Marginalized Graph Attention Hashing for Zero-Shot Image Retrieval

Meixue Huang[12]
huangmeixue@iie.ac.cn

Dayan Wu*[1]
wudayan@iie.ac.cn

Wanqian Zhang[12]
zhangwanqian@iie.ac.cn

Zhi Xiong[12]
xiongzhi@iie.ac.cn

Bo Li[1]
libo@iie.ac.cn

Weiping Wang[1]
wangweiping@iie.ac.cn

[1] Institute of Information Engineering
Chinese Academy of Sciences
Beijing, China

[2] School of Cyber Security
University of Chinese Academy of Sciences
Beijing, China

Abstract

Zero-shot image retrieval allows to precisely retrieve candidates relevant to unobserved queries, of which categories have never been seen during training. Recently, research interests arise in exploring hashing methods to solve this problem due to its storage and computational efficiency. However, existing methods only focus on leveraging semantic information, but omit to exploit the similarity structure of visual feature space for knowledge transfer. Besides, the domain shift problem across seen and unseen classes further degrades the performance. To tackle these issues, in this paper, we propose a novel deep zero-shot hashing method, named **M**arginalized **G**raph **A**ttention **H**ashing (MGAH). MGAH introduces the masked attention mechanism to construct a joint-semantics similarity graph, which captures the intrinsic relationship from different metric spaces, making it competent to transfer knowledge from seen classes into unseen classes. Furthermore, we elaborately design an Energy Magnified Softmax (EM-Softmax) loss, which is capable to alleviate the domain shift problem and encourage the generalization ability of hash codes. By using marginalized strategy, EM-Softmax produces the shared decision margin for hard samples, thus can avoid overfitting on seen classes and meanwhile cover more knowledge for the unseen ones. Extensive experiments demonstrate that MGAH delivers superior performance over the state-of-the-art zero-shot hashing methods.

1 Introduction

Zero-shot image retrieval (ZSIR) is a realistic and challenging problem in computer vision [3, 16], which assumes that the categories of queries are unseen during training. Given an

* The corresponding author

unobserved image, ZSIR attempts to find all relevant candidates of the same category as this query. Recent decade has witnessed the rapid development of deep feature learning [2, 31] for ZSIR. However, these real-valued image features in high dimension lead to huge storage and computational costs, which is impractical for large-scale retrieval systems [27, 28, 29, 32, 34, 37].

Recently, zero-shot hashing [4, 5, 10, 15] is widely utilized to alleviate the problem, in which high-dimensional image features are encoded into compact binary codes, thereby improving both storage and computational efficiency. The crucial importance of zero-shot hashing is to learn effective and compatible hash functions from seen categories so as to facilitate the retrieval performance on unseen concepts. To achieve this goal, numerous approaches [9, 33, 35, 36] have been proposed.

Although these methods have achieved promising performance, there still exist two challenging issues worthy of attention. First of all, semantic embeddings of seen classes are insufficient to learn such effective and compatible hash function, since they are generally learned from noisy text and not enough to span the visual feature space. Nonetheless, recent works [9, 35] merely focus on associating categories through their class semantics, which imposes restrictions on deducing helpful knowledge from seen concepts to the unseen ones. Secondly, with no training data of unseen classes, the different visual distribution on seen and unseen categories leads to the domain shift problem. This results in the learned hash codes being overfitted to the limited data in seen classes and thus less capable of effectively distinguishing samples of unseen categories. However, existing methods [33, 36] neglect this generic challenge, leading to generalize poorly to unseen classes.

To solve the aforementioned problems, we propose a novel zero-shot hashing framework, named **M**arginalized **G**raph **A**ttention **H**ashing (MGAH). The overall framework is illustrated in Figure 1. We utilize the masked attention mechanism to construct a joint-semantics similarity graph, which captures semantic relations from both semantic embeddings and visual features, making it more competent for knowledge transfer. Furthermore, in order to better generalize the learned hash function to unseen concepts, we elaborately design an Energy Magnified Softmax (EM-Softmax) loss that adopts marginalized strategy to optimize our network. EM-Softmax is a flexible learning objective that generates the shared decision margin for hard samples, which avoids overfitting on seen classes as well as covers more potential metric space for unseen categories. Hence, even though our model never meets unseen concepts, it still improves the retrieval performance for them.

Our contributions can be summarized as follows: **1)** To the best of our knowledge, MGAH is the first zero-shot hashing method to construct a joint-semantics similarity graph for knowledge transfer, which integrates the intrinsic relationship from different metric spaces and accordingly is capable to generate more semantic relevant hash codes. **2)** To alleviate the domain shift, we elaborately design an Energy Magnified Softmax loss, which utilizes marginalized strategy to optimize hash learning, so that the learned hash function generalizes well to unseen classes. **3)** Extensive experiments on three widely used zero-shot image retrieval datasets demonstrate that MGAH outperforms the state-of-the-arts.

