Self-Training Based Semi-Supervised and Semi-Paired Hashing Cross-Modal Retrieval

Rongrong Jing, Hu Tian, Xingwei Zhang, Gang Zhou, Xiaolong Zheng, Dajun Zeng
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
University of Chinese Academy of Sciences, Beijing 100049, China
{jingrongrong2019, tianhu2018, zhougang2020, zhangxingwei2019, xiaolong.zheng, dajun.zeng}@ia.ac.cn

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892301

Abstract—**The aim of cross-modal retrieval is to search for flexible results across different types of multimedia data. However, the labeled data is usually limited and not well paired with different modalities in practical applications. These issues are not well addressed in the existing works, which cannot consider the semantic information about unlabeled and unpaired data, synchronously. Self-training is a well-known strategy to handle semi-supervised problems. Motivated by the self-training, this paper proposes a self-training-based cross-modal hashing framework (STCH) to tackle the semi-supervised and semi-paired challenges. In the framework, graph neural networks are used to capture potential intra-modality and inter-modality similarities to produce pseudo labels. Then the inconsistent pseudo labels of different modalities are refined with a heuristic filter to enhance the model robustness. To train STCH, we propose an alternating learning strategy to conduct the self-train by predicting pseudo labels during the training procedure, which can be seamlessly incorporated into semi-supervised and supervised learning. In this way, the proposed method can leverage sufficient semantic information to enhance the semi-supervised effect and address the semi-paired problem. Experiments on the real-world datasets demonstrate that our approach outperforms related methods on hash cross-modal retrieval.**

Keywords—cross-modal retrieval, self-training, semi-supervised, semi-paired

I. INTRODUCTION

In recent years, cross-modal retrieval technology has achieved great success in real-world applications. Cross-modal retrieval is used for implementing a retrieval task across different modalities (e.g., texts vs. images and audio vs. texts). With the rapid growth of different types of media data such as texts, images, and videos on the Internet, cross-modal retrieval provides search results across various modalities can be helpful to the users to obtain comprehensive information about the target events or topics.

The main challenge of cross-modal retrieval is the modality gap. A common approach to bridge the heterogeneity gap is representation learning, which tries to generate new representations from different modalities in the shared subspace in which the similarity between them can be measured directly. The cross-modal representation method can be further categorized real-valued learning and hashing learning, depending on whether the representation is real-value or binary. Real-valued learning methods lack scalability and efficiency in the face of large-scale and high-dimensional multimedia data. Hash learning methods can tackle the above complex scenarios by importing the data into a hamming space, such that new generated features can be used for computing distance metrics while preserving the similarity structure in the original space.

However, multi-modal data quality in the real world is often unsatisfactory due to difficult collection and expensive data labeling. In addition to limited labeled data, the data of different modalities is not well-paired. For instance, many images in the web pages have no descriptions with tags or text. Similarly, the texts could not find their corresponding images because of inaccessible URL or poor quality.

Both supervised and unsupervised hashing cross-modal methods have their advantages and disadvantages. Supervised methods[1-5] usually exploit rich label information to connect modalities to learn the hash function. However, the performance of supervised methods intensely related to the amount of labeled data. Unsupervised methods[6-9] do not rely on the label information and can preserve the semantic correlation between different modalities utilizing pair-wise similarity or dissimilarity of different modalities, yet they have limited performance without semantic labels. Semi-supervised methods[10-14] combines a small amount of labeled data with a large amount of unlabeled data during training, which is essentially a trade-off between restricted labeled data and model performance.

Although the existing semi-supervised cross-modal hashing methods perform well, most of them require data pairs to measure semantic similarity, which is only appropriate to well-paired data. Consequently, it is challenging to measure the label similarity matrix when data of different modality is not well paired. To address the problem, Partial Multi-Modal Hashing (PMMH) [15] utilized graph Laplacian to preserve the intra-modal similarity and inter-modal similarity via latent subspace learning. Semi-Paired Discrete Hashing (SPDH)[16] constructed the similarity graph via anchor data pairs of cross-view to maintain the similarities of semi-paired data. Semi-Pair Hashing (SPH) [17] method maintained cross-view correlation and within-view similarity via anchor graph.

These existing studies explore the correlation between unpaired data under weakly supervised learning, ignoring the rich semantic information contained in labels. Even though some semi-supervised work [12, 13] explored the semantic information by predicting the labels of unlabeled data with the help of limited labeled data, the accuracy of pseudo labels depends on semantic information of limited labeled data and pair-wise correlation of unlabeled data. To fully exploit both inter-modal and inter-modal similarity, this paper propose a self-training-based cross-modal hashing (STCH) method, which can handle endless unlabeled queries to enhance the effect of semi-supervised learning.

