Agathe d'Aubenton-Carafa * (s243230)
Srijita Sarkar * (s242527)
Upasana Paul ᵀ(s242577)
Gokul Desu ᵀ (s242580)

* Human-Centered Artificial Intelligence, DTU
ᵀ Autonomous Systems, DTU

Fake Faces and Emotions

P18: Analysis of Synthetically Generated Images (Deep Learning: 02456)

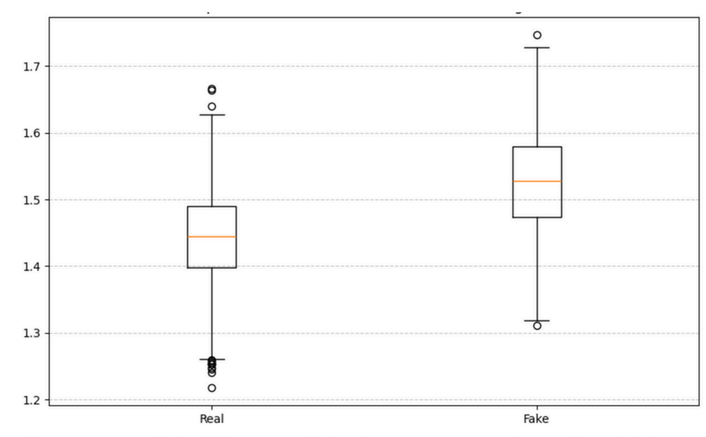
01. Introduction

One of the problematic uses of synthetically generated images is the propagation of false information and truth manipulation. In this project we aim to detect deepfakes of people's faces. We further train a model to recognise emotions from real facial images and test it on AI generated images to see how well a model trained on real data performs on hyper-realistic synthetic ones.

02. Dataset Description

- We use a dataset of 140K Real and Fake Faces available on Kaggle to detect synthetic facial data. It contains 70K real faces from the Flickr dataset collected by Nvidia and 70K fake faces sampled from the 1 million images generated by StyleGAN. The images have been resized into 64*64 px and the data has been split into train, test, and validation folders having 5000+500+500 real and 5000+500+500 fake images.
- For emotion detection, we have used the FER-2013 dataset, having 28,709 training images, 7178 validation images and 3,589 test images of real faces, also available on Kaggle. These are grayscale images resized to 48*48 px with the face more or less centered and occupying the same amount of space in each photo. Each facial expression belongs to one of seven categories: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral.
- To test our emotion detection model, we have created a dataset of synthetic images using the generator Canava and Adobe Firefly. We have 10 images belonging to each of the 7 emotions so a total of 70 AI-generated test images.

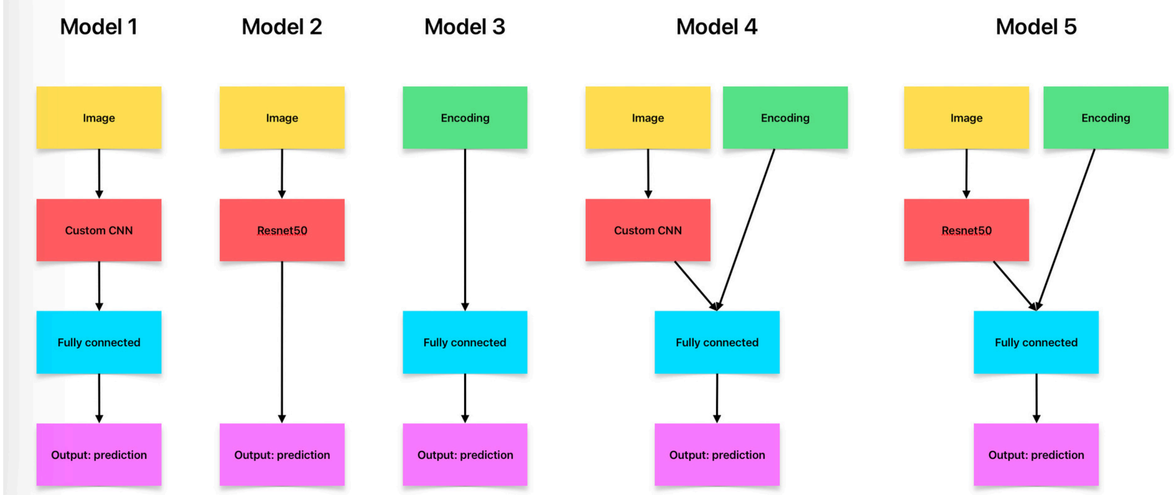
03. Fake images analysis



Repartition of the L2 norm of the feature vectors for real and fake images

To analyse the difference between real images and synthetically generated ones, we used the face_encoding function of the face-recognition libary that returns a 128 features vector caracterising the face on the picture.

