南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Category-Prompt Refined Feature Learning for Long-Tailed Multi-Label Image Classification

Jiexuan Yan
School of Big Data & Software Engineering, Chongqing University
Chongqing, China
jiexuanyan@stu.cqu.edu.cn

Sheng Huang*
School of Big Data & Software Engineering, Chongqing University
Chongqing, China
huangsheng@cqu.edu.cn

NanKun Mu
College of Computer Science, Chongqing University
Chongqing, China
nankun.mu@cqu.edu.cn

Luwen Huangfu
Fowler College of Business, San Diego State University
San Diego, California, USA
lhuangfu@sdsu.edu

Bo Liu
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
kfliubo@gmail.com

ACM MM 2024

Introduction

Long-Tailed Multi-Label image Classification:

In real-world applications, the distribution of different categories often follows a long-tailed pattern, where deep networks tend to underperform on tail classes. Meanwhile, unlike the classical single-label classification, practical scenarios frequently involve images associated with multiple, adding complexity and challenge to the task. To address these issues, an increasing number of works focus on the problem of Long-Tailed Multi-Label image Classification (LTMLC).

Strategies for LT-MLC:

- Resampling the number of samples for each category
- Re-weighting the loss for different categories
- Decoupling the learning of representation and classification head

Motivation

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

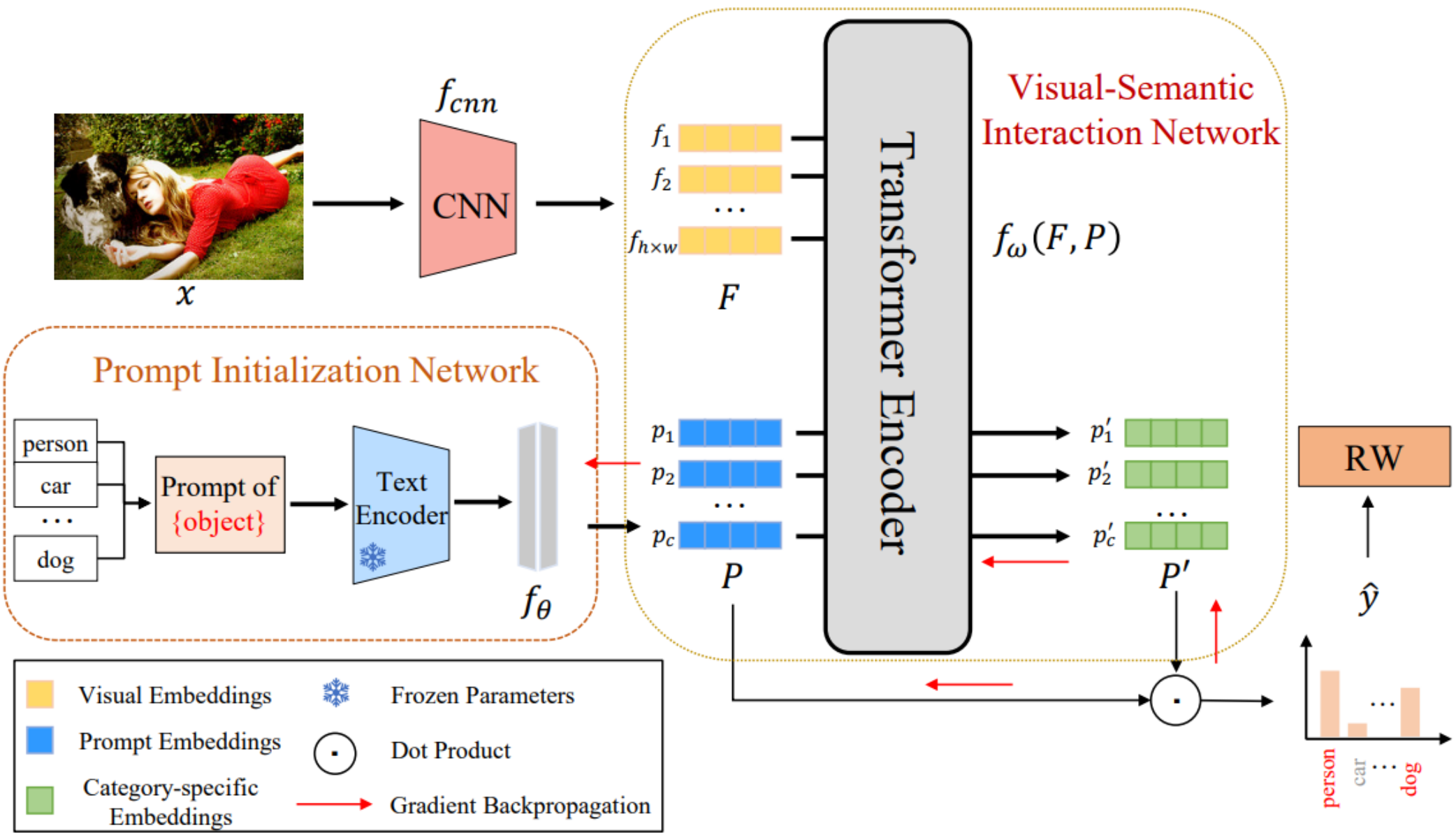
Challenges for LT-MLC:

