Engagement Detection with Multi-Task Training in E-Learning Environments

Onur Copur[1], Mert Nakip[2], Simone Scardapane[1], and Jürgen Slowack[3]

[1] Sapienza University of Rome, Roma, 00185, Italy
onurcopur12@gmail.com, simone.scardapane@uniroma1.it
[2] Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Gliwice, 44–100, Poland
mnakip@iitis.pl
[3] Barco NV, Kortrijk, 8500, Belgium
jurgen.slowack@barco.com

Abstract. Recognition of user interaction, in particular engagement detection, became highly crucial for online working and learning environments, especially during the COVID-19 outbreak. Such recognition and detection systems significantly improve the user experience and efficiency by providing valuable feedback. In this paper, we propose a novel Engagement Detection with Multi-Task Training (ED-MTT) system which minimizes mean squared error and triplet loss together to determine the engagement level of students in an e-learning environment. The performance of this system is evaluated and compared against the state-of-the-art on a publicly available dataset as well as videos collected from real-life scenarios. The results show that ED-MTT achieves 6% lower MSE than the best state-of-the-art performance with highly acceptable training time and lightweight feature extraction.

Keywords: Engagement detection, activity recognition, e-learning, triplet loss, multi-task training

1 Introduction

During the COVID-19 outbreak, nearly all of the learning activities, as other meeting activities, transferred to online environments [32]. Online learners participate in various educational activities including reading, writing, watching video tutorials, online exams, and online meetings. During the participation in these educational activities, participants show various engagement levels, e.g. boredom, confusion, and frustration [11]. To provide feedback to both instructors and students, online educators need to detect their online learners' engagement status precisely and efficiently. For example, the teacher can adapt and make lessons more interesting by increasing interaction, such as asking questions to involve non-interacting students. Since, in e-learning environments, students are not speaking most of the time, the engagement detection systems should extract valuable information from only visual input [29]. This makes the problem

non-trivial and subjective because annotators can perceive different engagement levels from the same input video. The reliability of the dataset labels is a big concern in this setting but often is ignored by the current methods [29,30,32]. Because of this, deep learning models overfit to the uncertain samples and perform poorly on validation and test sets.

In this work, we propose a system called Engagement Detection with Multi-Task Training (ED-MTT) to detect the engagement level of the participants in an e-learning environment. The proposed system first extracts features with *OpenFace* [2], then aggregates frames in a window for calculating feature statistics as additional features. Finally, it uses Bidirectional Long Short-Term memory (Bi-LSTM) [13] unit for generating vector embeddings from input sequences. In this system, we introduce a triplet loss as an auxiliary task and design the system as a multi-task training framework by taking inspiration from [22], where self-supervised contrastive learning of multi-view facial expressions was introduced. The reason for the triplet loss usage is based on the ability to utilize more elements for training via the combination of original samples. In this way, it avoids overfitting and makes the feature representation more discriminative. [9]. To the best of our knowledge, this is a novel approach in the context of engagement detection. The key novelty of this work is the multi-task training framework using triplet loss together with Mean Squared Error (MSE). The main contributions of this paper are as follows:

- Multi-task training with triplet and MSE losses introduces an additional regularization and reduces possibly over-fitting due to very small sample size.
- Using triplet loss mitigates the label reliability problem since it measures relative similarity between samples.
- A system with lightweight feature extraction is efficient and highly suitable for real-life applications.

Furthermore, we evaluate the performance of ED-MTT on a publicly available "Engagement in The Wild" dataset [7], which is comprised of separated training and validation sets. In our experimental work, we first analyze the importance of feature sets to select the best set of features for the resulting trained ED-MTT system. Then, we compare the performance of ED-MTT with 9 different works [1, 5, 15, 20, 24, 25, 27, 31, 32] from the state-of-the-art which will be reviewed in the next section. Our results show that ED-MTT outperforms these state-of-the-art methods with at least 6% improvement on MSE.

The rest of this paper is organized as follows: Section 2 reviews the related works in the literature. Section 3 explains the architectural design of ED-MTT. Section 4 presents experimental results for the performance evaluation of ED-MTT and comparison with the state-of-the-art methods. Section 5 conclude our work and experimental results.

