A CURIOUS NEW WORLD

NEDz

General Assembly

Learn how neural nets are changing our perceptions of reality

First, a task...

First, a task...

Score these people on trustworthiness

First, a task...

Pick your top three based on perceived trustworthiness



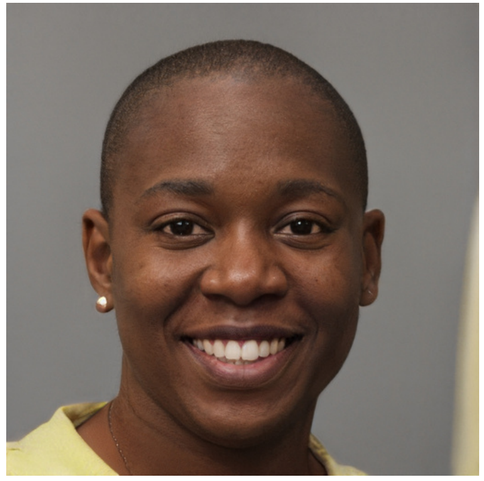





Next task...








Next task...

Find the people that aren't real...








Next task...

Find the people that aren't real...

Were they who you expected?

Next task...

Find the people that aren't real...

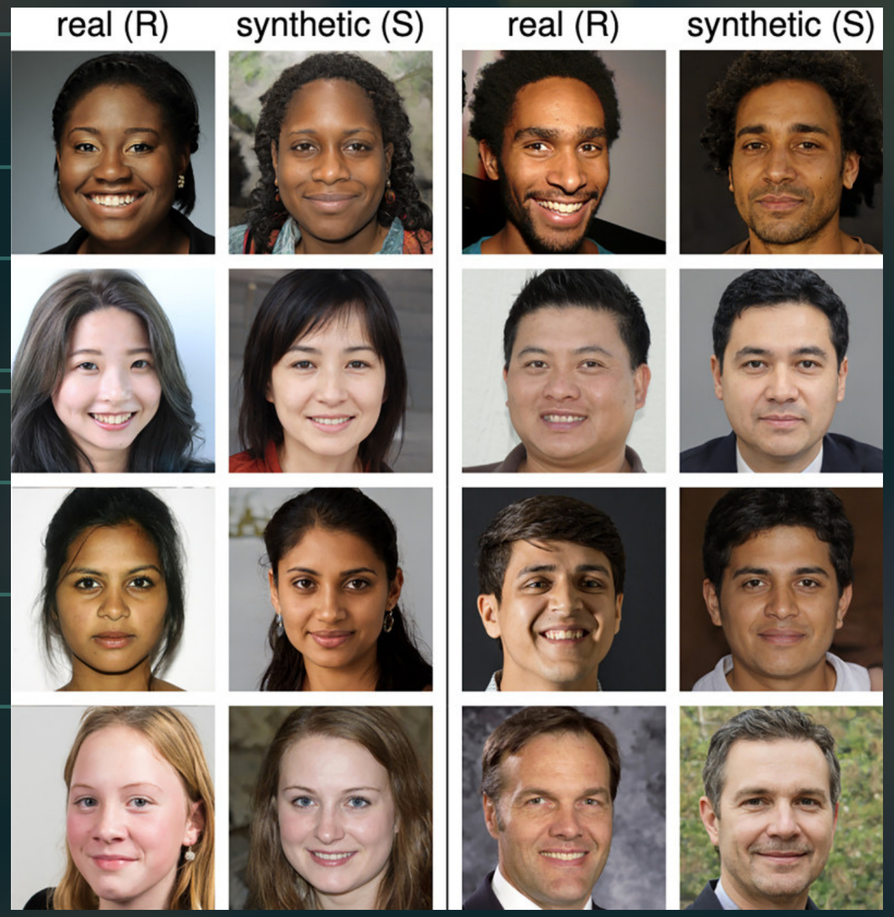
Were they who you expected?

Were they who you trusted??

The uncanny valley has been bridged...

- This was a small (and rather unscientific) adaptation of a study by Nightingale and Farid, published February of this year
- GAN images are "not just highly photorealistic, they are nearly indistinguishable from real faces and are judged more trustworthy."

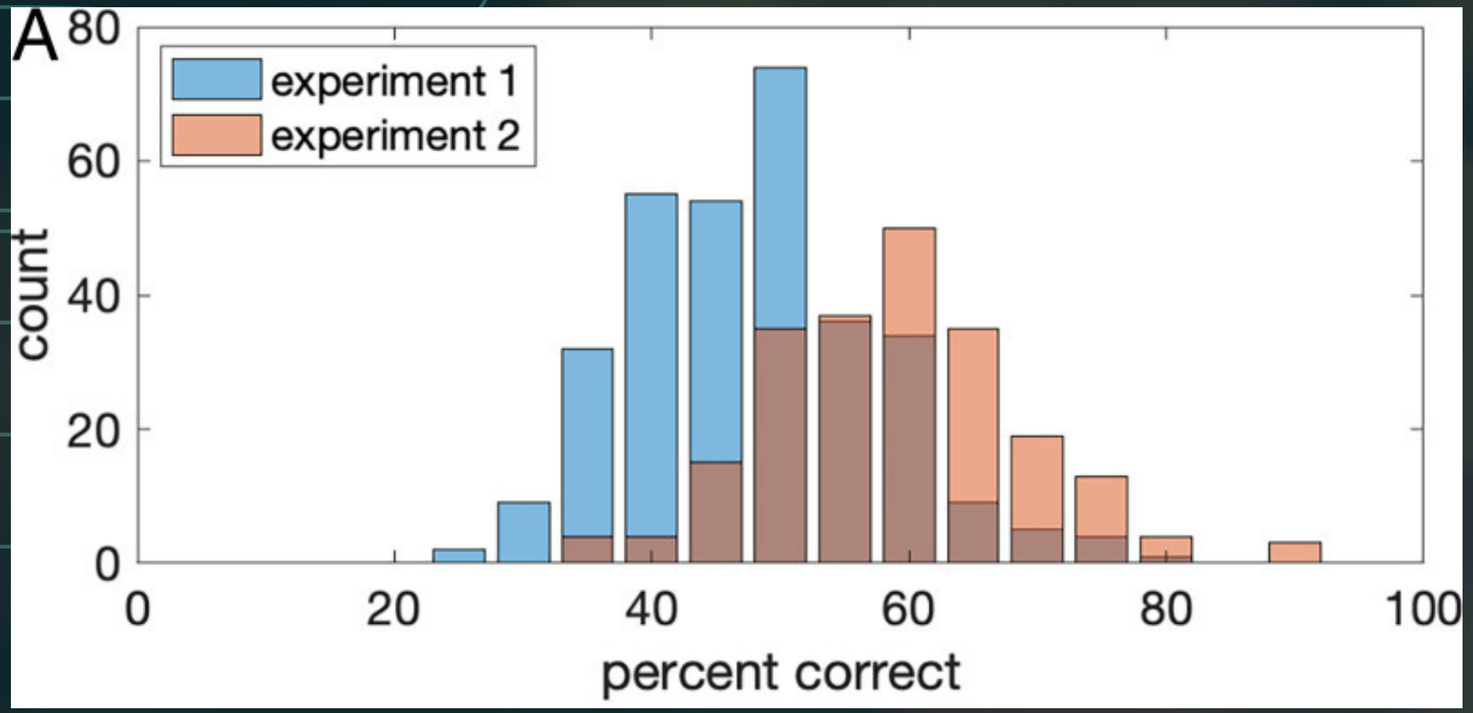
The uncanny valley has been bridged...



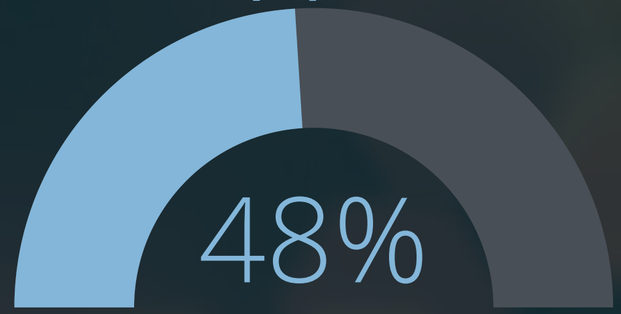
- This was a small (and rather unscientific) adaptation of a study by Nightingale and Farid, published February of this year
- GAN images are "not just highly photorealistic, they are nearly indistinguishable from real faces and are judged more trustworthy."

