Video-based Student Engagement Estimation via Time Convolution Neural Networks for Remote Learning

Khaled Saleh, Kun Yu and Fang Chen
Data Science Institute/UTS

MOTIVATION

Given the recent circumstances of the outbreak of COVID-19 pandemic globally, most of the schools and universities have shifted many of the learning materials and lectures to be delivered online. As a result, the necessity to have some quantifiable measures of how the students are perceiving and interacting with this “new normal” way of education became inevitable.

SCOPE

In this work, we are focusing on the engagement metric which was shown in the literature to be a strong indicator of how students are dealing with the information being presented to them

METHODS

In our formulation for student engagement estimation problem, we tackle it using a data-driven approach based on convolutional neural network (ConvNet). More specifically, We cast the problem as a regression task, where given an input sequence of video frames of students, the output is an estimated continuous value that represents the student's engagement level during that window of input video frames.

The following is our proposed framework (shown in Figure 1), which is internally contains two main modules, namely the spatial feature extractor module and the temporal modelling module.

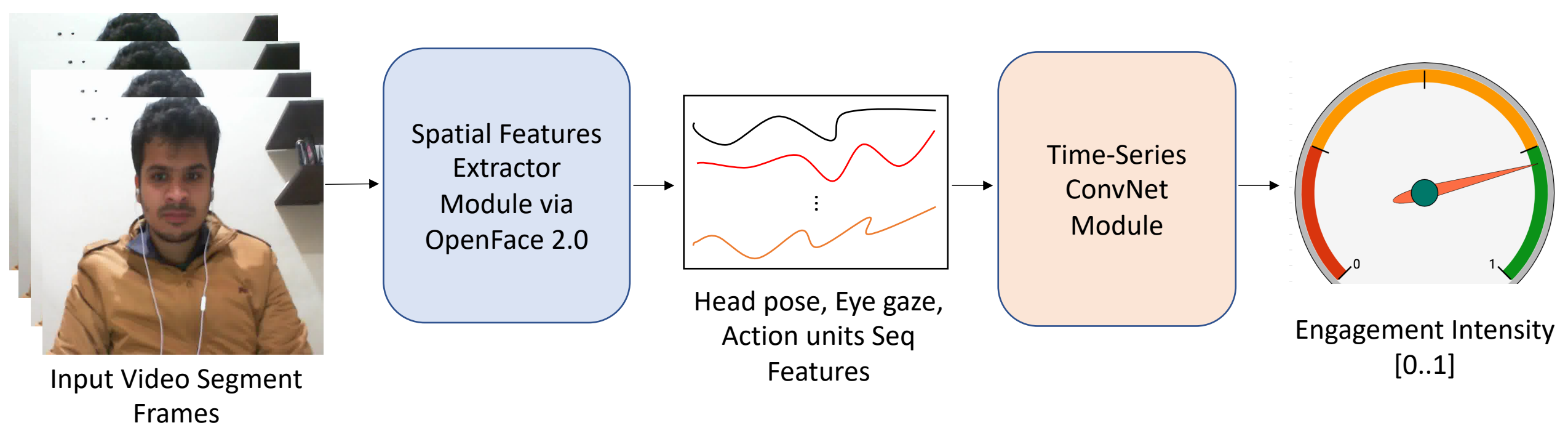
Input Video Segment Frames
Spatial Features Extractor Module via OpenFace 2.0
Head pose, Eye gaze, Action units Seq Features
Time-Series ConvNet Module
0
1
Engagement Intensity [0..1]

Figure 1: Our proposed framework for student engagement estimation.

The spatial feature extractor is based on OpenFace 2.0 [1] which is responsible for capturing and acquiring a frame-wise spatial representations of each frame from the input sequence. In this module, some interpersonal features such as head pose, facial landmarks, eye gaze and facial expression/action units are extracted. In our framework, we only relied on three types of features, namely head pose translation and rotation metrices in 3D, eye gaze direction in 3D and facial expressions intensity.

On the other hand, the temporal modelling module is responsible for modelling and capturing the temporal dependency between the input sequence spatial features. In this module (shown in Figure 2), we are adopting a network architecture based on time-series ConvNet architecture [2] that can model multivariate correlated time series data.

