Weakly Supervised Deep Hyperspherical Quantization for Image Retrieval

Jinpeng Wang[1,2*], Bin Chen[1*†], Qiang Zhang[3], Zaiqiao Meng[4], Shangsong Liang[2†], Shutao Xia[1]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]School of Computer Science and Engineering, Sun Yat-sen University
[3]University College London
[4]University of Cambridge
wjp20@mails.tsinghua.edu.cn, cb17@tsinghua.org.cn, qiang.zhang.16@ucl.ac.uk,
zm324@cam.ac.uk, liangshangsong@gmail.com, xiast@sz.tsinghua.edu.cn

Abstract

Deep quantization methods have shown high efficiency on large-scale image retrieval. However, current models heavily rely on ground-truth information, hindering the application of quantization in label-hungry scenarios. A more realistic demand is to learn from inexhaustible uploaded images that are associated with informal tags provided by amateur users. Though such sketchy tags do not obviously reveal the labels, they actually contain useful semantic information for supervising deep quantization. To this end, we propose **W**eakly-**S**upervised **D**eep **H**yperspherical **Q**uantization (**WSDHQ**), which is the first work to learn deep quantization from weakly tagged images. Specifically, **1**) we use word embeddings to represent the tags and enhance their semantic information based on a tag correlation graph. **2**) To better preserve semantic information in quantization codes and reduce quantization error, we jointly learn semantics-preserving embeddings and supervised quantizer on hypersphere by employing a well-designed fusion layer and tailor-made loss functions. Extensive experiments show that WSDHQ can achieve state-of-art performance on weakly-supervised compact coding.

With the explosive growth of media data on the web, many retrieval tasks need to handle large-scale and high-dimensional data. Due to high computational efficiency and low memory overhead, *learning to hash* (Wang et al. 2018), as a technique in Approximate Nearest Neighbor (ANN) search, has been applied in many applications. Briefly, the goal of hashing is to transform high-dimensional data into compact binary codes while preserving semantic information. Based on the ways of measuring distance between encoded data, hashing methods can be roughly categorized into two types. **1**) *Binary hashing* (Gionis et al. 1999; Salakhutdinov and Hinton 2009) transforms high-dimensional data into hash codes in Hamming space such that distances are computed with fast bitwise operators. **2**) *Quantization* (Jegou, Douze, and Schmid 2010; Ge et al. 2013; Babenko and Lempitsky 2014; Zhang, Du, and Wang 2014; Yang, Chen, and Xia 2020) divides high-dimensional data space into disjoint cells and approximately represents each point by its cell centroid. Since the pairwise distances between data points are pre-computed by inter-centroid distances and stored in a lookup table, the search speed is accelerated.

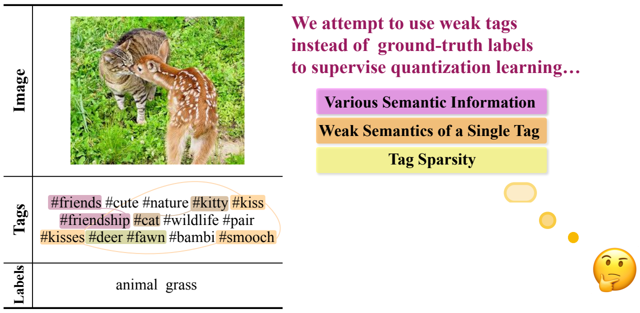


Image
Tags
#friends #cute #nature #kitty #kiss
#friendship #cat #wildlife #pair
#kisses #deer #fawn #bambi #smooch
Labels
animal grass
We attempt to use weak tags
instead of ground-truth labels
to supervise quantization learning…
Various Semantic Information
Weak Semantics of a Single Tag
Tag Sparsity

Figure 1: An example from NUS-WIDE dataset to illustrate the problem of weakly supervised quantization using tags.

Recently, deep learning has been integrated into hash models and yields superior performance over the shallow methods (Zhao et al. 2015; Huang, Chen, and Pan 2019; Jin et al. 2020). Empirically, deep quantization methods have more powerful representation capability than deep binary hashing methods for ANN search (Cao et al. 2016, 2017; Liu et al. 2018; Chen and Cheung 2019; Yuan et al. 2020). Nevertheless, there are two concerns about existing deep quantization methods. **1**) *Data hunger*: existing deep quantization methods heavily rely on high-quality supervision. However, despite the advent of large-scale annotated datasets such as ImageNet (Deng et al. 2009), the lack of well-labeled data in specific domains remains a critical bottleneck of deep learning. Collecting massive data with exact labels is usually labor-intensive and expensive, which hinders the deployment of deep quantization methods in practical large-scale applications, *e.g.* search engines and social media. **2**) *High norm variance of deep features*: deep networks often produce representation vectors with relatively high variance on the norm, which adversely degrades the quality of quantization (Wu et al. 2017; Eghbali and Tahvildari 2019).

To overcome the heavy dependence of manually annotated data, we consider taking the freely available web images with impure tags as training data, and study the novel

[*] Equal contribution.

[†] Corresponding authors.

problem of weakly supervised deep quantization, as illustrated in Fig. 1. Different from existing deep supervised quantization methods that leverage clear labels as supervision, we try to tackle the following challenges: **1**) *Tag sparsity*: the tag words are nearly unrestricted, so many synonymous tags share similar meanings; the total number of tags could be huge. **2**) *Weak semantics of a single tag*: a single tag can be semantically vague, leading to confusion in learning. **3**) *Various semantics in one image*: an image may contain multiple concepts. Besides, to further improve deep quantization, we also attempt to reduce the high norm variance of deep embeddings in the quantization model.

In this paper, we propose **W**eakly-**S**upervised **D**eep **H**yperspherical **Q**uantization (**WSDHQ**) for learning to quantize with weak tags. To our best knowledge, WSDHQ is the *first* work to address the problem of weakly-supervised deep quantization without using ground-truth labels. It explores the possibility of disconnecting the deep quantization evolution from the scaling of human-annotated datasets, given free and inexhaustible web and social media data. We make the following contributions in WSDHQ:

1) On the specific task of image quantization, we are the first to consider enhancing the weak supervision of tags. Concretely, we build a tag embedding correlation graph, to effectively enhance tag semantics and reduce sparsity.
2) To reduce the error of deep quantization, we remove the norm variance of the deep features by applying ℓ_2 normalization and maps visual representations onto a semantic hypersphere spanned by tag embeddings.
3) We further improve the ability of quantization model to better preserve semantic information into quantization codes by designing a novel adaptive cosine margin loss and a novel supervised cosine quantization loss, that directs the training of our model in an end-to-end manner.
4) Extensive experiments show that WSDHQ yields state-of-art retrieval results in weakly-supervised scenario.

Related Work

Deep Quantization. Deep quantization methods cooperating with CNNs (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016) have shown superior performance against traditional non-deep methods that use hand-crafted features (Jegou, Douze, and Schmid 2010; Ge et al. 2013; Kalantidis and Avrithis 2014; Babenko and Lempitsky 2014; Zhang, Du, and Wang 2014; Martinez et al. 2016, 2018). The goal of deep supervised quantization is to learn compact codes through deep networks that are faithful to given semantic information such as pointwise (Cao et al. 2017; Eghbali and Tahvildari 2019; Yuan et al. 2020), pairwise (Cao et al. 2016; Chen and Cheung 2019) or triplet labels (Yu et al. 2018; Liu et al. 2018). Despite the promising performance, existing methods largely rely on high-quality labels to learn satisfactory models. It limits the application of deep quantization on many real-world scenarios, where a lot of data is available without adequate ground-truths. Different from previous works, we attempt to solve the novel problem of weakly-supervised deep quantization in this paper.

