MaxSup: Overcoming Representation Collapse in Label Smoothing

Yuxuan Zhou[1,2*] Heng Li[3*] Zhi-Qi Cheng[3†] Xudong Yan[3]
Yifei Dong[3] Mario Fritz[2] Margret Keuper[1]

[1] University of Mannheim [2] CISPA Helmholtz Center for Information Security [3] University of Washington
* Equal contribution † Corresponding Author

GitHub: `https://github.com/ZhouYuxuanYX/Maximum-Suppression-Regularization`

HuggingFace: `https://huggingface.co/papers/2502.15798`

Background

What Is Label Smoothing (LS)?

Definition

- LS replaces a one-hot label $y \in \mathbb{R}^K$ with a softened target vector $s \in \mathbb{R}^K$:

$$s_k = (1 - \alpha) y_k + \frac{\alpha}{K}$$

- where $y_k = 1$ for the ground-truth class and 0 for all other classes.

Why LS Is Popular

- Mitigates overfitting (Szegedy et al., CVPR 2016)
- Improves calibration (Müller et al., NeurIPS 2019)

References:

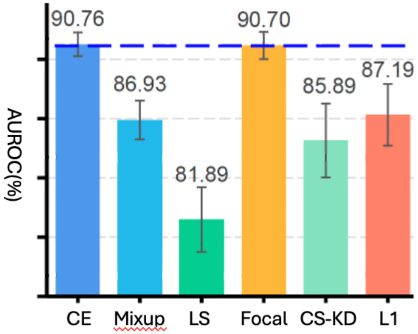
Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," CVPR 2016.

Müller et al., "When Does Label Smoothing Help?" NeurIPS 2019.

Background

Label Smoothing Regularization — Issue 1

Issue 1: Reinforces overconfidence in misclassified samples [Zhu et al., ECCV 2022]



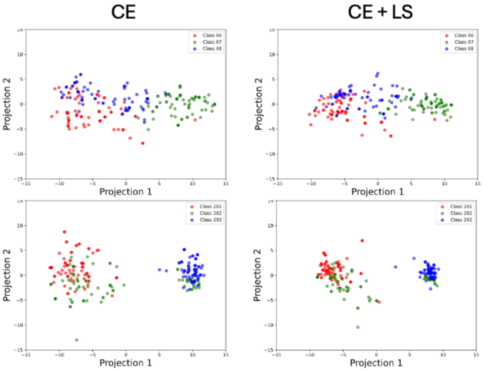
Misclassification detection by filtering out low-confidence predictions. [Zhu et al., ECCV 2022]

F. Zhu, Z. Cheng, et al., "Rethinking confidence calibration for failure prediction." ECCV 2022.

Background

Label Smoothing Regularization — Issue 2

Issue 2: Compresses features into overly tight clusters, hindering transfer learning [Kornblith et al., NeurIPS 2021]



Projected penultimate-layer activations of trained Vision Transformers on ImageNet validation.

S. Kornblith, T. Chen, et al., "Why do better loss functions lead to less transferable features?" NeurIPS 2021.

Method

Revisiting Label Smoothing (LS)

LS as a mixture of targets

$$s = (1-\alpha)\, y + \alpha u, \qquad u = \frac{1}{K}\mathbf{1}.$$

LS loss:

$$H(s,q) = -\sum_k s_k \log q_k.$$

Decomposition:

$$\begin{aligned} H(s,q) &= (1-\alpha)H(y,q) + \alpha H(u,q), \\ L_{\text{LS-reg}} &= H(s,q) - H(y,q) = \alpha\big(H(u,q) - H(y,q)\big). \end{aligned}$$

Interpretation

- LS mixes the one-hot target with a uniform target.
- Introduces a global smoothing force on predictions.

From probability space to logit space

- Let z_k denote the logit for class k, and $q_k = \mathrm{softmax}(z)_k$.
- For small perturbations, the change in cross-entropy can be related to logits.

Intuitive logit-space form of the LS regularizer

$$\begin{aligned} L_{\text{LS-reg}} &= H(s, q) - H(y, q) \\ &= \alpha\big(H(u, q) - H(y, q)\big) \\ &\approx \alpha\Big(z_{\text{gt}} - \frac{1}{K}\sum_{k=1}^{K} z_k\Big), \end{aligned}$$

where z_{gt} is the logit of the ground-truth class.

- This approximation highlights two effects of LS:
  - **Pushes up** the average logit $\frac{1}{K}\sum_k z_k$ of all classes.
  - **Pulls down** the ground-truth logit z_{gt} towards that average.
- Together, they lead to **feature compression** and weaker margins.

