LEVERAGING VALENCE AND ACTIVATION INFORMATION VIA MULTI-TASK LEARNING FOR CATEGORICAL EMOTION RECOGNITION

Rui Xia, Yang Liu

Computer Science Department, The University of Texas at Dallas, Richardson, TX 75080
{rx,yangl}@hlt.utdallas.edu

ABSTRACT

Deep learning technologies have been successfully applied to acoustic emotion recognition lately. In this work, we propose to apply multi-task learning for acoustic emotion recognition based on the Deep Belief Network (DBN) framework. We treat the categorical emotion recognition task as the major task. For the secondary task, we leverage two continuous labels, valence and activation. Two strategies are employed to achieve multi-task learning. First, we map the continuous labels into three categorical labels: $low, medium, high$, and use classification for the secondary task. Second, we project the continuous labels into $[-1, 1]$ range, and use regression for the secondary task. The combination of the loss functions from the major and secondary tasks is used in the objective function in multi-task learning. After iterative optimization, the values from the last hidden layer are used as features in the backend SVM classifier for emotion classification. Our experimental results show significant improvement over the baseline results using DBN, suggesting the benefit of utilizing additional information in a multi-task learning setup.

Index Terms— Emotion Recognition, Multi-task learning, Deep Belief Network

1. INTRODUCTION

Emotion recognition has attracted a lot of interest recently. Researchers have considered different modalities for this problem, such as facial expression, speech, and text. Several shared tasks related to emotion recognition have been developped in the past few years, with different task definitions. For example, in [1, 2], the task was to recognize categorical emotions; in [3, 4], systems were required to use regression to predict continuous values based on dimensions such as valence and activation; and in [5], the goal was to recognize emotions given selected movie clips, which is quite difficult even for the state-of-the-art systems.

Acoustic emotion recognition has been extensively studied in the past decade. Systems have used dynamic features with Hidden Markov Model (HMM) [6], supra-segmental features with support vector machines (SVM), and combined features from multiple modalities [7], just to name a few. With great success on automatic speech recognition [8] and image recognition [9], deep neural network (DNN) technologies have also been investigated and applied to acoustic emotion recognition. In [10], a Deep Belief Network (DBN) is built based on dynamic audio features and a temporal pooling method is applied to generate fixed length hidden features, which are fed into a soft-max layer for fine-tuning. In [11], auto-encoders are applied for transfer learning to alleviate the difference between emotion databases. In [12], the authors modified the auto-encoders and used two hidden representations (one for neutral speech, one for emotional speech) to extract more robust low dimensional emotional features. In [13], DBN is trained to capture the hidden dependency across video and audio modality.

Multi-task learning has been widely applied to many speech and natural language processing related problems [14, 15]. The advantage of multi-task learning is to improve system generalization by learning shared representations between appropriate tasks [16]. Classifiers learned based on the primary task can be better trained with the help of other related tasks. In some speech related tasks, multi-task learning has shown impressive performance once suitable minor tasks are found related to the major one. In [17], for phone state recognition, the authors explored using different auxiliary tasks, phone label, phone and state context, and showed better recognition performance. [18] [19] used multi-task learning in facial based emotion recognition, and considered facial verification and facial landmark prediction as the potential auxiliary tasks. To our knowledge there is very little work using multi-task learning for speech based emotion recognition.

In this work, we propose to use multi-tasking learning in the deep belief network (DBN) framework for speech emotion recognition. Typically, two kinds of labeling approaches are used to represent human emotions, categorical and continuous labels. The categorical labels can be interpreted as direct and common human mood such as anger and happiness. The continuous labels are based on a psychology theory that decomposes categorical emotions into continuous dimensions such as valence and activation, among others. We propose to treat valence and activation recognition as the secondary task in multi-task learning. Two different strategies are used to integrate the secondary task into the traditional categorical emotion recognition system in this study. First, valence and activation labels are clustered into three level labels, and thus the learning task is a classification task. The DBN is trained to simultaneously optimize the classification performance of the major emotion classification task and this secondary task. In the second method, we linearly normalize the valence and activation labels into the range of $[-1, +1]$, and use a regression task for it. The DBN system learns to lower the regression error of the secondary task while minimizing the classification error of the major task. We evaluate our proposed multi-task learning method using the USC-IEMOCAP database [20] and show significant improvement over the standard static features based method and the DBN system.