2 Related Work

We briefly introduce zero-shot hashing in both inductive setting and transductive setting. In this paper, we investigate inductive zero-shot hashing, i.e, the most general setting in the realistic scenario.

Inductive zero-shot hashing assumes that samples of unseen classes are inaccessible in the training phase. Representative inductive zero-shot hashing methods include [9, 33, 35, 36]. TSK [35] projects image features into word embedding space to generate effective hash codes. Instead of using word vectors, AH [33] utilizes semantic attributes as an intermediate layer between hash codes and image labels. In [36], an orthogonal projection constraint is adopted to enhance the discriminative power. SitNet [9] introduces a multi-task architecture which simultaneously employs a max-margin loss and a regularized center loss, thus preserving the semantic similarity among concepts for knowledge transfer.

Transductive zero-shot hashing refers to the setting that unseen class instances are available during training. [13, 19, 39] are remarkable transductive zero-shot hashing methods. In [13], a coarse-to-fine similarity mining method is proposed to find most presentative target examples of each unseen class. Besides, some methods [19, 39] also consider multi-label scenarios. The co-occurrence information is utilized in COSTA [19] to associate seen and unseen categories. Moreover, ICRH [39] designs an instance-concept coherence ranking algorithm to consider ranking relationship between relevant and irrelevant labels, which preserves multi-level semantic similarity for multi-label images.

However, the aforementioned methods only focus on leveraging semantic information, while the intrinsic structure in visual feature space is not fully exploited. Furthermore, the different data distribution on seen and unseen classes is ignored by these methods, which causes the domain shift problem and impedes the generalization ability of hash learning.

3 Proposed Approach

3.1 Problem Definition

Suppose we have N training images $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ from seen concept set C^s, where $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th sample, and d is the dimensionality of visual feature space. $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N}$ denotes the binary label, $\mathbf{Y}_{ij} = 1$ if the i-th training sample belongs to the j-th category, otherwise, $\mathbf{Y}_{ij} = 0$. Meanwhile, $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{N}$ is corresponding class semantics, which is obtained from either semantic attributes or distributed word representations. There are a total of M categories in the training set. In contrast with conventional hashing, our zero-shot scenario assumes that testing data belongs to an unseen concept set C^u, which shares no common concepts with C^s, i.e., $C^s \cap C^u = \phi$. By using training data $\mathbf{X}$ only from C^s, our goal is to learn a hash function: $\mathbf{X} \mapsto \mathbf{B} \in \{-1,1\}^{N \times K}$ that can accurately encode images of both seen and unseen classes to K-bits hash codes.

3.2 Binary Encoder

The binary encoder serves as hash function to generate desired hash codes. As is shown in Figure 1, we firstly adopt the pre-trained AlexNet [12] as feature extractor to obtain image representation $\mathbf{x}$, and then the hash layer with an element-wise $\text{sign}(\cdot)$ function is employed to produce K-bits binary code $\mathbf{b}$, which can be denoted as follows:

$$\mathbf{b} = \text{sign}(f(\mathbf{x}; \theta_h)) \in \{-1,1\}^K, \tag{1}$$

where $f(\cdot)$ is hash encoding function, and θ_h is the network parameters. However, the discrete constraint induces difficulty to optimization. Suggested by [1, 24], we adopt the scaled

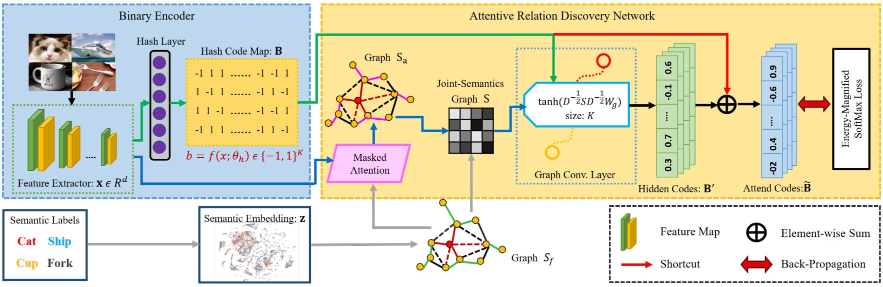


Figure 1: Overview of the proposed framework. See section 3 for a detailed description.

tanh$(\cdot)$ function to tackle this problem and rewirte Eq. (1) as follows:

$$\mathbf{b} = \tanh\left(\delta * f(\mathbf{x};\theta_h)\right) \in [-1,1]^K, \tag{2}$$

where $\delta > 0$ is the scale parameter, which gradually increases in the training phase so as to converge to the original binary encoding problem.

