Facial Affect Estimation in the Wild Using Deep Residual and Convolutional Networks

Behzad Hasani and Mohammad H. Mahoor
Department of Electrical and Computer Engineering
University of Denver, Denver, CO
behzad.hasani@du.edu and mmahoor@du.edu

Abstract

Automated affective computing in the wild is a challenging task in the field of computer vision. This paper presents three neural network-based methods proposed for the task of facial affect estimation submitted to the First Affect-in-the-Wild challenge. These methods are based on Inception-ResNet modules redesigned specifically for the task of facial affect estimation. These methods are: Shallow Inception-ResNet, Deep Inception-ResNet, and Inception-ResNet with LSTMs. These networks extract facial features in different scales and simultaneously estimate both the valence and arousal in each frame. Root Mean Square Error (RMSE) rates of 0.4 and 0.3 are achieved for the valence and arousal respectively with corresponding Concordance Correlation Coefficient (CCC) rates of 0.04 and 0.29 using Deep Inception-ResNet method.

1. Introduction

Affect is a psychological term for describing the external exhibition of internal emotions and feelings. Affective computing attempts to develop systems that can interpret and estimate human affects through different channels (*e.g.* visual, auditory, biological signals, etc.) [32]. Facial expressions are one of the primary non-verbal communication methods for expressing emotions and intentions.

There have been numerous studies for developing reliable automated Facial Expression Recognition (FER) systems in the past. However, current available systems are still far from desirable emotion perception capabilities required for developing robust and reliable Human Machine Interaction (HMI) systems. This is predominantly because of the fact that these HMI systems are needed to be in an uncontrolled environment (aka wild setting) where there are significant variations in the lighting, background, view, subjects' head pose, gender, and ethnicity [22].

Three models of Categorical, Dimensional, and FACS are proposed in the literature to quantify affective facial behaviors. In categorical model, emotion is chosen from a list of affective-related categories such as six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) defined by Ekman *et al*. [10]. In Dimensional model, a value is assigned to an emotion over a continuous emotional scale, such as "valence" and "arousal" defined in [25]. In Facial Action Coding System (FACS) model, all possible facial component actions are described in terms of Action Units (AUs) [11]. FACS model only describes facial movements and does not interpret the affective state directly.

The dimensional modeling of affect can distinguish between subtle differences in exhibiting of affect and encode small changes in the intensity of each emotion on a continuous scale, such as *valence* and *arousal* where valence shows how positive or negative an emotion is, and arousal indicates how much an event is intriguing/agitating or calming/soothing [25]. The First Affect-in-the-Wild challenge, focuses on estimation of valence and arousal (dimentional model of affect) in the wild.

In this paper, we propose different methods submitted to the First Affect-in-the-Wild challenge for estimating dimensional model values of affect (valence and arousal) in the wild using deep convolutional networks and Long Short-Term Memory (LSTM) units. We also report the results of our methods on the Aff-Wild database provided in the First Affect-in-the-Wild challenge.

The remainder of the paper is organized as follows: Section 2 provides an overview of the related work in this field. Section 3 explains the methods submitted to the challenge. Experimental results and their analysis are presented in Section 4 and finally the paper is concluded in Section 5.

2. Related work

Conventional algorithms for affective computing from faces use different engineered features such as Local Binary Patterns (LBP) [28], Histogram of Oriented Gradients (HOG) [6, 13], Histogram of Optical Flow (HOF) [7], and facial landmarks [4, 5]. These features often lack required generalizability in cases where there is high variation in lighting, views, resolution, subjects' ethnicity, etc. Also, most of these works are applied on the categorical model of affect which can be considered an easier task than the dimensional model estimation task.

Few number of studies have been conducted on the dimensional model of affect in the literature. Nicolaou *et al.* [23] trained bidirectional LSTM on multiple engineered features extracted from audio, facial geometry, and shoulders. They achieved Root Mean Square Error (RMSE) of 0.15 and Correlation Coefficient (CC) of 0.79 for valence as well as RMSE of 0.21 and CC of 0.64 for arousal.

He *et al.* [16] won the AVEC 2015 challenge by training multiple stacks of bidirectional LSTMs (DBLSTM-RNN) on engineered features extracted from audio (LLDs features), video (LPQ-TOP features), 52 ECG features, and 22 EDA features. They achieved RMSE of 0.104 and CC of 0.616 for valence as well as RMSE of 0.121 and CC of 0.753 for arousal.

