Creating Something from Nothing: Unsupervised Knowledge Distillation for Cross-Modal Hashing

Hengtong Hu[1,2]*, Lingxi Xie[3], Richang Hong[1,2]†, Qi Tian[3]
[1]School of Computer Science and Information Engineering, Hefei University of Technology,
[2]Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology,
[3]Huawei Inc.
huhengtong.hfut@gmail.com, 198808xc@gmail.com, hongrc@hfut.edu.cn, tian.qi1@huawei.com

Abstract

In recent years, cross-modal hashing (CMH) has attracted increasing attentions, mainly because its potential ability of mapping contents from different modalities, especially in vision and language, into the same space, so that it becomes efficient in cross-modal data retrieval. There are two main frameworks for CMH, differing from each other in whether semantic supervision is required. Compared to the unsupervised methods, the supervised methods often enjoy more accurate results, but require much heavier labors in data annotation. In this paper, we propose a novel approach that enables guiding a supervised method using outputs produced by an unsupervised method. Specifically, we make use of teacher-student optimization for propagating knowledge. Experiments are performed on two popular CMH benchmarks, i.e., the MIRFlickr and NUS-WIDE datasets. Our approach outperforms all existing unsupervised methods by a large margin.

1. Introduction

Recently, with the rapid increase of multimedia data, cross-modal retrieval [37, 46, 18, 47, 7, 1, 10, 25, 22] has attracted more and more attentions in both academia and industry. The goal is to retrieve instances from one modality using a query instance from another modality, *e.g.*, finding an image with a few textual tags. One of the most popular pipeline for this purpose, named cross-modal hashing (CMH) [1, 19, 23, 47, 7], involves mapping contents in different modalities into a common hamming space. By compressing each instance into a fixed-length binary code, the storage cost can be dramatically reduced and the time complexity for retrieval is constant since the indexing structure is built with hashing codes.

	WL?	ER?	KD?
DCMH [17]		✓	
SSAH [20]		✓	
UCH [21]	✓		
UGACH [45]	✓	✓	
UKD	✓	✓	✓

Table 1. The difference between our approach and some recent cross-modal hashing methods. Here, 'WL' indicates that training without using labels, 'ER' indicates that the method utilizes extensive relevance information rather than only the pairwise information, and 'KD' indicates utilizing knowledge distillation in the training process.

State-of-the-art CMH methods can be roughly categorized into two parts, namely, supervised and unsupervised methods. Both of them learn to shrink the gap between the distributions of two sets of training data (*e.g.*, using adversarial-learning-based approaches [20, 21, 13]), but they differ from each other in whether an instance-level annotation is provided during the training stage. From this perspective, the supervised CMH methods [1, 25, 22, 34, 44], receiving additional supervision, often produce more accurate results, and the unsupervised counterparts, while achieving lower performance, are relatively easier to deploy to real-world scenarios.

This paper combines the benefits of both methods by a simple yet effective idea, known as **creating something from nothing**. The core idea is straightforward: the supervised methods do not *really* require each instance to be labeled, but they use the labels to estimate the similarity between each pair of cross-modal data. Such information, in case of no supervision, can also be obtained from calculating the distance between their feature vectors, with the features provided by a trained unsupervised CMH method. Our approach, **unsupervised knowledge distillation** (UKD),

*This work was done when the first author was an intern at Huawei Noah's Ark Lab.
†Corresponding author.

contains an unsupervised CMH module followed by another supervised one, both of which can be freely replaced by new and more powerful models in the future.

Our research paves the way towards an interesting direction that using an unsupervised method to guide a supervised method, for which CMH is a good scenario to test on. We perform experiments on two popular cross-modality retrieval datasets, *i.e.*, MIRFlickr and NUS-WIDE, and demonstrate state-of-the-art performance, outperforming existing unsupervised CMH methods by a significant margin. Moreover, we delve deep into the benefits of supervision, and point out a few directions for future research.

The remainder of this paper is organized as follows. Section 2 briefly reviews the preliminaries of cross-modal retrieval and hashing, and Section 3 describes the unsupervised knowledge distillation approach. Experimental results are shown in Section 4 and conclusions are drawn in Section 5.

2. Related Work

2.1. Cross-Modal Retrieval and Hashing

Cross-modal retrieval aims to search semantically similar instances in one modality using a query from another modality [37, 39]. Throughout this paper, we consider the retrieval task between vision and language, *i.e.*, involving images and texts. To map them into the same space, two models need to be trained, one for each modality. The goal is to make the image-text pairs with relevant semantics to be close in the feature space. To train and evaluate the mapping functions, a dataset with image-text pairs is present. The dataset is further split into a training set and a query set, *i.e.*, the testing stage is performed on the query set.

In the past decade, many efforts were made on this topic [18, 46, 37]. However, most of them suffered from high computation costs in real-world, high-dimensional data. To scale up these models to real-world scenarios, researchers often compressed the output of these models into binary vectors of a fixed length [1, 19, 10, 24], *i.e.*, Hashing codes. In this situation, this task is often referred to as cross-modality hashing.

