

The Capstone Project of MLND - Face Recognition for My Family Members

I. Definition

A. Project Overview

The iCloud photos from iOS or Mac OS provided a comprehensive way to tag faces. All photos in the library will be scanned, thereafter faces will be identified and tagged accordingly based on the earlier defined tags/names.

I am a photography enthusiast and I have started to take photos from year 2004. As of today, it's over 200k shots and I've kept 40k of them. There are many desktop applications can do the job of face recognition, but it's going to be super fun if I can build the solution end to end.

Many other interesting use cases can be further built/extended by this, for example, auto greeting to the person sitting in front of the computer by calling his/her name. However, this won't be covered in this project.

B. Problem Statement

There are many great online blogs/projects discussed about the detailed mathematics and implementation of face recognition. For example, [this one](https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78) (<https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78>). I don't have plan to approach my problem in that way which may be too difficult and time consuming for me.

The first two questions came to my mind about this project were how to minimize the distraction factors in the photos and how to extract features understood by computers. All my photos are about something or somebody and I bet none of them is about a single face only. This project is about face recognition and all the other factors other than faces will be 'noise' and should be avoided before any machine learning kicking in. A 300 * 300 pixel color photo is quite good for human eye to distinguish faces. But it's a vector of size 300 * 300 * 3 = 270,000 which sounds too big to be machine learnt by today's computer. After some research and experiments I decided to use OpenCV to detect/extract faces and then used Google Inception V3 to extract features from the face photos.

Given the context it's face recognition for my family members and that makes this project to solve a multi-class classification problem.

C. Metrics

In this project both accuracy and speed will be looked into when evaluating the performance of the models. A model is useless if it performs only slightly better than a random guess. Speed is another important metric especially when we deal with large scale of data.

The metric of time is quite intuitive, both the training and testing time will be considered. The accuracy will be slightly more complicated. In the classification problem, we can look at precision, recall and f-beta score. The `precision_recall_fscore_support` from `sklearn` can do all this in one go. The detailed calculations are as follow. I used to have great difficulties to distinguish precision to recall until I found a way to describe them in two sentences under the context that God is the ground truth. Precision measures how much God agrees with me when I say it's positive. Recall measures how good I can find it's positive for whatever God says it's positive.

1. precision = Number of True Positive / (Number of True Positive + Number of False Positive)
2. recall = Number of True Positive / (Number of True Positive + Number of False Negative)
3. f-beta = $2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$, where here I used beta = 1.0

For a better intuition, confusion matrix will also be used to have a better visual feeling on the accuracy.

D. Workflow of the Approach

1. Analysis and data preparation. A brief experiment to display sample image, explanation on the techniques and algorithms will be used.
2. Implementation.
 - A. Scan all photos, detect faces and save them as new images with dimension of 299 * 299 pixels.
 - B. Hand pick eligible photos and save them into respective folders. There will be seven folders named `me` , `wife` , `daughter` , `son` , `dad` , `mum` , `brother` .
 - C. Apply Google Inception V3 model to extract the feature vectors of each face image.
 - D. Initial try on applying the different machine models.
3. Refinement.
 - A. Use image generator to generate more images.
 - B. Apply K-fold cross validation.
4. Performance metrics and benchmark. Linear classifier as the baseline. Evaluate the accuracy (F1 score) and time spent for both training and testing.
5. Conclusion.

II. Analysis and Data Preparation

A. Data Exploration

All my photos are in `D:\Pictures` , majority of them are in both `.jpg` and `.nef` format. The `.nef` is a raw image format for Nikon cameras and `.jpg` is the copy after image post-processing of raw file.

The output of below code chunk shows the root directory of my photo library. For each photo its full address always follows `D:\Pictures\[yyyy]\[yyyy.mm.dd] - [event name]\[yyyymmdd-hhmm][-index].jpg`

In [236]:

```
1 import os
2 from time import time
3 cur_dir = os.getcwd()
4 #print(cur_dir)
5 target_image_dir = os.path.join(cur_dir, 'images')
6 photo_dir = 'D:\Pictures'
7 os.listdir(photo_dir)
```

Out[236]:

```
['.SynologyWorkingDirectory',
 '2004',
 '2005',
 '2006',
 '2007',
 '2008',
 '2009',
 '2010',
 '2011',
 '2012',
 '2013',
 '2014',
 '2015',
 '2016',
 '2017',
 '2018',
 'Adobe Lightroom',
 'Camera Roll',
 'desktop.ini',
 'iCloud Photos',
 'Phone Photos',
 'Photography Works',
 'Readme.txt',
 'Saved Pictures',
 'SyncToy_6288703e-b478-4b80-9f48-ece12cdcb521.dat',
 'zbingjie',
 'zothers',
 '法蝶',
 '熊思宇和黄乐论辩论']
```

A glance of a recent photo directory. Each and every file use the time stamp as its file name. .nef is the raw image file, .xmp is generated by Adobe Lightroom, both are not applicable to this project. Only .jpg files are applicable to this project.

In [2]:

```
1 os.listdir(os.path.join(photo_dir, '2018'))
```

Out[2]:

```
['2018.01.01 - 冰洁爸妈证件照',
 '2018.01.01 - 彤彤和祺祺',
 '2018.01.24 - Changi Business Park Airbus A380',
 '2018.01.26 - Changi Business Park, A350, A320, B777, A380',
 '2018.01.26 - 冰洁, 祺祺',
 '2018.01.27 - Marina Bay',
 '2018.01.27 - 冰洁, 彤彤祺祺',
 '2018.01.27 - 祺祺在游泳, 哭']
```

```
In [3]: 1 os.listdir(os.path.join(photo_dir, '2018', '2018.01.01 - 彤彤和祺祺'))
```

```
Out[3]: ['20180101-0933-2.JPG',
 '20180101-0933-2.NEF',
 '20180101-0933-2.xmp',
 '20180101-0933.JPG',
 '20180101-0933.NEF',
 '20180101-0933.xmp',
 '20180101-1039.jpg',
 '20180101-1039.NEF',
 '20180101-1039.xmp',
 '20180101-1042-2.jpg',
 '20180101-1042-2.NEF',
 '20180101-1042-2.xmp',
 '20180101-1042.jpg',
 '20180101-1042.NEF',
 '20180101-1042.xmp',
 '20180101-1043-10.jpg',
 '20180101-1043-10.NEF',
 '20180101-1043-10.xmp',
 '20180101-1043-11.jpg',
 '20180101-1043-11.NEF',
 '20180101-1043-11.xmp',
 '20180101-1043-12.jpg',
 '20180101-1043-12.NEF',
 '20180101-1043-12.xmp',
 '20180101-1043-7.jpg',
 '20180101-1043-7.NEF',
 '20180101-1043-7.xmp',
 '20180101-1043-8.jpg',
 '20180101-1043-8.NEF',
 '20180101-1043-8.xmp',
 '20180101-1043-9.jpg',
 '20180101-1043-9.NEF',
 '20180101-1043-9.xmp',
 '20180101-1043.jpg',
 '20180101-1043.NEF',
 '20180101-1043.xmp',
 '20180101-1044-2.jpg',
 '20180101-1044-2.NEF',
 '20180101-1044-2.xmp',
 '20180101-1044.jpg',
 '20180101-1044.NEF',
 '20180101-1044.xmp',
 '20180101-1045-2.jpg',
 '20180101-1045-2.NEF',
 '20180101-1045-2.xmp',
 '20180101-1045.jpg',
 '20180101-1045.NEF',
 '20180101-1045.xmp']
```

All the photo files are in folders whose names are in year format. There are slightly more than 40k photos, below function will dump the full paths of them. 5 photos were randomly picked to show the full address.

```
In [4]: 1 def get_all_file_path(folder_addr):
2     """Return all jpg files' full path as a list
3     Args:
4         folder_addr (str): The folder address.
5     Returns:
6         all_jpg (list): A list of strings which are the full path of jpg files.
7     """
8     all_jpg = []
9     for root, dirs, files in os.walk(folder_addr):
10         # All the target photos are in D:\Pictures\20xx. Get the jpgs from them only.
11         path = root.split(os.sep)
12         if len(path) < 3:
13             continue
14         else:
15             year = path[2]
16             if year[:2] != '20':
17                 continue
18             #print((len(path) - 1) * '---', os.path.basename(root))
19             for file in files:
20                 if file[-3:].lower() == 'jpg':
21                     #print(len(path) * '---', file)
22                     all_jpg.append(os.path.join(root, file))
23     return all_jpg
24
25 all_jpg = get_all_file_path(photo_dir)
```

```
In [5]: 1 import random
2 print('Number of jpgs:', len(all_jpg))
3 for i in random.sample(range(len(all_jpg)), 5):
4     print(all_jpg[i])
```

```
Number of jpgs: 40783
D:\Pictures\2008\2008.05.30~2008.06.01 - 表妹六一表演\20080530-1058-54.JPG
D:\Pictures\2007\2007.03.03 - 9th SM2合影\20070303-1348-7.JPG
D:\Pictures\2006\2006.05.28 - 和家人在木兰天池\20060528-1110.JPG
D:\Pictures\2010\2010.07.15 - OCBC TC1 4th floor\20100715-1305-26.jpg
D:\Pictures\2007\2007.01.22 - PRCSU main com成员\20070122-2210-18.JPG
```

B. Sample Photo Visualization

This section demonstrates how to read and display an image file.

```
In [6]: 1 # Import required libraries for this section
2 %matplotlib inline
3 import numpy as np
4 import matplotlib.pyplot as plt
5 import math
6 import cv2
7 import time
8 DISPLAY = True
9 SAVE = True
10 NODISPLAY = False
11 NOSAVE = False
```

```
In [7]: 1 def get_numpy_from_file(file_path):
2     """Return the image file as a numpy array
3     Args:
4         file_path (str): The full address of the image file.
5     Returns:
6         image (numpy.ndarray): The numpy array of the image file. Shape of (width, height, number of channels).
7
8     The file_path may contain unicode characters, cannot use cv2.imread() directly. Below way works for both
9     unicode and non-unicode paths.
10    """
11    file_stream = open(file_path, 'rb')
12    bytes_arr = bytearray(file_stream.read())
13    numpy_ar = np.asarray(bytes_arr, dtype=np.uint8)
14    image = cv2.imdecode(numpy_ar, cv2.IMREAD_UNCHANGED)
15    print('Image numpy array shape:', image.shape, type(image))
16    return image
17
18 def display_from_numpy(image, fig_dim_x=10, fig_dim_y=10, plot_nrows=1, plot_ncols=1):
19     """Display the image from a given numpy array (the returned result from get_numpy_from_file() function).
20     Args:
21         image (numpy.ndarray): The numpy array representation of an image.
22         image (list[numpy.ndarray]): The list of numpy arrays of images.
23     Returns:
24         inline display of the image.
25     """
26     fig = plt.figure(figsize=(fig_dim_x, fig_dim_y))
27     if isinstance(image, list):
28         image_list = image
29     else:
30         image_list = []
31         image_list.append(image)
32     for i in range(len(image_list)):
33         ax = fig.add_subplot(plot_nrows, plot_ncols, i+1, xticks=[], yticks[])
34         ax.set_title('Sample Image')
35         ax.imshow(image_list[i])
36
37 def display_from_file(file_path):
38     """Display one image file inline.
39     Args:
40         file_path (str): The full address of the file.
41     Returns:
42         None: Display image inline.
43     """
44     image = get_numpy_from_file(file_path)
45     # Need to convert to RGB
46     image = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
47     display_from_numpy(image)
```

```
In [8]: 1 file_path = os.path.join(photo_dir, '2017', '2017.11.16 - Singapore Fintech Festival', '20171116-1923.jpg')
2 display_from_file(file_path)
```

Image numpy array shape: (4760, 7132, 3) <class 'numpy.ndarray'>



C. Algorithms and Techniques

a). Face Detector

In this project, I used [Haar Cascades Classifier](https://docs.opencv.org/trunk/d7/d8b/tutorial_py_face_detection.html) (https://docs.opencv.org/trunk/d7/d8b/tutorial_py_face_detection.html) from OpenCV for face detection. This classifier extracts thousands of features for each possible sub-window in the image. That will result in a huge amount of computation to determine whether this specific sub-windows is a face or not, not even mentioning this will be done for all the possible sub-windows. So the classification for each sub-windows will be gone through by stages. If at an earlier stage 10 features are used to determine whether the sub-window is a face or not, then classification will only continue if the result at this stage is positive. In this way, time won't be wasted to check the remaining thousands of features.

Another important technique used is the 'cascade' way. Instead of having one powerful classifier which could be too difficult to train, a bunch of weak classifiers are used. As long as all the weak classifiers are doing better than random guess, then majority vote can be used for the final decision/classification result.

The Haar Cascades Classifier is fast, but both the precision and recall rates are not perfect. Luckily I have a large raw dataset and I don't think it's a concern in this project.

b). Feature Extractor

The input image is with size of 299 * 299 pixels, 3 channels for a color pixel, that's a space with almost 270k dimensions. It's too big to be processed directly by the machine learning algorithms. Hence, Google Inception V3 will be applied first to extract the feature vectors with dimensionality of 2048. Maybe it's still too big in this project but we'll see.