3.3 Attentive Relation Discovery Network

Constructing joint-semantics similarity graph. As an initial step, we establish a fixed similarity graph pre-computed on class semantics, which preserves the original neighborhood relations from the perspective of semantic embedding space. Given a pair of class semantic $\mathbf{z}_i$ and $\mathbf{z}_j$, their relation is measured by cosine similarity:

$$\mathbf{S}_f(i,j) = \frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|}. \tag{3}$$

In light of the topology structure determined by S_f, the masked attention mechanism is leveraged to produce an adaptive similarity graph, which is directly driven by image visual features. In other words, for image $\mathbf{x}_i$, we only compute its attention score of images $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the first-order neighborhood of image $\mathbf{x}_i$ in the above graph $\mathbf{S}_f$. Firstly, to obtain detailed information, we utilize an identical fully-connected layer to learn the latent variable $\mathbf{h}$ as follows:

$$\mathbf{h} = g(\mathbf{x};\theta_s) \in \mathbb{R}^L, \tag{4}$$

where L refers to the dimensionality of $\mathbf{h}$ and θ_s indicates the network parameters. Suggested by [26], we then calculate the attention coefficient and normalize them to obtain the adaptive similarity graph $\mathbf{S}_a$, which can be presented by:

$$\mathbf{S}_a(i,j) = \text{sigmoid}\Big(\text{ReLU}(\mathbf{W}_a^\top[\mathbf{h}_i \| \mathbf{h}_j])\Big), \tag{5}$$

where $\|$ is concatenation operation and $\mathbf{W}_a$ is the network parameters. As observed in Eq. (5), the attention score $\mathbf{S}_a(i,j)$ indicates the importance of image $\mathbf{x}_j$ to image $\mathbf{x}_i$. Finally, we combine graph $\mathbf{S}_f$ and $\mathbf{S}_a$ to obtain the joint-semantics similarity graph $\mathbf{S}$ as follows:

$$\mathbf{S} = \alpha \mathbf{S}_f + (1-\alpha)\mathbf{S}_a, \tag{6}$$

where α is a trade-off parameter that balances the importance of the neighborhood relations from different metric spaces. Since information from different perspectives are generally complementary to each other, this graph $\mathbf{S}$ can provide more precise semantic relations than previous methods that only consider class semantics.

Integrating semantic relations into binary codes. To generate semantic relevant codes, the joint-semantics similarity graph $\mathbf{S}$ and original codes $\mathbf{B}$ are render into a graph convolutional layer [11], which is defined as:

$$\mathbf{B}' = \tanh\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}\mathbf{B}\mathbf{W}_g\right), \tag{7}$$

here, $\mathbf{W}_g$ refers to the linear transformation parameter, and $\mathbf{D} = \mathrm{diag}(\mathbf{S}\mathbb{1})$. We choose $\tanh(\cdot)$ function to restrict the output codes between -1 and 1. Furthermore, in order to better fulfill our ultimate goal of binary coding, we introduce a shortcut connection architecture to link the original binary codes $\mathbf{B}$ directly to the hidden codes $\mathbf{B}'$ and combine them with an element-wise sum, which is demonstrated as follows:

$$\widetilde{\mathbf{B}} = \lambda\mathbf{B} + (1-\lambda)\mathbf{B}', \tag{8}$$

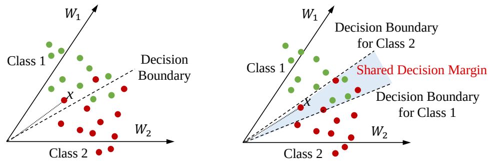
where λ is the trade-off factor to control the effect of original codes.

3.4 Energy Magnified Softmax Loss

In our zero-shot scenario, it's essential to learn an effective and compatible hash function from seen classes so as to boost the performance on unseen classes. However, the different distribution between seen and unseen categories seriously degrades this performance. From the viewpoint of domain adaption [7, 18], we elaborately design an Energy Magnified Softmax (EM-Softmax) loss with marginalized strategy to guide the learning of hash codes, which is formulated as follows:

$$\mathcal{L}_{em} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp\left(\beta * \mathbf{W}_j^\top \widetilde{\mathbf{b}}_i\right)}{\exp\left(\beta * \mathbf{W}_j^\top \widetilde{\mathbf{b}}_i\right) + \sum_{k\neq j}\exp\left(\mathbf{W}_k^\top \widetilde{\mathbf{b}}_i\right)}, \tag{9}$$

where $\mathbf{W}_j$ denotes the weight of j-th category, $*$ refers to the element-wise multiply operation, $\beta \geq 1$ is the energy factor that magnifies the prediction and encourages a flexible margin. Note that, the value of β gradually increases during training to encourage both discrimination and generalization ability.



