Practical Relative Order Attack in Deep Ranking

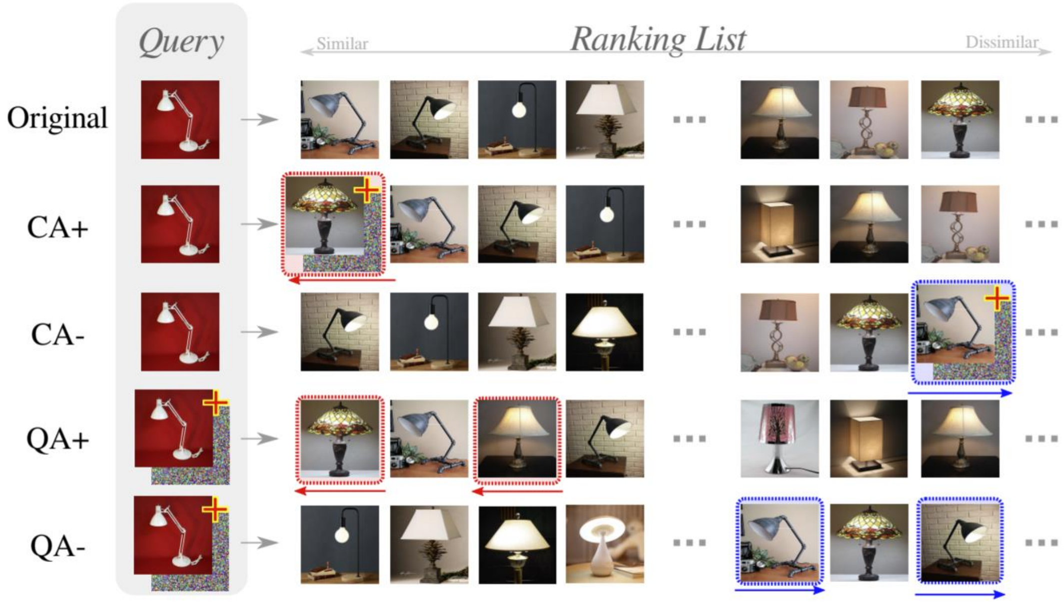
Mo Zhou, Le Wang, Zhenxing Niu, Qilin Zhang,
Yinghui Xu, Nanning Zheng, Gang Hua

arXiv

Github

Background

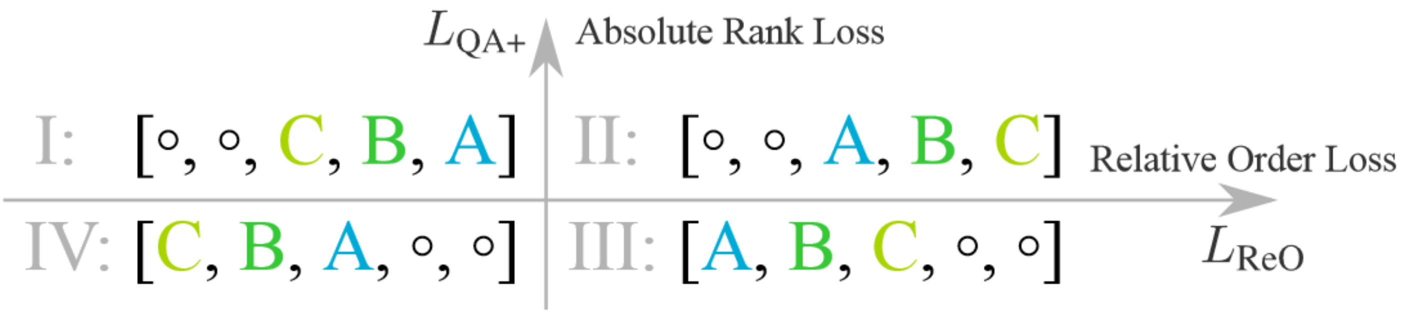
- Deep ranking (deep metric learning) models are vulnerable to adversarial attacks (the ranking result can be dramatically changed).



* Zhou, et al., Adversarial Ranking Attack and Defense, ECCV 2020.

