Hashing based Efficient Inference for Image-Text Matching

Rong-Cheng Tu[†], **Lei Ji** [‡§¶*]**, Huaishao Luo**[‖]**, Botian Shi**[≀]**,**
Heyan Huang[†]**, Nan Duan**[¶] and **Xian-Ling Mao**[†]
[†]School of Computer Science and Technology,
Beijing Institute of Technology, Beijing, China, Beijing, China
[‡]Institute of Computing Technology, CAS, Beijing, China
[§]University of Chinese Academy of Sciences, Beijing, China
[¶]Microsoft Research Asia, Beijing, China
[‖]Southwest Jiaotong University, Chengdu, China
[≀]Shanghai AI lab, Shanghai, China
[†]{turongcheng, hhy63, maoxl}@bit.edu.cn, [¶]{leiji,nanduan}@microsoft.com

Abstract

Image-text matching has been a popular research topic which bridges vision and language through semantic understanding. Recent works mainly focus on exploring the interactions between images and sentences to improve the performance without considering inference efficiency. Specifically, for the large scale databases, it is unacceptable to perform such time-consuming mechanisms between a query (text/image) and each candidate datapoint (image/text) in the whole retrieval set during inference. To tackle this problem, we propose a novel hashing based efficient inference module called HEI, which can be plugged into the existing framework to speed up the inference step without reducing the retrieval performance. In details, HEI learns to map the original datapoints into short binary hash codes and coarsely preserve the heterologous matching relationship. Thus, in the inference phase, the proposed HEI module uses the hash codes to quickly select a few candidate datapoints from the retrieval set for a given query. Then, the image-text matching model fine ranks the candidate set to find the matching datapoint. Extensive experiments on two widely used benchmark MS-COCO and Flickr30k with four baseline methods demonstrate the efficiency and effectiveness of our proposed HEI module.

1 Introduction

Language and vision understanding plays a fundamental role for human to understand the real world. A large amount of works has been proposed to bridge these two modalities. Image-text matching is one of the fundamental topics in this field, which benefits a series of downstream applications, such as visual captioning(Zhang et al., 2019; Wang et al., 2018) and visual grounding (Chen et al., 2018; Plummer et al., 2017). Specifically, given an image (text), its target is to match the closest textual description (image) for the image (text).

Early works (Karpathy and Fei-Fei, 2015; Wang et al., 2016; Niu et al., 2017; Faghri et al., 2017) achieve this goal by learning two modality-specific deep neural networks to directly map all the datapoints from the two modality into a common joint space without using attention mechanism, and then measures their similarities by feature distances in the joint space. Compared with these methods, recent works (Lee et al., 2018; Liu et al., 2019; Wang et al., 2019; Chen et al., 2020) mainly focus on incorporating variant attention mechanisms into the image-text matching models to explore the fine-grained interactions between vision and language. By using the attention mechanisms, the image-text matching models are able to filter out irrelevant information, and find the fine-grained cues to achieve a great matching performance. For example, CAMP (Wang et al., 2019) takes comprehensive and fine-grained cross-modal interactions into account, and also properly handles negative pairs and irrelevant information with an adaptive gating scheme to improve the matching performance.

Although existing attention mechanism based methods achieve great performance, they do not take the inference efficiency into account. Specifically, for the large scale databases, due to the attention mechanisms being time-consuming, it is unacceptable to perform such complex attention mechanisms between the query (text/image) and each candidate datapoint (image/text) in the whole retrieval set during inference. Thus, it is critical to improve the inference speed of these methods.

Intuitively, if a small candidate set containing positive datapoints can be quickly selected out, the image-text matching models can greatly speed up through only fine ranking such a small candidate set instead of the whole retrieval set. Then the key

[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 743–752
August 1–6, 2021. ©2021 Association for Computational Linguistics

challenge is how to quickly select such a small candidate set. Hashing is widely used in the field of data search with fast retrieval speed. Besides, although it can hardly perform the accurate matching, hashing is capable of quickly selecting a high quality candidate set containing the positive datapoints.

Hence, in this paper, we propose a novel hashing based efficient inference module, called HEI, which can be plugged into the existing attention mechanism based image-text matching framework to speed up the inference step without reducing the retrieval performance. Specifically, a matching score based hashing loss is proposed, which consists of two items: one is used to make Hamming similarity between hash codes of matching datapair be as large as possible; the other item is to make the Hamming similarity between hash codes of mismatching datapair no larger than their corresponding matching score. By minimizing the proposed hashing loss, the HEI module is optimized to map the original datapoints into short binary hash codes and coarsely preserve the heterologous matching relationship between datapoints. Thus, the trained HEI module can be used to speed up the inference step without reducing the retrieval performance. Extensive experiments on two widely used benchmark MS-COCO and Flickr30k with four baselines demonstrate the effectiveness of our proposed HEI module.

2 Related Work

2.1 Text-image Matching

Recently, many image-text matching methods have been proposed, which can be roughly grouped into one-to-one matching methods learning correspondence between the whole image and text, and many-to-many matching methods learning correspondence between the regions of image and the words of text.