2.2. Supervised Cross-Modal Hashing Methods

The fundamental challenges of cross-modal hashing lie in learning reliable mapping functions to bridge the modality gap. Supervised methods [25, 47, 38, 20, 39, 7] achieved this goal by exploiting semantic labels to capture rich correlation information among data from different modalities. Traditional supervised learning methods were mostly based on handcrafted features, and aimed to understand the semantic relevance in the common space. SePH [22] proposed a semantics-preserving hashing method which aimed to approximate the distribution of semantic labels with hash codes on the Hamming space via minimizing the KL-divergence. Wang *et al.* [34] proposed to leverage list-wise supervision into a principled framework of learning the hashing function.

With the rapid development of deep learning, researchers started to build supervised methods upon more powerful yet discriminative features. DCMH [17] proposed a deep cross-modal hashing method by integrating feature learning and binary quantization into one framework. SSAH [20] improved this work by proposing a self-supervised approach, which incorporated adversarial learning into cross-modal hashing. Zhang *et al.* [47] also investigated a similar idea by proposing an adversarial hashing network with an attention mechanism to enhance the measurement of content-level similarities. These supervised methods achieved superior performance, arguably by acquiring correlation information from the semantic labels of both images and texts. However, acquiring a large amount of such labels is often expensive and thus intractable, which makes the supervised approaches infeasible in the real-world applications.

2.3. Unsupervised Cross-Modal Hashing Methods

Compared with the supervised counterparts, unsupervised cross-modal hashing methods [8, 45, 13, 36] only relied on the correlation information from the paired data, making it easier to be deployed to other scenarios. These methods usually learned hashing codes by preserving inter- and intra-correlations. For example, Song *et al.* [32] proposed inter-media hashing to establish a common Hamming space by maintaining inter-media and intra-media consistency. Recently, several works introduced deep learning to improve unsupervised cross-modal hashing. UGACH [45] utilized a generative adversarial network to exploit the underlying manifold structure of cross-modal data. As an improvement, UCH [21] coupled the generative adversarial network to build two cycled networks in an unified framework to learn common representations and hash mapping simultaneously.

Despite the superiority in reducing the burden of data annotation, the accuracy of unsupervised cross-modal hashing methods is often below satisfaction, in particular, much lower than the supervised counterparts. The main reason lies in lacking the knowledge of pairwise similarity for the training data pairs. On the other hand, we notice that the output of an unsupervised model contains, though somewhat inaccurate, such semantic information. This motivates us to guide a supervised model by the output of an unsupervised model. This is yet another type of research which distills knowledge to assist model training.

3. Our Approach

In this work, we focus on the idea named **creating something from nothing**, *i.e.*, a supervised cross-modal hash-

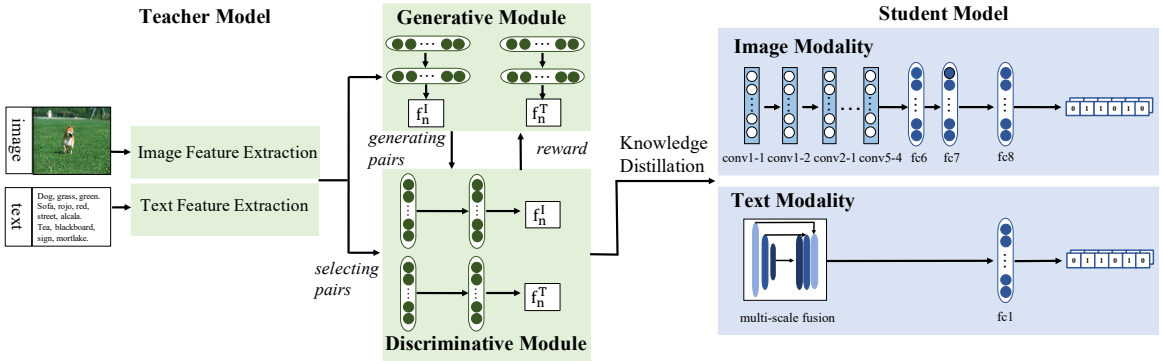

Figure 1. The proposed UKD framework which involves training a teacher model in an unsupervised manner, constructing the similarity matrix $\mathbf{S}$ by distilling knowledge from the teacher model, and using $\mathbf{S}$ to supervise the student model. Each dot represents an intermediate feature. Please zoom in to see the details of this figure.

ing method can be guided by the output of an unsupervised method, which reveals the similarity between training data pairs. Figure 1 shows the framework of the proposed UKD. In what follows, we first explain the motivation of our approach, and then introduce the proposed pipeline, **unsupervised knowledge distillation**, from two aspects, namely, how to distill similarity from an unsupervised model, and how to utilize it efficiently to optimize a supervised model.

