HMT-NET: HEAD-MIDDLE-TAIL NETWORK FOR LONG-TAILED MULTI-LABEL CLASSIFICATION

QIN WANG[1], XI-ZHAO WANG[1,2]

[1]Big Data Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, China.
[2]The Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, 518060, China
Correspondence to: Xizhao Wang (xizhaowang@ieee.org)

Abstract:

Real-world data often exhibits serious class-imbalance and is associated with multi-label settings. To manipulate the deep model adaptable to this scenario, existing works intend to perform conventional re-sampling for balanced predictions. However, the label co-occurrence and over-fitting in tail classes typically limit their generalization ability. In this work, we proposed a novel approach named Head-Middle-Tail Network (HMT-Net) by assembling the training of different class groups and assigning group-wise feature augmentation. Specifically, HMT-Net leverages class frequency to partition the label space into head, middle, and tail groups and applies group-specific perturbations to the corresponding features, where stronger perturbations are introduced for tail classes to enhance their representational diversity. Experiments on widely used benchmarks verify the performance of HMT-Net is comparable with the SOTAs.

Keywords:

Long-tailed learning; Class-imbalance; Multi-label classification;

1. Introduction

The rapid advancement of artificial intelligence has led to significant successes in deep learning models across various domains of biomedical informatics [?, ?, ?, ?]. However, there are two major limitations in the training of these models: first, the training data is often assumed to be class-balanced; and second, it is typically assumed that each sample corresponds to a single label. In reality, biomedical data rarely aligns with these assumptions; it is inherently imbalanced and often associated with multiple co-occurring labels [?]. Under these more complex and realistic conditions, existing models frequently experience performance degradation, particularly when it comes to identifying rare classes and capturing the interdependence among multiple labels. This situation underscores the need for more robust approaches that can effectively manage long-tail distributions and multi-label scenarios in biomedical settings.

Long-tailed multi-label problems are common in biomedical informatics. To address this, several studies have been proposed to enhance the representational capacity of models. For example, Debby et al. conducted a comprehensive review on structure-based protein-ligand binding affinity prediction, with a particular focus on molecular representation techniques [?]. Wang et al. introduced Intermolecular Contact Profiles (IMCPs), a concise and interpretable descriptor that significantly enhances feature expressiveness for protein-ligand complexes [?]. Wang et al. proposed proteo-chemometrics interaction fingerprints (PrtCmm IFPs) to enhance feature representation and improve predictive performance in protein–ligand binding affinity modeling [?]. In disease prediction using clinical records, models are typically expected to predict one or more disease labels. Although common diseases such as diabetes and hypertension are prominent in training data, rare diseases, such as rare genetic disorders, occur infrequently but must still be regarded as valid labels.

To address these challenges, recent studies have utilized re-sampling to improve skewed predictions by increasing the sampling probability of tail classes during mini-batch construction [?]. However, this line of approaches faces two significant limitations in multi-label scenarios. In multi-label settings, a single instance can belong to both head and tail classes simultaneously. When we replicate these instances to over-sample tail classes, we inadver-

tently increase the frequency of head classes as well, which fails to achieve the balance in label distribution. Also, re-sampling merely reinforces learning from the limited set of examples and does not fundamentally enhance the feature space of tail classes. This can lead to overfitting on tail classes, particularly when there is a severe lack of data.

In this paper, we propose a novel framework named Head-Middle-Tail Network (HMT-Net). By assigning independent parameter spaces for different class-groups, HMT-Net significantly reduces the gradient interference from head classes on the learning of tail classes. Additionally, it enhances the diversity of features for tail classes in a more natural manner while avoiding label co-occurrence. Our contributions can be summarized as follows:

- We introduce HMT-Net, a new architecture specifically designed for long-tailed multi-label learning, which simultaneously considers both label frequency and co-occurrence.
- We implement a group-aware parameter decoupling strategy that eliminates the feature entanglement between head and tail classes.
- Extensive experiments are conducted across multiple benchmark datasets, demonstrating the effectiveness of proposed HMT-Net.

The rest of this work is organized as follows: In section 2, a review of related works is discussed. In section 3, our proposed HMT-Net is described in detail. In section 4, the experimental analysis of our model and related comparisons are given. Finally, a summary is outlined in section 5.

