Examining of Shallow Autoencoder on Black-box Attack against Face Recognition

Vo Ngoc Khoi Nguyen* Takamichi Terada* Masakatsu Nishigaki† Tetsushi Ohki†
Shizuoka University, Shizuoka, Japan
* E-mail: nguyen,terada@sec.inf.shizuoka.ac.jp † E-mail: nisigaki,ohki@inf.shizuoka.ac.jp

***Abstract*—In this paper, we propose a Black-box Adversarial Examples (A.E.) attack that is effective for face recognition. Black-box A.E. for face recognition had multiple problems such as low probability of successful attack, limited attack targets, or large computational complexity which lead to impracticality in many real world scenarios. Therefore, we propose a more effective method of attacking face recognition system using Black-box A.E. by creating an attack substitute model suitable for face recognition based on the A.E. generation method of Huang et al.. For evaluation, this method and the public dataset are used to attack arbitrary and specific people registered in the face recognition system which points out the possibility of a Black-box Adversarial Attack against face recognition system.**

I. Introduction

Biometric authentication technology is widely used in academic research and daily life, including use in mobile terminals. In particular, face recognition systems (FRS) have made great strides with developing machine learning technology over the past few years and have achieved many significant results such as improved identification accuracy to a stage comparable to humans[1]. On the other hand, many Adversarial Examples (A.E.) methods have been proposed as attacks on machine learning algorithms. A.E. is crafted to fool the classifier model trained with machine learning algorithms by adding small perturbation to the input which cannot be perceived by human eyes.

Consider that the threat caused by A.E. also exists within the machine learning-based face recognition, studying the attack conducted by A.E. in terms of face recognition and its countermeasures is necessary to build a safer and more robust FRS. Attacks against FRS are divided into white-box attacks and black-box attacks based on whether the system's parameters are known or not to the attacker. Unlike white-box attacks, black-box attacks can be used in most of real world scenarios such as attacking a Machine Learning as a service (MLaaS)-based FRS. Therefore, this paper focuses more on A.E. based black-box attacks and its countermeasures.

There are several approaches to generate A.E. for Black-box attack against machine learning algorithms. For instance, the transfer-based attack methods first pre-train a local model and then generate A.E. using a white-box attack on said local model to attack a completely unknown target system[2]. Furthermore, score and label-based attack methods use the target system's output to calculate a loss function and use the result to approximate the target system's gradient through multiple queries[3][4].

The problem of transfer-based attack is that it can not achieve a high attack success rate, and it shows the limited effect on targeting arbitrary labels. On the other hand, score and label-based attacks achieve extremely high attack success rates but they may require many queries. To tackle these problems, Huang et al. (2020). [5] has proposed TREMBA (TRansferable EMbedding based Black-box Attack), which combines the approaches of transfer-based and label-based attack. As a result, Huang et al. came up with an A.E. generation method that can be applied to many models with a small number of queries while maintaining a high attack success rate.

However, many A.E. researches, including TREMBA, do not target face recognition and experiment is limited to virtual recognition. In many cases of virtual recognition, the main purpose is to identify the class to which the object belongs from the information of overall object's shape. It does not target the identification of parts or individuals in the class. On the other hand, since the face recognition system mainly targets the identification of individuals, it focuses not only on the face shape but also on the parts common to all face shapes such as the nose, eyes, and mouth, as well as considering the differences between these individuals. Therefore it is safe to assume that FRS performs differently than the virtual recognition system.

To address that problem, in this research, we focus on the learning process of the embedded network of Autoencoder and propose an attack method using local models with different depths by using the features of the face recognition system that only used face images as input. Note that we only use TREMBA as an example because this is the current State-of-the-art black-box attack method and we believe our method could apply to other attack using local models. We also conducted attack experiments using multiple Autoencoders with different depth and embedded vector dimensions. We showed that the proposed method could improve the attack success rate and the number of queries.

II. Related work

A. Black-box attack

Research on A.E. is being actively conducted and has a wide range of approaches. However, this section only reviews typical researches on Black-box attacks.