The one-to-one matching methods (Wang et al., 2016; Kiros et al., 2014; Zhang and Lu, 2018; Zheng et al., 2020) mainly aim to explore the heterologous relationship globally by mapping the whole images and the full texts into a common feature space. However, owing to these methods doing not explore the correspondence between image regions and text words, it might lead to learn sub-optimal features, which will damage the text-image matching performance.

By utilizing variant cross-modal attention mechanisms, many-to-many matching methods can explore the correspondence between image regions and text words, thus, these attention mechanism based methods can achieve the state-of-the-art performance. For instance, BFAN (Liu et al., 2019) is proposed to eliminate partial irrelevant words and regions from the shared semantic in image-text pairs to achieves state-of-the-art performance on several benchmark datasets. IMRAM (Chen et al., 2020) proposes a recurrent attention memory which incorporates a cross-modal attention unit and a memory distillation unit to refine the correspondence between image regions and text words. However, those attention mechanisms used by the many-to-many matching methods are usually complicated with high computation complexity. Hence, it is unacceptable to perform such time-consuming attention mechanisms between the query (text/image) and each candidate datapoint (corresponding to image/text) in the whole retrieval set during inference especially for the large scale databases.

Different from previous methods, our model explores hashing technology to improve the inference speed of the existing many-to-many text-image matching methods without reducing their performance.

2.2 Cross-Modal Hashing

The core of cross-modal hashing is to project the datapoints of different modalities into compact binary hash codes, meanwhile, preserve the semantic similarity of original datapoints. Then, in the cross-modal retrieval phase, the datapoints of the retrieval set can be sorted by the Hamming distance between their corresponding binary hash codes calculated by the ‘XOR’ operation, which has fast retrieval speed. Due to this advantage, a mount of cross-modal hashing methods have been proposed (Hu et al., 2020; Su et al., 2019; Lin et al., 2020; Tu et al., 2020; Shi et al., 2019). For example, SDCH (Lin et al., 2020) utilizes a semantic label branches to preserve semantic information of the learned features by integrating with inter-modal pairwise loss, cross-entropy loss and quantization loss.

However, due to these hashing methods belonging to approximate nearest neighbour (ANN) searching technology, they can hardly to accurately find the matching datapoint for a query. Hence, few works explore the hashing technology for text-image matching. To our best knowledge, this is the first work to explore the use of hashing to improve the inference speed of existing attention mecha-

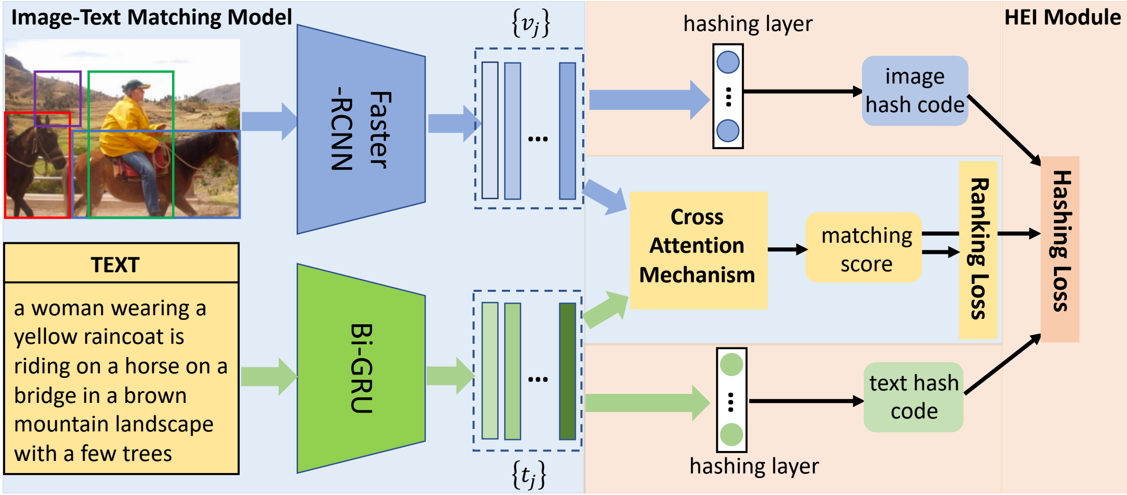


Image-Text Matching Model
TEXT
a woman wearing a yellow raincoat is riding on a horse on a bridge in a brown mountain landscape with a few trees
Faster -RCNN
Bi-GRU
$\{v_j\}$
$\{t_j\}$
hashing layer
image hash code
HEI Module
Cross Attention Mechanism
matching score
Ranking Loss
Hashing Loss
text hash code
hashing layer

Figure 1: The architecture of image-text method with the HEI module.

nism based image-text matching methods.

3 Methodology

As shown in Figure 1, different from the architecture of existing matching models, our framework introduces an extra hashing based efficient inference module, called HEI, which consists of an image modal hashing layer and a text modal hashing layer, and each hashing layer is a fully-connected layer with k units where k is the hash codes length.

