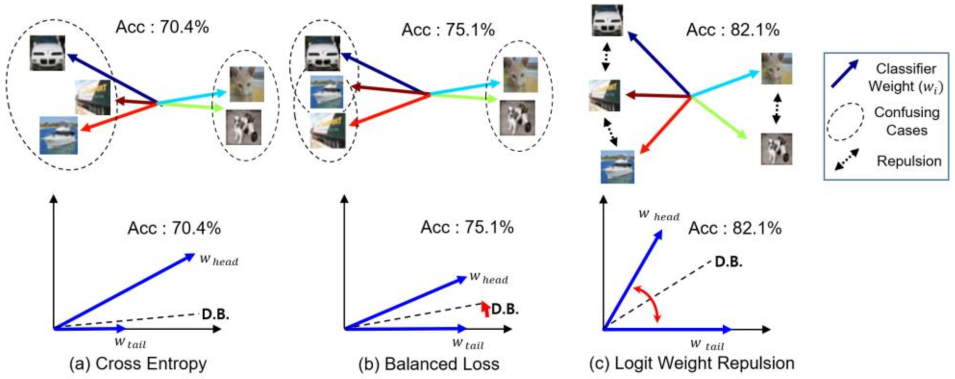
Novel Regularization via Logit Weight Repulsion for Long-Tailed Classification

TaeGil Ha, SeulKi Park, Jin Young Choi
Department of Electrical and Computer Engineering, Seoul National University
clfdydmlarnl@snu.ac.kr, seulki.park@snu.ac.kr, jychoi@snu.ac.kr

Perception Intelligence Laboratory

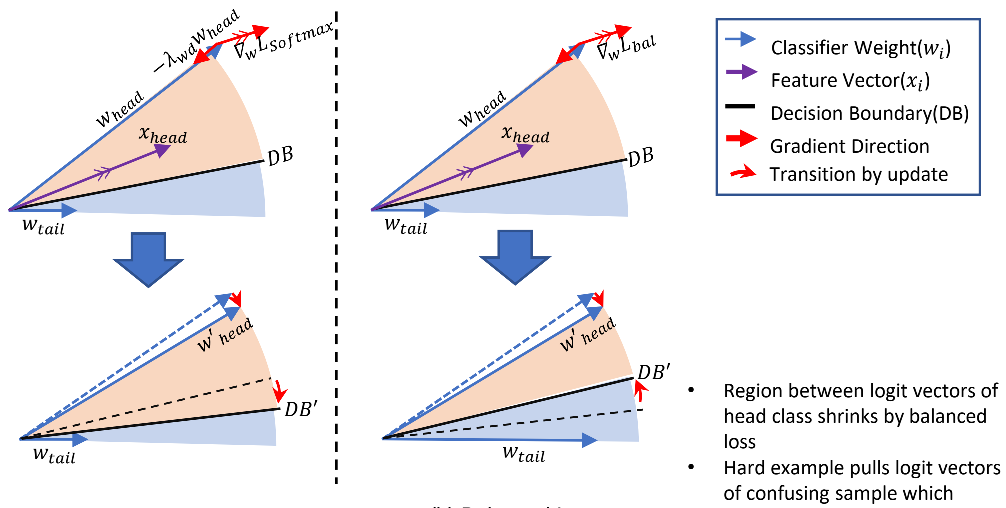
BMVC 2023

Motivation



- Problem & Key Idea
  - In scenarios where classes have imbalanced distributions, logit weight vectors sharing similar class features tend to influence each other, leading to subpar performance in the tail class.
  - Balanced loss seeks to correct the skewed norm of logit weight vectors, alleviating bias in favor of the head class. However, it introduces more pronounced pulling gradients compared to standard cross-entropy, which can negatively impact the performance of the head class.
  - To address this, we aim to create repulsion between logit weight vectors, ensuring a clear separation between them. This strategy enhances the performance of both the head and tail classes simultaneously.

Problematic Behavior of Balanced Loss



(a) Softmax Loss

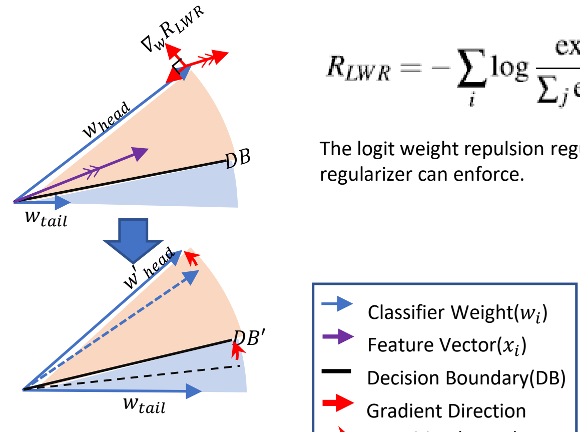
(b) Balanced Loss (Softmax + Adjustment)

- Region between logit vectors of head class shrinks by balanced loss
- Hard example pulls logit vectors of confusing sample which shrinks region between logit vectors
- All gradient terms are perpendicular to the logit weight direction or feature direction

$$L_{\text{bal}} = -C_i \log \frac{\exp(w_i \cdot x_i + \delta_i)}{\sum_j \exp(w_j \cdot x_i + \delta_{i,j})} + \lambda_{wd} ||w_i||_2^2$$

$$\Rightarrow \nabla_{w_i} L_{\text{bal}} = C_i x_i \left(1 - \frac{\exp(w_i \cdot x_i + \delta_{i,i})}{\sum_j \exp(w_j \cdot x_i + \delta_{i,j})}\right) + C_i \nabla_{w_i} \delta_i - C_i \sum_j \nabla_{w_i} \delta_{i,j} \frac{\exp(w_j \cdot x_i + \delta_{i,j})}{\sum_j \exp(w_j \cdot x_i + \delta_{i,j})} - 2\lambda_{wd} w_i$$

