Designing Generative Adversarial Networks for Privacy-enhanced Face Recognition

Sebastian Raschka
Asst. Prof. of Statistics
@ U. of Wisconsin-Madison

https://sebastianraschka.com

14th International (Virtual) Conference on Human System Interaction (HSI 2021)
Gdansk University of Technology, Poland – July 08-10, 2021

http://hsi2021.welcometohsi.org
Topics

1. Biometric Face Recognition
2. Extracting Soft-Biometric Attributes from Face Images
3. Hiding Soft-Biometric Attributes in Face Images
4. PrivacyNet: GAN-based Multi-attribute Face Privacy
Biometric (Face) Recognition

A. Identification
Determine identity of an unknown person 1-to-\(n\) matching

B. Verification
Verify claimed identity of a person 1-to-1 matching
Applications of Biometric (Face) Recognition
#### Soft-Biometric Attributes

<table>
<thead>
<tr>
<th>Identity</th>
<th>Meryl Streep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Female</td>
</tr>
<tr>
<td>Age</td>
<td>72</td>
</tr>
<tr>
<td>Race</td>
<td>Caucasian</td>
</tr>
<tr>
<td>Medical</td>
<td>Healthy</td>
</tr>
</tbody>
</table>
Topics

1. Biometric Face Recognition

2. Extracting Soft-Biometric Attributes from Face Images

3. Hiding Soft-Biometric Attributes in Face Images

4. PrivacyNet: GAN-based Multi-attribute Face Privacy
Ex. 1: How difficult is it to extract gender information from face images?

Identity: Meryl Streep

<table>
<thead>
<tr>
<th>Gender</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>72</td>
</tr>
<tr>
<td>Race</td>
<td>Caucasian</td>
</tr>
<tr>
<td>Medical</td>
<td>Healthy</td>
</tr>
</tbody>
</table>
Very Easy: ResNet-50 Applied to Gender Classification


Evaluation

```python
with torch.set_grad_enabled(False):  # save memory during inference
    print('Test accuracy: %.2f%%' % (compute_accuracy(model, test_loader,)

Test accuracy: 97.40%
```

Ex. 2: How difficult is it to extract **age** information from face images?

Identity: Meryl Streep

<table>
<thead>
<tr>
<th>Gender</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>72</td>
</tr>
<tr>
<td>Race</td>
<td>Caucasian</td>
</tr>
<tr>
<td>Medical</td>
<td>Healthy</td>
</tr>
</tbody>
</table>
Ordinal Regression for ordinal data: integrating label order info

- **Ranking**: Predict Correct order  
  (0 loss if order is correct, e.g., rank a collection of movies by "goodness")

- **Ordinal regression**: Predict correct (ordered) label  
  (E.g., age of a person in years; here, regard aging as a non-stationary process)

Excerpt from the UTKFace dataset  
https://susanqq.github.io/UTKFace/

18 < 29 < 41

Cao, Mirjalili, Raschka (2020)  
*Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation*  
Pattern Recognition Letters. 140, 325-331
Age Prediction Datasets

**AFAD**
- 165,501 face images
- age range: 15-40 years

**MORPH-2**
- 55,608 face images
- age range: 16-70 years
4.2. Convolutional Neural Network Architectures

To implement CORAL, we replaced the last age classification tasks. For the remainder of this paper, we

Known for achieving good performance on a variety of im-

teration, the 128x128x3 face images were center-cropped to a

pixels to augment the model training. During model evalua-

ing data and 20% test data. All images were resized to

considered face images with age labels between 21-60 years

that no additional steps were required. In this study, we

ther alignment was applied. The UTKFace database (AFAD; 165,501 faces

tered in the Asian Face Database (AFAD; 165,501 faces

in the age range 14-62 years. The CACD database (55,608 face images) was preprocessed

using facial landmark detection (Sagonas et al. 2010, 2017) via

Raschka et al. 2016, Niu et al. 2016, He et al. 2015), no fur-

m 2018), no fur-

2018)

than the different methods to allow for fair comparisons.

The source code is available at

https://github.com/

Ba et al. 2015, Raschka et al. 2016, Niu et al. 2016, He et al. 2015)

between the predicted rank labels and the ground truth are

higher than a threshold

All CNNs were trained for 200 epochs with stochastic gra-

on NVIDIA GeForce 1080Ti and Titan V graphics cards.

