Cross-modal Learning Using Privileged Information for Long-tailed Image Classification

Xiangxian Li [1] Yuze Zheng [1] Haokai Ma [1] Zhuang Qi [1] Xiangxu Meng [1] Lei Meng [1*]
[1] Shandong University
{xiangxian_lee, zhengyuze, mahaokai, z_qi}@mail.sdu.edu.cn
{mxx, lmeng}@sdu.edu.cn

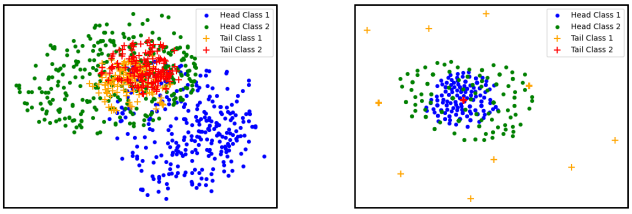
Abstract

The long-tailed distribution is widespread in data, learning from long-tailed images may lead the classification model to concentrate more on the head classes that occupied most samples, while paying less attention to the tail classes. Existing long-tail image classification methods try to alleviate the head-tail imbalance majorly by re-balancing the data distribution, assigning the optimized weights, and augmenting information, but they often get in trouble with the trade-off on the head and tail performance which mainly caused by the poor representation learning of tail classes. To address the above problems, we introduce descriptional words of images as cross-modal privileged information and propose a cross-modal enhanced method for long-tailed image classification, termed CMLTNet. The CMLTNet improves the learning of intra-class similarity of tail-class representations by the cross-modal alignment and captures the difference between head and tail classes in the semantic space by the cross-modal inference. After the fusion of the above information, CMLTNet achieves overall better performances than the benchmarking long-tailed learning and cross-modal learning methods on long-tailed cross-modal datasets NUS-WIDE and VireoFood-172. We further study the effectiveness of proposed modules through ablation experiments; from case study of feature distribution, we demonstrate that the model has learned better representation of tail classes, and in the experiments of model attention we find that CMLTNet may help to learn some rare concepts in the tail class through the mapping to the semantic space.

Keywords: ***Long-tailed classification, Cross-modal learning, Representation learning, Privileged information***

1. Introduction

The long-tailed phenomenon in data distribution means that most samples belong to a small number of head classes, while many tail classes only occupy a small part of samples. Learning image classification from long-tailed data tends to lead the model dominated by the head and get poor optimization on the tail. Therefore, existing works mainly try to re-balance the data distribution [37, 13] and re-assign the optimization weights to compensate the tail [7, 1, 6], but the lacking of diversity in tail-class information may consequently cause trade-off between the head and the tail performance. So recent works propose to apply data augmentation [4, 36, 24], adversarial training [16, 14], and transfer learning [19] to supplement information of tail classes. However, these methods in visual modality have the dilemma of injuring the head or exacerbating the imbalanced situation, since they still play a role of re-balancing and the representation learning has not been well improved, which hinders the way to solving problems like the interference of background noise which may be more serious in tail classes, so novel ideas are needed to alleviate the above problems.



(a) Visual feature distribution (b) Semantic feature distribution

Figure 1. Visualization of the distribution in visual and semantic feature spaces, (a) is the distribution of visual features, in which the distribution among classes is messy, and the tail features are mixed in the head classes; in the semantic space (b), the intra-class distribution of the head is more concentrated, and the tail features are distributed in specific regions that can be distinguished clearly.

Due to the popularity of multi-modal data, images in reality are usually accompanied by semantic information such as tags or description words, which are easier to distinguish main body of images, as shown in Figure 1. So introducing cross-modal semantic information as supplementary in the training process, i.e., the Learning Using Privileged Information (LUPI) paradigm [31, 30], is promising to improve the representation learning of model. Works in this area

*Lei Meng is the corresponding author.

are mainly divided into cross-modal constraint methods and cross-modal alignment methods. Cross-modal constraint methods utilize semantic information as the extra constraint on local or global feature extraction [2, 3, 23, 9]; and cross-modal alignment methods make the range [22, 27] or distribution [15, 17] of visual and semantic features more similar. However, existing works achieve limited performance gains since the uncontrollable constraints and the modal heterogeneity. In addition, since the long-tailed distribution also exists in semantic modality, the bias may be further exacerbated in the cross-modal learning.

To address the aforementioned problems, we propose a Cross-Modal learning method CMLTNet to improve the learning of visual representations in long-tailed image classification. Through the introduction of cross-modal semantic information, the visual representations are enhanced which achieves the upgrading of both the head and the tail classes. The overall idea of CMLTNet is shown in Figure 2, which consists of three main processes, the **Alignment** between cross-modal information, the **Inference** from visual to semantic space, and the cross-modal information **Fusion**. To make full use of the information in the semantic modality during training, we first propose feature-level alignment to form the cluttered visual features more similar to the focused and distinguishable semantic features. The alignment has limited effects due to the modal heterogeneity, so in another aspect, we encourage the model to learn to map from visual to semantic space, that is, visual-semantic inference, finding the meaningful semantic information from visual features to achieve communication between modalities. Finally, the representation learning of the model is enhanced from the fusion of distribution alignment and visual-semantic inference, which improves the intra-class similarity and inter-class discrimination learning to achieve better performance on long-tailed image classification.

