Adaptive Offline Quintuplet Loss for Image-Text Matching

Tianlang Chen[1][0000−0002−6355−6474], Jiajun Deng[2], and Jiebo Luo[1]

[1] University of Rochester, {tchen45,jluo}@cs.rochester.edu,
[2] Uiversity of Science and Technology of China, {djiajun1206}@gmail.com

Abstract. Existing image-text matching approaches typically leverage triplet loss with online hard negatives to train the model. For each image or text anchor in a training mini-batch, the model is trained to distinguish between a positive and the most confusing negative of the anchor mined from the mini-batch (*i.e.* online hard negative). This strategy improves the model's capacity to discover fine-grained correspondences and non-correspondences between image and text inputs. However, the above approach has the following drawbacks: (1) the negative selection strategy still provides limited chances for the model to learn from very hard-to-distinguish cases. (2) The trained model has weak generalization capability from the training set to the testing set. (3) The penalty lacks hierarchy and adaptiveness for hard negatives with different "hardness" degrees. In this paper, we propose solutions by sampling negatives offline from the whole training set. It provides "harder" offline negatives than online hard negatives for the model to distinguish. Based on the offline hard negatives, a quintuplet loss is proposed to improve the model's generalization capability to distinguish positives and negatives. In addition, a novel loss function that combines the knowledge of positives, offline hard negatives and online hard negatives is created. It leverages offline hard negatives as the intermediary to adaptively penalize them based on their distance relations to the anchor. We evaluate the proposed training approach on three state-of-the-art image-text models on the MS-COCO and Flickr30K datasets. Significant performance improvements are observed for all the models, proving the effectiveness and generality of our approach. Code is available at https://github.com/sunnychencool/AOQ.

Keywords: Image-text matching, Triplet loss, Hard negative mining

1 Introduction

Image-text matching is the core task in cross-modality retrieval to measure the similarity score between an image and a text. By image-text matching, a system can retrieve the top corresponding images of a sentence query, or retrieve the top corresponding sentences of an image query.

To train an image-text matching model to predict accurate similarity score, triplet loss is widely used [23,5,6,15,14]. Each given image or text of a training mini-batch is referred to as an *anchor*. For each image/text anchor, a text/image

that corresponds to the anchor is called a *positive* while one that does not correspond to the anchor is called a *negative*. The anchor and its positives/negatives belong to two modalities. A triplet loss is applied to encourage the model to predict higher similarity scores between the anchor and its positives (*i.e.* positive pairs) than those between the anchor and its negatives (*i.e.* negative pairs).

To utilize negative pairs to train the model, early approaches [23,5,10] adopt an all-in strategy. For each anchor, all its negatives in the mini-batch participate in the loss computing process. However, in most situations, the semantic meanings of an anchor and its negatives are totally different. With this strategy, the overall training difficulty is relatively low for the model to distinguish between positive and negative pairs. The model only needs to focus on each pair's global semantic meaning difference and may ignore the local matching details. Faghri et al. [6] propose a triplet loss with online hard negatives (*i.e.* online triplet loss) as a more effective training approach. Specifically, for each anchor in a mini-batch, the model computes its similarity score to all the negatives in the same mini-batch online, and selects the negative with the highest score to the anchor as online hard negative of the anchor. The new triplet loss guides the model to only distinguish between the positives and online hard negatives of the anchor. Compared with the all-in strategy, the models trained by this approach commonly achieve better performance in distinguishing between positives and confusing negatives that have similar semantic meanings to the anchor. This training approach is employed by all the state-of-the-art models [15,14,18,27].

Even with its effectiveness, we argue that the online triplet loss still have three drawbacks in negative selection strategy, distinguishing strategy, and penalization strategy: (1) for the negative selection strategy, the "hardness" degree of online hard negatives is still not sufficient. Given the MS-COCO dataset as example, the training set contains 500K corresponding image-text pairs. When we set the mini-batch size to 128 as in [15,14,18,27], for each online hard negative of an anchor mined from the mini-batch, we can prove that its similarity score rank expectation to the anchor in the whole training set is about 4000 (*i.e.* $\frac{500K}{128}$). The probability of its rank in the top-100 is only about 2.2%. In other words, a very hard negative with a top-100 similarity score rank for the anchor will rarely be sampled to train the model. This decreases the model's capacity to distinguish between the positives and those very confusing negatives. Increasing the mini-batch size could be helpful. However, the mini-batch computational complexity grows sharply. (2) For the distinguishing strategy, the triplet loss only focuses on obtaining the correct rank orders between the positives and negatives of the same anchor. However, it does not guide the model to rank among positive pairs and negative pairs that contain no common samples. Actually, this guidance is essential to improve the model's generalization capability from training to testing, especially when we apply the guidance on the very hard negative pairs. (3) For the penalization strategy, the triplet loss lacks a hierarchy. Ideally, the loss function should guide the model to maintain remarkable score gaps among the pairs of different classes. For example, the positive pairs should obtain far higher similarity scores than very hard negative pairs, and the very
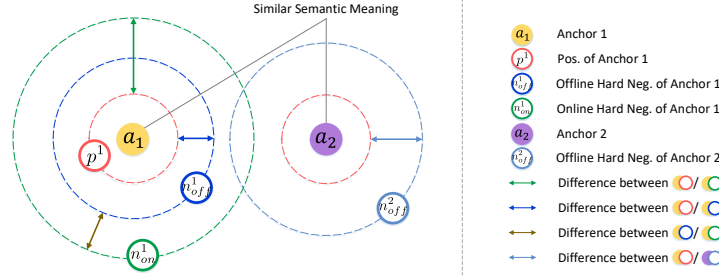
Fig. 1. Overview of the proposed training approach. For each anchor, we sample its positives, offline hard negatives and online hard negatives. The training approach gives adaptive penalties to enlarge the similarity score differences among positive pairs, offline hard negative pairs and online hard negative pairs (*i.e.* the blue, green and brown arrows). On the other hand, extra penalties are added to enlarge the similarity score difference between positive pairs and offline hard negative pairs with different anchors that share similar semantic meanings (*i.e.* the cyan arrow).

hard negative pairs should also obtain far higher similarity scores than ordinary hard negative pairs. When a pair's predicted score is close or beyond the boundary of its pair class, the loss function should give it a larger penalty to update the model. However, the current online triplet loss only defines positive and online hard negative pairs. More importantly, it gives an equal penalty to all the pairs when the margin conditions are not satisfied.

