

Training Deep Neural Networks with Different Datasets In-the-wild: The Emotion Recognition Paradigm

Dimitrios Kollias
Department of Computing
Imperial College London
United Kingdom
dimitrios.kollias15@imperial.ac.uk

Stefanos Zafeiriou
Department of Computing
Imperial College London
United Kingdom
&
Center for Machine Vision and Signal Analysis
University of Oulu
Finland
s.zafeiriou@imperial.ac.uk

Abstract—A novel procedure is presented in this paper, for training a deep convolutional and recurrent neural network, taking into account both the available training data set and some information extracted from similar networks trained with other relevant data sets. This information is included in an extended loss function used for the network training, so that the network can have an improved performance when applied to the other data sets, without forgetting the learned knowledge from the original data set. Facial expression and emotion recognition in-the-wild is the test bed application that is used to demonstrate the improved performance achieved using the proposed approach. In this framework, we provide an experimental study on categorical emotion recognition using datasets from a very recent related emotion recognition challenge.

Index Terms—deep neural network training; classification; clustering internal representations; extended loss function; domain adaptation; transfer learning; emotion recognition in-the-wild; .

I. INTRODUCTION

Many real life problems are represented by a variety of data sets which may possess different characteristics. In such cases learning to classify correctly one data set does not generalize well in the other sets. Emotion recognition based on facial expressions is such a problem, due to the variations in expression of emotions among different persons, as well as to the different ways of labeling emotional states by different annotators. It would be of great interest if we could use a training data set to design a deep neural network (DNN), taking into account some knowledge about another set, so as to improve generalization of the network when applied to the other data set, without forgetting the original knowledge of it.

In this paper we deal with categorical emotion recognition based on facial expressions, treated as a classification problem in the seven primary emotional states, i.e., happiness, anger, fear, disgust, sadness, surprise and neutral state. Our approach can be extended to the dimensional emotion recognition problem as well, through discretization of the 2D continuous Valence Arousal Space [19]. Moreover, we focus on

emotion recognition in-the-wild, i.e., on recognizing emotions expressed in real life, uncontrolled environments [25], [14], [5].

A related Grand Challenge, the EmotiW one [5], has been constantly organized during the last five years, providing the AFEW data sets, consisting of videos showing persons expressing their emotions in real life. The Challenge provide training and validation data sets for designing emotion recognition approaches and test data for evaluating the performance of these approaches.

Since the data are in the form of video sequences, we focus on Convolutional and Recurrent Neural Network (CNN-RNN) architectures. CNN-RNNs have achieved the best performances in all recent contests [8], [5], [6].

Deep Convolutional Neural Networks (CNNs) [15], [16], [2] include convolutional layers with feature maps composed of neurons with local receptive fields and shared weights and pooling layers, generating condensed representations. The resulting strength is the automatic hierarchical generation of rich internal representations, which are fed next to fully connected layers, for classification, or prediction, purposes.

Recurrent neural networks (RNNs) have the ability to model time varying asynchronous patterns in audio and video [24]. Since emotion events do not appear in a single frame, RNNs can follow and capture the events' sequential evolution. They include hidden layer(s) with long time dependencies. Their hidden units are functions of inputs and of hidden states, with their input being image pixel values, or features extracted from the images. Using the neuron Long Short-Term Memory (LSTM) model, one can overcome the backpropagation vanishing effect by introducing appropriate gates (input, update, output) and a cell state. BLSTMs are bidirectional LSTMs [3], processing input data in two directions. Other variants of LSTMs are the Gated Recurrent Units (GRUs) [4] that have two gates (forget, update) rather than three.

It should be mentioned that there is a big discrepancy be-



Fig. 1. Video frames extracted from the Aff-Wild Dataset

tween the EmotiW training and validation data sets. Different annotation strategies have been used in them, based on weak annotation labeling and clustering. As a consequence, there are different annotation behaviors and biases between the two sets. It should be mentioned that the generalization accuracy when training a network with either of these sets and testing on the other set is around 40%. The test data characteristics are close to the validation data.

In the paper we consider the problem of using the EmotiW 2017 training data and some knowledge about the validation data, to design a deep neural architecture with improved generalization accuracy on the validation data set. We propose a domain adaptation approach, by generating respective representations from the two datasets, i.e., the training and validation ones in our case, and by then trying to match the statistical distribution of these representations. As a consequence, a classification scheme, trained on the first data set, which would take into account the statistics of representations of the second dataset, would be able to improve its performance when applied to the latter one. Evidently, our first data set is the EmotiW 2017 training data and the second data set includes the respective validation data.

