Distilling Balanced Knowledge from a Biased Teacher

Seonghak Kim
Agency for Defense Development (ADD)

ADD Agency for Defense Development

Project page

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

Knowledge distillation (KD)

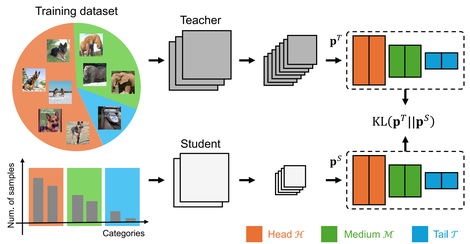
- Model compression: large teacher (T) → compact student (S)
- Conventional KD assumes **balanced datasets.**

Can a teacher trained on imbalanced data still offer trustworthy supervision?

Long-tailed dataset

- Common in real-world data
- Head class biased teacher
- Poor performance on tail classes (∵ insufficient exposure)

→ ***Failure of standard KD***



Preliminaries

Notation and definition

- For classification with C classes, predictive probability vector, $\mathbf{p} = [p_1, p_2, \dots, p_C] \in \mathbb{R}^C$

$$p_i = \sigma(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{C}\exp(z_j)}$$

- Under long-tailed distributions, $C \to \mathcal{G} \in \{\mathcal{H}, \mathcal{M}, \mathcal{T}\}$

$$\mathrm{KD} = \mathrm{KL}(\mathbf{p}^T \,\|\, \mathbf{p}^S) = \sum_{\mathcal{G}}\sum_{i\in\mathcal{G}} p_i^T \log\left(\frac{p_i^T}{p_i^S}\right)$$

- Cross-group probability, $\mathbf{p}_{\mathcal{G}} = [p_{\mathcal{H}}, p_{\mathcal{M}}, p_{\mathcal{T}}] \in \mathbb{R}^3$

$$p_{\mathcal{G}} = \frac{\sum_{i\in\mathcal{G}}\exp(z_i)}{\sum_{j=1}^{C}\exp(z_j)}$$

- Within-group probability, $\tilde{\mathbf{p}}_{\mathcal{G}} = \left[\tilde{p}_{\mathcal{G}_1}, \tilde{p}_{\mathcal{G}_2}, \dots, \tilde{p}_{\mathcal{G}_i}\right]_{i\in\mathcal{G}} \in \mathbb{R}^{|\mathcal{G}|}$

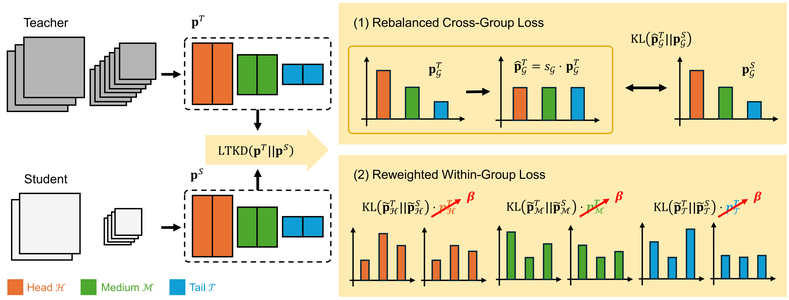
$$\tilde{p}_{\mathcal{G}_i} = \frac{\exp(z_i)}{\sum_{j\in\mathcal{G}}\exp(z_j)}$$

Revisiting KL divergence

$$\mathrm{KD} = \sum_{\mathcal{G}}\sum_{i\in\mathcal{G}} p_i^T \log\left(\frac{p_{\mathcal{G}}^T}{p_{\mathcal{G}}^S}\right) + \sum_{\mathcal{G}}\sum_{i\in\mathcal{G}} p_i^T \log\left(\frac{\tilde{p}_{\mathcal{G}_i}^T}{\tilde{p}_{\mathcal{G}_i}^S}\right)$$
$$= \sum_{\mathcal{G}} p_{\mathcal{G}}^T \log\left(\frac{p_{\mathcal{G}}^T}{p_{\mathcal{G}}^S}\right) + \sum_{\mathcal{G}} p_{\mathcal{G}}^T \sum_{i\in\mathcal{G}} \tilde{p}_{\mathcal{G}_i}^T \log\left(\frac{\tilde{p}_{\mathcal{G}_i}^T}{\tilde{p}_{\mathcal{G}_i}^S}\right)$$

$$\therefore \mathrm{KD} = \mathrm{KL}(\mathbf{p}_{\mathcal{G}}^T \,\|\, \mathbf{p}_{\mathcal{G}}^S) + \sum_{\mathcal{G}} p_{\mathcal{G}}^T \cdot \mathrm{KL}(\tilde{\mathbf{p}}_{\mathcal{G}}^T \,\|\, \tilde{\mathbf{p}}_{\mathcal{G}}^S)$$