Traditional Softmax loss EM-Softmax loss

Figure 2: Geometric interpretation of Energy-Magnified Softmax loss.

As observed in Eq. (9), EM-Softmax loss produces a relaxation classification criteria, which is loose to hard samples and accordingly avoid overfitting on seen classes. Intuitively, considering a sample $\widetilde{\mathbf{b}}_i$ belonging to j-th category, traditional softmax loss forces

$\mathbf{W}_j^\top \widetilde{\mathbf{b}}_i > \mathbf{W}_k^\top \widetilde{\mathbf{b}}_i \, (\forall k \neq j)$ to classify x_i correctly. In contrast, EM-Softmax loss requires $\beta * \mathbf{W}_j^\top \widetilde{\mathbf{b}}_i > \mathbf{W}_k^\top \widetilde{\mathbf{b}}_i \, (\forall k \neq j)$. Hence, our EM-Softmax loss can not only guarantee the separability between classes, but also improve the generalization ability.

Furthermore, EM-Softmax loss can produce a shared decision margin for hard samples, which is supposed to cover more potential metric space for unseen classes. The geometric interpretation is demonstrated in Figure 2. in which a binary classification case is analyzed. In EM-Softmax loss, the decision boundary for class 1 can be computed as: $(\beta \mathbf{W}_1 - \mathbf{W}_2) \widetilde{\mathbf{b}}_i = 0$ and the decision boundary for class 2 is $(\beta \mathbf{W}_2 - \mathbf{W}_1) \widetilde{\mathbf{b}}_i = 0$. Consequently, the EM-Softmax loss makes the decision boundaries for class 1 and class 2 different, while the decision boundaries are the same in traditional softmax loss. In essence, the EM-Softmax loss broadens the feasible region for each class and expects to encompass potential space for the unseen ones.

3.5 Out-of-Sample Extension

After MGAH is trained, the networks can be used to generate K-bits hash codes for any out-of-sample data. As discussed in section 3.2, the query image $\mathbf{x}^{(q)}$ is fed into the binary encoder to obtain its hash code:

$$\mathbf{b}^{(q)} = \mathrm{sign}(f(\mathbf{x}^{(q)}; \theta_h)) \in \{-1, 1\}^K . \tag{10}$$

When testing, only $f(\cdot)$ is required, which considerably eases the binary coding process.

4 Experiments

4.1 Datasets, Settings, Metrics, and Implementation Details

Datasets. Animals with Attributes2 [30] is a widely used zero-shot learning dataset which has $37,322$ images from 50 different animals. Besides, 85 human-annotated class attributes are provided. CIFAR-10 [12] consists of $60,000$ images which are manually labelled with 10 classes, with $6,000$ samples in each class. It is frequently utilized for evaluating the hashing approaches. ImageNet [6] is a large-scale image dataset organized according to WordNet hierarchy. The subset of ImageNet for ILSVRC2012 is used for our experiments, which contains over 1.2 million images manually labeled by $1,000$ concepts.

Settings. For Animals with Attributes2, we split 40 categories into seen concepts and the other 10 categories into unseen concepts following [30]. For CIFAR-10, we use one category as unseen concept and the other nine as seen categories, which leads to 10 seen-unseen splits. For ImageNet, we randomly select 90 categories as seen concepts as well as 10 categories as unseen concepts, giving us more than 130,000 images in total. For all datasets, we randomly sample $10,000$ images from seen concepts as the training set. When testing, $1,000$ images are randomly selected from the unseen concepts as the queries, while the remaining unseen category images and all seen category images form the retrieval database. Besides, since the initial model AlexNet is pretrained on ImageNet [6], it contain knowledge about the 100 categories (both seen and unseen) utilized in our experiment. To better evaluate the zero-shot performance, we retrain the model on the other 900 categories as initial model for experiments on ImageNet dataset.

Evaluation metrics. We adopted two widely used evaluation criteria for zero-shot hashing, i.e., mean Average Precision (mAP) and Precision within Hamming distance 2 (P@H ≤ 2). For above metrics, a larger value indicates better retrieval performance.

Implementation details. We implemented our method by PyTorch on NVIDIA RTX 2080Ti. To train our model, Adam optimizer is applied with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay is 0.0005. The initial learning rate is 0.0001 and exponentially decayed to $1e-6$ during training. The mini-batch size is 64. For hyper-parameters, we set $L = 512$, $\alpha = 0.3$, and $\lambda = 0.9$. To provide the similarity graph $\mathbf{S}$, we prepare class semantics $\mathbf{Z}$ as follows. For Animals with Attributes2 dataset, 85 annotated attributes for each category are scale-normalized as semantic vector of the concept. For CIFAR-10 dataset, the *word2vec* tool [20] is adopted which takes the names of seen concepts as inputs and gives us 300 dimensional word representations as class semantics. For ImageNet dataset, we map each category to a node in *WordNet* and use the *path-similarity* interface to directly set $\mathbf{S}_f(i, j)$.