To overcome the above drawbacks, we propose a new training approach that can be generally applied on *all* existing models. Specifically, we utilize a two-round training to additionally sample "harder" negatives offline. In the first round, we train the model by the original online triplet loss. After that, for each image and text anchor in the training set, the model predicts its similarity score to all its negatives in the training set and ranks them. In the second round, given each anchor in a mini-batch, we sample its offline hard negatives directly from its top negative list with the highest similarity score in the whole training set. In this process, multiple kinds of offline hard negative pairs are constructed which share/do not share common elements with the positive pairs. The model is trained by a combination of online triplet loss and offline quintuplet loss to overcome the first two drawbacks successfully. Furthermore, we modify the loss function and feed information of offline hard negative pairs into the online triplet loss term. The complete training loss achieves hierarchical and adaptive penalization for the positive pairs, offline hard negative pairs, and online hard negative pairs with different "hardness" degrees. The framework of the proposed training approach is shown in Figure 1.

Our main contributions are summarized as follows:

– We propose a novel and general training approach for image-text matching models. A new offline quintuplet loss is introduced that can effectively cooperate with the original online triplet loss.

- We skillfully feed the similarity score of offline hard negative pair into online loss term. It serves as a criterion to adaptively penalize different kinds of pairs. We analyze how it works mathematically.
- We evaluate our training approach on three state-of-the-art image-text matching models. Quantitative and qualitative experiments conducted on two publicly available datasets demonstrate its strong generality and effectiveness.

2   Related Work

Image-text matching has received much attention in recent years. Most of the previous works focus on the improvement of feature extraction and model design. Early image-text matching approaches [7,13,6,35] directly capture the visual-textual alignment at the level of image and text. Typically, they extract the global image feature by convolutional neural network (CNN), and extract the global text feature by language model such as Skip-gram model [22] or recurrent neural network (RNN). The image-text similarity score is then computed as the inner product [7,13,6] or cosine similarity [35] of the image and text features. The success of attention models for joint visual-textual learning tasks, such as visual question answering (VQA) [34,21,30,12] and image captioning [29,20,31,24,3], leads to the transition to capture image-text correspondence at the level of image regions and words [10,16,23,36]. Typically, these approaches extract the image region feature and word feature from the last pooling layer of CNN and temporal outputs of RNN. They focus on designing effective upper networks that can automatically find, align and aggregate corresponding regions and words to compute the final similarity score. Recently, Anderson et al. [1] extract the image object features by the combination of Faster R-CNN [25] and ResNet [8] for VQA. Based on [1], recent approaches [14,15,18,27,11] further construct the connection between words and image objects. They either propose new mechanisms for object feature extraction, such as feeding saliency information [11] or extracting joint features among objects by constructing object graph [15], or propose different cross-modality aggregation networks [14,27,18,2,9] to improve the aggregation process from object and word features to the final score.

Even though the network design is widely studied, relatively fewer works focus on the training approach. Early image-text matching approaches [7,13,5,32] commonly apply a standard triplet loss whose early form can be found in [28] for word-image embedding. On the other hand, Zhang et al. [35] improve the triplet loss and propose a norm-softmax loss to achieve cross-modal projection. For both losses, all the negatives of an anchor in the same mini-batch are utilized for loss computing. Significant improvement is observed as Faghri et al. [6] propose the triplet loss with online hard negatives. Online triplet mining is first introduced in [26] for face recognition. For image-text matching, it mines the online hard negatives of the anchors from the mini-batch and makes the model only pay attention to these confusing negatives. Almost all the current models [15,14,18,27] apply this online triplet loss. To the best of our knowledge, our work is the first that introduces offline hard negatives for image-text matching.
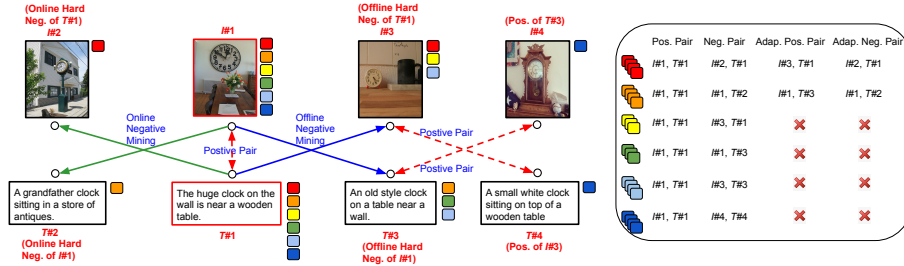
Fig. 2. Training process illustration. Given a positive image-text pair $(I\#1, T\#1)$, 6 margin-based ranking losses are applied to enlarge its similarity score differences from the online hard negative pairs $(I\#2, T\#1)$, $(I\#1, T\#2)$, the offline hard negative pairs $(I\#3, T\#1)$, $(I\#1, T\#3)$ (with the common anchor), and the derived offline hard negative pairs $(I\#3, T\#3)$, $(I\#4, T\#4)$ (without the common anchor). Adaptive penalization is imposed via the online losses to adaptively penalize positive and negative pairs with different strengths and directions. The involved samples of each loss are marked by the corresponding squares.

