Adversarial Samples in Deep Hash Learning

Cheng Deng
chdeng@mail.xidian.edu.cn

School of Electronical Engineering
Xidian University

Oct. 25, 2019

Content

I. Adversarial Samples

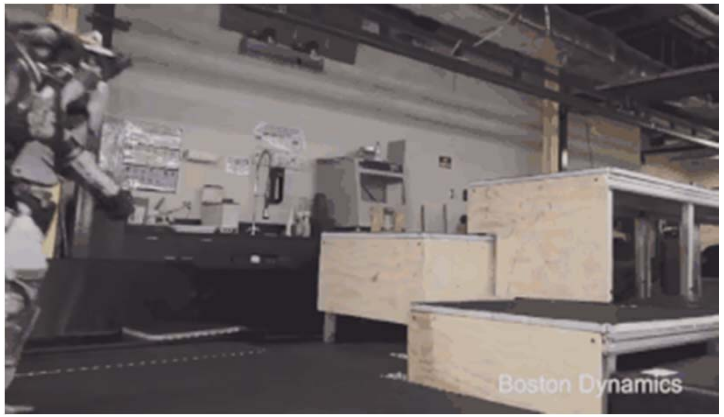
II. Deep Hash Learning

III. Adversarial Samples in Hashing

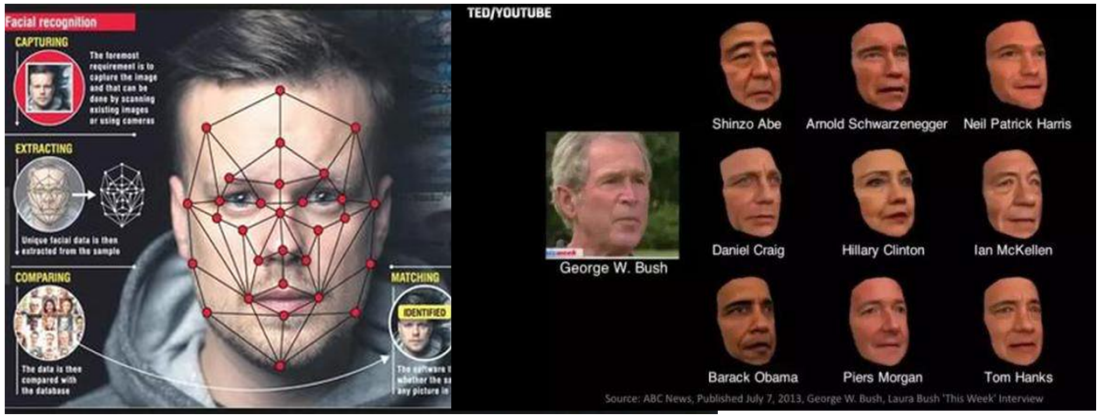
- Adversarial Examples for Hamming Space Search (HAG)
- Cross-Modal Learning with Adversarial Samples (CMLA)

Content

Background

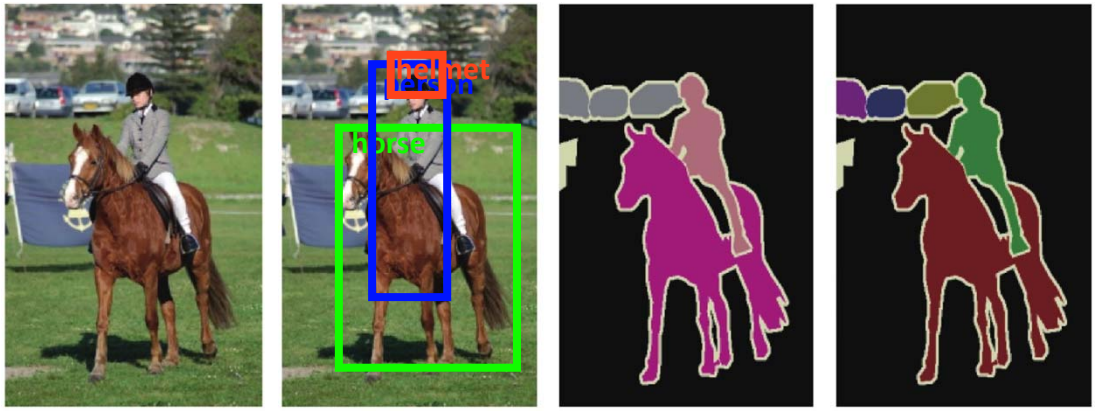


AI Robot



Face Recognition & Sythesis

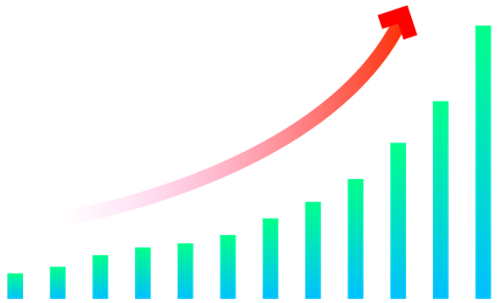
AI Recommended Systems



Object Recognition & Classification & Segmentation

Background

Boosting



Accuracy and Speed

Deep Network

Cat → 97%

Dog → 1%

Other → 2%

Deep Network

Cat → 4%

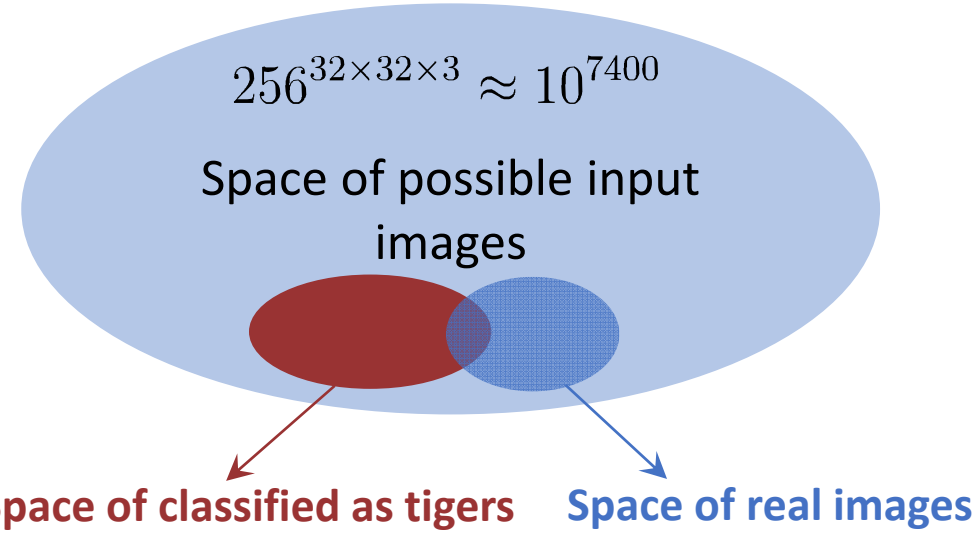
Dog → 96%

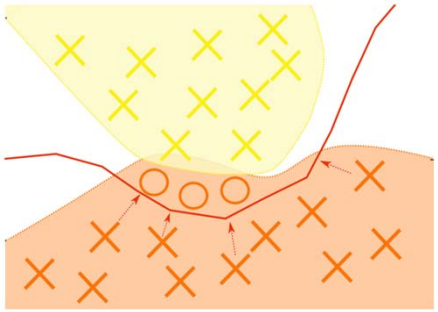
Other → 0%

Adversarial Samples

- **Attacking a network with adversarial samples**

  Questions: Will the forged image look like a tiger?





Adversarial Samples

- **Adversarial sample generation**

  Basic Algorithm: Fast Gradient Sign Method (FGSM)

$$X + \underbrace{\varepsilon sign(\nabla_x L(\theta, x, y))}_{\eta} = X_{adv}$$

"Tiger" 100% confidence

Benign image

Adversarial noise

"Wolf" 71% confidence

Adversarial image

I. Goodfellow, et al, Explaining and harnessing adversarial examples, 2014

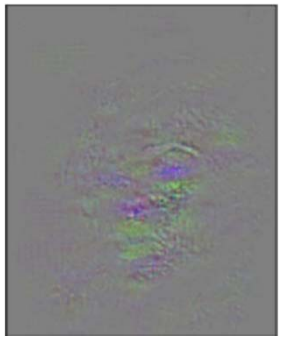
Adversarial Samples

Fast Flipping Attribute Method

Wearing lipstick

Clean

+



Perturbation

=

Not wearing lipstick

Adversarial

Retaining Biometric Utility of Face Images while Perturbing Gender

Female

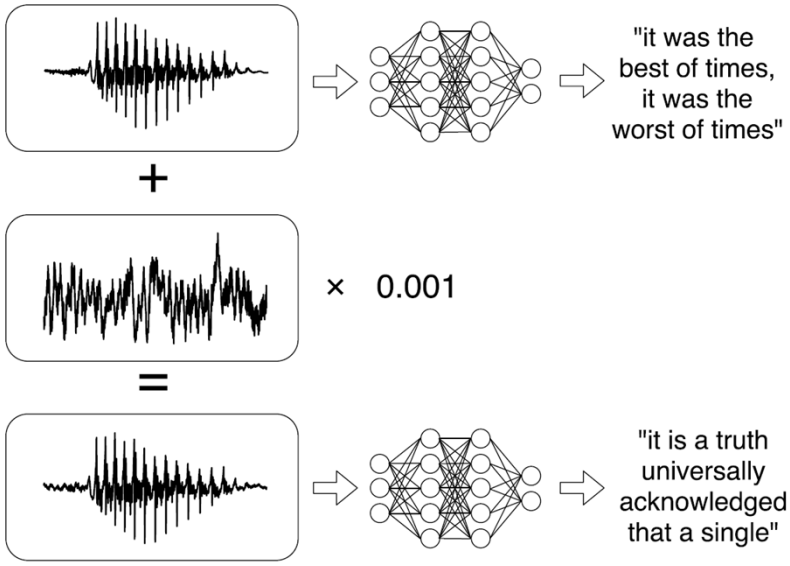
+

=

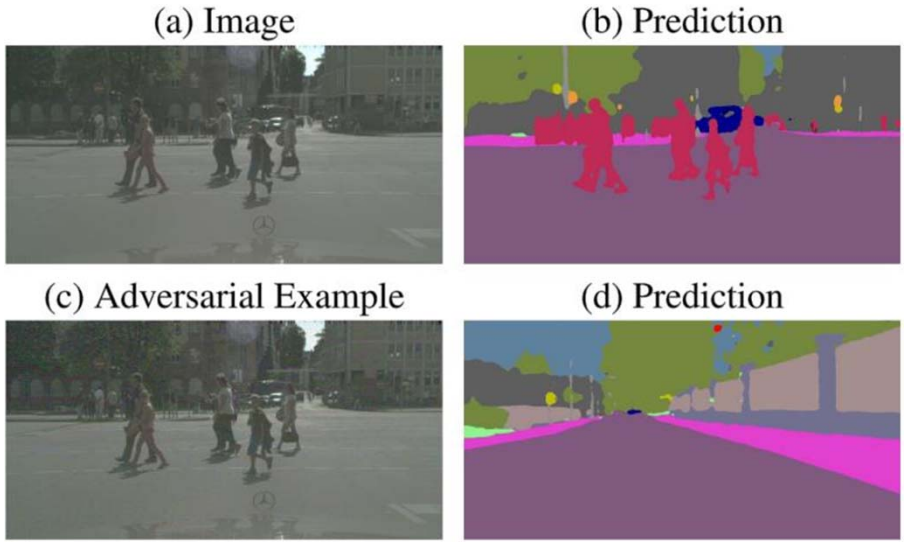
Male

Applications

- **Adversarial samples can mislead the classification of AI system**
  - ✓ Existing attack techniques: FGSM, Deepfool, JSMA, Black-box, CW, etc.
  - ✓ Existing defense techniques: adversary detection, adversarial retraining, obfuscated gradients, etc.



A small perturbation can make the result transferred to any desired target phrase.



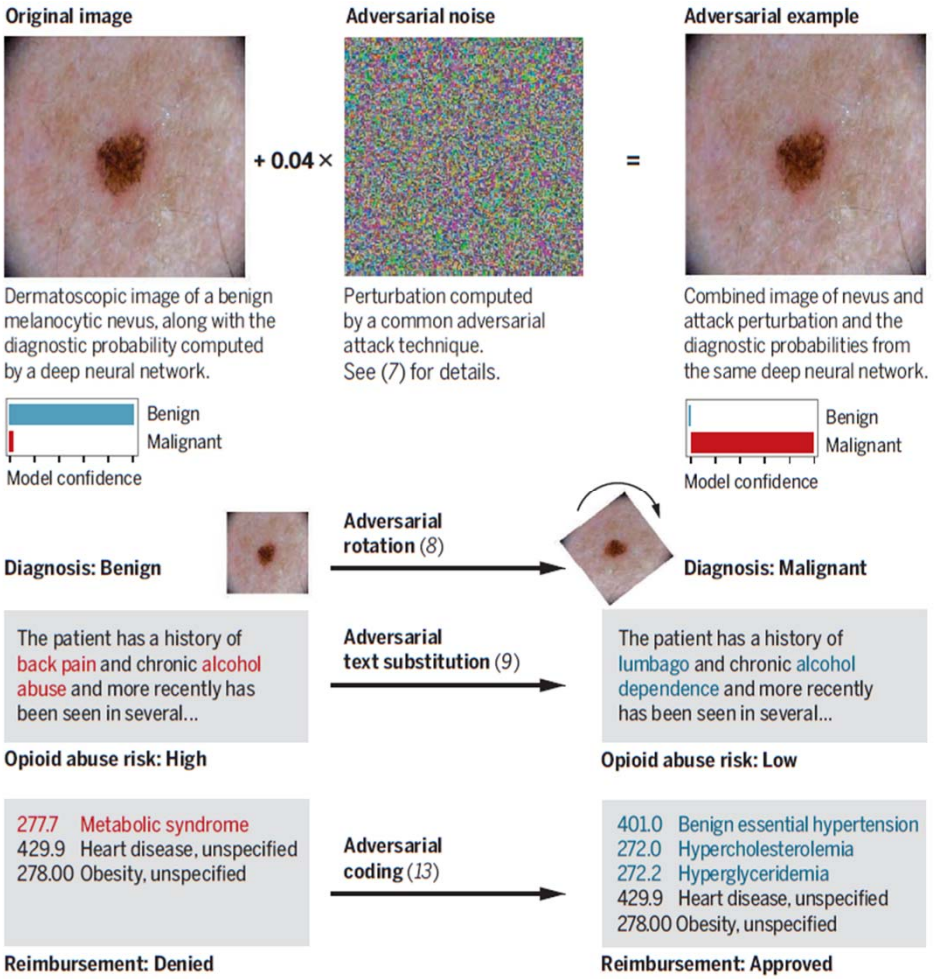
Attacking autonomous driving, AI mistakes can cause Life.