2. Related Work

Long-tailed learning has been a persistent challenge, traditionally focusing on single-label data. However, in recent years, the study of long-tailed multi-label learning has gained attention to better reflect real-world conditions.

2.1 Long-tailed Single-label Classification

Typical works for addressing the long-tailed single-label classification can be divided into three main approaches: (1) Re-sampling: The re-sampling-based methods encourage constructing balanced mini-batches by either over-sampling tail classes or under-sampling the head classes. For instance, [?] simply duplicates samples from the minority class to enhance the learning process for tail classes. [?] proposed removing partial instances in head-class spaces to yield balanced predictions. (2) Re-weighting. The re-weighting approaches aim to rescale the loss function according to the varying frequencies of different classes. [?] suggested adjusting costs in proportion to the inverse frequency of each class. (3) Two-stage fine-turning. This line of research divides feature learning and classifier learning into two distinct stages. In [?], a framework named ProCo performs contrastive learning using different global views of images, and a linear classifier is fine-tuned on top of the learned backbone.

2.2 Long-tailed Multi-label Classification

In real-world applications, an image could be associated with multiple labels. In such scenarios, long-tailed problems are more complicated when faced with label co-occurrence and the over-suppression of negative gradients. Consequently, long-tailed multi-label classification has attracted increasing attention in recent years. [?] proposed the DB loss, which innovates re-sampling in multi-label data while considering the negative gradient constraint. Inspired by re-weighting, some cost-sensitive methods are introduced such as DR loss [?]. Also, [?] assemble models to generating balanced predictions, to generate balanced predictions, where individual models handle classes with similar instance scales.

3. Methodology

Details for HMT-Net can be referred in Figure.2. Suppose the used dataset is $\mathcal{D} = \{(\mathrm{x}^i, \mathrm{y}^i)\}_{i=1}^N$, where N is the number of training samples, $\mathrm{x}^i \in \mathbb{R}^d$ represents the input feature, and $\mathrm{y}^i = \left[y_1^i, y_2^i, ..., y_C^i\right] \in \{0,1\}^C$ is the corresponding label with C denoting the total number of classes. Let $n_k = \sum_{i=1}^N y_k^i$ denote the number of training examples that contain class k. It should be noted that $N \leq \sum_{i=1}^C n_i$ since a single input could associated with multiple labels. The goal of multi-label classification is to learn a mapping $f : \mathbb{R}^d \rightarrow \{0,1\}^C$ that predicts the probability of each label for a given input.

3.1 Class Partition

Different from conventional multi-label classification methods, HMT-Net partitions the classes into head, mid-

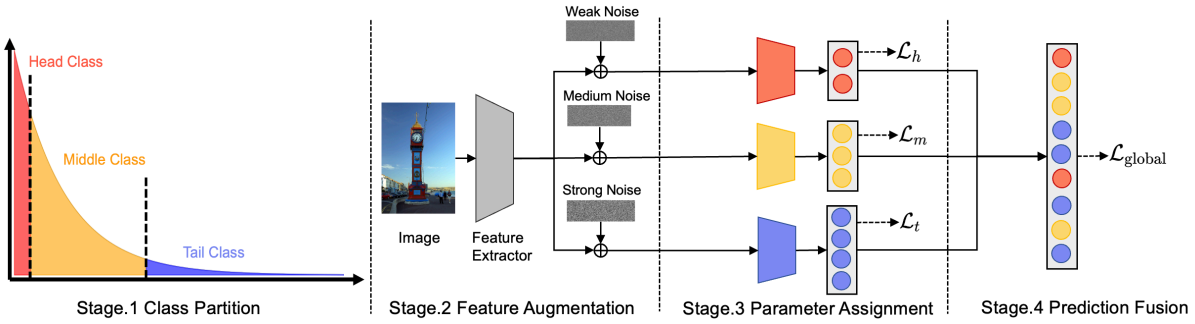

FIGURE 1. The framework of proposed BENet.

dle, and tail classes based on the number of training samples associated with each class. Specifically, let τ_h and τ_m be the thresholds for distinguishing head, middle, and tail classes, where $\tau_h > \tau_m$. Based on the thresholds, we define the sets of head, middle, and tail classes as follows: 1) Head class set: $\mathcal{C}_h = \{k \mid n_k \geq \tau_h\}$, containing classes with sufficient training samples. 2) Middle class set: $\mathcal{C}_m = \{k \mid \tau_m \leq n_k < \tau_h\}$, containing classes with a moderate number of training samples. 3) Tail class set: $\mathcal{C}_t = \{k \mid n_k < \tau_m\}$, containing classes with little training samples. With the above partition, we can subsequently allocate different parameter spaces to the classes in each group. The partition of different groups is commonly used in the single-label long-tailed problem. For the first time, HMT-Net offers a solution to apply the concept of class partition in multi-label long-tailed problems.

