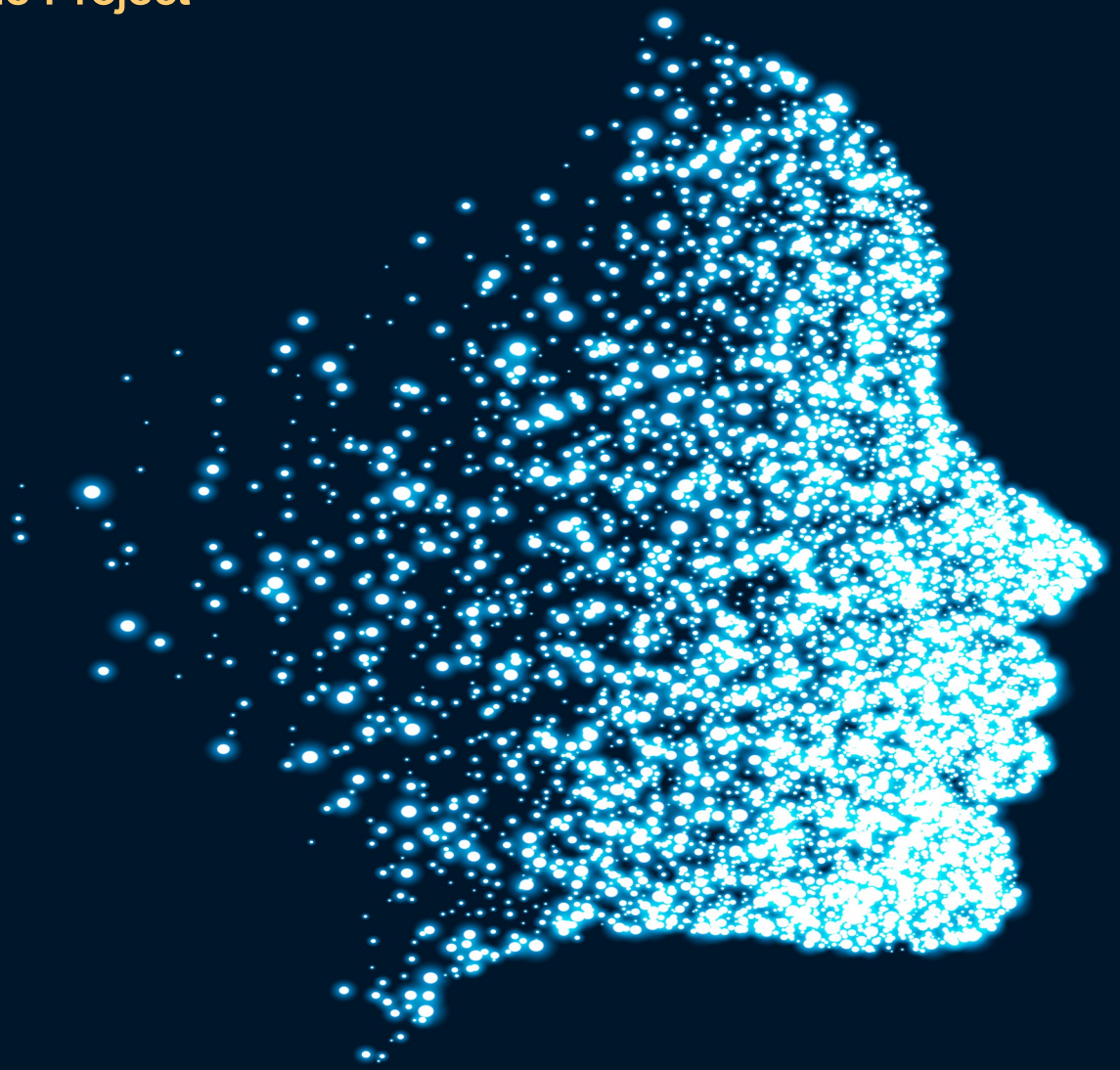
AN INTERPRETABLE MACHINE LEARNING MODEL FOR GENDER PREDICTION USING SHAP AND CLASS ACTIVATION MAPS

Capstone Project



- Renzo P. Castagnino

Outline

1. Introduction
    - Some News
    - Trend and Challenges
2. Main question
3. Data collection & Preprocessing
4. Training models
    1. VGG-16
    2. ResNet
    3. Inception v3
5. Interpretability models
    1. SHAP
    2. Class Activation Maps
6. Results
7. Conclusion
8. Future Research

Let’s Begin with some news!

THE NEW YORK TIMES
April 2019

One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority

In a major ethical leap for the tech world, Chinese start-ups have built algorithms that the government uses to track members of a largely Muslim minority group.

SenseTime is among the Chinese artificial intelligence companies developing facial recognition technology. Gilles Sabrié for The New York Times

By Paul Mozur

April 14, 2019

阅读简体中文版 · 閱讀繁體中文版

The Chinese government has drawn wide international condemnation for its harsh crackdown on ethnic Muslims in its western region, including holding as many as a million of them in detention camps.

Now, documents and interviews show that the authorities are also using a vast, secret system of advanced facial recognition technology to track and control the Uighurs, a largely Muslim minority. It is the first known example of a government intentionally using artificial intelligence for racial profiling, experts said.

The facial recognition technology, which is integrated into China's rapidly expanding networks of surveillance cameras, looks exclusively for Uighurs based on their appearance and keeps records of their comings and goings for search and review. The practice makes China a pioneer in applying next-generation technology to watch its people, potentially ushering in a new era of automated racism.

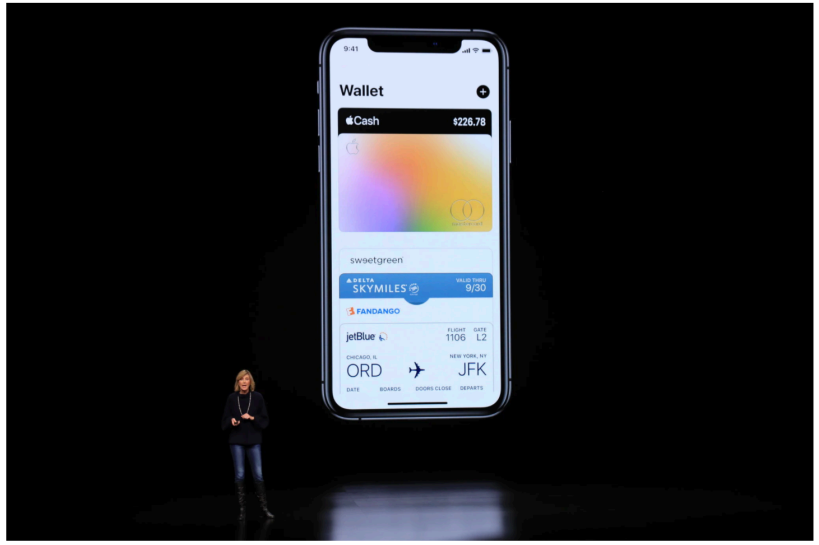
Source:https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html

THE NEW YORK TIMES
November 2019

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was "sexist" against women applying for credit.



Jennifer Bailey, vice president of Apple Pay. Regulators are investigating Apple Card's algorithm, which is used to determine applicants' creditworthiness. Jim Wilson/The New York Times

"My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time," Mr. Hansson wrote Thursday on Twitter. "Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does."

Mr. Hansson's tweets caught the attention of more than just his 350,000 followers.

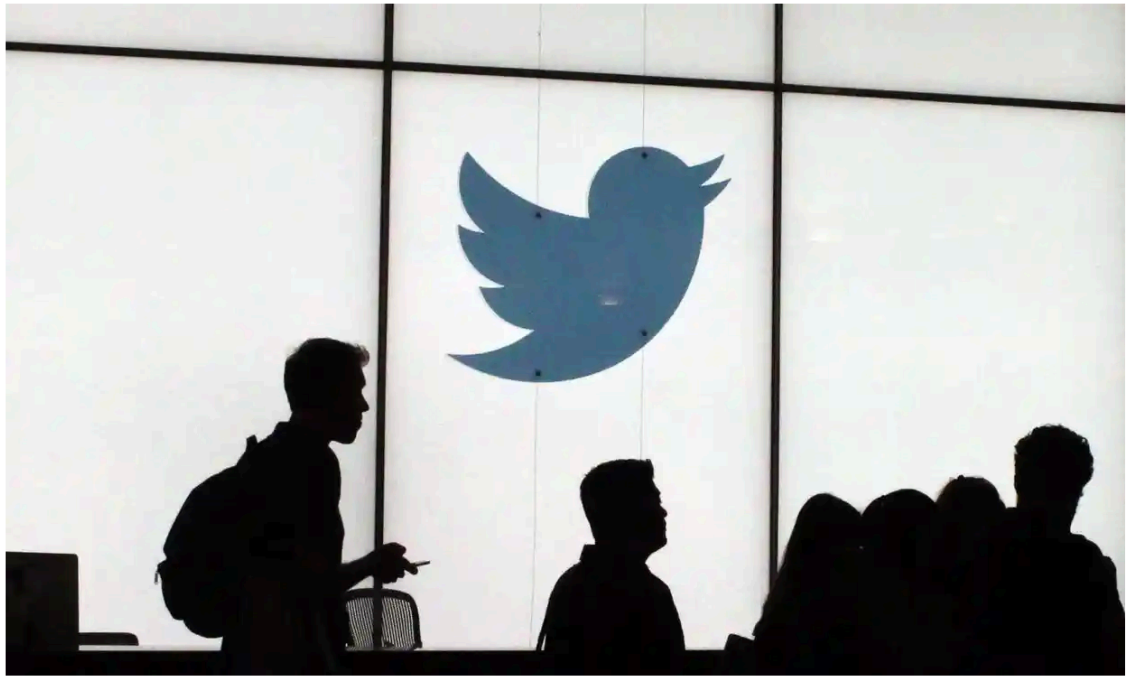
They struck a nerve with New York State regulators, who announced on Saturday that they would investigate the algorithm used by Apple Card to determine the creditworthiness of applicants.

Source: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

THE GUARDIAN
September 2020

Twitter apologises for 'racist' image-cropping algorithm

Users highlight examples of feature automatically focusing on white faces over black ones



▲ Twitter users began to spot flaws in the feature over the weekend. Photograph: Glenn Chapman/AFP/Getty Images

Twitter has apologised for a “racist” image cropping algorithm, after users discovered the feature was automatically focusing on white faces over black ones.

Source: https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm

Machine Learning has progressed dramatically over the past decades

However…

Deploying Machine Learning, you face the risk that it be discriminatory, biased, inequitable, exploitative, or opaque.

Source: https://bigthink.com/technology-innovation/machine-learning-ethics

We need to figure out how to train models that people can use safely, reliably, and in a way that they understand."

Source:https://news.mit.edu/2020/aleksander-madry-machine-learning-0308

THE IMPORTANCE OF
INTERPRETABILITY

Why not just trust the model and ignore why it made a certain decision?

"The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks."

(Doshi-Velez and Kim 2017)

The necessity for interpretability comes from an incompleteness in the problem formalization.