Jan Hendrik Metzen, et al. , Universal Adversarial Perturbations Against Semantic Image Segmentation, 2017.
N. Carlini and D. Wagner, Audio Adversarial Examples: Targeted Attacks on Speech-to-Text, 2018.

Applications

- **Adversarial examples in medical AI systems**

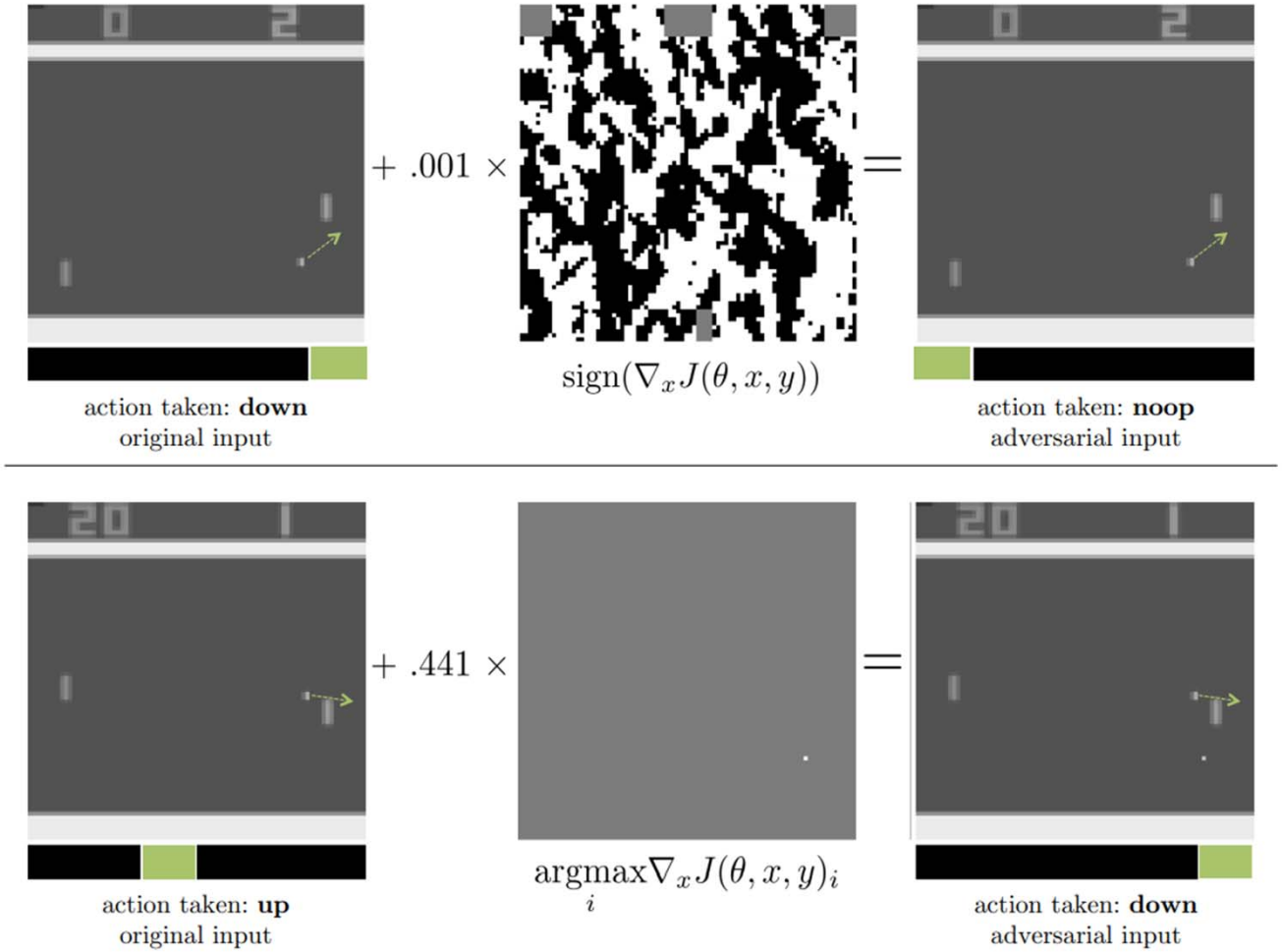
They might be executed ***without*** requiring any overtly fraudulent misrepresentation of the data



Samuel G. Finlayson, et al., Adversarial Attacks on Medical Machine Learning, Science, 2019

Applications

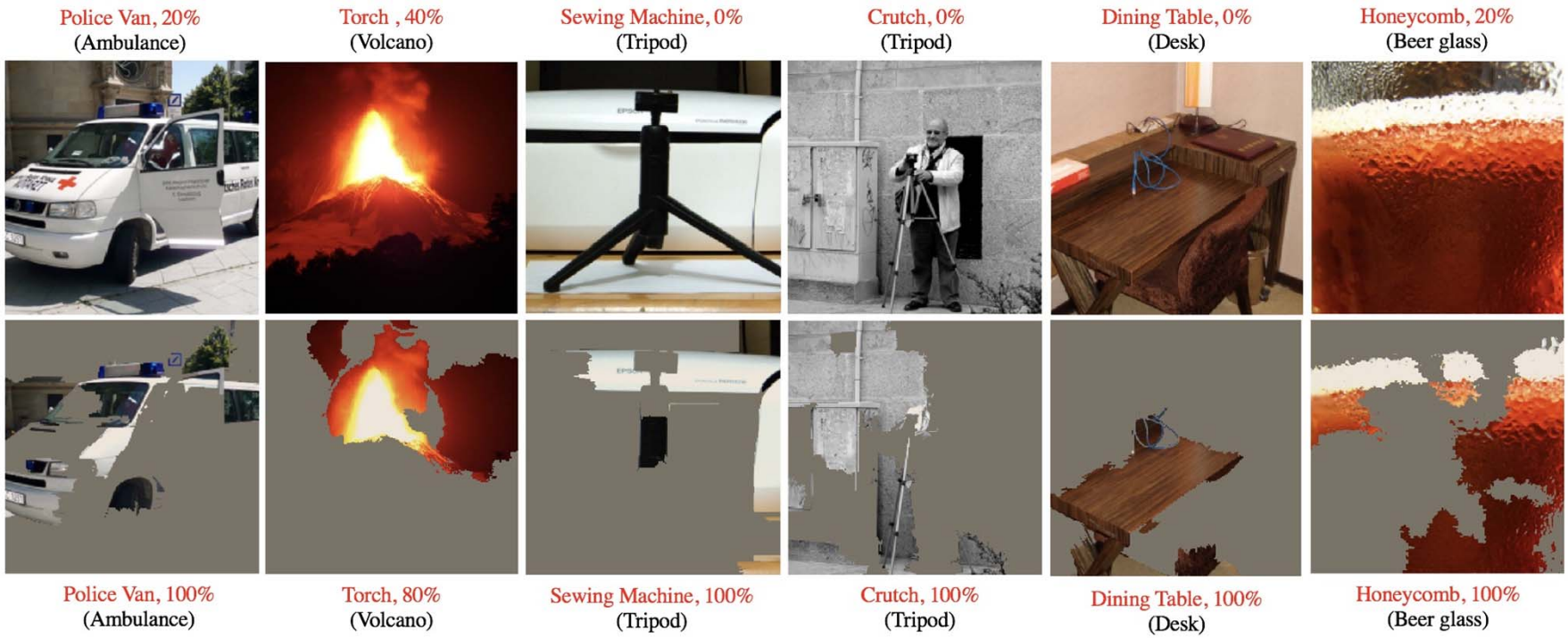
- **Adversarial Attacks in Reinforcement Learning**



Sandy Huang, et al. Adversarial Attacks on Neural Network Policies , 2017

Adversarial Sample Usages

- **Explore the biases of a neural network by analysing the distance of a sample to the decision boundary using adversarial samples**
- **The distance to the decision boundary is closely related to the magnitude of the perturbation necessary to make a sample cross it**



Pierre Stock and Moustapha Cisse, ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases, 2018

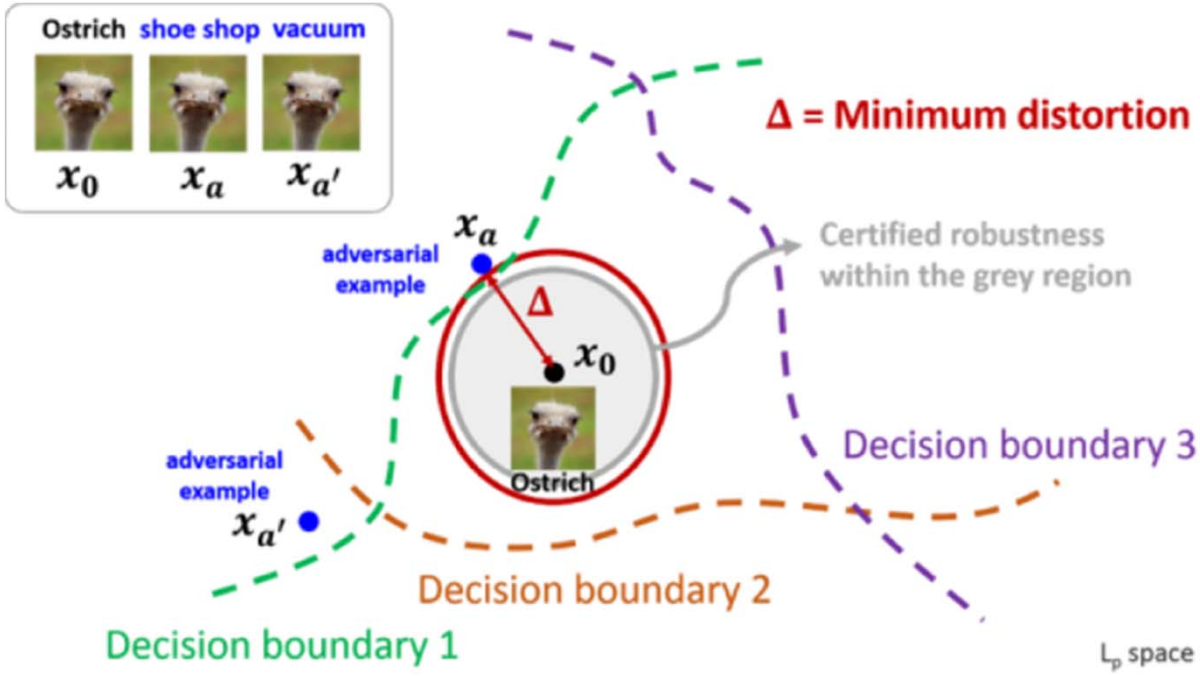
Adversarial Sample Usages

- **The adversarial perturbation exploits the ambiguity of the image by shifting the attention of the model towards regions supporting the adversarial prediction**

Left: an adversarial image of the true class *Jeep* predicted as Ambulance by the network.
Center: the explanation of the clean image for its prediction (*Jeep*).
Right: the explanation of the adversarial image for its prediction (*Ambulance*).