In experiments, we demonstrate the effectiveness of CMLTNet on two cross-modal long-tailed datasets NUS-WIDE and VireoFood-172. The experimental results show that our method can effectively enhance the prediction of the whole and especially the tail classes without the loss of the head. In ablation study, we analysis the effectiveness of cross-modal alignment and inference and the effects of different fusion strategies. Further, we exhibit the enhancement effect of CMLTNet on representation learning and the improvement of the model's attention to long-tailed data in cross-modal learning through case studies.

In summary, the main contributions of this paper are:

- We propose CMLTNet to effectively improve representation learning for long-tailed image classification, thus alleviating the issue in vision. This is a pioneering work which explores incorporating cross-modal privilege information.

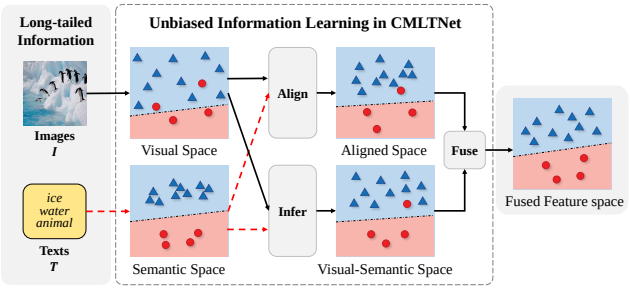


Figure 2. CMLTNet improves representation learning for both head (blue triangles) and tail (red circles) classes in long-tailed image classification. While the visual space has dispersed features and head dominates decision-making, the semantic space offers clear representations from description words. CMLTNet aligns feature distribution and maps visual space to semantic space, promoting semantic learning during training.

- We analyze the strengths and limitations of cross-modal learning methods on long-tailed image classification. On this basis, this paper proposes a model-agnostic "alignment-inference-fusion" framework and demonstrates its advantages in representation learning and filtering visual noise.

2. Related works

2.1. Long-tailed Image Classification

Works about long-tailed image classification mainly improve on the data level and the optimization level. Among the data-level methods, re-sampling methods [37, 13] assign weights of class-sampling to re-balance the data distribution; curriculum learning-based methods re-adjust the training process from easy to hard [32, 34]; methods above may cause over-fitting at the tail because the lacking of information diversity, so data augmentation-based methods are proposed to diverse the tail classes, such as samples interpolation [4] and background replacement [36, 24]. In addition, adversarial training-based methods [16, 14] are also effectively augmenting information by introducing perturbations. Works of optimization-level attempt to reduce the bias during model optimization, such as loss adjustment[18, 1, 6] and class re-weighting[7].

2.2. Cross-modal Learning for Image Classification

Cross-modal learning methods using semantic information as privileged information (LUPI) [31, 30] for image classification can be mainly divided into two approaches, the implicit cross-modal constraint-based methods and the explicit cross-modal alignment-based methods. The Cross-modal constraint-based methods introduce semantic information as the local[3, 23] or global[2, 12, 9] constraints by mapping the visual feature into semantic predictions, which enhance the extracting of semantic information from visual

features; while the methods of cross-modal alignment apply similarity loss such as KL-Divergence[22], covariance matrix [27] between visual features and privileged semantic features or between the feature distributions [15, 17] to guide the model to filter the noise in the visual feature space.

3. Method

3.1. Overview

To make full use of the cross-modal information in the training phase and improve the learning of long-tailed images, we constructs an "alignment-inference-fusion" learning framework in CMLTNet, as shown in Figure 3.

In the Visual Representation Enhancement module, cross-modal alignment is used to improve learning of intra-class representation. The modal heterogeneity limits alignment effects, so in the Cross-modal Representation Inference Module, semantic information is used as a constraint for mapping visual features to semantic space. This effectively learns semantically meaningful visual-semantic knowledge and reduces inter-class confusion from visual noise. Finally, the Cross-modal Information Fusion module fuses features learned from different channels to obtain debiased information, and thus improve long-tailed image classification results.

