

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and
Information Systems

School of Computing and Information Systems

4-2022

Fine-grained detection of academic emotions with spatial temporal graph attention networks using facial landmarks

Hua Leong FWA

Singapore Management University, hlfwa@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research



Part of the [Databases and Information Systems Commons](#)

Citation

FWA, Hua Leong. Fine-grained detection of academic emotions with spatial temporal graph attention networks using facial landmarks. (2022). *Proceedings of the 14th International Conference on Computer Supported Education, Virtual Conference, 2022 April 22-24. 2*, 27-34.

Available at: https://ink.library.smu.edu.sg/sis_research/7157

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylids@smu.edu.sg.

Fine-grained Detection of Academic Emotions with Spatial Temporal Graph Attention Networks using Facial Landmarks

Hua Leong Fwa^a

School of Computing and Information Systems, Singapore Management University, Singapore
hlfwa@smu.edu.sg

Keywords: spatial, temporal, affective states, facial landmarks, graph attention network, gated recurrent unit

Abstract: With the incidence of the Covid-19 pandemic, institutions have adopted online learning as the main lesson delivery channel. A common criticism of online learning is that sensing of learners' affective states such as engagement is lacking which degrades the quality of teaching. In this study, we propose automatic sensing of learners' affective states in an online setting with web cameras capturing their facial landmarks and head poses. We postulate that the sparsely connected facial landmarks can be modelled using a Graph Neural Network. Using the publicly available in the wild DAiSEE dataset, we modelled both the spatial and temporal dimensions of the facial videos with a deep learning architecture consisting of Graph Attention Networks and Gated Recurrent Units. The ablation study confirmed that the differencing of consecutive frames of facial landmarks and the addition of head poses enhance the detection performance. The results further demonstrated that the model performed well in comparison with other models and more importantly, is suited for implementation on mobile devices with its low computational requirements.

1 INTRODUCTION


The Covid-19 pandemic has transformed the way education is delivered globally. With face-to-face lessons curtailed, educational institutions have little choice but to transition to online lessons with the help of video conferencing and related technologies. Many educators and learners have however, critiqued that online lessons are impoverished in the affective and social interaction aspects and thus cannot completely replace traditional face-to-face delivery.

In a face-to-face classroom setting, teachers continuously monitor the emotional state of their learners to gauge their level of understanding and engagement. Contingent on the learners' cognitive state, the teachers can then adapt their tutoring style dynamically for example by slowing the pace of delivery or redirecting the learners' attention to the content material to optimize the learning. On the other hand, in an online lesson, learners hide behind the veil of a virtual screen. This is further exacerbated by a large class of online learners to a single teacher in a typical online session which constraints an individual learner's video to be a small window on the teacher's console, making it difficult for the teacher to infer their learn-

ers' affective state. To enhance the learning in online lesson delivery, it is thus essential to augment online lessons with automatic sensing of learners' affective or cognitive state.

Previous studies on emotion sensing have tapped on modalities such as audio (Sezgin et al., 2012; Guhan et al., 2020), body postures (D'Mello and Graesser, 2009), facial images and videos (Whitehill et al., 2008; Grafsgaard et al., 2013) as well as physiological signals such as ECG (Belle et al., 2012). The prevalence of web cameras and their use in various virtual learning platforms resulted in their extensive use for assessing learners' facial emotions. It is thus no coincidence that majority of recent works on automatic sensing of learners' emotions leverage on facial images and videos captured using web cameras. The affordability, unobtrusiveness and ubiquity of web cameras as compared to other sensors further adds to their use for emotion sensing.

Studies have shown that basic emotions such as happiness and sadness are not relevant to the learning process (Pekrun et al., 2002, 2007). Academic emotions or affect such as frustration, boredom, confusion and engagement are the ones which influence the learning process. The ability of expert human tutors to achieve enhanced learning outcomes has been widely attributed to their ability to sense the affect-

^a  <https://orcid.org/0000-0002-4472-2481>

tive states of the learners and to continually adapt their tutoring strategies in response to the dynamically changing affective states throughout the tutoring session. Sustaining the incidences of positive affect e.g. engagement and suppressing the incidences of negative affect e.g. boredom, the tutor (which can be human or computer) can enhance the interest and motivation of the learner in the learning process.