Koelstra *et al.* [19] trained Gaussian naive Bayes classifiers on EEG, physiological signals, and multimedia features by binary classification of low/high categories for arousal, valence, and liking on their proposed database DEAP. They achieved F1-score of 0.39, 0.37, and 0.40 on arousal, valence, and Liking categories respectively.

In recent years, Convolutional Neural Networks (CNNs) have become the most popular approach among researchers in the field of computer vision. Szegedy *et al.* [30] introduced GoogLeNet which contains multiple "Inception" layers that apply several convolutions on the feature map in different scales. Several variations of Inception have been proposed [18, 31]. Also, Inception layer is combined with residual unit introduced by He *et al.* [15] resulting considerable acceleration in the training of Inception networks [29].

Recurrent Neural Networks (RNNs) can learn temporal dynamics by mapping input sequences to a sequence of hidden states [9]. One of the problems of RNNs is that it is difficult for them to learn long-term sequences. This is mainly due to the vanishing or exploding gradients problem [17]. LSTMs [17] contain a memory unit which solves this problem by memorizing the context information for long periods of time. LSTM modules have three gates: 1) the input gate (i) 2) the forget gate (f) and 3) the output gate (o) which overwrite, keep, or retrieve the memory cell c respectively at the timestep t. Letting σ be the sigmoid function, ϕ be the hyperbolic tangent function, and $\circ$ denoting Hadamard product, the LSTM updates for the timestep t given inputs x_t, h_{t-1}, and c_{t-1} are as follows:

$$\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
g_t &= \phi(W_C \cdot [h_{t-1}, x_t] + b_C) \\
C_t &= f_t \circ C_{t-1} + i_t \circ g_t \\
h_t &= o_t \circ \phi(C_t)
\end{aligned} \tag{1}$$

Several works have used LSTMs and their extensions for different tasks. Fan *et al.* [12] won the EmotiW 2016 challenge by cascading 2D-CNN with LSTMs and combining the resulting feature map with 3D-CNNs for facial expression recognition task. Donahue *et al.* [9] proposed Long-term Recurrent Convolutional Network (LRCN) by combining CNNs and LSTMs. Byeon *et al.* [3] proposed an LSTM-based network by applying 2D-LSTMs in four direction sliding windows. As mentioned earlier, Nicolaou *et al.* [23] used bidirectional LSTMs and He *et al.* [16] used multiple stacks of bidirectional LSTMs (DBLSTM-RNN) for the dimensional model of affect.

In order to evaluate our methods, we calculate and report Root Mean Square Error (RMSE), Correlation Coefficient (CC), Concordance Correlation Coefficient (CCC), and Sign Agreement Metric (SAGR) metrics for our methods. In the following, we briefly review the definitions of these metrics.

Root Mean Square Error (RMSE) is the most common evaluation metric in a continuous domain which is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2} \tag{2}$$

where $\hat{\theta}_i$ and θ_i are the prediction and the ground-truth of i^{th} sample, and n is the number of samples. RMSE-based evaluation metrics can heavily weigh the outliers [2], and they do not consider covariance of the data.

Pearson's Correlation Coefficient (CC) overcomes this problem [23, 26, 27] and it is defined as:

$$CC = \frac{COV\{\hat{\theta}, \theta\}}{\sigma_{\hat{\theta}}\sigma_\theta} = \frac{E[(\hat{\theta} - \mu_{\hat{\theta}})(\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}\sigma_\theta} \tag{3}$$

Concordance Correlation Coefficient (CCC) is another metric [24, 33] which combines CC with the square difference between the means of two compared time series:

$$\rho_c = \frac{2\rho\sigma_{\hat{\theta}}\sigma_\theta}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2 + (\mu_{\hat{\theta}} - \mu_\theta)^2} \tag{4}$$

where COV is covariance function, ρ is the Pearson correlation coefficient (CC) between two time-series (e.g., prediction and ground-truth), $\sigma_{\hat{\theta}}^2$ and σ_θ^2 are the variance of each time series, $\sigma_{\hat{\theta}}$ and σ_{θ} are the standard deviation of each, and $\mu_{\hat{\theta}}$ and μ_{θ} are the mean value of each. Unlike CC, the predictions that are well correlated with the groundtruth but shifted in value are penalized in proportion to the deviation in CCC.

