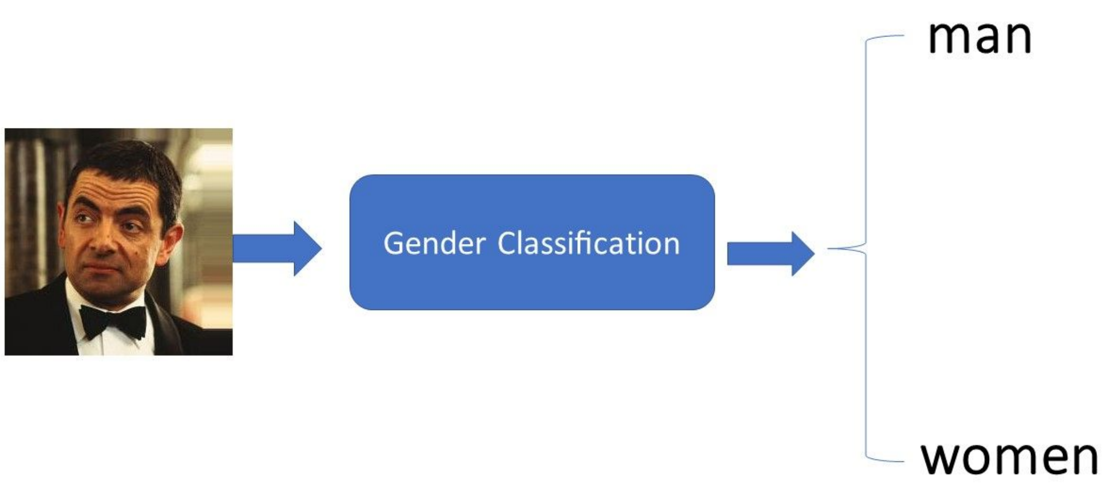
Capstone Project
Machine Learning Engineer Nanodegree

Saif Aljebury
7/8/2018

Definition

Project Overview

The gender recognition is essential and critical for many applicationsin the commercial domains such as applications of human-computer interactionand computer-aided physiological or psychological analysis, since it containsa wide range of information regarding the characteristics difference between male and female. Some have proposed various approaches for automatic genderclassification using the features derived from human bodies and/or behaviors.First, this paper introduces the challenge and application of gender classificationresearch. Then, the development and framework of gender classification are described. We compare these state-of-the-art approaches, including vision-based methods, biological information-based methods, and social network information-based methods, to provide a comprehensive review of gender classificationresearch. Next we highlight the strength and discuss the limitation of each method.Finally, this review also discusses several promising applications for future work.



Problem Statement

When a person tries to identify sex through the face, sometimes it is difficult to recognize it.And also when there is (for example a conference) and the number of attendance is very large, we need a technology that enables us to classify people on the basis of gender , After presenting the details of the training procedure setup we proceed to evaluate on standard benchmark sets. We report accuracies of 96% in the IMDB gender dataset. Along with this we also introduced the very recent real-time enabled guided backpropagation visualization technique.

Metrics

I would use the most common evaluation metric for classification problems, A confusion matrix is a technique for summarizing the performance of a classification algorithm.Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset.Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making. in a problem where there is a large class imbalance, a model can predict the value of the majority class for all predictions and achieve a high classification accuracy, the problem is that this model is not useful in the problem domain :

F1 Score

The F1 Score is the 2*((precision*recall)/(precision+recall)). It is also called the F Score or the F Measure. Put another way, the F1 score conveys the balance between the precision and the recall.

- The F1 for the All No Recurrence model is 2*((0*0)/0+0) or 0.
- The F1 for the All Recurrence model is 2*((0.3*1)/0.3+1) or 0.46.
- The F1 for the CART model is 2*((0.43*0.12)/0.43+0.12) or 0.19.

If we were looking to select a model based on a balance between precision and recall, the F1 measure suggests that All Recurrence model is the one to beat and that CART model is not yet sufficiently competitive.

