EFFICIENT FINE-GRAINED VISUAL-TEXT SEARCH USING ADVERSARIALLY-LEARNED HASH CODES

Yongzhi Li[†], *Yadong Mu*[†*], *Nan Zhuang*[†], *Xianglong Liu*[‡]

[†]Peking University; {yongzhili,zhuangn53,myd}@pku.edu.cn
[‡] Beihang University; xlliu@nlsde.buaa.edu.cn

2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428271

ABSTRACT

Cross-modal hashing for efficient visual-text search has attracted much research enthusiasm in recent years. The main argument of this work is that existing hashing methods mainly exploit a multi-label matching paradigm, ignoring various fine-grained semantics (high-order relationships, object attributes, etc.) in the multi-modal data. This paper explores cross-modal hashing from two rarely-explored aspects: first, we propose an efficient two-step hashing scheme that quickly screens irrelevant samples with global feature and then generate fine-grained feature guided by high-order concepts to rerank the survived candidates. Secondly, the robustness of the cross-modal hashing model, particularly under subtle tampering of fine-grained queries, is formally investigated. We propose a rephrase and adversarial training strategy for obtaining better performance and robustness. Comprehensive experiments and ablation studies on two large public datasets (MS-COCO and Flickr30K) demonstrate the proposed method's superiority in terms of both efficiency and accuracy.

Index Terms— Cross-modal Retrieval, Hashing, Adversarial Learning, Fine-grained Search

1. INTRODUCTION

With the popularity of Internet and social media, different kinds of large-scale multimedia data have become omnipresent on the Internet. Meanwhile, to mine the potential value in data and improve search engine's efficiency, cross-modal retrieval (CMR) has become a compelling topic in recent years. CMR aims to search semantically similar instances in one modality using a query from another modality (*e.g.*, image to text). Since different modality instances come from heterogeneous data sources with different distributions, it has posed new challenges to efficiently and effectively unify different modalities and bridge their semantic gaps.

This work mainly focuses on efficient cross-modal retrieval of text and images with complex and fine-grained semantics. A common practice is to bridge the modality gap via representation learning. Specifically, the goal is to learn optimal projections for different modalities into a common, modality-agnostic embedding subspace where the similarity between two samples can be directly calculated via simple metrics (*e.g.*, cosine similarity or Euclidean distance) [1]. Profiting from recent advance of deep learning, modern cross-modal retrieval methods have been characterized by the use of CNN for image and RNN for text [2, 3, 4]. Attention mechanism was also widely tailored into this task in some recent works [5, 6]. However, these methods are often heavy-weight in computations, and the attention must be executed in the runtime, which makes data indexing infeasible and brings slow response in a large-scale retrieval system.

Fig. 1. Illustration of the drawbacks of existing cross-modal hashing methods that essentially follow a multi-label matching paradigm. They tend to over-rate images that contain key objects mentioned in the queries (see the column of *False Result*), yet failing to capture fine-grained attributes or high-order relationships.

To solve that, a large number of hashing methods [7, 8, 9] have been proposed in recent years. By representing instances as binary hash codes and using Hamming distance to measure the semantic similarity, hashing-based retrieval systems can use bit operations to efficiently calculate the similarity between instances in a large-scale candidate set, consuming much smaller storage space. Existing cross-modal hashing methods either harness human-annotated semantic labels (*e.g.* 80 class labels in MS-COCO dataset) as in [10, 11], or utilize co-occurrence information of the input image-text pair as in [12]. Essentially, they represent the textual modality by some multi-label format and cast cross-modal hashing into a paradigm of multi-label matching. We strongly argue that this widely-adopted paradigm is incapable of capturing the

*Corresponding author.

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

fine-grained semantic information in the multi-modal data. Figure 1 illustrates several complex natural language queries. As shown, simple multi-label setting tends to lose the high-order relationships between objects (*e.g.*, man-ride-horse in the third example) and the attribute or numeral information (*e.g.*, green-shirt and a-woman in the second example), which may lead to sub-optimal results.

