Targeted Transferable Attack against Deep Hashing Retrieval

Fei Zhu
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
zhufei@iie.ac.cn

Wanqian Zhang*
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
zhangwanqian@iie.ac.cn

Dayan Wu
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
wudayan@iie.ac.cn

Lin Wang
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
wanglin5812@iie.ac.cn

Bo Li
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
libo@iie.ac.cn

Weiping Wang
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
wangweiping@iie.ac.cn

ACM Reference Format:
Fei Zhu, Wanqian Zhang, Dayan Wu, Lin Wang, Bo Li, and Weiping Wang. 2023. Targeted Transferable Attack against Deep Hashing Retrieval. In *ACM Multimedia Asia 2023 (MMAsia '23), December 6–8, 2023, Tainan, Taiwan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3595916.3626420

A EXPERIMENTS

A.1 Implementation Details

For fair comparison, we re-implement P2P [1], DHTA [1], ProS-GAN [9], THA [8] and NAG [10] with the same hyperparameters and experimental setting as in their original works. For NAG method, we randomly sample one image of target label for each query image to construct normal pairs. Since target label can affect the targeted transferability [13], we adopt the same target labels for ProS-GAN, THA, NAG and our TTA-GAN method for fair comparison. Besides, as P2P is a special case of DHTA, same target labels are designed for P2P and DHTA. We utilize Adam [4] optimizer for our IAO method with iteration number 1000 and learning rate $\eta = 1$ for FLICKR-25K and ImageNet datasets and $\eta = 2$ for NUS-WIDE dataset. We train our TTA-GAN with 100 epochs, 24 batch size and 1e-4 learning rate. The hyperparameter β is set to 0.03. We set α to 200 for FLICKR-25K, 100 for NUS-WIDE and gradually increase α from 1 to 100 for ImageNet. For the input transformations, we set $M = 3$. Besides, for rotation method, we set the maximum rotation degree (denoted as Θ) to 10 and perform random rotation with degree $d \sim \mathcal{U}(-\Theta, \Theta)$ for each image. For resizing method, we adopt the hyperparameters stated in the original paper. Besides, we adopt VGG11, RN50 and DN161 as white-box models and the other three as hold-out black-box models for our TTA-GANens method.

A.2 Attack on Adversarially Trained Models

Table 1 shows the targeted attack performance on adversarially trained models, i.e., VGG11, RN50 and DN161, with defense strategies ATRDH [8] and CgAT [7]. Specifically, VGG11_A and VGG11_C are deep hashing models trained with ATRDH and CgAT strategies respectively. From the results, we can derive that the targeted black-box performance on these models are unsatisfactory compared with normally trained models, because they augment training data with additional targeted and powerful untargeted adversarial examples. In spite of that, our method still achieves an improvement in most cases, which further demonstrates the superiority of our TTA-GAN method. Besides, our TTA-GANens method can boost the targeted transferability only with a slight improvement, which leaves us further exploration on methods with powerful attack performance on adversarially trained models.

A.3 Attack on Other Deep Hashing Methods

We also conduct experiments on other deep hashing methods including HashNet [2], CSQ [11], DSDH [5] and ADSH [3], to further show the universality of our TTA-GAN method and effectiveness on targeted transferability. As reported in Table 2, even on different deep hashing methods, our TTA-GAN method still achieves the best anchor t-MAP and targeted transferability. This phenomenon shows that our method can constantly improve the targeted black-box attack performance.

A.4 Multiple Input Transformations

In this subsection, we will analyze the correlation between input transformations and targeted transferability. First, we investigate the effect of rotation degree Θ, where Θ verifies from $0°$ to $70°$. As shown in Figure 1 (a), when the Θ increases from $0°$ to $10°$, the targeted transferability is improved. However, when $\Theta \geq 10°$, the targeted transferability starts to decline. We argue that this is because larger Θ will greatly change the semantic information of

*Corresponding author.

MMAsia '23, December 6–8, 2023, Tainan, Taiwan

ACM ISBN 979-8-4007-0205-1/23/12.
https://doi.org/10.1145/3595916.3626420

Table 1: The test t-MAP (%) of targeted attacks against adversarially trained models which are trained with DPSH method, 32 bits and different backbones on FLICKR-25K.