For certain problems or tasks it is not enough to get the answer (the what)

(Doshi-Velez and Kim 2017)

THE CURRENT PROJECT AIMS TO CREATE AN INTERPRETABLE MACHINE LEARNING MODEL FOR GENDER RECOGNITION.

SIMILAR RESEARCH

Towards Interpretable Face Recognition

Bangjie Yin[1*] Luan Tran[1*] Haoxiang Li[2†] Xiaohui Shen[3†] Xiaoming Liu[1]
[1]Michigan State University [2]Wormpex AI Research [3]ByteDance AI Lab

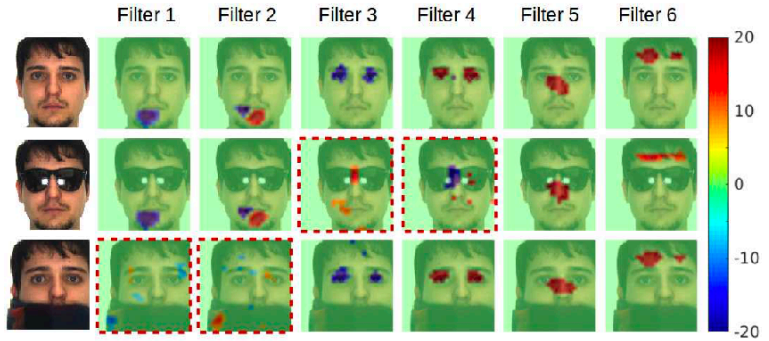


Figure 1. An example on the behaviors of an interpretable face recognition system: left most column is three faces of the same identity and right six columns are filter responses from six filters; each filter captures a clear and consistent semantic face part, e.g., eyes, nose, and jaw; heavy occlusions, eyeglass or scarf, alternate responses of corresponding filters and make the responses being more scattered, as shown in red bounding boxes.

Source: http://cvlab.cse.msu.edu/project-interpret-FR

Project Diagram

Data
(Images between Male & Female)

VGG-16

ResNet50

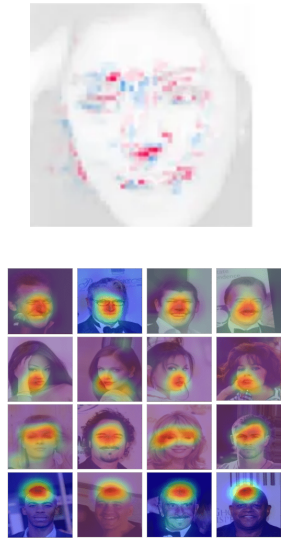
Inception v3

Pre-trained Models

SHAP

Class Activations Maps

Interpretability Models



Interpretable Results

THE DATA

CelebA Dataset

200k+ face images with attributes
Including gender

84,439

Female: ~~118,165~~

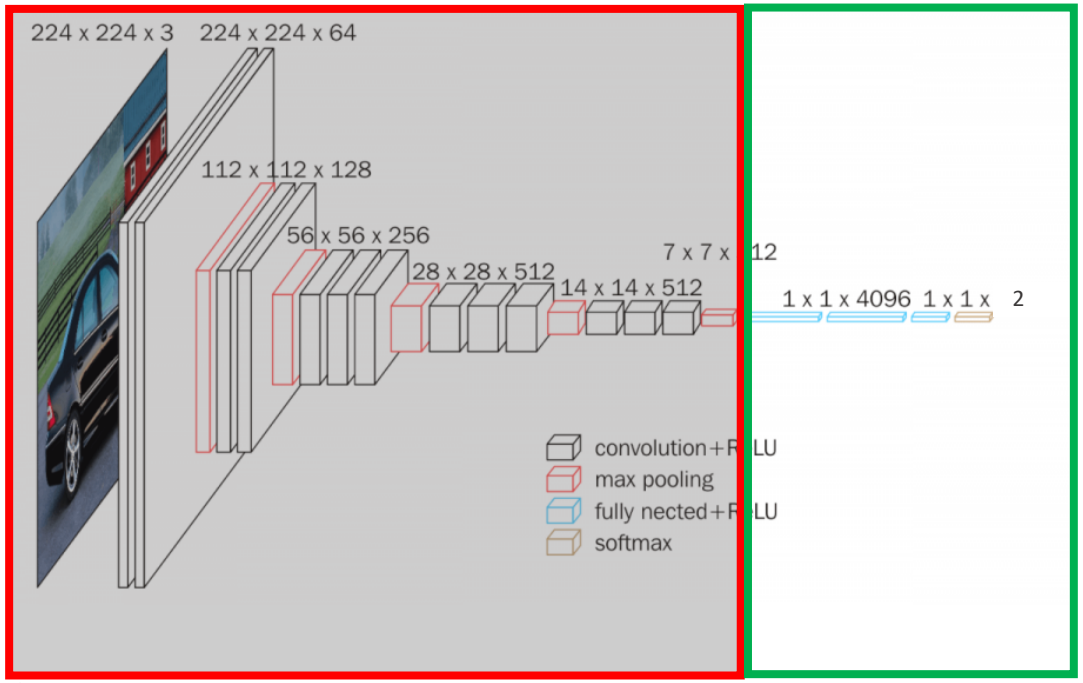
Male: 84,439

Source: https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

Pre-trained

MODELS

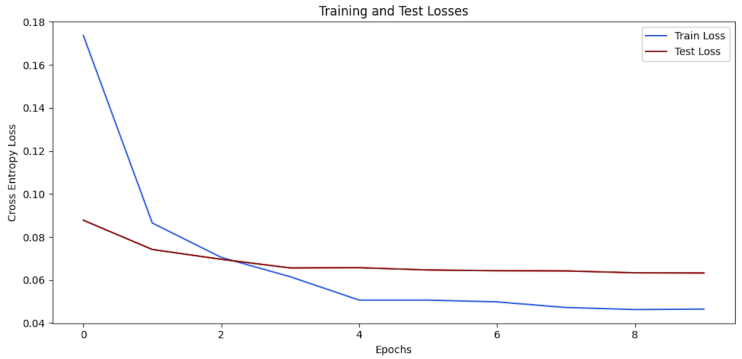
VGG-16

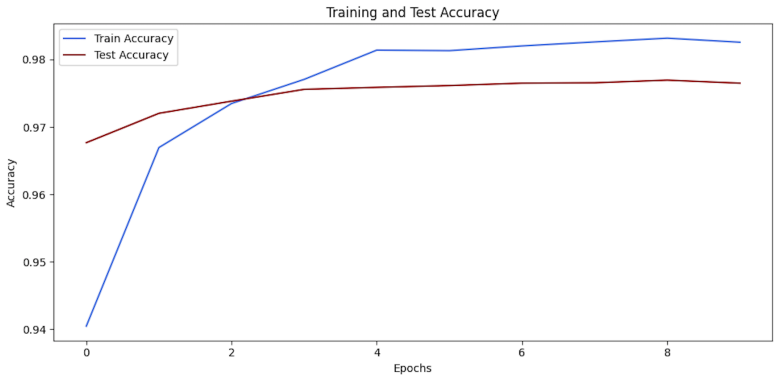


Source: Inception Module (source: original paper)

Frozen

Trainable



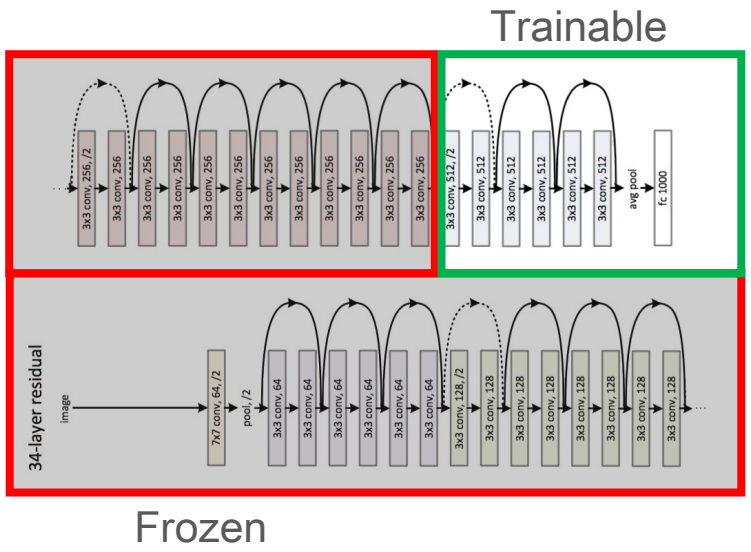


Hyperparameters:
Batch size = 64
LR = 0.001
Epochs = 10

Train Accuracy: 97.6%
Val. Accuracy: 98.2%

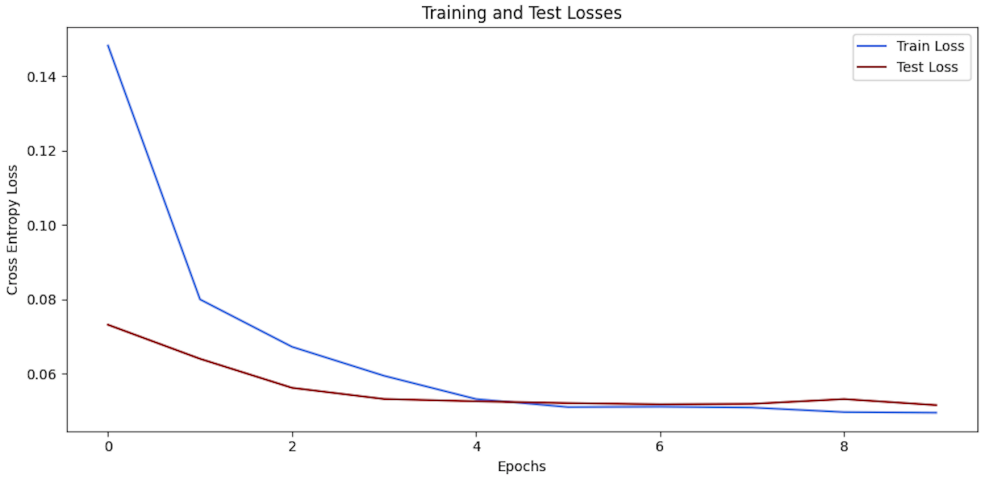
Train Loss: 0.05
Val Loss: 0.07

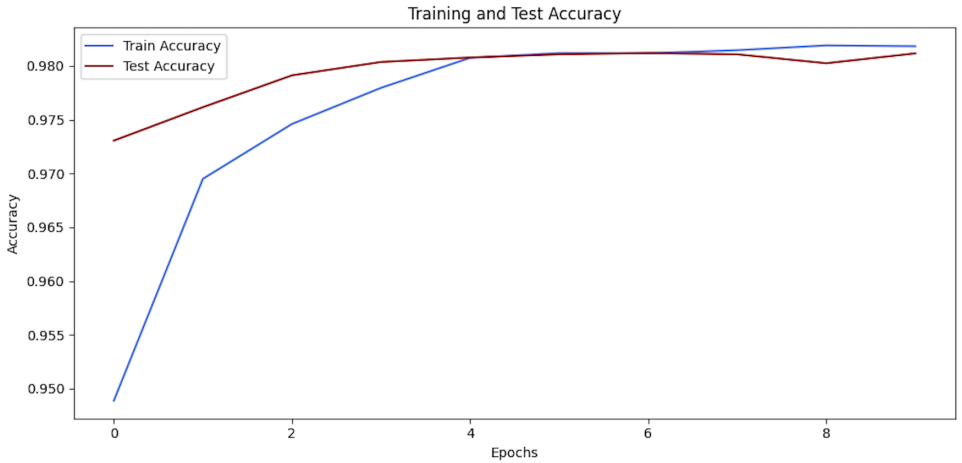
ResNet50



Source: Inception Module (source: original paper)