2 Related Works

One of the first attempts to investigate the relationships between facial features, conversational cues, and emotional expressions with engagement detection is presented by D'Mello et al. in [8]. The authors in [10, 28] used the Facial Action Coding System (FACS) which is a measure of discrete emotions with facial muscle movements, and point out the relation between specific engagement labels and facial actions. In Reference [28], Whitehill et al. showed that automated engagement detectors perform with comparable accuracy to humans. In [3], Booth et al. compared the performance of a Long-Short Term Memory (LSTM) based method with SVM and KNN methods with non-verbal features. In [6], Dewan et al. proposed a Local Directional Pattern (LDP) to extract person-independent edge features which are fed to a Deep Belief Network. Huang et al. [14] proposed a model, called Deep Engagement Recognition Network (DERN), which combines temporal convolution, bidirectional LSTM, and an attention mechanism to identify the degree of engagement based on the features captured by OpenFace [2]. Moreover in [18], Liao et al. proposed Deep Facial Spatiotemporal Network (DFSTN) which is is developed based on extracting facial spatial features and global attention for sequence modeling with LSTM. Finally, in [19, 21, 23], authors used models which are based on Convolutional Neural Networks (CNN) and Residual Networks (ResNet) [12]. All the works above considered the engagement detection problem as a multi-class classification problem. In contrast, in this paper, we follow a more recent line of research that considers engagement detection as a regression problem, where MSE loss is used to measure a continuous distance between predicted and ground truth engagement levels.

Yang et al. [31] also used MSE loss and developed a method that ensembles four separate LSTMs using facial features extracted from four different sources. In [20], Niu et al. combined the outputs of three Gated Recurrent Units (GRU) based on a 117-dimensional feature vector composed of eye gaze action units and head pose features. In [24], Thomas et al. used Temporal Convolutional Network (TCN) on the same set of features as in [20]. In previous works [29, 32], the most common ways to overcome over-fitting is data augmentation and cross-validation training. Some other works [1, 27] consider imbalanced sampling [17] and using weighted/ranked loss functions. Moreover, some works also consider spatial dropout and batch normalization as a regularization technique [5, 24]. All the previous studies focus on small sample sizes and imbalanced labels but none of them consider the reliability of the labels. On the other hand, in this paper, ED-MTT aims to handle both overfitting and label reliability at the same time via multi-task training with triplet loss.

3 Architectural Design for Engagement Detection with Multi-Task Training

We now present our architectural design as well as the multi-tasking with the combination of MSE and triplet loss for training, which are the main contributions of this work. To this end, Fig. 1 displays the training architecture of

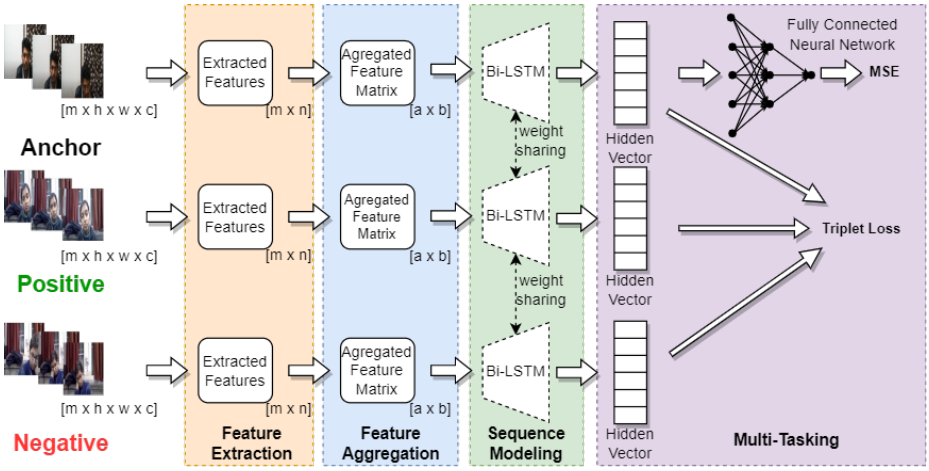

Fig. 1. The training architecture for ED-MTT.

ED-MTT that consists of four main parts: Feature Extraction, Frame Aggregation, Sequence Modeling, and Multi-Tasking. The inputs of this architecture are three batches of samples as Anchor, Positive and Negative. In each batch, each sample is the sequence of images which is obtained by segmenting a video into m frames each of size $h \times w \times c$, where h denotes the height in pixels, w denotes the width in pixels and c denotes the number of color channels of each frame, where RGB color space is used. During the training with this approach, each sample s in the anchor batch is assumed to have a labeled engagement level E^s between 0 and 1. For each s, E^s is assigned into either low engagement or high engagement classes. To this end, if $E^s < 0.5$, s is assigned into the low engagement class; otherwise, i.e. $E^s \geq 0.5$, s is assigned into the high engagement class. Then, for each sample s in the anchor batch, the positive batch contains a random sample from the same engagement class of s while the negative batch contains a random sample from the opposite engagement class of s.

