

Poster: One of a Kind: Correlating Robustness to Adversarial Examples and Face Uniqueness

1st Giuseppe Garofalo
imec-DistriNet, KU Leuven
giuseppe.garofalo@kuleuven.be

2nd Tim Van hamme
imec-DistriNet, KU Leuven
tim.vanhamme@kuleuven.be

3rd Davy Preuveneers
imec-DistriNet, KU Leuven
davy.preuveneers@kuleuven.be

4th Wouter Joosen
imec-DistriNet, KU Leuven
wouter.joosen@kuleuven.be

Abstract—Face authentication lacks key metrics to assess the robustness of users’ representation within the system. We fill the gap by investigating face uniqueness, which is the distinctiveness of a face within a population, as a proxy for robustness against adversarial examples. By generating malicious input that escapes face verification, a dodging attack, we show a correlation between the amount of perturbation needed for successfully attacking a user and their uniqueness within a dataset. Our experiments span over multiple networks under a realistic threat model, indicating that unique users are significantly more resilient to gradient-based attacks than non-unique ones.

Index Terms—component, formatting, style, styling, insert

1. Introduction

Unique faces are those that are decidedly different from the rest of the population while being easy to recognize [1]. In modern face recognition, biometric uniqueness is directly affected by the separation between two distributions: the scores originated from matching two samples of the same user, i.e. the *genuine distribution*, and the scores derived from matching samples of different users, i.e. the *impostor distribution*.

It is well known that different faces exhibit varying performance within a system [2], which is linked to their relative uniqueness within a dataset. These performance are mainly expressed in terms of False Acceptance Rate (FAR), which comes from the mislabeling of a user, and False Rejection Rate (FRR), which is failing to match two samples of the same user. Identifying groups of users who contribute disproportionately to a type of error can uncover their vulnerabilities, eventually improving their resilience. The Doddington’s Zoo [3], shown in Fig. 1, is the first attempt to categorize users based on their verification performance, dividing between the score distribution of classes that cause the errors (goats, lambs, and wolves) and the distribution of the average user (sheep). The existence of these classes was later confirmed for a number of biometric modalities, including faces, and expanded to new classes in a concept known as *biometric menagerie* [2].

However, studies on the menagerie are usually disconnected from those on the security of modern face recognition. The advent of deep learning has boosted face

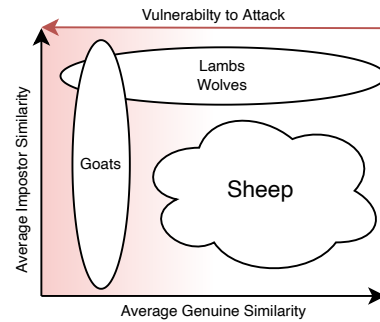


Figure 1. The four major classes of the Doddington’s Zoo with, overlapped in the background, the vulnerability to evasion attacks.

matching accuracy while broadening the threat surface. The assumption that training and test data are independent and identically distributed (i.i.d.) has proven to be hard to satisfy in practice, hence shifts in the data distribution affect algorithm performance and lead to poor generalization. In a malicious setting, imperceptible modifications known as adversarial examples can fool a system into assigning the wrong label to its input [4]. This shift represents a violation of the i.i.d. assumption, contributing to the overall FAR and FRR of a system in a way previously not envisioned by animal categorizations like the Doddington’s Zoo. Users who are contributing to the errors of a system (e.g. goats and lambs) need to be reconsidered in light of the novel threats posed by dataset shifts.

In our analysis we use a measure of uniqueness based on entropy to gather novel insights on the robustness of face recognition systems against adversarial examples. Motivated by modern tools [5], we construct a realistic yet simplified scenario of image publishing, where a user hides their identity before uploading a picture online. The attacker creates a human-imperceptible adversarial mask with the aim of fooling a face verification system that performs 1-vs-1 matching between a pair of images, which is called a dodging-attack. The amount of perturbation needed to escape matching will serve as an indicator of robustness towards an attack. We show that this notion of robustness is correlated with how well a user is embedded in the feature space, i.e. entropy-based uniqueness.

2. Methods

Our analysis can be divided in the following steps:

- 1) We compute the uniqueness of a set of users via Kullback-Leibler (KL) divergence estimation.
- 2) For a subset of the total users, we generate adversarial examples and derive attack robustness.
- 3) We correlate between uniqueness and robustness.
- 4) We quantify the unique and non-unique identities shared between networks.