Tackle Norm Variance for Quantization. It has been revealed that many adopted deep networks produce representations with relatively large norm variance, which leads to greater quantization error and unexpected performance degeneration (Wu et al. 2017; Eghbali and Tahvildari 2019). To reduce the norm variance, MSQ (Wu et al. 2017) quantized the data norms using an additional scalar quantizer before applying product quantization (PQ). Nevertheless, it is hard to balance the budget between two quantizers, and PQ is inferior to other quantization methods due to its orthogonality assumption on codebooks (Babenko and Lempitsky 2014). DSQ (Eghbali and Tahvildari 2019) removed the norm variance for deep representations by quantizing them on a unit-norm sphere, and learned to close the Euclidean distance between each image embedding and its unique class center.

In this paper, we employ a transformation layer with ℓ_2 normalization as DSQ did, to embed deep image representations onto a semantic hypersphere spanned by tag embeddings. Different from DSQ, we learn quantization in a multi-semantics (*i.e.*, multi-label and multi-class) settings by jointly optimizing two novel cosine losses, *i.e.*, adaptive cosine margin loss and supervised cosine quantization loss.

Weakly-supervised Hashing. There has been less attention paid to weakly-supervised hashing, where meta-data (weak tags) attached to web images is freely available to be an inexhaustible source of weak supervision (Gomez et al. 2018). To our knowledge, there has been no weakly-supervised quantization method for ANN search until now. Although there have been a few initial attempts (Zhang et al. 2016; Tang and Li 2018; Guan et al. 2018; Gattupalli, Zhuo, and Li 2019; Cui et al. 2020) in weakly-supervised binary hashing, they either rely on pure labels to rectify the model or directly train the model using raw tags without effective semantic enhancement, which leaves a lot of room for improvement. Zhang et al. (2016) used collaborative filtering to predict tag-label associations, where labels were requested. Guan et al. (2018) proposed a two-stage framework consisting of weakly supervised pre-training and fine-tuning using ground-truth labels. Technically, the above two methods are not ideal to be *truly* weakly-supervised hashing, because of the dependence on ground-truth information. Weakly Supervised Multimodal Hashing (WMH) (Tang and Li 2018) is the first holistically weakly-supervised hashing method, which constructs binary matrix of tags and formulates hashing as an eigenvalue problem. It tends to be vulnerable because WMH directly uses weak tags as binary labels. Weakly Supervised Deep Hashing using Tag Embeddings (WDHT) (Gattupalli, Zhuo, and Li 2019) employs Word2Vec (Mikolov et al. 2013) embeddings for tags. Although noises and vagueness are partially alleviated in WDHT, the tags of each image are roughly represented as the average vector of tag embeddings by simple weighting strategies, which is too rough to grasp various semantics.

Our WSDHQ is the *first* weakly supervised deep quantization method. Different from existing weakly supervised binary hashing methods, there are two improvements: **1**) WSDHQ investigates the semantic relations of weak tags, based on which, it further enhances tag semantics and reduces similar tags confusion so as to improve the supervision. **2**) WSDHQ preserves more fine-grained semantic representations during training, which helps to get better performance. We will discuss the effects of these two strategies (*i.e.*, loss designs) in the ablation experiment.

Proposed Approach

In this section, we first formulate the research problem and give a brief overview of our approach. Then we present the supervision enhancement and the quantization on hypersphere. Finally, we introduce the overall learning algorithm.

Problem Formulation

In a weakly-supervised image retrieval scenario, we are given a training set of N images with attached tag sets $\{\boldsymbol{x}_n, \mathcal{Y}_n\}_{n=1}^{N}$, where each image is represented as a P-dimensional vector $\boldsymbol{x}_n \in \mathbb{R}^P$ and is associated with a set of textual tags $\mathcal{Y}_n \subset \mathcal{Y}$, $\mathcal{Y}$ is a set containing all tags. While the web images with tags are easy to obtain from search engines or social media, the raw tags are not reliable enough to be ground-truth labels. Thus the goal of weakly-supervised deep quantization is to learn a compositional quantizer $q : \boldsymbol{x} \in \mathbb{R}^P \mapsto \boldsymbol{b} \in \{0,1\}^B$, which encodes each point $\boldsymbol{x}$ into a compact B-bit binary code $\boldsymbol{b}$ by preserving the weak semantics from tags via DNNs.

Overview of Our Approach

This paper enables efficient image retrieval by presenting **W**eakly **S**upervised **D**eep **H**yperspherical **Q**uantization (**WSDHQ**) in an end-to-end deep learning architecture, as shown in Fig. 2. WSDHQ consists of five main components: **1**) A standard CNN f, *e.g.* AlexNet (Krizhevsky, Sutskever, and Hinton 2012), to extract deep features of images. **2**) A word embedding model, *e.g.* Word2Vec (Mikolov et al. 2013), to represent each tag $y_i \in \mathcal{Y}$ as an embedding $\boldsymbol{t}_i$, while $\mathcal{T}$ and $\mathcal{T}_n$ are tag embedding sets *w.r.t.* raw tag sets $\mathcal{Y}$ and $\mathcal{Y}_n$. **3**) A correlation graph $\mathcal{G}$, to enhance the semantics of tags and merge synonymous tags. We denote semantically enhanced tag sets via $\mathcal{G}$ as $\widetilde{\mathcal{T}}$ and $\widetilde{\mathcal{T}}_n$. **4**) A transform layer g, to embed deep features $\boldsymbol{v}_n$ as $\boldsymbol{r}_n$ on a hypersphere, which is spanned by the normalized tags $\mathcal{S}$. **5**) A semantic hypersphere, on which the norm variances of data points are reduced to zeros. With the help of two well-customized cosine losses, our joint process of semantics-preserving learning for image embedding $\boldsymbol{r}_n$ and quantization on the hypersphere yield better compact codes for ANN search.

