Universal Perturbation Attack against Image Retrieval

Li et al.,

ICCV 2019

Presented by: Woo Jae Kim

Table of Contents

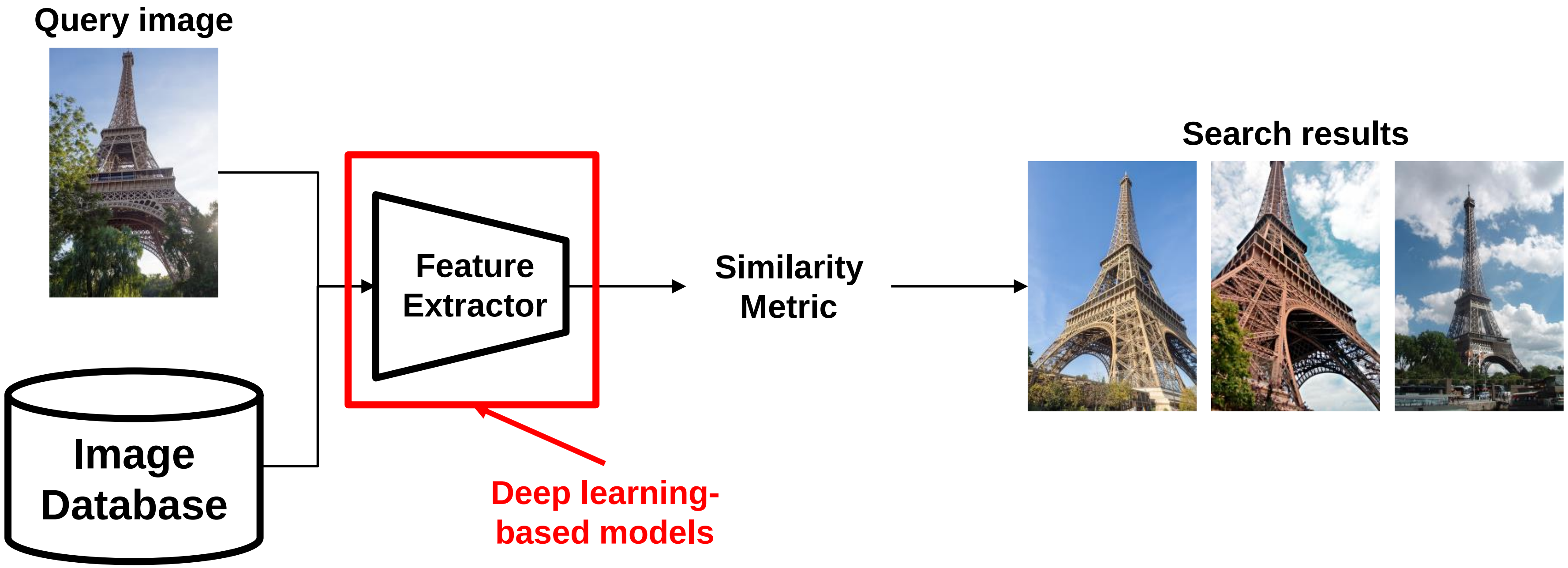
- **Backgrounds & Motivation**
- **Related Works**
- **Methods**
- **Experiments**
- **Conclusion**

BACKGROUNDS & MOTIVATION

1. Visual Search at eBay, SIGKDD 2017
2. Visual Fashion-Product Search at SK Planet, ArXiv 2017
3. Visual Search at Alibaba, SIGKDD 2018
4. Visual Search at Pinterest, SIGKDD 2015

DL-Based Image Retrieval

- **Image retrieval these days relies on <u>Deep Learning</u>**
  - **eBay -> ResNet-50 [1]**
  - **SK Planet, Alibaba → Inception-based network [2, 3]**
  - **Pinterest → AlexNet & VGG [4]**