The value of valence and arousal fall within the interval of [-1,+1] and correctly predicting their signs are essential in many emotion-prediction applications. Therefore, we use Sign Agreement Metric (SAGR) which is proposed in [23] to evaluate the performance of a valence and arousal prediction system with respect to the sign agreement. SAGR is defined as:

$$SAGR = \frac{1}{n}\sum_{i=1}^{n}\delta(sign(\hat{\theta}_i), sign(\theta_i)) \tag{5}$$

where δ is the Kronecker delta function, defined as:

$$\delta(a,b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} \tag{6}$$

3. Proposed methods

Inception and ResNet have shown remarkable results in various tasks [14, 21, 30, 34]. For the Affect-in-the-Wild challenge, we proposed Inception-ResNet based architectures followed by LSTM units (submission 3) for the task of affect estimation. Our proposed methods extract contextual information of the frames in an end-to-end deep neural network. In the following, we explain each of the methods presented in our submissions.

3.1. Shallow Inception-ResNet (submission 1)

For the first submission, we propose modified version of Inception-ResNet which originally presented in [29]. Our first module is shallower than the original Inception-ResNet containing only "stem" and single "Inception-ResNet" module. Most of the settings are the same as the ones presented in [29] while the input size of the network is changed from 299 × 299 to 49 × 49. Because of this reduction in the size of the input, we are not able to have a very deep network. Therefore, only one Inception-ResNet module is used in this method.

Figure 1 shows the structure of our shallow Inception-ResNet method. The input images with the size $49 \times 49 \times 3$ are followed by the "stem" layer. Afterwards, the stem is followed by Inception-ResNet-A, dropout, and a fully-connected layer respectively. In Figure 1, detailed specification of each layer is provided. All convolution layers are followed by a batch normalization layer and all batch normalization layers (except the ones that are indicated as "Linear" in Figure 1) are followed by a ReLU [20] activation function to avoid the vanishing gradient problem.

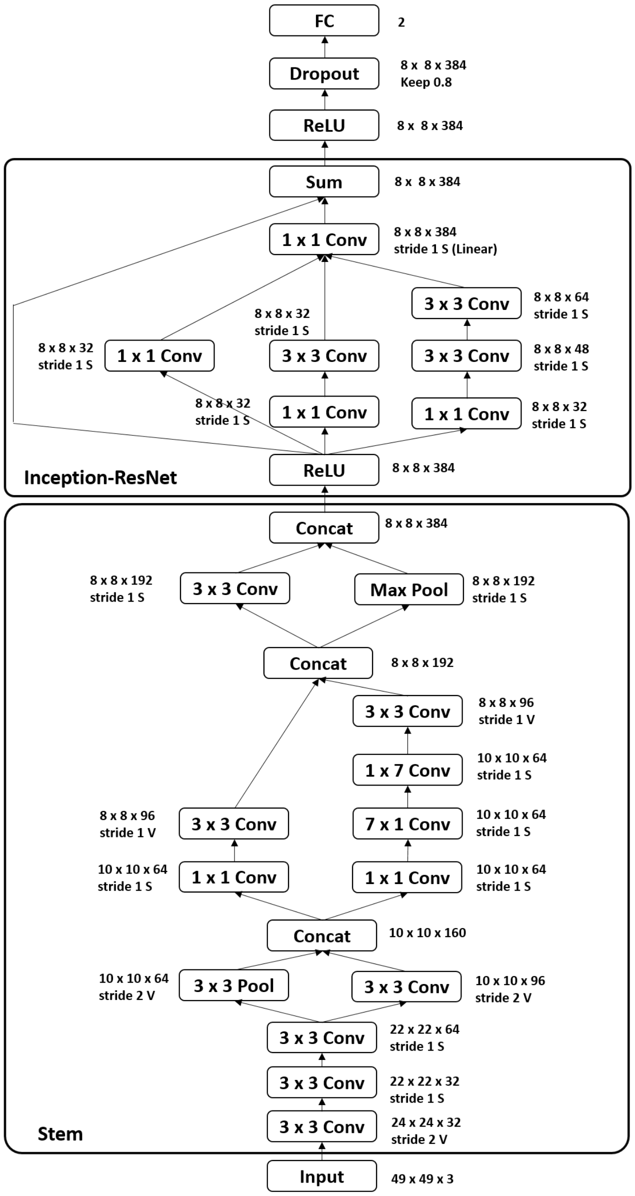


Figure 1. Network architecture for submission 1. The "V" and "S" marked layers represent "Valid" and "Same" paddings respectively. The size of the output tensor is provided next to each layer.

