Long-Tail Cross Modal Hashing

Zijun Gao\textsuperscript{1,2}, Jun Wang\textsuperscript{2,}\textsuperscript{*}, Guoxian Yu\textsuperscript{1,2}, Zhongmin Yan\textsuperscript{1,2}, Carlotta Domeniconi\textsuperscript{3}, Jinglin Zhang\textsuperscript{4}
\textsuperscript{1}School of Software, Shandong University, Jinan, China
\textsuperscript{2}SDU-NTU Joint Centre for AI Research, Shandong University, Jinan, China
\textsuperscript{3}Department of Computer Science, George Mason University, Fairfax, VA, USA
\textsuperscript{4}School of Control Science and Engineering, Shandong University, Jinan, China
zjgao@mail.sdu.edu.cn, \{kingjun, gxyu, yzm\}@sdu.edu.cn, carlotta@cs.gmu.edu, jinglin.zhang@sdu.edu.cn

Abstract

Existing Cross Modal Hashing (CMH) methods are mainly designed for balanced data, while imbalanced data with long-tail distribution is more general in real-world. Several long-tail hashing methods have been proposed but they cannot adapt for multi-modal data, due to the complex interplay between labels and individuality and commonality information of multi-modal data. Furthermore, CMH methods mostly learn the commonality of multi-modal data to learn hash codes, which may override tail labels encoded by the individuality of respective modalities. In this paper, we propose LiCMH (Long-tail CMH) to handle imbalanced multi-modal data. LiCMH firstly adopts auto-encoders to mine the individuality and commonality of different modalities by minimizing the dependency between the individuality of respective modalities and by enhancing the commonality of these modalities. Then it dynamically combines the individuality and commonality with direct features extracted from respective modalities to create meta features that enrich the representation of tail labels, and binaries meta features to generate hash codes. LiCMH significantly outperforms state-of-the-art baselines on long-tail datasets and holds a better (or comparable) performance on datasets with balanced labels.

Introduction

Hashing aims to map high-dimensional data into a series of low-dimensional binary codes while preserving the data proximity in the original space. The binary codes can be computed in a constant time and economically stored, which meets the need of large-scale data retrieval. In real-world applications, we often want to retrieve data from multiple modalities. For example, when we input key words to information retrieval systems, we expect the systems can efficiently find out related news/images/videos from database. Hence, many cross-modal hashing (CMH) methods have been proposed to deal with such tasks (Wang et al. 2016; Jiang and Li 2017; Yu et al. 2022a; Liu et al. 2019a).

Most CMH methods aim to find a low-dimensional shared subspace to eliminate the modality heterogeneity and to quantify the similarity between samples across modalities. More advanced methods leverage extra knowledge (i.e., labels, manifold structure, neighbors coherence) (Jiang and Li 2017; Yu et al. 2021; Liu et al. 2019b) to more faithfully capture the proximity between multi-modality data to induce hash codes. However, almost all of them are trained and tested on hand-crafted balanced datasets, while imbalanced datasets with long tail labels are more prevalent in real-world. Some recent studies (Chen et al. 2021; Cui et al. 2019) have reported that the imbalanced nature of real-world data greatly compromises the retrieval performance.

Real-world samples typically have a skewed distribution with long-tail labels, which means that a few labels (a.k.a. head labels) annotate to many samples, while the other labels (a.k.a. tail labels) have a large quantity but each of them is only annotated to several samples (Liu et al. 2020; Chen et al. 2021). It is a challenging task to train a general model from such distribution, because the head labels gain most of the attention while the tail ones are underestimated. Another inevitable problem of long-tail hashing is the ambiguity of the generated binary codes. Due to the dimension reduction, information loss is unavoidable. In this case, hash codes learned from data-poor tail labels more lack the discrimination, which seriously confuses the retrieval results.