Hyperparameters:
Batch size = 64
LR = 0.001
Epochs = 10

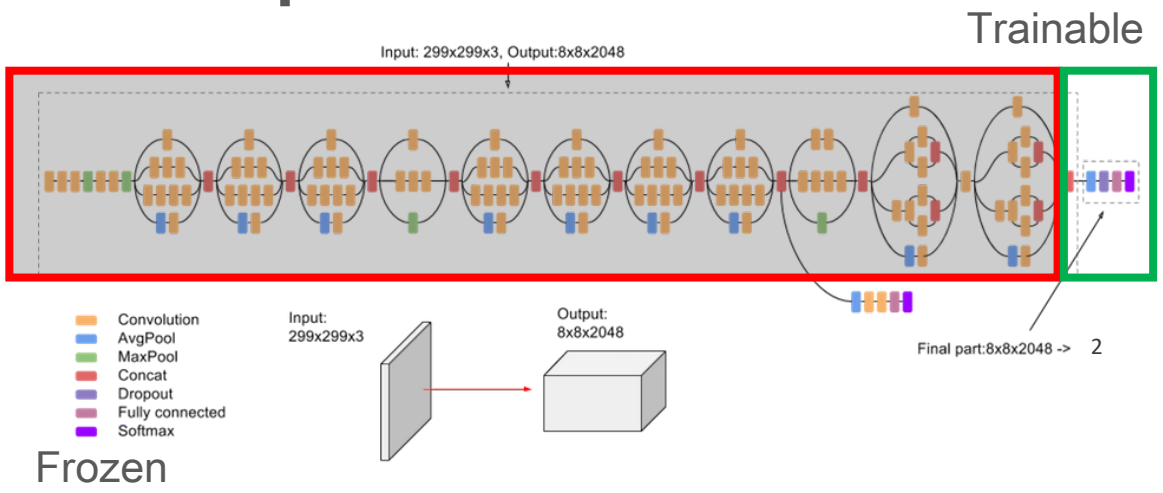




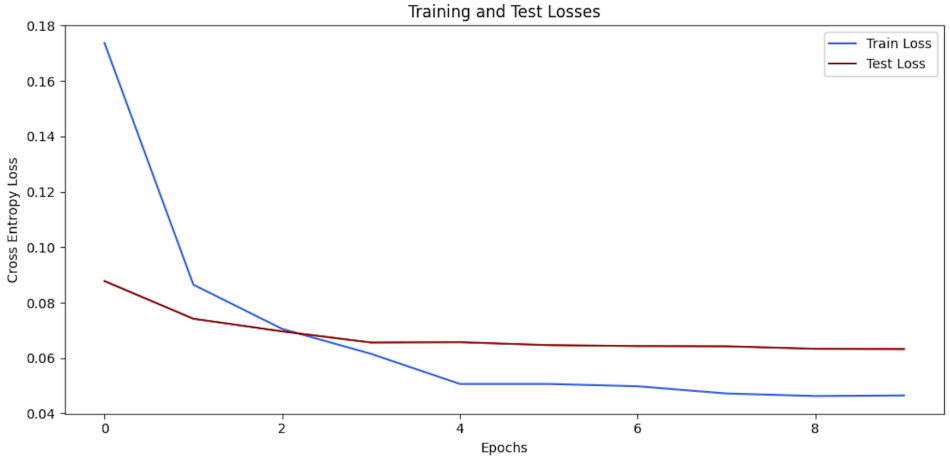
Train Accuracy: 98.1 %
Val. Accuracy: 97.9%

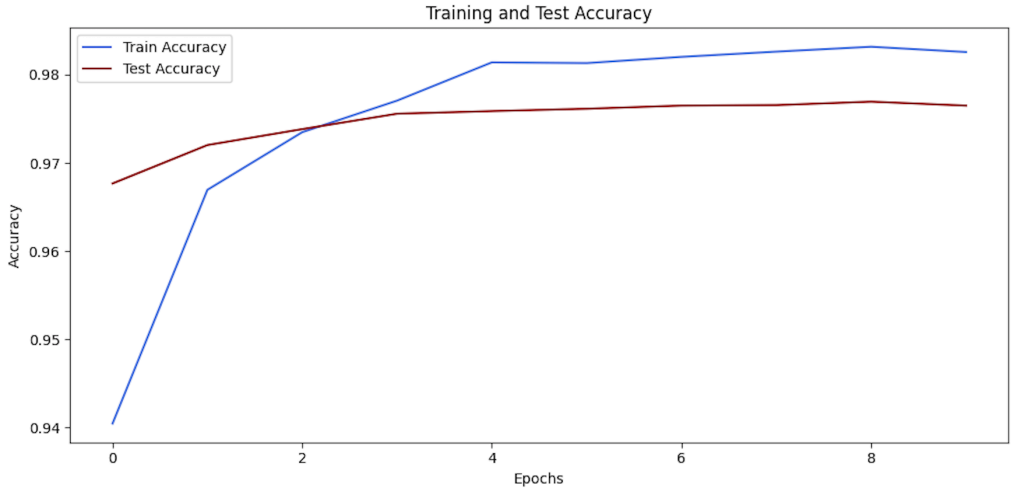
Train Loss: 0.058
Val Loss: 0.054

Inception v3



Source: Inception Module (source: original paper)





Hyperparameters:
Batch size = 64
LR = 0.001
Epochs = 10

Train Accuracy: 98.3%
Val. Accuracy: 97.5%

Train Loss: 0.051
Val Loss: 0.072

INTERPRETABILITY

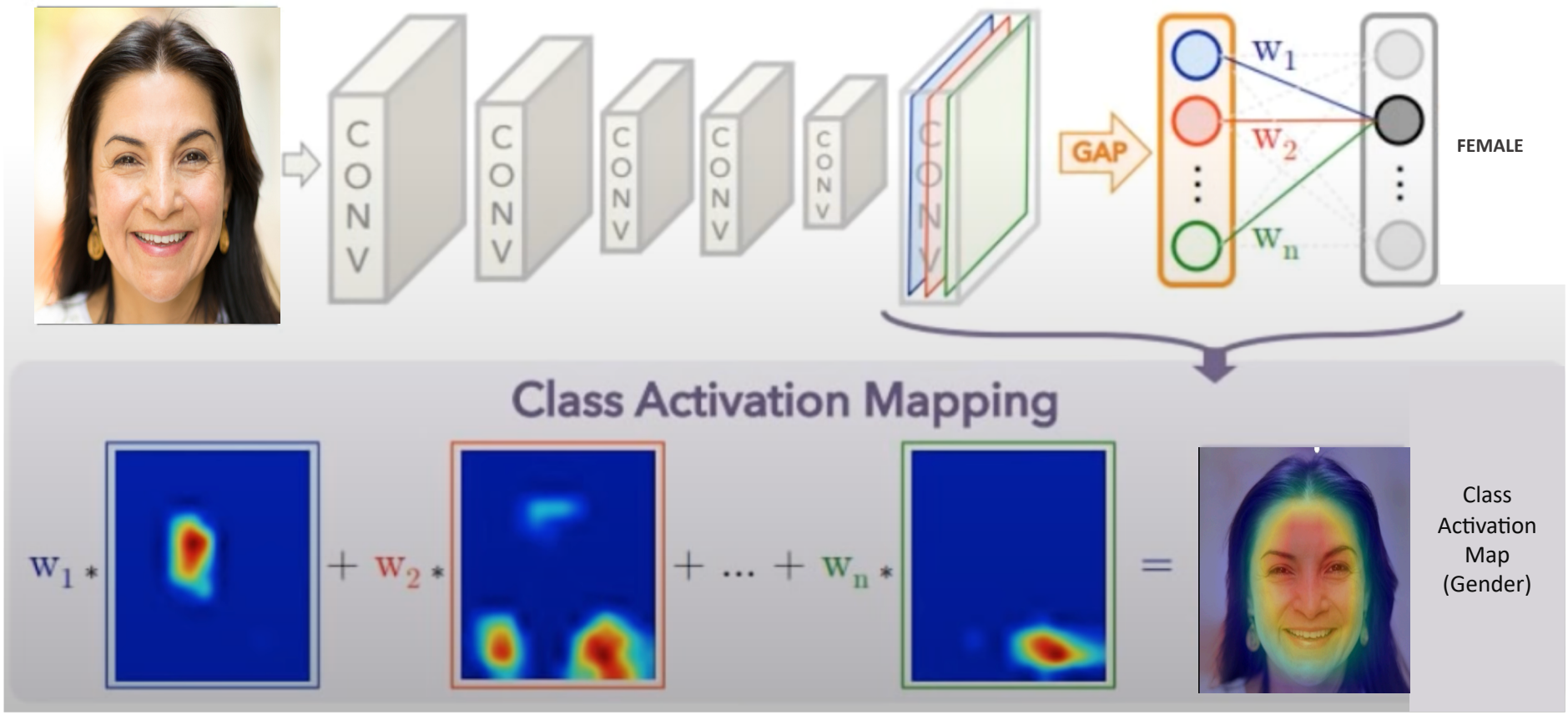
MODELS

Class Activation Maps

(CAM)

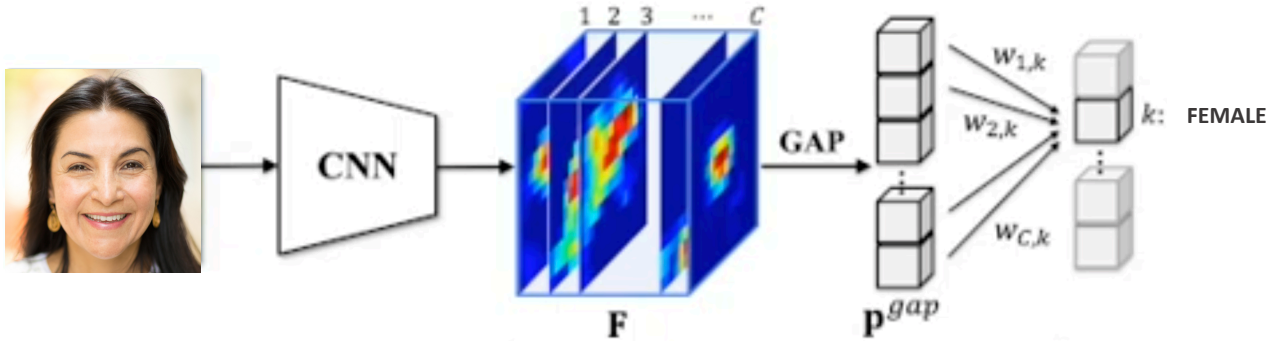
CLASS ACTIVATION MAPS

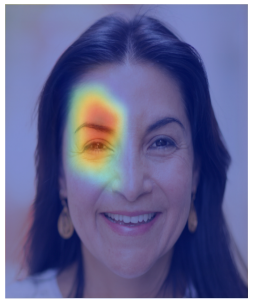
- The image shrinks, the number of features increases.
- ReLU: All features are positive or zero.