- It is of great importance to consider the semantic correlations between the head and tail classes in long-tailed learning. Leveraging such correlations can substantially improve the performance of tail classes with the support of head classes Re-weighting the loss for different categories.
- Real-world images often encompass a variety of objects, scenes, or attributes, adding complexity to the classification task. The aforementioned methods typically consider extracting the visual representation of images from a global perspective.

Category-Prompt Refined Feature Learning (CPRFL)

- CPRFL leverages CLIP's text encoder to extract category semantics, thereby enabling the establishment of semantic correlations between the head and tail classes.
- The extracted category semantics are utilized to initialize prompts for all categories, which interact with visual features in order to discern context-related visual information specific to each category.
- Finally, initial prompts lack visual-context information, resulting in a significant data bias between the semantic and visual domains during information interaction. CPRFL introduces a progressive Dual-Path Back-Propagation mechanism to iteratively refine the prompts.

Method


f_{cnn}
CNN
x
f_1
f_2
$f_{h\times w}$
F
Visual-Semantic Interaction Network
$f_\omega(F,P)$
Transformer Encoder
Prompt Initialization Network
person
car
dog
Prompt of {object}
Text Encoder
f_θ
p_1
p_2
p_c
P
p_1'
p_2'
p_c'
P'
RW
$\hat{y}$
person
car
dog
Visual Embeddings
Prompt Embeddings
Category-specific Embeddings
Frozen Parameters
Dot Product
Gradient Backpropagation

Category-Prompt Initialization

Specifically, we design a Prompt Initialize (PI) network, which consists of two fully connected layers followed by a nonlinear activation function. We map the pretrained CLIP's text embedding $\mathcal{W}$ to the initial category-prompts $\mathcal{P} = \{ p_1, p_2, ..., p_c \} \in \mathrm{R}^{c \times d}$.

$$\mathcal{P} = GELU(\mathcal{W}W_1 + b_1)W_2 + b_2$$

Here, $W_1 \in \mathrm{R}^{m \times t}$, $W_2 \in \mathrm{R}^{\mathrm{t} \times d}$, $t = \tau \times d$ and τ is the expansion coefficient controlling the dimension of hidden layers. Typically, τ is set to 0.5 in our experiments.