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

The main contributions of STCH are summarized as follows:

- A novel end-to-end deep hashing semi-supervised cross-modal retrieval is proposed, which can seamlessly handle paired and unpaired data under supervised and semi-supervised scenarios via an alternating learning strategy;
- We predict the pseudo label of unlabeled data and then utilize these data to enhance the semantic information of the dataset. Meanwhile, both unlabeled and unpaired samples can help subsequent training by the pseudo labels, which is not solved well under the standard situation due to their poor semantic information.
- We exploit the rich semantic information of multi-modal data. Image and text feature extractors can help extract abundant semantic features, and graph convolution networks (GCN) explore the latent intra-modality and inter-modality similarities.
- Extensive experiments demonstrate the effectiveness of our method in a variety of conditions.

The remaining chapters of this paper are organized as follows. We introduce related work in section II and describe the framework of STCH in detail in section III. The experimental results are analyzed in Section IV. We conclude this paper in section V.

II. Related Work

A. Hashing Cross-Modal Retrieval

Cross-modal retrieval has aroused the interest of researchers. The main challenge of cross-modal retrieval is measuring the semantic correlation of different modalities. In order to eliminate the diversity between heterogeneous modalities, many methods which learn multiple transformations and map different modalities into a common potential subspace have been proposed in recent years. Methods for cross-modal retrieval have evolved into two directions, one is based on real-value representation and another is based on binary representation. Cross-modal retrieval in real-value representation learning [13, 18-20] is more accurate than binary representation which is encoded in binary code of limited length. However, the binary representation method can significantly improve the training speed of the model and make the model more portable by importing the massive multi-modal data into a hamming space which is convenient to calculate the distance.

Supervised hashing cross-modal methods usually utilize rich label information to generate a connection between modalities to learn the hash function. Zhang et al. [1] proposed Semantic Correlation Maximization method (SCM), which integrates semantic labels into the hash learning process. SCM uses label vectors to obtain the semantic similarity matrix and reconstruct it by learning hash codes. Yu et al. [2] proposed Discriminative Coupled Dictionary Hashing (DCDH) to capture hidden semantic information in underlying multi-modal data.

Extending hash search to nonlinear patterns can further improve the effectiveness of the hash representation method. To obtain more complex data structures, Lin et al. [3] proposed a two-step supervised hash algorithm called SePH for cross-modal retrieval. In the training process, SePH first converts the semantic similarity of training data into a probability distribution, then approximates the probability distribution with learnable hash codes in Hamming space, and finally trains the model by minimizing KL divergence. SePH uses nuclear logistic regression and sampling strategies to learn nonlinear projections from features to hash codes in each view. Cao et al. [4] proposed Correlation Autoencoder Hashing (CAH), which maximizes the feature correlation and label semantic correlation jointly and then converts them into trainable hash codes through nonlinear autoencoders. Jiang et al. [5] proposed a Deep Cross-Modal Hashing method (DCMH), which integrates feature learning and hash code learning into the same framework.

Unsupervised hashing cross-modal methods preserve the semantic correlation between different modalities utilizing pair-wise similarity or dissimilarity of different modalities. Kumar et al. [6] proposed Cross-View Hashing (CVH) for multi-view data by extending spectral hashing. Collective Matrix Factorization Hashing (CMFH) proposed by Ding et at. [7] leverages collective matrix factorization to learn unified hash codes for cross-view similarity search. Considering the latent semantic information, Latent Semantic Sparse Hashing (LSSH) [8] generates the image and text representations by using sparse coding for images and matrix factorization for text and map them into a joint space. Wang et al. proposed Semantic Topic Multimodal Hashing (STMH) [9] to obtain the multimedia semantic concepts by exploring the data clustering patterns.

B. Self-Training

Self-training is a well-known strategy to handle the semi-supervised situation, whose effectiveness is proved in [21-23]. Pseudo-labeling as one simple and effective manner [21] of self-training is used widely for various applications, such as visual categorization [24] and person identification [25]. Pseudo-labeling regards the maximum predicted probability as the accurate label to participate in the subsequent iterative training process.

There are some studies to consider the hash cross-modal retrieval in a semi-supervised manner, which mainly utilizes the semantic information by predicting pseudo labels of unlabeled data. For instance, Semi-paired and Semi-supervised Multimodal Hashing (SSMH) [14] constructed a pseudo semantic correlation matrix by propagating semantic correlation of labeled data to all data. Semi-supervised Semantic Factorization Hashing (S3FH) [12] decomposed the predicted label matrix of unlabeled data into hash codes through the matrix factorization approach with minimizing decompose error. Mandal et al. [13] pre-trained a label prediction module by comparing the weakly supervised predicted labels of unlabeled data with its nearest neighbors, and labeled data with its actual label. However, the accuracy of pseudo labels of these methods strong depends on semantic information of limited labeled data and pair-wise correlation of unlabeled data. Thus, they cannot utilize pseudo label to enhance the data semantic information in the self-training manner.