RMSE values reported in this study were computed on

example

<table>
<thead>
<tr>
<th>Method</th>
<th>Random seed</th>
<th>MORMPH-2</th>
<th>AFAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>CE-CNN</td>
<td>0</td>
<td>3.26</td>
<td>4.62</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3.36</td>
<td>4.77</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.39</td>
<td>4.84</td>
</tr>
<tr>
<td>AVG ± SD</td>
<td></td>
<td>3.34 ± 0.07</td>
<td>4.74 ± 0.11</td>
</tr>
<tr>
<td>OR-CNN [16]</td>
<td>0</td>
<td>2.87</td>
<td>4.08</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2.81</td>
<td>3.97</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.82</td>
<td>3.87</td>
</tr>
<tr>
<td>AVG ± SD</td>
<td></td>
<td>2.83 ± 0.03</td>
<td>3.97 ± 0.11</td>
</tr>
<tr>
<td>CORAL-CNN (ours)</td>
<td>0</td>
<td>2.66</td>
<td>3.69</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2.64</td>
<td>3.64</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.62</td>
<td>3.62</td>
</tr>
<tr>
<td>AVG ± SD</td>
<td></td>
<td>2.64 ± 0.02</td>
<td>3.65 ± 0.04</td>
</tr>
</tbody>
</table>

Age prediction only off by 2 ½ to 3 ½ years on average

Table 1

Age prediction errors on the test sets. All models are based on the ResNet-34 architecture.

<table>
<thead>
<tr>
<th>Method</th>
<th>Random seed</th>
<th>MORPH-2</th>
<th>AFAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>CE-CNN</td>
<td>0</td>
<td>3.26</td>
<td>4.62</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3.36</td>
<td>4.77</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.39</td>
<td>4.84</td>
</tr>
<tr>
<td>AVG ± SD</td>
<td></td>
<td>3.34 ± 0.07</td>
<td>4.74 ± 0.11</td>
</tr>
<tr>
<td>OR-CNN [16]</td>
<td>0</td>
<td>2.87</td>
<td>4.08</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2.81</td>
<td>3.97</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.82</td>
<td>3.87</td>
</tr>
<tr>
<td>AVG ± SD</td>
<td></td>
<td>2.83 ± 0.03</td>
<td>3.97 ± 0.11</td>
</tr>
<tr>
<td>CORAL-CNN (ours)</td>
<td>0</td>
<td>2.66</td>
<td>3.69</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2.64</td>
<td>3.64</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.62</td>
<td>3.62</td>
</tr>
<tr>
<td>AVG ± SD</td>
<td></td>
<td>2.64 ± 0.02</td>
<td>3.65 ± 0.04</td>
</tr>
</tbody>
</table>

W Cao, V Mirjalili, and S Raschka (2020)

Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation

Pattern Recognition Letters. 140, 325-331

1. Biometric Face Recognition

2. Extracting Soft-Biometric Attributes from Face Images

3. Hiding Soft-Biometric Attributes in Face Images

4. PrivacyNet: GAN-based Multi-attribute Face Privacy
Biometric (Face) Recognition Can Be Useful
Soft-Biometric Attribute Mining Can Be Problematic in Absence of Consent

Identity: Meryl Streep
- Gender: Female
- Age: 72
- Race: Caucasian
- Medical: Healthy
Soft-biometric Attributes: Issues and Concerns

1. **Identity theft**: combining soft biometric info with publicly available data

2. **Profiling**: e.g., gender/race based profiling

3. **Ethics**: extracting data without users’ consent (e.g., intentional or via database breaches)
Preventing Automatic Extraction

Email-address harvesting

From Wikipedia, the free encyclopedia

Email harvesting or scraping is the process of obtaining lists of email addresses using various methods. Typically these are then used for bulk email or spam.