We realize these representations, by extracting and using internal features generated by deep convolutional and recurrent neural (CNN-RNN) architectures, separately trained with the validation and with the training EmotiW data sets. In this framework we propose and use a new loss function for training the CNN-RNN system with the EmotiW training set, including minimization of the difference in the statistics produced by this system and a respective one applied to the validation set.

The paper is organized as follows. Section II describes the facial expression and emotion recognition in-the-wild problem, focusing on categorical emotion recognition, and particularly, on classifying audiovisual data in one of the seven basic emotion categories. Section III presents the proposed approach, describing the new error function adopted in training, as well as a new formulation extracting internal features and representations from trained CNN-RNN architecture. Section IV provides the experimental study illustrating the performance obtained through the proposed novel approach for categorical emotion recognition. Conclusions and reference to planned future developments are provided in Section V of the paper.

II. EMOTION RECOGNITION IN-THE-WILD

A common way to represent emotions is through the categorical model, which uses the six basic emotion categories defined by [9], i.e., happiness, sadness, anger, disgust, surprise and fear, as well as the neutral category to represent human behaviors. This is due to several psychophysical experiments suggesting that the perception of emotions by humans is categorical [10].

Another way to think about human emotions is through the dimensional model [21] [23]. This model shows that emotions can be distributed on a two dimensional circular space which contains valence and arousal dimensions, with the center representing neutral valence and a medium level of arousal. Emotional states are represented in this model as any level of valence and arousal, or at a neutral level for one or both of them.

Both models have been recently used by various approaches targeting what is called emotion recognition in-the-wild, meaning recognizing emotions expressed by different persons in their everyday life, i.e., in uncontrolled environments. New datasets have been generated, by aggregating and annotating, according to one of these models, videos from YouTube and Movies [25], [5] and used in Emotion Recognition Challenges. Fig. 1 shows some examples of image frames, taken from the Aff-Wild Database [25], showing facial expressions captured in-the-wild.

In the following we focus on the EmotiW 2017 Challenge, which is the most recent one targeting categorical emotion recognition in-the-wild. This challenge was the fifth in a series, starting on 2013 [7], with all of them focusing on the topic of emotion recognition from audio-visual data in (real world) uncontrolled conditions.

The series of EmotiW challenges make use of data from the Acted Facial Expression in-the-wild (AFEW) dataset. This dataset is a dynamic temporal facial expressions data corpus consisting of close to real world scenes extracted from movies and reality television shows. In total it contains 1809 videos. The whole dataset is split into three sets: training set (773 video clips), validation set (383 video clips) and test set (653 video clips). It should be emphasized that both training and validation sets are mainly composed of real movie records, however 114 out of 653 video clips in the test set are real TV clips, increasing, therefore, the difficulty of the challenge. The number of subjects is more than 330, aged 1-77 years. The annotation is according to 7 facial expressions (Anger, Disgust,

Fear, Happiness, Neutral, Sadness and Surprise), as shown in Fig. 2. The challenges focus on audio-video classification of each clip into the seven basic emotion categories.

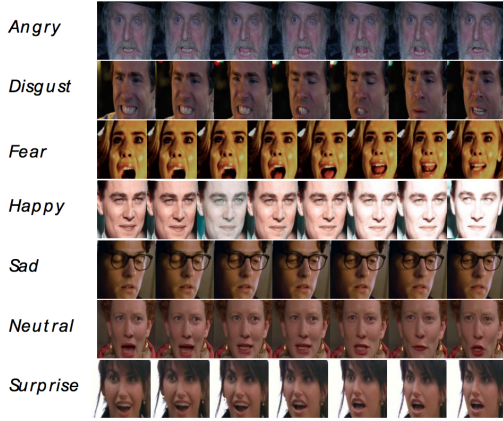


Fig. 2. Parts of Image Sequences from the AFEW dataset

III. THE PROPOSED APPROACH

Let us assume that we have two datasets that include videos of different persons expressing their emotions in-the-wild; these have been annotated with reference to the seven basic emotion categories model. We will use the one dataset for training a Convolutional and Recurrent Neural Network (CNN-RNN) architecture (let us call it NetA) and the other for testing NetA's performance on it. This is different from transfer learning, which encompasses catastrophic forgetting, i.e., deterioration of the performance of NetA on its original training set, after being fine-tuned with the new data set.

The proposed approach starts with the latter, test dataset. We train another CNN-RNN architecture (let us call it NetB) with these data, achieving the best possible emotion recognition accuracy. Such a network consists of the convolutional part, including one or more fully connected layers, using a ReLU type of neuron activation units, followed by the recurrent network part, which contains one or more layers, with B-LSTM, or GRU types of neurons.