This work represents the first attempt of its kind on fine-grained cross-modal retrieval in the hashing field. We argue that existing hashing methods essentially rely on matching a pre-specified set of tags (*i.e.*, objects), thus failing to capture most fine-grained semantics in the images and texts. To ensure the retrieved results consistent with human cognition, one may expect fine-grained high-order semantics (such as attributes and object-predicate-subject type information) can also be preserved during cross-modal indexing and retrieval. However, rich information means longer codes and lower retrieval efficiency. To attack such issues, we design a two-step cross-modal hashing method, consisting of *preliminary screening* and *fine-grained re-ranking*. Specifically, a shorter hash code is first generated based on coarse global features, quickly filtering out semantically irrelevant candidates from large reference sets and leaving a shortlisted set of high-confident candidates. After that, transformer encoder and bottom-up visual attention mechanism are combined to encode fine-grained semantic information in text and images, respectively, obtaining longer, fine-grained binary codes. The candidates that survive in the previous filtering step are further re-ranked, leading to more accurate results.

In addition, our work also formally investigates the robustness of cross-modal hashing models and explores an adversarially-learned scheme. As reported by [13], due to the modern median-sized cross-modal dataset, existing models tend to suffer from data bias, which will greatly affect the model's generalization ability and robustness. When facing the adversarial attack, as proved in [14] existing methods are often fragile and vulnerable. To solve this, we propose two approaches to enhance the model. First, a novel scheme is devised to automatically generate adversarial samples for each sentence to form an adversarial setting, which forces the text encoder to distinguish sentences with small text editing distance and thereby increases their sensitivity to fine-grained semantic information. Second, we develop a protocol to generate rephrasing sentences to expand the original text library's diversity and mitigate the data bias. Our contributions can be summarized as below:

1. We explore the fine-grained semantic matching problem for the first time in the visual-text cross-modal hashing field. A two-step strategy is developed to establish a competitive baseline for this novel task setting.

2. We introduce a rephrase-based augmentation strategy and adversarial learning paradigm in the training procedure, which further enhance the proposed model's performance and robustness.

2. RELATED WORK

Cross-modal hashing [15, 16] has been researched in the computer vision field for many years. Traditional methods vary in the learning criteria, including the minimization of the reconstruction or quantization error [17, 18], or similarity preservation with graph-based constraint [19]. With the development of deep learning technology, [20] utilized neural networks to encode different modality data, [10] adopted the metric learning framework to reduce the semantic gap. However, these existing methods only focus on coarse concept correlation, but fine-grained higher-order semantic information is not explored.

To accomplish fine-grained retrieval, some researchers in the semantic matching field tried to encode the images or text into continuous embedding vectors and calculate the similarities according to the Cosine or Euclidean distance in some common space [21, 2, 22]. In recent years, cross attention [23] and graph convolution [6] have also been introduced into this task to achieve more precise matching. Nevertheless, these methods are often heavy-weight in computations. Our method can achieve comparable performance but in a much more efficient way.

3. PROPOSED METHOD

3.1. Problem Formulation

Our method aims to encode multi-modal instances as hash codes and project them into a common space for efficient retrieval by the hamming distance. Given a dataset that contains a set of images $\mathbf{V}$ and caption texts $\mathbf{T}$, we utilize the image-text pair information provided by the annotations to supervise the learning procedure.

To tackle large data, we propose a two-step strategy with ***primary screening*** and ***fine-grained re-ranking***. The former extracts the global information of images or texts and encodes into a relatively short length code (*e.g.* tens of bytes) and filter out most irrelevant samples, which significantly reduces the search space. Then, for each survived candidate (200 samples in our experiments), we further harness some fine-grained and high-order semantics to learn a re-ranking model to get longer hashing code in order to be sufficiently discriminative. Moreover, rephrased and adversarial sentences are also generated and included during training to further improve the model performance and robustness. Figure 2 illustrates the entire framework.

3.2. Primary Screening

Efficiency is the focus of consideration in this step. Global visual or textual features are supposed to be sufficient to filter out a large number of irrelevant samples.

Global visual encoder. Let $\boldsymbol{v}$ be a raw input image in the dataset $\mathbf{V}$, we utilize a ResNet50 pre-trained on ImageNet as the global visual feature extractor to get the feature map f_v^g.

