 doi:10.20944/preprints202105.0303.v1

Transfer learning approach for analyzing attentiveness of students in an online classroom environment with emotion detection

Karan K V [1] and Vedant Bahel [2]

[1] Sri Sairam Engineering College, Chennai, India
[2] G H Raisoni College of Engineering, Nagpur, India
[1] karankv26102002@gmail.com

Abstract. There is a crucial need for advancement in the online educational system due to the unexpected, forced migration of classroom activities to a fully remote format, due to the coronavirus pandemic. Not only this, but online education is the future, and its infrastructure needs to be improved for an effective teaching-learning process. One of the major concerns with the current video call-based online classroom system is student engagement analysis. Teachers are often concerned about whether the students can perceive the teachings in a novel format. Such analysis was involuntarily done in the offline mode, however, is difficult in an online environment. This research presents an autonomous system for analyzing the students' engagement in the class by detecting the emotions exhibited by the students. This is done by capturing the video feed of the students and passing the detected faces to an emotion detection mode. The emotion detection model in the proposed architecture was designed by fine-tuning VGG16 pre-trained image classifier model. Lastly, the average student engagement index is calculated. We received considerable performance setting reliability of the use of the proposed system in real-time giving a future scope to this research.

Keywords: Emotion detection, CNN, VGG16, Education, Transfer learning, Engagement.

1 Introduction

In recent times, with the world going online, there have been advances in the social software technology in the field of education. Especially, with increasing digitalization, many researchers have worked on intelligent tools to improve the educational system. Educational Data Mining (EDM) is a popular field of research focusing on how data science can be used on data from educational settings [1,2, 28]. In [2], authors discuss how educational data can be used for multiple applications like student performance prediction, course recommendation, early dropout predictions, and some more. In this research, we focus on a similar application of analysing student's engagement in an online video conferencing-based classroom system from their video feed using computer vision.

 doi:10.20944/preprints202105.0303.v1

The interest in e-learning is on an upward trend especially in the period of pandemic and it seems to increase higher in the future. There aren't necessary tracking systems available for the educational institutions to track the engagement of the students during their lectures and sessions, which makes teachers helpless for the progress of their students. Thus, the application discussed in this paper is in more need than any time before.

In on-campus classroom learning, teachers were able to receive continuous feedback on their teaching by witnessing student's reactions to what they are learning. Such feedback often helps to understand the state of the students about specific concepts in the class and allows teachers to take necessary steps. For example, if the teacher senses that the class seems to be confused about certain concepts, the teacher could sense and revisit them. But in online classroom systems, that seems to be lacking. In [3], Raes et. al discusses the difficulties that the teachers face to engage students in a remote learning environment as compared to the face-to-face learning environment. In the study of Weitze [4], both students and teachers state that remote students learned less, were generally more passive, and often behaved like they were watching TV and not attending a lesson. The above statements marks that facial expressions are the vital identifiers of human feelings. Detecting facial emotion is quite an easy task for the human brain. But the same becomes challenging when we achieve this task with a computer algorithm. With the recent advancement in computer vision and machine learning models, it is possible to detect facial emotions from images and videos synchronously. Our system tries to detect the facial emotions of the students i.e, confused, happy, neutral, sleepy and displays the emotional index of the class as a whole. In this way, teachers will be able to know/understand the state of the class, making them feel comfortable and ensuring the reach of the knowledge shared with the students. We have used computer vision via transfer learning.

Transfer learning is roughly defined as the task of improving a machine learning model by leveraging knowledge from existing models in the same domain [17]. Such methods of fine-tuning pre-trained deep learning models are extremely beneficial when there is less data for the current tasks [18]. Transfer learning can be improved by these three measures. First is the initial performance achievable with the transferred knowledge before any further learning. Second is the difference between the time taken to finish the learning with transferred knowledge to the time taken for achieving the same from scratch. Third is the final performance level achievable in the target task with transferred knowledge compared to the final level without transfer [17]. Figure 1 shows the level of accuracy and prediction with and without transfer.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 May 2021 doi:10.20944/preprints202105.0303.v1

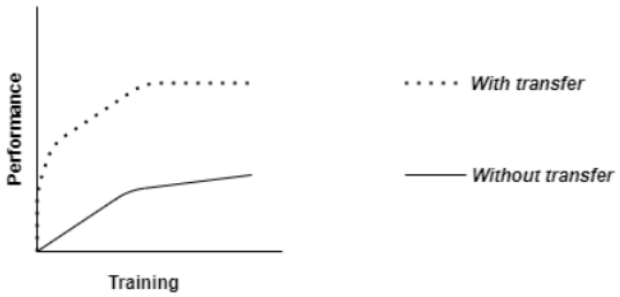

Fig. 1. General observed performance of model with and without transfer learning.