3.1. Supervised and Unsupervised Baselines

Throughout this paper, we consider the case that the training set contains paired data, *i.e.*, $\mathcal{D} = \left\{\mathbf{v}_n^{\mathrm{I}}, \mathbf{v}_n^{\mathrm{T}}\right\}_{n=1}^{N}$, where N is the number of image-text pairs. Here, $\mathbf{v}_n^{\mathrm{I}} \in \mathbb{R}^{D_{\mathrm{I}}}$ be an image and $\mathbf{v}_n^{\mathrm{T}} \in \mathbb{R}^{D_{\mathrm{T}}}$ be a text, where the superscripts I and T denote 'images' and 'texts', and D_{I} and D_{T} denote the dimensionality of the feature spaces, respectively. D_{I} and D_{T} can be different, *e.g.*, as in our experiments. The models that map them into the same space are denoted as $\mathbf{f}_n^{\mathrm{I}} \doteq f^{\mathrm{I}}\left(\mathbf{v}_n^{\mathrm{I}}; \boldsymbol{\theta}^{\mathrm{I}}\right) \in \mathbb{R}^K$ and $\mathbf{f}_n^{\mathrm{T}} \doteq f^{\mathrm{T}}\left(\mathbf{v}_n^{\mathrm{T}}; \boldsymbol{\theta}^{\mathrm{T}}\right) \in \mathbb{R}^K$, respectively, where K is the dimensionality of the common feature space, and $\boldsymbol{\theta}^{\mathrm{I}}$ and $\boldsymbol{\theta}^{\mathrm{T}}$ are model parameters. The compressed hashing code for images and texts are denoted by $\mathbf{b}_n^{\mathrm{I}} \doteq \mathrm{sgn}\left(\mathbf{f}_n^{\mathrm{I}}\right)$ and $\mathbf{b}_n^{\mathrm{T}} \doteq \mathrm{sgn}\left(\mathbf{f}_n^{\mathrm{T}}\right)$, respectively, *i.e.*, both $\mathbf{b}_n^{\mathrm{I}}$ and $\mathbf{b}_n^{\mathrm{T}}$ fall within $\{-1, +1\}^K$.

The key to cross-modal hashing lies in recognizing which pairs of image-text data are semantically relevant while other are not, so that the model can learn to pull the features of relevant pairs closer in the common space. A straightforward idea is to define all paired image and text instances to be relevant and all others irrelevant. However, this strategy produces a very small positive set and a much larger negative set, which often causes data imbalance during the training stage. A generalized yet more effective solution is to define a similarity matrix $\mathbf{S} \in \{0, 1\}^{N \times N}$, so that when $S_{i,j} = 1$ defines a positive pair (i, j) and vice versa. The original sampling strategy is equivalent to $\mathbf{S} \equiv \mathbf{I}$.

Given $\mathbf{S}$, the objective of training involves minimizing the total distance with respect to $\boldsymbol{\theta}^{\mathrm{I}}$ and $\boldsymbol{\theta}^{\mathrm{T}}$, *i.e.*,

$$\boldsymbol{\theta}^{\mathrm{I},\star}, \boldsymbol{\theta}^{\mathrm{T},\star} = \arg\min_{\boldsymbol{\theta}^{\mathrm{I}}, \boldsymbol{\theta}^{\mathrm{T}}} = \sum_{i,j} S_{i,j} \cdot \left|\mathbf{f}_i^{\mathrm{I}} - \mathbf{f}_j^{\mathrm{T}}\right|. \quad (1)$$

Therefore, the definition of $\mathbf{S}$ forms the major challenge of the learning task. According to whether extra labels of images and texts, besides the paired information, are used, existing methods can be categorized into either supervised or unsupervised learning. In the supervised setting, instance-level annotations (*e.g.*, classification tags) are used to measure whether two instances are relevant, while in the unsupervised setting, no additional labels are available and thus the raw features are the only source of judgment. Obviously, the former provides more accurate estimation on $\mathbf{S}$ than the latter and, consequently, stronger models for cross-modal hashing. However, collecting additional annotations, even at the instance level, can be a large burden especially when the dataset is very large. Hence, we focus on improving the performance of unsupervised learning methods which are easier to be deployed to real-world scenarios.

3.2. Unsupervised Knowledge Distillation

Our idea originates from the fact that, as shown above, the difference between supervised and unsupervised cross-modal hashing algorithms is not big, but supervised methods often report much higher accuracy than the unsupervised counterparts. Moreover, the supervised algorithms do not require *real* supervision, namely, the manually labeled image/text tags, but only need to know, or estimate, the similarity between any pair of data, *i.e.*, elements in $\mathbf{S}$. Beyond the unsupervised baseline that estimates $\mathbf{S}$ using raw image/text features (extracted from a pre-trained deep network or computed using bag-of-words statistics), we seek for the possibility that a cross-modal retrieval model, after trained in an unsupervised manner, can produce a more accurate estimation of $\mathbf{S}$. We illustrate an example in Figure 2. Later,

Function	new	image	text	P@1000	P@5000
$S_{i,j} = \left(2 - \left\|\mathbf{v}_i^{\mathrm{I}} - \mathbf{v}_j^{\mathrm{I}}\right\|_2\right)/2$		✓		74.6%	64.4%
$S_{i,j} = \left(2 - \left\|\mathbf{v}_i^{\mathrm{T}} - \mathbf{v}_j^{\mathrm{T}}\right\|_2\right)/2$			✓	57.1%	55.6%
$S_{i,j} = \left(2 - \left\|\mathbf{f}_i^{\mathrm{I}} - \mathbf{f}_j^{\mathrm{I}}\right\|_2\right)/2$	✓	✓		84.6%	74.6%
$S_{i,j} = \left(2 - \left\|\mathbf{f}_i^{\mathrm{T}} - \mathbf{f}_j^{\mathrm{T}}\right\|_2\right)/2$	✓		✓	75.9%	67.9%
$S_{i,j} = \left(4 - \left\|\mathbf{f}_i^{\mathrm{I}} - \mathbf{f}_j^{\mathrm{I}}\right\|_2 - \left\|\mathbf{f}_i^{\mathrm{T}} - \mathbf{f}_j^{\mathrm{T}}\right\|_2\right)/4$	✓	✓	✓	83.9%	73.4%