Then it is compressed with a pixel-wise average pooling layer and passed to feature projection layers (implemented by MLP) to get a global semantic embedding e_v^g. A tailored bit-balanced hashing layer $Hash(\cdot)$ is followed to read e_v^g and

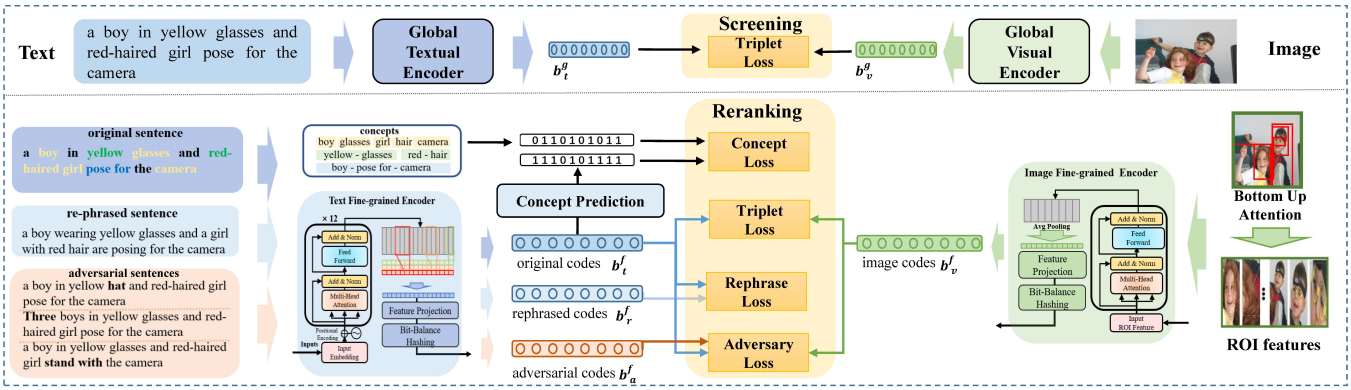

Fig. 2. The architectural diagram of our model. Note that the text encoders in primary screening and fine-grained re-ranking are the same architecture but with different parameters and output dimensions. Zoom in and view in color for better details.

produces a global visual binary code b_v^g:

$$v_m = med(softmax(e)), \quad b = sign(e - v_m), \qquad (1)$$

where $med(\cdot)$ represents the median function. During training $sign(\cdot)$ is approximated by $tanh(\cdot)$ for differentiation.

Global textual encoder. For text encoding, unlike most previous methods that only use bag-of-words or TF-IDF features as input and fully connected layers as the encoder, we expect the encoder can retain accurate and contextual semantic information in the text. We decide to borrow BERT [24] to mine the complex semantic context and produce the global textual embedding e_t^g. Then, the same bit-balance hash function as described in Eq. 1 is adopted to produce the textual binary code b_t^g. The text encoders in primary screening and fine-grained re-ranking have the same architecture but with different parameters and output dimensions. We will explain the detailed computation procedure in the next section.

3.3. Fine-grained Re-ranking

The hash codes obtained in the previous stage can only reflect global semantic information, which is not enough when facing complex queries. Therefore, more refined and longer hash codes b_v^f, b_t^f are required to explore the higher-order semantic information and achieve more accurate matching results.

Image fine-grained encoding. In order to obtain more refined visual semantic information, we need to find some meaningful regions or objects in the image and then explore the potential internal relations or attributes. Specifically, bottom-up-attention [25] is firstly employed to produce a series of salient regions and object proposals, with each represented by a pooled ROI feature r_i. To further refine the obtained features, a multi-heads self-attention mechanism is then applied to explore the relationship between objects and the higher-order semantic information.

Similar to the Transformer model [26], the self-attention layer is comprised of multi-head self-attention sub-layer $MultiHead(\cdot)$ and position-wise feed-forward sub-layer $FFN(\cdot)$. Residual connections followed by layer normalization are also applied around each of two sub-layers, through which semantic information can be propagated to higher layers. Computation details are explained in supplementary materials.