Let us now focus on the outputs of the last layer of the trained RNN part of NetB. In fact, it is through these output values that the network produces its final outputs, corresponding to the seven emotion categories. Assuming that a satisfactory performance is obtained by training NetB, it would be desirable to have NetA when trained with its respective data set - generating output values, in its last before the output layer, which are close to the ones produced by NetB. If this happened, this would mean that training of NetA also managed to bring its own outputs closer to the ones generated by NetB. This is a desirable task, since both NetA and NetB target recognition of persons' emotions in different, randomly selected environments.

Our proposed approach is based on clustering the above-mentioned extracted internal representations of NetB into

seven clusters, corresponding to the targeted emotion categories and using the derived cluster centers as desired outputs for the respective representations generated at the corresponding layer during the training of NetA.

In particular, let us denote, for the m -th input sample, by

$$U_m = \{u_m^1, \dots, u_m^n\} \quad (1)$$

a vector with the CNN-RNN last (before the output) hidden layer neurons outputs, assuming it contains n neurons. We perform clustering of the U values in seven clusters, corresponding to the seven basic emotion categories. The k -means algorithm [17] can be used for this task.

Let

$$Z_i = \{z_i^1, \dots, z_i^n\} \quad (2)$$

denote the seven cluster centroids ($i = 1, \dots, 7$), in the n -dimensional representation space. These centroids, with their labels, constitute the information used in the proposed approach for adapting the training of NetA towards a clustered representation related to the best performance of NetB on its respective data set.

In the following, we introduce the above centroids, in the form of additional desired responses during training of NetA. Let us denote by

$$O_m = \{o_m^1, \dots, o_m^7\} \quad (3)$$

a vector with the seven NetA outputs, also corresponding to the basic emotion categories and by

$$X_m = \{x_m^1, \dots, x_m^n\} \quad (4)$$

a respective vector with the NetA last (before the output) hidden layer neurons outputs. In (2)-(4), m also denotes the m -th training input data sample of NetA.

Originally, training of NetA is done through minimization, with respect to the network weights, of the mean squared error between NetA outputs O_m and desired outputs, say D_m , defined as follows:

$$E_a = \frac{1}{7M} \sum_{m=1}^M \sum_{i=1}^7 (d_m^i - o_m^i)^2 \quad (5)$$

for m covering all M data samples and the squared difference in (5) being computed between the respective seven components of D_m and O_m .

In our formulation, we include a second term in error minimization, derived as follows. Let us define, using (2) and (4), for $i = 1, \dots, 7$, vector Y_m^i and the values V_m^i :

$$Y_m^i = Z_i - X_m = \{z_i^1 - x_m^1, \dots, z_i^n - x_m^n\} \quad (6)$$

$$V_m^i = Y_m^i * (Y_m^i)^T \quad (7)$$

where T denotes the transpose of the vector.

We request minimization of the V_m^i value corresponding to the desired output category and maximization of the rest V_m^i values corresponding to the other categories.

To achieve this, we normalize these V_m^i values, either using a softmax f function, or linearly, by dividing each one with the sum of all of them. Moreover, we subtract each normalized value from unity, so that the minimum distances correspond to maximum output values.

This leads us to define a new Mean Squared Error criterion, between the computed values and the corresponding desired outputs D_m , as follows:

$$E_b = \frac{1}{7M} \sum_{m=1}^M \sum_{i=1}^7 (d_m^i - [1 - f(V_m^i)])^2 \quad (8)$$

where, similarly to (5) the squared difference is computed over all training data samples.

We target minimization of both Error Criteria in (5) and (8) during training of NetA, through the following combined Loss Function,

$$E_{tot} = \lambda E_a + (1 - \lambda) E_b \quad (9)$$

by selecting appropriate values for the tuning parameter λ in $[0, 1]$.

Fig. 3 shows the proposed approach, which can be repeated with any new NetB, providing different facets of NetA that can be used for analyzing corresponding datasets in-the-wild.

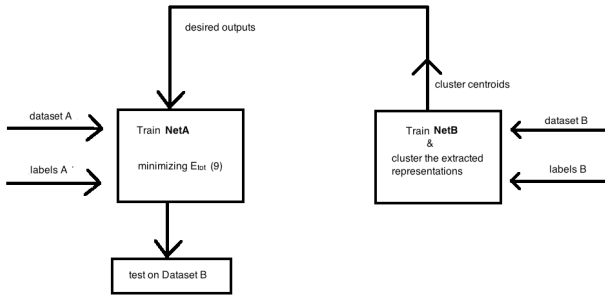


Fig. 3. The proposed procedure for generating different facets of a Deep Neural Architecture designed for Emotion Recognition in-the-wild

In the experimental study which follows, we will illustrate that, using (9) to train a CNN-RNN network for categorical emotion recognition, provides improved performance of the trained NetA, when applied on the validation EmotiW data set. We will also show that using (9) for NetA training leads the network to better perform on the training set as well.