Table 2. Comparison among different functions to measure the similarity between image-text pairs. All the results are computed using features extracted from a UGACH [45] model trained on the MIRFlickr dataset. Here we consider four properties: 'new' means that the new feature space, learned by the teacher model, is used; 'image' and 'text' means the corresponding features used, and 'indiv' means image and text features are used individually. P@1000/P@5000 indicate the accuracy rates among the top 1000/5000 retrieved pairs.

Figure 2. Knowledge distilled from an unsupervised model (best viewed in color). Compared to the retrieval results in the original feature space, our approach produces more accurate information about the tag of an image and, more importantly, better estimation on the relevance of image-text pairs.

we will show in experiments, with the help of *oracle* annotations, that the updated estimation of $\mathbf{S}$ is indeed more accurate in terms of finding relevant pairs.

Note that the estimated $\mathbf{S}$ can be used to train either supervised or unsupervised models, with the formulations detailed above. When $\mathbf{S}$ is used for unsupervised learning, the only effect is to provide a better sampling strategy, so as to increase the portion of *true-positive* image-text pairs in the chosen training set. This alleviates the risk that the model learns to pull the features of actually irrelevant pairs. When it is used for supervised learning, we are actually *creating something from nothing*, *i.e.*, guiding a supervised model with the output of an unsupervised model.

The proposed framework, **unsupervised knowledge distillation** (UKD), works as follows. After the teacher model has been trained, we obtain both $f^{\mathrm{I}}\left(\cdot;\boldsymbol{\theta}^{\mathrm{I}}\right)$ and $f^{\mathrm{T}}\left(\cdot;\boldsymbol{\theta}^{\mathrm{T}}\right)$ for image and text feature embedding, respectively. It remains to determine each element of $\mathbf{S}$. Without loss of generality, we assume that the feature vectors extracted from either modality, *i.e.*, $\mathbf{f}_n^{\mathrm{I}}$ or $\mathbf{f}_i^{\mathrm{T}}$, have a ℓ_2-norm. This is to ease the following calculations.

First, we point out that $S_{i,i} \equiv 1$ for all i. When $i \neq j$, $S_{i,j}$ takes four vectors, $\mathbf{f}_i^{\mathrm{I}}$, $\mathbf{f}_i^{\mathrm{T}}$, $\mathbf{f}_j^{\mathrm{I}}$ and $\mathbf{f}_j^{\mathrm{T}}$, into consideration. The design of $S_{i,j}$ can have various forms. For example, it can consider both image and text features so that $S_{i,j} = \left(4 - \left\|\mathbf{f}_i^{\mathrm{I}} - \mathbf{f}_j^{\mathrm{I}}\right\|_2 - \left\|\mathbf{f}_i^{\mathrm{T}} - \mathbf{f}_j^{\mathrm{T}}\right\|_2\right)/4$, where $|\mathbf{f}_1 - \mathbf{f}_2|_2$ is the Euclidean distance between two vectors which lies in the range of $[0, 2]$ for two normalized vectors. Also, it is possible for $S_{i,j}$ to consider only single-modal information, *e.g.*, $S_{i,j} = \left(2 - \left\|\mathbf{f}_i^{\mathrm{I}} - \mathbf{f}_j^{\mathrm{I}}\right\|_2\right)/2$ in which only image features are used for measuring similarity.

Here, we take several definitions of $S_{i,j}$ into consideration, and compare their performance in finding true-positive pairs. Results are shown in Table 2. We can observe several important properties that are useful for similarity measurement. **First**, the features trained for cross-modal hashing are indeed better than those without being fine-tuned; Second, measuring similarity in the image feature space is more accurate than that in the text feature space; Third, directly combining image and text similarity into one does not improve accuracy beyond using image similarity alone, though we expect that text features to provide auxiliary information. Motivated by these results, we use image features and text features to retrieve two lists of relevant pairs and then merge them into one. This strategy reports a precision of 76.1% at top-5000 instances, surpassing that using image and text features alone. *We fix this setting throughout the remainder of this paper.*

3.3. Models and Implementation Details

We first illustrate the supervised and unsupervised methods we have used. We take DCMH [17] as an example of supervised learning. Here we utilize the framework of DCMH but modify its architecture for higher accuracy. This model contains two deep neural networks, designed for the image modality and the text modality, respectively. The image modality network consists of 19 layers, the first eighteen layers are the same as those in VGG19 network [31], and the last layer maps features into the Hamming space. For the

text modality, a multi-scale fusion model from SSAH [20] which consists of multiple average pooling layers and a 1×1 convolutional layer is used to extract the text features. Then, a hash layer follows to map the text features into the Hamming space.