3.2. Visual Representation Enhancement Module

As mentioned in Section 3.1, the main target of the Visual Representation Enhancement Module is to make the extracted visual features closer to the semantic features at the distribution level during classification. For input images $\mathcal{V} = \{v_i | i = 1, ..., N\}$ and their corresponding description words $\mathcal{S} = \{s_i | i = 1, ..., N\}$, the model first extracts visual features $\mathbf{F}_v = \rho_v(\mathcal{V})$ and semantic features $\mathbf{F}_s = \rho_s(\mathcal{S})$ through visual feature extractor $\rho_v(.)$ and semantic feature extractor $\rho_s(.)$. Then the model tries to find a shared space that minimizes the distance of distribution of $\mathbf{F}_v$ and $\mathbf{F}_s$:

$$\min\{\text{Distance}(\alpha_v(\mathbf{F}_v), \alpha_s(\mathbf{F}_s))\} \tag{1}$$

where $Distance(.)$ means the measurement of distance like L_p Norm; α_v and α_s are shared space mapping for visual and semantic features, we applied Linear projection followed by ReLu activation in the CMLTNet.

In CMLTNet, the aligned features mapped by shared space $\mathbf{F}_{va} = \alpha_v(\mathbf{F}_v)$ and $\mathbf{F}_{sa} = \alpha_s(\mathbf{F}_s)$ achieves the goal of Equation 1 through KL-Divergence, making visual features closer to semantic features in shared space:

$$\mathcal{L}_{explicit} = \text{KLD}(\text{Softmax}(\mathbf{F}_{va}), \text{Softmax}(\mathbf{F}_{sa})) \tag{2}$$

In the above process, visual and semantic features are mapped into shared space to form alignment features $\mathbf{F}_{va} = \alpha_v(\mathbf{F}_v)$ and $\mathbf{F}_{sa} = \alpha_s(\mathbf{F}_s)$, by imposing classification constraints, the features of two modalities are further optimized in the direction of improving the classification, thus forming Implicit constraints, the loss function is:

$$\mathcal{L}_{implicit} = \mathcal{L}_{cls}(f(\mathbf{F}_{va}), \mathcal{C}) + \mathcal{L}_{cls}(f(\mathbf{F}_{sa}), \mathcal{C}) \tag{3}$$

where $\mathcal{L}_{cls}$ is the classification loss which can be the Cross-Entropy in the single-label classification task and the Binary Cross-Entropy in the multi-label classification. The $\mathcal{C}$ means the labels of samples and $f(.)$ is the shared class mapping for both visual and semantic features.

3.3. Cross-modal Representation Learning Module

The visual representation is enhanced by alignment, but the modal heterogeneity limits the effectiveness, thus the visual noise and error propagation still serious. Therefore, we design a cross-modal representation learning method infer features in visual modality to semantic modality.

To extract semantically meaningful visual information, we need to find a cross-modal transfer mapping from visual to visual-semantic features, i.e., $\mathbf{F}_{v\to s} = \text{Trans}(\mathbf{F}_v)$ which can well correspond to description words $\mathcal{S}$. The target of cross-modal inference is:

$$\min\{\text{Error}(\text{g}(\mathbf{F}_{v\to s}), \mathcal{S})\} \tag{4}$$

where $Error(.)$ means the error of words predictions and $\text{g}(.)$ is the predicted mapping of words. In CMLTNet, we applied two blocks of Linear projections followed by a LeackyReLu activation as the modal-transfer mapping $\text{Trans}(.)$, and the $\text{g}(.)$ is the Linear projection.

To achieve the goal of Equation 4, the semantic words $\mathcal{S}$ are used as targets for word predictions, and semantic features $\mathbf{F}_s$ are also used for improving the cross-modal transfer mapping:

$$\mathcal{L}_{transfer} = BCE(\text{g}(\mathbf{F}_{v\to s}), \mathcal{S}) + \beta_t \cdot MSE(\mathbf{F}_{v\to s}, \mathbf{F}_s) \tag{5}$$

where $BCE(.,.)$ is the Binary Cross-Entropy loss, $MSE(.,.)$ is the Mean square Error loss, and β_t is the coefficient of transfer loss, and the range is shown in Section .

The semantic predictions $\mathbf{P}_{v\to s} = \text{g}(\mathbf{F}_{v\to s})$ contains words probability in the given images, and for using these information to enhance the representations learned in visual space, CMLTNet encodes $\mathbf{P}_{v\to s}$ as embeddings:

$$\mathbf{E}_s = \theta(\text{Embed}(\text{Topk}(\mathbf{P}_{v\to s}))) \tag{6}$$

where $Topk(.)$ is the operation to chose top-k predicted words, $Embed(.)$ is the word embedding, and $\theta(.)$ is the operation of embedding fusion, we will show the results of using linear and mean embedding fusion in Section 4.4.