1) Transfer-based attack: Papernot et al. showed that A.E. has the same transferability as a typical machine learning model and that A.E. generated by one model can be utilized to attack different models[2]. Transfer-based attack methods are black-box attacks that utilize the transferability of A.E.. An attacker attacks a local model which is accessible creates an A.E., and transfers the created A.E. to an unknown target network. This method has the advantage that it is not necessary to inquire about the target model. However, this method has been found to achieve a low attack success rate and much lower on attacks aiming at particular classes (persons), such as impersonation attacks on FRS.

2) Score-based attack: Score-based attacks are attacks conducted on the assumption that the attacker can access the score that is the output of the target system. In many cases, the score is a value expressed in the form of confidence or confidence probability. Many score-based attacks approximate the correct gradient from the score obtained using the sampling method. Chen et al. proposed AutoZoom, which is one of the typical methods of score-based attacks[3]. AutoZoom degenerated sampling space and succeeded in reducing the number of queries to the target network required for the attack using autoencoder and bilinear transformation. In addition, Ilyas et al. succeeded in further reducing the number of queries to the network and attack failure rate by incorporated data and time prior [6]. Moon et al. showed that it is possible to create an effective A.E. with a smaller number of queries by utilizing combinatorial optimization without using the gradient approximation method at all[7].

3) Label-based attack: Label-based attacks are attacks conducted on the assumption that only the label that is the output of the target system can be accessed and is the strictest assumption in terms of the Black-box attack. On the other hand, in face recognition and identification services provided by MLaaS, it is obvious that only the identity result is of interest to end-users. Therefore, it can be assumed that it is the most commonly used setting in these services. Brendel et al. succeeded in freely manipulating output labels of the Google Cloud Vision API by generating A.E. using the two prior knowledge, data bias and gradient of a local model[8]. TREMBA is a method proposed by Huang et al. that succeeded in conducting an efficient Black-box attack on the attack target network by combining a transfer-based and label-based attack method [5]. TREMBA used autoencoder to generate perturbations that take into account the characteristics of the attack target. The Black-box attack performed by using these perturbations has achieved a higher attack success rate and improve the number of queries required.

B. Existing problems

This study aims to carry out label-based attacks assuming that most commercially available authentication models produce only labels. Brendel et al. succeeded in generating an effective A.E. for the face recognition model. However, the attack's success rate is low and requires a huge number of queries. TREMBA, which combines transfer-based and label-based attack methods, outperforms Brendel et al. in terms of attack success rate and can generate semantic perturbations using autoencoder as a local model. Semantic perturbation is defined as a perturbation in which the characteristics of the original image can be estimated to a certain extent even from a human perspective, and Huang et al. showed that semantic perturbation has high transferability[5].

In this paper, we will perform Black-box attacks on a machine learning model based on TREMBA focusing on different embedded network that is particularly effective when targeting a FRS. To be specific, we will conduct concealer attack and spoofing attack to evaluate proposed method's effectiveness.

III. Method

A. Background

1) Neural network: In this paper, a FRS is a neural network for m-class identification $F(\boldsymbol{x}) = y$ that accepts a n-dimension image $\boldsymbol{x} \in \mathbb{R}^n$ as input and produces an output $y \in [1, m]$. We use the notation from Papernot et al.[2]: define F to be the full neural network including the softmax function, $Z(\boldsymbol{x})$ to be the output of all layers except the softmax, and

$$F(\boldsymbol{x}) = \arg\max\left(softmax(Z(\boldsymbol{x}))\right) = y. \tag{1}$$

2) Assumption: In this attack, we assume that rather than the training dataset itself, the attacker only has access to a similar dataset which has same classes but different images. Furthermore, the attacker has no information about parameters of the target FRS and can only access to the output label of the FRS.

3) Attacker's goal: The attacker's goal is to successfully cause misidentification against the FRS. In detail, there are two types of attacks on the biometric authentication system: (1) a *concealer attack* that avoids being identified as the attacker, and (2) a *spoofing attack* that fool the FRS to identify the attack as other specific person.