III. Method

In this section, we introduce our proposed self-training framework and alternating learning strategy for semi-supervised and semi-paired hashing cross-modal retrieval.

A. *Problem Formulation*

Suppose two modalities of images and texts are represented as $X=\{x_i\}_{i=1}^{N_x}$ and $Y=\{y_i\}_{i=1}^{N_y}$, where N_x and N_y are the number of training samples from images and text, respectively. The first N_m samples are correlated with each other, and the unpaired samples are $X'=\{x_i'\}_{i=N_m}^{N_x}$ and $Y'=\{y_i'\}_{i=N_m}^{N_y}$, respectively. The partial labels are denoted as $L_x\in\mathbb{R}^{d_c\times N_{l_x}}$ and $L_y\in\mathbb{R}^{d_c\times N_{l_y}}$, respectively, where N_{l_x} and N_{l_y} is the number of labeled data and d_c is the dimension of one-hot code for c categories. The unlabeled data is denoted as $\hat{X}$ and $\hat{Y}$, respectively. In brief, the input of STCH are four multimodal data types, paired and labeled $\{X,Y\}$, unpaired and labeled $\{X',Y'\}$, paired and unlabeled $\{\hat{X},\hat{Y}\}$, unpaired and unlabeled $\{\widehat{X}',\widehat{Y}'\}$. Each modality of sample is learned by model as binary code $B=\{0,1\}^k$ of length k-bit. The goal of hash cross-modal retrieval is that search for the same-category data from another modality in the hamming space through calculating the similarity of hash codes.

B. *Self-training Framework*

As shown in Fig.1, the proposed cross-modal end-to-end framework includes (a) an image feature extractor $F^x=f^x(\theta^x,X)$, (b) a text feature extractor $F^y=f^y(\theta^y,Y)$, (c) two classification blocks based on GCN $G^*=f^{g^*}(\theta^{g^*},F^*)$, $*\in\{x,y\}$, (d) a pseudo label filter $F^p=f^p(G^*)$, and (d) a hash code learning block $H=f^h(X,Y,G^*)$.

The image feature extractor F^x is based on the CNN-F [26] structure which comprises eight learnable layers, five of which are convolutional, and the last three layers are fully-connected. In text feature extractor F^y, the 1386-dimension text vector, represented by bag-of-words (BOW), is fed into the two fully-connected layers according to the setting of [5].

For the features O^x and O^y extracted from image and text, we calculate the similarity matrix with cosine distance. The assumption is that the feature representation of data with high similarity is also similar. In particular, image features and text features are employed to enhance the semantic mining to establish the graph $G^*=(V,E,W)$,$*\in\{X,Y\}$, where the node set V contains both unlabeled and labeled data and the edge set E represent the intra-modality affinities among the samples. The weight matrix indicates the intra-modality relationship, defined as

$$\mathrm{W}_{ij}^*=\frac{o_i^*\cdot o_j^*}{|o_i^*|\cdot|o_j^*|}\subseteq[\rho,1], \tag{1}$$

where ρ is the threshold to filter the strong negative relationship and retain strong positive relationship.

Then we adopt GCN [27] as the classifier to predict the pseudo label L^p due to the adequate performance of semi-supervised node classification task and aggregating the intra-modality similarity. Formally, the normalized similarity matrix of intra-modality is calculated by

$$A=D^{-\frac{1}{2}}(W+I)D^{\frac{1}{2}}, \tag{2}$$

where I is the identity matrix and D is the degree matrix. Compared with traditional graph-based regularization methods[18-20], GCN adopt an efficient layer-wise propagation