In this project I'm extracting the `avg_pool` layer whose output shape is (, 2048).

c). Machine Learning Models

Four models will be evaluated. SGD Classifier, K-nearest Neighbour, Logistic Regression and Deep Learning models.

1. SGD Classifier.

Both SGD Classifier and Deep Learning using Stochastic Gradient Descent optimizer work quite similar as the 'SGD' way. When one or a batch of samples feeding into the model, a loss (the distance between output and expected output) will be computed. Then this loss value will be used to update the parameters/weights of the model to let it lean towards the expected output.

The SGD classifier prefers data been preprocessed to be normalized with zero mean and unit variance. In this report, the preprocessed data will be normalized from [0, 255] to [0, 1]. Not exactly as what's favored by SGD Classifier, but I standardize it to [0, 1] for all the models here. Default hyper parameters will be used.

2. K-nearest Neighbour.

Instead of constructing a generalized model, KNN is storing all the training data. When performing the prediction, distance (eg, euclidean distance) from the testing data point to all the training data point will be computed. Then based on the defined K value, using majority vote to decide the class of the testing data point. After a few tries, I set K=7.

KNN is simple but it can be very time consuming during testing as distance to all training data are needed. In addition, due to the curse of high dimensionality, it may not be effective always.

3. Logistic Regression.

Despite the 'regression' word in the name, it's indeed a classification model. Logistic regression takes the features into a logit function with output value in the range of [0, 1], which quantifies the probability of correctly predicting the class. When all the training points fed into the model, we get the average of the probability and that's the likelihood the training data is correctly predicted. So the objective in logistic regression training is to maximize this likelihood. There are many different algorithms, e.g. Newton-Raphson, Iteratively re-weighted least squares and etcs.

Logistic regression is widely used industrial widely. But it doesn't perform well when the feature space is large and I have no idea whether it'll do a great job in this project. Default hyper parameters will be used.

4. Deep Learning.

Without a doubt, deep learning has gained great exposure and evolution over the past few years, especially in the domains of computer vision, natural language processing. In this project, the feature extractor is taking almost 95% of the Google Inception V3 layers. Instead of taking the remaining 5% layers, which will do a job to classify objects into 1000 classes, I will add another few dense layers to do the customization to suit my purpose of classifying objects into 7 classes. Essentially, this is just a transfer learning approach.

Deep learning takes a lot of computation resources and time to do the gradient descent to find the best weights of the neurons. And in this project, I don't need to worry about the convolutional layers except the last few dense layers added by me.

d). Creating More Data

Despite I have a large photo library, there is still a concern on the quantity of the images with good quality. If each class only has an average of 5 images, probably I won't have any result as expected. So if it's needed, techniques like using [image generator](https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html) (<https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html>) will be explored in this project.

D. Benchmark

The SGD classifier will be set as the baseline of the benchmark. Based on few random tries, the F1 score I got was around 75%. And I'm expecting I can push the score to around 85%.

III. Implementation**A1. Detect Faces, Save as New Files - Single Image**

There are many supporting functions and `process_faces()` is the one with many flags to represent whether need to display or save the given image file.

In [9]:

```

1 # pathlib available from python 3.5
2 from pathlib import Path
3 def get_faces(image, scaleFactor, minNeighb):
4     """Perform face detection and return the detected faces as a list of (x,y,w,h).
5     Args:
6         image (numpy.ndarray): The numpy array of an image.
7         scaleFactor (float): The scaling factor to be used by the detectMultiScale() function.
8         minNeighb (int): The number of minimum neighbors to be used by the detectMultiScale() function.
9     Returns:
10        faces (list of tuples): The list of the face locations.
11    """
12    # Convert to RGB then to grayscale
13    image = np.copy(image)
14    image = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
15    gray = cv2.cvtColor(image, cv2.COLOR_RGB2GRAY)
16    # Extract the pre-trained face detector from an xml file
17    face_cascade = cv2.CascadeClassifier('detector_architectures/haarcascade_frontalface_default.xml')
18    # Detect the faces in image
19    faces = face_cascade.detectMultiScale(gray, scaleFactor, minNeighb)
20    return faces
21
22 def draw_bounding_box(image, faces):
23     """Draw the bounding box of faces on the image.
24     Args:
25         image (np.ndarray): Numpy array of the image.
26         faces (list of tuples): The list of the face locations.
27     Returns:
28         image_with_detections (np.ndarray): A image with bounding box on faces, in numpy array format,
29             after converting to RGB.
30         image_faces (list[np.ndarray]): List of face images.
31     """
32     # Use np.copy() to create duplicate images to avoid alteration of the original image.
33     image_copy = np.copy(image)
34     image_with_detections = np.copy(image)
35     image_copy = cv2.cvtColor(image_copy, cv2.COLOR_BGR2RGB)
36     image_with_detections = cv2.cvtColor(image_with_detections, cv2.COLOR_BGR2RGB)
37     # The list of detected faces
38     image_faces = []
39     # Get the bounding box for each detected face
40     for (x,y,w,h) in faces:
41         # Add a red bounding box to the detections image
42         if w > 200:
43             line_width = w//20
44         else:
45             line_width = 3
46         image_faces.append(image_copy[y:(y+h), x:(x+w)])
47         cv2.rectangle(image_with_detections, (x,y), (x+w,y+h), (255,0,0), line_width)
48     return image_with_detections, image_faces
49
50 def create_get_target_dir_file(file_path):
51     """Create the target directory if it's not existent, return the target directory as string
52     Args:
53         None: Global variables.
54     Returns:
55         target_dir (str): The address of the target directory.
56     """
57     # Create the full path of the target images by replacing the photo_dir string into target_image_dir string.
58     target_file = file_path.replace(photo_dir, target_image_dir)
59     target_dir = os.path.dirname(target_file)
60     target_path = Path(target_dir)
61
62     # Create parents of directory, don't raise exception if the directory exists
63     target_path.mkdir(parents=True, exist_ok=True)
64     return target_dir, target_file
65
66 def save_faces(file_path, image_faces):
67     """Save each face image into new files in target_iameg_dir.
68     Args:
69         file_path (str): The full path of the original photo.
70         image_faces (list[np.ndarray]): The list of face images, in numpy array format.
71     Returns:
72         None
73     """
74     if len(image_faces) == 0:
75         return
76
77     target_dir, target_file = create_get_target_dir_file(file_path)
78
79     # Resize and save each face image.
80     for i, face in enumerate(image_faces):
81         face = cv2.resize(face, (299, 299))
82         os.chdir(target_dir)
83         file_name = os.path.basename(target_file)
84         cv2.imwrite(file_name + '-face-' + str(i) + '.jpg', cv2.cvtColor(face, cv2.COLOR_BGR2RGB))
85         print(os.path.join(target_dir, file_name + '-face-' + str(i) + '.jpg'), 'saved.')
86
87 def process_faces(file_path, display=NODISPLAY, save=NOSAVE, scaleFactor=1.3, minNeighb=5):
88     """Process the input image file by extracting face(s). Display and save based on the flags.
89     Args:
90         file_path (str): The full path of the input image file.
91         display (bool): Default is NODISPLAY/False.
92         save (bool): Default is NOSAVE/False.
93         scaleFactor (float): The scaling factor used by face detection function.
94         minNeighb (int): The number of minimum neighbors.
95     Returns:
96         None: Perform display or save actions based on the flags.
97     """

```

```

98     print('Image path', file_path)
99     image = get_numpy_from_file(file_path)
100    faces = get_faces(image, scaleFactor, minNeighb)
101    print('Number of faces detected:', len(faces))
102
103    image_with_detections, image_faces = draw_bounding_box(image, faces)
104
105    if save:
106        save_faces(file_path, image_faces)
107
108    if display:
109        # Display the image with the detections
110        display_from_numpy(image_with_detections)
111
112    # Return to this project's current working directory
113    os.chdir(cur_dir)

```

In [10]:

```

1 # Load in color image for face detection
2 file_path = os.path.join(photo_dir, '2017\\2017.11.16 - Singapore Fintech Festival', '20171116-1923.jpg')
3 process_faces(file_path, DISPLAY, SAVE)

```

Image path D:\Pictures\2017\2017.11.16 - Singapore Fintech Festival\20171116-1923.jpg
 Image numpy array shape: (4760, 7132, 3) <class 'numpy.ndarray'>
 Number of faces detected: 5
 D:\Google Drive\Study\Machine Learning Engineer Nanodegree - Udacity\Projects\MLND-Projects\projects\facial_recognition_family_members\images\2017\2017.11.16 - Singapore Fintech Festival\20171116-1923.jpg-face-0.jpg saved.
 D:\Google Drive\Study\Machine Learning Engineer Nanodegree - Udacity\Projects\MLND-Projects\projects\facial_recognition_family_members\images\2017\2017.11.16 - Singapore Fintech Festival\20171116-1923.jpg-face-1.jpg saved.
 D:\Google Drive\Study\Machine Learning Engineer Nanodegree - Udacity\Projects\MLND-Projects\projects\facial_recognition_family_members\images\2017\2017.11.16 - Singapore Fintech Festival\20171116-1923.jpg-face-2.jpg saved.
 D:\Google Drive\Study\Machine Learning Engineer Nanodegree - Udacity\Projects\MLND-Projects\projects\facial_recognition_family_members\images\2017\2017.11.16 - Singapore Fintech Festival\20171116-1923.jpg-face-3.jpg saved.
 D:\Google Drive\Study\Machine Learning Engineer Nanodegree - Udacity\Projects\MLND-Projects\projects\facial_recognition_family_members\images\2017\2017.11.16 - Singapore Fintech Festival\20171116-1923.jpg-face-4.jpg saved.

Sample Image



Read and display the extracted 5 faces from the above image.

In [10]:

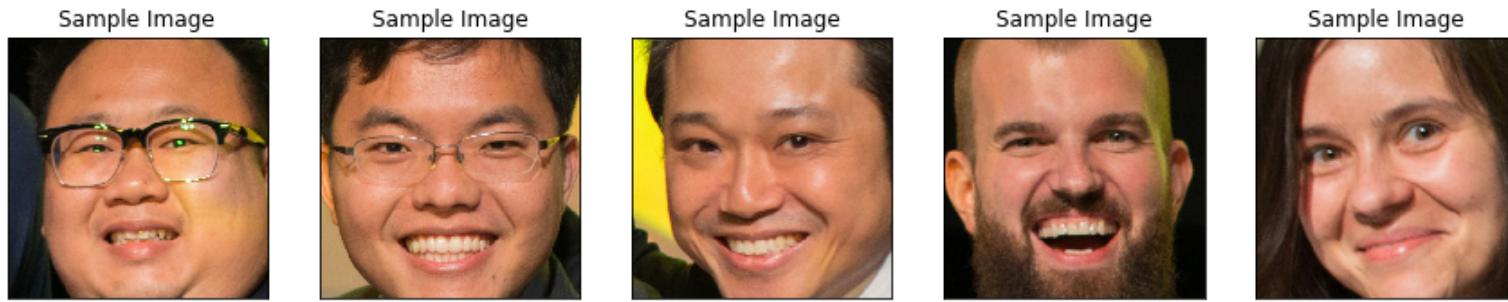
```

1 def get_file_path_from_folder(folder_addr):
2     """Return all jpg files' full paths as a list.
3     Args:
4         folder_addr (str): The folder address.
5     Returns:
6         all_jpg (list): A list of strings which are the full path of jpg files.
7     """
8     all_jpg = []
9     for root, dirs, files in os.walk(folder_addr):
10         for file in files:
11             if file[-3:].lower() == 'jpg':
12                 #print(len(path) * '---', file)
13                 all_jpg.append(os.path.join(root, file))
14
15     return all_jpg
16
16 def get_numpy_from_folder(folder_addr):
17     """Return all jpg files in a folder as numpy arrays.
18     Args:
19         folder_addr (str): The folder address.
20     Returns:
21         image_nparrays (list[numpy.ndarrays]): The list of numpy arrays of jpg images in a folder.
22     """
23     all_jpg_addr = get_file_path_from_folder(folder_addr)
24     image_nparrays = []
25     for jpg_addr in all_jpg_addr:
26         image = get_numpy_from_file(jpg_addr)
27         image = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
28         image_nparrays.append(image)
29
30     return image_nparrays

```

```
In [12]: 1 image_numpys = get_numpy_from_folder(os.path.join(cur_dir, 'images\\2017\\2017.11.16 - Singapore Fintech Festival'))
2 display_from_numpy(image_numpys, fig_dim_x=15, fig_dim_y=5, plot_nrows=1, plot_ncols=5)
```

Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>



A2. Detect Faces, Save as New Files - All Images

In the earlier section, all the 40k image full pathes were store in `all_jpg` variable. So just need to be patient to wait for the results, it'll take quite several hours to finish.

```
In [ ]: 1 #####
2 ### RUN WITH CAUTION#####
3 #####
4
5 # Scan through all 40k photos and extract faces
6 start = time.time()
7 for i in range(len(all_jpg)):
8     process_faces(all_jpg[i], NODISPLAY, SAVE)
9 end = time.time()
```

```
In [13]: 1 def get_count_files(folder_addr):
2     n = 0
3     for root, dirs, files in os.walk(folder_addr):
4         for file in files:
5             n = n + 1
6     return n
```

