Deep Learned Cumulative Attribute Regression

Shashank Jaiswal[1], Joy Egede[2] and Michel F. Valstar[1]
[1]School of Computer Science, The University of Nottingham, UK
[2]School of Computer Science, The University of Nottingham, Ningbo, China

***Abstract*— Learning regression-based machine learning models for computer vision problems is a challenging task due to noisy features, variation in pose and illumination, occlusion, etc. Typically the problem is compounded by the non-uniform distribution of labels in the training data, resulting in parts of the label space that suffer from data sparsity and a problem of label imbalance in general. Deep Convolutional Neural Networks (CNN) have shown remarkable success on a number of computer vision tasks such as object classification and face recognition. However, they too suffer from sparse and imbalanced training datasets for regression problems, even when those datasets are very large. Cumulative Attributes have previously been proposed to address the issue of label imbalance, but to date this concept has not been integrated with Deep Learning. In this work, we propose a CNN-based framework for learning regression models by using Cumulative Attributes as intermediate features. We evaluate our method on a number of tasks which includes pain intensity estimation, Facial Action Unit intensity estimation and age estimation. Our results show that the proposed method is robust to imbalance and sparsity present in the training datasets, and performs significantly better than the current methods where CNNs are learnt directly for regression.**

I. INTRODUCTION

Automatic face analysis is an increasingly popular area of research due to its potential applications in a wide range of fields. Recently, there has been a surge of research activity focussing on its application in healthcare. Pain estimation ([46], [11]), age estimation ([14], [18]) and facial expression recognition ([21], [51]) are some examples where it has been applied for healthcare purposes. Regression models are often used for learning these tasks. It involves learning a mapping from an input set of features to a real-valued output. This scalar output is usually modelled as a continuous variable using regression techniques ([3], [39], [30]). However, a number of existing approaches also treat it as a discreet variable and pose it as a multi-class classification problem ([34], [37], [18]). Training regression models for visual tasks is a difficult problem for a number of reasons. One of the major challenges is finding a set of optimal features which is descriptive enough to capture all the relevant information but which is robust to noise, occlusion variation in lighting condition and viewing angles at the same time. Recent advances in deep learning approaches has made this problem of crafting the right features much easier to solve, by making it possible to actually learn task-specific features in a data-driven manner. The rich hierarchical features learnt using deep Convolutional Neural Networks (CNNs) have shown remarkable success in a number of tasks such as object classification ([25], [49]), face recognition [42], facial expression recognition [20], etc.

Another major challenge which has received considerably less attention is the problem of learning from imbalanced and sparse datasets. Most databases currently available for facial Action Units (AU) intensity estimation (e.g. DISFA[37] and McMaster[31]), pain intensity estimation (e.g. McMaster[31]) and age estimation (e.g. FGNet[41]) have a sparse and highly imbalanced distribution of intensity labels. For example, in the McMaster database the majority (82%) of the video frames have neutral expression with pain intensity level zero. Even when considering only non-zero intensity levels, the distribution is highly imbalanced with low intensity pain occurring in many more frames compared to high intensity pain levels, whose occurrence is relatively rare. Learning regression models on such imbalanced datasets can cause training difficulties resulting in models which are biased towards intensity levels which occur more frequently. Balancing the dataset for each intensity level by sub-sampling is not only difficult but it also results in a waste of precious training data.

One solution to this problem was proposed by Chen et al. [8], who presented the Cumulative Attribute (CA) approach. CA does not need any additional labels, or change the distribution of the data, but applies a transformation of the original labels into a binary-valued attribute vector which are then used as the training data of a regressor instead of the original features. These attributes not only provide good discrimination between different values of labels but due to their cumulative nature, labels that are close to each other are also close together in the attribute space. The CA features are also able to utilize all data samples for discriminating each attribute even if there is no sample available for that attribute alone. This is due to the fact that the attribute can assigned positively/negatively by the neighbouring samples. This makes Cumulative Attributes suitable for learning regression models even in case of sparse and imbalanced datasets. CA does not need the label space to be a metric space, it also works on ordinal data, such as Facial Action Unit intensities. The original approach was evaluated for crowd density and age estimation using hand-crafted features [8].

Deep learning currently gives state-of-the-art performance for classification tasks. However, for regression tasks its performance is negatively affected by sparse and imbalanced datasets. It would of course be possible to learn features using CNNs, then apply the Cumulative Attribute technique

on this to create a distribution-insensitive representation, and then follow this by a regressor. However, this would involve three stages of learning, rather than a single end-to-end learning approach. In this work, we present a deep learning based framework for learning regression models from sparse and imbalanced data by integrating the concept of Cumulative Attributes in the Deep Learning formulation. The proposed method builds upon the the work in [8], extending it to CNN framework where the features and inference engine are learnt in a single end-to-end Deep Learned network. This method uses deep CNNs to learn Cumulative Attributes as an intermediate representation for learning regression models. This representation is robust to sparsity and data imbalance, and results in significantly improved performance compared to conventional approaches where a CNN model is trained to directly predict the output. In contrast to [8], we also experiment with different loss functions for learning Cumulative Attributes and show that Logarithmic loss performs significantly better than Euclidean loss (originally used in [8]).