2. DEEP BELIEF NETWORK

In this paper, we use the deep belief network (DBN) framework for emotion recognition, where DBN is applied to the original raw audio features to extract better feature representation, which is then fed into another classifier, e.g., support vector machines (SVM). We

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

briefly describe DBN in this section. The DBN is constructed by stacking more than one Restricted Boltzmann Machines (RBMs), which is one special case of undirected graphical models [21]. This framework learns to extract meaningful hidden hierarchical representation from the training data.

The pre-training stage typically is done in a greedy layer-wise manner. In this work, Gaussian-RBM is used as the layer component in order to model the real-valued acoustic inputs. Given input data v as visible nodes and hidden variable h as hidden nodes, the joint-probability of Gaussian-RBM is as follow:

$$p(v) = \frac{\sum_h e^{-E(v,h)}}{Z} \tag{1}$$

where Z is the partition function, and the energy function $E(v, h)$ is calculated as:

$$E(v,h) = \sum_{i \in vis} \frac{1}{2\sigma_i^2}(v_i - b_i)^2 - \sum_{j \in hid} c_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} \tag{2}$$

where w_{ij} is the connection matrix between the hidden and visible nodes, σ_i is the standard deviation of visible unit i, b and c are bias vectors for the visible and hidden nodes respectively. The learning process is to minimize the empirical negative log-likelihood of training data. However, the exact solution is computationally intractable since the sum in the partition function needs to run over an exponential number of joint combinations of the visible and hidden nodes. The approximate algorithm called Contrastive Divergence (CD) introduced in [22] can be applied to update parameters efficiently.

In this DBN framework, we use the Noisy Rectified Linear Unit (NReLU) as the non-linear activation function instead of using sigmoid. The NReLU is one version of ReLU with Gaussian noise introduced in [23]. Prior studies have shown that systems with (NReLU) have better performance on several tasks (e.g., [24]). We also add an upper-bound to avoid having hidden nodes with large values. The bounded NReLU function is as follows:

$$min(a, max(x + \mathcal{N}(0, sigmoid(x)))) \tag{3}$$

where a is as the upper-bound (in this work a is equal to 3) and x is the input of network. In this set-up, hidden nodes are sampled as $h = NReLU(x)$. The conditional probability distribution of v is:

$$p(v_i|h) = \mathcal{N}(v_i; \sum_j w_{ij} h_j + c_j, \sigma^2) \tag{4}$$

After pre-training, the parameters of DBN are used as the initial values and further tuned in the subsequent supervised fine-tuning stage. Here one soft-max layer is added on top of the last hidden layer for classification. Given the values of the last hidden layer h_l, the probability of the k^{th}

$$p(y_k|h_l) = \frac{e^{h_l^T W_k + B_k}}{\sum_j e^{h_l^T W_j + B_j}} \tag{5}$$

where W and B are the parameters of the soft-max layer. The stochastic gradient descent algorithm is applied to iteratively minimize the objective function over the training data.

3. PROPOSED METHOD

Multi-task learning can improve generalization of a model by learning from related tasks that share the same feature representation with the main task. In this work, we use categorical emotion classification as the major task, and investigate if it can benefit from multi-task learning. It is important to choose appropriate related tasks for effective multi-task learning. According to the psychology theory, human emotions can be mapped into dimensional continuous spaces, such as valence and activation (referred to arousal in some studies). Valence describes the pleasantness or unpleasantness, and activation is to measure the stimulus degree of human activities. Intuitively categorical emotion labels have close relation with valence and activation labels, therefore, we evaluate using valence and activation labels for the secondary task for multi-task learning.

