Adversarial Ranking Attack and Defense Supplementary Material

Anonymous ECCV submission

Paper ID 2274

A Adversarial Example Visualization

Some adversarial ranking examples are presented in this section. Every figure contains three rows of pictures. **The first row** shows c, r, $\tilde{c} = c + r$ for CA, or q, r, $\tilde{q} = q + r$ for QA. **The second row** shows the query and the original retrieval results, as well as the chosen candidate c and its immediately adjacent candidates. **The third row** shows the effects of the attack on the ranking list, *i.e.*, either the chosen candidate c is replaced with $\tilde{c}$ for CA, or the query q is replaced with $\tilde{q}$ for QA. **The digits** above every picture is the value of $\text{Rank}_X(q, \cdot) - 1$ [1].

Pictures with a " ☆ " mark on the top-left corner are adversarial examples. The "←" indicates the chosen candidate whose rank will be raised. The "→" indicates the chosen candidate whose rank will be lowered.

A.1 MNIST Dataset

CA+. See Fig. 1,2. **CA-**. See Fig. 3,4.
QA+. See Fig. 5,6. **QA-**. See Fig. 7,8.

A.2 Fashion-MNIST

CA+. See Fig. 9,10. **CA-**. See Fig. 11,12.
QA+. See Fig. 13,14. **QA-**. See Fig. 15,16.

A.3 Stanford Online Products

CA+. See Fig. 17,18. **CA-**. See Fig. 19,20.
QA+. See Fig. 21,22. **QA-**. See Fig. 23,24.
CA+ ($w = 2$). See Fig. 25, 26. **QA+** ($m = 2$). See Fig. 27, 28.
CA+ ($w = 10$). See Fig. 29.

[1] In mathematical context the rank counts from 1, but in implementation it counts from 0 instead. Hence the offset.

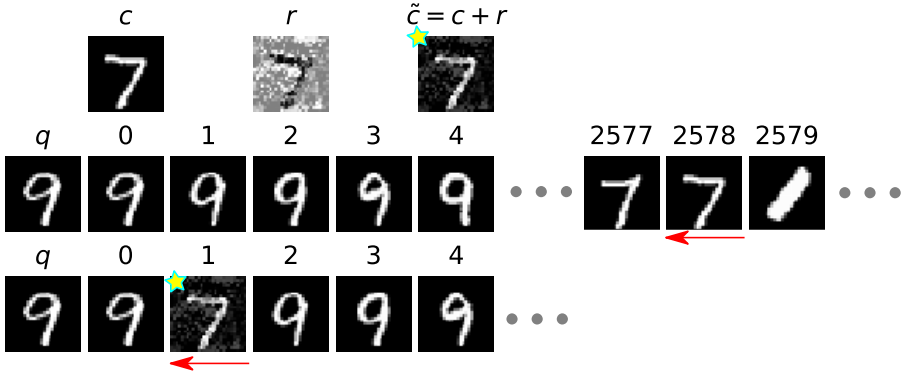

Fig. 1. CA+ on MNIST. Example 1. For query "9", the candidate "7" is ranked at the 2576-th position in the original ranking list (row 2). After adversarial ranking attack, the rank of the perturbed candidate "7" is raised to the 1-st position (row 3). The original candidate "7", the learned perturbation r, and the perturbed candidate "7" are illustrated in the row 1.

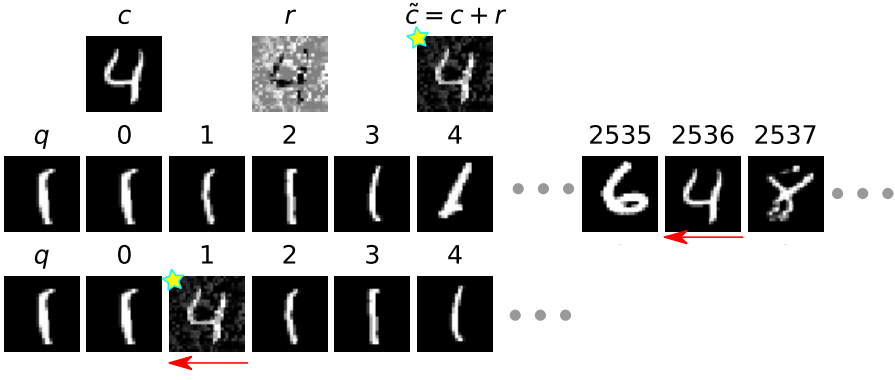

Fig. 2. CA+ on MNIST. Example 2.

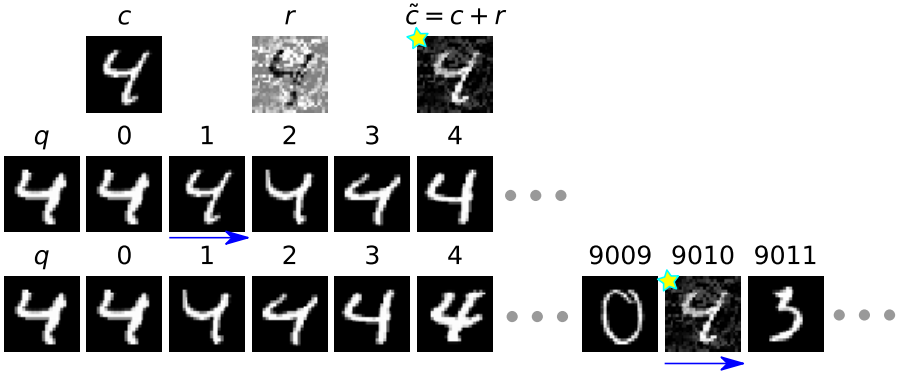

Fig. 3. CA- on MNIST. Example 1.

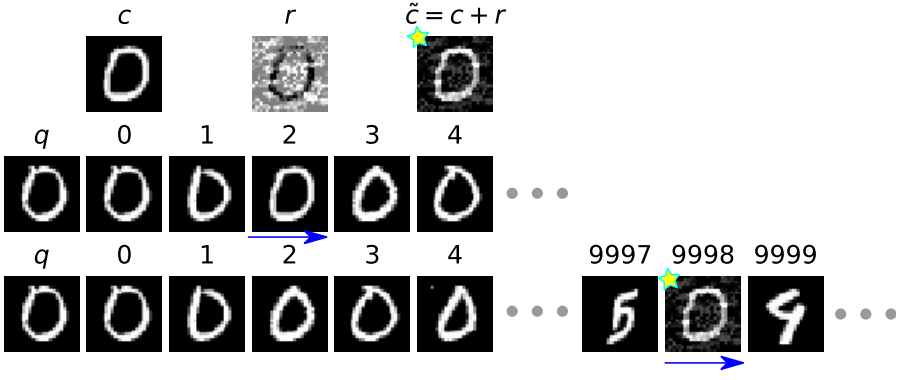

Fig. 4. CA- on MNIST. Example 2.

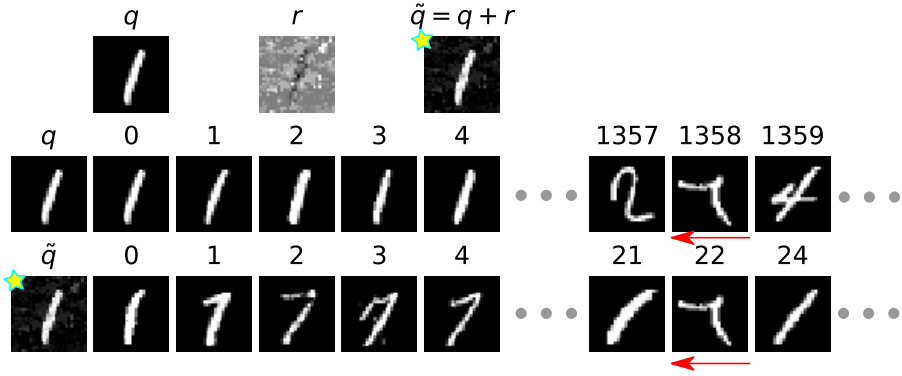

Fig. 5. QA+ on MNIST. Example 1.

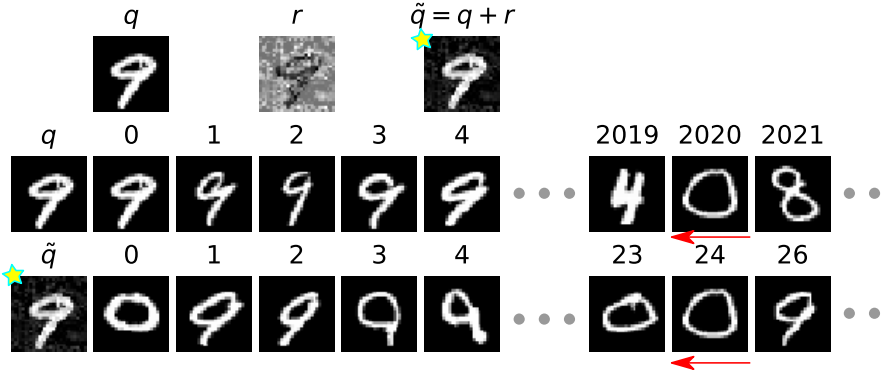

Fig. 6. QA+ on MNIST. Example 2.

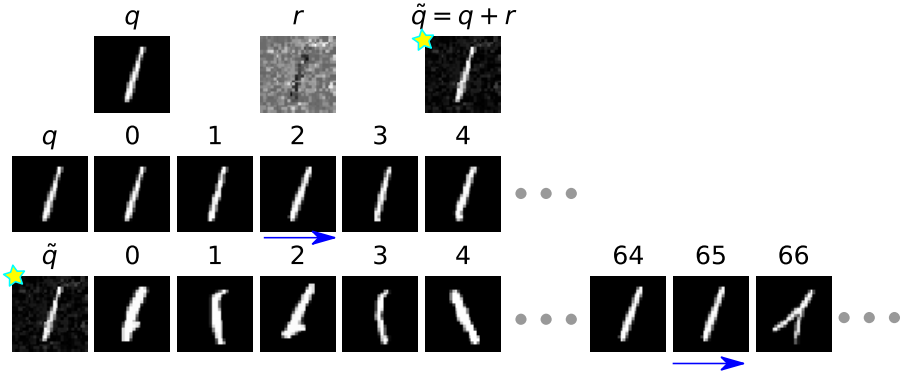

Fig. 7. QA- on MNIST. Example 1.

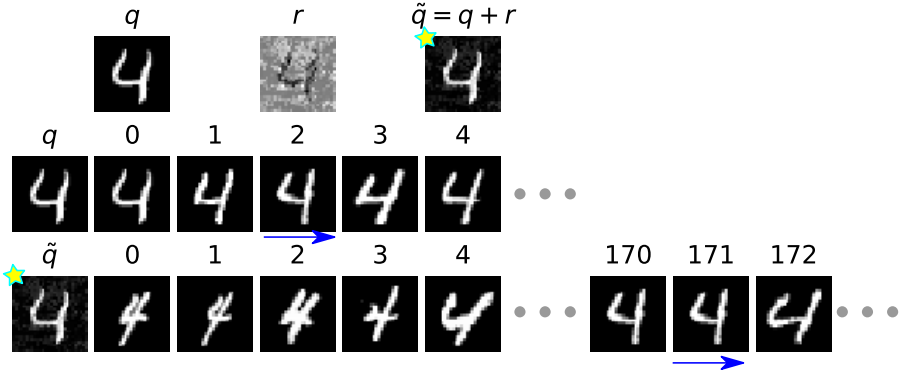

Fig. 8. QA- on MNIST. Example 2.