2.1. Computing Uniqueness

Balazia et. al [1] approximate the KL divergence by using a distance estimator $D(x, S)$ that measures the average dissimilarity between a vector x and a set of vectors S in the embedding space. If Therefore, assuming $|G| = |I|$, the *uniqueness* U is equal to

$$KL(p_g|p_i) \approx U(G, I) = \frac{1}{|G|} \sum_{g \in G} \log \frac{D(g, I)}{D(g, G)} \quad (1)$$

where $D(g, G)$ measures the average distance of a template g from embeddings of the genuine distribution G , and $D(g, I)$ performs the same measurement towards the embeddings of other users, the impostor distribution I . It is possible to separate the contribution of the genuine scores, i.e. intra-class, from the one relative to the distance between genuine and impostor, i.e. inter-class:

$$U(G, I) = InterU(G, I) + IntraU(G) \quad (2)$$

$$InterU(G, I) = \frac{1}{|G|} \sum_{g \in G} \log D(g, I) \quad (3)$$

$$IntraU(G) = -\frac{1}{|G|} \sum_{g \in G} \log D(g, G) \quad (4)$$

2.2. Computing Attack Robustness

We define the robustness of an image, and by extension of a user, by computing the amount of perturbation in the input space needed for the adversarial example x_{adv} to cross the verification threshold θ , causing the mislabeling. We define this measure as Lowest Perturbation Budget (LPB). Given an embedding network f and a perturbation budget ϵ , we maximize the distance D :

$$\arg \max D(f(x_{adv}), f(x)), \text{ s.t. } \|x_{adv} - x\|_p < \epsilon \quad (5)$$

Since ϵ is a parameter to be decided before carrying out the attack, the LPB of an image x , and by extension of a user, is found by performing a binary search:

$$LPB(x) = \min \epsilon, \text{ s.t. } D(f(x_{adv}), f(x)) < \theta \quad (6)$$

TABLE 1. NETWORKS PERFORMANCES ON THE LFW DATASET.

Model	ACC	$TAR_{0.05\%}$	U	IntraU	InterU
FT	99.82	99.73	0.527	-2.939	3.467
FTo	99.75	99.60	0.535	-2.932	3.467
MF	99.43	98.20	0.437	-3.029	3.466
IR50-S	99.60	98.17	0.680	-2.786	3.466
IR50-C	99.68	99.17	0.512	-2.955	3.466
FN	99.23	85.80	0.730	-2.735	3.465

2.3. Correlation Analysis

Having computed the $U_{k,f}$ and $LPB_{k,f}$ for each user k , from embeddings computed using a model f , we are interested in the strength of the association between the two variables. We perform a correlation analysis by computing Kendall's τ score. Fixing a model f , given n pairs (U_x, LPB_x) ,

$$\tau = \frac{n_c - n_d}{\binom{n}{2}} \quad (7)$$

where n_c denotes the number of concordant pairs and n_d is the number of discordant pairs. A pair is concordant if given two users i and j , (U_i, U_j) and (LPB_i, LPB_j) have the same ordinal relationship (vice versa they are discordant). Therefore, τ measures the pairwise ordinal concordance between two variables, which is their monotonic relationship.

3. Results

In this section, the experimental setup is followed by an analysis of the results.

Experimental setup. Inspired by anti-facial recognition tools [5], we create adversarial examples by using the widely adopted gradient-based strategy FGSM and its iterative version BIM, both under the l_2 and l_{inf} norms. The attack is white-box and does not include a defensive strategy, which goes beyond the scope of our analysis. The attacked embedding networks are representative of the spectrum of SoTA face recognition solutions: FaceTransformer with and without overlapping patches (FT and FTo), MobileFace (MF), Inception-ResNet trained with a softmax and CosFace loss function (IR50-S and IR50-C), and FaceNet (FN). They share similarities that allow to selectively exclude the contribution of certain covariates when we focus on one single aspect of the models. *RobFR* [6] provides the backbone implementations on top of which we perform our white-box attack¹.

As test dataset, we pick the Labelled Faces in the Wild (LFW) [7]. LFW has the advantage of a big sample size, variability, and does not overlap with the training dataset of the attacked networks. From LFW we derive two subsets: *lfw-U* which is used to compute uniqueness, and *lfw-R* which is a further refined sample list to compute LPB scores.

Analysis. Table 1 displays the Uniqueness U and its intra-class and inter-class components. Surprisingly, the

1. <https://github.com/ShawnXYang/Face-Robustness-Benchmark>.

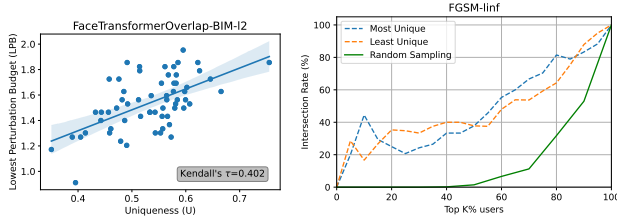


Figure 2. On the left, linear regression between LPB and U with $CI=95\%$. On the right, percentage of shared identities across models as we increase the most or least unique users.

network with the highest system uniqueness, i.e. FaceNet, is the worst performing one on LFW. This is because U delves deep into the score distribution, providing information that go beyond the average case.