Though valence and activation are defined in the continuous space, it is not easy for humans to assign continuous values in annotation. In practice, annotators are often asked to give a discrete value from a range, e.g., $[1, 5]$, for each unit. Using labels from multiple annotators for a unit, the average value can be calculated and used as the continuous label.

For multi-task learning, we adopt the DBN framework as described above, but now the learning objective is to optimize the performance for both the major task and the secondary task. Figure 1 illustrates our proposed method. The same hidden layers are shared for the major and the secondary tasks. The DBN is tuned to optimize the objective functions that take both of them into account. We evaluate two mechanisms to integrate valence and activation recognition as the secondary task into the DBN: treating them as hard categories and thus a classification task; or continuous values and thus a regression task. The following explains the two methods in more details. Note that in both methods, model parameters are pre-trained first, and only in the fine-tuning stage, the secondary task is utilized.

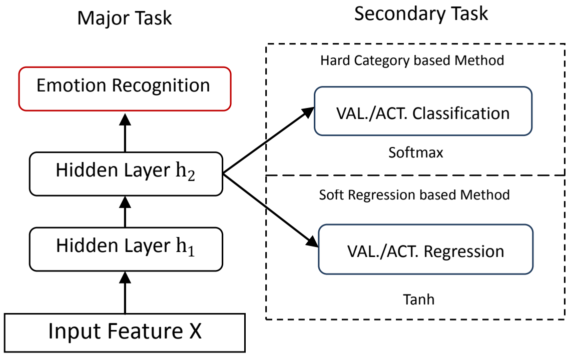

Fig. 1. Proposed multi-tasking learning framework.

3.1. Hard category based method

In this method, we cluster the continuous labels into categorical labels, and treat classification of the categorical levels of valence and activation as the secondary task. Three levels are used, $low, medium, high$, similar to prior work in [25]. Now each sentence i in the training corpus has the following feature and label set, $[x_i, (y_{emo,i}, y^c_{act,i}, y^c_{val,i})]$, where x_i is the feature set of sentence i, $y_{emo,i}$, $y^c_{act,i}$ and $y^c_{val,i}$ represent the associated categorical emotion, activation and valence label respectively.

As shown in Figure 1, we keep the same structure as the original DBN, but add another soft-max layer on top of the last hidden layer to predict the valence and activation categories. Our goal is to learn the model parameters to optimize for the major and the secondary tasks. Given N training sentences, the objective function of this system is as follows:

$$J_h = \sum_{i}^{N} -log(P(y_{emo,i}|h_{l,i})) - \alpha * (\sum_{i}^{N} log(P(y^c_{act,i}|h_{l,i}) + \sum_{i}^{N} log(P(y^c_{val,i}|h_{l,i})) \tag{6}$$

where $h_{l,i}$ represents the last hidden layer in the DBN when enrolling features of sentence i, the posterior probability P is calculated using Eq 5, and hyper-parameter α is the weight of the secondary task. Using this objective function, the model is trained to minimize the negative posterior probability of the major task (categorical emotion recognition), and the secondary task (categorical valence and activation recognition).

3.2. Soft regression based method

In this method, we use continuous values for valence and activation, rather than mapping them to categorical levels. First the continuous dimension labels are linearly regressed to the range of $[-1, 1]$. When the original labels used in annotation are from 1 to 5,[1] we use the the following function, the same as level-of-interest sub-challenge task in [26]:

$$y^r_{val|act} = y^o_{val|act}/2.5 - 1 \tag{7}$$

where $y^r_{val|act}$ and $y^o_{val|act}$ represent the regressed and the original continuous labels for valence or activation.