3.1 Problem formulation and notations

Without loss of generality, suppose there are datasets with M images $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{M}$ and N texts $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{N}$. Given an image $\boldsymbol{x}_i$ with its region-level visual features denoted as $\boldsymbol{V}_i = [\boldsymbol{v}_1^i, \cdots, \boldsymbol{v}_m^i]$, and a text $\boldsymbol{y}_j$ with its word-level textual features denoted as $\boldsymbol{T}_j = [\boldsymbol{t}_1^j, \cdots, \boldsymbol{t}_n^j]$, the goal of image-text matching is to calculate a matching score $\boldsymbol{s}_{ij}$ for the image $\boldsymbol{x}_i$ and the text $\boldsymbol{y}_j$ based on their features $\boldsymbol{V}_i$ and $\boldsymbol{T}_j$. Moreover, if the image $\boldsymbol{x}_i$ and the text $\boldsymbol{y}_j$ are matching, the matching score $\boldsymbol{s}_{ij}$ should be large, otherwise $\boldsymbol{s}_{ij}$ should be small.

Furthermore, the goal of hashing based efficient inference module is to learn the two modality-specific hashing layer which can map their corresponding modal datapoints into binary hash codes with the heterologous matching relationship preserved.

3.2 Cross-modal Feature Representation

3.2.1 Image region-level visual features

To obtain the region-level visual features $\boldsymbol{V}_i = [\boldsymbol{v}_1^i, \cdots, \boldsymbol{v}_m^i]$ of the image $\boldsymbol{x}_i$, we first employ the Faster R-CNN (Ren et al., 2016) model using ResNet-101 (He et al., 2016) as the backbone, which is pre-trained on the Visual Genomes dataset (Krishna et al., 2017) by (Anderson et al., 2018), to extract the top m region proposals of the image. Then, by average-pooling the spatial feature map, a feature vector $\boldsymbol{v}_j^{i\prime} \in \mathcal{R}^{2048}$ for the j^{th} region proposal is calculated. Finally, We obtain the d-dimentional region features with a linear projection layer:

$$\boldsymbol{v}_j^i = \boldsymbol{W}_v \boldsymbol{v}_j^{i\prime} + \boldsymbol{b}_v \tag{1}$$

where $\boldsymbol{W}_v$ and $\boldsymbol{b}_v$ are to-be-learned parameters, and $\boldsymbol{v}_j^i$ is the visual feature for the j^{th} region of image $\boldsymbol{x}_i$.

3.2.2 Text word-level textual features

To obtain the textual features of a input text $\boldsymbol{y}_j$ with n words, we first embed each word $\boldsymbol{w}_i$ of the input text $\boldsymbol{y}_j$ into a 300-dimensional vector $\boldsymbol{t}_i^{j\prime}$. Then, to enhance the word-level feature with sufficient context information, we use a single-layer bi-directional GRU (Cho et al., 2014) with d-dimensional hidden to summarize information from both forward and backward directions of the input text $\boldsymbol{y}_j$:

$$\begin{aligned} \overrightarrow{\boldsymbol{h}_i^j} &= \overrightarrow{GRU}(\overrightarrow{\boldsymbol{h}_{i-1}^j}, \boldsymbol{t}_i^{j\prime}), \\ \overleftarrow{\boldsymbol{h}_i^j} &= \overleftarrow{GRU}(\overleftarrow{\boldsymbol{h}_{i-1}^j}, \boldsymbol{t}_i^{j\prime}). \end{aligned} \tag{2}$$

where $\overrightarrow{\boldsymbol{h}_i^j}$ and $\overleftarrow{\boldsymbol{h}_i^j}$ denote hidden states from the forward GRU and the backward GRU, respectively. Then, the textual feature of the word $\boldsymbol{w}_i^j$ in the text $\boldsymbol{y}_j$ is defined as:

$$\boldsymbol{t}_i^j = \frac{\overrightarrow{\boldsymbol{h}_i^j} + \overleftarrow{\boldsymbol{h}_i^j}}{2} \tag{3}$$

3.3 General Attention Framework

Existing attention mechanism based image-text matching methods mainly learn to associate shared semantics between the region-level feature $\boldsymbol{V}_i$ of image $\boldsymbol{x}_i$ and word-level feature $\boldsymbol{T}_j$ of text $\boldsymbol{y}_j$ through variant cross-attention mechanisms to calculate the matching score $\boldsymbol{s}_{ij}$, which can be formulated as follows:

$$\boldsymbol{s}_{ij} = CAM(\boldsymbol{V}_i, \boldsymbol{T}_j; \boldsymbol{W}) \tag{4}$$

where $CAM(\cdot; \boldsymbol{W})$ denotes the cross-modal attention mechanism and $\boldsymbol{W}$ is the set of learnable parameters. For example, in BFAN (Liu et al., 2019), $CAM(\cdot; \boldsymbol{W})$ denotes the Focal attention mechanism proposed in the original paper.

Then to maximize matching scores of the matching image-text pairs and minimize the ones of the mismatching datapairs, a hinge-based triplet ranking loss with emphasis on the hard negatives are used as the loss function. Specifically, given a pair of matching image-text $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$, we denote their matching score as $\boldsymbol{s}_{ij}$; $\bar{j} = argmax_{t \neq j} \boldsymbol{s}_{it}$ denotes the hard negative when using the image to match text; $\bar{i} = argmax_{t \neq i} \boldsymbol{s}_{tj}$ denotes the hard negative when using the text to match image, then the ranking loss is formulated as:

$$\mathcal{L}_{rank} = [\alpha - \boldsymbol{s}_{ij} + \boldsymbol{s}_{i\bar{j}}]_+ + [\alpha - \boldsymbol{s}_{ij} + \boldsymbol{s}_{\bar{i}j}]_+ \tag{5}$$

where α is the margin for the ranking loss, and $[a]_+ = max(0, a)$.