Contact

Personal email:

mail@sebastianraschka.com

Work email:

sraschka@wisc.edu

mail _at_ sebastianraschka .dot. com
Can/do we need to take similar measures to prevent soft-biometric attribute harvesting?
One solution: Storing face representation vectors with sensitive information removed


Very useful approach, but can have limitation for certain application domains, because
- not interpretable by humans
- not compatible with arbitrary face matching software
Goal: Selective Privacy

1. Perturb soft-biometric (e.g., gender) information
2. Ensure realistic face images
3. Retain biometric face recognition utility
Face Matcher

Gender Classifier
Autoencoder to perturb image

\[ \phi(X) = X' \]

Face Matcher

Gender Classifier
General architecture of the semi-adversarial network (SAN)

General architecture of the semi-adversarial network (SAN)

Objective 1: Realistic images

Objective 2: Retain matching utility

Objective 3: Confound gender
Semi-adversarial network

Objective 1: Realistic images

Objective 2: Retain matching utility

Objective 3: Confound gender
Convolutional neural networks
SAN Examples

Original Inputs

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>99%</td>
<td>98%</td>
</tr>
<tr>
<td>97%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Outputs

<table>
<thead>
<tr>
<th>Female</th>
<th>Male</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>69%</td>
<td>99%</td>
<td>71%</td>
<td>58%</td>
</tr>
</tbody>
</table>
Replacing Detachable Parts for Evaluation

Use methods *unseen* during training
3.2. Retaining matching accuracy

face mixing scheme proposed in [21] result in higher error in the case of G-COTS software, perturbations made by the same-gender prototype appear much closer to the original images compared to the reconstructed images from neutral- or opposite-gender prototypes. Overall, we were able to retain a good matching performance even when using neutral- or opposite-gender prototypes are not directly subject to this loss term (see Section 2.3). As a result, the reconstructions that are produced via our SAN method are very close to the value obtained from the unperturbed original images.

Table 2. Error rates in gender prediction using IntraFace and G-COTS software and the resulting ROC curves are obtained from outputs of the mixing approach proposed in [21] are heavily impacted, resulting in de-identified curves obtained from outputs of the mixing approach proposed in [21].

Figure 6. ROC curves comparing the performance of IntraFace and opposite-gender prototype. On the other hand, the ROC curve of the reconstructed images coming from same-gender prototype, in the same datasets: (a) MUCT, (b) LFW, and (c) AR-face.

Figure 7. ROC curves showing the performance (true and false matching rates) of M-COTS biometric matching software on the original dataset.

4. Conclusions

In this work, we focused on developing a semi-adversarial network for imparting soft-biometric privacy to face images. In particular, our semi-adversarial network perturbs an input face image such that gender prediction is very close to the value obtained from the unperturbed original image. Use neutral- or opposite-gender prototypes are not directly subject to this loss term (see Section 2.3). As a result, the reconstructions that are produced via our SAN method are very close to the value obtained from the unperturbed original image.

Face matching performance

Multi-subject comparisons

Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers

Improvements to construct a more diverse set of SAN models for better generalizability

Figure 1: Diversity in an ensemble SAN can be enhanced through its auxiliary gender classifiers (see Figure 2). When the auxiliary gender classifiers lack diversity, ensemble SAN cannot generalize well to arbitrary gender classifiers.

Figure 4: Face prototypes computed for each group of attribute labels. The abbreviations at the bottom of each image refer to the prototype attribute-classes, where Y=young, O=old, M=male, F=female, W=white, B=black.

FlowSAN: Privacy-enhancing Semi-Adversarial Networks to Confound Arbitrary Face-based Gender Classifiers

(A) $I_{\text{orig}} \rightarrow \text{SAN}_1 \rightarrow I'_1 \rightarrow \text{SAN}_2 \rightarrow I'_2 \rightarrow \cdots \rightarrow \text{SAN}_n \rightarrow I'_n$

(B) $I_{\text{orig}} \rightarrow I'_1 \rightarrow I'_2 \rightarrow I'_3 \rightarrow I'_4 \rightarrow I'_5$

Improvements to better control the perturbations and enhance the removal of soft-biometric information

Gender Prob. P(Male): 80% 56% 34% 14% 6%

Matching Acc. w/ original: 98% 98% 97% 94% 91%

V Mirjalili, S Raschka, A Ross (2019)
FlowSAN: Privacy-enhancing Semi-Adversarial Networks to Confound Arbitrary Face-based Gender Classifiers
IEEE Access 2019, 10.1109/ACCESS.2019.2924619
1. Biometric Face Recognition
2. Extracting Soft-Biometric Attributes from Face Images
3. Hiding Soft-Biometric Attributes in Face Images
4. PrivacyNet: GAN-based Multi-attribute Face Privacy
Selective and collective perturbations for imparting multi-attribute privacy to face images