IV. EXPERIMENTAL STUDY

A. The Deep Neural Networks used in the experiments

The Deep Neural Network architecture proposed in this study, is an end-to-end model including both CNN and RNN components. CNNs are used to extract high level features from the input images, while the RNN exploits the sequential nature of the input data to provide the final predictions.

The CNN subsystem matches the VGG-Face architecture [18], a model trained for face verification. Both Bidirectional LSTM and GRU cells were considered as units of the RNN subsystem. The output of the first fully connected layer of the CNN is fed as an input to the RNN, which in turn outputs the final prediction of the system.

Table I shows the configuration of the architecture. It is composed of 9 blocks. For each convolutional layer the parameters are denoted as (channels, kernel, stride) and for the max pooling layer as (kernel, stride). It also shows the respective number of units of each fully connected layer.

TABLE I
THE CNN-RNN ARCHITECTURE

block 1	2× conv layer 1× max pooling	(64, 3 × 3, 1 × 1) (2 × 2, 2 × 2)
block 2	2× conv layer 1× max pooling	(128, 3 × 3, 1 × 1) (2 × 2, 2 × 2)
block 3	3× conv layer 1× max pooling	(256, 3 × 3, 1 × 1) (2 × 2, 2 × 2)
block 4	3× conv layer 1× max pooling	(512, 3 × 3, 1 × 1) (2 × 2, 2 × 2)
block 5	3× conv layer 1× max pooling	(512, 3 × 3, 1 × 1) (2 × 2, 2 × 2)
block 6	fully connected 1 dropout layer	4096
block 7	RNN layer 1 dropout layer	128
block 8	RNN layer 2	128

Fig. 4 shows the architecture of the proposed system. The weights from the convolutional and pooling layers of the VGG-Face are initialized from a pre-trained implementation. During the training phase, these parts remain fixed, while only the fully connected layer at the end is actually trained.

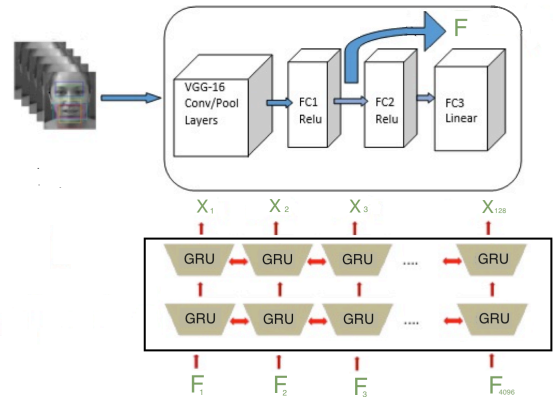


Fig. 4. The features F extracted from CNN FC1 layer are passed to the RNN network; the latter provides the final decision

The next layer is a fully connected one, using the Rectified Linear Unit as its activation function. The CNN, using a linear-activated fully connected layer (FC3), can also provide classification to the 7 basic emotion categories.

The CNN outputs a vector of 4096 high-level features which are extracted from each video frame (denoted by F in Fig. 4).

The RNN processes the $F_1, F_2, \dots, F_{4096}$, corresponding to the number of CNN outputs used and delivers the final representations X_1, X_2, \dots, X_{128} , defined in (4).

The hyper-parameter values which were used to train the CNN-RNN networks, were selected after extensive experimentation: a sequence length size of 80 consecutive data frames was used so as to provide the RNN part of the network with the ability to detect meaningful time correlations in the data; a constant learning rate of 0.001; 4096 hidden units in the fully connected layer of the CNN part; dropout after this CNN layer with a value of 0.5; 128 hidden units of the GRU type in two RNN layers. The weights were initialized from a Truncated Normal distribution with a zero mean and a variance equal to 0.1, while the biases were initialized to 1. Training was performed on a single GeForce GTX TITAN X GPU.

B. The Training and Validation Data

As already mentioned, we used the training data of the EmotiW 2017 dataset to train the NetA CNN-RNN architecture and the validation set of the same Challenge to train the NetB one.

These two datasets possess quite different characteristics, because they are generated from different video clips in-the-wild. As shown in all three methods that produced the best emotion recognition results in the EmotiW 2017 Grand Challenge [12] [13] [22], the performances obtained by deep neural networks, which learned the EmotiW training data, on the validation and test data were quite similar. So, our target is to provide an as good as possible performance on the validation set, while achieving the best possible performance on the training set as well.