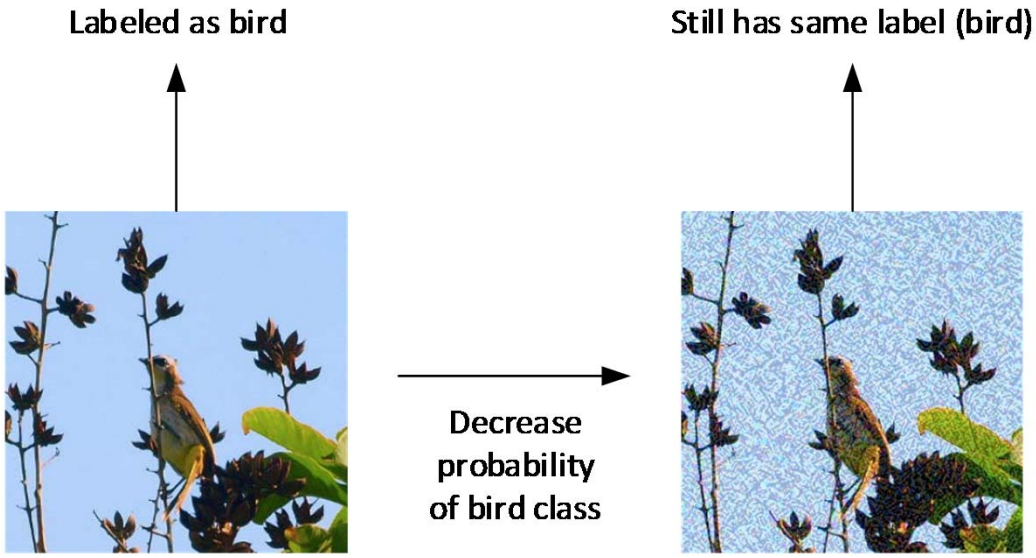
Adversarial Sample Usages

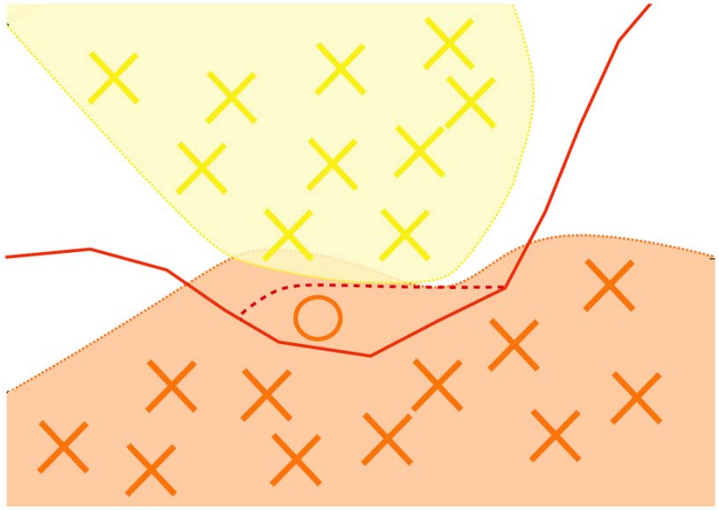
- **Measuring Robustness**
  - ✓ **Accuracy and attack success rate**
  - ✓ **Transferability**
  - ✓ **Distortion and minimal perturbation:** Amount of noise required for attack to succeed
  - ✓ **Loss sensitivity:** Rough estimate of the Lipschitz continuity of the model
  - ✓ **CLEVER:** Lower bound on adversarial distortion based on extreme value theory



Adversarial Sample Usages

- **Defense: Adversarial Training**

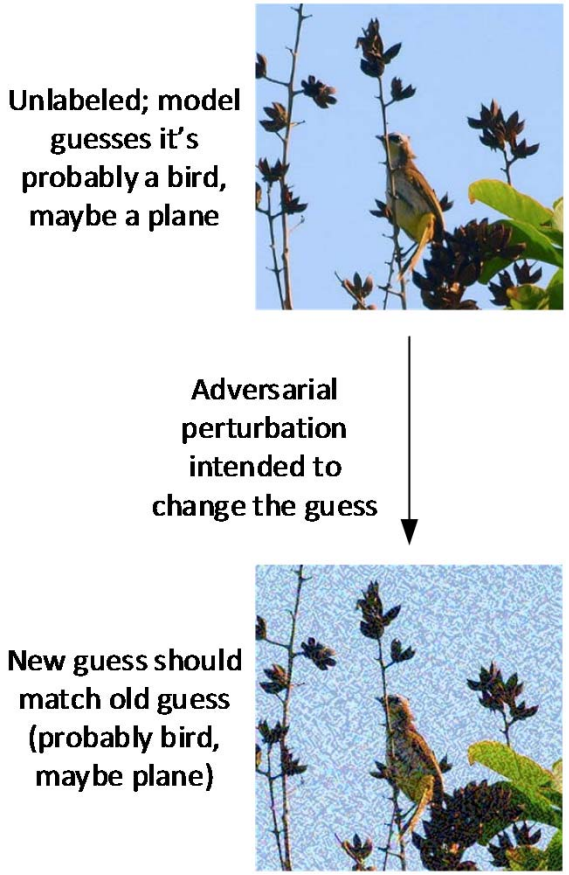


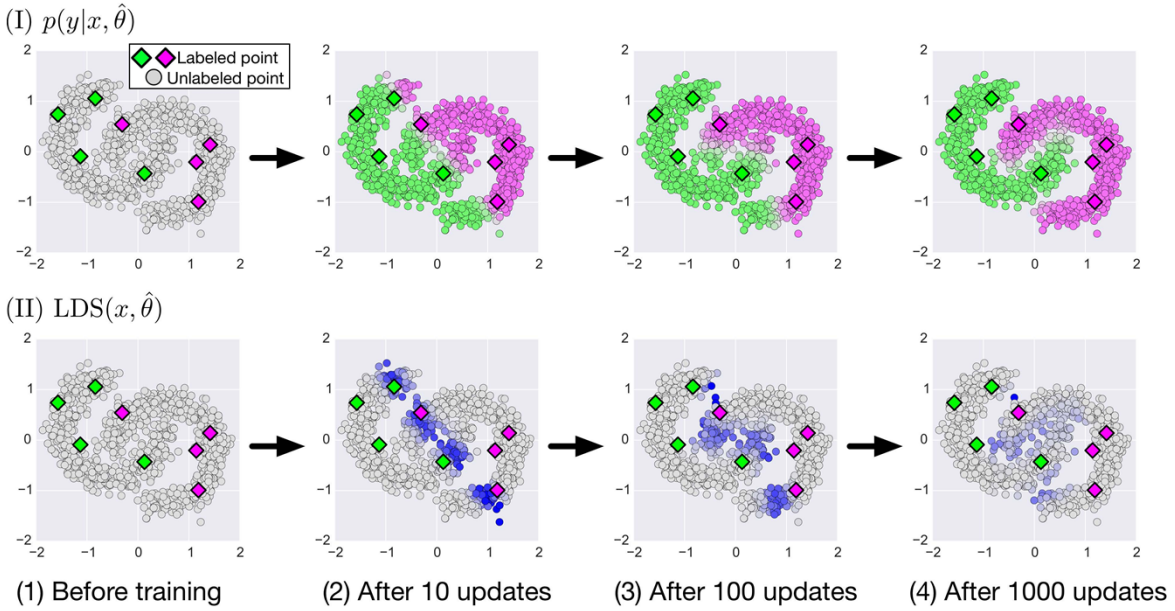


Adapt the classifier to attack directions by including adversarial data at training

Adversarial Sample Usages

- **Adversarial Training for Semi-Supervised Learning**





To make the same prediction on an unlabeled image and its adversarial counterpart

Takeru Miyato, et al., Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning, 2018

Content

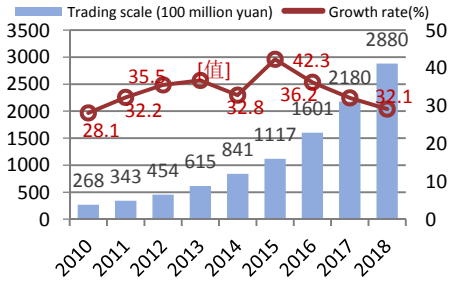
Background

The Exploding Big Data

Big Data from Media

Approximately 61,744 GB data is generated per second worldwide

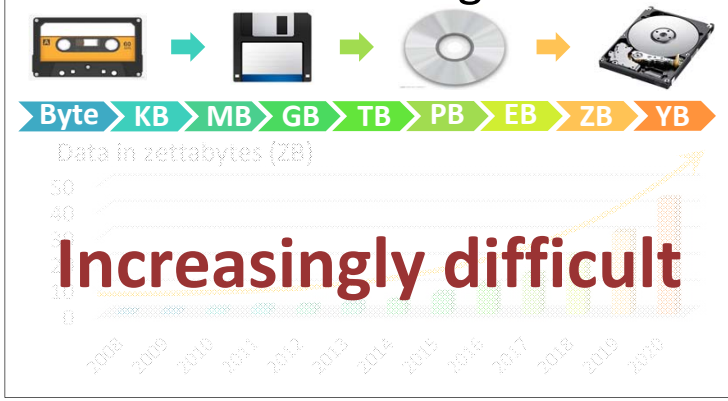
Big Data from Education



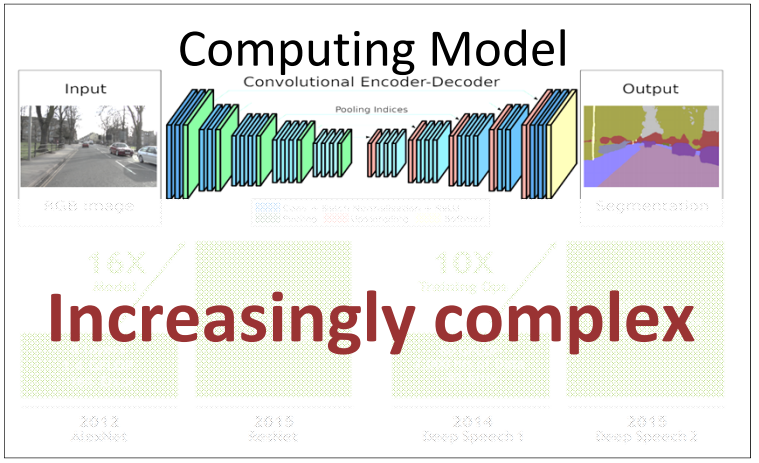
Netease's open audio and video courses reached 150,000

Great Challenges in Big Data

Data Storage



Computing Model



Background

The Bell Tower of Xi'an (simplified Chinese: 西安钟楼; traditional Chinese: 西安鐘樓; pinyin: Xī'ān Zhōnglóu), built in 1384 during the early Ming Dynasty, is a symbol of the city of Xi'an and one of the grandest of its kind in China. The Bell Tower also contains several large bronze-cast bells from the Tang Dynasty. The tower base is square and it covers an area of 1,377 m² (14,820 sq ft). The tower is a brick and timber structure and close to 40 m (130 ft) high. It is located in the center of Xi'an, at the intersection of the four streets of the east, west, south and north. It is the largest and most preserved one amongst the many bell towers left over from ancient China.

Supervised hashing
SSH, MLH, KSH, FastH, SDH, COSDISH…

Unsupervised hashing
PCAH, SH, ITQ, AGH, IMH, DGH…

0 1 1 0 0 0 1 0

$$x : d \times 1 \xrightarrow{d \gg r} b \in \{0,1\}^r$$

Binary code generation

0	1	0	1	0	0	1	0	3
0	1	0	0	1	1	1	0	2
0	1	0	0	0	1	1	1	1
0	1	0	1	0	1	1	1	

XOR operation

Hamming distance calculation

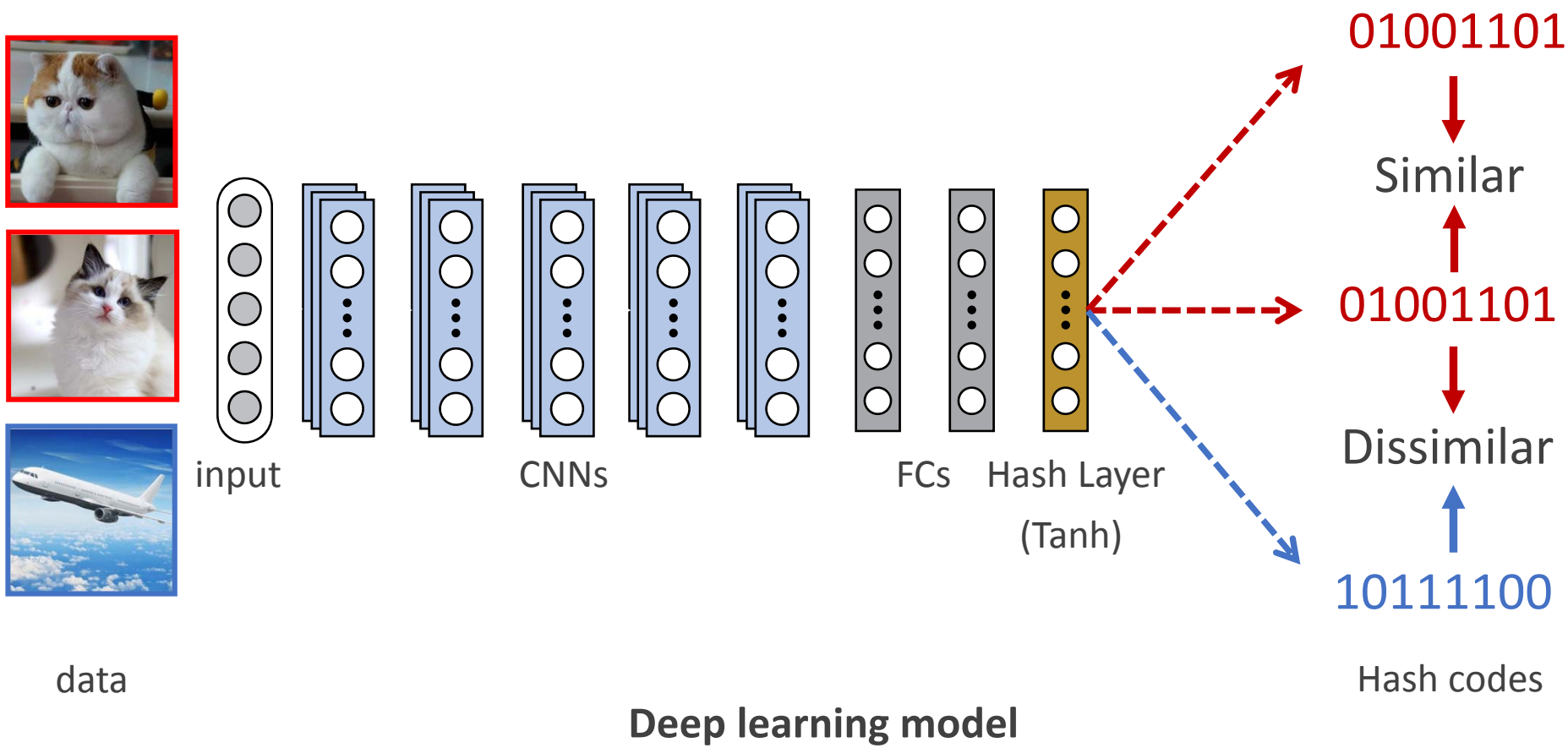
Data Scale: High-dimensional → Low-dimensional