Visual-Semantic Information Interaction

To facilitate visual-semantic information interaction between category-prompts and visual features, we concatenate the initial category-prompts $\mathcal{P} \in \mathrm{R}^{c \times d}$ with the visual features $\mathcal{F} \in \mathrm{R}^{v \times d}$, forming a combined set of embeddings Z. These embeddings are then input to the VSI network for the visual-semantic information interaction. Within the VSI network, each embedding $z_i \in Z$ undergoes calculation and updating through the multi-head self-attention mechanism inherent to the Transformer encoder.

$$\alpha_{ij}^{p} = softmax\left((W_q p_i)^T (W_k z_i)/\sqrt{d}\right),$$

$$\bar{p}_i = \sum_{j=1} (\alpha_{ij}^{p} W_v z_j),$$

$$p_i' = GELU(\bar{p}_i W_r + b_3) W_o + b_4,$$

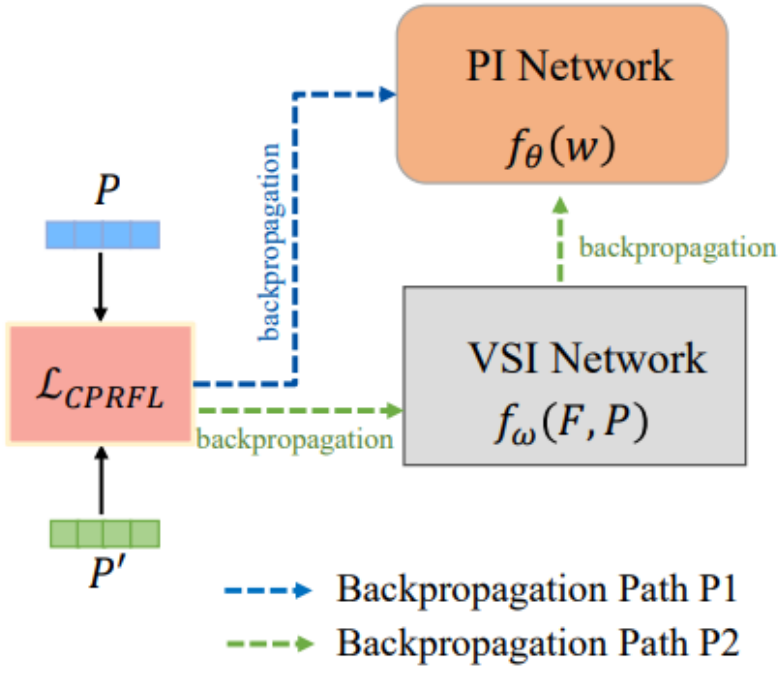
The resulting output of the VSI network and the category-specific visual features are denoted as $Z' = \{ f_1', f_2', f_3', \cdots, f_v', p_1', p_2', \cdots, p_c' \}$ and $\mathcal{P}' = \{ p_1', p_2', \cdots, p_c' \}$.

