# Hiding in Plain Sight: Adversarial Neural Net Facial Recognition

Crystal Qian '17, David Dobkin (Advisor)
Princeton University, Department of Computer Science

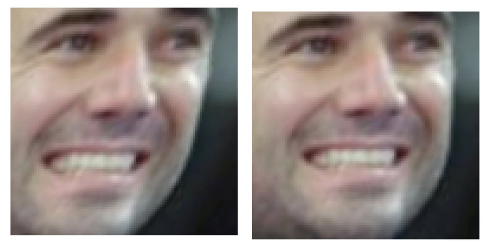
## Background

Deep neural networks (DNNs) excel at pattern-recognition tasks, particularly in visual classification. Implementations of DNN-based facial recognition systems approach and even exceed human-level performance on certain datasets. However, recent studies have revealed that imperceptible image perturbations can result in object misclassification in neural network-based systems. We explore the effects of image-agnostic perturbation methods at various stages of the facial recognition pipeline on network prediction errors, specifically training perturbations of the widely-used Labeled Faces in the Wild (LFW) dataset on FaceNet.

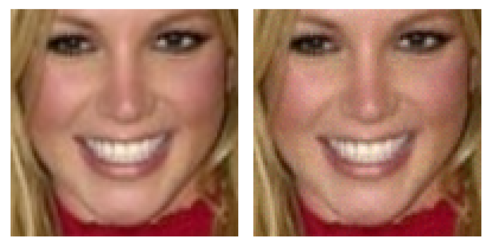
## Approach

We chose FaceNet as our recognition system because of its strong performance on the LFW dataset (99.63% accuracy). Our LFW dataset is condensed to 6,715 images of 610 people instead of 13,233 images of 5,759 people, filtered so that all people in our dataset have at least 4 images for cross-validation.

We target the alignment, representation, and classification stages of the recognition pipeline. Faces are aligned by the outer eyes and nose (left) and inner eyes and bottom lip (right).



Images are perturbed with Poisson noise and Gaussian noise. Poisson noise is calculated as follows, with k=1 and lambda sampled from the image. The image on the left is unaltered; the image on the right has applied Poisson noise.

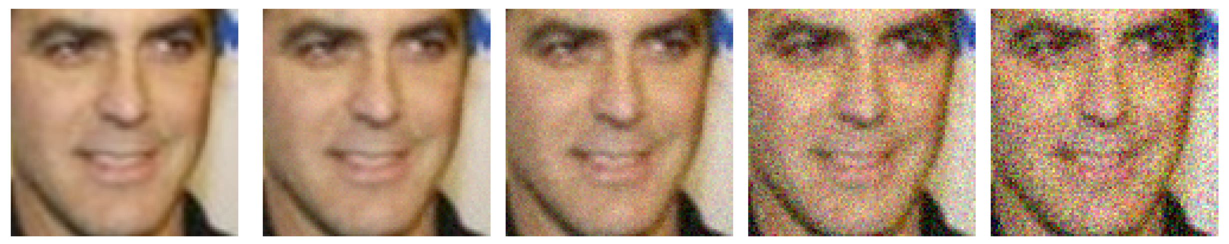


$$f(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Poisson noise is used in comparison with a similar Gaussian perturbation to compare the results of applied and additive noise. The additive Gaussian distribution is calculated as follows, with mu as the mean pixel value and sigma as the standard deviation.

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

Left to right: unaltered, sigma * sigma = 16, 100, 500, 1000.

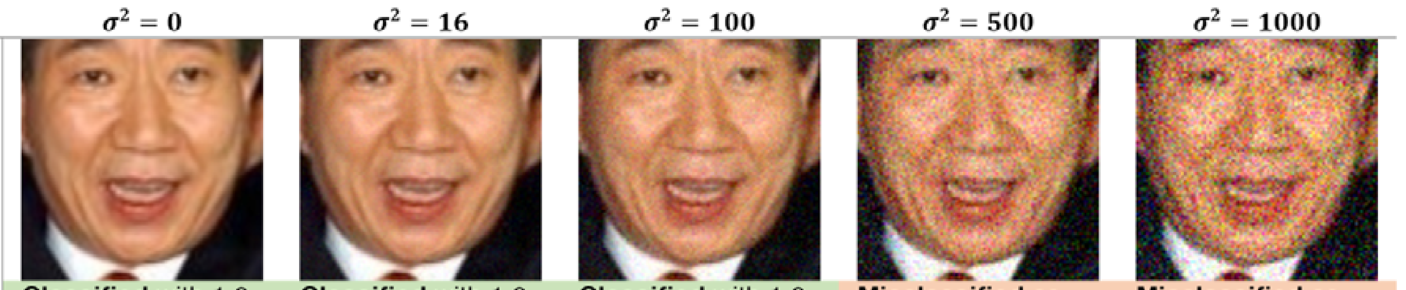


## Experiments

Can random noise significantly decrease the accuracy rate of various neural networks with minimal perturbation? Do networks trained on different classifiers respond similarly to perturbation? Do all types of noise: additive, multiplicative, applicative, etc. applied in the same amount result in the same degree of accuracy? Below is a summary of the parameters we tested on, as well as case studies at various levels of Gaussian noise on inner alignment. Notice the inconsistencies in classification confidence scores.