Concealer attack is a type of attack in which the input $\boldsymbol{x}+\delta$ is misidentified as a label $y' \neq y$ by giving a small perturbation δ to the input $\boldsymbol{x}$.

On the other hand, spoofing attack is a type of attack aimed at causing the FRS to misidentified the input $\boldsymbol{x}+\delta$ as a specific label t different from y. In other words, spoofing attack aims to find δ such that $F(\boldsymbol{x}+\delta) = t$. Here, the perturbation δ is bounded by its l_p norm: $\|\delta\|_p \leq \epsilon$ with a small $\epsilon > 0$.

B. Label-based attack based on TREMBA

In this study, we carry out a label-based attack based on TREMBA. Therefore, we first describe the A.E. generation method in TREMBA. A.E. generation in TREMBA can be divide into the following two steps.

- Step 1: Train autoencoder network to generate adversarial perturbations.
- Step 2: Find A.E. in the low-dimensional embedded latent space of autoencoder.

Algorithm 1 -Searching for A.E. on the embedded space (Partially cited and modified from Algorithm 1 of [5])

Input: Target system F_t; Input $\boldsymbol{x}$; Output y; Encoder E; Decoder D; Standard deviation σ; Learning rate η; Batch size b; Iterations T; Bound for adversarial perturbation ϵ. Sample ν_k from Gaussian distribution $\mathcal{N}(\nu_k|\boldsymbol{z}_j, \sigma^2)$.
Output: A.E.perturbation

$\boldsymbol{z}_0 = E(\boldsymbol{x})$
for $j = 1$ to T do **do**
  Generate Gaussian noise $\nu_1, \nu_2, .., \nu_b \sim \mathcal{N}(\boldsymbol{z}_{t-1}, \sigma^2)$
  $\mathcal{L}_i \leftarrow \mathcal{L}_{untarget}(x, y)$ or $\mathcal{L}_{target}(x, t)$
  $\boldsymbol{z}_j \leftarrow \boldsymbol{z}_{j-1} - \frac{\eta}{b}\sum_{k=1}^{b} \mathcal{L}_i \nabla_{\boldsymbol{z}_{j-1}} \log \mathcal{N}(\nu_i|\boldsymbol{z}_{j-1}, \sigma^2)$
end for
return $\delta = \epsilon \tanh\left(D(\boldsymbol{z}_T)\right)$

1) Generating adversarial perturbation: Let G be the Generator network created from the encoder E and decoder D which is used to generate perturbation. The encoder E takes a n dimensional vector $\boldsymbol{x} \in \mathbb{R}^n$ as input and produces a low-dimensional embedded latent vector $\boldsymbol{z} = E(\boldsymbol{x})$ as output. Note that we assume $\boldsymbol{x}$ as facial image and G is trained so that the FRS would be fooled. The decoder D takes $\boldsymbol{z}$ as input and produces perturbation $\delta = \epsilon \tanh D(\boldsymbol{z})$, which has the same dimension as $\boldsymbol{x}$, as output.

The training set is defined as $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_i$ denotes the input and y_i denotes its correct label. For concealer attack, the Generator G is trained by minimizing the hinge loss function used in the C&W attack[9]:

$$\mathcal{L}_{untarget}(x_i, y_i) = \max\Big(Z(\epsilon \tanh(G(x_i)) + x_i)_{y_i} - \max Z(\epsilon \tanh(G(x_i)) + x_i)_j, -\kappa \Big) \quad (2)$$

As for spoofing attack,

$$\mathcal{L}_{target}(x_i, t) = \max\Big(\max Z(\epsilon \tanh(G(x_i)) + x_i)_j - Z(\epsilon \tanh(G(x_i)) + x_i)_t, -\kappa \Big) \quad (3)$$

where t denotes the targeted class. By setting $\delta = \epsilon \tanh(D(\boldsymbol{z}))$, we can guarantee $\|\delta\|_p \leq \epsilon$.