They are mined offline from the whole training set. Motivated by [4] for person re-identification, we propose a quintuplet loss based on offline hard negatives to effectively cooperate with an online triplet loss, leading to significant improvement. It should be noticed that Liu et al. [19] explicitly feed adaptive penalty weight into triplet loss for image-text matching. However, they use it to solve the hubness problem, while we implicitly feed hierarchical information into the model to enlarge the similarity score differences among different pair classes.

3   Methods

In this section, we formally present our training approach for image-text matching. In Section 3.1, we introduce the margin-based standard and online triplet losses that are used in previous works. In Section 3.2, we present offline quintuplet loss as an effective complement to online triplet loss to significantly improve the performance. In Section 3.3, we propose our final loss function with adaptive penalization and mathematically show how it works. The overall training process and the involved pairs are illustrated in Figure 2.

3.1   Triplet Loss for Image-text Matching

Given an input image-text pair, image-text matching models aim to predict the pair's similarity score as a criterion for cross-modality retrieval. To achieve this, positive pairs (*i.e.* corresponding image-text pairs) and negative pairs (*i.e.* non-corresponding image-text pairs) are constructed. The model is trained to predict higher similarity score for the positive pairs than the negative ones.

Because the metrics of cross-modality retrieval are based on the ranking performance of multiple candidates on a single query, triplet loss is widely applied

to train the model. It holds a common sample for each positive pair and negative pair as an *anchor*. The other sample in the positive pair is called the anchor's *positive* while the other sample in the negative pair is called the anchor's *negative*. In essence, triplet loss encourages the model to predict higher similarity scores from the anchor to its positives. This is consistent with the retrieval process of finding the corresponding candidates of a query with the high similarity scores.

Early image-text matching works [7,13,5,32] typically apply a standard triplet loss without hard negative mining. Given a training mini-batch that contains a set of positive pairs, the standard triplet loss is defined as:

$$\mathcal{L}_{std} = \sum_{(i,t)\in P} (\sum_{\bar{t}\in T/t} [\gamma - S(i,t) + S(i,\bar{t})]_+ + \sum_{\bar{i}\in I/i} [\gamma - S(i,t) + S(\bar{i},t)]_+) \quad (1)$$

Here γ is the margin of the triplet loss, $[x]_+ \equiv max(x,0)$. I, T and P are the image, text and positive pair sets of the mini-batch, respectively. i and t are the anchors of the two terms, respectively. (i,t) represents the positive pair, while $(i,\bar{t})$ and $(\bar{i},t)$ represent the negative pairs available in the mini-batch.

On the other hand, to overcome the drawback of standard triplet loss mentioned in Section 1, Faghri et al. [6] present triplet loss with online hard negatives (*i.e.* online triplet loss). In particular, for a positive pair (i,t) in a mini-batch, the hard negatives of the anchor i and t are given by $\bar{t}_{on} = argmax_{c\in T/t}S(i,c)$ and $\bar{i}_{on} = argmax_{b\in I/i}S(b,t)$, respectively. The online triplet loss is defined as:

$$\mathcal{L}_{online} = \sum_{(i,t)\in P} ([\gamma - S(i,t) + S(i,\bar{t}_{on})]_+ + [\gamma - S(i,t) + S(\bar{i}_{on},t)]_+) \quad (2)$$

Compared with the standard triplet loss, online triplet loss forces the model to only learn to distinguish between the positive and the most confusing negative of an anchor in the mini-batch. This guides the model to not only consider the overall semantic meaning difference of a pair, but also discover correspondences and non-correspondences from the details hidden in local regions and words.

3.2   Offline Quintuplet Loss

One problem of online triplet loss in Section 3.1 is that the "hardness" degree of most online hard negatives is still not sufficient, especially when the training involves a large-scale training set and a relatively small batch size. As mentioned in Section 1, the rank of an anchor's online hard negative in the whole training set is commonly not very high. Qualitatively, as shown in Figure 3, the online hard negatives of an anchor typically contain a few related words, objects or scenes to the anchor. However, there exist obvious non-correspondences between the anchor and the negatives. Indeed, the model only needs to find these non-correspondences and strengthen their influence, which is sufficient for the score difference between the positive pair and negative pair to exceed the margin γ in Equation 2. However, during inference, when the model encounters "harder"
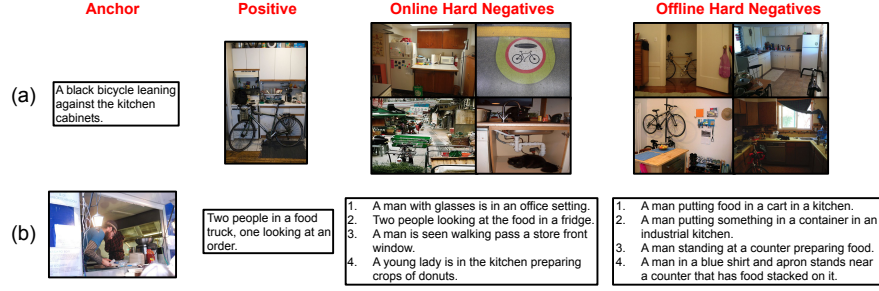
Fig. 3. Two example anchors, their corresponding positives, their sampled online hard negatives and offline hard negatives.

negatives like the offline hard negative examples of Figure 3, the model may not be able to distinguish them from the positives. The non-corresponding parts of these "harder" negatives to the anchor are subtle, and their influence on the predicted score can be offset by the perfectly corresponding parts.