Furthermore, the outputs of the architecture in Fig. 1 are the MSE and Triplet Loss which are combined to train the Bi-LSTM model. Note that during inference, the engagement level prediction is the output of the fully connected neural network. While creating a multi-task learning problem through triplet loss, which aims to prevent overfitting due to the very few samples available for engagement detection during e-learning, we are able to perform regression for continuous engagement levels using MSE. In the rest of this section, we explain each part of the training architecture.

3.1 Feature Extraction

In order to narrow down the feature space by extracting the important features from the sequence of video frames, we first determine the features that are related to the engagement level of a subject. Accordingly, as done in [14, 20, 24, 29, 31],

we consider 29 features which are related to eye gaze, head pose, head rotation, and facial action units. We extract these features with OpenFace which provides many different facial features [2] and can be described as

$$\mathbf{Y}^s_{m\times n} = \text{OpenFace}(\mathbf{X}^s_{mhwc}), \tag{1}$$

where $\mathbf{X}^s_{mhwc}$ is the tensor of frame sequences at sample s, and $\mathbf{Y}^s_{m\times n}$ is the matrix of sequence of features at sample s, where the (i, j)-th element of $\mathbf{Y}^s_{m\times n}$ the feature i for frame j.

In the result of feature extraction, the eye gaze-related features are, *gaze_0_x, gaze_0_y, gaze_0_z* which are eye gaze direction vectors in world coordinates for the left eye and *gaze_1_x, gaze_1_y, gaze_1_z* for the right eye in the image. The head pose-related features are *pose_Tx, pose_Ty, pose_Tz* representing the location of the head with respect to the camera in millimeters (positive Z is away from the camera). *pose_Rx, pose_Ry, pose_Rz* indicates the rotation of the head in radians around x,y,z axes. This can be seen as pitch (Rx), yaw (Ry), and roll (Rz). The rotation is in world coordinates with the camera being the origin. Finally, the following 17 facial action unit intensities varying in the range $0 - 5$ are used: *AU01_r, AU02_r, AU04_r, AU05_r, AU06_r, AU07_r, AU09_r, AU10_r, AU12_r, AU14_r, AU15_r, AU17_r, AU20_r, AU23_r, AU25_r, AU26_r, AU45_r*.

3.2 Feature Aggregation over Time Windows

We now explain the aggregation of feature statistics over time windows with multiple video frames. In this way, the number of features (which was equal to n at the end of the Feature Extraction phase) is increased to b in order to provide more information to the Sequence Model.

Let the operation of the "Feature Aggregation over Time Windows" be shown as

$$\mathbf{Z}^s_{a\times b} = \text{Aggregate}(\mathbf{Y}^s_{m\times n}), \tag{2}$$

where $\mathbf{Z}^s_{a\times b}$ is the matrix of the b feature statistics for a aggregated frames. Let z be the number of frames in each time window that are considered for feature aggregation, where $m = a \times z$. Then, in each of a windows, we compute the *mean*, *variance*, *standard deviation*, *minimum*, and *maximum* of each feature over the consecutive z frames resulting in b feature statistics, where $b = 5 \times n$.

3.3 Sequence Modeling Combined with Multi-Tasking

Multi-task learning aims to learn multiple different tasks simultaneously while maximizing performance on one or all of the tasks [4]. The suggested architecture contains two tasks: The first task is predicting the multi-level engagement label by optimizing the MSE loss between actual and predicted labels. The second task is learning hidden vector embeddings by optimizing the triplet loss.

As shown in Fig. 1, during sequence modeling, we use three parallel (siamese) Bi-LSTM models with weight sharing to compute the hidden vectors for Triplet

Loss and for MSE loss as cascaded to the Fully Connected Neural Network. However, note that training is performed for only one Bi-LSTM model since the Bi-LSTM models in Fig. 1 are used with weight sharing for triplet loss. We call the Bi-LSTM model for the aggregated feature matrix $\mathbf{Z}^s_{a\times b}$ as

$$T^s_v = \text{Bi-LSTM}(\mathbf{Z}^s_{a\times b}), \tag{3}$$

where T^s_v is the hidden vector, which is the hidden state of the last layer of Bi-LSTM model. Thus, the length of this vector, denoted by v, is equal to twice the number of hidden units of the last layer of the Bi-LSTM.