In our first experiment we trained a deep CNN-RNN network on the validation data. The best trained (NetB) CNN-RNN achieved a training accuracy of 99%. However, its generalization on the EmotiW 2017 training data was only 33.4%. Nevertheless, we adopted this configuration as NetB, because obtaining an as high as possible accuracy on the validation data has been crucial for the Challenge goals, as was above mentioned. From this network we extracted the U representations defined in (1), for all data in the validation set. About 19,000 frames were provided by the Challenge, extracted from the AFEW videos, with the facial areas cropped. The images were resized to the resolution of $96 \times 96 \times 3$. Their values were normalized to the range $[-1, 1]$. We provided all these as inputs to the network.

Then, we applied k -means clustering on these representations, with $k = 7$, defining seven well separated clusters corresponding to the seven basic emotion categories. We derived the cluster centroids defined in (4), with $n = 128$ and used these in our following experiments. Next, we focused our attention on training the NetA network, using the EmotiW train dataset; about 45,000 video frames, extracted, cropped and pre-processed as in the former case, were used for training.

First, we trained a similar CNN-RNN architecture with the training data, minimizing the error criterion in (5) and selected the architecture which provided the best accuracy on the

validation data set. This architecture provided a best accuracy of 38,4% on the validation data, with its performance on the training data being only 60,6%; illustrating the significant differences between the training and validation data sets.

C. Application of the Proposed Training Procedure

In the following, we implemented the proposed approach described in Section III. We trained a CNN-RNN architecture minimizing the new error criterion defined in (9), using the seven cluster centroids obtained in our former experiment as desired outputs in (9) and testing different values of the parameter λ . Using $\lambda = 1$ derives the network described in the former paragraph; a value of $\lambda = 0$ provides a network trying to replicate the 7 cluster centroids at the outputs of its last RNN hidden layer; a value of $\lambda = 0.5$ pays the same attention to both criteria in (9).

Table II summarizes the obtained accuracy on the validation, as well as training data sets, for different values of λ . It can be easily shown that the best results have been obtained when $\lambda = 0$. The proposed criterion really helped to achieve better results than the original mean squared error (MSE) criterion. This verifies that the role of validation data is of high importance in this categorical emotion recognition in-the-wild problem.

TABLE II
ACCURACIES OF THE PROPOSED ARCHITECTURE FOR THE EMOTIW
TRAINING AND VALIDATION SET FOR DIFFERENT VALUES OF λ

Value of λ	Accuracy on Set	
	Validation	Training
0	0.446	0.69
0.25	0.426	0.67
0.5	0.422	0.622
0.75	0.398	0.611
1	0.384	0.606

Tables III and IV compare the obtained accuracy between $\lambda = 0$ and $\lambda = 1$ of each category of the validation and training sets, respectively. It can be easily shown that the best results have been obtained when $\lambda = 0$. Table V shows the confusion matrix for the validation set when $\lambda = 0$.

It should be mentioned that the performance of the network trained with $\lambda = 0$ is much higher than the baseline 38,81% reported in [5] and better than all other original architectures in the three winning methods in the Audio-video emotion recognition EmotiW 2017 Grand Challenge [12] [13] [22], as shown in Table VI.

Here it can be noted that our network was trained to classify only video frames (and not audio) and then video classification based on frame aggregation was performed. Moreover, no data-augmentation, post-processing of the results or ensemble methodology have been conducted, as was done in the above three winning methods. Also, for training the network we used only the cropped faces provided by the challenge. Our results can be even better, if we use the above methodologies, as well as our own detection and/or normalization procedure, but this is not the main scope of this paper.

TABLE III
ACCURACIES PER CATEGORY FOR THE EMOTIW VALIDATION SET WHEN $\lambda = 0$ AND $\lambda = 1$

Value of λ	Accuracy on Validation Set							
	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise	Total
0	0.492	0.578	0.075	0.217	0.667	0.623	0.217	0.446
1	0.466	0.56	0	0.046	0.635	0.431	0.111	0.384

TABLE IV
ACCURACIES PER CATEGORY FOR THE EMOTIW TRAINING SET WHEN $\lambda = 0$ AND $\lambda = 1$

Value of λ	Accuracy on Training Set							
	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise	Total
0	0.842	0.854	0.08	0.273	0.923	0.928	0.444	0.69
1	0.833	0.802	0.007	0.014	0.901	0.707	0.164	0.606