Assume we have a set of ROI features and forms as a matrix $M_R = [r_1, \cdots, r_{n_r}] \in \mathbb{R}^{n_r \times d_f}$ where n_r is the number of proposals, and d_f is the dimension of ROI feature vectors. The matrix is fed into a multi-head self-attention layer to extract high-order and interactive information between objects. Residual connection and layer-norm are followed to get the context feature output $O^v = [o_1^v; \cdots; o_{n_r}^v] \in \mathbb{R}^{n_r \times d_f}$:

$$O^v = LayerNorm(M_R + (MultiHead(M_R))). \qquad (2)$$

Then, the position-wise feed-forward network and layer normalization are applied to further adjust and encode the high-order information, whose output is:

$$z_i^v = LayerNorm(o_i^v + FFN(o_i^v)), i = 1, \cdots, n_r. \qquad (3)$$

After getting a set of high-order continuous representations $[z_1^v, \cdots, z_{n_r}^v]$, average pooling is adopted to aggregate the representations into a compact embedding $e_v^f = \frac{1}{n_r}\sum_i^{n_r} z_i^v$. Then same hashing function described in Eq.1 is applied to get a visual fine-grained bit-balanced binary code b_v^f.

Text fine-grained encoding. Given a sentence $\boldsymbol{t}$, it is tokenized to a set of word tokens and represented as a sequence of 1-hot vectors. Then each of the tokens is feed into a pre-trained BERT encoder to obtain a series of contextual embeddings $Z = \{z^1, \cdots, z^{n_t}\}$, where n_t is the number of tokens.

To further exploit the local context information of the sequential features, we adopt 1-dim convolutional neural networks and applied three kinds of kernels to explore uni-gram, bi-gram, and tri-gram information in the sentence, respectively. For the k_{th} embedding in the sequence, the output of kernel size $\boldsymbol{s}$ is:

$$p_{s,k} = ReLU(Conv1D_s(z_{k:k+s})), s \in \{1, 2, 3\}, \qquad (4)$$

Zero-padding is applied to keep the length consistent. Choosing $s \in \{1, 2, 3\}$, an element-wise max-pooling is adopted to get a fixed length output $q_s = max\{p_{s,1}, \cdots, p_{s,n_t}\}$. The obtained three feature vectors q_1, q_2, q_3 are then concatenated together and passed to a feature projection layer to get the text continuous embedding e_t^f. Also, the bit-balance layer is followed to produce the final binary code b_t^f.

The optimization of the distance between b_v^f and b_t^f is piloted by cross-modal consistency. More details are deferred to the section "joint learning paradigm".

3.4. High-order Concepts Guidance

While simple distance constraint cannot get a good enough code representation, we hope the obtained code also have sufficient high-order and fine-grained semantic modeling capabilities. As we argued, most previous methods only tend to utilize the human-annotated low-level tag information to supervise the training procedure, which can not capture the full semantics of a complex text. Thus we further leverage high-order phrases (relationship triplets and attribute tuples) to offer a comprehensive semantic understanding beyond tags. Specifically, after getting the fine-grained code b_t^f, we further designed a concept prediction module (CPM) to model the higher-order semantic contained in it. To achieve this, we parsed all captions in the training set and constructed a concept library not only containing a set of low-level nouns but also higher-order concepts (e.g., $\langle apple \rangle, \langle red, apple \rangle, \langle apple, on, table \rangle$ for "a red apple is on the table") to supervise the training procedure. The construction details are described in supplementary materials. Given b_t^f, a Multi-layer Perceptron (MLP) network and Sigmoid function are stacked to formulate the CPM and produce the concept prediction vector X.

$$X = \sigma(MLP(b_t^f)),\ X = \{x_1, \cdots, x_K\} \in \{0,1\}^K, \quad (5)$$

where K is the predefined concept numbers, and σ is the sigmoid activation function. The obtained prediction is compared with the corresponding ground truth Y, thereby enhancing the semantic modeling ability of the hash code.

3.5. Rephrased & Adversarial Learning

Besides, due to the inevitable bias in the dataset (e.g., two objects may co-occur with each other in most cases like table and plate) and limited scale, we found the learned sentence encoders usually pay attention to only part of the sentence (e.g., high-frequency keywords) thus ignore other useful information. To tackle this issue, we propose a rephrased-based data augmentation strategy and an adversarial learning mechanism to force the encoder to reflect subtle deviations from the fine-grained meaning of the text data.