For the continuous labels, the soft-max layer is not applicable, therefore we use a different objective function. Given the values of the last hidden layer $h_{l,i} \in R^m$, the predicted value $y^p_{val|act,i}$ for sentence i is calculated using the following formula:

$$y^p_{val|act,i} = \tanh(W_{r,val|act} h_{l,i} + b_{r,val|act}) \tag{8}$$

where $W_{r,val|act} \in R^m$ and $b_{r,val|act}$ are the weights and bias, for valence or activation respectively. Note that $\tanh(z) = (e^z - e^{-z})/(e^z + e^{-z})$ is applied as a non-linear activation function here, since its output ranges from -1 to 1. Using the predicted value $y^p_{val|act}$ for N training sentences, the mean squared error between $y^p_{val|act}$ and $y^r_{val|act}$ is calculated as the loss function:

$$L_{con,val|act} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2} * (y^p_{val|act,i} - y^r_{val|act,i})^2 + \gamma * ||W_{r,val|act}||_2^2, \tag{9}$$

where $||W_{r,val|act}||_2^2 = \sum_{j=1}^{m} W[j]^2_{r,val|act}$ is L2 regularization with γ as hyper-parameter (in this paper, γ is equal to 0.001). Then this loss function is used in the objective function for the secondary task. The objective function for the whole system becomes:

$$J_s = -\sum_{i}^{N} log(P(y_{emo,i}|h_{l,i})) + \beta * (L_{con,val} + L_{con,act}) \tag{10}$$

where β is the hyper-parameter to control the influence of the loss of the regression function for the continuous labels. This objective function aims to minimize the negative posterior probability of the emotion label and the mean squared error between the predicted and the reference valence and activation values.

4. EXPERIMENTS

4.1. Data

We use the Interactive Emotional Dyadic Motion Capture (USC-IEMOCAP) database [20] in this study. This corpus has approximately 12 hours of audiovisual data, including video, speech, motion capture of face, and text transcriptions [20]. It has 10 professional actors (5 male and 5 female) acting in two different scenarios: scripted play and spontaneous dialog, in their dyadic interaction. Each interaction is around 5 minutes in length, and is segmented into sentences. These sentences are labeled by at least 3 annotators. We use four emotion categories in this study: angry, happy, sad, and neutral. Note that we merged 'happy' and 'excited' in the original annotation into one class: happy. Only the utterances with the majority agreement are used. In total, there are 5531 utterances selected from the entire data set in this experiment. The class distribution is: 20% angry, 29.6% sad, 19.6% happy, and 30.8% neutral. In this corpus, self-assessment manikins are used as the tool to annotate the valence and activation. The annotation scale is from 1 to 5. Three annotators, including the participant him/herself and two other annotators, were asked to label continuous dimensions.

4.2. Features

We use the static features extracted with openSMILE [27] as the input signal in the DBN framework. There are 1,584 features in total, as used in the INTERSPEECH 2010 Paralinguistic Challenge [26]. Since the feature values have very different ranges, we standardized all the features before using them as the input to DBN. Details of the features can be found in [26]. Table 1 summarizes these features.

Table 1. *Acoustic feature sets: 38 low-level descriptors (LLD) and 21 functionals.*

Descriptors	Functionals
PCM loudness	Position max./min.
MFCC [0-14]	arith. mean, std. deviation
log Mel Freq. Band [0-7]	skewness, kurtosis
LSP Frequency [0-7]	lin. regression coeff. 1/2
F0	lin. regression error Q/A
F0 Envelope	quartile 1/2/3
Voicing Prob.	quartile range 2-1/3-2/3-1
Jitter local	percentile 1/99
Jitter consec. frame pairs	percentile range 99-1
Shimmer local	up-level time 75/90

4.3. Experimental setup

We use the leave-one-speaker-out cross validation setup. For the hard category method, we cluster continuous labels (in the range of $[1, 5]$) to three groups: $low, medium, high$, corresponding to the range of $[1, 2]$, $(2, 3.5]$, and $(3.5, 5]$ respectively. For the continuous regression method, we use Eq 7 to map the labels to $[-1, 1]$.

