Adversarial Attack on Deep Cross-Modal Hamming Retrieval Supplementary Material

Chao Li[1,*] Shangqian Gao[2,*] Cheng Deng[1,†] Wei Liu[3] Heng Huang[2,4]
[1]Xidian University [2]University of Pittsburgh [3]Tencent Data Platform [4]JD Explore Academy
{chaolee.xd, chdeng.xd, henghuanghh}@gmail.com, shg84@pitt.edu, wl2223@columbia.edu

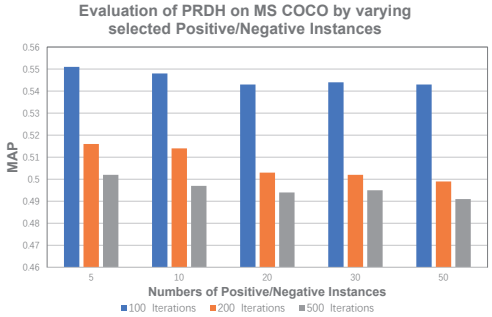
Table 1: Ablation study of AACH on MS COCO.

Tasks	Methods	Target Models			
		DCMH	PRDH	SSAH	CMHH
I → T	Reg	0.792	0.783	0.805	0.756
	AACH-S	0.244	0.235	0.265	0.237
	AACH-Q	0.472	0.499	0.469	0.434
	AACH	0.461	0.497	0.453	0.411
T → I	Reg	0.779	0.773	0.787	0.772
	AACH-S	0.220	0.229	0.238	0.226
	AACH-Q	0.494	0.464	0.461	0.421
	AACH	0.475	0.457	0.451	0.405

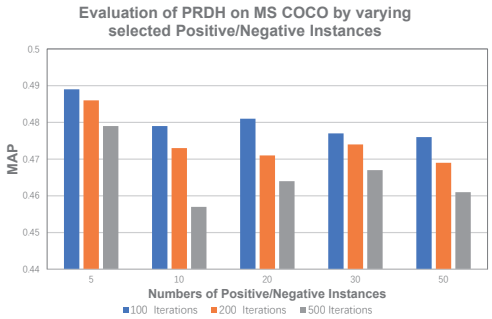
Results

We further evaluate our proposed AACH by implementing two ablation experiments "AACH-S" and "AACH-Q". "AACH-S" denotes the variant of AACH, which removes the surrogate model and attacks the target network by previously obtaining label information as well as model architectures and parameters. "AACH-Q" denotes that we execute AACH without considering the quantization loss. Results shown in Table 1 demonstrate the superiority of our proposed quantization loss constraint. AACH-S achieves good attack performance due to adopting the white-box setting.

To select positive/negative instances for the triplet construction, we further evaluate the proposed AACH with a different number of instances. Taking the target model of PRDH as an example, Fig. 1 shows the result that MAP evaluations on MS COCO by varying numbers of the selected instance. We find that improving the selected positive/negative instances can further facilitate the attack capacity of our method. However, it can be seen that this improvement is not obvious when over 10 instances are selected. Therefore, considering both the computation efficiency and good performance, 10 positive/negative instances are used in our experiments.



(a) Image-query-Text Retrieval



(b) Text-query-Image Retrieval

Figure 1: Selection of the number of positive/negative instances in the triplet construction.

Following our experiments, the evaluation of the attack transferability on MS COCO is shown in Table 2. The target model for each bit (e.g., 16 bits, 32 bits, and 64 bits) is first obtained, respectively. Then, we learn adversarial examples based on the 32-bit surrogate model to attack 16-bit and 32-bit target models. We find that the adversarial examples learned by our AACH have good transferability across target models with different code lengths.

PR and precision @Top1000 curves of different target models before and after being attacked on MS COCO are

*Equal contribution.
†Corresponding author.

Table 2: Attack transferability comparison in terms of MAP scores of two retrieval tasks on MS COCO. The adversarial examples are learned from the target model designed for 32-bit binary codes, aiming to attack target models of other bits."R" denotes regular retrieval, and "A" denotes attacking retrieval using our proposed AACH.