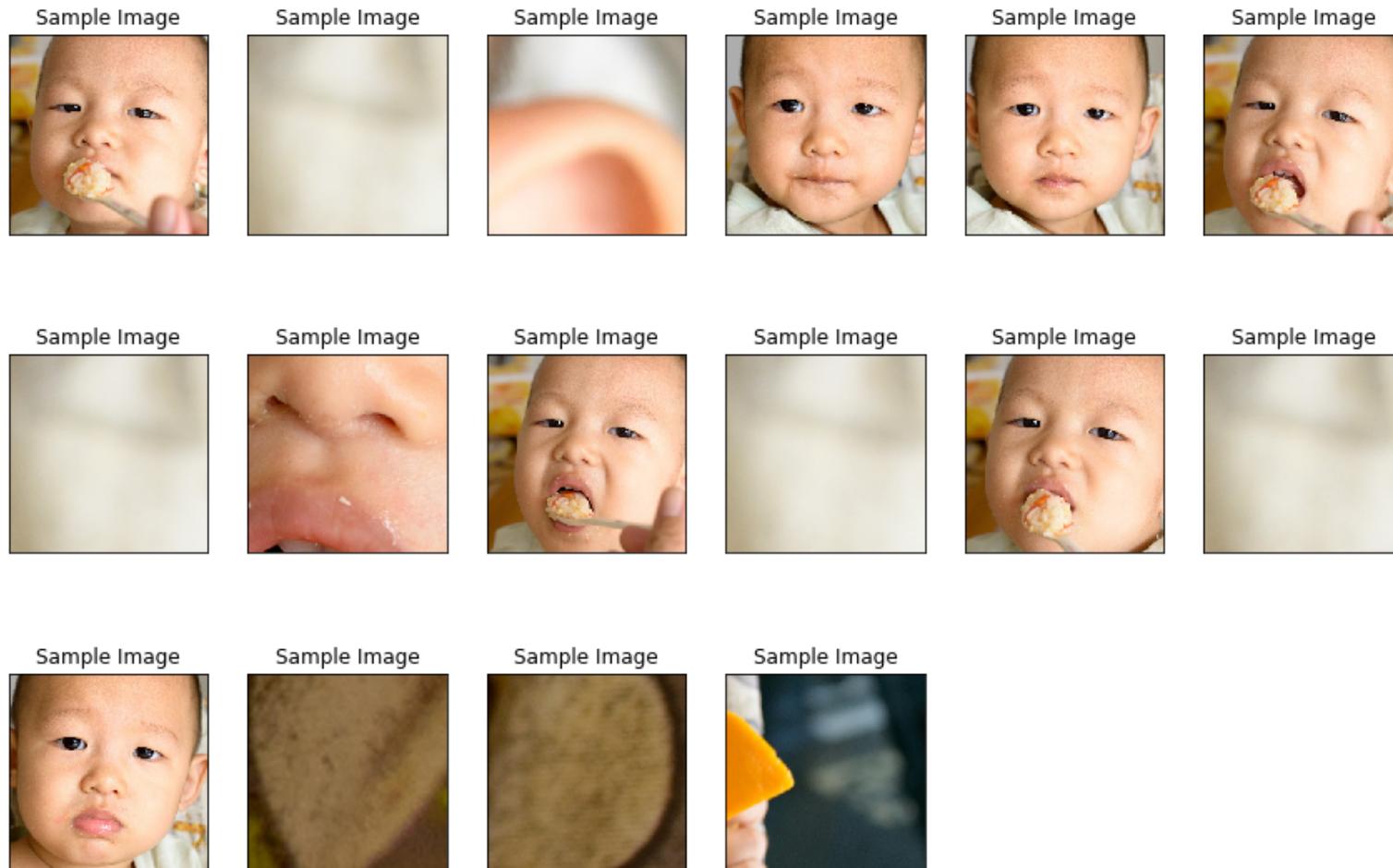
```
In [98]: 1 print('Extracting faces took %s hours' % round((end-start)/3600, 1))
2 print(get_count_files('images'), 'face images extracted')
```

Extracting faces took 13.2 hours
 98059 face images extracted

About 13 hours has been taken and around 100k face images were generated. Below sample images are from one folder. It's quite clear that the openCV is not 100% reliable, some of the photos are not face images.

```
In [97]: 1 image_numpys = get_numpy_from_folder(os.path.join('images', '2017', '2017.10.25 - 棋祺吃饭'))
2 display_from_numpy(image_numpys, fig_dim_x=15, fig_dim_y=10, plot_nrows=3, plot_ncols=6)
```

Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>



B1. Preprocess - Manually Label the Face Images by Putting Them Into Different Folders

There are almost 100k faces extracted. Most of them are irrelevant, and many of them are objects other than faces. I hand-picked about 500 of the face images and saved them into 7 categories/folders.

```
In [11]: 1 categories = os.listdir('./images')
2 face_jpgs = get_file_path_from_folder('./images/')
3 print('Categories:', categories)
4 print('Total number of images:', len(face_jpgs))
```

Categories: ['Brother', 'Dad', 'Daughter', 'Me', 'Mum', 'Son', 'Wife']
 Total number of images: 419

Below are some high level statistics about how many images in each category.

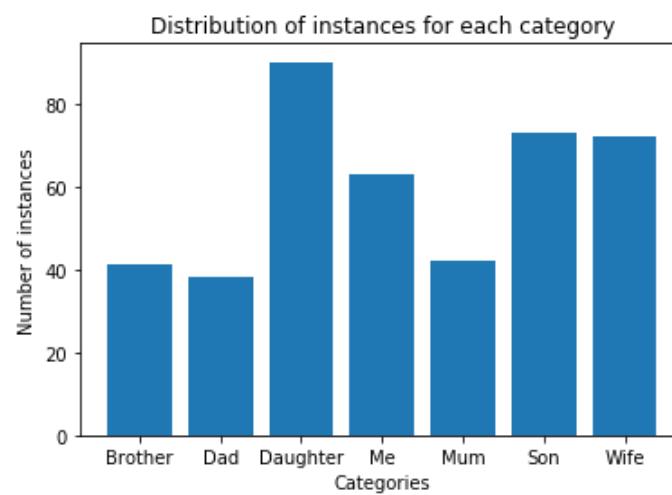
```
In [12]: 1 def count_each_category(categories, files):
2     """Count the number of file for each category.
3     Args:
4         categories (list): A list of categories.
5         files (list[str]): A list of strings which are the relative address of face images.
6     Returns:
7         stats_dict (dict): A dictionary of (category: number).
8     """
9     stats_dict = {}
10    # Initialize the dictionary.
11    for category in categories:
12        stats_dict[category] = 0
13    # Increment the value of the matching item.
14    for file in files:
15        # Convert the path string into Path object.
16        file_path = Path(file)
17        # str(file_path.parent) will return 'images\Brother', need to use os.sep as the delimiter
18        # for cross platform use.
19        file_category = str(file_path.parent).split(os.sep)[1]
20        stats_dict[file_category] += 1
21    return stats_dict
```

```
In [13]: 1 stats_dict = count_each_category(categories, face_jpgs)
2 print(stats_dict)

{'Daughter': 90, 'Dad': 38, 'Brother': 41, 'Son': 73, 'Mum': 42, 'Me': 63, 'Wife': 72}
```

Below bar plot shows the distribution of samples for each category. 'Daughter' category has the most images, 92. 'Son' has the second most, 73, so on and so forth. The distribution of samples are not well balanced. This observation will lead to few special 'treatment' in the later sections.

```
In [14]: 1 import matplotlib.pyplot as plt
2 %matplotlib inline
3 plt.bar(stats_dict.keys(), stats_dict.values())
4 plt.xlabel('Categories')
5 plt.ylabel('Number of instances')
6 plt.title('Distribution of instances for each category')
7 plt.show()
```



B2. Preprocess - Loading the Images and Split Them into Train, Validate and Test Datasets

```
In [106]: 1 # Function to Load the images as a list of file names and one hot code categories
2 from sklearn.datasets import load_files
3 from sklearn.model_selection import train_test_split
4 from keras.utils import np_utils
5 from glob import glob
6
7 # Read all the files and return 2 numpy arrays.
8 # One is the address of the files and the other one is the one hot encode of the category.
9 def load_dataset(folder_addr):
10     """Load all the files in the given directory. The name of each subdirectory will be the category name.
11     Args:
12         folder_addr (str): The folder address in which there are subfolders.
13     Returns:
14         face_files (list[str]): A list of face file address strings.
15         face_targets (numpy.ndarray): Numpy array of categories, value from 0 to 6, without one-hot encoding.
16     """
17     data = load_files(folder_addr)
18     face_files = np.array(data['filenames'])
19     # face_targets = np_utils.to_categorical(np.array(data['target']), 7)
20     face_targets = np.array(data['target'])
21
22     # Many of the times Windows and Mac are producing 'desktop.ini' or '.DS_Store' file automatically.
23     # Need to omit them.
24     invalid_files_idx = []
25     for i in range(len(face_files)):
26         if 'desktop.ini' in face_files[i] or 'DS_Store' in face_files[i]:
27             invalid_files_idx.append(i)
28     face_files = np.delete(face_files, invalid_files_idx)
29     face_targets = np.delete(face_targets, invalid_files_idx).tolist()
30
31     return face_files, face_targets
```

```
In [88]: 1 # Load the List of images and categories
2 faces, targets = load_dataset('./images')
3 face_names = [item[9:-1] for item in glob('./images/*')]
4 print('There are %d face categories.' % len(face_names))
5 print(face_names)
6 print('There are %d total faces.' % len(faces))
7 print(count_each_category(categories, faces))
```

There are 7 face categories.
['Brother', 'Dad', 'Daughter', 'Me', 'Mum', 'Son', 'Wife']
There are 419 total faces.
{'Daughter': 90, 'Dad': 38, 'Brother': 41, 'Son': 73, 'Mum': 42, 'Me': 63, 'Wife': 72}

A random check on the file name and category name.

```
In [17]: 1 def parse_image_category(file_addr, category_code, is_one_hot=False):
2     """Given the category in index or one-hot code, return the index and name of the category of an image.
3     Args:
4         file_addr (str): The address of the image whose name contains the category name.
5         category_code (int or numpy.ndarray): Integer index of the category or numpy array after one-hot encoding.
6         is_one_hot (bool): Default it's False, set to True if the passed in category_code is in one-hot format.
7     Returns:
8         category_index (int): The index of the category, from 0 to 6.
9         face_names[category_index] (str): The name of the category.
10    """
11    # Print file path and category in one hot code
12    print(file_addr, category_code)
13    if is_one_hot:
14        category_index = np.argmax(category_code)
15    else:
16        category_index = category_code
17    return category_index, face_names[category_index]
18
19 parse_image_category(faces[100], targets[100])
```

./images\ Dad\ 20160701-1737.jpg-face-0.jpg 1

Out[17]: (1, 'Dad')

From this point, the data set is finally ready for the coming machine learning pipelines.

C. Extract Feature Vectors

```
In [81]: 1 from keras.applications.inception_v3 import InceptionV3
2 from keras.models import Model
3 from keras.layers import Dense, GlobalAveragePooling2D, Input
4 from keras import backend as K
5 from keras.applications.imagenet_utils import preprocess_input, decode_predictions
6 from keras.callbacks import ModelCheckpoint, EarlyStopping, LambdaCallback, ReduceLROnPlateau
7 from keras.models import load_model
8 from keras.preprocessing import image
```

```
In [227]: 1 def path_to_tensor(img_path):
2     """Given one image path, return it as a numpy array.
3     Args:
4         img_path (str): The full path string of a image.
5     Returns:
6         np.expand_dims(x, axis=0) (numpy.ndarray): A 4D numpy array of a image after expanding from 3D to 4D.
7     """
8     img = image.load_img(img_path, target_size=(299,299))
9     x = image.img_to_array(img)
10    return np.expand_dims(x, axis=0)
11
12 def paths_to_tensor(img_paths):
13     """Given the image paths, return them as a vertically stacked numpy array.
14     Args:
15         img_paths (list[str]): The full paths of the images in a list.
16     Returns:
17         np.vstack(list_of_tensors) (numpy.ndarray): 4D numpy array of all the images, after normalization.
18     """
19     list_of_tensors = [path_to_tensor(img_path) for img_path in img_paths if 'desktop.ini' not in img_path]
20     return np.vstack(list_of_tensors).astype('float32')/255
```

```
In [228]: 1 # Load the inception v3 model, include the dense layers
2 base_model = InceptionV3(weights='imagenet', include_top=True)
3 vector_out = base_model.get_layer('avg_pool')
4 feature_model = Model(inputs=base_model.input, outputs=vector_out.output)
```

```
In [229]: 1 def get_features(feature_model, tensors):
2     """Using the given model to convert a tensor/image to a feature vector
3     Args:
4         feature_model (Keras Model): In this project, it's the Google Inception V3 model without the last dense layer.
5         tensors (numpy.ndarray): Group of images in a 4D numpy array.
6     Returns:
7         feature_outputs (numpy.ndarray): Feature vectors of the group of images, dimension of (x, 2048)
8     """
9     tensor_inputs = np.expand_dims(tensors, axis=0)
10    feature_outputs = feature_model.predict(tensor_inputs)
11    return feature_outputs
```