3.2. Deep Inception-ResNet (submission 2)

As mentioned earlier, in order to have a deeper network to extract more abstract features and having more number of parameters to learn, we changed the properties of the previously mentioned network.

Figure 2 shows the structure of our Deep Inception-ResNet method. Comparing tho the previous method, we change all of the "valid" paddings to "same" paddings to save the feature map size. Also, strides are changed in this method. Same as before, the input images with the size

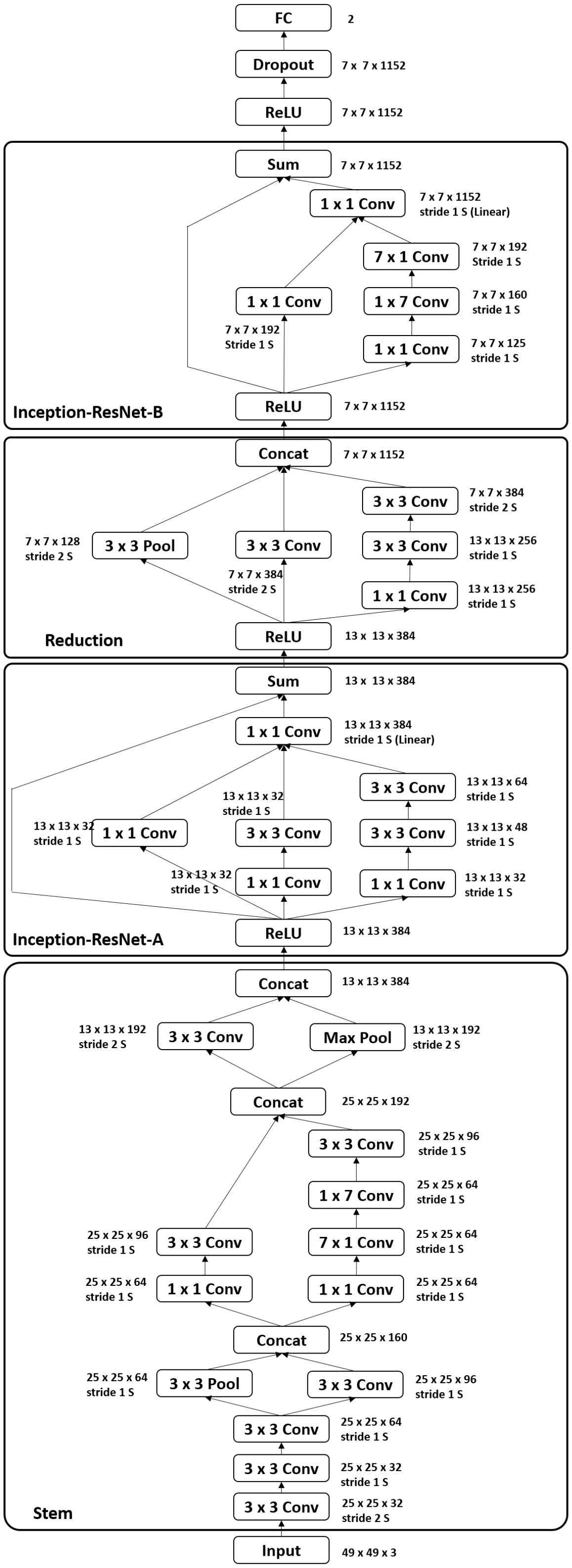


Figure 2. Network architecture for submission 2. The "S" marked layers represent "Same" padding. The size of the output tensor is provided next to each layer.

$49 \times 49 \times 3$ are followed by the "stem" and "Inception-ResNet-A" layers. Afterwards, to deepen the network, we include "Reduction" (which reduces the grid size from 13×13 to 7×7) followed by "Inception-ResNet-B", dropout, and fully-connected layers.