To overcome the problem, we additionally mine "harder" negatives in an offline fashion. In particular, it involves a two-round training. In the first round, the model is trained by the online triplet loss. After that, it performs global similarity score prediction – for each image/text in the training set, the model predicts its similarity score to all its non-corresponding texts/images in the training set, ranks them by their scores and stores the list of the top-h. In the second round, for each anchor in a mini-batch, its offline hard negatives are uniformly sampled from the top-h negatives of the anchor in the whole training set. The model is trained from scratch again by the following loss function:

$$
\mathcal{L} = \sum_{(i,t) \in P} (([\gamma_1 - S(i,t) + S(i,\bar{t}_{on})]_+ + [\gamma_2 - S(i,t) + S(i,\bar{t}_{off})]_+) \\
+ ([\gamma_1 - S(i,t) + S(\bar{i}_{on},t)]_+ + [\gamma_2 - S(i,t) + S(\bar{i}_{off},t)]_+))
\tag{3}
$$

Here $\bar{t}_{off}$ and $\bar{i}_{off}$ are the offline hard negatives of i and t, γ_1 and γ_2 are the margins of the online and offline triplet losses. It should be noticed that for models with relatively low inference speed, the above mentioned global similarity score prediction step can be time-consuming. In Section 4, we demonstrate that a model can safely utilize the prediction of another efficient model to mine offline hard negatives, which still sharply benefits the training process.

Because the offline hard negatives are very confusing, to make them benefit the training, we should set γ_2 to a lower margin than γ_1, *e.g.* 0. However, in this situation, if the positive and offline hard negative pairs share a same anchor, the model will merely learn how to find the subtle non-corresponding parts of the offline hard negative pair, but still does not learn how to deal with the situation when the negative pair's perfect matching parts offset the score influence of non-corresponding parts. We attribute it to the fact that the positive and offline hard negative get close similarity score for their corresponding parts to the same

anchor. The model only needs to find the non-corresponding parts of the negative pair to satisfy the margin condition of γ_2. Also, as claimed in [4], this setting weakens the model's generalization capability from training to testing.

Considering this, we additionally derive two offline hard negative pairs and modify Equation 3 for the second-round training as follows:

$$
\mathcal{L} = \sum_{(i,t)\in P} (([\gamma_1 - S(i,t) + S(i,\bar{t}_{on})]_+ + [\gamma_2 - S(i,t) + S(i,\bar{t}_{off})]_+ + [\gamma_2 - S(i,t) + S(\bar{i}_{off},\bar{t}_{off})]_+)
$$
$$
+ ([\gamma_1 - S(i,t) + S(\bar{i}_{on},t)]_+ + [\gamma_2 - S(i,t) + S(\bar{i}_{off},t)]_+ + [\gamma_2 - S(i,t) + S(\widetilde{\bar{i}}_{off},\widetilde{\bar{t}}_{off})]_+)) \tag{4}
$$

Here $\widetilde{\bar{i}}_{off}$ and $\widetilde{\bar{t}}_{off}$ are the corresponding image and text of $\bar{t}_{off}$ and $\bar{i}_{off}$, respectively. Because $\bar{t}_{off}$ and $\bar{i}_{off}$ are offline hard negatives of corresponding i and t, both $(\bar{i}_{off},\bar{t}_{off})$ and $(\widetilde{\bar{i}}_{off},\widetilde{\bar{t}}_{off})$ can be also regarded as offline hard negative pairs (we re-sample $\bar{i}_{off}$ and $\bar{t}_{off}$ if they occationally correspond to each other). The samples of each pair are non-corresponding but share very similar semantic meanings to each other, and also to i and t. This two new terms guide the model to distinguish between positive and negative pairs without common elements. In Section 4, we prove the effectiveness of deriving the new terms based on $\bar{i}_{off}$, $\bar{t}_{off}$ instead of $\bar{i}_{on}$, $\bar{t}_{on}$. The complete offline loss terms based on anchor i and t contain 4 and 5 elements. Following [4], we define it as an offline quintuplet loss.

3.3   Adaptive and Hierarchical Penalization

In Section 3.2, we introduce offline hard negatives which cooperate with online hard negatives to train the model as Equation 4. During the training process, it is natural that we should give different penalty weights to negative pairs with different "hardness" degrees. For example, if the similarity score between a positive pair and a hard negative pair is close, both pairs should obtain higher penalty weight which guides the model to distinguish between them better. However, when we derive each loss term with respect to its contained pairs' similarity scores, the gradients are always constant. This indicates that when the margin condition is not satisfied, the penalty weight is consistent regardless of the closeness degree between the positive and negative pairs.

One simple solution is modifying each loss term to a form of square so that the penalty weight is related to the score difference between the positive and negative pairs. However, we find that the improvement is limited as there are no hierarchical knowledge provided by the loss function. Ideally, we expect that the positive pairs to obtain higher scores than offline hard negative pairs, and that the offline hard negative pairs obtain higher scores than online hard negative pairs. To this end, we feed the information of offline hard negatives into the online loss term. The final loss function for the second-round training is as follows:

$$
\mathcal{L} = \sum_{(i,t)\in P} (((\beta - \frac{S(i,\bar{t}_{off}) - S(i,\bar{t}_{on})}{\alpha})[\gamma_1 - S(i,t) + S(i,\bar{t}_{on})]_+ + [\gamma_2 - S(i,t) + S(i,\bar{t}_{off})]_+ + [\gamma_2 - S(i,t) + S(\bar{i}_{off},\bar{t}_{off})]_+)
$$
$$
+ ((\beta - \frac{S(\bar{i}_{off},t) - S(\bar{i}_{on},t)}{\alpha})[\gamma_1 - S(i,t) + S(\bar{i}_{on},t)]_+ + [\gamma_2 - S(i,t) + S(\bar{i}_{off},t)]_+ + [\gamma_2 - S(i,t) + S(\widetilde{\bar{i}}_{off},\widetilde{\bar{t}}_{off})]_+))
$$
$$
\tag{5}
$$

Here α and β are hyper-parameters. In Section 4, we present that they can be set to consistent values for different models on different datasets.