For each sentence $\boldsymbol{t}$, we generate two kinds of corresponding augmentation samples. Specifically, one is sentences with subtle modification from the source (e.g., replacing the object nouns, attributes, or changing the relationship in $\boldsymbol{t}$ to produce different semantics), we called them adversarial samples and note as t_{adv}. While the other kind is the rephrase sentences (e.g., replacement of synonyms or restatements to keep same semantics) which are denoted as t_{rep}. We display some examples in the peach box and light blue box in Figure 2 and explain the detailed generation procedure in the supplementary materials. Note b_a^f and b_r^f as the codes of t_{adv} and t_{rep} respectively. Intuitively, we expect the model is sensitive enough and distinguish the subtle difference between the original sentence and the adversarial samples, which will help the learning of fine-grained semantics. Thus, the distance between b_t^f and b_a^f should be sufficiently large. On the other hand, to avoid the model converging to a limited number of high-frequency keywords, the distance between b_t^f and b_r^f should be close enough.

3.6. Joint Learning Paradigm

Let b_v denote the binary code of image $\boldsymbol{v}$, and b_t denote the code of text $\boldsymbol{t}$. While $b_{a,i}^f$ represents the binary code of $t_{adv,i}$ and $b_{r,j}^f$ represents the binary code of $t_{rep,j}$.

To constrain the score distance between positive and negative pairs, a bi-directional triplet ranking loss with hard example mining strategy is adopted in the training process. For a positive pair $(\boldsymbol{v}, \boldsymbol{t})$, the triplet loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{trip}(v,t) =& max[0, m_1 - s(b_v, b_t) + s(b_v, b_t^{'})] \\ &+ max[0, m_1 - s(b_v, b_t) + s(b_v^{'}, b_t)],\end{aligned} \quad (6)$$

where $m_1 = 0.2$ is the margin parameter, $s(\cdot)$ is the similarity function which is instanced by cosine similarity, $b_t^{'}$, $b_v^{'}$ denote the hardest negative sentence and image for $\boldsymbol{v}$ and $\boldsymbol{t}$ in the mini-batch respectively.

Besides, as mentioned above, to ensure the distance between b_a^f and b_v^f larger than the ground-truth pair, and b_r^f close to b_t^f, we further define the adversarial loss and rephrase loss in the following way:

$$\begin{aligned}\mathcal{L}_{adv}(v,t,t_{adv}) &= \sum_{i=1}^{n_i} max[0, m_2 + s(b_{a,i}^f, b_v^f) - s(b_t^f, b_v^f)], \\ \mathcal{L}_{rep}(t,t_{rep}) &= \sum_{j=1}^{n_j} (1 - s(b_t^f, b_{r,j}^f)),\end{aligned} \quad (7)$$

where, n_i, n_j are the quantity for adversarial and rephrased sentences, respectively, $m_2 = 0.5$ is a margin hyper-parameter to penalize the adversarial pairs.

A multi-label cross-entropy loss is further adopted to read the predicted concept label X and ground-truth labels Y to supervise the high-order concepts learning.

$$\mathcal{L}_{con} = -\frac{1}{K} \sum_{c=1}^{K} [y_c \cdot log(x_c) + (1 - y_c) \cdot log(1 - x_c)], \quad (8)$$

where x_c and y_c are the predicted and ground-truth label for the c_{th} concept.

We synthesize all above loss functions, leading to the final loss formulation:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{trip} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{rep} + \lambda_4 \mathcal{L}_{con}, \quad (9)$$

where all λs are hyper-parameters. We train the screen model and re-ranking model separately with full training set, and $\lambda_{2,3,4}$ are set to zero when training the screening model. In the inference phase, we replace $s(\cdot)$ with the hamming distance to measure the similarity between two binary codes.

4. EXPERIMENTS

4.1. Datasets and Experiments Settings

We chose two public datasets for evaluation. **MS-COCO** [29] is a large-scale dataset containing 123,287 images, and each image annotated with five text descriptions. We follow [5] to

Table 1. Comparison with state-of-the-art hashing methods.