Split the dataset into train, validate and test datasets. The data are not well balanced based on the earlier bar plot diagram. Hence, stratified split function will be used here.

```
In [230]: 1 train_faces, test_faces, train_targets, test_targets = train_test_split(faces, targets, test_size=0.15, random_state=1, stratify=targets)
2 train_faces, validate_faces, train_targets, validate_targets = train_test_split(train_faces, train_targets, test_size=0.2, random_state=1, stratify=targets)
3 print('There are %d training faces.' % len(train_faces))
4 print(count_each_category(categories, train_faces))
5 print('There are %d validate faces.' % len(validate_faces))
6 print(count_each_category(categories, validate_faces))
7 print('There are %d test faces.' % len(test_faces))
8 print(count_each_category(categories, test_faces))
```

There are 284 training faces.
{'Daughter': 61, 'Dad': 25, 'Brother': 28, 'Son': 49, 'Mum': 29, 'Me': 43, 'Wife': 49}
There are 72 validate faces.
{'Daughter': 15, 'Dad': 7, 'Brother': 7, 'Son': 13, 'Mum': 7, 'Me': 11, 'Wife': 12}
There are 63 test faces.
{'Daughter': 14, 'Dad': 6, 'Brother': 6, 'Son': 11, 'Mum': 6, 'Me': 9, 'Wife': 11}

```
In [231]: 1 # Read the images as numpy arrays
2 train_tensors = paths_to_tensor(train_faces)
3 test_tensors = paths_to_tensor(test_faces)
4 validate_tensors = paths_to_tensor(validate_faces)
5 print("Train tensor shape.", train_tensors.shape)
6 print('Test tensor shape.', test_tensors.shape)
7 print('Validate tensor shape.', validate_tensors.shape)
```

Train tensor shape. (284, 299, 299, 3)
Test tensor shape. (63, 299, 299, 3)
Validate tensor shape. (72, 299, 299, 3)

```
In [232]: 1 train_features = get_features(feature_model, train_tensors)
2 validate_features = get_features(feature_model, validate_tensors)
3 test_features = get_features(feature_model, test_tensors)
4 print('Train features shape:', train_features.shape, '\nValidate feature_modelres shape: ', validate_features.shape,
5      '\nTest features shape:', test_features.shape)
```

Train features shape: (284, 2048)
Validate feature_modelres shape: (72, 2048)
Test features shape: (63, 2048)

D. Different Models - Attempt to Train and Test on Various Models

'Traditional' machine learning models of SGDClassifier, LogisticRegression, KNeighborsClassifier will be explored first. Then followed by deep neural network.

a). 'Traditional' Models

```
In [233]: 1 from sklearn.linear_model import SGDClassifier, LogisticRegression
2 from sklearn.neighbors import KNeighborsClassifier
3 clf_sgd = SGDClassifier(random_state=0)
4 clf_knn = KNeighborsClassifier(n_neighbors=7)
5 clf_log = LogisticRegression(random_state=0)
```

```
In [234]: 1 %time model_sgd = clf_sgd.fit(train_features, train_targets)
2 %time model_knn = clf_knn.fit(train_features, train_targets)
3 %time model_log = clf_log.fit(train_features, train_targets)
```

Wall time: 23 ms
Wall time: 17 ms
C:\Users\Xiaowei\Anaconda3\envs\tfkeras\lib\site-packages\sklearn\linear_model\stochastic_gradient.py:128: FutureWarning: max_iter and tol parameters have been added in <class 'sklearn.linear_model.stochastic_gradient.SGDClassifier'> in 0.19. If both are left unset, they default to max_iter=5 and tol=None. If tol is not None, max_iter defaults to max_iter=1000. From 0.21, default max_iter will be 1000, and default tol will be 1e-3.
"and default tol will be 1e-3." % type(self), FutureWarning)

Wall time: 470 ms

```
In [239]: 1 # start = time.time()
2 %time predict_test_sgd = model_sgd.predict(test_features)
3 %time predict_test_knn = model_knn.predict(test_features)
4 %time predict_test_log = model_log.predict(test_features)
5 # end = time.time()
6 # print('%.2gs' %(end - start))
```

Wall time: 2 ms
Wall time: 63.1 ms
Wall time: 1 ms

Due to the imbalanced data as mentioned earlier, the average of precision, recall and fbeta score will be set as weighted .

```
In [240]: 1 from sklearn.metrics import fbeta_score, precision_recall_fscore_support, confusion_matrix
2 import pandas as pd
3 def get_score_numpy(test_targets, predict_targets, average = 'weighted'):
4     score = precision_recall_fscore_support(test_targets, predict_targets, average=average)
5     score = np.array(score)
6     score = score[score != None]
7     return score
```

```
In [241]: 1 score_sgd = get_score_numpy(test_targets, predict_test_sgd)
2 score_knn = get_score_numpy(test_targets, predict_test_knn)
3 score_log = get_score_numpy(test_targets, predict_test_log)
```

```
In [242]: 1 data = pd.DataFrame(np.stack((score_sgd, score_knn, score_log)),
2                         columns=['Precision', 'Recall', 'F1'], index=['SGD', 'KNN', 'Log'])
3 data
```

Out[242]:

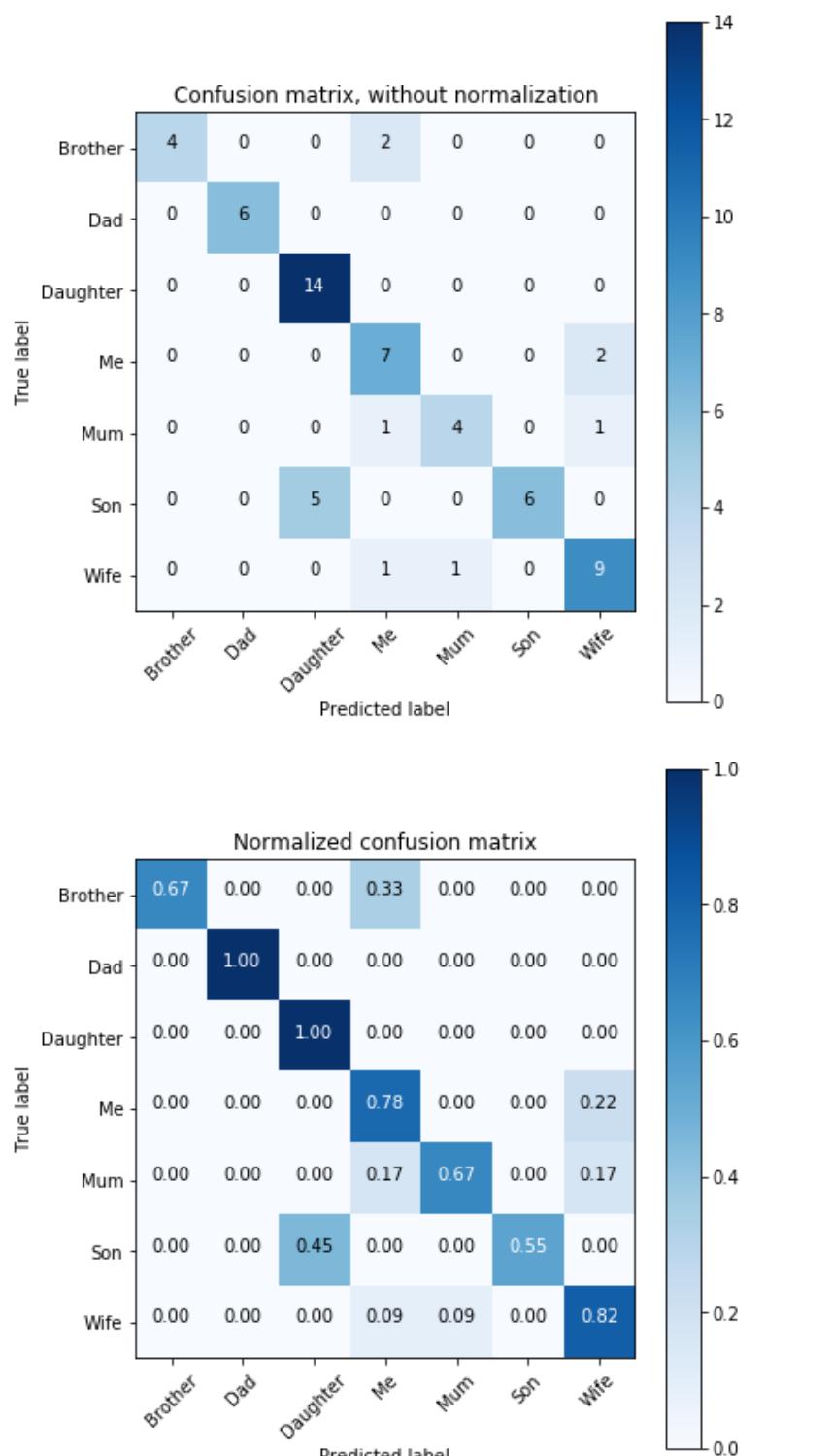
	Precision	Recall	F1
SGD	0.792347	0.777778	0.775751
KNN	0.717082	0.68254	0.683151
Log	0.826874	0.793651	0.78914

Based on above table, logistic regression performed the best based on every angel of precision, recall and f1 scores. It is fast on testing but not training.

Below confusion matrix plots provide more intuitive ways of how good or bad the logistic regression model is doing for each label. The first plot is based on the absolute number of samples. The second plot is after normalization, which shows a more accurate estimation.

```
In [197]: 1 import itertools
2 def plot_confusion_matrix(cm, classes,
3                           normalize=False,
4                           title='Confusion matrix',
5                           cmap=plt.cm.Blues):
6     """
7         This function prints and plots the confusion matrix.
8         Normalization can be applied by setting `normalize=True`.
9     """
10    if normalize:
11        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
12    #         print("Normalized confusion matrix")
13    else:
14        pass
15    #         print('Confusion matrix, without normalization')
16
17    #     print(cm)
18
19    plt.imshow(cm, interpolation='nearest', cmap=cmap)
20    plt.title(title)
21    plt.colorbar()
22    tick_marks = np.arange(len(classes))
23    plt.xticks(tick_marks, classes, rotation=45)
24    plt.yticks(tick_marks, classes)
25
26    fmt = '.2f' if normalize else 'd'
27    thresh = cm.max() / 2.
28    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
29        plt.text(j, i, format(cm[i, j], fmt), horizontalalignment="center", color="white" if cm[i, j] > thresh else "black")
30
31    plt.tight_layout()
32    plt.ylabel('True label')
33    plt.xlabel('Predicted label')
```

```
In [195]: 1 cnf_matrix = confusion_matrix(test_targets, predict_test_log)
2 plt.figure(figsize=(6,6))
3 plot_confusion_matrix(cnf_matrix, classes=face_names, title='Confusion matrix, without normalization')
4 plt.figure(figsize=(6,6))
5 plot_confusion_matrix(cnf_matrix, classes=face_names, normalize=True, title='Normalized confusion matrix')
6 plt.show()
```



b). Deep Learning Model

So far all the chosen models performed training and prediction in fractions of a second. The best accuracy is about 80% from logistic regression while KNN only got the lowest 68%, even though it's much better than random guess of $1/7 = 16.7\%$. Below is an exploration on using DNN model. After a few tries I decided to have 3 dense layers as it provides a balance between accuracy and time.

```
In [97]: 1 Inp = Input(shape=(2048,))
2 x = Dense(300, activation='relu')(Inp)
3 x = Dense(50, activation='relu')(x)
4 output = Dense(7, activation='softmax')(x)
5 model_dnn = Model(inputs=Inp, outputs=output)
6 model_dnn.summary()
```

Layer (type)	Output Shape	Param #
input_3 (InputLayer)	(None, 2048)	0
dense_1 (Dense)	(None, 300)	614700
dense_2 (Dense)	(None, 50)	15050
dense_3 (Dense)	(None, 7)	357

Total params: 630,107
Trainable params: 630,107
Non-trainable params: 0

As a classification problem, deep learning models will require the output in one-hot format. If not, the class 1 and 2 will be misunderstood as ordinal instead of nominal.

```
In [98]: 1 train_targets_one_hot = np_utils.to_categorical(train_targets)
2 validate_targets_one_hot = np_utils.to_categorical(validate_targets)
3 test_targets_one_hot = np_utils.to_categorical(test_targets)
```