Robustness of Deep Learning

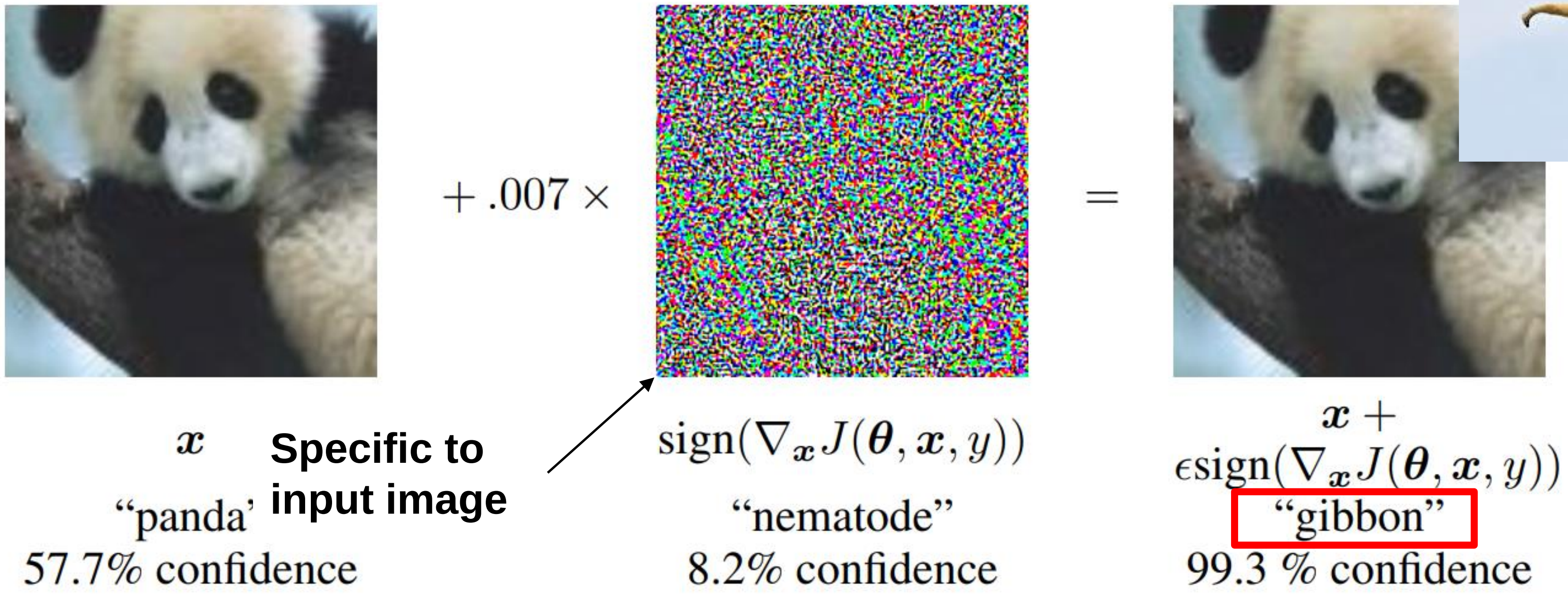
- **However, deep learning is not robust**
- **It is susceptible to specific types of noise**
- **This noise is called <u>"adversarial attack"</u>**

https://www.youtube.com/watch?v=pc2ssNY98LA&feature=emb_logo&ab_channel=PoloClubofDataScience

KAIST

Adversarial Attack

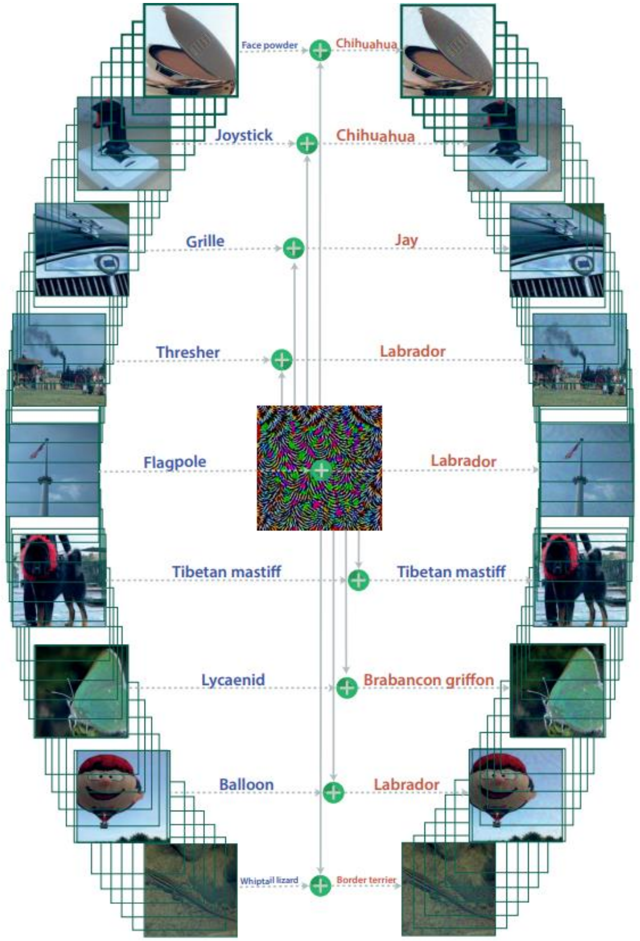
- **Then, what is adversarial attack?**
- **Imperceptible perturbation maliciously designed to fool machine learning models**



gibbon

Universal Adversarial Attack

- **Universal Adversarial Perturbation (UAP) – A single perturbation can be added to any image to fool machine learning model**
- **Strengths**
  - **Can attack images on-the-fly**
  - **Can attack unknown images**
- **Focuses on classification task**