Long-tailed knowledge distillation (LTKD)



$$\mathrm{LTKD} = \alpha \cdot \mathrm{KL}\left(\tilde{\mathbf{p}}_{\mathcal{G}}^T \,\middle\|\, \mathbf{p}_{\mathcal{G}}^S\right) + \beta \cdot \sum_{\mathcal{G}} \mathrm{KL}\left(\tilde{\mathbf{p}}_{\mathcal{G}}^T \,\middle\|\, \tilde{\mathbf{p}}_{\mathcal{G}}^S\right)$$

Rebalanced cross-group loss ***Reweighted within-group loss***

Rebalanced cross-group loss

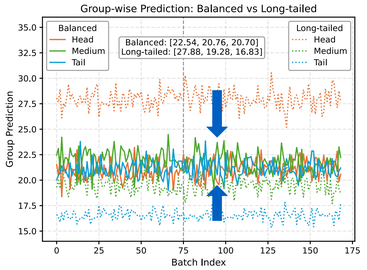
- **Rebalancing before distillation** w/ **scaling factors** for each group

$$s_{\mathcal{H}} = \frac{p_{\text{avg}}^B}{p_{\mathcal{H}}^B}, s_{\mathcal{M}} = \frac{p_{\text{avg}}^B}{p_{\mathcal{M}}^B}, s_{\mathcal{T}} = \frac{p_{\text{avg}}^B}{p_{\mathcal{T}}^B}$$

$$\mathbf{p}_{\text{batch}} = \left[p_{\mathcal{H}}^B, p_{\mathcal{M}}^B, p_{\mathcal{T}}^B\right],\ p_{\text{avg}}^B = \text{Mean}\left(p_{\mathcal{H}}^B, p_{\mathcal{M}}^B, p_{\mathcal{T}}^B\right)$$

- After **normalization** (for valid probability)

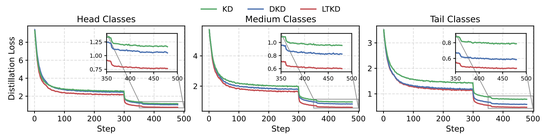
$$\tilde{\mathbf{p}}_{\mathcal{G}}^T = \left[\frac{s_{\mathcal{H}} p_{\mathcal{H}}^T}{\sum_{\mathcal{G}} s_{\mathcal{G}} p_{\mathcal{G}}^T}, \frac{s_{\mathcal{M}} p_{\mathcal{M}}^T}{\sum_{\mathcal{G}} s_{\mathcal{G}} p_{\mathcal{G}}^T}, \frac{s_{\mathcal{T}} p_{\mathcal{T}}^T}{\sum_{\mathcal{G}} s_{\mathcal{G}} p_{\mathcal{G}}^T}\right]$$



Reweighted within-group loss

$$\sum_{\mathcal{G}} p_{\mathcal{G}}^T \cdot \mathrm{KL}(\tilde{\mathbf{p}}_{\mathcal{G}}^T \,\|\, \tilde{\mathbf{p}}_{\mathcal{G}}^S) = p_{\mathcal{H}}^T \mathrm{KL}(\tilde{\mathbf{p}}_{\mathcal{H}}^T \,\|\, \tilde{\mathbf{p}}_{\mathcal{H}}^S) + p_{\mathcal{M}}^T \mathrm{KL}(\tilde{\mathbf{p}}_{\mathcal{M}}^T \,\|\, \tilde{\mathbf{p}}_{\mathcal{M}}^S) + p_{\mathcal{T}}^T \mathrm{KL}(\tilde{\mathbf{p}}_{\mathcal{T}}^T \,\|\, \tilde{\mathbf{p}}_{\mathcal{T}}^S)$$

- Weaker supervision for tail classes → suboptimal convergence



- **Equal importance** to all groups by replacing $p_{\mathcal{G}}^T$ with **uniform constant** $\boldsymbol{\beta}$

Experiments

Main results on ImageNet-LT

- **Consistent improvements** on the **large-scale, imbalanced datasets** as well as general benchmarks such as CIFAR-100-LT and TinyImageNet-LT