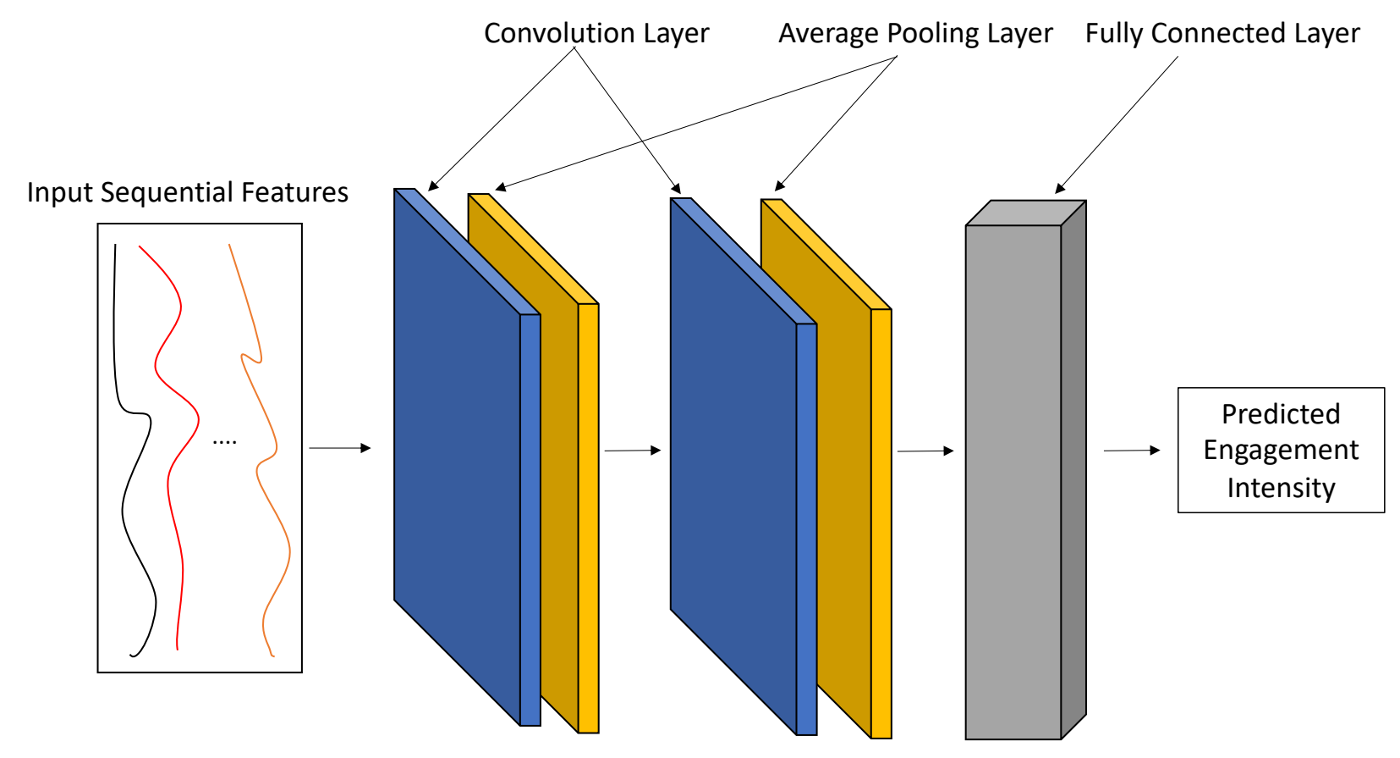

Figure 2: The architecture of our Time-Series ConvNet Module.

The time-series ConvNet architecture is 1D ConvNet that utilises 1D convolution layers by sliding a convolution filter over the input time series. In return, it can preserve the integrity of the temporal dimension of the input time series instead of the spatial dimensions of 2D images in case of 2D ConvNet architecture.

DATASET

In order to train and evaluate the performance of our proposed framework, we relied on the Engagement Prediction in the Wild dataset [3] (shown in Figure 3) in our experiments.