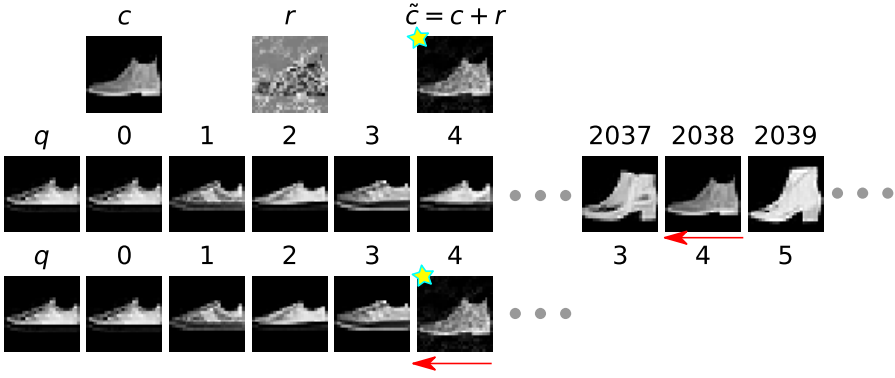

Fig. 9. CA+ on Fashion-MNIST. Example 1.

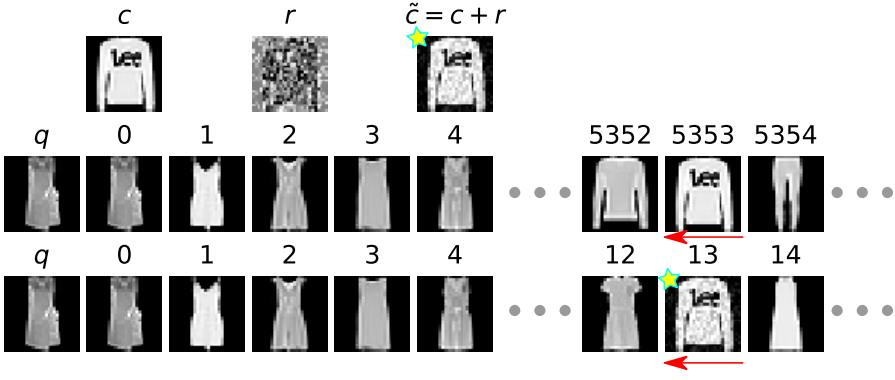

Fig. 10. CA+ on Fashion-MNIST. Example 2.

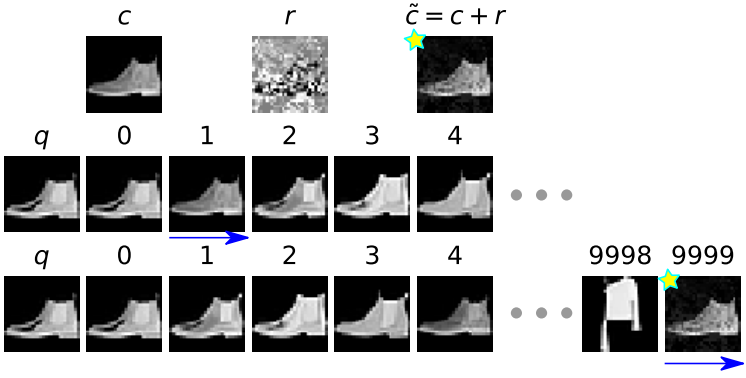

Fig. 11. CA- on Fashion-MNIST. Example 1.

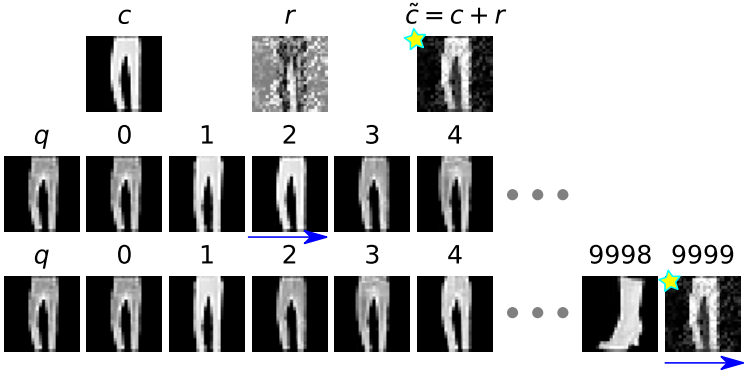

Fig. 12. CA- on Fashion-MNIST. Example 2.

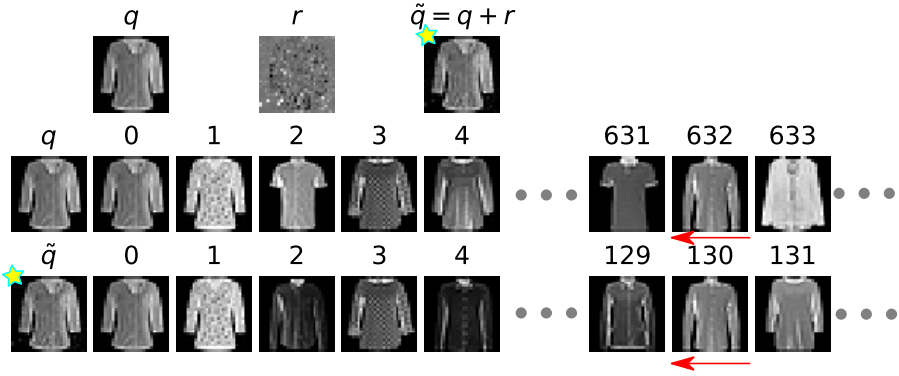

Fig. 13. QA+ on Fashion-MNIST. Example 1.

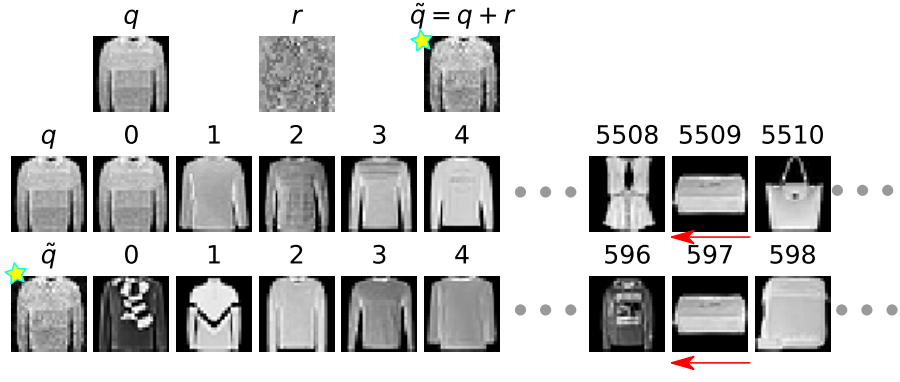

Fig. 14. QA+ on Fashion-MNIST. Example 2.

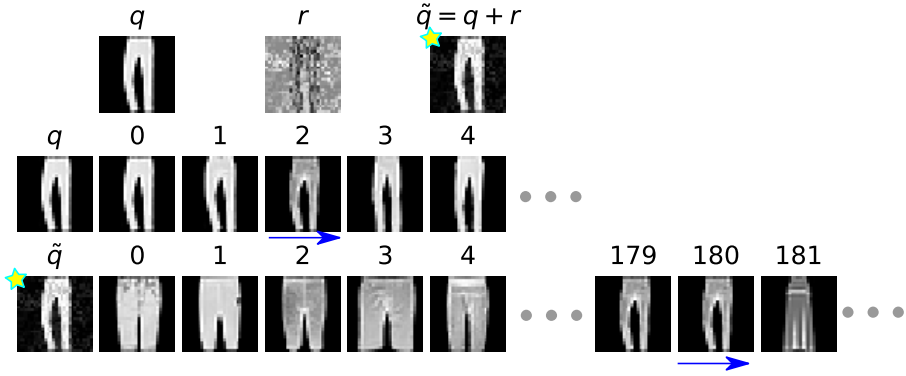

Fig. 15. QA- on Fashion-MNIST. Example 1.

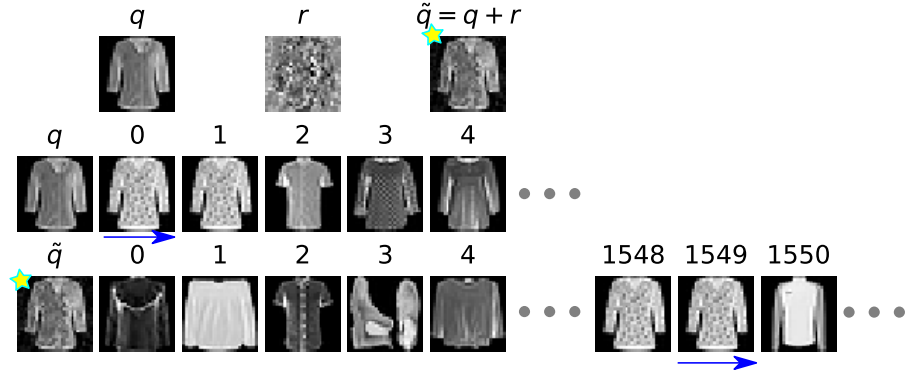

Fig. 16. QA- on Fashion-MNIST. Example 2.

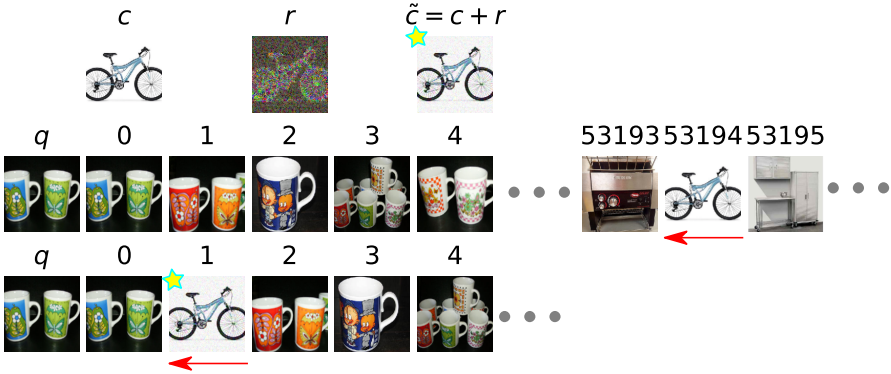

Fig. 17. CA+ on SOP. Example 1.

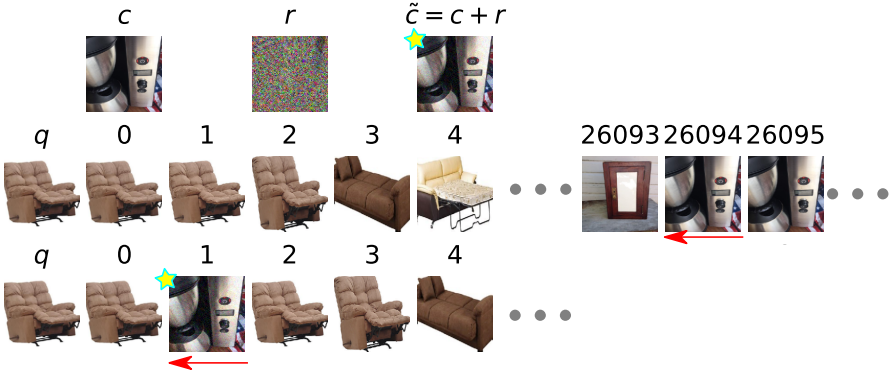

Fig. 18. CA+ on SOP. Example 2.

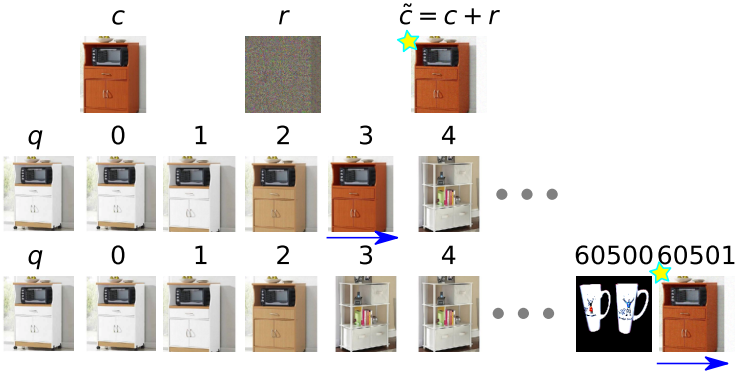

Fig. 19. CA- on SOP. Example 1.