Insight

- Previous attacks focus on absolute rank
- Attack on relative order remains under-explored



Absolute rank: the absolute positions of selected candidates
Relative order: the relative positions among selected candidates

Contributions

- **Order Attack (OA),** which alters the relative order among selected candidates through adversarial attack.
- White-Box OA: a triplet-style implementation
- Black-Box OA: a Short-range Ranking Correlation (SRC) metric as a surrogate objective approximating the triplet-style formulation.
- Real-world attack demo: including a major online retailing e-commerce platform and a major search-by-image platform.

Order Attack (OA)

Original Query Image → Deep Ranking Model ⇒ Original Ranking Result

$A < B < C < D < E < \ldots\ldots$

Top-1 Result

Adversarially Perturbed Query Image → Deep Ranking Model ⇒ Ranking Result w.r.t. Perturbed Query

$A < \underline{E < D < C < B} < \ldots\ldots$

Changed the relative order into the specified permutation [1, 5, 4, 3, 2].

Impacts the *Click-Through Rate (CTR)* hence indirectly influence the sales.

White-Box OA

Formulation

▷ **Order Attack** finds an adversarial perturbation

$$\boldsymbol{r}\ (\|\boldsymbol{r}\|_\infty \leqslant \varepsilon \text{ and } \tilde{\boldsymbol{q}} = \boldsymbol{q} + \boldsymbol{r} \in \mathcal{I}),$$

so that the adversarial query $\tilde{\boldsymbol{q}}$ results in $\boldsymbol{c}_{p_1} \prec \boldsymbol{c}_{p_2} \prec \cdots \prec \boldsymbol{c}_{p_k}$ based on the **attacker-specified permutation** $\mathbf{p} = [p_1, p_2, \ldots, p_k]$

White-Box OA

Implementation

- ▷ The inequality chain prescribed by the permutation

$$f(\tilde{\boldsymbol{q}}, \boldsymbol{c}_{p_1}) < f(\tilde{\boldsymbol{q}}, \boldsymbol{c}_{p_2}) < \cdots < f(\tilde{\boldsymbol{q}}, \boldsymbol{c}_{p_k})$$

can be decomposed into a series of inequalities, i.e.,

$$f(\tilde{\boldsymbol{q}}, \boldsymbol{c}_{p_i}) < f(\tilde{\boldsymbol{q}}, \boldsymbol{c}_{p_j}), \quad i, j = 1, 2, \ldots, k, \; i < j.$$

- ▷ Reformulation of the inequalities into triplet loss form leads to the **relative order loss** function

$$L_{\mathrm{ReO}}(\tilde{\boldsymbol{q}}; \mathbb{C}, \mathbf{p}) = \sum_{i=1}^{k} \sum_{j=i}^{k} \left[f(\tilde{\boldsymbol{q}}, \boldsymbol{c}_{p_i}) - f(\tilde{\boldsymbol{q}}, \boldsymbol{c}_{p_j}) \right]_+.$$

which can be combined with a previously proposed semantics-preserving loss term to keep the selected candidates within the topmost part of ranking.

White-Box Experiments

2021 ICCV OCTOBER 11-17 VIRTUAL

	$k=5$					$k=10$					$k=25$				
ε	0	$\frac{2}{255}$	$\frac{4}{255}$	$\frac{8}{255}$	$\frac{16}{255}$	0	$\frac{2}{255}$	$\frac{4}{255}$	$\frac{8}{255}$	$\frac{16}{255}$	0	$\frac{2}{255}$	$\frac{4}{255}$	$\frac{8}{255}$	$\frac{16}{255}$
Fashion-MNIST $N=\infty$															
τ_S	0.000	0.286	0.412	0.548	0.599	0.000	0.184	0.282	0.362	0.399	0.000	0.063	0.108	0.136	0.149
mR	2.0	4.5	9.1	12.7	13.4	4.5	7.4	10.9	15.2	17.4	12.0	16.1	17.6	18.9	19.4
Stanford Online Products $N=\infty$															
τ_S	0.000	0.396	0.448	0.476	0.481	0.000	0.263	0.348	0.387	0.398	0.000	0.125	0.169	0.193	0.200
mR	2.0	5.6	4.9	4.2	4.1	4.5	12.4	11.2	9.9	9.6	12.0	31.2	28.2	25.5	25.4

Table 1: White-box order attack on Fashion-MNIST and SOP datasets with various settings.

k: number of selected candidates   ε: perturbation budget

τ_S: Kendal's ranking correlation   mR: mean rank of selected candidates

* τ_S is equivalent to Kendall's ranking correlation in white-box scenario.

