Inferring the Students' Attention in a Machine Learning Approach: A Feasibility Study

Jiachun Li,Ruiqi Li, Yuting Zhou, Junlin Xian, Xiong Zhang, Xiaojun Hei
School of Information Science and Engineering
Huazhong University of Science and Technology, Wuhan, China
Email:jiachunli@hust.edu.cn,jlxian@hust.edu.cn,heixj@hust.edu.cn

Abstract—This paper proposes a multi-model intelligence teaching system, which uses ordinary student data as a research sample, including video and audio. Through comprehensive analysis and the machine learning method, feedback on the students' listening status in real time can effectively help teachers adjust and improve the teaching strategy. This can help achieve personalized teaching content recommendation and counsel for students. Through machine learning, the instant feedback of the multi-model smart classroom can provide good help to students' academic development.

Keywords-smart classroom, sensor, face recognition, feedback

I. Introduction

Recently, Y. Kim et.al used data such as eyeball trace, arm gesture etc. to evaluate and promote the teaching process with a Restricted Boltzmann Machine[1]. In 2018, Anusha James et.al looked at utilizing video and audio data to classify the activeness of interaction, thus elevating effect of teachers to students[2]. R. Martinez-Maldonado et.al described a system to monitor the situation of group discussion[3]. The work of N. Gligoric et.al focused on assist teaching through the IoT(Internet of Things) equipment[4]. However, this paper focuses on designing a real-time teaching assist system with a multi-model information analysis method. And we will consider designing a specific scene for people like the autistic student.

II. Problem elicited

At present, the recognition of students' listening status is still a difficult problem, it is hard to achieve full coverage by manual observation. Therefore, we propose a multi-model analysis of students' learning status with the data from video and audio.

Our conceptual classroom is located at Huazhong University of Science and Technology, where many sensors, cameras and wireless microphones are placed to capture audio and video information from students. We apply the edge computing network in the complex scene of the future classroom, aiming to improve the atmosphere, optimizing the students learning experience.

III. System model

The system uses an edge computing network, and the main module is a convolutional neural network(CNN). The CNN network is responsible for connecting edge nodes and collecting vectors obtained by sensors and other structures. Various sensor structures capture information such as student gesture, sounds, faces, heart rate etc., and convert the original audio and video information into processed audio and video vectors to quantify the students' attendance status. The output of the CNN network is the state of "attentive" and "inattentive". The fluctuation curve on the time axis reveals the number of children who concentrate on listening to the class.

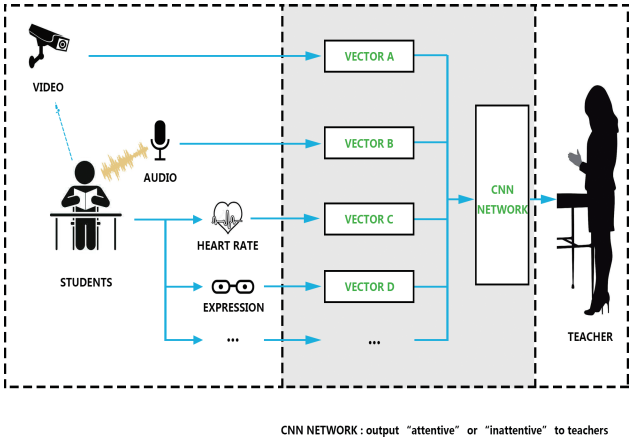

Fig. 1. System model

IV. System Design

A. Video CNN network

The classifier is trained by constructing the CNN network to judge the students' listening state in real time. In the experience, we collected the face information by the video for network training. This is based on the implementation of the facial feature point landmark: we pre-processed the input image, accurately identified the positions of different features of the face, and made 32 pixels. We trained the 400 pictures captured by the video manually, and labeled two different states: attentive and inattentive. So we can form a trusted input vector to the CNN network, which can calculate the probabilities of the "attentive" and the "inattentive" state, and make the change rule of the students' attention concentration with the time axis, including the instant feedback on the seat lecture information. We can automatically implement

the following training process during class, which can be applied in real time when teaching.