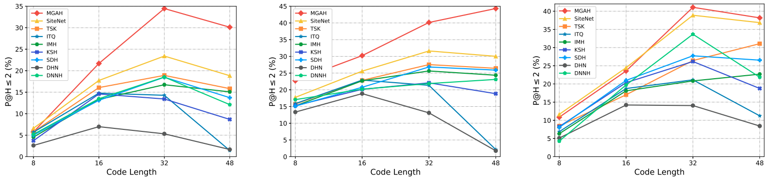
4.2 Comparison with Existing Methods

Baselines. We compare our MGAH with both shallow methods and deep methods. Shallow methods include ITQ [8], IMH [22], KSH [17], SDH [21], TSK [35], while deep methods include DHN [38], DNNH [14] and SitNet [9]. For a fair comparison, we utilize AlexNet [12] to extract deep features for all shallow baselines.

Result on mAP. We compare our MGAH with all baseline methods, the mAP results are presented in Table 1, from which we can get the following observations. (1) MGAH achieves inspiring performance compared to the previous hashing methods. (2) Compared to the conventional setting in [14, 21], the performance of several supervised approaches drops significantly in the zero-shot scenario. The main reason is that these methods overfit seen concepts, leading to a poor generalization on unseen concepts. (3) The unsupervised approaches like IMH [22] gain competitive performance compared to some supervised methods. This is because unsupervised methods encode images with the distributional properties that somewhat preserves data structures.

Method	Animals with Attributes2				CIFAR-10				ImageNet			
	8 bits	16 bits	32 bits	48 bits	8 bits	16 bits	32 bits	48 bits	8 bits	16 bits	32 bits	48 bits
ITQ	0.0539	0.0936	0.1176	0.1392	0.1382	0.1613	0.2072	0.2321	0.1294	0.1801	0.2413	0.2648
IMH	0.0676	0.0887	0.1217	0.1417	0.1302	0.1611	0.1891	0.2002	0.1109	0.1694	0.2196	0.2597
KSH	0.0631	0.0846	0.1169	0.1379	0.1189	0.1519	0.1839	0.2104	0.1502	0.2018	0.2388	0.2761
SDH	0.0543	0.0969	0.1281	0.1471	0.1323	0.1678	0.2009	0.2092	0.1688	0.1992	0.2469	0.2936
DHN	0.0317	0.0572	0.0734	0.0986	0.1556	0.1722	0.1874	0.2034	0.1297	0.1984	0.1979	0.2037
DNNH	0.0841	0.1112	0.1356	0.1587	0.1774	0.2012	0.2205	0.2326	0.1589	0.2268	0.2767	0.2964
TSK	0.0827	0.1193	0.1403	0.1528	0.1491	0.1881	0.2203	0.2401	0.1803	0.2231	0.2571	0.3096
SitNet	0.1138	0.1465	0.1673	0.1869	0.2198	0.2308	0.2574	0.2669	**0.1905**	**0.268**4	0.3192	0.3546
MGAH	**0.1169**	**0.1780**	**0.1943**	**0.2451**	**0.2490**	**0.2757**	**0.3087**	**0.3307**	0.1744	0.2651	**0.3230**	**0.3697**

Table 1: The mAP comparison results for different number of bits on three datasets.



(a) Animals with Attributes2 (b) CIFAR-10 (c) ImageNet

Figure 3: The P@H $\leq$ 2 comparison results for different number of bits on three datasets.

Result on precision curves. For further comparisons, we evaluate all the hashing methods with P@H ≤ 2, which are shown in Figure 3. We can observe that: (1) MGAH outperforms the baseline approaches with significant margins in most cases. It's worth noting that the results of 16-bits code length in MGAH are superior to those of 32-bits in other methods, demostrating its inspiring performance. (2) As code length varies from 8 to 32, the P@H ≤ 2 increases rapidly for all hashing methods. This is because when code length is short, more codes are required to guarantee the descriptive power. (3) When code length is larger than 32, the performance significantly deteriorates. As Hamming space becomes larger, fewer data points will fall within Hamming distance 2, thereby declining the P@H ≤ 2 performance.

Comparison with DOPH. DOPH [36] is one of the state-of-the-art zero-shot hashing method, which adopts orthogonal projection constraint to project the information from different modalities into a common binary hash space. To fairly compare with DOPH, we follow the experiment settings used in DOPH. Table 2 shows the experiment results. In particular, due to the pre-trained GoogLeNet [25] is adopted in DOPH, we also employ the same architecture as backbone, which denotes as "MGAH†". It can be seen that our method outperforms DOPH in most cases. The reason is that DOPH relies on the orthogonal projection constraint to generate hash codes, while our MGAH explicitly exploits semantic similarity to facilitate the transfer ability, thus can generalize well to unseen concepts.