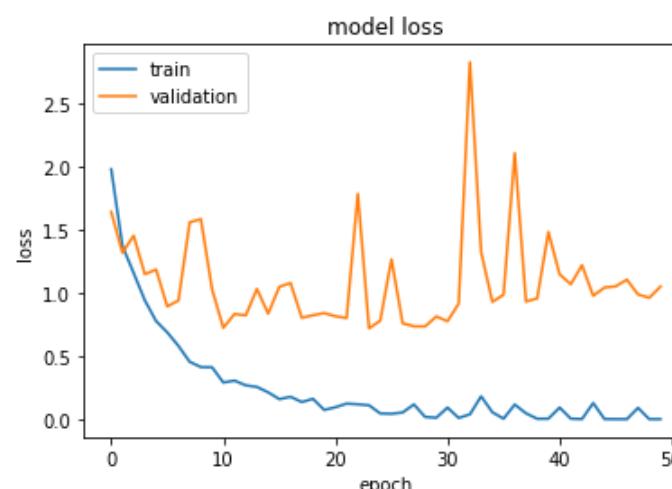
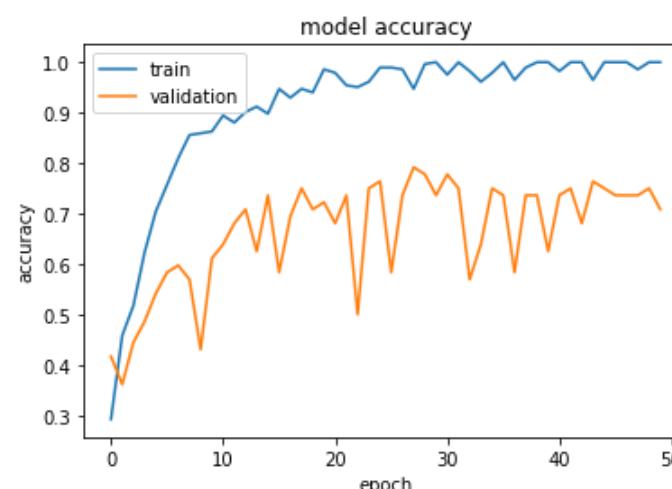
```
In [99]: 1 model_dnn.compile(optimizer='rmsprop', loss='categorical_crossentropy', metrics=['accuracy'])
2 model_dnn.save_weights('init_weights.h5')
3 checkpointer = ModelCheckpoint(filepath='model_dnn.h5', verbose=1, save_best_only=True)
```

```
In [33]: 1 %%time
2 model_dnn.load_weights('init_weights.h5')
3 hist_1 = model_dnn.fit(train_features, train_targets_one_hot,
4                         validation_data=(validate_features, validate_targets_one_hot),
5                         epochs=50, verbose=1, batch_size=20,
6                         callbacks=[checkpointer])
```

Train on 284 samples, validate on 72 samples
Epoch 1/50
200/284 [=====>.....] - ETA: 0s - loss: 2.2020 - acc: 0.2650Epoch 00000: val_loss improved from inf to 1.6452
7, saving model to model_dnn.h5
284/284 [=====] - 0s - loss: 1.9837 - acc: 0.2923 - val_loss: 1.6453 - val_acc: 0.4167
Epoch 2/50
220/284 [=====>.....] - ETA: 0s - loss: 1.4043 - acc: 0.4682Epoch 00001: val_loss improved from 1.64527 to 1.
32060, saving model to model_dnn.h5
284/284 [=====] - 0s - loss: 1.3760 - acc: 0.4577 - val_loss: 1.3206 - val_acc: 0.3611
Epoch 3/50
200/284 [=====>.....] - ETA: 0s - loss: 1.2114 - acc: 0.5150Epoch 00002: val_loss did not improve
284/284 [=====] - 0s - loss: 1.1602 - acc: 0.5176 - val_loss: 1.4567 - val_acc: 0.4444
Epoch 4/50
200/284 [=====>.....] - ETA: 0s - loss: 0.9996 - acc: 0.6200Epoch 00003: val_loss improved from 1.32060 to 1.
15075, saving model to model_dnn.h5
284/284 [=====] - 0s - loss: 0.9435 - acc: 0.6232 - val_loss: 1.1507 - val_acc: 0.4861
Epoch 5/50
200/284 [=====>.....] - ETA: 0s - loss: 0.8163 - acc: 0.6700Epoch 00004: val_loss did not improve
284/284 [=====] - 0s - loss: 0.7795 - acc: 0.7042 - val_loss: 1.1881 - val_acc: 0.5417
Epoch 5/50

```
In [34]: 1 ## TODO: Visualize the training and validation Loss of your neural network
2 import matplotlib.pyplot as plt
3 def plt_hist(hist):
4     print(hist.history.keys())
5     # summarize history for accuracy
6     plt.plot(hist.history['acc'])
7     plt.plot(hist.history['val_acc'])
8     plt.title('model accuracy')
9     plt.ylabel('accuracy')
10    plt.xlabel('epoch')
11    plt.legend(['train', 'validation'], loc='upper left')
12    plt.show()
13    # summarize history for loss
14    plt.plot(hist.history['loss'])
15    plt.plot(hist.history['val_loss'])
16    plt.title('model loss')
17    plt.ylabel('loss')
18    plt.xlabel('epoch')
19    plt.legend(['train', 'validation'], loc='upper left')
20    plt.show()
```

```
In [35]: 1 plt_hist(hist_1)
dict_keys(['val_loss', 'acc', 'loss', 'val_acc'])
```



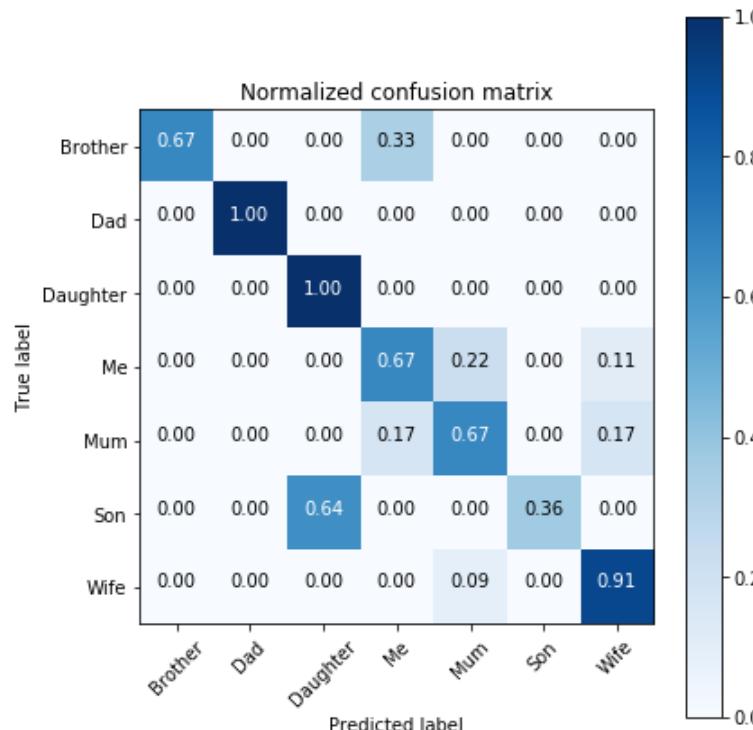
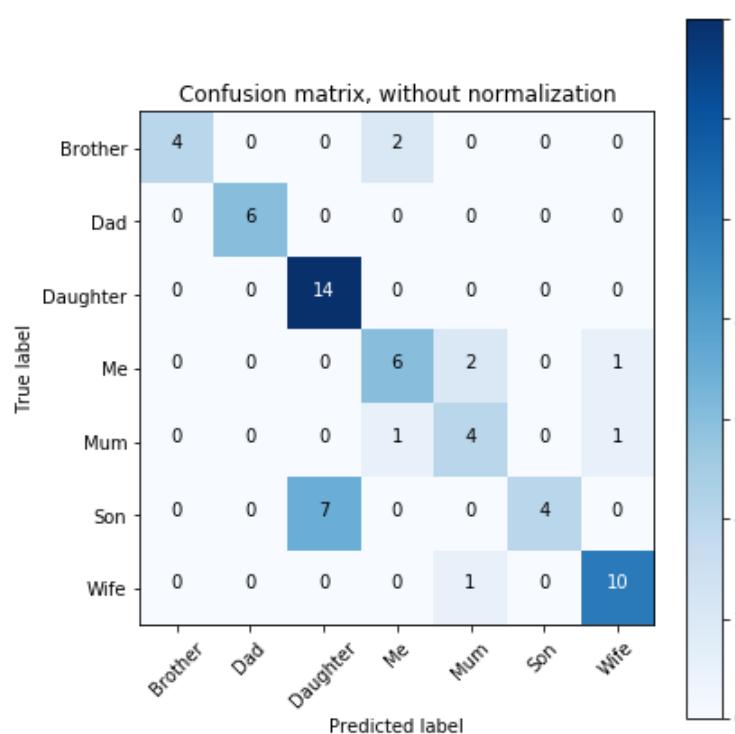
Epoch of 50 looks like a reasonable number of iterations. The validation accuracy is stable and the loss is starting to grow, that's a sign of overfitting. Techniques of reducing learning rate, early stopping could be further explored, but not covered in this project.

Below is the metrics of the DNN model.

```
In [186]: 1 model_dnn = load_model('./model_dnn.h5')
2 %time predict_test_dnn = model_dnn.predict(test_features)
3 predict_test_dnn = predict_test_dnn.argmax(axis=-1)
4 print('Precision, Recall, F1')
5 print(get_score_numpy(test_targets, predict_test_dnn))
```

Wall time: 217 ms
Precision, Recall, F1
[0.80839002267573701 0.76190476190476186 0.74800304365521764]

```
In [187]: 1 cnf_matrix = confusion_matrix(test_targets, predict_test_dnn)
2 plt.figure(figsize=(6,6))
3 plot_confusion_matrix(cnf_matrix, classes=face_names, title='Confusion matrix, without normalization')
4 plt.figure(figsize=(6,6))
5 plot_confusion_matrix(cnf_matrix, classes=face_names, normalize=True, title='Normalized confusion matrix')
6 plt.show()
```



Based on the current split of the dataset, logistic regression performed the best. KNN reached merely 70% while SGD classifier, logistic regression and DNN reached accuracy of 75% to 80%. DNN took significant longer time on training, as well as prediction.

But what if it's just a coincidence, what if I had more images? The next section will look into the refinement of the models from two ways.

The first one is to use image generator to have a bigger dataset. The other one will focus on using cross validation to determine the best model.

IV. Refinement

A. Create More Images

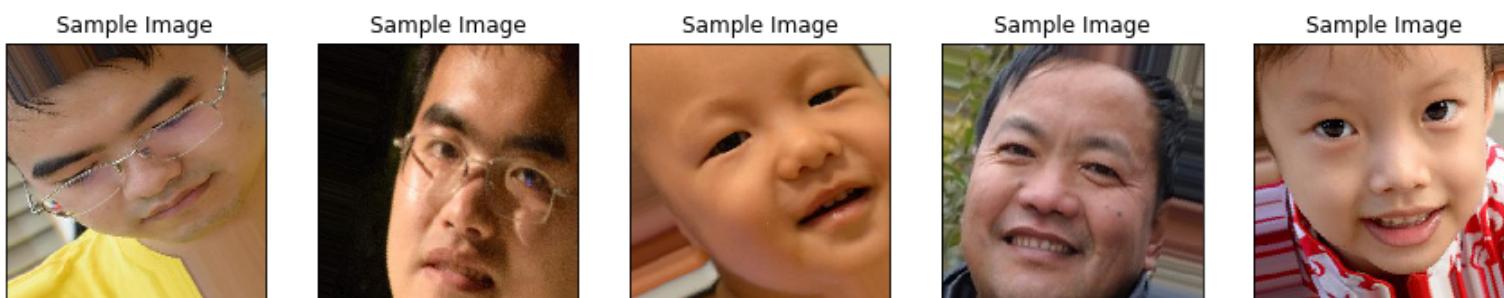
Use [image generator](https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html) (<https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html>) to create more images. This is a very useful technique when having little data.

