CS574

Computer vision and machine learning

(Assignment #2)

Supervised By - Arijit Sur

Created by:-
(Group 07)
Aditya(160101010)
Avinash(160101018)
Phool Chandra(160101051)
Savinay(160101062)
Kevin(160101063)

Introduction

Age estimation comes under facial image classification. It can be defined as estimation of a person age from facial images. For Research work, it can be defined as age of person based on persons face biometric features, precisely on the basis of 2-D images of face. Morphology by itself is a study of form. Therefore, the craniofacial morphology is a study of shape of the face and skull. Changes in texture of face are defined as changes in face associated with muscle and skin elasticity. The age of person effects structure and its appearance of a person in many different ways. The changes of face are related to face texture and craniofacial morphology. Some facial feature appear only in people of certain age group and change occur in certian age group.Usually some changes occur only during adulthood like changes in skin texture.

Fusion Network for Face-based Age Estimation

Introduction

Convolutional Neural Networks(CNN) have been core framework for age-related research. CNNs do not pay enough attention to facial regions that carry age-specific feature for this particular task and consider face as a typical object. Here a novel CNN architecture (Fusion Network) for face based age estimation is proposed. Specifically, FusionNets take the age-specific facial patches and face as successive n + 1 inputs (n facial patches + 1 face). The aligned face contain major information, which is primary input that is fed to the lowest layer to have the longest learning path. In this work, they tackle the age estimation problem by focusing on the representation learning and they modify the network structure to extract feature which are more representative by paying more focus to information-rich regions.

Fusion Network

The method proposed here consists of three components:-
1 The facial patch selection
2 The convolutional network
3 The age regression
Selected patches are subsequently fed sequentially into the convolutional network, together with the face. Regression method is used for the final prediction calculation.

1. **Facial Patch Selection**
   Age-specific feature from aligned faces is extracted through Bio-inspired Features. Faces are convolved with a bank of Gabor filters. They convolve each face with a total of 8 orientations and 8 bands of Gabor filters which generates a k-dimensional feature vector. In the experiments, k is greater than 10,000 with each element encoding one potential input for the subsequent CNN. As k is very large to use this high-dimensional feature vector in the feeding sequence so we need k’ features from the Bio-inspired Features feature vector to form a subset where k’ << k. The multi-class AdaBoost is used for selecting the subset k’ from

the high-dimensional feature vector. For a dataset with m samples, they pick the k' most informative features from a k-dimensional vector by using the weak classifier h,

$$F_j = argmin_k\left(\sum_{i=1}^{m} w_i^{k'} e\left(h_k(x_i), y_i\right)\right)$$ where, $$e\left(h_k(x_i), y_i\right) = \begin{cases} 0 & h_k(x_i) = y_i \\ 1 & otherwise \end{cases}$$

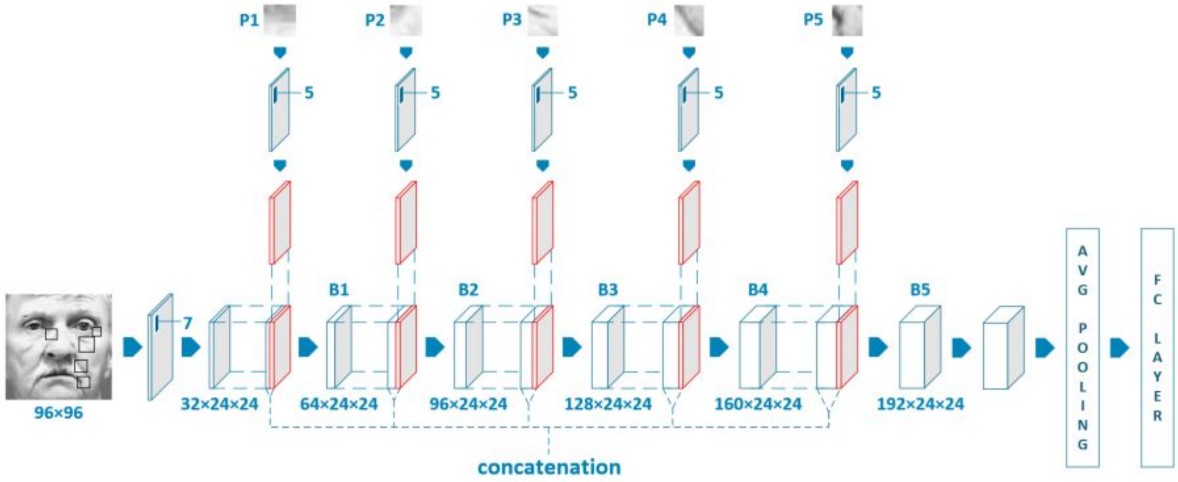
where F_j is the j-th selected feature and j ∈ [1, k']. x_i is the high dimensional feature vector after the i-th sample is filtered by Gabor filters and y_i is the associated age label. In addition, $w_i^{k'}$ is the weight in AdaBoost, which is updated and normalized after each F_j is found.