On the other hand, we investigate UGACH [45], a representative unsupervised learning method as the teacher model. It consists of a generative module and a discriminative module. The discriminator receives the data selected by the generator as negative instances, and take the data sampled using $\mathbf{S}$ as positive instances. Then a triplet loss is used to optimize to obtain better discriminate ability for the discriminator. Both the generative and discriminative modules have a two-pathway architecture, each of which has two fully-connected layers. We set the dimension of representation layer to 4096 in our experiments. The dimension of the hash layer is same as the hash code length.

For the supervised model, we take the raw pixels as inputs. In pre-processing, we resize all images into 256×256 and crop a 224×224 patches randomly. We select relevant instances for the student model by using the teacher model with the highest precision (128 bits in all experiments). We set the number of relevant instances to be $10{,}000$ for the supervised student model, and 20 for the unsupervised student model. We train our approach in a batch-based manner and set the batch size to 256. We train the model using an SGD optimizer with a weight decay of 0.01. For the compared methods, we apply the same implementations as provided in the original work.

3.4. Relationship to Previous Work

Our method is related to knowledge distillation [29, 43, 33], which was proposed to extract knowledge from a teacher model to assist training a student model. Hinton *et al.* [15] suggested that there should be some 'dark knowledge' that can be propagated during this process. Recently, many efforts were made to study what the dark knowledge is [41, 40], and/or how to efficiently take advantage of such knowledge [11, 42, 35, 2]. In particular, DarkRank [5] distilled knowledge for deep metric learning by matching two probability distributions over ranking, while our approach utilized knowledge by selecting relevant instances. On the other hand, both [42] and [27] transferred knowledge to improve the student models by designing a distillation loss, while our approach enables guiding a supervised method by an unsupervised method, in which no extra loss is used.

We also notice the connection between our approach and the self-learning algorithms for semi-supervised learning, *e.g.*, medical image analysis [48]. The shared idea is to start with a small part of labeled data (in our case, labeled image-text pairs) and try to explore the unlabeled part (in our case, other image-text pairs with unknown relevance), but the methods to gain additional supervision are different. Also, the idea that 'training a stronger model at the second time' is related to the coarse-to-fine learning approaches [14, 49] which often adopted iteration for larger improvements.

Our approach shares the same idea with some prior work that guided a supervised model with the output of an unsupervised model. DeepCluster [3] groups the features with a standard clustering algorithm and uses the subsequent assignments as supervision to update the weights of the network. Gomez *et al.* [12] performed self-supervised learning of visual features by mining a large scale corpus of multimodal (text and image) documents. Differently, our approach makes use of teacher-student optimization to combine the supervised and unsupervised models. Experiment results show the effectiveness of knowledge distillation.

4. Experiments

4.1. Datasets, Evaluation, and Baselines

We evaluate our approach on two benchmark datasets: MIRFlickr and NUS-WIDE. MIRFlickr-25K [16] consists of $25{,}000$ images downloaded from Flickr. Each image is associated with text tags and annotated with at least one among 24 pre-defined categories. Following UGACH [45], we use $20{,}015$ image-text pairs in our experiments, where $2{,}000$ are preserved as the query set and the rest are used for retrieving. We represent each image by a 4096-dimensional feature vector, extracted from a pre-trained VGGNet [31] of 19 layers, and each text by a 1386-dimensional bag-of-words features.

NUS-WIDE [6] is much larger than MIRFlickr, which contains $269{,}498$ images and the associated text tags from Flickr. It defined 81 categories, but there are considerable overlaps among them. Still, following UGACH [45], 10 largest categories and the corresponding $186{,}577$ image-text pairs are used in the experiments. We preserve 1% of data as the query database and use the rest as the retrieval set. Each image is represented by a 4096-dimensional feature vector extracted from the same VGGNet, and each text by a 1000-dimensional bag-of-words vector.

Following the convention, we adopt the mean Average Precision (mAP) criterion to evaluate the retrieval performance of all methods. The mAP score is computed as the mean value of the average precision scores for all queries.

We compare our approach against 9 previous methods. 4 of them used additional supervision (CMSSH [1], SCM [44], DCMH [17], and SSAH [20]), and while 5 others (CVH [19], PDH [28], CMFH [9], and CCQ [26]), and UGACH [45]) did not. Following our direct baseline, UGACH, we use a 19-layer VGGNet [31] pre-trained on the ImageNet dataset [30] to extract deep features and, for a fair comparison, use them to replace the features used in other baselines, including those using handcrafted features.