Black-Box OA

- Short-range Ranking Correlation (SRC) as a surrogate objective.
- Measures the alignment between the specified permutation and the actual ranking result.
- Inspired by Kendall's tau.

Algorithm 1: Short-range Ranking Correlation τ_S.

Input: Selected candidates $\mathbb{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_k\}$,
permutation vector $\mathbf{p} = [p_1, p_2, \dots, p_k]$,
top-N retrieval $\mathbb{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ for $\tilde{\boldsymbol{q}}$.
Note that $\mathbb{C} \subset \mathbb{D}$, $\mathbb{X} \subset \mathbb{D}$, and $N \geqslant k$.

Output: SRC coefficient τ_S.

Permute candidates as $\mathbb{C}_\mathbf{p} = \{\boldsymbol{c}_{p_1}, \boldsymbol{c}_{p_2}, \dots, \boldsymbol{c}_{p_k}\}$;
Initialize score matrix $\boldsymbol{S} = \mathbf{0}$ of size $k \times k$;
for $i \leftarrow 1, 2, \dots, k$ **do**
  for $j \leftarrow 1, 2, \dots, i-1$ **do**
    if $\boldsymbol{c}_i \notin \mathbb{X}$ [1] or $\boldsymbol{c}_j \notin \mathbb{X}$ **then**
      $S_{i,j} = -1$ `// out-of-range`
    else if $[\mathrm{R}_{\mathbb{C}_\mathbf{p}}(\boldsymbol{c}_i) > \mathrm{R}_{\mathbb{C}_\mathbf{p}}(\boldsymbol{c}_j)$ and $\mathrm{R}_\mathbb{X}(\boldsymbol{c}_i) > \mathrm{R}_\mathbb{X}(\boldsymbol{c}_j)]$ or $[\mathrm{R}_{\mathbb{C}_\mathbf{p}}(\boldsymbol{c}_i) < \mathrm{R}_{\mathbb{C}_\mathbf{p}}(\boldsymbol{c}_j)$ and $\mathrm{R}_\mathbb{X}(\boldsymbol{c}_i) < \mathrm{R}_\mathbb{X}(\boldsymbol{c}_j)]$ **then**
      $S_{i,j} = +1$ `// concordant`
    else if $[\mathrm{R}_{\mathbb{C}_\mathbf{p}}(\boldsymbol{c}_i) > \mathrm{R}_{\mathbb{C}_\mathbf{p}}(\boldsymbol{c}_j)$ and $\mathrm{R}_\mathbb{X}(\boldsymbol{c}_i) < \mathrm{R}_\mathbb{X}(\boldsymbol{c}_j)]$ or $[\mathrm{R}_{\mathbb{C}_\mathbf{p}}(\boldsymbol{c}_i) < \mathrm{R}_{\mathbb{C}_\mathbf{p}}(\boldsymbol{c}_j)$ and $\mathrm{R}_\mathbb{X}(\boldsymbol{c}_i) > \mathrm{R}_\mathbb{X}(\boldsymbol{c}_j)]$ **then**
      $S_{i,j} = -1$ `// discordant`