The main contributions of this paper are threefold: firstly we present a novel CNN architecture which learns cumulative attributes in its intermediate layer resulting in robust models for a number of regression tasks. Secondly we show that the Logarithmic loss function outperforms the Euclidean loss function in three problems, including one that the technique was originally designed for. Thirdly, we show that the proposed method attains superior performance across a range of tasks, to wit facial Action Unit intensity estimation, pain intensity estimation and age estimation.

II. RELATED WORK

A. Attribute Learning

One of the first approaches for attribute learning was presented by Ferrari et al. [12] who learnt classifiers for manually defined visual attributes. Since then, attribute learning has received considerable attention and a number of approaches have been proposed in the past few years (e.g. [19], [57], [44], [59]). Many of the existing approaches on attribute learning target classification tasks either using manually defined attributes (e.g. [27], [12]) or using data-driven learning methods (e.g. [57], [44]). Relatively few approaches have been proposed for attribute learning in visual regression tasks. Chen et al. [8] proposed Cumulative Attributes which is defined as a transformation of the original ground truth labels into a binary valued vector. In [9], spectral clustering was used to automatically discover attribute space for visual regression tasks. Zhang et al. proposed a CNN architecture to jointly learn facial landmark locations with facial attributes related to expression, gender and appearance [59]. However, the attributes required additional manual annotations at the training stage. In contrast to most of the existing approaches for attribute learning, our work presents learning of regression tasks in a CNN based framework without the requirement of any additional annotations.

B. Pain estimation

Automatic pain recognition has attracted considerable research attention mostly due to its potential application to health care. Pain recognition is based on the analysis of pain indicators such as facial expression changes [22], [11], [38], [45], body movements [2], cry characteristics [6] and changes in biomedical signals [55], [26]. Following the introduction of the Prkachin and Solomon Pain Intensity (PSPI) scale [43], a number of studies have focused on classifying facial expressions of pain. Pioneering work [1], [32] on automatic pain recognition adopted a binary classification approach i.e. pain versus no pain. In [4] and [29], this was extended to recognizing posed pain from real pain.

Modelling pain recognition as a classification problem assumes that pain levels are unrelated discrete classes whereas neighbouring pain levels are related and do in fact share characteristics. Consequently, a number of studies [33], [22], [58], [45], [38], [23], [11] have proposed regression frameworks for continuous pain estimation. Most of these [33], [22], [58], [45], [38] use only static facial features while others [23], [13], [60], [11] have experimented with a combination of static and temporal features. Results from the latter show that temporal features improve the performance of pain recognition models. Following the huge success that deep learning frame works have recorded in computer vision applications, a few studies have considered its application to pain estimation. Egede et al. [11] proposed a 2-level RVM framework that combines hand-crafted features and deep learned features for continuous pain estimation. Zhou et al. [61] proposed recurrent convolutional neural network which takes a time-windowed frame sequence as input. Martinez et al. [36] use recurrent neural networks in combination with Hidden Conditional Random Fields (HCRF) to predict sequence-level pain intensity.

Compared to many machine learning models, the performance of automatic pain recognition systems is still limited due to the high data imbalance in available pain datasets. A number of methods have been proposed to deal with this problem. Both [46] and [60] reduce the original 16-point PSPI scale to a 5-point scale by categorizing the original pain levels in a data balancing manner. Majumda et al. [35] used the Synthetic Minority Oversampling Technique (SMOTE) approach for data balancing. Egede et al. [10] proposed cumulative attributes for pain estimation, learning the attributes using a multi-output ridge regression model and doing inference of the pain levels with a subsequent Relevance Vector Machine. In contrast to using regression algorithms based on hand-crafted features, we automatically learn these attributes using CNNs and use the learned attributes for continuous pain estimation.

C. AU intensity estimation

Various approaches have been proposed for automatic prediction facial Action Unit (AU) intensities. Some pose it as a regression problem and approach it using SVR[3] , ordinal regression [47], kernel Regression [39]. In [24], a latent tree based probabilistic graphical model was proposed, aimed

at learning higher order relationship between input features and multiple AU intensities. In many other approaches, AU intensity estimation is treated as a classification problem where each intensity level is treated as a separate class. Such kind of approaches use classifiers such as Support Vector Machines (SVM) (e.g. [34], [37]), Markov Random Fields (MRF) (e.g. [15]), Dynamic Bayesian Networks [28]. Most of these approaches use hand-crafted appearance features such as HoG[39]), LBP[24], LGBP[52]. Many approaches also use shape features computed from locations of facial landmarks (e.g. [47], [39], [52]).