Triplet loss is a loss function where a baseline (anchor) sample is compared with a positive and negative sample. The distance between the anchor and the positive sample is minimized and the distance between the anchor and the negative is maximized. We use the triplet loss function which is presented in [26] and defined as

$$\begin{aligned} \ell(\text{Anchor}, \text{Positive}, \text{Negative}) = L = \{l_1, \dots l_s \dots, l_S\}^\top, \\ l_s = \max\{d(\text{Anchor}_s, \text{Positive}_s) - d(\text{Anchor}_s, \text{Negative}_s) + \text{margin}, 0\}, \end{aligned} \tag{4}$$

where S is the number of samples in a batch, d is the euclidean distance, and $margin$ is a non-negative margin representing the minimum difference between the positive and negative distances that are required for the loss to be 0. Moreover, Anchor_s, Positive_s and Negative_s denote the Anchor, Positive and Negative batches for sample s, respectively.

In addition to the triplet loss, we also minimize the MSE loss which measures the error for the engagement regression. To this end, we cascade the Bi-LSTM model to the Fully Connected Neural Network whose output is the engagement level. Recall that the engagement regression is the main task during the real-time application. Accordingly, during training, the minimization of MSE can be considered as the main task while the minimization of Triplet loss is the auxiliary task.

4 Experimental Results

4.1 Dataset

For the performance evaluation of the proposed technique, we use both training and validation datasets published at "Emotion Recognition in the Wild" (EmotiW 2020) challenge [7] where the engagement regression is a sub-task. The dataset is comprised of 78 subjects (25 females and 53 males) whose ages range from 19 to 27. Each subject is recorded while watching an approximately 5 minutes long stimulus video of a Korean Language lecture. This procedure results in a collection of 195 videos, where the environment varies over videos and the subjects are not disturbed during recording. The engagement level of each video recording is labeled by a team of five between 0 and 3 resulting in the distribution shown in Fig. 2.

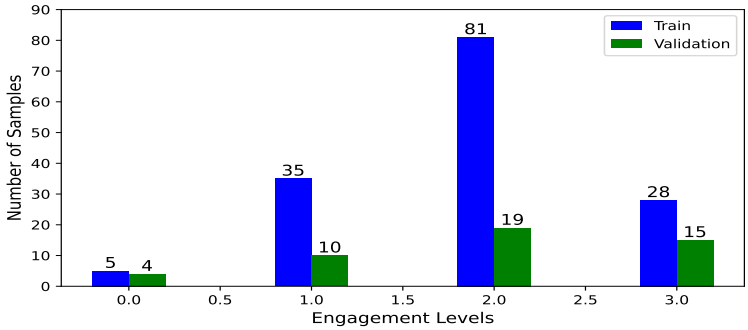

Fig. 2. The distribution of the engagement classes for each of the training and validation sets. In this figure, we see that the dataset is highly imbalanced, in particular there is a lack of low level engagement class samples.

4.2 Experimental Setup and Hyperparameter Settings

We implemented ED-MTT by using PyTorch on Python 3.7.12. The experiments are executed on the Google Colab platform where the operating system is Linux-5.4.144, and the GPU device is Tesla P100-PCIE-16GB. The model is trained via the adam optimizer [16] for 500 epochs with 5×10^{-5} initial learning rate and batch size of 16.

Furthermore, during our experiments, we first fixed the number of aggregated frames $a = 100$. At the input of Bi-LSTM, we used a batch normalization with an imbalanced sampler from the "imbalanced-learn" library of Python [17]. Then in order to determine the architectural hyperparameters of the sequential model, we performed a random search for the number of Bi-LSTM layers, the size of the hidden state as well as the number of neurons at each of two fully connected neural network layers. The random search sets are as follows: $\{1, 2, 3\}$ for the number of Bi-LSTM layers, $\{128, 256, 512, 1024\}$ for the size of hidden state of each Bi-LSTM layer[4], $\{256, 128, 64\}$ for the first layer of the fully connected neural network, and $\{32, 16, 8\}$ for the second layer of fully connected neural network. At the end of this search, the resulting architecture is comprised of 2 Bi-LSTM layers each of whose hidden state size is 1024, and two sequential fully connected layers with 64 and 32 neurons respectively.

