Face Recognition in the Wild

Pranav Kompally, Juluri Sai Chandu, Sai Dheeraj Malkar

Abstract

Face Recognition is a very popular problem in the Machine learning community. Over the years, researchers have suggested various methods and techniques to tackle the same. In this project, we highlight a few approaches and compare how their perform on the Celebrity Face Dataset.

1 Introduction

We aim to analyse the ability for machine learning, deep learning and transformer models to recognize faces in the celebrity face dataset.

2 Dataset

For this task, we put the Celebrity Face Dataset (CFD)[1] to use. The dataset contains images of 18 celebrity segregated in various directories containing 100 images each. These images consists a good mixture of photos captured in different angles over a period of time, therefore ensuring a diversity.

3 Dataset Preparation

The dataset has been handled and prepared accordingly for SVMs and Deep Learning Models.

3.1 Preprocessing of SVM

All the images have been resized to 64 x 64 pixels. The images are flattened from a 2D array into a 1D array. The labels for each image have been encoded into unique integers. A dictionary is curated with mappings that relate the images with the labels. The set of images and labels are then converted into NumPy arrays which facilitate Machine learning tasks and model training.

3.2 Processing for CNNs and Vision Transformers

For training, images are resized to 256 x 256 pixels, randomly cropped, and flipped horizontally to augment the dataset and prevent overfitting. Both training and validation images are then converted to tensors and normalized with mean and standard deviation values for model compatibility. Validation images are resized and center-cropped without random augmentations for consistency in evaluation.

4 Approach and Implementation

The design and implementation of the project can be accessed with this link: `https://drive.google.com/drive/folders/1HOMwjeC6OhZjQYiqxo_lbvOIE6YGaWtc?usp=sharing`

4.1 Support Vector Machines in Action

We build a custom Support Vector Classifier [2] using numpy. For evaluation purposes, we tried fitting using various kernels such as RBF, Linear and Poly. For extracting features and handling dimensionality, feature extracting techniques such as HOG (Histogram of Oriented Gradients) among many.

4.2 Legacy Convolutional Neural Networks

Legacy CNNs have been once proven to provide benchmarking results in image classification tasks. To put their mettle to test, 5-layer two-dimensional CNN model that uses LeakyReLU as activation function in each layer. We train for 400 epochs on Adam Optimzer at a learning rate of 0.0001. While there are various complex configurations, we chose to model an architecture that is functionally learning features with the minimum number of paratmeters.

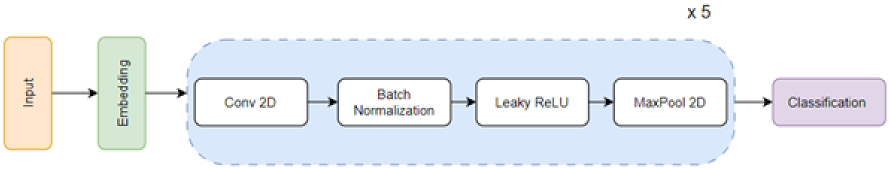


Figure 1

4.3 Vision Transformers (ViT)

Vision Transformers[3] have been in the news lately for their ability to understand pixel configuration in every images. Vision Transformers are data hungry architectures when it comes to image recognition tasks. With the current distribution of the CFD, we observe their performance and compare the results with that of the CNNs to see if they outperform even with limited set of examples. We use the Google's pretrained ViT[4] model that was trained on ImageNet dataset. We finetune the model on CFD and the images have been specifically resized to 224 x 224 pixels which the model demands as input.

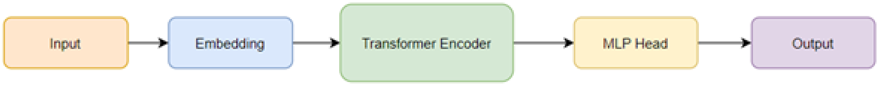


Figure 2

5 Metrics for Evaluation

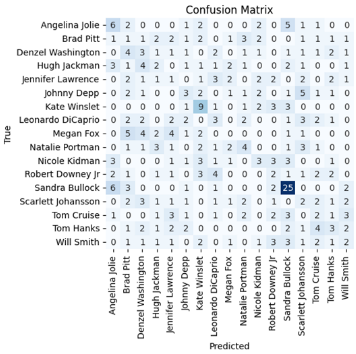
For the ease of comparing the performance of multiple models on the same dataset, we utilise validation accuracy, precision and then the confusion matrix to distinguish the model's ability to generalize on test set.

5.1 Did Supervise ML perform as expected?

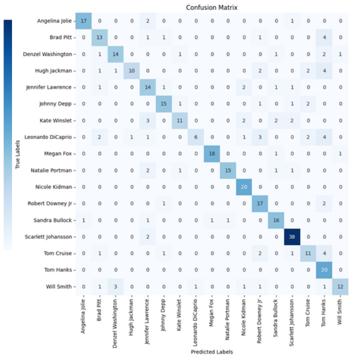
SVMs disappoint in this case. The Linear SVM yields a test accuracy of $\approx$20%. The confusion matrix of Figure 1 represents SVMs performance on the test set. You will notice that SVMs don't seem to capture much features to learn due to data sparsity. With as low as 140 images (post augmentations) per class for training, SVMs struggle to fit a classification criterion for 18 classes. And thus, perform poorly!