Models	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
	Results on MS-COCO					
DCMH[27]	4.0	17.0	29.0	4.0	14.0	27.0
CMHH[8]	5.0	8.0	24.0	4.0	15.0	20.0
SCAHN[28]	19.0	47.0	67.0	17.0	48.0	68.0
DJSRH[12]	51.0	86.0	93.0	41.0	85.0	93.0
S(256bit)	60.0	91.0	94.0	60.0	91.0	94.0
R(2048bit)	62.0	92.0	96.0	56.0	90.0	95.0
S+R	**69.0**	**94.0**	**99.0**	**61.0**	**92.0**	**99.0**
Models	**Results on Flickr30K**					
DCMH[27]	2.0	6.0	12.0	2.0	8.0	9.0
CMHH[8]	2.0	6.0	12.0	3.0	9.0	17.0
SCAHN[28]	3.0	13.0	24.0	3.0	11.0	21.0
DJSRH[12]	25.0	57.0	68.0	29.0	55.0	66.0
S(256bit)	58.0	88.0	94.0	60.0	89.0	94.0
R(2048bit)	70.0	91.0	**96.0**	62.0	93.0	95.0
S+R	**74.0**	**93.0**	**96.0**	**72.0**	**95.0**	**96.0**

prepare the training, validation, and testing set. All images are split into three parts which contain 113,287, 5,000, 5,000 samples, respectively. The testing results are reported by averaging over five folds of 1,000 test images. **Flickr30K** [30] contains 31,783 images collected from the Flickr website also with five captions each. We followed the split in [2] and [31] that used 1,000 images for testing and 1,000 images for validation. The rest are used for training.

For evaluation metrics, we argue the commonly used mean Average Precision (mAP) in the hashing field is based on a multi-label scheme, which is not sufficient to reflect the model performance in the fine-grained scenario (with complex attributes and relations). So we propose to follow the common practice in the cross-modal matching field and measure the performance by recall at top-K (R@K), which is defined as the fraction of queries for which the ground-truth item is retrieved in the closest K points to the query. K is set to 1, 5, and 10 in all experiments. We present more implementation details in the supplementary materials.

4.2. Comparisons with State-of-the-art Methods

We first compare with four state-of-the-art and open-source cross-modal hashing methods on two benchmarks. We follow [9] and randomly select 10,000 samples to form the training set. Due to the poor performance of the baseline methods, we only report the results on a test set with 100 samples, which is shown in the upper part of Table 1. We first present the results of only using the screening model and only the rerank model (noted as "S" and "R"), and also report the results of re-ranking after the primary screening (noted as "S+R"). It is clear that even the screening model dominates all the baseline methods in all indicators. This also shows previous hashing methods are basically unable to handle fine-grained queries with complex semantics.

For better evaluation, we also compare with several powerful *continuous embedding-based* cross-modal matching methods and achieve state-of-the-art performance, which are shown in the supplementary materials.

4.3. Adversarial Attack in Sentence Retrieval

To evaluate the proposed method's robustness, we select the fine-grained reranking model (2048bit) and attack it by adversarial samples in the sentence retrieval scenario. We first extend each caption with 30 adversarial samples (10 noun-typed, 10 numeral-typed, and 10 relation-typed). Therefore, each image has 30×5 contrastive adversarial samples in total. The candidate retrieval set for each image now becomes 5000 + 150. We design four kinds of adversarial learning protocol: noun-type, numeral-type, relation-type, and mixed-types. Under each specific kind, we use 10 corresponding type adversarial samples for each sentence in the training phase. Besides, we also use all the 30 samples in the mixed setting. Results are shown in Table 2.

Comparing the results without attack (the 1_{st} row in Table 2), the reranking model has a noticeable performance drop in each indicator, especially when facing the relation-type attack, the R@1 drop 54.9 points ($64.4 \rightarrow 9.5$). In comparison, the drop on the noun-type attack is much more slightly, which shows that the model tends to focus on frequent key nouns and ignore the exploration of potential higher-order semantic information (*i.e.*, relationships).