Recently, a number CNN based approaches have also been proposed for facial AU intensity estimation. In [16], a CNN is learnt to jointly predict the occurrence and intensities of multiple AUs. Another multi-task CNN based architecture was proposed in [62] which was trained to predict pose and pose-dependent AU intensity simultaneously. In [53] a joint CNN-CRF model was proposed to learn multiple AU intensities and their correlations. Most of these approaches have focused on developing new architectures which are specialized for estimating AU intensities and learning the correlations between the multiple AUs. However, none of these methods have focused on the problem of learning regression models using imbalanced and sparse dataset. Highly imbalanced and sparse distribution of labels are quite common in databases available for AU intensity estimation (e.g. DISFA[37] and McMaster[31]).

D. Age estimation

Automatic age estimation from face images has been an active area of research in the past few years due to its potential applications in the analysis of demographics, video surveillance, etc. Age estimation methods usually consists of a feature extraction step and then a classification or regression step. Majority of the existing approaches use handcrafted features such as Gabor [18], HoG[48] or LBP[48]. A number of approaches pose age estimation as a classification problem where age is considered as set of discreet classes which are independent of each other. For example [18] used Support Vector Machines as classifier for age estimation. Many other approaches use regression techniques such as SVR [30] and Partial Least Squares [17]. In order to take advantage of the relative ordering of the age labels apart from learning their exact values, ordinal regression approaches have also been proposed. For e.g. in [5] a Rank-SVM algorithm was proposed for age estimation. Similarly, in [7], an ordinal hyperplane ranking algorithm was proposed to learn relative ordering information in the age labels. Apart from the traditional handcrafted feature based approaches, a few CNN based approaches have also been proposed in recent years (e.g. [54], [56]). In [40] an ordinal regression approach was used to train a CNN for age estimation.

None of the existing CNN based approaches have focused on the problem of sparse and imbalanced label distribution. Chen et al.[8] proposed the use of Cumulative Attributes in an attempt to solve such problems. However, their approach used handcrafted features to learn the Cumulative Attributes. In addition, their approach used Euclidean loss function to learn the Cumulative Attributes. This work adapts the learning of Cumulative Attributes in a CNN based framework. This has the advantage of learning the attributes directly from the input image instead of using handcrafted features. Additionally we experiment with different loss functions and show that Logarithmic loss is more suitable for learning the Cumulative Attributes.

III. DEEP LEARNED CUMMULATIVE ATTRIBUTES

A. Cumulative Attributes

Consider training data consisting of images X_i each labelled with a real-valued quantity y_i and $i \in \{1,2,...,N\}$, where N is the total number of images. The objective is to learn a regression model which is trained to correctly predict the labels y_i of a set of data not seen during training, given their corresponding input images X_i. Cumulative Attributes is defined as an intermediate representation C_i obtained by transforming the original labels y_i into a vector whose elements are given by:

$$C_i^j = \begin{cases} 1, & \text{if } j <= y_i \\ 0, & \text{if } j > y_i \end{cases} \quad (1)$$

where C_i^j represents the jth element of the vector C_i. The size the vector C_i is usually defined as the maximum possible value of y in the given dataset (assuming y can only take integer values). Therefore, for any label y_i, the Cumulative Attribute C_i is a vector whose first y_i elements are 1 and the rest of the elements are set to 0. In this way, the continuous (scaler) values represented by labels y_i are transformed into a vector with binary values. This transformation makes it suitable for posing the learning objective as a classification problem instead of a regression problem. In a straightforward approach where each unique value of y is treated as a separate class, the error in predicting an example with ground truth label say 4.0 as 10.0 is the same as the error in predicting the same as 5.0. On the other hand, in case of Cumulative Attributes, the error increases proportionally to the difference between ground truth and prediction.

B. Deep CNN learning with Cumulative Attributes

The training for CNN regression models with Cumulative Attributes as intermediate features, is done as a two step procedure. In the first step, the CNN is trained to output the Cumulative Attribute vector. In the second step, a regression layer is added which is trained to produce the final real-valued output y (See Fig.1(a)). Below we describe these steps in more detail.

