# Detecting Photos with Digital Facial Enhancement

By Emily Ross and Ketan Agrawal
emiross@stanford.edu | agrawalk@stanford.edu

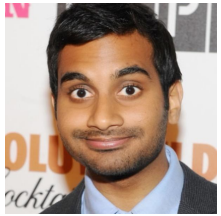
## Introduction

- Social issue: celebrities and "influencers" promote unrealistic beauty standards by posting digitally-enhanced photos as if they were real
- **Goal:** develop a classifier that can differentiate between photos that have been edited with "beautification" apps, specifically Facetune2, and those that have not.
- Though there have been models developed to detect generic photo warping,[1] there has not been research into training for Facetune edit recognition, specifically.
- Companies like Instagram are taking active steps[2] to protect the emotional well-being of users, illuminating a clear use case for such a tool.

## Data

- Since there are no datasets created for the purpose of detecting Facetune2, it proved a significant challenge to generating the data.
- An automated version of iOS's Switch Control Accessibility feature was used to apply "Auto" and smile-warping edits to 5875 photos from the CyberExtruder Ultimate Face Matching Data Set,[3] producing 5,880 positive examples.

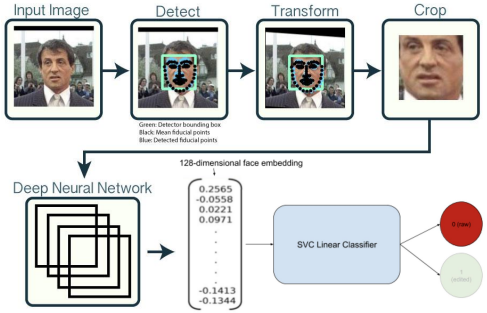
Negative Example: 600x600 pixel raw image



Positive Example: 600x600 pixel image passed through Facetune2

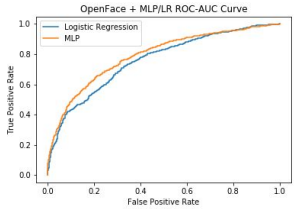
## Implementation

- Fed examples into pretrained OpenFace[4] convolutional network to retrieve 128-dimensional embeddings.
- **[EXP1]:** Used embeddings to train a scikit-learn Linear Support Vector Classifier
- **[EXP2]:** Used embeddings to train scikit-learn Multi-Layered Perceptron Neural Network




## Results

| Model | P | R | F1 | AUC |
|---|---|---|---|---|
| OpenFace + LR | 0.71 | 0.68 | 0.69 | 0.76 |
| OpenFace + MLP | 0.76 | 0.67 | 0.71 | 0.79 |




## Analysis

- ROC curve indicates that the MLP has only a slightly better ratio of true to false positives. Marginal benefit of NN over linear classifier is minimal when predicting from OpenFace embeddings.
- If MLP/LR give more weight to recognizing the "Auto" brightening feature, this could explain misclassified black and white photographs
- False negatives possibly result from photos where subject isn't smiling originally, so Facetune edits are subtle

Example A false-positive | Example A's true Facetune | Example B false-negative | Example B's true raw

## Future Work

- Run classifier/neural network on embeddings other than OpenFace, perform transfer learning using the UC Berkeley researchers' model
- Create saliency maps to understand why misclassifications occur

## References

[1] Wang, Sheng-Yu, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A. Efros. "Detecting Photoshopped Faces by Scripting Photoshop." *arXiv preprint arXiv:1906.05856* (2019).

[2] So, Adrienne. "Instagram Will Test Hiding 'Likes' in the US Starting Next Week." Wired. Condé Nast, November 8, 2019.

[3] CYBEREXTRUDER. (2019). CyberExtruder Ultimate Face Matching Data Set. [Directory of .jpg images]. Retrieved from https://cyberextruder.com/face-matching-data-set-download/

[4] B. Amos, B. Ludwiczuk, M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.