Enhance Weak Supervision with Tag Correlation

When using attached tags as supervision, we first consider the issues of *weak semantics of single tags* and *tag sparsity*.
Semantic Correlation Graph. We first extract the pretrained Word2Vec embedding $\boldsymbol{t}_i \in \mathbb{R}^D$ for each informative tag $y_i \in \mathcal{Y}$, where D is the dimension of textual embedding space, then form the tag embedding sets $\mathcal{T}$ and $\{\mathcal{T}_n\}_{n=1}^N$ *w.r.t.* $\mathcal{Y}$ and $\{\mathcal{Y}_n\}_{n=1}^N$. Let $\boldsymbol{T} \in \mathbb{R}^{D\times|\mathcal{T}|}$ be the tag embedding matrix. Next we obtain a k-nearest neighbor set $\mathrm{NN}^k(i)$ by cosine similarity for each tag $\boldsymbol{t}_i$. A semantic correlation graph $\mathcal{G}$ is then constructed on the total tag set $\mathcal{T}$ by an adjacency matrix $\boldsymbol{A} = (a_{ij}) \in \{0,1\}^{|\mathcal{T}|\times|\mathcal{T}|}$, where

$$a_{ij} = \begin{cases} 1, & \text{if } j \in \mathrm{NN}^k(i) \text{ and } \frac{\boldsymbol{t}_i^\top \boldsymbol{t}_j}{\|\boldsymbol{t}_i\|\|\boldsymbol{t}_j\|} \geq \tau, \text{ or just } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

and τ is the correlation threshold that determines whether a neighbor tag is related to an anchor tag in semantics. Note that each row $\boldsymbol{a}_i$ in $\boldsymbol{A}$ indicates the semantic similarities between $\boldsymbol{t}_i$ and other tags on $\mathcal{G}$.
Semantic Enhancement on Graph. Since the neighbor information from $\mathcal{G}$ can effectively enhance the semantic representations of single tags and alleviate their semantic biases, we can obtain an semantics-enhanced tag embedding matrix $\widetilde{\boldsymbol{T}} = \boldsymbol{T}\tilde{\boldsymbol{A}}^\top$ by aggregating the embeddings from its neighbor tags, where $\tilde{\boldsymbol{A}}$ is the row-wise normalization of $\boldsymbol{A}$.
Reduce Sparse Tags. After semantic enhancement, the representations of similar tags are expected to move closer and aggregate in small regions, so those sparse and redundant tags can be further detected and merged. The merging process follows classic density-based clustering algorithm (Ester et al. 1996). Any $\boldsymbol{t}'_i, \boldsymbol{t}'_j$ on $\mathcal{G}$ are *density-reachable* to each other if $d(\boldsymbol{t}'_i, \boldsymbol{t}'_j) < \epsilon$, where ϵ is the merging threshold and $d(\cdot,\cdot)$ is a distance metric, *e.g.* the ℓ_2 distance. Each time we pick up one unprocessed point $\boldsymbol{t}'_i$ from $\mathcal{G}$ in order and compute its density-reachable set $\mathcal{R}(i,\epsilon) = \{\boldsymbol{t}'_j \mid \boldsymbol{t}'_j$ is density-reachable from $\boldsymbol{t}'_i\}$. Once $|\mathcal{R}(i,\epsilon)| > 1$, we merge and replace all the elements in $\mathcal{R}(i,\epsilon)$ by the average point. Finally, we obtain refreshed embedding sets $\widetilde{\mathcal{T}}$ and $\{\widetilde{\mathcal{T}}_n\}_{n=1}^N$ by removing same embeddings in each tag set.

Quantization on Semantic Hypersphere

We adopt a standard CNN $f : \mathbb{R}^P \mapsto \mathbb{R}^V$ for extracting V-dimensional deep features $\boldsymbol{v}_n = f(\boldsymbol{x}_n)$ of image $\boldsymbol{x}_n$. $\mathcal{S} = \{\boldsymbol{t}'_i/\|\boldsymbol{t}'_i\|_2\}_{i=1}^{|\widetilde{\mathcal{T}}|}$ and $\mathcal{S}_n = \{\boldsymbol{t}'_i/\|\boldsymbol{t}'_i\|_2\}_{i=1}^{|\widetilde{\mathcal{T}}_n|}$ are the normalized total tag set and the normalized tag set of image $\boldsymbol{x}_n$, which span the D-dimensional hyperspherical semantic space. We set up a transform layer $g : \mathbb{R}^V \mapsto \mathbb{R}^D$ with ℓ_2 normalization to remove norm variance of deep features $\boldsymbol{v}_n$ and map $\boldsymbol{v}_n$ onto the semantic hypersphere as $\boldsymbol{r}_n = g(\boldsymbol{v}_n) = \frac{\sigma(\boldsymbol{W}_g \boldsymbol{v}_n)}{\|\sigma(\boldsymbol{W}_g \boldsymbol{v}_n)\|_2}$, where $\sigma(\cdot)$ is the activation, *e.g.* the hyperbolic tangent (tanh), and $\boldsymbol{W}_g$ is the parameter matrix of g. For brevity, we use $h = f \circ g$ to denote the fusion of network f and layer g such that $\boldsymbol{r}_n = h(\boldsymbol{x}_n)$.
Semantics-preserving Learning on Hypersphere. We propose an adaptive cosine margin loss $\mathcal{L}_n$ to enable semantics-preserving learning on hypersphere, namely

$$\mathcal{L}_n = \sum_{\boldsymbol{s}_i^+ \in \mathcal{S}_n} \sum_{\boldsymbol{s}_j^- \in \mathcal{N}_n} \Big[\Delta_{ij} - \cos\theta_{(\boldsymbol{s}_i^+, \boldsymbol{r}_n)} + \cos\theta_{(\boldsymbol{s}_j^-, \boldsymbol{r}_n)}\Big]_+ ,$$

where $[\cdot]_+ = \max(0,\cdot)$, $\cos\theta_{(\boldsymbol{v},\boldsymbol{v}')} = \langle \boldsymbol{v}, \boldsymbol{v}' \rangle$, s.t. $\|\boldsymbol{v}\|_2 = \|\boldsymbol{v}'\|_2 = 1$, $\mathcal{N}_n$ is the negative semantic set of $\boldsymbol{x}_n$ and Δ is the cosine margin. This metric loss encourages the image embedding $\boldsymbol{r}_n$ to move closer to positive semantic embeddings while pushing it from negative embeddings.

Ideally, we would like to utilize all negative semantic embeddings of $\boldsymbol{x}_n$, *i.e.*, $\mathcal{N}_n = \mathcal{S} \backslash \mathcal{S}_n$ for learning, but this can

