Table 1: Comparison in terms of MAP scores of two retrieval tasks between natural training and CMLA attack on IAPRTC12 with different hash code lengths.

Task	Method	Natural training		CMLA attack	
		16	32	16	32
I $\rightarrow$ T	DCMH	0.454	0.470	0.299	0.302
	SSAH	0.478	0.494	0.334	0.346
T $\rightarrow$ I	DCMH	0.480	0.496	0.298	0.304
	SSAH	0.488	0.509	0.289	0.291

Table 2: Trade-off of SSAH between natural retrieval and performance under adversarial attack on MIRFlickr-25K with 32-bit hash code length.

Task	SSAH Adv-Trained	# Adversarial Samples used			
		0	100	500	2000
I $\rightarrow$ T	natural	0.805	0.799	0.785	0.770
	under attack	0.665	0.681	0.709	0.784
T $\rightarrow$ I	natural	0.805	0.801	0.784	0.773
	under attack	0.589	0.623	0.667	0.788

Thank all the reviewers for their valuable comments. We have fixed all the mistakes and made responses to all questions. Given your constructive suggestions, we have confidence on improving our work further.

To Reviewer 1 :

1: Considering the structure difference between image and text, CMLA learns different perturbations for two modalities, where two perturbations are updated iteratively. The correlation between different modalities is mainly learned during cross-modal hash codes generation and is then treated as a supervision signal to learn the optimal perturbation for each modality. **2:** By replacing the second term of Eq. (5) with $\sum_{i,j=1}^{n}\left\|(1-S_{ij})\Theta_{ij}-\log\left(1+e^{\Theta_{ij}}\right)\right\|^2$, CMLA can learn adversarial samples to attack both the cross-modal correlations and intra-modal similarities. However, given a cross-modal system, adversarial samples with errors in both single-modal and cross-modal are suspicious and can be easily detected. On the contrary, adversarial samples with errors merely in cross-modal but correct in single-modal are much harder to be discovered, which are more deceptive adversarial samples. This is the major reason why we want to keep intra-modal similarity. Therefore, these two types of adversarial samples are different. By utilizing adversarial samples from CMLA, the robustness of model is improved. **3:** During learning adversarial samples for a cross-modal task, the correlation between different modalities is leveraged as a guidance to generate adversarial samples with high deception. While single-modal learning only focuses on intra-modal relationship, which can be seen as a sub-problem of cross-modal counterparts. **4:** Thanks for the valuable suggestion. We further evaluate our CMLA on another cross-modal dataset IAPRTC12 holding richer data semantics, where 1000 and 4000 data points are respectively selected as a query set and a training set. Each text is represented as a 2912-dimensional bag-of-words vector, and each text-image pair belongs to at least one of 255 concepts. Due to the limited space, partial results are shown in Table 1 on this page. The entire results have been obtained and will be placed into the final version, which can again demonstrate the effectiveness of the proposed CMLA. Moreover, an additional experiment on adversarial samples is done. Please refer to our response to Reviewer 2. Other cross-modal tasks are out of the scope for this paper but can be a good guidance for future work.

To Reviewer 2 :

1: We further highlight the contributions of CMLA in the following three aspects. First, instead of simply learning adversarial samples attacking a neural network, our main contribution is to exploit adversarial samples across different modalities. Second, we simultaneously integrate inter- and intra- modality similarity regularizations across different modalities into the learning of adversarial samples, which has a great difference from a single-modal task. Finally, the task of cross-modal hashing, for the first time, is adopted to demonstrate adversarial sample learning, obviously showing the effectiveness of the proposed CMLA. **2:** Thanks for your careful checking. Following your suggestion to clearly illustrate this, we rewrite Eq. (2) as $\max_{\delta^*} D\left(H\left(x^*+\delta^*;\theta^*\right), H\left(x^*;\theta^*\right)\right), s.t. \|\delta^*\|_p \leq \epsilon, * \in \{v,t\}$, where hash codes H^* are generated from hash layer $\mathcal{H}$ by learning a deep network θ^*, and $D(\cdot,\cdot)$ is a distance measure. Considering the binarization of hash codes, a large divergence between $\mathcal{H}\left(x^*+\delta^*;\ \theta^*\right)$ and $\mathcal{H}\left(x^*;\ \theta^*\right)$ means a long Hamming distance between the generated hash codes, thus resulting in effective perturbations. This problem is further specified in Eq. (5), where $\max_{\delta^*} D(\cdot,\cdot)$ is equal to $\min_{\delta^*} \mathcal{J}$. **3:** In this paper, we maximize the distance for $S_{ij}=1$, while the case of $S_{ij}=0$ means that two data points are semantically dissimilar, so this relationship should be kept. Therefore, we don't design an individual constraint in Eq. (3). **4:** It seems that the reviewer may have misunderstood the robustness. During performing adversarial training, the robustness is about the defense to the adversarial samples. For a better illustration, we additionally evaluate our CMLA using different quantities of adversarial samples. The result is shown in Table 2 on this page. As the quantity of adversarial samples used in training increases, the performance under attacking also increases (robustness is increased) while the natural performance decreases. Such a trade-off is widely observed in adversarial training for regular classification tasks.

To Reviewer 3 :

1: Thanks for your interest in our work. **2:** In Eq. (5), the equality constraints here are just a simple replacement of Γ_{ij} and Θ_{ij}, so they do not introduce any error-prone signals. Thus, back-propagation is sufficient to optimize Eq. (5). **3:** Thanks for your valuable suggestion. Up to now, the paper referred to by the reviewer still cannot be searched. We would be glad to cite and compare this work with our CMLA in our final version if it can be totally published before that. **4:** Agree. Following your valuable suggestion, these three works will be cited to further enrich our work.