Even with help, participants struggled

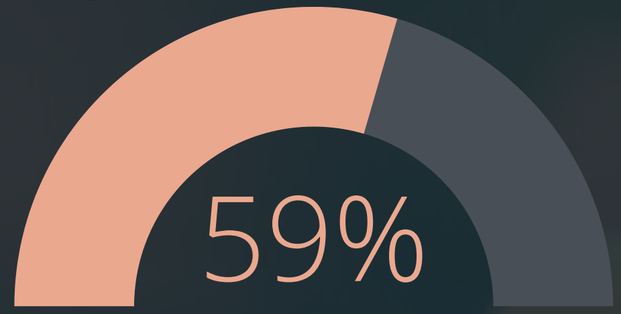
Participants presented with 128 images



315 participants
no help provided

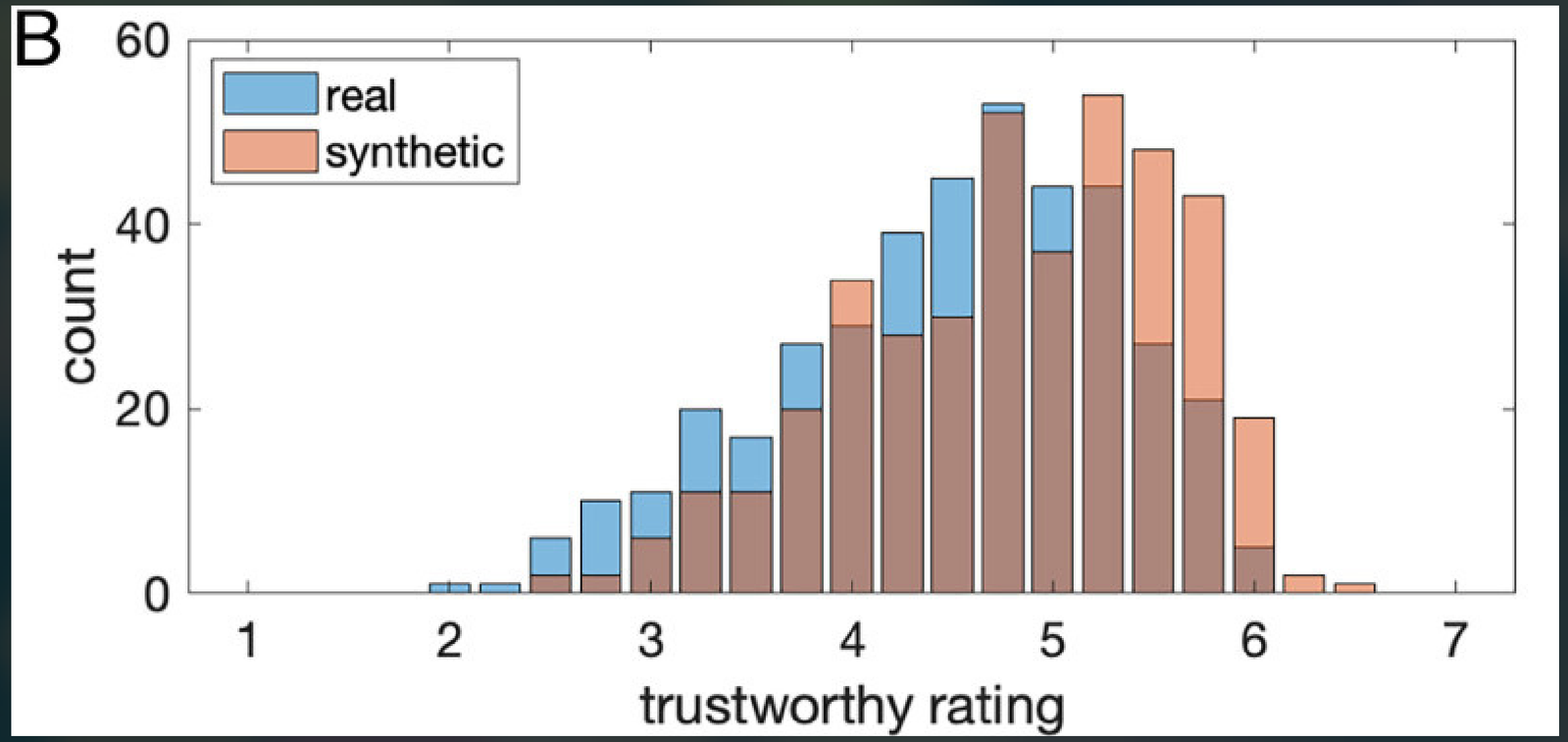


219 participants
tips and feedback



Synthetic faces are more trustworthy...

223 participants presented with 128 images



Avg. Score: Real: 4.48 Synth: 4.82

It is easier than ever to deceive and be deceived

The Question

We know distinguishing between real and fabricated images can be extraordinarily difficult for humans; how difficult is it for a machine?

Can we use this same technology in a reasonable manner to protect ourselves against those that would misuse the technology?

What are we dealing with?

a quick look at styleGAN

Architecture

Generator Vs. Discriminator

Generator Architecture:
distinguish high level attributes like pose and identity, and stochastic variation like freckles and hair, and use discovered latent variables to reproduce convincing images
Discriminator:

Train

2000s

Allowed faster communication, web browsing, and video streaming in smartphones

Deploy

Current

Significantly faster speeds with lower requirements to support IoT devices

The Project

Produce a Convolutional Neural Net model that outperforms human capability in identifying faces produced by the styleGAN system.

IS IT POSSIBLE?

Knowing that styleGAN leverages a pair of AIs competing against one another, is it really possible to create a model that can reliably identify these images?

HOW SOPHISTICATED?

Is this something that can be managed on local machines, and with limited background in neural nets?

WHAT SEPARATES REAL & FAKE?

What particular features, if any, drive the ability for a neural net to identify synthetic faces? Are we able to isolate these for better understanding?

The Process

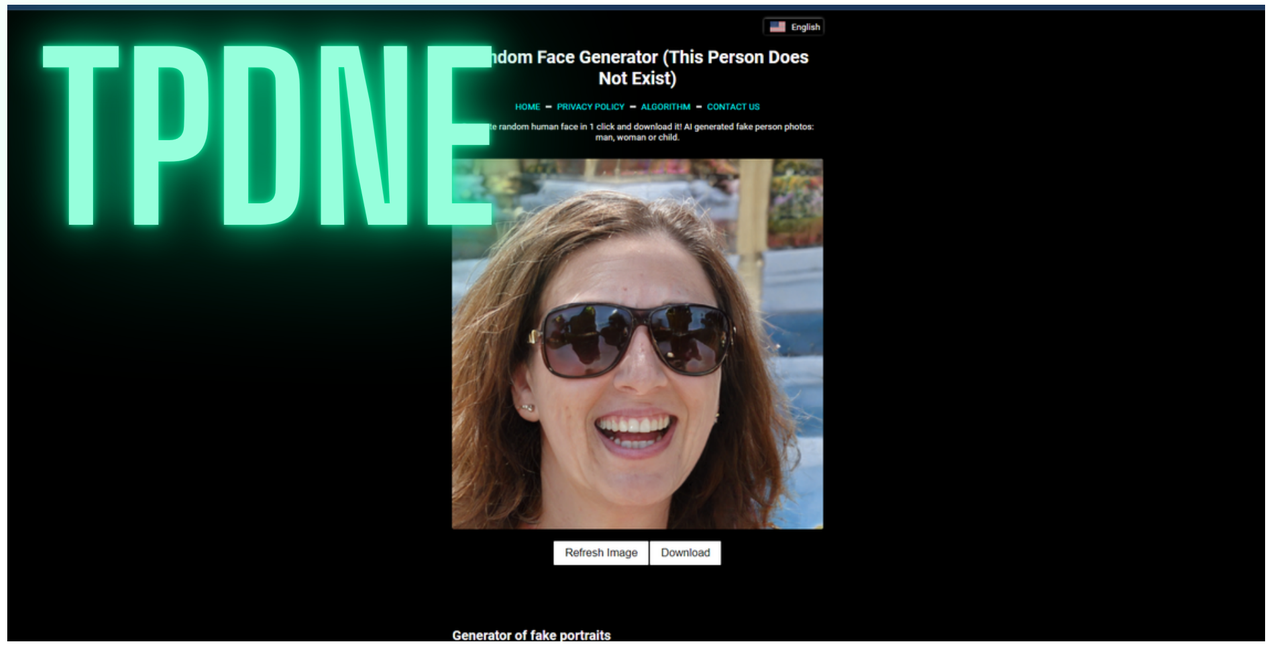
building a model

Data Sources

In this particular case, we are targeting images generated by the StyleGAN model.

