Query-Free Attacks on Industry-Grade Face Recognition Systems under Resource Constraints

Di Tang

Chinese University of Hong Kong
XiaoFeng Wang

Indiana University
Kehuan Zhang

Chinese University of Hong Kong

Abstract

To attack a deep neural network (DNN) based Face Recognition (FR) system, one needs to build *substitute* models to simulate the target, so the adversarial examples discovered could also mislead the target. Such *transferability* is achieved in recent studies through querying the target to obtain data for training the substitutes. A real-world target, likes the FR system of law enforcement, however, is less accessible to the adversary. To attack such a system, a substitute with similar quality as the target is needed to identify their common defects. This is hard since the adversary often does not have the enough resources to train such a model (hundreds of millions of images for training a commercial FR system).

We found in our research, however, that a resource-constrained adversary could still effectively approximate the target's capability to recognize *specific* individuals, by training *biased* substitutes on additional images of those who want to evade recognition (the subject) or the victims to be impersonated (called Point of Interest, or PoI). This is made possible by a new property we discovered, called *Nearly Local Linearity* (NLL), which models the observation that an ideal DNN model produces the image representations whose distances among themselves truthfully describe the differences in the input images seen by human. By simulating this property around the PoIs using the additional subject or victim data, we significantly improve the transferability of black-box impersonation attacks by nearly 50%. Particularly, we successfully attacked a commercial system trained over 20 million images, using 4 million images and 1/5 of the training time but achieving 60% transferability in an impersonation attack and 89% in a dodging attack.

1 Introduction

With its commercial success, deep learning (DL) based face recognition (FR) is haunted by the security risks posed by the adversary adaptive to new AI inventions. Prior research shows that *adversarial examples* can be found to mislead even the state-of-the-art recognition algorithms [4, 7, 27, 29], causing them to misclassify these examples. More specifically, such adversarial examples are images derived from those correctly recognized by a DL model, for the purpose of inducing classification errors while maintaining the level of changes low so they can appear less distinguishable from the original images by humans. Indeed, a recently approach [2] alters merely 16 pixels to ensure misclassification on 32×32 images.

Attacking a strawman. On the other hand, such adversarial learning risks need to be put into perspective. Still we are less clear how *realistic* the discovered threats could be, given that most of them are reliant on a *white-box* assumption about the *target* (the FR system they aim at), that is, the availability of full information about the target's parameters. In practice, however, an industry-grade system's parameters are often commercial secret and cannot be easily acquired by unauthorized parties.

A more realistic way to understand a DL system's security properties is the *black-box* approach, in which the adversary *queries* the target, utilizes the features inferred through the queries to learn a *substitute* model and then searches the substitute for the adversarial examples that also work on the target. Such an approach is based upon *transferability* of adversarial examples across different models [13]: some examples mislabeled by one DL model are also found to be misclassified by another. To ensure a high transferability, existing approaches aggressively **query the target** to obtain adequate input-output samples for accurately simulating the target. As a prominent example, a recent black-box attack needs to interact with the target for at least 1,000 times [17].

With all the progresses being made, still a big gap exists between hypothetic attacks proposed and credible threats with practical impacts. Particularly, querying security-critical FR systems is often expensive or even infeasible in practice: e.g., an FR ATM can immediately alert a card holder to a potential fraud once an impersonation attempt fails, making further probes less likely to continue. Another problem of prior transferability studies is the simple dataset used to train their models, e.g., the MNIST database [12] includes only tens of thousands of images for recognizing handwritten digits. A real-world FR system, however, is typically trained over tens or even hundreds of millions of images for identifying millions of people. Less clear is whether what is learnt from such small-scale studies over toy examples is indeed applicable to real FR systems.

Cross-class transferability. To better understand the security guarantee of real-world FR systems, we revisited transferability in our research, assuming that the adversary *cannot* get any feedback from the target and has limited resources. In the study, we trained multiple common deep neural networks (DNN), including VGG, GoogLeNet and ResNet, and evaluated the transferability of adversarial examples across these models, under various settings (shadower substitute networks, different structures and fewer data) to simulate a resource-constrained attacker. This research sheds new light on transferability: e.g., for ResNet, the transferability from a 50-layer substitute to a 101-layer target is about 16.8%, compared with 24.6% between the 101-layer substitute and the same target, in an impersonation attack. More interesting is the significant impact of **training data sizes**: the transferability has dropped from 24.6% in an impersonation attack to 14.5% when the substitute was trained on a dataset one order of magnitude smaller than that of the target, and further to 7.1% for a training set two orders of magnitude smaller. Intuitively, substitutes learned with fewer data or a shallower model would have a looser boundary and therefore is less likely to ensure a misclassification on the target (which is better trained with more data). Overall, however, we only witnessed a very limited success on transferability, particularly when it comes to the impersonation attack, only around 20% under different settings.

To attack an industry-grade FR system without querying it, a set of high-quality substitutes need to be built to find *common* defects of the systems similar to or even better trained than the target. However, constructing such substitutes is hard, particularly with limited resource. In our research, we studied what the adversary could do to narrow this gap and enhance his odds of success. A unique observation we have is that even though the target generally has a more precise decision boundary, the substitute could still *partially* approach this boundary, at points of interest (PoI): for example, a criminal may leverage a large number of his own and his victim's photos to boost the substitute's accuracy with regard to the identification of (just) these two individuals, for the purpose of finding the right makeup to cheat an FR ATM into authenticating him as the victim. This attack, which we call *asymmetric cross-class image transfer* or *EXCIT*, is found to be completely feasible in our research, due to a new property called *nearly local linearity* (NLL) discovered in our study.

More specifically, under a well-trained DL model, the difference between a pair of images' *representations* (e.g., the Cosine distance between the vectors produced by the DL model) should be nearly linear to their similarity as seen by human eye: in other words, when these images become increasingly dissimilar, the difference between their representations grows (nearly) proportionally. This NLL property, as discovered in our research, can be approximated at a given PoI (e.g., the fugitive) or a pair of PoIs (e.g., the victim and the impersonator) during the training of a substitute model: using the victim or the attacker's additional images, we can minimize the distances between the scores they receive from the model and what are expected according to NLL. We found that such a model can effectively simulate a better-trained target's behaviors around the PoIs.

In our research, we implemented EXCIT and first evaluated it under the settings of our transferability study. We observed that the new technique vastly enhanced the effectiveness of the attacks, particularly for impersonations, from 20% to 50%, even when the adversary only used 10% of the training data and half of the layers (thus saving the training time by 5 orders of magnitude). Further we ran this approach against industry-grade systems including ColorReco, Facevis, Face++ and SenseTime (the SenseTime system trained over tens of millions of photos). Using 4 million images collected from the web (the largest scale for this type of research), EXCIT was found to significantly elevate the chance of successful cross-model attacks, from 11% to 60%, without any communication with the target before the attack.

Contributions. The contributions of the paper are outlined as follows:

• *The NLL property and understanding of transferability*. Our large-scale study reveals the impacts the training data size can have on the successful transferring of an adversarial instance from one model to another. More importantly, we discovered the nearly linear relation between input images and their representations (in terms of their differences) under an ideal model, which enables our query-free attack and might lead to better understanding of the fundamental defects in DL models.

• *New techniques for query-free attacks*. Based upon the new discovery, we designed a new attack technique that finds adversarial instances against a well-trained target model *without querying the target and using limited re-*

sources. At the center of the technique is leverage of additional victim and attacker images and the NLL property to train a substitute capable of simulating the target model around PoIs, even when the adversary only possess a small amount of training data and much less computing resources. This makes an important step toward understanding the realistic threat of adversarial learning.

• *Implementation and evaluation*. We implemented the technique and evaluated it over industry FR systems.

2 Background

2.1 Deep Learning and Face Recognition

Deep Neural Network. Deep Neural network (DNN) is a function that projects the input domain onto an output domain for classification and other purposes. Following prior research [15], a DNN for image processing can be formalized below:

$$F(x) = softmax(Z(x)) = y,$$

where x is the image serving as the input to the DNN, y is its output, typically a vector of probabilities for the image to be in different classes, and $Z(x)$ is the "logits", a function describing all DNN layers except the "softmax" layer, whose outputs are unscaled log probabilities serving as the inputs to "softmax".