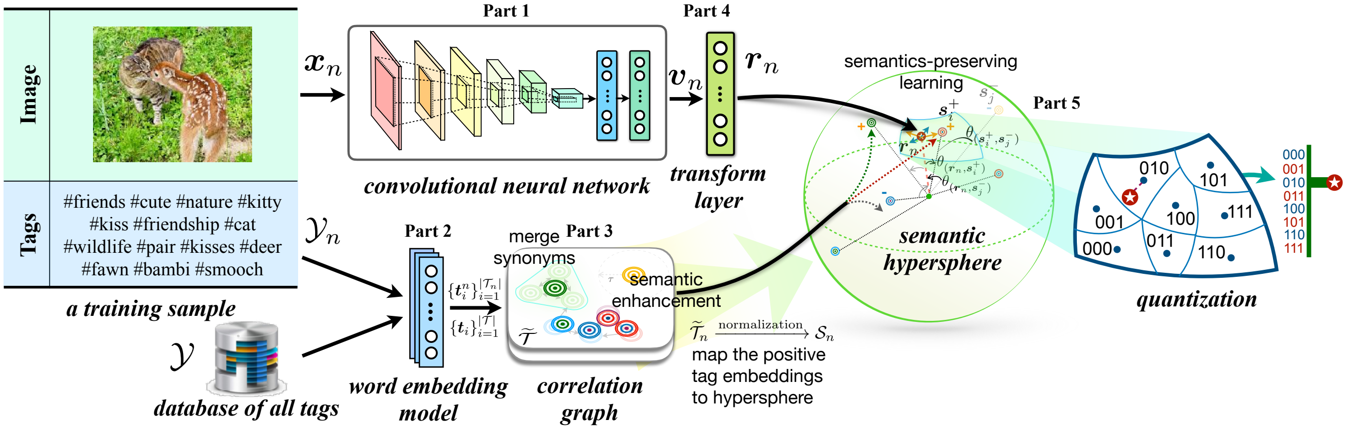


Image
Tags
#friends #cute #nature #kitty #kiss #friendship #cat #wildlife #pair #kisses #deer #fawn #bambi #smooch
a training sample
$\mathcal{Y}$
database of all tags
$\boldsymbol{x}_n$
$\mathcal{Y}_n$
Part 1
convolutional neural network
Part 4
$\boldsymbol{v}_n$
$\boldsymbol{r}_n$
transform layer
Part 2
word embedding model
$\{t_i^n\}_{i=1}^{|\mathcal{T}_n|}$
$\{t_i\}_{i=1}^{|\mathcal{T}|}$
merge synonyms
Part 3
semantic enhancement
$\widetilde{\mathcal{T}}$
correlation graph
$\widetilde{\mathcal{T}}_n \xrightarrow{\text{normalization}} \mathcal{S}_n$
map the positive tag embeddings to hypersphere
semantics-preserving learning
Part 5
$\boldsymbol{s}_i^+$
$\boldsymbol{s}_j^-$
$\theta_{(\boldsymbol{s}_i^+,\boldsymbol{s}_j^-)}$
$\theta_{(\boldsymbol{r}_n,\boldsymbol{s}_i^+)}$
$\theta_{(\boldsymbol{r}_n,\boldsymbol{s}_j^-)}$
semantic hypersphere
010
101
001
100
111
000
011
110
000
001
010
011
100
101
110
111
quantization

Figure 2: The proposed **W**eakly **S**upervised **D**eep **H**yperspherical **Q**uantization (**WSDHQ**) consists of five main parts: 1) a standard CNN, 2) a word embedding model, 3) a correlation graph, 4) a transform layer and 5) a semantic hypersphere.

lead to unacceptably high computational cost as $|\mathcal{S}|$ is usually huge in real world. In practice, we only involve K_n most tricky negative semantic embeddings as our negative set, *i.e.*,

$$\mathcal{N}_n = \operatorname*{arg\,max}_{\boldsymbol{s}_i^- \in \mathcal{S}\setminus\mathcal{S}_n,\, |\mathcal{N}_n|=K_n} \cos\theta_{(\boldsymbol{s}_i^-,\boldsymbol{r}_n)}.$$

To keep the image representations consistent with precise semantic information, we adopt an adaptive margin strategy Δ_{ij} that depends on the discrepancy between positive and negative semantics. For example, a smaller margin $\Delta_{(\text{cat, dog})}$ is acceptable while $\Delta_{(\text{kiss, dog})}$ should adapt to be larger, since the semantic representation of negative semantics "*dog*" is closer to that of positive semantics "*cat*" than positive semantics "*kiss*". Specifically, we define

$$\Delta_{ij} = 2^{1-\gamma} \cdot \left(1 - \cos\theta_{(\boldsymbol{s}_i^+,\boldsymbol{s}_j^-)}\right)^{\gamma},$$

where γ is a hyper-parameter such that smaller γ leads to larger adaptive margin under same semantic similarity.

Supervised Cosine Quantization. We propose a supervised quantizer on hypersphere to enable efficient image retrieval. Specifically, each image embedding $\boldsymbol{r}_n$ will be quantized with a set of M codebooks $\boldsymbol{C} = [\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_M]$, each $\boldsymbol{C}_m = \{\boldsymbol{c}_{m1}, \boldsymbol{c}_{m2}, \cdots, \boldsymbol{c}_{mK}\}$ contains K codewords, and each codeword $\boldsymbol{c}_{mk}$ is a D-dimensional centroid vector obtained by k-means. The codeword assignment vector $\boldsymbol{b}_n$ is segmented into M 1-of-K indicator vectors $\boldsymbol{b}_n = [\boldsymbol{b}_{1n}; \boldsymbol{b}_{2n}; \cdots; \boldsymbol{b}_{Mn}]$ *w.r.t.* M codebooks, and each one-hot indicator vector $\boldsymbol{b}_{mn}$ indicates which one of K codewords in the m-th codebook is used to compose the approximation for $\boldsymbol{r}_n$. The proposed quantizer encodes each image embedding $\boldsymbol{r}_n$ as the sum of M codewords, each of which comes from its codebook C_m assigned by an indicator vector $\boldsymbol{b}_n$, *i.e.*, $\boldsymbol{r}_n \approx \hat{\boldsymbol{r}}_n \equiv \sum_{m=1}^{M} \boldsymbol{C}_m \boldsymbol{b}_{mn}$. Then we integrate semantic supervision into quantization learning, with the cosine quantization loss formulated as

$$\mathcal{Q}_n = \sum_{\boldsymbol{s}_i \in \mathcal{S}} \left(\cos\theta_{(\boldsymbol{s}_i,\boldsymbol{r}_n)} - \cos\theta_{(\boldsymbol{s}_i,\hat{\boldsymbol{r}}_n)}\right)^2.$$

Approximate Nearest Neighbor Search. Approximate nearest neighbor (ANN) search by maximum inner-product similarity (MIPS) is a powerful tool for quantization methods (Cao et al. 2017; Liu et al. 2018). Note that on the unit hypersphere, the cosine similarity between two points can be equivalently transformed into inner-product. Hence, with the image database of N quantized binary codes $\{\boldsymbol{b}_n\}_{n=1}^{N}$, we adopt *Asymmetric Quantizer Distance* (AQD) (Jegou, Douze, and Schmid 2010) as the metric, which computes the cosine of the angle on hypersphere between a given query $\boldsymbol{q}$ and the reconstruction of a database point $\boldsymbol{x}_n$ as

$$\text{AQD}(\boldsymbol{q}, \boldsymbol{x}_n) = \cos\theta_{(\boldsymbol{r}_q,\hat{\boldsymbol{r}}_n)} = \boldsymbol{r}_q^{\top}\left(\sum_{m=1}^{M} \boldsymbol{C}_m \boldsymbol{b}_{mn}\right),$$

where $\boldsymbol{r}_q$ is the hyperspherical embedding *w.r.t.* query $\boldsymbol{q}$. We set up a query-specific lookup table of $M \times K$ items for $\boldsymbol{q}$, which stores the pre-computed results of inner-product between $\boldsymbol{q}$ and all codewords in $\boldsymbol{C}$. Hence, the AQD can be efficiently computed by summing chosen items from lookup table according to the quantization code $\boldsymbol{b}_n$.

Learning Algorithm

Overall Objective. WSDHQ enables efficient image retrieval in an end-to-end architecture, which jointly learns semantics-preserving hyperspherical embedding and supervised quantization in a total objective as

$$\min_{\mathcal{W},\boldsymbol{C},\boldsymbol{B}} \sum_{n=1}^{N} \left(\mathcal{L}_n + \lambda \mathcal{Q}_n\right), \qquad (1)$$

where λ is a positive hyper-parameter to balance the adaptive cosine margin loss $\mathcal{L}$ and the semantically supervised quantization loss $\mathcal{Q}$, and $\mathcal{W}$ denotes the learnable network parameters. Through optimizing objective (1), WSDHQ preserves the semantic information of tags into hyperspherical embeddings while effectively reducing quantization error.