Facial emotion recognition can be broadly categorized into static and dynamic facial emotion recognition. Static facial emotion recognition techniques use static facial images as inputs while dynamic facial emotion recognition techniques use videos as inputs. Static facial emotion recognition techniques model only the spatial dimensions while on the other hand, dynamic facial emotion recognition model not only the spatial but also the temporal dimensions. By including temporal dimensions in the model, dynamic facial emotion techniques are known to offer higher accuracies of facial emotion detection as compared to static facial emotion recognition techniques.

In this paper, we focus on the detection of academic emotions as they are more relevant in a classroom or online teaching scenario. The detection of the academic emotions is however challenging as they are more subtle as compared to the basic emotions and thus the significance of this research. In an online learning context, with the successful detection of the academic emotional states of a learner e.g. when a learner is disengaged, bored or frustrated, a human or computer tutor can then enact appropriate pedagogical interventions to avert detrimental effects such as the learner giving up on the learning altogether.

Micro facial expressions changes is correlated to facial emotions and the detection and monitoring of facial landmarks movements is one of the established methods for tracking facial expressions. The network of facial landmark points can be formulated as a grid and this makes it amenable to be represented as a graph where each landmark point can be modelled to be spatially related to other landmark points. This motivates our study on the modelling of facial landmark changes through Graph Attention Network (GAT) (Veličković et al., 2018) using the Dataset for Affective States in E-Environment (DAiSEE) dataset (Gupta et al., 2016).

In short, our contributions can be summarized as follows:

- We proposed a novel spatial temporal GAT based model to detect fast changing intensity levels of academic emotions which include frustration, boredom, engagement and confusion.
- We captured facial landmark points and head poses as inputs as opposed to facial images, result-

ing in a computationally lighter academic emotion detection model which is more adequate for deployment on mobile devices.

- Our empirical results showed that the proposed graph based model outperforms Convolutional Neural Network (CNN) based models such as InceptionNet.

2 RELATED STUDIES

Prior studies have shown a tight coupling between head pose and engagement. A study by Asteriadis (Asteriadis et al., 2009) documented the use of head, eye and hand movements to gauge children's level of interest and attention in the context of reading an electronic document. Gaze Tutor (D'Mello et al., 2012), an intelligent tutoring system used a commercial eye tracker to identify the engagement of learners from their eye gaze patterns and attempts to re-engage the learner with dialogue moves. Tracking eye gaze allows one to capture instances of inattention where the head is stationary while eye focus on somewhere off-screen but tracking head pose is still relevant in numerous instances where user turns head away from the screen.

A CNN architecture was proposed in a study by Murshed et al. (2019) to classify engagement levels for the DAiSEE dataset. The proposed CNN architecture is then compared against three other popular CNN architectures after re-categorizing the original four levels of engagement into three levels – not engaged, normally engaged and highly engaged. The authors merged the labels of bored, frustrated and confused into the non-engaged category. The proposed model achieved an average accuracy level of 92.33% for the DAiSEE dataset. The authors did not model the temporal aspect but instead aggregated the facial frames into sliding window width of 5.

The Dataset for Affective States in E-Environment (DAiSEE) is the first multi-label academic emotions video classification dataset collected in the wild by (Gupta et al., 2016). It consists of 9068 snippets from 112 users and annotated with the affective states and intensity levels of boredom, frustration, confusion and engagement. The authors used InceptionNet, C3D and Long-term Recurrent Convolution Network (LRCN) deep learning models to classify the affective states and found the LRCN model to be the best performing model (with reported accuracies of 53.7%, 57.9%, 72.3% and 73.5% for boredom, engagement, confusion and frustration). The results of this study confirmed that temporal classifiers tend to outperform static classifiers for the classification of

affective states related to learning.

Liao et al. (2021) proposed the Deep Facial Spatiotemporal Network (DFSTN) comprising of a pre-trained SE-ResNet-50 for extracting the facial spatial features and passing them into an LSTM network with Global Attention module to generate an attentional hidden state. With the modelling of both spatial and temporal features, the authors managed to attain an accuracy of 58.84% on the four levels of engagement as labelled in the DAiSEE dataset.