2. Network Architecture

All of the blocks shown in the figure below are residual blocks, and each block after concatenation (B1 to B5) contains bottleneck layers. As it has been found that have found that lowering the number of feature maps right before the global pooling largely reduces the performance so they do not apply feature reduction to B5. Before each convolutional layer they applied a batch normalization layer to improve the training speed and overall accuracy. After the convolutional stage, a global average pooling layer and a fully-connected layer are attached to generate the final output of the network. In the Fusion Network compared to the multipath CNN, all the features from different inputs have a longer and more efficient learning path and the common age-specific features among the inputs are being extracted and emphasized. The DenseNet was an inspiration for use of concatenation. In the FusionNet the formulation is based on blocks and after concatenations the output of each residual block can be represented as:

$$x_i = B_i(\ [x_{i-1}, s_i]\ )$$

Where i ∈ [1,5] since they decide to use 5 input patches in our network and B_i [·] denotes the synthesized learning function of the i-th block. s_i is the feature map learned from the i-th input patch and x_{i-1} is the output from the previous residual block.



3. **Age Regression**

   When the number of classes becomes larger the discretization error becomes smaller for the regressed signal so regression approach is used by them to calculate the final prediction. After processing features by the fully-connected layer, all the negative values from the output vector is firstly eliminated and then it is feed to a Softmax function to form a probability distribution. Then, the distribution is normalized to sum up to 1. The final prediction is the summation of the products of the probabilities by the corresponding age labels.

$$E(O) = \sum_{i=1}^{j} p_i y_i$$

   Where j is the number of classes and p_i denotes the normalized probability for the i-th class, y_i is the associated age label.

Conclusion

They compare their approach with other recent state-of-the-art CNN-based models: DEX, OR-CNN, and Ranking-CNN. They evaluated data with same data partition ratio to have a fair competition. The FusionNet significantly outperforms other state-of-the-art models by achieving the lowest MAE of 2.82 which help us to conclude that this network has a much more efficient feature extraction architecture.

In this paper, they presented the Fusion Network to tackle the face-based age estimation problem. This model takes other age-specific facial patches as inputs with the face. This input facial patches act as shortcut connections in the network, which amplify the learning efficiency for age-specific features. Experiments show that this network significantly outperforms other CNN-based state-of-the-art methods on the MORPH II benchmark.

Deeply Learned Feature for Age Estimation

Introduction

In the paper Convolutional Neural Network based Deep Learning investigated. Deep learning model is used for age feature extraction. In difference with the other models feature maps obtained in different layers is used instead of top layer. Addition to that manifold learning algorithm is combined in proposed scheme to improve performance.
Evaluation of different classification and regression schemes is done in estimating age using the deep learned aging pattern (DLA).

Deep Learned Aging Pattern (DLA)

Training is done on deep model of CNN for multiclass age estimation. The leaned structure is used to predict the age given a testing facial image.