Training the Cumulative Attribute layer:
In order to learn the Cumulative Attributes (CA) in a Convolutional Neural Network (CNN), we define the output layer of any CNN to represent the CA layer, so the number of output nodes is equal to the length of the CA vector. The objective of the CNN is to predict the CA vector obtained

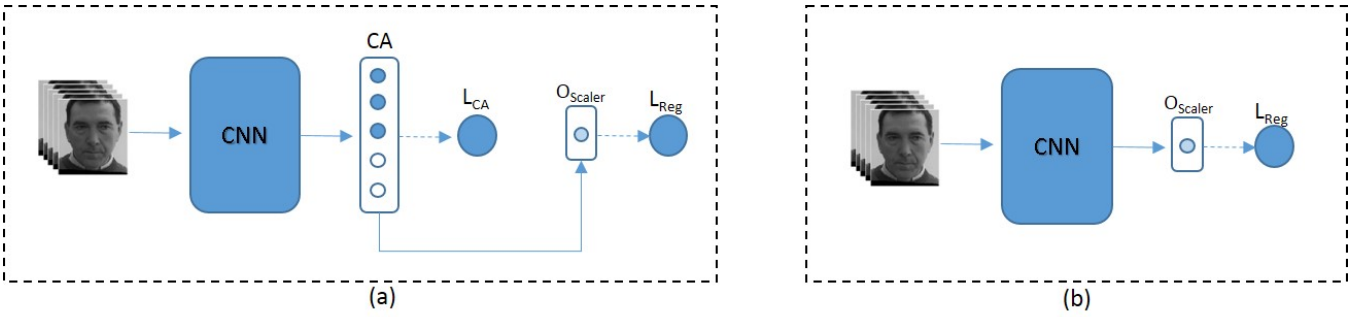

Fig. 1. Comparison of our proposed CA-CNN approach (left) with non-cumulative attribute based CNN (right). Here L_{CA} represents the Loss layer for learning the CA layer, O_{scaler} is the scaler output prediction from the network and L_{Reg} represents the loss layer for learning the regression layer. For CA-CNN approach, the Cumulative attributes are learnt from the CNN and its output is used for learning the final regression layer. For the non-cumulative attribute CNN (NCA-CNN), the regression layer is learnt directly from the CNN.

by transforming the original label as described in Sec.III-A, for each training example. We compare and contrast two loss functions used to train the CNNs, the Euclidean and Log-loss.

CA-Euclidean loss: This loss function uses the Euclidean distance metric to compute the loss L_{euc} given as:

$$L_{euc} = \frac{1}{N}\sum_i\sum_j (P_i^j - C_i^j)^2 \tag{2}$$

where P_i^j is the predicted output of the CNN at the jth output node, for the ith training example. Similar loss function was used in [8] for learning Cumulative Attributes. The only difference here is that there is no regularization term here. The regularization on CNN weights is done using the dropout technique as described in [50].

CA-Log loss: This loss function uses Logarithmic loss L_{log} which can be computed from the predicted output and the ground-truth label as follows,

$$L_{log} = -\frac{1}{N}\sum_i\sum_j Log(Z_i^j + \eta) \tag{3}$$

where,

$$Z_i^j = \begin{cases} P_i^j, & \text{if } C_i^j == 1 \\ 1 - P_i^j, & \text{if } C_i^j == 0 \end{cases} \tag{4}$$

and η is regularization parameter added to avoid the loss becoming infinite. We set this parameter to be 10^{-4}. The CNN can be trained for predicting the CA vector by using the backpropagation and mini-batch gradient descent method.

Training the regression layer:
Although the CNN can be trained to predict the CA vector, which can then be used to predict the original label values y, it is desirable to have a single network that directly outputs y. For this reason, an additional fully-connected layer (with output size equal to 1) is added to the CNN. This layer takes input from the CA layer and is trained to predict y. For training the regression layer, a Euclidean loss is computed as given below,

$$L_{reg} = \sum_i (y_i - \widehat{y}_i)^2 \tag{5}$$

where $\widehat{y}_i$ is the output from the regression layer for the ith training example. Training is again done using the backpropagation and mini-batch gradient descent methods. During training, the weights in all layers prior to the regression layer are frozen. Only the weights in the regression layer are allowed to update during the training. After the regression layer has been trained, the weights in all the layers can be fine-tuned by doing an end-to-end training of the entire network.

C. CNN architecture

In order to show how generally applicable our proposed Deep Learned Cumulative Attribute method is, we will evaluate it on three different regression tasks: pain intensity estimation, facial Action Unit (AU) intensity estimation and Age estimation. For each of these tasks, facial images are used as input. The CNN architectures that we used for each of these tasks are described below.

Facial AU intensity estimation: For this task, we used the dynamic shape and appearance encoding architecture as proposed by Jaiswal & Valstar [20]. This architecture consists of 2 input streams: the first stream takes in as input a short sequence of facial image regions, while the second one takes a sequence of binary shape masks corresponding to the same facial regions, computed from facial landmarks. The facial region is defined according to domain knowledge of the facial AU being modelled. The network consists of 3 Convolutional layers and 2 fully connected layers. For more details please refer to [20].

Pain intensity estimation: For this task, we used a similar architecture as used for facial AU intensity estimation. However, in this case instead of using small facial regions of the face as input, the entire face bounding box is used for computing the input sequence of images and binary shape masks. This was done because the expression for pain is not localized to only one particular region of the face as compared to facial AUs which usually cause local appearance and/or shape changes.