Task	Method	MIRFlickr-25K				NUS-WIDE			
		16	32	64	128	16	32	64	128
image → text	CMSSH [1]	0.611	0.602	0.599	0.591	0.512	0.470	0.479	0.466
	SCM [44]	0.636	0.640	0.641	0.643	0.517	0.514	0.518	0.518
	DCMH [17]	0.677	0.703	0.725	-	0.590	0.603	0.609	-
	SSAH [20]	0.797	0.809	0.810	-	0.636	0.636	0.637	-
	CVH [19]	0.602	0.587	0.578	0.572	0.458	0.432	0.410	0.392
	PDH [28]	0.623	0.624	0.621	0.626	0.475	0.484	0.480	0.490
	CMFH [9]	0.659	0.660	0.663	0.653	0.517	0.550	0.547	0.520
	CCQ [26]	0.637	0.639	0.639	0.638	0.504	0.505	0.506	0.505
	UGACH [45]	0.676	0.693	0.702	0.706	0.597	0.615	0.627	0.638
	UKD-US	0.695	0.703	0.705	0.707	0.606	0.621	0.634	0.643
	UKD-SS	**0.714**	**0.718**	**0.725**	**0.720**	**0.614**	**0.637**	**0.638**	**0.645**
text → image	CMSSH [1]	0.612	0.604	0.592	0.585	0.519	0.498	0.456	0.488
	SCM [44]	0.661	0.664	0.668	0.670	0.518	0.510	0.517	0.518
	DCMH [17]	0.705	0.707	0.724	-	0.620	0.634	0.643	-
	SSAH [20]	0.782	0.797	0.799	-	0.653	0.676	0.683	-
	CVH [19]	0.607	0.591	0.581	0.574	0.474	0.445	0.419	0.398
	PDH [28]	0.627	0.628	0.628	0.629	0.489	0.512	0.507	0.517
	CMFH [9]	0.611	0.606	0.575	0.563	0.439	0.416	0.377	0.349
	CCQ [26]	0.628	0.628	0.622	0.618	0.499	0.496	0.492	0.488
	UGACH [45]	0.676	0.692	0.703	0.707	0.602	0.610	0.628	0.637
	UKD-US	0.704	0.707	0.715	0.714	0.621	0.625	0.640	0.647
	UKD-SS	**0.715**	**0.716**	**0.721**	**0.719**	**0.630**	**0.656**	**0.657**	**0.663**

Table 3. The mAP scores of our approach and state-of-the-art competitors, in two datasets and four different code lengths. In each half, the four rows above the horizontal line contain supervised learning algorithms, while the right rows below contain unsupervised ones.

4.2. Unsupervised Student vs. Supervised Student

In Table 3, we list the accuracy, in terms of mAP, of our approach as well as other methods for comparison. On two benchmark datasets MIRFlickr and NUS-WIDE. We use 'image → text' to denote the task that images are taken as the query to retrieval the instances in the text database, and 'text → image' the task in the opposite direction. Our approach is denoted by 'UKD-US' and 'UKD-SS', with 'US' and 'SS' indicating 'unsupervised-student' and 'supervised-student', respectively.

We observe interesting results. Regarding the image → text task, UKD-SS outperforms UKD-US significantly on the MIRFlickr dataset, but the advantage on the NUS-WIDE dataset becomes much smaller. This is explained by noting that the impact brought by supervision is different between these two datasets. We consider SSAH [20] and UGACH [45], the supervised and unsupervised models we used as the students. SSAH typically outperforms UGACH by 9% on MIRFlickr, but the number is quickly shrunk to 1%–4% on NUS-WIDE. This is partially due to the larger variance of the images in NUS-WIDE, which makes it difficulty for the labeled tags to provide accurate and valuable supervision. From this perspective, the reduced advantage of UKD-SS over UKD-US is reasonable, considering that SSAH is the upper-bound of UKD-SS.

On the other hand, by introducing extra supervision, (in particular, by checking the distance between the features extracted from an unsupervised model), considerable noise (*e.g.*, inaccurate similarity measurement) is also introduced to the supervised student model. Hence, there is a trade-off between the quality and impact of these self-annotated pairs. Most often, the latter can be measured by the advantage of the supervised student model over the unsupervised one, if both can be obtained in a small reference dataset.

4.3. Comparison to the State-of-the-Arts

From Table 3, one can observe that our approach, UKD, significantly outperforms all existing unsupervised cross-modal hashing methods on both datasets, and under any length of hash code. In particular, compared to our baseline (UGACH, which is also the strongest model that ever reported results with VGGNet-19 features), UKD enjoys 3.9%, 2.5%, 2.1% and 1.3% gains (averaged over image → text and text → image) under 16, 32, 64 and 128 bits on the MIRFlickr dataset, and the corresponding numbers on the NUS-WIDE dataset are 2.3%, 3.4%, 2.0% and 1.7%, respectively. Given such a high baseline, these improvements clearly demonstrate the effectiveness of distilling knowl-

Task	Method	MIRFlickr-25K			
		16	32	64	128
image → text	GEN-0	0.676	0.693	0.702	0.706
	GEN-1	0.695	0.703	0.705	0.707
	GEN-2	0.698	0.705	0.708	0.712
text → image	GEN-0	0.676	0.692	0.703	0.707
	GEN-1	0.704	0.707	0.715	0.714
	GEN-2	0.705	0.712	0.716	0.719

Table 4. Results of training in generations for unsupervised student model on MIRFlickr-25K. 'GEN-0' and 'GEN-1' are identical to the UGACH and UKD-US models reported in Table 3, respectively.

Method	Task	MIRFlickr-25K			
		16	32	64	128
UKD-SS	image → text	0.711	0.704	0.711	0.720
	text → image	0.692	0.702	0.705	0.706

Table 5. Results of using a 16-bit teacher to guide the supervised student model on MIRFlickr-25K.

edge from the teacher model, although it is trained in an unsupervised manner. Moreover, the accuracy gain in more significant in the low-bit scenarios, arguably because richer information is provided by the teacher model which has 128 bits. On the other hand, the amount of supervision saturates with the increasing number of compressed bits. We also tried to use full-precision models to serve as the teacher, but achieved marginal gain.