To better understand how the proposed loss function works, we focus on the first part (line) of Equation 5 which is symmetrical to the second part, and compute its gradient with respect to $S(i,t)$, $S(i,\bar{t}_{off})$ and $S(i,\bar{t}_{on})$ as follows:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial S(i,t)} &= (\frac{S(i,\bar{t}_{off}) - S(i,\bar{t}_{on})}{\alpha} - \beta)\mathbb{I}(\gamma_1 - S(i,t) + S(i,\bar{t}_{on}) > 0) - \mathbb{I}(\gamma_2 - S(i,t) + S(i,\bar{t}_{off}) > 0) \\
&\quad - \mathbb{I}(\gamma_2 - S(i,t) + S(\bar{i}_{off},\bar{t}_{off}) > 0), \\
\frac{\partial \mathcal{L}}{\partial S(i,\bar{t}_{off})} &= (\frac{S(i,t) - S(i,\bar{t}_{on})}{\alpha} - \frac{\gamma_1}{\alpha}))\mathbb{I}(\gamma_1 - S(i,t) + S(i,\bar{t}_{on}) > 0) + \mathbb{I}(\gamma_2 - S(i,t) + S(i,\bar{t}_{off}) > 0), \\
\frac{\partial \mathcal{L}}{\partial S(i,\bar{t}_{on})} &= (\frac{2S(i,\bar{t}_{on}) - S(i,t) - S(i,\bar{t}_{off})}{\alpha} + \beta + \frac{\gamma_1}{\alpha})\mathbb{I}(\gamma_1 - S(i,t) + S(i,\bar{t}_{on}) > 0)
\end{aligned}
\tag{6}
$$

Here $\mathbb{I}(A)$ is the indicator function: $\mathbb{I}(A) = 1$ if A is true, and 0 otherwise.

When the margin conditions are not satisfied, the gradient of $\mathcal{L}$ with respect to $S(i,\bar{t}_{on})$ becomes larger when $S(i,\bar{t}_{on})$ is close to the average of $S(i,\bar{t}_{off})$ and $S(i,t)$, which indicates a larger penalty to make $S(i,\bar{t}_{on})$ lower. For the gradient of $\mathcal{L}$ with respect to $S(i,t)$, the second and third terms indicate a negative constant which pushes $S(i,t)$ to be higher than $S(i,\bar{t}_{off})$. In addition, the first term indicates an additional adaptive penalty for $S(i,t)$ to be far away from $S(i,\bar{t}_{on})$. When $S(i,\bar{t}_{on})$ is remarkably lower than $S(i,\bar{t}_{off})$, the penalty drops since $S(i,\bar{t}_{on})$ is sufficiently lower. As for the gradient of $\mathcal{L}$ with respect to $S(i,\bar{t}_{off})$, it is subtle as the second term indicates a positive constant that penalizes $S(i,\bar{t}_{off})$ to be lower than $S(i,t)$. However, this penalty could be neutralized when $S(i,t)$ and $S(i,\bar{t}_{on})$ are close to each other. In this situation, it prevents the penalty from incorrectly making $S(i,\bar{t}_{off})$ lower than $S(i,\bar{t}_{on})$.

Overall, the proposed loss function applies adaptive and hierarchical penalties to the positive, offline hard negative and online hard negative pairs based on the differences among their predicted scores. Essentially, the pairs that are close to the boundary of its pair class obtain larger penalty weights, the inter-class score gaps can thus be enlarged among these three kinds of pairs. In Section 4, we demonstrate its strong effectiveness to improve the model's performance.

4   Experiments

Extensive experiments are performed to evaluate the proposed training approach. The performance of retrieval is evaluated by the standard recall at K (R@K). It is defined as the fraction of queries for which the correct item belongs to the top-K retrieval items. We first present the datasets, experiment settings and implementation details. We then compare and analyze the performance of the proposed approach with others quantitatively and qualitatively.

4.1   Dataset and Experiment Settings

We evaluate our model on two well-known datasets, MS-COCO and Flickr30K. The original MS-COCO dataset [17] contains 82,783 training and 40,504 validation images. Each image is annotated with five descriptions. Following the splits

of [18,14,15], we divide the dataset into 113,283 training images, 5,000 validation images and 5,000 test images. Following [6,14,15], we report the results by averaging over 5 folds of 1K test images or testing on the full 5K test images. Flickr30k [33] consists of 31K images collected from the Flickr website. Each image also corresponds to five human-annotated sentences. Following the split of [18,14,15], we randomly select 1,000 images for validation and 1,000 images for testing and use other images to train the model.

To evaluate the effectiveness and generality of the proposed approach, we apply it to the following current state-of-the-art image-text matching models:

– **SCAN** [14]. The first model that captures image-text correspondence at the level of objects and words. The word and object features are extracted by bidirectional GRU and the combination of Faster R-CNN [25] and ResNet-101 [8], respectively. Stacked cross attention is fed into the network to discover the full latent alignments using both objects and words as context.
– **BFAN** [18]. A novel Bidirectional Focal Attention Network based on SCAN that achieves remarkable improvement. Compared with SCAN, it focuses additionally on eliminating irrelevant fragments from the shared semantics.
– **VSRN** [15]. The current state-of-the-art image-text matching models without leveraging extra supervision (the model in [11] is trained by extra saliency-annotated data). It generates object representation by region relationship reasoning and global semantic reasoning.