Age estimation: For this task, we used a CNN architecture consisting of 6 Convolutional layers and 3 fully connected layer. All convolutional layers consists of filters of size

3x3. The first 2 convolutional layers consists of 64 filters followed by a max-pool layer. The next two convolutional layers consists of 128 filters followed by another max-pool layer. The last two convolutional layers consists of 256 filters followed by a third max-pool layer. The convolutional layers are followed by 2 fully-connected layers consisting of 3072 filters each. The last fully-connected layer is the output layer whose size depends upon the desired number of cumulative attributes. For the FG-Net database the size of the output layer was set to 69. In contrast to the above architectures, this network consists of only 1 input stream which accepts a registered full face image of size 70x70 pixels (grayscale). No shape masks were used as input to this architecture because the objective here was to learn features which are invariant to changes in facial expressions which could dominate the variation in shape. Also, since most databases available for evaluating age estimation models consist of only static images, we use only one image as input to CNN instead of a sequence of images as in the above two architectures.

IV. EXPERIMENTS

We evaluated our proposed method on two commonly used benchmark datasets: the UNBC-McMaster Shoulder Pain database for facial Action Unit intensity and pain intensity estimation, and the FG-Net Ageing database for age estimation. Below we give a brief description of the databases we used for training and testing our models.

A. Evaluation databases

UNBC-McMaster Shoulder Pain Database. The UNBC McMaster database [31] is a collection of facial expressions of patients suffering from shoulder pain. It consist of 200 videos from 25 subjects with an average of seven frames per subject. Each subject has two categories of videos. In the first set, the patients are asked to move their arms while in the second set, a physiotherapist moves the patient's arms within bearable limits. All videos are annotated for action units (AU) and pain intensity on a frame level. The Prkachin and Solomon (PSPI) [43] 16 point pain scale is used for the frame-wise pain annotation. The action units are scored on a 5-point intensity scale. ranging from A (min =1) to E (max = 5). Sequence level pain annotation are also included for all videos based on Observer Pain intensity (OPI) ratings.

TABLE I
PERCENTAGE DISTRIBUTION OF PAIN FRAMES IN THE MCMASTER DATABASE

Pain levels	0	1-2	3-4	5-6	7-10	11-16
% dist.	82.71	10.87	4.57	1.06	0.47	0.32

While the McMaster shoulder pain database provides a good platform for pain estimation from facial expressions, it suffers from high data imbalance. Only 17% of the frames contain pain expression compared to 83% of no-pain signal frames. Table I shows the percentage distribution of pain-signal frames in the database. A similar pattern is observed for the AUs with some high level AUs having little or no representation as shown in Table II. Cumulative attributes target datasets with sparse data representation for certain classes which also have shared characteristics. An incremental feature relationship exists as we move from the low level AU or pain intensities to higher levels. Thus, the McMAster database is an ideal platform to evaluate our proposed CA methodology.

TABLE II
AU FRAME DISTRIBUTION IN THE MCMASTER DATABASE

AUs	A	B	C	D	E	Total
4	202	509	225	74	64	1074
6	1776	1663	1327	681	110	5557
7	1362	991	608	305	100	5557
9	93	151	68	76	35	423
10	171	208	63	61	22	525
12	2145	1799	2158	736	49	6887
20	286	282	118	0	20	706
25	767	803	611	138	88	2407
26	431	918	265	478	1	2093
43	2434	-	-	-	-	2434

FG-Net Aging Dataset The FG-Net Aging dataset [41] consists of 1002 images from 82 subjects with ages ranging from 0 to 69 years. Each subject has an average of 12 age-separated images. Figure 2 shows the age distribution in the database. Similar to the McMaster pain database, the FG-Net data set exhibits high data sparsity for higher ages with the number of examples decreasing drastically as we move up the age hierarchy. This is more evident for ages 50 and higher. Though not a desirable database feature, it allows us to show how deep learned CA features can improve performance of learning algorithms when compared to deep learned features that don't use CA features (called NCA below, for non-Cumulative Attribute features).

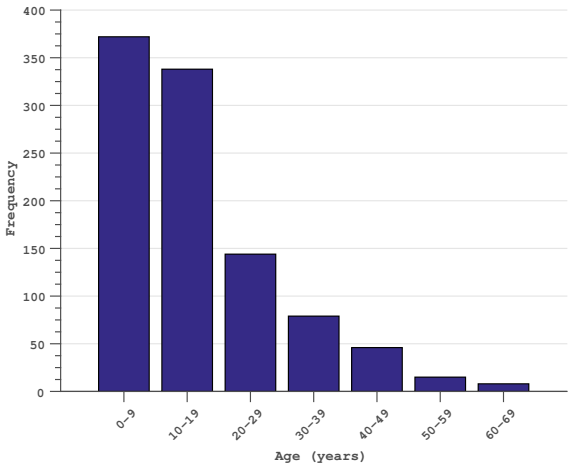

Fig. 2. Age distribution in the FG-Net Aging dataset

B. Evaluation

Below we describe the evaluation protocol and results obtained for each task. Note that our evaluation protocol is designed to explicitly and convincingly show the benefit of using Cumulative Attributes in end-to-end learning over traditional Deep Learning approaches that either directly regress the real-valued output or approximate it by a discretising the output and casting the problem as a multiclass task. We do not attempt to engineer our system so as to attain state of the art results for the three use-cases.