Metric	Method	Animals with Attributes2				CIFAR-10				ImageNet			
		8 bits	16 bits	32 bits	48 bits	8 bits	16 bits	32 bits	48 bits	8 bits	16 bits	32 bits	48 bits
mAP	DOPH	0.1578	0.1822	0.2407	0.2884	**0.3402**	0.3351	0.3491	0.3438	**0.1579**	**0.2748**	0.3413	0.4288
	MGAH†	**0.1611**	**0.2467**	**0.3036**	**0.3582**	0.2716	**0.3378**	**0.3862**	**0.4128**	0.1459	0.2660	**0.3526**	**0.4476**
P@H ≤ 2	DOPH	0.0671	0.2152	0.2743	**0.4318**	**0.3227**	0.3781	0.3844	0.3609	0.0669	**0.2204**	0.4186	**0.6327**
	MGAH†	**0.0738**	**0.2747**	**0.4734**	0.4168	0.2515	**0.3813**	**0.4830**	**0.5404**	**0.0736**	0.2196	**0.5028**	0.4598

Table 2: The mAP and P@H ≤ 2 results in comparison to DOPH.

4.3 Ablation Study

Effect of different components. We further investigate the impact of different components on the performance of MGAH. We reconstruct the network architecture as follows: (1) MGAH-1: we remove Energy Magnified Softmax loss from MGAH and train the network with traditional softmax loss. (2) MGAH-2: we remove the fixed similarity graph $\mathbf{S}_f$ from MGAH. (3) MGAH-3: we remove the adaptive similarity graph $\mathbf{S}_a$ from MGAH. (4) MGAH-4: we perform weighted average on binary codes $\mathbf{B}$ and similarity graph $\mathbf{S}$, instead of employing the graph convolutional layer. (5) MGAH-5: we remove the shortcut connection architecture from MGAH. The mAP results are presented in Table 3. Compare results on MGAH-1 and MGAH, we can find out that EM-Softmax loss achieves an average of 3.34% improvement. The results of MGAH-2 and MGAH-3 demonstrates the improvement brought by the joint-semantics similarity graph $\mathbf{S}$. The performance of MGAH-4 shows that the graph convolutional layer is more effective to incorporate semantic relations into hash codes than weighted average. Moreover, the shortcut connection can further improve performance, which can be observed from results on MGAH-4.

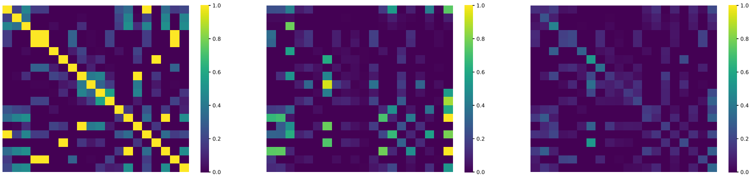
Effect of the EM-Softmax loss. To evaluate the impact of different margins in EM-Softmax loss, the mAP results are presented in Table 3. When β increases, the performance first upgrades and then degrades with $\beta > 4$. This is because the shared margin is too large to impede the discriminative power for distinguishing unseen categories. Besides, MGAH

Method	Animals with Attributes2			
	8 bits	16 bits	32 bits	48 bits
MGAH-1	0.1090	0.1343	0.1732	0.1840
MGAH-2	0.0870	0.1657	0.1885	0.2184
MGAH-3	0.1137	0.1618	0.1817	0.2269
MGAH-4	0.0837	0.1457	0.1793	0.2180
MGAH-5	0.0766	0.1374	0.1882	0.2161
MGAH	**0.1169**	**0.1780**	**0.1943**	**0.2451**

Margin	Animals with Attributes2			
	8 bits	16 bits	32 bits	48 bits
$\beta = 1$	0.1090	0.1343	0.1732	0.1840
$\beta = 2$	0.0933	0.1315	0.1737	0.2124
$\beta = 3$	0.1040	0.1438	0.1809	0.2127
$\beta = 4$	0.0971	0.1585	0.1857	0.2150
$\beta = 5$	0.0984	0.1646	0.1811	0.2060
MGAH	**0.1169**	**0.1780**	**0.1943**	**0.2451**

Table 3: Ablation Study on Animals with Attributes2. Contributions of different components (left) and different margins of the EM-Softmax loss (right).

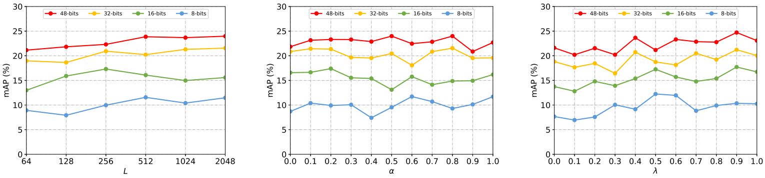
achieves inspiring performance compared to other variants, because the gradual growth of β improves both discrimination and generalization ability of our model.