There are three sets of variables in objective (1): **1)** the network parameters $\mathcal{W}$, **2)** N binary codes $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_N] = [\boldsymbol{B}_1; \boldsymbol{B}_2; \cdots; \boldsymbol{B}_M]$, where $\boldsymbol{B}_m = [\boldsymbol{b}_{m1}, \boldsymbol{b}_{m2}, \cdots, \boldsymbol{b}_{mN}]$ is the subcode matrix of all points *w.r.t.* the m-th codebook $\boldsymbol{C}_m$ and **3)** M codebooks $\boldsymbol{C} =$

$[\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_M]$. We adopt an commonly used alternating optimization paradigm (Liu et al. 2018), which iteratively optimizes one variable set while fixing the others.

Learning $\mathcal{W}$. Many back-propagation algorithms can be adopted to optimize the network parameters $\mathcal{W}$. With the advance of automatic differentiation techniques in mainstream machine learning libraries (Abadi et al. 2016; Paszke et al. 2019), it is easy to optimize $\mathcal{W}$ within a few code lines.

Learning $\boldsymbol{B}$. We update N quantization codes $\boldsymbol{B}$ by fixing $\mathcal{W}$ and $\boldsymbol{C}$ as known variables. Since each $\boldsymbol{b}_n$ is independent with $\{\boldsymbol{b}_{n'}\}_{n' \neq n}$, the optimization of $\boldsymbol{B}$ can be split into N subproblems, e.g., we optimize $\boldsymbol{b}_n$ as

$$\min_{\boldsymbol{b}_n} \left(\boldsymbol{r}_n - \sum_{m=1}^{M} \boldsymbol{C}_m \boldsymbol{b}_{mn} \right)^\top \boldsymbol{\Sigma}_{\boldsymbol{S}} \left(\boldsymbol{r}_n - \sum_{m=1}^{M} \boldsymbol{C}_m \boldsymbol{b}_{mn} \right), \tag{2}$$

$$\text{s.t. } \|\boldsymbol{b}_{mn}\|_0 = 1,\ \boldsymbol{b}_{mn} \in \{0,1\}^K,$$

where $\boldsymbol{\Sigma}_{\boldsymbol{S}} = \sum_{\boldsymbol{s}_i \in \boldsymbol{S}} \boldsymbol{s}_i \boldsymbol{s}_i^\top$ is the covariance matrix of semantic embeddings, which reflects the latent distribution of queries as all visual features will be eventually embedded to the hypersphere spanned by these embeddings. Objective (2) is generally an NP-hard problem of discrete combinatorial optimization, while some stochastic local search algorithms can provide a promising solution. We take the widely used Iterated Conditional Modes (ICM) algorithm (Besag 1986; Zhang, Du, and Wang 2014) to solve it.Given fixed $\{\boldsymbol{b}_{m'n}\}_{m' \neq m}$, we update $\boldsymbol{b}_{mn}$ by exhaustively checking all codewords in $\boldsymbol{C}_m$ and finding the codeword such that objective (2) is minimized. In each encoding iteration, the M indicators $\{\boldsymbol{b}_{mn}\}_{m=1}^{M}$ are calculated alternatively in this way. Moreover, we inject some stochastic relaxations (Zeger et al. 1992; Martinez et al. 2018) into ICM to avoid local optimum so as to improve the performance of quantization. Specifically, for the i-th encoding iteration, we add an iteration-decaying perturbation $\boldsymbol{\pi}(i) = (T(i)/M) \cdot \boldsymbol{\epsilon}$ to codebooks $\boldsymbol{C}$ and get perturbed codebooks $\tilde{\boldsymbol{C}} = \boldsymbol{C} + \boldsymbol{\pi}(i)$, where $T(i) = \sqrt{1 - (i/I)}$ is the temperature scheduled for the i-th iteration among all I encoding iterations, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \operatorname{diag}(\operatorname{cov}(\boldsymbol{R}))$ is the diagonal covariance proportional to $\boldsymbol{R}$. Finally, we replace the codebook $\boldsymbol{C}_m$ in objective (2) with $\tilde{\boldsymbol{C}}_m$ before learning $\{\boldsymbol{b}_{mn}\}_{n=1}^{N}$.

Learning $\boldsymbol{C}$. We update M codebooks $\boldsymbol{C}$ by fixing $\mathcal{W}$ and $\boldsymbol{B}$ as known variables, and rewrite the objective (1) as

$$\min_{\boldsymbol{C}} \operatorname{tr}\left((\boldsymbol{R} - \boldsymbol{C}\boldsymbol{B})^\top \boldsymbol{\Sigma}_{\boldsymbol{S}} (\boldsymbol{R} - \boldsymbol{C}\boldsymbol{B}) \right), \tag{3}$$

where $\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N]$ is the image embedding matrix. The optimal solution of objective (3) comes in a closed form as $\boldsymbol{C} = \boldsymbol{R}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)^{-1}$. Note that the binary matrix $\boldsymbol{B}$ satisfies structurally regular conditions: **1**) Diagonal $\boldsymbol{B}_i^\top \boldsymbol{B}_i$ can be computed as histograms of the codes in $\boldsymbol{B}_i$, and $\boldsymbol{B}_i^\top \boldsymbol{B}_j = (\boldsymbol{B}_j^\top \boldsymbol{B}_i)^\top$ can be computed as bivariate histograms of the codes in $\boldsymbol{B}_i$ and $\boldsymbol{B}_j$. **2**) $\boldsymbol{R}\boldsymbol{B}_i$ can be computed by treating the columns of $\boldsymbol{B}_i$ as binary vectors that select the columns of $\boldsymbol{R}$ to sum together. With the help of these properties, the optimal solution of $\boldsymbol{C}$ can be solved with time complexity $\mathcal{O}(MND)$ and $\mathcal{O}(M^2N)$, compared with their naïve computations of $\mathcal{O}(MKND)$ and $\mathcal{O}(M^2K^2N)$ (Martinez et al. 2018).

Experiments

In this section, we conduct extensive experiments to evaluate our proposed WSDHQ model with several state-of-art shallow and deep hashing methods on two web image datasets.

Setup

To our best knowledge, there are only two large-scale and commonly-used web image datasets (**MIR-FLICKR25K**, **NUS-WIDE**) that contain image-tag-label triplets. (*E.g.*, **ImageNet** does not contain tag information.) Thus we conduct our empirical evaluations on them.

MIR-FLICKR25K (Huiskes and Lew 2008) is a dataset of 25,000 Flickr images associated with 1,386 tags. The authors labelled the images with 38 semantic concepts in total, which are only used for evaluation in our experiments. 2,000 images are randomly sampled as test queries and the rest are used as retrieval database and training images.

NUS-WIDE (Chua et al. 2009) is a large-scale web image dataset also collected from Flickr, which contains 269,648 images with 5,018 tags provided by users. The authors have manually annotated each image with a pre-defined set of 81 ground-truth labels, which are only used for evaluation. We collect a subset of 193,752 images with the 21 most frequent labels for experiments. We follow (Cao et al. 2017; Liu et al. 2018) to randomly sample 5,000 images as queries and remain the rest as the database, from which we further sample 10,000 images and their tag sets as training data.