Analysis

Data Exploration

In this project we have used the publicly available image dataset for gender classification from (IMDB),This file contains 7GB pictures of artists and celebrities[here] Total number of images in the file: (460723)The images in the file contain different sizes we need to format our image for supplying to model,The file contains unbalanced data which means we have a problem,as multi-class classification problems..The file contains only faces with a file metadata , also provide a version with the cropped faces (with 40% margin)

Exploratory Visualization

Algorithms and Techniques

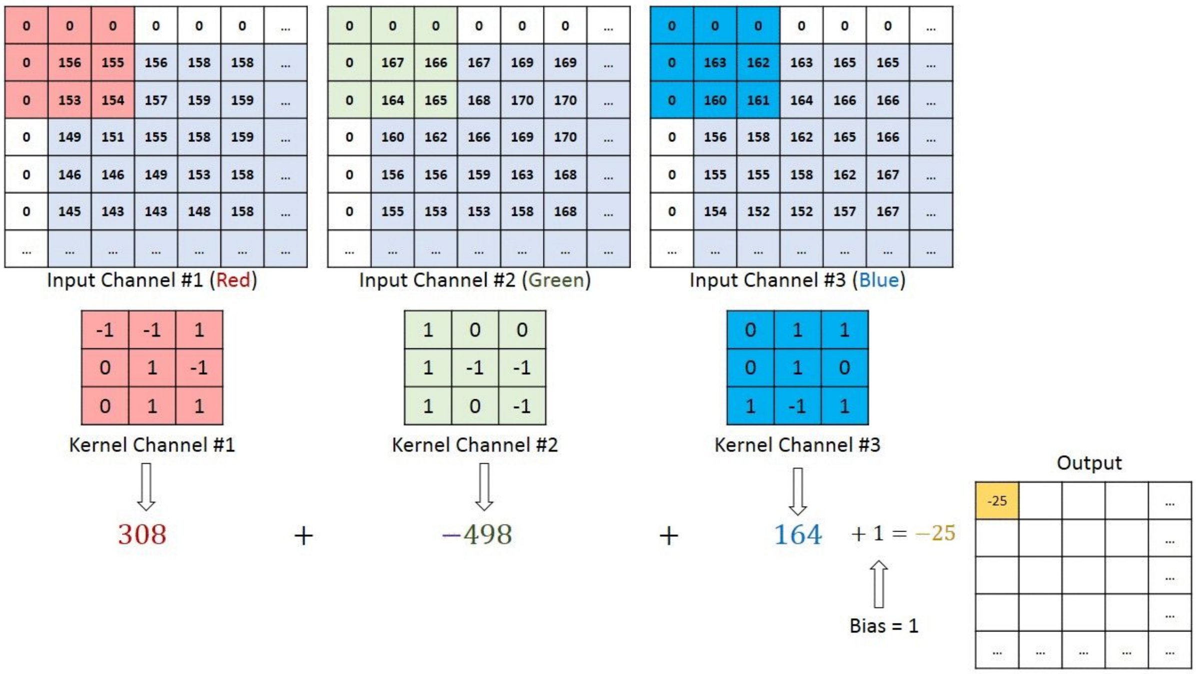
Since we have a problem classification the images the best technique is (CNN)The first thing to know about convolutional networks is that they don’t perceive images like humans do. Therefore, you are going to have to think in a different way about what an image means as it is fed to and processed by a convolutional network.

Convolutional networks perceive images as volumes, i.e. three-dimensional objects, rather than flat canvases to be measured only by width and height. That’s because digital color images have a red-blue-green (RGB) encoding, mixing those three colors to produce the color spectrum humans perceive. A convolutional network ingests such images as three separate strata of color stacked one on top of the others. So a convolutional network receives a normal color image as a rectangular box whose width and height are measured by the number of pixels along those dimensions, and whose depth is three layers deep, one for each letter in RGB. Those depth layers are referred as *channels*.

As images move through a convolutional network, we will describe them in terms of input and output volumes, expressing them mathematically as matrices of multiple dimensions in this form: 30x30x3. From layer to layer, their dimensions change for reasons that will be explained below. You will need to pay close attention to the precise measures of each dimension of the image volume, because they are the foundation of the linear algebra operations used to process images.

Now, for each pixel of an image, the intensity of R, G and B will be expressed by a number, and that number will be an element in one of the three, stacked two-dimensional matrices, which together form the image volume. Those numbers are the initial, raw sensory features being fed into the convolutional network and the ConvNets purpose is to find which of those numbers are significant signals that actually help classify images more accurately (just like other feedforward networks we have discussed).

Rather than focus on one pixel at a time, a convolutional net takes square patches of pixels and passes them through a *filter*. That filter is also a square matrix smaller than the image itself, and equal in size to the patch. It is also called a *kernel*, which will ring a bell for those familiar with support-vector machines, and the job of the filter is to find patterns in the pixels.