The work have 6 layers, For input level they have used 2D grayscale face image of size 60 by 60 pixels as input to first convolutional layer where kernel size is 5*5. Once filter is applied of size a x b to feature map of size h x w, the output size will be (h - a + 1) x (h - b +1). So the size of feature maps in layer L2 after convolution is 56 x 56. A feature map is obtained by applying an activation function(Logistic Function) at the end of L2 layer. Within a feature map, all the units share the same set of weights for the filter. The obtained feature maps in the convolution layer go through a pooling layer.

After pooling to the feature map obtained in L2, L3 layer is obtained. To reduce spatial resolution 2 x 2 subsampling was applied on each of feature maps in L2. L4 is obtained by using kernel size of 7 x 7 on feature maps from L3. L5 is obtained by pooling operation of each feature maps from L4. L6 consist of 80 feature maps of size 1 x1.

Each unit is connected to all feature maps in layer L5. Output layer is fully connected to L6. Feature maps obtained in these layers are constructed to extract low-level features, including edges and texture. Face images with labels were used for Supervised learning These filtres used to extract the feature which might be useful for age estimation. A small neighbourhood in one layer is connected to the units in the successive layer. This is for extracting features from local receptive fields. Another advantage of applying CNN in this work is filter weights are shared by the units in the same feature map. This can greatly reduce the number of parameters for training.

The difference here is other works usually use output of top layer in deep learning model as feature representation in their problem here in this paper where representations in different layers respond to particular activations, we investigate the use of these extracted features in the problem of age estimation and whether combining these features is an effective way for predicting age.

After training the CNN model, features can be extracted from different layers. Principal component analysis is applied to reduce feature dimension. Extracted features are concatenated to get aging pattern.

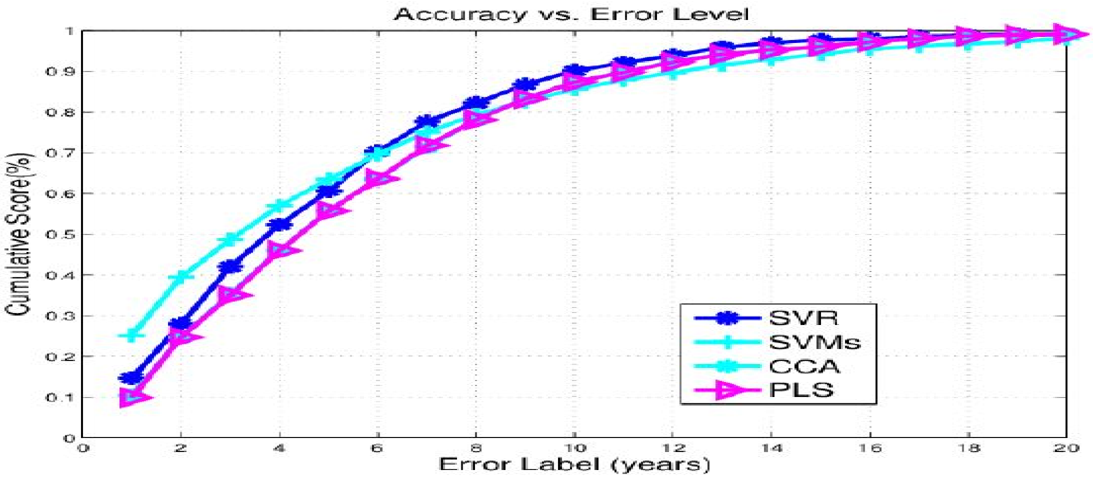
Manifold learning has been applied in this work for its good performance in learning aging patterns and capturing the underlying face aging structure.

In this work, both SVMs for age classification and SVR for age regression are used and evaluated.

Conclusion

The given approach works well than the state of the art algorithm experiments for comparing both algorithms under same condition. The Mean Absolute Error (MAE) for proposed approach on MORPH data is 4.77 compared to the current best result 5.69. MAE for FG-NET dataset is 4.26 is less than 4.67 of current best.

Based on the applied manifold learning algorithms, we can categorize the proposed schemes into three different methods, which are DLA+MFA, DLA+OLPP, DLA+LSDA. Efficiency of theses methods can be see in below graph



Cumulative scores of the algorithms with different settings on MORPH dataset.

