Feature Fusion from Head to Tail for Long-Tailed Visual Recognition

Mengke Li[1,2] Zhikai Hu[3] Yang Lu[4] Weichao Lan[3] Yiu-ming Cheung[3] Hui Huang[2*]

[1]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), China [2]Shenzhen University, China [3]Hong Kong Baptist University, China [4]Xiamen University, China

Introduction & Motivation

Problem:

- The imbalanced distribution of long-tailed data presents a considerable challenge for deep learning models, as it causes them to prioritize the accurate classification of head classes but largely disregard tail classes.
- The biased decision boundary, resulting from insufficient semantic information in tail classes, is a key factor contributing to their low recognition accuracy.

Existing methods:

- Primarily focus on training a new model to acquire a relatively balanced embedding space and/or assembling multiple diverse networks.
- Neglecting the optimization of the classifier, also referred to as the decision boundary, is crucial and hinders the full realization of the potential of the acquired backbone, as illustrated in Figure 1 (a).

Motivation:

- Enhance the diversity of tail class samples to prevent overfitting.
- Simulate the potential unseen samples to adjust the decision boundary.

Figure 1. Decision boundaries comparison.

Overview of H2T:

Head-to-tail fusion (H2T) enriches the sparse tail class semantics and calibrate the bias in tail classes, which grafts partial semantics from the head class on the tail class. Transferring the head semantics can effectively fill the tail semantic area and the category overlap, which compels the decision boundary to shift closer to a more optimal one, as shown in Figure 1 (b).

- More details at: https://arxiv.org/abs/2306.06963
- Project homepage: https://vcc.tech/research/2024/H2T
- Contact: huihuang@szu.edu.cn; limengke@gml.ac.cn;

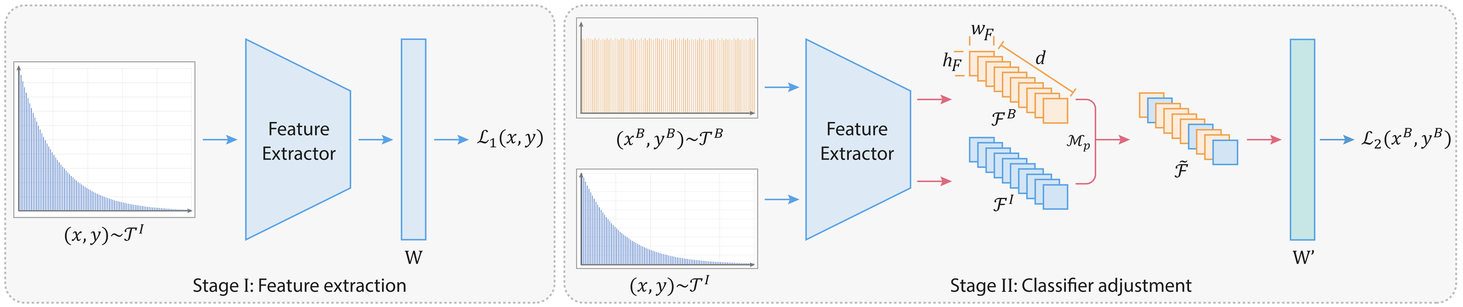

Figure 2. Framework of head-to-tail fusion (H2T).

Fusing Head Features to Tail

Methodology:

We fuse the features of head classes to the tail to exploit the abundant closet semantic information. This operation can enrich the tail classes and expand their embedding space distribution. The proposed framework is illustrated in Figure 2. The fusion process is formulated as:

$$\tilde{\mathcal{F}} = \mathcal{M}_p \otimes \mathcal{F}_t + \overline{\mathcal{M}}_p \otimes \mathcal{F}_h. \quad (1)$$

- $\mathcal{F}_t$ and $\mathcal{F}_h$: Feature maps of tail and head classes, respectively.
- $\mathcal{M}_p$: mask stacked with multiple 1 matrices and 0 matrices.
- $\overline{\mathcal{M}p}$: The complement of $\mathcal{M}p$, that is, the indices of 0 matrices in $\overline{\mathcal{M}}p$ correspond to the 1 matrix in $\mathcal{M}p$, and vice versa.
- d: The total number of all the 1 and 0 matrices.
- The subscript p of the mask matrices represents the fusion ratio.