Fortunately, the images used to train the StyleGAN are readily and freely available in the Flickr Faces High Quality (FFHQ) dataset, numbering over 70,000 high resolution, 1024 x 1024 pixel, images.

Synthetic images were harvested from the website: "this-person-does-not-exist.com", as it reliably generates new images using the StyleGAN model on request, with a native resolution of 1024 x 1024 pixels.



Training

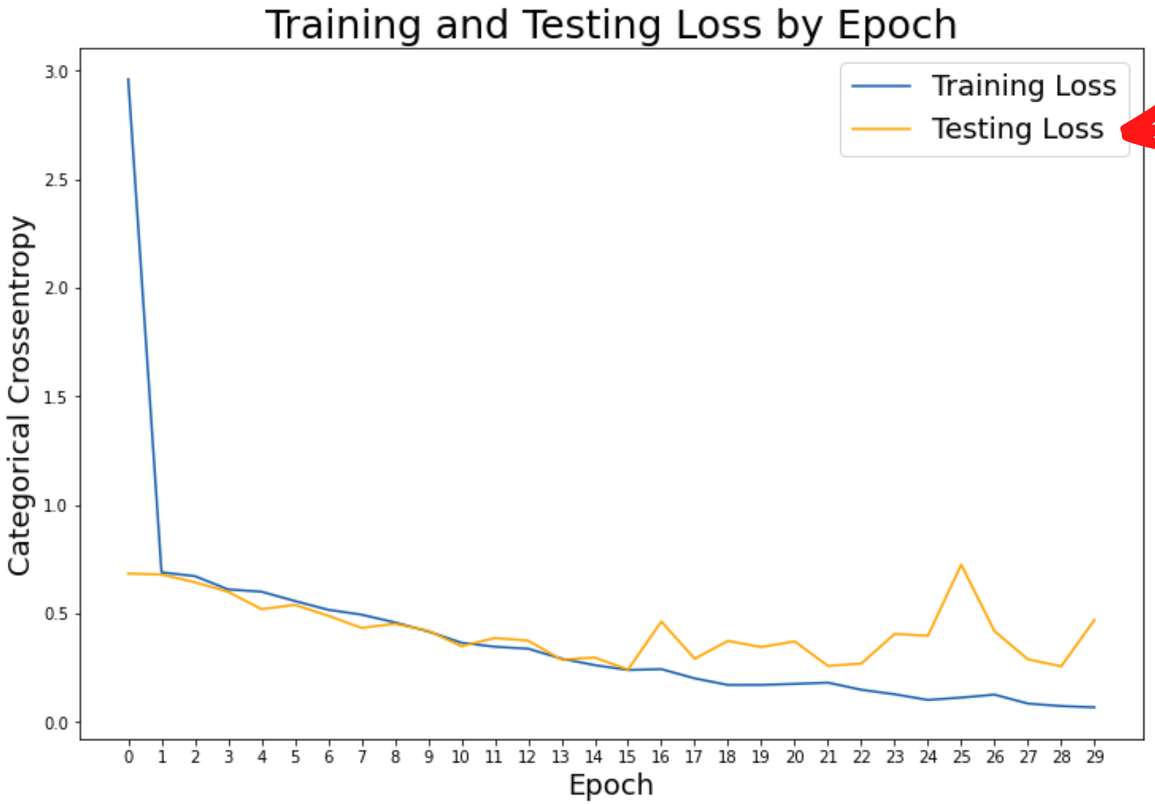

Populate Train/Validate/Test Image Groups
Preprocess Images and Create Batches
Feed Into NN Architecture

Neural Net


Two Convolutional Layers with Pooling
Two Fully Connected Layers with Dropout
Binary Classification (Fake or Real)

This simple neural net substantially outperforms humans

The model finds synthetic faces 9 times out of 10



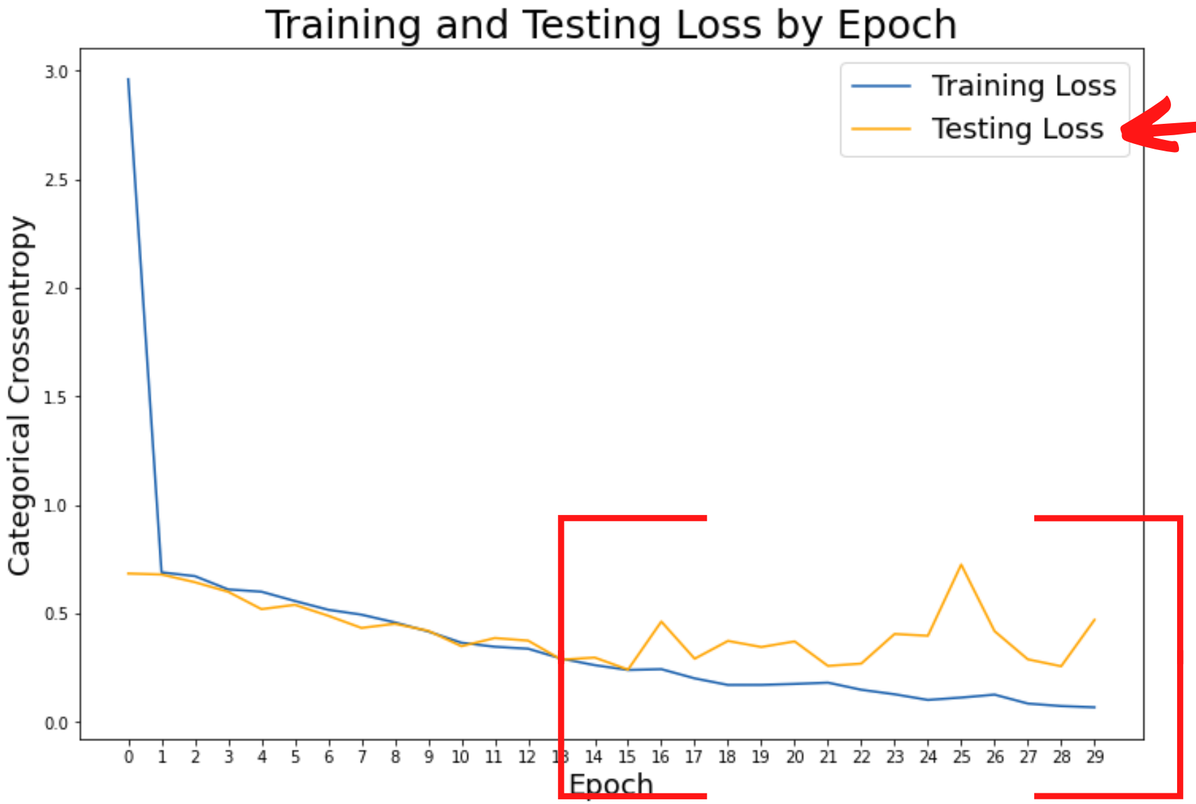
(Validation Loss)

Over final 15 epochs:

Training Accuracy: 95%

Validation Accuracy: 88%

The model finds synthetic faces 9 times out of 10



(Validation Loss)

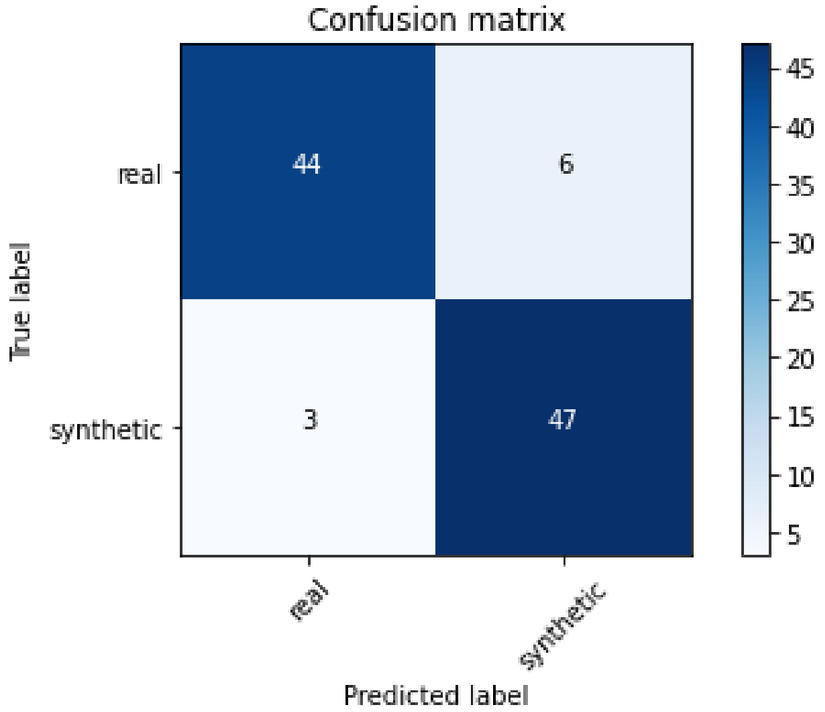
Over final 15 epochs:

Training Accuracy: 95%

Validation Accuracy: 88%

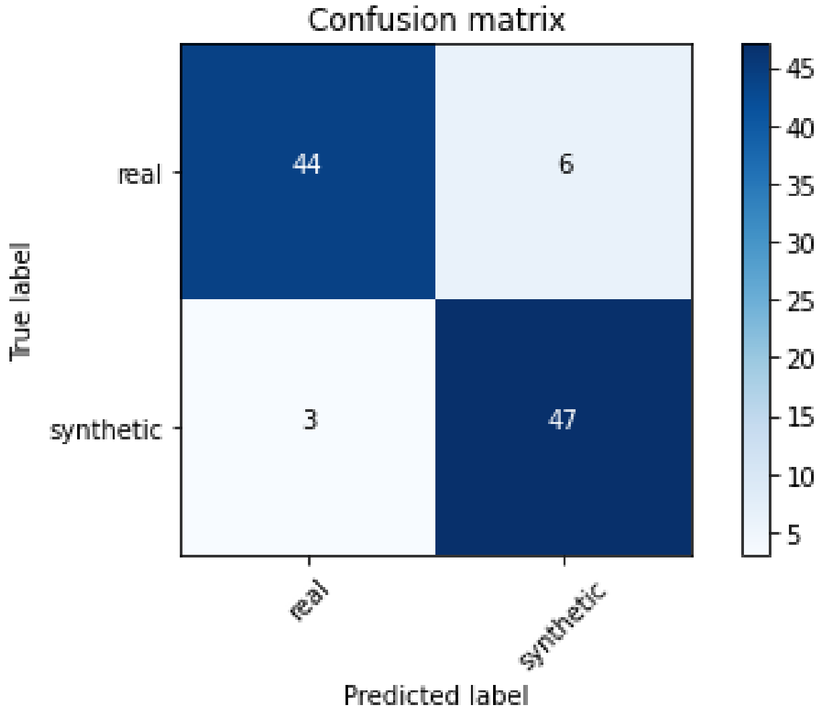
Signs of Overfitting after epoch 15

The model tended to misclassify real faces



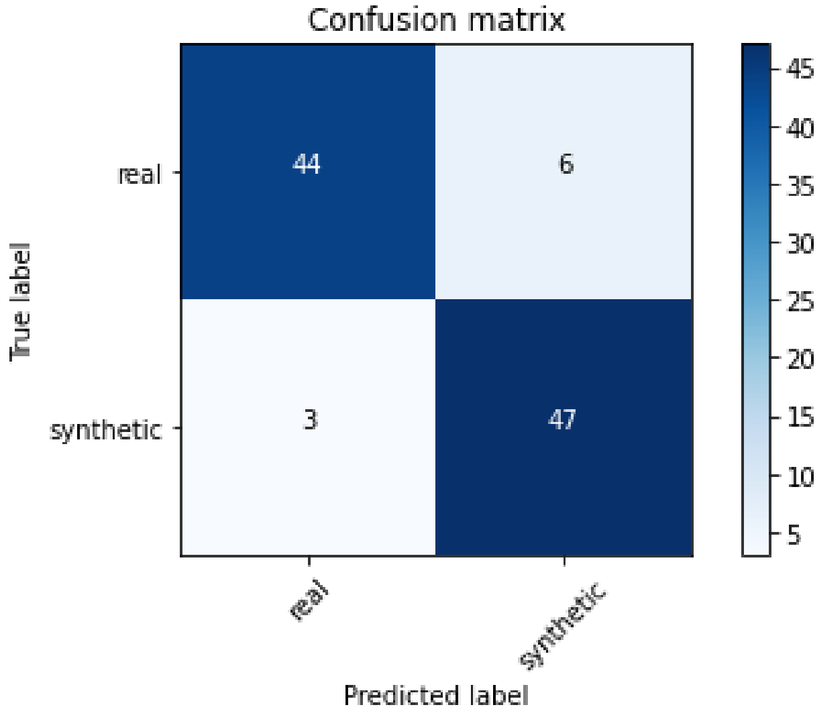
Test Accuracy: 91%

The model tended to misclassify real faces



Test Accuracy: 91%

Let's take a look at these misclassifications



Test Accuracy: 91%

Let's take a look at these misclassifications

Index: 5; False Positive

Index: 29; False Positive

Index: 51; False Negative

Index: 18; False Positive

Index: 38; False Positive

Index: 79; False Negative

Index: 27; False Positive

Index: 40; False Positive

Index: 83; False Negative

Let's take a look at these misclassifications

Index: 5; False Positive

Index: 29; False Positive

Index: 51; False Negative

Index: 18; False Positive

Index: 38; False Positive

Index: 79; False Negative

Index: 27; False Positive

Index: 40; False Positive

Index: 83; False Negative

What is it about these images that threw off the model?

It is difficult to know what exactly is going on inside a neural net, but we can at least get some hints...