Following standard evaluation protocols adopted in previous works (Cao et al. 2017; Liu et al. 2018; Gattupalli, Zhuo, and Li 2019), we use three evaluation metrics: Mean Average Precision (**MAP**), Precision-Recall curves (**PR**), and Precision curves *w.r.t.* the number of top returned results (**P@N**). To make fair comparisons, all methods use identical training and test sets, and we follow previous works to adopt MAP@5000 for both datasets. Given a query, the ground truth is defined as: if a result shares at least one common label with the query, it is relevant; otherwise it is irrelevant.

We compare the retrieval performance of the proposed **WSDHQ** model with several state-of-art hashing methods, including: **1**) Five shallow unsupervised methods, **LSH** (Gionis et al. 1999), **SH** (Weiss, Torralba, and Fergus 2009), **SpH** (Lee 2012), **ITQ** (Gong et al. 2013) and **AQ** (Babenko and Lempitsky 2014). **2**) One deep unsupervised method, **DeepBit** (Lin et al. 2016). **3**) Two shallow weakly-supervised methods, **WMH** (Tang and Li 2018) and **WDH** (Cui et al. 2020). **4**) One deep weakly-supervised method, **WDHT** (Gattupalli, Zhuo, and Li 2019).

We use AlexNet (Krizhevsky, Sutskever, and Hinton 2012) to extract 4096-dimensional deep $fc7$ features from each image for shallow models. For deep models, we directly use raw image pixels as input and adopt AlexNet ($conv1 \sim fc7$, pre-trained on ImageNet) as the backbone network. We take the Word2Vec (Mikolov et al. 2013) as word embedding model and represent each tag with a 300-dimensional pre-trained embedding.

Dataset	MIR-FLICKR25K				NUS-WIDE			
	8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
LSH	0.524	0.570	0.562	0.572	0.376	0.392	0.413	0.418
SH	0.592	0.609	0.617	0.604	0.498	0.505	0.477	0.492
SpH	0.556	0.582	0.579	0.586	0.463	0.448	0.464	0.461
ITQ	0.641	0.623	0.654	0.633	0.536	0.545	0.556	0.563
AQ	0.637	0.645	0.658	0.661	0.524	0.567	0.587	0.592
DeepBit	0.628	0.632	0.623	0.608	0.542	0.555	0.558	0.552
WMH	0.656	0.684	0.672	0.671	0.558	0.592	0.605	0.601
WDH	0.669	0.678	0.694	0.685	0.577	0.602	0.618	0.627
WDHT	0.704	0.733	0.737	0.724	0.652	0.670	0.682	0.692
WSDHQ	**0.744**	**0.751**	**0.765**	**0.772**	**0.716**	**0.722**	**0.738**	**0.731**

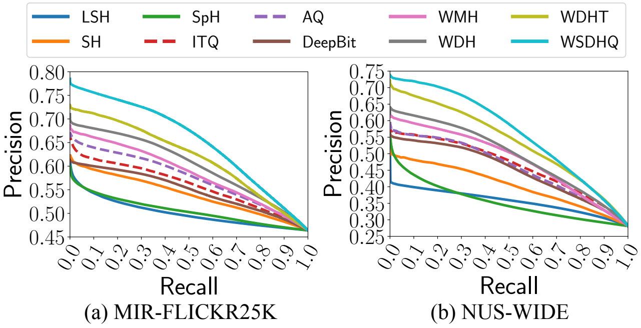
Table 1: Mean Average Precision (MAP) Results for Different Number of Bits on the Two Benchmark Image Datasets.

We implement WSDHQ based on TensorFlow (Abadi et al. 2016). For the semantic correlation graph, we set the maximum number of neighbors $k = 20$ for each tag, the correlation threshold τ as 0.75 and the merging threshold ϵ as 0.1. We set the number of tags for negative tags selected in the adaptive cosine margin loss as $K_n = 1000$. We fine-tune all layers copied from pre-trained model and train the transform layer via back-propagation from scratch. We adopt a mini-batch Adam with default parameters as optimizer. Besides, we select learning rate from $10^{-5} \sim 10^{-2}$, the hyper-parameter λ from $10^{-5} \sim 10^{-1}$ and γ from $[0.3, 0.5, 0.7, 1, 2, 3, 4]$ via cross-validation. Following (Cao et al. 2016, 2017; Liu et al. 2018; Eghbali and Tahvildari 2019), we adopt $K = 256$ codewords for each codebook, thus the binary index for each image of all M codebooks requires $B = M \log_2 K = 8M$ bits (*i.e.*, M bytes).

Results

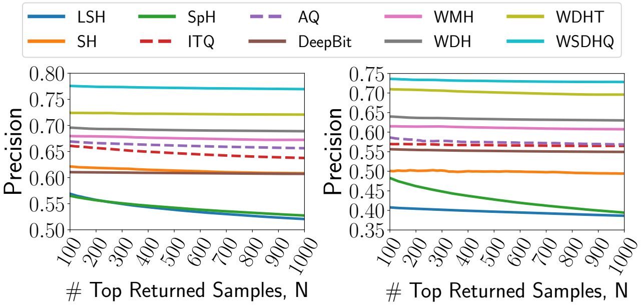
The MAP results of all methods are reported in Table 1, which shows that the proposed WSDHQ model substantially outperforms all the comparison methods. Specifically, compared to AQ (shallow quantization with deep features as input), the best unsupervised hashing method, WSDHQ achieves absolute increases of **10.8%**, **15.9%** in the average MAP on MIR-FLICKR25K and NUS-WIDE, respectively. Compare to WDHT (deep binary hashing), the state-of-art weakly-supervised hashing method, WSDHQ outperforms WDHT by appreciable margins of **3.4%** and **5.3%** in average MAP on two datasets, respectively.

We discover several interesting insights from the MAP results. **1**) The weak tags can actually be utilized as supervision. Weakly-supervised methods (*e.g.* WSDHQ and WDHT) significantly outperform unsupervised methods (*e.g.* AQ and DeepBit). **2**) The quantization often shows superiority over binary hashing under the same code length, for instance, the shallow unsupervised quantization method AQ achieves better performance than shallow unsupervised binary hashing competitors SpH, SH as well as LSH, and our WSDHQ also outperforms deep weakly-supervised binary hashing WDHT. **3**) Deep quantization methods (*e.g.* WSDHQ) can learn better codes by jointly preserving semantics and reducing the quantization error, significantly outperforming the shallow counterparts ITQ and AQ with deep features.



(a) MIR-FLICKR25K (b) NUS-WIDE

Figure 3: Precision-recall curves on the MIR-FLICKR25K and NUS-WIDE datasets with binary codes @ 32 bits.



(a) MIR-FLICKR25K (b) NUS-WIDE

Figure 4: Precision@top-N curves on the MIR-FLICKR25K and NUS-WIDE datasets with binary codes @ 32 bits.

The retrieval performance in terms of Precision-Recall curves and Precision curves *w.r.t.* different numbers of top returned samples are shown in Figures 3 and 4, respectively. These metrics are widely used in practical image search systems. The proposed WSDHQ model significantly outperforms all comparison methods by large margins in the two metrics. In particular, WSDHQ achieves much higher precision than all compared baselines at low recall levels or when the number of top returned samples is small. This is desirable for precision-oriented retrieval in practical systems, in which users usually pay more attention to the top-N returned results with a relatively small N.