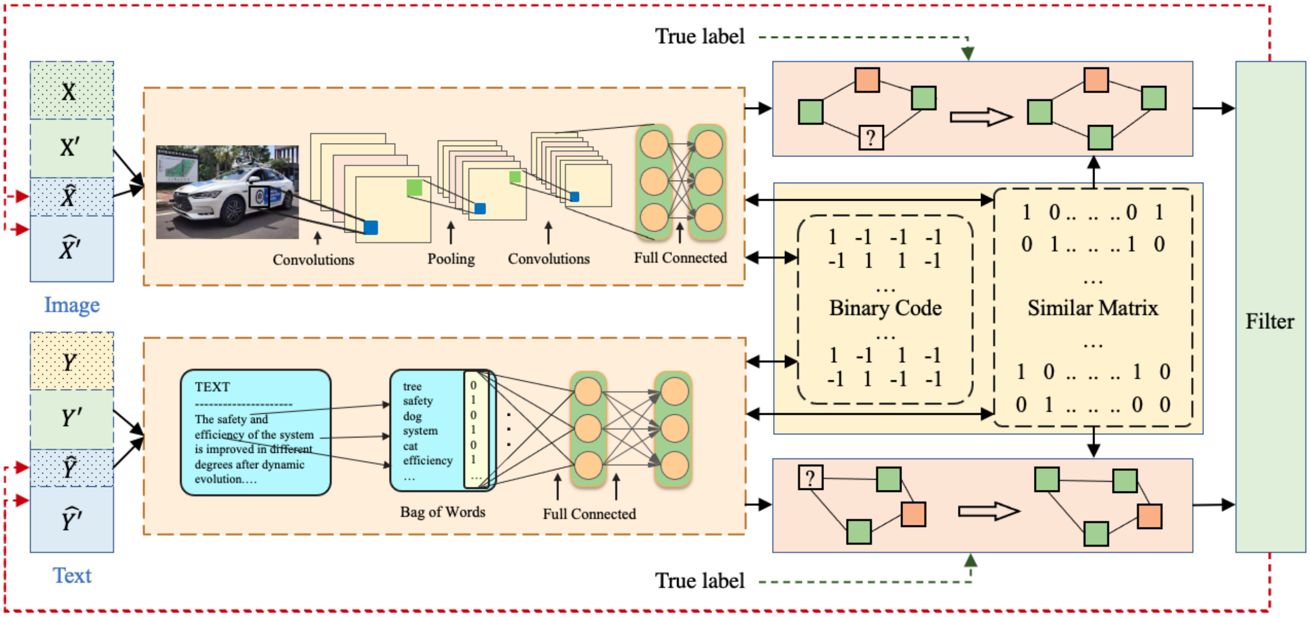

Fig. 1. The framework of our proposed STCH. We proposed an alternating learning strategy to conduct the self-train by predicting pseudo labels during the training procedure. The input of STCH are four multimodal data types, paired and labeled $\{X,Y\}$, unpaired and labeled $\{X',Y'\}$, paired and unlabeled $\{\hat{X},\hat{Y}\}$, unpaired and unlabeled $\{\widehat{X}',\widehat{Y}'\}$. The green arrows represent the forward data flow while the red arrows represent the back forward data flow. The black arrows depict the interactions between modules.

rule and each spectral graph convolution operation integrates weighted-average features from the local neighbor nodes as its own. We use a two-layer GCN as the set-up of [27] with the following layer-wise propagation rule:

$$G^{*}=\operatorname{softmax}\big(A\operatorname{ReLU}\big(AF^{*}W^{(0)}\big)W^{(1)}\big), \tag{3}$$

where $W^{(0)}$ and $W^{(1)}$ are weight matrixes.

Considering the different intra-modal similarity of image and text, and model robustness, we set two classification blocks for image and text separately rather than combine features of different modalities with the adjustable weighted hyper-parameter of different modalities. The latter approach avoids the case that predicted pseudo label differ between the two modalities, which we utilize to enhance model robustness by the filter strategy. The pseudo label filter F^{p} select the R pseudo labels that seem most credible to update the label matrix. In order to select the credible pseudo labels and enhance the robustness of model, we set the filter rules set Q: (1) When the prediction accuracy of labeled data is less than $\mathscr{q}_1$, no one is selected. (2) When the matched accuracy of paired data is less than $\mathscr{q}_2$, no one is selected. (3) When the conditions (1) and (2) are satisfied, R unlabeled samples are randomly selected. $\mathscr{q}_1$ is set 0.5 and $\mathscr{q}_2$ is set 0.5.

In the hash code learning H, pseudo label L^{p} from G^{*} and modality features o^{*} will be used to generate the hash code B as follows:

$$B^{*}=\operatorname{sign}\big(F^{*}(X;Y;\theta^{*})\big), \qquad *\in\{x,y\}. \tag{4}$$

Inspired by [5], we set $B=B^{x}=B^{y}$. With each module working together, we obtain the hash code with rich semantic information for cross-modal retrieval task.

C. Alternating Learning Strategy

Under the above framework, we adopt an alternating learning strategy to learn the parameters of STCH, which is a natural selection to take advantage of the superiority of self-training scheme. During the back propagation, we will learn the parameters of one module using stochastic gradient descent with fixing the parameters of all other modules. Under this alternating learning strategy, the pseudo label can be updated and revised in each iteration.