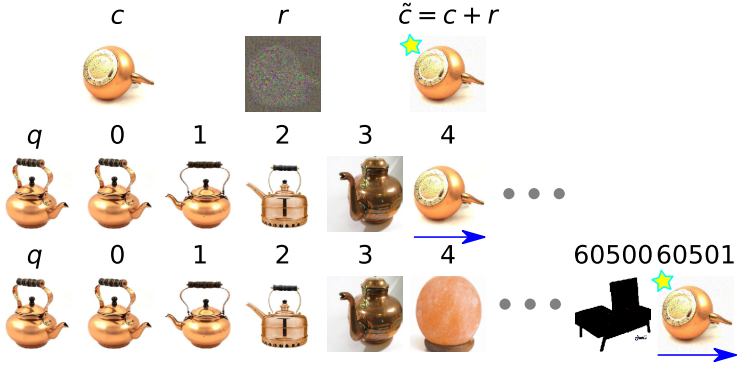

Fig. 20. CA- on SOP. Example 2.

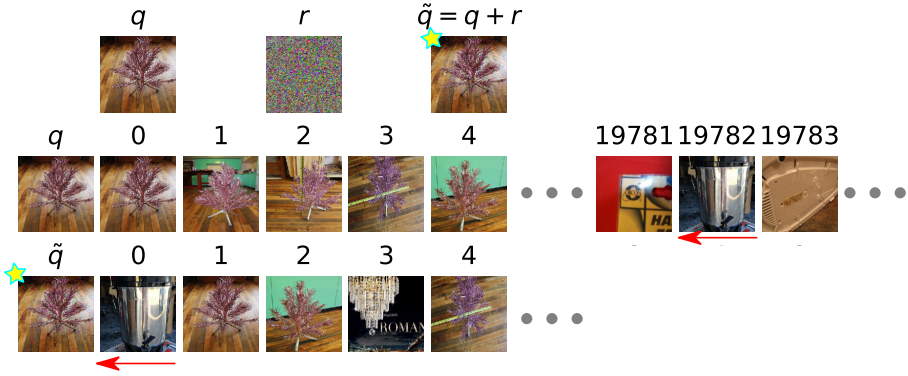

Fig. 21. QA+ on SOP. Example 1.

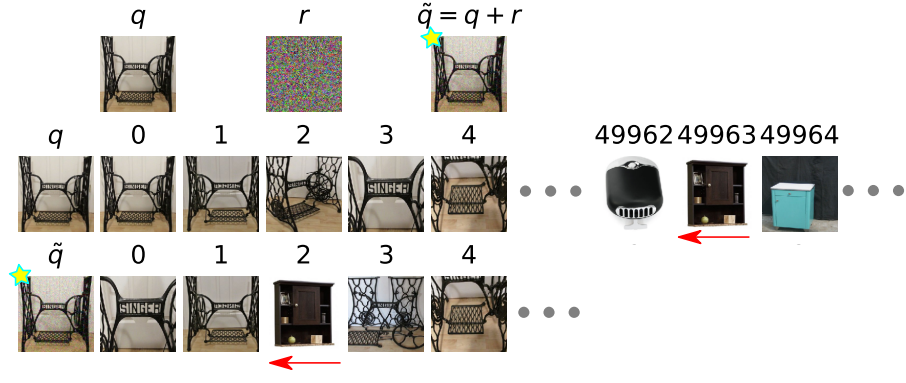

Fig. 22. QA+ on SOP. Example 2.

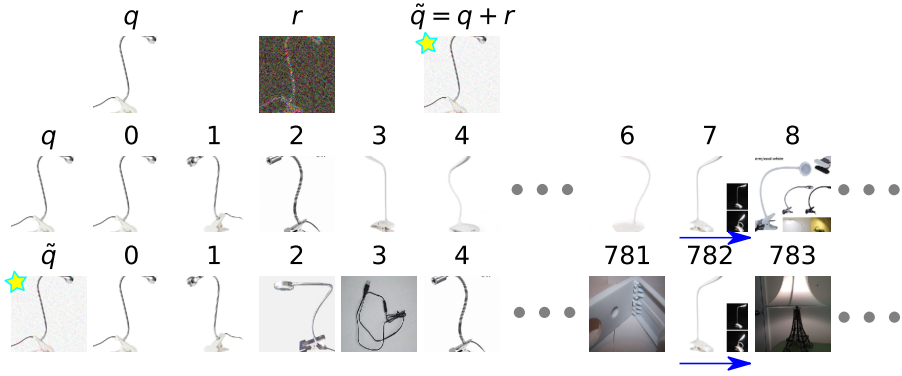

Fig. 23. QA- on SOP. Example 1.

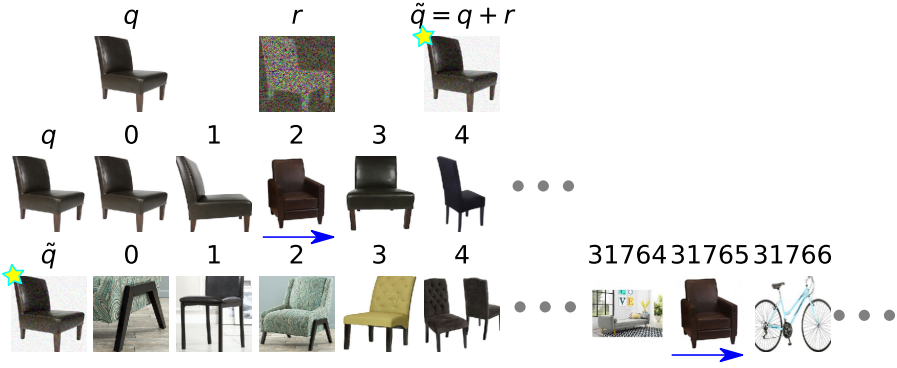

Fig. 24. QA- on SOP. Example 2.

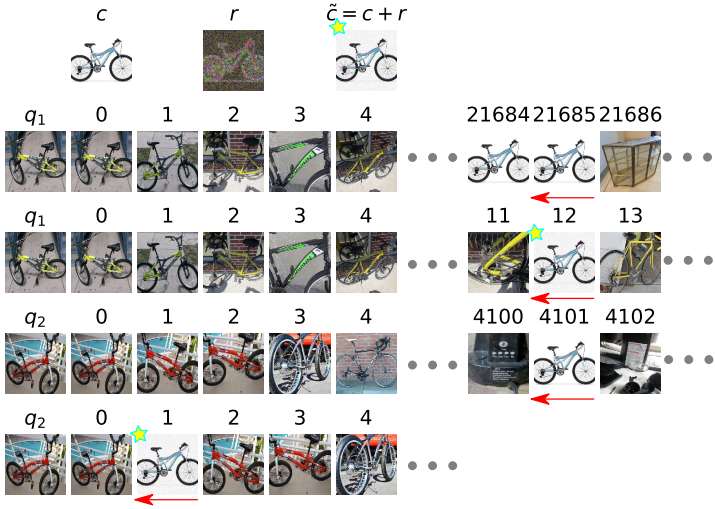

Fig. 25. CA+ ($w = 2$) on SOP. Example 1.

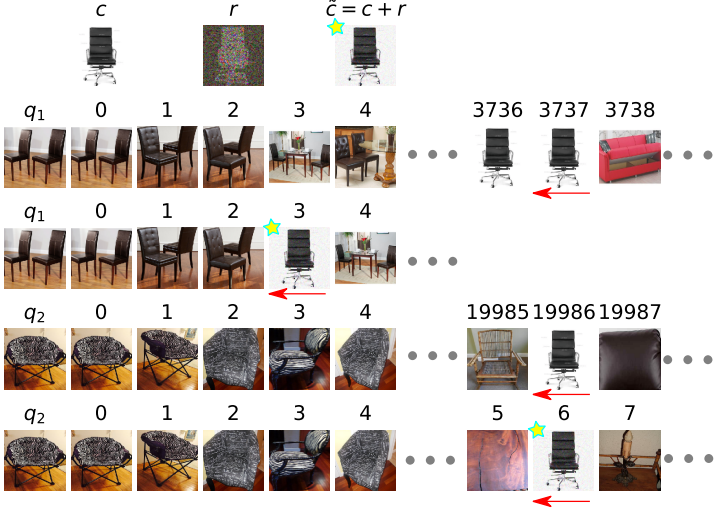

Fig. 26. CA+ ($w = 2$) on SOP. Example 2.

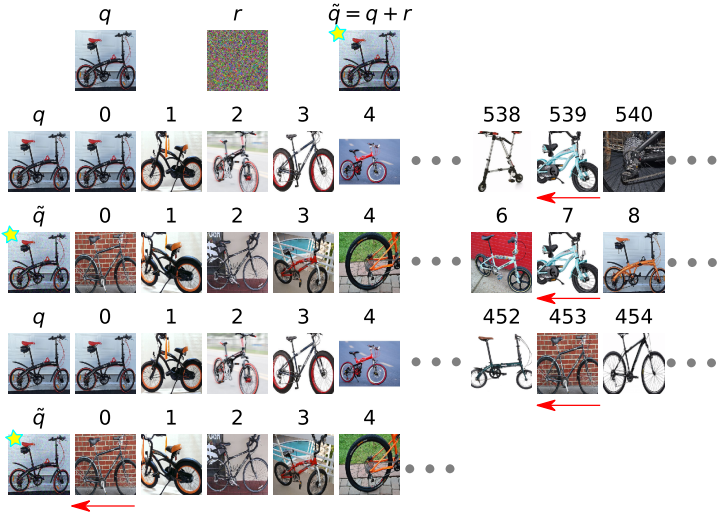

Fig. 27. QA+ ($m = 2$) on SOP. Example 1.

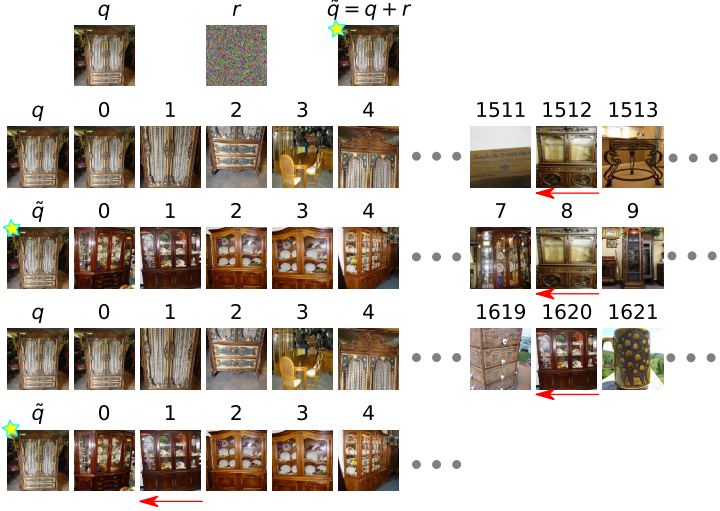

Fig. 28. QA+ ($m = 2$) on SOP. Example 2.

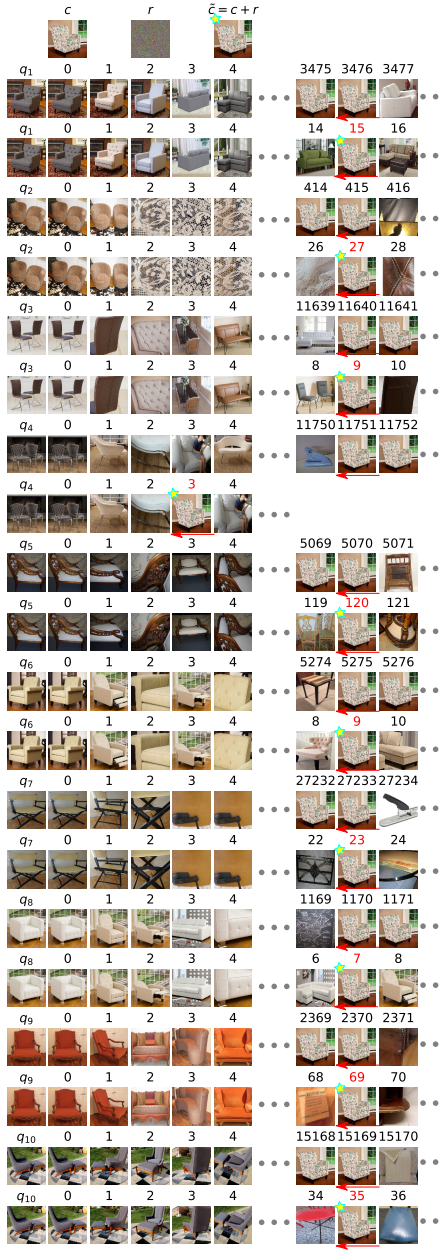

Fig. 29. CA+ ($w = 10$) on SOP.