Benchmark

What I want to do is compare the accuracy results of (standard convolution neural network) with the results (and depth-wise separable convolutions.), The time it takes to process images in both cases

Methodology

Data Preprocessing

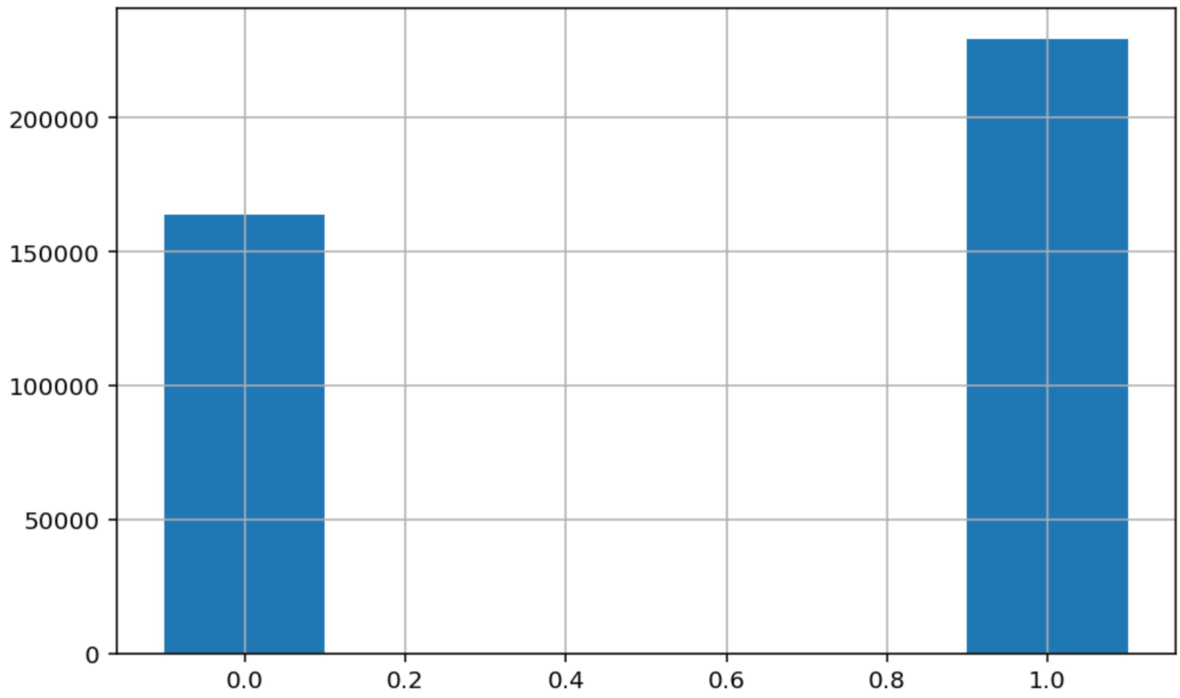
Process data processing through the following methods :

- Crop parts of the image
- The images are split into square shaped segments; random noise is used for padding
- The images are converted to grayscale
- The pixel values get transformed to 16-bit floats
- The pixel values get divided by the standard deviation of the pixel values

Exploratory Visualization :

Fig. 2 A plot showing how the legible. in the matlab file ,Show **gender:** us that images containing values (0) are a woman and values (1) man,*NaN* if unknown

face_location: location of the face. To crop the face in Matlab run



(Fig. 2)

- **face_location:** location of the face. To crop the face in Matlab run

- **face_score:** detector score (the higher the better). *Inf* implies that no face was found in the image and the *face_location* then just returns the entire image
- **second_face_score:** detector score of the face with the second highest score. This is useful to ignore images with more than one face. *second_face_score* is *NaN* if no second face was detected.

Implementation:

Our first model reliesontheideaofeliminatingcompletelythefullyconnected layers[8]. The second step is to train each class with an optimizer [ADMS] and then remove the layers connected with each other we are using. Global Average Pooling was achieved by having in the last convolutional layer the same number of feature maps as number of classes, and applying a softmax activation function to each reduced feature map. Our initial proposed architecture is a standard fully-convolutional neural network composed of 9 convolution layers: ReLUs [6], Batch Normalization [7] and Global Average Pooling. This model contains approximately 600,000 parameters. It was trained on the IMDB gender dataset, which contains 460,723 RGB images where each image belongs to the class “woman” or “man”, and it achieved an accuracy of 96% in this dataset.