Deep Regression Forests for Age Estimation

Introduction

Age estimation from facial images is a nonlinear regression problem. The main challenge of this problem is the facial feature space w.r.t. ages are heterogeneous, due to the large variation in facial appearance across different persons of the same age and the non-stationary property of ageing patterns.

To solve this problem, we need a nonlinear mapping function between facial image features and the real chronological age. However, to learn such a mapping is a challenging task.

Some previous model whose difficulties overcome this paper:-

1. Divide-and-conquer is a good strategy to learn the non-stationary age changes in human faces, but the existing methods make hard partitions according to ages. Consequently, they may not find homogeneous subsets for learning local regressors.
2. To model such heterogeneous data, Kernel-based global non-linear mapping can be used but Learning non-stationary kernel is inevitably biased by the heterogeneous data distribution and thus easily causes overfitting.
3. Traditional regression forests make hard data partitions, based on heuristics such as using a greedy algorithm where locally-optimal hard decisions are made at each split node.
4. Several researchers formulated age estimation as an ordinal regression problem because the relative order among the age labels is also important information. They trained a series of binary classifiers to partition the samples according to ages, and estimated ages by summing over the classifier outputs. Thus, ordinal regression is limited by its lack of scalability.

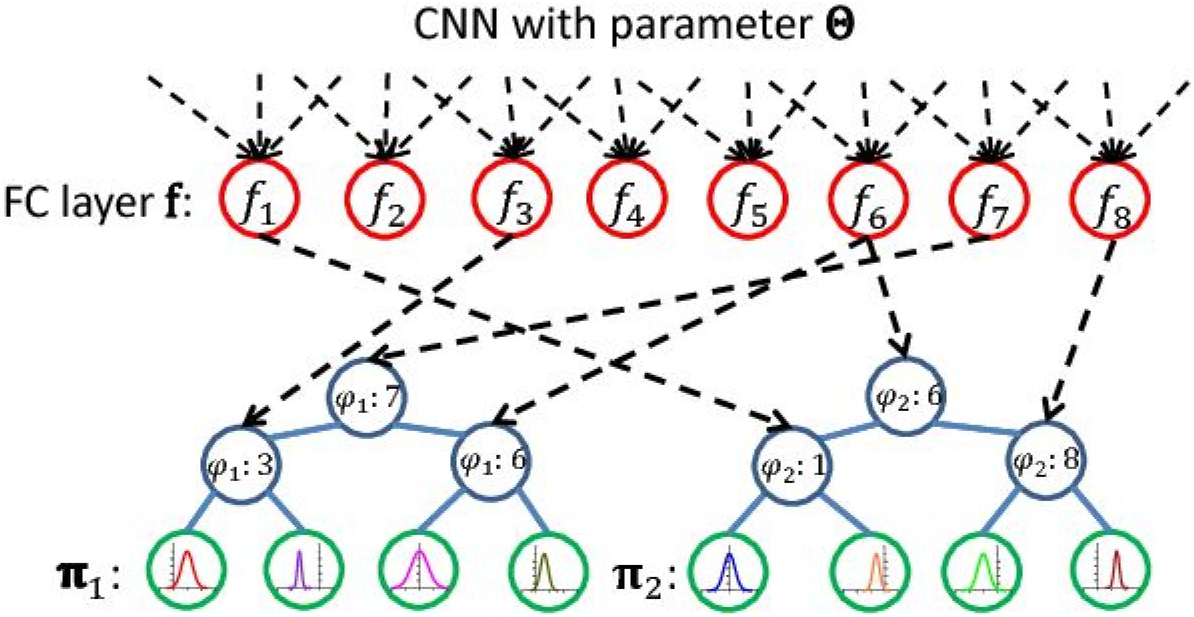
Deep Regression Forests (DRFs), an end-to-end model, for age estimation. DRFs connect the split nodes to a fully connected layer of a convolutional neural network (CNN) and deal with heterogeneous data by jointly learning input-dependant data partitions at the split nodes and data abstractions at the leaf nodes. This joint learning follows an alternating strategy: First, by fixing the leaf nodes, the split nodes, as well as the CNN parameters, are optimized by Back-propagation; Then, by fixing the split nodes, the leaf nodes are optimized by iterating a step-size free and fast-converging update rule derived from Variational Bounding. We verify the proposed DRFs on three standard age estimation benchmarks and achieve state-of-the-art results on all of them.