4.4. Does Iteration Help?

Motivated by the consistent improvement from the teacher to the student, a question is straightforward: is it possible to further improve the performance if we continue distilling knowledge from the student, so as to guide a 'new student'? We investigate this option, and results are summarized in Table 4. We find that, compared to the significant gain brought by the first knowledge distillation, the gain of the second round is mostly marginal, *e.g.*, the average gain on the image → text task is 0.33% compared to 0.60% of the first round.

We owe this to the limited improvement of our student model in intra-modal learning – recall that we have used intra-modal similarities to choose relevant pairs. Unlike the accuracy of cross-modal retreival performance, that of intra-model retrieval, from the teacher to the student, is hardly improved. This is to say, the new batch of image-text pairs for either supervised or unsupervised learning do not have a clear advantage over the previous batch, and so the quality of training data mostly remains unchanged.

Task	Method	MIRFlickr-25K		
		16	32	64
image → text	UGACH [45]	0.603	0.607	0.616
	UCH [21]	0.654	0.669	0.679
	UKD-US	0.667	0.674	0.677
	UKD-SS	0.678	0.680	0.679
text → image	UGACH [45]	0.590	0.632	0.642
	UCH [21]	0.661	0.667	0.668
	UKD-US	0.676	0.683	0.680
	UKD-SS	0.688	0.687	0.694

Table 6. Accuracy (mAP) comparison on MIRFlickr-25K, with UGACH and UCH as the baselines. To observe how a stronger teacher model (128-bit) teaches a weaker student model, we only report 16-bit, 32-bit and 64-bit results.

4.5. Diagnostic Experiments

- **Knowledge Distillation with a Weaker Teacher**

In order to show that UKD can work under a relatively weaker teacher signal, we use a 16-bit model of UGACH [45] as the teacher. As shown in Table 5, we still achieve consistent accuracy gain beyond the baseline. However, the gain is reduced compared to using a 128-bit teacher, since the benefit of UKD is mostly determined by the quality of the similarity matrix $\mathbf{S}$, and a weaker teacher often leads to a weaker $\mathbf{S}$, *e.g.*, the precision of the top-ranked list of pairs is reduced.

- **Transferring to Other Features**

To verify that our approach is generalized to other features, we apply it to UCH [21], a recently published unsupervised cross-modal hashing method, using features extracted from a pre-trained CNN-F model [4] (same as in the original paper). Table 6 shows the comparison between UCH and our approach in terms of mAP values on MIRFlickr. Note that our baseline is still UGACH, with the features replaced, since the authors of UCH did not provide the code. One can see that both UKD-US and UKD-SS outperform UGACH (and also, UCH), and UKD-SS works better than UKD-US, *i.e.*, the same phenomena we have observed previously.

- **Sensitivity to the Number of Selected Pairs**

Next, we analyze how the performance of cross-modal hashing is related to the number of relevant pairs selected during the training process. In Figure 3, one can observe a trend of accuracy gain as the number of selected pairs increases, but when the number goes to a relatively large value, it tends to saturate and even goes down a little bit. This is related to the total number of relevant pairs in the dataset and, of course, the ability of the model in choosing relevant pairs.

We also compare our approach with the baseline in terms

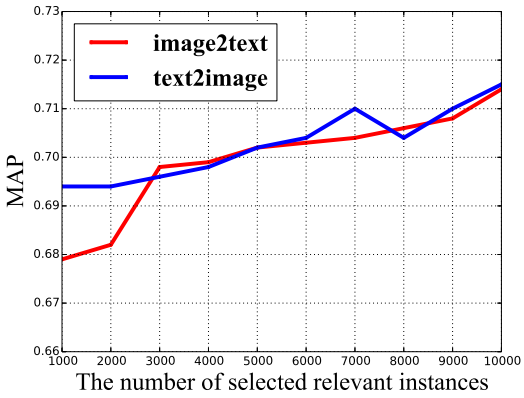

Figure 3. The mAP value with respect to the number of relevant pairs selected (tested on the MIRFlickr dataset, teacher is a 128-bit model, student is a 16-bit model).

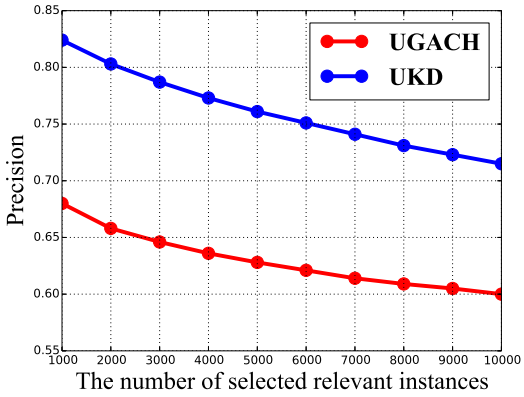

Figure 4. The top-K precision curves with respect to the number of relevant pairs selected (tested on the MIRFlickr dataset, teacher is a 128-bit model).

of the precision of the top-ranked, selected instance pairs. From Figure 4, we can see that UKD enjoys a significant advantage over UGACH, our direct baseline. Nevertheless, we see a rapid drop in precision when the number of selected pairs grows, implying that non-top-ranked pairs can introduce noise to the model. Again, this is a tradeoff between quantity and quality.

