Supplementary Material: Geometrically Adaptive Dictionary Attack on Face Recognition

Junyoung Byun, Hyojun Go, Changick Kim
Korea Advanced Institute of Science and Technology (KAIST)
{bjyoung, gohyojun15, changick}@kaist.ac.kr

1. Overview

In this supplementary material, we describe hyperparameter settings for attack methods used in our experiments. We also illustrate the perturbation norm curves that visually show each method’s query efficiency and additional qualitative results for comprehensive comparisons. We also list the experimental results of the decision-based black-box attacks against a deeper target model, CurricularFace ResNet-100 [8, 6].

2. Additional details of experimental setting

We use Pytorch framework [10] for our experiments and borrow the code for face recognition from face.evoLVe library. We use the 112 × 112 aligned datasets provided by face.evoLVe library for the LFW [7] and CPLFW [11] datasets.

3. Implementation and hyperparameter settings of the attacks

Sign-OPT (SO) [3]. We adopt the code of Sign-OPT provided by the authors without special tuning of the hyperparameters of the attack.

HSJA [1]. We implement HSJA by using Adversarial Robustness Toolbox (ART) library [9]. From its default setting, we increase the maximum iterations to 64 to follow the authors’ experimental settings.

EA [5]. We implement EA based on the code of EA provided by the authors. We set the dimension of the perturbation search space as 60 × 60 × 3 and the coefficient of the distance in calculation of σ as 0.03 instead of 0.01 for faster convergence. These settings also apply to the EA’s variants (EAD, EAG, EAGD).

SFA [2]. We adopt the code of SFA provided by the authors without special tuning of the hyperparameters. We set the dimension reduction ratio as 2 for reducing the perturbation search space. These settings also apply to the SFA’s variants (SFAD, SFAG, SFAGD).

4. More experimental results on the ArcFace ResNet-50 model

We illustrate the perturbation norm curves that visually show the query efficiency of each attack method in Fig. 1 and Fig. 2. We show additional examples for more extensive qualitative comparison in Fig. 3 to Fig. 17. In detail, Fig. 3 and Fig. 4 show the results of dodging attacks on the LFW dataset. Fig. 5 through Fig. 8 display the examples of impersonation attacks on the LFW dataset. Fig. 9 to Fig. 12 show the results of dodging attacks on the CPLFW dataset. Fig. 13 to Fig. 17 show the examples of impersonation attacks on the CPLFW dataset.

5. Experimental results on the CurricularFace ResNet-100 model

We conduct the decision-based black-box attacks against the CurricularFace ResNet-100 which is trained on the refined MS1MV2 dataset [4]. We arrange the experimental results in Table 1. Clearly, GADA greatly improves the query efficiency of EA.

As the model is deeper and more accurate, the robustness to attack increases, so the minimum perturbation norm is higher in all datasets than that of the results of the ArcFace ResNet-50 model [4, 6]. We illustrate the perturbation norm curves that visually show each attack method’s query efficiency in Fig. 18 and Fig. 19.

1https://github.com/ZhaoJ9014/face.evoLVe
2https://github.com/cmhcbb/attackbox
3https://github.com/Jianbo-Lab/HSJA
4https://github.com/thu-ml/realsafe/blob/master/realsafe/attack/evolutionary_worker.py
5https://github.com/wubaoyuan/Sign-Flip-Attack
6We use the pretrained model from https://github.com/HuangYG123/CurricularFace
Table 1: Evaluation of decision-based black-box attacks against the CurricularFace ResNet-100 [8, 6] with the two datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minimum perturbation norm with query budget</td>
<td>Avg. # queries for perturbation with norm</td>
</tr>
<tr>
<td></td>
<td>1K</td>
<td>2K</td>
</tr>
<tr>
<td>EA [5]</td>
<td>14.16</td>
<td>7.23</td>
</tr>
<tr>
<td>EAD</td>
<td>11.16</td>
<td>6.15</td>
</tr>
<tr>
<td>EAG</td>
<td>8.68</td>
<td>4.65</td>
</tr>
<tr>
<td>EAGD</td>
<td>7.06</td>
<td>4.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minimum perturbation norm with query budget</td>
<td>Avg. # queries for perturbation with norm</td>
</tr>
<tr>
<td></td>
<td>1K</td>
<td>2K</td>
</tr>
<tr>
<td>EA [5]</td>
<td>17.41</td>
<td>10.29</td>
</tr>
<tr>
<td>EAG</td>
<td>13.09</td>
<td>7.63</td>
</tr>
</tbody>
</table>

Table 1: Evaluation of decision-based black-box attacks against the CurricularFace ResNet-100 [8, 6] with the two datasets.

References


Figure 1: Perturbation norm curves of the decision-based attacks against ArcFace ResNet-50 model with the LFW dataset [7].

Figure 2: Perturbation norm curves of the decision-based attacks against ArcFace ResNet-50 model with the CPLFW dataset [11].
Figure 3: Qualitative results of dodging attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.

Figure 4: Qualitative results of dodging attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.
Figure 5: Qualitative results of impersonation attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.

Figure 6: Qualitative results of impersonation attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.
Figure 7: Qualitative results of impersonation attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.

Figure 8: Qualitative results of impersonation attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.
Figure 9: Qualitative results of dodging attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.

Figure 10: Qualitative results of dodging attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.
Figure 11: Qualitative results of dodging attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.

Figure 12: Qualitative results of dodging attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.
Figure 13: Qualitative results of impersonation attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.

Figure 14: Qualitative results of impersonation attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.
Figure 15: Qualitative results of impersonation attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.

Figure 16: Qualitative results of impersonation attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.
Figure 17: Qualitative results of impersonation attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The $\ell_2$ norm of perturbation is displayed under each image.

Figure 18: Perturbation norm curves of the decision-based attacks against the CurricularFace ResNet-100 with the LFW dataset [7].
Figure 19: Perturbation norm curves of the decision-based attacks against the CurricularFace ResNet-100 with the CPLFW dataset [11].