Our second model is inspired by the Xception [8] architecture. This architecture combines the use of residual modules [10] and depth-wise separable convolutions [9]. Residual modules modify the desired mapping between two subsequent layers, so that the learned features become the difference of the original feature map and the desired features. Consequently, the desired features H(x) are modified in order to solve an easier learning problem F(X) such that:

$$H(x) = F(x) + x$$

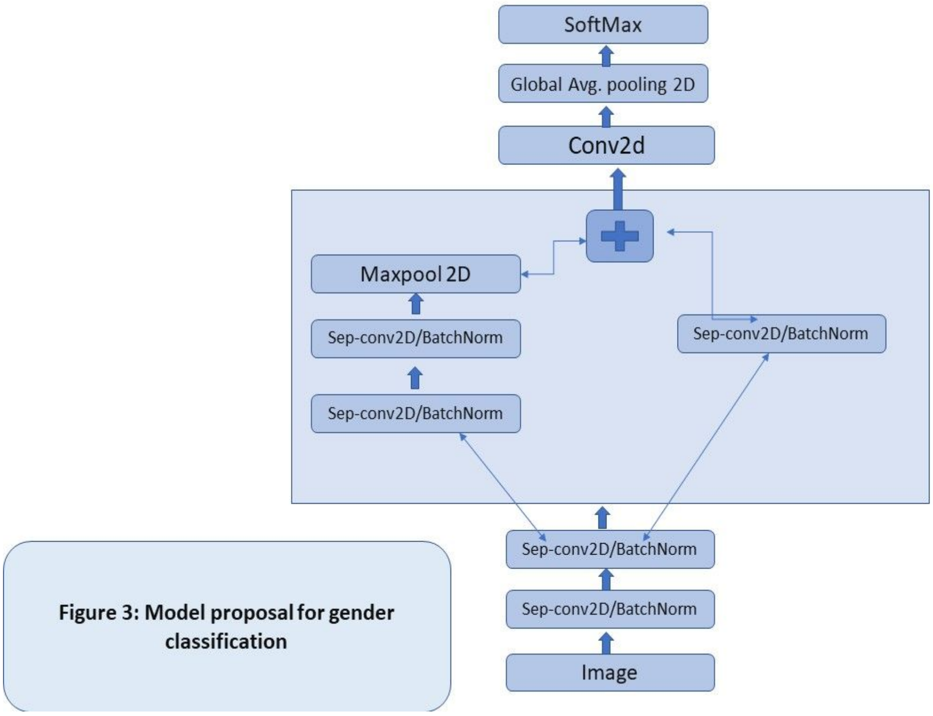


Fig. 3: Our mini-Xception model for gender classification.

Our final architecture is a fully-convolutional neural network that contains 4 residual depth-wise separable convolutions where each convolution is followed by a batch normalization operation and a ReLU activation function. The last layer applies a global average pooling and a soft-max activation function to produce a prediction. This architecture has approximately 60000 parameters, which corresponds to a reduction of 10× when it is compared to our initial naive implementation,and 80× when it is compared to the original CNN

Refinement

Figure 3 displays our complete final architecture which we refer as mini-Xception. This architectures obtains an accuracy of 95% in gender classification task. Which corresponds to a reduction of one percent with respect to our initial implementation.