During its operations, a DNN first converts its inputs into a *representation*, a high number of parameters that capture the features of the inputs, and then hands it over to the last a few layers of the network to generate the output. For image classification and FR in specific, a DNN can be further described as follows:

$$Z(x) = C \circ R(x),$$

where $R(X)$ outputs the representation of the input and $C(\cdot)$ is the classification function that produces the "logits" based upon the representation. A well-trained DNN is characterized by its capability to generate similar representations for similar inputs. This avoids the pitfall when two similar inputs actually are mapped to very different representations and as a result, are assigned into different classes. Note that in our research, the similarity between two representations is measured by the cosine distance between them.

Face recognition systems. Since the introduction of deep convolutional neural networks (CNN) [11], FR technologies have been evolving rapidly. As a prominent example, DeepFace [28] close the gap between the recognition capabilities of human beings and machines. Further, DeepID3 [25] attained a 99.53% accuracy on the LFW dataset [6] that exceeds the human performance, 99.2%. More recently, FaceNet [22] exploited a deep architecture to achieve a 99.63% accuracy on the same dataset.

More generically in the image processing area, three DNN models have been extensively used. VGG-16 [24] running 16 cascaded convolution layers was reported to achieve state-of-the-art recognition results in the ImageNet Large-Scale Visual Recognition Challenge 2014 [20] (ILSVRC-2014), together with GoogLeNet [26], which involves 22 layers and an Inception architecture invented by Google for combining information from multi-views. Empowered by the pervasiveness of GPU and Batch Normalization technologies [8], ResNet-152 [5] winning the ILSVRC-2015 classification task is armed with 152 layers and capable of transferring shadow features to deep layers.

2.2 Adversarial Learning

The potential of deploying DNN to real-world systems (e.g., self-driving cars) faces the security challenges of *adversarial learning*, an attack that manipulates the inputs to a DNN to cause misclassification. This attack was first discussed by Szegedy et al. [27], who point out the existence of *adversarial examples*, i.e., perturbed input x' similar to the original input x but misclassified by the DNN into a different category. Such attacks can be targeted or not. In the latter case, the attacker seeks adversarial examples misclassified into *any* categories except the one they belong to. For instance, in FR, the adversary wants to *dodge* a face detection system by slightly changing his appearance from x to x', causing the classification result $\text{argmax}_i F(x')_i \neq \text{argmax}_i F(x)_i$. Here, the DNN outputs a vector that describes the probabilities for the input belonging to different individuals. During a targeted attack, the adversary intends to impersonate a given individual t, by seeking a makeup x' causing $\text{argmax}_i F(x')_i = t$.

Attack methods. To find adversarial examples, people need to define the similarity between two images (the inputs), x and x', based upon a distance metric. Prior research on adversarial learning uses the L_p distance, with p being 0, 2 or ∞:

$$\|x - x'\|_p = (\sum_{i=1}^{n} |x_i - x_i'|^p)^{\frac{1}{p}}.$$

Here $x_i - x_i'$ is the subtraction between the i-th pixel of two input images. Minimizing the L_0 distance, we can get x' with the smallest number of pixels differing from those on the original input x. The *Jacobian-based Saliency Map* (JSMA) [18] is an attack optimized under the L_0 distance. It iteratively picks pixels that have the most impact on the results and modifies them, until either a given threshold (an upper bound for the number of pixels) is reached or an adversarial example is found.

Minimizing the L_2 distance, we can obtain x' that has the least modification, in terms of Euclidean distance, across all pixels on x and x'. The first attempt using this distance is L-BFGS [27] that minimizes the L_2 distance under the box-constraint, i.e., $x' \in [0,1]^n$, where n is the number of pixels. It exploited the classical gradient descend method to find the optimal solution with a pre-defined learning rate lr:

$$x' = x + lr \cdot \bigtriangledown_x F(x), \quad (1)$$

Minimizing L_∞ distance, we can find x' with the smallest maximum-changes to the pixels. Under this distance, the optimization algorithm seeks a region of pixels with similar intensities to modify. An example of the prior attack is *Fast Gradient Sign Method* (FGSM) [4], which iteratively updates x' to produce an adversarial example by stepping away a small stride along with the direction of $\bigtriangledown_x F(x')$.

In our research, we use L_2 distance to measure the changes to an image for finding its adversarial examples.

Transferability. As mentioned earlier, transferability is the key to practical adversarial learning, when the adversary cannot directly access the internal parameters of the target model. Prior research [13] demonstrates that around 20% adversarial examples discovered from one of the three models (ResNet-152, VGG-16 and GoogLeNet) are also misclassified by other two models under a dodging attack. A more recent study [16] further shows that transferability can happen even across different machine learning techniques: DNN, Logistic Regression (LR), Support Vector Machine (SVM) and Nearest Neighbors (kNN). Particularly, more than 60% of adversarial examples discovered in LR or SVM were found to be still effective on the other model. When it comes to the impersonation attack, also 20% adversarial examples were reported to work across different DNN models [13]. These examples were found using an ensemble-based approach that descends along the summation of the gradients of several models.

A primary limitation of these prior studies is that they are all based upon relatively "small" datasets, such as ILSVRC-2014 including 1000 categories. Compared with the industry-grade FR systems such as Facevisa, ColorReco, which are trained to classify tens or even hundreds of thousands of identities, what has been learned from these studies can be less conclusive. Also importantly, the prior research either considers that the substitute is built upon similar or even identical datasets as the target , or at the very least, assumes that the adversary is capable of continuously querying the target to collect data (query results) for training the substitute. As discussed earlier, in many cases, these assumptions are still a far cry from reality. Our research instead looked into the transferability over a large dataset, when the adversary cannot query the target and does not have enough data to train his substitute model.

2.3 Threat Model

We consider an adversary that intends to perform a dodging attack or a impersonation attack on a target FR model that he *cannot* query. The adversary does not have access to the internal parameters of the target but has limited information about its architecture (e.g., ResNet, VGG or GoogLeNet) and its depth (e.g., about 100 layers for ResNet, though the precise number of layers is still unknown to him). All such information about a commercial system is often made available through various public sources, such as research papers (e.g., the design of Face++ was described in the paper [3]), technical reports and other online documents.

The target model studied in our research is assumed to be trained over a large amount of data, tens or even hundreds of millions of images, as those commercial FR systems are. On the other hand, the adversary does not have that level of resources, though we do assume that he can still acquire millions of images publicly available online, as we did in the study. Further, the adversary can obtain thousands of images of himself and the victim he want to impersonate, and also sufficient resources from the cloud to train the substitute model over the data. We believe that these assumptions are all realistic, as demonstrated in our research: particularly, all the computing power required for training our attack model can be purchased from Amazon at an approximate cost of 10,000 dollars. Specially, if the adversary can not obtain sufficient images of the victim from the internet, they can follow the victim and record videos to get enough images that are taken in various scenarios and from different angles.

3 Understanding Transferability across Asymmetric Models

To understand whether a target model the adversary cannot query is still subject to attacks, we need to find out the challenges in simulating the target's operations, under the limited resources and information. For this purpose, we conducted the largest study on transferability, using a dataset with 4 million images. Our research reveals the importance of training data size to a successful cross-model attack.

3.1 Settings

Our study utilized *MegaFace Challenge 2* [14], a dataset including 672K identities and their above 4 million photos, and *Caffe* [9], an open-source deep learning framework, to train FR DNN models in our experiments. All such experiments were conducted on a 8-GPUs server with each GPU armed with 12GB memory.

In our studies, we assume our target model is $F^*(\cdot)$ and it outputs a vector, $F^*(x)$, for a input image x. Our study covers both dodging attacks and the impersonation attacks. Here, we say that an adversarial example x', that is similar to x, causes a dodging attack if $\text{argmax}_i F^*(x')_i \neq \text{argmax}_i F^*(x)_i$, where $F^*(\cdot)_i$ represents the i-th element of the output vector. In the case of an impersonation attack, we have $F^*(x')_t > 0.5$, where t is the victim to be impersonated and we ensure $t \neq o$, the true owner or x.