Several efforts have been made toward long-tail single-modal hashing via knowledge transfer (Liu et al. 2020) and information augmentation (Chu et al. 2020; Wang et al. 2020b; Kou et al. 2022). LEAP (Liu et al. 2020) augments each instance of tail labels with certain disturbances in the deep representation space to provide higher intra-class variation to tail labels. OLTR (Liu et al. 2019c) and LTHNet (Chen et al. 2021) propose a meta embedding module to transfer knowledge from head labels to tail ones. In addition, these long-tail hashing methods cannot be directly adapted for multi-modal data, due to the complex interplay between tail labels and heterogeneous data modalities. The head labels can be easily represented by the commonality of multiple modalities, but the tail labels are often represented by individuality of a particular modality. For example, in the left part of Figure 1, tail label ‘Hawaii’ can only be read out from the text-modality. On the other hand, head label ‘Sea’ can be consolidated from the commonality of image and text modality, and head label ‘Island’ comes from the image modality alone. From this example, we can conclude that both the individuality and commonality should be used for effective CMH on long-tail data, which also account for the complex interplay between head/tail labels and multi-
Figure 1: The schematic framework of LtCMH. The direct image/text features $F^x/F^y$ are first extracted by CNN-F and FC (fully-connected) networks with two layers. LtCMH then uses different auto-encoders to mine the commonality $C^*$ and individuality $P^x/P^y$ of multi-modal data via cross-modal regularization and Hilbert-Schmidt independence criterion (HSIC). After that, it creates meta features $M^x/M^y$ by fusing $C^*$, $P^x/P^y$ and $F^x/F^y$. Next, LtCMH binarizes meta features into hash codes. The head label ‘Sea’ can be consolidated from the commonality of image and text modality, while tail label ‘Hawaii’ can only be obtained from the text-modality. The complex relations between labels and heterogeneous modalities can be more well explored by the commonality and individuality.

To address these problems, we propose LtCMH (Long-tail CMH) to achieve CMH on imbalanced multi-modal data, as outlined in Figure 1. Specifically, we adopt different auto-encoders to mine individuality and commonality information from multi-modal data in a collaborative way. We further use the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al. 2005) to extract and enhance the individuality of each modality, and cross-modal regularization to boost the commonality. We model the interplay between head/tail labels and multi-modal data by meta features dynamically fused from the commonality, individuality and direct features extracted from respective modalities. The meta features can enrich tail labels and preserve the correlations between different modalities for more discriminate hash codes. We finally binarize the meta features to generate hash codes. The contributions of this work are summarized as follows:

(i) We study CMH on long-tail multi-modal data, which is a novel, practical and difficult but understudied topic. We propose LtCMH to achieve effective hashing on both long-tail and balanced datasets.

(ii) LtCMH can mine the individuality and commonality information of multi-modal data to more comprehensively model the complex interplay between head/tail labels and heterogeneous data modalities. It further defines a dynamic meta feature learning module to enrich labels and to induce discriminate hash codes.

(iii) Experimental results show the superior robustness and performance of LtCMH to competitive CMH methods (Wang et al. 2020a; Li et al. 2018; Yu et al. 2021) on benchmark datasets, especially for long-tail ones.

Related Work

Cross Modal Hashing

Based on using the semantic labels or not, existing CMH can be divided into two types: unsupervised and supervised. Unsupervised CMH methods usually learn hash codes from data distributions without referring to supervised information. Early methods, such as cross view hashing (CVH) (Kumar and Udupa 2011) and collective matrix factorization hashing (CMFH) (Ding, Guo, and Zhou 2014), typically build on canonical correlation analysis (CCA). UDCMH (Wu et al. 2018) keeps the maximum structural similarity between the deep hash codes and original data. DJSRH (Su, Zhong, and Zhang 2019) learns a joint semantic affinity matrix to reconstruct the semantic affinity between deep features and hash codes. Supervised CMH methods additionally leverage supervised information such as semantic labels and often gain a better performance than unsupervised ones. For example, DCMH (Jiang and Li 2017) optimizes the joint loss function to maintain the label similarity between unified hash codes and feature representations of each modality. BiNCMH (Sun et al. 2022) uses bi-direction relation reasoning to mine unrelated semantics to enhance the similarity between features and hash codes.

Although these CMH methods perform well on balanced datasets, they are quite fragile on long-tail datasets, in which samples annotated with head labels gain more attention than those with tail labels when jointly optimizing correlation between modalities (Zhou et al. 2020). Given that, we try to enrich the representation of tail labels by mining the individuality and commonality of multi-modal data. This enriched representation can more well model the relation between head/tail labels and multi-modal data.