Pain estimation: For pain intensity estimation, the models were evaluated on the McMaster database using a leave-one-subject-out cross-validation strategy in which all video frames from one subject are kept aside for testing and the training is done using the frames from all other subjects. This process is repeated for each subject and the results are averaged over all subjects. We trained 3 separate kinds of models, one model was trained in which no Cumulative Attributes (NCA) were used and the CNNs were trained to directly predict the pain intensities (See Fig.1(b)). A second category of model was trained in which the Cumulative Attribute layer was used with Euclidean loss (CA-Euclidean loss). The third and final category of model was trained in which Cumulative Attribute layer was used with Log loss (CA-Log loss).

Table III shows the performance of all the three types of models in terms of Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE). In this table, it can be observed that in terms of both PCC and RMSE measures the models using Cumulative Attributes perform significantly better than the NCA models which shows the advantage of learning Cumulative Attributes in the intermediate layer. It can also be observed that in terms of PCC measures, the CA-Log loss models perform significantly better than the CA-Euclidean loss models indicating that Log loss is more suitable for learning the CA layer. Although, in terms of RMSE measure, the performance of the CA-Log loss model is slightly less as compared to CA-Euclidean loss, it should be noted that RMSE is not a good measure for highly imbalanced and sparse datasets like McMaster where majority of the pain intensity labels are zero.

For pain estimation, we also experimented with treating each pain intensity independently and posing it as a multi-class classification problem. Similar to the regression experiments, three kinds of models were trained: CNN without CA layer(NCA), CNN with CA-Log loss and CNN with CA-Euclidean loss. The only difference here is that instead of training a regression layer with a scaler output, a multi-output classification layer was trained with softmax loss. Table IV shows the result for this set of experiments. Here again, the CA models perform better than NCA models with CA-Log loss outperforming CA-Euclidean loss in terms of PCC measure.

Facial AU intensity estimation: The facial AU intensity models were also trained and evaluated on the McMaster database. For this set of experiments, a five-fold subject independent cross-validation approach was used for evaluating the performance. Similar to experiments for pain intensity, we trained three kinds of models: NCA, CA-Euclidean loss and CA-Log loss.

TABLE III
RESULTS OF PAIN ESTIMATION ON MC MASTER DATASET (REGRESSION).

	NCA	CA-Euclidean loss	CA-Log loss
PCC	0.44	0.47	0.53
RMSE	1.30	1.20	1.23

TABLE IV
RESULTS OF PAIN ESTIMATION ON MC MASTER DATASET (MULTI-CLASS).

	NCA	CA-Euclidean loss	CA-Log loss
PCC	0.33	0.36	0.41
RMSE	1.28	1.17	1.19

Table V shows the performance of facial AU intensity models in terms of Pearson Correlation Coefficient (PCC). In this table it can be observed that in general CA models perform better than the NCA models. However, it is interesting to note that CA-Log loss models perform significantly better than the CA models learnt using Euclidean loss. It should again be noted that RMSE is not a good measure in this case as majority of the intensity labels are zero.

TABLE V
RESULTS (PCC) FOR AU INTENSITY ESTIMATION ON MC MASTER DATASET.

AU	NCA	CA-Euclidean loss	CA-Log loss
4	0.02	0.01	0.02
6	0.40	0.40	0.4
7	0.36	0.39	0.45
9	0.04	0.04	0.05
10	0.03	0.13	0.41
12	0.42	0.36	0.34
20	0.0003	0.04	0.02
25	0.33	0.18	0.42
26	0.11	0.14	0.11
43	0.22	0.32	0.25
Mean	0.19	0.20	0.25

Age estimation: The models for age estimation were evaluated on the FGNet database. The performance was evaluated using a 21 fold subject independent cross-validation. Similar to the experiments for pain and AU intensity estimation, we evaluated three kinds of models: NCA, CA-Euclidean loss and CA-Log loss.

Table VI shows the performance of age estimation models in terms of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC) ans Cumulative Score (CS) defined in [14]. We used a threshold of 5 for calculating the CS measure. From the table, it can be observed that CA models outperform NCA models. It can also be observed that CA models trained with Log loss

outperform CA models trained with Euclidean loss, by a large margin.

TABLE VI
RESULTS OF AGE ESTIMATION ON FGNET DATABASE.

