Practical Relative Order Attack in Deep Ranking

Mo Zhou[1] Le Wang[1] Zhenxing Niu[2] Qilin Zhang[3] Yinghui Xu[2] Nanning Zheng[1] Gang Hua[4]

[1] Xi'an Jiaotong University [2] Alibaba Group
[3] HERE Technologies [4] Wormpex AI Research

arXiv Github

2021 ICCV OCTOBER 11-17 VIRTUAL

Introduction

[Background]

- Deep Ranking Models are *vulnerable to adversarial attacks*, where an imperceptible perturbation can trigger dramatic changes in the ranking result.

[Insight]

- Previous attempts focus on manipulating **absolute ranks** of certain candidates, but the possibility of adjusting their **relative order** remains under-explored.

[Order Attack]

- *In this paper,* we formulate **Order Attack**, which covertly alters the **relative order** among a selected set of candidates according to an attacker-specified permutation, with limited interference to other unrelated candidates.

[White-box Order Attack]

- Order Attack under the white-box assumption is implemented as a triplet-style loss imposing an inequality chain to reflect the specified permutation.

[Black-box Approximation]

- To make Order Attack applicable in real-world *black-box attack scenario*, we propose a **Short-range Ranking Correlation** metric as a surrogate objective to approximate the white-box method.

[Quantitative Experiments]

- Both the white-box and black-box Order Attack are evaluated on Fashion-MNIST and Stanford-Online-Products datasets.

[Efficacy on Real-world E-commerce API]

- Black-box Order Attack is also successfully implemented on a major e-commerce platform through its API.

Order Attack

Original Query Image → Deep Ranking Model ⇒ Original Ranking Result: A < B < C < D < E < ......

Top-1 Result

Adversarially Perturbed Query Image → Deep Ranking Model ⇒ Ranking Result w.r.t. Perturbed Query: A < E < D < C < B < ......

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 Order Attack

- **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** $\mathbf{p} = [p_1, p_2, \ldots, p_k]$

- 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

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

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

Black-Box Order Attack

- Gradient is inaccessible in Black-box attack scenario. Thus white-box attack is infeasible.
- We propose "**Short-range ranking correlation**" metric to measure the alignment between desired order and the actual order, in order to approximate the white-box loss. It is inspired by Kendall's tau.
- This metric can be used as a **surrogate objective** for black-box order attack.
- Can be optimized by various black-box optimization methods, such as PSO, NES, and SPSA.

Algorithm 1: Short-range Ranking Correlation τ_S.

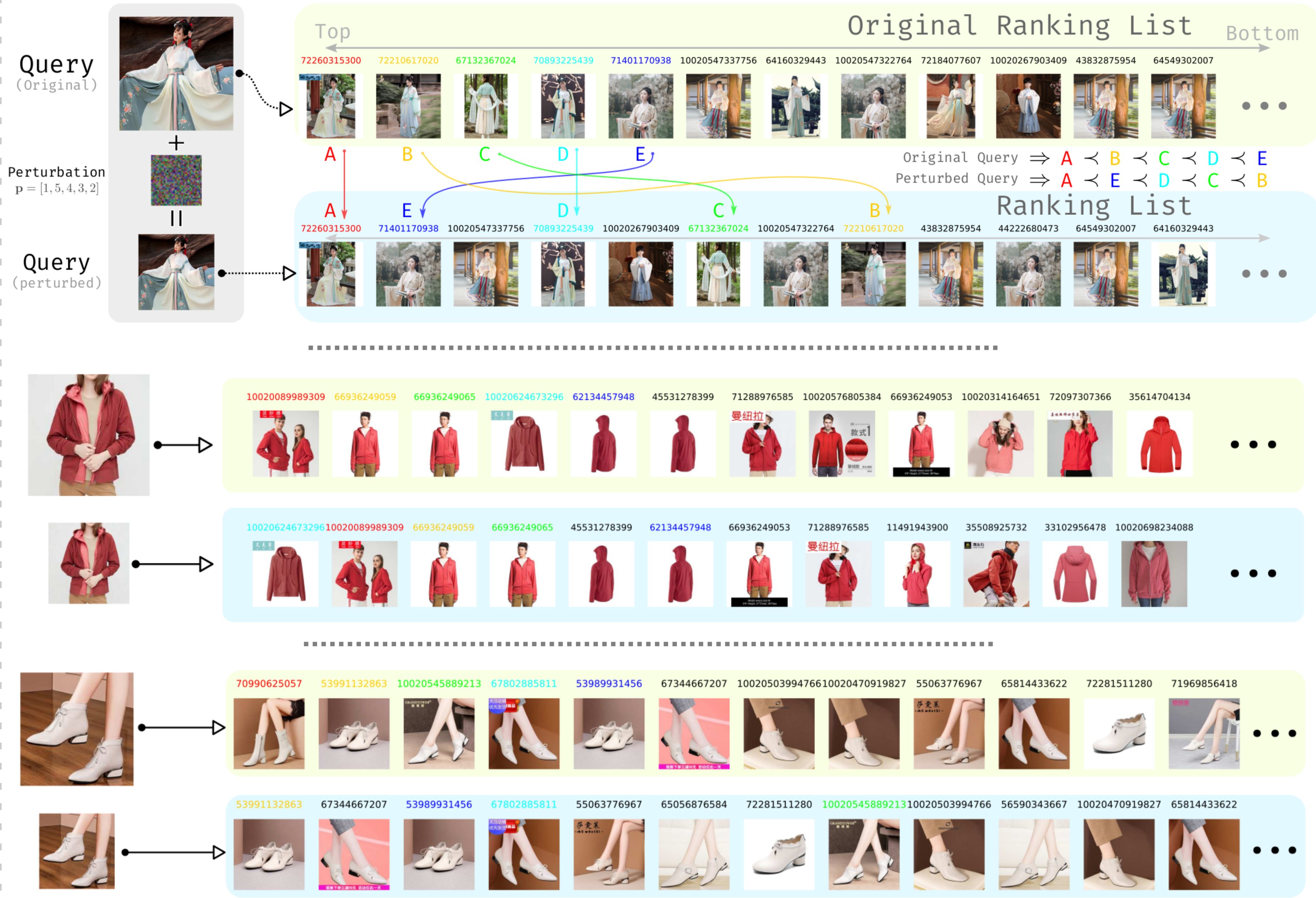
Input: Selected candidates $\mathbb{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_k\}$, permutation vector $\mathbf{p} = [p_1, p_2, \ldots, p_k]$, top-$N$ retrieval $\mathbb{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ for $\tilde{\boldsymbol{q}}$. Note that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

Output: SRC coefficient τ_S.

Permute candidates as $\mathbb{C}_\mathbf{p} = \{\boldsymbol{c}_{p_1}, \boldsymbol{c}_{p_2}, \ldots, \boldsymbol{c}_{p_k}\}$;
Initialize score matrix $\boldsymbol{S} = \mathbf{0}$ of size $k \times k$;
for $i \leftarrow 1, 2, \ldots, k$ **do**
  for $j \leftarrow 1, 2, \ldots, i-1$ **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 $\tau_S = \sum_{i,j} S_{i,j} / \binom{k}{2}$

Practical Order Attack

- To illustrate the viability of the **black-box OA in practice**, we showcase successful attacks against the "JingDong SnapShop" , a major retailing e-commerce platform based on content-based image retrieval.
- The following are qualitative results on JingDong Snapshop API:



- The following are quantitative results on **JD Snapshop API** and **Microsoft Bing Visual Search API** (both are real-world search-by-image APIs):

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.

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.