A.4 "Imperceptibility"

Larger ε increases attacking effectiveness, at the cost of making the adversarial perturbation more perceptible. Following previous works such as [2], we use $\varepsilon = 0.3$ as the maximum attack strength.

See Fig 30 for CA+ with $\varepsilon = 0.3$.
See Fig 31 for CA+ with $\varepsilon = 0.1$.
See Fig 32 for CA+ with $\varepsilon = 0.03$.
See Fig 33 for CA+ with $\varepsilon = 0.01$.

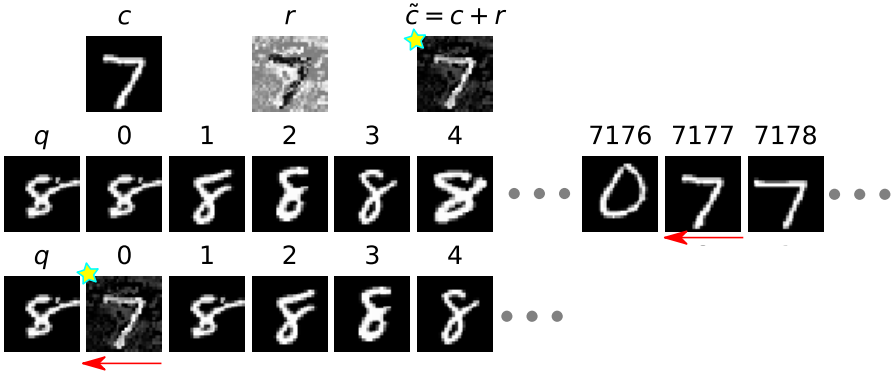

Fig. 30. "Imperceptibility": CA+ on MNIST. $\varepsilon = 0.3$.

A.5 Semantic-Preserving for QA

Conducting Query Attack without preserving the query semantics will often lead to irrelevant retrieval results at the top of the ranking list, as shown in Fig. 34, which raises red flags and possibly reveals the attack. Therefore, the value of the Semantics-Preserving term in QA is substantial, as it keeps the retrieval results as "normal" as possible while achieving the attacking goal.

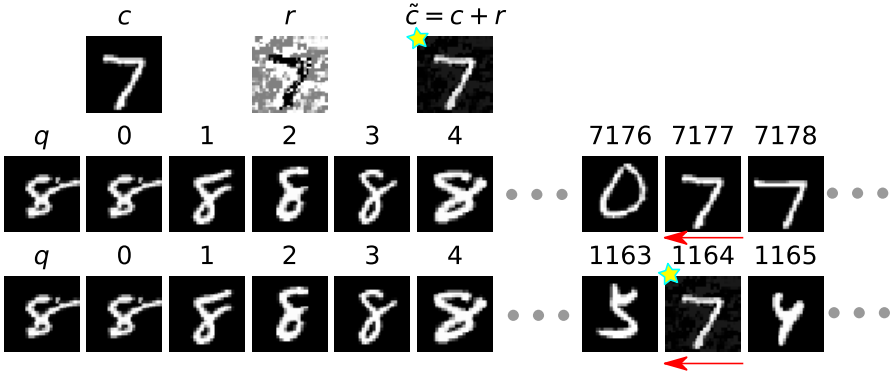

Fig. 31. "Imperceptibility": CA+ on MNIST. $\varepsilon = 0.1$.

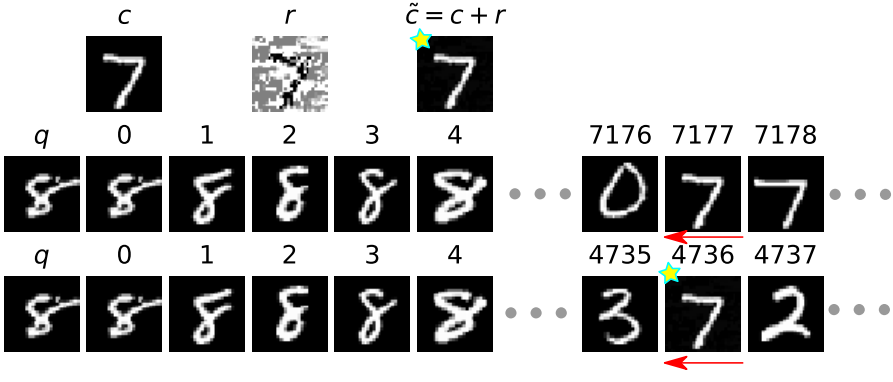

Fig. 32. "Imperceptibility": CA+ on MNIST. $\varepsilon = 0.03$.

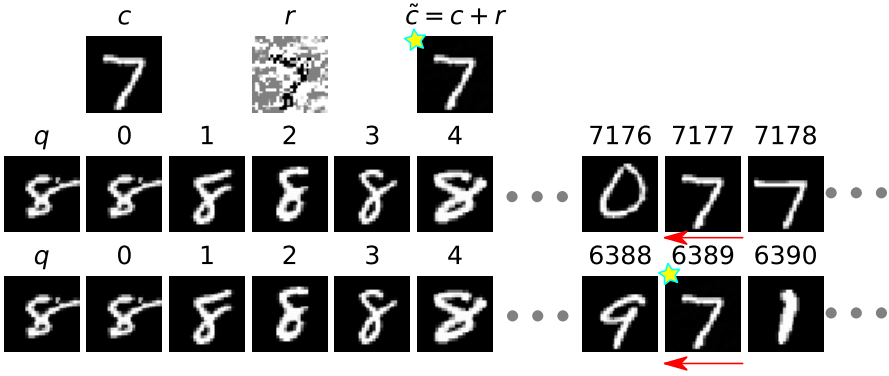

Fig. 33. "Imperceptibility": CA+ on MNIST. $\varepsilon = 0.01$.

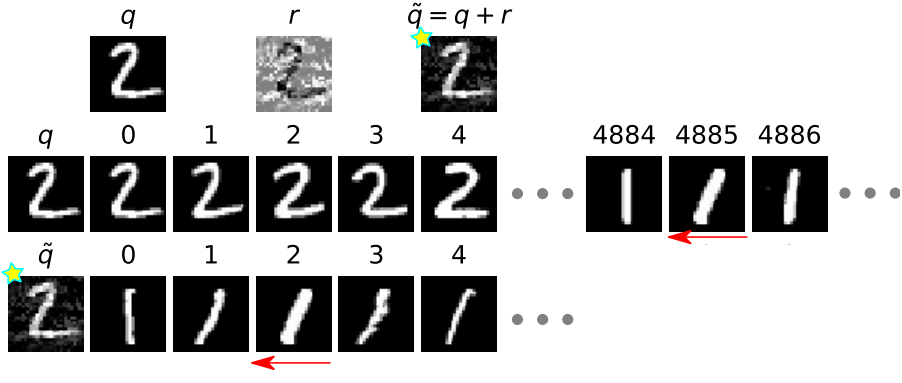

Fig. 34. QA+ without preserving query semantics.

B Complete Results of Attack & Defense

Some experiments and details on MNIST, Fashion-MNIST and SOP datasets are omitted in the manuscript due to limited space. In this section, we present the complete experimental results on these datasets, including the average rank of C_{SP} during QA. Besides, we also conduct attacking experiments on ranking models trained with different combinations of loss functions and distance metrics. For brevity, we denote the (Cosine distance, Triplet loss) setting as (CT), the (Euclidean distance, Contrastive loss) as (EC). Models trained with our defense method will have a trailing "D" in notation, *e.g.*the (CT) model with our defense will be denoted as (CTD).

See Tab. 2 for complete results on MNIST.

See Tab. 3 for complete results on Fashion-MNIST.

See Tab. 1 for results on SOP.

ε	CA+				CA-				SP-QA+				SP-QA-			
	$w=1$	2	5	10	$w=1$	2	5	10	$m=1$	2	5	10	$m=1$	2	5	10
(ET) Euclidean Distance, Triplet Loss (R@1=63.1%)																
0	50	50	50	50	1.9	1.9	1.9	1.9	50	50	50	50	0.5	0.5	0.5	0.5
0.01	0.0	0.8	2.0	2.6	99.7	99.6	99.4	99.3	4.8, 0.2	7.0, 0.7	16.3, 1.9	25.8, 2.3	54.9, 0.3	40.2, 0.6	27.1, 0.8	21.9, 0.9
0.03	0.0	0.3	1.0	1.5	100.0	100.0	100.0	100.0	1.6, 0.1	3.3, 0.5	10.0, 2.0	19.2, 2.7	68.1, 0.3	52.4, 0.6	36.6, 0.8	30.1, 1.0
0.06	**0.0**	**0.2**	**1.0**	**1.5**	**100.0**	**100.0**	**100.0**	**100.0**	**1.1**, 0.2	**2.7**, 0.6	**8.8**, 1.9	**17.6**, 3.3	**73.8**, 0.4	**57.9**, 0.7	**40.3**, 0.8	**32.4**, 1.0
(ETD) Euclidean Distance, Triplet Loss, Defensive (R@1=46.4%)																
0	50	50	50	50	2.0	2.0	2.0	2.0	50	50	50	50	0.5	0.5	0.5	0.5
0.01	7.5	12.2	16.5	18.0	66.4	62.6	59.3	57.8	16.1, 2.1	24.8, 2.7	36.1, 2.7	41.4, 2.5	26.7, 0.6	18.1, 0.7	12.2, 0.7	10.2, 0.7
0.03	0.7	4.5	8.7	10.4	91.7	90.2	89.1	88.4	7.9, 2.9	14.5, 4.2	27.2, 5.4	35.6, 5.3	43.4, 1.3	31.7, 1.5	21.9, 1.7	18.1, 1.8
0.06	**0.1**	**3.8**	**7.9**	**9.7**	**97.3**	**96.8**	**96.4**	**96.2**	**6.9**, 4.3	**12.5**, 5.8	**24.3**, 7.0	**33.4**, 6.9	**51.4**, 2.8	**39.0**, 3.2	**28.0**, 3.5	**23.5**, 3.6

Table 1. Experiments on SOP dataset.