3.2 Feature Augmentation

To enhance the number of training samples for tail classes in multi-label long-tailed classification, existing methods generally utilize re-sampling strategies to increase the likelihood of selecting tail-class instances during the mini-batch construction. However, these methods merely replicate existing samples, resulting in unchanged feature spaces. Consequently, re-sampling-based models often overfit the already-known tail-class samples during training.

To address this limitation, it's essential not only to increase the number of training samples for the tail classes but also to improve the diversity of their feature representations. For the first time, HMT-Net realizes this goal by introducing specialized perturbations for feature spaces, and the perturbations guide the model to learn potential representations of tail classes. Specifically, let $\mathbf{f} \in \mathbb{R}^{d'}$ denote the high-level feature extracted by a backbone. In the former settings, we have already categorized all classes into three groups: head ($\mathcal{C}_h$), middle ($\mathcal{C}_m$), and tail ($\mathcal{C}_t$). Let n_c denote the number of samples in class $c \in \mathcal{C}$. For each group $g \in \{h, m, t\}$, the perturbation coefficient σ_g is defined as Eq.(1), i.e.,

$$\sigma_g = \frac{\lambda}{\frac{1}{|\mathcal{C}_g|}\sum_{c \in \mathcal{C}_g} n_c} \tag{1}$$

where λ is a scaling hyperparameter and $|\mathcal{C}_g|$ is the number of classes in group g. The coefficient guarantees that the perturbation strength is inversely proportional to the average sample count in each group. Based on this coefficient, we define group-specific perturbations as Gaussian noise, as shown in Eq.(2), i.e.,

$$\boldsymbol{\delta}_g \sim \mathcal{N}(0, \sigma_g^2 \mathbf{I}), \quad g \in \{h, m, t\} \tag{2}$$

The perturbed high-level features for each group are defined as Eq.(3), i.e.,

$$\tilde{\mathbf{f}}_g = \mathbf{f} + \boldsymbol{\delta}_g \tag{3}$$

These augmented features for each group are passed through their respective classifier branches, allowing the model to learn distinctive representations.

3.3 Parameter Assignment

Based on Eq.(3), we apply three levels of Gaussian perturbation to the extracted high-level features. However, in its basic form, these perturbations are applied uniformly across all feature dimensions, without considering the specific group to which the associated features belong. To address this issue, HMT-Net introduces a structure that assigns different classification heads to manage various perturbed features. Each classification head predicts labels

specific to its group, effectively filtering and representing the related features. Specifically, the classification scores of group $g \in \{h, m, t\}$ is defined as Eq.(4), i.e.,

$$\mathbf{s}_g = \mathcal{F}_g(\tilde{\mathbf{f}}_g), \mathbf{s}_g \in \mathbb{R}^{|\mathcal{C}_g|} \tag{4}$$

where $\mathcal{F}_g$ is the classification head for group g, and $\mathbf{s}_g$ is the predicted score vector over the classes in $\mathcal{C}_g$. The next step involves merging the scores specific to each group for inference. Through this structured perturbation-to-head mapping, HMT-Net enables perturbations of different intensities to be explicitly associated with the class groups.