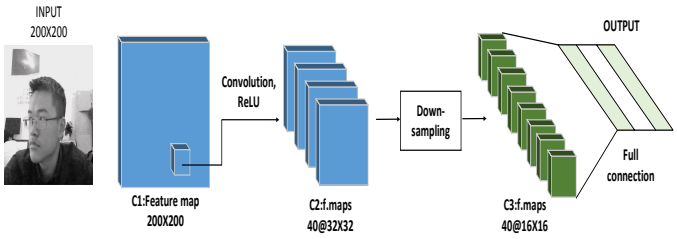

Fig. 2. CNN network

The dataset based on the orientation of the face is constructed to identify the state. The basic unit of training network consists of a convolutional layer, an active layer, a convolutional layer and a pooling layer. The Dropout method with a parameter of 0.5 is used to avoid overfitting. The training uses the SGD method for gradient descent, and the Batch size is 32. The training image is a 200*200 black and white image, which has 40 iterations. The network contains 32 convolution kernels, and the pooling layer is 2 units. We used multiple methods to capture and evaluate video information.

Parameter Specification	Value
Batch size	32
Number of classes	2
Number of epoch	40
Image row and column	200 x 200
Number of channels	1
Number of filters	32
Pooling size	2
Number of convolution	3

Fig. 3. Parameter in training

B. TextCNN network

We define several class atmosphere states derived by [2]. To obtain real time atmosphere state, an analysis of audio information collected by microphone installed in smart classroom is required. The procedure of audio information is divided as speech recognition, text segmentation, word vector transfer, sentence classification.

Speech recognition is referring as a major problem which Artificial Intelligence area has been researching in for decades. Nowadays plenty of open source speech recognition engine is adaptable for students and researchers. We test our system on Google Cloud Speech-to-Text API[5], which works on over 120 languages. Results shows adequate accuracy on Mandarin recognition, while sometimes it is hard for the machine to recognize speech of students from speech of teacher. In that case we make our attempts on different microphone position in order to separate speech of students, and finally gain obvious advances on speech accuracy and efficiency.

Different from English, Chinese sentence is composed by monosyllable characters while one or several characters form a meaningful word. Under this situation analysis on Chinese sentence should start from separating characters in a sentence to words. The course is known as Chinese Text Segmentation. In our system we use “pkuseg” tool proposed by Peking University recently, and modulate working environment condition to be teaching situation by employing specific dictionary[6].

Convolutional neural networks have achieved remarkable results in both computer vison and natural language processing area. Yoshua Bengio proposed a succinct network to generate vectors representing corresponding words, and made a successful attempt on eliminating the curse of dimensionality in 2003. In 2014, Yoon Kim proposed a deep learning method for sentence classification problem, as known as TextCNN network. On top of word vectors trained by Mokolov et al., the model improved upon the state of art on the task.