Method

Max Suppression Regularization

Key idea

- LS applies a *global* smoothing force to *all* non-ground-truth classes.
- Instead, we propose to penalize **only the most competitive wrong class**.

Most competitive wrong class

$$k^* = \arg\max_{k \neq y} p_k,$$

where p_k is the predicted probability for class k and y is the ground-truth label.

- If the prediction is already confident and correct, p_{k^*} is small $\Rightarrow$ almost no extra penalty.
- If a wrong class is highly competitive, p_{k^*} is large $\Rightarrow$ strong suppression on that class.

Method

Max Suppression Regularization

MaxSup loss

- Start from standard cross-entropy:

$$\mathcal{L}_{\mathrm{CE}} = H(y, p) = -\log p_y.$$

- Add a penalty on the maximum non-ground-truth probability:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda\, \phi\left(\max_{k \neq y} p_k\right),$$

where $\lambda > 0$ is a weight and $\phi(\cdot)$ is an increasing function (e.g., identity or hinge).

- No change to data or network architecture.
- Computational overhead is negligible: only one `max` operation over non-GT classes.
- Behaves as a **local, margin-aware** regularizer instead of a global smoother.

G. Xia et al., "Towards Understanding Why Label Smoothing Degrades and How to Fix It," ICLR 2025.

Experiments

Ablation on Loss Components

- Model: DeiT-Small on ImageNet-1K (no CutMix, Mixup, KD).
- We ablate LS, MaxSup, and their components.

Loss	Formulation	Accuracy
Cross Entropy	$H(\boldsymbol{y}, \boldsymbol{q})$	74.21
+ Label Smoothing	$\frac{\alpha}{K} \sum_{z_m < z_{gt}} (z_{gt} - z_m) + \frac{\alpha}{K} \sum_{z_n > z_{gt}} (z_{gt} - z_n)$	75.91
+ Regularization	$\frac{\alpha}{M} \sum_{z_m < z_{gt}} (z_{gt} - z_m)$	75.98
+ Error Amplification	$\frac{\alpha}{N} \sum_{z_n > z_{gt}} (z_{gt} - z_n)$	73.63
+ MaxSup	$\frac{\alpha}{K} \sum_{z_n > z_{gt}} (z_{gt} - z_n)$	76.12

Experiments

Impact on Representation Learning

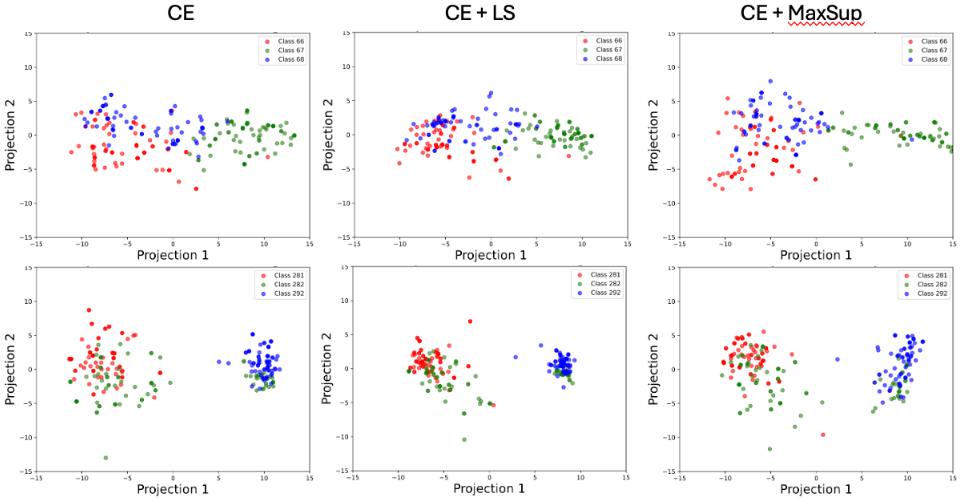
- MaxSup improves feature quality while maintaining or improving accuracy.
- Evaluated on ResNet-50 / ImageNet-1K.

Method	Intra-Class Variation (d_{within})		Inter-Class Separation (R^2)	
	Train	Validation	Train	Validation
Cross Entropy	**0.311**	**0.331**	0.403	0.445
+ Label Smoothing	0.263	0.254	<u>0.469</u>	<u>0.461</u>
+ MaxSup	<u>0.293</u>	<u>0.300</u>	**0.519**	**0.497**

Feature quality of ResNet-50 on ImageNet-1K.

Experiments

Impact on Representation Learning



Projected penultimate-layer activations of trained Vision Transformers on ImageNet validation.

Projected penultimate-layer activations of trained Vision Transformers on ImageNet validation.