Universal Adversarial Perturbations, CVPR 2017

Problems of UAP on Image Retrieval

- **However, UAP on classification (UAP-C) cannot be used in image retrieval**

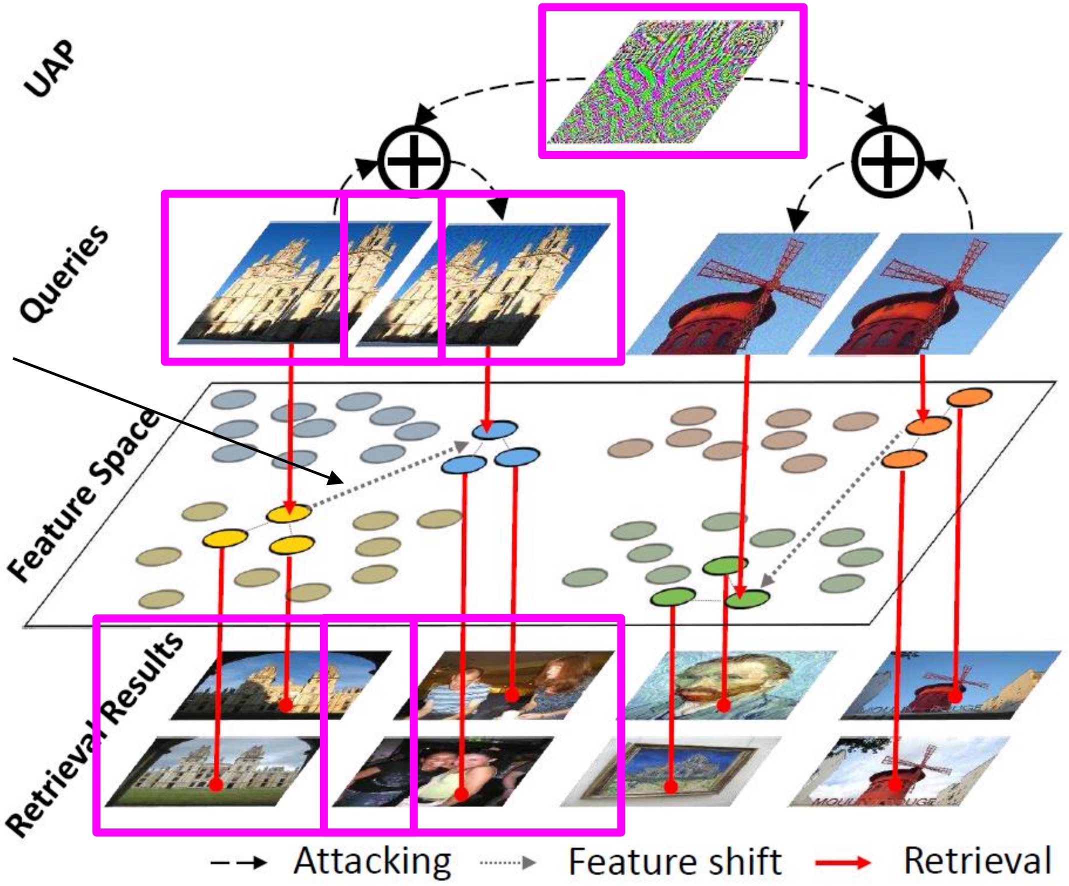
  1. **UAP-C requires datasets with *labeled categories***
  2. **UAP-C only fools *top-1 prediction***
  3. **UAP-C assumes *fixed size inputs***
  4. **Classification model produces *continuous probability* as output**

KAIST

Goals

- **Build UAP specific to image retrieval task (UAP-IR)**
  - **Disrupt the neighborhood relationship among features**