3.4 Prediction Fusion

In realization, for each group, we apply the well-known Binary Cross-Entropy (BCE) loss between the predicted scores and the corresponding ground-truth label vector $\mathbf{y}_g \in \{0, 1\}^{|\mathcal{C}_g|}$, as shown in Eq.(5), i.e.,

$$\mathcal{L}_g = -\frac{1}{|\mathcal{C}_g|} \sum_{i=1}^{|\mathcal{C}_g|} [y_{g,i} \log(s_{g,i}) + (1 - y_{g,i}) \log(1 - s_{g,i})] \tag{5}$$

where $s_{g,i}$ is the predicted probability for the i-th class in group g, and $y_{g,i}$ is the corresponding ground-truth binary label. To facilitate collaborative inference across the entire label space, we combine the outputs from the three heads into a global prediction vector $\hat{\mathbf{s}} \in \mathbb{R}^{|\mathcal{C}|}$. This process involves mapping each dimension of $\mathbf{s}_g$ back to its corresponding class index within the complete label set $\mathcal{C}$. The fused prediction $\hat{\mathbf{s}}$ is compared with the original ground-truth vector $\mathbf{y} \in \{0, 1\}^{|\mathcal{C}|}$ to compute a global BCE loss, as shown in Eq.(6), i.e.,

$$\mathcal{L}_{\text{global}} = -\frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} [y_j \cdot \log(\hat{s}_j) + (1 - y_j) \cdot \log(1 - \hat{s}_j)] \tag{6}$$

The total loss term is formulated as a weighted combination of the global loss and the group-specific losses, as shown in Eq.(7), i.e.,

$$\mathcal{L}_{\text{all}} = \eta \mathcal{L}_{\text{global}} + (1 - \eta) \sum_{g \in \{h,m,t\}} \mathcal{L}_g \tag{7}$$

where $\eta \in [0, 1]$ is a balancing hyperparameter.

4. Experiment

In this section, we assess the effectiveness of HMT-Net through comparative experiments and parameter analysis on long-tailed multi-label datasets.

4.1 Datasets

Two commonly used long-tailed multi-label benchmarks are employed: VOC-LT and COCO-LT. They are subsets artificially sampled from the Pascal Visual Object Classes Challenge (VOC) and Microsoft COCO (MS-COCO), following a Pareto distribution defined by the probability density function $pdf(x) = \alpha \frac{x_{min}^{\alpha}}{x^{\alpha+1}}$. Following [?], VOC-LT is constructed with $\alpha = 6$ and comprises 1,142 images across 20 classes. Within this set, the class with the fewest samples contains only 4 images, while the largest class has 775 samples. We utilize the test set VOC2007, which consists of 4,952 test images, for model evaluation. Similarly, COCO-LT is also created with $\alpha = 6$ and includes 1,909 images spanning 80 classes. The class with the most images contains 1,128, whereas the class with the fewest has merely 6 images. For evaluation, we use the testing set from COCO2017, which comprises 5,000 images.

4.2 Basic Settings

The Resnet50 pretrained on ImageNet is used as backbone. The SGD optimizer is employed with a momentum of 0.9, and weight decay is configured at 1×10^{-4}. The initial learning rate is set to 0.02, utilizing a warm-up learning rate schedule for the first 500 iterations at a ratio of $\frac{1}{3}$. Training with re-sampling runs for 8 epochs, with learning rate decays at the 5th and 7th epochs. Training without re-sampling runs for 80 epochs, with learning rate decays at the 55th and 70th epochs. The re-sampling strategy follows the guidelines in [?]. This training implementation uses PyTorch version 1.10, and the networks are trained on NVIDIA Tesla V100.

4.3 Comparison

Classes with more than 100 samples are identified as head classes, those with 20 to 100 samples are considered medium classes, and categories with fewer than 20 samples are marked as tail classes. We evaluate the mean average precision (mAP) across all classes and also report the mAP for each group of classes.

Results on VOC-LT. The experimental results for VOC-LT are presented in Table 1. HMT-Net demonstrates the best performance among all compared methods, showcasing the superiority of our approach. Specifically, compared to the previous state-of-the-art model, DB Focal,

TABLE 1. The mAP comparison results on VOC-LT

Methods	Head	Mid	Tail	All
ERM	68.91	80.20	65.31	70.86
RW	67.58	82.81	73.96	74.70
RS	70.95	82.94	73.05	75.38
Focal Loss [?]	69.41	81.43	71.56	73.88
LSEP [?]	69.00	79.83	70.88	72.99
ML-GCN [?]	70.14	76.41	62.39	68.92
LDAM [?]	68.73	80.38	69.09	70.73
CB Focal [?]	70.30	83.53	72.74	75.24
Circle Loss [?]	70.00	82.00	73.88	75.20
DB Focal [?]	72.67	83.17	78.75	78.29
ASL Loss [?]	70.70	82.26	76.29	76.40
ZLPR Loss [?]	71.00	82.67	72.38	75.10
BalPoE [?]	69.00	82.17	71.00	73.76
HMT-Net (Ours)	73.44	83.86	79.33	78.92