CLASS ACTIVATION MAPS

- We consider one class at a time (the predicted class)
- **Example**: W = w[:, human_face_index] # Size 512
- CAM = F[0] * w[0] + F[1] * w[1] + … + F[512] * w[512] (Weighted sum of images)
- Rescale the output image to the original size (i.e. 224 x 224 for ResNet), and the 2 images over each other

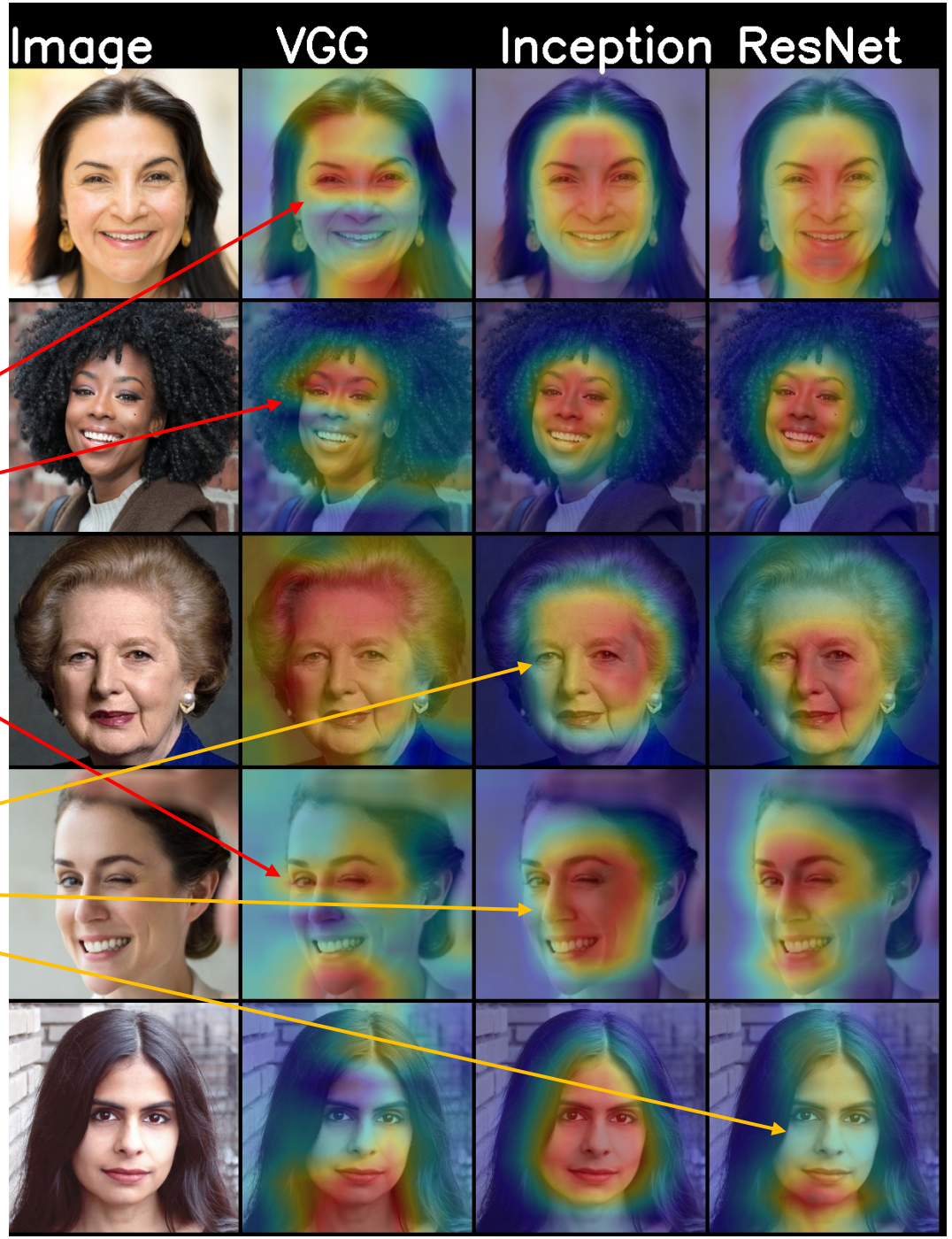




CAM RESULTS FOR FEMALE GENDER

VGG16 focus more on specific features. i.e. eyes, Chin

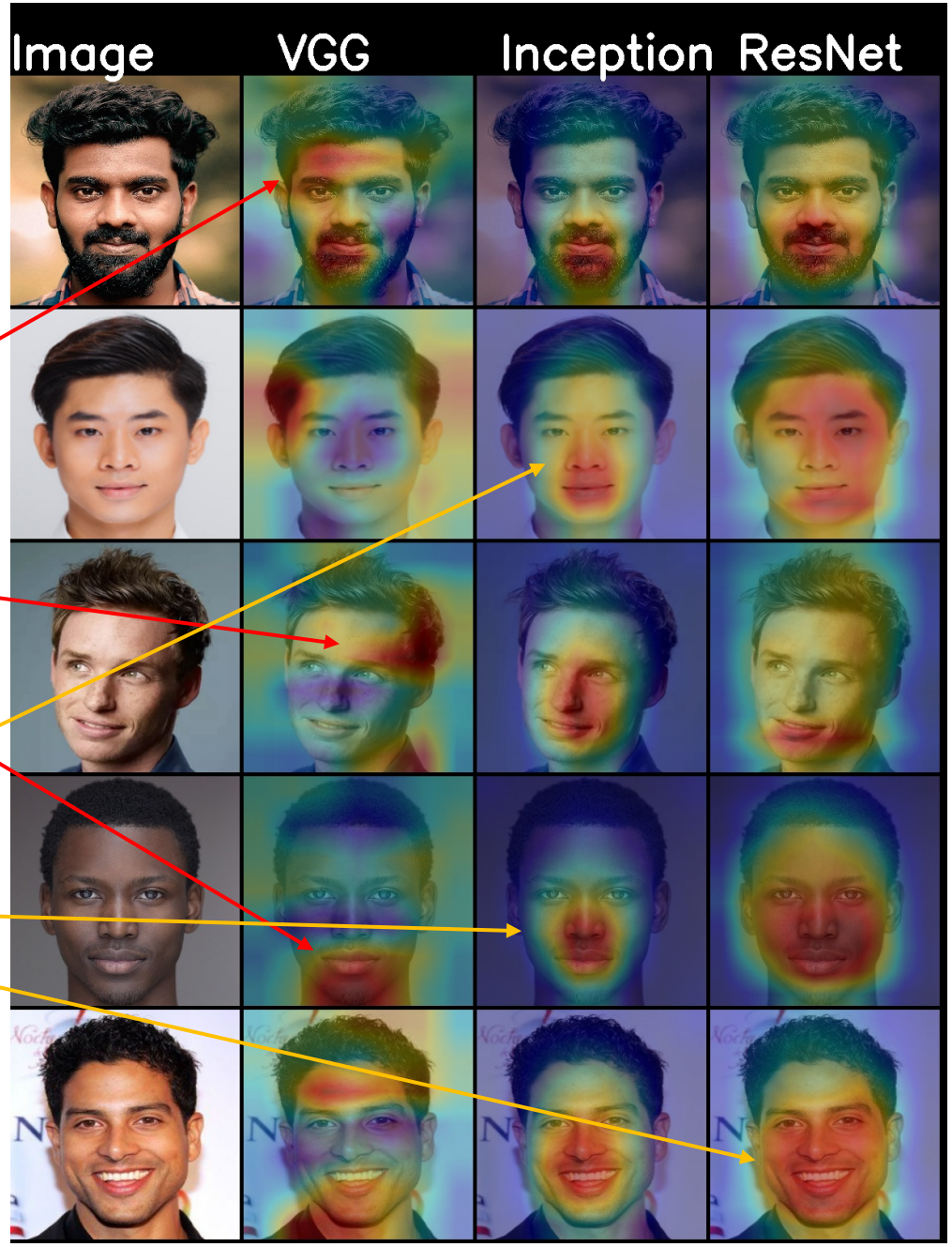
Inception and Resnet Focus on the overall face



CAM RESULTS FOR MALE GENDER

VGG16 focus more on specific features. i.e. eyes, chin.

Inception and Resnet Focus on the overall face



SHAP

SHAP (SHapley Additive exPlanations) is a method to explain individual predictions.

SHAP is based on the game theoretically optimal Shapley Values.

Lundberg and Lee (2016)

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

Source: https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

SHAP

- SHAP is based on Game Theory.
- The game reproduces the outcome of the model.
- The players are the features in the model.
- One game → One observation.
- SHAP quantifies the contribution that each feature brings to the prediction made by the model.

Adam

Prediction 75%

Math. A → +25%

SAT verbal . 550 → -10%

AP chem . TRUE → +20%

AP Eng . FALSE → 0%

Base Rate - 40%

Ben

Prediction 75%

Math . B- → -5%

SAT verbal . 780 → +25%

AP chem . FALSE → 0%

AP Eng . TRUE → +15%

Base Rate - 40%

Source: https://www.datarobot.com/webinars/fast-path-ai/

SHAP

THIS PROJECT WILL USE THE GRADIENT-BASED EXPLANATION METHOD

$$\mathsf{IG}_i(x, x') ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

(here $\frac{\partial F(x)}{\partial x_i}$ *is the gradient of* F *along the* i^{th} *Dimension at x).*

- Integrated Gradient (IG) approximate the Shapley values.
- IG is faster Shapley-value-based method.
- Computes the gradients of the model output on a few different inputs (typically 50).