In particular, given the batch of raw data, F^{x} and F^{y} extract the image and text features O^{x}, O^{y}, respectively. To preserve the modality similarity, the similarity loss function of negative log likelihood is calculated as follows:

$$\mathcal{J}_s=-\sum_{i}\sum_{j}\big(\mathrm{S}_{ij}^{xy}\times\Psi_{ij}-\log\big(1+e^{\Psi_{ij}}\big)\big), \quad \forall i=1,2,\dots,N_X, \forall j=1,2,\dots,N_Y, \tag{5}$$

where $\Psi_{ij}=(o_i^{x})^{T}o_j^{y}$, and label similarity matrix S defined as follows:

$$S_{ij}=\begin{cases} +1, L_{x_i}=L_{y_j} \text{ or } x_i\in\wp(y_j) \text{ or } y_j\in\wp(x_i), \\ \varphi\left(\hat{L}_{x_i},\hat{L}_{y_j}\right), \hat{x}_i \text{ and } \hat{y}_j \text{ has pseudo labels,} \\ 0,\ \hat{x}_i \text{ and } \hat{y}_j \text{ have no labels,} \\ -1,\ L_{x_i}\neq L_{y_j}, \end{cases} \tag{6}$$

where $\wp(x)$ represents the pairs set of sample x, $\varphi(l_x, l_y)$ represents cosine similarity function. As for both unpaired and unlabeled data $\widehat{x_i'}$ and $\widehat{y_j'}$, $S_{ij}=0$ at the training beginning. Once they have pseudo labels, we can utilize their semantic information to enhance the subsequent training procedure.

```
Algorithm 1: STCH
Input: Training set $X=\left[x_1,x_2,\dots x_{N_{l_x}},\dots,x_{N_x}\right]$, $Y=\left[y_1,y_2,\dots y_{N_{l_y}},\dots,y_{N_y}\right]$, and labels $L^{x}=\left[l_1^{x},l_2^{x},\dots l_{N_{l_x}}^{x}\right]$, $L^{y}=\left[l_1^{y},l_2^{y},\dots l_{N_{l_y}}^{y}\right]$, similarity matrix $S$, prediction frequency $\ell$
Output: Parameters of image and text feature extractor, and GCN-based classifier $\theta^{x},\theta^{y},\theta^{g^{x}},\theta^{g^{y}}$, hash code matrix $B$
Initialize image feature extractor with pretrained CNN-F model
Initialize $\theta^{x},\theta^{y},\theta^{g^{x}},\theta^{g^{y}}$
for $iter = 1,2,\dots,\ max\_iters$ do
    Sample a mini-batch $x,y$ from data $X,Y$.
    Learn the image and text features by $O^{x},O^{y}=f^{x}(\theta^{x},X),f^{y}(\theta^{y},Y)$.
    Calculate $B$ by (4).
    Calculate $S$ by (6).
    Update $\theta^{x},\theta^{y}$ by using back propagation with object function as (5).
    for $iter = 1,2,\dots,\ell$ do
        Learn the predicted labels $\{L^{p},\widehat{L^{p}}\},E^{x},E^{y}$ by $O^{x},O^{y}=f^{g_x}(\theta^{g_x},O^{x}),\ f^{g_y}(\theta^{g_y},O^{y})$.
        Update $\theta^{g_x},\theta^{g_y}$ by using back propagation with object function as (10).
        Filter the pseudo labels $\hat{L}$ according rules set $Q$
        Update $L_x,L_y$,$S$ by (6) and expand the training data
    end for
end for
return $\theta^{x},\theta^{y},\theta^{g^{x}},\theta^{g^{y}},B$.
```

The hash code B is generated by hash code learning block through Equation (4). To preserve the semantic information of the hash code and balance each bit of hash code on training points, we construct a constraint loss as follows:

$$\mathcal{J}_c=\beta(\|\mathrm{O}^{x}-B\|_F^2+\|\mathrm{O}^{y}-B\|_F^2) +\gamma(\|\mathrm{B}^{x}\|_F^2+\|\mathrm{B}^{y}\|_F^2). \tag{7}$$

We update the F^{x} and F^{y} successively with the object function as

$$\min_{\theta^{*}}\mathcal{J}^{*}=\mathcal{J}_s+\mathcal{J}_c, \qquad *\in\{x,y\} \tag{8}$$

After obtaining the features of the images and texts, we exploit GCN as the semi-supervised classifier to predict pseudo label $\widehat{L_x},\widehat{L_y}$. At the same time, we obtain the characteristic representation e^{x}, e^{y} for each sample. We use the cross-entropy loss for labeled data and Mean Squared Error (MSE) for paired data to compute the prediction error as

TABLE I MAP OF HASHING SEMI-SUPERVISED METHODS ON MIRFLICKR AND NUS WIDE BENCHMARK

Dataset	Flickr 25k						Nus Wide					
Task\ Methods	I → T			T → I			I → T			T → I		
	16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
DCMH s	**0.741**	**0.746**	**0.748**	**0.782**	**0.790**	**0.793**	**0.590**	**0.601**	**0.607**	**0.638**	**0.653**	**0.659**
SePH s	0.711	0.719	0.723	0.744	0.726	0.732	0.603	0.613	0.621	0.598	0.603	0.611
JRL ss	0.562	0.563	0.584	0.588	0.589	0.598	0.551	0.558	0.573	0.455	0.458	0.473
JRL s	0.572	0.578	0.597	0.591	0.591	0.602	0.557	0.561	0.580	0.479	0.488	0.497
GSS-SL ss	0.525	0.532	0.547	0.551	0.558	0.567	0.460	0.461	0.487	0.421	0.427	0.441
GSS-SL s	0.543	0.542	0.562	0.562	0.566	0.582	0.479	0.483	0.501	0.437	0.444	0.460
STCH ss	0.587	0.590	0.612	0.634	0.631	0.643	0.511	0.513	0.531	0.502	0.509	0.523
STCH s	**0.741**	**0.746**	**0.748**	**0.782**	**0.790**	**0.793**	**0.591**	**0.601**	**0.607**	**0.638**	**0.653**	**0.659**