Perturbing neighborhood relationship among features



RELATED WORKS

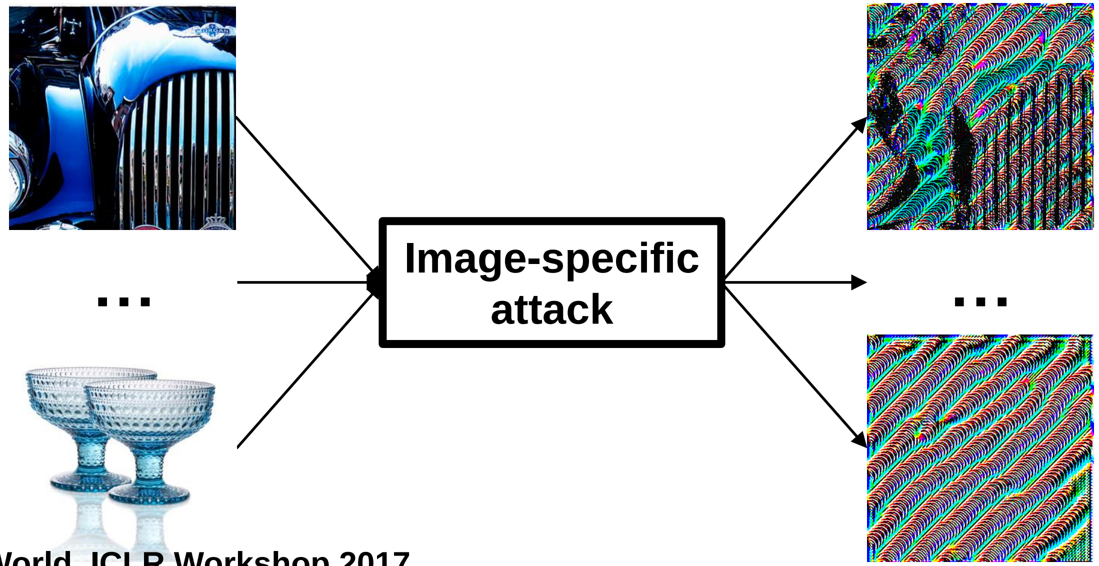
Image-Specific Attack on Classification

- **Gradient-Based Attacks**
  - **In classification task, classification loss is minimized by using <u>gradient descent</u>**
  - **Gradient-based attacks "maximize" the loss by <u>adding gradient to image x</u>**

$$x^{adv} = x + \text{sign}\big(\nabla J(w)\big)$$

Cross-entropy loss of x^{adv}

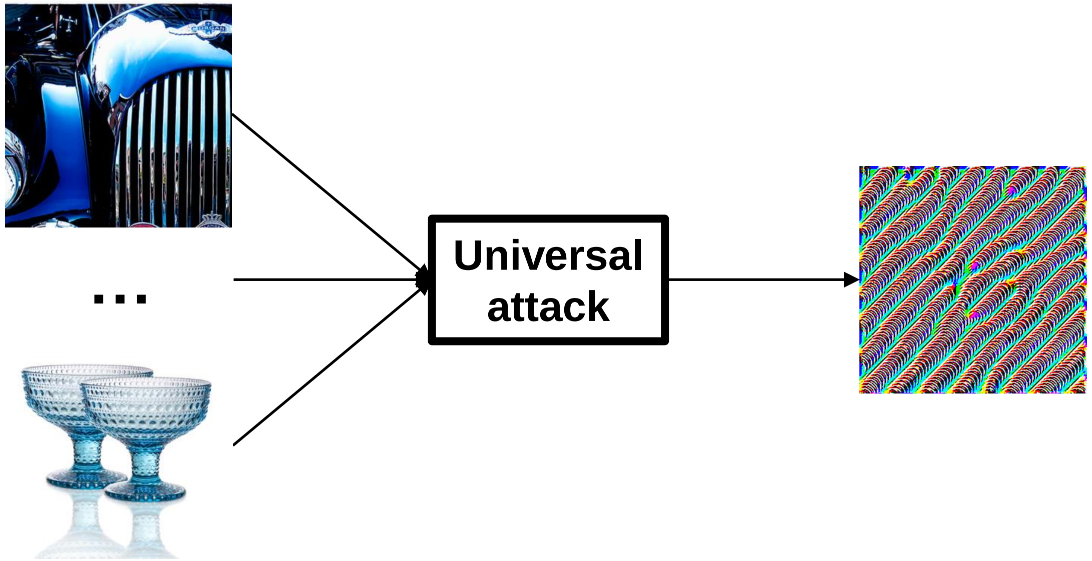
- **Image-Specific Attacks**
  - **Different perturbations are generated for each image**