Long-Tail Data Learning

Long-tail problem is prevalence in real-world data mining tasks and thus has been extensively studied (Chawla 2009). There are three typical solutions: re-sampling, re-weighting and transfer-learning. Re-sampling based solutions use different sampling strategies to re-balance the distribution of different labels (Buda, Maki, and Mazurowski 2018; Byrd and Lipton 2019). Re-weighting solutions give different weights to the head and tail labels of the loss function (Huang et al. 2016). Transfer-learning based techniques learn general knowledge from the head labels, and then use these knowledge to enrich tail labels. Liu et al. (2019c) proposed a meta feature embedding module to combine direct features with memory features to enrich the representation of tail labels. Liu et al. (2020) proposed the concept of ‘feature cloud’ to enhance the diversity of tail ones.

Most of these long-tail solutions manually divide samples into head labels and tail ones, which compromises the gen-
eralization ability. In contrast, our LtCMH does not need to divide head and tail labels beforehand, it leverages both individuality and commonality as memory features, rather than directly stacking direct features (Liu et al. 2019c; Chen et al. 2021), to reduce information overlap. As a result, LtCMH has a stronger generalization and feature representation capability.

**The Proposed Methodology**

**Problem Overview**

Without loss of generality, we firstly present LtCMH based on two modalities (image and text). LtCMH can also be applied to other data modality or extended to ≥ 3 modalities (Liu et al. 2022). Suppose $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, y_2, \cdots, y_n\}$ are the respective image and text modality with $n$ samples, and $L = \{l_1, l_2, \cdots, l_n\}$ is the label matrix of these samples from $c$ distinct labels. A dataset is called long-tail if the numbers of samples annotated with these $c$ labels conform to Zipf’s law distribution (Reed 2001): $z_a = \frac{a^{-\mu}}{\sum_{a \in X} a^{-\mu}}$, where $z_a$ is the number of samples with label $a$, $\sum_{a \in X} a^{-\mu}$ is the control parameter of balance factor (IF for short, $\text{IF} = \frac{z_1}{z_c}$). The goal of CMH is to learn hash functions $(h^x, h^y)$ from $X$ and $Y$, and to generate hash codes via $b^x = h^x(x)$ and $b^y = h^y(y)$, where $b^x/b^y \in \{0,1\}^k$ is the length-$k$ binary hash code vector. For ease presentation, we denote $v$ as any modality, $x/y$ as the image/text modality indicator.

The whole framework of LtCMH is illustrated in Figure 1. LtCMH firstly captures the individuality and commonality of multi-modal data, and takes the individuality and commonality as the memory features. Then it introduces individuality-commonality selectors on memory features, along with the direct features extracted from respective modality, to create meta features, which not only enrich the representation of tail labels, but also automatically balance the head and tail labels. Finally, it binaries meta features into binary codes through the hash learning module. The following subsections elaborate on these steps.

**Individuality and Commonality Learning**

Before learning the individuality and commonality of multi-modal data, we adopt Convolutional Neural Networks (CNN) to extract direct visual features $F^x = CNN(X) \in \mathbb{R}^{n \times d_v}$ and fully connected networks with two layers (FC) to extract textual features $F^v = FC(Y) \in \mathbb{R}^{n \times d_v}$. Other feature extractors can also be adopted for modality adaption. Due to the prevalence of samples of head labels, these direct features are more biased toward head labels and represent tail ones. To address this issue, long-tail single-modal hashing solutions learn label prototypes to summarize visual features and utilize them to transfer knowledge from head to tail labels (Wei et al. 2022; Tang, Huang, and Zhang 2020; Chen et al. 2021). But they can not account for the complex interplay between head/tail labels and individuality and commonality of multi-modal data, as we exemplified in Figure 1 and discussed in the Introduction. Some attempts have already explored the individuality and commonality of multi-view data to improve the performance of multi-view learning (Wu et al. 2019; Yu et al. 2022b), and Tan et al. (2021) empirically found the classification of less frequent labels can be improved by the individuality. Inspired by these works, we advocate to mine the individuality and commonality of multi-modal data with long-tail distribution. We then leverage the individuality and commonality to create meta features, which can enrich the representation of tail labels and model the interplay between labels and multi-modal data.