Nonetheless, when we added some adversarial samples in the training process, the situation clearly improves. It demonstrates that the proposed adversarial learning strategy can significantly improve the model robustness. Overall, training the model with one type gains the best performance on robustness against the adversarial attack of the type itself. Training with relation-typed adversarial samples helps improve the robustness against a noun-typed attack, which suggests the necessity to explore higher-order information in the cross-modal retrieval field. We can also find that using more adversarial samples in the training procedure helps the model to obtain better robustness and accuracy on top1 results when facing all kinds of attacks.

4.4. Ablation Study

To explore the effectiveness of the proposed rephrased sentences and high-order concepts, we conduct ablation experiments on the Flickr30k dataset. We choose the rerank models with output lengths of 1024bit and 2048bit as the basis and sequentially subtract the rephrase data and concept prediction branch from them. From the results in Table 3, we can clearly find that the model's performance and generalization ability have decreased in most indicators without the rephrased training samples and high-order concepts guidance.

We also explore the effect of code length on the retrieval result, deferred to the supplementary materials.

4.5. Efficiency Evaluation

To evaluate the retrieval efficiency of the proposed method, we select three representative and open-source embedding-based image-text matching methods for comparison. We use a database containing 100,000 samples as a benchmark and record the total time cost for 5000 queries to rule out sampling randomness. We also calculate the storage space requirement. We report the encoding time of the query data and the retrieval time in the database separately. The results are shown in Table 4. It can be found that our proposed screen model has greatly improved retrieval efficiency and reduced

Table 2. Detailed results on each type of adversarial attack for sentence retrieval scenario on Flickr30K. 'Mixed' represents using all three kinds of adversarial samples. +Mixed_adv10 means randomly select 10 adversarial samples for each sentence in the training process.

Models	Noun adversarial			Num adversarial			Rela adversarial			Mixed adversarial		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Rerank(W/O Attack)	68.2	93.5	97.5	54.5	86.5	93.4	64.4	86.6	92.9	44.9	75.2	84.4
Rerank(With Attack)	33.9	74.9	89.3	31.3	65.3	82.5	9.50	43.3	77.8	6.30	27.5	53.6
+Noun_adv	**52.1**	**79.3**	87.6	27.9	61.5	78.9	16.5	61.8	82.1	10.3	43.8	69.2
+Num_adv	29.6	71.2	86.4	**58.3**	**83.7**	**90.6**	8.40	40.9	73.4	6.00	30.8	62.5
+Rela_adv	40.0	76.4	87.0	26.3	62.1	78.5	**49.6**	80.2	88.6	20.5	57.1	75.5
+Mixed_adv10	49.2	81.0	**88.8**	50.1	81.6	89.1	44.3	**80.7**	**89.1**	40.3	79.5	**88.6**
+Mixed_adv30	51.9	79.1	88.3	51.6	79.3	88.3	49.1	78.7	88.1	**43.9**	**79.8**	88.4

Table 3. Ablation study results on Flickr30K.

Models	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Full Rerank(2048bit)	**64.4**	86.6	**92.9**	**44.9**	**75.2**	**84.4**
-Rephrase	62.1	**87.3**	92.7	44.1	74.7	83.4
-Concepts	60.7	86.1	92.7	44.2	74.9	83.9
Full Rerank(1024bit)	**62.3**	**86.7**	**93.4**	**43.1**	**74.0**	**83.4**
-Rephrase	60.0	85.6	91.9	42.6	72.7	81.8
-Concepts	57.2	83.8	91.2	42.2	73.1	82.0

Table 4. Retrieval efficiency comparison, T_{enc}, I_{enc} represent the encoding time for texts and images, respectively. "Search" shows the time cost to rank all the results. All experiments are conducted on a Linux server with one NVIDIA TITAN-X GPU and two Intel Xeon E5-2697-v4 CPUs.