(a) $\mathbf{S}_f$ (b) $\mathbf{S}_a$ (c) $\mathbf{S}$

Figure 4: Similarity graphs on Animals with Attributes2 with 32-bits. (a) fixed graph $\mathbf{S}_f$, (b) adaptive graph $\mathbf{S}_a$, and (c) joint-semantics graph $\mathbf{S}$.

Effect of the joint-semantics similarity graph. We show the effectiveness of the joint-semantics similarity graph in Figure 4. 20 random samples are selected from a training batch to plot the similarity graphs. In contrast to graph $\mathbf{S}_f$, graph $\mathbf{S}$ is asymmetric, which allows for assigning different importance to the same image pair. Moreover, compared with graph $\mathbf{S}_a$, graph $\mathbf{S}$ is much smoother that may enable a leap in model generalization ability.

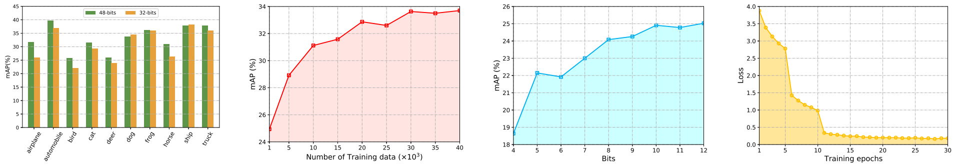


(a) L (b) α (c) λ

Figure 5: Parameter sensitivity analysis of L, α and λ on Animals with Attributes2.

4.4 More Analysis

Sensitivity study. We further analyze the effect of hyper-parameter L, α and λ. Figure 5 shows the mAP results under different values of these parameters. When L is small, the retrieval performance is poor since low-dimensional features cannot fully describe the details of visual images. When α varies from 0 to 0.3, the performance remains in a relatively stable range, while fluctuates slightly with $\alpha > 0.4$. When λ increases, the performance first upgrades and degrades with $\lambda > 0.4$.



(a) (b) (c) (d)

Figure 6: Further analysis on MGAH. (a) different unseen categories, (b) training numbers, (c) extremely short codes, and (d) convergency analysis.

Impact of different unseen categories. We list the corresponding mAP of different unseen categories on CIFAR-10 with 32-bits in Figure 6(a). Intuitively, if an unseen class is semantically closer to other seen categories, more relevant semantic knowledge can be transferred from our joint-semantics similarity graph, thus boosting the retrieval performance.

Impact of training numbers. We further investigate the impact of training numbers, which is presented in Figure 6(b). As we can see, when the size increases from 1000 to 10,000, the mAP rapidly increases, which implies that as training samples grow, our model is able to learn more visual knowledge.

Effect of extremely short codes. Inspired by [23], we illustrate the retrieval performance with extremely short code length in Figure 6(c). We can find out that, MGAH works well even when the code length is set to $K = 5$. The reason is that the joint-semantic similarity graph complements the information discarded by the binary codes.

Convergency analysis. We conduct empirical study on the convergence property using ImageNet dataset. In Figure 6(d), we plot the loss in Eq. (9) w.r.t. the number of training epochs. We can observe that, as margin β increases, the loss descends dramatically, which demonstrates the efficiency of our algorithm.

5 Conclusion

In this work, we propose a novel deep hashing method, named Marginalized Graph Attention Hashing (MGAH), for zero-shot image retrieval. The joint-semantics similarity graph well captures semantic relations from both semantic embedding and visual feature space, which is essential for recognizing unseen categories. Furthermore, the elaborately designed Energy Magnified Softmax loss employs marginalized strategy to generate a shared decision margin, which encourages the transfer ability of our hash function. Experiments on three widely used datasets compared with state-of-the-art methods show that MGAH achieves inspiring performance.

Acknowledgements

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDC02050200) and the National Key Research and Development Program of China (No. 2018YFC0825102, No.2019YFC0850202).

References

[1] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu. Hashnet: Deep learning to hash by continuation. In *ICCV*, pages 5608–5617, 2017.

[2] Binghui Chen and Weihong Deng. Energy confused adversarial metric learning for zero-shot image retrieval and clustering. In *AAAI*, volume 33, pages 8134–8141, 2019.

[3] Binghui Chen and Weihong Deng. Hybrid-attention based decoupled metric learning for zero-shot image retrieval. In *CVPR*, pages 2750–2759, 2019.

[4] Yudong Chen, Zhihui Lai, Yujuan Ding, Kaiyi Lin, and Wai Keung Wong. Deep supervised hashing with anchor graph. In *ICCV*, pages 9796–9804, 2019.