ε	CA+				CA-				SP-QA+				SP-QA-			
	$w=1$	2	5	10	$w=1$	2	5	10	$m=1$	2	5	10	$m=1$	2	5	10
(CC) Cosine Distance, Contrastive Loss (R@1=98.6%)																
0	50	50	50	50	2.2	2.2	2.2	2.2	50	50	50	50	0.5	0.5	0.5	0.5
0.01	42.2	43.3	44.6	44.7	5.6	5.3	5.2	5.1	45.0, 0.5	46.9, 0.5	48.4, 0.5	49.1, 0.5	1.2, 0.5	1.2, 0.5	1.1, 0.5	1.1, 0.5
0.03	31.3	33.5	35.4	36.1	9.3	9.2	9.1	9.1	37.6, 2.3	41.7, 2.5	45.9, 2.5	47.7, 2.5	2.8, 1.8	2.7, 1.8	2.7, 1.8	2.6, 1.8
0.1	9.7	12.5	14.6	15.6	52.1	52.2	52.4	52.5	18.5, 8.2	27.1, 9.7	38.0, 9.9	43.4, 9.9	5.6, 4.1	5.4, 4.1	5.4, 4.1	5.4, 4.1
0.3	**5.8**	**10.0**	**11.9**	**12.9**	**99.0**	**99.1**	**99.0**	**99.1**	**14.3**, 9.7	**23.3**, 12.5	**36.3**, 13.8	**42.6**, 13.9	**6.1**, 4.4	**5.8**, 4.4	**5.7**, 4.4	**5.7**, 4.4
(CT) Cosine Distance, Triplet Loss (R@1=99.1%)																
0	50	50	50	50	2.1	2.1	2.1	2.1	50	50	50	50	0.5	0.5	0.5	0.5
0.01	44.6	45.4	47.4	47.9	3.4	3.2	3.1	3.1	45.2, 0.0	46.3, 0.0	47.7, 0.0	48.5, 0.0	0.9, 0.0	0.7, 0.0	0.6, 0.0	0.6, 0.1
0.03	33.4	37.3	41.9	43.9	6.3	5.9	5.7	5.6	35.6, 0.3	39.2, 0.3	43.4, 0.3	45.8, 0.3	1.9, 0.2	1.4, 0.2	1.1, 0.2	1.1, 0.2
0.1	12.7	17.4	24.4	30.0	15.4	14.9	14.8	14.7	14.4, 2.2	21.0, 2.3	30.6, 2.3	37.2, 2.3	5.6, 1.2	4.4, 1.2	3.7, 1.2	3.5, 1.2
0.3	**2.1**	**9.1**	**13.0**	**17.9**	**93.9**	**93.2**	**93.0**	**92.9**	**6.3**, 3.6	**11.2**, 5.7	**22.5**, 7.7	**32.1**, 7.7	**8.6**, 1.6	**6.6**, 1.6	**5.3**, 1.5	**4.8**, 1.5
(EC) Euclidean Distance, Contrastive Loss (R@1=99.0%)																
0	50	50	50	50	1.8	1.8	1.8	1.8	50	50	50	50	0.5	0.5	0.5	0.5
0.01	31.9	33.5	34.6	34.9	8.2	8.0	7.9	7.8	41.3, 2.0	44.0, 2.2	47.2, 2.4	48.4, 2.5	2.6, 1.3	2.4, 1.3	2.2, 1.3	2.2, 1.3
0.03	15.8	17.4	18.7	19.1	10.7	10.6	10.5	10.5	27.2, 4.8	34.6, 5.4	42.0, 5.6	45.6, 5.8	4.1, 2.5	3.9, 2.5	3.7, 2.5	3.7, 2.5
0.1	6.7	10.0	12.1	12.7	82.1	81.6	81.7	82.0	15.7, 9.8	25.4, 12.0	37.7, 13.1	43.2, 13.1	5.6, 3.2	5.3, 3.2	4.9, 3.2	4.9, 3.2
0.3	**4.8**	**9.9**	**12.1**	**12.7**	**99.8**	**99.8**	**99.8**	**99.8**	**14.4**, 9.4	**23.9**, 12.5	**36.7**, 13.6	**42.6**, 13.8	**5.9**, 3.2	**5.4**, 3.2	**5.0**, 3.2	**4.9**, 3.1
(ET) Euclidean Distance, Triplet Loss (R@1=99.2%)																
0	50	50	50	50	1.6	1.6	1.6	1.6	50	50	50	50	0.5	0.5	0.5	0.5
0.01	39.0	40.4	40.6	40.8	3.2	3.0	2.8	2.8	45.5, 0.1	47.4, 0.1	48.1, 0.1	49.0, 0.1	1.0, 0.1	0.9, 0.1	0.8, 0.1	0.8, 0.1
0.03	23.5	25.6	26.5	27.0	7.0	6.6	6.3	6.2	36.4, 0.6	40.9, 0.7	45.1, 0.8	47.1, 0.9	2.7, 0.7	2.3, 0.7	2.0, 0.7	2.0, 0.8
0.1	8.1	10.6	12.1	12.7	13.7	13.3	13.0	12.9	14.1, 4.8	23.4, 5.7	34.6, 6.2	40.8, 6.5	6.2, 1.7	5.1, 1.7	4.4, 1.7	4.2, 1.7
0.3	**1.8**	**8.5**	**10.9**	**11.7**	**77.4**	**75.7**	**75.0**	**74.7**	**7.5**, 5.2	**15.0**, 7.4	**28.1**, 9.0	**36.3**, 8.9	**7.4**, 1.7	**5.8**, 1.7	**4.9**, 1.7	**4.6**, 1.7
(CCD) Cosine Distance, Contrastive Loss, Defensive (R@1=97.5%)																
0	50	50	50	50	2.2	2.2	2.2	2.2	50	50	50	50	0.5	0.5	0.5	0.5
0.01	49.3	49.0	49.3	49.3	2.4	2.4	2.3	2.3	49.7, 0.0	49.6, 0.0	49.8, 0.0	49.9, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0
0.03	47.0	47.9	48.1	48.0	2.8	2.7	2.7	2.7	48.3, 0.0	49.1, 0.0	49.1, 0.0	49.4, 0.0	0.6, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0
0.1	42.3	43.3	44.0	44.4	4.2	4.0	3.9	3.8	44.9 ,0.1	46.7 ,0.1	48.1 ,0.1	48.8 ,0.1	0.8 ,0.1	0.7 ,0.1	0.7 ,0.1	0.7 ,0.1
0.3	**32.0**	**34.2**	**36.1**	**36.7**	**7.0**	**7.0**	**6.5**	**6.4**	**37.4**, 0.6	**41.1**, 0.6	**44.9**, 0.5	**47.2**, 0.5	**1.9**, 0.5	**1.6**, 0.5	**1.5**, 0.5	**1.5**, 0.5
(CTD) Cosine Distance, Triplet Loss, Defensive (R@1=98.3%)																
0	50	50	50	50	2.0	2.0	2.0	2.0	50	50	50	50	0.5	0.5	0.5	0.5
0.01	48.9	49.3	49.4	49.5	2.2	2.2	2.2	2.1	49.9, 0.0	49.5, 0.0	49.5, 0.0	49.7, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0
0.03	47.4	48.4	48.6	48.9	2.5	2.5	2.4	2.4	48.0, 0.0	48.5, 0.0	49.2, 0.0	49.5, 0.0	0.6, 0.0	0.6, 0.0	0.5, 0.0	0.5, 0.0
0.1	42.4	44.2	45.9	46.7	3.8	3.6	3.5	3.4	43.2, 0.1	45.0, 0.1	47.4, 0.1	48.2, 0.1	1.0, 0.1	0.8, 0.1	0.7, 0.1	0.7, 0.1
0.3	**30.7**	**34.5**	**38.7**	**40.7**	**7.0**	**6.7**	**6.5**	**6.5**	**33.2**, 0.5	**37.2**, 0.5	**42.3**, 0.5	**45.1**, 0.5	**2.4**, 0.4	**1.9**, 0.4	**1.6**, 0.4	**1.5**, 0.4
(ECD) Euclidean Distance, Contrastive Loss, Defensive (R@1=97.9%)																
0	50	50	50	50	1.3	1.3	1.3	1.3	50	50	50	50	0.5	0.5	0.5	0.5
0.01	47.4	47.6	48.0	47.8	2.3	2.2	2.1	2.1	48.7, 0.1	49.1, 0.1	49.4, 0.1	49.7, 0.1	0.9, 0.1	0.8, 0.1	0.7, 0.1	0.7, 0.1
0.03	42.7	43.6	44.0	44.2	4.5	4.2	4.0	4.0	46.3, 0.6	48.1, 0.6	48.8, 0.6	49.2, 0.6	1.8, 0.6	1.6, 0.6	1.5, 0.6	1.4, 0.6
0.1	31.7	33.7	34.9	35.3	10.0	9.6	9.4	9.3	39.2, 2.8	43.2, 2.9	46.6, 2.9	47.9, 2.9	3.3, 1.3	2.9, 1.3	2.7, 1.3	2.6, 1.3
0.3	**19.6**	**23.0**	**25.4**	**26.3**	**35.6**	**35.2**	**35.7**	**36.0**	**27.3**, 7.1	**34.4**, 7.4	**42.2**, 7.1	**45.4**, 6.9	**4.5**, 1.8	**3.9**, 1.8	**3.6**, 1.8	**3.4**, 1.8
(ETD) Euclidean Distance, Triplet Loss, Defensive (R@1=99.0%)																
0	50	50	50	50	1.4	1.4	1.4	1.4	50	50	50	50	0.5	0.5	0.5	0.5
0.01	47.5	48.2	48.1	48.2	1.7	1.7	1.6	1.6	48.5, 0.0	48.8, 0.0	49.4, 0.0	49.7, 0.0	0.6, 0.0	0.6, 0.0	0.5, 0.0	0.5, 0.0
0.03	43.4	43.8	44.1	44.5	2.4	2.3	2.2	2.2	46.6, 0.0	47.8, 0.0	48.9, 0.1	49.3, 0.1	0.8, 0.1	0.7, 0.1	0.6, 0.1	0.6, 0.1
0.1	29.8	31.3	32.6	33.0	5.7	5.2	5.0	4.9	39.6, 0.3	42.9, 0.3	45.9, 0.4	47.9, 0.4	2.1, 0.4	1.7, 0.4	1.5, 0.4	1.4, 0.4
0.3	**10.8**	**13.2**	**15.0**	**15.6**	**14.4**	**13.9**	**13.5**	**13.4**	**19.7**, 3.2	**28.1**, 3.7	**37.4**, 4.1	**42.6**, 4.3	**6.5**, 1.7	**5.3**, 1.7	**4.7**, 1.8	**4.5** 1.8

Table 2. Complete Experiment Results on MNIST. The first value for each QA (*i.e.* SP-QA) experiment result is the average rank of the chosen candidates, while the other value is the average rank of the C_{SP} used for SP.