$$\begin{gathered}\mathcal{J}_p^* = -\tfrac{1}{n}\left(L_* \cdot log\left(\widehat{L_*}\right) + (1 - L_*) \cdot log\left(1 - \widehat{L_*}\right)\right) \\ + \tfrac{1}{m}\textstyle\sum_{x_i \in \wp(y_j)\ or\ y_j \in \wp(x_i)}\left(\widehat{L_{x_\iota}} - \widehat{L_{y_J}}\right)^2, \\ \forall i = 1,2,\dots,N_{X'}, \forall j = 1,2,\dots,N_{Y'},\end{gathered} \tag{9}$$

where $N_{X'}, N_{Y'}$ is the number of labeled samples and paired samples, respectively. In addition to exploring the intra-modality by GCN, we also measure the inter-modality to assist the paired images and text features with the same semantics by similarity loss (5), where $\Psi_{ij} = (e_i^x)^T e_j^y$. Therefore, we update the G^* with the object function to train the classifier for better pseudo label prediction as

$$\min_{\theta^{g^*}} \mathcal{J}^{g^*} = \mathcal{J}_p^* + \mathcal{J}_s\,, \qquad * \in \{x, y\} \tag{10}$$

After filtering the pseudo labels according the rules Q, we obtain the credible pseudo labels $\widehat{L_x}, \widehat{L_y}$ for partial unlabeled data, which we regard as newcome labeled data. Then we update L_x, L_y by appending $\widehat{L_x}, \widehat{L_y}$, similarity matrix S by (6). Until now, an iteration has finished. After training on enough epochs, all unlabeled data will be tagged via our proposed alternating learning strategy to conduct the self-training. We summarize the overall training algorithm in Algorithm 1.

IV. EXPERIMENTS

In this section, we evaluate the effectiveness of proposed STCH based on self-training for partial unpaired and unlabeled data, and explore the performance of different training strategy.

A. Data Sets

To evaluate our model, we conduct our experiments on two public multi-modal data sets MirFlickr-25K and Nus Wide.

1) MIRFLICKR-25K [28]: The Flickr25k obtained from the Flickr website [1] contains 25,000 image-text pairs, which belongs to 24 categories. Same as [5], 20015 samples with at least 20 textual tags are selected for experiment. We select 8000 pairs as training set, 2000 pairs as query set, and the left as retrieval set. Bag-of-words (BOW) are exploited to represent the text in 1386-dimension vectors. Imgase are resized to $3 \times 256 \times 256$ and then extracted as 4,096-dimension features of each image with pretrained CNN-F net.

2) NUS WIDE [29]*:* Nus-wide obtained from the National University of Singapore website[2] has 269,648 image-text pairs, belonging to 81 concepts. Although there were not enough samples for some concepts, according to [19], we selected samples belonging to the 10 most common labels, with at least 5000 samples per label. Then we have 69,993 samples left. We randomly select 100 pairs of each label for query, 80 pairs of each label for training, and the rest are retrieval sets, which are same as [19]. Each text is represented as 1386-dimension vectors by BOW while each image is represented as 4096-dimension vectors by pretrained CNN-F.

Due to the data from Mirflikr-25k and Nus Wide are labeled and paired, we construct four special dataset. We select $\alpha * N_t$ samples from training set as paired samples, where $\alpha \in [0,1]$, N_t is the number of training set. The left modal samples are shuffled severally to break correlation. To construct the unlabeled data, $\beta * N_t$ labels will be masked regardless of whether paired or not.

B. Experimental Setting

We employ the mean average precision (mAP)[30] to evaluate the performance of cross-modal retrieval models through calculating the average precision of R documents, as follows:

$$AP = \frac{1}{T}\sum_{r=1}^{R} P(r)\delta(r), \tag{11}$$

where T represents the number of related documents in the retrieval dataset and $P(r)$ represent the precision of the first r documents retrieved. $\delta(r) = 1$ means the r^{th} document is