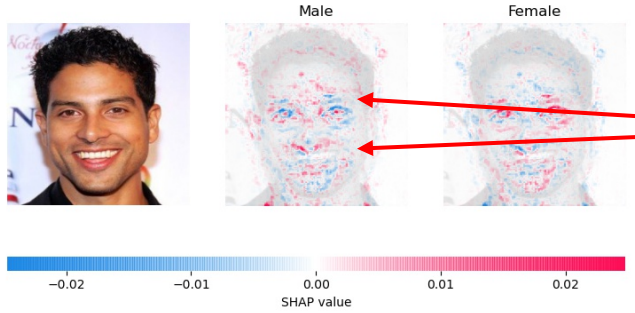
SHAP RESULTS

LOW SHAP VALUES HIGH SHAP VALUES

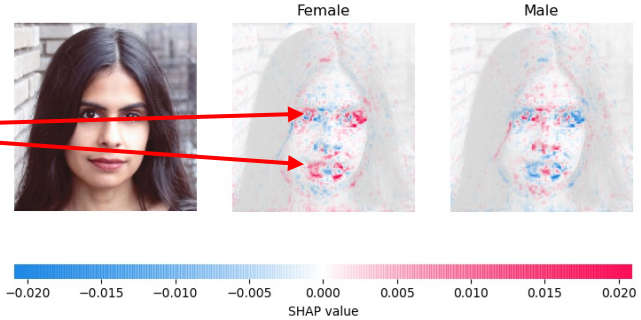
MALE

FEMALE

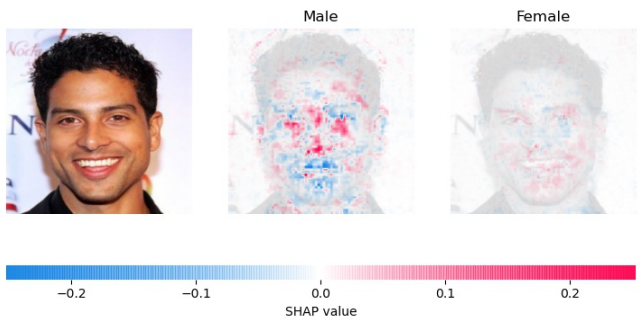
VGG-16

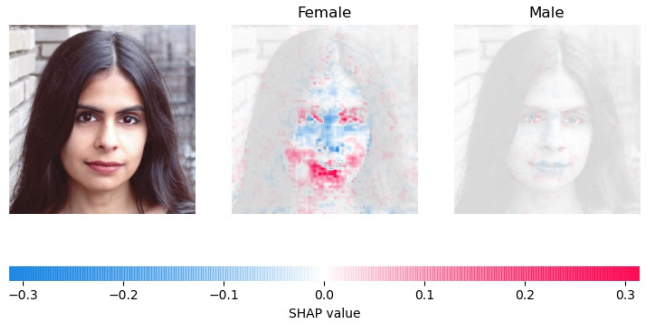


Focus on specific features

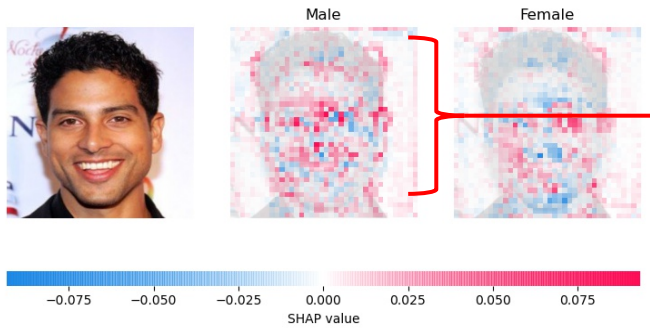


INCEPTION v3

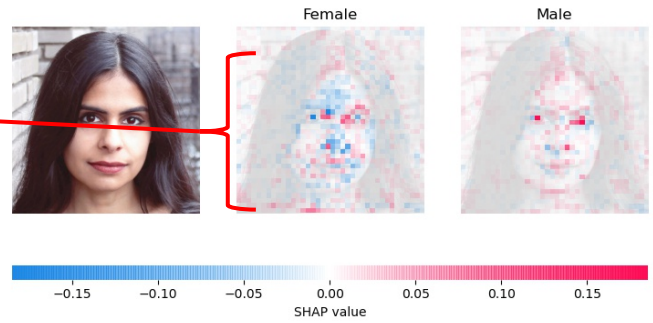




RESNET



Focus on overall image



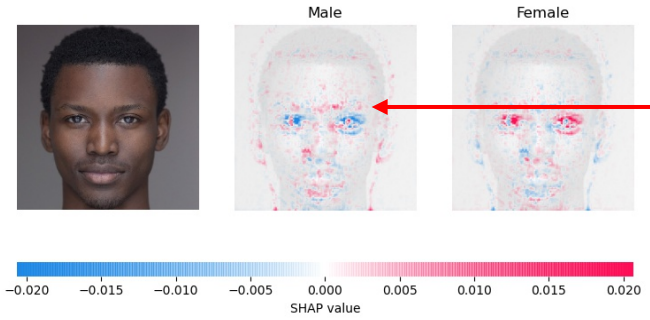
SHAP RESULTS

LOW SHAP VALUES HIGH SHAP VALUES

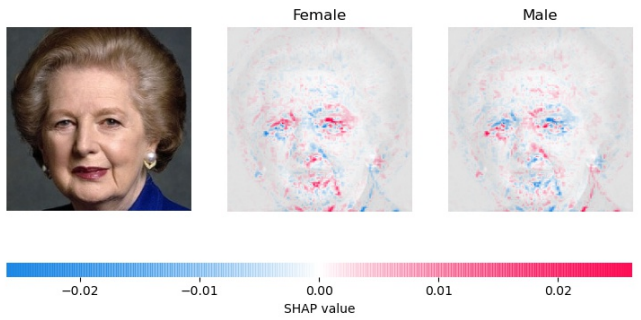
MALE

FEMALE

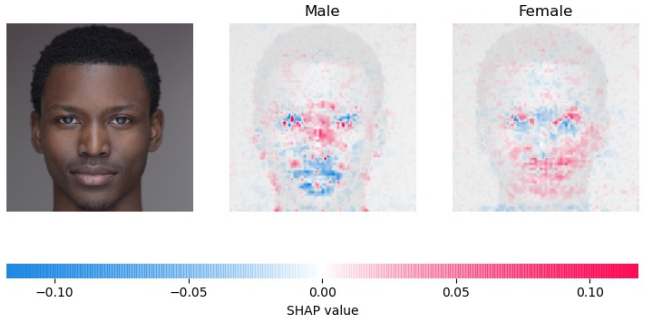
VGG-16

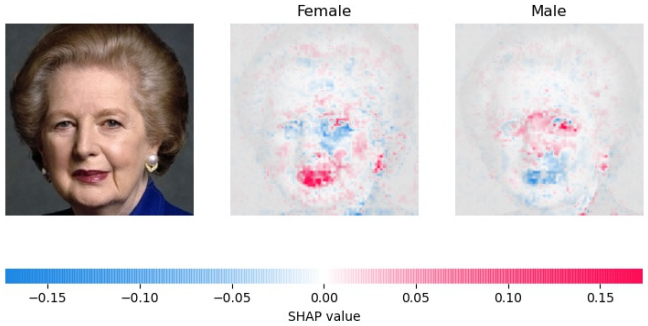


VGG16 is showing low SHAP (blue) values instead of high values (red)