	Method	VGG11$_A$	RN50$_A$	DN161$_A$	VGG11$_C$	RN50$_C$	DN161$_C$
VGG11	P2P	63.20	60.71	62.11	76.75	61.99	58.39
	DHTA	62.83	60.58	62.09	77.47	61.98	58.38
	THA	72.03	**69.75**	70.92	78.72	71.07	**67.52**
	ProS-GAN	71.72	69.21	70.91	83.78	71.02	67.45
	TTA-GAN	**72.13**	69.29	**71.07**	**87.10**	**71.11**	67.51
RN50	P2P	62.58	60.63	62.08	64.73	61.99	58.39
	DHTA	62.60	60.57	62.09	64.96	61.99	58.40
	THA	71.59	**69.80**	70.93	77.25	71.07	67.51
	ProS-GAN	71.51	68.99	70.84	76.08	71.00	67.52
	TTA-GAN	**72.31**	69.45	**71.17**	**83.95**	**71.25**	**67.55**
DN161	P2P	62.57	60.51	62.07	64.58	61.96	58.39
	DHTA	62.65	60.59	62.08	64.93	61.97	58.39
	THA	71.60	**69.78**	70.92	77.26	71.06	67.52
	ProS-GAN	71.34	69.09	70.85	74.59	70.96	67.54
	TTA-GAN	**72.03**	69.63	**71.19**	**83.08**	**71.22**	**67.56**
TTA-GANens		72.60	69.46	71.26	87.11	71.29	67.56

Table 2: The results of targeted attacks against other deep hashing methods with 32 bits on FLICKR-25K. VGG11 is adopted as the backbone of white-box models.

	Method	Per.	A.	AlexNet	VGG11	RN50	RN152	Inc-v3	DN161
HashNet	P2P	0.90	85.17	64.60	85.19*	68.78	67.75	67.08	68.75
	DHTA	0.81	88.85	64.57	88.70*	69.29	68.11	67.21	69.26
	THA	2.16	93.01	73.18	**92.95***	76.71	75.50	74.82	76.59
	ProS-GAN	2.53	93.21	74.23	92.42*	81.41	80.13	76.71	81.74
	TTA-GAN	2.70	**93.62**	**77.46**	91.83*	**86.77**	**85.62**	**83.90**	**86.42**
CSQ	P2P	0.86	77.96	61.34	77.57*	62.96	63.16	62.49	63.81
	DHTA	0.78	80.52	61.29	80.03*	63.01	63.24	62.50	63.91
	THA	2.13	91.33	69.80	**91.07***	70.70	71.06	70.01	71.85
	ProS-GAN	3.00	91.84	70.33	90.43*	73.89	74.65	71.85	74.50
	TTA-GAN	2.90	**92.42**	**73.24**	89.24*	**77.54**	**76.94**	**75.45**	**77.37**
DSDH	P2P	0.89	85.41	65.18	85.45*	68.77	67.94	66.36	68.98
	DHTA	0.81	89.27	65.30	89.12*	69.01	67.93	66.60	69.22
	THA	2.15	93.35	73.98	**93.35***	76.20	75.59	73.92	76.38
	ProS-GAN	2.55	93.57	75.29	92.83	82.32	80.24	76.38	83.27
	TTA-GAN	2.69	**93.77**	**78.78**	92.11*	**87.45**	**85.24**	**81.59**	**87.53**
ADSH	P2P	0.54	90.48	71.89	94.35*	73.35	73.86	72.05	72.62
	DHTA	0.54	90.48	71.87	94.35*	73.34	73.89	72.07	72.66
	THA	1.92	93.05	79.92	93.53*	79.04	79.43	77.86	78.81
	ProS-GAN	2.35	93.28	79.72	92.92*	80.08	80.36	78.24	82.55
	TTA-GAN	2.40	**94.55**	**81.71**	**94.69***	**85.81**	**84.55**	**85.81**	**86.35**

images, resulting in inaccurate update direction, and thus hinder the generation of adversarial feature. To verify this, we adopt the Hamming distance between hash codes of image before and after transformations to represent the changes of semantic information. As can be seen in Figure 1 (a), with the increase of Θ, the Hamming distance is gradually improved. Besides, although SI [6] method has been proved to be effective in image classification task [12], as in Figure 1 (b), we experimentally show that the combination with SI method will lower the targeted transferability. This is because the combination with SI method will greatly change the semantic information of generated adversarial example.