	NCA	CA-Euclidean loss	CA-Log loss
RMSE	9.41	9.33	8.56
MAE	7.1	7.15	5.91
PCC	0.68	0.70	0.75
CS	0.44	0.44	0.58

V. CONCLUSIONS

We introduced a CNN based framework for learning regression models which can handle sparse and imbalanced datasets. The proposed framework employs an intermediate Cumulative Attribute layer from which the final output layer is learnt. The Cumulative Attribute layer was learnt using two different loss function: Euclidean-loss and Log-loss. Experiments show that in general the CA-CNN framework performs better than Non-Attribute based CNNs. It was also shown that CA layer trained with Log-loss significantly outperforms CA layer trained with Euclidean loss, on a number tasks which includes pain intensity, facial AU intensity and age estimation.

ACKNOWLEDGEMENTS

The research reported in this paper was conducted by the NIHR Nottingham Biomedical Research Centre. This work has been funded by the National Institute for Health Research. The work of Egede is additionally supported by the International Doctoral Innovation Centre (IDIC), Ningbo Education Bureau, Ningbo Science and Technology Bureau and The University of Nottingham. The views represented are the views of the authors alone and do not necessarily represent the views of the Department of Health in England, NHS, or the National Institute for Health Research.

REFERENCES

[1] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F. Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M. Prkachin, and Patricia E. Solomon. The painful face – Pain expression recognition using active appearance models. *Image and Vision Computing*, 27(12):1788–1796, nov 2009.

[2] M. S. H. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A. C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. J. Watson, A. C. d. C. Williams, M. Pantic, and N. Bianchi-Berthouze. The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal emopain dataset. *IEEE TAC*, 7(4):435–451, Oct 2016.

[3] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE, 2015.

[4] Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, and Kang Lee. Automatic Decoding of Facial Movements Reveals Deceptive Pain Expressions. *Current Biology*, 24(7):738–743, mar 2014.

[5] Dong Cao, Zhen Lei, Zhiwei Zhang, Jun Feng, and Stan Z Li. Human age estimation using ranking svm. In *CCBR*, pages 324–331. Springer, 2012.

[6] Chuan-Yu Chang and Jia-Jing Li. Application of deep learning for recognizing infant cries. In *2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2. IEEE, 2016.

[7] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on*, pages 585–592. IEEE, 2011.

[8] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2467–2474, 2013.

[9] Ke Chen, Kui Jia, Zhaoxiang Zhang, and Joni-Kristian Kämäräinen. Spectral attribute learning for visual regression. *Pattern Recognition*, 66:74–81, 2017.

[10] J. Egede and M. Valstar. Cumulative attributes for pain estimation. In *2017 ACM International Conference on Multimodal Interaction*, in press.

[11] J. Egede, M. Valstar, and B. Martinez. Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. In *2017 12th IEEE International FG 2017*, pages 689–696, May 2017.

[12] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, pages 433–440, 2008.

[13] Corneliu Florea, Laura Florea, and Constantin Vertan. *Learning Pain from Emotion: Transferred HoT Data Representation for Pain Intensity Estimation*, pages 778–790. Springer International Publishing, Cham, 2015.

[14] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 29(12):2234–2240, 2007.

[15] Isabel Gonzalez, Werner Verhelst, Meshia Oveneke, Hichem Sahli, and Dongmei Jiang. Framework for combination aware au intensity recognition. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 602–608. IEEE, 2015.

[16] Amogh Gudi, H Emrah Tasli, Tim M den Uyl, and Andreas Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–5. IEEE, 2015.

[17] Guodong Guo and Guowang Mu. Joint estimation of age, gender and ethnicity: Cca vs. pls. In *Automatic face and gesture recognition (fg), 2013 10th ieee international conference and workshops on*, pages 1–6. IEEE, 2013.

[18] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S Huang. Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 112–119. IEEE, 2009.

[19] Chen Huang, Chen Change Loy, and Xiaoou Tang. Unsupervised learning of discriminative attributes and visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5175–5184, 2016.

[20] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[21] Shashank Jaiswal, Michel F Valstar, Alinda Gillott, and David Daley. Automatic detection of adhd and asd from expressive behaviour in rgbd data. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 762–769. IEEE, 2017.

[22] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous Pain Intensity Estimation from Facial Expressions. In *Advances in Visual Computing*, pages 368–377. Springer Science + Business Media, 2012.

[23] Sebastian Kaltwang, Sinisa Todorovic, and Maja Pantic. Doubly sparse relevance vector machine for continuous facial behavior estimation. 2015.

[24] Sebastian Kaltwang, Sinisa Todorovic, and Maja Pantic. Latent trees for estimating intensity of facial action units. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 296–304, 2015.

[25] M. Kchele, P. Thiam, M. Amirian, F. Schwenker, and G. Palm. Methods for person-centered continuous pain intensity assessment

from bio-physiological channels. *IEEE Journal of Selected Topics in Signal Processing*, 10(5):854–864, Aug 2016.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[27] Li-Jia Li, Hao Su, Yongwhan Lim, and Fei-Fei Li. Objects as attributes for scene classification. In *ECCV Workshops (1)*, pages 57–69, 2010.