To find adversarial examples, in our study, we chose L_2 distance as our metric to optimize the following objective function [2]:

$$\text{minimize} \quad \frac{1}{2}\|tanh(w)-x\|_2^2 + c \cdot f(tanh(w)).$$

For the dodging attack, f is defined as

$$f(x') = max(Z(x')_o - max\{Z(x')_i : i \neq o\}, -\kappa).$$

For the impersonation attack, f becomes

$$f(x') = max(max\{Z(x')_i : i \neq t\} - Z(x')_t, -\kappa).$$

Here, the adversarial example we found is $x' = \tanh(w^*)$, where w^* is the optimal solution of above function. In that function, c is a parameter that balances the importance of two components, the first component minimizing the L_2 distance between the adversarial example x' and the original image x, the second component modeling the goal of this attack, either dodging or impersonation. Also, κ is a threshold indicating when the attack goal is achieved. We set $c = 20$, $\kappa = 20$ for both the dodging attack and the impersonation attack in our experiments.

And we further improve the performance of above function by exploiting the standard ensemble-based approach. Specifically, we will use $K = 4$ substitutes and assemble them to solve the following function:

$$\text{minimize} \quad \frac{1}{2}\|tanh(w)-x\|_2^2 + c \cdot \sum_{k=1}^{K} f^k(tanh(w)). \tag{2}$$

where $f^k(\cdot)$ is the k-th substitutes.

3.2 Impacts of Structural Features

Structures. As mentioned earlier, to understand the impacts of DNN structures on transferability, we looked into the 3 most prominent structures: VGG[1], GoogLeNet [2] and ResNet[3]. In our study, we trained a target model and 4 substitutes for each structure using 600K photos from 100K identities from the MegaFace Challenge 2 dataset. Such a training set was selected to ensure that it did not overlap with those of other models except for the subjects and victims involved in the attacks. Over these models, we analyzed the transferability of the dodging and impersonation attacks, using an ensemble learning that integrates the common adversary examples found in 4 substitutes trained independently to find those causing the target to misclassify. More specifically, in the experiments, we studied both the dodging attacks, using 635 photos from 100 individuals randomly selected from our dataset, and also the impersonation attacks, based upon randomly chosen 600 photo pairs (each pair with the images of two different individuals). These images were inside the training sets of both the target and substitutes.

The results are presented in Table 1. As we can see from the table, for dodging, we observed a transferability about 95%, and for impersonation, it became 20%. This finding is pretty much in line with what is reported in the prior research, indicating that transferring adversarial examples across different models are feasible, though less effective in the case of impersonation.

Depths. Further we looked into the impacts of depths on transferability. For this purpose, we utilized ResNet, since the depth of its structure can be easily adjusted. More specifically, we built 4 ResNets with 50, 65, 80 and 101 cascaded convolutional layers respectively. The compositions of their structures are also presented in Table 1. Using these structures, we also trained models on different 100K identities' photos (about 600K photos for each model).

In this study, we ran those models as substitutes to attack the target (ResNet-101), both dodging and impersonation. As expected, the complexity of the network (its depth) indeed affects transferability: more layers make the DNN more capable and enhance transferability. Again, transferability tends to be low for the impersonation attack, around 16% when using ResNet-50 to attack ResNet-101.

3.3 Impacts of Data Size

Training size and transferability. An important observation is that a real-world adversary typically cannot get as many photos as a large organization uses to train its industry-grade FR system. An important question we were asking is what impacts a relatively smaller dataset could have on the chance of a successful cross-model attack. For this purpose, we trained VGG, GoogLeNet and ResNet-101 models on three datasets, with 10K identities (60K photos), 100K identities (600K photos) and 300K identities (1.5M photos) respectively. Again, all these individuals were randomly drawn from our dataset and we made sure that there was no overlap across substitutes' training set and target model's training set. In the experiment, we utilized the substitutes of the same structure to attack (dodging and impersonating) the one

[1] https://github.com/davidgengenbach/vgg-caffe

[2] https://github.com/BVLC/caffe/tree/master/models/bvlc-googlenet

[3] https://github.com/KaimingHe/deep-residual-networks

Table 1: Transferability among Different Structures: Cell (i, j) corresponds to the transferability of adversarial examples that are generated from the ensemble of four model i for dodging/impersonation attack on model j.

	ResNet-101		GoogLeNet		VGG-16	
	Dodging	Impersonation	Dodging	Impersonation	Dodging	Impersonation
ResNet-50	95.1%	16.8%	-		-	
ResNet-65	95.6%	18.0%	-		-	
ResNet-80	96.3%	23.2%	-		-	
ResNet-101	98%	24.6%	96.7%	20.2%	96.4%	20.3%
GoogLeNet	95%	16.5%	97.8%	23.1%	94.6%	18.5%
VGG-16	93.4%	17%	94.4%	18.1%	97.2%	22.3%

Table 2: Structures of Different Depths: All ResNet layers can be divided into 5 stages, starting with a convolutional layer, followed by four stages, each including a different number of "bottleneck" blocks. Different stage has different output size.

	ResNet-50	ResNet-65	ResNet-80	ResNet-101
Stage 1	1 Conv	1 Conv	1 Conv	1 Conv
Stage 2	3 Blocks	3 Blocks	3 Blocks	3 Blocks
Stage 3	4 Blocks	4 Blocks	4 Blocks	4 Blocks
Stage 4	5 Blocks	10 Blocks	15 Blocks	22 Blocks
Stage 5	3 Blocks	3 Blocks	3 Blocks	3 Blocks

built upon 300K identities' photos, The results are presented in Table 3. As we can see here, training data size turns out to significantly affect transferability and such an influence is also consistent across different structures. Compared with the structural impacts, transferability became lower when we reduced the data size from 300K to 100K (1.5M to 600K images). Compared with the impact of depth, the attack was less likely to succeed when we downsized the data (300K to 100K) than when we removed layers (from 101 to 50).

Further we ran the models with 10K identities to exploit those with 100K, as illustrated in the right column of Table 3. An interesting observation is that the transferability from 10K models to 100K models actually is lower (e.g., 14.5%) than that for 100K to 300K (e.g., 17.5%), even though the difference in the training data is even larger in the latter case (200K) than the former (90K). Intuitively, with the increase in training data, a substitute becomes closer to a perfect model and adding more data then can be less effective in improving the model's precision than the time when the model only learns from a small set of data and therefore much less accurate. Further analysis of the observation leads to the conclusion that the enhancement of transferability will slow down when the data size goes up (see Section 4.2).

Discussion. Our study shows that although both structural features and training data size affect transferability, apparently the impact of the latter is more prominent. In practice, the structural information of many commercial FR systems can often be found, from research articles, public paper and other sources. On the other hand, a deeper network with more data certainly need more computing resources to train. For example, on our system with 8 GPUs, training a ResNet-101 model took 9 hours for a data set of 60K images, while only half of the time was needed for a 50-layer model over the same images. Most importantly, collecting a large number of high-quality images is often a challenge for the adversary: for example, SenseTime Ltd's model is reported to be built from above 20M images and the dataset of this scale could *not* be found on the Internet, up to our knowledge. Therefore, we believe that whether transferability could be enhanced in the presence of a relatively small set of training data is critical question for assessing the practical impacts of adversarial learning on FR systems.

Also to attack a real-world system without querying it, the adversary needs to estimate his chance of success based upon the features of a given adversarial example, for example, the percentage of the pixels modified. This can *not* be easily done since the adversary does not have access to the target system and therefore cannot figure out the probability of success by testing his adversarial examples on the target. However, our study described above shows that the transferability between the substitute and the target can actully be gagued using the transferability between a model learnt from a smaller dataset and the substitute. This is because the probability of success in the latter case is expected to be higher than that in the former. A more formal analysis of this observation is presented in Section 4.2.

4 Query-Free Asymmetric Attack

To enhance transferability, ideally we need to make the substitute very similar to or even more accurate than the target model. Although this sounds like a mission impossible for most real-world adversaries, given their limited resources, particularly a much smaller training set they are able to acquire, still something can be done to narrow the gap between the two models. A key observation is that a unique resource the adversary often has is abundant photos of the subject (often himself) in a dodging attack and also those of the victim in an impersonation

Table 3: Transferability among Different Dataset Sizes: Cell (i, j) corresponds to the transferability of adversarial examples that are generated from the ensemble of four model i for dodging/impersonation attack on model j.