- **Qualitative Studies**

Finally, we qualitatively compare the results of our approach and the baseline. Figure 5 shows two typical examples. The text $\rightarrow$ image query (*dog*) is relatively simple, but in the original paired training set, there are no sufficient amount of labeled data for the algorithm to learn the vision-language correspondence. This is compensated with the enlarged set found by an unsupervised teacher model. In comparison, the image $\rightarrow$ text query contains complicated semantics that are even more difficult to learn, but our

Figure 5. Qualitative comparison (top: a text query with top-5 retrieved instances; bottom: an image query with top-5 retrieved instances) between our approach and UGACH (16-bit hashing), our direct baseline. Red frames and words indicate relevant images or words in the retrieved results. Note that the image query is much more difficult, as it contains semantically complicated concepts which even requires aesthetic perception to understand.

model, by making use of image-level similarity, mines extra training data from other sources (see the examples in Figure 5 which is also related to these tags). Consequently, the prediction of our approach is much better.

5. Conclusions

In this paper, we propose a novel approach to improve cross-modal hashing which enables guiding a supervised method using the outputs produced by an unsupervised method. We make use of teacher-student optimization for propagating knowledge. Superior performance can be achieved for the supervised student model by utilizing the extensive relevance information exploited from the outputs of the unsupervised teacher model. We evaluate our approach on two benchmarks MIRFlickr and NUS-WIDE, and the experiment results show that our method outperforms the state-of-the-art methods.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under grant 61722204, 61932009, and in part by the National Key Research and Development Program of China under grant 2019YFA0706200, 2018AAA0102002.

References

[1] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3594–3601. IEEE, 2010.

[2] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[4] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[5] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao 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*, page 48. ACM, 2009.

[7] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018.

[8] Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2075–2082, 2014.

[9] Guiguang Ding, Yuchen Guo, Jile Zhou, and Yue Gao. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Transactions on Image Processing*, 25(11):5427–5440, 2016.

[10] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16. ACM, 2014.

[11] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. *arXiv preprint arXiv:1805.04770*, 2018.

[12] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4230–4239, 2017.

[13] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018.

[14] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning for image captioning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[16] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43. ACM, 2008.

[17] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3232–3240, 2017.

[18] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[19] Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[20] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4242–4251, 2018.

[21] Chao Li, Cheng Deng, Lei Wang, De Xie, and Xianglong Liu. Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval. *arXiv preprint arXiv:1903.02149*, 1, 2019.

[22] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3864–3872, 2015.

[23] Hong Liu, Rongrong Ji, Yongjian Wu, Feiyue Huang, and Baochang Zhang. Cross-modality binary code learning via fusion similarity hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7380–7388, 2017.

[24] Wei Liu, Cun Mu, Sanjiv Kumar, and Shih-Fu Chang. Discrete graph hashing. In *Advances in neural information processing systems*, pages 3419–3427, 2014.

[25] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. Supervised hashing with kernels. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2074–2081. IEEE, 2012.

[26] Mingsheng Long, Yue Cao, Jianmin Wang, and Philip S Yu. Composite correlation quantization for efficient multimodal retrieval. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 579–588. ACM, 2016.

[27] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[28] Mohammad Rastegari, Jonghyun Choi, Shobeir Fakhraei, Daume Hal, and Larry Davis. Predictable dual-view hashing.

In *International Conference on Machine Learning*, pages 1328–1336, 2013.

[29] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao 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*, pages 785–796. ACM, 2013.

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[34] Jun Wang, Wei Liu, Andy X Sun, and Yu-Gang Jiang. Learning hash codes with listwise supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3032–3039, 2013.

[35] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng. Progressive teacher-student learning for early action prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3556–3565, 2019.

[36] Lin Wu, Yang Wang, and Ling Shao. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 28(4):1602–1612, 2018.

[37] Yiling Wu, Shuhui Wang, and Qingming Huang. Online asymmetric similarity learning for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4269–4278, 2017.

[38] Yiling Wu, Shuhui Wang, and Qingming Huang. Learning semantic structure-preserved embeddings for cross-modal retrieval. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 825–833. ACM, 2018.

[39] Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web*, 22(2):657–672, 2019.

[40] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan L Yuille. Training deep neural networks in generations: A more tolerant teacher educates better students. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5628–5635, 2019.

[41] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

[42] Lu Yu, Vacit Oguz Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, and Arnau Ramisa. Learning metrics from teachers: Compact networks for image embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2907–2916, 2019.

[43] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[44] Dongqing Zhang and Wu-Jun Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[45] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[46] Ting Zhang and Jingdong Wang. Collaborative quantization for cross-modal similarity search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2036–2045, 2016.

[47] Xi Zhang, Hanjiang Lai, and Jiashi Feng. Attention-aware deep adversarial hashing for cross-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 591–606, 2018.

[48] Yuyin Zhou, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot Fishman, and Alan Yuille. Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–140. IEEE, 2019.

[49] Yuyin Zhou, Lingxi Xie, Wei Shen, Yan Wang, Elliot K Fishman, and Alan L Yuille. A fixed-point model for pancreas segmentation in abdominal ct scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 693–701. Springer, 2017.