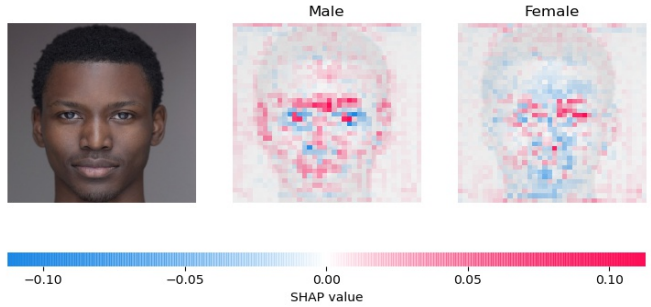


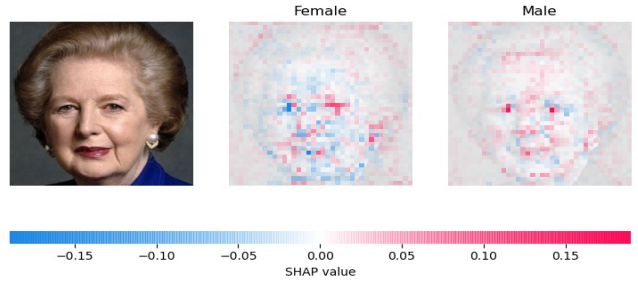
INCEPTION





RESNET

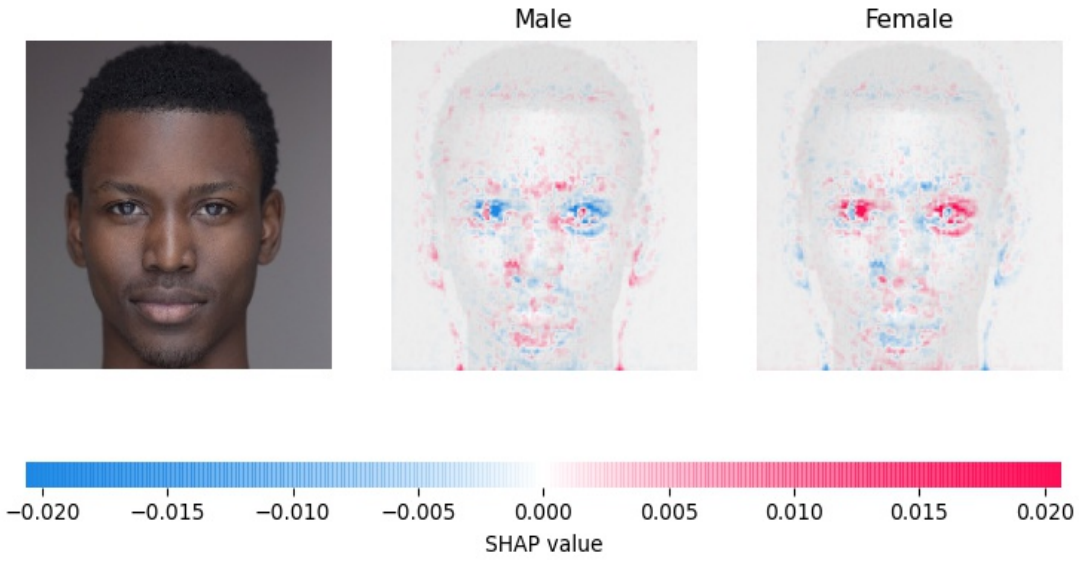


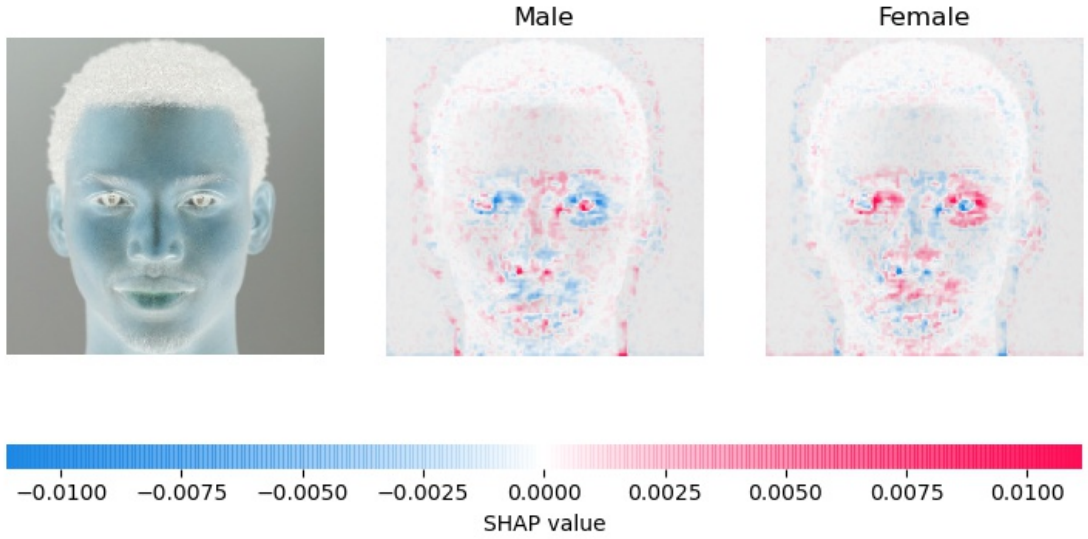


SHAP RESULTS

LOW SHAP VALUES

HIGH SHAP VALUES



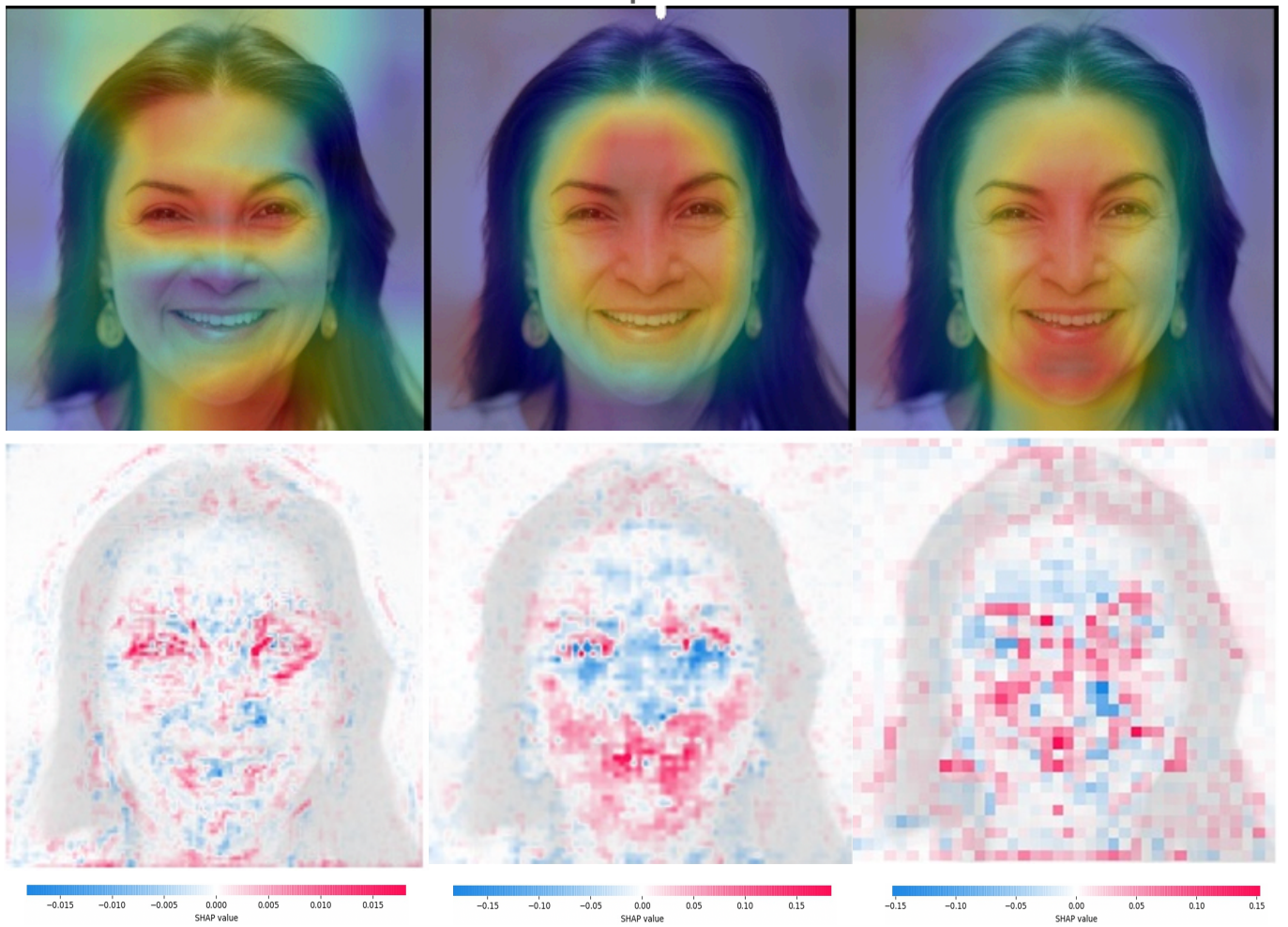


SHAP RESULTS

COMPARING CAM VS SHAP

(Gender – Female)

VGG-16 Inception v3 ResNet

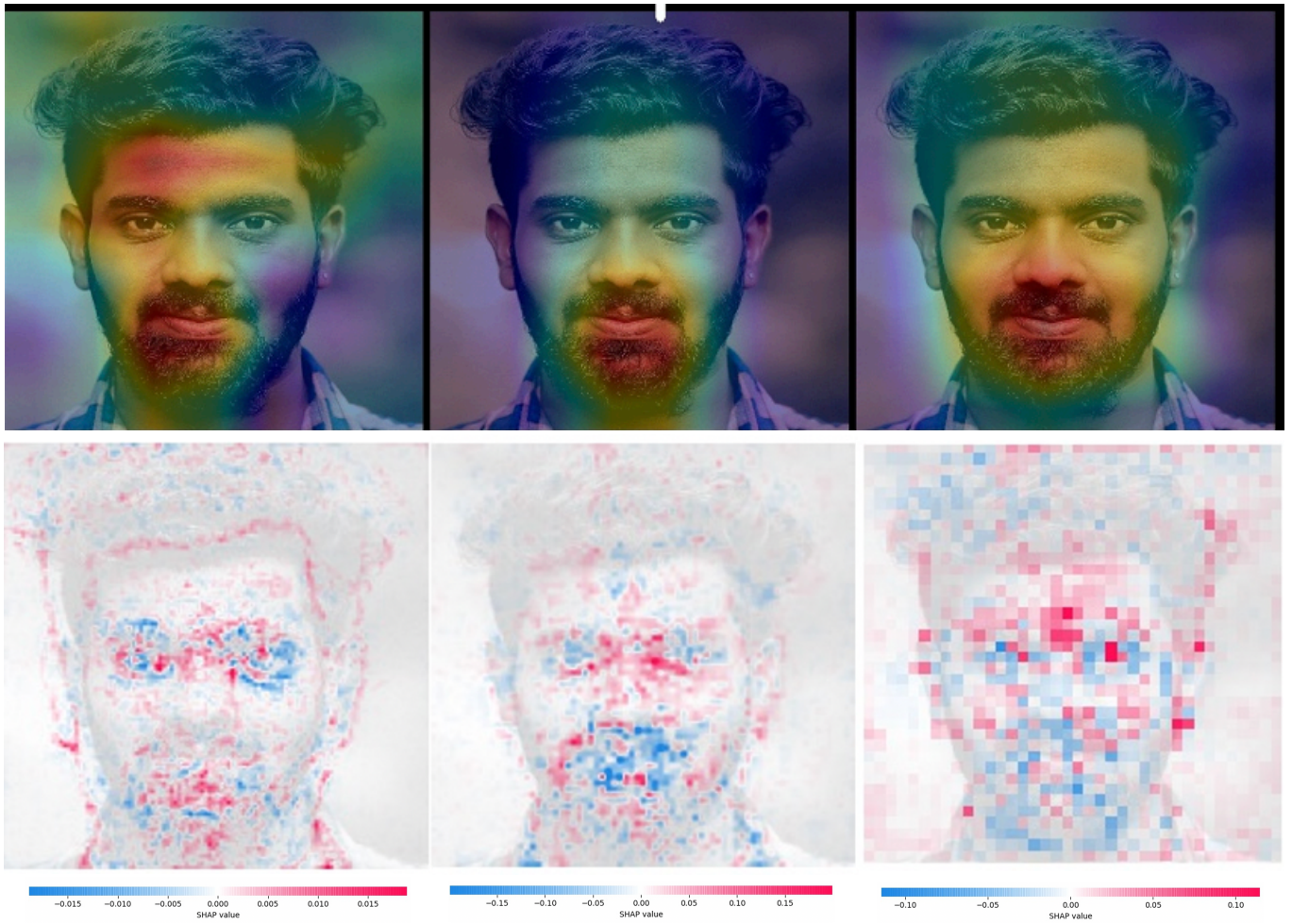


LOW SHAP VALUES HIGH SHAP VALUES

COMPARING CAM VS SHAP

(Gender – Male)