return $\tau_S = \sum_{i,j} S_{i,j} / \binom{k}{2}$

Black-Box Experiments

Fashion-MNIST Dataset

Algorithm	$k=5$				$k=10$				$k=25$			
	$\varepsilon=\frac{2}{255}$	$\varepsilon=\frac{4}{255}$	$\varepsilon=\frac{8}{255}$	$\varepsilon=\frac{16}{255}$	$\varepsilon=\frac{2}{255}$	$\varepsilon=\frac{4}{255}$	$\varepsilon=\frac{8}{255}$	$\varepsilon=\frac{16}{255}$	$\varepsilon=\frac{2}{255}$	$\varepsilon=\frac{4}{255}$	$\varepsilon=\frac{8}{255}$	$\varepsilon=\frac{16}{255}$
None	0.0, 2.0	0.0, 2.0	0.0, 2.0	0.0, 2.0	0.0, 4.5	0.0, 4.5	0.0, 4.5	0.0, 4.5	0.0, 12.0	0.0, 12.0	0.0, 12.0	0.0, 12.0
Fasion-MNIST $N=\infty$												
Rand	0.211, 2.1	0.309, 2.3	0.425, 3.0	0.508, 7.7	0.172, 4.6	0.242, 5.0	0.322, 6.4	0.392, 12.7	0.084, 12.3	0.123, 13.1	0.173, 15.8	0.218, 25.8
Beta	0.241, 2.1	0.360, 2.6	0.478, 4.6	0.580, 19.3	0.210, 4.8	0.323, 5.7	0.430, 9.6	0.510, 30.3	0.102, 12.4	0.163, 13.8	0.237, 19.7	0.291, 42.7
PSO	0.265, 2.1	0.381, 2.3	0.477, 4.4	0.580, 21.1	0.239, 4.8	0.337, 5.7	0.424, 9.7	0.484, 34.0	0.131, 12.7	0.190, 14.6	0.248, 21.7	0.286, 54.2
NES	0.297, 2.3	**0.416, 3.1**	**0.520, 8.7**	**0.630, 46.3**	**0.261, 5.0**	0.377, 6.6	0.473, 14.3	0.518, 55.6	**0.142, 13.0**	0.217, 15.9	0.286, 28.3	0.312, 74.3
SPSA	**0.300, 2.3**	0.407, 3.2	0.465, 7.1	0.492, 16.3	0.249, 5.0	**0.400, 6.6**	**0.507, 12.8**	**0.558, 27.5**	0.135, 12.9	**0.236, 16.3**	**0.319, 27.1**	**0.363, 46.4**
Fashion-MNIST $N=50$												
Rand	0.207	0.316	0.424	0.501	0.167	0.242	0.321	0.378	0.083	0.123	0.165	0.172
Beta	0.240	0.359	0.470	0.564	0.204	0.323	0.429	0.487	0.103	0.160	0.216	0.211
PSO	0.266	0.377	0.484	0.557	0.239	0.332	0.420	0.458	0.134	0.183	0.220	0.203
NES	**0.297**	**0.426**	**0.515**	**0.584**	**0.262**	0.378	0.463	0.458	**0.141**	0.199	0.223	0.185
SPSA	0.292	0.407	0.468	0.490	0.253	**0.397**	**0.499**	**0.537**	0.131	**0.214**	**0.260**	**0.275**
Fashion-MNIST $N=k$												
Rand	0.204	0.289	0.346	0.302	0.146	0.181	0.186	0.124	0.053	0.062	0.049	0.021
Beta	0.237	0.342	0.372	0.275	0.183	0.236	0.218	0.106	0.072	0.079	0.058	0.020
PSO	0.252	0.342	0.388	0.284	**0.198**	0.240	0.219	0.081	**0.080**	0.082	0.046	0.013
NES	**0.274**	**0.360**	0.381	0.282	**0.198**	0.234	0.213	0.113	0.071	0.076	0.055	0.016
SPSA	**0.274**	**0.360**	**0.412**	**0.427**	0.188	**0.251**	**0.287**	**0.298**	0.067	**0.086**	**0.091**	**0.095**

Table 3: Black-box OA on Fashion-MNIST dataset. In the $N=\infty$ experiments, (τ_S, mR) are reported in each cell, while only τ_S is reported in the cells when N equals 50 or k. A larger k and a smaller N make the attack harder.