Ablation Study

We investigate four variants of WSDHQ: **1**) **WSDHQ**$_{\mathcal{G}}$, a variant of WSDHQ that removes the semantic correlation graph $\mathcal{G}$ as well as the steps of semantics enhancement and synonyms merging (*i.e.*, sparse tags reducing) on $\mathcal{G}$. **2**) **WSDHQ**$_{\mathbf{N}}$, a variant which does not involve ℓ_2 normalization for both tag embeddings and image embeddings. Note that this variant conducts semantics-preserving learning no longer on hypersphere, we replace all the cosine similarity $\cos\theta_{(\boldsymbol{v},\boldsymbol{v}')}$ by inner-product $\langle \boldsymbol{v}, \boldsymbol{v}' \rangle$ between any two D-dimensional vectors $\boldsymbol{v}$ and $\boldsymbol{v}'$ in loss functions $\mathcal{L}_n$ and $\mathcal{Q}_n$ for all training samples. During retrieval, the AQD between any query $\boldsymbol{q}$ and database image $\boldsymbol{x}_n$ is computed with inner-product, *i.e.*, $\text{AQD}(\boldsymbol{q}, \boldsymbol{x}_n) = \boldsymbol{r}_q^\top \hat{\boldsymbol{r}}_n$, where $\boldsymbol{r}_q \in \mathbb{R}^D$ is the

Dataset	MIR-FLICKR25K				NUS-WIDE			
	8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
WSDHQ$_{\mathcal{G}}$	0.728	0.736	0.749	0.751	0.708	0.711	0.725	0.717
WSDHQ$_{\mathrm{N}}$	0.724	0.727	0.740	0.755	0.703	0.716	0.723	0.726
WSDHQ$_{\mathcal{L}}$	0.711	0.740	0.744	0.753	0.698	0.701	0.713	0.701
WSDHQ$_2$	0.725	0.736	0.731	0.747	0.706	0.707	0.702	0.694
WSDHQ	**0.744**	**0.751**	**0.765**	**0.772**	**0.716**	**0.722**	**0.738**	**0.731**

Table 2: Mean Average Precision (MAP) Results of WSDHQ and Its Variants on Two Benchmark Datasets.

transformed embedding of $\boldsymbol{q}$ and $\hat{\boldsymbol{r}}_n \in \mathbb{R}^D$ is the quantization reconstruction of image $\boldsymbol{x}_n$. **3**) **WSDHQ**$_{\mathcal{L}}$, a variant of WSDHQ which replaces the adaptive cosine margin loss for semantics-preserving learning with the mini-batch-wise hinge loss in WDHT (Gattupalli, Zhuo, and Li 2019) as

$$\mathcal{L}_n = \sum_{\bar{\boldsymbol{s}}_j \in \bar{\mathcal{S}}^b \backslash \{\bar{\boldsymbol{s}}_n\}} \max\left(0, \delta + \bar{\boldsymbol{s}}_j^\top \boldsymbol{r}_n - \bar{\boldsymbol{s}}_n^\top \boldsymbol{r}_n\right),$$

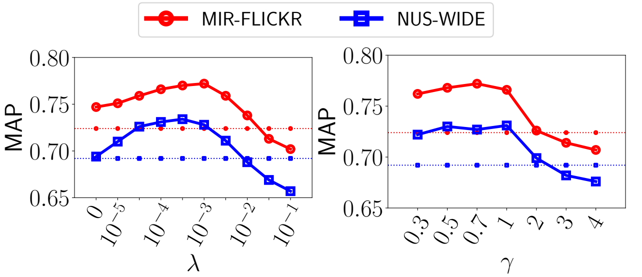
where $\bar{\boldsymbol{s}}_n = \frac{1}{|\mathcal{S}_n|} \sum_{\boldsymbol{s}_i \in \mathcal{S}_n} \boldsymbol{s}_i$ is the average point of all the semantic embeddings in $\mathcal{S}_n$, $\bar{\mathcal{S}}^b$ is the average semantic embedding set of the b-th mini-batch, and $\delta = 0.2$ is the predefined margin recommended by the authors of WDHT. **4**) **WSDHQ**$_2$, the two-stage variant which separately learns semantics-preserving embeddings and quantization codes. The MAP results on two datasets are reported in Table 2.

Enhancing Semantics on Correlation Graph. WSDHQ outperforms WSDHQ$_{\mathcal{G}}$ by 2.3% and 1.6% in the average MAP on two datasets, which indicates that the tag processing on $\mathcal{G}$ plays an important role in weakly-supervised deep quantization, because redundant and synonymous tags in a tag set may lead to semantic bias and degrade performance.

Learning and Quantizing on Hypersphere. We find that the WSDHQ with ℓ_2 normalization for norm variance removal boosts 2.9% and 1.4% on the average MAP from WSDHQ$_{\mathrm{N}}$ on two datasets. It shows that the transformation to hypersphere contributes to learn better image representations and reduce the quantization error.

Loss for Visual Embedding Learning. Another observation is that WSDHQ outperforms WSDHQ$_{\mathcal{L}}$ by 2.8% and 3.3% in average MAP on two datasets. WSDHQ$_{\mathcal{L}}$ adopts a mini-batch-wise hinge loss in previous state-of-art weakly-supervised (binary) hashing method WDHT. The result shows that this loss is not optimal for weakly-supervised quantization (more generally, weakly-supervised hashing), because tag embedding averaging for each image is merely a coarse-grained estimation of true semantics, which fails to model the true tag distribution in semantic space. Moreover, the hinge loss may mistakenly push two semantically similar image embeddings far away from each other, since the pairwise similarity is hard to well measure with weak semantics. By contrast, the adaptive cosine margin loss in WSDHQ is more granular. Instead of direct assertions of semantic inter-relationship for training images, it wisely focuses on the pointwise intra-relationships between images and tags in hyperspherical semantic space, thus better preserving semantic information into image embeddings.

Joint Process of Embedding Learning and Quantization. By jointly learning semantics-preserving image embeddings as well as controlling the quantization error, WSDHQ outperforms WSDHQ$_2$ by 3.2% and 3.5% in average MAP on two datasets. This verifies that end-to-end quantization can improve the quantization quality of deep image embeddings.



(a) MAP @ 32 bits *w.r.t.* the trade-off hyper-parameter of two losses

(b) MAP @ 32 bits *w.r.t.* the scaling hyper-parameter in adaptive margin

Figure 5: The MAP results of WSDHQ @ 32 bits *w.r.t.* γ and λ on two datasets. The values on dotted lines are the MAP results of WDHT (*i.e.*, the best baseline) @ 32 bits.

Parameter Sensitivity

We study the sensitivities of **1**) λ, the trade-off hyper-parameter between losses $\mathcal{L}_n$ and $\mathcal{Q}_n$, and **2**) γ, the scaling hyper-parameter of the adaptive margin in $\mathcal{L}_n$. The MAP results are shown in Fig. 5. WSDHQ surpasses the best baseline within a wide range of λ and γ. The hill-shape trend curves of MAP *w.r.t.* λ and γ confirm the importance of selecting appropriate hyper-parameters in WSDHQ. WSDHQ degenerates into its two-stage variant WSDHQ$_2$ that gets inferior results when $\lambda \to 0$, which justifies the effectiveness of joint learning of deep visual embeddings and deep supervised quantization. A moderately smaller γ that leads to relatively larger margins helps the WSDHQ to learn more discriminative quantization codes, while we should avoid degeneration or even non-convergence caused by too small γ with abnormally large margins.