Method

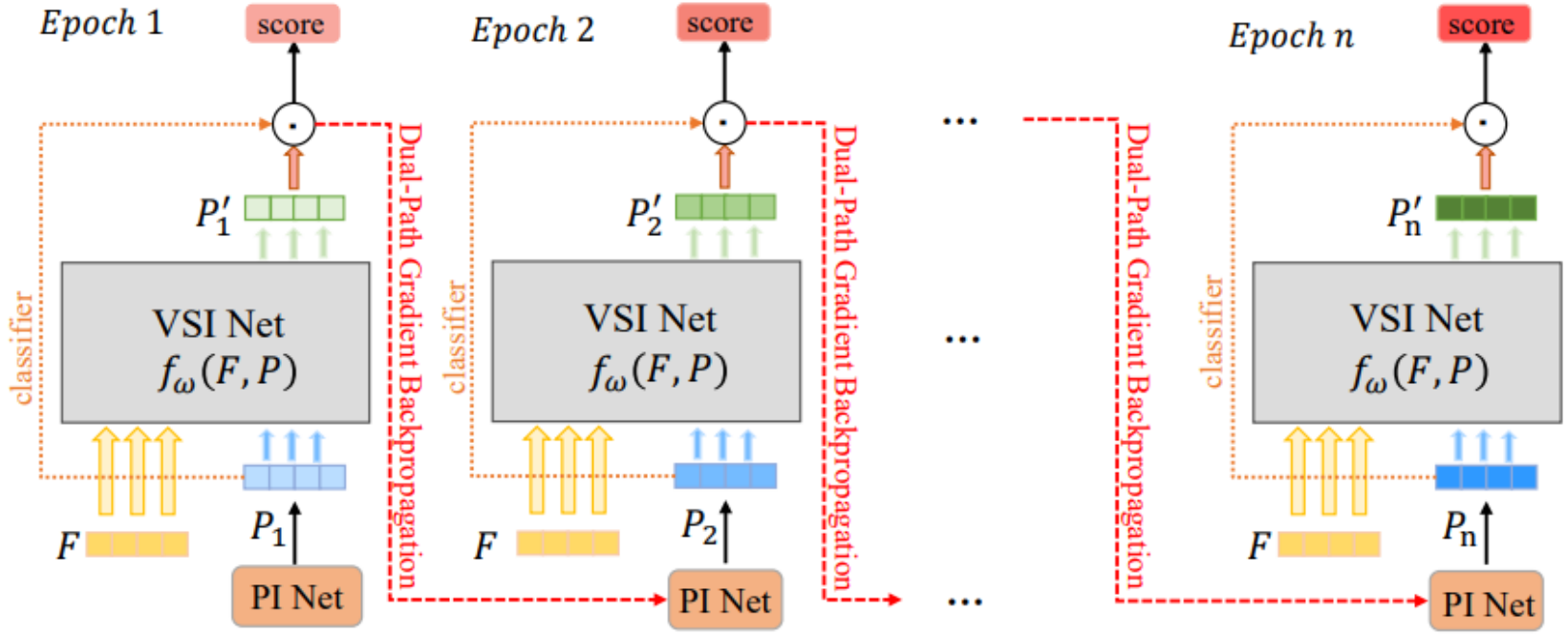
Category-Prompt Refined Feature Learning

The classification probability si for class i can be calculated by

$$s_i = sigmoid(p_i' \cdot p_i).$$



(a) Dual-Path Gradient Back-Propagation



(b) Category-Prompt Refined Feature Learning

Optimization

To further address the negative-positive sample imbalance inherent in multiple categories, we adopt the Asymmetric Loss (ASL) as our optimization objective, which is a variant of focal loss with different γ values for positive and negative samples.

$$\mathcal{L}_{cls} = \mathcal{L}_{ASL} = \sum_{x_i \in X} \sum_{j=1}^{c} \begin{cases} (1 - s_j^i)^{\gamma^+} log(s_j^i), & s_j^i = 1, \\ (\tilde{s}_j^i)^{\gamma^-} log(1 - \tilde{s}_j^i), & s_j^i = 0, \end{cases}$$

where c is the number of classes. $\tilde{s}_j^i$ is the hard threshold in ASL, denoted as $\tilde{s}_j^i = max(s_j^i - \mu, 0)$.