All the three models are originally trained by triplet loss with online hard negatives. We replace it with the proposed training approach for comparison.

4.2   Implementation Details

To perform a fair comparison, for SCAN, BFAN and VSRN, we completely preserve their network structures and model settings (*e.g.* training batch size, feature dimension and other model-related hyper-parameter settings) as described in their original work. We only replace the online triplet loss by the proposed one to train them. For all the situations, the margins for online and offline ranking losses γ_1 and γ_2 are set to 0.2 and 0, the hyper-parameters β and α in Equation 5 are set to 1.5 and 0.3. The top list size h is set to 300 and 60 to sample offline hard negative texts and images (the training texts are 5 times as many training images for both datasets). As mentioned in Section 3.2, for VSRN, it takes 3,400s/620s to perform global similarity score prediction on MS-COCO/Flickr30K. However, for SCAN and BFAN, they hold complex upper networks which make this step extremely time-consuming. Therefore, we skip the first-round training of SCAN and BFAN. The similarity scores predicted by VSRN are also used as a basis for the second-round training of SCAN and BFAN to sample offline hard negatives. We consider this setting valid because, after the second-round training, the final prediction is still made by SCAN or BFAN without the participating of VSRN, which can be regarded as a teacher model. For the first-round training on MS-COCO/Flickr30K, as [15], VSRN is trained by a start learning rate of 0.0002 for

15/10 epochs, and then trained by a lower learning rate of 0.00002 for another 15/10 epochs. For the second-round training on both datasets, SCAN, BFAN and VSRN are trained by a start learning rate of 0.0005, 0.0005 and 0.0002 for 10 epochs, and then trained by a lower learning rate of 0.00005, 0.00005 and 0.00002 for another 5, 5 and 10 epochs, respectively.

4.3 Results on MS-COCO and Flickr30K

Table 1. Quantitative evaluation results of image-to-text (sentence) retrieval and text-to-image (image) retrieval on MS-COCO 1K/5K test set. The baseline models (first row) are trained by the triplet loss with online hard negatives. "+ OffTri", "+ OffQuin", "+ AdapOffQuin" represent training the model by Equation 3, 4, 5, respectively.

Model	Sentence Retrieval			Image Retrieval			
	R@1	R@5	R@10	R@1	R@5	R@10	rsum
1K Test Images							
SCAN [14]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
SCAN + OffTri	73.1	94.8	98.2	59.3	88.3	94.8	508.5
SCAN + OffQuin	73.6	95.0	98.4	59.6	**88.6**	**95.0**	510.2
SCAN + AdapOffQuin	**74.1**	**95.2**	**98.5**	**59.8**	88.6	95.0	**511.2**
BFAN [18]	74.9	95.2	98.3	59.4	88.4	94.5	510.7
BFAN + OffTri	75.8	95.6	98.4	60.1	88.8	94.7	513.4
BFAN + OffQuin	76.3	95.7	98.4	60.5	89.0	94.8	514.7
BFAN + AdapOffQuin	**77.3**	**96.0**	**98.5**	**61.2**	**89.2**	**95.0**	**517.2**
VSRN [15]	76.2	94.8	98.2	62.8	89.7	95.1	516.8
VSRN + OffTri	76.8	95.2	98.4	63.1	89.9	95.2	518.6
VSRN + OffQuin	76.9	95.3	98.4	63.3	90.2	95.5	519.7
VSRN + AdapOffQuin	**77.5**	**95.5**	**98.6**	**63.5**	**90.5**	**95.8**	**521.4**
5K Test Images							
SCAN [14]	50.4	82.2	90.0	38.6	69.3	**80.4**	410.9
SCAN + AdapOffQuin	**51.2**	**82.5**	**90.1**	**39.4**	**69.7**	80.4	**413.3**
BFAN [18]	52.9	82.8	90.6	38.3	67.8	79.3	411.7
BFAN + AdapOffQuin	**57.3**	**84.5**	**91.7**	**40.1**	**69.2**	**80.1**	**422.9**
VSRN [15]	53.0	81.1	89.4	40.5	70.6	81.1	415.7
VSRN + AdapOffQuin	**55.1**	**83.3**	**90.8**	**41.1**	**71.5**	**82.0**	**423.8**

Table 1 shows the performance comparison of models trained by different approaches on MS-COCO. We can see that all the three models are significantly improved on all the settings when trained by our proposed training approach. As mentioned in Section 4.2, for all the models, the offline hard negatives in their second-round training are sampled from the prediction of the first-round trained VSRN. It indicates that the proposed training approach is insensitive to the model consistency of the two-round training. When the global similarity score prediction step is intractable for the current model, we can train it by sampling offline hard negatives based on the prediction of another more efficient model. Overall, we achieve the most significant improvement on BFAN. In particular, on the more reliable 5K test set, it outperforms the baseline by 8.3% and 4.7% in top-1 sentence retrieval and top-1 image retrieval.
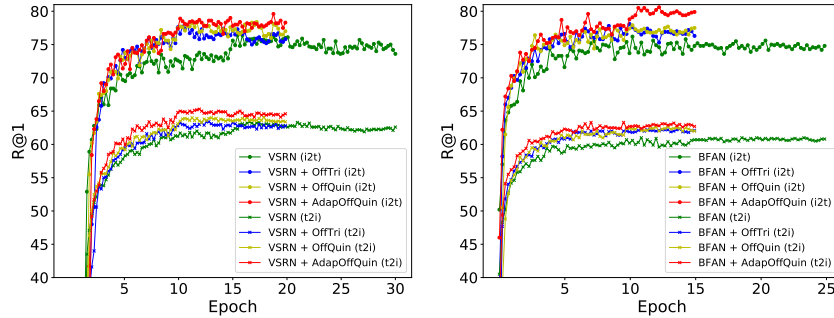
Fig. 4. Plotting training epoch against R@1 on the MS-COCO validation set for different training approaches applied on VSRN and BFAN. For the proposed approaches, the training curves correspond to the second-round training. "t2i" and "i2t" represents image retrieval and sentence retrieval, respectively.