Various pre-trained neural network models like ResNet, MobileNet, VGG, Inception, etc have found to out-perform traditional methods for a variety of tasks [19]. In this research, we have used VGG16 pre-trained neural network model since the model is trained on 3.31 million images of person identity. Thus, the nature of this model suits best for our target task of emotion recognition [20].

2 Literature Review

Facial expression is the most common way of conveying the mood or feelings, not only for human beings but also for many other living organisms. There are a lot of attempts in developing the technologies that analyse facial expression as it has many applications in fields like Lie-detector, medicine, and robotics [5-7]. In a recent study on facial recognition technology (FERET) dataset, Sajid. et .al found that the impact of facial asymmetry is the marker of age estimation [8]. Since the twentieth century, Ekman. et. Al [9] defined seven moods or expressions in which a human grows irrespective of culture, tribe, and ethnicity. They were anger, sadness, fear, disgust, contempt, happiness, and surprise. In a recent study on Facial emotion recognition using convolutional neural networks (FERC), Mehendale [10] described a way of improving the accuracy of facial emotion detection by using single-level CNN and by novel background removal techniques.

There has been substantial work on determining the user emotions, one of which is done by McDuff. et al [11] at MSR, developed techniques to determine three aspects of human emotion: valence (checking the positiveness or negativeness), arousal (degree of emotion), and engagement level. These mining algorithms use data captured from hardware sensors like microphones, web cameras, GPS, etc. In addition, they used interaction data such as web URLs visited, documents opened, applications used, emails sent and received, and calendar appointments. They used the inferences mainly to allow users to reflect on their mood and recall events in the last week. Dewan et.al examined the engagement of the remote audience through computer vision. They just measured two scenarios, namely bored and engaged, and displayed the result of the engagement of the audience using OpenCV library [16] and computer vision.

 doi:10.20944/preprints202105.0303.v1

With the increase in the companies that offer video-based learning services such as tuitions and exam preparation purposes, video-based learning has become a trend due to various benefits [12], most importantly getting audio and video communications at the same time. Theories suggest that there are two subsystems involved here [13,14]. One is the processing of visual objects and the other is verbal processing. They both happen separately in our brains and can only process limited information [13], which in turn distracts students. Some methods have been proposed by researchers using video learning analytics to better understand how students learn with videos. For example, Kim et al. [15] investigated the learners' in-video dropout rates and interaction peaks by analysing their video-watching patterns (pausing, playing, replaying, and quitting) in online lecture videos. But to date, there is limited research on how students interact with video lectures, and therefore, it is difficult for instructors to determine the effectiveness of the learning design, especially at a fine grain level.

Transfer learning has been recently used commonly by a lot of researchers where Artificial intelligence and machine learning models are in use. Hazarika et.al proposed a way to recognize the state of emotion in a conversational text through transfer learning and natural language processing [23]. Transfer learning has played a critical role in the success of modern-day computer vision systems. Knetsch et al used computer vision and deep-learning techniques for the analysis of drone-acquired forest images on invasive species through transfer learning. He also mentions that the usage of transfer learning in his analysis improved the accuracy by 2.7% [24]. This, evidently in literature there have been multiple approaches in using transfer learning for a wide variety of tasks in the domain of computer vision. Additionally, researchers have also implemented this approach specifically for emotion recognition. However, there is hardly any approach to use emotion recognition for teaching-learning environments to improve the online learning system, which is the goal of this research.