Experiments

Datasets	VOC-LT				COCO-LT			
Methods	total	head	medium	tail	total	head	medium	tail
ERM	70.86	68.91	80.20	65.31	41.27	48.48	49.06	24.25
RW	74.70	67.58	82.81	73.96	42.27	48.62	45.80	32.02
ML-GCN [9]	68.92	70.14	76.41	62.39	44.24	44.04	48.36	38.96
OLTR [34]	71.02	70.31	79.80	64.95	45.83	47.45	50.63	38.05
LDAM [4]	70.73	68.73	80.38	69.09	40.53	48.77	48.38	22.92
CB Focal [11]	75.24	70.30	83.53	72.74	49.06	47.91	53.01	44.85
BBN [56]	73.37	71.31	81.76	68.62	50.00	49.79	53.99	44.91
DB Focal [51]	78.94	73.22	84.18	79.30	53.55	51.13	57.05	51.06
ASL [41]	76.40	70.70	82.26	76.29	50.21	49.05	53.65	46.68
LTML [16]	81.44	75.68	85.53	82.69	56.90	54.13	60.59	54.47
CDRS+AFL [45]	78.96	73.35	85.03	78.63	55.35	52.45	59.48	52.46
Bilateral-TPS [28]	81.58	75.88	84.11	83.95	56.38	55.93	58.26	54.29
PG Loss [29]	80.37	73.67	83.83	82.88	54.43	51.23	57.42	53.40
COMIC [53]	81.53	73.10	89.18	84.53	55.08	49.21	60.08	55.36
CAE-Net [5]	81.61	74.00	85.35	85.28	57.64	52.37	61.18	57.63
CPRFL-GloVe(ours)	85.14	**82.50**	90.42	83.17	65.18	65.12	69.97	58.91
CPRFL-CLIP(ours)	**86.28**	81.84	**90.51**	**86.43**	**66.69**	**66.35**	**70.99**	**61.33**

Experiments

Datasets		VOC-LT					COCO-LT			
Methods		total	head	medium	tail		total	head	medium	tail
CPRFL-GloVe(ours)		85.14	**82.50**	90.42	83.17		65.18	65.12	69.97	58.91
CPRFL-CLIP(ours)		**86.28**	81.84	**90.51**	**86.43**		**66.69**	**66.35**	**70.99**	**61.33**
MLC-NC		**84.37**	72.75	**88.15**	**90.31**		**60.52**	49.69	**64.94**	**64.21**

Ablation Study

Table 2: The mAP (%) performance of the proposed CPRFL with different multi-label classification losses on two long-tailed multi-label datasets. Bold indicates the best scores.

Dataset	VOC-LT			
Loss Functions	total	head	medium	tail
BCE	79.60	82.37	88.95	70.52
MLS	80.75	82.44	90.37	72.26
CB Focal [11]	83.87	80.32	90.48	81.54
DB No-Focal [51]	84.27	**82.64**	89.06	81.89
DB Focal [51]	86.18	81.81	90.01	86.30
ASL [41]	**86.28**	81.84	**90.51**	**86.43**

Dataset	COCO-LT			
Loss Functions	total	head	medium	tail
BCE	59.62	61.48	66.09	49.47
MLS	59.91	63.36	66.19	48.59
CB Focal [11]	64.20	64.50	69.36	57.12
DB No-Focal [51]	64.28	64.31	69.62	57.21
DB Focal [51]	65.12	63.74	69.97	59.91
ASL [41]	**66.69**	**66.35**	**70.99**	**61.33**