Results

Model Evaluation and Validation

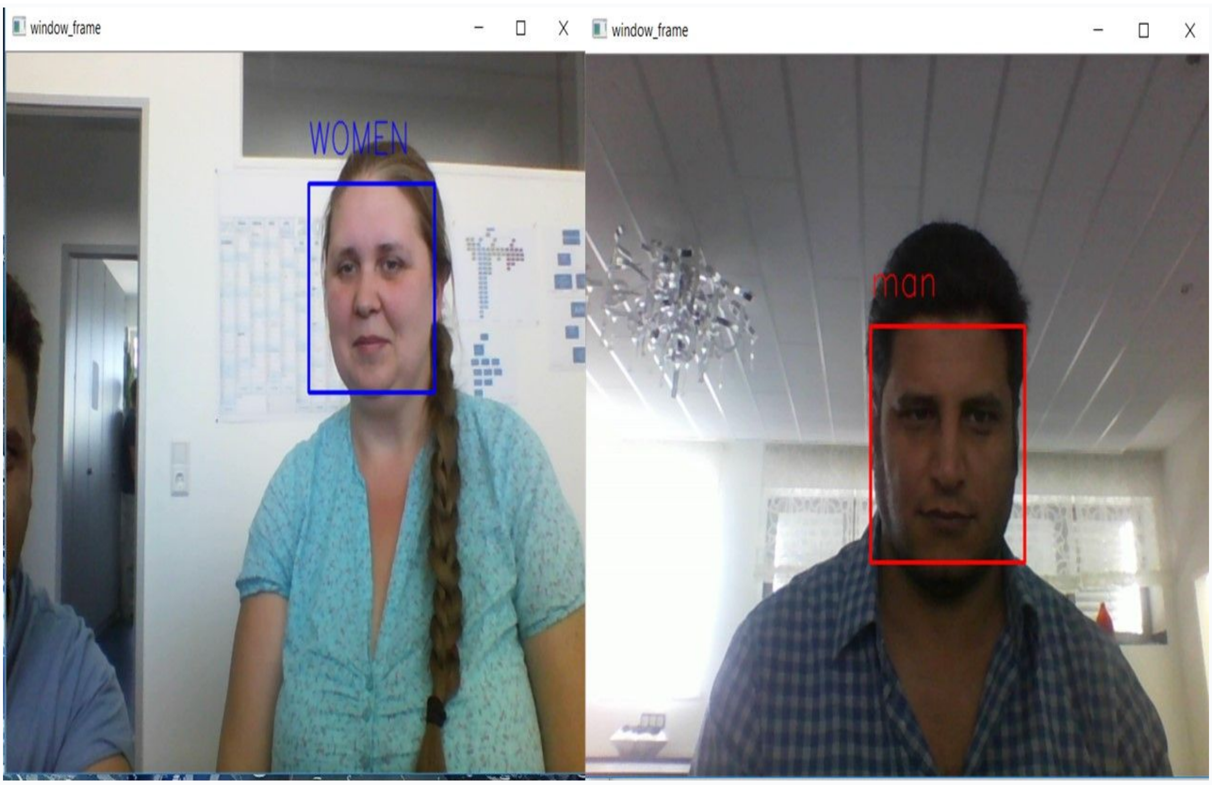
Justification

We have obtained positive results through the computer camera(10) if it is possible to separate the genders (man and woman). It can be observed that Network (CNN) was able to identify certain characteristics in the face of the human through which we have been able to sort out the gender. Taking into account that ordinary man finds it difficult sometimes to differentiate between men and women because of the overlap in the features of the beard, hair, age, etc., the network (CNN) finds it difficult to differentiate between gender.which is illustrated by Figure 4

Conclusion

Free-Form Visualization

below are some of image classified by mini-Xception model.



The image in Fig.12 was an example of correctly classified image. It is correctly classified as(mand and women)

Misclassifications may be due to the following:

• Less clear image

• Too many objects(Clutter) in the image

Reflection

The process used for this project can be summarized using the following steps:

1. An initial problem and relevant, public datasets were found

2. The data was downloaded and preprocessed (segmented)

3. A benchmark was created for the classifier

4. The classifier was trained using the data (multiple times, until a good set of parameters were found)

5. Considered the highest degree of accuracy measure

6. Create a network (CNN)

7. I have set the form on depending on the test frequency and the number of parameters

8. Use of extracted weights from(CNN) For the technology of facial recognition (CV2)

I found steps 7,8,9 to be difficult. This is because of the fact that training and testing took large time despite running on GPU instances. This may be due to model complexity and huge size of data.

Improvement

Machine learning models are biased in accordance to their training data. In our specific application we have empirically found that our trained CNNs for gender classification are biased towards western facial features and facial accessories. We hypothesize that this misclassfications occurs since our training dataset consist of mostly western: actors, writers and cinematographers as observed in Figure 2. Furthermore, as discussed previously, the use of glasses might affect the emotion classification by interfering with the features learned. However, the use of glasses can also interfere with the gender classification. This might be a result from the training data having most of the images of persons wearing glasses assigned with the label “man”. We believe that uncovering such behaviours is of extreme importance when creating robust classifiers, and that the use of the visualization techniques such as guided back-propagation will become invaluable when uncovering model biases.

REFERENCES

[1] Ian Goodfellow et al. Challenges in Representation Learning: A report on three machine learning contests, 2013.[2] Andrew G. Howard et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2818–2826, 2016

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[5] Dario Amodei et al. Deep speech 2: End-to-end speech recognition in english and mandarin. CoRR, abs/1512.02595, 2015.

[6] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 315–323, 2011.

[7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning, pages 448–456, 2015.

[8] Franc¸ois Chollet. Xception: Deep learning with depthwise separable convolutions. CoRR, abs/1610.02357, 2016

[9] Andrew G. Howard et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.