Finally, after optimizing the matching model, given a query datapoint, it will be used to calculated the matching score with each datapoint in the retrieval set to find the most matching one by the cross-attention mechanism. However, the cross-attention mechanism is time-consuming which means it unacceptable to calculate a matching score between the query and each point in retrieval set with the attention mechanism during inference. Thus, we propose a hashing based efficient inference module to improve the inference speed.

3.4 Hashing based Efficient Inference module

Specifically, the input of the HEI module is the fragment-level feature of datapoint, i.e., the region-level feature $\boldsymbol{V}_i$ for an image modal input $\boldsymbol{x}_i$ or the word-level feature $\boldsymbol{T}_j$ for a text modal input $\boldsymbol{y}_j$. We further aggregate the fragment-level feature $\boldsymbol{V}_i$ ($\boldsymbol{T}_i$) into an instance-level feature $\hat{\boldsymbol{v}}_i$ ($\hat{\boldsymbol{t}}_i$) for an image (text) datapoint $\boldsymbol{x}_i$ ($\boldsymbol{y}_i$):

$$\hat{\boldsymbol{v}}_i = \sum_{j=1}^{m} a_j \boldsymbol{v}_j^i; \quad a_j = \frac{\boldsymbol{v}_j^{iT} \boldsymbol{w}_v}{\sum_{k=1}^{m} \boldsymbol{v}_k^{iT} \boldsymbol{w}_v}. \tag{6}$$

$$\hat{\boldsymbol{t}}_i = \sum_{j=1}^{n} q_j \boldsymbol{t}_j^i; \quad q_j = \frac{\boldsymbol{t}_j^{iT} \boldsymbol{w}_t}{\sum_{k=1}^{n} \boldsymbol{t}_k^{iT} \boldsymbol{w}_t}. \tag{7}$$

where $\boldsymbol{w}_t$ and $\boldsymbol{w}_v$ denote learnable parameter. Then by forwarding the instance-level feature $\hat{\boldsymbol{v}}_i$ ($\hat{\boldsymbol{t}}_i$) into the image (text) modal hashing layer, the hash codes $\boldsymbol{b}_i^v$ ($\boldsymbol{b}_i^t$) of image $\boldsymbol{x}_i$ (text $\boldsymbol{y}_i$) can be generate as:

$$\begin{aligned} \boldsymbol{b}_i^v &= sgn(\mathcal{H}_x(\hat{\boldsymbol{v}}_i; \boldsymbol{\Theta}_v)) \in \{-1, 1\}^k \\ \boldsymbol{b}_i^t &= sgn(\mathcal{H}_y(\hat{\boldsymbol{t}}_i; \boldsymbol{\Theta}_t)) \in \{-1, 1\}^k \end{aligned} \tag{8}$$

where $\mathcal{H}_x(\hat{\boldsymbol{v}}_i; \boldsymbol{\Theta}_v)$ denotes the image modal hashing layer and $\boldsymbol{\Theta}_v$ denotes the set of parameters in the image hashing layer; k is the length of hash codes; $\mathcal{H}_y(\hat{\boldsymbol{t}}_i; \boldsymbol{\Theta}_t)$ represents the text modal hashing layer and $\boldsymbol{\Theta}_t$ represents the set of parameters in the text hashing layer; $sgn(\cdot)$ is an element-wise sign function, which returns 1 if the element is positive and returns -1 otherwise.

Furthermore, the core of hashing based efficient inference module is to learn two modality-specific hashing layer to map the datapoints into binary hash codes which are used to select a few candidate datapoints from the retrieval set for an query. To achieve this goal, the learned hash codes should coarsely preserve the heterologous matching relationship between datapoints, i.e., if two datapoints are matching, then the Hamming distance between their corresponding binary hash codes should be small, otherwise it should be large.

Considering that the Hamming distance between $\boldsymbol{b}_i^v$ and $\boldsymbol{b}_j^t$ can be difined as: $D_H(\boldsymbol{b}_i^v, \boldsymbol{b}_j^t) = 0.5(k - \boldsymbol{b}_i^{vT} \boldsymbol{b}_j^t)$, where k denotes the code length. It means when $\frac{1}{k}\boldsymbol{b}_i^{vT} \boldsymbol{b}_j^t$ is close to 1, the Hamming distance $D_H(\boldsymbol{b}_i^v, \boldsymbol{b}_j^t)$ is close to 0; and when $\frac{1}{k}\boldsymbol{b}_i^{vT} \boldsymbol{b}_j^t$ is close to -1, the Hamming distance $D_H(\boldsymbol{b}_i^v, \boldsymbol{b}_j^t)$

is close to k. Thus, $\frac{1}{k}\boldsymbol{b}_i^{vT}\boldsymbol{b}_j^t$ can be used to denote the Hamming similarity between $\boldsymbol{b}_i^v$ and $\boldsymbol{b}_j^t$, and measure the heterologous matching relationship preserved by $\boldsymbol{b}_i^v$ and $\boldsymbol{b}_j^t$. Furthermore, as the mathching score $\boldsymbol{s}_{ij} \in [0,1]$ of image $\boldsymbol{x}_i$ and text $\boldsymbol{y}_j$ computed by the cross-attention mechanism may preserve the heterologous matching relationship to a certain extent, then we can use it as soft target to supervise the learning of similarity between hash codes of mismatching data pairs. Owing to $\frac{1}{k}\boldsymbol{b}_i^{vT}\boldsymbol{b}_j^t \in [-1,1]$, we re-scale $\boldsymbol{s}_{ij}$ as $\hat{\boldsymbol{s}}_{ij} = 2\boldsymbol{s}_{ij} - 1 \in [-1,1]$.