Ablation Study

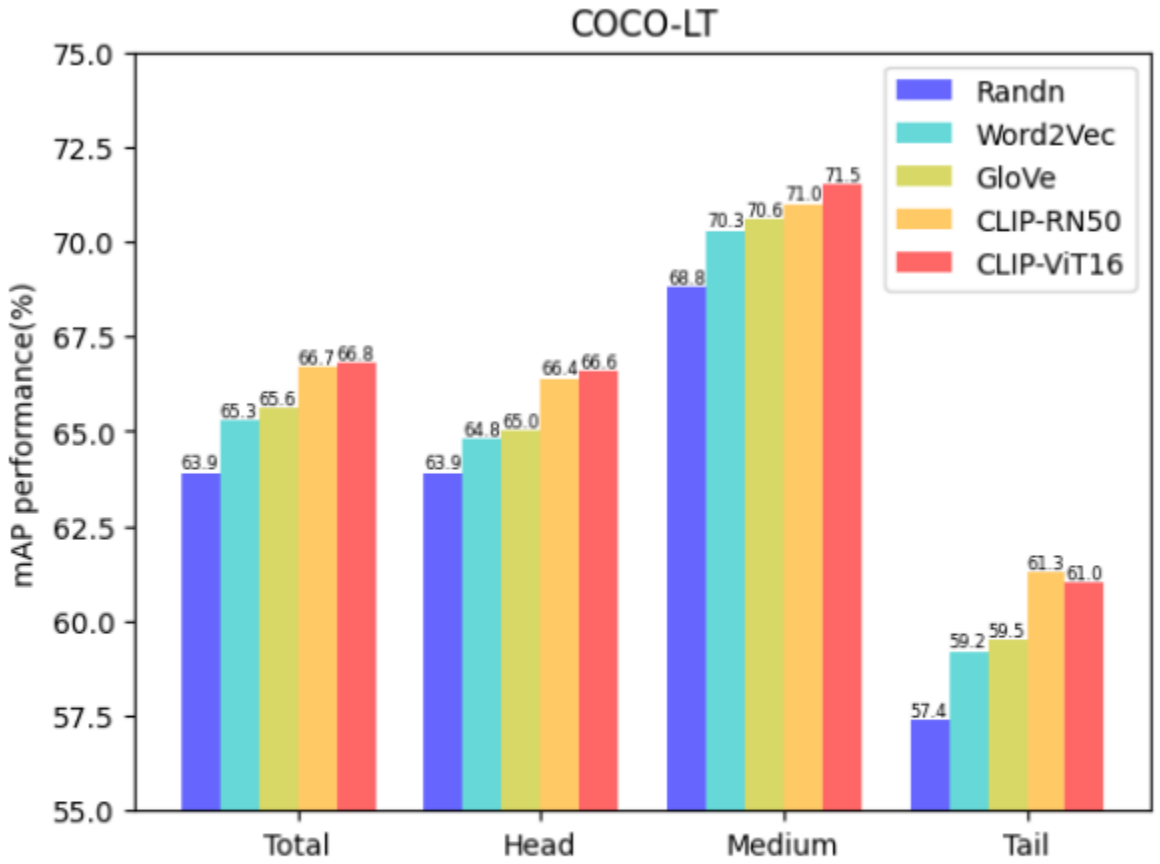


Figure 3: The mAP (%) performance with various types of category semantics for prompt initialization on COCO-LT dataset.

Ablation Study

Table 3: The ablation analysis on different components of the proposed CPRFL. Here "VSI" denotes Visual-Semantic Interaction, "PI" denotes Prompt Initialization, "RW" denotes Re-Weighting strategy, "avg.△" denotes average performance improvement. Bold indicates the best scores.

	VSI	PI	RW	VOC-LT				avg.△	COCO-LT				avg.△
				total	head	medium	tail		total	head	medium	tail	
Baseline				73.88	69.41	81.43	71.56		49.46	49.80	54.77	42.14	
	√			83.87	80.32	90.53	81.54	+9.99	63.87	63.55	69.01	57.36	+14.40
	√	√		84.24	80.78	90.47	82.16	+10.34	64.27	64.30	69.62	57.20	+14.80
	√	√	√	86.28	81.84	90.51	86.43	**+12.20**	66.69	66.35	70.99	61.33	**+17.30**

