Robust Face Identification
Category: Computer Vision

Leo Mehr, Luca Schroeder
{leomehr, lucschr}@stanford.edu

October 10 2019

1 Motivation

Recent advances in facial recognition technology have revolutionized a plethora of applications, ranging from device authentication (e.g. Apple's Face ID) and e-commerce (e.g. Mastercard's 'selfie' payment technology) to public safety (e.g. police identifying criminals at large public events). And in the near future, facial recognition could be deployed in even higher-stake situations, such as targeting on military rifles [1] and management of pain medication for patients [2].

Thus, the risks of attacks on face identification technology and the consequences of misidentification grow ever larger—and so it is crucial to understand how brittle face identification currently is and how robust it can be made. These questions are particularly relevant in the wake of recent research on adversarial examples that has shown the vulnerability of DNN approaches to even small perturbations in natural images.

For our project, we propose to:

1. build a DNN for multi-class classification of individuals by facial recognition,
2. assess the robustness of the network through a range of state-of-the-art white- and black-box attacks,
3. and, if the attacks are successful, develop defenses against attacks to make the network more robust.

We anticipate the primary challenges of this project to be: (1) building and tuning a high-quality facial recognition neural network and (2) implementing a variety of attacks/defenses and rigorously assessing their efficacies.

2 Data and Methods

We plan to leverage the recently released IMDb-Face dataset [3], which contains 1.7 million faces and 59k identities, derived from 2.0 million raw IMDb images and designed to minimize the label noise common in other face recognition datasets. Each face image is labeled with the identity of the celebrity appearing in it.

Our DNN will take a photo with a bounding box for the query face as input, and output an $N + 1$ dimensional vector of probabilities corresponding to each of the N celebrities and one for "unknown face". To start, we will implement the well-known VGG16 architecture [4], which has also performed well in the Face Recognition domain [5].

For attacks we hope to experiment with a number of recent approaches, such as: one-pixel attacks with differential evolution [6], fast gradient sign method [7], rotation and image filter application, the $L_1/L_2/L_\infty$ distance metric attacks introduced by Carlini and Wagner [8], etc. Similarly, there are a number of potential defenses to consider, ranging from training on adversarial examples and defensive distillation [9], to detecting and rejecting adversarial examples as done by SafetyNet [10] or attempting to un-perturb or denoise them in the style of DefenseGAN [11]. We will start with 1 attack and 1 defense (if applicable) and expand our scope as time permits.

3 Evaluation and Success

Our primary metric will be classifier accuracy. For our base DNN, we will want this to be as high as possible. For our attacks to be successful, we want to maximize the difference:

$$\Delta = \text{Accuracy(normal test set)} - \text{Accuracy(adversarial test set)}$$

For our defenses to be successful, we will then want to shrink this Δ gap as much as possible to build a robust face recognition system.

It will be interesting to explore the interaction between different attacks and defenses and different classes of attacks (white-/black-box) and defenses. Thus one table that would be interesting to produce is a matrix that gives classification accuracy for each defense under each attack. It will also be interesting to consider the runtime characteristics of each attack and whether adversarial examples can be constructed in real-time e.g. to attack face recognition applications in live video.

References

[1] Matthew Cox. Army's next infantry weapon could have facial-recognition technology. *Military.com*, 2019. [Online; accessed 10-October-2019].

[2] Clarice Smith. Facial recognition enters into healthcare. *Journal of AHIMA*, 2018. [Online; accessed 10-October-2019].

[3] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. *ECCV*, 2018.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2015.

[5] Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv preprint 1804.06655*, 2018.

[6] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019.

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

[8] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privacy*, 2017.

[9] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *IEEE Symposium on Security and Privacy*, 2016.

[10] Jiajun Lu, Theerasit Issaranon, and David A. Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[11] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *ICLR*, 2018.