2) Search over low-dimensional space: TREMBA uses NES[10] to approximate the gradient of a properly defined surrogate loss (hereinafter referred to as the local loss function $\mathcal{L}$) in order to find a valid A.E.. NES does not directly calculate the loss gradient after adding perturbation but can update perturbation δ using their updating algorithm, which updates the parameters of search distribution by following the natural gradient towards higher expected fitness.

The detailed procedure is presented in Algorithm 1.

C. Using shallow autoencoders on attack against FRS

As we stated in section II-B, most of the existing Black-box A.E. researches consider a visual recognition model as an attack target and there are some remaining problems when applying these methods to face recognition scenario. In addition, the attack method using TREMBA described in section III-B may not be optimized for spoofing attacks in face recognition even though it is an A.E. generation method with a high probability of successful attack and only requires a small number of queries. Face recognition assumes a face image as an input. Since all face images have a common structure regardless of the person such as eyes, nose and mouth, there is a possibility that the attack can be made to be more efficient by using an autoencoder suitable for feature extraction of faces. In this study, we assume that changing the depth of the autoencoder can affect the extraction of facial features that are useful for generating A.E., and create three attack networks with different depths of the autoencoder to perform attacks. In this research, these autoencoders are called S1, S2 and S3. Since the main purpose of our research is to examine the hypothesis that a simpler structure is easier to attack FRS, we do not add any major changes to the origial autoencoder but only trim of a layer in each network. In order to find an optimized network that is particularly effective for FRS, S1, S2 and S3 are created as networks with different depths and latent vector dimensions. The networks used in TREMBA, S1, S2 and A3 are shown in Fig. 1 (a) to (d), respectively. As in Fig. 1, TREMBA(5-0.8K) means that TREMBA's autoencoder consists of a 5-layer Encoder, which produces 800-dimension latent vector, and a 5-layer Decoder. In the same way, S1, S2, S3 consists of 4, 3, 2-layer Decoder and Encoder, which produces approximately 51000, 102000 and 409000-dimension latent vector.

IV. Experiment

In this section, we evaluate attacks by A.E. created by multiple autoencoders created as mentioned in III-C. In the evaluation, the A.E. created by the attack network is evaluated by two indexes, effectiveness and efficiency. Effectiveness is an indicator of whether the created A.E. can circumvent system authentication and successfully impersonate any target. Efficiency is an index that indicates the amount of calculation. In this experiment, it is necessary to input A.E. into the FRS and repeat updating perturbation based on the obtained results multiple times before the attack is successful. Inputting a face image into the attack target FRS once counted as 1 query and attack's efficiency is evaluated by the number of queries required for the attack to succeed. This study uses TREMBA as the baseline method and compares three proposed autoencoder networks with TREMBA's result using these two indicators.

A. Experimental environment and dataset

Table I shows the environment of this experiment. In our experiments, we used CASIA-WebFace [11] dataset. This dataset contains 453,453 face images of 10,575 identities. To eliminate the imbalances between each identity, 1,021 identities were selected and 81,680 images of the selected identities (80 images per identity) were used. We used MTCNN[12] to