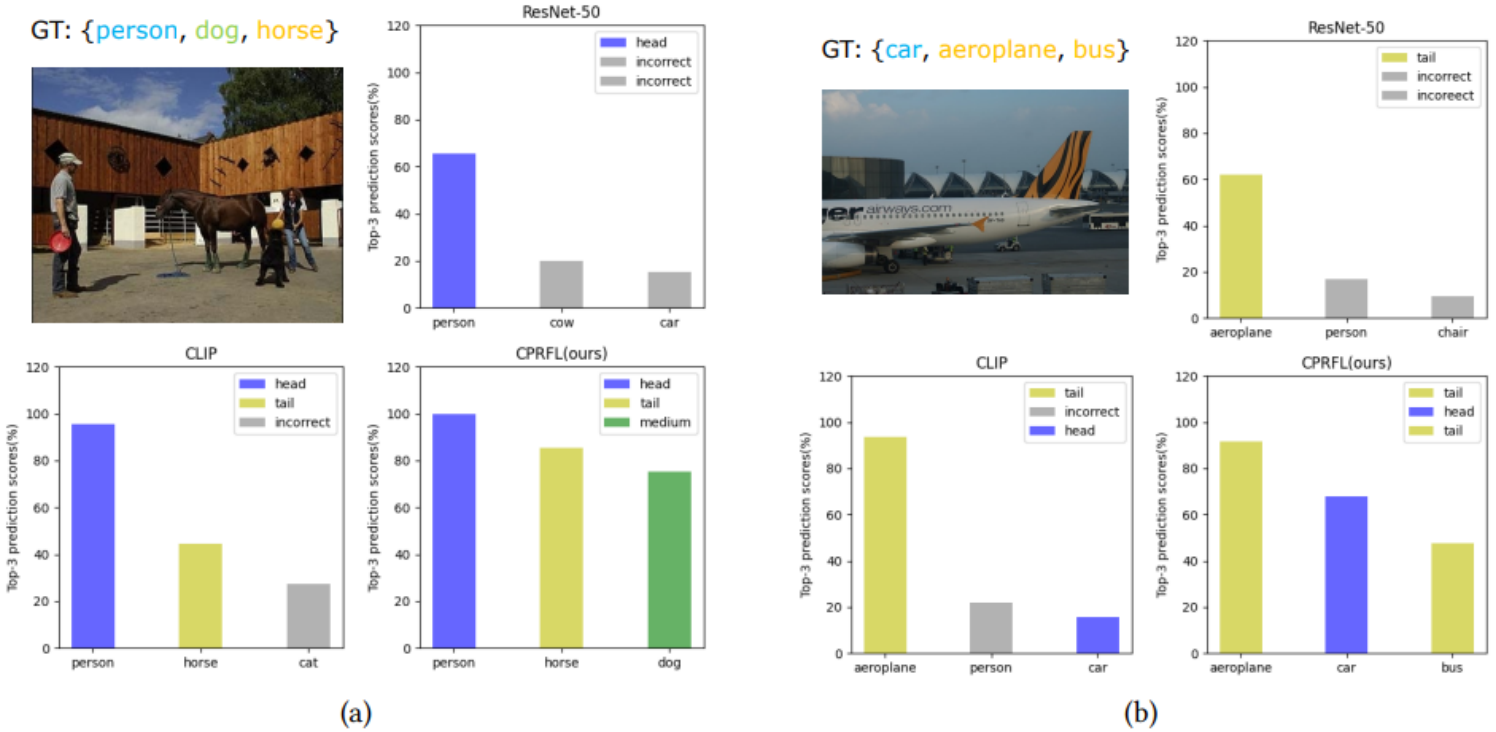


Figure 4: Visualization examples of Top-3 predicated categories by ResNet-50, CLIP and our CPRFL.

Conclusion

To tackle the challenges of head-to-tail imbalance and multi-object recognition in long-tailed multi-label image classification, we propose a novel and effective approach, termed Category-Prompt Refined Feature Learning (CPRFL). CPRFL capitalizes on CLIP's text encoder to extract category semantics, leveraging its robust semantic representation capability. This allows for the establishment of semantic correlations between the head and tail classes. The derived category semantics are then utilized as category-prompts, facilitating the decoupling of category-specific visual representations. Through a series of dual-path gradient back-propagations, we refine these prompts to effectively mitigate the visual-semantic domain bias. Simultaneously, the refinement process aids in purifying the category-specific visual representations under the guidance of the refined prompts. To our knowledge, this is the pioneering work to leverage category semantic correlations for mitigating head-to-tail imbalance in LTMLC, offering an innovative solution tailored to the unique characteristics of the data.

Thanks