Because of the information overlap of multi-modal high-dimensional data, we want to mine the shared and specific essential information of different modalities in the low-dimensional embedding space. Auto-encoders (AE) (Goodfellow, Bengio, and Courville 2016) is an effective technique that can map high-dimensional data into an informative low-dimensional representation space. AE can encode unlabeled/incomplete data and disregard the non-significant and noisy part, so it has been widely used to extract disentangled latent representation. Here, we propose a new structure for learning the individuality and commonality of multi-modal data. The proposed individuality-commonality AE (as shown in Figure 1) is similar to a traditional AE in both the encoding end of the input and the decoding end of the output. The new ingredients of our AE are the regularization of learning individuality and commonality, and the decoding of each modality via combining the individuality of this modality and commonality of multiple modalities.

As for the encoding part, we use view-specific encoders $\{f_{en}^v\}$ to initialize the individuality information matrix of each modality as follows:

$$P^v = f_{en}^v(F^v)$$

where $P^x$ and $P^y$ encode the individuality of image and text modality, respectively. To learn the commonality of multi-modal data, we concatenate $F^x$ and $F^y$ and then input them into commonality encoder $f_{enc}$ as:

$$C^v = f_{enc}([F^x, F^y])$$

Although $C^v$ can encode the commonality of multi-modal data by fusing $F^x$ and $F^y$, it does not concretely consider the intrinsic distribution of respective modalities. Hence, we further define a cross-modal regularization to optimize $C^v$ and to enhance the commonality using the shared labels of samples and data distribution within each modality as:

$$J_1(C^v) = \frac{1}{2} \sum_{a,b=1}^{c} \left\| C^v_{ab} - C^v_{ba} \right\|^2 \left( R_{ab}^v + R_{ba}^v \right)$$

$$= tr(C^v((D^v - R^v) + (D^v - R^v))^T)(C^v)^T)$$

Here $C^v_{ab}$ is the $a$-th column of $C^v$, $tr(\cdot)$ denotes the matrix trace operator, $D^v$ is a diagonal matrix with $D_{aa}^v = \sum_{b=1}^{c} R_{ab}^v$. $R^v$ quantifies the similarity between different
labels, it is defined as follows:

\[
R_{ab}^v = e^{-\frac{H^v_{ab}}{2}}
\]

\[
H^v_{ab} = \frac{\sum_{f_i^a \in X^a} \min_{f_j^b \in X^b} d(f_i^a, f_j^b)}{|X^a| + |X^b|}
+ \frac{\sum_{f_j^b \in X^b} \min_{f_i^a \in X^a} d(f_i^a, f_j^b)}{|X^a| + |X^b|}
\]

where \( H^v_{ab} \) is the average Hausdorff distance between two sets of samples separately annotated with label \( a \) and \( b \), \( d(f_i^a, f_j^b) \) is the Euclidean distance between \( f_i^a \) and \( f_j^b \). \( |X^v| \) counts the number of samples annotated with \( a \). \( \sigma^v \) is set to the average of \( H^v \).

Eq. (3) aims to learn the commonality of each label across modalities, it jointly considers the intrinsic distribution of samples annotated with a particular label within each modality, thus \( C^* \) can bridge \( X \) and \( Y \) and enable CMH. We want to remark that other distance metrics can also be used to setup \( H^v \). Our choice of Hausdorff distance is for its intuitiveness and effectiveness on qualifying two sets of samples (Hausdorff 2005). For the variants of Hausdorff distance, we choose the average Hausdorff distance because it considers more geometric relations between samples of two sets than the maximum and minimum Hausdorff distances (Zhou et al. 2012). By optimizing Eq. (3) across modalities, we can enhance the quality of extracted commonality of multi-modal data.

\( P^x \) and \( P^y \) may be highly correlated, because they are simply obtained from the individuality auto-encoders without any contrast and collaboration, and samples from different modalities share the same labels. As such, the individuality of each modality cannot be well preserved and tail labels encoded by the individuality are under-represented. Here, we minimize the correlation between \( P^x \) and \( P^y \) to capture the intrinsic individuality of each modality. Meanwhile, \( C^* \) can include more shared information of multi-modal data.