Thus, based on these observations, we proposed a matching score based hashing loss:

$$\begin{aligned}\mathcal{L}_1 =&\frac{1}{|N_i^+|}\sum_{j\in N_i^+}(\frac{1}{k}\boldsymbol{b}_i^{vT}\boldsymbol{b}_j^t - 1)^2 \\ &+ \frac{1}{\sum\limits_{j\in N_i^-}\boldsymbol{I}_{ij}}\sum_{j\in N_i^-}\boldsymbol{I}_{ij}(\frac{1}{k}\boldsymbol{b}_i^{vT}\boldsymbol{b}_j^t - \hat{\boldsymbol{s}}_{ij})^2\end{aligned} \tag{9}$$

$$\boldsymbol{I}_{ij} = \begin{cases} 1, & \frac{1}{k}\boldsymbol{b}_i^{vT}\boldsymbol{b}_j^t > \hat{\boldsymbol{s}}_{ij}, \\ 0, & otherwise. \end{cases} \tag{10}$$

where N_i^+ denotes the set of text datapoints which are matching with the image $\boldsymbol{x}_i$, and N_i^- denotes the set of text datapoints which are not matching with the image $\boldsymbol{x}_i$; $\boldsymbol{s}_{ij}$ denotes the matching score between the image $\boldsymbol{x}_i$ and text $\boldsymbol{y}_j$;

It can be found that the first item of $\mathcal{L}_1$ is to make $\frac{1}{k}\boldsymbol{b}_i^{vT}\boldsymbol{b}_j^t$ be close to 1, i.e., make the Hamming distance between the hash codes of matching datapair be small. The second item of $\mathcal{L}_1$ is constructed to penalize the mismatching datapair that the Hamming similarity between their hash codes is larger than their matching score s_{ij}, i.e., the goal of $\mathcal{L}_1$ is to make the Hamming distance between their hash codes be large. Thus, by minimizing the hashing loss $\mathcal{L}_1$, the learned binary hash codes can coarsely preserve the heterologous matching relationship.

Furthermore, as the $sgn(\cdot)$ function is indifferentiable at zero and the derivation of it will be zeros for a non-zero input, the parameters of hashing model will not be updated with the back-propagation algorithm when minimizing the hashing loss function $\mathcal{L}_1$. Thus, we directly discard the $sgn(\cdot)$ function to ensure the parameters of our hashing model can be updated, and use $\tanh(\cdot)$ to approximate the $sgn(\cdot)$ function to make each element of output of hashing layer can be close to "+1" or "-1". Then the final hashing loss function can be formulated as follows:

$$\begin{aligned}\mathcal{L}_h =&\frac{1}{|N_i^+|}\sum_{j\in N_i^+}(\frac{1}{k}\hat{\boldsymbol{b}}_i^{vT}\hat{\boldsymbol{b}}_j^t - 1)^2 \\ &+ \frac{1}{\sum\limits_{j\in N_i^-}\boldsymbol{I}_{ij}}\sum_{j\in N_i^-}\boldsymbol{I}_{ij}(\frac{1}{k}\hat{\boldsymbol{b}}_i^{vT}\hat{\boldsymbol{b}}_j^t - \hat{\boldsymbol{s}}_{ij})^2.\end{aligned} \tag{11}$$

where $\hat{\boldsymbol{b}}_i^v = tanh(\mathcal{H}_x(\hat{\boldsymbol{v}}_i;\boldsymbol{\Theta}_v))$ and $\hat{\boldsymbol{b}}_j^t = tanh(\mathcal{H}_y(\hat{\boldsymbol{t}}_j;\boldsymbol{\Theta}_t))$

3.5 Inference

After training the image-text matching model and HEI module well, we can generate the hash codes $\{\boldsymbol{b_i^v}\}_{i=1}^N$ ($\{\boldsymbol{b_i^t}\}_{i=1}^M$) for all the images $\{\boldsymbol{x_i}\}_{i=1}^M$ (text $\{\boldsymbol{y_i}\}_{i=1}^N$) in the retrieval set using the HEI module. When given a query image $\boldsymbol{x}_q$ (text $\boldsymbol{y}_q$), we also use HEI module map it into a hash code $\boldsymbol{b_q^v}$ ($\boldsymbol{b_q^t}$), and calculate the Hamming distances between $\boldsymbol{b_q^v}$ ($\boldsymbol{b_q^t}$) and each code in $\{\boldsymbol{b_i^t}\}_{i=1}^M$ ($\{\boldsymbol{b_i^v}\}_{i=1}^N$). Then, we sort the texts (images) in the retrieval set in ascending order according to the Hamming distances, and select a few of the top texts (images) as the candidate set. Finally, we only need to do the fine-grained matching in the candidate set to find the matching datapoints.