[5] Qi Dai, Jianguo Li, Jingdong Wang, and Yu-Gang Jiang. Binary optimized hashing. In *ACM MM*, pages 1247–1256, 2016.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[7] Zhengming Ding and Hongfu Liu. Marginalized latent semantic encoder for zero-shot learning. In *CVPR*, pages 6191–6199, 2019.

[8] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *TPAMI*, 35(12):2916–2929, 2012.

[9] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Sitnet: Discrete similarity transfer network for zero-shot hashing. In *IJCAI*, pages 1767–1773, 2017.

[10] Qing-Yuan Jiang and Wu-Jun Li. Asymmetric deep supervised hashing. In *AAAI*, 2018.

[11] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[13] Hanjiang Lai. Transductive zero-shot hashing via coarse-to-fine similarity mining. In *ICMR*, pages 196–203, 2018.

[14] Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan. Simultaneous feature learning and hash coding with deep neural networks. In *CVPR*, pages 3270–3278, 2015.

[15] Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang. Feature learning based deep supervised hashing with pairwise labels. *IJCAI*, 2016.

[16] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*, pages 3662–3671, 2019.

[17] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. Supervised hashing with kernels. In *CVPR*, pages 2074–2081. IEEE, 2012.

[18] Fuchen Long, Ting Yao, Qi Dai, Xinmei Tian, Jiebo Luo, and Tao Mei. Deep domain adaptation hashing with adversarial learning. In *SIGIR*, pages 725–734, 2018.

[19] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, pages 2441–2448, 2014.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[21] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. Supervised discrete hashing. In *CVPR*, pages 37–45, 2015.

[22] Fumin Shen, Chunhua Shen, Qinfeng Shi, Anton Van den Hengel, Zhenmin Tang, and Heng Tao Shen. Hashing on nonlinear manifolds. *TIP*, 24(6):1839–1851, 2015.

[23] Shupeng Su, Chao Zhang, Kai Han, and Yonghong Tian. Greedy hash: Towards fast optimization for accurate hash coding in cnn. In *NIPS*, pages 798–807, 2018.

[24] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *ICCV*, pages 3027–3035, 2019.

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[26] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *ICLR*, 2018.

[27] Dayan Wu, Zheng Lin, Bo Li, Mingzhen Ye, and Weiping Wang. Deep supervised hashing for multi-label and large-scale image retrieval. In *ICMR*, pages 150–158, 2017.

[28] Dayan Wu, Jing Liu, Bo Li, and Weiping Wang. Deep index-compatible hashing for fast image retrieval. In *ICME*, pages 1–6. IEEE, 2018.

[29] Dayan Wu, Qi Dai, Jing Liu, Bo Li, and Weiping Wang. Deep incremental hashing network for efficient image retrieval. In *CVPR*, 2019.

[30] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 41 (9):2251–2265, 2018.

[31] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *CVPR*, pages 9384–9393, 2019.

[32] Zhi Xiong, Dayan Wu, Wen Gu, Haisu Zhang, Bo Li, and Weiping Wang. Deep discrete attention guided hashing for face image retrieval. In *ICMR*, pages 136–144, 2020.

[33] Yahui Xu, Yang Yang, Fumin Shen, Xing Xu, Yuxuan Zhou, and Heng Tao Shen. Attribute hashing for zero-shot image retrieval. In *ICME*, pages 133–138. IEEE, 2017.

[34] Dejie Yang, Dayan Wu, Wanqian Zhang, Haisu Zhang, Bo Li, and Weiping Wang. Deep semantic-alignment hashing for unsupervised cross-modal retrieval. In *ICMR*, pages 44–52, 2020.

[35] Yang Yang, Yadan Luo, Weilun Chen, Fumin Shen, Jie Shao, and Heng Tao Shen. Zero-shot hashing via transferring supervised knowledge. In *ACM MM*, pages 1286–1295, 2016.

[36] Haofeng Zhang, Yang Long, and Ling Shao. Zero-shot hashing with orthogonal projection for image retrieval. *Pattern Recognition Letters*, 117:201–209, 2019.

[37] Wanqian Zhang, Dayan Wu, Bo Li, Xiaoyan Gu, Weiping Wang, and Dan Meng. Fast and multilevel semantic-preserving discrete hashing. In *BMVC*, 2019.

[38] Han Zhu, Mingsheng Long, Jianmin Wang, and Yue Cao. Deep hashing network for efficient similarity retrieval. In *AAAI*, 2016.

[39] Qin Zou, Zheng Zhang, Ling Cao, Long Chen, and Song Wang. Transductive zero-shot hashing for multi-label image retrieval. *CoRR abs/1911.07192*, 2019.