Adversarial Examples in the Physical World, ICLR Workshop 2017

Universal Attack on Classification

- **Universal Adversarial Perturbations (UAP)**
  - **<u>Single perturbation</u> is added to any image to form adversarial image**
  - **Also optimized to <u>maximize classification loss</u> of adversarial image**



Universal Adversarial Perturbations, CVPR 2017

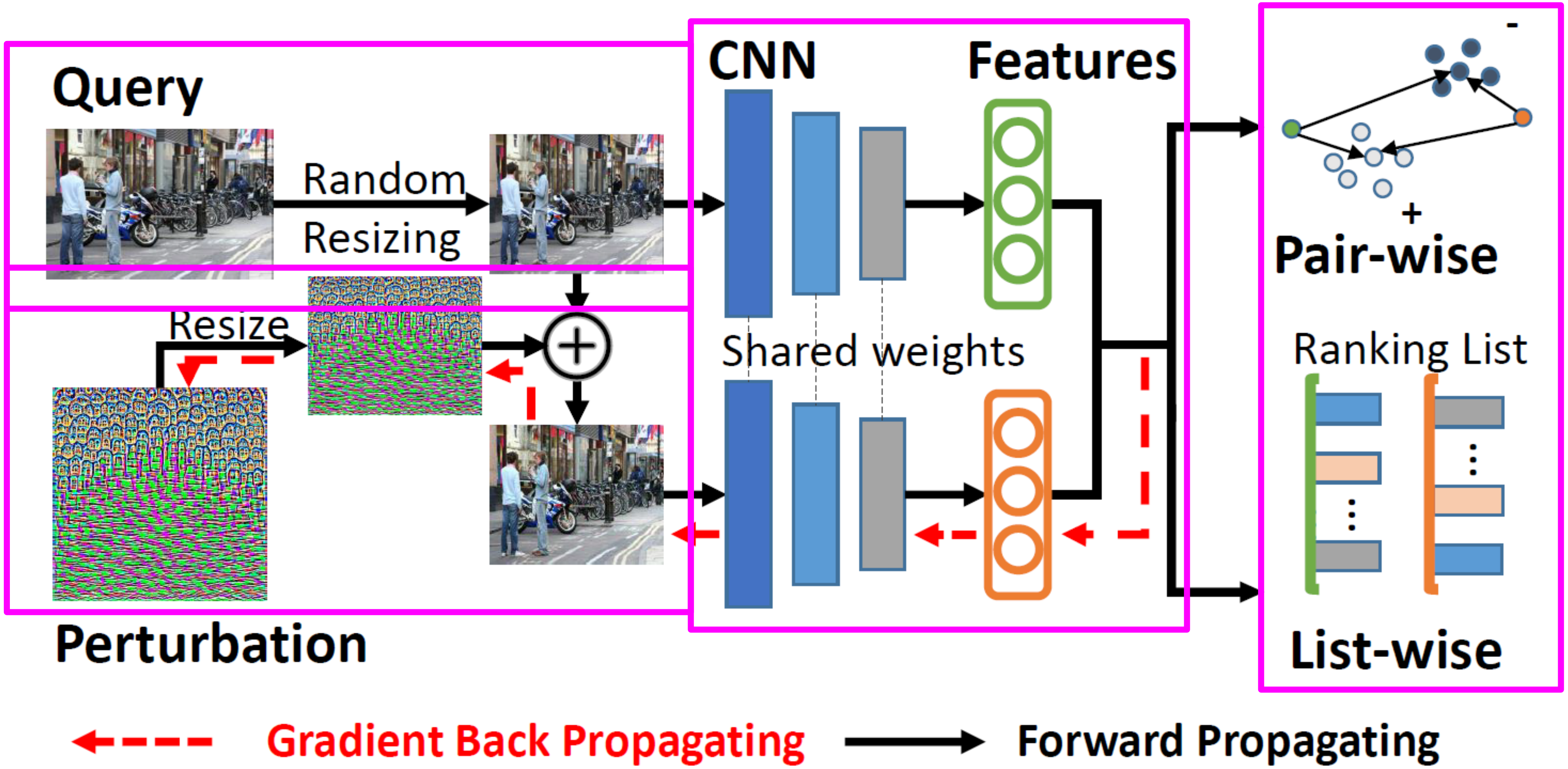
METHODS

Main Contributions

- **Unlike UAP on classification, UAP on image retrieval:**
  1. **Corrupts <u>relationship among features</u>**
     - **Pair-wise loss**
     - **List-wise loss**