Selective = **which** attributes to conceal

Collective = **how many** attributes to conceal

V Mirjalili, S Raschka, and A Ross (2020)
PrivacyNet: Semi-Adversarial Networks for Multi-attribute Face Privacy
PrivacyNet replaces the convolutional autoencoder with a GAN-based model with cycle consistency loss.
CycleGAN


Does not require paired images from source and target domains

Zebras ⇔ Horses

zebra → horse

horse → zebra

Figure 1: Given any two unordered image collections \(X\) and \(Y\), our algorithm learns to automatically “translate” an image from one into the other and vice versa: (left) Monet paintings and landscape photos from Flickr; (center) zebras and horses from ImageNet; (right) summer and winter Yosemite photos from Flickr. Example application (bottom): using a collection of paintings of famous artists, our method learns to render natural photographs into the respective styles.

Abstract

Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. However, for many tasks, paired training data will not be available. We present an approach for learning to translate an image from a source domain \(X\) to a target domain \(Y\) in the absence of paired examples. Our goal is to learn a mapping \(G: X → Y\) such that the distribution of images from \(G(X)\) is indistinguishable from the distribution \(Y\) using an adversarial loss. Because this mapping is highly under-constrained, we couple it with an inverse mapping \(F: Y → X\) and introduce a cycle consistency loss to enforce \(F(G(X)) ⇡ X\) (and vice versa). Qualitative results are presented on several tasks where paired training data does not exist, including collection style transfer, object transfiguration, season transfer, photo enhancement, etc. Quantitative comparisons against several prior methods demonstrate the superiority of our approach.

1. Introduction

What did Claude Monet see as he placed his easel by the bank of the Seine near Argenteuil on a lovely spring day in 1873 (Figure 1, top-left)? A color photograph, had it been invented, may have documented a crisp blue sky and a glassy river reflecting it. Monet conveyed his impression of this same scene through wispy brush strokes and a bright palette.

What if Monet had happened upon the little harbor in Cassis on a cool summer evening (Figure 1, bottom-left)? A brief stroll through a gallery of Monet paintings makes it possible to imagine how he would have rendered the scene: perhaps in pastel shades, with abrupt dabs of paint, and a somewhat flattened dynamic range.

We can imagine all this despite never having seen a side by side example of a Monet painting next to a photo of the scene he painted. Instead, we have knowledge of the set of Monet paintings and of the set of landscape photographs. We can reason about the stylistic differences between these...
Abstract


1. Introduction

We can reason about the stylistic differences between these Monet paintings and landscape photos from Flickr; indeed, our algorithm learns to automatically “translate” an image from ImageNet; perhaps in pastel shades, with abrupt dabs of paint, and a possible to imagine how he would have rendered the scene: a glassy river reflecting it. Monet conveyed his impression of Yosemite; a river, perhaps in the summer and winter.

Figure 1: Given any two unordered image collections from ImageNet; our algorithm learns to render natural photographs into the respective styles.

CycleGAN


Real/Generated

Image from domain A

Real/Generated

Image from domain B

$D_A$

$D_B$

$G_{AB}$

$G_{BA}$
Conditional GAN

Conditional GAN

PrivacyNet Architecture

4 Subnetworks

Target Labels $c_t$
Input Image $X$

Generator $G$

Output Image $X'$

Discriminator $D_{src}$
Attribute Classifier $D_{attr}$
Auxiliary Face Matcher $M$

Real / Synthesized
Gender, Age, Race
Match score $\mathcal{L}_{G,m}$

Original Labels $V$
Normalized face descriptor $\mathcal{E}$

Distance between normalized features

The total loss terms for training the discriminator $D_{tot} = \lambda_{D,src} \mathcal{L}_{G,src} + \lambda_{D,attr} \mathcal{L}_{G,attr}$

where,

$\mathcal{L}_{G,src} = \mathbb{E}_{X \sim V} \left[ - \log \left( D_{src}(X) \right) \right]$

and

$\mathcal{L}_{G,attr} = \mathbb{E}_{X \sim V} \left[ - \log \left( D_{attr}(X) \right) \right]$