4 Experiments

4.1 Datasets

We evaluate the performance of the proposed HEI module on two public used datasets: **Flickr30K** (Plummer et al., 2015) and **MS-COCO** (Lin et al., 2014). Specifically, Flickr30k contains 31783 images collected from the Flickr website. Each image is accompanied with five human annotated sentences descriptions. Following the setting of previous works (Wang et al., 2019; Liu et al., 2020), this dataset is split into 29,000 images, 1,000 images, and 1,000 images for training set, validation set, and testing set respectively. We report the performance evaluation of image-text retrieval on 1000 testing set. MS-COCO is another large-scale image-caption benchmark which consists of about 123,287 images with each image also roughly annotated with five sentence descriptions. Following the widely used split (Karpathy et al., 2014; Chen et al., 2020), we use 113,287 images for training, 1000 images for validation and 5000 images for testing. Moreover, we evaluate our method on both the 5 folds of 1K test images and the full 5K test images for MS-COCO.

Method	Text Retrieval				Image Retrieval			
	R@1	R@5	R@10	Time(s)	R@1	R@2	R@10	Time(s)
BFAN (Liu et al., 2019)	0.692	0.914	0.962	76.72	0.500	0.772	0.848	31.29
BFAN-random	0.582	0.814	0.883	37.00	0.291	0.419	0.451	17.00
BFAN-HEI	0.692	0.912	0.962	22.31	0.499	0.772	0.846	11.38
CAMP (Wang et al., 2019)	0.675	0.914	0.954	568.57	0.527	0.787	0.850	514.52
CAMP-random	0.599	0.823	0.873	295.61	0.318	0.436	0.461	280.06
CAMP-HEI	0.676	0.909	0.950	168.15	0.526	0.782	0.844	166.21
IMRAN (Chen et al., 2020)	0.710	0.920	0.963	1858.82	0.531	0.799	0.862	692.16
IMRAN-radom	0.594	0.842	0.892	956.44	0.314	o.432	0.465	353.75
IMRAN-HEI	0.710	0.920	0.964	574.93	0.532	0.797	0.858	219.72
GSMN (Liu et al., 2020)	0.733	0.918	0.964	518.32	0.524	0.792	0.863	146.70
GSMN-random	0.611	0.839	0.890	149.65	0.302	0.426	0.458	72.37
GSMN-HEI	0.734	0.919	0.967	99.30	0.524	0.790	0.860	51.02

Table 1: Comparison in terms of R@N scores and time cost of two retrieval tasks on Flickr30K

Method	Text Retrieval				Image Retrieval			
	R@1	R@5	R@10	Time(s)	R@1	R@2	R@10	Time(s)
BFAN (Liu et al., 2019)	0.753	0.961	0.989	62.84	0.610	0.896	0.954	32.45
BFAN-random	0.657	0.899	0.930	29.87	0.361	0.486	0.502	16.15
BFAN-HEI	0.753	0.962	0.989	18.21	0.610	0.896	0.953	12.24
CAMP (Wang et al., 2019)	0.711	0.953	0.977	567.86	0.581	0.882	0.948	515.18
CAMP-random	0.648	0.890	0.935	302.29	0.340	0.469	0.486	284.65
CAMP-HEI	0.713	0.953	0.977	169.24	0.580	0.882	0.946	168.39
IMRAN (Chen et al., 2020)	0.784	0.964	0.991	1808.01	0.644	0.912	0.960	662.59
IMRAN-random	0.688	0.904	0.935	868.64	0.370	0.484	0.496	332.95
IMRAN-HEI	0.784	0.964	0.991	504.64	0.644	0.912	0.960	215.15
GSMN (Liu et al., 2020)	0.758	0.960	0.992	465.75	0.607	0.897	0.955	205.89
GSMN-random	0.658	0.903	0.946	120.56	0.302	0.426	0.458	99.23
GSMN-HEI	0.758	0.960	0.992	80.38	0.607	0.894	0.953	70.69

Table 2: Comparison in terms of R@N scores and time cost of two retrieval tasks on MS-COCO 1K

4.2 Evaluation

Following the setting in (Chen et al., 2020; Liu et al., 2020), we evaluate the performance of our proposed approach by reporting Recall@K (K = 1, 5, 10) values for bi-directional matching tasks, i.e. matching texts given an image query (Text Retrieval) and matching images given a text query (Image Retrieval). The Recall computes the proportion of correct image or text being retrieved among top K results. In addition, we also record the inference time in seconds to evaluate the efficiency of our proposed HEI.

4.3 Baselines

To evaluate the performance of our proposed HEI, some state-of-the-art attention mechanism based image-text matching methods are selected as our baselines, including BFAN (Liu et al., 2019), CAMP (Wang et al., 2019), IMRAN (Chen et al., 2020) and GSMN (Liu et al., 2020). It should be noted that the proposed HEI focuses on exploring a novel and efficient hashing based efficient inference module that can be universally plugged into existing attention mechanism based image-text methods to speed up the inference speed rather than redesigning a new cross-modal attention mechanism to improve their matching performance.