  2. **Adapts to <u>input images of various sizes</u>**

Overall Pipeline



Objective Functions

- (baseline) Label-wise loss
  - Disrupts the classification loss
  - Same as the UAP on classification task (not proposed by this paper)

$$L = -\mathcal{H}\left(f\left(x^{adv}\right), y_{gt}\right)$$

  - where:
    - $\mathcal{H}$ = cross-entropy loss
    - f = target classifier
    - x^{adv} = adversarial query
    - y_{gt} = ground truth class

Objective Functions

- **Pair-wise loss**
  - **Disrupts the Triplet Loss – switch "positive" and "negative" images**
  - **Original Triplet Loss:**

$$L = \| f_i - \underline{f_p} \|_2^2 - \| f_i - \underline{f_n} \|_2^2 + \alpha$$

  - **Disturbed Triplet Loss:**

$$L = \| f'_i - \underline{f_n} \|_2^2 - \| f'_i - \underline{f_p} \|_2^2 + \alpha$$

  - **where:**
    - f_i = given query feature
    - f'_i = adversarial query feature
    - f_n = negative cluster feature
    - f_p = positive cluster feature
    - α = margin parameter

Objective Functions

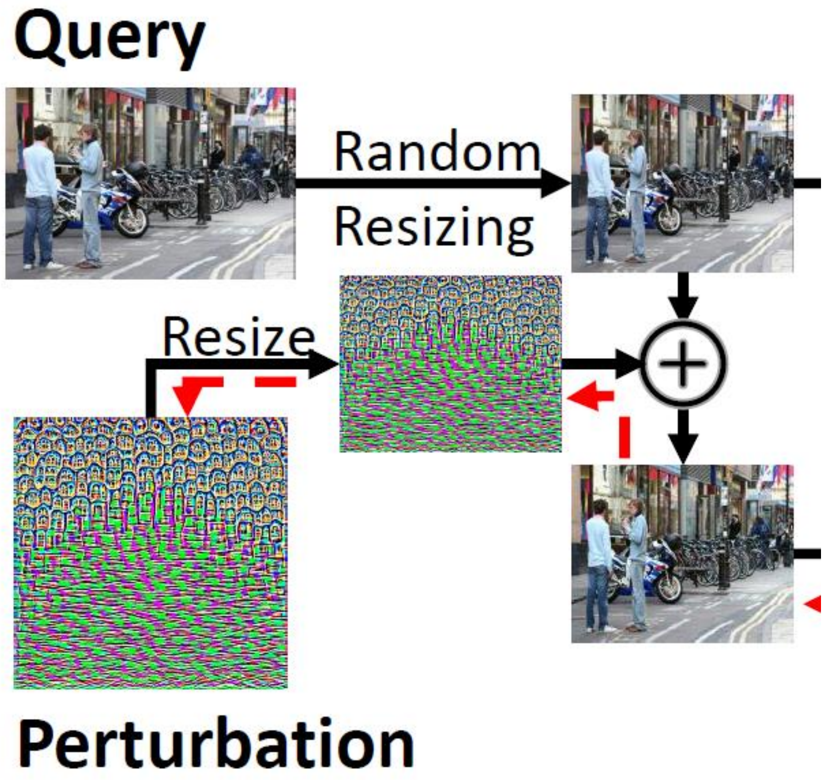
- **List-wise Loss**
  - **Perturb the entire ranking list**
  - **Disturbs normalized Discounted Cumulative Gain (NDCG) metric**
    - **Used to measure relevance of retrieved ranking list**
    - **Higher NDCG ➔ more relevant search results**
  - **Minimize NDCG to 0**

$$DCG = \sum_{i=1}^{|g|} \frac{2^{y_i} - 1}{\log_2(i+1)}$$

  - **where**
    - $\{y_i\}_{i=1}^{|g|}$ **= relevance of search results**
    - $|g|$ **= # elements in search results**

Random Resizing

- **Random resizing to attack queries of various sizes**



EXPERIMENTS

Target Models

- 3 feature extractors
  - AlexNet (A), VGG (V), ResNet (R)
- 2 pooling layers
  - GeM and MAC

} 6 models

- Feature extractors are:
  - Pretrained on ImageNet
  - Fine-tuned on SfM-120k dataset
- Attacks are evaluated on:
  - ROxford5k and RParis6k

Results of UAP Attack

O = No attack
C = label-wise loss (baseline)
P = pair-wise loss
L = list-wise loss

mAP = mean average precision (↓)
mP@10 = mean precision @ 10 (↓)
mDR = dropping (attack success) rate (↑)