		300K		100K	
		Dodging	Impersonation	Dodging	Impersonation
ResNet-101	300K	98.8%	25.2%	-	
	100K	81.3%	17.5%	-	
	10K	34.5%	7.1%	71.2%	14.5%
GoogLeNet	300K	97.3%	24.1%	-	
	100K	79%	16.2%	-	
	10K	32%	5.1%	69.4%	13.3%
VGG-16	300K	97.5%	23.9%	-	
	100K	77%	16.3%	-	
	10K	32.3%	6.3%	66.7%	12.9%

attack. Leveraging such images, we could train a model *biased* toward the subject or the subject and victim pair. Even though such a model can be overfit and therefore its overall accuracy could be way below that of the target model, all we care about here is just the target's behavior around the subject and/or the victim, which we could potentially simulate in the substitute using the resource (extra photos).

However, effective use of such resource turns out to be challenging. Table 4 shows the experimental results when we directly duplicate those photos of subjects and the victims to 600K photos and add them to the original 600K (photos) training data for augmenting the transferability under the VGG, GoogLeNet and ResNet models in attacking the targets trained over 1.5M photos of 300K identities. From the table, we do not see a significant improvement in the effectiveness of the attack, compared with those without such data augmentation.

Table 4: Results for Naively Augmented Data: the original transferability is in the bracket.

	Dodging	Impersonation
ResNet-101	83%(81.3%)	18.2%(17.5%)
GoogLeNet	80.2%(79%)	15.8%(16.2%)
VGG-16	77.3%(77%)	16.8%(16.3%)

Intuitively, the subject and victim's photos alone are insufficient for simulating the *relations* (with regard to facial features) established by a better trained model between them and between the subject and other identities in the dataset. Such relations need to be built upon other images between the pairs (e.g., those more similar to the subject than the victim) and the way a well-trained DNN scores them according to their features. Following we show how such relations can be modeled using a new property discovered in our research, called *Nearly Local Linearity* (NLL), the key technique behind our EXCIT attack, which helps augment the substitute model for simulating the target's behaviors around the subject and/or the victim, boosting the transferability from below 30% (for impersonation) to above 60% on commercial systems (Section 5.3).

4.1 Nearly Local Linearity

Essentially, an adversarial example exploits an imperfectly trained model, inducing its *nonlinear* behavior, that is, a small perception change causing a big representation change.

Observation. A key observation we have is that the representations produced by an ideal DNN FR should accurately model the human perceptions: when two images look very different to the human, the distance between the representations should be large, and when the images appear to be similar, the distance should become small. So our idea to enhance the substitute model, with regard to a given subject or a subject-victim pair, is to approximate such nearly linear relations in the substitute, between the subject and the victim or the subject and other identities, for simulating the behaviors of a better trained model around these data points. To this end, right metrics need to be chosen to measure the human perception and the distance between the representations. In our research, we found that L_2 (as used in the prior research [2]) and Cosine distances can serve these purposes.

In our research, we trained three ResNet-101 models on three datasets with 10K identities (60K photos), 100K identities (600K photos) and 600K identities (4M photos) respectively. From each dataset, we selected, uniformly at random, 10K photo pairs, with the photos from two different identities in one pair. Then between each pair (a,b), we synthesized a series of 99 images by equidistant interpolation. Formally, the k-th image can be represented as $x^k = a + \frac{k}{100}(b-a)$. Specially, $x^0 = a$ and $x^{100} = b$.

Then, we ran all three models on these interpolated images to get their representations. Altogether, 10^6 representations were produced from the 10K image pairs. Further we calculated the mean for the Cosine distances between the representation of each image x^k, denoted by $R(x^k)$, and that of b, $R(b)$, across all their peers.

Fig 1 illustrates the relations between the L_2 distance (divided by $\|b-a\|_2$) between x^k and a and their representation's mean Cosine distance, together with ρ, which is used to measure the mean slope of each model: $\rho = 10^{-6}\sum_{k,(a,b)} k^{-1} cos(R(x^k), R(b))$.

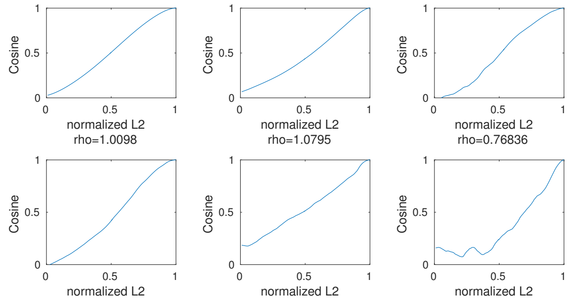

Figure 1: NLL on ResNet-101: There are 3 columns, from left to right, representing 600K, 100K, 10K model respectively. The first row demonstrates the average results, and the second row demonstrates the results of a randomly selected image pair.

As we can see from the figures, the relation between the L_2 distance and the Cosine distance approaches linear with the increase of the training data size. Particularly, it becomes almost linear for the ResNet model trained over 600K individuals (4 million images), with $\rho = 1.01$. Under the identical experimental settings, we observed the same L_2 and Cosine distance relation in the VGG and GoogLeNet models, as illustrated in Fig 2. This indicates that the relations between the subject and interpolated images, and between the subject and the other victim can be captured by this NLL property.

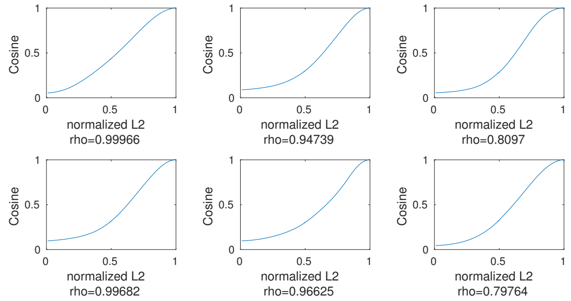

Figure 2: NLL property: There are 3 columns, from left to right, representing 600K, 100K, 10K model respectively. The first row demonstrates results of VGG-16 models, and the second row demonstrates the results of GoogLeNet models.

Concept. Formally, we define the NLL property as follows:

$$\frac{|R(b)\cdot R(x^{\lambda})|}{\|R(b)\|_2\|R(x^{\lambda})\|_2} \approx \lambda = 1 - \frac{\|b-x^{\lambda}\|_2}{\|b-a\|_2} \qquad (3)$$

where

$$x^{\lambda} = a + \lambda(b-a), \lambda \in [0,1]$$

4.2 The EXCIT Attack

The discovery of the Nearly Local Linearity property in the ideal DNN FR model enables us to train a substitute to not only leverage the extra resource at the adversary's disposal but also integrate such resource into the model by approximating the relations between these additional photos and the existing images in the dataset. For this purpose, we need to synthesize a set of "transitional" images as those interpolated photos mentioned earlier, and redefine the optimization goals when training the substitute to connect these photos to others in an expected way. These are the key steps for our EXCIT attack, as elaborated below.

Subject-oriented data augmentation. To synthesize "transitional" images and enrich the training dataset, we designed a subject-oriented data augment algorithm (see Algorithm 1). Our approach takes the original dataset $\mathscr{D}$, a subject-victim (identity) pair (o,t) and the number of images to synthesize n as its input and outputs the augmented dataset $\mathscr{D}_{aug}$. For a dodging attack, o and t will be set to the same input so the algorithm will look for transitional images between all o's images $\mathscr{A}$ and randomly-selected other images $\mathscr{B}$ in the dataset. Given the constraint of the total number of synthetic images n and the requirement that for each image pair, at least 10 synthetic photos need to be injected, we have $|\mathscr{B}| \leq \frac{n}{10}/|\mathscr{A}|$. For each photo pair (a,b) from $\mathscr{A}\times\mathscr{B}$, our approach randomly ($\lambda \sim U(0,1)$) interpolate $\frac{n}{|\mathscr{A}||\mathscr{B}|}$ transitional images.

Training NLL-augmented models. With enriched data, we can train a substitute to approximate the NLL property for given PoIs (the subject or the subject-victim pair). In this way, the substitute is expected to be closer to or even surpass a better-trained model at the PoIs, which will elevate the transferability of the adversary example our discovered. To build such substitutes, we first train the model on the original dataset $\mathscr{D}$ using the standard softmax loss function. After convergence, we fine-tune the substitute on our augmented dataset with n synthetic images, using the following loss function to minimize the distance between the output of the image x, $F(x)$ and its *expected* value, which given a pair of identities u and v, has $1-\lambda$ as the value for the uth element on the output vector and λ as the vth element and all other values set to 0.

$$-\sum_x[(1-\lambda)logF(x)_u + \lambda logF(x)_v]$$

Here, the loss function uses the Kullback–Leibler divergence to compare the distribution of $F(x)$ (with $F(x)_u$ and $F(x)_v$ being its uth and vth elements respectively) against the output vector (with its corresponding elements being $1-\lambda$ and λ).