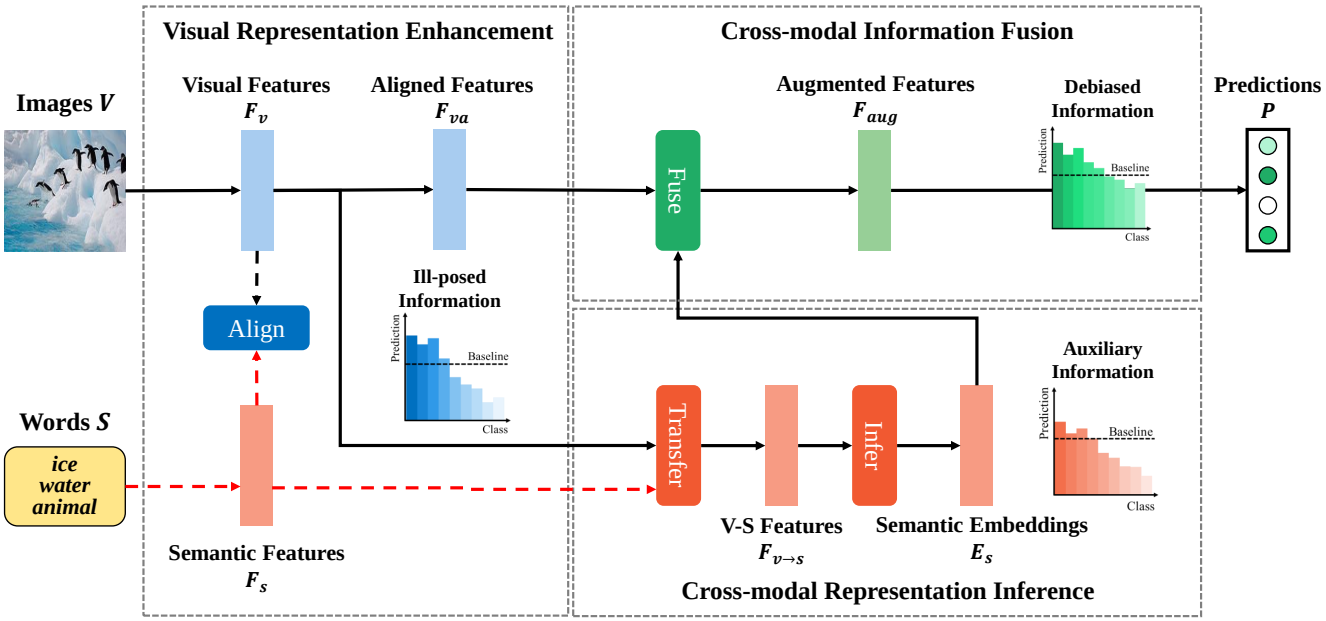


Figure 3. The schematic diagram of CMLTNet, description words $\mathcal{S}$ of the images $\mathcal{V}$ are introduced in the training phase, on the one hand, they are formed as cross-modal semantic feature $\mathbf{F}_s$ to help the alignment learning from visual features $\mathbf{F}_v$ to visual aligned features $\mathbf{F}_{v_a}$ in a shared space; on the other hand, words $\mathcal{S}$ help visual features $\mathbf{F}_v$ to better transfer to semantic space and infer semantic embeddings $\mathbf{E}_s$. Finally the learned features $\mathbf{F}_{v_a}$ and embeddings are fused as augmented features $\mathbf{F}_{aug}$. Through the training of the CMLTNet framework, the bias information in learning long-tail images is alleviated, thereby improving the image classification ability in the test.

The learning process of embedding is constrained by its prediction of the class:

$$\mathcal{L}_{embed} = \mathcal{L}_{cls}(f_e(\mathbf{E}_s), \mathcal{C}) \tag{7}$$

where $f_e(.)$ is the class mapping of semantic embeddings.

Therefore, the overall loss of cross-modal inference is:

$$\mathcal{L}_{infer} = \mathcal{L}_{transfer} + \mathcal{L}_{embed} \tag{8}$$

3.4. Cross-modal Information Fusion Module

The representations of visual features are strengthened by alignment, but there is still an ill-posed between head and tail classes. After cross-modal inference, visual noise in the vision is filtered, but the loss of information brings a dropping of performance. Therefore, we propose to fuse the two parts of the features to combine the advantages of the two modalities:

$$\mathbf{F}_{aug} = Fusion(\phi(\mathbf{F}_{va}), \phi(\mathbf{E}_s)) \tag{9}$$

where $Fusion(.,.)$ is a feature-level operation like feature concatenation, add, min, and max operation, and $\phi(.)$ is a linear layer followed by LeakyReLU activation.

The classification constraint is applied in the fusion:

$$\mathcal{L}_{fusion} = \mathcal{L}_{cls}(f_f(\mathbf{F}_{aug}), \mathcal{C}) \tag{10}$$

where $\mathcal{L}_{cls}$ is CE loss for single-label classification or BCE loss for multi-label classification, and $f_f(.)$ is the fused feature to class mapping.