For this purpose, we use HSIC (Gretton et al. 2005) to approximately quantify the correlation between \( P^x \) and \( P^y \), for its simplicity and effectiveness on measuring linear and nonlinear interdependence/correlation between two sets of variables. The correlation between \( P^x \) and \( P^y \) is approximated as:

\[
J_2(P^v) = HSIC(P^x, P^y) = (n-1)^{-1}tr(K^x A K^y A)
\]

\[
s.t. K^u_{ab} = \kappa(P^x_a, P^y_b) = e^{-\frac{||P^x_a - P^y_b||^2}{\sigma^v}}
\]

where \( K^x \) and \( K^y \) are the kernel-induced similarity matrix from \( P^x \) and \( P^y \), respectively. \( A \) is a centering matrix: \( A = I - ee^T/n \), where \( e = (1, \cdots, 1)^T \in \mathbb{R}^n \) and \( I \) is the identity matrix.

For the decoding part, we use the extracted commonality \( C^* \) shared across modalities and the individuality of each modality to reconstruct this original modality as follows:

\[
J_3(C^*, P^v) = \sum_{v=xy} \frac{||f_{de}^v - f_{de}^v((C^*, P^v))||_F^2}{nd_v}
\]

where \( \{f_{de}^v\} \) is the corresponding decoder of each modality.

Then we can define the loss function of the individuality-commonality AE as:

\[
\min_{\theta_s} \text{Loss}1 = \alpha J_1(C^*) + \beta J_2(P^v) + J_3(C^*, P^v)
\]

\( \theta_s \) are the parameters of the individuality-commonality AE, \( \alpha \) and \( \beta \) are the parameters to control the weight of individuality and commonality. To this end, LtCMH can capture the commonality and individuality of different modalities, which will be used to create dynamic meta features in the next subsection.

**Dynamic Meta Features Learning**

For long-tail datasets, the head labels have abundant representations while the tail labels don’t (Cui et al. 2019). As a result, head labels can be easily distinguished but tail labels can not. To enrich label representation and transfer knowledge of head labels to tail ones, we propose the dynamic meta memory embedding module using direct features of \( F^x \) and \( F^y \), and the commonality \( C^* \) and individuality \( P^v \).

The modality heterogeneity is a typical issue in cross-modal information fusion, which means that different modalities may have completely different feature representations and statistical characteristics for the same semantic labels. This makes it difficult to directly measure the similarity between different modalities. However, multi-modal data that describe the same instance usually have close semantic meanings. In the previous subsection, we have learnt a commonality information matrix \( C^* \) by mining the geometric relations between samples of two labels in the embedding space. Therefore, \( C^* \) can be seen as a bridge between different modalities. Meanwhile, different modalities often have their own individual information, so we learn \( P^v \) to capture the individuality of each modality. We take \( C^* \) and \( P^v \) as the memory features, and obtain the meta features of a modality as follows:

\[
M^v = F^v + E_1 \odot C^* + E_2 \odot P^v
\]

Data-poor tail labels need more memory features to enrich their representations than data-rich head labels, we design two adaptive selection factors \( E_1 \) and \( E_2 \) on \( C^* \) and \( P^v \) for this purpose, which can be adaptive computed from \( F^v \) to derive selection weights as \( E_1 = \text{Tanh}(FC(F^v)) \) and \( E_2 = \text{Tanh}(FC(F^v)) \). \( \odot \) is the Hadamard product.

To match the general multi-modal data in a soft manner, we adopt the lightweight “Tanh+FC” to avoid complex and inefficient parameters adjusting. Different from long-tail hashing methods (Kou et al. 2022) that use prototype networks (Snell, Swersky, and Zemel 2017) or simply stack the direct features to learn meta features for each label, our \( M^v \) is created from both individuality and commonality, and direct features of multi-modal data. In addition, the dynamic factors can integrate both the enhanced commonality across modalities and individuality of each modality for cross-modal hashing, thus \( M^v \) has a good interpretability and enables a better performance without the clear division of head and tail labels.
Hash Code Learning