Algorithm 1: Subject-oriented data augmentation algorithm.

```
Input: $\mathscr{D}$, $(o,t)$, $n$
Output: $\mathscr{D}_{aug}$
$\mathscr{A} = \{a : a \in \mathscr{D}, o = argmax_i\ R(a)_i\}$;
if $o = t$ then
    $m = \frac{n}{10}/|\mathscr{A}|$;
    Select a subset $\mathscr{B}$ from $\mathscr{D}$, s.t., $\mathscr{B} \cap \mathscr{A} = \emptyset$ and
      $|\mathscr{B}| = m$;
end
else
    $\mathscr{B} = \{b : b \in \mathscr{D}, t = argmax_i\ R(b)_i\}$;
end
$\mathscr{D}_{aug} = \mathscr{D}$;
for $a$ in $\mathscr{A}$ do
    for $b$ in $\mathscr{B}$ do
        for $k = 1$ to $\frac{n}{|\mathscr{A}||\mathscr{B}|}$ do
            Sample $\lambda$ from $U(0,1)$;
            $c = (1-\lambda)a + \lambda b$;
            $\mathscr{D}_{aug} = \mathscr{D}_{aug} \cup \{c\}$;
        end
    end
end
```

Finding adversarial examples. After generating multiple substitutes enhanced by NLL, we need to effectively assemble them to find common adversarial examples. This can be done directly, using the standard ensemble-based approach (Eq 2). However, a question is how to effectively use our substitutes trained on NLL augmented dataset. The standard approach is to average the gradients discovered from individual substitute model. However, we found a way to outperform the standard approach by computing a *weighted* average and optimize the following function:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|tanh(w) - x\|_2^2 \\ & + c \cdot \sum_{k=1}^{K} f^k(tanh(w))(1 - cos(R_k(tanh(w)), x)). \end{aligned} \tag{4}$$

Specifically, since each substitute we use already approximates the NLL property around the victim, we can determine the "quality" of the gradient it provides based upon the Cosine distance between its representation of the updated image (the intermediate result for finding the adversarial example) and that of the victim's image: the smaller it is, the more similar these images would be based upon their L_2 distance. Therefore, we set higher weights in favor of the gradients from the substitutes producing smaller representation differences. In the meantime, we discard the gradient values in the dimension where different substitutes cannot agree on the direction.

This weighted search algorithm is illustrated in Algorithm 2. In each step, it modifies the intermediate result x' to move along a certain direction (20-th to 21-th line), as indicated by the sum of weighted average derivative calculated from the substitutes (9-th line). Also, it drops the average on the dimension where the substitutes do not agree on the direction. More specifically, the algorithm sets the dimensions whose average derivative is small while it deviation is large (15-th to 19 line). For this purpose, we normalize the deviation of every dimension by the max value in this dimension (13-th line) and the average by the max average value (16-th line). This allows us to choose one uniform threshold for every dimension, which is set to 0.3. The algorithm for the dodging attack is similar.

Algorithm 2: Modified ensemble-based algorithm for impersonation attacks.

```
   Input: $\{R^k(\cdot)\}$, $\{f^k(\cdot)\}$, $a$, $b$, $c$, $K$, $lr$
   Output: $x'$
1  $d = Dim(x')$;
2  $x' = a$;
3  $w = tanh^{-1}(a)$;
4  while Not converge do
5      for k = 1 to K do
6          $\alpha_k = 1 - cos(R^k(x'), R^k(b))$;
7          $s_k = \alpha_k \bigtriangledown_w f^k(x')$;
8      end
9      $g = \sum_i s_i / \sum_j \alpha_j$;
10     for j = 1 to d do
11         $h_j = \{s_{1j}, s_{2j}, ...\}$;
12         $p_j = mean(h_j)$;
13         $q_j = std(h_j)/max(h_j)$;
14     end
15     for j = 1 to d do
16         if $\frac{p_j}{q_j}/max(\{p_1, p_2, ...\}) \leq 0.3$ then
17             $g_j = 0$;
18         end
19     end
20     $w = w - lr(c \cdot g + \bigtriangledown_w \frac{1}{2}\|x' - x\|_2^2)$;
21     $x' = tanh(w)$;
22 end
```

4.3 Analysis

To find out how EXCIT enhances the transferability in a query-free, black-box attack, we analyzed our implementation using a ResNet-101 model trained on 1.5M images of 300K identities as the target, and a set of ResNet-101 models trained on 600K images of 100K identities (no overlap with the target's training set except attack subjects and victims) as substitutes. The latter were further augmented with the NLL property under different attack settings, using the similar subject and victim data for the data-size study reported in Section 3.2. Specifically, in the dodging attacks, the same set of 100 identities and their 635 images were utilized. For each identity, 4 substitutes were trained over about 300K transitional images automatically synthesized. In the impersonation attacks, we selected 600 photo pairs from 10 identity pairs (subject-victim). For all identities involved in impersonation attacks, we ensure that each of them have at least 100 photos in our data set. And for a certain identity pair, each photo pair (not only those selected photo pairs) between them were augmented with 30 transitional photos for training the 4 substitutes. Under both attack settings, the target was attacked with the common adversarial examples for all 4 substitutes. The results were compared with our findings in the data-size study (Section 3.3).

Transferability. As we can see from Fig 3 and Fig 4, in both attacks, EXCIT improved the transferability, which were evident for dodging (from 81.2% to 89.6%) and dramatic for impersonation (from 17.5% to above 49%). Interesting here is that for the dodging attack, our approach even close to the attack using the substitutes trained on the same data size of the target, indicating that our EXCIT model was actually effective in recognizing the subject. This is further supported by the findings for the impersonation attack, in which none of the substitute models without the NLL enhancement could come even close to our performance, even for those as well-trained as the target. Actually, even for a ResNet-101 trained on 1.5M images of 300K identities, we found that our substitutes got a transferability of 49%.

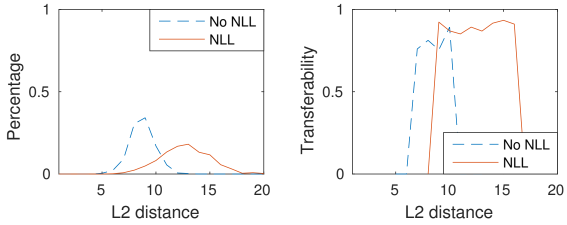

Figure 3: Dodging performance: The left figure shows the distribution of modifications made by approaches with and without NLL enhance. The right figure shows the transferabilities of them.

Further our study shows that EXCIT also works on other DNN structures: we trained a VGG model and a GoogLeNet over the 600K images to perform an impersonation attack on their corresponding targets with the same structures but trained on the 1.5M images, and found that (Table 5) both attacks achieved around 45% transferability, way above the 16% reported in our data-size study (Section 3.3).

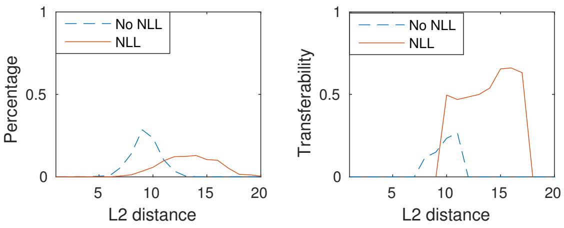

Figure 4: Impersonation performance: The left figure shows the distribution of modifications made by approaches with and without NLL enhance. The right figure shows the transferabilities of them.

Table 5: Impersonation Transferability of NLL-enhanced Approach on Different Structures.The number in the bracket is the transferability using 300K-substitutes to attack the target 300K-model.

	ResNet-101	GoogLeNet	VGG-16
with NLL	49% (71.3%)	45.8%	43.1%
without NLL	17.5% (22.7%)	16.1%	16.4%

Distance from the subject. Without querying the target, naturally the adversarial examples discovered by EXCIT, through simulating a "better" model, tend to be farther away from the subject. What we want to know, however, is for a given distance from the subject, whether the examples found by our approach still have a higher probability of success, compared with the attack without the NLL enhancement. For this purpose, we compare both methods' average transferability under various L_2 distances. The results are given in Table 6. A more detailed study, by restricting the searches within a certain distances is given in the Appendix B. From these results, we observe that NLL improves the transferability in every distance.