3.5. Training Strategy

3.5.1 Multi-stage Training

In order to improve the training efficiency, the training of CMLTNet can be divided into the following stages according to the aforementioned process:

- **Stage 1**: Training the feature extractor and shared space mapping net in visual and semantic modality, using $\mathcal{L}_{implicit}$ and $\mathcal{L}_{explicit}$ as constraint, with an adjustable factor γ_i on $\mathcal{L}_{implicit}$.
- **Stage 2**: Freezing the networks in Stage 1, and training the transfer network, visual-semantic mapping network, and the embedding using $\mathcal{L}_{infer}$ as loss.
- **Stage 3**: Freezing the networks in Stage 1 and Stage 2, and training the linear net and class mapping in fusion by the constraints of $\mathcal{L}_{fusion}$.

3.5.2 One-stage Training

CMLTNet can be trained end-to-end by combine the above losses, but in this case, parameter adjustment needs to be careful, we will provide some choices in Section 4.2.2:

$$\mathcal{L} = \gamma_i \cdot \mathcal{L}_{implicit} + \mathcal{L}_{explicit} + \gamma_t \cdot \mathcal{L}_{infer} + \mathcal{L}_{fusion} \tag{11}$$

where γ_i and γ_t are weight factors of losses.

Table 1. Statistical details of NUS-WIDE and VireoFood-172. IR is the short of Imbalance Ratio and the # indicates the categories.

Datasets	#Classes	#Words	IR (Train)	IR (Test)
NUS-WIDE	81	1000	1083.62	1465.70
VireoFood-172	172	353	5.57	5.50

4. Experiments

4.1. Datasets

Experiments are taking on two cross-modal long-tailed datasets as shown in Table 1, where the Imbalance Ratio (IR) [1, 7] measures the degree of imbalance in datasets, i.e., IR $= max_i n_i / min_i n_i$, which means the ratio of the sample amount in most sampled class and the least sampled class.

NUS-WIDE [5]: a multi-label classification dataset containing images in 81 classes. Each image corresponds to several texts, and the total number of word classes is 1000. We follow previous works[5, 28, 29] to split the train/test set and remove samples missing labels or text. Finally, 203,598 samples remain, including 121,962 training samples and 81,636 testing samples, with an IR of 1083 in the training set and 1465 in the test set.

VireoFood-172 [2]: a single-label classification dataset with a total of 99,225 images corresponding to 172 categories. Each image corresponds to multiple texts, and the total number of classes is 353. Among them, there are 66,071 samples in the training set and 33,154 samples in the test set. The IR in the training set is 5.57 and the IR in the test set is 5.50.

4.2. Experimental Settings

4.2.1 Evaluation Protocol

Following the previous works about multi-label long-tailed classification [33, 10], the mean Average Precision (mAP) is adapted on the multi-label dataset NUS-WIDE to evaluate the performance of algorithms. We report the performances in three disjoint class-subsets that divided by the frequency of occurrences in training set like the settings in [20]: Head classes (classes each with over 5000 occurrences), Medium classes (classes each with 2000 to 5000 occurrences) and Tail classes (classes each under 2000 occurrences).

We use the Accuracy score for evaluating the classification performance of algorithms on single-label dataset VireoFood-172 as previous works [16, 20] did. As mentioned in the protocol settings of NUS-WIDE, we also divide classes of VireoFood-172 into three disjoint class-subsets: Head classes (classes each with over 500 occurrences), Medium classes (classes each with 300 to 500 occurrences) and Tail classes (classes under 300 occurrences).

4.2.2 Implementation Details.

For all the algorithms, we set batch size as 64, and the interval of learning rate decay is 4 epochs and each model decays 3 times for 0.1 then training for an extra epoch. The optimizer is Adam with weight decay selected from [1e-3, 5e-4, 2e-4, 1e-4], the learning rate is chosen in the range from 5e-5 to 5e-3. For the comparative long-tailed learning methods, we chose the β in Class-Balanced (CB) [7] Resample, Reweight, and LDAM-DRW [1] from 0.9 to 0.9999; For comparative corss-modal learning methods, we set the dimension of latent space as 2048. As for the parameters in CMLTNet and its variants, the coefficient of losses β_{align}, $\beta_{transfer}$ are selected from [0.1, 0.2, 0.5, 1.0, 1.5, 2.0]; and for the one-stage training, the weight of loss β_t, γ_i and γ_t are chosen in [0.1, 0.5, 1.0, 2.0], and the dimension of latent space is 300.