ε	CA+				CA-				SP-QA+				SP-QA-			
	$w=1$	2	5	10	$w=1$	2	5	10	$m=1$	2	5	10	$m=1$	2	5	10
(CC) Cosine Distance, Contrastive Loss (R@1=88.7%)																
0	50	50	50	50	2.0	2.0	2.0	2.0	50	50	50	50	0.5	0.5	0.5	0.5
0.01	29.8	32.8	34.2	34.7	9.8	9.3	9.0	8.8	36.1, 1.6	40.3, 1.6	44.8, 1.5	46.9, 1.5	3.1, 1.4	2.8, 1.4	2.5, 1.4	2.5, 1.4
0.03	12.5	16.4	19.1	20.3	46.0	45.2	44.7	44.5	18.5, 4.6	25.7, 5.6	35.6, 5.9	41.6, 5.9	4.9, 2.4	4.4, 2.4	4.0, 2.4	3.9, 2.5
0.1	4.8	10.1	13.4	14.9	96.0	96.0	96.0	96.0	10.3, 7.0	17.3, 9.7	30.0, 11.1	38.2, 11.7	7.1, 3.7	6.4, 3.7	5.9, 3.7	5.7, 3.7
0.3	**3.7**	**9.6**	**12.8**	**14.2**	**100.0**	**100.0**	**100.0**	**100.0**	**9.0**, 6.3	**15.7**, 9.3	**28.5**, 11.3	**37.5**, 11.6	**7.2**, 3.5	**6.4**, 3.5	**5.8**, 3.5	**5.6**, 3.4
(CT) Cosine Distance, Triplet Loss (R@1=88.8%)																
0	50	50	50	50	1.9	1.9	1.9	1.9	50	50	50	50	0.5	0.5	0.5	0.5
0.01	36.6	39.9	43.2	44.8	5.6	5.1	4.9	4.8	39.4, 0.2	42.0, 0.2	45.3, 0.2	47.1, 0.2	2.1, 0.2	1.6, 0.2	1.2, 0.2	1.1, 0.2
0.03	19.7	25.4	31.7	35.6	15.5	14.8	14.4	14.3	21.7, 1.5	28.2, 1.6	35.7, 1.7	40.6, 1.7	5.6, 0.8	4.1, 0.8	3.3, 0.8	2.9, 0.7
0.1	3.7	10.5	17.3	22.7	87.2	86.7	86.3	86.3	7.1, 2.4	12.4, 4.5	23.6, 6.9	32.5, 6.8	10.9, 1.9	8.3, 1.9	6.7, 1.9	6.0, 1.8
0.3	**1.3**	**9.4**	**16.0**	**21.5**	**100.0**	**100.0**	**100.0**	**100.0**	**6.3**, 3.0	**10.8**, 5.2	**21.8**, 7.8	**31.7**, 8.3	**12.6**, 1.9	**9.4**, 1.9	**7.5**, 1.9	**6.6**, 1.8
(EC) Euclidean Distance, Contrastive Loss (R@1=87.6%)																
0	50	50	50	50	1.3	1.3	1.3	1.3	50	50	50	50	0.5	0.5	0.5	0.5
0.01	20.9	23.6	25.7	26.4	12.3	11.7	11.3	11.1	30.6, 3.7	36.7, 3.8	42.8, 3.9	45.8, 3.8	3.8, 1.6	3.3, 1.6	3.0, 1.6	2.9, 1.6
0.03	7.0	11.5	14.8	16.0	74.1	72.9	71.8	71.4	15.7, 8.0	25.4, 9.3	36.9, 9.1	42.4, 8.8	4.6, 2.0	4.0, 2.0	3.6, 2.0	3.6, 2.1
0.1	6.2	12.2	17.0	18.7	100.0	100.0	100.0	100.0	14.2, 9.3	22.6, 12.3	34.8, 13.2	41.2, 13.0	8.3, 4.7	7.8, 4.8	7.3, 4.8	7.1, 4.8
0.3	**5.7**	**12.0**	**16.5**	**18.2**	**100.0**	**100.0**	**100.0**	**100.0**	**12.8**, 9.0	**21.1**, 11.9	**33.8**, 13.5	**40.4**, 13.4	**9.1**, 5.1	**8.4**, 5.2	**7.9**, 5.0	**7.5**, 4.9
(ET) Euclidean Distance, Triplet Loss (R@1=88.3%)																
0	50	50	50	50	1.5	1.5	1.5	1.5	50	50	50	50	0.5	0.5	0.5	0.5
0.01	33.2	34.8	34.7	36.1	6.5	5.9	5.5	5.3	41.7, 0.3	44.4, 0.3	47.3, 0.3	48.4, 0.3	2.5, 0.3	1.9, 0.3	1.6, 0.3	1.4, 0.3
0.03	14.0	17.6	20.3	21.3	18.4	17.2	16.4	16.0	22.2, 2.5	30.0, 2.7	39.4, 2.7	43.6, 2.6	6.6, 1.3	5.1, 1.4	4.2, 1.4	3.9, 1.4
0.1	1.7	9.5	14.0	15.6	88.3	86.7	85.2	84.5	8.0, 3.9	15.0, 6.3	27.6, 7.9	36.5, 8.1	10.5, 1.9	8.0, 1.9	6.4, 2.0	6.0, 2.0
0.3	**0.3**	**9.0**	**13.8**	**15.5**	**100.0**	**100.0**	**100.0**	**100.0**	**6.7**, 3.3	**12.6**, 5.7	**25.4**, 7.9	**34.8**, 8.3	**11.9**, 1.9	**8.9**, 1.9	**7.1**, 1.9	**6.4**, 1.9
(CCD) Cosine Distance, Contrastive Loss, Defensive (R@1=82.2%)																
0	50	50	50	50	2.0	2.0	2.0	2.0	50	50	50	50	0.5	0.5	0.5	0.5
0.01	47.0	47.7	47.8	47.9	2.3	2.3	2.3	2.2	48.9, 0.0	48.7, 0.0	49.2, 0.0	49.6, 0.0	0.6, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0
0.03	42.2	43.3	44.1	44.3	3.1	3.0	2.9	2.8	45.2, 0.0	47.2, 0.0	48.0, 0.0	49.1, 0.0	0.8, 0.0	0.7, 0.0	0.6, 0.0	0.6, 0.0
0.1	29.1	31.4	32.9	33.8	8.0	7.1	6.4	6.2	34.7, 0.1	39.5, 0.1	43.8, 0.1	46.5, 0.1	3.3, 0.1	2.1, 0.1	1.5, 0.1	1.3, 0.1
0.3	**11.8**	**14.8**	**17.4**	**18.4**	**28.7**	**25.5**	**23.1**	**22.4**	**13.3**, 0.7	**20.0**, 0.9	**31.0**, 1.0	**38.3**, 1.0	**21.3**, 1.0	**14.3**, 1.0	**10.3**, 1.0	**8.6**, 1.1
(CTD) Cosine Distance, Triplet Loss, Defensive (R@1=79.6%)																
0	50	50	50	50	1.2	1.2	1.2	1.2	50	50	50	50	0.5	0.5	0.5	0.5
0.01	48.9	48.9	49.3	49.3	1.4	1.4	1.4	1.4	49.4, 0.0	49.9, 0.0	49.9, 0.0	50.0, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0
0.03	47.1	47.9	48.3	48.3	2.0	1.9	1.8	1.8	48.3, 0.0	49.1, 0.0	49.5, 0.0	49.8, 0.0	0.7, 0.0	0.6, 0.0	0.6, 0.0	0.6, 0.0
0.1	42.4	43.5	44.5	44.8	4.6	4.2	4.0	3.9	45.4, 0.3	47.2, 0.2	48.7, 0.2	49.2, 0.2	1.4, 0.2	1.2, 0.2	1.1, 0.2	1.0, 0.2
0.3	**32.5**	**35.4**	**37.5**	**38.2**	**11.2**	**10.5**	**10.1**	**10.0**	**39.3**, 1.5	**42.6**, 1.5	**46.5**, 1.3	**47.8**, 1.3	**3.9**, 1.4	**3.3**, 1.4	**3.0**, 1.4	**2.9**, 1.4
(ECD) Euclidean Distance, Contrastive Loss, Defensive (R@1=80.4%)																
0	50	50	50	50	1.4	1.4	1.4	1.4	50	50	50	50	0.5	0.5	0.5	0.5
0.01	45.3	45.8	46.3	46.6	3.0	2.8	2.7	2.6	48.0, 0.2	48.7, 0.2	49.3, 0.3	49.7, 0.3	1.2, 0.2	1.0, 0.2	0.9, 0.2	0.9, 0.2
0.03	37.6	39.5	40.4	40.7	7.5	7.0	6.8	6.7	43.9, 1.5	46.4, 1.6	48.4, 1.7	49.1, 1.9	2.7, 0.7	2.2, 0.7	1.9, 0.7	1.9, 0.7
0.1	20.3	24.9	27.8	28.6	32.6	32.1	32.0	32.5	28.4, 6.6	36.2, 7.2	43.0, 7.4	45.9, 7.4	5.3, 1.5	4.4, 1.6	3.8, 1.6	3.7, 1.6
0.3	**7.3**	**16.0**	**21.1**	**22.5**	**78.0**	**79.0**	**80.3**	**81.2**	**14.0**, 10.2	**24.6**, 11.9	**36.5**, 11.8	**42.2**, 11.2	**7.0**, 1.9	**5.7**, 1.9	**4.9**, 1.9	**4.6**, 1.9
(ETD) Euclidean Distance, Triplet Loss, Defensive (R@1=84.4%)																
0	50	50	50	50	1.3	1.3	1.3	1.3	50	50	50	50	0.5	0.5	0.5	0.5
0.01	46.0	46.7	46.9	46.7	2.0	1.9	1.8	1.8	48.1, 0.0	49.1, 0.0	49.3, 0.0	49.6, 0.0	0.8, 0.0	0.7, 0.0	0.6, 0.0	0.6, 0.1
0.03	38.9	40.6	41.2	41.4	4.3	3.8	3.5	3.4	45.2, 0.2	46.2, 0.1	48.1, 0.1	48.9, 0.1	1.9, 0.2	1.5, 0.2	1.2, 0.2	1.1, 0.2
0.1	23.0	26.3	28.0	28.8	15.9	14.9	14.2	13.9	31.8, 1.8	38.0, 1.8	43.3, 1.6	46.2, 1.6	5.6, 1.0	4.3, 1.0	3.6, 1.0	3.3, 1.0
0.3	**7.1**	**13.9**	**18.4**	**19.9**	**54.1**	**51.8**	**50.0**	**49.2**	**10.6**, 5.6	**19.7**, 7.1	**32.5**, 8.0	**39.6**, 7.6	**9.2**, 1.6	**7.0**, 1.6	**5.6**, 1.6	**5.1**, 1.6

Table 3. Complete Experiments on Fashion-MNIST Dataset. Each QA result comprises the average rank of the chosen candidates and the C_{SP} used for SP, respectively.

C Complete Results on Transferability

C.1 Fashion-MNIST

In addition to the transferability experiment on MNIST dataset, we also conduct the same transferability experiment on the Fashion-MNIST dataset, as shown in Tab. 4.

CA+ Transfer (Black Box), $w = 1$			
From \ To	LeNet	C2F1	Res18
LeNet	50→16.0	41.0	44.3
C2F1	38.6	50→1.3	40.3
Res18	39.2	34.3	50→1.7

CA- Transfer (Black Box), $w = 1$			
From \ To	LeNet	C2F1	Res18
LeNet	2.5→84.3	1.9→8.1	1.6→6.0
C2F1	2.5→7.8	1.9→100.0	1.7→7.7
Res18	2.5→9.5	1.9→14.4	1.7→80.0

SP-QA+ Transfer (Black Box), $m = 1$			
From \ To	LeNet	C2F1	Res18
LeNet	50→18.0	47.2	49.3
C2F1	48.1	50→6.4	49.2
Res18	48.1	44.8	50→13.7

SP-QA- Transfer (Black Box), $m = 1$			
From \ To	LeNet	C2F1	Res18
LeNet	0.5→13.5	0.5→1.7	0.5→1.5
C2F1	0.5→1.1	0.5→12.5	0.5→1.6
Res18	0.5→0.9	0.5→1.3	0.5→8.0

Table 4. Transferability experiment on Fashion-MNIST dataset.

C.2 "Self-Transfer" Attack on MNIST

In the adversarial ranking example transferability experiments, we transfer adversarial examples between neural networks with different architectures. In this section, we transfer adversarial examples between neural networks based on the same architecture but with different parameters.

We train three vanilla C2F1 models (denoted as C2F1-1, C2F1-2, and C2f1-3) on MNIST dataset and two models with our defense (denoted as C2F1-D1, C2F1-D2). All these models have exactly the same architecture, but different parameters. Transferability experimental results between these models are present in Tab. 5.

CA+ Transfer (Black Box), $w = 1$					
From \ To	C2F1-1	C2F1-2	C2F1-3	C2F1-D1	C2F1-D2
C2F1-1	50→2.1	27.9	25.3	45.3	45.1
C2F1-2	24.4	50→2.7	23.7	44.5	44.0
C2F1-3	22.6	24.4	50→2.2	45.0	44.5
C2F1-D1	36.7	37.7	37.4	50→30.9	38.6
C2F1-D2	37.3	36.8	37.3	39.6	50→30.3

CA- Transfer (Black Box), $w = 1$					
From \ To	C2F1-1	C2F1-2	C2F1-3	C2F1-D1	C2F1-D2
C2F1-1	2.1→93.7	2.2→9.4	2.1→11.1	2.0→3.2	1.9→3.1
C2F1-2	2.1→13.5	2.1→88.1	2.2→12.3	2.0→3.4	1.9→3.3
C2F1-3	2.1→13.4	2.2→10.1	2.2→92.4	2.0→3.2	1.9→3.2
C2F1-D1	2.1→9.3	2.2→8.3	2.1→9.1	2.0→7.0	2.0→4.9
C2F1-D2	2.1→9.2	2.2→8.4	2.1→9.1	2.0→4.9	1.9→7.0

SP-QA+ Transfer (Black Box), $m = 1$					
From \ To	C2F1-1	C2F1-2	C2F1-3	C2F1-D1	C2F1-D2
C2F1-1	50→6.4	39.4	38.1	48.6	48.1
C2F1-2	36.6	50→6.3	36.4	47.3	48.4
C2F1-3	36.9	37.8	50→6.2	48.7	48.3
C2F1-D1	41.5	40.3	41.4	50→32.9	40.6
C2F1-D2	41.7	38.9	41.7	41.4	50→32.7

SP-QA- Transfer (Black Box), $m = 1$					
From \ To	C2F1-1	C2F1-2	C2F1-3	C2F1-D1	C2F1-D2
C2F1-1	0.5→8.8	0.5→1.4	0.5→1.9	0.5→0.6	0.5→0.6
C2F1-2	0.5→2.3	0.5→9.0	0.5→2.2	0.5→0.6	0.5→0.6
C2F1-3	0.5→1.9	0.5→1.5	0.5→8.7	0.5→0.6	0.5→0.6
C2F1-D1	0.5→5.8	0.5→3.1	0.5→4.8	0.5→2.4	0.5→1.3
C2F1-D2	0.5→7.1	0.5→3.7	0.5→5.9	0.5→1.3	0.5→2.4

Table 5. Transfer experiment between models based on the same architecture but with different parameters.