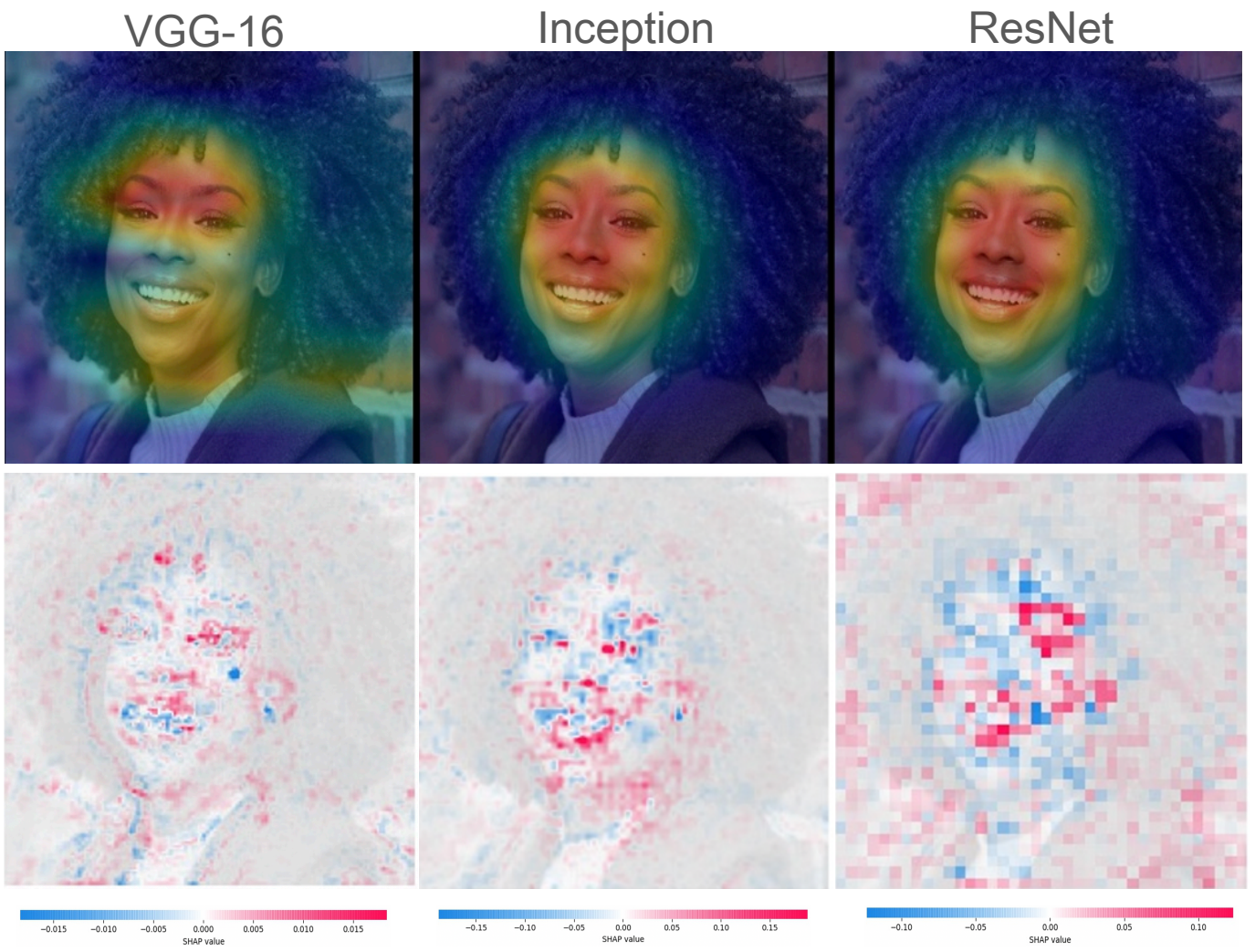
VGG-16 Inception ResNet



LOW SHAP VALUES HIGH SHAP VALUES

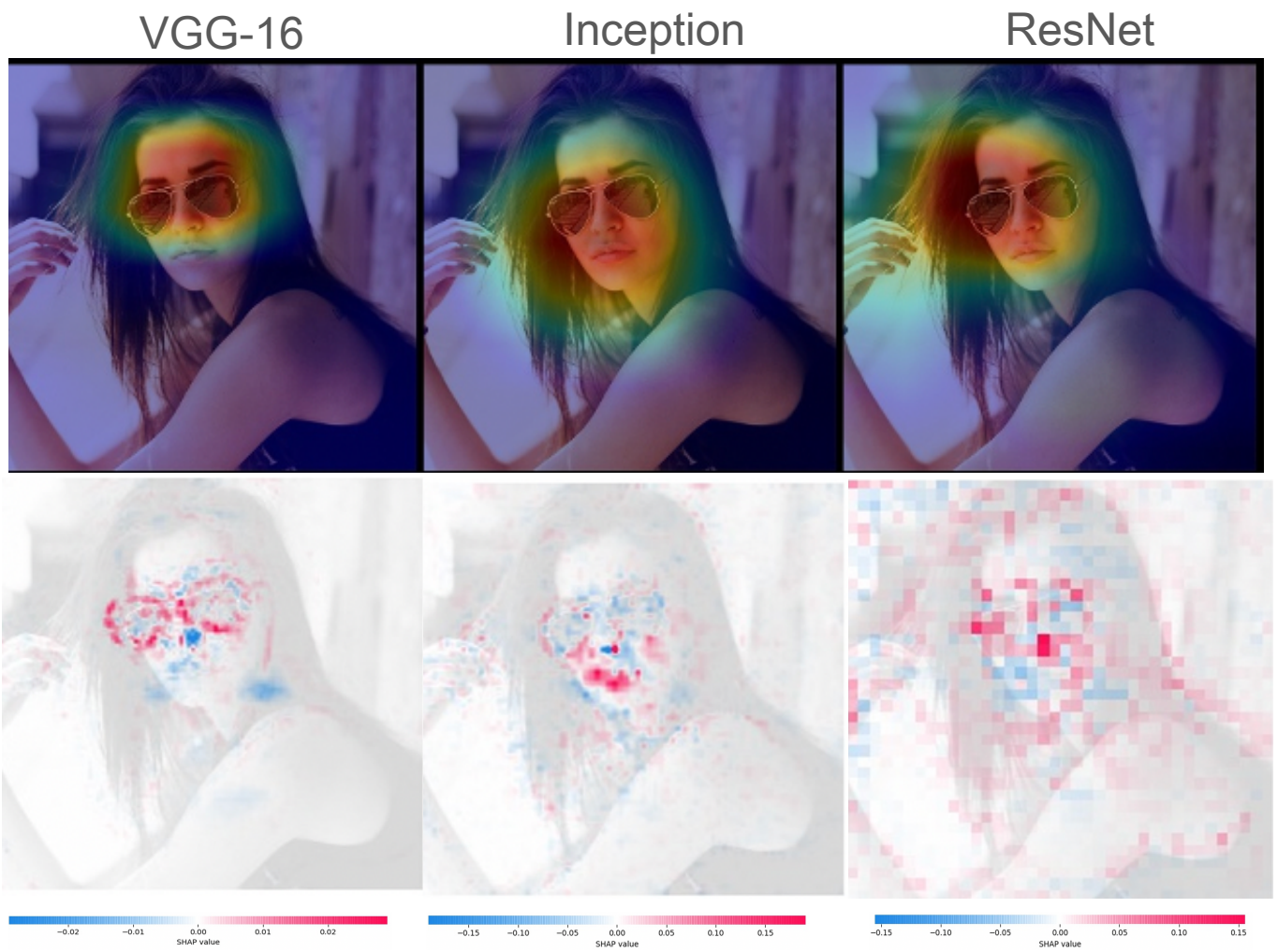
COMPARING CAM VS SHAP

(Gender – Female)



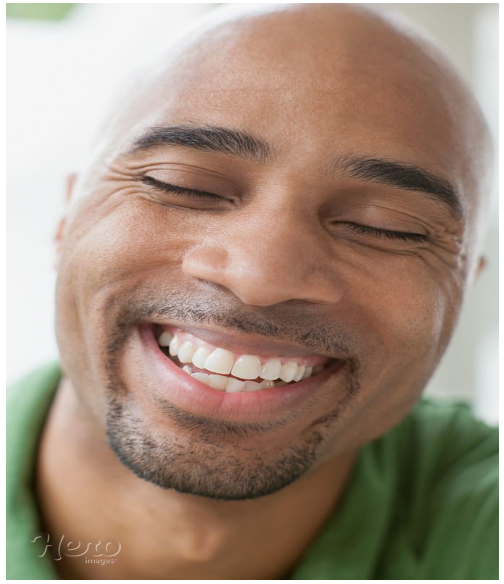
COMPARING CAM VS SHAP

(wearing sunglasses)

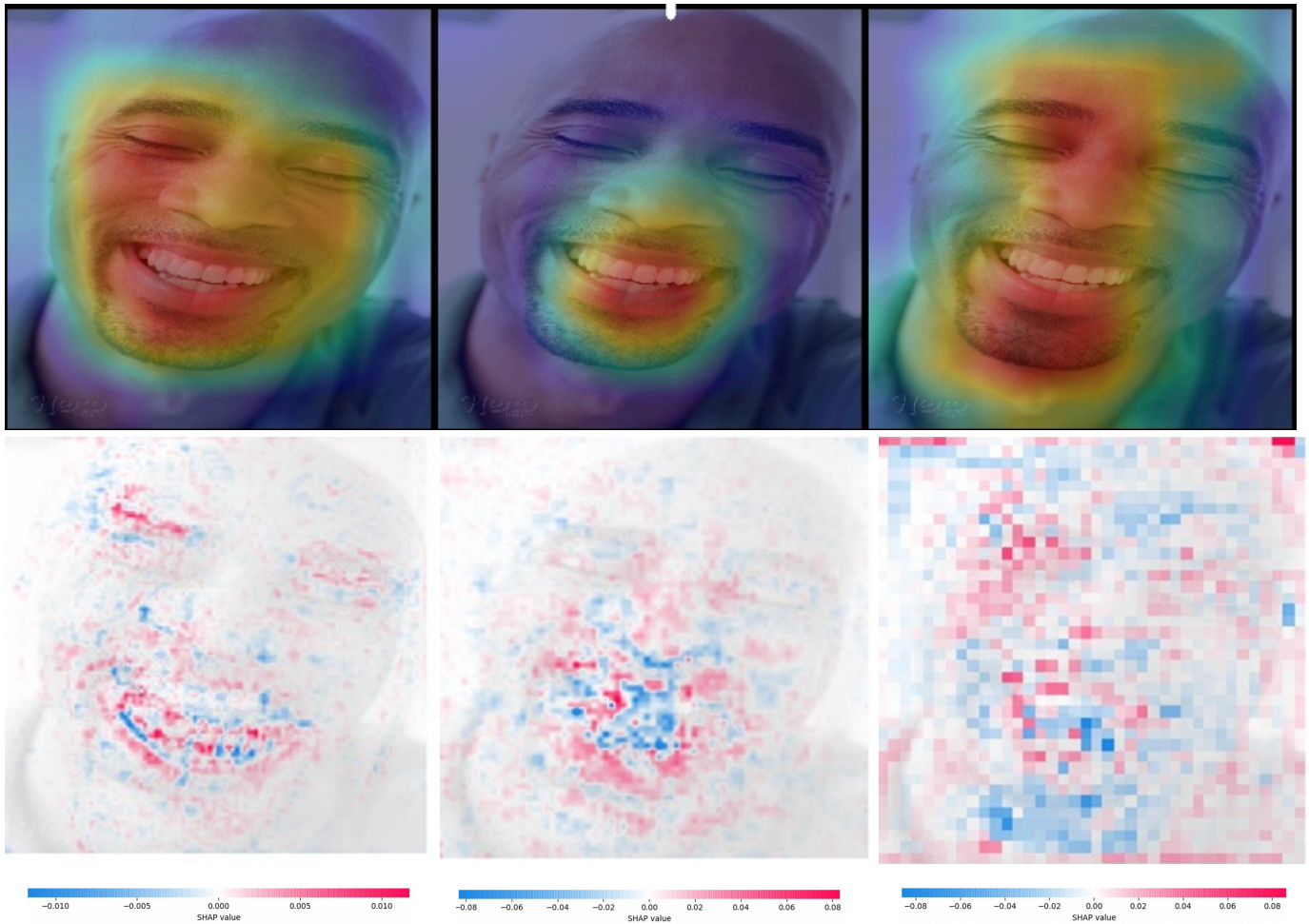


COMPARING CAM VS SHAP

(smiling)