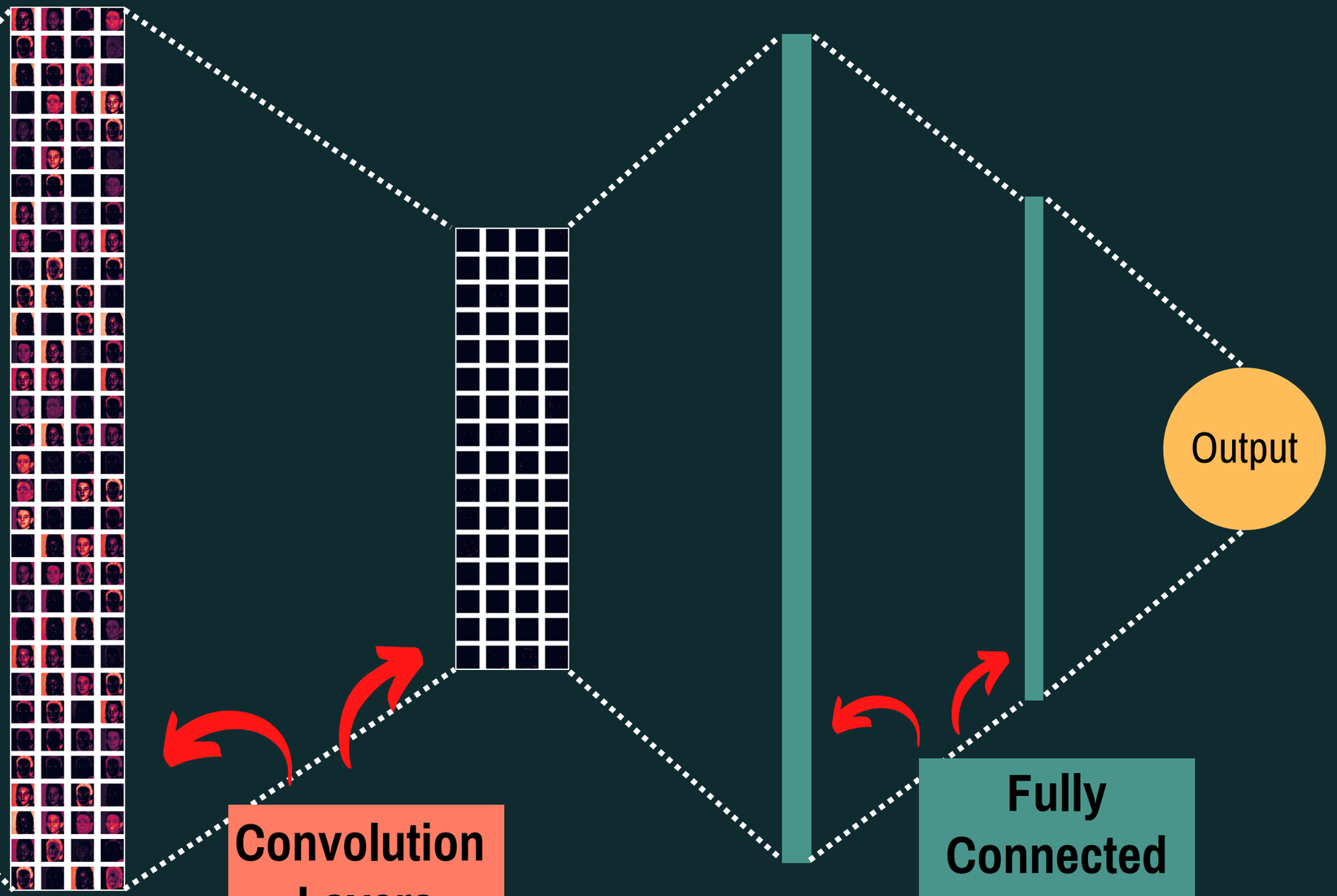
Let's start with this guy

He produces these convolutional feature maps

False Positive

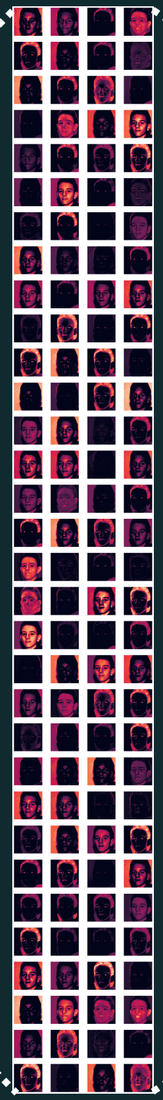
He produces these convolutional feature maps

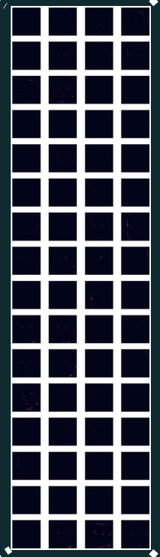
False Positive

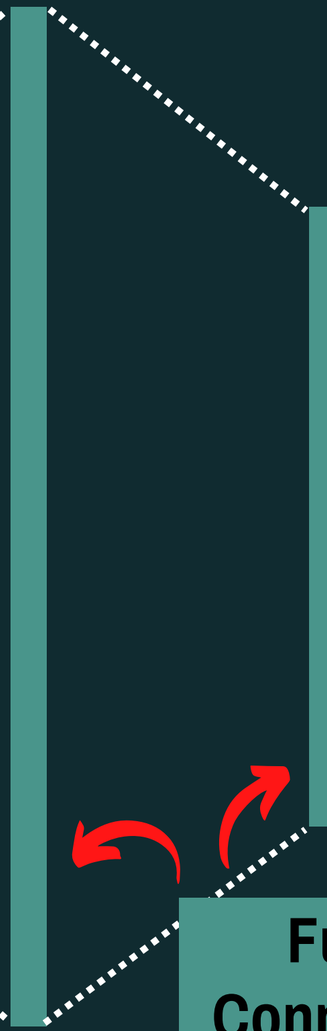


A closer look at these feature maps...

False Positive







Output

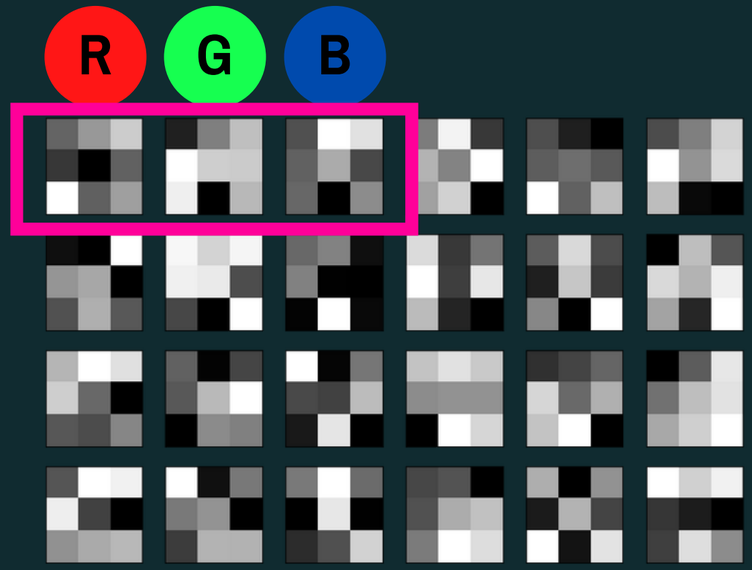
Convolution Layers

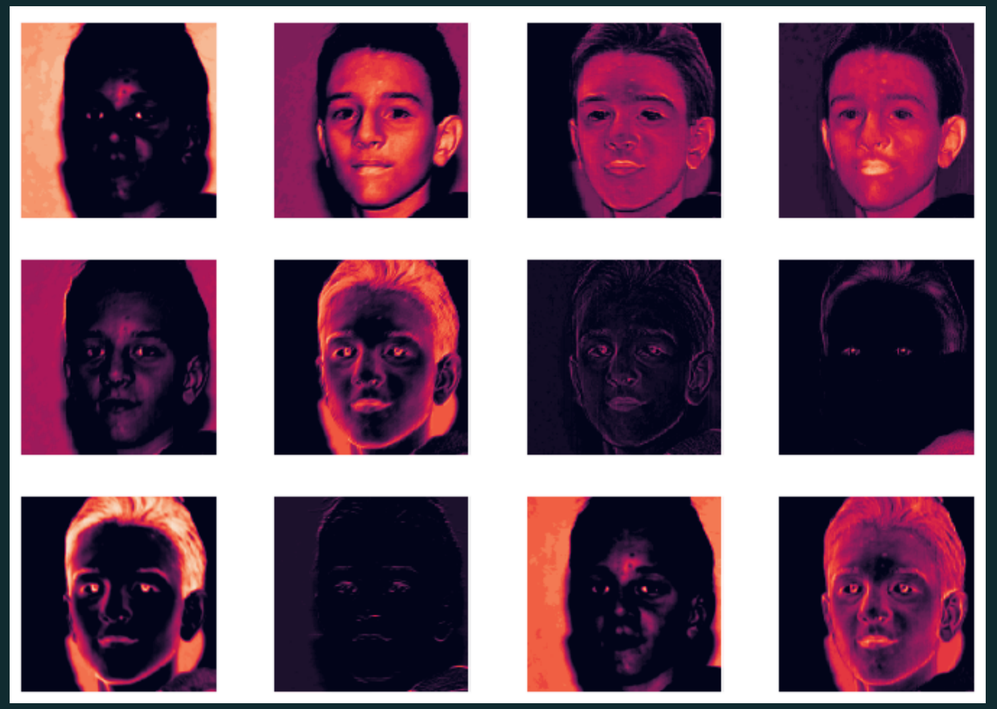
Fully Connected Layers

A closer look at these feature maps...

These are a summing up of the pixel values

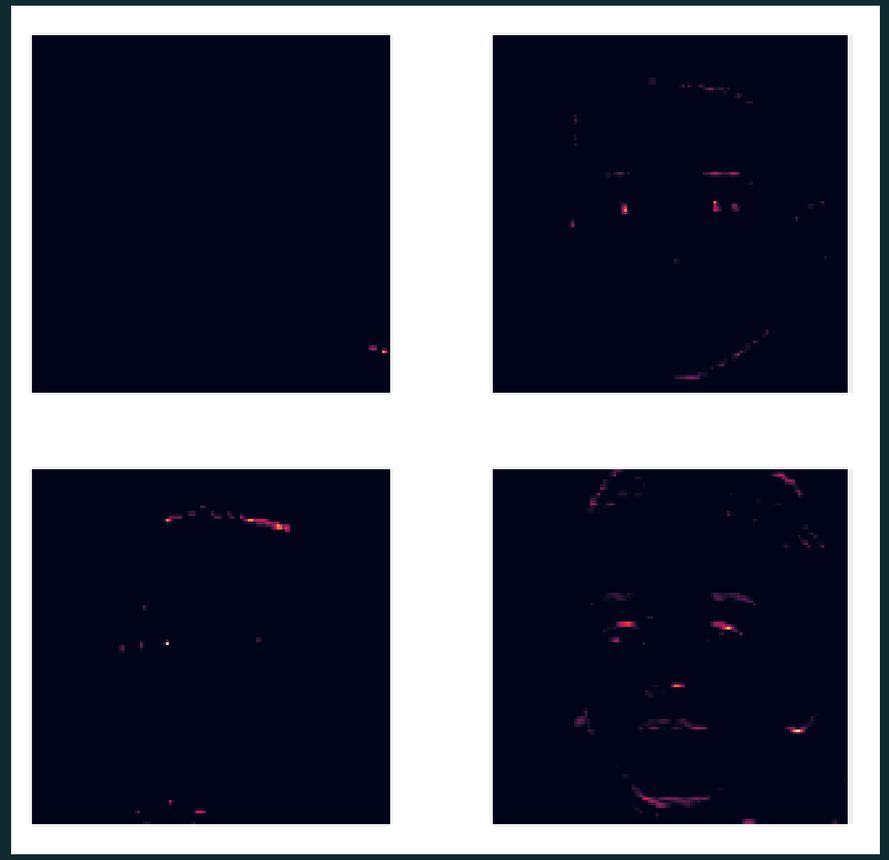
After filters like these:

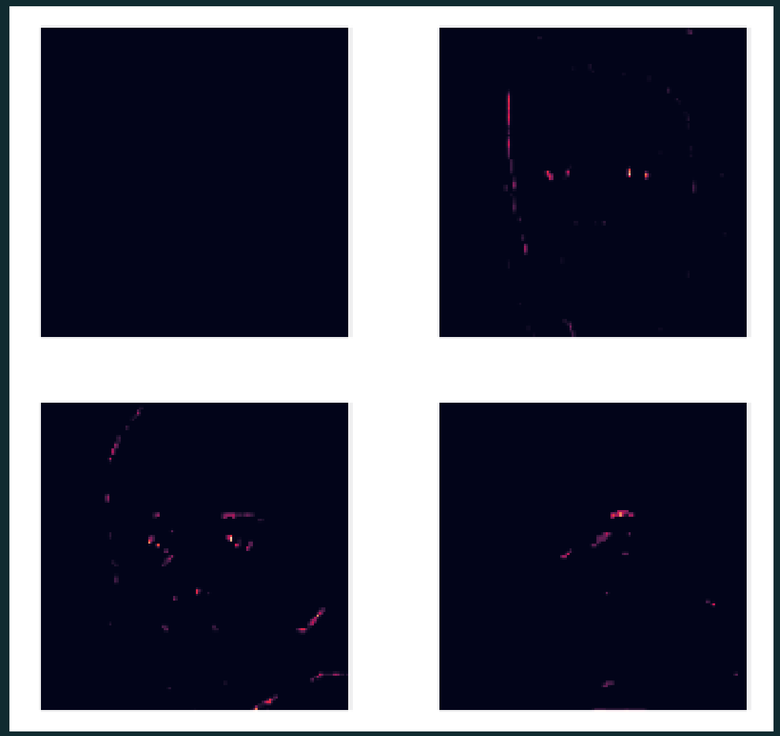




Have been passed (convoluted) over the image

The second layer is similar...

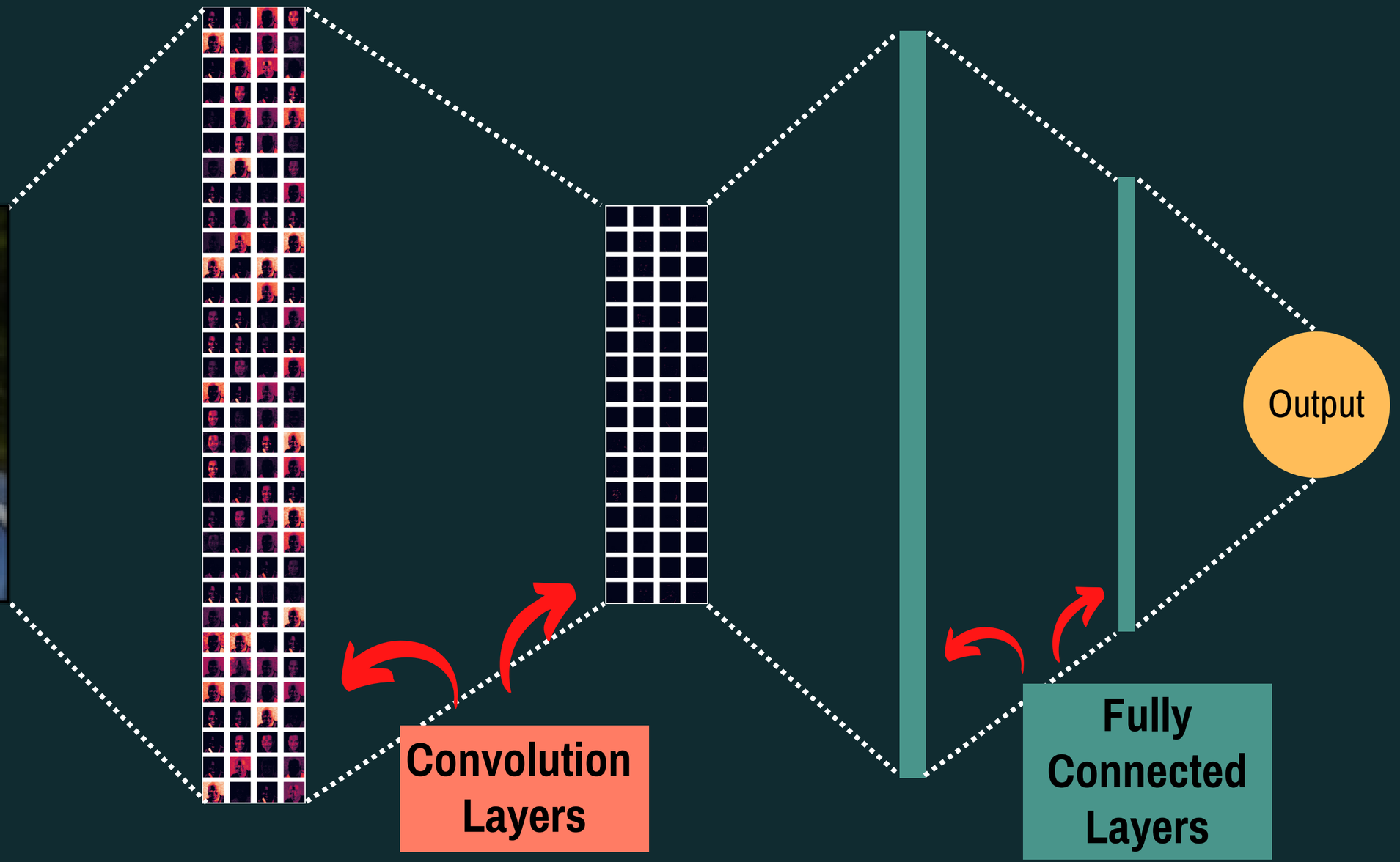




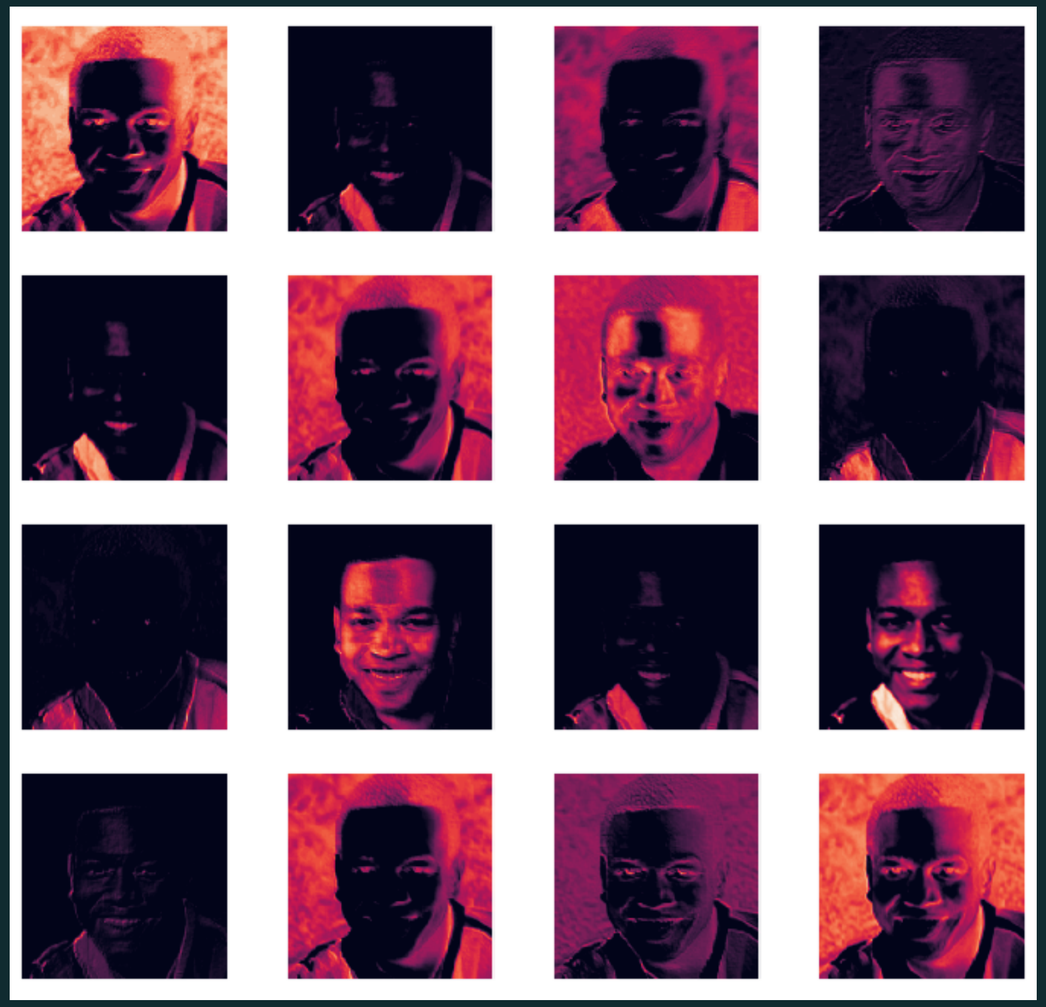
But becomes much harder to interpret.