Conclusions

In this paper we propose the Weakly-Supervised Deep Hyperspherical Quantization (WSDHQ) for efficient image retrieval. Different from current deep quantization methods, WSDHQ enables learning quantization codebook from weakly tagged web images without using ground-truth labels. The weak supervision is enhanced via semantic enhancement for tags and sparse tag reduction based on the investigated tag correlation. Our developed joint learning of deep visual embeddings and semantics-preserving quantizer on hypersphere also yield a significant performance boost to WSDHQ on large-scale image retrieval. Extensive experiments demonstrate state-of-art retrieval performance on two well-known and widely-tested datasets. Our work encourages the exploration of weakly-supervised deep quantization by leveraging web and social media data, which promotes quantization to adapt real-world scenario. Future work includes improving weakly supervised deep quantization by detecting and completing the probably missing semantic information in the given tag sets during training.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China under Grant 2018YFB1800204 and the National Natural Science Foundation of China under Grant 61771273 and Grant 61906219.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI*, 265–283.

Babenko, A.; and Lempitsky, V. 2014. Additive quantization for extreme vector compression. In *CVPR*, 931–938.

Besag, J. 1986. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)* 48(3): 259–279.

Cao, Y.; Long, M.; Wang, J.; and Liu, S. 2017. Deep visual-semantic quantization for efficient image retrieval. In *CVPR*, 1328–1337.

Cao, Y.; Long, M.; Wang, J.; Zhu, H.; and Wen, Q. 2016. Deep quantization network for efficient image retrieval. In *AAAI*.

Chen, J.; and Cheung, W. K. 2019. Similarity Preserving Deep Asymmetric Quantization for Image Retrieval. In *AAAI*, volume 33, 8183–8190.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y.-T. 2009. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*.

Cui, H.; Zhu, L.; Cui, C.; Nie, X.; and Zhang, H. 2020. Efficient weakly-supervised discrete hashing for large-scale social image retrieval. *Pattern Recognition Letters* 130: 174–181.

Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K.; and Li, F. F. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Eghbali, S.; and Tahvildari, L. 2019. Deep spherical quantization for image search. In *CVPR*, 11690–11699.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, 226–231.

Gattupalli, V.; Zhuo, Y.; and Li, B. 2019. Weakly Supervised Deep Image Hashing Through Tag Embeddings. In *CVPR*, 10375–10384.

Ge, T.; He, K.; Ke, Q.; and Sun, J. 2013. Optimized product quantization. *TPAMI* 36(4): 744–755.

Gionis, A.; Indyk, P.; Motwani, R.; et al. 1999. Similarity search in high dimensions via hashing. In *VLDB*, volume 99, 518–529.

Gomez, R.; Gomez, L.; Gibert, J.; and Karatzas, D. 2018. Learning to learn from web data through deep semantic embeddings. In *ECCV Workshops*, 514–529.

Gong, Y.; Lazebnik, S.; Gordo, A.; and Perronnin, F. 2013. Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-Scale Image Retrieval. *TPAMI* 35(12): 2916–2929.

Guan, Z.; Xie, F.; Zhao, W.; Wang, X.; Chen, L.; Zhao, W.; and Peng, J. 2018. Tag-based Weakly-supervised Hashing for Image Retrieval. In *IJCAI*, 3776–3782.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Huang, L.-K.; Chen, J.; and Pan, S. J. 2019. Accelerate Learning of Deep Hashing With Gradient Attention. In *ICCV*.

Huiskes, M. J.; and Lew, M. S. 2008. The MIR Flickr Retrieval Evaluation. In *Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*.

Jegou, H.; Douze, M.; and Schmid, C. 2010. Product quantization for nearest neighbor search. *TPAMI* 33(1): 117–128.

Jin, S.; Zhou, S.; Liu, Y.; Chen, C.; Sun, X.; Yao, H.; and Hua, X.-S. 2020. SSAH: Semi-Supervised Adversarial Deep Hashing with Self-Paced Hard Sample Generation. In *AAAI*, 11157–11164.

Kalantidis, Y.; and Avrithis, Y. 2014. Locally optimized product quantization for approximate nearest neighbor search. In *CVPR*, 2321–2328.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Lee, Y. 2012. Spherical Hashing. In *CVPR*, 2957–2964.

Lin, K.; Lu, J.; Chen, C.-S.; and Zhou, J. 2016. Learning Compact Binary Descriptors With Unsupervised Deep Neural Networks. In *CVPR*, 1183–1192.

Liu, B.; Cao, Y.; Long, M.; Wang, J.; and Wang, J. 2018. Deep triplet quantization. In *ACM MM*, 755–763.

Martinez, J.; Clement, J.; Hoos, H. H.; and Little, J. J. 2016. Revisiting additive quantization. In *ECCV*, 137–153.

Martinez, J.; Zakhmi, S.; Hoos, H. H.; and Little, J. J. 2018. LSQ++: Lower running time and higher recall in multi-codebook quantization. In *ECCV*, 491–506.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. In *ICLR Workshops*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 8024–8035.

Salakhutdinov, R.; and Hinton, G. 2009. Semantic hashing. *International Journal of Approximate Reasoning* 50(7): 969–978.

Tang, J.; and Li, Z. 2018. Weakly Supervised Multimodal Hashing for Scalable Social Image Retrieval. *TCSVT* 28(10): 2730–2741.

Wang, J.; Zhang, T.; j. song; Sebe, N.; and Shen, H. T. 2018. A Survey on Learning to Hash. *TPAMI* 40(4): 769–790.

Weiss, Y.; Torralba, A.; and Fergus, R. 2009. Spectral hashing. In *NIPS*, 1753–1760.

Wu, X.; Guo, R.; Suresh, A. T.; Kumar, S.; Holtmann-Rice, D. N.; Simcha, D.; and Yu, F. 2017. Multiscale quantization for fast similarity search. In *NIPS*, 5745–5755.

Yang, J.; Chen, B.; and Xia, S.-T. 2020. Mean-Removed Product Quantization for Large-scale Image Retrieval. *Neurocomputing* .

Yu, T.; Yuan, J.; Fang, C.; and Jin, H. 2018. Product quantization network for fast image retrieval. In *ECCV*, 186–201.

Yuan, L.; Wang, T.; Zhang, X.; Tay, F. E.; Jie, Z.; Liu, W.; and Feng, J. 2020. Central Similarity Quantization for Efficient Image and Video Retrieval. In *CVPR*, 3083–3092.

Zeger, K.; Vaisey, J.; Gersho, A.; et al. 1992. Globally optimal vector quantizer design by stochastic relaxation. *TSP* 40(2): 310–322.

Zhang, H.; Zhao, N.; Shang, X.; Luan, H.; and Chua, T.-s. 2016. Discrete image hashing using large weakly annotated photo collections. In *AAAI*.

Zhang, T.; Du, C.; and Wang, J. 2014. Composite Quantization for Approximate Nearest Neighbor Search. In *ICML*, volume 2, 3.

Zhao, F.; Huang, Y.; Wang, L.; and Tan, T. 2015. Deep semantic ranking based hashing for multi-label image retrieval. In *CVPR*, 1556–1564.