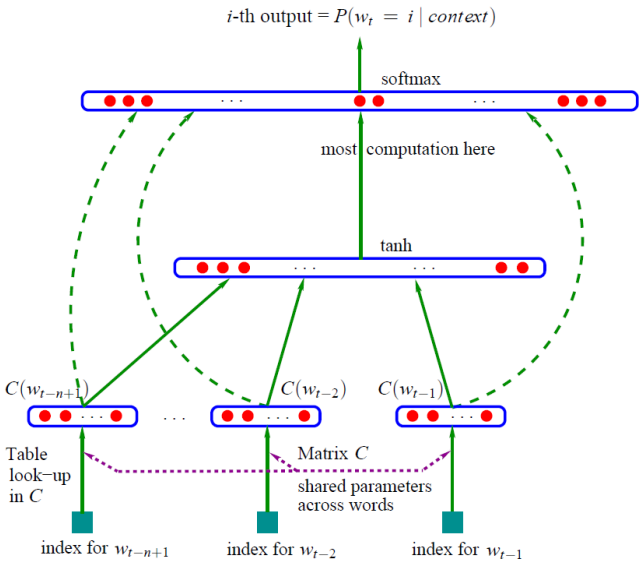

Fig. 4. TextCNN model

The architecture of this model is shown in figute. Let n amout of k-dimensional word vectors compose n, x, k representation of the n-word sentence to be disposed, after which the matrix is processed by convolutin layer, activation unit, pooling layers and fully connected layers which form the hidden layers. We figure out the probabilities of each classes of sentence in this way[1].

Similar to the processing of the video, the textcnn network is supplemented and the translated statement is

input into the trained textcnn network to obtain the score of the feature vector. Through the weighted summation, the degree of association between the sentence and the classroom is obtained, and the state acquired by the video cnn network is integrated to judge the state of the student's lecture.

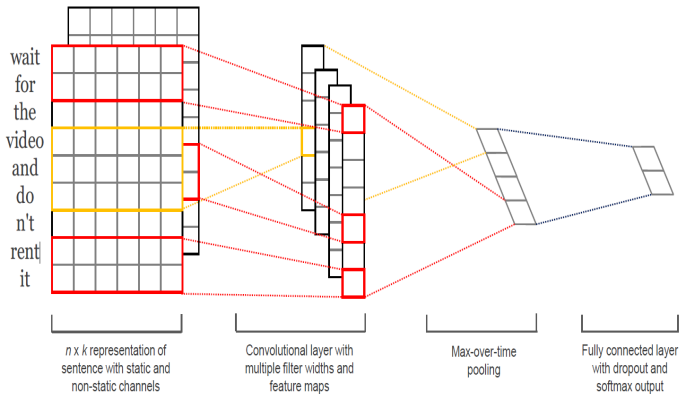

Fig. 5. model architecture with two channels

The components include: hidden layer, convolution layer, maximum pooling layer, softmax layer. To avoid overfitting, we use dropout rate: 0.5, mini batch size: set to 50. Finally, we output the probability of each feature category, and the degree of association with the current classroom can be obtained by comparing the magnitudes of the probabilities to further confirm the state of the lecture. When the output synthesis of both networks is judged to be attentive, the status of the classmate is confirmed as concentrating in the class.

V. Performance Evaluation

During the experiments, the CPU used i5-7300hq, memory 8G, and the graphics card used Nvida GeForce 1050ti. By training 400 sets of data, we obtained the loss function and the validation function curve.

Loss function maps a value of variables onto a real number intuitively, which represents some "cost" associated with the event. By manually observing the results of the listening and non-listening states, we compared the output of the CNN network and divided the matching by the total amount of training as the accuracy rate. The delay is a time from inputting the data to generating the "attentive" or "inattentive" states to the teacher's interface. After being optimized, it is shortened to the instant operation. We will also try algorithmic acceleration based on Kinect and FPGA to reduce the latency.

Figure 6 shows that, when the speech signal collected by the sensor is processed, the accuracy of speech recognition and word segmentation varies with the sampling rate. We continued to reduce the error rate of recognition by increasing the sampling rate, focusing on the extraction of high-frequency vocabulary. It can be seen that choosing a suitable sampling rate is a significant aid to the translation of the voice signal.

When analyzed only by video, the accuracy curve is shown in figure 7, after several times of training, the confirmation function of loss function keeps decreasing, which reveals the effect is good, and the comprehensive result also reflects the proper treatment of overfitting. Combining the two kinds of data from video and audio, the loss function is more fitted. The performance and accuracy are greatly improved, which is increased from 71.3 percent to 87.4 percent.