The loss term for optimizing the performance of the biometric face matcher $L_{G,m}$ is defined as

where

$\lambda_{D,src}$ and $\lambda_{D,attr}$ are hyperparameters representing the relative weights for the corresponding loss terms. The individual terms of the total loss for the discriminator $(\mathcal{L}_{G,src} + \mathcal{L}_{G,attr})$ are as follows:

$\mathcal{L}_{G,src} = \mathbb{E}_{X, V} \left[ \log (1 - D_{src}(X, V)) \right]$

and

$\mathcal{L}_{G,attr} = \mathbb{E}_{X, V} \left[ - \log \left( \mathbb{E}_{G \sim V} \left[ D_{attr}(G, V) \right] \right) \right]$

where $G_{G,m}$ is the matched generated face image.
PrivacyNet's Cycle Consistency

![Diagram showing PrivacyNet's Cycle Consistency]

Input Image \( X \) \rightarrow \text{Original Labels} \rightarrow \text{Target Labels} \rightarrow \text{Output Image} \( X' \)

\( C_0 \) \text{Original Labels}

\( C_t \) \text{Target Labels}
Discriminator Loss Function

\[
\mathcal{L}_D,\text{tot} = \mathcal{L}_{D,\text{src}} + \lambda_{D,\text{attr}} \mathcal{L}_{D,\text{attr}}
\]

where, \( \mathcal{L}_{D,\text{src}} \) and \( \mathcal{L}_{D,\text{attr}} \) are defined as

\[
\mathcal{L}_{D,\text{src}} = -\mathbb{E}_{X \sim p_{\text{data}}} \log D_{\text{src}}(X) + \mathbb{E}_{X \sim p_{\text{data}}} \log (1 - D_{\text{src}}(G(X)))
\]

and

\[
\mathcal{L}_{D,\text{attr}} = -\mathbb{E}_{X \sim p_{\text{data}}} \log D_{\text{attr}}(X) + \mathbb{E}_{X \sim p_{\text{data}}} \log (1 - D_{\text{attr}}(G(X)))
\]

Next, the loss terms for attribute classification are defined as

\[
\mathcal{L}_{G,\text{attr}} = \mathbb{E}_{X \sim p_{\text{data}}} \left[ D_{\text{attr}}(X) \right] - \mathbb{E}_{X \sim p_{\text{data}}} \left[ D_{\text{attr}}(G(X)) \right]
\]

where, \( \mathcal{L}_{G,\text{attr}} \) represents the expected value of the loss function over the distribution of the original face image, and \( \mathcal{L}_{G,\text{attr}} \) is the loss function over the distribution of the synthesized image.

Finally, the auxiliary face matcher loss is defined as

\[
\mathcal{L}_{G,\text{m}} = \mathbb{E}_{X \sim p_{\text{data}}} \left[ M(X) \right] - \mathbb{E}_{X \sim p_{\text{data}}} \left[ M(G(X)) \right]
\]

where, \( M(X) \) is a metric face matcher.
Generator Loss Function

\[ \mathcal{L}_{G,tot} = \mathcal{L}_{G,src} + \lambda_{G,\text{attr}} \mathcal{L}_{G,\text{attr}} + \lambda_m \mathcal{L}_{G,m} + \lambda_{\text{rec}} \mathcal{L}_{G,\text{rec}} \]
target age group is specified as A0 (young), A1 (middle-aged), or A2 (old).

Baseline-GAN models. The rows are marked by their selected attributes: G: gender, R: race, and A: age, where the specific
Fig. 8: ROC curves showing the performance of unseen face matchers on the original images for PrivacyNet, the baseline-GAN model, face mixing [34] approach and the controllable face privacy [35] method. The results show that ROC curves of PrivacyNet have the smallest deviation from the ROC curve of original images suggesting that the performance of face matching is minimally impacted, which is desired.
Limitation

\[ y_{\text{age}} = \begin{cases} 
0 & \text{age} \leq 30 \\
1 & 30 < \text{age} \leq 45 \\
2 & 45 < \text{age} 
\end{cases} \]
Suggested future solutions now that we can hide/change the age in face images ...
Thank You!

https://sebastianraschka.com
@rasbt
Sebastian Raschka