3 Dataset

The primary objective of this research was to create a model that can detect the facial expressions of the students in the class. There are a lot of varied facial expressions that a human being can exhibit. However, we considered those facial expressions that are crucial when the learning activity of a student is concerned. This was also decided based on what teachers find most relevant in the classroom learning sphere. Psychological research shows that some positive emotions of students, such as concentration, happiness, and satisfaction promote their learning interests, motivation as well as cognitive activities while negative emotions, such as boredom, sadness, anxiety, etc. can have a bad influence on students' commitment and patience [21].
In this paper, we have considered 4 classes of emotion namely: confused, sleepy, happy, and neutral. The facial dataset for each class was collected automatically by scrapping publicly available google images. Web scraping is the process of using bots to extract content and data from a website. A variety of bots are used for web scraping for the applications like recognizing unique HTML site structures, extracting and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 May 2021 doi:10.20944/preprints202105.0303.v1

transforming content, store scraped data, and extracting the data from API [26]. Web scraping of publicly available google images is one of the efficient ways to collect the data to train the model. We have used python to scrape the image through the firefox web driver. The benefit of this approach is that (a) it retrieves thousands of images "from the wild" and (b) we can automatically label the images using the keywords in the query [25]. We have used Selenium and Beautiful Soup libraries for scraping purposes. The code used for scraping these google images can be found in this Github link: https://github.com/karankv26/Google-image-webscraper. The class size for each class of emotion is discussed in Table 1. Though the size of the dataset is relatively lesser than the general requirement of deep learning models. However, the transfer learning approach considered in this research has proven to show good results even with a limited size of data [22].

Table 1. Number of training and validation images in each class in the dataset

Class Name	Happy	Confused	Sleepy	Neutral
# training images	126	228	168	161
# validation images	18	18	18	18

The data is made publicly available for widening the scope of this research and potential future work to improve the current system. Dataset can be accessed at https://github.com/karankv26/Google-image-webscraper/tree/main/dataset.

4 Method & Implementation

The first module of the proposed project pipeline is to capture the snapshot of an active online classroom screen in a grid format. That image is passed to the face detection model which detects individual faces and extracts them as separate image files. The face detection model is based on OpenCV.

Further in the pipeline, the individual identified facial images are passed to the emotion detection model. In this research we have considered the VGG16 pre-trained model and have fine-tuned it for our dataset. VGG16 is a convolutional neural network model proposed by Simonyan and Zisserman from the University of Oxford in a paper named "Very Deep Convolutional Networks for Large-Scale Image Recognition" [27]. The model achieves 92% top-5 test accuracy in ImageNet, which consists of over 14 million images belonging to 1000 classes. It increases the depth of the architecture with very small (3 × 3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. This model was trained for weeks and was using NVIDIA Titan Black GPUs.

One of the significant ways to avoid overfitting is to use a larger dataset. However, over scraping of the dataset (that was done) produced garbage data with irrelevant images. Thus, we considered image augmentation to increase the size of the dataset by various transformations of the scraped relevant dataset. Some of the image augmenta-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 May 2021 doi:10.20944/preprints202105.0303.v1

tion techniques practised in this research are rotation, width_shift, height_shift, shear, zoom and horizontal flip.

Finally, the emotion detector model finds individual emotions and based on that the class emotion index is calculated using the following formula.

$$Emotion\ Index\ (emotion = e) = \frac{\#\ of\ students\ exhibiting\ 'e'}{total\ \#\ of\ students}$$
$$, where\ emotion = \{happy, confused, sleepy, neutral\}$$

5 Result & Discussion

The proposed pipeline starts with a face detection module, details of which has been discussed in the previous section. When a grid of faces (as expected from an online classroom environment) is passed to the face-detection module, the result obtained is shown in the Fig. 2.

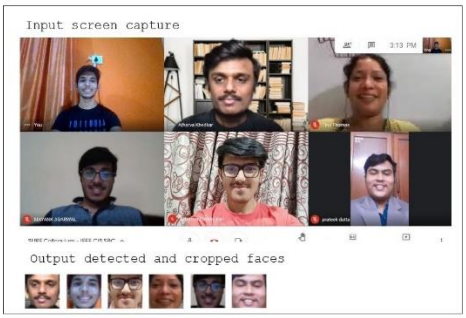

Fig. 2. Working of face detection and individual cropping tool.

Further, these images are passed to emotion detector model. The Fig. 3. shows the training and validation accuracy received for the proposed fin tuned model with ran on 10 epochs. The training accuracy appeared to roughly flatten (parallel to the x-axis) at an accuracy of 74% (approx) after the 4th epoch which marks the appropriate place to stop training to avoid overfitting. However, the validation accuracy continued to improve.

