Logit Mixing Training for More Reliable and Accurate Prediction

Duhyeon Bang*, Kyungjune Baek*, Jiwoo Kim, Yunho Jeon, Jin-Hwa Kim, Jiwon Kim Jongwuk Lee, and Hyunjung Shim†

* indicates equal contribution † indicates a corresponding author

Motivation & Goal

- **Problem:** DNNs are poor at understanding inter-class relationship because model training strictly enforces to predict the one-hot labels.
- **Goals:** We devise DNNs to utilize inter-class relationships by rejecting improbable classes.
- **Key idea :** We adopt logits as weak supervision for learning inter-class relationship.

Logit vs. One-hot Labels

- **Logits** can reveal the inter-class relationship while one-hot labels do not provide any relationship between classes.

=

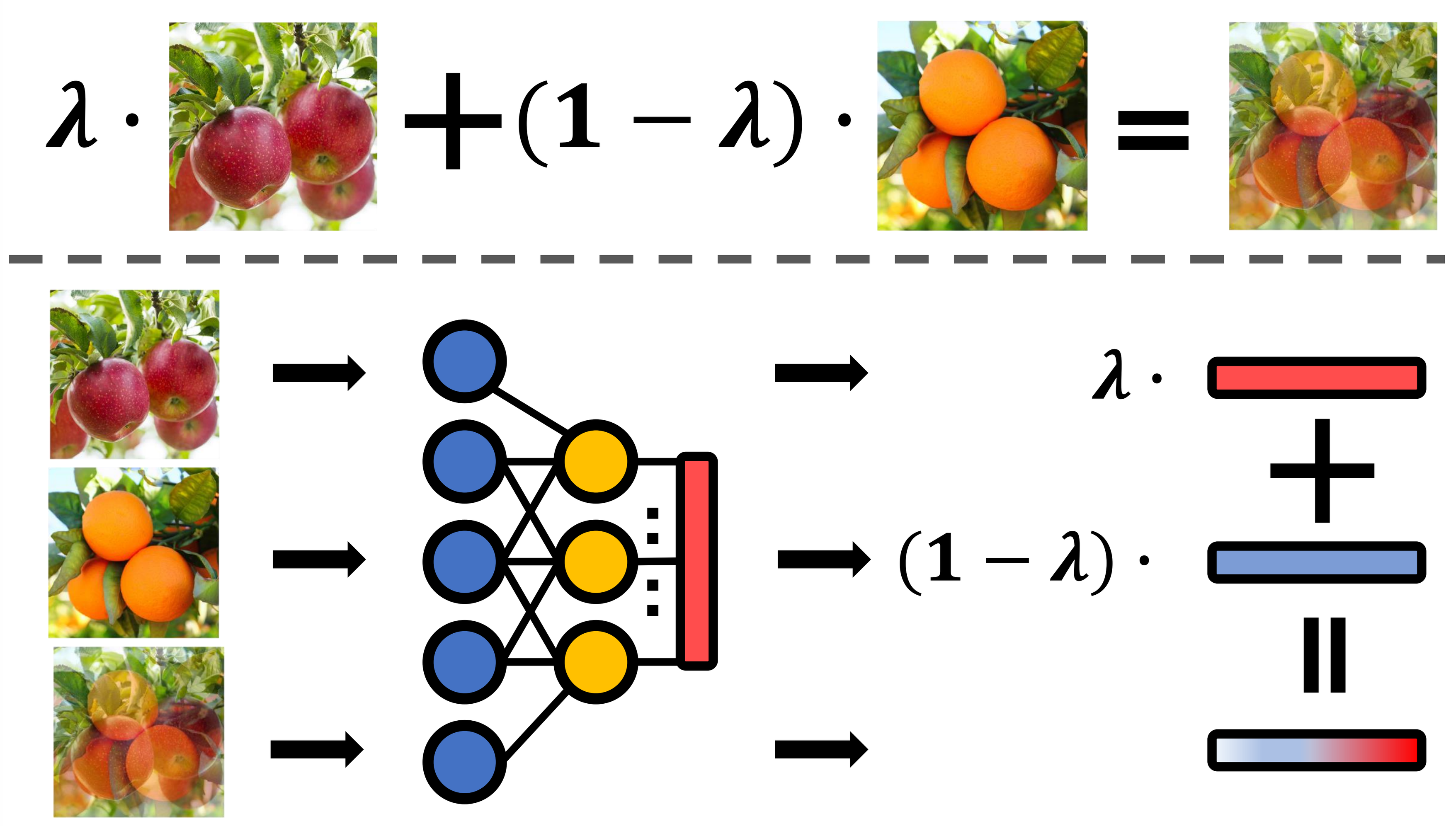
	Apple	Orange		
Logit:	[1.5	0.5	0 . . . 0]	→ **Apple & Orange: Similar**
Label:	[1	0	0 . . . 0]	→ **Apple & Orange: ?**

Logits vs. Probability Vector

- **Probability vector** (a.k.a post-softmax output) can capture the positive inter-class relationship, however, it can distort the true relationship.