In this project, I put the new images in folder `images2` and 10 new images will be generated for each original image. Generating new images is basically to do some distortion on the images. I think it's reasonable to narrow, widen, rotate, zoom and flip the images a little, in a reasonable range. Hence, values of the most parameters I set to 20%.

```
In [78]: 1 from keras.preprocessing.image import ImageDataGenerator, array_to_img, img_to_array, load_img
2 import shutil
3
4 datagen = ImageDataGenerator(
5     rotation_range=40,
6     width_shift_range=0.2,
7     height_shift_range=0.2,
8     shear_range=0.2,
9     zoom_range=0.2,
10    horizontal_flip=True,
11    fill_mode='nearest')
12
13 def get_target_dir(face_file):
14     """ Given the address of the original image, create directory for the new generated images if needed.
15         Also return the new directory address.
16     Args:
17         face_file (str): The address of the original face file, may or may not be in full address.
18     Returns:
19         target_dir (str): The address of the directory for the generated images.
20     """
21     target_file = face_file.replace('images', 'images2')
22     target_dir = os.path.dirname(target_file)
23     target_path = Path(target_dir)
24
25     # Create parents of directory, don't raise exception if the directory exists
26     target_path.mkdir(parents=True, exist_ok=True)
27     return target_dir
28
29 def generate_images(faces):
30     """ Given the address of face images, use image generator to generate new images.
31     Args:
32         faces ([str]): The list of face images addresses.
33     Returns:
34         None: Generate new images and return None.
35     """
36     if os.path.exists('images2'):
37         shutil.rmtree('images2', ignore_errors=True)
38         # sleep for 2 seconds to allow OS finish the previous deletion action.
39         time.sleep(2)
40     for n, face in enumerate(faces):
41         if n % 100 == 0:
42             print('Generating images, ' + str(n) + ' of ' + str(len(faces)) + ' faces')
43         target_dir = get_target_dir(face)
44
45         img = load_img(face)
46         x = img_to_array(img) # this is a Numpy array with shape (3, 150, 150)
47         x = x.reshape((1,) + x.shape) # this is a Numpy array with shape (1, 3, 150, 150)
48
49         # the .flow() command below generates batches of randomly transformed images
50         # and saves the results to the target directory
51         i = 0
52         for batch in datagen.flow(x, batch_size=1, save_to_dir=target_dir, save_prefix='gen', save_format='jpg'):
53             i += 1
54             if i > 10:
55                 break # otherwise the generator would loop indefinitely
```

```
In [79]: 1 image_npys = get_numpy_from_folder('sample_generated')
2 display_from_numpy(image_npys, fig_dim_x=15, fig_dim_y=5, plot_nrows=1, plot_ncols=5)
```

Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>
 Image numpy array shape: (299, 299, 3) <class 'numpy.ndarray'>



B. With Stratified K-fold CV

For a more 'stable' performance of each model, K-fold CV will be used in this project. So that each and every image will have the chance to be used for both training and testing. In this way, we will have more 'averaged' and thus more 'stable' performance evaluation.

```
In [77]: 1 from sklearn.model_selection import StratifiedKFold
2 n_splits = 10
3 skf = StratifiedKFold(n_splits=n_splits)
4 skf.get_n_splits(faces, targets)
```

Out[77]: 10

In [104]:

```
1 def perform_training(clf, features, targets):
2     ''' Perform training, return the trained model and time taken in seconds.
3     Args:
4         clf (sklearn-model): The sklearn model.
5         features (numpy.array): Training input, 4D numpy array.
6         targets (numpy.array): Training output, 1D numpy array.
7     Returns:
8         model (sklean-model): The trained model.
9         end - start: Time taken, in seconds.
10    ...
11    start = time.time()
12    model = clf.fit(features, targets)
13    end = time.time()
14    return model, round(end - start, 3)
15
16 def perform_testing(model, features):
17     ''' Perform testing, return the accuracy and time taken in seconds.
18     Args:
19         model (sklearn-model): The sklearn model.
20         features (numpy.array): Testing input, 4D numpy array.
21     Returns:
22         prediction ([int]): The list of predicted labels.
23         end - start: Time taken, in seconds.
24    ...
25    start = time.time()
26    predictions = model.predict(features)
27    end = time.time()
28    return predictions, round(end - start, 3)
29
30 def perform_dnn_training(model, features, targets):
31     ''' Perform training, return the trained model and time taken in seconds.
32     Args:
33         clf (sklearn-model): The sklearn model.
34         features (numpy.array): Training input, 4D numpy array.
35         targets (numpy.array): Training output, 1D numpy array.
36     Returns:
37         model (sklean-model): The trained model.
38         end - start: Time taken, in seconds.
39    ...
40     targets = np_utils.to_categorical(targets)
41     model.compile(optimizer='rmsprop', loss='categorical_crossentropy', metrics=['accuracy'])
42     start = time.time()
43     model.fit(features, targets, epochs=50, verbose=0, batch_size=20)
44     end = time.time()
45     return model, round(end - start, 3)
46
47 def perform_dnn_testing(model, features):
48     ''' Perform testing, return the accuracy and time taken in seconds.
49     Args:
50         model (keras.model): The keras model
51         features (numpy.array): Testing input, 4D numpy array.
52     Returns:
53         predictions ([int]): The list of predicted labels.
54         end - start: Time taken, in seconds.
55    ...
56    start = time.time()
57    probas = model.predict(features)
58    predictions = probas.argmax(axis=-1)
59    end = time.time()
60    return predictions, round(end - start, 3)
```

```
In [149]: 1 all_test_targets = []
2 all_predict_targets = []
3 all_train_times = []
4 all_test_times = []
5 n = 0
6 for train_index, test_index in skf.split(faces, targets):
7     n = n + 1
8     print('Round ' + str(n) + ' of ' + str(n_splits))
9     x_train, x_test = faces[train_index], faces[test_index]
10    y_train, y_test = targets[train_index], targets[test_index]
11
12    generate_images(x_train)
13    # After generating the images, reload the training dataset, testing dataset remains the same
14    x_train, y_train = load_dataset('./images2')
15    face_names = [item[10:-1] for item in glob('./images2/*')]
16
17    x_train_tensors = paths_to_tensor(x_train)
18    x_test_tensors = paths_to_tensor(x_test)
19    x_train_features = get_features(feature_model, x_train_tensors)
20    x_test_features = get_features(feature_model, x_test_tensors)
21
22    # Define and initialize models.
23    clf_sgd = SGDClassifier(random_state=0)
24    clf_knn = KNeighborsClassifier(n_neighbors=7)
25    clf_log = LogisticRegression(random_state=0)
26    predict_targets = []
27    train_times = []
28    test_times = []
29    for clf in [clf_sgd, clf_knn, clf_log]:
30        model, train_time = perform_training(clf, x_train_features, y_train)
31        cur_predict_targets, test_time = perform_testing(model, x_test_features)
32        all_predict_targets.append(cur_predict_targets)
33        all_test_targets.append(y_test)
34        train_times.append(train_time)
35        test_times.append(test_time)
36
37    # DNN model must use this way to 're-initialize' the weights. The 'init_weights.h5' was
38    # created in the earlier section for the first try of DNN.
39    model_dnn.load_weights('init_weights.h5')
40    model_dnn, train_time = perform_dnn_training(model_dnn, x_train_features, y_train)
41    cur_predict_targets, test_time = perform_dnn_testing(model_dnn, x_test_features)
42    all_predict_targets.append(cur_predict_targets)
43    all_test_targets.append(y_test)
44    train_times.append(train_time)
45    test_times.append(test_time)
46
47    all_train_times.append(train_times)
48    all_test_times.append(test_times)
```

Round 1 of 10
Generating images, 0 of 373 faces
Generating images, 100 of 373 faces
Generating images, 200 of 373 faces
Generating images, 300 of 373 faces

C:\Users\Xiaowei\Anaconda3\envs\tfkeras\lib\site-packages\sklearn\linear_model\stochastic_gradient.py:128: FutureWarning: max_iter and tol parameters have been added in <class 'sklearn.linear_model.stochastic_gradient.SGDClassifier'> in 0.19. If both are left unset, they default to max_iter=5 and tol=None. If tol is not None, max_iter defaults to max_iter=1000. From 0.21, default max_iter will be 1000, and default tol will be 1e-3.
"and default tol will be 1e-3." % type(self)), FutureWarning)

V. Results and Benchmark

In this section, results from the refinement section will be analyzed.

A. Pre-processing

Convert the predicted targets and test targets into a list with size 4.

```
In [212]: 1 print('The length of all_predict_targets:', len(all_predict_targets))
2 print('Shape of the element in all_predict_targets:', all_predict_targets[0].shape)
```

The length of all_predict_targets: 40
Shape of the element in all_predict_targets: (46,)

```
In [213]: 1 flat_predict_targets = all_predict_targets.copy()
2 flat_test_targets = all_test_targets.copy()
3 for i in range(4, len(flat_predict_targets)):
4     r = i % 4
5     flat_predict_targets[r] = np.concatenate((flat_predict_targets[r], flat_predict_targets[i]))
6     flat_test_targets[r] = np.concatenate((flat_test_targets[r], flat_test_targets[i]))
7
8 flat_predict_targets = flat_predict_targets[:4]
9 flat_test_targets = flat_test_targets[:4]
```

```
In [214]: 1 score_sgd = get_score_numpy(flat_test_targets[0], flat_predict_targets[0])
2 score_knn = get_score_numpy(flat_test_targets[1], flat_predict_targets[1])
3 score_log = get_score_numpy(flat_test_targets[2], flat_predict_targets[2])
4 score_dnn = get_score_numpy(flat_test_targets[3], flat_predict_targets[3])
```

```
In [215]: 1 cnf_matrix_sgd = confusion_matrix(flat_test_targets[0], flat_predict_targets[0])
2 cnf_matrix_knn = confusion_matrix(flat_test_targets[1], flat_predict_targets[1])
3 cnf_matrix_log = confusion_matrix(flat_test_targets[2], flat_predict_targets[2])
4 cnf_matrix_dnn = confusion_matrix(flat_test_targets[3], flat_predict_targets[3])
```

B. Summary of the Performance

```
In [219]: 1 score = pd.DataFrame(np.stack((score_sgd, score_knn, score_log, score_dnn)),
2                           columns=['Precision', 'Recall', 'F1'], index=['SGD', 'KNN', 'Log', 'DNN'])
3 score
```

```
Out[219]:
```

	Precision	Recall	F1
SGD	0.795078	0.792363	0.791667
KNN	0.732169	0.72315	0.72579
Log	0.834699	0.828162	0.825837
DNN	0.800736	0.794749	0.79257

```
In [257]: 1 np_train_times = np.vstack(all_train_times)
2 np_test_times = np.vstack(all_test_times)
```

```
In [258]: 1 np_train_times
```

```
Out[258]: array([[ 0.328,   0.742,  13.27 ,  47.497],
 [ 0.303,   0.828,  13.486,  48.112],
 [ 0.309,   0.813,  13.3  ,  46.077],
 [ 0.302,   0.762,  12.771,  45.417],
 [ 0.309,   0.767,  13.059,  47.12 ],
 [ 0.31 ,   0.775,  12.868,  46.597],
 [ 0.307,   0.771,  12.797,  46.33 ],
 [ 0.315,   0.779,  12.722,  47.501],
 [ 0.314,   0.77 ,  13.565,  46.88 ],
 [ 0.305,   0.744,  13.371,  46.109]])
```

```
In [266]: 1 time = pd.DataFrame(np.stack((np.mean(np_train_times, axis=0), np.mean(np_test_times, axis=0))),
2                           columns=['SGD', 'KNN', 'Log', 'DNN'], index=['Train (seconds)', 'Test (seconds)').T
3 print('Average time for training and testing.')
4 time
```

Average time for training and testing.

```
Out[266]:
```

	Train (seconds)	Test (seconds)
SGD	0.3102	0.0004
KNN	0.7751	0.5107
Log	13.1209	0.0003
DNN	46.7640	0.3856

The linear model (SGDClassifier) is doing quite ok in the first try and refinement section. It is almost the fastest one on both training and testing. And its performance is the baseline of this project's benchmark.

KNN doesn't perform well in terms of accuracy (F1 score). I guess it's due to the [curse of dimensionality](https://en.wikipedia.org/wiki/Curse_of_dimensionality#Nearest_neighbor_search) (https://en.wikipedia.org/wiki/Curse_of_dimensionality#Nearest_neighbor_search). It also takes a significant longer time on testing.

Logistic regression did a better job on accuracy (F1 score) but took a longer time on training.

DNN performed almost the same in terms of F1 score. But it took significant longer time on both training and testing.