Cheaper storage cost

Lower computational load

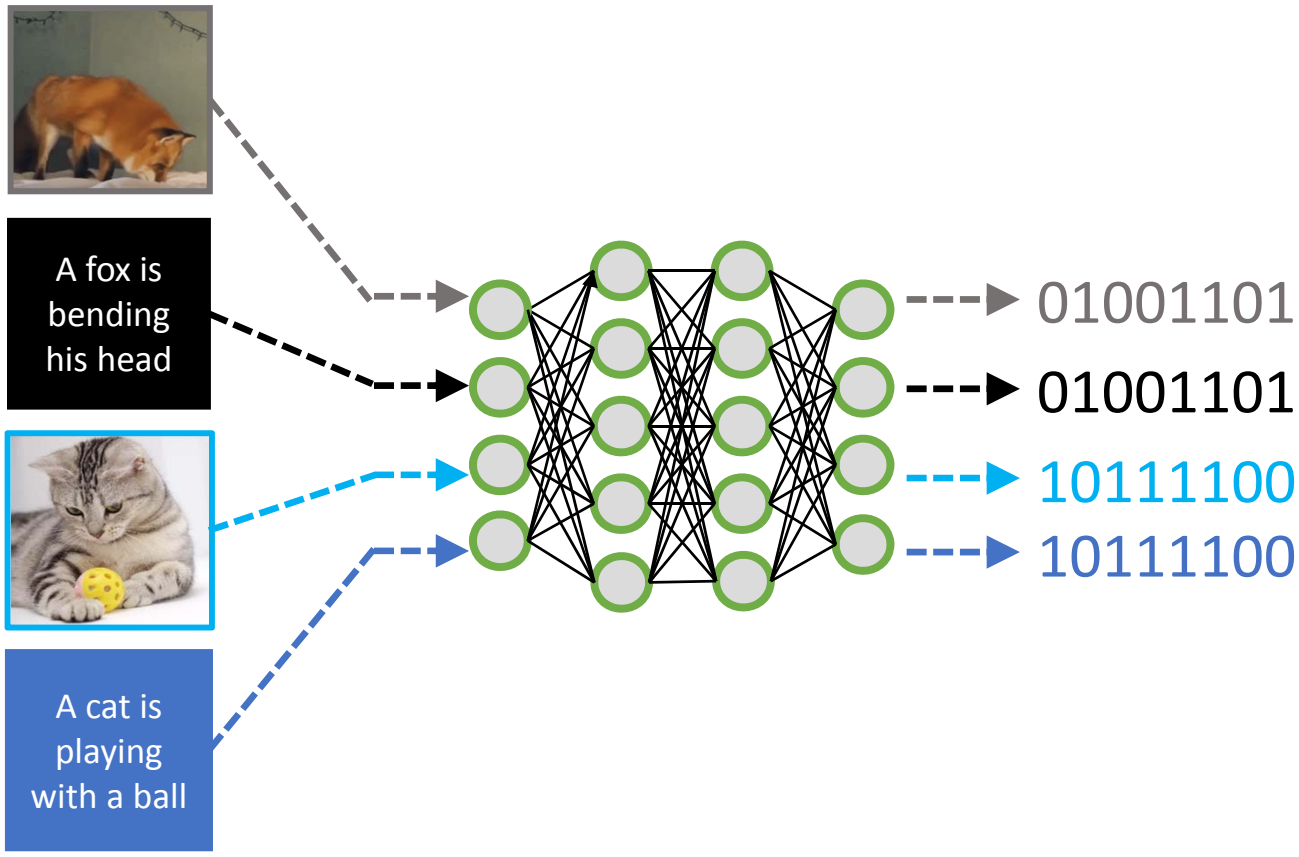
Deep Hashing

- **Single-Modal Hash Learning**



Deep Hashing

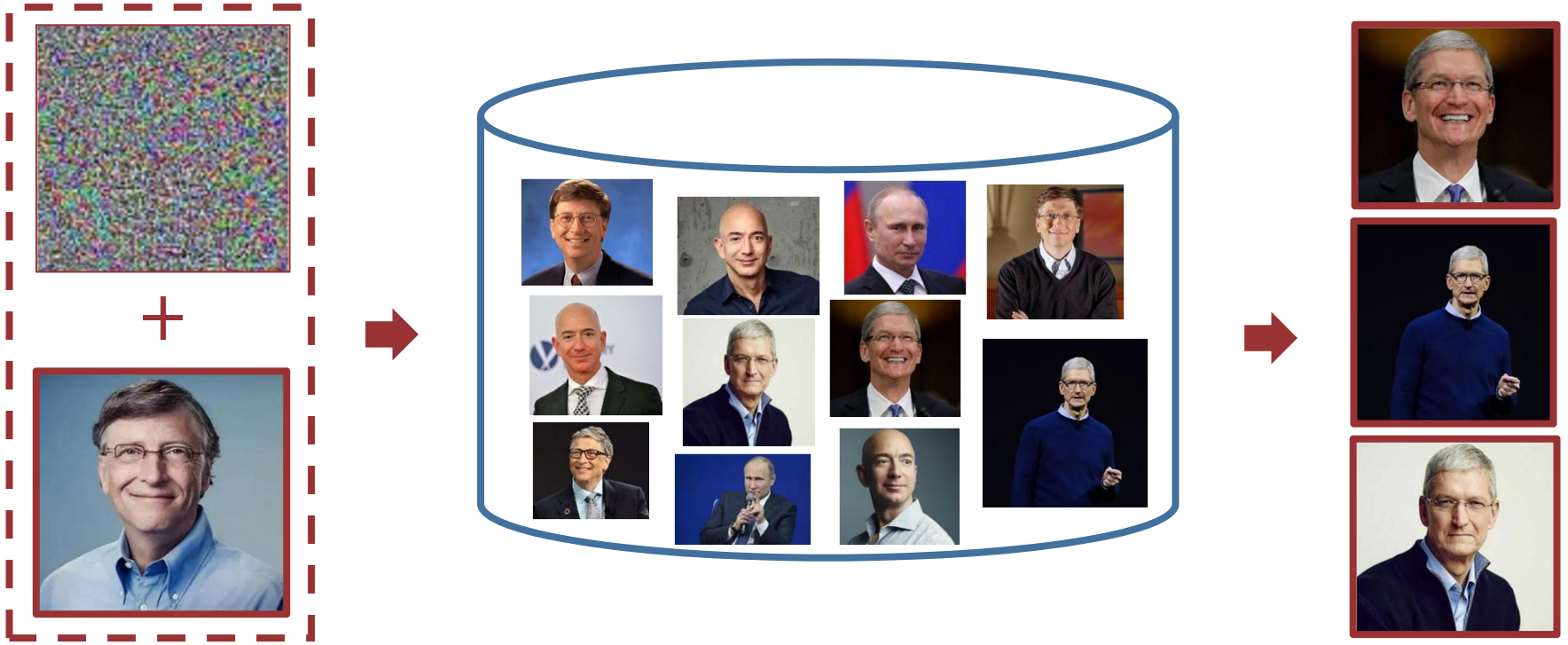
- **Cross-Modal Hash Learning**



Content

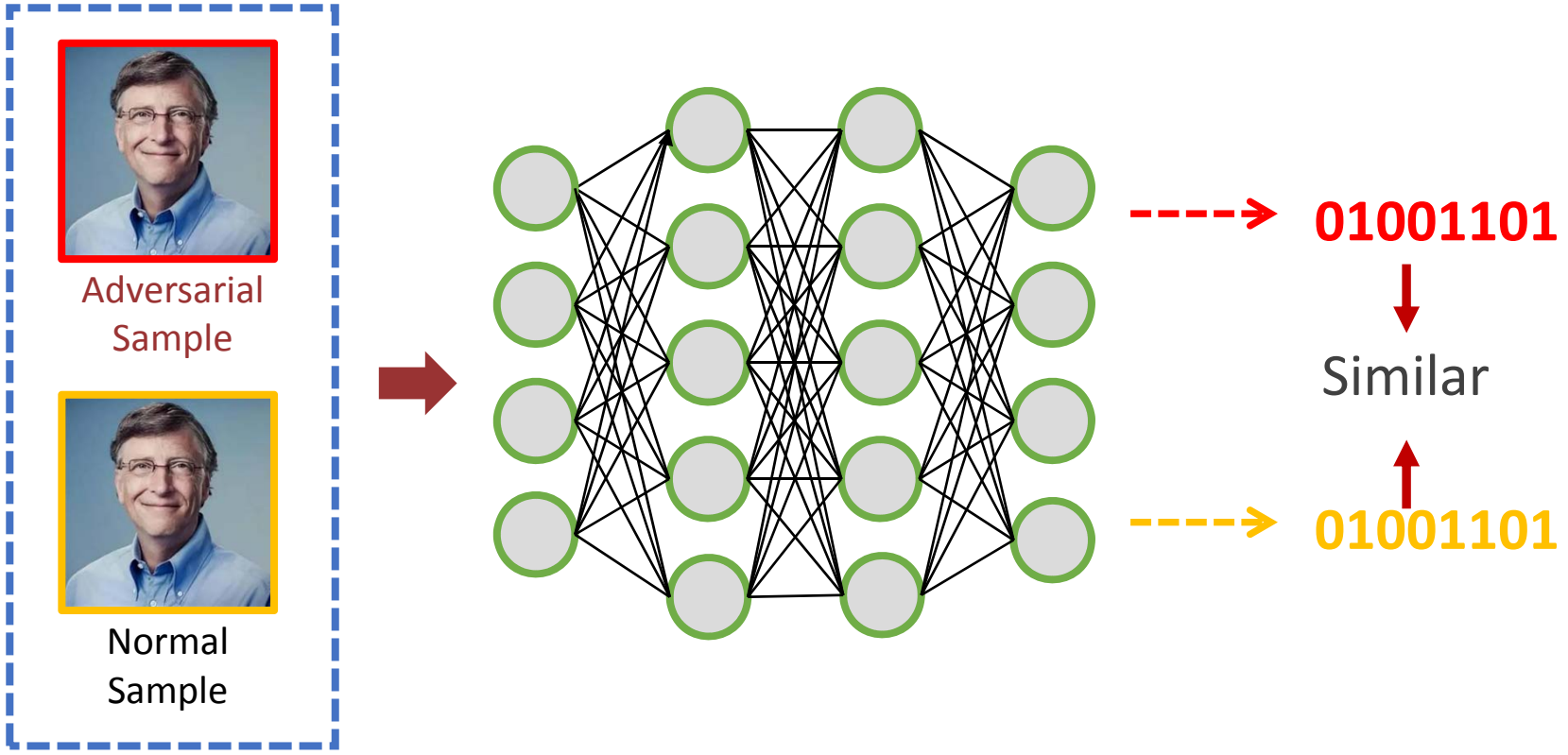
Adversarial Samples in Deep Hashing

- Potential **security** problem for information retrieval



Adversarial Samples in Deep Hashing

- **Robust** information retrieval through **adversarial training**



Adversarial Examples for Hamming Space Search

TCYB2018: Erkun Yang, Tongliang Liu, Cheng Deng, Dacheng Tao

Motivation

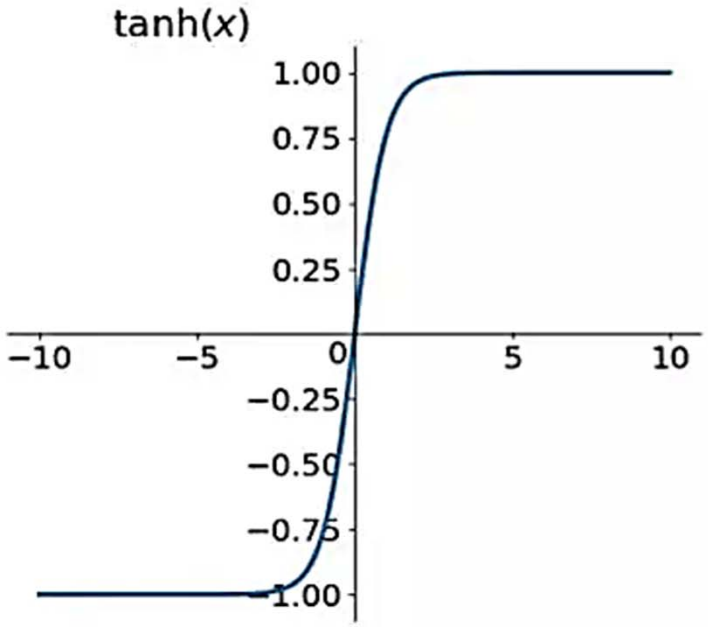
- **Our approach handles the problems**
  - ✓ **Problem 1:** ***The lack of query image's label*** caused retrieving semantic irrelevant results rapidly become infeasible

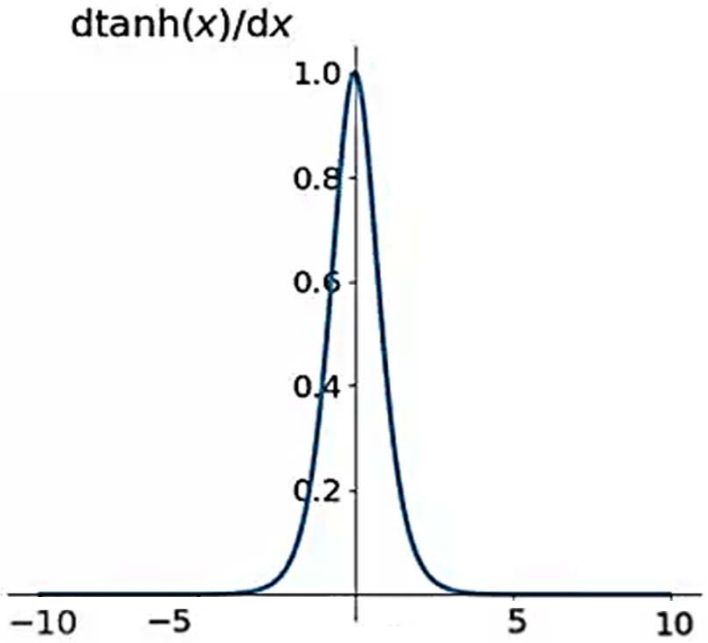
We opt to design a novel objective to force the hash codes for adversarial samples to be dissimilar from those of the original examples