4.3 Performance Evaluation

We now evaluate the performance of ED-MTT for engagement detection on a publicly available "Engagement in The Wild" dataset. During performance evaluation, we first aim to select the subset of facial and head position features with respect to their effects on the performance of our system. To this end, Table 1 displays the performance of the model under different combinations of feature sets, where the combinations are selected empirically to achieve high

[4] Note that the size of the hidden state is constant across all Bi-LSTM layers.

Table 1. Performance of The Model Under Different Combinations of Feature Sets

Eye Gaze	Head Pose	Head Rotation	Action Units	MSE
✓	✗	✗	✗	0.08347
✗	✓	✗	✗	0.07784
✗	✗	✓	✗	0.05723
✗	✗	✗	✓	0.05044
✗	✗	✓	✓	0.06578
✗	✓	✓	✗	0.07238
✓	✗	✓	✗	0.06915
✓	✓	✗	✗	0.06036
✓	✓	✓	✗	0.06973
✓	✗	✓	✓	0.05681
✓	✓	✗	✓	**0.04271**
✓	✓	✓	✓	0.05431

performance. Recall that the number of features in each feature set is as follows: 6 features in *Eye Gaze*, 3 features in *Head Pose*, 3 features in *Head Rotation*, and 17 features in *Action Units*. According to our observations on the results presented in this table, we may draw the following conclusions:

- The best performance is achieved by using all features except *Head Rotation* features. Accordingly, in the rest of our results, we use the combination of *Eye Gaze*, *Head Pose*, and *Action Unit* features.
- The most effective individual feature set is *Action Units*.
- The MSE loss significantly decreases when *Action Unit* features are included in the selected features.

Furthermore, in Fig. 3, we present the comparison of ED-MTT against the state-of-the-art engagement regression methods that are evaluated on the Engagement in The Wild dataset. In this figure, the MSE scores of the state of the art methods are taken from their original papers. The results show that ED-MTT achieves the best performance with 0.0427 MSE loss on the validation set. Although the performances of all methods are highly competitive with each other, the ED-MTT improved the best performance (Chang et. al. [5]) in the literature by 6%. In addition, the training time of ED-MTT is around 38 minutes for 149 samples for 500 epochs.

Fig. 4 displays the box plot of the predicted engagement levels on the validation sets which are classified with respect to the ground truth engagement labels in the dataset. In this figure, from median and percentiles of predicted engagement levels, one may see that the continuous predictions of ED-MTT distinctly reflects the four level of engagement classes in the ground truth labels. Moreover, ED-MTT can easily distinguish between classes 0, 0.33, and 0.66 while the difference between 0.66 and 1.0 is more subtle.

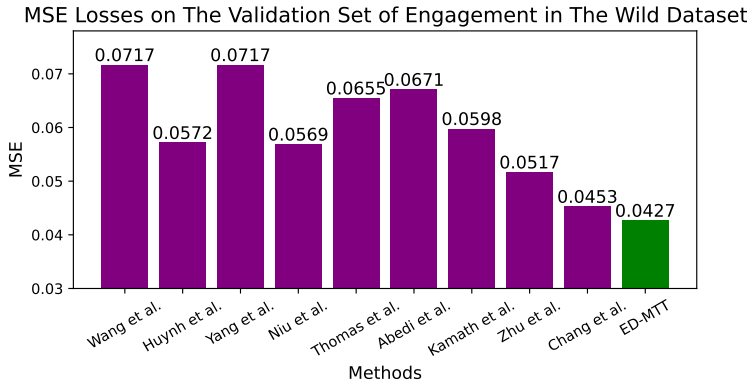

Fig. 3. The performance comparison of ED-MTT against the state-of-the-art methods, where the MSE scores are presented as in the original papers. *Wang et al.* [27] score: 0.0717, *Huynh et al.* [25] score: 0.0572, *Yang et al.* [31] score: 0.0717, *Niu et al.* [20] score: 0.0569, *Thomas et al.* [24] score: 0.0655, *Abedi et al.* [1] score: 0.0671, *Kamath et al.* [15] score: 0.0598, *Zhu et al.* [32] score: 0.0517, *Chang et al.* [5] score: 0.0453, ***ED-MTT* score: 0.0427**