The contribution of this paper is three folds:
1) It proposes Deep Regression Forests, an end-to-end model, to deal with inhomogeneous data by jointly learning input-dependant data partition at split nodes and data abstraction at leaf nodes.
2) Based on Variational Bounding, the convergence of our update rule for leaf nodes in DRFs is mathematically guaranteed.
3) We apply DRFs on three standard age estimation benchmarks and achieve state-of-the-art results.

Deep Regression forests formulation:-

Let X = R^{dx} and Y = R^{d_y} denote the input and output spaces, respectively. We consider a regression problem, where for each input sample x ∈ X, there is an output target y ∈ Y.

A decision regression tree consists of a set of split nodes N and a set of leaf nodes L. Each split node $\boldsymbol{n}$ ∈ N defines a split function $S_n(\cdot; \Theta)$: X → [0, 1] parameterized by **Θ** to determine whether a sample is sent to the left or right subtree. Each leaf node *l*∈ L contains a probability density distribution $\pi_l(y)$ over Y.



Above FC = Fully Connected
The red colour nodes are the nodes of a fully-connected layer of CNN. They denote the output units of the function f which is parameterized by Θ.

Two index functions φ_1 and φ_2 are assigned to these two trees respectively. The blue and green circles are split nodes and leaf nodes, respectively. The black dash arrows indicate the correspondence between the split nodes of these two trees and the output units of the FC layer. Note that, one output unit may correspond to the split nodes belonging to different trees. Each tree has independent leaf node distribution π (denoted by distribution curves in leaf nodes). The output of the forest is a mixture of the tree predictions. f (·; Θ) and π are learned jointly in an end-to-end manner.

As each tree in the forest F has its own leaf node distribution π, we update them independently. In our implementation, we do not conduct this update scheme on the whole dataset S but on a set of mini-batches B.

Implementation Details

Parameters Setting The model-related hyperparameters (and the default values we used) are: number of trees (5), tree depth (6), number of output units produced by the feature learning function (128), iterations to update leaf-node predictions (20), number of mini-batches used to update leaf node predictions (50). We decrease the learning rate (×0.5) every 10k iterations. The network training based hyper-parameters (and the values we used) are: initial learning rate (0.05), mini-batch size (16), maximal iterations (30k).

Performance Comparison

We run DFSs for different well known data sets like MORPH, CACD and output statics are:

Dataset	MAE	CS
MORPH	2.17%	91.3%
GF-NET	3.85	80.6%

DRF learn nonlinear regression between heterogeneous facial feature space and ages. In DRFs, by performing soft data partition at split nodes, the forests can be connected to a deep network and learned in an end-to-end manner. The data partition at split nodes is learnedby Back-propagation and data abstraction at leaf nodes is optimized by iterating a step-size free and fast-converged update rule derived from Variational Bounding. The end-to-end learning of split and leaf nodes ensures that partition function at each split node is input-dependent and the local input-output correlation at each leaf node is homogeneous. Experimental results showed that DRFs achieved state-of-the-art results on three age estimation benchmarks.

Age Estimation Using Ordinal regression with CNN

Introduction:

Before this paper, age estimation is done by the multiclass classification method and metric regression method.

In multiclass classification, all class labels are independent of one another but in real age labels have a strong relationship and those labels form a well-ordered set. Which is not in the multiclass method.

In metric regression takes age label as a numerical value but due to non-stationary aging pattern, it is not good method because learning non-stationary kernels for a regression problem usually difficult because it will easily cause over-fitting in training process.

The aging pattern is non-stationary because facial aging effects appear as changes in the shape of the face during childhood and changes in skin texture during adulthood. This property makes it a non-stationary pattern.

To address the non-stationary property of aging patterns, age estimation can be cast as an ordinal regression problem.