C. Detailed Look into Score

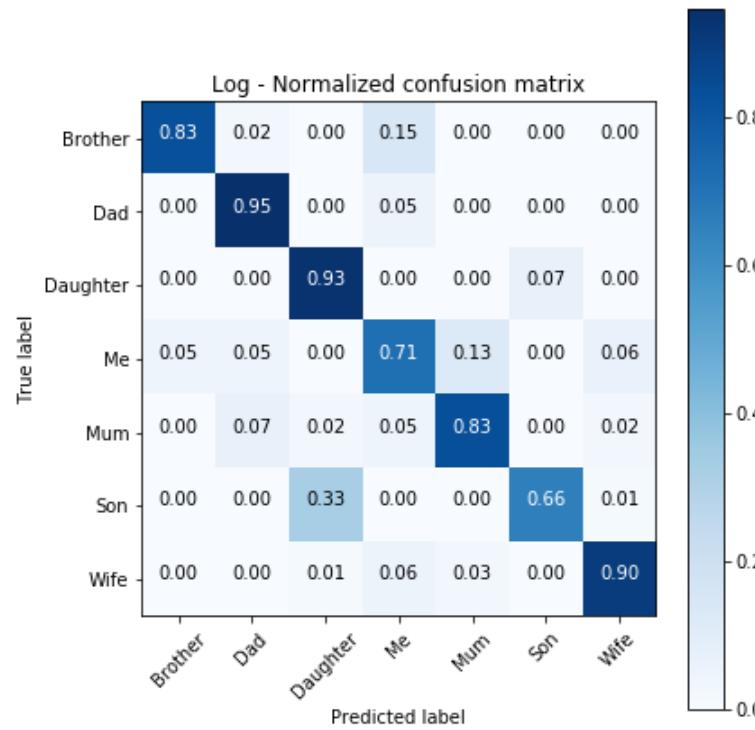
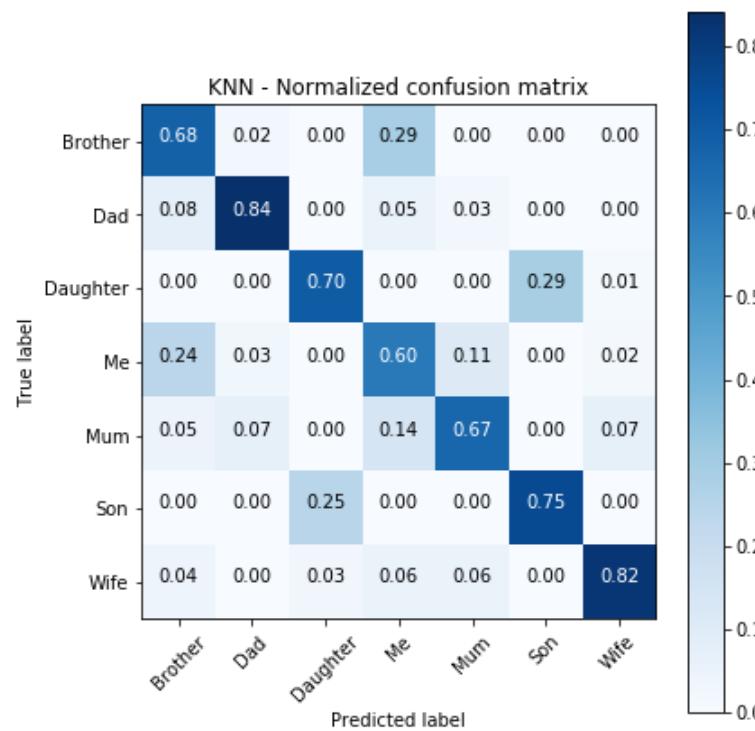
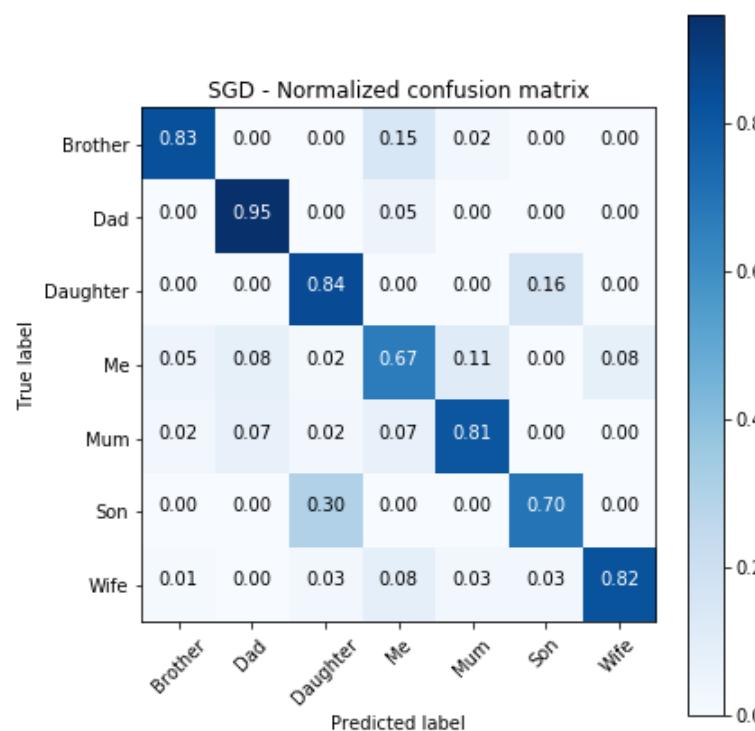
Logistic Regression returned the highest 0.83 of F1 score. And below is the confusion matrix plots (after normalization) of the 4 models. From them we can clearly see the precision and recall rates.

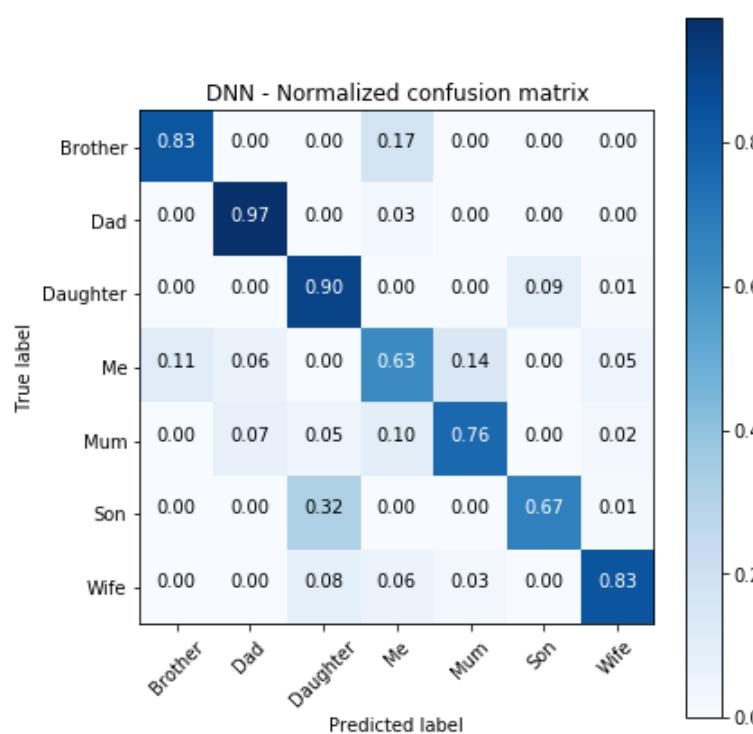
In [226]:

```

1 plt.figure(figsize=(6,6))
2 plot_confusion_matrix(cnf_matrix_sgd, classes=face_names, normalize=True, title='SGD - Normalized confusion matrix')
3
4 plt.figure(figsize=(6,6))
5 plot_confusion_matrix(cnf_matrix_knn, classes=face_names, normalize=True, title='KNN - Normalized confusion matrix')
6
7 plt.figure(figsize=(6,6))
8 plot_confusion_matrix(cnf_matrix_log, classes=face_names, normalize=True, title='Log - Normalized confusion matrix')
9
10 plt.figure(figsize=(6,6))
11 plot_confusion_matrix(cnf_matrix_dnn, classes=face_names, normalize=True, title='DNN - Normalized confusion matrix')
12
13 plt.show()

```

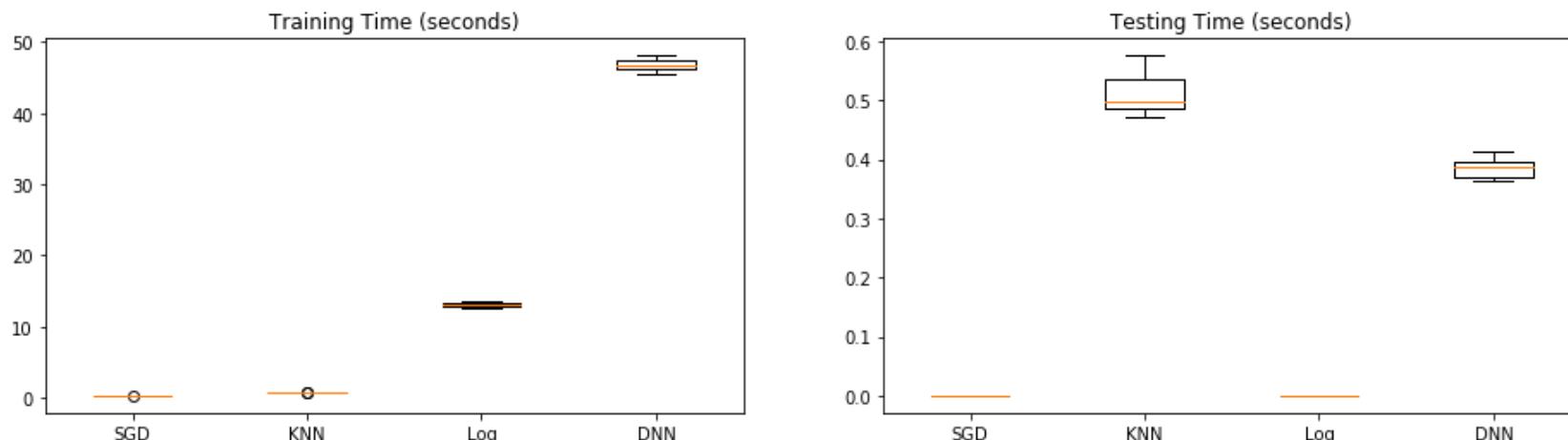




All the models seem to have slight difficulty to distinguish between my son and daughter.

D. Detailed Look into Time

```
In [224]: 1 fig = plt.figure(figsize=(16, 4))
2 ax = fig.add_subplot(1, 2, 1)
3 ax.boxplot(np_train_times)
4 ax.set_title('Training Time (seconds)')
5 ax.set_xticklabels(['SGD', 'KNN', 'Log', 'DNN'])
6
7 ax = fig.add_subplot(1, 2, 2)
8 ax.boxplot(np_test_times)
9 ax.set_title('Testing Time (seconds)')
10 ax.set_xticklabels(['SGD', 'KNN', 'Log', 'DNN'])
11 plt.show()
```



SGD Classifier is the fastest on both training and testing. KNN is fast on training but it took the longest time for testing because it needs to compute the distance to all the training points during testing. Logistic Regression takes a long time on training but the testing is super fast and indeed it's fast on testing. DNN took the longest time for training and its testing time is the 2nd longest.

VI. Conclusion

Among all the 4 models, logistic regression is the one giving highest accuracy and also acceptable training and testing time.

KNN doesn't perform so well in terms of accuracy. I guess it's due to the curse of high dimensionality (https://en.wikipedia.org/wiki/Curse_of_dimensionality#Nearest_neighbor_search).

The DNN model in this project is exactly a transfer learning approach. Its accuracy is quite good, but it takes a long time to train and it's also costly. In this project I'm using GTX 1070 Ti, a GPU costs around USD 500. If the training is under the same CPU environment, I won't doubt it'll take at least 10 more times of current time.

The accuracy of 79% from SGD classifier is the base of the benchmark. Logistic regression model has an average accuracy of almost 83% which pushed the boundary by few percent. The only cost is that logistic regression model takes longer time to train. If time is a big concern, probably SGD classifier is the best in terms of both accuracy and time. Otherwise, logistic regression still remains the best accuracy.

I will choose logistic regression with default parameters (or refer to below code chunk) as my final winning model in this project. It doesn't show a perfect result on the accuracy, especially there are only 7 categories in this project. I doubt it'll work well in the real-world scenario that people need to perform face recognition among thousands to millions of faces. However, it still shows the effectiveness of feature engineering by leveraging transfer learning and applying 'traditional' machine learning models.

One possible way to improve the model may be using transfer learning to detect the facial key points (centre and corners of eyes, mouth, nose and etc.) first. Then build mathematic models to represent the relative positions of the key points. Normalization may be needed to make the face centred and 'staring' at the camera. Then the recognition problem will become to compute similarity of the key points map.

```
In [271]: 1 clf_log.get_params
```

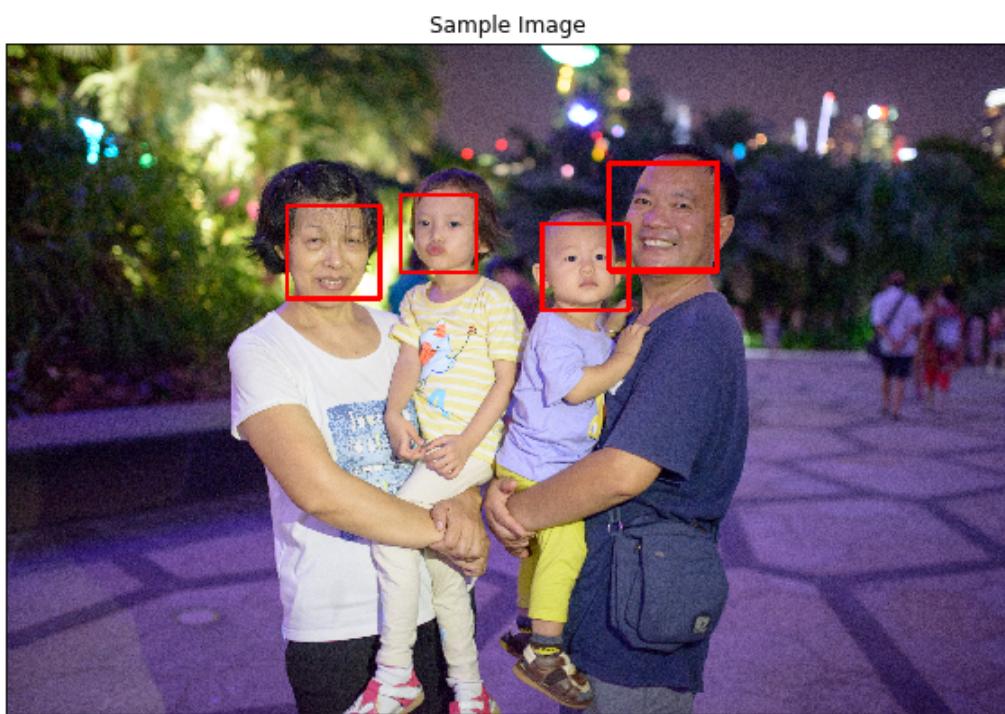
```
Out[271]: <bound method BaseEstimator.get_params of LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
   intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
   penalty='l2', random_state=0, solver='liblinear', tol=0.0001,
   verbose=0, warm_start=False)>
```

VII. Fun Part - in Real World

There are two family photos `test_1.jpg` and `test_2.jpg` in the root of working directory. Both of them were not used for training or testing in the earlier sections. Let's first perform a face detection using the `process_faces()`.