A combination of ResNet and Temporal Convolutional Network (TCN) hybrid deep learning architecture was used in a study by Abedi and Khan (2021) for engagement detection in the DAiSEE videos. The pre-trained 2D ResNet extracts the spatial features while the TCN analyses the temporal changes in the video frames. A weighted cross entropy loss function is employed to tackle the issue of imbalanced dataset (low proportion of lower levels of engagement as compared to the higher levels of engagement). An accuracy of 63.9% is achieved on the four levels of engagement of the DAiSEE dataset. A point to note is that the previous two studies only focus on the detection of engagement and not the other academic emotions.

The study by Ngoc et al. (2020) inspires us to model facial landmarks as a graph attention network for the detection of academic emotions. In their study, the authors constructed a graph representation of facial landmarks using Graph Neural Network (GNN) and employed a Gated Linear Unit (GLU) as the temporal block to predict seven basic emotions, namely anger, contempt, disgust, fear, happiness, sadness, and surprise. The authors justify that facial landmark points have a sparse relation with spatially adjacent landmark points and in addition, a continuous relation of facial landmarks on the temporal axis as captured in a facial video makes them amenable to being modelled with a temporal graph network. The proposed network achieved an accuracy of 96.02%, 69.64% and 32.64% on the CK+ dataset, MMI and AFEW dataset (for classification of basic emotions and not academic emotions) respectively. This compares well with other state of the art benchmarks and justifies the use of graph neural network to model facial landmark points for facial emotion recognition.

Inspired by the previous work and the success of attention mechanism in previous studies, we employed the GAT for modelling the spatial features of facial landmarks. The previous study (Liao et al., 2021) used a pre-trained SE-Resnet-50 for extracting facial spatial features from facial images. In contrast, for our study, we used GAT which attends over the neighbours of each facial landmark node in a graph

structure for computing a hidden representation for individual frame of facial landmarks. The temporal dimensions of hidden representations of facial landmarks and head poses are then modelled with the Gated Recurrent Unit (Cho et al., 2014). We postulate that the inclusion of head pose will improve the performance of the detection of academic emotions. We also intend to incorporate this affect sensing module in a mobile-based intelligent tutoring system. The computation complexity of the model is integral to implementation on mobile devices as mobile devices typically have lower processing power and storage.

3 METHODOLOGY

This section describes the processing pipeline and proposed model architecture which uses both the GAT and Gated Recurrent Unit (GRU) as seen in Fig. 1.

3.1 Facial landmarks and head pose extraction

This study uses the DAiSEE dataset. Each DAiSEE video is 10 seconds in duration and is captured at a frame rate of 30 frames per seconds. The videos were first segregated into the individual frames which resulted in a total of 300 frames or static images per video. We next used the Multi-Task Cascaded Convolutional Neural Network (MTCNN) (Zhang et al., 2016) algorithm to crop out the facial images. MTCNN is employed here as it is balanced in both speed and performance as compared to other algorithms.

The cropped facial images were resized to 112 by 112 pixels before being passed to the Practical Facial Landmarks Detector (PFLD) (Guo et al., 2019) for extraction of the facial landmarks. PFLD is used here as it achieves good performance across challenging datasets with low computational requirements. Out of the 98 facial landmark points extracted by the pre-trained PFLD model, we excluded the facial landmark points that outlined the outer shape of the face as they were not correlated to affective states (Ngoc et al., 2020). In total, 65 facial landmark points that outlined the brow, eye, nose and mouth were used. The head pose angles of roll, pitch and yaw were also extracted from PFLD model.

We extracted facial landmark points from each 10 seconds video and for a video rate of 30 frames per second, this equates to a total of 300 frames of facial landmarks. Only 30 frames (1 frame picked at a fixed interval of every 10 frames) were used as input into the model. For videos with less than 15 frames,