| Alignment Methods | Noise Generators | Classification Systems |
|---|---|---|
| Outer eyes and nose | Poisson | Linear SVM |
| Inner eyes and bottom lip | Gaussian, $\sigma^2 = 16$ | Radial SVM, $\gamma = 2$ |
| | Gaussian, $\sigma^2 = 100$ | Decision Tree, max depth 20 |
| | Gaussian, $\sigma^2 = 500$ | Gaussian Naive Bayes |
| | Gaussian, $\sigma^2 = 1000$ | Deep Belief Network, 300 epochs (learning decay .3, learning rate .3) |

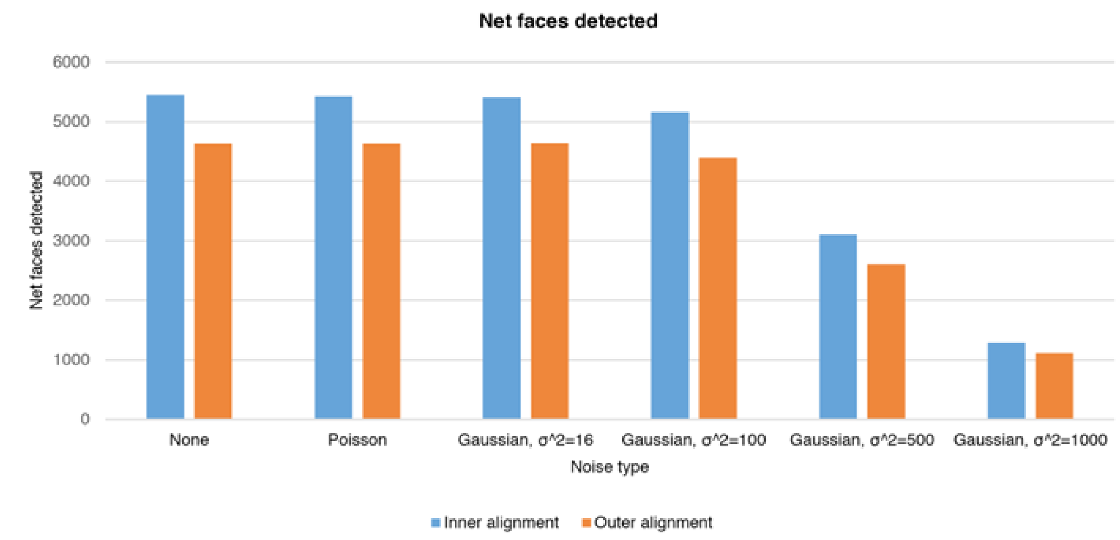
| | $\sigma^2 = 0$ | $\sigma^2 = 16$ | $\sigma^2 = 100$ | $\sigma^2 = 500$ | $\sigma^2 = 1000$ |
|---|---|---|---|---|---|
| ***Decision Tree*** | **Misclassified** as Vladimir Putin with .016 confidence. | **Misclassified** as Vladimir Putin with .016 confidence. | **Misclassified** as Michael Douglas with .033 confidence. | **Misclassified** as Michael Douglas with .033 confidence. | **Misclassified** as Michael Douglas with .033 confidence. |
| ***Linear SVM*** | **Misclassified** as Paul Bremer with .028 confidence. | **Misclassified** as Paul Bremer with .024 confidence. | **Misclassified** as Paul Bremer with .0316 confidence. | **Misclassified** as Paul Bremer with .043 confidence. | **Misclassified** as Paul Bremer with .035 confidence. |
| ***Radial SVM*** | **Classified** with .024 confidence. | **Classified** with .024 confidence. | **Classified** with .027 confidence. | **Classified** with .029 confidence. | **Classified** with .024 confidence. |
| ***Gaussian Naïve Bayes*** | **Classified** with 1.0 confidence. | **Classified** with 1.0 confidence. | **Classified** with 1.0 confidence. | **Classified** with 1.0 confidence. | **Classified** with 1.0 confidence. |
| ***DBN*** | **Classified** with .925 confidence. | **Classified** with .907 confidence. | **Classified** with .755 confidence. | **Classified** with .831 confidence. | **Classified** with .854 confidence. |