Methods	T_{enc} (S)	I_{enc} (S)	Search (S)	Storage (MB)
Screen(256bit)	6.055	**11.74**	**32.17**	**3.052**
Rerank(2048bit)	6.157	410+1.991	374.2	24.41
Screen+Rerank	12.21	423.7	32.17+1.34	27.46
VSE++(1024dim)[2]	**0.527**	15.62	1682	390.6
CMPM(512dim)[32]	0.646	21.41	877.1	195.3
SAEM(256dim)[5]	6.788	410+0.926	465.3	97.65

storage requirements compared to previous methods. For example, the search efficiency has improved by 52 times compared to VSE++ [2], while storage requirements reduced by 127 times. It should be noted that the 410s time in the image encoding part is cost by Faster-RCNN to extract ROI features. If it is replaced with a faster object detector, the efficiency will be further improved. The 1.34 second in the third row is the re-ranking time on the top 200 screened candidates.

5. CONCLUSION

In this work, we first explored the importance of complex queries, fine-grained semantics, and higher-order information in the hashing field. The well-designed two-step retrieval paradigm retained high accuracy while effectively improving retrieval efficiency. The novelly generated rephrase sentences and adversarial sentences further improved the model performance and robustness. Abundant ablation studies illustrated the effectiveness of all the devised components and strategies. Comprehensive experiments on two datasets suggested our method's superiority in both efficiency and accuracy.

Acknowledgement: This work is supported by National Natural Science Foundation of China (61772037) and Beijing Natural Science Foundation (Z190001).

6. REFERENCES

[1] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM MM*, 2010.

[2] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler, "Vse++: Improved visual-semantic embeddings," *arXiv:1707.05612*.

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

[4] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov, "Devise: A deep visual-semantic embedding model," in *NeurIPS*, 2013.

[5] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang, "Learning fragment self-attention embeddings for image-text matching," in *ACM MM*, 2019.

[6] Yongzhi Li, Duo Zhang, and Yadong Mu, "Visual-semantic matching by exploring high-order attention and distraction," in *CVPR*, June 2020.

[7] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou, "Cross-modal deep variational hashing," in *ICCV*, 2017.

[8] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang, "Cross-modal hamming hashing," in *ECCV*, 2018.

[9] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *CVPR*, 2018.

[10] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang, "Semantics-preserving hashing for cross-view retrieval," in *CVPR*, 2015.

[11] Botong Wu, Qiang Yang, Wei-Shi Zheng, Yizhou Wang, and Jingdong Wang, "Quantized correlation hashing for fast cross-modal search," in *IJCAI*, 2015.

[12] Shupeng Su, Zhisheng Zhong, and Chao Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *ICCV*, 2019.

[13] Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun, "Learning visually-grounded semantics from contrastive adversarial samples," *arXiv:1806.10348*, 2018.

[14] Erkun Yang, Tongliang Liu, Cheng Deng, and Dacheng Tao, "Adversarial examples for hamming space search," *IEEE transactions on cybernetics*, 2018.

[15] X. Liu, J. He, C. Deng, and B. Lang, "Collaborative hashing," in *CVPR*, 2014.

[16] Yi Zhen and Dit-Yan Yeung, "Co-regularized hashing for multimodal data," in *NeurIPS*, 2012.

[17] Fangxiang Feng, Xiaojie Wang, and Ruifan Li, "Cross-modal retrieval with correspondence autoencoder," in *ACM MM*, 2014.

[18] Mingsheng Long, Yue Cao, Jianmin Wang, and Philip S Yu, "Composite correlation quantization for efficient multimodal retrieval," in *SIGIR*, 2016.

[19] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *SIGMOD*, 2013.

[20] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *SIGKDD*, 2016.

[21] Guy Lev, Gil Sadeh, Benjamin Klein, and Lior Wolf, "Rnn fisher vectors for action recognition and image annotation," in *ECCV*, 2016.

[22] 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 *CVPR*, 2018.

[23] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He, "Stacked cross attention for image-text matching," in *ECCV*, 2018.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, 2018.

[25] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[27] Qing-Yuan Jiang and Wu-Jun Li, "Deep cross-modal hashing," in *CVPR*, 2017.

[28] Xinzhi Wang, Xitao Zou, Erwin M Bakker, and Song Wu, "Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval," *Neurocomputing*, 2020.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[30] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[31] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[32] Ying Zhang and Huchuan Lu, "Deep cross-modal projection learning for image-text matching," in *ECCV*, 2018, pp. 686–701.