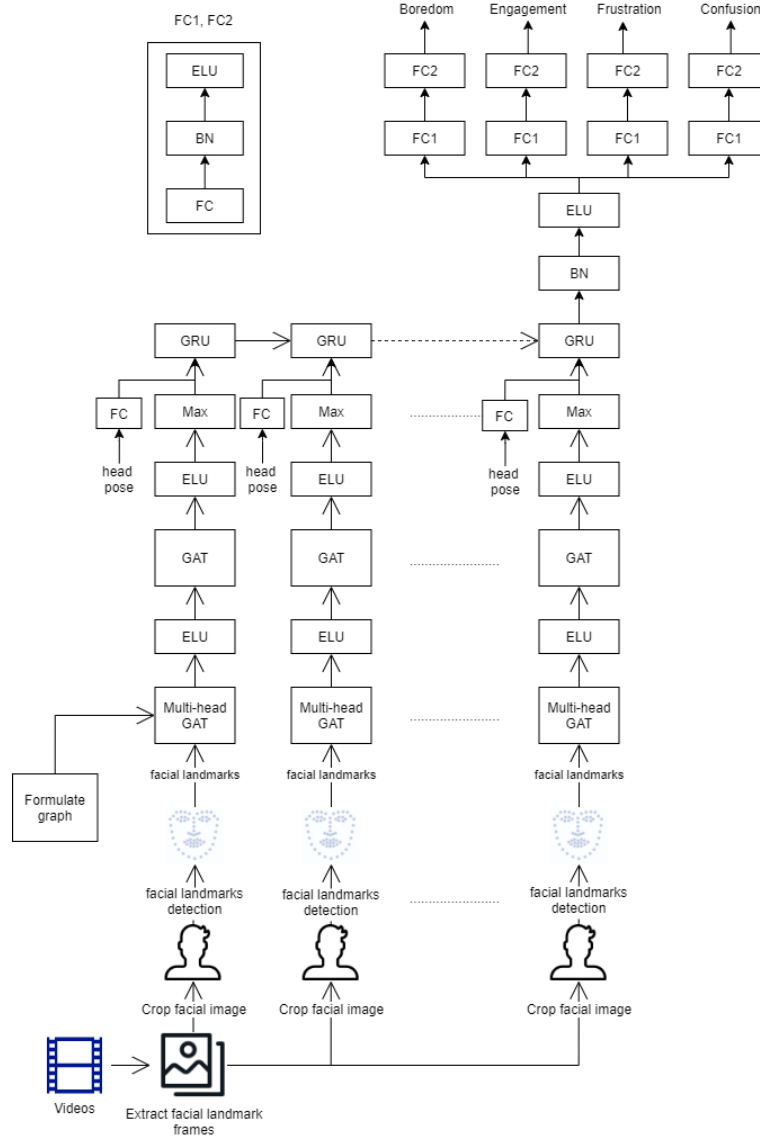


Figure 1: Modelling workflow and architecture

we interpolated additional frames between consecutive frames to make up 30 frames. Some of the videos with less than 15 frames of facial landmarks extracted were dropped. We postulate that the rate of movement of each facial feature contributes to the tracking of emotions. Thus across consecutive frames, we calculated the difference of the vertex coordinates for each facial vertex which symbolizes the rate of change of the facial features and input the rate of change of facial features into the model instead of the raw facial landmark coordinates.

The facial landmarks are formulated into the shape $N \times T \times V \times 2$ where N denotes the batch size, T denotes the temporal width of 29 frames, V denotes

the facial landmark vertex and the last dimension of 2 stores the x and y coordinates of the facial landmark point. We also formulated head poses into features with shape $(N \times T \times H)$ where N denotes the batch size, T denotes the temporal width. The last dimension of $H=3$ stores the roll, pitch and yaw values.

Across the 9068 DAiSEE videos, there were 144 videos in which facial landmarks cannot be extracted (no face can be detected in the video). Another 47, 49 and 17 videos were excluded from the train, validation and test videos respectively as there were less than 15 frames of facial landmarks which can be extracted. In total, only 2.8% of the entire dataset of videos were discarded.

3.2 Graph construction

Graphs are data structures that allow us to model a set of objects (vertices) and the relationships between them (edges) (Zhou et al., 2020). To transform the facial landmarks into a graph, each facial landmark point constitutes a vertex on our graph and the edges are formed with the Delaunay method. The Delaunay triangulation method (Lee and Schachter, 1980) has been used for construction of the graph structure linking the vertices in a number of previous studies (Fabian Benitez-Quiroz et al., 2016; Ngoc et al., 2020) that used facial landmarks for the detection of facial action units.

3.3 Spatial Temporal Graph Attention Network (STGAT)

The STGAT model is illustrated in Fig.1. The input to each GAT layer is a set of node features,

$$h = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_N\}, \vec{h}_i \in R^F \quad (1)$$

where $N = 65$, denoting the number of facial landmarks and $F = 2$ as each node feature consists of the x and y coordinates of the facial landmark point.