TABLE V
CONFUSION MATRIX FOR THE EMOTIW VALIDATION SET WHEN $\lambda = 0$

	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise
Neutral	0.492	0.111	0.032	0.047	0.175	0.111	0.032
Anger	0.094	0.578	0.108	0.063	0.016	0.078	0.063
Disgust	0.125	0.175	0.075	0.15	0.05	0.175	0.25
Fear	0.109	0.109	0.152	0.217	0.131	0.109	0.173
Happy	0.175	0.016	0.016	0.016	0.667	0	0.11
Sad	0.066	0	0.115	0.082	0.033	0.623	0.081
Surprise	0.109	0.087	0	0.261	0.174	0.152	0.217

TABLE VI
TOTAL ACCURACIES OF THE BEST ARCHITECTURES OF THE 3 WINNING METHODS OF THE EMOTIW 2017 GRAND CHALLENGE REPORTED ON THE VALIDATION SET VS OUR OWN MODEL

Group	Architecture	Total Accuracy			
		Original	SSE Learning Strategy	After Fine-Tuning	Data augmentation
[12]	DenseNet-121	0.414	0.457		
	HoloNet	0.41	0.465	-	-
	ResNet-50	0.418	0.426		
[13]	VGG-Face	0.379		0.483	-
	FR-Net-A	0.337		0.446	-
	FR-Net-B	0.334	-	0.488	-
	FR-Net-C	0.376		0.452	-
	LSTM + FR-NET-B	-		0.465	0.504
[22]	Weighted C3D (no overlap)				0.421
	LSTM C3D (no overlap)				0.432
	VGG-Face				0.414
	VGG-LSTM 1 layer				0.486
Our	CNN-RNN with $\lambda = 0$	0.446	-	-	-

To illustrate the good performance of the proposed training approach, we computed t-Distributed Stochastic Neighbor Embedding (t-SNE) [11] on:

- the seven Z centroids defined in (2), each composed of 128 neuron output values, provided as desired outputs for training NetA; these are marked with dots,
- the seven centroids, computed by similarly clustering all internal representations X - defined in (4) and composed of 128 elements as well - extracted from NetA, during its training through minimization of the error in (9), with $\lambda = 0$; these are marked with stars,
- the seven respective centroids computed by clustering the internal representations of the network trained through minimization of the error in (5) or equivalently in (9) with $\lambda = 1$; these are marked with crosses.

Figs. 5-9 show the respective cluster centroids, visualized in a three-dimensional space, for some of the emotion categories. Table VII shows: i) the mean distance of each of the centroids marked with stars from the respective Z centroids marked with dots ($\lambda = 0$) and ii) the mean distance of each of the centroids marked with crosses from the respective Z centroids marked with dots ($\lambda = 1$).

It can be easily verified that in all cases NetA trained with the proposed procedure manages to bring the cluster centroids, corresponding to the EmotiW training data, closer to the respective cluster centroids of the EmotiW validation dataset. This leads to achieving higher accuracies in both datasets.

V. CONCLUSIONS AND FURTHER WORK

In this paper we have defined a new error criterion for training a deep neural network with different datasets in-

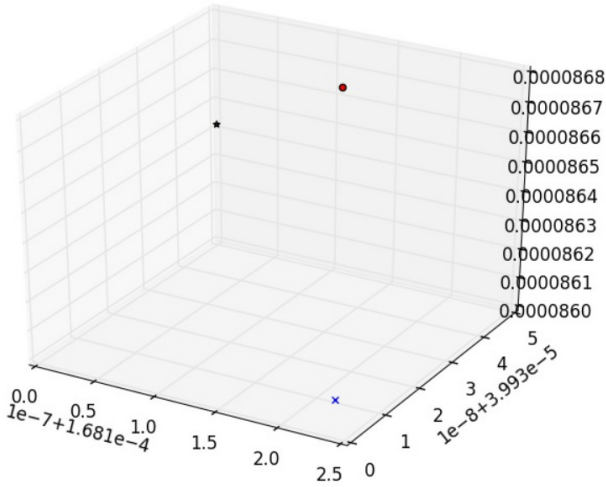


Fig. 5. Improving Neutral cluster centroid proximity through (9)

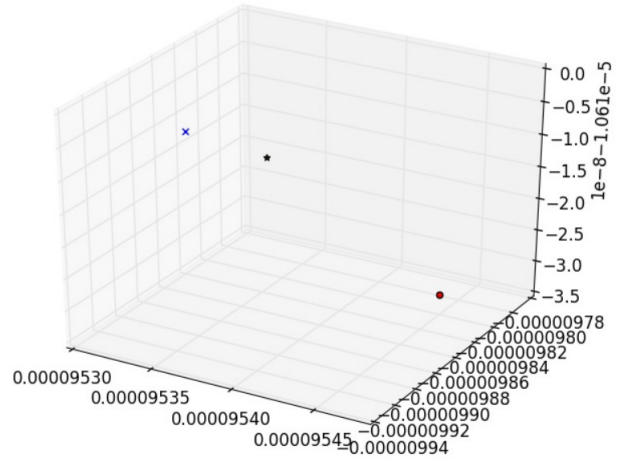


Fig. 7. Improving Disgust cluster centroid proximity through (9)