Motivation

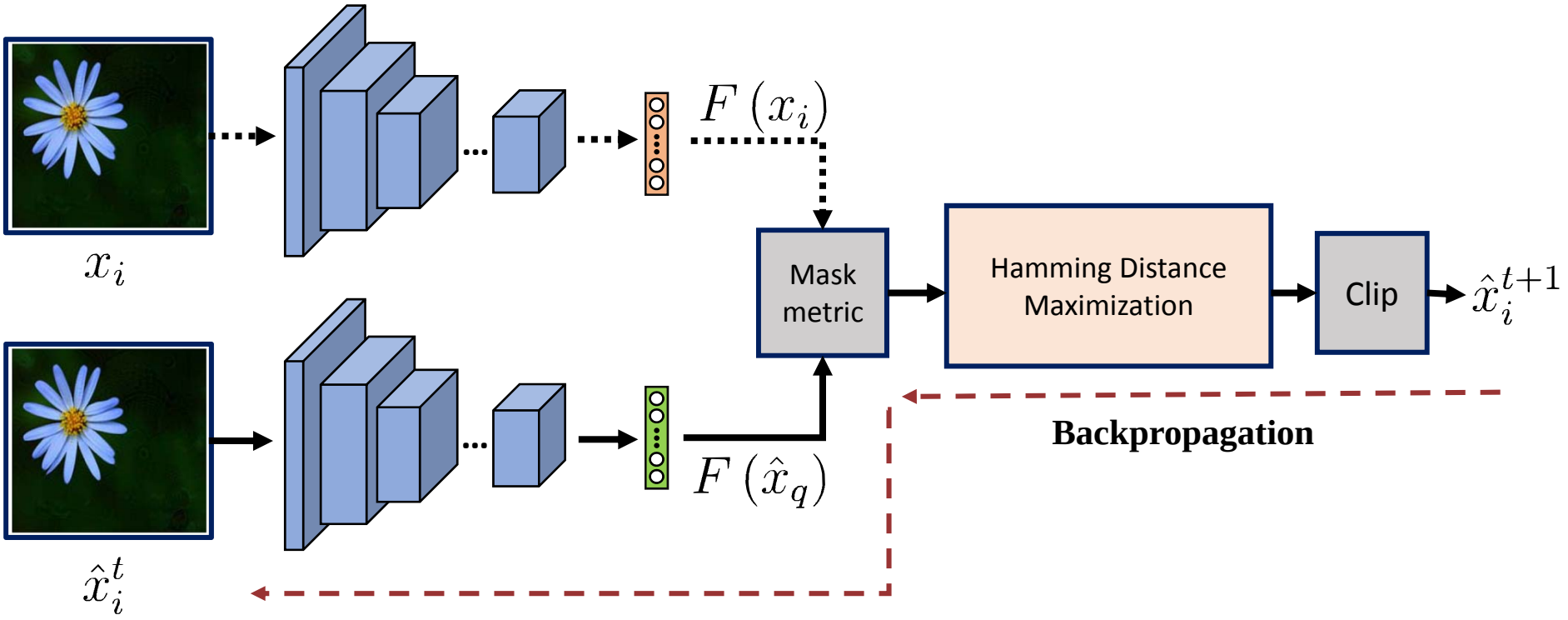
- **Our approach handles the problems**
  - ✓ **Problem 2:** The existing methods have the ***vanishing gradient problem***

We proposed a novel updating strategy for the hash activation function





Proposed Method



The Framework of HAG

Proposed Method

- **Overall Learning Objectives in each iteration**

✓ The targeted hashing model

$$b_i = E(x_i) = sign(H(x_i))$$

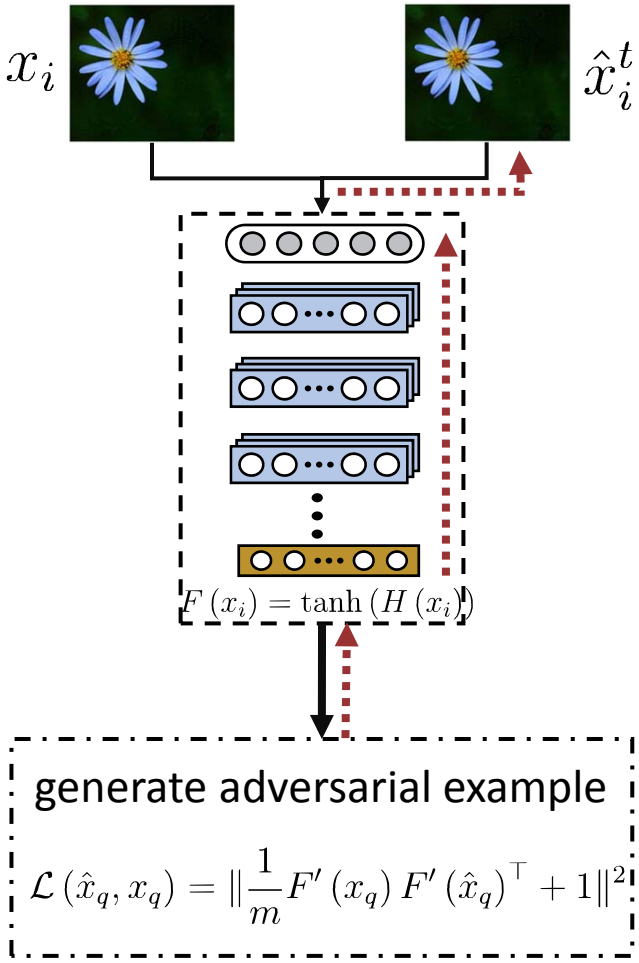
b_i can be approximated by

$$F(x_i) = \tanh(\alpha H(x_i))$$

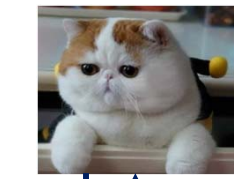
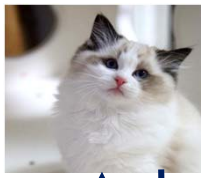
✓ Generate adversarial example

- the perturbation should be small
- $\hat{x}_q$ should be dissimilar to x_q

$$\min_{\hat{x}_q} \mathcal{L}(\hat{x}_q, x_q) = \|\frac{1}{m} F'(x_q) F'(\hat{x}_q)^\top + 1\|^2$$



Proposed Method

Semantic Similar

01001101 01001101


Semantic Similar

01001101 10111100

maximize


Semantic Dissimilar

01001101 10111100

Semantically similar pairs will have similar hash codes, vice versa

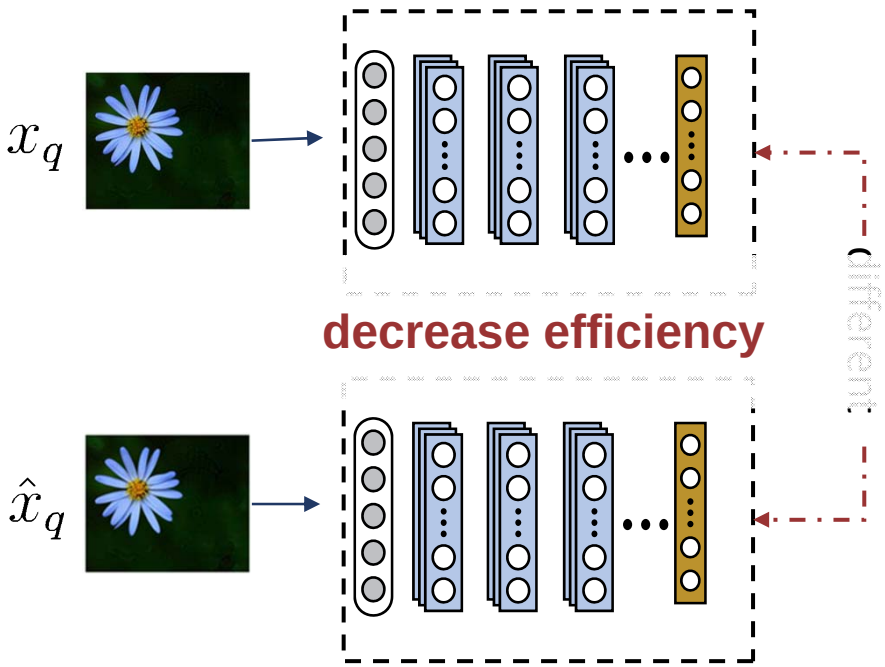
maximize the Hamming distance

$$\max d(x, y) = \sum x_i \oplus y_i$$

minimize the inner products of hash codes

$$\min_{\hat{x}_q} \mathcal{L}(\hat{x}_q, x_q) = \| \frac{1}{m} F'(x_q) F'(\hat{x}_q)^\top + 1 \|^2$$

Proposed Method



- The mask for the q^{th} sample

$$w_{qi} = \begin{cases} 1, \text{if} \left| F_i\left(\hat{x}_q\right) - sign\left(F_i\left(x_q\right)\right) \right| < 1 + t \\ 0, \text{otherwise} \end{cases}$$

- The corresponding vectors

$$F'\left(x_q\right) = w_q \odot F\left(x_q\right)$$

$$F'\left(\hat{x}_q\right)^\top = w_q \odot F\left(\hat{x}_q\right)^\top$$

$$\min_{\hat{x}_q} \mathcal{L}\left(\hat{x}_q, x_q\right) = \left\| \frac{1}{m} F\left(x_q\right) F\left(\hat{x}_q\right)^\top + 1 \right\|^2 \rightarrow \min_{\hat{x}_q} \mathcal{L}\left(\hat{x}_q, x_q\right) = \left\| \frac{1}{m} F'\left(x_q\right) F'\left(\hat{x}_q\right)^\top + 1 \right\|^2$$

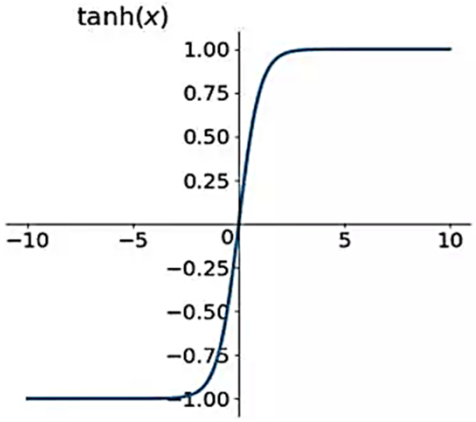
Proposed Method

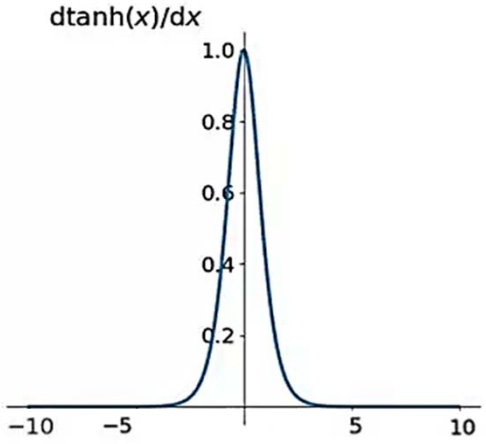
- **Vanishing gradient problem**
  - ✓ Zero gradients in the hash activation layer will greatly hinder the optimization of the adversarial examples

$$\frac{\partial L}{\partial \hat{x}_q} = \frac{\partial L}{\partial F\left(\hat{x}_q\right)} \frac{\partial F\left(\hat{x}_q\right)}{\partial H\left(\hat{x}_q\right)} \frac{\partial H\left(\hat{x}_q\right)}{\partial \hat{x}_q}$$

$$= \frac{\partial L}{\partial F\left(\hat{x}_q\right)} \frac{\partial \tanh\left(H\left(\hat{x}_q\right)\right)}{\partial H\left(\hat{x}_q\right)} \frac{\partial H\left(\hat{x}_q\right)}{\partial \hat{x}_q}$$

$$\frac{\partial \tanh\left(H\left(\hat{x}_q\right)\right)}{\partial H\left(\hat{x}_q\right)} = 1 - \left(\tanh\left(H\left(\hat{x}_q\right)\right)\right)^2$$



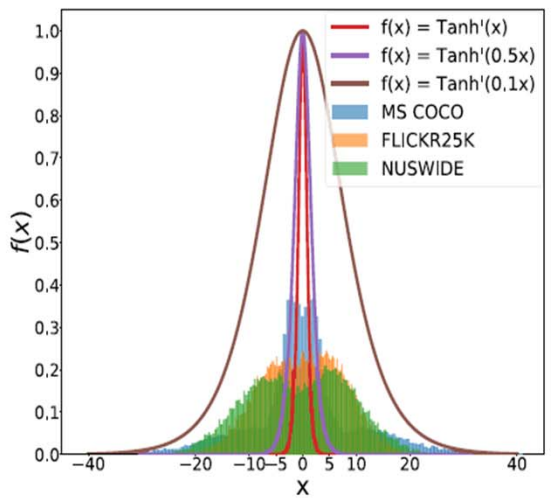


$F\left(x_i\right) = \tanh\left(H\left(x_i\right)\right)$ will cause **vanishing gradient** problem