4.4 Implementation Details

All our experiments are implemented in PyTorch and conducted on a NVIDIA Tesla V100 GPU. For representing visual modality, the amount of regions in each image is $m = 36$, and the dimensionality

Method	Text Retrieval				Image Retrieval			
	R@1	R@5	R@10	Time(s)	R@1	R@2	R@10	Time(s)
BFAN	0.499	0.795	0.888	9461.96	0.369	0.657	0.772	3797.68
BFAN-random	0.426	0.705	0.807	5521.66	0.231	0.381	0.428	2903.52
BFAN-HEI	0.496	0.791	0.884	3355.99	0.369	0.657	0.772	2177.12
CAMP	0.433	0.751	0.863	18646.21	0.324	0.633	0.753	16344.43
CAMP-random	0.361	0.661	0.763	11515.34	0.211	0.373	0.423	9615.76
CAMP-HEI	0.434	0.750	0.862	8198.13	0.323	0.632	0.751	6277.92
IMRAN	0.525	0.812	0.898	39030.95	0.391	0.684	0.795	16877.68
IMRAN-random	0.447	0.721	0.814	19423.71	0.244	0.390	0.435	8505.07
IMRAN-HEI	0.525	0.813	0.898	8265.47	0.390	0.683	0.794	3315.97
GSMN	0.494	0.793	0.888	25261.93	0.359	0.655	0.769	12226.86
GSMN-random	0.414	0.696	0.795	6453.45	0.229	0.383	0.435	7461.81
GSMN-HEI	0.493	0.793	0.888	3561.13	0.359	0.654	0.767	4607.29

Table 3: Comparison in terms of R@N scores and time cost of two retrieval tasks on MS-COCO 5K

of the final region representation vectors is set as 1024. Moreover, the dimensionality of hidden state (i.e., $\overrightarrow{\boldsymbol{h}}_i^j$ and $\overleftarrow{\boldsymbol{h}}_i^j$ in Formula (2)) in the GRU is also set as 1024. The length of hash codes is set as 64. In the training phase, we first train the base cross-modal attention module for 20 epochs, then train the HEI module jointly. We adopt SGD with a mini-batch size of 128 and a learning rate within 10^{-2} to 10^{-3} to optimize the HEI modul. The optimization algorithm for the base cross-modal attention module is the same with the ones defined in the original method, for example, when plugging HEI module into GSMN, the optimization algorithm for cross-modal attention module is Adam.

4.5 Main results

We conduct extensive experiments on Flickr30K and MS-COCO. The image-text matching results on Flickr30K, MS-COCO dataset with 1K test points and 5K test points are shown in Table 1, 2 and 3, respectively. ”method”-HEI denotes the method using the proposed HEI module, for example, BFAN-HEI means plugging HEI into BFAN to speed up the inference speed. Similarly, ”method”-random denotes randomly selecting 50% datapoints from retrieval set as the candidate set to speed up the inference speed of the method.

Based on the results shown in these tables, the following observations can be got: (1) our proposed HEI module can greatly improve the matching efficiency of all the four baselines almost without reducing the matching performance, and even slightly improve the performance of some baselines. For example, as shown in Table 1, comparing GSMN-HEI with GSMN, GSMN-HEI achieves an increase of 0.1% on the R@1 metric in the text retrieval task, and greatly reduces the inference time from 518.32 seconds to 99.30 seconds. The reason why plugging the HEI module can slightly improve the performance maybe that for some query, there are some false positive datapoints which can misguide the image-text model, but the Hamming distance between the hash codes of queries and the ones of false positive datapoints are large, i.e., the false positive datapoints will not be selected as the candidate points. Thus, without the effect of the false positive datapoints, the image-text model can find the matching points successfully and improve the retrieval performance. (2) The proposed HEI module can map datapoints into hash codes with the original heterologous matching relationship coarsely preserved. For instance, as shown in Table 1, 2 and 3, all the methods with the proposed HEI module achieve not only better performance than the methods with the randomly selected candidate, but also lower inference time. It means that the number of datapoints in the candidate set selected by our proposed HEI module is smaller but the possibility of the candidate set containing the matching datapoint is higher.

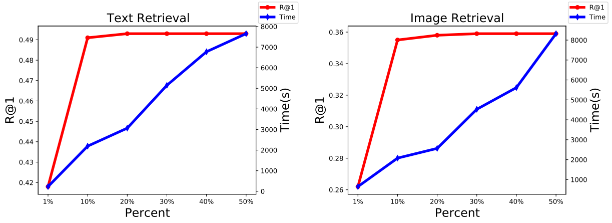
4.6 Discussion

4.6.1 Ablation study

To further investigate the impact of the length of hash codes, we construct there variants of HEI module with code length being 16, 32, and 128 bits with two baselines on Flickr30K, respectively. The re-