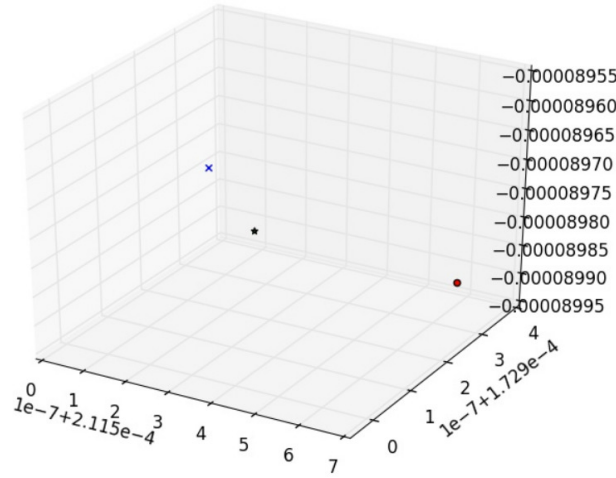


Fig. 6. Improving Anger cluster centroid proximity through (9)

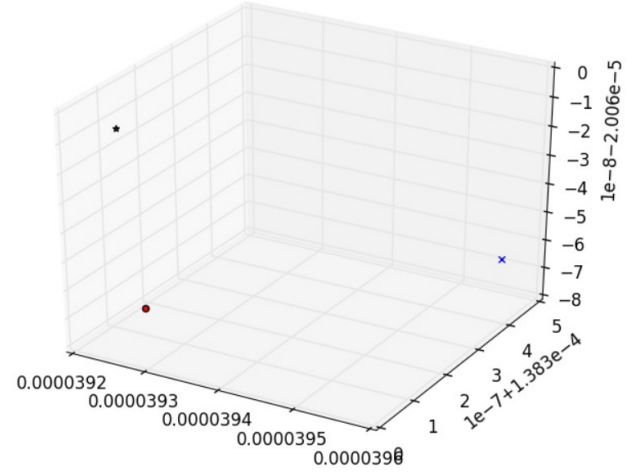


Fig. 8. Improving Sad cluster centroid proximity through (9)

the-wild, so as to achieve a high network performance in both datasets. We used the categorical emotion recognition paradigm, using datasets collected from movies and tv clips, as a testbed where the proposed approach can provide improved deep neural network performances.

In particular, we extracted internal representations at the final hidden layer of a deep CNN-RNN architecture and created seven clusters corresponding to the primary emotion categories, i.e. anger, disgust, fear, happiness, sadness, surprise and the neutral state. We then used the above cluster centroids as desired outputs for training a new deep CNN-RNN with another dataset for emotion recognition in-the-wild. We defined an appropriate loss function to be used for this domain adaptation procedure.

We were able to illustrate, through an experimental study, that the latter network was able to have an improved perfor-

mance on its training set by about 14%, as well as on the data set of the former network, by about 12%.

Compared to transfer learning or retraining of a DNN with the new dataset, the proposed approach provides the following advantages: i) it reduces catastrophic forgetting, since the desired outputs in DNN training also include the cluster centroids of the former dataset and ii) much less resources are necessary, since our method does not require availability of the former DNN architecture/weights, nor of any large amount of former data.

In our future work, the proposed procedure will be extended, to obtain a block component form, in which the proposed error criterion is iteratively minimized by each of the above networks. Future work also includes extending the proposed approach to deal with dimensional emotion recognition for estimation of the Valence and Arousal values. To do so, we

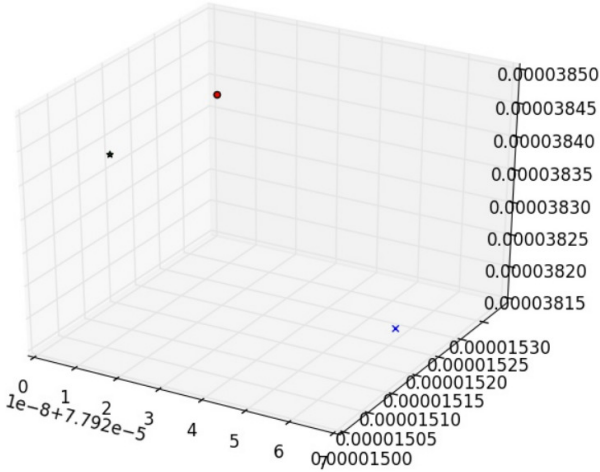


Fig. 9. Improving Happy cluster centroid proximity through (9)