The DBN is constructed by stacking two Gaussian-RBMs. In the pre-training stage, we first initialize parameters of Gaussian-RBM with small numbers. The learning rate is set as 0.001 and the number of training epochs is 10. During training, Contrastive Divergence with one-step (CD-1) is used for sampling. In pre-training, we used the mini-batch mode with 64 instances in each mini-batch. In the fine-tuning stage, we use 8 iterations with 0.02 as the learning rate. We utilize Theano to implement our system. The detail of Theano

[1] For different annotation schemes, similar scaling methods can be used to map the range to $[-1, 1]$.

can be found in [28]. After pre-training and fine-tuning, the training and testing instances are enrolled into the DBN. Then the values of the last hidden layer are extracted as new features. We use SVMs with linear kernels as the classifier for emotion recognition.

4.4. Results

Table 2 shows the emotion recognition results based on the unweighted average accuracy (UA). This metric has been widely used in many emotion related challenges. It is the average accuracy across all the emotion classes. We compare our method to several systems. First is the baseline approach that uses the original static feature set. The second one uses the output of the last hidden layer from DBN as features. This is a baseline DBN method, without the multi-task learning strategy. Since our method uses additional information from valence and activation annotation, for a fair comparison, we design two systems that also leverage valence and activation information. We build a classifier based on the static features to predict valence and activation labels, using the three level class. For each utterance, the posterior probabilities of activation and valence from the classifier are incorporated as additional features. Note that for the training instances, we used a cross-validation setup to generate the predicted valence and activation labels. Since we have two different feature sets, the original static features and features generated by the DBN framework, two new feature sets can be formed by concatenating the posterior probabilities of valence and activation with them. These correspond to $Staticfeatures + Pred._{act,val}$ and $DBNfeatures + Pred._{act,val}$ in the table. The last two rows show the emotion recognition results based on our proposed multi-task learning using two strategies, hard category and soft regression based respectively.

Table 2. *Emotion classification results (in %).*

System		UA
Static features		59.7
DBN framework		60.5
$Staticfeatures + Pred._{act,val}$		60.7
$DBNfeatures + Pred._{act,val}$		61.1
Multi-task learning	Hard category	**62.2**
	Soft regression	**62.5**

From Table 2, we can see there is a performance gain using the features derived from the DBN over the original static features, showing the benefit of better representation via DBN. Using our proposed methods (last two rows) achieves the highest accuracy, which demonstrates the effectiveness of considering the valence and activation labels during training with the multi-task learning strategy. The improvement is statistically significant based on one tailed z-test ($p < 0.05$). Between using the hard categories vs. continuous values, the latter is slightly better, suggesting that there is some loss of information when using the hard categories for the valence and activation values. Regarding the systems that use the predicted valence and activation information as additional features, we can see there is a performance improvement comparing to the corresponding original systems without using the extra features. However, it does not bring as much gain as our proposed method. This is in part because the performance of activation and valence prediction is not great, but more importantly, this shows that multi-task learning is an effective way to use the valence and activation annotation.

We also investigate the impact on system performance by varying parameters α and β, which control the strength of the cost function for the secondary task. Figure 2 shows the curve of UA by changing the weights from 0.0 to 1.2. When the weight is 0, the results are the same as the basic DBN framework without using multi-task learning. As the weight increases, system performance first improves, and then starts to drop when the weight of the secondary task is too large. The best performance is achieved when α is 0.4 and β is 1.0 (corresponding to the results shown in Table 2).

