Practical Machine Learning Project 2 Report

Olteanu Fabian Cristian

FMI, AI Master, Year 1

---

1. Dataset

The dataset that was used is "Real vs Fake Face Classification"[1]. It contains a total of 1709 images that are split in train and validation folders (the data from the test folder was not used). The train split is composed of 1197 pictures, out of which 532 are labeled as fake and 665 as real. Additionally, the validation split has a total of 512 items, 228 of which are fake and 284 are real, so the dataset is mostly balanced. As such, no data augmentation method was employed.

The images are labeled as being either real or AI generated (via GAN [2]). Although a supervised learning would be better suited to achieve this goal, it is interesting to consider the usage of algorithm that employ unsupervised learning. In this case, the Mean-Shift algorithm was used.

2. Data preprocessing

The first step taken was to embed from the data (feature extraction) using a pre-trained face detection model (FaceNet, David Sandberg - model "20180402-114759" [3]). "20180402-114759" was trained on another dataset composed of 3.31 million images. This model calculates embeddings of 512 values for each image processed. A very useful tool that was taken advantage of was the keras_facenet [4] Python library, which is a wrapper of the discussed model that streamlined the process of retrieving the embeddings for the images at hand.

After these embeddings were extracted, Principal Component Analysis was conducted on the computed arrays, reducing the dimensionality of these (from 512 to 2), in order to illustrate the data and execute the Mean-Shift clustering algorithm.

3. The Mean-Shift Algorithm

The algorithm used on the data is part of the SkLearn Python library [5]. SkLearn also offers a way of estimating its hyperparameter, the bandwith. Taking all of this into account, the bandwith was estimated on the training split, using a quantile value of 0.0855.

After the model was fit with the training data, a scatter plot was computed where the clusters computed by the algorithm are shown (Fig 1).

In addition to the scatterplot, in an attempt to understand the results, the images taking part of the clusters were grouped based on the model's output. In figures 2 through 5, examples of snippets of 30 images from their respective clusters can be observed.