Within each GAT layer, we apply a linear transformation to encode the input features into embedded features.

$$z_i^{(l)} = W^{(l)} h_i \quad (2)$$

Next, a pair-wise un-normalized attention score is computed between two neighbours using the equation below and a drop out is then applied during training. The embedded features of the two nodes are concatenated and then dot product with a learnable weight vector a before applying a LeakyReLU to the product.

$$e_{ij}^{(l)} = \text{LeakyReLU} \left(\vec{a}^{(l)T} \left(z_i^{(l)} || z_j^{(l)} \right) \right) \quad (3)$$

The next equation applies a softmax to normalize the attention scores on each node's incoming edges.

$$\alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum_{k \in N(i)} \exp(e_{ik}^{(l)})} \quad (4)$$

Finally, the embeddings from neighbours are aggregated together and scaled by the attention scores.

$$h_i^{(l+1)} = \sigma \left(\sum_{j \in N(i)} \alpha_{ij}^{(l)} z_j^{(l)} \right) \quad (5)$$

We use Exponential Linear Units (ELU) as the activation function σ . W and a are weight matrices, l

denotes the layer and $N(i)$ denotes the set of one-hop neighbours of node i .

The max pooling operation then selects the maximum value across the facial landmark vertices to generate the embedding for the entire facial graph. The facial graph embedding is then concatenated with the head pose embeddings (after passing through a fully connected layer) before being passed to the GRU to model temporal features. GRU resolves the problem of vanishing gradient typically associated with a standard Recurrent Neural Unit. The output from the last GRU is then passed through Batch Normalization layer (BN), ELU activation layer to 2 layers of Fully Connected (FC) or linear, BN and ELU layers (FC1 and FC2) to produce the predicted scores for 4 levels of boredom, engagement, frustration and confusion.

4 MODELLING

With the use of Optuna library (Akiba et al., 2019) for hyperparameter tuning, we trained the STGAT using the train dataset and optimized the model for the best set of hyperparameters using the validation dataset. To train STGAT, we use the Root Mean Squared Propagation (RMSProp) (Tieleman et al., 2012) optimizer to train for a maximum of 40 epochs. An early stopping algorithm was employed to stop the experiments if the loss on the validation set did not reduce for consecutive 4 epochs.

We ran a total of 50 Optuna trials and saved the model which gave the lowest loss for each trial as well as their corresponding set of hyperparameter values. The hyperparameters include the learning rate, the hidden layer dimensions for GATs and GRU, FCs and the dropout rate for the drop out layers. A batch size of 64 and learning rate of 1×10^{-3} (selected through Optuna trials) was used for model inference. Table 1 shows the final configuration for the model that is used to generate the accuracies using the test dataset.

The overall accuracy was calculated by averaging the detection accuracies across the four states of boredom, frustration, engagement and confusion. The model which gave the best overall accuracy score among the trials was selected to be used for the evaluation of the performance using the test dataset.

4.1 Loss

We used a combination of the softmax and center loss (Wen et al., 2016) as the final loss function. We noted that with the softmax loss alone, the proposed model was not able to discriminate between the four levels of academic emotions. With the center loss function,

Table 1: STGAT model configuration

Configuration	No. of heads	Dropout	Input dim.	Output dim.
Multi-head GAT	6	0.5	2	80
GAT	1	0.5	480	48
GRU	-	-	48	112
FC	-	-	3	112
FC1	-	-	224	128
FC2	-	-	128	4

Table 2: Accuracies of STGAT face only, STGAT raw landmark models versus STGAT model