5.2 Legacy CNNs still got the juice?

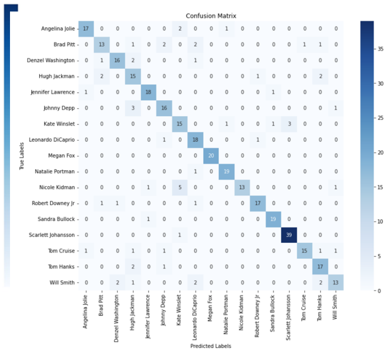
CNNs seem to have still got their juice. The custom CNN network yields a test accuracy of 82% with a precision of 81.45 after training for 400 epochs. The confusion matrix in figure 2 depicts



(a) Figure 3



(b) Figure 4



(c) Figure 5

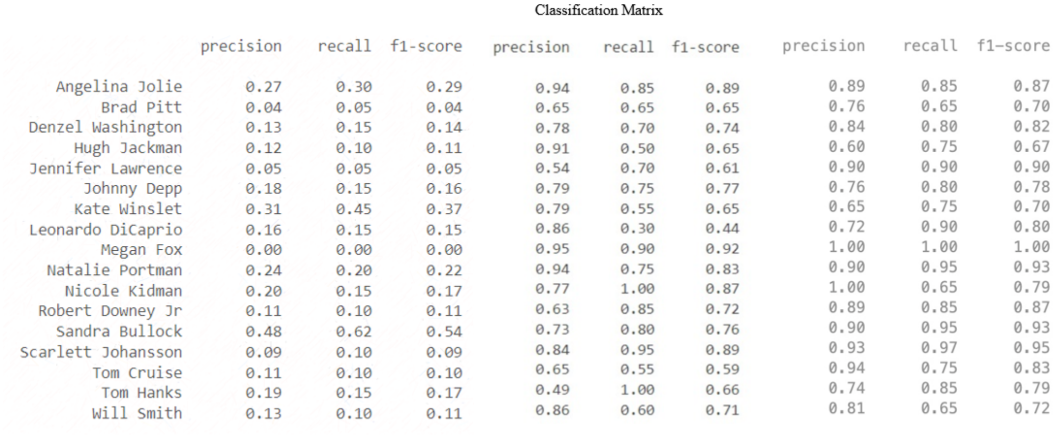
Table 1: Evaluation Metrics

Model	Precision	Accuracy
SVM	17.48	20%
Legacy CNN	81.45	82%
Vision Transformer	85.55	87%

the contrast in how accurately the CNNs perform than SVMs. While there were transformations applied to the training set. The number of examples were pretty much limited. CNNs capture the dimensionality of the images that SVMs struggle with.

5.3 Vision Transformers for the win?

Like CNNs, ViT in this case has achieved a similar test accuracy of 87% after 400 epochs . ViT unlike CNN has Transformer Encode at its core to learn the sequential arrangement of pixels in the image. With such complexity, we've observed that ViT has merely outperformed Legacy CNNs by a small margin. Precision at 85.55, ViT still beats in CNNs in practice considering the evaluation metric results.



Classification Matrix

```
                   precision    recall  f1-score   precision   recall  f1-score    precision    recall  f1-score

    Angelina Jolie      0.27      0.30      0.29        0.94     0.85      0.89         0.89      0.85      0.87
         Brad Pitt      0.04      0.05      0.04        0.65     0.65      0.65         0.76      0.65      0.70
 Denzel Washington      0.13      0.15      0.14        0.78     0.70      0.74         0.84      0.80      0.82
      Hugh Jackman      0.12      0.10      0.11        0.91     0.50      0.65         0.60      0.75      0.67
 Jennifer Lawrence      0.05      0.05      0.05        0.54     0.70      0.61         0.90      0.90      0.90
       Johnny Depp      0.18      0.15      0.16        0.79     0.75      0.77         0.76      0.80      0.78
      Kate Winslet      0.31      0.45      0.37        0.79     0.55      0.65         0.65      0.75      0.70
 Leonardo DiCaprio      0.16      0.15      0.15        0.86     0.30      0.44         0.72      0.90      0.80
         Megan Fox      0.00      0.00      0.00        0.95     0.90      0.92         1.00      1.00      1.00
   Natalie Portman      0.24      0.20      0.22        0.94     0.75      0.83         0.90      0.95      0.93
     Nicole Kidman      0.20      0.15      0.17        0.77     1.00      0.87         1.00      0.65      0.79
  Robert Downey Jr      0.11      0.10      0.11        0.63     0.85      0.72         0.89      0.85      0.87
    Sandra Bullock      0.48      0.62      0.54        0.73     0.80      0.76         0.90      0.95      0.93
Scarlett Johansson      0.09      0.10      0.09        0.84     0.95      0.89         0.93      0.97      0.95
        Tom Cruise      0.11      0.10      0.10        0.65     0.55      0.59         0.94      0.75      0.83
         Tom Hanks      0.19      0.15      0.17        0.49     1.00      0.66         0.74      0.85      0.79
        Will Smith      0.13      0.10      0.11        0.86     0.60      0.71         0.81      0.65      0.72
```

Figure 6: SVM   Figure 7: Legacy CNN   Figure 8: Vision Transformer

6 Conclusion

From the above experiments, we can observe and conclude the following.

- Although SVMs have shown to perform well on image classification tasks; on the CFD we find them struggling to make sense of the features of 18 celebrities suggesting their need for more distinguishing feature attributes.
- Legacy CNNs give Fine-tuned Vision Transformers (ViT) a tough fight after training for 400 epochs. Employing Custom build CNNs have shown to yield accurate classification results for most of the classes as seen in the Classification Matrix.
- ViTs being data demanding have managed to capture sufficient number of features from fewer number of images, thus performing the best of all three models.

References

[1] `https://www.kaggle.com/datasets/vishesh1412/celebrity-face-image-dataset`

[2] M. Arya Chandra and S. S. Bedi. Survey on SVM and their application in image classification. *International Journal of Information Technology*, 13(5):1-11, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, NeurIPS 2020.

[4] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph E. Gonzalez, Kurt Keutzer, Peter Vajda. Visual Transformers: Token-based Image Representation and Processing for Computer Vision.