Same as before, all convolution layers are followed by a batch normalization layer and all batch normalization layers (except the ones that are indicated as "Linear" in the Figure) are followed by a ReLU activation function to avoid the vanishing gradient problem.

3.3. Inception-ResNet & LSTM (submission 3)

As explained earlier, LSTMs have shown remarkable results in different emotion recognition/estimation tasks [9, 12, 16, 23]. Therefore, we incorporate LSTMs in our next method in two directions to estimate the valence and arousal intensity in the challenge.

Figure 3 shows the network used for the third submission. The network has the same settings as the second submission. The only difference here is that after the dropout layer, we vectorize the feature map on two dimensions (one on the width of the feature map and the other one on its height). This is inspired by the work in [3] where LSTMs are used in four different directions. Adding LSTMs will take the complex spatial dependencies of adjacent pixels into account [3]. We investigated that 200 hidden units for each LSTM unit is a reasonable amount for this task. At the end, the resulting feature vectors of these two LSTMs are concatenated together and are followed by a fully-connected layer (Figure 3).

All of the proposed methods are implemented using a combination of TensorFlow [1] and TFlearn [8] toolboxes on NVIDIA Tesla K40 GPUs. We used asynchronous stochastic gradient descent with weight decay of 0.0001, and learning rate of 0.01. Mean square error used for loss function.

4. Database & results

In this section, we briefly review Aff-Wild database provided in the First Affect-in-the-Wild challenge. We then report the results of our experiments on both validation and test sets using the metrics provided in section 2.

4.1. Aff-Wild database

Aff-Wild database contains 300 videos of different subjects watching videos of various TV shows and movies. The videos contain subjects from different genders and ethnicities with high variations in head pose and lightning. Provided videos in this database are annotated with valence and arousal values for each frame. 254 videos of this database are selected for training and the rest 46 videos are used for evaluating the participants in the challenge.

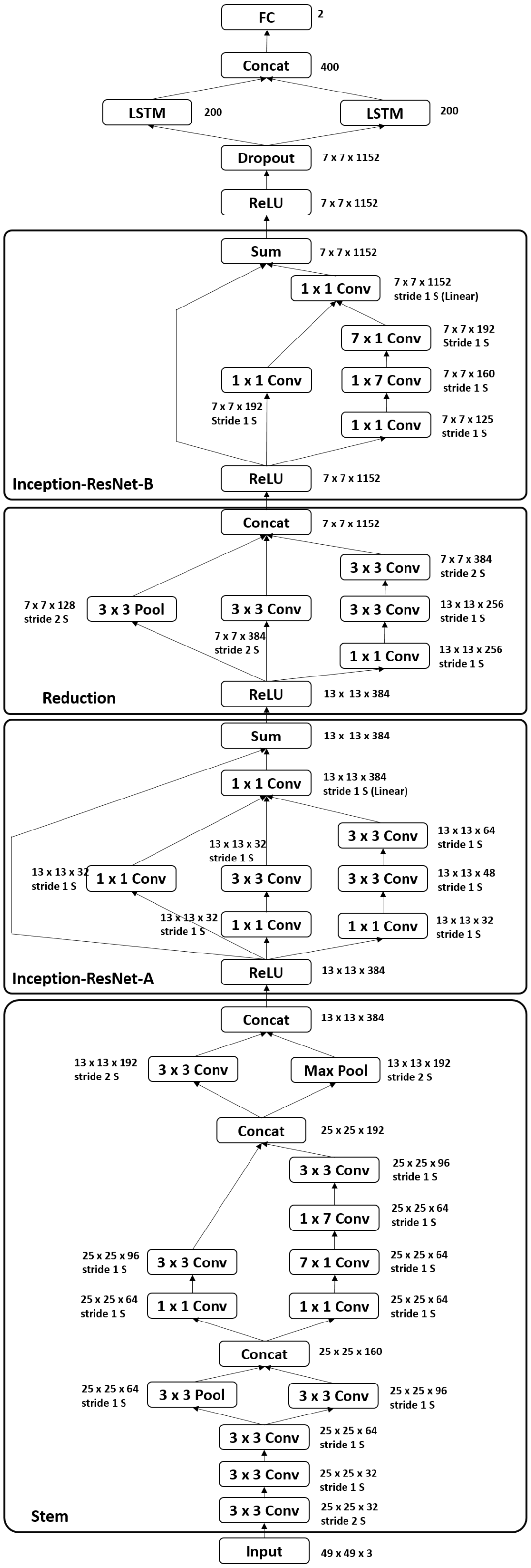


Figure 3. Network architecture for submission 3. The "S" marked layers represent "Same" padding. The size of the output tensor is provided next to each layer.