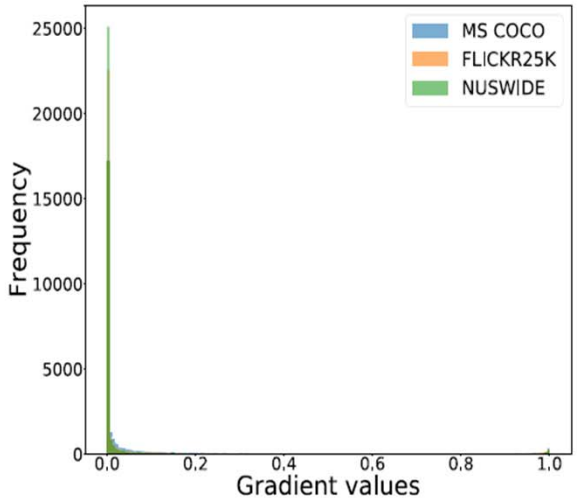
Proposed Method

- **The hash activation function**

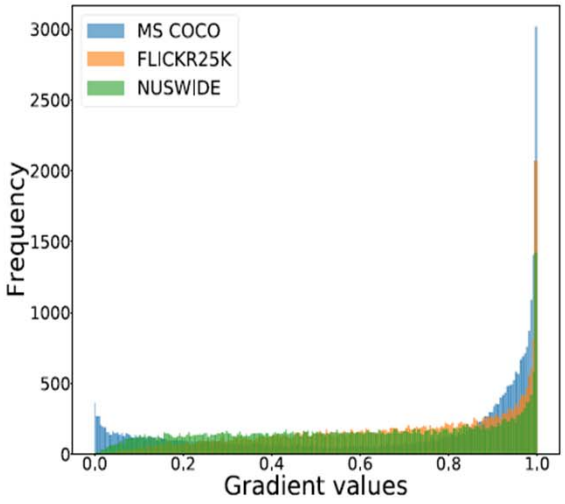
$$F(x_i) = \tanh(H(x_i)) \longrightarrow F(x_i) = \tanh(\alpha H(x_i))$$



(a)



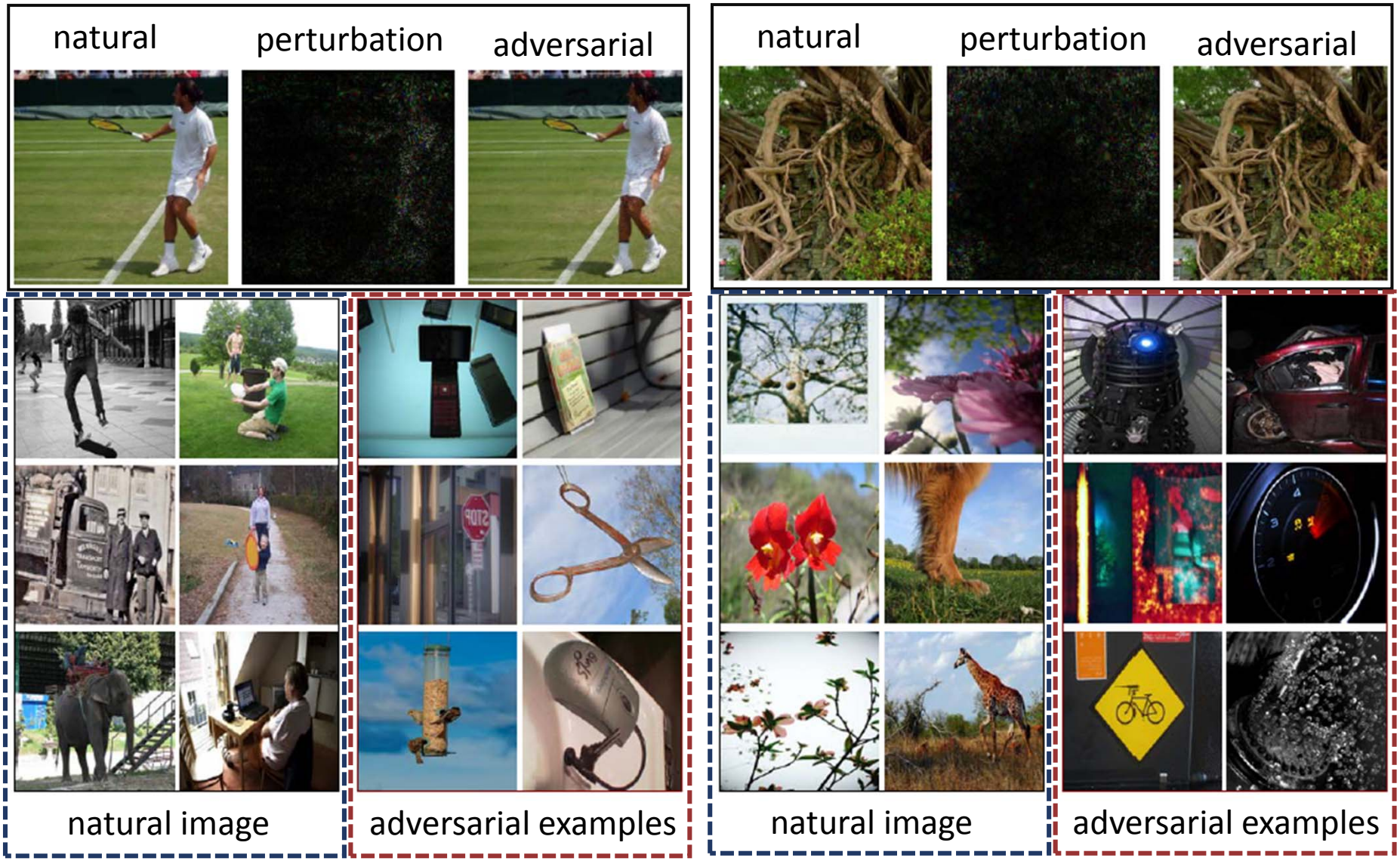
(b)



(c)

Method	FLICKR25K			
	16 bits	32 bits	64 bits	128 bits
HAG*	0.3004	0.1963	0.2410	0.1851
HAG	**0.1657**	**0.1164**	**0.1190**	**0.1315**

Results


natural
perturbation
adversarial
natural
perturbation
adversarial
natural image
adversarial examples
natural image
adversarial examples

Results

Table: Map for a different number of bits on the datasets at different iterations

Training Iteration	MS-COCO				FLICKR25K			
	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
0	0.7114	0.7365	0.7460	0.7494	0.9280	0.9281	0.9262	0.9295
10	0.6920	0.7163	0.7237	0.7353	0.8969	0.8954	0.9069	0.9130
100	0.4208	0.4451	0.4589	0.4890	0.5535	0.5221	0.5425	0.5953
500	0.2115	0.2072	0.2046	0.2323	0.1836	0.1496	0.1501	0.1712
1000	0.1617	0.1398	0.1392	0.1698	0.1611	0.1274	0.1164	0.1293
1500	0.1125	0.0979	0.1010	0.1144	0.1731	0.1188	0.1075	0.1352
2000	0.1081	0.0973	0.0916	0.0998	0.1657	0.1164	0.1190	0.1315

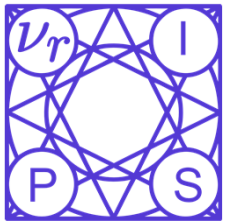
MAP is one of the most widely used criteria for measuring the performance of modern retrieval systems

HAG can **successfully attack** targeted hashing models !

Results

Transfer MAP results on MS-COCO datasets

Perturbations	**DPH-VGG16**	**DPH-VGG16***	**DPH-VGG19**	**DPH-VGG19***	**HashNet-ResNet50**	**HashNet-ResNet50***
None	0.7365	0.7460	0.7279	0.7458	0.8042	0.8536
DPH-VGG16 (r_1)	0.0973	0.1138	0.4392	0.5374	0.7640	0.8043
DPH-VGG16* (r_2)	0.1036	0.0916	0.4166	0.5117	0.7673	0.8008
DPH-VGG19 (r_3)	0.5015	0.4968	0.0790	0.1197	0.7603	0.8040
DPH-VGG19* (r_4)	0.4707	0.4931	0.0988	0.1059	0.7573	0.8035
$r_1 + r_2$	0.0808	0.0753	0.1890	0.2535	0.7034	0.7227
$r_1 + r_3$	0.1002	0.1030	0.0830	0.1110	0.6995	0.7413
$r_1 + r_4$	0.0944	0.0984	0.0902	0.1106	0.7126	0.7361
$r_1 + r_2 + r_3 + r_4$	0.0776	0.0720	0.0703	0.0803	0.6252	0.6609

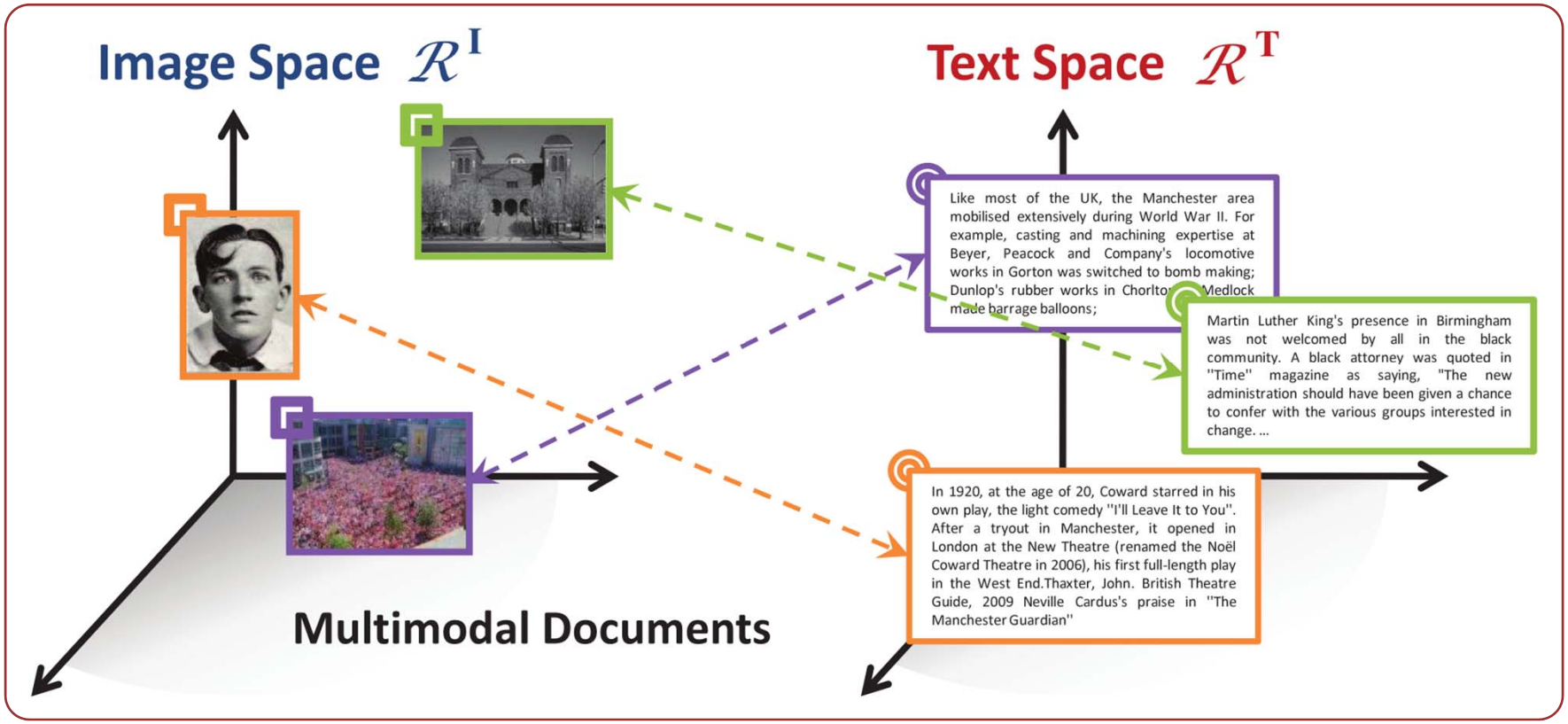


Cross-Modal Learning with Adversarial Samples

NeurIPS 2019: Chao Li, Cheng Deng, Shangqian Gao, De Xie, Wei Liu

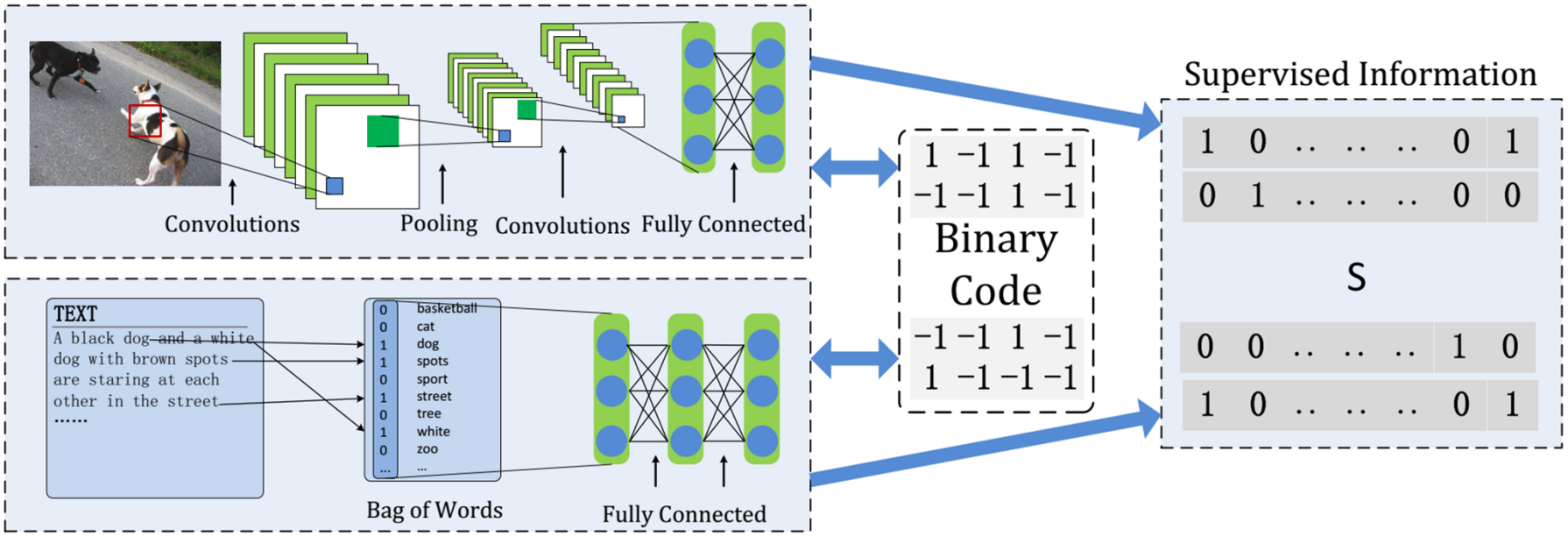
Motivation

Cross-Modal Learning aims to explore ***the Relation***
between one modality (e.g., image) and another (e.g., text)