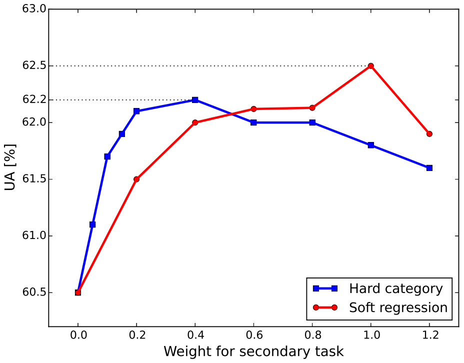

Fig. 2. Tuning hyper-parameters α and $beta$ for emotion recognition

The results above show that emotion recognition performance can benefit from the multi-task learning framework by considering both activation and valence labels. We conduct further experiments by treating activation and valence separately as the secondary task, in order to evaluate their individual contributions. Similar DBN framework and objective functions are used; the only difference is that we only use either activation or valence in this experiment. Table 3 shows the results. It shows that enrolling activation or valence alone can also improve the system performance compared to the baseline (first two rows in Table 2). However, the results are worse than considering both, which is expected. Table 3 also shows that using activation performs slightly better than valence. Finally we performed system combination where we concatenated the feature sets learned by using either valence or activation in the secondary task, and found the results are not as good as our proposed method. This shows that learning model parameters in the multi-task learning framework has advantages in that the system can jointly optimize multiple tasks.

Table 3. *Emotion classification results (in %).*

System		UA
Hard category	activation	62.0
	valence	61.9
Soft regression	activation	62.1
	valence	61.7

5. CONCLUSION AND FUTURE WORK

In this paper, we propose to apply multi-task leaning on acoustic emotion recognition based on Deep Belief Network. The secondary task is based on valence and activation recognition. We evaluated two different ways of representing the secondary task, a classification or a regression task. Our experiments show significant improvement using multi-task learning over the original single task learning framework. In the future, we plan to investigate whether we can further improve the system by modeling the relationship of valence and activation.

6. REFERENCES

[1] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge.," in *Proceedings of INTERSPEECH*, 2009.

[2] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *Proceedings of ACII*. 2011.

[3] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al., "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH*, 2013.

[4] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of ICMI*, 2012.

[5] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon, "Emotion recognition in the wild challenge 2013," in *Proceedings of ICMI*, 2013.

[6] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Hidden Markov Model-based Speech Emotion Recognition," in *Proceedings of ICME*, 2003.

[7] Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, Aravind Namandi Vembu, and Rohit Prasad, "Emotion recognition using acoustic and lexical features.," in *Proceedings of INTERSPEECH*, 2012.

[8] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, 2012.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[10] Samira Ebrahimi Kanou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al., "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of ICMI*, 2013.

[11] Jun Deng, Rui Xia, Zixing Zhang, Yang Liu, and Björn Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Proceedings of ICASSP*, 2014.

[12] Rui Xia, Jun Deng, Björn Schuller, and Yang Liu, "Modeling gender information for emotion recognition using denoising autoencoders," in *Proceedings of ICASSP*, 2014.

[13] Yelin Kim, Honglak Lee, and Emily Mower Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proceedings of ICASSP*, 2013.

[14] Xiao Li, Ye-Yi Wang, and Gökhan Tür, "Multi-task learning for spoken language understanding with shared slots.," in *Proceedings of INTERSPEECH*, 2011.

[15] Ronan Collobert and Jason Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of ICML*, 2008.

[16] R Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *Proceedings of Machine Learning*, 1997, pp. 41 – 75.

[17] Michael L Seltzer and Jasha Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proceedings of ICASSP*, 2013.

[18] Terrance Devries, Kumar Biswaranjan, and Graham W Taylor, "Multi-task learning of facial landmarks and expression," in *Computer and Robot Vision (CRV), 2014 Canadian Conference on*, 2014.

[19] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas, "Learning active facial patches for expression analysis," in *Proceedings of CVPR*, 2012.

[20] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP:interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[21] Yoshua Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[22] Geoffrey E Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[23] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of ICML*, 2010.

[24] George E Dahl, Tara N Sainath, and Geoffrey E Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proceedings of ICASSP*, 2013.

[25] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Björn Schuller, and Shrikanth Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 184–198, 2012.

[26] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proceedings of INTERSPEECH*, 2010.

[27] Florian Eyben, Martin Wöllmer, and Björn Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 18th ACM International Conference on Multimedia, MM 2010*, 2010.

[28] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010, Oral Presentation.