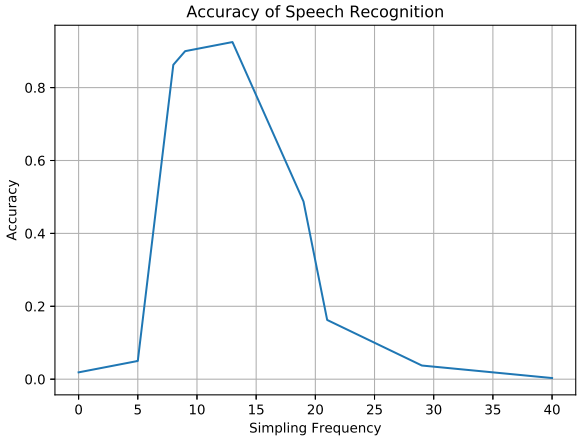

Fig. 6. audio result

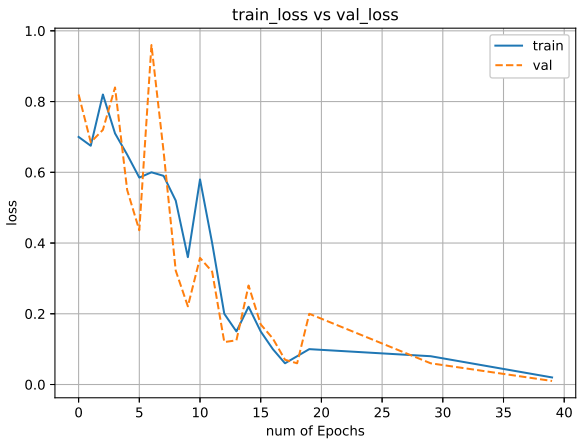

Fig. 7. loss function

VI. Conclusion and future work

This paper adopts multi-modal analysis method, which is a smart classroom aid system developed for students. It can provide good feedback for teaching and can also provide personalized learning recommendations for students. The audio and video analysis tools are combined with the techniques related to machine learning face recognition,

and the experimental algorithms are continuously optimized.

We're also going to use Kinect for algorithm acceleration. Kinect can collect high-quality images every second, which is easy for face detection and tracking, human-computer interaction, 3D modeling, and it is also effective to extract features, improving the accuracy of class status judgment. We will apply more statistical methods to verify the system, and try to apply as early as possible in actual teaching for autism.

VII. 3D Kinect Face Alignment

Kinect face recognition is something we're working on. Kinect has obvious advantages in depth image acquisition and 3D image processing compared with 2D images captured by the camera before. After several iterations and improvement of the number of feature points calibration, the number of labeled points this time increased from 32 to 80 (from the dlib library function) to obtain the depth image of face samples. Compared with 2D images, depth images have more data points for training, and the obtained feature vectors and classifier outputs are more accurate, which is conducive to improving the experimental accuracy.

References

[1] Y. Kim, T. Soyata, and R. F. Behnagh, "Towards emotionally aware ai smart classroom: Current issues and directions for engineering and education," IEEE Access, vol. 6, pp. 5308–5331, 2018.

[2] A. James, M. Kashyap, Y. H. V. Chua, T. Maszczyk, R. Bull, and J. Dauwels, "Inferring the climate in classrooms from audio and video recordings: A machine learning approach," in IEEE International Conference on Teaching, Assessment and Learning for Engineering, 4-7 December 2018.

[3] R. Martinez-Maldonado, K. Yacef, and J. Kay, "Data mining in the classroom: Discovering groups' strategies at a multi-tabletop environment," in Educational Data Mining 2013, 2013.

[4] N. Gligoric, A. Uzelac, S. Krco, I. Kovacevic, and A. Nikodijevic, "Smart classroom system for detecting level of interest a lecture creates in a classroom," Journal of Ambient Intelligence and Smart Environments, vol. 7, no. 2, pp. 271–284, 2015.

[5] https://cloud.google.com/speech-to-text/

[6] https://github.com/lancopku/PKUSeg-python/