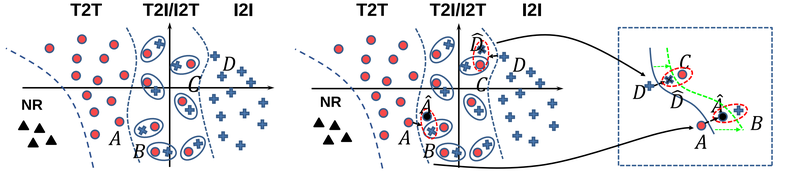
Cross-Modal Learning with Adversarial Samples

Chao Li[1,2], Cheng Deng[1], Shangqian Gao[2], De Xie[1], Wei Liu[3]
1 School of Electronic Engineering, Xidian University
2 Electrical and Computer Engineering, University of Pittsburgh 3 Tencent AI Lab

Introduction

Cross-Modal Learning



(a) Cross-modal Search Space (b) Adversarial Sample Learning (c) Adversarial Training

Cross-Modal Hashing Attack

$$\Delta(o^*, H^*) := \min_{\delta^*} \|\delta^*\|_p$$

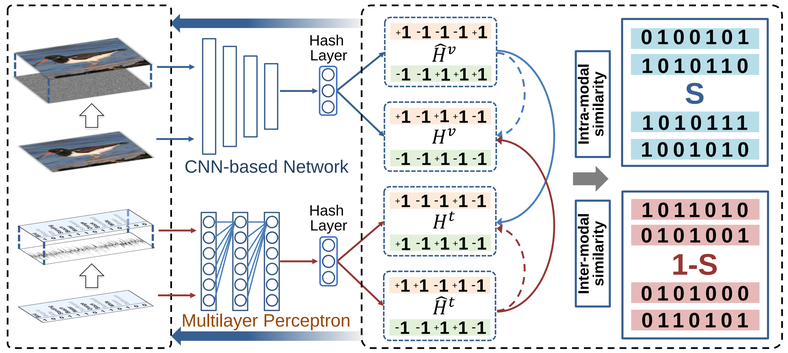
$$s.t. \max_{\delta^*} D\big(H^*(o^* + \delta^*; \theta^*), H^*(o^*; \theta^*)\big), \quad \|\delta^*\|_p \le \varepsilon$$

$$B^* = \mathrm{sign}(H^*), * \in \{v, t\}$$

Contribution

- We propose a simple yet effective cross-modal learning method by exploring adversarial samples, where adversarial sample is defined in two aspects: learned perturbation is designed to fool a deep cross-modal network while the distortion on an individual modality will not impact the performance within its own modality.
- A novel cross-modal adversarial sample learning method is presented. We decrease the inter-modality similarity and simultaneously keep intra-modality similarity in one optimization, with which more deceptive samples can be learned.
- We additionally apply the proposed CMLA into cross-modal hashing learning. Experiments on two widely used cross-modal retrieval benchmarks show the effectiveness of our CMLA in attacking a target retrieval system and further improving its robustness.

Framework



Propose CMLA

Take adversarial attack samples learning for image modality as an example:

$$\min_{\delta^v} J^v = \sum_{i,j=1}^{n} \big(\log\big(1 + e^{-\Gamma_{ij}}\big) + S_{ij}\Gamma_{ij}\big)$$

Inter-modal similarity

$$+ \sum_{i,j=1}^{n} \big(\log\big(1 + e^{\Theta_{ij}}\big) - S_{ij}\Theta_{ij}\big)$$

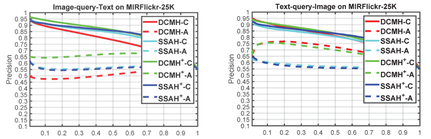
Intra-modal similarity

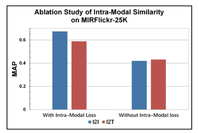
$$+ \sum_{i=1}^{n} \|\hat{o}_i^v - o_i^v\|_p$$