Fused feature $\tilde{\mathcal{F}}$ is then passed through a pooling layer and classifier to predict the corresponding logits $\mathbf{z} = [z_0, z_1, \cdots, z_{C-1}]$. Different loss functions, such as CE loss, MisLAS, or GCL, to name a few, can be adopted. The backbone ϕ can be the single model as well as the multi-expert model. We exploit the two-stage training and apply H2T in stage II. The framework is shown in Figure 2.

Rationale Analysis:

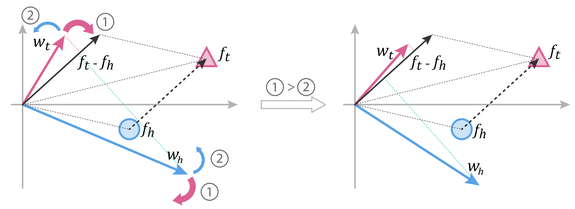

Figure 3. Rationale analysis of H2T.

H2T generates two forces ① and ② among features. Force ① > ② makes the tail sample to "pull" closer to w_t and "push" further away from w_h, leading to the adjustment of decision boundary and enlargement of the tail class space.

Comparison Results

Table 1. Comparison results on imageNet-LT.

Method	Head	Med	Tail	All
Single Model				
Decouple	62.93	49.77	33.26	52.18
MisLAS	62.53	49.82	34.74	52.29
GCL	62.24	48.62	52.12	**54.51**
BSCE+CMO	62.00	49.10	36.70	52.30
Decouple+H2T	63.26	50.43	34.11	52.74
MisLAS+H2T	62.42	51.07	35.36	52.90
GCL+H2T	62.36	48.75	52.15	**54.62**
Multi-Expert Model				
RIDE	69.59	53.06	30.09	55.72
RIDE+CMO	66.40	53.90	35.60	**56.20**
ResLT	59.39	50.97	41.29	52.66
RIDE+H2T	67.55	54.95	37.08	**56.92**
ResLT+H2T	62.29	52.29	35.31	53.39

Table 2. Comparison results on iNaturalist 2018.

Method	Head	Med	Tail	All
Single Model				
Decouple	72.88	71.15	69.24	70.49
MisLAS	72.52	72.08	70.76	71.54
GCL	66.43	71.66	72.47	71.47
BSCE+CMO	68.80	70.00	72.30	70.90
DR+H2T	71.73	72.32	71.30	**71.81**
MisLAS+H2T	69.68	72.49	72.15	**72.05**
GCL+H2T	67.74	71.92	72.22	71.62
Multi-Expert Model				
RIDE	76.52	74.23	70.45	72.80
RIDE+ CMO	68.70	72.60	73.10	72.80
ResLT	64.85	70.64	72.11	70.69
RIDE+H2T	75.69	74.22	71.36	**73.11**
ResLT with H2T	68.41	72.31	72.09	71.88

Visualization & Ablation Experiments

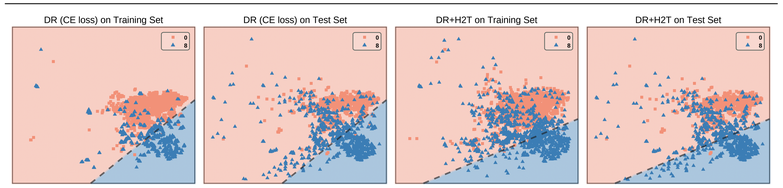

Figure 4. Decision boundary comparison of Class 0 and 8 without H2T v.s. with H2T.

Figure 4 shows the t-SNE visualization of the distribution in embedding space and the decision boundary, which demonstrates our motivation (i.e., H2T can enrich tail classes and calibrate the decision boundary).

Concluding Remarks

By virtue of this fusion operation, our proposed H2T has two-fold advantages:

- Produce relatively abundant features to augment tail classes.
- Generate two opposing forces that restrain each other, preventing excessive sacrifices in head class accuracy.

Limitation: Slightly compromising the performance of the head class.