Supplementary Material for "Adaptive Offline Quintuplet Loss for Image-Text Matching"

Tianlang Chen[1], 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

In this document, we provide additional materials to supplement our paper "Adaptive Offline Quintuplet Loss for Image-Text Matching". In the first section, we perform additional ablation studies to verify the robustness and efficiency of the proposed training approach. In the second section, we show additional qualitative examples to compare the performance of the models trained by different approaches.

1 Ablation Study

In the section, we perform extra experiments to demonstrate the robustness and efficiency of our proposed training approach. All the experiments are performed based on VSRN (the last one is based on both VSRN and BFAN) on a single GeForce GTX 1080 Ti GPU.

First, one may ask whether simply increasing the training mini-batch size can replace our proposed training approach since it can also increase the "hardness" of the negative samples. To verify this, we increase the training batch size of VSRN to 192 – the maximum batch size that can be allocated by a single GeForce GTX 1080 Ti GPU. As shown in Table 1, simply improving the batch size cannot lead to better performance, demonstrating the effectiveness and validity of the proposed training approach.

Table 1. Performance of VSRN with the mini-batch size set to 192.

	Sentence Retrieval			Image Retrieval		
Model	R@1	R@5	R@10	R@1	R@5	R@10
1K Test Images						
VSRN	75.7	95.0	98.3	62.4	89.7	95.2

In Section 3.3 of the main paper, we feed the information of offline hard negatives into the online loss term and present a new loss function (*i.e.* Equation 5 of the main paper) for the second-round training. Hyper-parameters α and β are used to adjust the degree of adaptive penalization. Table 2 shows the performance of training VSRN with different α and β on the MSCOCO 1K test set. Overall, the performance difference is little when using different α and β of specific ranges to train the model. The adaptive penalization is not very sensitive

Table 2. Performance of selecting different α and β for adaptive penalization.

Model	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
1K Test Images						
VSRN ($\beta = 1.5$, $\alpha = 0.2$)	77.1	95.3	98.7	63.8	90.6	95.8
VSRN ($\beta = 1.5$, $\alpha = 0.3$)	77.5	95.5	98.6	63.5	90.5	95.8
VSRN ($\beta = 1.5$, $\alpha = 0.5$)	77.5	95.5	98.5	63.3	90.3	95.6
VSRN ($\beta = 1.0$, $\alpha = 0.3$)	77.0	95.2	98.5	63.2	90.1	95.5
VSRN ($\beta = 2.0$, $\alpha = 0.3$)	77.4	95.4	98.7	63.3	90.3	95.7

to the selection of α and β and constantly makes a positive effect on the training process, indicating the robustness of the proposed approach in Section 3.3.

In addition, to evaluate the training efficiency of the proposed training approach (the model's inference efficiency is unrelated to the training approaches), we compare the per-batch training time of VSRN with different training approaches on a fixed mini-batch size of 128. As shown in Table 3, when we employ our proposed approach, the training speed drops since the model needs to additionally compute the similarity score between the anchor and its sampled offline hard negatives. Overall, the per-batch training speed of VSRN with our proposed training approach is about 1.5 times slower than the per-batch training speed of VSRN with the basic online triplet loss. Considering that VSRN needs to be trained for 30 epochs (first round) by the online triplet loss and 20 epochs (second round) by the proposed training approach, the total training time of the second round is acceptable.

Table 3. Training efficiency comparison among different training approaches.

Model	Per-batch Training Time (Second)
1K Test Images	
VSRN	1.096
VSRN + OffTri	1.489
VSRN + OffQuin	1.557
VSRN + AdapOffQuin	1.604

In the end, for our cross-modality retrieval task, a corresponding positive image-text pair may perform well on one modality but poorly on the other (*e.g.* ranks top against the negative pairs that share the same image, but obtains low rank against the negative pairs that share the same text). We prove that our training approach does not exacerbate this unbalance. On the full MS-COCO 5K test set that contains 5,000 images, 25,000 texts, and 25,000 positive image-text pairs, for each pair, the trained models predict its rank against the 4,999 negative pairs that share the same text and 24,995 negative pairs that share the same image as r_i and r_t. For fair weighting between r_i and r_i with different negative pair numbers, the cross-retrieval rank of each positive pair is defined as: $\max(5r_i - 4, r_t)$. It records the lower rank of the positive pair against the two

kinds of negative pairs. Figure 1 shows the 25,000 positive pairs' cross-retrieval rank frequency distribution of different rank intervals. It can be seen that for both VSRN and BFAN, the number of positive pairs with the cross-retrieval rank of 1 (*i.e.* the positive pair's score is higher than the scores of all the 4,999 text-shared and 24,995 image-shared negative pairs) increases significantly when the proposed approach is applied. Meanwhile, the number drops for the pairs with cross-retrieval rank larger than 200, indicating a comprehensive improvement for the overall ranking of positive pairs in the test set.

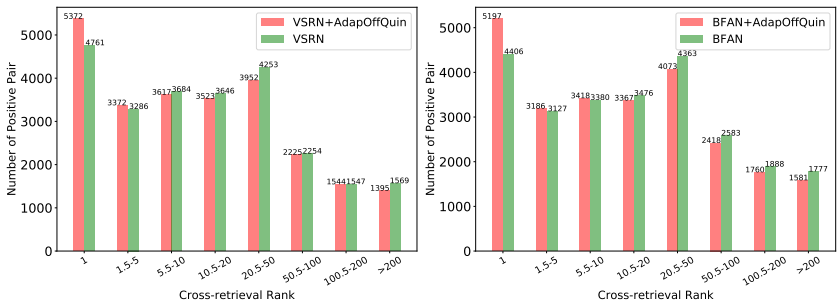

Fig. 1. Comparison of positive pairs' cross-retrieval rank frequency on the MS-COCO test set for different training approaches applied on VSRN and BFAN.