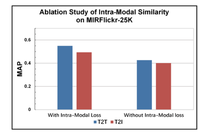
$$s.t. \Gamma_{ij} = \frac{1}{2}\big(\hat{H}_i^v\big)\big(H_j^t\big)^T, \Theta_{ij} = \frac{1}{2}\big(\hat{H}_i^v\big)\big(H_j^v\big)^T$$

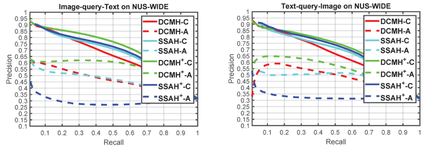
Experiments

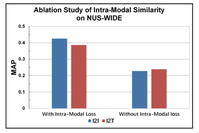
Task	Iteration		MIRFlickr-25K				NUS-WIDE			
			DCMH	DCMH$^+$	SSAH	SSAH$^+$	DCMH	DCMH$^+$	SSAH	SSAH$^+$
Image Query v.s. Text Database	0	MAP	0.749	0.816	0.805	0.815	0.607	0.679	0.660	0.675
	100	MAP	0.579	0.631	0.679	0.681	0.526	0.609	0.587	0.591
		D	0.039	0.041	0.034	0.038	0.031	0.033	0.032	0.025
	200	MAP	0.563	0.599	0.671	0.699	0.499	0.583	0.534	0.543
		D	0.023	0.038	0.028	0.032	0.026	0.031	0.029	0.026
	500	MAP	0.521	0.554	0.665	0.674	0.457	0.578	0.460	0.502
		D	0.019	0.029	0.020	0.023	0.025	0.028	0.026	0.024
Text Query v.s. Image Database	0	MAP	0.797	0.820	0.805	0.809	0.614	0.691	0.677	0.685
	100	MAP	0.615	0.619	0.603	0.611	0.523	0.628	0.501	0.523
		D	0.048	0.037	0.031	0.021	0.037	0.035	0.042	0.025
	200	MAP	0.587	0.577	0.595	0.605	0.447	0.549	0.454	0.474
		D	0.027	0.033	0.025	0.019	0.035	0.031	0.035	0.023
	500	MAP	0.561	0.564	0.589	0.593	0.371	0.533	0.351	0.427
		D	0.019	0.021	0.023	0.017	0.030	0.027	0.017	0.019

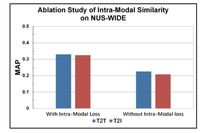


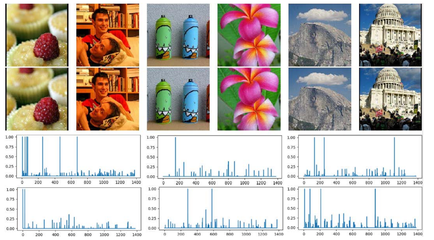












Task	Method	MIRFlickr-25K				NUS-WIDE			
		16	32	48	64	16	32	48	64
Image Query v.s. Text Database	DCMH	0.711	0.723	0.735	0.759	0.578	0.587	0.611	0.628
	DCMH$^+$	0.779	0.781	0.803	0.801	0.647	0.649	0.666	0.677
	SSAH	0.783	0.784	0.788	0.785	0.615	0.640	0.658	0.661
	SSAH$^+$	0.784	0.788	0.787	0.789	0.621	0.635	0.662	0.670
Text Query v.s. Image Database	DCMH	0.771	0.775	0.782	0.786	0.610	0.603	0.611	0.620
	DCMH$^+$	0.793	0.801	0.803	0.799	0.655	0.673	0.675	0.676
	SSAH	0.790	0.788	0.792	0.791	0.638	0.659	0.660	0.664
	SSAH$^+$	0.789	0.789	0.794	0.793	0.641	0.664	0.668	0.667

Code and Homepage

https://github.com/ChaoLi1991/CMLA

https://chaoli1991.github.io