In this paper, they transformed ordinal regression into a series of simpler binary classification subproblems. the benefit of this kind of transformation is that new generalization bounds for ordinal regression can be easily derived from known bounds for binary classification.

In this paper Convolutional neural network(CNN) is used for solving those binary classification subproblems.
Our CNN has multiple output layers where each output layer corresponds to a binary classification subproblem, called Multiple Output CNN.

But in other ordinal regression approaches, the processes of extracting features and learning a regression model are separated and optimized independently.
The unclear mechanism of how human perceives the different aging pattern and make it difficult to design good features for age estimation

According to their, implement End-to-End learning with CNN for age estimation and simultaneously, it optimizes regression modeling and feature learning.

.
In this paper, a new age data set is published called Asian Face Age Dataset (AFAD), it includes more than 160K age labels and Asian facial images. Because other age datasets have problems like fewer numbers of Asian facial images of the same person or dataset is small.

Their Approach:

In this paper, ordinal regression is transformed into a series of sub-problems of binary classification. In particular, $K - 1$ simpler binary classification sub-problems is obtained after transformation from an ordinal regression problem with K ranks. A binary classifier for reach rank $r_k \in \{r_1, r_2, \cdots, r_{K-1}\}$ is constructed to judge the value the rank of a sample y_i is greater than r_k. After that, the rank of an unseen sample is predicted based on the classification results of the $K - 1$ binary classifiers on this sample.

This approach contains 3 steps:

a) Give original training data D = $\{x_i, y_i\}$ and converts it to binary classification subproblem and k'th binary classification has to input data $D^k = \{ x_i, y^k_i, w_i^k \}$

y_i^k = 1, if($y_i > r_k$)
y_i^k = 0 otherwise

And $w_i^k = | C_{y,k} - C_{y,k+1} |$

b) After training data transformation all k-1 binary classification subproblems trained by this training data and they adopt one CNN to collectively implement these binary classifiers.
In this approach, CNN has multiple output structures where each output corresponds to a binary classifier.

c) The rank for an unseen sample x' predicted as follows

$$h(x') = r_q$$
$$q = 1 + \sum_{k=1}^{k-1} f_k(x') \qquad \text{where } f_k(x') \in \{0, 1\}$$

The benefit of Approach:

1) All K − 1 classification subproblems are simultaneously solved with our multiple output CNN. Due to that, all the classification subproblem share the same mid-level representations in such a CNN, the correlation of distinct classification subproblem could be explored, which is beneficial to improve the final performance.

2) we can automatically learn better features from facial images that are better than handcraft features if we use ordinal regression with solved by using an End-to-End deep learning method.

The architecture of the Multiple Output CNN :

In this approach, the network has 3 convolutional, 3 local response normalization and 2 max-pooling layers followed by a layer with 80 neurons. For input face image of size

60 x 60 x 3 is given as input. On input image 20 kernels of size 5 x 5 x 3 with stride of 1 pixel is applied. The feature map of size 28 x 28 x 20 is obtained after local response normalization and max pooling operations. At 2nd and 3rd layer similar operations like mentioned above were conducted with kernel size. To generate a mid level representation a fully connected layer with neurons 80 is used. The network branches out K−1 output layers, The k-th task is to predict whether the age of the i-th facial image is larger than the rank r_k. For each task, the softmax normalized cross entropy loss is employed as loss function.

Conclusion

For both MORPH II and AFAD datasets, the whole dataset into two parts: 80% of the whole data is used for training, and the remaining 20% of the data is used for testing. There is no overlap between the training and testing data. The MAE obtained for MORPH II Dataset is 3.27 and AFAD dataset is 3.34 which is much better than other age estimation methods. This MAE value help us conclude ordinal regression based methods outperform the metric regression based methods in general. The integration of ordinal regression and deep learning methods could boost the performance significantly.

This Ordinal Regression approach presents an End-to-End CNN learning method, which transforms ordinal regression into a series of binary classification sub-problems, which are collectively solved with the proposed multiple output CNN learning algorithm. The performance shown by this approach demonstrates the potential of ordinal regression in age estimation.