[28] Yongqiang Li, S Mohammad Mavadati, Mohammad H Mahoor, Yongping Zhao, and Qiang Ji. Measuring the intensity of spontaneous facial action units with dynamic bayesian network. *Pattern Recognition*, 48(11):3417–3427, 2015.

[29] Gwen C. Littlewort, Marian Stewart Bartlett, and Kang Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797–1803, nov 2009.

[30] Jianyi Liu, Yao Ma, Lixin Duan, Fangfang Wang, and Yuehu Liu. Hybrid constraint svr for facial age estimation. *Signal Processing*, 94:576–582, 2014.

[31] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Face and Gesture 2011*, pages 57–64, March 2011.

[32] Patrick Lucey, Jeffrey F Cohn, Iain Matthews, Simon Lucey, Sridha Sridharan, Jessica Howlett, and Kenneth M Prkachin. Automatically detecting pain in video through facial action units. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(3):664–674, 2011.

[33] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, Sien Chew, and Iain Matthews. Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database. *Image and Vision Computing*, 30(3):197–205, 2012.

[34] Mohammad H Mahoor, Steven Cadavid, Daniel S Messinger, and Jeffrey F Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 74–80. IEEE, 2009.

[35] A. Majumder, L. Behera, and V. K. Subramanian. Gmr based pain intensity recognition using imbalanced data handling techniques. In *2016 International Conference on Signal and Information Processing (IConSIP)*, pages 1–5, Oct 2016.

[36] D. L. Martinez, O. Rudovic, and R. Picard. Personalized automatic estimation of self-reported pain intensity from facial expressions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2318–2327, July 2017.

[37] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[38] Nikolay Neshov and Agata Manolova. Pain detection from facial characteristics using supervised descent method. In *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015 IEEE 8th International Conference on*, volume 1, pages 251–256. IEEE, 2015.

[39] Jeremie Nicolle, Kevin Bailly, and Mohamed Chetouani. Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE, 2015.

[40] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4920–4928, 2016.

[41] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. Overview of research on facial ageing using the fg-net ageing database. *IET Biometrics*, 5(2):37–46, 2016.

[42] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[43] Kenneth M. Prkachin and Patricia E. Solomon. The structure reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, oct 2008.

[44] Jie Qin, Yunhong Wang, Li Liu, Jiaxin Chen, and Ling Shao. Beyond semantic attributes: Discrete latent attributes learning for zero-shot recognition. *IEEE Signal Processing Letters*, 23(11):1667–1671, 2016.

[45] Neeru Rathee and Dinesh Ganotra. A novel approach for pain intensity detection based on facial feature deformations. *Journal of Visual Communication and Image Representation*, 33:247–254, 2015.

[46] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. *Automatic Pain Intensity Estimation with Heteroscedastic Conditional Ordinal Random Fields*, pages 234–243. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[47] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE transactions on pattern analysis and machine intelligence*, 37(5):944–958, 2015.

[48] Farshad Samadi. Human age-group estimation based on anfis using the hog and lbp features. *Electrical and Electronics Engineering: An International Journal (ELEIJ)*, 2(1):21–29, 2013.

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

[50] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[51] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

[52] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–8. IEEE, 2015.

[53] Robert Walecki, Vladimir Pavlovic, Björn Schuller, Maja Pantic, et al. Deep structured learning for facial action unit intensity estimation. *arXiv preprint arXiv:1704.04481*, 2017.

[54] Xiaolong Wang, Rui Guo, and Chandra Kambhamettu. Deeply-learned feature for age estimation. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 534–541. IEEE, 2015.

[55] Philipp Werner, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald C Traue. Automatic pain recognition from video and biomedical signals. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4582–4587. IEEE, 2014.

[56] Dong Yi, Zhen Lei, and Stan Z Li. Age estimation by multi-scale convolutional network. In *Asian Conference on Computer Vision*, pages 144–158. Springer, 2014.

[57] Felix X Yu, Liangliang Cao, Rogerio S Feris, John R Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 771–778, 2013.

[58] Zuhair Zafar and Nadeem Ahmad Khan. Pain Intensity Evaluation through Facial Action Units. In *2014 22nd International Conference on Pattern Recognition*. Institute of Electrical & Electronics Engineers (IEEE), aug 2014.

[59] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2016.

[60] R. Zhao, Q. Gan, S. Wang, and Q. Ji. Facial expression intensity estimation using ordinal information. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3466–3474, June 2016.

[61] J. Zhou, X. Hong, F. Su, and G. Zhao. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1535–1543, June 2016.

[62] Yuqian Zhou, Jimin Pi, and Bertram E Shi. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 872–877. IEEE, 2017.