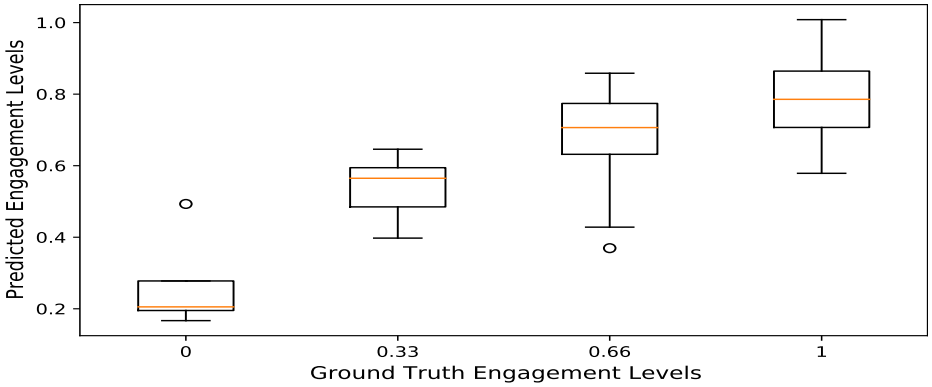

Fig. 4. The box plot of predicted engagement levels for each class in the ground truth engagement levels

4.4 Qualitative Results

Finally, ED-MTT is also tested on a preliminary real-life engagement detection tasks for which the prediction results are presented in Fig. 5. These results show that the proposed model, ED-MTT, trained on Engagement in The Wild dataset is able to provide highly successful predictions in real-life use-cases, which are totally different than the cases in the training set. According to our observations on the prediction results for a total (approximately) 12 minutes long videos including 8 people, the model can successfully distinguish different levels of engagement (very low, low, high, and very high engagement levels). However, the predicted engagement levels lie between 0.2 and 0.92, which forces to determine smaller quantization intervals to classify engagement levels in real-life use-cases.

Fig. 5. Sample images with the following predicted engagement levels by ED-MTT: 0.35 (top left), 0.53 (top middle), 0.86 (top right), 0.47 (bottom left), 0.61 (bottom middle), and 0.82 (bottom right).

5 Conclusion

Online working and learning environments are currently more essential in our lives, especially after the COVID-19 era. In order to improve the user experience and efficiency, advanced tools, such as recognition of user interaction, became highly crucial in these digital environments. For e-learning, one of the most important tools might be the engagement detection system since it provides valuable feedback to the instructors and/or students.

In this paper, we developed a novel engagement detection system called "ED-MTT" based on multi-task training with triplet and MSE losses. For engagement regression task, ED-MTT uses the combination of *Eye Gaze*, *Head Pose*, and *Action Units* feature sets and is trained to minimize MSE and triplet loss together. This training approach is able to improve the regression performance due to the following reasons; 1) multi-task training with two losses introduces an additional regularization and reduces over-fitting due to very small sample size, 2) triplet loss measures relative similarity between samples to mitigate the label reliability problem. 3) minimization of MSE ensures that the main loss considered for the regression problem is minimized alongside the triplet loss.

The performance of ED-MTT is evaluated and compared against the performances of the state-of-the-art methods on the publicly available Engagement in The Wild dataset which is comprised of separated training and validation sets. Our results showed that the novel ED-MTT method achieves 6% lower MSE than the lowest MSE achieved by the state-of-the-art while the training of ED-MTT takes around 38 minutes for 149 samples for 500 epochs. We tested the performance of ED-MTT for real-life use cases with 8 different participants, and the prediction results for majority of these cases were shown to be highly successful.