D Complete Results for Universal Perturbation

D.1 MNIST

See Tab. 6.

Model	I-CA+ ($w = 1$)		I-CA- ($w = 1$)		I-QA+ ($m = 1$)		I-QA- ($m = 1$)	
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
(CC)	50 → 13.7	50 → 14.1	0.6 → 37.6	0.6 → 33.4	50 → 18.1	50 → 18.9	2.4 → 41.6	2.4 → 39.3
(CT)	50 → 18.1	50 → 18.5	0.6 → 9.5	0.7 → 9.4	50 → 20.5	50 → 21.0	2.1 → 7.6	2.2 → 7.4
(EC)	50 → 9.1	50 → 10.3	1.9 → 94.6	2.0 → 79.4	50 → 3.9	50 → 6.9	3.3 → 87.1	3.5 → 74.4
(ET)	50 → 11.5	50 → 12.6	2.1 → 10.6	2.1 → 9.6	50 → 21.6	50 → 23.6	3.2 → 28.6	3.2 → 19.5

Table 6. Universal Adversarial Ranking Perturbation on MNIST. Each pair of result presents the original rank of chosen candidate(s), and the rank of the chosen candidate(s) after adding adversarial perturbation to the candidate or to the query. "Seen" samples are those used for generating the universal perturbation, while "Unseen" samples are a set of other non-overlapping samples.

D.2 Fashion-MNIST

As shown in Tab. 7, Image-agnostic adversarial perturbation has better effect on seen data from Fashion-MNIST, but the gap between the effect on seen samples and that on unseen samples is slightly larger, which may due to the higher intra-class variance of Fashion-MNIST than MNIST.

Model	I-CA+ ($w = 1$)		I-CA- ($w = 1$)		I-QA+ ($m = 1$)		I-QA- ($m = 1$)	
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
(CC)	50 → 7.0	50 → 7.3	0.6 → 91.8	0.6 → 84.9	50 → 4.4	50 → 4.9	2.1 → 87.5	2.1 → 84.4
(CT)	50 → 9.8	50 → 9.9	0.6 → 72.3	0.6 → 69.7	50 → 8.2	50 → 8.4	2.0 → 76.3	2.0 → 72.9
(EC)	50 → 5.8	50 → 9.6	2.0 → 97.5	1.9 → 83.7	50 → 1.8	50 → 7.1	2.9 → 87.9	2.8 → 78.4
(ET)	50 → 5.7	50 → 8.5	2.0 → 84.4	1.9 → 69.9	50 → 3.3	50 → 6.3	3.1 → 88.0	3.0 → 78.0

Table 7. Image-agnostic Adversarial Perturbation on Fashion-MNIST.

E Semantics Preserving for QA & Parameter Search for ξ

As discussed previously, the Query Attack (**QA**) may drastically change the semantics of the query q. To alleviate this problem, the Semantics-Preserving (SP) term is added to the naive **QA** to help preserve the query semantics. Predictably, it is more difficult to perform **QA** with a large ξ, as the ranks of C_{SP} are almost not allowed to be changed.

To investigate the actual influence of the balancing parameter ξ, we provide parameter search on it with MNIST dataset. In particular, We set ξ to $0, 10^0, 10^2, 10^4$, and compare their results. Note that when $\xi = 0$, the QA becomes *naive* QA as the SP term is eliminated. With a strong SP constant, *e.g.*$\xi = 10^4$, the semantics of the chosen query is almost not allowed to be changed, hence result in extreme difficulty of attack.

As shown in Tab. 8, setting ξ to 0 could greatly boost the attacking effect, but consequently the ranks of C_{SP} will be drastically changed. In contrast, when ξ is set to the excessive value 10^4 for a perfectly stealth QA, the attack can still raise the rank of chosen candidate from 50% to 37.9% in QA+ with $m = 1$, or lower the rank of chosen candidate from 0.5% to 1.9% in QA- with $m = 1$. During these attacks, the ranks of C_{SP} are kept within 0.1 despite of the extreme difficulty. It means the query semantics can be preserved. In practice, we empirically set the parameter ξ as 1 for QA+, or as 10^2 for QA- for the balance between attack effectiveness and preserving query semantics.

ξ	SP-QA+				SP-QA-			
	m=1	2	5	10	m=1	2	5	10
(CC) Cosine, Contrastive								
0	1.4, 43.9	13.8, 38.1	30.7, 37.9	39.4, 36.3	94.7, 93.3	94.4, 93.5	94.2, 93.7	94.1, 93.8
10^0	14.3, 9.7	23.3, 12.5	36.3, 13.8	42.6, 13.9	40.6, 37.6	40.6, 38.1	40.4, 38.1	39.8, 37.7
10^2	25.4, 2.0	33.2, 2.1	41.7, 2.1	45.4, 2.1	6.1, 4.4	5.8, 4.4	5.7, 4.4	5.7, 4.4
10^4	48.1, 0.3	49.2, 0.2	49.6, 0.2	49.8, 0.3	1.2, 0.3	1.1, 0.3	1.1, 0.3	1.1, 0.3
(CT) Cosine, Triplet								
0	0.2, 33.6	6.3, 23.7	18.5, 26.5	29.6, 25.7	94.1, 89.4	93.2, 90.3	92.6, 90.9	92.3, 91.2
10^0	6.3, 3.6	11.2, 5.7	22.5, 7.7	32.1, 7.7	55.5, 35.6	52.4, 37.6	50.2, 39.3	49.4, 40.0
10^2	14.1, 0.6	20.8, 0.7	31.2, 0.7	38.1, 0.7	8.6, 1.6	6.6, 1.6	5.3, 1.5	4.8, 1.5
10^4	37.9, 0.1	42.6, 0.1	46.3, 0.1	47.8, 0.1	1.9, 0.1	1.4, 0.1	1.2, 0.1	1.1, 0.1
(EC) Euclidean, Contrastive								
0	0.7, 44.5	13.4, 39.9	31.0, 41.8	39.7, 41.6	94.0, 92.6	93.9, 93.0	93.9, 93.4	93.9, 93.5
10^0	14.4, 9.4	23.9, 12.5	36.7, 13.6	42.6, 13.8	39.9, 37.6	39.0, 37.2	39.2, 37.7	38.7, 37.3
10^2	30.1, 1.2	37.3, 1.1	43.9, 0.9	46.7, 0.9	5.9, 3.2	5.4, 3.2	5.0, 3.2	4.9, 3.1
10^4	50.1, 0.8	50.2, 0.8	49.8, 0.8	50.0, 0.8	1.6, 0.7	1.6, 0.7	1.7, 0.8	1.6, 0.8
(ET) Euclidean, Triplet								
0	0.6, 37.4	8.2, 35.1	22.6, 38.2	33.2, 38.8	93.0, 88.4	92.5, 89.9	92.3, 90.9	92.1, 90.9
10^0	7.5, 5.2	15.0, 7.4	28.1, 9.0	36.3, 8.9	50.9, 37.3	49.2, 39.4	47.7, 40.6	47.2, 40.9
10^2	20.3, 0.5	28.6, 0.5	38.2, 0.5	43.1, 0.4	7.4, 1.7	5.8, 1.7	4.9, 1.7	4.6, 1.7
10^4	46.2, 0.1	48.2, 0.1	49.1, 0.1	49.6, 0.1	1.7, 0.1	1.4, 0.1	1.2, 0.1	1.1, 0.1

Table 8. Parameter search on Semantics-Preserving balancing parameter ξ with MNIST. We report two two mean ranks in each cell: one for the chosen candidate(s) C, the other for the C_{SP} used for SP.

F Analysis on Our Defense

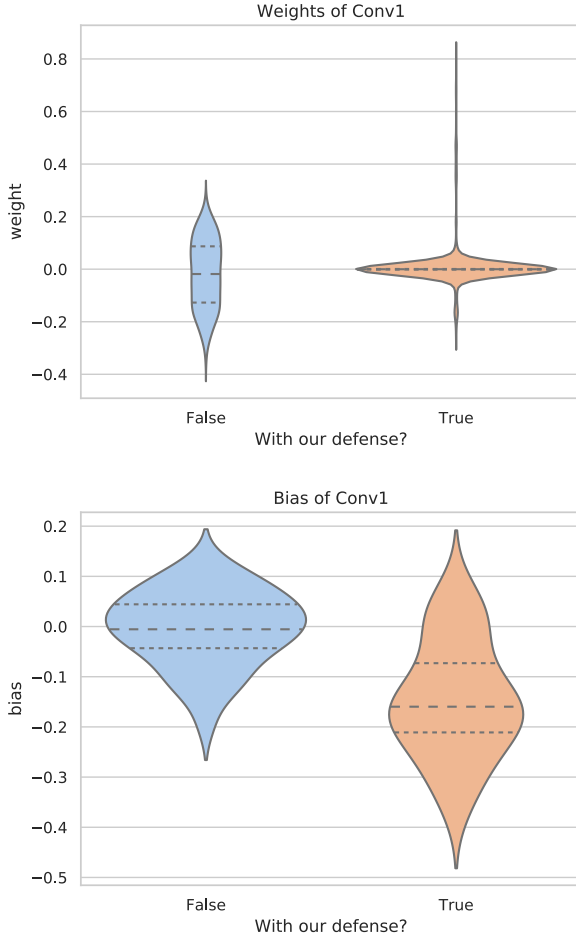

Fig. 35. Weights and bias of the first convolutional layer (conv1) with/without our defense.

Ian *et al.* [1] proposed a linear explanation of adversarial examples:

> Consider the dot product between a weight vector w and an adversarial example $\tilde{x} = x + r$ ($\|r\|_\infty < \varepsilon$):
>
> $$w^T \tilde{x} = w^T x + w^T r. \quad (1)$$

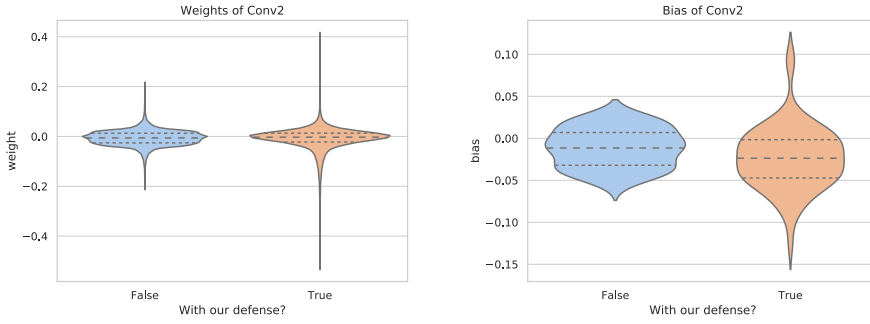

Fig. 36. Weights and bias of the second convolutional layer (conv2) with/without our defense.

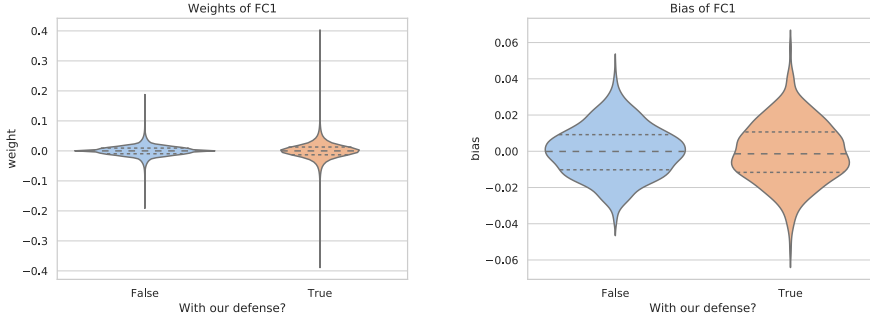

Fig. 37. Weights and bias of the first fully-connected layer (fc1) with/without our defense.