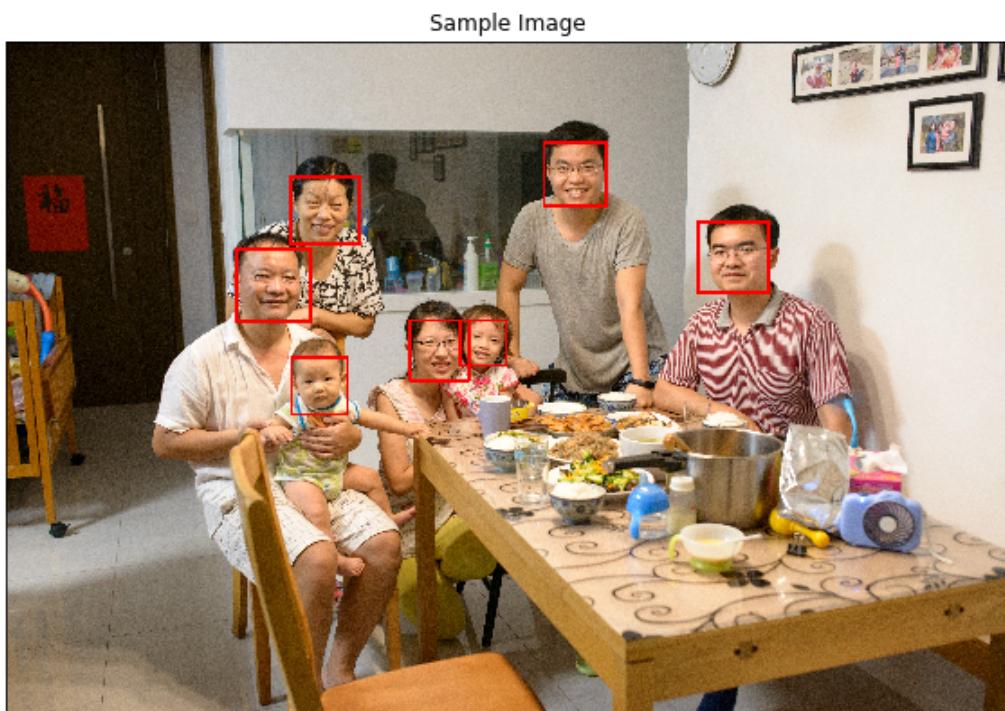
```
In [48]: 1 process_faces('./test_1.jpg', DISPLAY, NOSAVE, 1.35, 5)
```

```
Image path ./test_1.jpg
Image numpy array shape: (4766, 7141, 3) <class 'numpy.ndarray'>
Number of faces detected: 4
```



```
In [49]: 1 process_faces('./test_2.jpg', DISPLAY, NOSAVE, 1.32, 7)
```

```
Image path ./test_2.jpg
Image numpy array shape: (4912, 7360, 3) <class 'numpy.ndarray'>
Number of faces detected: 7
```



Now we can train a logistic regression model by using the `./images2` dataset.

```
In [99]: 1 faces, _ = load_dataset('./images')
2 generate_images(faces)
```

```
Generating images, 0 of 419 faces
Generating images, 100 of 419 faces
Generating images, 200 of 419 faces
Generating images, 300 of 419 faces
Generating images, 400 of 419 faces
```

```
In [60]: 1 faces, targets = load_dataset('./images2')
2 all_train_tensors = paths_to_tensor(faces)
3 all_train_targets = targets
4 all_train_features = get_features(feature_model, all_train_tensors)
5
6 clf_log = LogisticRegression(random_state=0)
7 %time model_log = clf_log.fit(all_train_features, all_train_targets)
```

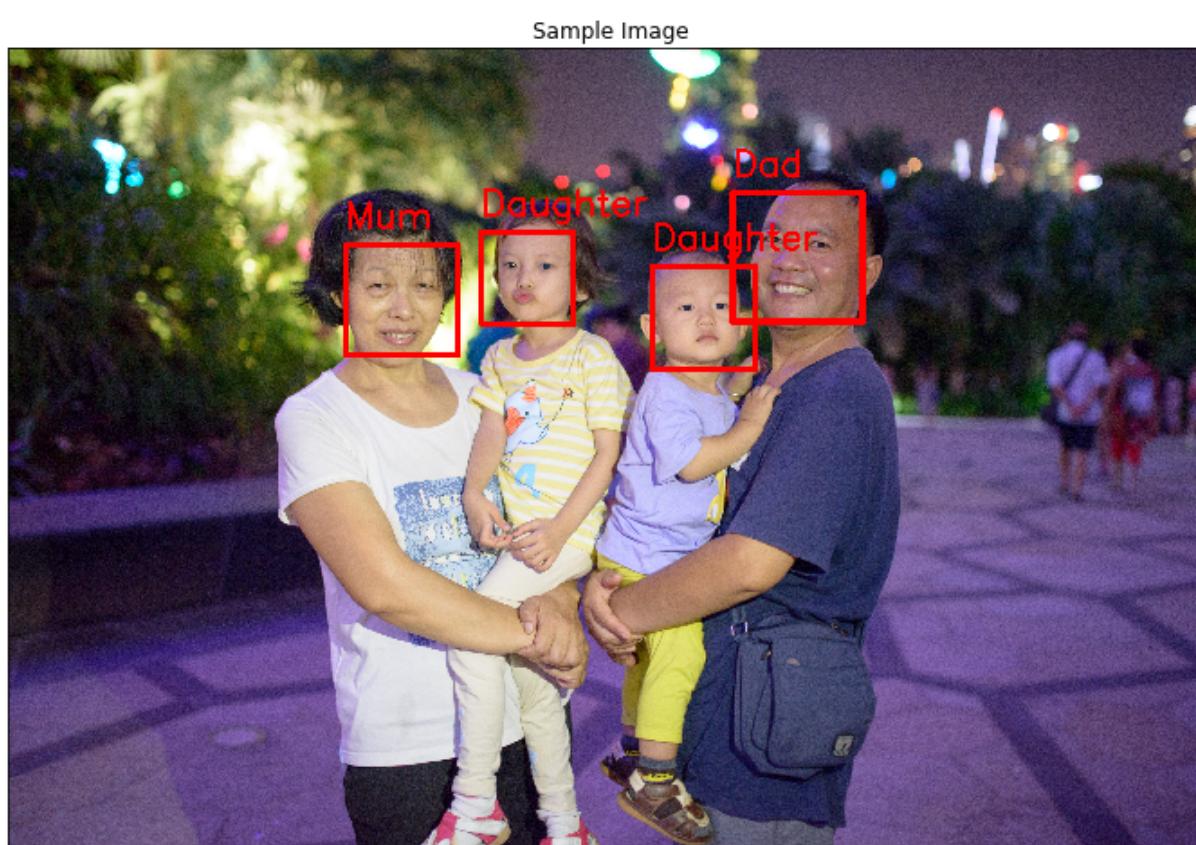
Wall time: 15.8 s

```
In [87]: 1 import pickle
2 pickle.dump(clf_log, open('clf_log', 'wb'))
3 clf_log = pickle.load(open('clf_log', 'rb'))
```

```
In [79]: 1 def get_name(model, image_face):
2     ''' Use the model to make prediction and return the result as the name of the face.
3     Args:
4         model (sklearn.model): A trained model.
5         image_face (numpy.array): A face image in 3D numpy array.
6     Returns:
7         face_names[predict_idx] (str): The name of the face.
8         ...
9     # Convert the 3 channel RGB to 4-d tensor and normalize it
10    image_face_tensors = image_face.reshape(-1, 299, 299, 3)/255
11    image_face_features = get_features(feature_model, image_face_tensors)
12    predict_idx = model.predict(image_face_features)[0]
13    return face_names[predict_idx]
14
15 def process_faces_names(file_path, scaleFactor=1.3, minNeighb=5):
16     """Process the input image file by extracting face(s). Draw bounding box and detected name.
17     Args:
18         file_path (str): The full path of the input image file.
19         scaleFactor (float): The scaling factor used by face detection function.
20         minNeighb (int): The number of minimum neighbors.
21     Returns:
22         None: Display the image.
23         """
24     print('Image path', file_path)
25     image = get_numpy_from_file(file_path)
26     #     image = cv2.fastNLMeansDenoisingColored(image, None, 5, 5, 7, 15)
27     faces = get_faces(image, scaleFactor, minNeighb)
28     print('Number of faces detected:', len(faces))
29
30     image_with_detections, image_faces = draw_bounding_box(image, faces)
31
32     for i, (x,y,w,h) in enumerate(faces):
33         cur_face = cv2.resize(image_faces[i], (299,299))
34         name = get_name(model_log, cur_face)
35
36         # Write the returned name on the image
37         cv2.putText(image_with_detections, name,
38                     (x,y-100),cv2.FONT_HERSHEY_SIMPLEX, 7, (255,0,0),20,cv2.LINE_AA)
39
40     display_from_numpy(image_with_detections, 12, 12)
```

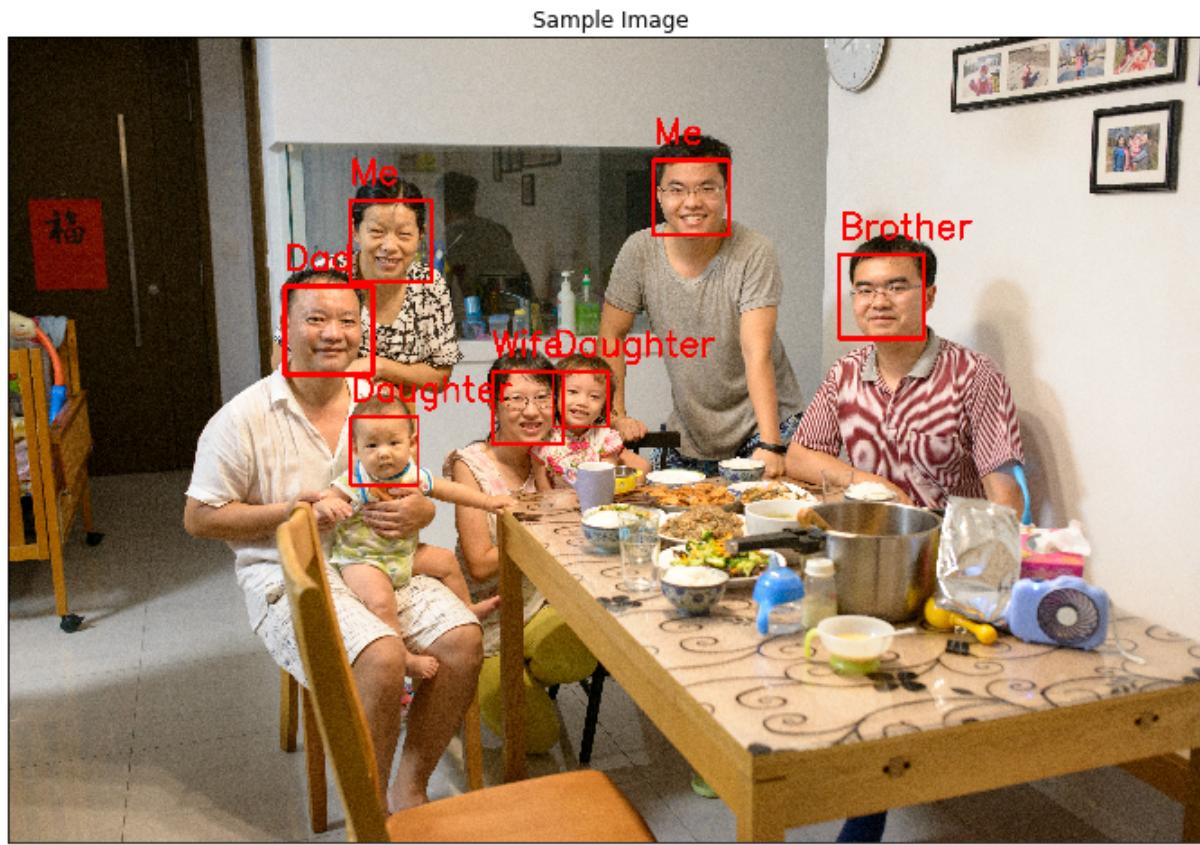
```
In [80]: 1 process_faces_names('./test_1.jpg', 1.35, 5)
```

Image path ./test_1.jpg
 Image numpy array shape: (4766, 7141, 3) <class 'numpy.ndarray'>
 Number of faces detected: 4



```
In [81]: 1 process_faces_names('./test_2.jpg', 1.32, 7)
```

```
Image path ./test_2.jpg
Image numpy array shape: (4912, 7360, 3) <class 'numpy.ndarray'>
Number of faces detected: 7
```



The result is not perfect, both photos have misclassified face(s). The accuracy is around 70%. Not good, but not too bad, because we are family and we are born to look similar. In addition, it's no surprise that my 'son' is misclassified as 'daughter' twice. Based on the earlier confusion matrix plot, the recall rate of my 'son' is 66% and most of the time it's misclassified to 'daughter'.

```
In [ ]: 1
```