4.3. Performance Comparison

Comparison experiments are conducted among visual models, cross-modal learning methods, and long-tailed learning methods. The Visual Backbones include pretrained basic networks ResNet-18, ResNet-50 [11], and VGG [26], two improved networks WRN [35] and WISeR [21] based on ResNet-50, and the recent Transform-based backbone ViT-B [8]; Long-tailed Learning methods include Focal loss [18], Class-balanced (CB) [7] resample and reweight, and LDAM-DRW [1]; The Cross-modal learning methods include our in-house implemented constraint-based methods ARCH-D [2], CMRR [3], and CMFL [9], and align-based methods ATNet [22], methods above use pretrained ResNet-50 as the backbone. From the Table 2 we have following observations:

- There is an obvious head-to-tail deviation in the visual Backbones, including the convolutional network and transform-based network, and the improved backbones bring limited improvements on the tail on NUS-WIDE with higher IR. The performance of medium classes is higher than the head on the VireoFood-172 since there are more confusing classes in the head.
- CMLTNet has achieved comparable or better performance with benchmarking long-tailed learning and cross-modal learning methods under the same visual backbone and simultaneously improving the performance of head, medium, and tail classes.
- CMLTNet achieves more stable performance gains across datasets in different domains compared with other methods. And it is worth noting that on the NUS-WIDE, CMLTNet using ResNet-18 can outperform most cross-modal learning methods using ResNet-50, indicating that CMLTNet combines beneficial information in different modalities in an effective way.

Table 2. Comparative results on multi-label NUS-WIDE report the mAP scores, and Accuracy scores are reported on the single-label VireoFood-172. In the table, **All**, **Head**, **Med**, and **Tail** represent the results on the whole, the head, the medium, and the tail classes.

Method	Model	NUS-WIDE				VireoFood-172			
		All	**Head**	**Med**	**Tail**	**All**	**Head**	**Med**	**Tail**
Visual Backbone	ResNet-18	0.421	0.681	0.508	0.332	0.782	0.784	0.785	0.767
	ResNet-50	0.444	0.692	0.536	0.357	0.817	0.817	0.824	0.798
	VGG	0.436	0.694	0.531	0.346	0.811	0.805	0.820	0.801
	WRN	0.451	0.711	0.546	0.361	0.825	0.817	0.830	0.823
	WISeR	0.451	0.711	0.544	0.362	0.828	0.832	0.829	0.819
	ViT	0.455	0.709	0.544	0.367	0.836	0.829	0.846	0.830
Long-tailed Learning	Focal(ResNet-50)	0.452	0.714	**0.569**	0.356	0.821	0.821	0.827	0.801
	CB Resample(ResNet-50)	0.467	0.691	0.518	0.397	0.812	0.802	0.821	0.811
	CB Reweight(ResNet-50)	0.459	0.695	0.534	0.379	0.820	0.817	0.826	0.810
	LDAM-DRW(ResNet-50)	0.470	0.701	0.548	0.392	0.833	0.826	0.840	0.825
Cross-modal Learning	ATNet(ResNet-50)	0.458	0.693	0.531	0.380	0.829	0.824	0.837	0.814
	ARCH-D(ResNet-50)	0.450	0.695	0.532	0.366	0.825	0.824	0.833	0.804
	CMRR (ResNet-50)	0.450	0.686	0.508	0.375	0.819	0.815	0.824	0.812
	CMFL(ResNet-50)	0.478	0.706	0.564	0.398	0.831	0.829	0.833	0.816
	CMLTNet(ResNet-18)	0.478	0.702	0.539	0.405	0.792	0.785	0.800	0.781
	CMLTNet(ResNet-50)	0.486	0.707	0.548	0.413	0.833	0.825	0.842	0.823
	CMLTNet(ViT)	**0.494**	**0.715**	0.557	**0.420**	**0.843**	**0.837**	**0.850**	**0.832**

- For the long-tailed learning methods, to bring an overall improvement, the head classes need to get more optimization; otherwise, it is easy to weaken the head while improving the tail. The CB Resample on VireoFood-172 is an example, while the tail is increased by 1.2%, the medium and head are weakened by 0.3% and 1.8%, which finally brings a 0.6% reduction to overall. The focal loss encountered the problem of increasing the head-to-tail gap on NUS-WIDE.

- Cross-modal learning methods generally improve the tail predictions, but effects are different between datasets. For example, since the deception words are diversity inner class on NUS-WIDE, the effectiveness of align-based ATNet is limited. On the VireoFood-172, the cross-modal constraint-based methods ARCH-D and CMRR bring less improvement on the tail.

4.4. Ablation Study

To analyze the mechanism by which CMLTNet takes performance gains, we gradually added modules to the base model for ablation study, the results are shown in Table 3.

- Compared with the base model, after the alignment (+A), the model achieves better performance in overall classes. In addition, the rise of the tail class is higher than that of the head class and the middle class, for example, the rise of the head, med, and tail classes in NUS-WIDE are 4%, 0.8%, and 8%, which means that after the introduction of cross-modal information, the tail class has better information supplementation effectively alleviates the imbalance problem.

Table 3. Ablation study of CMLTNet uses ResNet-50 as the Base model. In the table, +A means cross-modal alignment, and +I means cross-modal inference which contains two words embedding combination methods, mean of features (M) and linear projection (L), and F is the feature fusion.