		Oxford5k	ROxford5k						Paris6k	RParis6k						
			E	M	H	E	M	H		E	M	H	E	M	H	
Eval		mAP				mP@10			mAP				mP@10			mDR
A-MAC	O	57.11	45.23	32.96	10.43	57.25	55.43	15.36	65.64	63.99	46.93	20.06	88.00	91.29	58.29	
	C	46.99	36.13	27.89	7.86	49.58	48.36	12.71	57.91	52.96	40.33	16.27	80.86	83.00	48.86	15.47%
	P	29.61	24.52	17.99	4.92	32.06	30.86	**6.67**	42.89	38.71	30.43	11.13	52.86	54.71	29.14	44.35%
	L	**27.88**	**21.59**	**16.31**	**4.06**	**28.33**	**28.57**	7.50	**41.15**	**37.40**	**29.28**	**10.00**	**49.29**	**51.43**	**25.00**	**48.33%**
A-GeM	O	59.86	50.21	36.72	14.29	58.10	53.60	23.32	73.66	70.65	51.89	22.80	87.71	88.86	57.86	
	C	35.49	30.07	22.00	7.03	33.62	31.71	10.16	48.27	42.60	33.80	12.55	46.57	50.00	27.00	43.51%
	P	29.31	22.85	17.57	5.56	**25.65**	24.79	8.36	40.71	35.17	29.44	10.71	38.86	41.71	20.14	54.12%
	L	**26.48**	**22.45**	**17.12**	**5.29**	25.78	**24.25**	**8.03**	**37.17**	**32.28**	**27.42**	**10.23**	**34.86**	**37.14**	**18.29**	**56.88%**
V-MAC	O	81.45	75.07	57.15	29.96	78.60	78.33	45.57	88.31	86.39	69.60	44.97	93.57	96.86	84.71	
	C	42.70	37.15	30.14	14.87	35.59	36.14	20.43	34.15	29.88	27.37	12.48	18.57	18.86	12.43	61.80%
	P	37.60	32.33	26.99	14.49	35.15	35.29	20.57	**23.76**	**21.02**	**20.12**	**9.21**	**13.86**	**15.57**	**9.86**	66.94%
	L	**35.57**	**29.83**	**24.97**	**13.13**	**32.79**	**32.29**	**19.71**	25.38	22.13	20.99	9.23	15.29	17.14	10.43	**67.96%**
V-GeM	O	85.24	76.43	59.17	32.26	80.52	81.29	49.71	86.28	84.66	67.06	42.40	95.14	97.57	83.00	
	C	46.08	38.98	31.59	14.20	36.45	36.29	19.57	44.51	38.05	34.44	15.39	27.14	27.29	17.57	57.60%
	P	43.71	37.84	30.92	**15.36**	36.76	37.00	21.86	**30.92**	28.12	**25.78**	11.91	17.43	17.43	12.86	62.64%
	L	**41.94**	**37.13**	**30.00**	15.39	**34.40**	**34.00**	**21.43**	32.29	**27.39**	25.95	**11.69**	**16.86**	**16.86**	**10.86**	**63.72%**
R-MAC	O	81.69	73.85	56.14	29.80	78.33	79.86	46.57	83.55	81.56	63.91	39.06	93.52	96.71	79.57	
	C	58.52	50.65	37.50	15.59	56.47	54.29	24.71	67.57	61.51	49.43	25.01	70.00	72.43	49.57	31.27%
	P	35.31	30.34	24.73	13.37	36.62	36.43	20.71	35.66	32.61	27.23	12.12	32.57	34.86	21.29	59.71%
	L	**34.08**	**28.68**	**23.30**	**12.09**	**34.26**	**32.95**	**19.86**	**34.63**	**30.71**	**26.16**	**11.50**	**28.00**	**29.71**	**18.43**	**62.60%**
R-GeM	O	86.24	80.63	63.13	38.51	82.72	83.14	54.57	90.66	90.33	74.06	51.69	94.96	98.29	88.29	
	C	68.45	59.30	45.57	21.38	66.25	62.52	34.86	79.00	73.48	59.05	33.36	84.00	87.00	68.71	23.76%
	P	34.81	30.50	24.33	13.79	28.97	28.43	19.71	33.76	31.67	26.54	11.28	**27.86**	29.43	17.00	66.69%
	L	**31.73**	**29.21**	**23.17**	**13.01**	**27.21**	**27.29**	**18.00**	**32.07**	**29.60**	**25.18**	**10.35**	**27.86**	**28.86**	**16.14**	**68.47%**