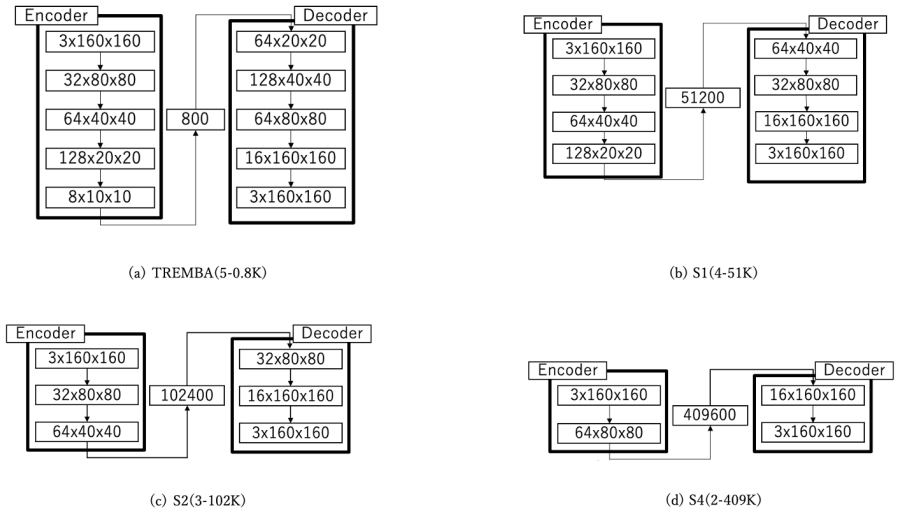

Fig. 1: Structure of autoencoders
(a)TREMBA's network (b)S1 (c)S2 (d)S3

TABLE I: Experimental environment

Language	Python 3.7.1
GPU	Geforce GTX TITAN X (11GB) × 2
CUDA cores	3584
CPU	Intel(R) Core(TM) i7-5930K CPU
Memories	64GB
OS/Kernel	Ubuntu 16.04 / Linux 4.15.0-74-generic

TABLE II: Result of concealer attack

Autoencoder	Attack success rate	Average queries
TREMBA	100.0%	26.74
S1	100.0%	**14.88**
S2	100.0%	36.56
S3	100.0%	34.19

TABLE III: Result of spoofing attack

Autoencoder	Attack success rate	Average queries
TREMBA	85.71%	4718.22
S1	93.85%	**1920.09**
S2	**96.15%**	3078.76
S3	93.85%	4399.48

align faces with 160x160 pixels images. Since the experiment consists of two processes, training the FRS and executing the attack, we split the images for each identity into 40:40 for training and attack set. In the training process, we used FaceNet[13] (InceptionResNetv1[14]) pretrained by the VGG2 dataset[15] as the target model and fine-tuned to be a 1,021 classes classifier using the training set. The accuracy of the fine-tuned FRS was 97.06%.

B. Attack Scenarios

In the evaluation of concealer attack, we find a perturbation δ that would cause FRS to misidentify the attack image using the loss function Eq(2). The attack is considered successful if there exists a perturbation δ such that $F(\boldsymbol{x}+\delta) = y' \neq y$ where y is the correct label. The attack is considered unsuccessful if no suitable perturbation is found after 50,000 queries. The attack was performed using all images in the attack set, and the attack success rate is calculated as (Number of successful attacks / Total number of images) $\times 100$. In addition, the average number of queries was calculated as (Number of queries when succeeds/ Number of successful attacks).

In the evaluation of spoofing attack, we find a perturbation δ that would cause FRS to misidentify the attack image using the loss function Eq(3). The attack is considered successful if there exists a perturbation δ such that $F(\boldsymbol{x}+\delta) = t \neq y$ where y is the correct label and t is a specific target. The attack is considered unsuccessful if no suitable perturbation is found after 50,000 queries. The attack success rate and the average number of queries were calculated by the same procedure as the concealer attack evaluation.

C. Results

Table II show the results of concealer attack experiments using TREMBA and S1, S2 and S3. As shown in Table II, TREMBA and S1, S2 and S3 all achieves a 100% success attack rate which shows that using TREMBA in face recognition scenario is extremely effective.

Table III show the results of spoofing attack experiments using TREMBA and S1, S2 and S3. From the table III, it can be seen the success rate of spoofing attack by TREMBA is 85.71% which is lower than other proposed methods. On the other hand, it can be seen that S2 shows a very high success rate of 96.14%. Regarding the average number of queries, S1, S2 and S3 all achieved an average number of queries less than TREMBA, and S2, which has the highest success rate, only requires 65% of queries compared to TREMBA to conduct a successful attack.