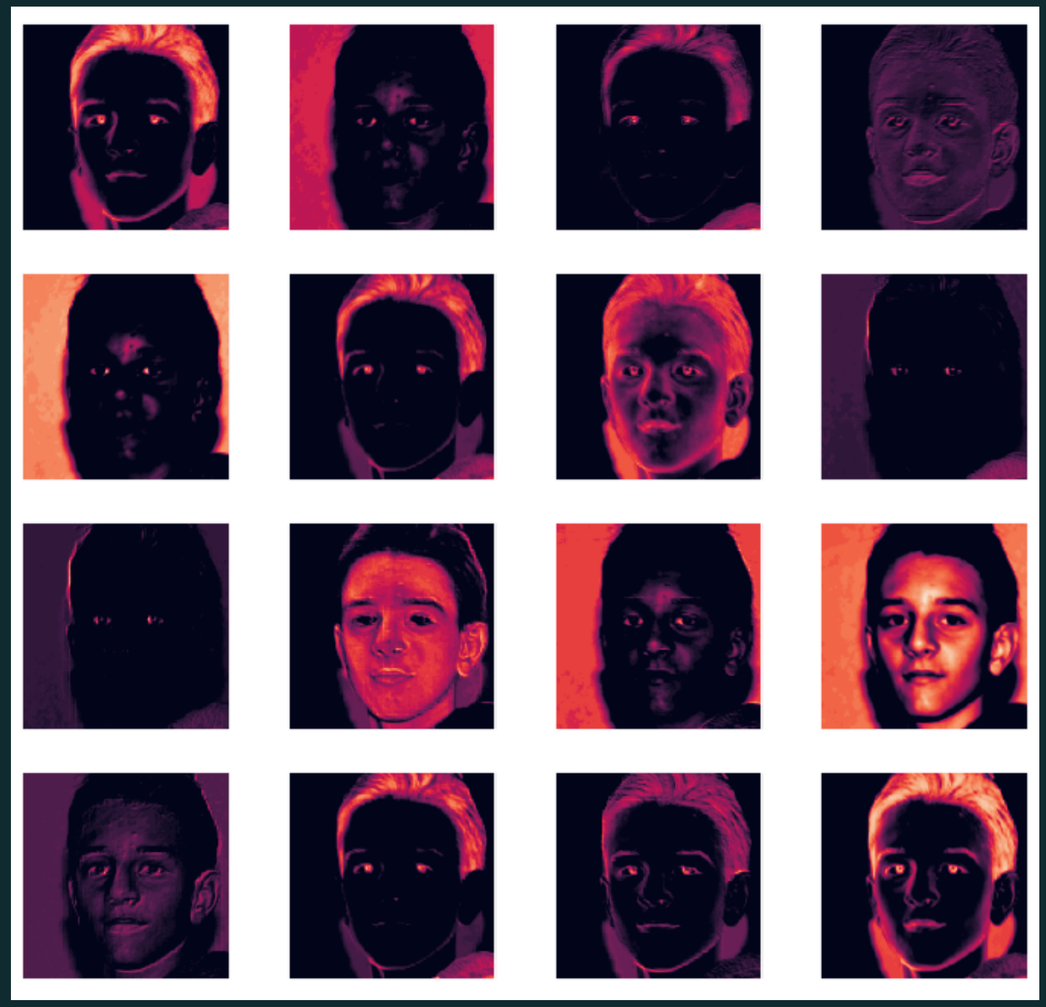
Looking at one of our false negatives...

False Negative



We can place them side-by-side to compare....



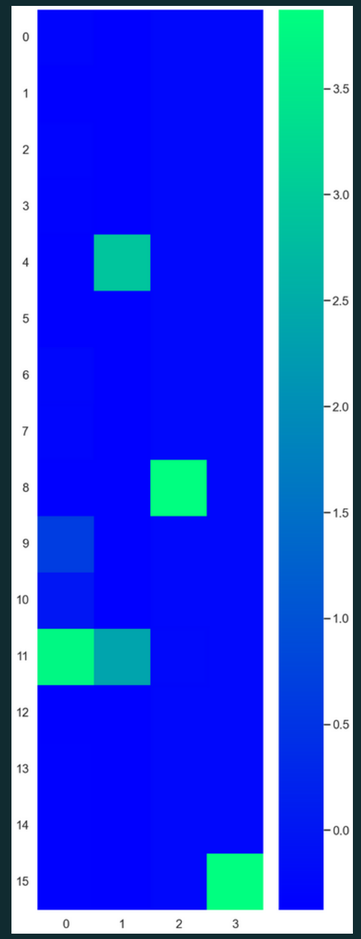


But can do little more than speculate.

However...

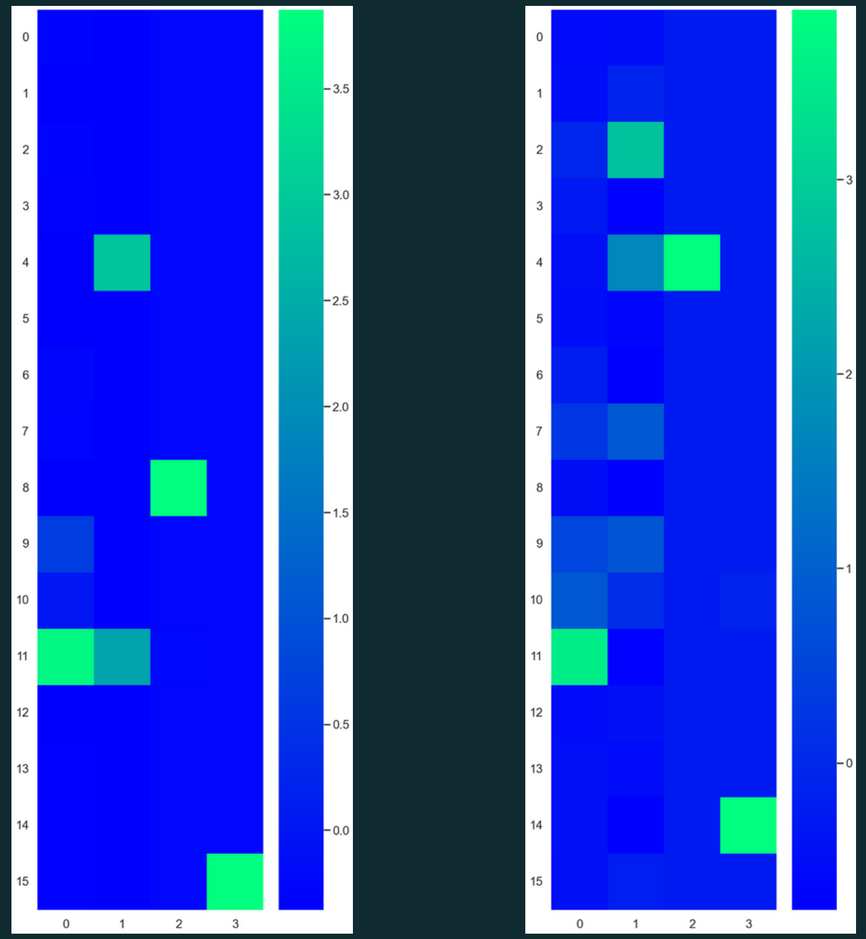
If we look at average feature map 'activation' over specific groups...