Table 6: Transferability under Different Distance.

	< 10	< 15	< 20
Dodging with NLL	87.7%	88.3%	89.6%
Dodging without NLL	79.5%	81.2%	81.2%
Impersonation with NLL	50.5%	47.2%	49%
Impersonation without NLL	14.5%	17.5%	17.5%

Training cost. To understand how EXCIT helps a resource-limited adversary, we evaluated the cost for training substitutes over our 8-GPU server (with the 12GB memory for each GPU). As illustrated in Table 9, it took about 200 hours to train a ResNet-101 over 1.5M images and more than 500 hours to build the model over 4M images, while constructing a substitute on 600K images used 75 hours. Note that we could fully parallelize the training of 4 substitutes, but could not do this for a target model over the same amount of computing resources, due to the communication overheads.

To further analyze the cost EXCIT could reduce, we trained 4 EXCITs of ResNet-50 and 4 EXCITs of ResNet-101 over 60K photos to perform impersonation attacks against a ResNet-101 target trained over 600K photos and 4M images. As we can see in Tabel 7, with the small amount of training data, our approach elevated the transferability of these attacks to 30-40% for the 600K target and around 20% for the 4M target, while the training time stayed at 5 to 9 hours per substitute.

Table 7: Impersonation Transferability with NLL-enhanced Approach of 60K Photos' Models. Cell (i, j) is the result of using model i to attack model j.

	600K	4M
ResNet-50 60K	38.5%	16.8%
ResNet-101 60K	44.2%	23.1%

Also we compared the efficiency of our approach against the attacks without the NLL enhancement. In the latter case, the only way to improve the transferability is to train more substitutes to find their common adversary examples. In our study, we conducted three experiments using 4, 8 and 16 substitutes respectively, each model trained over 60K images. The results of ensemble adversarial learning over these models are presented in Table 8. As we can see here, when attacking the 4M target, even with 16 substitutes, the attack could not achieve the same level of transferability as the 4 NLL-enhanced substitutes, even for the those with only 50 layers. In this case, the cost of EXCIT, in terms of training time, is no more than 16.8% of the direct attack (with 16 substitutes).

Table 8: Impersonation Transferability of Ensemble-based Models (no NLL) Trained over 60K Photos.

4 models	8 models	16 models
7.2%	12.7%	16.5%

Table 9: Cost for Training Different Models.

Depth	Images	Time	Memory
50	60K	5h	8x4.5G
101	60K	9h	8x6.5G
101	600K	75h	8x7.5G
101	4M	>500h	8x8G

5 Evaluation on Real-World Systems

We evaluated our approach, NLL-enhanced attacks, on four real world systems. Three of them are online, with APIs available for the public, and the last one is a commercial system without open access, one of the products from SenseTime Ltd. We performed both dodging and impersonation attacks against them. The details of our experiments and our findings are elaborated below.

5.1 Experimental Settings

Unlike the models built in our analysis, which were trained over the subject and victim's images and output a vector to specify whether an input image belongs to these identities, a real world FR system takes two photos as inputs and calculates a score about the similarity of the individuals in these two photos. Here is how we determined whether an adversarial example worked on these systems:

In a successful dodging attack, we expect that the target system outputs a low score ($< Th_{dod}$) for two images: one is the subject's original photo and the other is the adversarial example generated by our approach from the original photo. In a successful impersonation attack, the target system is supposed to output a high score ($> Th_{imp}$) for two images: the victim's photo and the adversarial example generated by our approach from the subject's image, indicating that they all belong to the same individual.

In our experiments, we first trained 4 ResNet-101 models on randomly sampled 3M photos from 600K identities in the MegaFace Challenge 2 dataset[4]. From all identities, we selected 10 individuals as the subjects in our dodging attack. For the impersonation, we exploited 10 subject-victim pairs. Every identity involved has at least 100 images in the dataset. In the experiments, we randomly chose 10 of each individual's images for the dodging attack and 10 photo pairs for each subject-victim pair to execute the impersonation attack. For each of these subjects or subject-victim pairs, we further enhanced the 4 models using the NLL augmentation and then ran our algorithm (Algorithm 2) to find adversarial examples to attack them.

5.2 Attack on Online APIs

The three online APIs attacked in our research are ColorReco[5], FaceVisa[6] and Face++[7]. These models were trained with a large amount of data. For example, Face++ was built upon 5M photos of 20K identities [30] and FaceVisa was upon 2M photos. Also they all demonstrated a high recognition accuracy over the Labeled Faces in the Wild (LFW) dataset [6]: 99.4% for ColorReco, 99.5% for FaceVisa and 99.5% for Face++. In our experiments, we ran a python script to automatically upload our test photo pairs to ColorReco and FaceVisa. For Face++, we had to do it manually due to the requirement of CAPTCHA solving.

The success rates of our attacks are presented in Table 10. Note that in the experiments, the thresholds

[4] We did not use all 4M for each substitute in an attempt to make these substitutes diverse.

[5] http://www.colorreco.com/faceCompare

[6] http://www.facevisa.com/web/index/demo

[7] https://www.faceplusplus.com/face-comparing/#demo

for different APIs are different and defined by the APIs themselves. Besides, we considered that an attack failed if the target FR system could not detect face from the adversarial example submitted, even for the dodging attack. As we can see from the table, our approach achieved a higher accuracy in the dodging attack, compared with the attacks without the NLL enhancement (Table 5). A much bigger boost, however, is observed for the impersonation attack, in which EXCIT raised the success rates for all three systems from around 20% to 67-82%. Fig 5 further illustrates the distributions for the scores of our submitted photo pairs.

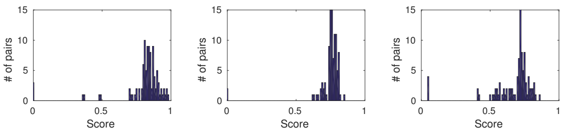

Figure 5: Distributions of scores for impersonation attacks: from left to right, they are the results of ColorReco, Facevisa and Face++ respectively.

Table 10: Success rate against online APIs.

	ColorReco	Facevisa	Face++
dodging	98%	96%	93%
Th_{dod}	0.75	0.64	0.623
impersonation	74%	82%	67%
Th_{imp}	0.80	0.74	0.691

5.3 Attack on Industrial System

Commercial FR systems are often better trained and more capable than the free FR APIs, which are mostly used for online demo. Such industry-grade systems are typically characterized by a large number of layers, and being trained over a massive amount of data on clusters of GPUs. The services they provide are not open to the public and only available for purchase. In our research, we obtained the commercial SDKs from SenseTime Ltd. through our collaborations. SenseTime's products are known to be among the leading FR systems [25]. So the system we analyzed represents the state-of-the-art in FR technologies. It was trained over 20M photos for 1M individuals, using a ResNet-like model, though the details of the structure are commercial secrets. The model tested in our study was estimated to require at least 14,000 hours (50 epochs) to train, over our GPU server. By comparison, all 4 EXCIT substitutes used in our attack were trained for 2,500 hours in total. With less than 1/5 of the time spent on training the models, our approach achieved a high success rate for the 100 individual selected for the dodge attack and 100 pairs for the impersonation attack: in the former case, 89% of transferability was achieved, compared to 70% without the NLL enhancement, and the in the latter, we raised the transferability from 11% to 60%. Fig 6 further shows examples for the successful attacks. We have reported our findings to SenseTime and are helping them improve their system.

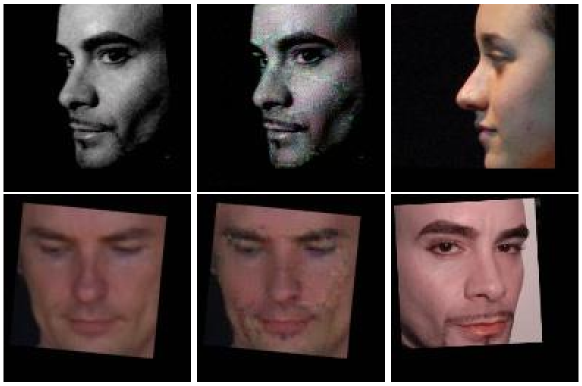

Figure 6: Successful impersonation attacks on SenseTime. The three columns are the subject photos, our generated adversarial examples and the target photos respectively. The modification of the first case is 13.98 and the second case is 7.32.