Black-Box Experiments

Stanford Online Products Dataset

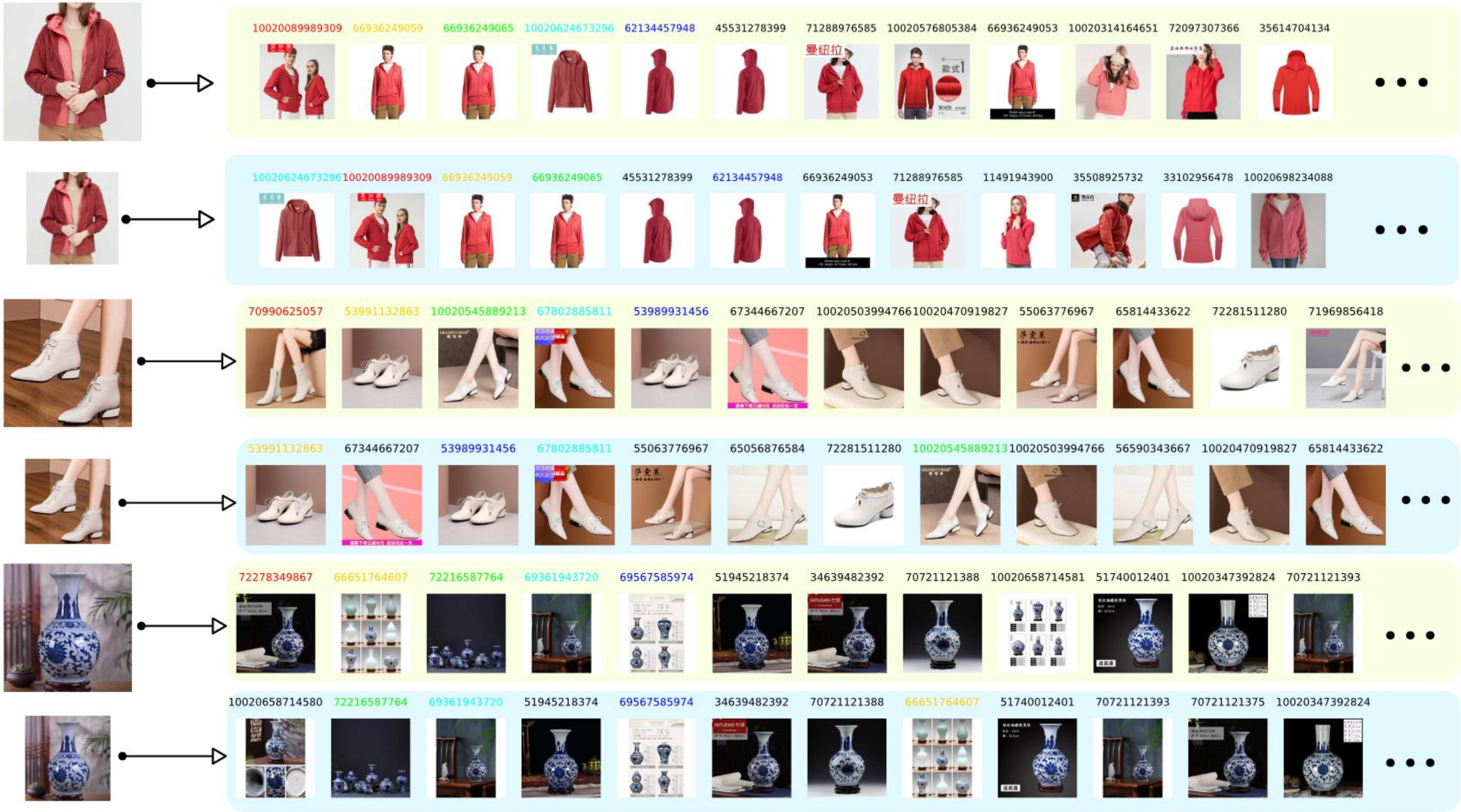
Algorithm	$k=5$				$k=10$				$k=25$			
	$\varepsilon=\frac{2}{255}$	$\varepsilon=\frac{4}{255}$	$\varepsilon=\frac{8}{255}$	$\varepsilon=\frac{16}{255}$	$\varepsilon=\frac{2}{255}$	$\varepsilon=\frac{4}{255}$	$\varepsilon=\frac{8}{255}$	$\varepsilon=\frac{16}{255}$	$\varepsilon=\frac{2}{255}$	$\varepsilon=\frac{4}{255}$	$\varepsilon=\frac{8}{255}$	$\varepsilon=\frac{16}{255}$
None	0.0, 2.0	0.0, 2.0	0.0, 2.0	0.0, 2.0	0.0, 4.5	0.0, 4.5	0.0, 4.5	0.0, 4.5	0.0, 12.0	0.0, 12.0	0.0, 12.0	0.0, 12.0
Stanford Online Product $N=\infty$												
Rand	0.187, 2.6	0.229, 8.5	0.253, 85.8	0.291, 649.7	0.167, 5.6	0.197, 13.2	0.208, 92.6	0.222, 716.4	0.093, 14.1	0.110, 27.6	0.125, 146.7	0.134, 903.7
Beta	0.192, 3.3	0.239, 15.3	0.265, 176.7	0.300, 1257.7	0.158, 6.2	0.186, 19.9	0.207, 139.0	0.219, 992.5	0.099, 15.5	0.119, 37.1	0.119, 206.5	0.132, 1208.5
PSO	0.122, 2.1	0.170, 3.0	0.208, 13.3	0.259, 121.4	0.135, 4.8	0.177, 6.5	0.206, 22.8	0.222, 166.5	0.104, 12.7	0.122, 16.7	0.137, 49.5	0.140, 264.2
NES	**0.254, 3.4**	0.283, 15.6	**0.325, 163.0**	**0.368, 1278.7**	**0.312, 7.2**	**0.351, 26.3**	0.339, 227.1	0.332, 1486.7	**0.242, 18.0**	**0.259, 51.5**	0.250, 324.1	0.225, 1790.8
SPSA	0.237, 3.5	**0.284, 11.9**	0.293, 75.2	0.318, 245.1	0.241, 7.8	0.325, 22.2	**0.362, 112.7**	**0.383, 389.0**	0.155, 18.1	0.229, 41.9	**0.286, 185.6**	**0.306, 557.8**
Stanford Online Product $N=50$												
Rand	0.180	0.216	0.190	0.126	0.163	0.166	0.119	0.055	0.092	0.055	0.016	0.003
Beta	0.181	0.233	0.204	0.119	0.153	0.168	0.116	0.054	0.084	0.057	0.021	0.003
PSO	0.122	0.173	0.183	0.153	0.135	0.164	0.137	0.081	0.093	0.083	0.042	0.011
NES	**0.247**	0.283	0.246	0.152	**0.314**	0.295	0.195	0.077	**0.211**	**0.136**	0.054	0.013
SPSA	0.241	**0.287**	**0.297**	**0.303**	0.233	**0.298**	**0.298**	**0.292**	0.125	0.130	**0.114**	**0.103**
Stanford Online Product $N=k$												
Rand	0.148	0.100	0.087	0.026	0.094	0.044	0.018	0.001	0.023	0.009	0.002	0.001
Beta	0.136	0.106	0.053	0.025	0.076	0.040	0.010	0.004	0.021	0.004	0.001	0.001
PSO	0.102	0.098	0.059	0.031	0.088	0.049	0.022	0.007	0.040	0.015	0.006	0.001
NES	**0.185**	0.139	0.076	0.030	**0.173**	0.097	0.036	0.008	**0.071**	**0.027**	0.007	0.005
SPSA	0.172	**0.154**	**0.141**	**0.144**	0.107	**0.104**	**0.085**	**0.069**	0.026	0.025	**0.017**	**0.016**