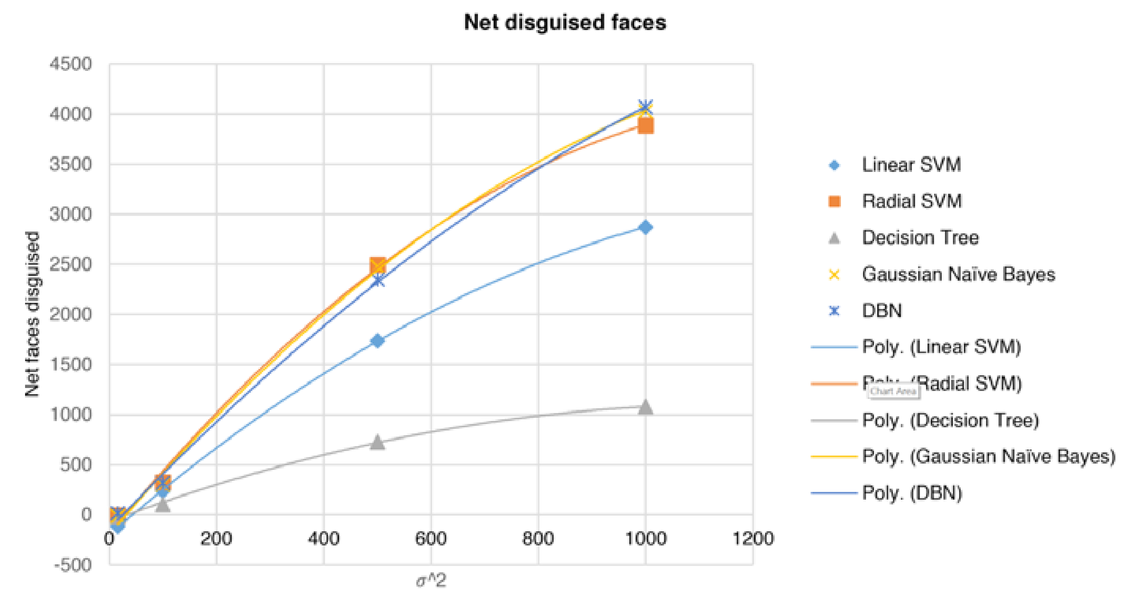
| | $\sigma^2 = 0$ | $\sigma^2 = 16$ | $\sigma^2 = 100$ | $\sigma^2 = 500$ | $\sigma^2 = 1000$ |
|---|---|---|---|---|---|
| ***Decision Tree*** | **Classified** with 1.0 confidence. | **Classified** with 1.0 confidence. | **Classified** with 1.0 confidence. | **Misclassified** as Jackie Chan with .056 confidence. | **Misclassified** as Jackie Chan with .056 confidence. |
| ***Linear SVM*** | **Classified** with .087 confidence. | **Classified** with .073 confidence. | **Classified** with .076 confidence. | **Classified** with .054 confidence | **Classified** with .026 confidence |
| ***Radial SVM*** | **Classified** with .069 confidence. | **Classified** with .065 confidence. | **Classified** with .068 confidence. | **Misclassified** as Junichiro Koizumi with .0316 confidence. | **Misclassified** as Junichiro Koizumi with .0226 confidence. |
| ***Gaussian Naïve Bayes*** | **Classified** with 1.0 confidence. | **Classified** with 1.0 confidence. | **Classified** with 1.0 confidence. | **Classified** with .999 confidence. | **Misclassified** as Tung Chee-hwa with .515 confidence. |
| ***DBN*** | **Classified** with .939 confidence. | **Classified** with .923 confidence. | **Classified** with .973 confidence. | **Classified** with .756 confidence. | **Classified** with .362 confidence. |

## Results

**Alignment:** Faces aligned by the inner eyes and upper lip were more readily detected than those aligned by the outer eyes and nose; however, alignment had no effect on relative classification accuracies.




**Classification:** The different classifiers showed varying success in correctly labelling faces. When the recognition scores were normalized by success of detection, the classifier had no effect on relative classification accuracies.




**Perturbation:** The most interesting results showed that adding small amounts of random noise to faces at times revealed (R) more images than disguised (D). Faces that were originally not detected or correctly classified tended to become correctly classified with minor perturbation.

The more perceptible noise is added to our dataset, the more faces are misclassified. However, the relationship between these changes is inconsistent on an individual basis; adding noise can increase classification confidence or expose faces in many cases.

| | $\sigma^2 = 16$ | | | $\sigma^2 = 100$ | | | $\sigma^2 = 500$ | | | $\sigma^2 = 1000$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | D | R | T | D | R | T | D | R | T | D | R | T |
| Decision Tree | 287 | 304 | 1664 | 340 | 234 | 1513 | 863 | 135 | 739 | 1134 | 56 | 250 |
| Linear SVM | 325 | 440 | 4612 | 454 | 217 | 4281 | 1862 | 94 | 2112 | 2896 | 25 | 722 |
| Radial SVM | 285 | 288 | 3618 | 628 | 300 | 3266 | 2604 | 107 | 1735 | 3916 | 29 | 632 |
| Gaussian Naïve Bayes | 261 | 272 | 4830 | 585 | 284 | 4518 | 2597 | 119 | 2341 | 4057 | 29 | 791 |
| DBN | 278 | 260 | 4911 | 585 | 268 | 4612 | 2648 | 118 | 2399 | 4100 | 30 | 859 |