4.2. Results

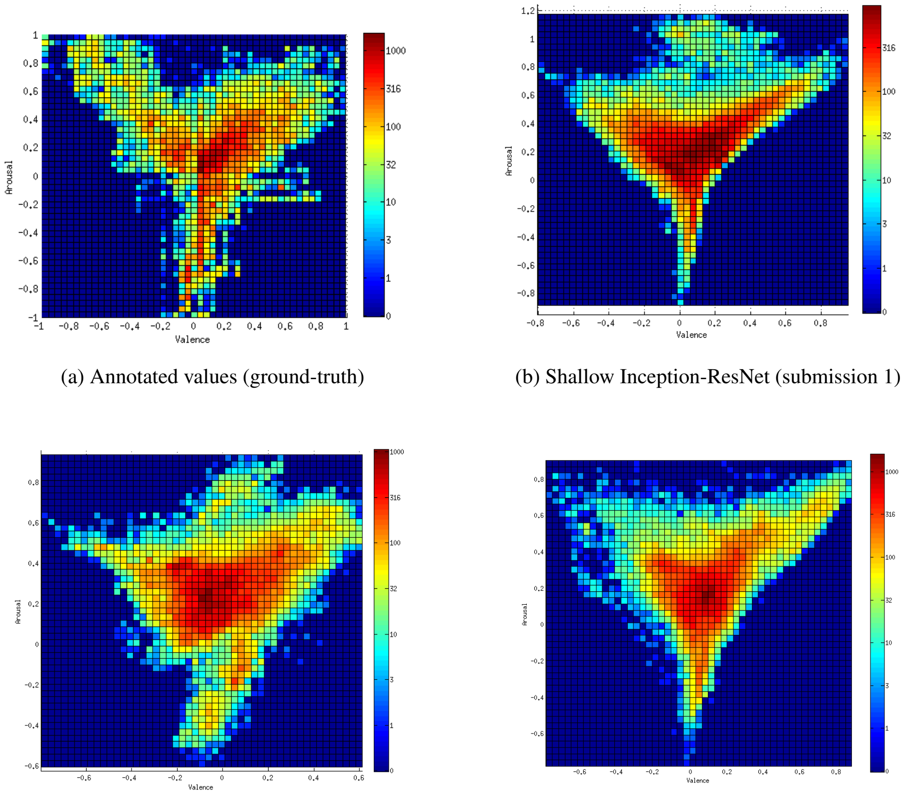
We first extract the faces from the frames using the bounding boxes that are released alongside the database. Afterwards, we resize the faces to 49×49 pixels and divide the training data into training and validation sets by assigning 10 percent of the subjects to the validation set and the rest of them to the training set.

Figure 4 shows the histograms of annotated values and predicted values on the validation set for all submissions. It can be seen that the database is heavily imbalanced. The number of annotated frames with values close to zero for valence and arousal (center of the circumplex) is considerably higher than other regions. Therefore, our methods are also biased toward this region. Figures 4b and 4c show that submissions 1 and 2 were not able to correctly estimate instances with high arousal and low valence values. Figure 4d shows that submission 3 performs better in this region but as table 1 indicates, this submission generally does not perform well on estimating arousal values comparing to others. However, all of the methods show mostly similar patterns on the validation set and no unusual predicted values can be seen throughout any of the methods.

We evaluate our proposed methods with RMSE, CC, CCC, and SAGR metrics defined in the section 2. The reported results are the calculated values on the validation set. Table 1 shows different metrics calculated for the validation and test sets of the provided database. The CC and SAGR metric are not reported to the authors for the test set, therefore these metrics are not reported in Table 1.

Results on validation set show more accurate estimation for arousal in all submissions. This can be in part due to the less dynamic range of values for arousal in the training data. By looking at the results on our validation set, it can be seen that almost all of the metrics show the superiority of submission 2 comparing to other submissions. CC and CCC metrics show that there is more correlation between the results of submission 2 and the ground-truth comparing to other methods.