[1] http://press.liacs.nl/mirflickr/

[2] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html

TABLE II. MAP OF DIFFERENT HASHING SEMI-PAIRED METHODS ON MIRFLICKR AND NUS WIDE BENCHMARK

Dataset	Flickr 25k		Nus Wide	
Methods\ Tasks	I → T	T → I	I → T	T → I
	64bits	64bits	64bits	64bits
SPDH α_1	0.623	0.605	0.647	0.610
SPDH α_2	0.614	0.597	0.632	0.600
PMMH α_1	0.572	0.584	0.568	0.579
PMMH α_2	0.568	0.559	0.554	0.553
IMH α_1	0.687	0.635	0.715	0.635
IMH α_2	0.679	0.642	0.706	0.641
STCH $\alpha_{1\ \mathrm{s}}$	0.689	0.638	0.704	0.625
STCH $\alpha_{2\ \mathrm{ss}}$	0.644	0.625	0.568	0.584
STCH $\alpha_{2\ \mathrm{s}}$	**0.697**	**0.645**	**0.720**	**0.643**

related, while the $\delta(r) = 0$ means conversely. Following [12], we set $R = 50$.

To verify the performance of STCH, seven state-of-the-arts methods are compared. JRL[20] and GSS-SL[19] are selected for semi-supervised method, DCMH[5] and SePH[3] are chosen for supervised method, and SPDH[16], PMMH[15], IMH[31] are selected for unpaired data. To be fair, supervised and semi-supervised methods only can use β of labeled data of training set according to [12], where we set $\beta = 40\%$ for semi-supervised learning while $\beta = 100\%$ for supervised learning.

We use ReLU[31] as a unique nonlinear activation function for all semantic features and Tanh to approximate symbolic hash codes. Adam optimizer [32] is adopted to update the parameters, and the initial learning rate was set to 1e-4 with 0.9 momentum. The weight decay parameters of 1e-5 is also applied to avoid over-fitting. All the baselines are finetuned to report the best performance in related literatures.

C. Comparison of Semi-Supervised Methods

To verify the performance of semi-supervised and supervised paradigms, we compare STCH with GSS-SL, JRL, DCMH, and SePH under semi-supervised and supervised learning.

Table I shows the mAP of STCG and other cross-modal retrieval algorithms with different length of hash bits, where subscript of the model represents supervised(s), semi-supervised(ss), and unsupervised(us) methods. It can be seen that the retrieval accuracy is higher when length of hash code are longer, because longer hash code can carry more semantic information. The comparison results indicate that our proposed STCH outperforms other state-of-art baselines. It is noted that when $\beta = 100\%$, STCH evolves into DCMH.

The performance of semi-supervised approach STCH was supposed to be slightly lower than the supervised methods with the same training labeled data. However, our approach still exceeds the performance of GSS-SL.

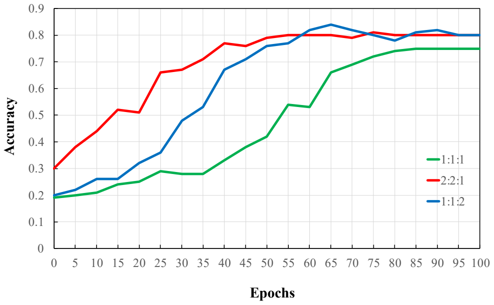

Fig.2. THE PREDICTION ACCURACY OF LABELED DATA IN THE TRAINING PROCESS.

D. Comparison of Semi-Paired Methods

As shown in Table II, we compare STCH with the state-of-art algorithms SDPH, PMMH, and IMH, which can work for partial paired situations. We set the two paired data proportion $\alpha_1 = 0.3$ and $\alpha_2 = 0.6$. We found that the results of IMH, PMMH, and SPDH are relatively unstable. For instance, the mAP of PMMH decrease as α increases while retrieval text by image. It's worth noting that, as a rule of thumb, the more paired data we have, the more information we can obtain. However, we found through experiments that excessive information will interfere with model predictions when the number of paired data reaches a limit.

E. Comparison of Alternating Learning Strategies

In order to clearly show the training process of self-training, we compare the performance of STCH with the different alternating learning strategies. As shown in Algorithm 1, the image and text feature extractors and GCN-based classifiers take over training with the different frequencies. In order to investigate the interaction between modules, we set the different strategies with various training frequency comparisons.

In the Table III, we represent the final mAP under different strategies. We assume that if the feature extractors are trained more, the classifier will be relatively weak and generation speed of pseudo labels will be limited. On the contrary, if the classifier are trained more, the text and image features could not support the classification process. Experiment shows that the optimal training ratio is 2:2:1, which may imply the importance of semantic feature representation.

TABLE III. THE PERFORMANCE OF DIFFERENT TRAINING FREQUENCY OF FEATURE EXTRACTOR AND CLASSIFIER

Proportion	I->T	T->I
1:1:1	0.5650	0.6215
1:1:2	0.5655	0.6233
1:1:3	0.5668	0.6245
2:2:1	**0.6086**	**0.6423**
3:3:1	0.6005	0.6346

Furthermore, we compare the accuracy of predicted labels with its ground truth on Flickr-25k with k as 64 in the training process. As shown in Fig.2, it is clear that the accuracy of classification and the similarity of cross-modal are rising with the increase of training epochs. The label learning abilities are different due to the significant training strategies.