04. Fake images detection

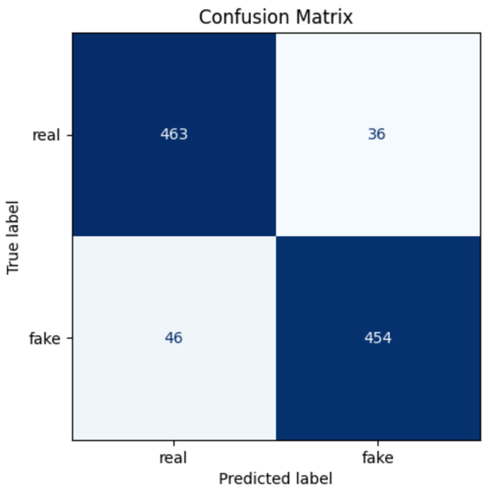


Models used

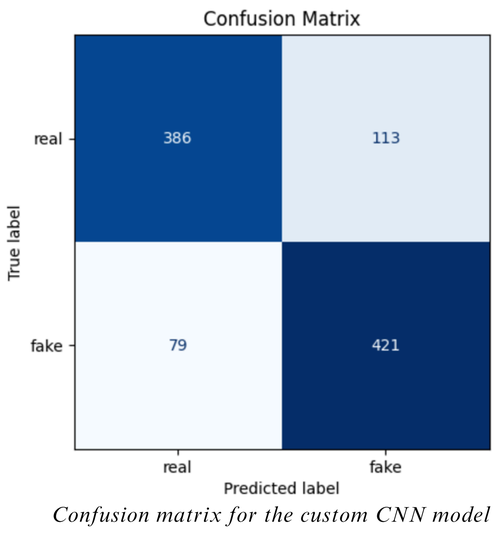
	Custom CNN	Resnet50	Encoding	CNN + Encoding	Resnet50 + Encoding
Training accuracy	98.52%	98.58%	98.97%	99.21%	99.56%
Test accuracy	80.78%	86.29%	91.78%	87.99%	91.69%

Parameters
Batch Size : 32
Optimizer : Adam
Loss : BCEWithLogits
Learning rate : 0.001
Epochs : 20

Model Performances

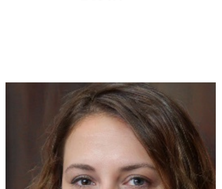


Confusion matrix for the encoding model



Confusion matrix for the custom CNN model

Real



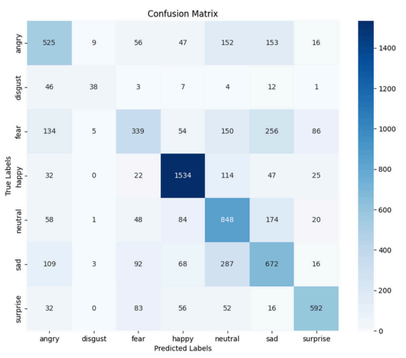
Fake

06. Emotion Detection on Real & Fake Faces

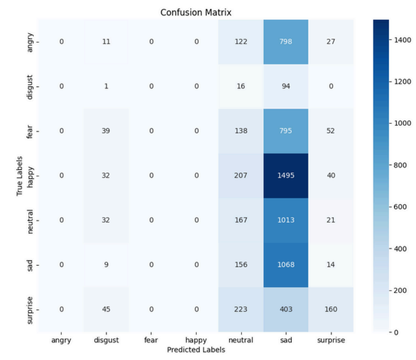
We have used two models both of which rely on the pre-trained VGG16 architecture, the architectures of which are shown below, to train and test on real face images. After that, we tested it on our own dataset of AI-generated hyper-realistic face images belonging to the different emotions

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 48, 48, 3)	30
vgg16 (Functional)	(None, 1, 1, 512)	14,714,688
flatten (Flatten)	(None, 512)	0
dense (Dense)	(None, 512)	262,656
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 7)	3,591