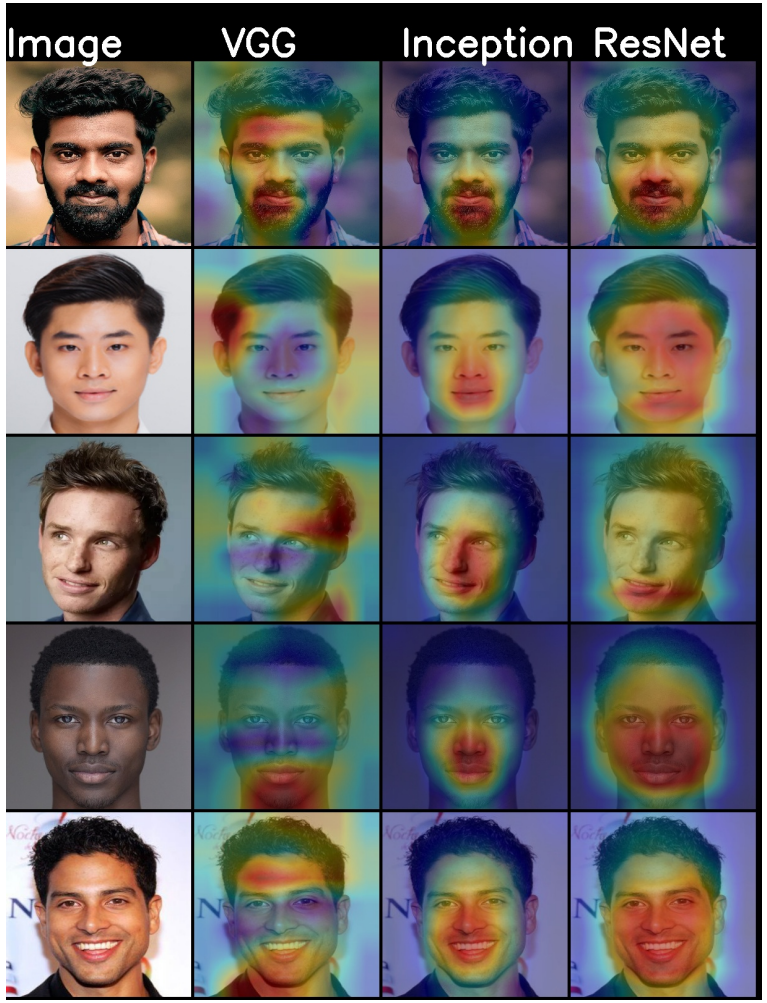
VGG-16 Inception ResNet



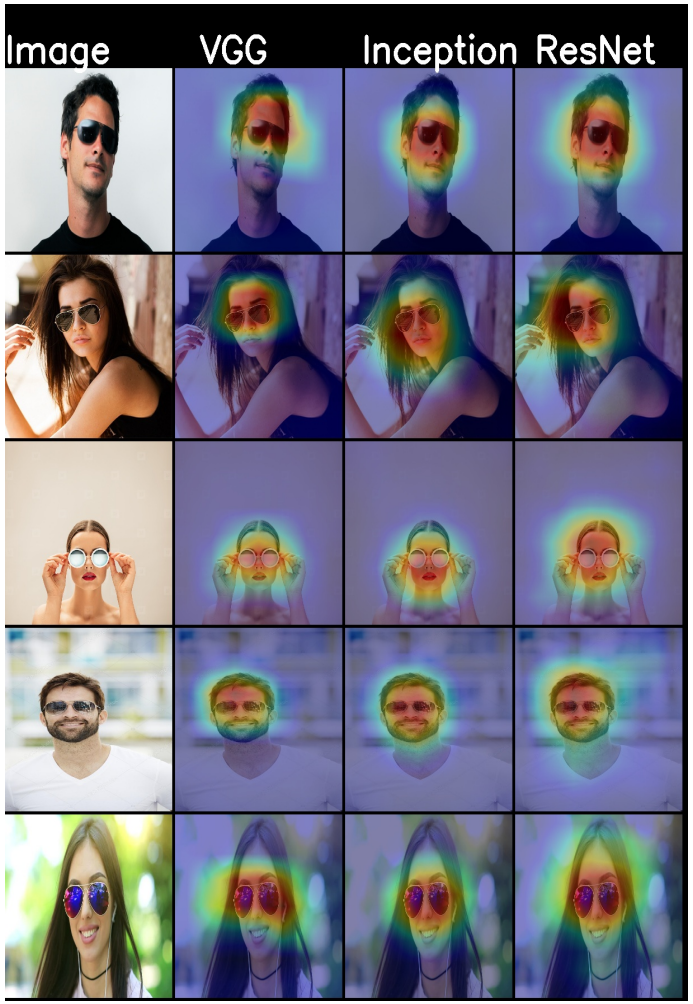
LOW SHAP VALUES HIGH SHAP VALUES

IDENTIFYING BIAS

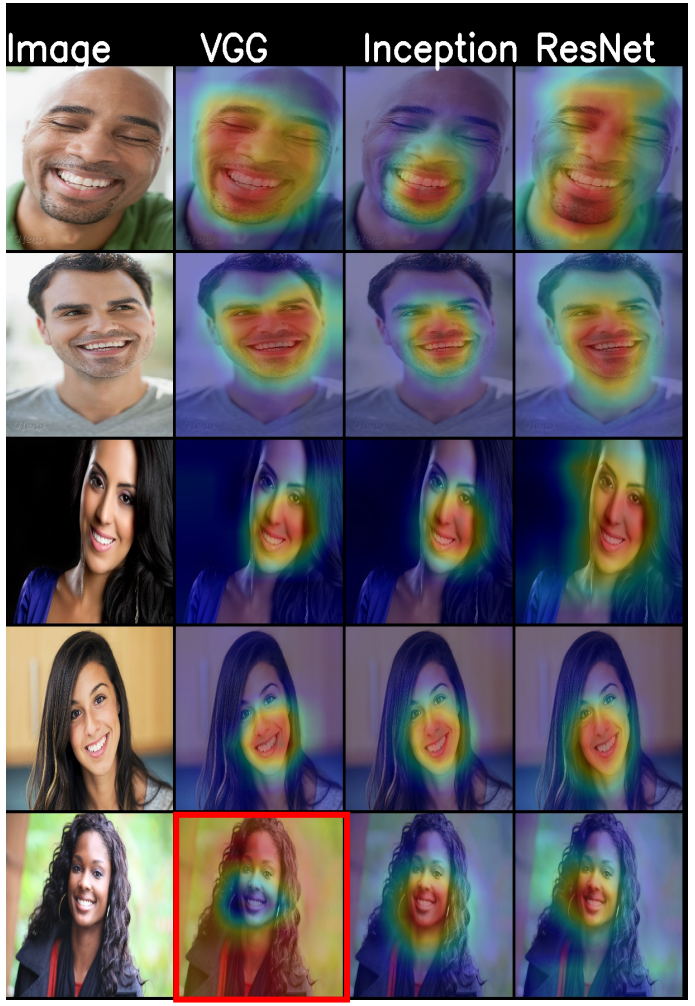
GENDER MALE



WEARING SUNGLASSES

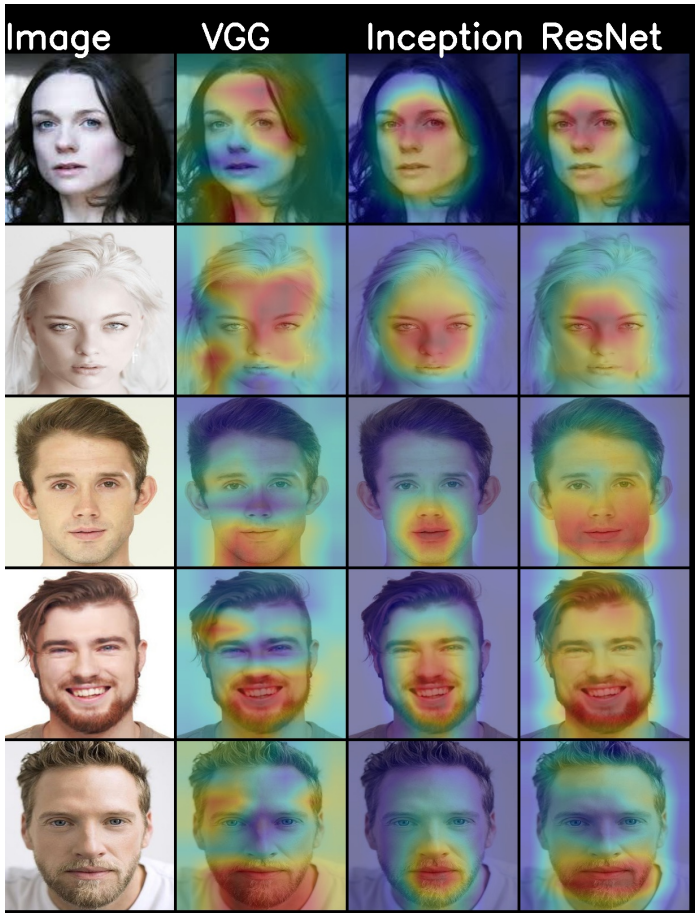


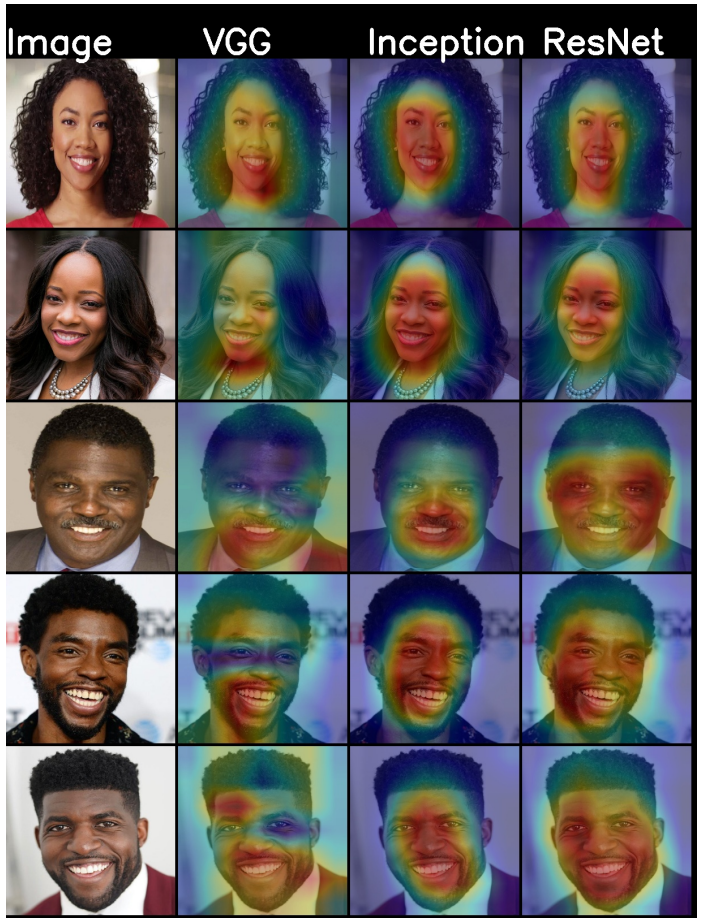
SMILING



The most important feature,
turns out to be the less important

IDENTIFYING BIAS





HARD TO CONCLUDE THAT THE MODELS ARE BIASED

CONCLUSIONS

- CAM can help in visualizing the predicted class, highlighting the features detected by the CNN.
- Interpretability can give certainty to complex models to make correct predictions in an ethical way.
- It can help to spot bias in the model.
- Can Help you explain your model.
- In some countries, regulation can make it mandatory.

REFERENCES

- Zhou, B., Khosla, A., Lapedriza, A., & Oliva, A. (2016). Learning Deep Features for Discriminative Localization. Retrieved from http://cnnlocalization.csail.mit.edu/Zhou_Learning_Deep_Features_CVPR_2016_paper.pdf
- Petsiuk, Vitali, et al. *RISE: Randomized Input Sampling for Explanation of Black-Box Models*, 2016.
- Ka ̈rkka ̈inen Kimmo, and Jungseock Joo. *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age*, Aug. 2019.
- https://news.mit.edu/2020/aleksander-madry-machine-learning-0308
- https://www.pewresearch.org/internet/2019/09/05/the-challenges-of-using-machine-learning-to-identify-gender-in-images/
- https://science.sciencemag.org/content/349/6245/255.full
- https://shap.readthedocs.io/en/latest/example_notebooks/gradient_explainer/Multi-input%20Gradient%20Explainer%20MNIST%20Example.html
- https://blog.fiddler.ai/2019/08/should-you-explain-your-predictions-with-shap-or-ig/
- https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30
- http://cvlab.cse.msu.edu/project-interpret-FR