Experiments

Impact on Linear Transfer

- Linear-probe transfer from ImageNet representations.
- MaxSup:
  - enhances inter-class separation,
  - preserves intra-class variation,
  - improves in-distribution accuracy.

Method	ImageNet	Linear Transfer					
		CIFAR10	CIFAR100	CUB	Flowers	Food	Pets
Cross Entropy	76.41 (0.10)	**91.74**	**75.35**	**70.21**	**90.96**	**72.44**	<u>92.30</u>
+ Label Smoothing	76.91(0.11)	90.14	71.28	64.50	84.84	67.76	91.96
+ Online Label Smoothing	77.23(0.21)	90.29	73.13	<u>67.86</u>	87.47	69.34	92.21
+ MaxSup	**77.69**(0.07)	<u>91.00</u>	<u>73.93</u>	67.29	<u>88.84</u>	<u>70.94</u>	**92.93**

Classification accuracy of ResNet-50 models trained on ImageNet; linear transfer via L2-regularized multinomial logistic regression.

Experiments

Impact on Fine-Grained Classification

- Evaluate on CUB and Cars datasets.
- MaxSup improves performance over CE and CE+LS.

Method	CUB	Cars
Cross Entropy	80.88	90.27
+ Label Smoothing	81.96	91.64
+ Online Label Smoothing	<u>82.33</u>	<u>91.96</u>
+ Zipf Label Smoothing	81.40	90.99
+ MaxSup	**82.53**	**92.25**

Classification on CUB and Cars datasets.

Experiments

Impact on Long-Tailed Classification

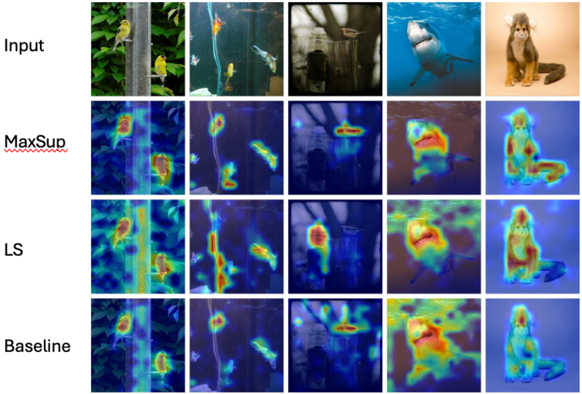
- Long-tailed CIFAR-10 with various imbalance levels.
- MaxSup improves accuracy across imbalance ratios.

Dataset	Split	Imbalance Ratio	Method	Overall	Many	Medium	Low
LT CIFAR-10	validation	50	Focal Loss	77.4	76.0	**89.7**	0.0
			LS	<u>81.2</u>	<u>81.6</u>	77.0	0.0
			MaxSup	**82.1**	**82.5**	<u>78.1</u>	0.0
	test		Focal Loss	76.8	75.3	**90.4**	0.0
			LS	<u>80.5</u>	<u>81.1</u>	<u>75.4</u>	0.0
			MaxSup	**81.4**	**82.3**	73.4	0.0
	validation	100	Focal Loss	75.1	71.8	88.3	0.0
			LS	<u>76.6</u>	**80.6**	**60.7**	0.0
			MaxSup	**77.1**	<u>80.1</u>	<u>65.1</u>	0.0
	test		Focal Loss	74.7	71.6	**87.2**	0.0
			LS	**76.4**	**80.8**	59.0	0.0
			MaxSup	**76.4**	<u>79.9</u>	<u>62.4</u>	0.0

Comparison of classification performance (%) across imbalance levels for different loss strategies on long-tailed CIFAR-10 using ResNet-32.

Experiments

Impact on Decision Making



(a) Label Smoothing is severely distracted by the pole.

(b) Label Smoothing is severely distracted by the tube, and Baseline almost overlooks the gold fish at bottom.

(c) Label Smoothing completely focuses on the wrong position, whereas Baseline is distracted by the surrounding objects.

(d) Label Smoothing and Baseline are both severely distracted by the waves.

(e) Label Smoothing fails to consider the tail of the monkey, and Baseline mostly focus on the head forehead.

Visualization of decision behaviour for Input, Baseline, LS, and MaxSup.

Q&A

Thank you!

Questions?

- **Key takeaway:** LS introduces global smoothing that harms representation geometry.
- **Our solution:** MaxSup applies local, margin-aware suppression on the top competitor.
- **Benefits:** Better features, better transfer, better detection, zero extra cost.

GitHub: `https://github.com/ZhouYuxuanYX/Maximum-Suppression-Regularization`
HuggingFace: `https://huggingface.co/papers/2502.15798`