Tasks	Bits	MS COCO					
		DCMH		PRDH		SSAH	
		R	A	R	A	R	A
I → T	16	0.589	0.465	0.592	0.483	0.601	0.462
	32	0.617	0.461	0.621	0.497	0.628	0.453
	64	0.622	0.466	0.620	0.511	0.631	0.457
T → I	16	0.575	0.447	0.580	0.424	0.594	0.423
	32	0.593	0.475	0.610	0.457	0.646	0.451
	64	0.604	0.455	0.611	0.443	0.650	0.466

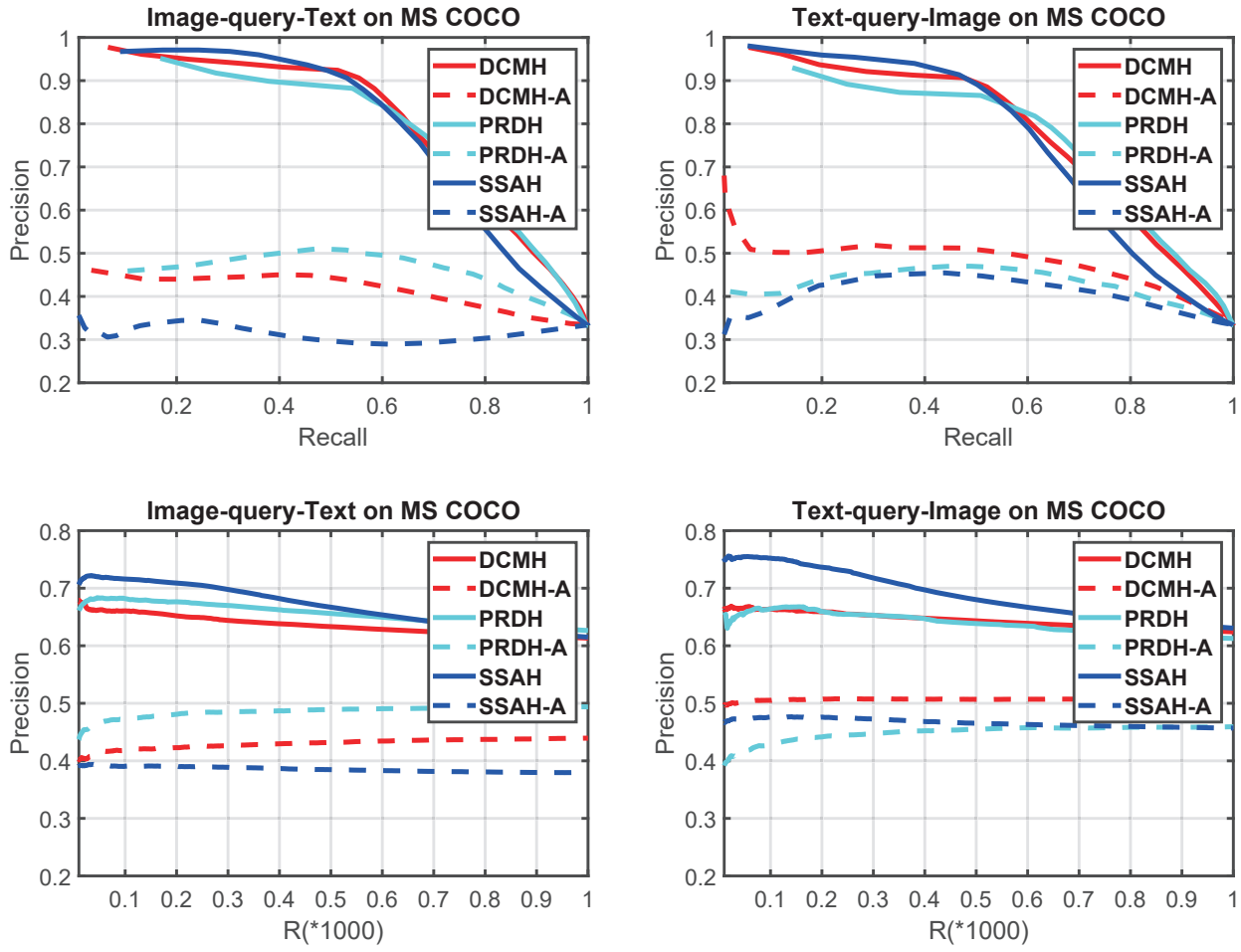


Figure 2: PR and precision @Top1000 curves evaluated on MS COCO with 32-bit binary codes. '*-A' means that the target model is attacked by the proposed AACH.

also provided, which are shown in Fig. 2. Again, similar conclusions to that in our original paper can be achieved that our proposed AACH can significantly decrease the retrieval accuracy of all target models.