After obtaining the dynamic meta features, the information across modalities is reserved and each sample’s representation is enriched. Alike DCMH (Jiang and Li 2017), we use $S \in \mathbb{R}^{n \times k}$ to store the similarity of $n$ training samples. $S_{ij} = 1$ means that $x_i$ and $y_j$ are with the same label, and $S_{ij} = 0$ otherwise. Based on $M^x$, $M^y$ and $S$, we can define the likelihood function as follows:

$$p(S|M^x, M^y) = \begin{cases} \sigma(\phi^{xy}) & S_{ij} = 1 \\ 1 - \sigma(\phi^{xy}) & S_{ij} = 0 \end{cases}$$ (9)

where $\phi^{xy} = 1/2(M^x)^T M^y$ and $\sigma(\phi^{xy}) = \frac{1}{1+e^{-\phi^{xy}}}$. The smaller the angle between $M^x$ and $M^y$ is, the larger the $\phi^{xy}$ is, which makes it a higher probability that $S_{ij} = 1$ and vice versa. Then we can define the loss function for learning hash codes as follows:

$$\min_{B, \theta_x, \theta_y} \text{Loss2} = -\sum_{i,j=1}^n (S_{ij} \phi^{xy}_{ij} - \log(1 + e^{\phi^{xy}_{ij}})) + \gamma(\|B - M^x\|^2_p + \|B - M^y\|^2_p) + \eta(\|M^x1\|^2_p + \|M^y1\|^2_p)$$

s.t. $B \in \{+1, -1\}^{c \times n}$

(10)

$\phi^{xy}_{ij}$ means the $i$-th row and $j$-th column of $\phi^{xy}$. $B$ is the unified hash codes and 1 is a vector with all elements being 1. $\theta_x$ and $\theta_y$ are the parameters of image and text meta feature learning module. The first term is the negative log likelihood function, it aims at minimizing the inner product between $M^x$ and $M^y$ and preserving the semantic similarities of samples from different modalities. The Hamming distance between similar samples will be reduced and between dissimilar ones will be enlarged. The second term aims to minimize the difference between the unified hash codes and meta features of each modality. Due to the similarity preservation of $M^x$ and $M^y$ in $S$, enforcing the unified hash codes closer to $M^x$ and $M^y$ is expected to preserve the cross modal similarity to match the goal of cross modal hashing. The last term pursues balanced binary codes with fixed length for a larger coding capacity.

Optimization

There are three parameters $\theta_x$, $\theta_y$, $B$ in our hash learning loss function in Eq. (10), it is difficult to simultaneously optimize them all and find the global optimum. Considering that, we adopt a canonically-used alternative strategy that optimizes one of them with others fixed, and give the optimization as below.

Optimize $\theta_x$ with fixed $\theta_y$ and $B$: We first calculate the derivative of the loss function $\text{Loss2}$ with respect to $M^x$ for image modality and then we take the back-propagation (BP) and stochastic gradient decent (SGD) to update $\theta_x$ until convergence or the preset maximum epoch. $\frac{\partial \text{Loss2}}{\partial M^x}$ can be calculated as:

$$\frac{\partial \text{Loss2}}{\partial M^x} = \frac{1}{2} \sum_{i,j=1}^n (\sigma(\phi_{ij})M^y_{ij} - S_{ij}M^y_{ij}) + 2\gamma(M^x_{ij} - B_{ij}) + 2\eta M^x1$$ (11)
The first two are shallow solutions, and latter four are deep ones. They all focus on the commonality of multi-modal data to learn hash codes. CMFH, JIMFH and DGCPN are unsupervised solutions, while the others are supervised ones. We also take the recent long-tail single-modal hashing method LTHNet (Chen et al. 2021) as another baseline. For fair evaluation with LTHNet, we solely train and test LtCMH and LTHNet on the image-modality. The parameters of compared methods are fixed as reported in original papers or selected by a validation process. All experiments are independently repeated for ten times. The code of LtCMH is shared at www.sdu-idea.cn/codes.php?name=LtCMH. We implement LtCMH in Python 3.7 with the MindSpore deep learning framework.