Models	Boredom	Engagement	Confusion	Frustration
STGAT face only	41.8%	49.6%	67.5%	78.0%
STGAT raw landmark	40.0%	49.6%	67.5%	78.0%
STGAT	46.2%	49.6%	67.5%	78.0%

the model needs to learn the center for the deep features in each class and minimize the distance between the features and its class center, resulting in better discrimination. The softmax loss function is defined as

$$L_s = - \sum_{i=1}^m \log \frac{\exp(W_{y_i}^T x_i + b_{y_i})}{\sum_{j=1}^n \exp(W_j^T x_i + b_j)} \quad (6)$$

where $x_i \in R^d$ denotes the i^{th} deep feature, belonging to the y^{th} class and $W_j \in R^d$ denotes the j^{th} column of the weights $W \in R^{d \times n}$ and $b \in R^n$ is the bias term. The center loss function is denoted by

$$L_c = \frac{1}{2} \sum_{i=1}^m \|x_i - c_{y_i}\|_2^2 \quad (7)$$

where c_{y_i} is the y_i^{th} class center of deep features. The final loss function is then denoted by

$$L = L_s + \lambda L_c \quad (8)$$

where λ (which is set to 0.2 in our experiment) is used for balancing the softmax and center loss functions.

4.2 Ablation study

Other than facial landmarks, head pose was also acquired from PFLD model. To investigate whether the addition of head pose improves the performance of our model, we constructed a model that uses only facial landmarks (referred to as STGAT face only model) which is similar to the STGAT model outlined in Fig. 1 but without the input of head poses to the GRU layer.

To verify that the differencing of facial landmark coordinates across consecutive facial landmark frames enhances the performance of STGAT, we compare it against another model (referred to as STGAT raw landmark) that takes in the raw facial landmark coordinates and head poses as input.

We followed the same workflow described in Methodology section for both STGAT face only model and STGAT raw landmark where the models were trained using the train dataset and optimized for the best set of hyperparameters using the validation dataset. The same optimizer and loss function were used and the hyperparameters were selected by using Optuna to run for 50 trials.

As seen in Table 2, although the accuracies for engagement, confusion and frustration are similar for both STGAT and STGAT face only models, the detection accuracy for boredom improves by 4.4% from 41.8% to 46.2% with the inclusion of head pose. Similarly, comparing the STGAT and STGAT raw landmark models, the detection for boredom improves by 6.2% for STGAT versus STGAT raw landmark model.

5 RESULTS AND DISCUSSION

The performance of STGAT as compared to other models is shown in Table 3. The accuracies for the detection of academic emotions for the first 5 models are retrieved from the study by Gupta et al. (2016). As compared to the other models, STGAT outperforms InceptionNet Video and performs on par with C3D Training but lags behind LRCN. It is worth noting however that the authors stated that they recreated the DAiSEE videos by treating every alternate frame as continual affective states for LRCN to achieve a higher performance.

In addition, LRCN consists of a combination of CNNs (to model the spatial dimension) connected to Long Short Term Memory modules (LSTMs) for modelling of the temporal dimensions which is com-

Table 3: Accuracy of models across all four levels of boredom, engagement, frustration and confusion

Models	Boredom	Engagement	Confusion	Frustration
InceptionNet Frame	36.5%	47.1%	70.3%	78.3%
InceptionNet Video	32.3%	46.4%	66.3%	77.3%
C3D Training	47.2%	48.6%	67.9%	78.3%
C3D FineTuning	45.2%	56.1%	66.3%	79.1%
LRCN	53.7%	57.9%	72.3%	73.5%
STGAT	46.2%	49.6%	67.5%	78.0%

putationally heavier as compared to our proposed model which consists of a light pre-trained facial landmark interference PFLD model, a relatively small GAT (only 65 nodes) and GRUs. To put in perspective, ResNet-50 a CNN that is 50 layers deep, trained on more than one million images from the ImageNet database, has close to 2.6 million parameters and is about 98 MB in size when compressed. ResNet-50 was used in studies by Abedi and Khan (2021) and Liao et al. (2021). In contrast, the proposed STGAT model has less than 115 thousand parameters and only 480 KB in size. With the inclusion of PFLD’s 1.27 million parameters and 6.9 MB in size, it is still about 14 times smaller than ResNet-50. Running on an Intel i3 with 4 core processor, 16 GB RAM and a NVidia GeForce GTX1660 super GPU with a batch size of 64, STGAT took 1.46 ms for face cropping and facial landmarks inference. Added to that, STGAT took about 4.8 ms to output the affective states giving a total inference time of about 6.26 ms. Both training and inference of deep learning models containing millions of parameters demand computational resources that are beyond the constraints of typical mobile devices (Wang et al., 2018). Specifically, the memory and battery resource of mobile devices are key considerations. For model inference, a basic requirement is the model must be small enough to at least fit into the limited on device memory and even if the model can be compressed to fit, a computationally heavy model will also rapidly deplete the battery of the mobile device. Both the computational and storage considerations are pivotal for implementation on mobile devices. A lighter model results in faster detection of the fast changing affective states of learners and depletes battery slower while a smaller sized model allows it to be deployed on a wide range of mobile devices including those with lesser on device memory.