References

1. Abedi, A., Khan, S.: Affect-driven engagement measurement from videos. arXiv preprint arXiv:2106.10882 (2021)
2. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018). pp. 59–66 (2018)
3. Booth, B.M., Ali, A.M., Narayanan, S.S., Bennett, I., Farag, A.A.: Toward active and unobtrusive engagement assessment of distance learners. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 470–476 (2017)
4. Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)
5. Chang, C., Zhang, C., Chen, L., Liu, Y.: An ensemble model using face and body tracking for engagement detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 616–622 (2018)
6. Dewan, M.A.A., Lin, F., Wen, D., Murshed, M., Uddin, Z.: A deep learning approach to detecting engagement of online learners. In: 2018 IEEE SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI. pp. 1895–1902 (2018)
7. Dhall, A., Sharma, G., Goecke, R., Gedeon, T.: Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 784–789 (2020)
8. D'Mello, S.K., Craig, S.D., Graesser, A.C.: Multimethod assessment of affective experience and expression during deep learning. International Journal of Learning Technology **4**(3-4), 165–187 (2009)
9. Dong, X., Shen, J.: Triplet loss in siamese network for object tracking. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
10. Grafsgaard, J., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.: Automatically recognizing facial expression: Predicting engagement and frustration. In: Educational Data Mining 2013 (2013)
11. Gupta, A., D'Cunha, A., Awasthi, K., Balasubramanian, V.: Daisee: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885 (2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. Huang, T., Mei, Y., Zhang, H., Liu, S., Yang, H.: Fine-grained engagement recognition in online learning environment. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). pp. 338–341 (2019)
15. Kamath, S., Singhal, P., Jeevan, G., Annappa, B.: Engagement analysis of students in online learning environments. In: Misra, R., Shyamasundar, R.K., Chaturvedi, A., Omer, R. (eds.) Machine Learning and Big Data Analytics (Proceedings of International Conference on Machine Learning and Big Data Analytics (ICMLBDA) 2021). pp. 34–47. Springer International Publishing, Cham (2022)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

17. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research **18**(17), 1–5 (2017)
18. Liao, J., Liang, Y., Pan, J.: Deep facial spatiotemporal network for engagement prediction in online learning. Applied Intelligence pp. 1–13 (2021)
19. Murshed, M., Dewan, M.A.A., Lin, F., Wen, D.: Engagement detection in e-learning environments using convolutional neural networks. In: 2019 IEEE (DASC/PiCom/CBDCom/CyberSciTech). pp. 80–86. IEEE (2019)
20. Niu, X., Han, H., Zeng, J., Sun, X., Shan, S., Huang, Y., Yang, S., Chen, X.: Automatic engagement prediction with gap feature. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 599–603 (2018)
21. Rao, K.P., Rao, M.C.S.: Recognition of learners' cognitive states using facial expressions in e-learning environments. Journal of University of Shanghai for Science and Technology pp. 93–103 (2020)
22. Roy, S., Etemad, A.: Self-supervised contrastive learning of multi-view facial expressions. In: Proceedings of the 2021 International Conference on Multimodal Interaction. pp. 253–257 (2021)
23. Thiruthuvanathan, M., Krishnan, B., Rangaswamy, M.A.D.: Engagement detection through facial emotional recognition using a shallow residual convolutional neural networks. International Journal of Intelligent Engineering and Systems **14**, 236–247 (2021)
24. Thomas, C., Nair, N., Jayagopi, D.B.: Predicting engagement intensity in the wild using temporal convolutional network. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 604–610 (2018)
25. Thong Huynh, V., Kim, S.H., Lee, G.S., Yang, H.J.: Engagement intensity prediction withfacial behavior features. In: 2019 International Conference on Multimodal Interaction. pp. 567–571 (2019)
26. Vassileios Balntas, Edgar Riba, D.P., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 119.1–119.11. BMVA Press (2016)
27. Wang, K., Yang, J., Guo, D., Zhang, K., Peng, X., Qiao, Y.: Bootstrap model ensemble and rank loss for engagement intensity regression. In: 2019 International Conference on Multimodal Interaction. pp. 551–556 (2019)
28. Whitehill, J., Serpell, Z., Lin, Y., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagementfrom facial expressions. IEEE Transactions on Affective Computing **5**(1), 86–98 (2014)
29. Wu, J., Yang, B., Wang, Y., Hattori, G.: Advanced multi-instance learning method with multi-features engineering and conservative optimization for engagement intensity prediction. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 777–783 (2020)
30. Wu, J., Zhou, Z., Wang, Y., Li, Y., Xu, X., Uchida, Y.: Multi-feature and multi-instance learning with anti-overfitting strategy for engagement intensity prediction. In: 2019 International Conference on Multimodal Interaction. pp. 582–588 (2019)
31. Yang, J., Wang, K., Peng, X., Qiao, Y.: Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In: Proceedings of the 20th ACM international conference on multimodal interaction. pp. 594–598 (2018)
32. Zhu, B., Lan, X., Guo, X., Barner, K.E., Boncelet, C.: Multi-rate attention based gru model for engagement prediction. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 841–848 (2020)