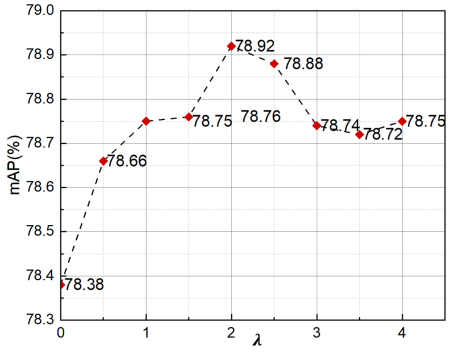

FIGURE 2. Long-tailed multi-label problems in medical areas.

our proposed method improves mAP score by 0.77% for head classes, 0.69% for middle classes, and 0.58% for tail classes. Additionally, when compared to the latest dual-expert system for long-tailed multi-label learning, BalPoE, our framework shows significant improvements of 4.44% for head classes, 1.69% for middle classes, and 8.33% for tail classes. Notably, HMT-Net achieves the highest performance on tail classes, reaching an mAP of 79.33%.

Results on COCO-LT. The results on COCO-LT are presented in Table.2. HMT-Net achieves an overall mAP of 53.80% and demonstrates the best performance among all compared methods. Compared with DB Focal, HMT-Net improves the mAP on tail classes by 0.84%, indicating more balanced predictions. BalPoE also employs different parameter spaces for various distributions, while HMT-Net novelly addresses the overlooked middle classes and emphasizes the feature perturbations on tail classes. Consequently, HMT-Net outperforms BalPoE by a large margin, achieving improvements of 3.78% in overall mAP, 2.64% in head classes, 2.50% in middle classes, and notable 6.40% in tail classes.

TABLE 2. The mAP comparison results on COCO-LT

Methods	Head	Mid	Tail	All
ERM	48.48	49.06	24.25	41.27
RW	48.62	45.80	32.02	42.27
RS [?]	47.58	50.55	41.70	46.97
Focal Loss [?]	49.80	54.77	42.14	49.46
LSEP [?]	46.18	50.91	42.88	47.05
ML-GCN [?]	44.04	48.36	38.96	44.24
LDAM [?]	48.77	48.38	22.92	40.53
CB Focal [?]	47.91	53.01	44.85	49.06
Circle Loss [?]	48.64	55.52	50.28	52.00
DB Focal [?]	50.91	56.58	51.52	53.45
ASL Loss [?]	49.05	53.65	46.68	50.21
ZLPR Loss [?]	47.59	53.73	47.00	49.90
BalPoE [?]	48.45	54.18	45.96	50.02
HMT-Net (Ours)	51.07	56.68	52.36	53.80

4.4 Further Analysis

To investigate the effect of perturbation intensity on model performance, we vary the value of λ, which controls the magnitude of the perturbation, and assess the mAP scores on VOC-LT. The results are presented in Figure.2. When the perturbation is set to 0, the model achieves an mAP of 78.38%, which already exceeds the performance of current methods. This improvement can be attributed to our proposed three-branch architecture, which effectively reduces the gradient dominance of head classes over tail classes during training. As the value of λ increases, the mAP improves significantly, reaching a peak of 78.92% when $\lambda = 2$. However, further increasing the perturbation introduces excessive noise, which disrupts the original feature space.

5. Conclusions

In this paper, we tackled the challenge of long-tailed multi-label learning in biomedical informatics, where conventional re-sampling struggles to resolve the feature redundancy and gradient interference. The proposed HMT-Net leverages group-aware parameter decoupling and targeted feature augmentation to independently optimize head, middle, and tail classes. Extensive experiments on multiple benchmarks demonstrate that HMT-Net outperforms existing methods, especially in accurately predicting tail classes. These findings highlight the potential of HMT-Net for improving diagnosis accuracy in real-world biomedical applications. In future work, we plan to explore adaptive grouping strategies to enhance the transferability of HMT-Net.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grants 62376161 and U24A20322), in part by the Stable Support Project of Shenzhen City (No. 20231122124602001), and in part by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: STG5/E- 103/24-R).