Model	NUS-WIDE				VireoFood-172			
	All	**Head**	**Med**	**Tail**	**All**	**Head**	**Med**	**Tail**
Base	0.444	0.692	0.536	0.357	0.817	0.817	0.824	0.798
+A	0.466	0.698	0.540	0.388	0.821	0.821	0.834	0.822
+I(M)	0.355	0.594	0.411	0.279	0.780	0.790	0.797	0.708
+I(L)	0.401	0.628	0.446	0.332	0.801	0.802	0.812	0.771
+A+I(M)+F	0.473	0.703	0.543	0.397	0.829	0.823	0.837	0.816
+A+I(L)+F	**0.486**	**0.707**	**0.548**	**0.413**	**0.833**	**0.825**	**0.842**	**0.823**

- As for the cross-modal inference (+I), the visual noise is filtered through the inference, which makes the gap between the head and the tail smaller, but it also causes a loss of information that may be beneficial for classification, so the accuracy of the direct prediction of cross-modal inference is not high (whether mean or linear projection in embedding), but it contains semantic information which can supplement the aligned features.

- The supplemental effects can be seen from the effect of cross-modal fusion (+F). Compared with the aligned prediction, the performances of the head, middle, and tail are further improved.

4.5. In-depth Analysis of Fusion Strategy

In this section, we discuss the fusion strategy choosing of the CMLTNet, the performance of fusion strategy of cross-modal features are shown in Table 4.We find although the best that whether using features summation (Add), features

Table 4. The performance of CMLTNet using different fusion strategies. **Align**, **Inference** and **Cross-modal Fusion** represent the performance using aligned visual features, semantic embeddings, and fused augmented features.

Class	Align	Inference	Cross-modal Fusion			
			Add	Con	Max	Min
All	0.466	0.401	0.482	0.483	**0.486**	0.484
Head	0.698	0.628	**0.709**	0.707	0.707	0.707
Med	0.540	0.446	**0.550**	0.549	0.548	0.549
Tail	0.388	0.332	0.407	0.410	**0.413**	0.411

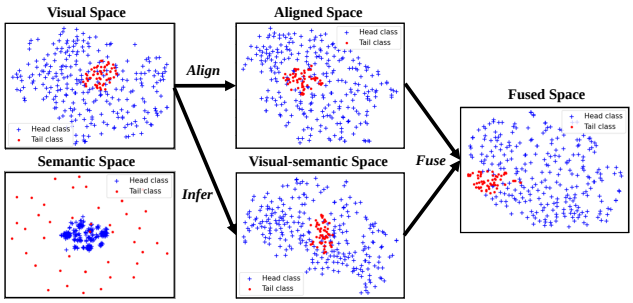


Figure 4. A t-SNE visualization of the feature distribution in the latent embedding spaces, where blue crosses represent head class samples and red dots represent tail class samples.

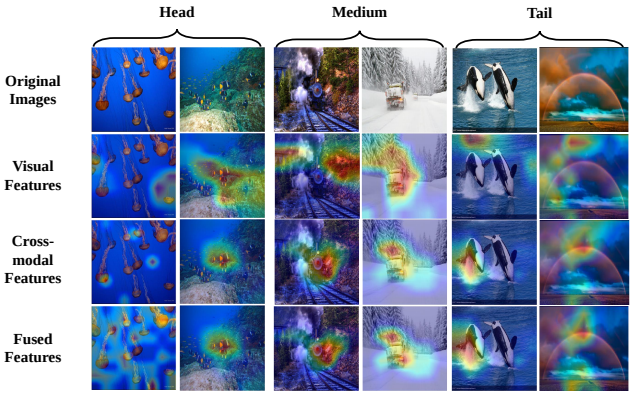


Figure 5. Visualization of model attention, Visual Features, Cross-modal Features and Fused Features represent the the model attention in different feature spaces.

concatenation (Con), features maximum (Max), or features minimum (Min), the head, medium, and tail predictions can be stably improved compared with the , which demonstrates that CMLTNet can extract the useful information from the two modalities through privileged information learning.

4.6. Case Study

4.6.1 Representation Learning in Feature Spaces

Through the above analysis, we found that each module of CMLTNet played a positive role in alleviating the imbalance problem. In this section we delve further into the feature space to understand how it improves representation learning. We randomly chose two confusing head and tail classes from VireoFood-172 (the imbalance ratio between them is 4.8) and use t-SNE to observe their distribution in the feature space.

The results are shown in Figure 4, features are heavily mixed in the visual space, which is due to the bias of the model in the optimization. However, in the semantic space, feature dimension is relatively lower, so the head class and the tail class are better distinguished. At the same time, the distribution of features is the aggregation of multiple clusters with distinct semantic features.

Through the alignment operation of CMLTNet, we find that the features of both head and tail classes, like that in the semantic space, tend to form small clusters, which makes some mixed head and tail features be distinguished. On the other line, after semantic inference, the head and tail are gathered into their respective spaces, and there is a clear demarcation between classes. Finally, in the fusion space, the features combine the characteristics of the above two spaces, which makes the intra-class aggregation and the inter-class separation at the same time, so that both the head and the tail get better representation learning.