It should be noticed that we do not specially handle the false negative problem in the dataset – for some common scenes that occur many times (e.g., "surfing man"), there are offline negatives that should be considered positive. We instead implicitly avoid frequently sampling them by setting the top list size "h" to be not too small. For most anchors, there are no more than five false negatives in the dataset. The final setting in Section 4.2 of the main paper can safely maintain a very low rate of false negatives and prevent them from having a bad effect. Moreover, we also try to sample the negatives from a normal distribution instead of a uniform distribution to reduce the probability of sampling the most top ones that could be false negatives. We found that it does not lead to further improvement when "h" has already been set to a suitable value. The performance difference between sampling from a normal distribution or a uniform distribution is little.

2 Additional Qualitative Results

In this section, we provide a great number of image retrieval and sentence retrieval examples to compare the models trained by the baseline approach and the proposed one. Qualitative sentence and image retrieval results are shown as in Figure 2 and Figure 3, respectively. From Figure 2, the models trained by the proposed approach achieve better performance to differentiate between

Fig. 2. Qualitative sentence retrieval comparison between the baseline training approach and ours on the MS-COCO test set.

the corresponding sentences and the confusing non-corresponding sentences of an image query. In particular, they perform better to find the detailed non-correspondences of non-corresponding image-text pairs from the object number

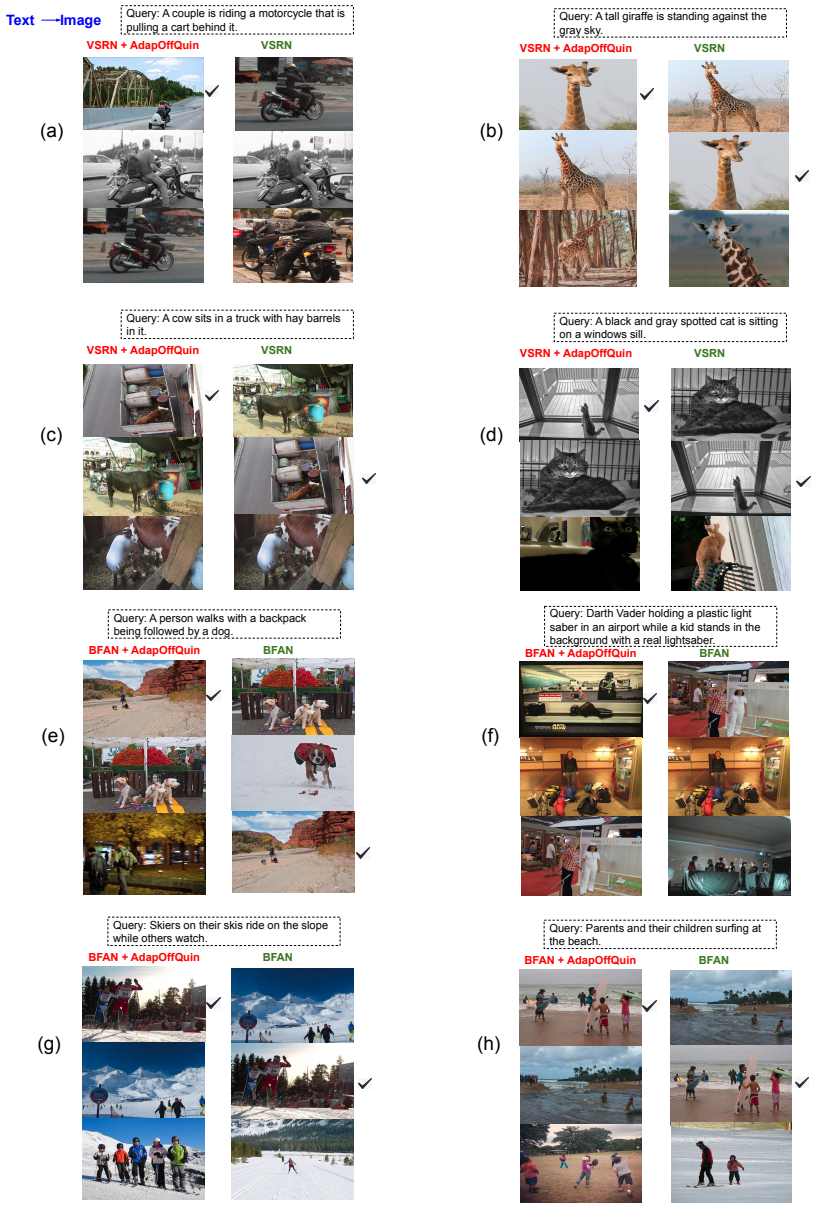

Fig. 3. Qualitative image retrieval comparison between the baseline training approach and ours on the MS-COCO test set.

(*e.g.* "the man" in (a)), object attribute (*e.g.* "goat" in (e)) and object relation (*e.g.* "holding a laptop" in (h)). As for image retrieval, as shown in Figure 3, they can successfully identify the images that miss the corresponding informa-

tion (*e.g.* "in a truck" in (c), "real lightsaber" in (f), "skis ride" in (g)) or contain the false information (*e.g.* "gray sky" in (b)).