Table 2 shows the performance comparison on Flickr30K. It should be noted that Flickr30K is much smaller than MS-COCO as it contains fewer very confusing negative image-text pairs to be served as high-quality offline hard negative pairs. However, significant improvements are still observed for all the models. In Section 4.4, we show that our proposed training approach has strong robustness for the quality of offline hard negatives.

Table 2. Quantitative evaluation results of sentence retrieval and image retrieval on the Flickr30K test set.

Model	Sentence Retrieval			Image Retrieval			
	R@1	R@5	R@10	R@1	R@5	R@10	rsum
1K Test Images							
SCAN [14]	67.4	90.3	**95.8**	48.6	77.7	85.2	465.0
SCAN + AdapOffQuin	**70.3**	**92.0**	95.5	**50.0**	**79.2**	**86.2**	**473.2**
BFAN [18]	68.1	91.4	95.9	50.8	78.4	85.8	470.4
BFAN + AdapOffQuin	**73.2**	**94.5**	**97.0**	**54.0**	**80.3**	**87.7**	**486.7**
VSRN [15]	71.3	90.6	**96.0**	54.7	81.8	88.2	482.6
VSRN + AdapOffQuin	**72.8**	**91.8**	95.8	**55.3**	**82.2**	**88.4**	**486.3**

We look deeper into different training approaches by examining VSRN and BFAN's training behaviours[3] on the widely-used MS-COCO 1K validation set [6,15,14] (*i.e.* the first fold of the 5K validation set). As shown in Figure 4, both models' performance obtains continuous improvement as we feed different proposed mechanisms into the training process. When the models are trained by Equation 5, they converge significantly faster than the baselines as it takes less than 10 epochs for them to outperform the highest R@1 of their baselines.

---

[3] The final BFAN model is an ensemble of two independently trained models BFAN-equal and BFAN-prob [18], here we show the behaviours of BFAN-prob.

Fig. 5. Qualitative image retrieval and sentence retrieval comparison between the baseline training approach and ours on the MS-COCO test set.

4.4   Ablation Study and Visualization

Table 3. Performance of different training approach variants on MS-COCO 1K test set. "OnlyOffline" represents the model that is only trained by the offline term. "Finetune" represents the model that is fine-tuned in the second-round instead of re-trained from scratch. "OnlineQuin" indicates that we apply online quintuplet loss instead of offline in Equation 4 (*i.e* replace $S(\bar{i}_{off}, \bar{t}_{off})$), $S(\widetilde{i}_{off}, \widetilde{t}_{off})$) with $S(\bar{i}_{on}, \bar{t}_{on})$)), $S(\widetilde{i}_{on}, \widetilde{t}_{on})$)) to train the model. "w/o OfflineAdap" represents that we replace $S(i, \bar{t}_{off})$ and $S(\bar{i}_{off}, t)$ by $S(i, t)$ for the new added terms in Equation 5 to train the model. Performance of selecting different top list size h for offline hard negative text sampling is also studied. The values in parentheses indicate the performance difference between the models trained by the variant and by the proposed approach with the final settings.

Model	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
1K Test Images						
BFAN (OnlyOffline)	1.1(-76.2)	2.5 (-93.5)	4.9 (-93.6)	0.5 (-60.7)	1.4 (-87.8)	2.6(-92.4)
VSRN (OnlyOffline)	0.7(-76.8)	2.1(-93.4)	3.8(-94.8)	0.4(-63.1)	1.2(-89.3)	2.3(-93.5)
BFAN (Fine-tune)	74.3 (-3.0)	94.7 (-1.3)	98.2 (-0.3)	58.7 (-2.5)	88.1 (-1.1)	94.2(-0.8)
VSRN (Fine-tune)	74.5 (-3.0)	94.3 (-1.2)	98.1 (-0.5)	62.0 (-1.5)	89.3(-1.2)	94.8(-1.0)
BFAN (OnlineQuin)	75.3 (-2.0)	95.8 (-0.2)	98.5 (+0.0)	59.8 (-1.4)	88.6 (-0.6)	94.6(-0.4)
VSRN (OnlineQuin)	76.4 (-1.1)	94.9 (-0.6)	98.2 (-0.4)	62.8 (-0.7)	89.9(-0.6)	95.2(-0.6)
BFAN (w/o OfflineAdap)	76.6 (-0.7)	95.8(-0.2)	98.4 (-0.1)	60.8 (-0.4)	89.1 (-0.1)	94.8(-0.2)
VSRN (w/o OfflineAdap)	77.1 (-0.4)	95.4(-0.1)	98.4 (-0.2)	63.4 (-0.1)	90.2 (-0.3)	95.5(-0.3)
VSRN (h = 200)	77.1 (-0.4)	95.3(-0.2)	98.4 (-0.2)	63.3 (-0.2)	90.4 (-0.1)	95.6(-0.2)
VSRN (h = 500)	77.4 (-0.1)	95.6(+0.1)	98.6 (+0.0)	63.5 (+0.0)	90.4 (-0.1)	95.7(-0.1)
VSRN (h = 1000)	77.3 (-0.2)	95.4(-0.1)	98.6 (+0.0)	63.3 (-0.2)	90.3 (-0.2)	95.6(-0.2)