V. CONCLUSION

Deep hashing cross-modal retrieval has attracted lots of attention while retrieving from large-scale multi-modal data. However, existing methods cannot solve well when the limited multi-modal data is partial labeled and partial paired. We proposed a self-training-based cross-modal hashing framework to tackle the semi-supervised and semi-paired challenges. We utilized the graph neural network to explore implied intra-modality and inter-modality similarities to produce pseudo labels, and we filter the inconsistent pseudo labels of different modalities to enhance the model robustness. To train STCH effectively, we developed an alternating learning strategy to conduct the self-train by predicting pseudo labels during the training procedure, which can be seamlessly incorporated into semi-supervised and supervised learning. Experiments on the real-world datasets demonstrate that our approach outperforms related methods on hash cross-modal retrieval.

ACKNOWLEDGMENT

We want to thank Runsong Jiang for many discussions. This work is supported the Ministry of Science and Technology of China under Grant Nos: 2020AAA0108401.

REFERENCES

[1] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proceedings of the AAAI conference on artificial intelligence*, 2014, vol. 28, no. 1.

[2] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 395-404.

[3] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3864-3872.

[4] Y. Cao, M. Long, J. Wang, and H. Zhu, "Correlation autoencoder hashing for supervised cross-modal search," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 2016, pp. 197-204.

[5] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3232-3240.

[6] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Twenty-second international joint conference on artificial intelligence*, 2011.

[7] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2075-2082.

[8] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 415-424.

[9] D. Wang, X. Gao, X. Wang, and L. He, "Semantic topic multimodal hashing for cross-media retrieval," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[10] J. Duan, Y. Luo, Z. Wang, and Z. Huang, "Semi-supervised cross-modal hashing with graph convolutional networks," in *Australasian Database Conference*, 2020: Springer, pp. 93-104.

[11] Z. Shen, D. Zhai, X. Liu, and J. Jiang, "Semi-Supervised Graph Convolutional Hashing Network For Large-Scale Cross-Modal Retrieval," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020: IEEE, pp. 2366-2370.

[12] J. Wang, G. Li, P. Pan, and X. Zhao, "Semi-supervised semantic factorization hashing for fast cross-modal retrieval," *Multimedia Tools and Applications,* vol. 76, no. 19, pp. 20197-20215, 2017.

[13] D. Mandal, P. Rao, and S. Biswas, "Semi-supervised cross-modal retrieval with label prediction," *IEEE Transactions on Multimedia,* vol. 22, no. 9, pp. 2345-2353, 2019.

[14] D. Wang, B. Shang, Q. Wang, and B. Wan, "Semi-paired and semi-supervised multimodal hashing via cross-modality label propagation," *Multimedia Tools and Applications,* vol. 78, no. 17, pp. 24167-24185, 2019.

[15] Q. Wang, L. Si, and B. Shen, "Learning to hash on partial multi-modal data," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[16] X. Shen, F. Shen, Q.-S. Sun, Y. Yang, Y.-H. Yuan, and H. T. Shen, "Semi-paired discrete hashing: Learning latent hash codes for semi-paired cross-view retrieval," *IEEE transactions on cybernetics,* vol. 47, no. 12, pp. 4275-4288, 2016.

[17] X. Shen, Q.-S. Sun, and Y.-H. Yuan, "Semi-paired hashing for cross-view retrieval," *Neurocomputing,* vol. 213, pp. 14-23, 2016.

[18] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Transactions on Multimedia,* vol. 21, no. 5, pp. 1276-1288, 2018.

[19] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Transactions on Multimedia,* vol. 20, no. 1, pp. 128-141, 2017.

[20] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 24, no. 6, pp. 965-978, 2013.

[21] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, 2013, vol. 3, no. 2, p. 896.

[22] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems,* vol. 17, 2004.

[23] B. Zoph *et al.*, "Rethinking pre-training and self-training," *Advances in neural information processing systems,* vol. 33, pp. 3833-3845, 2020.

[24] D. Mugnai, F. Pernici, F. Turchini, and A. D. Bimbo, "Soft pseudo-labeling semi-supervised learning applied to fine-grained visual classification," in *International Conference on Pattern Recognition*, 2021: Springer, pp. 102-110.

[25] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," *arXiv preprint arXiv:2001.01526,* 2020.

[26] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531,* 2014.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907,* 2016.

[28] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 39-43.

[29] 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*, 2009, pp. 1-9.

[30] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 251-260.

[31] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proceedings of the 2013 ACM SIGMOD international conference on management of data*, 2013, pp. 785-796.

[32] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375,* 2018.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.