**Result Analysis**

We adopt the typical mean average precision (MAP) as the evaluation metric, and report the average results and standard deviation in Table 2 (long-tailed), Table 3 (balanced) and Table 4 (single-modality). From these results, we have several important observations:

(i) LtCMH can effectively handle long-tail multi/single-modal data, this is supported by the clear performance gap between LtCMH and other compared methods. DCMH and SSAH have greatly compromised performance on long-tailed datasets, because they use semantic labels to guide hash codes learning, while tail labels do not have enough training samples to preserve the modality relationships in both the common semantic space and Hamming space. Although unsupervised CMH methods give hash codes without referring to the skewed label distributions, they are also misled by the overwhelmed samples of head labels. Another cause is that they all focus the commonality of multi-modal data, while LtCMH considers both the commonality and individuality. We note each compared method has a greatly improved performance on the balanced datasets, since they all target at balanced multi-modal data. Compared with results on long-tail datasets, the performance drop of LtCMH is the smallest, this proves its generality. LtCMH sometimes slightly loses to SSAH on the balanced datasets, that is because the adversarial network of SSAH can utilize more semantic information in some cases.

(ii) LtCMH can learn more effective meta features and achieve better knowledge transfer from head labels to tail labels than LTHNet and MetaCMH. The latter two stack direct features as memory features to transfer knowledge, they perform better than other compared methods (except LtCMH), but the stacked features may cause information overlap and they only enrich tail labels within each modality, and suffer the inconsistency caused by modality heterogeneity. LtCMH captures both individuality and commonality of multi-modal data, and the enhanced commonality helps to keep cross-modal consistency. We further separately measure the performance of LtCMH, MetaCMH and LTHNet on the head labels and tail ones, and report the results in Figure 2 and Supplementary file. We find that LtCMH gives better results than MetaCMH and LTHNet on both head and tail labels, which further suggest the effectiveness of our meta features.

(iii) The consideration of label information and data distribution improve the performance of CMH. We notice that most supervised methods perform better than unsupervised ones. Deep neural networks based solutions also perform better than shallow ones. Besides deep feature learning, SSAH and MetaCMH leverage both the label information and the data distribution, they obtain a better performance than other compared methods. LtCMH leverages these information sources in a more sensible way, and thus achieves a better performance on both long-tail and balanced datasets.

**Ablation Experiments**

To gain an in-depth understanding of LtCMH, we introduce five variants: LtCMH-w/oC, LtCMH-w/oI, LtCMH-w/oT, LtCMH-w/oMI, LtCMH-w/oMT, which separately disregard the individuality, commonality and the both, dynamic meta features from the image modality and text modality. Figure 3 shows the results of these variants on Flicker25K and NUS-WIDE. We have several important observations.

(i) Both the commonality and individuality are important for LtCMH on long-tail datasets. This is confirmed by clearly reduced performance of LtCMH-w/oC and LtCMH-w/oI, and by the lowest performance of LtCMH-w/oII. We find LtCMH-w/oI performs better than LtCMH-w/oC, since commonality captures the shared and complementary information of multi-modal data, and bridges two modalities for cross-modal retrieval. (ii) Both the meta features from the text- and image-modality are helpful for hash code learning. We note a significant performance drop when meta features of target query modality are unused.

Besides, we also study the impact of input parameters $\alpha$, $\beta$, $\gamma$ and $\eta$, which control the learning loss of commonality, individuality, consistency of hash codes across modalities, balance of hash codes. We find that a too small $\alpha$ or $\beta$ cannot ensure to learn commonality and individuality across modalities well, but a too large of them brings down the validity of AE. A too small $\gamma$ and $\eta$ cannot keep consistency of hash codes across modalities and generate balanced hash codes. Given these results, we set $\alpha=0.05$, $\beta=0.05$, $\gamma=1$, $\eta=1$.