J. C. Pereoa, et al., On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval, TPAMI, 2014

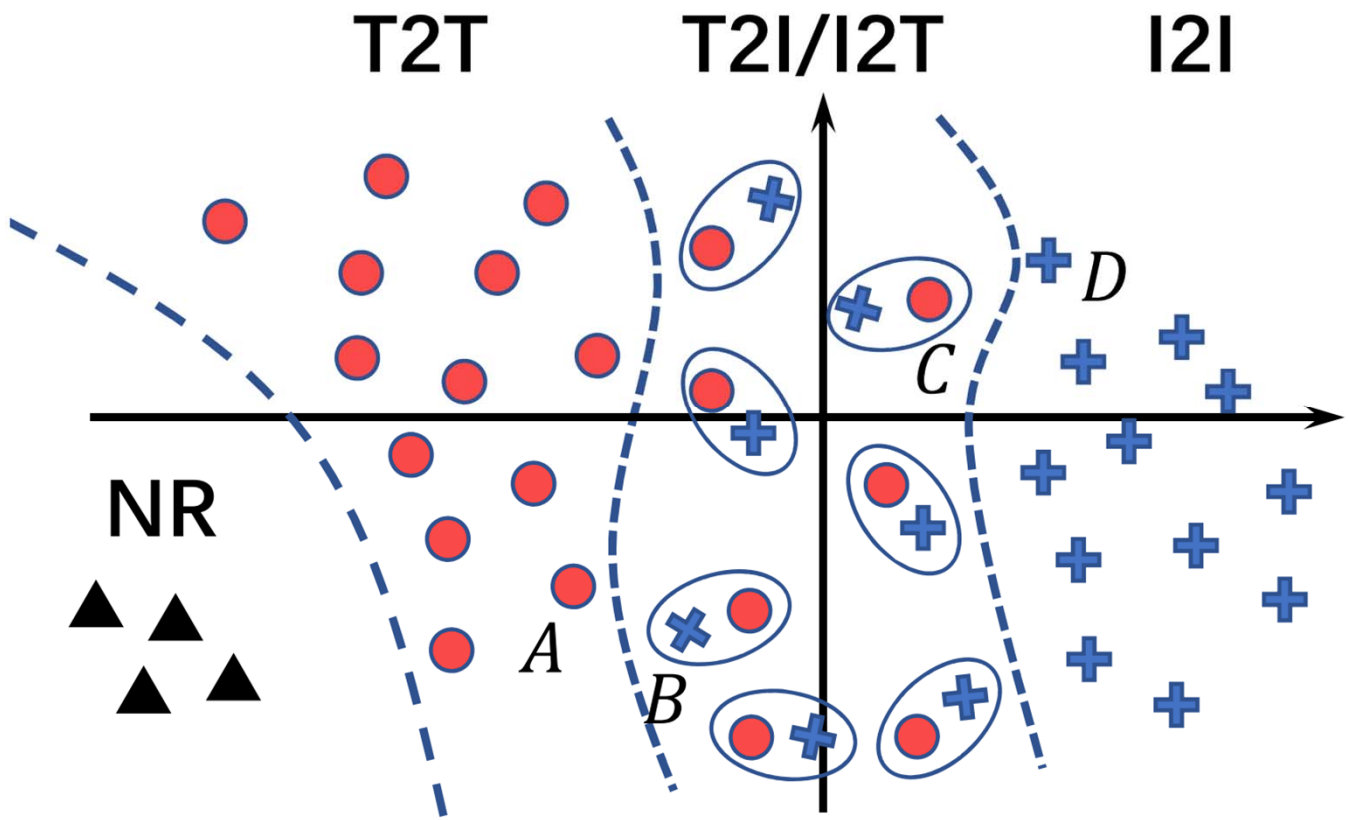
Motivation



Data points from two modalities are mapped from the original spaces into a ***Hamming Space of Binary Codes***

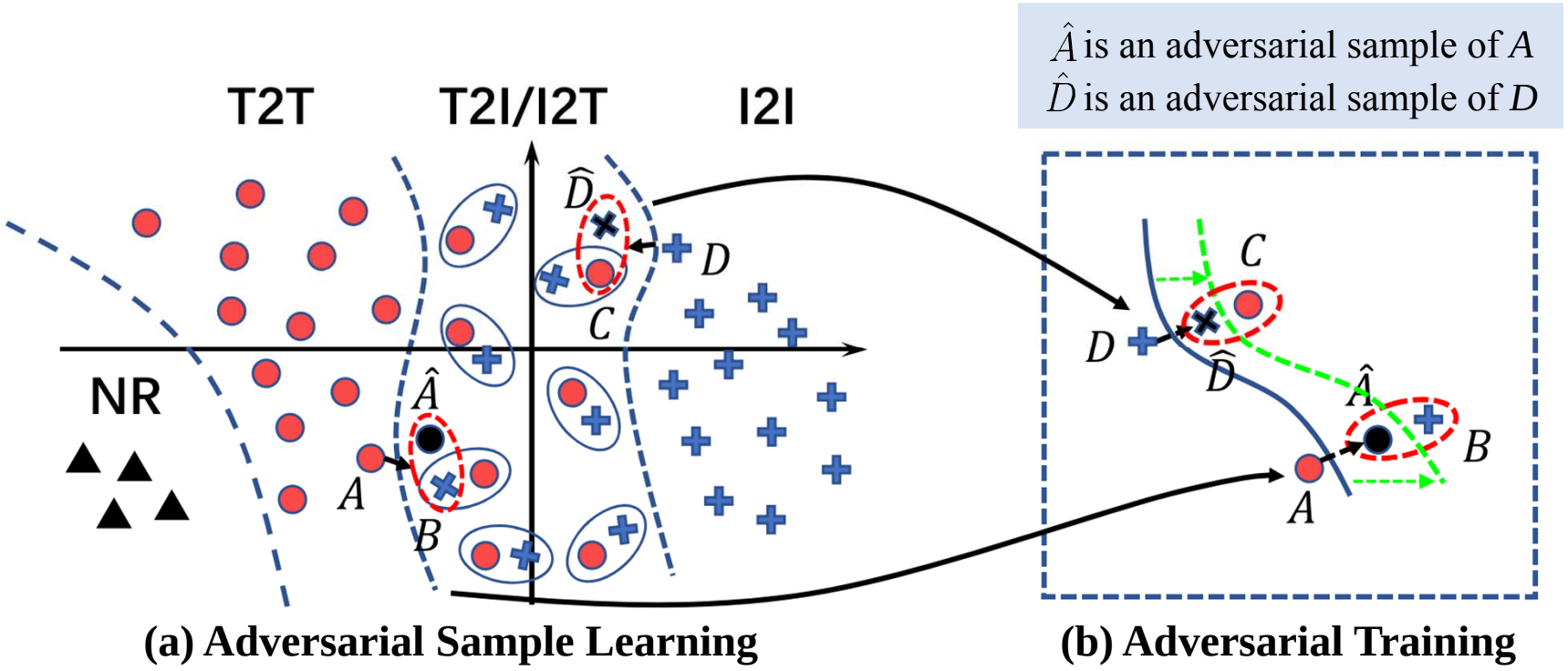
Q.Y. Jiang and W.-J. Li, Deep Cross-Modal Hashing, CVPR2017

Motivation



Cross-Modal Search Space

Proposed Method

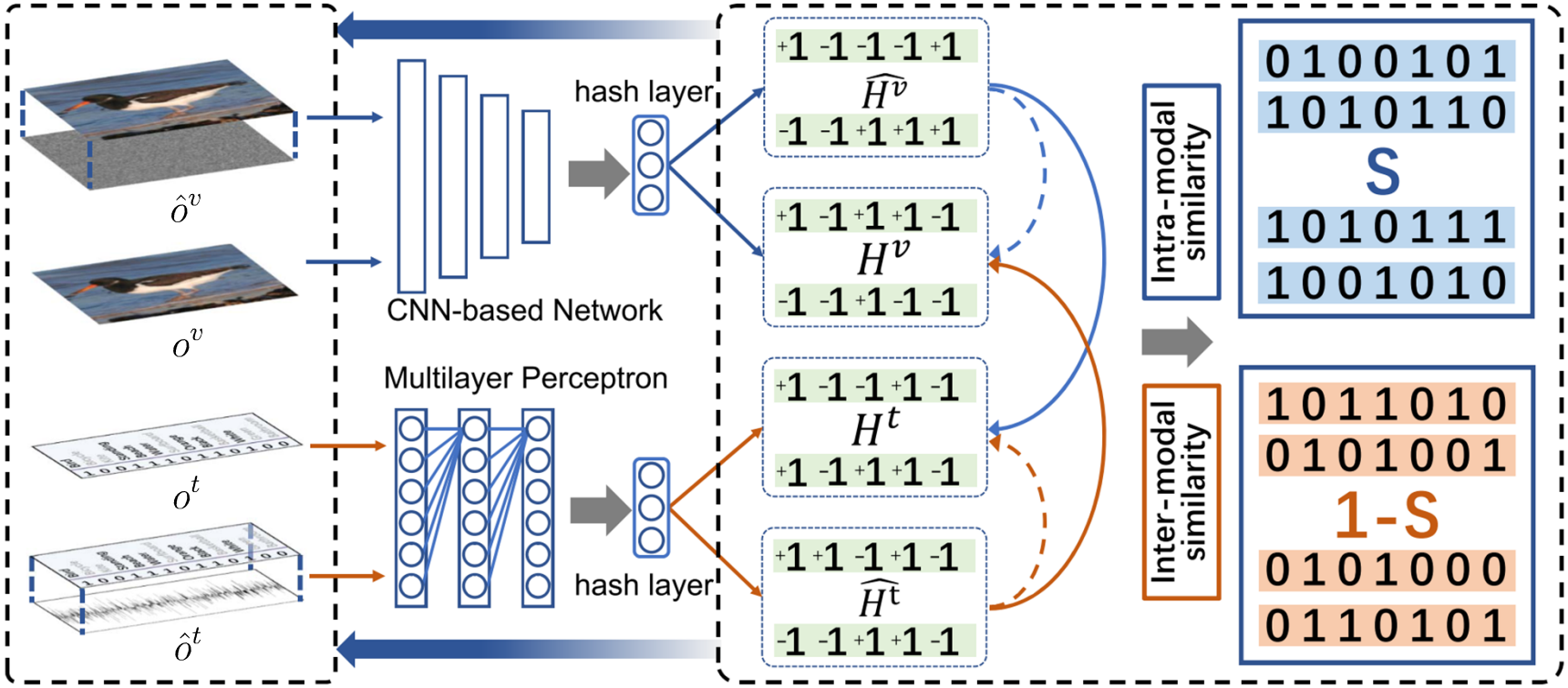


(a) Adversarial Sample Learning **(b) Adversarial Training**

Adversarial sample is defined in two aspects:

- Learned perturbation is designed to fool a deep cross-modal network
- The learned adversarial samples won't impact the retrieval performance within its own modality