Like...



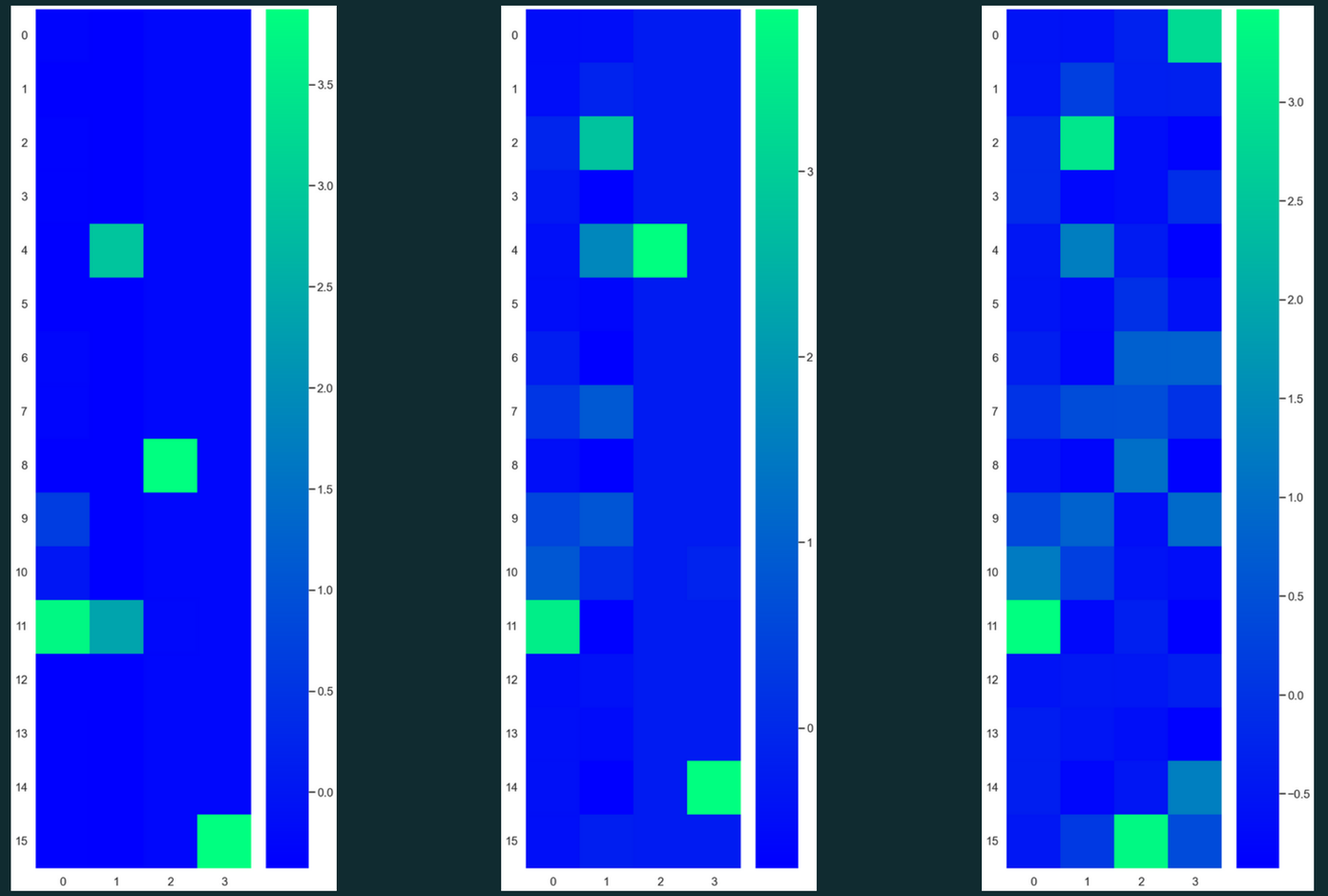
The group where the model was most confident the image was fake (99.99%)

Like...



The group where the model was most confident the image was fake (99.99%)

Like...



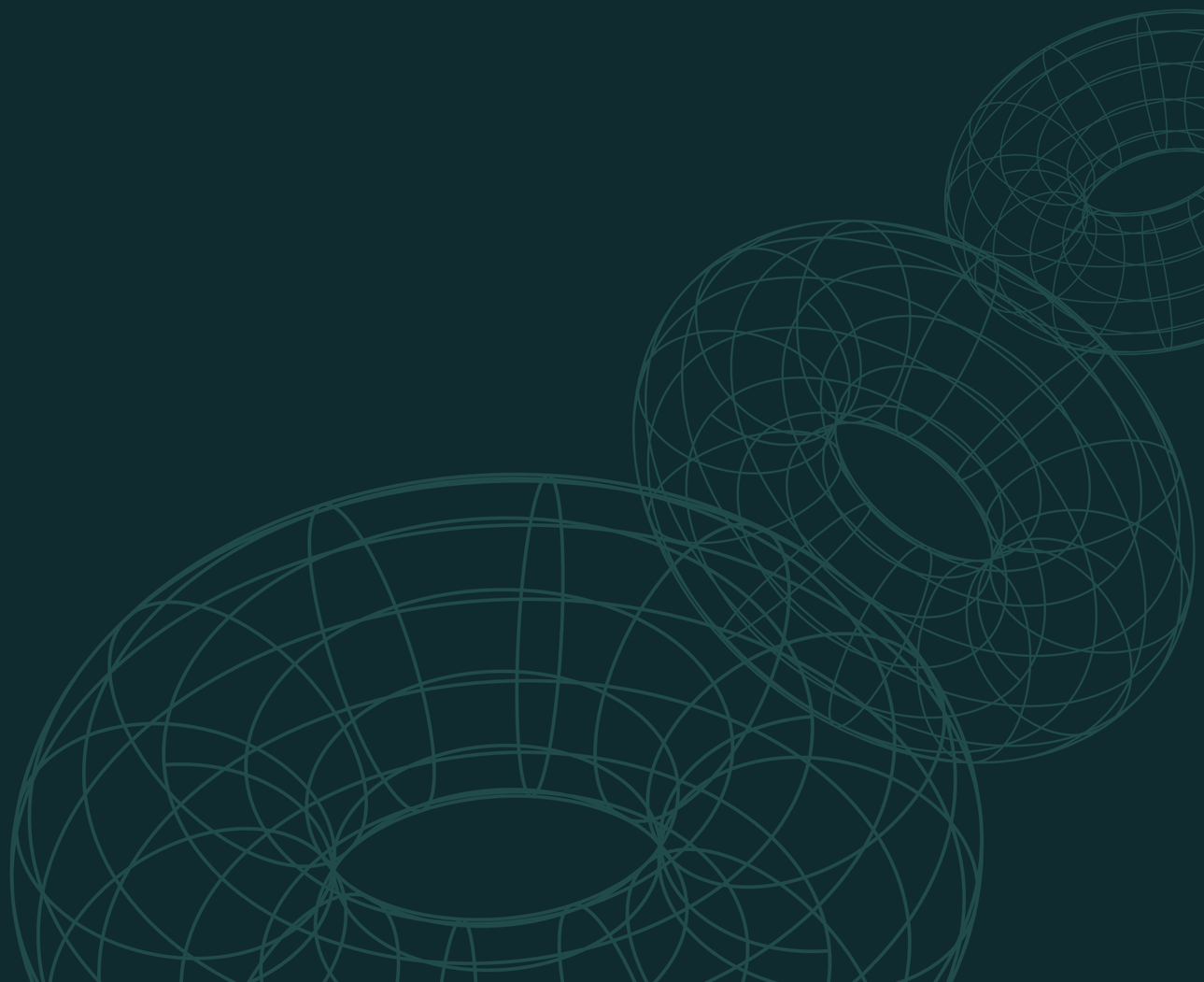
The group where the model was most confident the image was fake (99.99%)

"Any sufficiently advanced technology is equivalent to magic."

SIR ARTHUR C. CLARKE

Do you have any questions?

Send it to us!
We hope you learned something new.



Nightingale, Sophie J, and Hany Farid. “AI-synthesized faces are indistinguishable from real faces and more trustworthy.” Proceedings of the National Academy of Sciences of the United States of America vol. 119,8 (2022): e2120481119. doi:10.1073/pnas.2120481119

Free Resources

Use these free recolorable icons
and illustrations in your Canva design