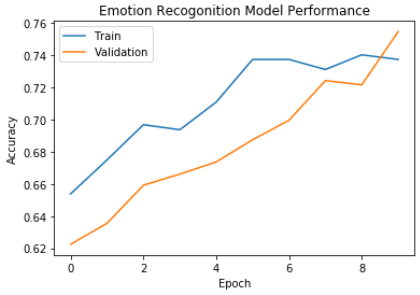

Fig. 3. Training and validation performance on 10 epochs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 May 2021 doi:10.20944/preprints202105.0303.v1

To investigate further, we considered running the model for 15 epochs to see a further movement. The Fig. 4 shows the training and validation accuracy for the same model when run for 15 epochs. In this case, the performance received was not very smooth. However, both the training and validation accuracy were still found to be roughly improving. To analyse better, we ran the model with the same configuration again (refer to figure). This time the performance found was relatively smoother, with flattening performance for both the training and validation curve at an averagely adjusted accuracy of 77.5%.

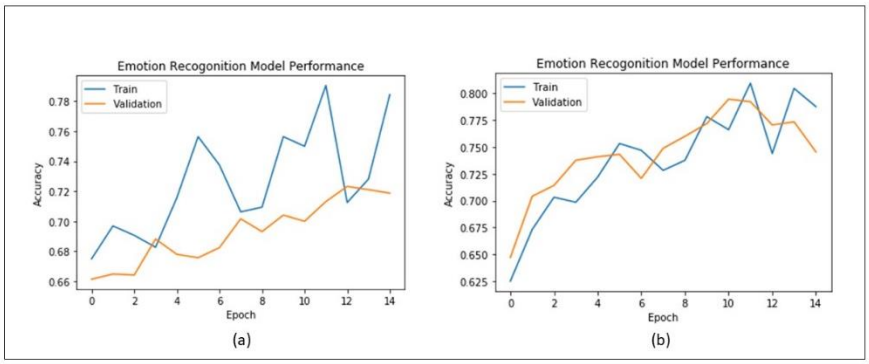

Fig. 4. Training and validation performance on 15 epochs (a) trial 1 (b) trial 2.

Overall, the best accuracy obtained for the model was 80% and 78% in training and validation, respectively. The reliable average accuracy received was 77.5% for both.

Finally, the detection results are used to find the final emotion index using the formula given in the section 5.

6 Conclusion

This paper proposes an architecture for analysis of student's engagement in video-based online classroom systems. The architecture starts with the detection of the faces of the student from their incoming video feed. Later, the detected faces are cropped individually and passed to the emotion detection model. The emotion detection model is a fine-tuned transfer learning model based on VGG16 as the pre-trained model. The reliable validation accuracy received for this task was found to be 77.5%. Later, these individual detected emotions are used to find an emotion index based on the number of students exhibiting a certain emotion. In future, we wish to extent the scope of the research, focusing on improving the accuracy and testing the architecture in real-time.

References

1. Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *Ieee Access*, *5*, 15991-16005.
2. Bahel, V., Bajaj, P., & Thomas, A. (2019, December). Knowledge Discovery in Educational Databases in Indian Educational System: A Case Study of GHRCE, Nagpur. In *2019*

International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 235-239). IEEE.