Table 5: Black-box OA on Stanford Online Product dataset. In the $N=\infty$ experiments, (τ_S, mR) are reported in each cell, while only τ_S is reported in the cells when N equals 50 or k. A larger k and a smaller N make the attack harder.

Practical OA Demo

- A major e-commerce platform: JD Snapshop

Practical OA Demo



Practical OA Demo

2021 ICCV OCTOBER 11-17 VIRTUAL

- Quantitative Results on JD Snapshop

Algorithm	ε	k	Q	T	Mean τ_S	Stdev τ_S	Max τ_S	Min τ_S	Median τ_S
SPSA	1/255	5	100	204	0.390	0.373	1.000	-0.600	0.400
SPSA	1/255	10	100	200	0.187	0.245	0.822	-0.511	0.200
SPSA	1/255	25	100	153	0.039	0.137	0.346	-0.346	0.033

Table 6: Quantitative $(k, 50)$-OA Results on JD Snapshop.

Practical OA Demo

- Quantitative Results on Bing Visual Search API

Algorithm	ε	k	Q	T	Mean τ_S	Stdev τ_S	Max τ_S	Min τ_S	Median τ_S
SPSA	$8/255$	5	100	105	0.452	0.379	1.000	-0.400	0.600
SPSA	$8/255$	10	100	95	0.152	0.217	0.733	-0.378	0.156
SPSA	$8/255$	25	100	93	0.001	0.141	0.360	-0.406	0.010

Table 7: $(k, 50)$-OA Results on Bing Visual Search API.

- Thanks!

arXiv

Github