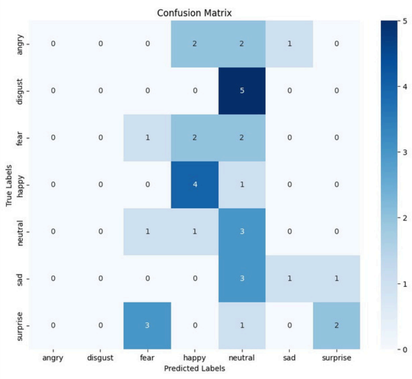
Model 1 summary



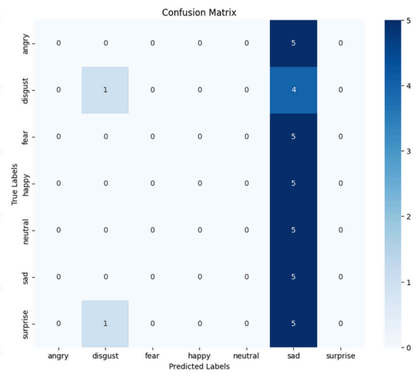
Model 1 Confusion Matrix for Real Data



Model 2 Confusion Matrix for Real Data



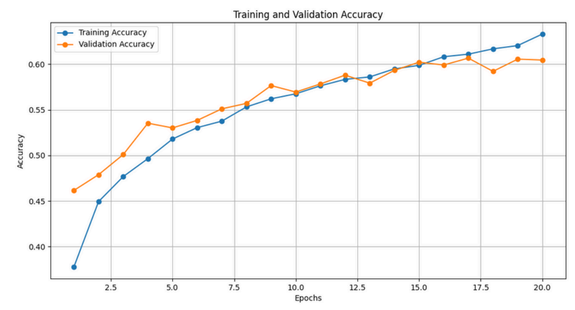
Model 1 Confusion Matrix for Fake Data



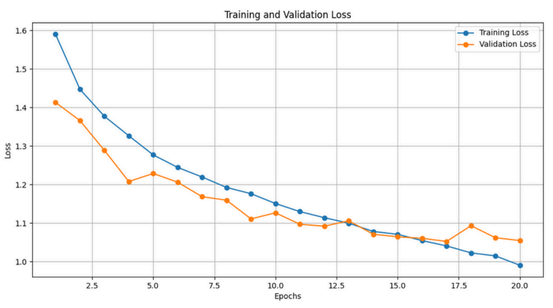
Model 2 Confusion Matrix for Fake Data

Layer (type)	Output Shape	Param #
input_layer_2 (InputLayer)	(None, 48, 48, 1)	0
conv2d_1 (Conv2D)	(None, 48, 48, 3)	30
vgg16 (Functional)	(None, 1, 1, 512)	14,714,688
dropout_3 (Dropout)	(None, 1, 1, 512)	0
flatten_1 (Flatten)	(None, 512)	0
batch_normalization_4 (BatchNormalization)	(None, 512)	2,048
dense_4 (Dense)	(None, 32)	16,416
batch_normalization_5 (BatchNormalization)	(None, 32)	128
activation_3 (Activation)	(None, 32)	0
dropout_4 (Dropout)	(None, 32)	0
dense_5 (Dense)	(None, 32)	1,056
batch_normalization_6 (BatchNormalization)	(None, 32)	128
activation_4 (Activation)	(None, 32)	0
dropout_5 (Dropout)	(None, 32)	0
dense_6 (Dense)	(None, 32)	1,056
batch_normalization_7 (BatchNormalization)	(None, 32)	128
activation_5 (Activation)	(None, 32)	0
dense_7 (Dense)	(None, 7)	231

Model 2 summary



Model 1 Train and Validation Accuracy



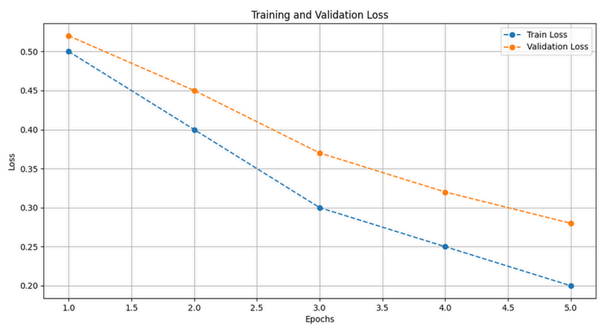
Model 1 Train and Validation Loss

	Model 1	Model 2
Train Accuracy	65.25%	85.73%
Test Accuracy for Real Faces	63.36%	85.69%
Test Accuracy for Fake Faces	30.56%	85.71%

Model Performances for Emotion Detection

Both Models 1 and 2 perform better for dominant classes such as 'happy' but poorly for minority classes. In model 2, even though accuracy is high, the real images confusion matrix shows that emotions like 'angry' and 'fear' are heavily misclassified as 'sad' and there is no diagonal dominance. This is due to class imbalance in the FER dataset.

Model 2 Train and Validation Accuracy



Model 2 Train and Validation Loss

Examples of Synthetic Images for each Emotion

Model 1 achieves low accuracy on the synthetic image dataset and struggles to differentiate between the classes, though its prediction on real data is acceptable. The confusion matrix for fake data is scattered with no diagonal dominance. Thus, the model is suited to detect emotions in real face images but fails for fake hyper-realistic data. This could be because images generated by AI might not always show clear emotions through facial features.

Model 2 has high test accuracy on generated images but is probably overfitted to dominant classes during training. It generalizes poorly to diverse classes of emotions though it seems to have learnt more robust features than Model 1, from the training set of real images.

References:

1. R. Wang, Z. Yang, W. You, L. Zhou and B. Chu, "Fake Face Images Detection and Identification of Celebrities Based on Semantic Segmentation," in IEEE Signal Processing Letters, vol. 29, pp. 2018-2022, 2022, doi: 10.1109/LSP.2022.3205481
2. C-C Hsu, Y-X Zhuang, and C-Y Lee, "Deep Fake Image Detection Based on Pairwise Learning" Applied Sciences 10, no. 1: 370, 2020, doi: 10.3390/app10010370
3. S Tariq, S Lee, H Kim, Y Shin, and S S. Woo, "GAN is a friend or foe? a framework to detect various fake face images." In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19). Association for Computing Machinery, New York, NY, USA, 1296–1303, doi: 10.1145/3297280.3297410