Table 2 shows a moderate-to-strong correlation between the uniqueness of a user U_k and the average lowest perturbation budget needed to escape verification LPB_k . IntraU closely resembles the correlation of the overall U . This is expected since IntraU explains most of variance of U (see Table 1). Differently, the correlation analyses between LPB and InterU show little negative correlation and are, in most cases, not statistically significant (p-value greater than 0.05). Nonetheless, FaceTransformerOverlap scores have a weak negative correlation in the case of the BIM attack that is significant and worth of further investigations. Fig. 2 (left) shows the U score as a function of LPB for FaceTransformerOverlap and the (BIM, l_2) attack.

The results have an interpretation from a Doddington’s Zoo perspective (Fig. 1). Goats are users whose intraU is particularly low and are therefore especially susceptible to the class of attacks under study. On the other side of the spectrum, lambs, sheep and wolves are all eligible to containing a set of users with increased robustness compared to the average case. These users are the ones with higher U (and IntraU). Animal groups have been used to create adapting strategies that move verification threshold in order to cope with, e.g., large intra-class variance [8]. While adaptive thresholding can positively affect the FAR of the system in the benign case, moving the threshold triggers a cascade effect on the perturbation budgets needed to escape verification. Hence these strategies should account for the consequences on the robustness of the users.

Fig. 2 (right) plots the intersection rate between all the models as a function of the most unique and least unique K% users for the (FGSM, l_{inf}) configuration². To highlight the significance of the intersection rates, a baseline is added to the graph showing the intersection rates for a random sampling of users repeated 6 times (as many as the number of considered models).

4. Conclusion

Our analysis underlines the strong existing correlation between the resilience of a user against *dodging attacks* and the their distinctiveness within the population. This gives a clear indication whether a face is harder to protect

TABLE 2. KENDALL’S τ BETWEEN THE U SCORES AND LPB OF THE USERS IN THE l_{fw} - U LIST. A MODERATE-TO-STRONG POSITIVE CORRELATION IS FOUND FOR U AND IntraU.

Model	Attack	Norm	τ_U	τ_{IntraU}	τ_{InterU}
FT	BIM	l2	0.38	0.39	-0.17
FT	BIM	linf	0.34	0.34	-0.15
FT	FGSM	l2	0.38	0.38	0.00
FT	FGSM	linf	0.44	0.44	0.02
FTto	BIM	l2	0.40	0.41	-0.26
FTto	BIM	linf	0.35	0.36	-0.26
FTto	FGSM	l2	0.39	0.39	-0.08
FTto	FGSM	linf	0.46	0.46	-0.11
MF	BIM	l2	0.45	0.46	-0.16
MF	BIM	linf	0.40	0.41	-0.19
MF	FGSM	l2	0.41	0.42	-0.08
MF	FGSM	linf	0.45	0.45	-0.09
IR50-S	BIM	l2	0.43	0.43	0.03
IR50-S	BIM	linf	0.39	0.40	0.04
IR50-S	FGSM	l2	0.43	0.42	0.05
IR50-S	FGSM	linf	0.47	0.46	0.07
IR50-C	BIM	l2	0.49	0.49	-0.17
IR50-C	BIM	linf	0.45	0.45	-0.17
IR50-C	FGSM	l2	0.41	0.39	-0.10
IR50-C	FGSM	linf	0.47	0.45	-0.09
FN	BIM	l2	0.44	0.45	-0.20
FN	BIM	linf	0.41	0.42	-0.16
FN	FGSM	l2	0.37	0.39	-0.16
FN	FGSM	linf	0.43	0.43	-0.14

using Anti-Facial recognition [5], emerging privacy tools that rely on adversarial perturbations. Our exploration of the embedding space can be expanded beyond the class of dataset shift we take into consideration for our experiments, to account for different attacks, like impersonation, and benign covariate shifts, like changes in lighting conditions. The implications of our findings range from a better characterization of biometric system performance to a new understanding of what makes a face, therefore a user, more resilient against gradient-based adversarial attacks.

References

- [1] M. Balazia, S. Happy, F. Br mond, and A. Dantcheva, “How unique is a face: An investigative study,” in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 7066–7071.
- [2] N. Yager and T. Dunstone, “The biometric menagerie,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 2, pp. 220–230, 2008.
- [3] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, “Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation,” National Inst of Standards and Technology Gaithersburg Md, Tech. Rep., 1998.
- [4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.
- [5] E. Wenger, S. Shan, H. Zheng, and B. Y. Zhao, “Sok: Anti-facial recognition technology,” *arXiv preprint arXiv:2112.04558*, 2021.
- [6] X. Yang, D. Yang, Y. Dong, W. Yu, H. Su, and J. Zhu, “Delving into the adversarial robustness on face recognition,” *arXiv preprint arXiv:2007.04118*, 2020.
- [7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.
- [8] A. Mhenni, E. Cherrier, C. Rosenberger, and N. E. B. Amara, “Analysis of doddington zoo classification for user dependent template update: Application to keystroke dynamics recognition,” *Future Generation Computer Systems*, vol. 97, pp. 210–218, 2019.

2. Notably, the results hold for all the combinations (method, norm).