T-S Pairs	ResNet34 − ResNet18						ResNet50 − MobileNetV1					
γ	5		10		20		5		10		20	
Group	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All
Teacher	57.75	67.61	50.61	63.85	43.00	59.91	59.62	69.06	52.69	65.57	44.82	61.46
Student	54.48	64.73	47.45	61.18	39.98	57.25	55.42	65.46	49.49	62.55	41.68	58.94
KD [16]	56.14	66.27	49.35	63.03	41.72	59.15	56.58	66.51	50.42	63.70	42.45	59.91
DKD [52]	56.83	66.84	49.95	63.50	42.65	59.82	58.54	68.09	52.39	65.04	45.04	61.46
ReviewKD [6]	57.27	66.96	50.80	63.72	43.48	60.15	58.59	68.10	53.06	65.31	45.35	61.84
DIST [17]	56.79	66.66	50.28	63.56	42.78	59.94	57.29	66.94	50.89	64.09	43.70	60.71
CAT-KD [12]	55.92	66.14	49.58	63.20	42.20	59.67	57.08	66.96	51.00	64.07	43.55	60.54
LTKD	**58.33**	**67.23**	**52.55**	**64.29**	**45.88**	**60.80**	**60.17**	**68.48**	**54.40**	**65.52**	**48.55**	**62.22**
Δ	**+1.06**	**+0.27**	**+1.75**	**+0.57**	**+2.40**	**+0.65**	**+1.58**	**+0.38**	**+1.34**	**+0.21**	**+3.20**	**+0.38**

Ablation study

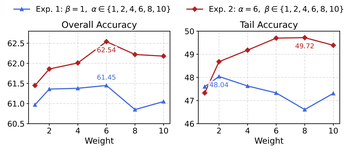
- **Cross-group loss:** Non-rebalanced $\left(\mathbf{p}_{\mathcal{G}}^T\right)$ *vs.* rebalanced $\left(\tilde{\mathbf{p}}_{\mathcal{G}}^T\right)$

Models	Teacher		ResNet32×4				VGG13				WRN–40–2			
	Student		ResNet8×4		ShuffleNetV1		VGG8		MobileNetV2		WRN–40–1		ShuffleNetV1	
Loss	Cross	Within	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All
Biased	✓	✗	38.30	55.77	30.01	**48.82**	35.55	53.30	23.42	40.29	34.64	53.97	30.23	49.31
Ours	✓	✗	**40.51**	**56.55**	**30.68**	48.81	**37.26**	**53.70**	**23.97**	**40.30**	**35.23**	**54.35**	**32.44**	**49.91**
	Δ		**+2.21**	**+0.78**	**+0.67**	-0.01	**+1.71**	**+0.40**	**+0.55**	**+0.01**	**+0.59**	**+0.38**	**+2.21**	**+0.60**

- **Complementary cross- and within-group components**

Models	Teacher		ResNet32×4				VGG13				WRN–40–2			
	Student		ResNet8×4		ShuffleNetV1		VGG8		MobileNetV2		WRN–40–1		ShuffleNetV1	
Loss	Cross	Within	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All
Baseline	✗	✗	36.81	57.41	36.13	55.03	37.29	56.13	29.05	47.69	35.63	56.71	38.30	56.61
	✓	✗	40.51	56.55	30.68	48.81	37.26	53.70	23.97	40.30	35.23	54.35	32.44	49.91
Ours	✗	✓	42.34	59.78	39.10	56.84	38.54	56.83	31.99	49.20	39.67	58.11	40.45	57.65
	✓	✓	**49.70**	**62.54**	**45.94**	**59.62**	**45.77**	**58.86**	**38.33**	**52.03**	**45.74**	**59.91**	**48.42**	**60.94**

- **Hyperparameters** & **Number of groups** $\boldsymbol{n(\mathcal{G})}$



Continuous reweighting

$n(\mathcal{G})$	3	4	5	10	20	25	50	100
R32×4−R8×4	51.08	51.08	51.10	**51.14**	50.99	50.34	50.06	50.41
VGG13−VGG8	47.66	47.85	48.06	48.26	**48.69**	48.58	48.19	47.82
WRN402−SV1	48.60	48.98	49.03	49.27	49.54	**49.90**	49.64	47.95
R50−MV2	42.45	43.01	42.99	43.40	**43.43**	43.42	42.51	41.18