6 Related Works

Our approach utilizes a synthesized dataset to fine-tune our substitutes, for the purpose of approximating the NLL property at PoIs. Synthesized data have also been used in the prior research [17], to support completely different techniques and to different purpose. Specifically, the prior research uses synthesized data to query the target model to train the substitute (1000 times to achieve an 84.24% success rate in transferring adversarial examples). This is *exact* the attack scenario our *query-free* approach is designed to avoid. Without communicating with the target model, the only thing we can do is to build a substitute as well-trained as the target, so as to captures their common structural weaknesses to enhance transferrability. Such an attack is found to be completely feasible through simulating the target's behaviors around PoIs, based upon the NLL property we discovered and additional images collected from the victims and the attackers.

Our approach also exploited the ensemble method. The original ensemble-based approach is proposed by Liu et al. [13]. Their work focus on the transferability among DNN models with different structures, whereas the data size is the factor that really matters in face recognition domain, as have been demonstrated before. Compared with them, our method can increase the transferability from a model trained with insufficient data to a model trained with plenty of data, which goes beyond their method's capability. Another ensemble method is proposed by Sarkar et al. [21]. They trained a DL model to find "universal" perturbations fooling all their pretrained target models. Particularly, their object function combines both the sum of targets' (mis)classification loss and the scale of the finding perturbations. The difference from their method is similar with above. In our attacking scenario, their method may not work.

7 Discussion

Understanding NLL. Unlike existing cross-model attacks, EXCIT does not even interact with the target, so there is no way for our approach to exploit the specific defects of the target. The reason we can still find highly transferable examples similar to the subject is that by simulating better-trained models around PoIs, our approach is likely to discover some common (potentially structural) defects fundamental to a certainly type of DNN, and in the meantime, avoid exploring the subspace unlikely to contain adversarial examples, given the reduction of training-specific weaknesses (e.g., lack of sufficient data) around PoIs. Under an NLL-enhanced substitutes, we could even discover transferable examples with changes restricted to given facial features: e.g., around the eyes (Appendix C), which allows the prior attack [23] (using printout glasses to evade detection) to work in a query-free, black-box setting.

In the meantime, our understanding of transferability is still limited. Still less clear are the questions such as whether there exist adversarial examples inherent to certain DNN structures or even the fundamental design of artificial neural networks. Further studies on these issues are certainly important.

Our current definition of NLL describes a relation between Cosine distance and the L_2 distance. However, such a relation may not be general, particularly when it comes to non-FR problems: as an example, we found that the NLL-enhanced models still improved transferability but less significantly over Cifar10 [10], a dataset for image classification. This could be attributed to the unique features of FR: e.g., differences between two faces can be added to to another one to form a new face, which makes the "transitional" images easy to construct; also FR tasks are characterized by a large number of training categories, compared with other tasks (e.g., 1K categories for ILSVRC vs. 600K entities for Megaface), forcing the DNN models to map input images to a high-dimensional sphere with a maximum space utilization. All these features make NLL more effective for FR. What is less clear, however, is how to expand the concept to enhance the transferability of other tasks, which should be investigated in future research.

Defense. Defensive *distillation* [19] has been demonstrated to be effective against most of previous attacks. However, as pointed out by Carlini et al [2], a modified version of existing attacks will break them defense. We also found that *distillation* does not work on EXCIT either: more specifically, We implemented the defense on our 300K target with temperature $T = 20$, and ran four 100K substitutes to attack it. The result is that *distillation* can only reduce our transferability from 89.6% to 88.8%, for dodging, and from 49% to 45.6%, for impersonation.

Alternatively, we can consider to insert "secret" into the commercial system. Specifically, train the system on a custom dataset where all the photos are covered by a secret pattern. Since queries are not supposed to be made to the target during an attack, the secret added to a commercial system could help mitigate the EXCIT threat. In general, however, defense against adversarial learning is known to be hard [1]. Further research is needed to find an effective way to defeat our attack.

Cost of the attack. As mentioned earlier, training the substitutes to attack SenseTime's FR system took 2,500 hours on our server. An estimate cost for such resources is about 10000 dollars on Amazon AWS. The computing time here can be shortened through parallelization, since all 4 substitutes can be trained together. Also, the computing cost could be reduced when the adversary attempts to impersonate multiple victims or hide multiple subjects. In this case, only one set of substitutes need to be trained over our dataset, which can later be augmented with NLL for different subjects or subject-victim pairs to support dodging or impersonation attacks.

8 Conclusion

In this paper, we present our new understanding of DNN-based FR systems, in terms of their vulnerability to the transferable attack under a resource-constrained adversary. Our research shows that limited resources, particularly smaller training sets, can have a significant impact on the effectiveness of the attack. This is important since a real-world adversary typically cannot query the target frequently and needs to build substitutes as capable as, or even more powerful than the target model under his limited resources. Narrowing such a resource gap, however, turns out to be feasible through a novel technique we developed. Specifically, we found that the adversary could make an effective use of the extra information (images) about subjects and victims in his possession, by approximating the relations of these PoIs with other images in the training set characterized by a Near local linearity property we discovered. As a result, we can grossly elevate the transferability in both a dodging and an impersonation attack by training NLL-enhanced models by nearly 50% in attacking industry-grade systems. With our new techniques and findings, still more effort needs to be made to better understand transferability and mitigate the threat it poses.

References

[1] CARLINI, N., AND WAGNER, D. Adversarial examples are not easily detected: Bypassing ten detection methods. *arXiv preprint arXiv:1705.07263* (2017).

[2] CARLINI, N., AND WAGNER, D. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on* (2017), IEEE, pp. 39–57.

[3] FAN, H., CAO, Z., JIANG, Y., YIN, Q., AND DOUDOU, C. Learning deep face representation. *arXiv preprint arXiv:1403.2802* (2014).

[4] GOODFELLOW, I. J., SHLENS, J., AND SZEGEDY, C. Explaining and harnessing adversarial examples. *stat 1050* (2015), 20.

[5] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[6] HUANG, G. B., RAMESH, M., BERG, T., AND LEARNED-MILLER, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[7] HUANG, S., PAPERNOT, N., GOODFELLOW, I., DUAN, Y., AND ABBEEL, P. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284* (2017).

[8] IOFFE, S., AND SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* (2015), pp. 448–456.

[9] JIA, Y., SHELHAMER, E., DONAHUE, J., KARAYEV, S., LONG, J., GIRSHICK, R., GUADARRAMA, S., AND DARRELL, T. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia* (2014), ACM, pp. 675–678.

[10] KRIZHEVSKY, A., AND HINTON, G. Learning multiple layers of features from tiny images.

[11] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105.

[12] LECUN, Y., BOTTOU, L., BENGIO, Y., AND HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[13] LIU, Y., CHEN, X., LIU, C., AND SONG, D. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770* (2016).

[14] NECH, A., AND KEMELMACHER-SHLIZERMAN, I. Level playing field for million scale face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).

[15] PAPERNOT, N., AND MCDANIEL, P. On the effectiveness of defensive distillation. *arXiv preprint arXiv:1607.05113* (2016).

[16] PAPERNOT, N., MCDANIEL, P., AND GOODFELLOW, I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277* (2016).

[17] PAPERNOT, N., MCDANIEL, P., GOODFELLOW, I., JHA, S., CELIK, Z. B., AND SWAMI, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (2017), ACM, pp. 506–519.

[18] PAPERNOT, N., MCDANIEL, P., JHA, S., FREDRIKSON, M., CELIK, Z. B., AND SWAMI, A. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on* (2016), IEEE, pp. 372–387.

[19] PAPERNOT, N., MCDANIEL, P., WU, X., JHA, S., AND SWAMI, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on* (2016), IEEE, pp. 582–597.

[20] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M., BERG, A. C., AND FEI-FEI, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV) 115*, 3 (2015), 211–252.

[21] SARKAR, S., BANSAL, A., MAHBUB, U., AND CHELLAPPA, R. Upset and angri: Breaking high performance image classifiers. *arXiv preprint arXiv:1707.01159* (2017).

[22] SCHROFF, F., KALENICHENKO, D., AND PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 815–823.

[23] SHARIF, M., BHAGAVATULA, S., BAUER, L., AND REITER, M. K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016), ACM, pp. 1528–1540.

[24] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[25] SUN, Y., LIANG, D., WANG, X., AND TANG, X. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873* (2015).