Balancing for speed versus performance, we chose to use the pre-trained PFLD 0.25x model for facial landmarks detection. The larger pre-trained PFLD models gives better detection accuracies but results in slower inference. In addition, PFLD is also not the current state of the art model for facial landmark detection. Thus, we hypothesize that with better

facial landmark detection, the performance of STGAT should further improve.

6 CONCLUSION

In conclusion, we have demonstrated the feasibility of detecting academic emotions by proposing a graph based model using GAT and GRU that modelled both the spatial and temporal dimensions of facial landmarks. The ablation study also showed that the addition of head poses into STGAT and the differencing of consecutive frames of facial landmarks further enhance the accuracy of detection specifically for boredom. Although the results showed that the overall performance of STGAT is not state of the art, the proposed model has low computational requirements and is adequate for implementation on mobile devices. We contend that with the availability of better performing facial landmark detection models, the performance of facial landmarks graph neural network based model should improve further and this will be a direction for future research.

REFERENCES

- Abedi, A. and Khan, S. S. (2021). Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network. In *18th Conference on Robots and Vision, CRV 2021*, pages 151–157.
- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.
- Asteriadis, S., Tzouveli, P., Karpouzis, K., and Kollias, S. (2009). Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment. *Multimedia Tools and Applications*, 41(3):469–493.
- Belle, A., Hargraves, R. H., and Najarian, K. (2012). An automated optimal engagement and attention detection sys-

- tem using electrocardiogram. *Computational and mathematical methods in medicine*, 2012.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- D'Mello, S. and Graesser, A. (2009). Automatic detection of learner's affect from gross body language. *Applied Artificial Intelligence*, 23(2):123–150.
- D'Mello, S., Olney, A., Williams, C., and Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, 70(5):377–398.
- Fabian Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570.
- Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., and Lester, J. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*.
- Guhan, P., Agarwal, M., Awasthi, N., Reeves, G., Manocha, D., and Bera, A. (2020). Abc-net: Semi-supervised multimodal gan-based engagement detection using an affective, behavioral and cognitive model. *arXiv preprint arXiv:2011.08690*.
- Guo, X., Li, S., Yu, J., Zhang, J., Ma, J., Ma, L., Liu, W., and Ling, H. (2019). Pfld: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*.
- Gupta, A., D'Cunha, A., Awasthi, K., and Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*.
- Lee, D.-T. and Schachter, B. J. (1980). Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242.
- Liao, J., Liang, Y., and Pan, J. (2021). Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence*, 51(10):1–13.
- Murshed, M., Dewan, M. A. A., Lin, F., and Wen, D. (2019). Engagement detection in e-learning environments using convolutional neural networks. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 80–86. IEEE.
- Ngoc, Q. T., Lee, S., and Song, B. C. (2020). Facial landmark-based emotion recognition via directed graph neural network. *Electronics (Switzerland)*, 9(5).
- Pekrun, R., Frenzel, A. C., Goetz, T., and Perry, R. P. (2007). The control-value theory of achievement emotions: An integrative approach to emotions in education. In *Emotion in education*, pages 13–36. Elsevier.
- Pekrun, R., Goetz, T., Titz, W., and Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist*, 37(2):91–105.
- Sezgin, M. C., Gunsel, B., and Kurt, G. K. (2012). Perceptual audio features for emotion detection. *Eurasip Journal on Audio, Speech, and Music Processing*, 2012(1):1–21.
- Tieleman, T., Hinton, G., et al. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.
- Veličković, P., Casanova, A., Liò, P., Cucurull, G., Romero, A., and Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*.
- Wang, J., Cao, B., Yu, P., Sun, L., Bao, W., and Zhu, X. (2018). Deep learning towards mobile applications. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 1385–1393. IEEE.
- Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer.
- Whitehill, J., Bartlett, M., and Movellan, J. (2008). Automatic facial expression recognition for intelligent tutoring systems. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6. IEEE.
- Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.
- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.