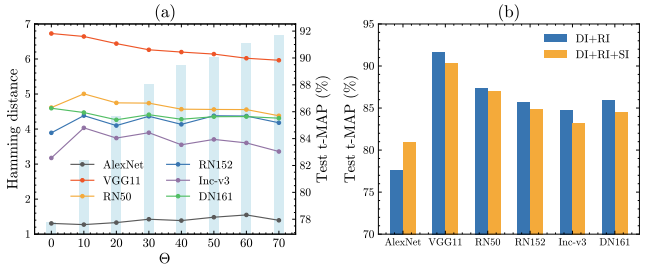

Figure 1: The ablation experiments on input transformations.

A.5 Further Analysis

Transferability across code lengths and deep hashing methods. In this subsection, we conduct experiments on transferring the generated adversarial example to different code lengths and deep hashing methods respectively. We report the targeted transferability in Table 3 and Table 4 respectively. From the results, we can observe that the targeted attack on black-box models with different code lengths (or deep hashing methods) achieves comparable targeted white-box attack performance. This indicates that the transferability of these two settings mainly depends on the white-box attack performance. Even though our method can not achieve the best white-box attack performance, the targeted white-box attack performance can be further boosted with the supervision of our IAO method.

Table 3: The test t-MAP (%) of targeted transferability across different code lengths.

Bits	Method	16 bits	32 bits	64 bits
32 bits	P2P	83.91	85.17*	85.13
	DHTA	87.22	88.46*	88.59
	THA	91.11	92.62*	92.84
	ProS-GAN	91.05	92.45*	92.57
	TTA-GAN	90.35	91.60*	91.88

Table 4: The test t-MAP (%) of targeted transferability across different deep hashing methods.

Hash	Method	DPSH	HashNet	CSQ	DSDH	ADSH
DPSH	P2P	85.17*	84.76	77.62	84.77	84.42
	DHTA	88.46*	88.12	79.66	88.23	85.88
	THA	92.62*	92.38	85.91	92.28	92.22
	ProS-GAN	92.45*	92.24	85.98	92.06	93.06
	TTA-GAN	91.60*	91.30	87.98	91.32	93.46

REFERENCES

[1] Jiawang Bai, Bin Chen, Yiming Li, Dongxian Wu, Weiwei Guo, Shu-tao Xia, and En-hui Yang. 2020. Targeted attack for deep hashing based retrieval. In *Proceedings of the European Conference on Computer Vision.*

[2] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu. 2017. Hashnet: Deep learning to hash by continuation. In *Proceedings of the IEEE International Conference on Computer Vision.*

[3] Qing-Yuan Jiang and Wu-Jun Li. 2018. Asymmetric deep supervised hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

[4] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations.*

[5] Qi Li, Zhenan Sun, Ran He, and Tieniu Tan. 2017. Deep supervised discrete hashing. *Proceedings of the Advances in Neural Information Processing Systems* (2017).
[6] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. 2020. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In *Proceedings of the International Conference on Learning Representations*.
[7] Xunguang Wang, Yiqun Lin, and Xiaomeng Li. 2023. CgAT: Center-Guided Adversarial Training for Deep Hashing-Based Retrieval. In *Proceedings of the ACM Web Conference 2023*.
[8] Xunguang Wang, Zheng Zhang, Guangming Lu, and Yong Xu. 2021. Targeted attack and defense for deep hashing. In *Proceedings of the International Conference on Research and Development in Information Retrieval*.
[9] Xunguang Wang, Zheng Zhang, Baoyuan Wu, Fumin Shen, and Guangming Lu. 2021. Prototype-supervised adversarial network for targeted attack of deep hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[10] Yanru Xiao and Cong Wang. 2021. You see what I want you to see: Exploring targeted black-box transferability attack for hash-based image retrieval systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[11] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. 2020. Central similarity quantization for efficient image and video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[12] Zheng Yuan, Jie Zhang, and Shiguang Shan. 2022. Adaptive image transformations for transfer-based adversarial attack. In *Proceedings of the European Conference on Computer Vision*.
[13] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. 2021. On success and simplicity: A second look at transferable targeted attacks. In *Proceedings of the Advances in Neural Information Processing Systems*.