	Dog	Cat		
Logit	[1.5	0.5	0 ..... 0]	→ **Dog & Cat: Positive**
Logit	[0.5	-0.5	0 ..... 0]	→ **Dog & Cat: Negative**
Softmax ↓				
Probability	[0.73	0.27	0 ..... 0]	→ **Dog & Cat: Positive (?)**

Method



$$\mathcal{L}_{sim} = \|(\lambda f(x_1) + (1-\lambda) f(x_2)) - f(x_{mix})\|_2$$

Objective Function

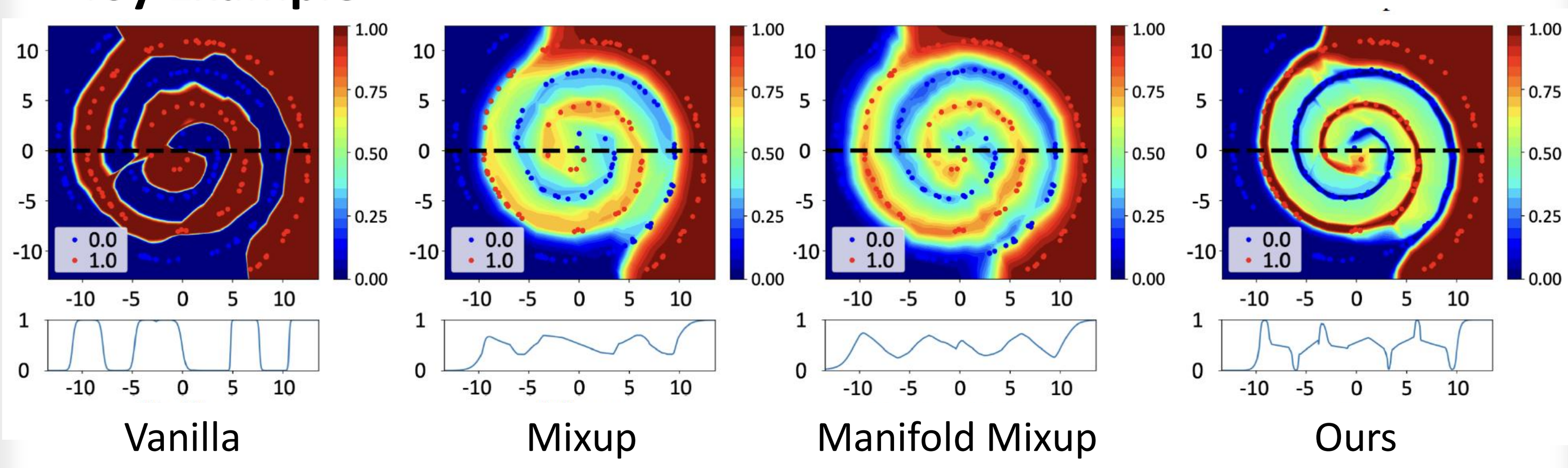
Similarity loss
$$\mathcal{L}_{sim} = \|(\lambda f(x_1) + (1-\lambda) f(x_2)) - f(x_{mix})\|_2$$

Cross-entropy with original labels
$$\mathcal{L}_{cls} = \mathcal{H}(\tilde{y}_1, y_1) + \mathcal{H}(\tilde{y}_2, y_2)$$

Cross-entropy with mixed labels
$$y_{mix} = \lambda y_1 + (1-\lambda) y_2$$
$$\mathcal{L}_{mix} = \lambda \mathcal{H}(\tilde{y}_{mix}, y_1) + (1-\lambda)\mathcal{H}(\tilde{y}_{mix}, y_2)$$

Experimental Results

- **Toy Example**



- **Ablation Study on Losses**

Name	Loss	Accuracy(%)
Vanilla	$\mathcal{L}_{cls}$	78.32 ± 0.07
Mixup	$\mathcal{L}_{mix}$	79.82 ± 0.08
LogitMix$_m$	$\mathcal{L}_{cls} + \mathcal{L}_{sim} + \mathcal{L}_{mix}$	$\mathbf{81.59 \pm 0.09}$
(−) Mixup	$\mathcal{L}_{cls} + \mathcal{L}_{sim}$	80.11 ± 0.09
(−) Cross entropy	$\mathcal{L}_{mix} + \mathcal{L}_{sim}$	80.51 ± 0.08
(−) Similarity loss	$\mathcal{L}_{mix} + \mathcal{L}_{cls}$	80.08 ± 0.49

