Adversarial Attacks on Facial Recognition Models

Andrew Milich and Michael Karr

{amilich, mkarr} @ stanford.edu

Overview

We analyzed the sensitivity of a facial recognition deep neural network (DNN) to adversarial images. Both attack mechanisms tested reduced accuracy.

Attack mechanisms

- Add random noise to images
- Recognize facial landmarks (eyes, nose, ears, mouth) using another DNN and add noise near them

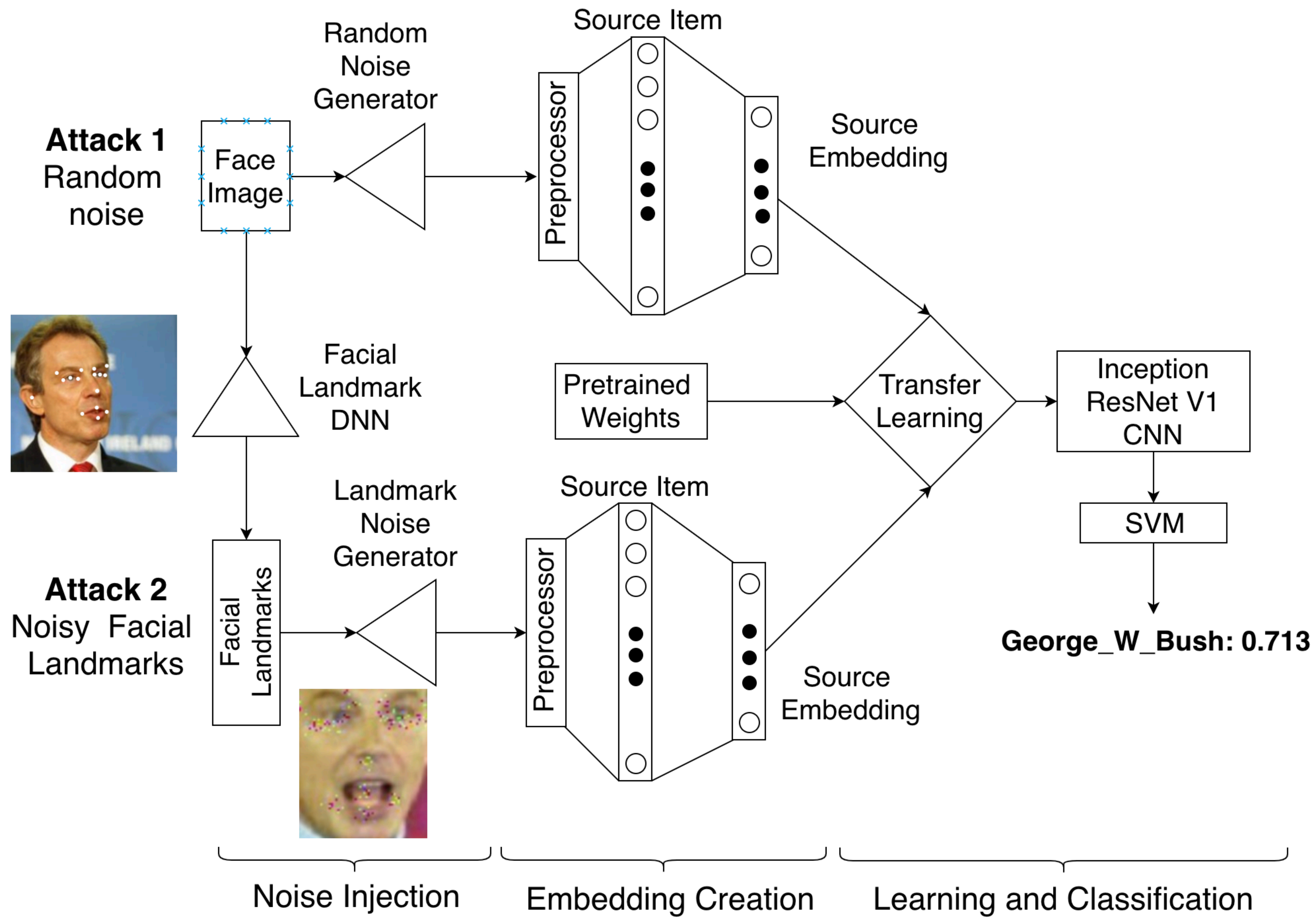
Defense mechanisms

- Train DNN facial recognition model on subset of adversarial images [4]

Implications

- Evading facial recognition models
- DNN sensitivity to "single pixel" attacks [6]

Model and Adversarial Example Creation



Discussion

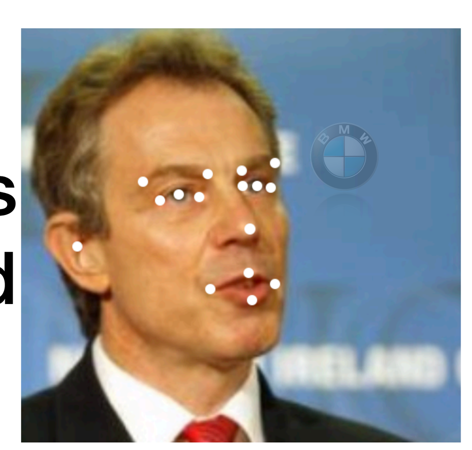
- Random noise lowers model classification accuracy
- Clustering noise around landmarks further reduces model performance, but less so for classes with more training images (George Bush has 500+ training samples)
- We rely on two transfer learning steps: One for facial recognition, and another for landmark recognition. Imperfect transfer learning could reduce model accuracy.
- Adversarial training by adding randomly perturbed images to the training set did not consistently increase performance, likely because of our use of randomness

Future Work

- Generate perturbations that minimize likelihood of classification as correct class
- Create physical "adversarial patch" for evading facial recognition [1]

Data

Labeled Faces in the Wild (LFW): 13,235 images of 5,750 individuals. We trained our facial recognition model on subjects with 10+ photos [3].



Facial Keypoints Dataset: Kaggle dataset of 7,049 images with facial landmarks identified by *(x,y)* positions [2].

Features

- Facial recognition: Inception Resnet V1 model outputs 128-dimensional embeddings that are classified by an SVM [5].
- Facial landmark recognition: Our DNN uses convolutional, dropout, and fully connected layers to recognize ears, eyes, eyebrows, nose, and mouth.

Model Accuracy

	Raw	Random Noise	Noisy Landmarks	Adv. Training
George Bush	0.98	0.91	0.88	0.94
Bill Clinton	0.99	0.75	0.58	0.63
Hamid Karzai	1.0	0.67	0.50	0.67
Tony Blair	0.97	0.69	0.71	0.62
John Negroponte	1.0	0.63	0.625	0.50

References

1. Brown, Tom B. et al. "Adversarial Patch." ArXiV. https://arxiv.org/abs/1712.09665.
2. Kaggle. "Basic Fully Connected NN." https://www.kaggle.com/madhawav/basic-fully-connected-nn/data.
3. "Labeled Faces in the Wild." http://vis-www.cs.umass.edu/lfw/.
4. Makelov, Aleksandar et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." ArXiv, June 2017. https://arxiv.org/pdf/1706.06083.pdf.
5. Murray, Cole. "Building a Facial Recognition Pipeline with Deep Learning in Tensorflow." Hacker Noon, https://hackernoon.com/building-a-facial-recognition-pipeline-with-deep-learning-in-tensorflow-66e7645015b8.
6. Su, Jiawei et al. "One pixel attack for fooling deep neural networks." ArXiv, February 2018. https://arxiv.org/abs/1710.08864.