> The adversarial perturbation causes the activation to grow by $w^T r$. We can maximize this increase subject to the max norm constraint on r by assigning $r = \varepsilon \text{sign}(w)$. If w has k dimensions and the average magnitude of an element of the weight vector is h, then the activation will grow by εkh. Since $\|r\|_\infty$ does not grow with the dimensionality of the problem but the change in activation caused by perturbation by r can grow linearly with k, then for high dimensional problems, we can make many infinitesimal changes to the input that add up to one large change to the output.

In our experiments, we analysed the parameters in models trained on MNIST, as shown in Fig. 35, 36, 37. These violin plots suggest that

– the weights in the first convolution layer of the defensive model are closer to 0 and have smaller variance than those of the vanilla model. That means

the h in the above quotation is decreased with our defense, and it will be harder for the r to incur a large increase in activation εkh.
- the bias in the first convolution layer of the defensive model tend to be negative values instead of being nearly "zero-mean". These negative bias could help further suppress the increase in activation caused by purturbation r.

Therefore, Ian *et al.*'s [1] theory could explain why our defense works, as the resulting differences in network parameters may help reduce the probability for adversarial perturbation to increase the layer outputs into the local linear area of ReLU.

G Alternative Attack

G.1 Distance-based Ranking Attack

In order to implement the proposed ranking attack, alternative attacking objectives are possible. Some related works such as **Feature Adversary** [3] generates untargeted adversarial examples against classifiers by maximizing the distance shift of representation vectors off their original locations. This may inspire an alternative version of CA or QA objective functions which are directly based on distance. For example, such alternative objective for CA+ and QA- could be as follows:

$$r = \arg\min_{r \in \Gamma} \sum_{q \in Q} d(q, c + r) \tag{2}$$

$$r = \arg\max_{r \in \Gamma} \sum_{c \in C} d(q + r, c). \tag{3}$$

However, it must be pointed out that our method significantly differs from feature adversary [3]: **(1)** Feature adversary concerns the *pairwise* similarity of source-target representations, while our image ranking problem concerns the ranking order of multiple candidates; **(2)** Feature adversary attempts to reduce the ℓ_2 distance as much as possible, while our triplet-like loss attempts to make positive candidates closer to query than the negative ones, which well fits the objective of ranking order optimization. Such *relative* distance optimization becomes more important since our attack simultaneously involves multiple queries and multiple candidates; **(3)** The ℓ_2 distance based methods suffer from inevitable disadvantages. Specifically, distance-based objectives are suboptimal, because they disregard the relative positions among the candidates and queries.

As shown in the top-left part of Fig. 38, the solution set for for distance-based CA+ (the green dotted line) contains suboptimal solutions. Similarly, in the bottom-left part of Fig. 38 the distance-based objective for QA- tends to maximize the sum of distance neglecting the ranking result, and further optimization (moving $\tilde{q}$ along the green arrow) will not change the ranking result. In contrast, our proposed inequality-based method does not suffer from these issues, as shown in the top-right and bottom-right parts.

We also implemented such distance-based method and compared it with our triplet-like method, as shown in Tab. 9. Experimental results suggest that our method always outperforms distance-based method by a margin. Especially for QA-, the distance-based objective is very difficult to optimize because the distance-based Semantics-Preserving term contradicts with the other term in the loss function. In summary, distance-based method is not well-suited for our proposed *adversarial ranking attack*, especially in the scenario of QA-.

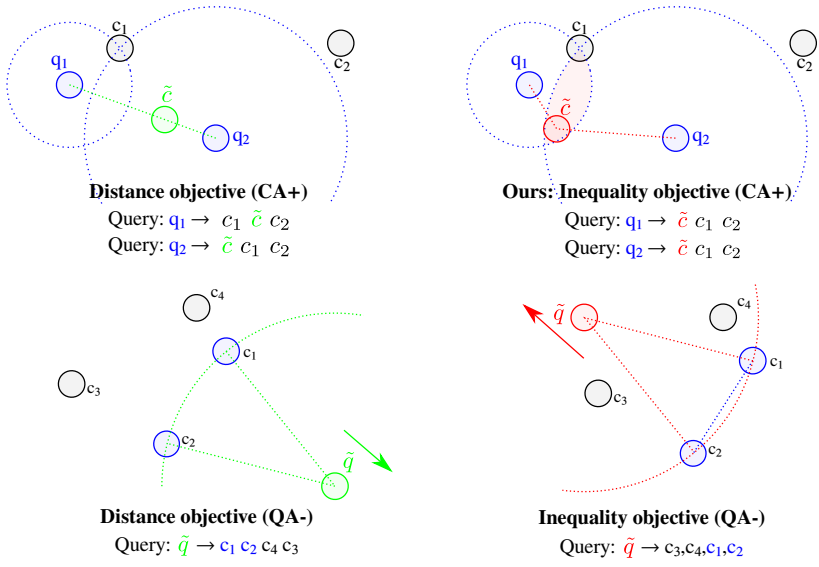

Fig. 38. Distance-based CA+ (top-left) and QA- (bottom-left) objective functions do not lead to desired attacking result compared to our inequality-based method (top-right and bottom-right). In the top-left diagram for CA+, q_1 and q_2 are the chosen queries, while $\tilde{c}$ is the adversarial candidate found by the distance objective. In the bottom-left diagram of QA-, c_1 and c_2 are the chosen candidates, while $\tilde{q}$ is the adversarial query found by the distance objective. It is noted that the distance objective is not optimal as shown in the top-left part, where the solution set for minimizing the distance objective contains suboptimal results. Distance-based objective may even fail to change the ranking result as shown in the bottom-left part, where optimizing the objective further cannot change the ranking result. In contrast, our proposed inequality-based method does not suffer from these issues, as shown in the top-right and bottom-right parts.

ε	CA+				CA-				QA+				QA-			
	$w=1$	2	5	10	$w=1$	2	5	10	$m=1$	2	5	10	$m=1$	2	5	10
0.3	3.0	9.5	15.9	22.2	86.0	85.2	84.7	84.6	7.4	20.2	34.9	41.7	0.8	0.8	0.8	0.8

Table 9. Attack Based on L-2 Distance Loss with MNIST.

H Alternative Defense

Apart from the defense provided in the manuscript, we also tried some other loss functions for adversarial training. In literature, there is no predominant choice for the adversarial training loss function. The only common trait among these choices is that all of them involve adversarial examples. Inspired by previous works, we also implement some alternative defenses for ranking systems as follows.

H.1 Straightforward Adaptation of Madry Defense

Madry [2] formularised improving neural network classifier robustness as a min-max optimization problem, where the inner maximization seeks to generate adversarial examples that lead to maximum cross-entropy loss L_{CE}, while the outer minimization tunes the neural network parameters θ to suppress the cross-entropy loss:

$$\min_{\theta} \left\{ \mathbb{E}_{(x,y)\sim D} \left[\max_{r \in \Gamma} L_{\text{CE}}(x + r, y) \right] \right\} \tag{4}$$

where (x, y) is a pair of image and ground-truth class label.

Similarly, we follow the idea and use a similar defense for ranking models:

$$\min_{\theta} \left\{ \mathbb{E}_{(q,c_p,c_n)\sim D} \left[\max_{r \in \Gamma} L_{\text{triplet}}(q + r, c_p, c_n) \right] \right\} \tag{5}$$

where the inner maximization aims to generate strongest adversarial examples that could lead to triplet ranking error and a large loss value, while the outer minimization seeks network parameters that could reduce such error.

However, during experiments, we observe that such defensive loss function always diverges, possibly due to the adversarial examples generated by the inner problem being too "strong". We leave further investigation into this problem for future work.

H.2 Straightforward Adaptation of Ian Defense

Ian [1] proposed the following loss function for adversarial training:

$$L_{\text{Ian}}(x, y) = \alpha L_{\text{CE}}(x, y) \tag{6}$$

$$+ (1 - \alpha) L_{\text{CE}}(x + \varepsilon \text{sign}(\nabla_x L_{\text{CE}}(x, y))) \tag{7}$$

where the first term is a normal Cross-Entropy loss, and the second term is the cross-entropy loss with untargeted adversarial example that aims to increase the loss value. Constant α is a balancing parameter.

When adapted to a deep ranking system, this defense method also suffers from the diverging issue similar to Madry's defense.

H.3 Directly Suppressing Shift Distance

As discussed in the manuscript, another possible adversarial training method could be to directly suppress the maximum shift distance of embedding vectors, *i.e.*:

$$L_{\text{trip-es}} = L_{\text{triplet}}(q, c_p, c_n) + \sum_{x \in \{q, c_p, c_n\}} \big(\max_{r \in \Gamma} d(x + r, x) \big), \tag{8}$$

where the most severe distance shift incurred by adversarial perturbations is explicitly suppressed, in addition to a standard ranking loss term.

Experimental results (Tab. 10) show that cosine distance-based ranking models are more robust with this defense. However, we note that this loss may numerically explode on an Euclidean distance-based embedding model, as a strong adversary can gradually cause very large embedding shift distance. To mitigate such divergence issue, we also tried to add a balancing parameter to greatly scale down the second term of the loss function, but the instability problem was still not alleviated.

Due to the lack of universality, we leave this alternative defense in supplementary material as a pure discussion, and possible improvements as future work.

ε	CA+				CA-				SP-QA+				SP-QA-			
	$w=1$	2	5	10	$w=1$	2	5	10	$m=1$	2	5	10	$m=1$	2	5	10
(CCD) Cosine Distance, Contrastive Loss, Defensive (R@1=95.3%)																
0	50	50	50	50	1.2	1.2	1.2	1.2	50	50	50	50	0.5	0.5	0.5	0.5
0.01	49.1	49.3	49.6	49.5	1.3	1.3	1.3	1.3	49.7, 0.0	49.9, 0.0	50.0, 0.0	49.8, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0
0.03	48.0	48.2	48.5	48.6	1.6	1.5	1.5	1.5	48.7, 0.0	49.2, 0.0	49.7, 0.0	49.7, 0.0	0.6, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0
0.1	43.1	44.4	45.2	45.5	2.4	2.3	2.1	2.1	45.3, 0.1	47.4, 0.1	48.6, 0.1	49.4, 0.1	0.8, 0.1	0.7, 0.1	0.6, 0.1	0.6, 0.1
0.3	**33.3**	**35.8**	**37.4**	**38.0**	**5.6**	**5.1**	**4.8**	**4.7**	**38.2**, 0.3	**42.3**, 0.3	**45.7**, 0.3	**47.6**, 0.3	**2.1**, 0.4	**1.7**, 0.4	**1.5**, 0.4	**1.5**, 0.4
(CTD) Cosine Distance, Triplet Loss, Defensive (R@1=97.4%)																
0	50	50	50	50	1.5	1.5	1.5	1.5	50	50	50	50	0.5	0.5	0.5	0.5
0.01	49.3	49.6	49.5	49.7	1.6	1.6	1.6	1.6	49.5, 0.0	49.7, 0.0	50.0, 0.0	50.0, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0
0.03	48.2	48.3	48.8	48.8	1.8	1.8	1.8	1.7	49.3, 0.0	48.9, 0.0	49.5, 0.0	49.7, 0.0	0.6, 0.0	0.5, 0.0	0.5, 0.0	0.5, 0.0
0.1	44.7	45.4	46.3	46.4	2.8	2.6	2.5	2.5	46.3, 0.1	47.4, 0.1	48.6, 0.1	49.2, 0.1	0.7, 0.1	0.7, 0.1	0.6, 0.1	0.6, 0.1
0.3	**35.5**	**38.5**	**40.3**	**40.9**	**5.7**	**5.3**	**5.1**	**5.0**	**39.3**, 0.4	**43.1**, 0.3	**46.0**, 0.3	**47.5**, 0.3	**1.8**, 0.4	**1.6**, 0.4	**1.4**, 0.4	**1.4**, 0.4

Table 10. The loss that directly suppresses embedding shift distance with MNIST dataset.

References

1. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
2. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
3. Sabour, S., Cao, Y., Faghri, F., Fleet, D.J.: Adversarial manipulation of deep representations. arXiv preprint arXiv:1511.05122 (2015)