3. Raes, A., Vanneste, P., Pieters, M., Windey, I., Van Den Noortgate, W., & Depaepe, F. (2020). Learning and instruction in the hybrid virtual classroom: An investigation of students' engagement and the effect of quizzes. *Computers & Education*, *143*, 103682.
4. Weitze, Charlotte Lærke. "Pedagogical innovation in teacher teams: An organizational learning design model for continuous competence development." In *EXCEL 2015: The 14th European Conference on E-Learning*, pp. 629-638. Academic Conferences and Publishing International, 2015.
5. Ali N, Zafar B, Riaz F, Dar SH, Ratyal NI, Bajwa KB, Iqbal MK, Sajid M (2018) A hybrid geometric spatial image representation for scene classification. PLoS ONE 13(9):e0203339
6. Ali N, Zafar B, Iqbal MK, Sajid M, Younis MY, Dar SH, Mahmood MT, Lee IH (2019) Modeling global geometric spatial information for rotation invariant classification of satellite images. PLoS ONE 14:7
7. Ali N, Bajwa KB, Sablatnig R, Chatzichristofs SA, Iqbal Z, Rashid M, Habib HA (2016) A novel image retrieval based on visual words integration of SIFT and SURF. PLoS ONE 11(6):e0157428
8. Sajid M, Iqbal Ratyal N, Ali N, Zafar B, Dar SH, Mahmood MT, Joo YB (2019) The impact of asymmetric left and asymmetric right face images on accurate age estimation. Math Probl Eng 2019:1–10
9. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Personal Soc Psychol 17(2):124
10. Ninad Mehendale, Ninad's Research Lab, Thane, India, K. J. Somaiya College of Engineering, Mumbai, India, Facial Emotion Recognition using convolutional neural networks (FERC).
11. McDuff, D., et al. "AffectAura: An Intelligent System for Emotional Memory," in Proc. CHI. 2012
12. Marija Sablić, Ana Mirosavljević, and Alma Škugor. 2020. Video-Based Learning (VBL)—Past, Present, and Future: an Overview of the Research Published from 2008 to 2019. Technology, Knowledge, and Learning (07 Jul 2020). https://doi.org/10.1007/s10758-020-09455-5
13. R. Mayer and R.E. Mayer. 2005. The Cambridge Handbook of Multimedia Learning. Cambridge University Press
14. Allan Paivio. 1991. Dual coding theory: Retrospect and current status. Canadian Journal of Psychology 45, 3 (1991), 255.
15. Juho Kim, Philip J. Guo, Daniel T. Seaton, Piotr Mitros, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Understanding In-Video Dropouts and Interaction Peaks In-online Lecture Videos. In Proceedings of the First ACM Conference on Learning @ Scale Conference (Atlanta, Georgia, USA) (L@S '14). Association for Computing
16. G. Bradski and A. Kaehler, "Learning OpenCV: Computer Vision with the OpenCV Library". O'Reilly Press, 2008
17. Torrey, L., & Shavlik, J. (2010). Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques (pp. 242-264). IGI global.
18. Li, X., Grandvalet, Y., Davoine, F., Cheng, J., Cui, Y., Zhang, H., ... & Yang, M. H. (2020). Transfer learning in computer vision tasks: Remember where you come from. *Image and Vision Computing*, *93*, 103853.

19. Bahel, V., & Pillai, S. (2020). Detection of COVID-19 Using Chest Radiographs with Intelligent Deployment Architecture. In *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach* (pp. 117-130). Springer, Cham.
20. Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018, May). Vggface2: A dataset for recognizing faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (pp. 67-74). IEEE.
21. Sun Bo, Liu Yongna, Chen Jiubing, Luo Jihong and Zhang Di. (2015). Emotion Analysis Based on Facial Expression Recognition in Smart Learning Environment.Modern Distance Education Research, 2, 96-103.
22. Hutchinson, M. L., Antono, E., Gibbons, B. M., Paradiso, S., Ling, J., & Meredig, B. (2017). Overcoming data scarcity with transfer learning. *arXiv preprint arXiv:1711.05099*.
23. Conversational transfer learning for emotion recognition Devamanyu Hazarikaa, Soujanya Poria, Roger Zimmermann, Rada Mihalcea School of Computing, National University of Singapore, Singapore.Computer Science & Engineering, University of Michigan, USA, Information Systems Technology and Design, Singapore University of Technology and Design, Singapore. Received 28 November 2019, Revised 20 May 2020, Accepted 13 June 2020, Available online 1 July 2020.
24. Kentsch, S.; Lopez Caceres, M.L.; Serrano, D.; Roure, F.; Diez, Y. Computer Vision and Deep Learning Techniques for the Analysis of Drone-Acquired Forest Images, a Transfer Learning Study. *Remote Sens.* 2020, *12*, 1287. https://doi.org/10.3390/rs12081287
25. Henrys, Kasereka, Importance of Web Scraping in E-Commerce and E-Marketing (January 19, 2021). Available at SSRN: https://ssrn.com/abstract=3769593 or http://dx.doi.org/10.2139/ssrn.3769593
26. Nigam H., Biswas P. (2021) Web Scraping: From Tools to Related Legislation and Implementation Using Python. In: Raj J.S., Iliyasu A.M., Bestak R., Baig Z.A. (eds) Innovative Data Communication Technologies and Application. Lecture Notes on Data Engineering and Communications Technologies, vol 59. Springer, Singapore. https://doi.org/10.1007/978-981-15-9651-3_13.
27. VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION Karen Simonyan ∗ & Andrew Zisserman + Visual Geometry Group, Department of Engineering Science, University of Oxford {karen,az}@robots.ox.ac.uk
28. Bahel, Vedant, Shreyas Malewar, and Achamma Thomas. "Student Interest Group Prediction using Clustering Analysis: An EDM approach." In *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pp. 481-484. IEEE, 2021.