In this experiment, the depth of the autoencoder was set to

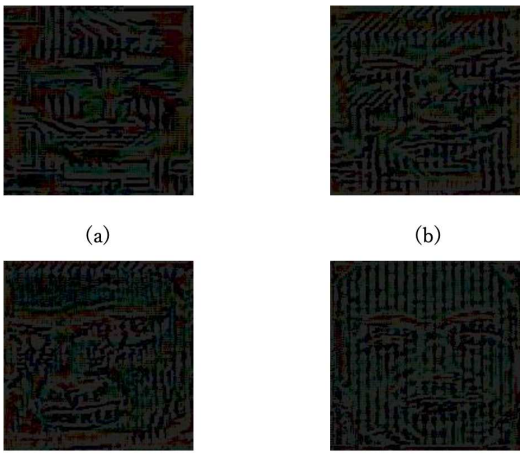

Fig. 2: (a)(b)(c)(d) is perturbations generated by TREMBA, S1, S2, S3 respectively

become shallower in the order TREMBA, S1, S2, S3. Since the attack success rate of TREMBA, which has the deepest network structure, is the lowest, it can be assumed that it is more effective to use autoencoder with a shallow network structure in face recognition scenario. On the other hand, S3, which has the shallowest network structure, could not achieve a better result than S2. From above observations, we can say that the attack success rate does not always increase and average number of queries does not always decrease as the network structure become shallower.

Huang et al. [5] states that searching for A.E. on the latent vector space $\boldsymbol{z}$ is more effective since these space likely to contain adversarial patterns. Based on this statement, it is considered that the dimension number 800 of the latent vector used by TREMBA's autoencoder is too small to find A.E. in face recognition, which leads to occasional fail and high average queries. On the other hand, the search space in S3 is too large (409600) which also leads to high average queries. From above arguments, it is suggested that it is necessary to consider not only the depth of the network but also the number of dimensions of the embedded vector $\boldsymbol{z}$ in the middle layer when constructing an attack network that is effective for face recognition.

V. Discussion

A. Transferability between recognition networks

Huang et al. [5] shows that it is possible to generate a A.E. against visual recognition system by using autoencoder and cause false authentication for other target network as well by producing semantic perturbations. We have also succeeded in producing such perturbations. Fig. 2 (a) $\sim$ (d) is an example of the perturbation obtained in this experiment and the outer shape of a human face can be perceived by naked eyes. From this example, it can be assumed that the Generator G has succeeded in acquiring the characteristics of the attack target, which later on contributes to the improvement of attack success rate and average queries. In this research, we have shown that configuration of autoencoder is effective for the attack against FRS. However, we have not reviewed transferability of A.E. generated by these autoencoders on other different networks. We leave it as a future work.

B. Attack against defended network

Huang et al. [5] has also succeeded in breaking through the authentication system with Madry et al. [16]'s method as a countermeasure against A.E. In fact, since many commercially available FRS have a built-in defense system, especially against A.E., proving the proposed method of this research is effective for defended system is also necessary. We leave it as a future work.

VI. Conclusion

In this study, we reviewed the Black-box A.E. generation method for FRS and proposed a method of constructing autoencoder that increases success attack rate while decreasing the number of average queries. From the experiment, we obtained results suggesting that the depth of the autoencoder built against FRS and dimension of embedded latent vector in the middle layer have a great influence on the attack success rate and average queries. However, we only achieved these results by reconstructing the autoencoder but finding the optimum construction method for FRS. From the results of this study, it was suggested that setting the shallow autoencoder different from the conventional TREMBA is effective for biometric authentication, especially for blackbox attacks on FRS. Future tasks are to clarify the relationship between width and depth and attack performance, and to quantitatively show the effectiveness of this method for face recognition.

Therefore, it is possible that we can achieve better results by examining appropriate indicators including the depth of the network and the dimension of the embedded vector. On the other hand, we leave defending against proposed attack method as a future task.

References

[1] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[2] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, 2016.

[3] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, Nov 2017.

[4] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks, 2020.

[5] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *International Conference on Learning Representations*, 2020.

[6] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information, 2018.

[7] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization, 2019.

[8] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, 2018.
[9] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017.
[10] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, and Jürgen Schmidhuber. Natural evolution strategies, 2011.
[11] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch, 2014.
[12] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.
[13] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.
[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.
[15] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2017.
[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.