**Conclusion**

In this paper, we study how to achieve CMH on the prevalence long-tail multi-modal data, which is a practical and important, but largely unexplored problem in CMH. We pro-
Table 2: Results (MAP) of each method on long-tailed Flickr25K and NUS-WIDE. The best results are in boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>Flickr25K</th>
<th>NUS-WIDE</th>
<th>Flickr25K</th>
<th>NUS-WIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16-bits</td>
<td>32-bits</td>
<td>64-bits</td>
<td>16-bits</td>
</tr>
<tr>
<td>CMIH</td>
<td>.597±.037</td>
<td>.597±.036</td>
<td>.597±.037</td>
<td>.409±.041</td>
</tr>
<tr>
<td>JIMF</td>
<td>.621±.038</td>
<td>.635±.050</td>
<td>.635±.045</td>
<td>.503±.051</td>
</tr>
<tr>
<td>DGCPN</td>
<td>.732±.010</td>
<td>.742±.004</td>
<td>.751±.008</td>
<td>.625±.003</td>
</tr>
<tr>
<td>DCMH</td>
<td>.710±.022</td>
<td>.721±.017</td>
<td>.735±.014</td>
<td>.573±.027</td>
</tr>
<tr>
<td>SSAH</td>
<td>.738±.018</td>
<td>.750±.010</td>
<td>.779±.009</td>
<td>.630±.001</td>
</tr>
<tr>
<td>MetaCMH</td>
<td>.708±.007</td>
<td>.716±.005</td>
<td>.724±.013</td>
<td>.612±.011</td>
</tr>
<tr>
<td>LtCMH-w/oC</td>
<td>.745±.018</td>
<td>.753±.012</td>
<td>.781±.011</td>
<td>.635±.010</td>
</tr>
<tr>
<td>LtCMH-w/oI</td>
<td>.741±.009</td>
<td>.758±.014</td>
<td>.763±.004</td>
<td>.594±.004</td>
</tr>
<tr>
<td>LtCMH-w/oIC</td>
<td>.770±.008</td>
<td>.795±.004</td>
<td>.802±.023</td>
<td>.608±.029</td>
</tr>
</tbody>
</table>

Table 3: Results (MAP) of each method on balanced Flickr25K and NUS-WIDE. The best results are in boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>Flickr25K</th>
<th>NUS-WIDE</th>
<th>Flickr25K</th>
<th>NUS-WIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16-bits</td>
<td>32-bits</td>
<td>64-bits</td>
<td>16-bits</td>
</tr>
<tr>
<td>CMIH</td>
<td>.598±.048</td>
<td>.602±.055</td>
<td>.601±.065</td>
<td>.412±.072</td>
</tr>
<tr>
<td>JIMF</td>
<td>.650±.042</td>
<td>.662±.064</td>
<td>.657±.059</td>
<td>.584±.081</td>
</tr>
<tr>
<td>DGCPN</td>
<td>.729±.015</td>
<td>.741±.008</td>
<td>.749±.014</td>
<td>.631±.008</td>
</tr>
<tr>
<td>DCMH</td>
<td>.738±.020</td>
<td>.752±.020</td>
<td>.760±.019</td>
<td>.638±.010</td>
</tr>
<tr>
<td>SSAH</td>
<td>.750±.028</td>
<td>.795±.023</td>
<td>.799±.008</td>
<td>.655±.002</td>
</tr>
<tr>
<td>MetaCMH</td>
<td>.741±.009</td>
<td>.758±.014</td>
<td>.763±.004</td>
<td>.594±.004</td>
</tr>
<tr>
<td>LtCMH-w/oC</td>
<td>.770±.008</td>
<td>.795±.004</td>
<td>.802±.023</td>
<td>.608±.029</td>
</tr>
</tbody>
</table>

Figure 3: Results of LtCMH and its variants on long-tailed Flickr25K and NUS-WIDE.

Table 4: Results of LtCMH and LTHNet on long-tailed Flickr25K and NUS-WIDE. Better results are in boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>Flickr25K</th>
<th>NUS-WIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16-bits</td>
<td>32-bits</td>
</tr>
<tr>
<td>LTHNet</td>
<td>.602</td>
<td>.573</td>
</tr>
<tr>
<td>LtCMH</td>
<td>.651</td>
<td>.647</td>
</tr>
<tr>
<td>NUS-WIDE</td>
<td>.401</td>
<td>.411</td>
</tr>
<tr>
<td></td>
<td>.473</td>
<td>.508</td>
</tr>
</tbody>
</table>
Acknowledgements
This work is supported by NSFC (No. 62031003 and 62072380) and CAAI-Huawei MindSpore Open Fund.

References