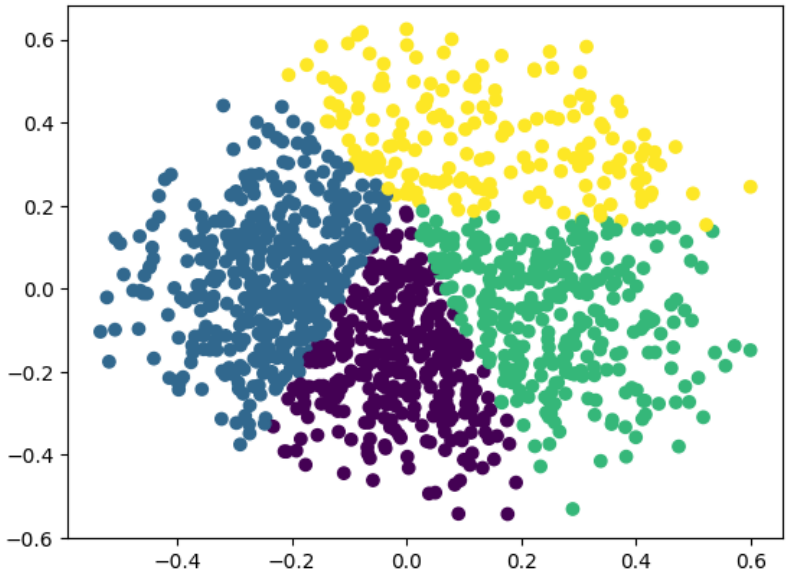

Figure 1. Scatterplot of the Computed Clusters

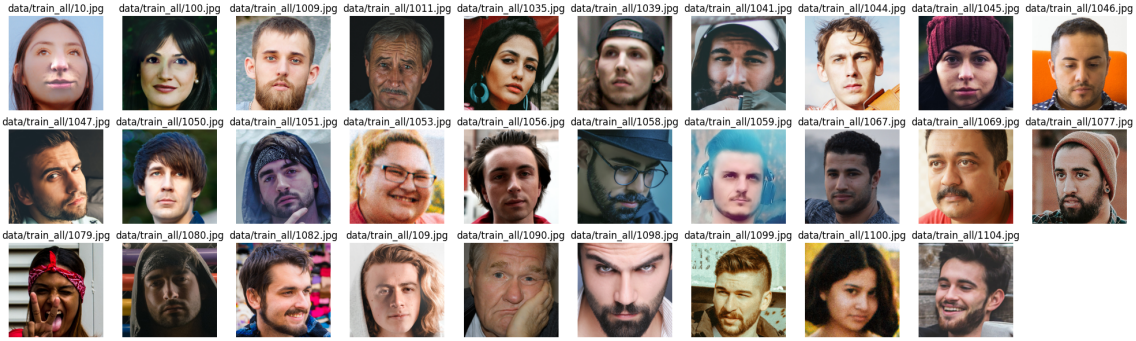

Figure 2. 30 Images from Cluster One

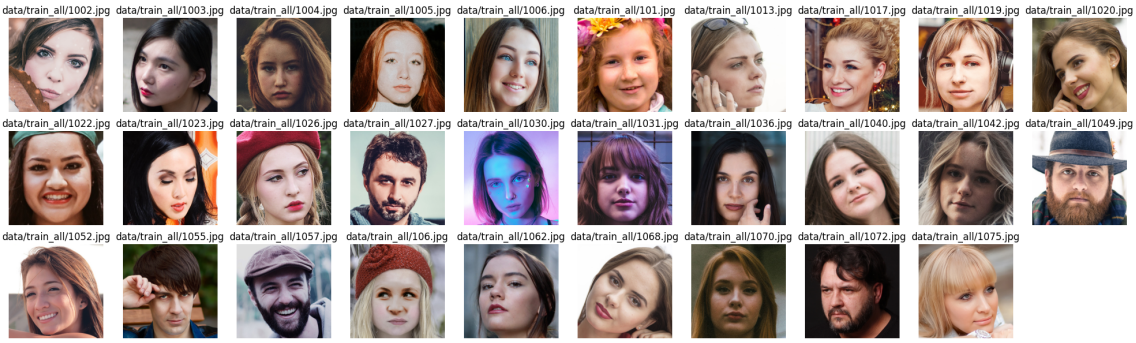

Figure 3. 30 Images from Cluster Two

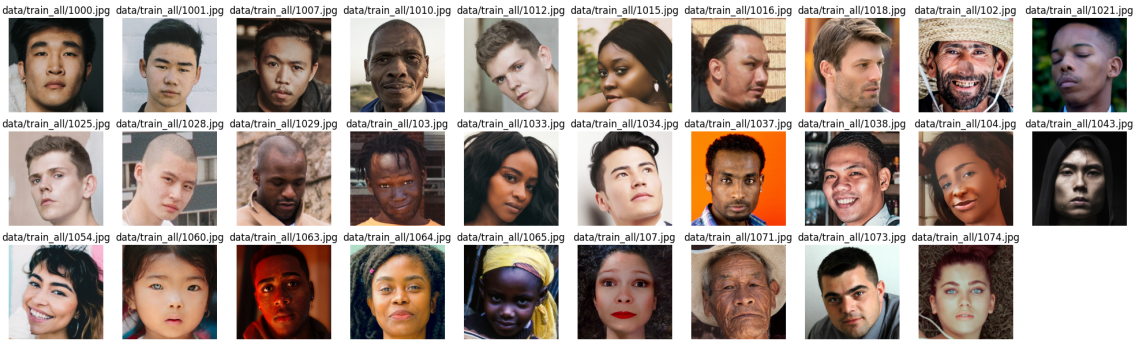

Figure 4. 30 Images from Cluster Three

Figure 5. 30 Images from Cluster Four

3.1. Scoring

An attempt was made to create a scoring metric (comparing the results to the ground truth, the labels of the dataset). In order to achieve this, an interpretation of the four clusters was required, so as to discern between real and AI-generated images. Based on the four snippets that were presented above, the final interpretation was the following:

- Cluster one is mainly composed of fake men,
- cluster two of real women,
- cluster three of real men,
- cluster four of fake women/children.

After comparing with the ground-truth, the accuracy scores were the following:

Train accuracy	Validation accuracy
0.532	0.519

4. BIRCH Algorithm

The balanced iterative reducing and clustering using hierarchies[6] algorithm was also used on the encodings obtained from the pre-processing of the images taking part of the dataset, courtesy of the SkLearn library[7].

In this case, the parameters used for the model were: 4 clusters after the final clustering step and a threshold (by which the radius of the subcluster obtained by merging a new sample and the closest subcluster) value of 0.08. After fitting the training data, the scatterplot from the figure below was obtained.

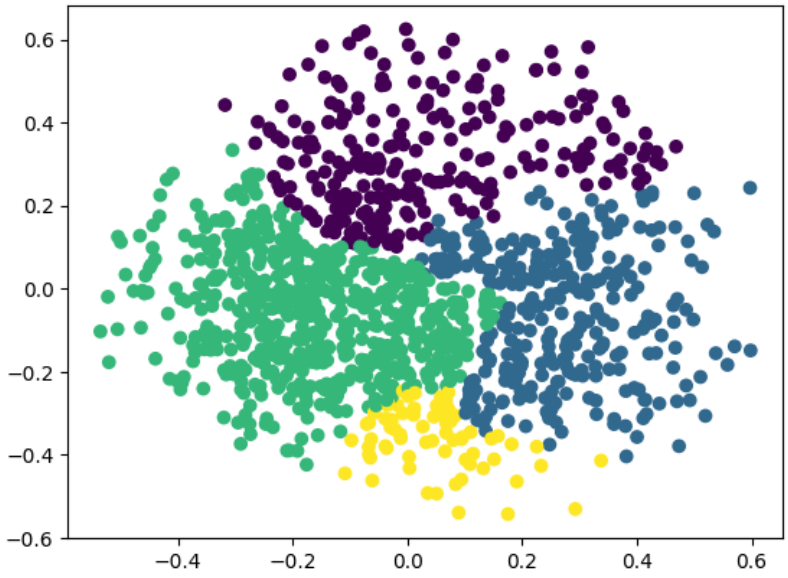

Figure 6. 30 Images from Cluster Four

Similar to the Mean-Shift section, interpretations for the four clusters were conducted and the conclusions were:

- Cluster one seemed to be composed mostly of fake children/young women,
- cluster two of real women,
- cluster three of real women,
- cluster four of real, older men.

4.1. Scoring

Following the same idea as in the last section, the following scores were computed:

Train accuracy	Validation accuracy
0.543	0.535

5. Conclusion

In the case of both Mean-Shift algorithms, the results were, although interesting, considering the unfitness of the algorithms for the task at hand, somewhat unsatisfactory, based on the metrics used to measure their performance.

References

[1] https://www.kaggle.com/datasets/underscOre/fake-vs-real-face-classification

[2] https://en.wikipedia.org/wiki/Generative_adversarial_network

[3] https://github.com/davidsandberg/facenet

[4] https://pypi.org/project/keras-facenet/

[5] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html

[6] https://en.wikipedia.org/wiki/BIRCH

[7] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.Birch.html