The test set results in Table 1 also show the superiority of submission 2. In all three methods, the estimation for valence is considerably less accurate comparing to the estimated values for arousal while the results in the validation set do not show such drastic difference. Nevertheless, submissions 1 and 3 show almost the same results on the test set while submission 2 shows less error in terms of RMSE and also shows more correlation with ground-truth in terms of CCC metric. The reduction of correlation in submission 3 can be in part due to the fact that the input feature map of LSTM units does not contain the notion of time which shapes an unfitting input for the LSTMs. Using 3D convolutional neural networks would provide such temporal information within the feature map but this temporal processing of input sequences is not experimented in this work.



(a) Annotated values (ground-truth)

(b) Shallow Inception-ResNet (submission 1)

(c) Deep Inception-ResNet (submission 2)

(d) Inception-ResNet with LSTMs (submission 3)

Figure 4. Histograms of valence and arousal values in the validation set for annotated values (a), submission 1 (b), submission 2 (c), and submission 3 (d). Best viewed in color.

submission	validation set								test set			
	RMSE		CC		CCC		SAGR		RMSE		CCC	
	valence	arousal	valence	arousal	valence	arousal	valence	arousal	valence	arousal	valence	arousal
#1	0.29	0.37	0.29	0.22	0.28	0.19	0.53	0.72	0.41	0.33	0.03	0.19
#2	0.27	0.36	0.44	0.26	0.36	0.19	0.57	0.74	0.40	0.30	0.04	0.29
#3	0.28	0.36	0.33	0.17	0.31	0.14	0.55	0.70	0.40	0.33	0.04	0.17

Table 1. Results of submissions on validation and test sets

5. Conclusion

In this paper, we presented three methods submitted to the First Affect-in-the-Wild Challenge: Shallow Inception-ResNet, Deep Inception-ResNet, and Inception-ResNet with LSTMs. These Inception-ResNet based methods are engineered specifically for the task of facial affect estimation by extracting facial features in different scales and they estimate both valence and arousal values for each frame simultaneously. We used four metrics to evaluate our methods on our validation set: RMSE, CC, CCC, and SAGR. On the test set, Inception-ResNet with LSTMs network achieved the best performance with noticeably good estimation in terms of RMSE and CCC rates especially on arousal values.

6. Acknowledgement

This work is partially supported by the NSF grants IIS-1111568 and CNS-1427872. We gratefully acknowledge the support from NVIDIA Corporation with the donation of the Tesla K40 GPUs used for this research.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al.

Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 4

[2] S. Bermejo and J. Cabestany. Oriented principal component analysis for large margin classifiers. *Neural Networks*, 14(10):1447–1461, 2001. 2

[3] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with lstm recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3547–3555, 2015. 2, 4

[4] T. F. Cootes, G. J. Edwards, C. J. Taylor, et al. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001. 2

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 2

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 1

[7] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, pages 428–441. Springer, 2006. 1

[8] A. Damien et al. Tflearn. https://github.com/tflearn/tflearn, 2016. 4

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2, 4

[10] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 1

[11] P. Ekman and W. V. Friesen. Facial action coding system. 1977. 1

[12] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 445–450, New York, NY, USA, 2016. ACM. 2, 4

[13] B. Hasani, M. M. Arzani, M. Fathy, and K. Raahemifar. Facial expression recognition with discriminatory graphical models. In *2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–7, Dec 2016. 1

[14] B. Hasani and M. H. Mahoor. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. *arXiv preprint arXiv:1703.06995*, 2017. 3

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[16] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80. ACM, 2015. 2, 4

[17] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 2

[19] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. 2

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[21] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016. 3

[22] A. Mollahosseini, B. Hasani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor. Facial expression recognition from world wild web. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016. 1

[23] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011. 2, 3, 4

[24] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1335–1336. ACM, 2015. 2

[25] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. 1

[26] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011. 2

[27] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012. 2

[28] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. 1

[29] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016. 2, 3

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015. 2, 3

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[32] J. Tao and T. Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005. 1

[33] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016-depression, mood, and emotion recognition workshop and challenge. *arXiv preprint arXiv:1605.01600*, 2016. 2

[34] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen. Holonet: Towards robust emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 472–478, New York, NY, USA, 2016. ACM. 3