Method	Text Retrieval				Image Retrieval			
	R@1	R@5	R@10	Time(s)	R@1	R@2	R@10	Time(s)
BFAN-HEI$_{16bits}$	0.687	0.903	0.958	20.85	0.485	0.742	0.810	11.02
BFAN-HEI$_{32bits}$	0.685	0.907	0.957	21.03	0.498	0.765	0.838	11.17
BFAN-HEI$_{128bits}$	0.690	0.913	0.961	21.46	0.500	0.771	0.846	11.55
BFAN-HEI	0.692	0.912	0.962	21.31	0.499	0.772	0.846	11.38
GSMN-HEI$_{16bits}$	0.719	0.911	0.956	99.13	0.506	0.756	0.821	50.84
GSMN-HEI$_{32bits}$	0.731	0.916	0.966	99.22	0.519	0.779	0.847	50.92
GSMN-HEI$_{128bits}$	0.734	0.919	0.965	99.51	0.520	0.788	0.855	51.23
GSMN-HEI	0.734	0.919	0.967	99.31	0.524	0.790	0.860	51.02

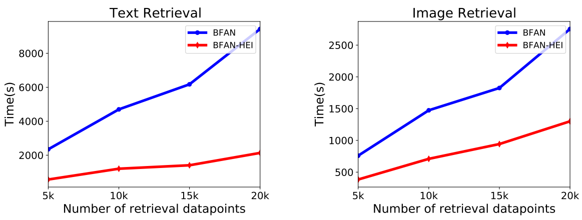
Table 4: Comparison in terms of R@N scores and time cost of two retrieval tasks on Flickr30K



(a)

(b)

Figure 2: The Figure 2(a) and (b) denotes the results of BFAN-HEI on MS-COCO(5k) in text retrieval task and image retrieval task, respectively. Moreover, in each figure, the axis X denotes selecting how much percentage of points in the retrieval set as candidate set, and the axis Y for the red line is in the left which is the value of R@1, and the axis Y for the blue line is in the right which denotes the inference time taken for the transaction in seconds.



(a)

(b)

Figure 3: Compare the inference time of BFAN and the one of BFAN-HEI in large retrieval set under the condition that BFAN-HEI achieves the same performance of BDAN

sults are shown in Table 4. From these results, it can be found: (1) The length of hash codes rarely influence the inference time that each baseline with a different hash code length of HEI consumes nearly the same inference time. This is because that the speed of the “XOR” operation between hash codes is far faster than the ones of the cross-modal attention mechanism. Thus, it implicitly demonstrates the availability of speeding up the inference speed of baselines by using the proposed HEI to fast select the candidate set. (2) The matching performance first increases as the hash code length varies from 16 to 64, and then tend to be stable when the length varies from 64 to 128. Thus, for the other experiments, the hash code length is set as 64.

4.6.2 Efficiency and performance

We also conduct experiments to further investigate the trade-off between inference efficiency and matching performance.

As the results shown in Figure 2, with the size of candidate increasing, the matching performance of BFAN-HEI (the red lines) increase rapidly and then tend to stable, and BFAN-HEI consumes more inference time (the blue lines). It can be found that when selecting only 20% of datapoints in the retrieval set as the candidate set by the proposed HEI module, BFAN-HEI can already achieve the best performance, and greatly reduce the inference time. Thus, it demonstrates the effectiveness of our proposed HEI module.

4.6.3 Scalability for the large retrieval set

To further investigate the scalability of the proposed HEI module for the large retrieval set, when doing experiments on the MS-COCO (1K) with the BFAN baseline, we directly use training data to expand the data volume of the retrieval set. The curves of inference time w.r.t. the volume of retrieval set are shown in Figure 3. It can be found that with volume of the retrieval set increasing, our proposed HEI module can still be used to speed up the inference speed without reducing the matching performance, which demonstrates the scalability of our proposed HEI module for large retrieval sets.

5 Conclusion

In this paper, we have proposed a novel Hashing based Efficient module, called HEI, which can be plugged into the existing image-text matching methods to speed up the inference step without reducing the matching performance. Extensive experiments on two widely used benchmark MS-COCO and Flickr30k with four baseline methods demonstrate the efficiency and effectiveness of our proposed HEI module.

Acknowledgments

The work is supported by National Key RD Plan (No.2018YFB1005100), National Natural Science Foundation of China (No. 61751201, 61602197 and 61772076), Natural Science Fund of Beijing (No. Z181100008918002) and the funds of Beijing Advanced Innovation Center for Language Resources (No. TYZ19005).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12655–12663.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hengtong Hu, Lingxi Xie, Richang Hong, and Qi Tian. 2020. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3123–3132.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27:1889–1897.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216.

Qiubin Lin, Wenming Cao, Zhihai He, and Zhiquan He. 2020. Semantic deep cross-modal hashing. *Neurocomputing*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 3–11.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2017. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1881–1889.

Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1928–1937.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Yufeng Shi, Xinge You, Feng Zheng, Shuo Wang, and Qinmu Peng. 2019. Equally-guided discriminative hashing for cross-modal retrieval. In *IJCAI*, pages 4767–4773.

Shupeng Su, Zhisheng Zhong, and Chao Zhang. 2019. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3027–3035.

Rong-Cheng Tu, Xian-Ling Mao, Bing Ma, Yong Hu, Tan Yan, Wei Wei, and Heyan Huang. 2020. Deep cross-modal hashing with hashing functions and unified hash codes jointly learning. *IEEE Transactions on Knowledge and Data Engineering*.

Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. 2018. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7190–7198.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5764–5773.

Wei Zhang, Bairui Wang, Lin Ma, and Wei Liu. 2019. Reconstruct and represent video contents for captioning via reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–701.

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23.