- **Image Classification**

Dataset	Network	Metric	Vanilla	Mixup	LogitMix$_m$	CutMix	LogitMix$_c$	PuzzleMix	LogitMix$_p$
CIFAR100	VGG16	Acc	74.30	75.02	76.22 (+1.20)	75.34	76.10 (+0.76)	75.92	**76.38** (+0.46)
		ECE	0.176	0.060	**0.035** (-0.025)	0.051	0.062 (+0.011)	0.121	0.100 (-0.021)
		OE	0.154	0.035	0.025 (-0.010)	0.022	**0.008** (-0.014)	0.011	0.049 (+0.038)
	ResNet50	Acc	78.32	79.82	81.59 (+1.77)	80.57	81.02 (+0.45)	82.57	**83.76** (+1.19)
		ECE	0.087	0.040	**0.014** (-0.026)	0.078	0.073 (-0.005)	0.092	0.215 (+0.123)
		OE	0.073	0.028	0.003 (-0.025)	0.064	0.060 (-0.004)	0.015	**0.000** (-0.015)
	ResNeXt50	Acc	79.18	81.10	81.63 (+0.53)	81.16	81.46 (+0.30)	81.40	**82.13** (+0.73)
		ECE	0.069	0.042	**0.021** (-0.021)	0.059	0.032 (-0.027)	0.092	0.220 (+0.128)
		OE	0.057	0.001	**0.000** (-0.001)	0.047	0.023 (-0.024)	0.017	0.001 (-0.016)
	MobileNetV2	Acc	69.69	69.98	73.90 (+3.92)	68.82	69.91 (+1.09)	75.77	**75.99** (+0.22)
		ECE	0.061	0.091	**0.048** (-0.043)	0.050	0.049 (-0.001)	0.097	0.100 (+0.003)
		OE	0.042	**0.000**	**0.000** (0.000)	**0.000**	**0.000** (0.000)	0.022	0.009 (-0.013)
	ShuffleNetV2	Acc	72.17	74.17	75.53 (+1.36)	73.60	73.73 (+0.13)	76.18	**76.75** (+0.57)
		ECE	0.079	0.060	0.042 (-0.018)	**0.016**	0.023 (+0.007)	0.126	0.094 (-0.032)
		OE	0.060	**0.000**	**0.000** (-0.000)	0.002	**0.000** (-0.002)	0.014	0.001 (-0.013)
TinyImageNet	ResNet50	Acc	66.6	68.34	**70.71** (+2.37)	69.08	69.87 (+0.79)	69.71	70.15 (+0.44)
		ECE	0.098	0.032	0.030 (-0.002)	**0.029**	0.034 (+0.005)	0.121	0.131 (+0.010)
		OE	0.076	0.022	0.010 (-0.012)	0.015	**0.005** (-0.010)	0.012	0.012 (0.000)
	MobileNetV2	Acc	57.62	59.55	62.12 (+2.57)	53.54	57.66 (+4.12)	64.08	**65.30** (+1.22)
		ECE	0.073	0.091	**0.032** (-0.059)	0.094	0.082 (-0.012)	0.112	0.104 (-0.008)
		OE	0.045	0.019	**0.000** (-0.019)	**0.000**	**0.000** (0.000)	0.034	0.016 (-0.018)
ILSVRC2015	ResNet50	Acc	76.13	77.37	78.38 (+1.01)	78.43	**78.51** (+0.08)	75.63	77.47 (+1.84)
		ECE	0.370	0.041	0.028 (-0.013)	0.028	**0.020** (-0.008)	0.120	0.117 (-0.003)
		OE	0.030	0.003	**0.001** (-0.002)	0.029	0.029 (0.000)	0.053	0.056 (+0.003)

- **Text Classification / Regression**

Model	MNLI-mm	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
BERT$_{BASE}$ [Devlin *et al.*, 2018]	**84.73**	**91.25**	91.43	**93.12**	57.82	89.43	**87.75**	68.95	83.06
Mixup on BERT$_{BASE}$ [Sun *et al.*, 2020]	84.29	91.15	91.36	**93.12**	58.82	89.44	87.50	67.87	82.94
LogitMix on BERT$_{BASE}$	**84.73** (0.00)	**91.25** (0.00)	**91.58** (+0.15)	**93.12** (0.00)	**58.85** (+0.03)	**89.45** (+0.01)	87.50 (-0.25)	**70.04** (+1.09)	**83.32** (+0.26)
BERT$_{LARGE}$ [Devlin *et al.*, 2018]	85.99	90.20	92.20	92.89	60.88	89.9	87.75	73.29	84.14
Mixup on BERT$_{LARGE}$ [Sun *et al.*, 2020]	86.02	90.09	92.42	92.66	61.86	**90.02**	88.24	73.29	84.33
LogitMix on BERT$_{LARGE}$	**86.10** (+0.08)	**90.95** (+0.75)	**92.60** (+0.18)	**93.58** (+0.69)	**63.89** (+2.03)	89.98 (-0.04)	**89.22** (+0.98)	**74.37** (+1.08)	**85.09** (+0.76)

Conclusion

- We propose to utilize logits for the better understanding on the inter-class relationship.
- We analyze the effect of three different losses and the robustness for the choice of the hyperparameter.
- We verify that our LogitMix can be combined with various mixing-based augmentation methods.
- We show that LogitMix effectively improves the classification and the calibration performance for image and text datasets.

References

[Zhang et al., 2018] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. ICLR, 2018.

[Yun et al., 2019] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoo Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. ICCV, 2019.

[Kim et al., 2020] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. ICML, 2020.