4.6.2 Visual Attention of Different Features

Previous experiments have shown the improvement of CMLTNet on representation learning. In this section, we further analyze whether the CMLTNet learn semantic-meaningful information from features on the head, middle, and tail classes by GradCAM [25] visualization, as shown in Fig. 5.

we find that the interference of visual noise on the model makes it easy to be attracted by the background, especially in the tail with insufficient information diversity. By cross-modal inference, the attention of the model is more focused than that in visual modality, and reduces the problem of attention to the background; the modal fusion combines the attention of both modality. We can also find that the visual modality pay less attention to some rare concepts such as ”rainbow” and ”whale” in the tail class, but the semantic modality can learn them better, which explains the effectiveness of CMLTNet in alleviating the long tail problem.

5. Conclusion

This paper introduces CMLTNet, which enhances long-tailed classification based on cross-modal privilege information. Through heterogeneous feature alignment, cross-modal transfer and fusion enhancement representation learning, it strengthens the focus on minority classes, improves overall prediction ability, and provides a ”alignment-inference-fusion” framework for enhancing classification using cross-modal information.

This work is a preliminary step of cross-modal learning for long-tailed classification, in the future, we consider further enriching the diversity at the sample level by methods such as contrastive learning, and introduce causal inference in feature learning to improve the extraction of key information and further enhance the learning of tail features.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China (Grant no. 62006141), the National Key R&D Program of China (Grant no. 2021YFC3300203), the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073), and the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048).

References

[1] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 1, 2, 5

[2] J. Chen and C.-W. Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 32–41, 2016. 2, 5

[3] J.-j. Chen, C.-W. Ngo, and T.-S. Chua. Cross-modal recipe retrieval with rich food attributes. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1771–1779, 2017. 2, 5

[4] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, and D.-C. Juan. Remix: rebalanced mixup. In *European Conference on Computer Vision*, pages 95–110. Springer, 2020. 1, 2

[5] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 5

[6] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724, 2021. 1, 2

[7] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 1, 2, 5

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 5

[9] A. George and S. Marcel. Cross modal focal loss for rgbd face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2021. 2, 5

[10] H. Guo and S. Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15089–15098, 2021. 5

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[12] S. Jiang, W. Min, L. Liu, and Z. Luo. Multi-scale multi-view deep feature aggregation for food recognition. *IEEE Transactions on Image Processing*, 29:265–276, 2019. 2

[13] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2019. 1, 2

[14] J. Kim, J. Jeong, and J. Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13896–13905, 2020. 1, 2

[15] S. Li, B. Xie, J. Wu, Y. Zhao, C. H. Liu, and Z. Ding. Simultaneous semantic alignment network for heterogeneous domain adaptation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3866–3874, 2020. 2, 3

[16] X. Li, H. Ma, L. Meng, and X. Meng. Comparative study of adversarial training methods for long-tailed classification. In *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, pages 1–7, 2021. 1, 2, 5

[17] X. Li, Z. Xu, K. Wei, and C. Deng. Generalized zero-shot learning via disentangled representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1966–1974, 2021. 2, 3

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 5

[19] J. Liu, Y. Sun, C. Han, Z. Dou, and W. Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2020. 1

[20] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 5

[21] N. Martinel, G. L. Foresti, and C. Micheloni. Wide-slice residual networks for food recognition. In *2018 IEEE Winter Conference on applications of computer vision (WACV)*, pages 567–576. IEEE, 2018. 5

[22] L. Meng, L. Chen, X. Yang, D. Tao, H. Zhang, C. Miao, and T.-S. Chua. Learning using privileged information for food recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 557–565, 2019. 2, 3, 5

[23] W. Min, L. Liu, Z. Luo, and S. Jiang. Ingredient-guided cascaded multi-attention network for food recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1331–1339, 2019. 2

[24] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6887–6896, 2022. 1, 2

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. pages 1–14, 2015. 5

[27] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 2, 3

[28] J. Tang, X. Shu, , Z. Li, G.-J. Qi, and J. Wang. Generalized deep transfer networks for knowledge propagation in heterogeneous domains. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 12(4s), 2016. 5

[29] J. Tang, X. Shu, G.-J. Qi, Z. Li, M. Wang, S. Yan, and R. Jain. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5

[30] V. Vapnik, R. Izmailov, et al. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, 2015. 1, 2

[31] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009. 1, 2

[32] Y. Wang, W. Gan, J. Yang, W. Wu, and J. Yan. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5017–5026, 2019. 2

[33] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer, 2020. 5

[34] L. Xiang, G. Ding, and J. Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263. Springer, 2020. 2

[35] S. Zagoruyko and N. Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 5

[36] Y. Zhang, X.-S. Wei, B. Zhou, and J. Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3447–3455, 2021. 1, 2

[37] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 1, 2