[26] SZEGEDY, C., LIU, W., JIA, Y., SERMANET, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V., AND RABINOVICH, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1–9.

[27] SZEGEDY, C., ZAREMBA, W., SUTSKEVER, I., BRUNA, J., ERHAN, D., GOODFELLOW, I., AND FERGUS, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[28] TAIGMAN, Y., YANG, M., RANZATO, M., AND WOLF, L. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), pp. 1701–1708.

[29] XU, W., EVANS, D., AND QI, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017).

[30] ZHOU, E., CAO, Z., AND YIN, Q. Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv preprint arXiv:1501.04690* (2015).

A Transferability prediction

Given an adversarial example discovered, the attacker needs to have some idea how likely the example could also mislead the target model. Also in the presence of multiple examples, the most promising one would be given preference. One way to estimate the transferability of an example is to train multiple models as capable as the target, called *target simulators* or simply *simulators*, run substitutes to attack them and then collect the statistics about the relation between the features of the adversarial examples discovered in substitutes and the transferability of the examples. Given such a relation, the attacker can look at the features of an example to estimate the likelihood that it could fool the target. This simple approach, however, does not work in practice, as we do *not* have the resources (e.g., a large number of images) to build such powerful simulators.

Therefore in our research, we took a different path including three steps: firstly, we train simulators on the data we have; secondly, we estimate the difference between simulators and the target, thirdly, we plus the estimated difference to our simulators' outputs to predict the target's output. In this process, estimation the difference is challenge, cause we don't know how the target looks like. But we can know the scale of the training set of the target model. Thus we built a function $g(m_\alpha, m_\beta)$ to estimate the difference between models trained on m_α photos and m_β photos. The detailed description of $g(\cdot)$ is given in Appendix A. Specifically, our study shows that when training data grows, the substitute becomes similar to the target, and the impact of their data size difference becomes less prominent. Next, plussing the average difference to every simulator's output, we derive what the target model would output for adversarial examples and can choose the best one with the largest likelihood that it will fool the target.

To build $g(m_\alpha, m_\beta)$ measuring the difference between two models trained on m_α photos and m_β photos, we need to find a "bridge" to connect them. A nature idea is leveraging their loss that measures how far way they are from the perfect model and further implementing the triangle inequality to estimate the difference between themselves. While the classic *softmax* loss is inappropriate here, cause it not satisfies the triangle inequality. Thus we used the Cosine distance again. We define the Cosine distance loss of a model $R(\cdot)$ as following:

$$L_{cos}(R) = \sum_{a,b} l_{cos}(a,b,R)$$

$$\text{where} \quad l_{cos}(a,b,R) = \begin{cases} 1 - |cos(R(a),R(b))|, \text{same identity;} \\ |cos(R(a),R(b))|, \text{different identity.} \end{cases} .$$

And we count the mean and the standard deviation of Cosine distance loss for models trained on different data size. The results are showed on Fig 7. Now, we can infer the mean (μ_β) and the standard deviation (δ_β) of the target model trained on m_β photos, according to the fitted curve. Further, we assume the target model's and simulators' losses obey the normal distribution $\mathcal{N}(\mu_\beta, \delta_\beta^2)$ and $\mathcal{N}(\mu_\alpha, \delta_\alpha^2)$ respectively, and roughly estimate the difference by assuming it also obey the normal distribu-

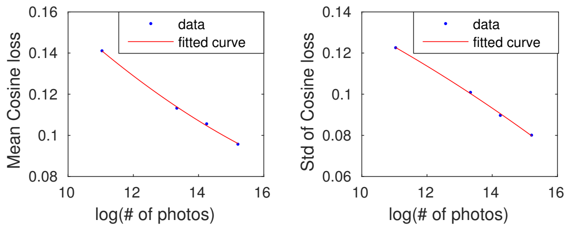

Figure 7: Cosines distance loss of models trained on different size of data.

tion $\mathcal{N}(\mu_\beta - \mu_\alpha, \delta_\beta^2 + \delta_\alpha^2)$. Thus, for a certain simulator $R(\cdot)$, we can calculate out what is the likelihood that $1 - |cos(R(a), R(b))|$ plus the estimated difference will surpass 0.5 for dodging attack or lower than 0.5 for impersonation attack.

Table 11: Statistics of $L_{cos}(R_{m_\alpha}) - L_{cos}(R_{m_\beta})$.

m_α	m_β	μ	δ
62258	1541811	0.0188	0.1176
62258	4019407	0.0371	0.1225
1541811	4019407	0.0163	0.1001

Besides, in Table 11, we list true values of the distance of some pairs of (m_α, m_β), and observe that, along with the increase of data size, the simulator becomes similar to the target, and the impact of their data size difference becomes less prominent (small μ and δ).

B Performance in Difference Distance

Without querying the target, naturally the adversarial examples discovered by EXCIT, through simulating a "better" model, tend to be farther away from the subject. What we want to know, however, is for a given distance from the subject, whether the examples found by our approach still have a higher probability of success, compared with the attack without the NLL enhancement. For this purpose, we need to modify the objective function of the DNN to limit its search within a given distance (in terms of $L-2$ in our research) constraint, as follows:

$$\text{minimize} \quad \exp(\|tanh(w) - x\|_2 - \gamma) + f(w).$$

Here we use an *exp* function that penalizes the L_2 distance when exceeding γ. More specifically, we calculated its derivative as follows:

$$\exp(L_2 - \gamma) \cdot \frac{\tanh(w) - x}{\|\tanh(w) - x\|_2} \bigtriangledown_w \tanh(w) + 1 \cdot \bigtriangledown_w f(w)$$

where L_2 represents $\|\tanh(w) - x\|_2$. As we can see here, When $L_2 > \gamma$, the component involving the *exp* function grows quickly, moving the objective function away from the optimality. Therefore, in the optimal situation, L_2 should not exceed γ much.

In our research, we evaluated our approach using the objective function when $\gamma = 5$, $\gamma = 20$ and $\gamma = 30$, and exploiting 4 EXCITs trained on 600K photos of 100K identities to attack the target trained on 1.5M photos of 300K identities. The results are presented in Fig 8 and Table 12. As we can see, under various distances, the adversarial examples found by our approach are always much more transferable than the one without the NLL enhancement. In the meantime, our approach tends to pick up the examples away from the subject, given the fact that the NLL property moves the decision boundary of the substitute (with regard to the subject and the victim) closer to the ideal one, making it harder to find the adversarial examples close to the subject's image, though once such an image is found, it is more likely to lead to a successful attack.

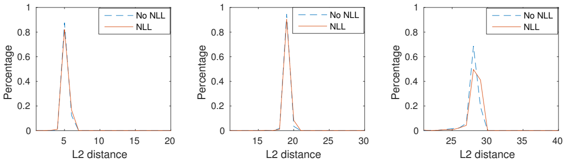

Figure 8: Distributions of modifications under different γ.

Table 12: Impersonation Transferability of NLL-enhanced Approach under Different Distances.

$\gamma =$	5	20	30
with NLL	6.3%	51.3%	74%
without NLL	3.8%	21.8%	49.5%

C Restrict the modification to certain region

The trivial method to restrict modifications within a certain region is to quench those derivatives out of the region, while finding the adversarial examples. However, in this setting, finding adversarial examples becomes harder than before. So it is need to totally release the constrain on the magnitude of modifications. We demonstrate two examples restricting modifications around eyes on Fig 9. We observe that the modifications become severe: the L_2 distance between generated adversarial example and the original photo of the first case is 24.63, and of the second case is 27.04.

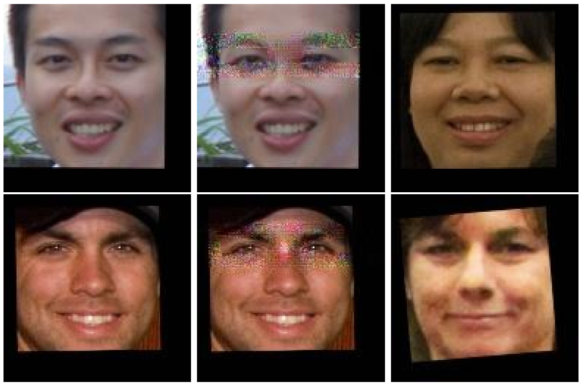

Figure 9: Successful impersonation attacks within restricted region.