TABLE VII

DISTANCES BETWEEN TRAINING AND VALIDATION CLUSTER CENTERS

Category	Mean Distance from respective validation cluster center	
	$\lambda = 0$	$\lambda = 1$
Neutral	0.015	0.067
Anger	0.015	0.061
Disgust	0.052	0.058
Fear	0.055	0.065
Happy	0.012	0.077
Sad	0.025	0.065
Surprise	0.055	0.067

will discretize the 2D Valence Arousal Space, e.g. in 400 (20×20) classes and then apply the proposed approach between different in-the-wild datasets, such as AffWild, RECOLA [20] and OMG-Emotion Behavior Dataset [1].

VI. ACKNOWLEDGMENT

The work of Stefanos Zafeiriou has been partially funded by the FiDiPro program of Tekes (project number: 1849/31/2015). The work of Dimitris Kollias was funded by a Teaching Fellowship of Imperial College London. We also wish to thank Dr Abhinav Dhall for providing us with the datasets of all recent EmotiW Challenges [5]- [8].

REFERENCES

- [1] Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., Wermter, S.: The omg-emotion behavior dataset. arXiv preprint arXiv:1803.05434 (2018)
- [2] Bengio, Y.: Learning deep architectures for ai. *Foundations and trends® in Machine Learning* **2**(1), 1–127 (2009)
- [3] Byeon, W., Breuel, T.M., Raue, F., Liwicki, M.: Scene labeling with 1stm recurrent neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3547–3555 (2015)
- [4] Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
- [5] Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., Gedeon, T.: From individual to group-level emotion recognition: EmotiW 5.0. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 524–528. ACM (2017)

- [6] Dhall, A., Goecke, R., Joshi, J., Hoey, J., Gedeon, T.: EmotiW 2016: Video and group-level emotion recognition challenges. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 427–432. ACM (2016)
- [7] Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T.: Emotion recognition in the wild challenge 2013. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 509–516. ACM (2013)
- [8] Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: EmotiW 2015. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 423–426. ACM (2015)
- [9] Ekman, P., Friesen, W.V., Ellsworth, P.: *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier (2013)
- [10] Ekman, P., Rosenberg, E.L.: *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA (1997)
- [11] Hinton, G.E., Roweis, S.T.: Stochastic neighbor embedding. In: *Advances in neural information processing systems*, pp. 857–864 (2003)
- [12] Hu, P., Cai, D., Wang, S., Yao, A., Chen, Y.: Learning supervised scoring ensemble for emotion recognition in the wild. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 553–560. ACM (2017)
- [13] Knyazev, B., Shvetsov, R., Efremova, N., Kuharenko, A.: Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. arXiv preprint arXiv:1711.04598 (2017)
- [14] Kollias, D., Nicolaou, M., Kotsia, I., Zhao, G., Zafeiriou, S.: Recognition of affect in the wild using deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop* (2017)
- [15] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, pp. 1097–1105 (2012)
- [16] LeCun, Y., Kavukcuoglu, K., Fierberg, C.: Convolutional networks and applications in vision. In: *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pp. 253–256. IEEE (2010)
- [17] Lloyd, S.: Least squares quantization in pcm. *IEEE Transactions on information theory* **28**(2), 129–137 (1982)
- [18] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: *BMVC*, vol. 1, p. 6 (2015)
- [19] Plutchik, R.: *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division (1980)
- [20] Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pp. 1–8. IEEE (2013)
- [21] Russell, J.A.: Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology* **36**(10), 1152 (1978)
- [22] Vielzeuf, V., Pateux, S., Jurie, F.: Temporal multimodal fusion for video emotion classification in the wild. arXiv preprint arXiv:1709.07200 (2017)
- [23] Whissel, C.: *The dictionary of affect in language, emotion: Theory, research and experience: vol. 4, the measurement of emotions*, r. Plutchik and H. Kellerman, Eds., New York: Academic (1989)
- [24] Yao, A., Shao, J., Ma, N., Chen, Y.: Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 451–458. ACM (2015)
- [25] Zafeiriou, S., Kollias, D., Nicolaou, M., Papaioannou, A., Zhao, G., Kotsia, I.: Aff-wild: Valence and arousal ‘in-the-wild’ challenge. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop* (2017)