Results of Transfer Attack

- **Attack success rates (↑) on unknown models**

Target models under attack

Source models for attack

	A-MAC	A-GeM	V-MAC	V-GeM	R-MAC	R-GeM
A-MAC	**48.33**	34.94	13.60	10.78	8.57	11.27
A-GeM	38.18	**56.88**	14.31	12.00	7.64	12.22
V-MAC	14.68	15.26	**67.96**	60.16	18.46	19.32
V-GeM	15.66	16.30	66.16	**63.72**	18.24	19.87
R-MAC	16.38	15.53	23.59	19.62	62.60	58.25
R-GeM	14.27	14.29	23.94	22.35	**67.91**	**68.47**

Effects of Resizing

- **Attack success rates (↑) w/ various resizings**

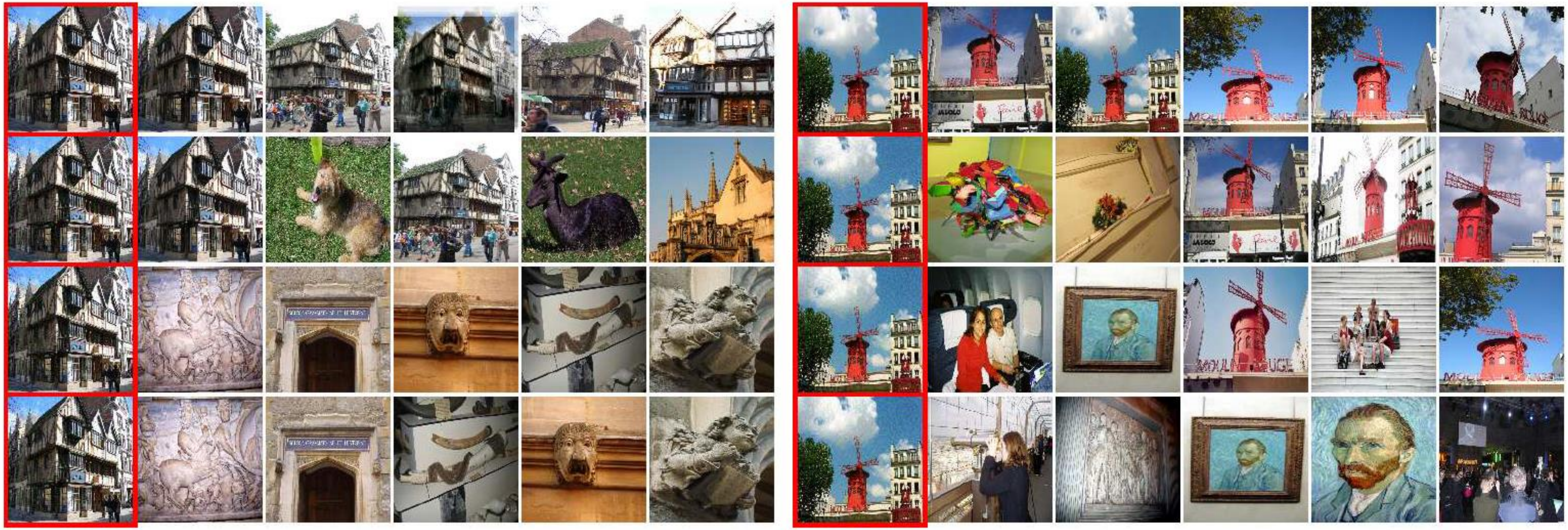
Range for image resizing: [$size_{min}$, $size_{max}$]

Range	[362, 362]	[1024, 1024]	[128, 1024]	[256, 1024]	[512, 1024]	[768, 1024]
A-GeM	16.89%	24.69%	53.21%	56.88%	51.41%	39.21%
V-GeM	25.87%	30.42%	61.93%	63.72%	55.02%	42.08%

Visualization

- **Retrieval results**

No attack
Label-wise loss (baseline)
Pair-wise loss
List-wise loss



(a) Oxford5k (b) Paris6k

CONCLUSION

Strengths & Weaknesses

- **Strengths**
  - **First proposed pair-wise loss and list-wise loss to disrupt feature relationships**
  - **Achieved high attack success rates on image retrieval compared to baseline label-wise loss**

- **Weaknesses**
  - **Show poor attack success rates on unknown models**
    - **e.g. AlexNet → VGG**
  - **Lacks analysis on more current retrieval models**
    - **Attention module**
    - **Different pooling layers**