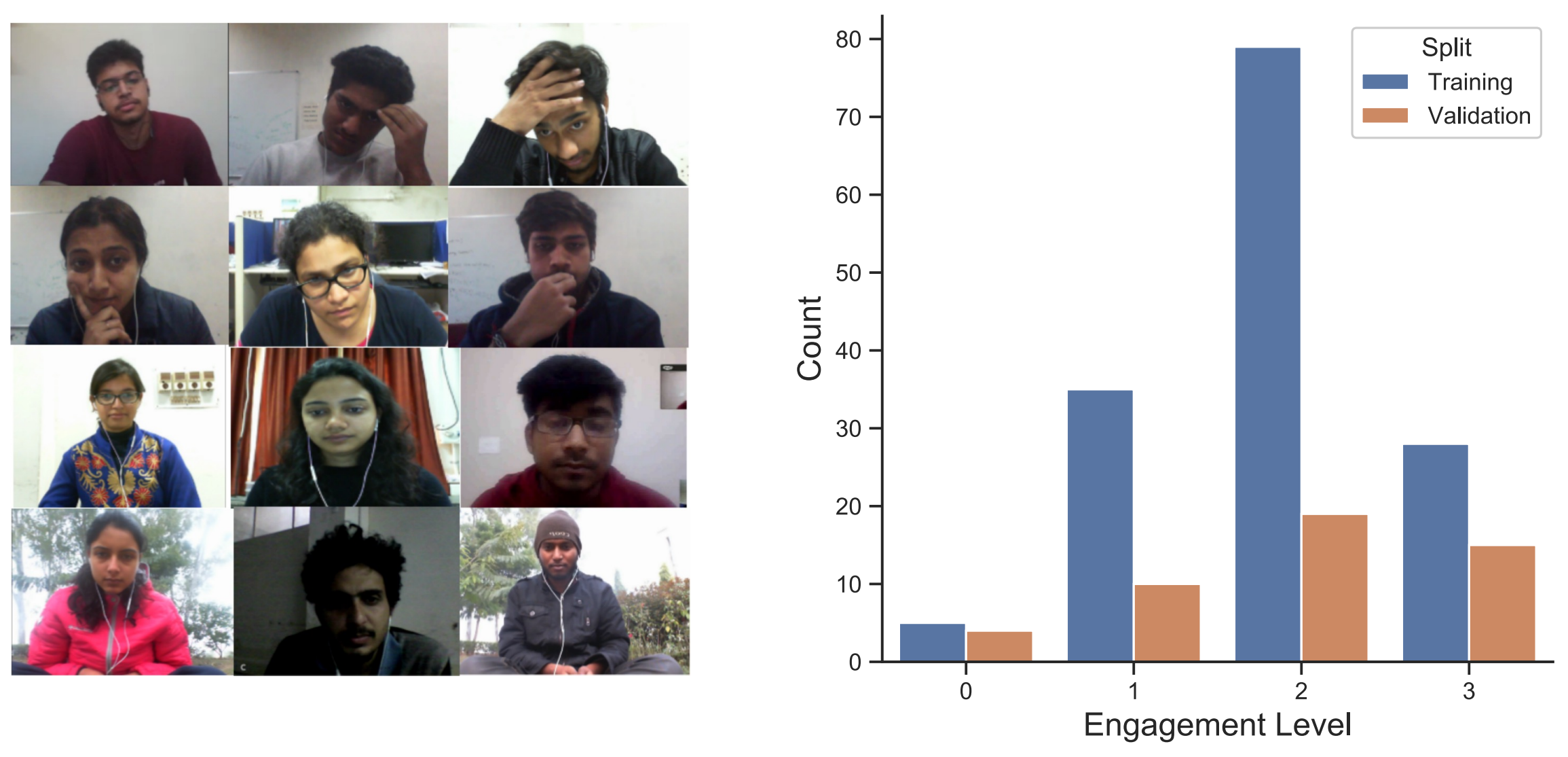

Figure 3: The Engagement Prediction in the Wild dataset [3] used in our experiments.

The dataset are a recorded videos of students who were taking MOOC and each video captures the upper body including the faces of students for a duration of 5 minutes. The dataset covers a wide variety of background environments with different wildness attributes such as varying illumination.

The total number of videos in the dataset are 262 videos, with 147 videos for training, 48 videos for validation and 67 videos for testing. Each video in the dataset was holistically labelled with values that corresponds to student engagement levels based on video coder labellers.

RESULTS

We report the performance of our proposed framework (in Table 1) in terms of mean-squared error (MSE) and compare it against two different approaches from the literature (for more information about the inner details of each approach, please check out the main paper).

Model	Testing (MSE)
Baseline	0.19
LSTM(Face+Body) [3]	0.08
Time-Series ConvNet (Ours)	**0.07**

Table 1: Performance of our proposed framework over the testing split of the Engagement Prediction in the Wild dataset [3].

As it can be noticed from the table, our proposed approach achieved the lowest MSE score (with 0.07) in comparison to the other two compared models. The closest score to our proposed model was the `LSTM(Face+Body)' model which scored 0.08 in MSE. From these scores, we can deduce that the Time-Series ConvNet was comparable (and even scored better MSE score) with the LSTM models which is well known for modelling sequential data. One of the main reasons for that, in LSTM it is assumed that there is some temporal dependency between the input sequential observations to the model. However, in our task of the students’ engagement estimation, this assumption does not hold because students’ behaviours and emotions are independent over time and can change drastically at any time throughout the observed input frames.

References

[1] OpenFace 2.0: Facial Behavior Analysis Toolkit Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency, IEEE International Conference on Automatic Face and Gesture Recognition, 2018.

[2] Bendong Zhao, Huanzhang Lu, Shengfeng Chen, Junliang Liu, and Dongya Wu. 2017. Convolutional neural networks for time series classification. Journal of Systems Engineering and Electronics 28, 1 (2017), 162–169.

[3] Abhinav Dhall, Garima Sharma, Roland Goecke and Tom Gedeon, “EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal based Challenges”, ACM International Conference on Multimodal Interaction 2020.

https://www.uts.edu.au/data-science-institute

UTS CRICOS PROVIDER CODE: 00099F