)	48.6 (2)	44.4 (5)	**50.8 (1)**
	ResNet152	47.3 (2)	44.7 (5)	42.3 (9)	44.6 (6)	34.5 (11)	45.5 (4)	43.0 (8)	39.7 (10	46.1 (3)	43.8 (7)	**50.0 (1)**
	ResNet18†	53.6 (6)	53.5 (8)	54.0 (5)	53.5 (7)	44.1 (11)	**54.7 (1)**	49.7 (10)	50.5 (9)	54.5 (3)	54.4 (4)	54.7 (2)
	ResNet152†	62.4 (6)	63.2 (2)	**63.7 (1)**	63.0 (3)	49.6 (11)	62.6 (4)	58.8 (10)	59.9 (9)	61.9 (7)	62.5 (5)	61.6 (8)
	Avg. gain	—	− 1.01	− 1.18	− 0.87	− 11.5	− 0.60	− 3.70	− 3.04	+ 0.81	− 0.41	+ 1.83

Thanks!

See details in our paper~ ☺

Deng-Bao Wang