Logit Weight Repulsion Regularizer



(c) Ours
(Balanced Loss + Repulsion)

$$R_{LWR} = -\sum_i \log \frac{\exp(\overline{w_i} \cdot \overline{w_j})}{\sum_j \exp(\overline{w_i} \cdot \overline{w_j})} \Rightarrow -\nabla_{w_i} R_{LWR} = -\sum_{j \neq i} C_{i,j} \left(\frac{\overline{w_j} - (\overline{w_i} \cdot \overline{w_j})\overline{w_i}}{||w_i||_2} \right)$$

The logit weight repulsion regularizer introduces a gradient direction that neither the classification loss nor weight decay regularizer can enforce.

Training Details

Repulsion Threshold θ

$$R_{LWR} = -\sum_i \left(\log \frac{\exp(1_{i,j,\theta} \overline{w_i} \cdot \overline{w_j})}{\sum_j \exp(1_{i,j,\theta} \overline{w_j} \cdot \overline{w_j})} \right)$$

Where

$$1_{i,j,\theta} = \begin{cases} 1 & \text{if } \overline{w}_i \cdot \overline{w}_j > \theta \\ 0 & \text{otherwise.} \end{cases}$$

Scheduled Training

$$L_{bal} + \lambda(t) R_{LWR}$$

Where

$$\lambda(t) = (1 - t/T)^2$$

Experiments

Compatibility with Various Balanced Losses

Methods	All	Few	Medium	Many
CE-DRW [5]	47.6	28.0	44.9	57.6
$+R_{LWR}$	50.4(+2.8)	30.0(+2.0)	47.7(+2.8)	60.4(+2.8)
BS [31]	48.7	24.0	46.2	60.5
$+R_{LWR}$	51.5(+2.8)	30.7(+6.7)	49.2(+3.0)	62.6(+2.1)
Focal [20]	38.0	11.2	31.0	56.3
$+R_{LWR}$	41.0(+3.0)	13.7(+2.5)	34.1(+3.1)	58.6(+2.3)
Logit Adjustment (LA) [26]	48.0	29.1	44.4	58.5
$+R_{LWR}$	50.1(+2.1)	31.5(+2.4)	46.9(+2.5)	60.4(+1.9)

Ablation Studies

Methods	Few	Medium	Many	All
Balanced Softmax [31]	25.0	46.7	63.9	46.3
$+R_{LWR}$	28.8	**51.6**	64.0	49.3
$+R_{LWR}$ w/ proj	30.0	49.9	65.3	49.5
$+\lambda(t)R_{LWR}$ w/ proj	**32.2**	50.0	**66.6**	**50.7**

θ	Few	Medium	Many	All
−1.0	30.0	50.3	65.3	50.1
0	29.9	51.3	65.5	50.0
0.2	30.3	**51.4**	65.5	50.2
0.4	**32.2**	50.0	**66.6**	**50.7**
0.6	31.3	49.7	65.5	49.9
1.0	25.0	46.7	63.9	46.3

Comparison with the State-of-the-arts

Methods	Architecture	All	Few	Medium	Many
Focal [20]	ResNet-50	38.0	11.2	31.0	56.3
BBN§ [46]	ResNeXt	41.2	40.8	43.3	40
Softmax	ResNet-50	41.6	5.8	33.8	64.0
UNO-IC§ [34]	ResNeXt	45.7	9.3	38.7	**66.3**
OLTR§ [22]	ResNeXt	46.7	19.5	45.5	58.2
LFME§ [41]	ResNeXt	47.0	22.0	43.5	60.6
ESQL§ [32]	ResNeXt	47.3	15.7	44	62.5
cRT# [17]	ResNet-50	47.3	26.1	44.0	58.8
CE-DRW [5]	ResNet-50	47.6	28.0	44.9	57.6
LWS# [17]	ResNet-50	47.7	29.3	45.2	57.1
LA [26]	ResNet-50	48.0	29.1	44.4	58.5
BS [31]	ResNet-50	48.7	24.0	46.2	60.5
LDAM-DRW# [5]	ResNet-50	49.8	**30.7**	46.9	60.4
Ours(R_{LWR} + BS [31])	ResNet-50	**51.5**	**30.7**	**49.2**	62.6

Methods	All IF=100	All IF=50	All IF=10	IF=100 Few	IF=100 Medium	IF=100 Many	IF=100 All
Softmax	38.6	44.0	56.4	8.7	37.6	65.3	38.6
CBL† [10]	39.6	45.4	58.0	-	-	-	39.6
IB-Loss [28]	39.8	46.4	50.4	20.4	44.9	50.3	39.8
Focal [20]	41.9	48.2	59.8	10.9	41.3	**68.7**	41.9
BBN†[46]	42.6	47.0	59.1	-	-	-	42.6
UNO-IC† [34]	43.1	-	58.6	-	-	-	43.1
SEQL† [32]	43.4	-	-	-	-	-	43.4
LFME† [41]	43.8	-	-	28.0	-	59.5	43.8
BS[31]	46.3	51.2	61.5	25.0	46.7	63.9	46.3
LA [26]	46.5	-	-	24.4	47.1	63.6	46.5
LDAM-DRW[5]	46.6	51.2	59.6	22.8	48.5	64.4	46.6
Ours(R_{LWR} + BS [31])	**50.7**	**54.3**	**63.6**	**32.2**	**50.0**	66.6	**50.7**