First, we validate whether the offline hard negatives can completely replace online hard negatives to train the model. Specifically, we remove the online loss term in Equation 4 to train VSRN and BFAN. As shown in Table 3, the training process fails as it is too difficult for the model to directly learn to

distinguish between the positive pairs and these extremely confusing negative pairs. Also, we demonstrate the usefulness of re-training the model from scratch in the second round. As shown in Table 3, when we apply Equation 5 to fine-tune the model that has already been trained by the online triplet loss and get trapped in a local optimum, it cannot obtain additional improvement. In Equation 4, we create two new terms based on offline negatives. Indeed, we can instead apply them based on online negatives. However, the performance of "OnlineQuin" models are remarkably worse than the models train by Equation 4, this supports our claim of the second problem in Section 1. On the other hand, in Equation 5, we feed the offline hard negative information into the online term for hierarchical penalization. To validate its effectiveness, we replace $S(i, \bar{t}_{off})$ and $S(\bar{i}_{off}, t)$ by $S(i, t)$ for the new added terms in Equation 5 to break this hierarchical relation. α and β are re-adjusted to achieve the best performance on the validation set. The performance drops to the same level of using Equation 4 to train the models, indicating the effectiveness. In the end, for VSRN, we present the model's performance when selecting different top list size h for offline hard negative text sampling (we always keep it 5 times larger than the top list size for offline hard negative image sampling). We can find that even when h is set to 1000 which indicates significant drops of "hardness" degree of offline hard negatives, the model still achieves great performance. This is consistent with the excellent performance on Flickr30K and proves the robustness of our training approach on smaller datasets when very confusing hard negative pairs are limited.

Figure 5 shows the qualitative comparison between the models trained by different approaches on MS-COCO. For sentence retrieval, given an image query, we show the top-5 retrieved sentences. For image retrieval, given a sentence query, we show the top-3 retrieved images. The correct retrieval items for each query are ticked off. Overall, our training approach guides the model to better find and attend to the detailed non-correspondences of negative image-text pairs such as "snow covered field", "rhiho", "blowing out a candle" and "poster".

5   Conclusion

We present a novel training approach for image-text matching. It starts by mining "harder" negatives offline from the whole training set. Based on the mined offline hard negatives, an effective quintuplet loss is proposed to complement the online triplet loss to better distinguish positive and negative pairs. Furthermore, we take the distance relations among positive, offline hard negative and online hard negative pairs into consideration and effectively achieve adaptive penalization for different pairs. Extensive experiments demonstrate the effectiveness and generality of the proposed approach.

6   Acknowledgment

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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. pp. 6077–6086 (2018)
2. Chen, T., Luo, J.: Expressing objects just like words: Recurrent visual embedding for image-text matching. arXiv preprint arXiv:2002.08510 (2020)
3. Chen, T., Zhang, Z., You, Q., Fang, C., Wang, Z., Jin, H., Luo, J.: "factual"or"emotional": Stylized image captioning with adaptive learning and attention. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 519–535 (2018)
4. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 403–412 (2017)
5. Eisenschtat, A., Wolf, L.: Linking image and text with 2-way nets. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4601–4611 (2017)
6. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improved visual-semantic embeddings. arXiv preprint arXiv:1707.05612 **2**(7), 8 (2017)
7. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems. pp. 2121–2129 (2013)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Huang, Y., Wang, L.: Acmm: Aligned cross-modal memory for few-shot image and sentence matching. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5774–5783 (2019)
10. Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal lstm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2310–2318 (2017)
11. Ji, Z., Wang, H., Han, J., Pang, Y.: Saliency-guided attention network for image-sentence matching. arXiv preprint arXiv:1904.09471 (2019)
12. Kim, J.H., Lee, S.W., Kwak, D., Heo, M.O., Kim, J., Ha, J.W., Zhang, B.T.: Multimodal residual learning for visual qa. In: Advances in neural information processing systems. pp. 361–369 (2016)
13. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
14. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 201–216 (2018)
15. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4654–4662 (2019)
16. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1890–1899 (2017)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

18. Liu, C., Mao, Z., Liu, A.A., Zhang, T., Wang, B., Zhang, Y.: Focus your attention: A bidirectional focal attention network for image-text matching. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 3–11 (2019)
19. Liu, F., Ye, R., Wang, X., Li, S.: Hal: Improved text-image matching by mitigating visual semantic hubs. arXiv preprint arXiv:1911.10097 (2019)
20. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 6 (2017)
21. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems. pp. 289–297 (2016)
22. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
23. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 299–307 (2017)
24. Pedersoli, M., Lucas, T., Schmid, C., Verbeek, J.: Areas of attention for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1242–1250 (2017)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
26. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
27. Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., Shao, J.: Camp: Cross-modal adaptive message passing for text-image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5764–5773 (2019)
28. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: learning to rank with joint word-image embeddings. Machine learning **81**(1), 21–35 (2010)
29. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)
30. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016)
31. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4651–4659 (2016)
32. You, Q., Zhang, Z., Luo, J.: End-to-end convolutional semantic embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5735–5744 (2018)
33. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
34. Yu, D., Fu, J., Mei, T., Rui, Y.: Multi-level attention networks for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4709–4717 (2017)
35. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 686–701 (2018)

36. Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Shen, Y.D.: Dual-path convolutional image-text embedding with instance loss. arXiv preprint arXiv:1711.05535 (2017)