Proposed Method

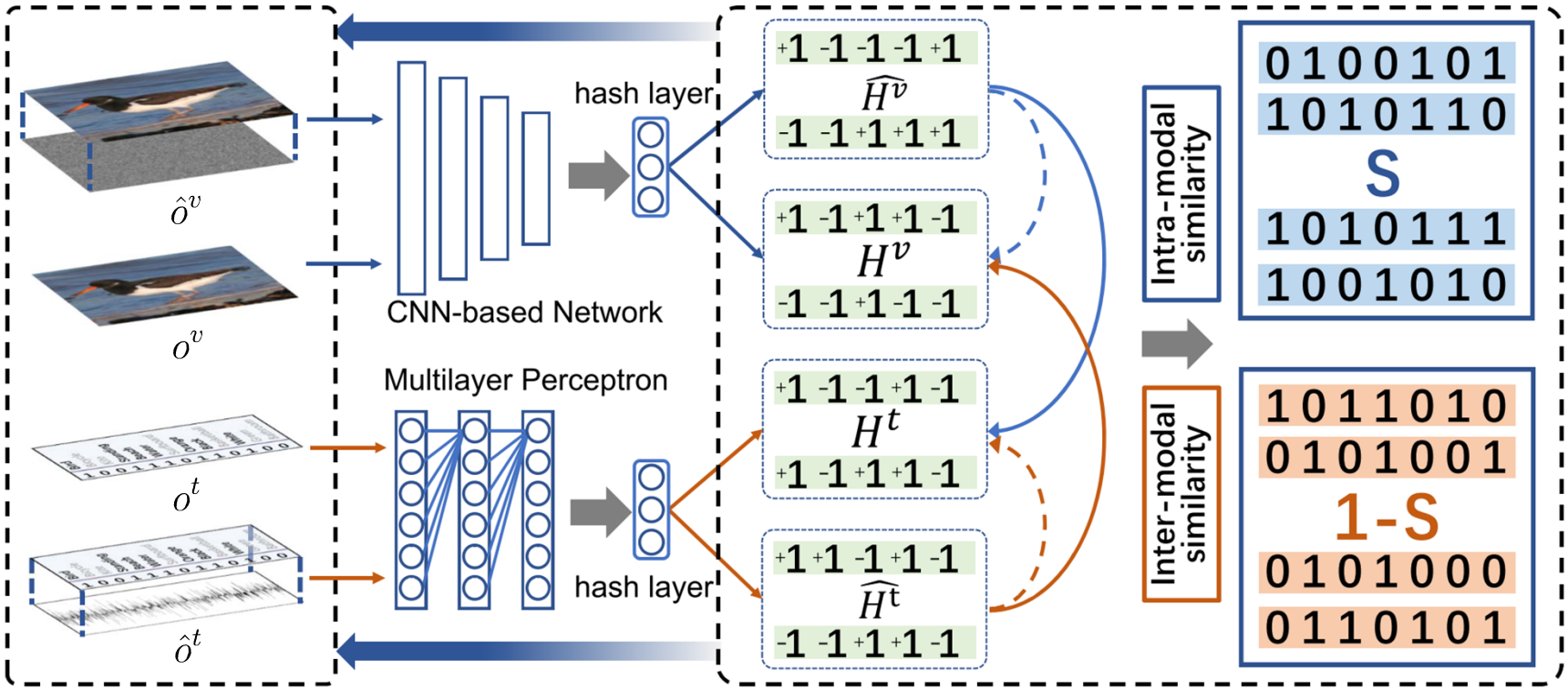


The aim of an ***adversarial attack*** is to find the minimum perturbations Δ^* in v and t that result in the change of retrieval accuracy:

$$\Delta\left(o*,\mathcal{H}*\right):=\min_{\delta*}\|\delta*\|p,$$

$$s.t.\arg\max_{\theta}*\left(\mathcal{H}\left(o*+\delta*;\theta*\right)\right)\neq\arg\max_{\theta}*\left(\mathcal{H}\left(o*;\theta*\right)\right),*\in\{v,t\}.$$

Proposed Method



Inter-modal similarity :

$$\min_{\delta *} \mathcal{J}_{inter} = \sum_{i,j=1}^{n} \left\| (1 - S_{ij}) \, \Gamma_{ij} - \log\left(1 + e^{\Gamma_{ij}}\right) \right\|^2 + \left\| \hat{o}^v - o^v \right\|_p, s.t. \Gamma_{ij} = \frac{1}{2} \left(\hat{H}_i^v \right) \left(H_j^t \right)^T .$$

Intra-modal similarity :

$$\min_{\delta *} \mathcal{J}_{inter} = \sum_{i,j=1}^{n} \left\| S_{ij} \Theta_{ij} - \log\left(1 + e^{\Theta_{ij}}\right) \right\|^2, s.t. \Theta_{ij} = \frac{1}{2} \left(\hat{H}_i^v \right) \left(H_j^v \right)^T .$$

Results

Comparison in terms of ***MAP scores*** with different lengths of hash codes

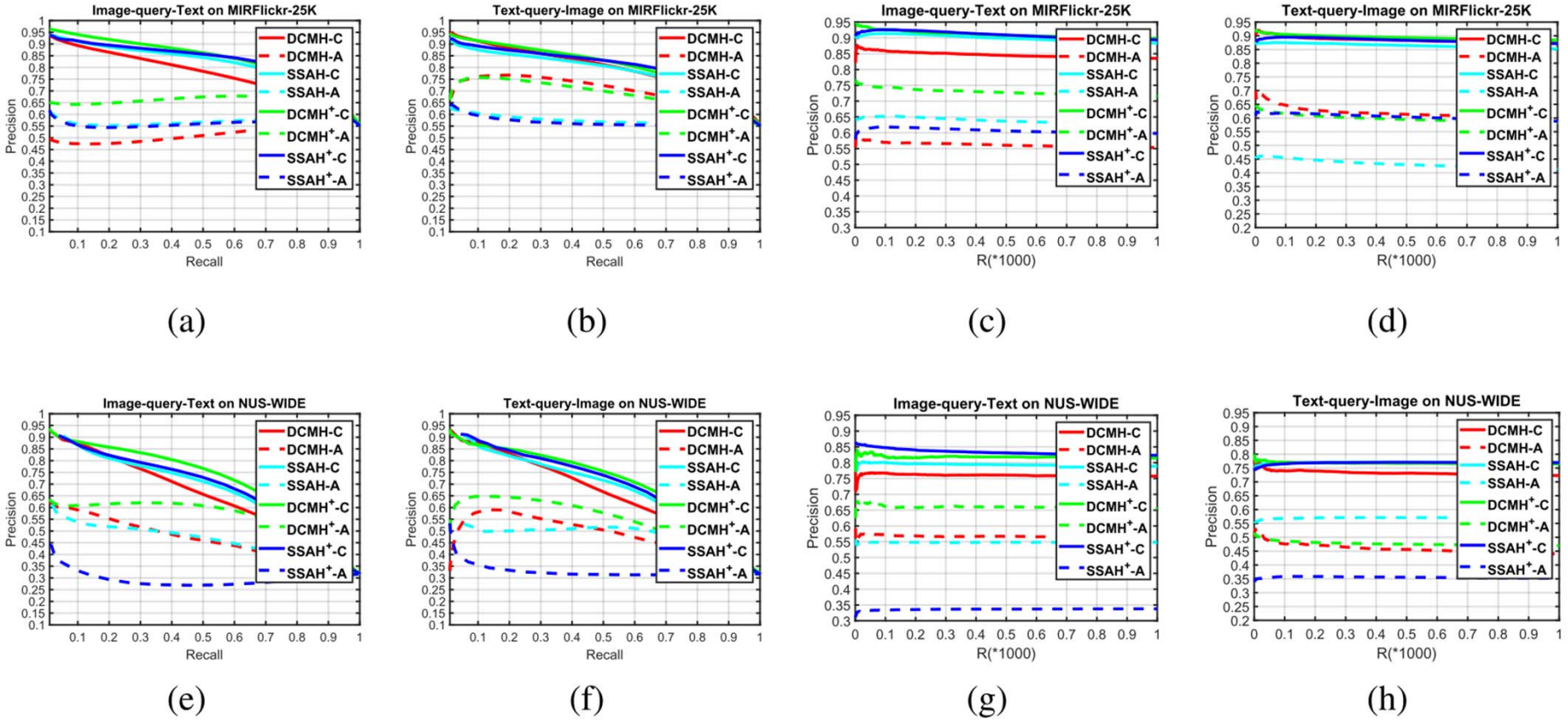
Task	Method	MIRFlickr-25K				NUS-WIDE			
		16	32	48	64	16	32	48	64
Image Query v.s. Text Database	DCMH	0.736	0.749	0.756	0.761	0.595	0.607	0.620	0.641
	DCMH$^+$	0.805	0.816	0.825	0.828	0.658	0.679	0.686	0.683
	SSAH	0.797	0.805	0.807	0.807	0.645	0.660	0.670	0.672
	SSAH$^+$	0.804	0.815	0.826	0.829	0.660	0.675	0.690	0.694
Text Query v.s. Image Database	DCMH	0.796	0.797	0.804	0.806	0.601	0.614	0.623	0.645
	DCMH$^+$	0.810	0.820	0.820	0.819	0.679	0.691	0.693	0.690
	SSAH	0.798	0.805	0.807	0.804	0.661	0.677	0.681	0.684
	SSAH$^+$	0.808	0.809	0.814	0.815	0.671	0.685	0.693	0.697

Comparison in terms of ***MAP scores*** and ***distortions(D)*** with 32-bit code length

Task	Iteration		MIRFlickr-25K				NUS-WIDE			
			DCMH	DCMH$^+$	SSAH	SSAH$^+$	DCMH	DCMH$^+$	SSAH	SSAH$^+$
Image Query v.s. Text Database	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	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

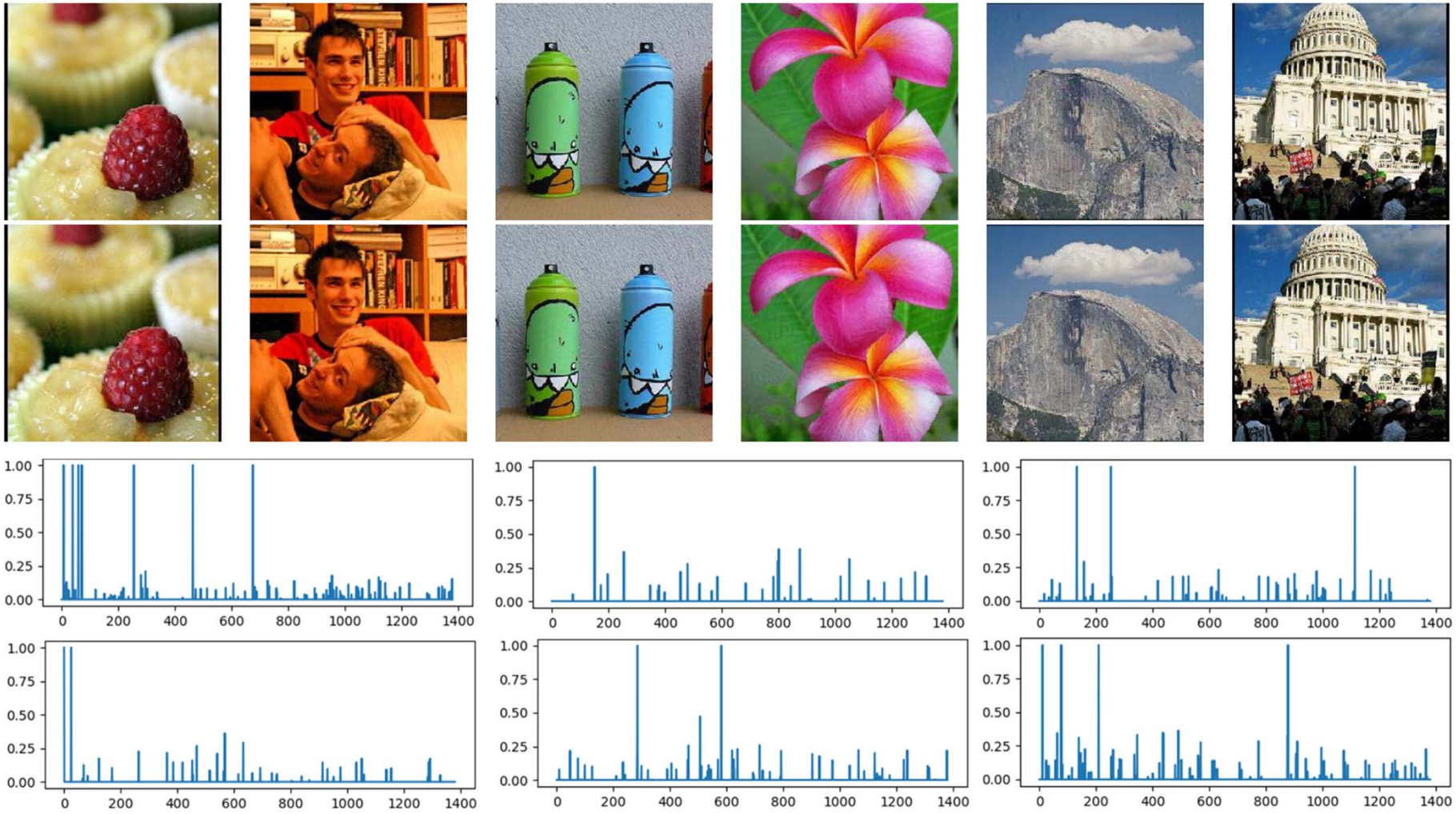
$$D = \sqrt{\frac{\sum (\hat{o} * - o*)^2}{M}}$$

Results



PR and ***Precision@1000 Curves*** evaluated on MIRFlickr-25K and NUS-WID datasets with 32 bits

Results

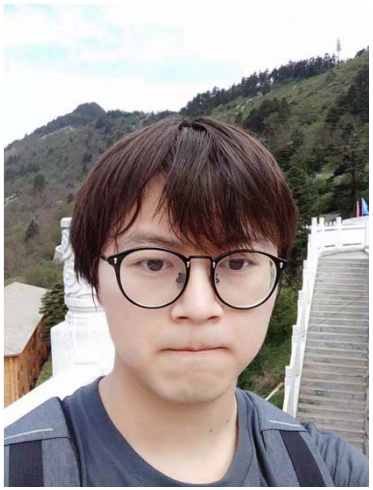


Adversarial Samples of different modalities learned by the proposed ***CMLA***

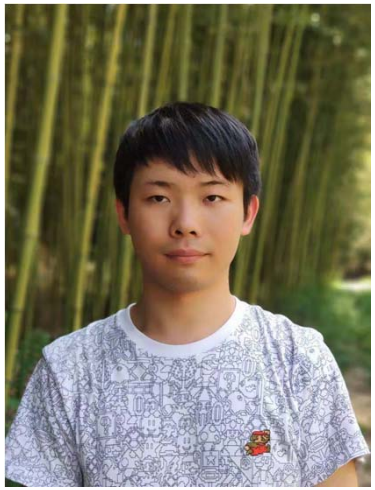
Acknowledgement

Erkun Yang

Chao Li



Zhipeng Wang



Xiaozhe Zhang

Thanks for Your Attention!