Attention-based Video Classification for Engagement Detection

Xiaoying Yang
xyang123@stanford.edu

Zihao Song
zs0226@stanford.edu

Shenghan Chen
csh3@stanford.edu

June 3, 2022

Abstract

Classifying engagement level of remote students is becoming more crucial as remote learning prevails. A variety of machine learning architectures have been applied on this topic, including different CNN models, transformer models, and spatialtemporal models. As baseline, our group ran ResNet models for engagement, boredom, and confusion classification. Taking temporal relationship between frames into account, our group implemented a spatio-temporal hybrid model, Res-TCAN, and a CNN-Transformer model. With the former, we introduced the attention mechanism to a state-of-the-art spatio-temporal model; With the latter, we adopted the idea from Vision Transformer(ViT) to complement CNN layers which produce a CNN-Transformer hybrid model to achieve better results with fewer computational cost. We applied data augmentation techniques and facial cropping for preprocessing. To test our models, we used 2 types of testing datasets. Trained Faces consists of testing instances in which some frames of the same person (other than the testing frames) are included in the training set, while Untrained Faces consists of unseen faces. Among models, the Res-TCAN achieved the best accuracy at 0.82 predicting Trained Faces within a smaller dataset, while the CNN-Transformer model achieved the best accuracy at 0.67 predicting Untrained Faces. In addition, our models perform well on classifying trained faces. However, if an unseen face appears, its inference accuracy is not good enough.

1 Introduction

A clear understanding of student engagement in class is crucial for effective instruction. It can provide real-time feedback so the instructor can adjust accordingly for better outcomes. The latest trend of remote learning, however, has posed challenges to such a feedback. Luckily, the use of computing devices in most online classes, coupled with recent breakthroughs in Computer Vision, entails a webcam-based, automated approach to the real-time tracking of student engagement.

We define our task as a video classification task where the input is video clips of student faces captured during online classes, and the output is the predicted engagement level for each clip. In this project, with DAiSEE dataset [4], we start with a ResNet model to perform engagement classification. We then explore a spatio-temporal hybrid model, Res-TCAN, that explicitly models the temporal dependencies with the help of the attention mechanism. In order to take the relationship between frames into account, we build a CNN-Transformer hybrid model to achieve better accuracy and efficiency.

2 Related Work

2.1 Diversified model combination of Engagement Learning

Extensive explorations have been performed on engagement classification problems with a combination of popular neural nets (CNN, DenseNet, Resnet) and computer vision techniques (SIFT, HOG, SVM)[2]. We applied some feature detection techniques like facial recognition in our project, which largely improved our model accuracy. CNN-based networks is a major direction of experiments, including state-of-art models ShuffleNet v2 model, ResNet v2 and Inception v3 models[7]. Experiments on other CNN-based models include all convolutional network (All-CNN), network-in-network (NiN-CNN), and very deep convolutional network (VD-CNN), which have simpler architectures and more efficient [9]. Our group decide to experiment residual network as our baseline model, inspired by an experimental application on disease detection[10].

2.2 Spatiotemporal models

[1] propose an end-to-end hybrid architecture with Residual Network (ResNet) and Temporal Convolutional Network (TCN) to model students' engagement level as a spatio-temporal classification problem. The ResNet extracts spatial features from consecutive frames; the TCN analyzes the temporal changes in consecutive frames and outputs the predicted class. To deal with the problem of highly imbalanced classes in the dataset, techniques such as weighted cross entropy loss function and a customized sampling strategy are employed, and higher accuracy is achieved in minority classes at the cost of more false alarms and reducing the overall accuracy of the classifier.

Inspired by the recent successes of transformers, we seek to explore the benefits of the attention mechanism added to this then start-of-the art architecture. Two other works provide such inspiration for adding the attention mechanism to Residual Network and Temporal Convolutional Network, respectively. The Residual Attention Network [12] enhances the original ResNet by inserting attention modules between residual units; the Temporal Convolutional Attention-based Network [5] adds to TCN temporal attention layers that connect units of each time step.

2.3 Transformer models

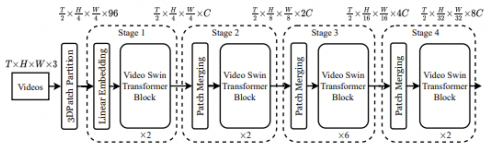

Figure 1: Architecture of Video Swin Transformer

[8] presents a pure-transformer backbone architecture for video recognition that is found to surpass the factorized models in efficiency. The approach implemented is a spatiotemporal adaptation of Swin Transformer. The overall architecture is shown in Figure 1.

A 3D Shifted Window based MSA Module is used instead of a global self-attention module which is unsuitable for video tasks as this would lead to enormous computation and memory costs. They followed Swin Transformer by introducing a locality inductive bias to the self-attention module, which is later shown to be effective for video recognition.

Inspired by this, we plan to build hybrid CNN-transformer model or even pure-transformer for comparison. This paper could offer detailed ideas and implementations including CNN-based model, hybrid model and pure-transformer model for video classification.

[3] provided another solution as Vision Transformer(ViT) that generates excellent results compared to state-of-the-art convolutional networks while the computational costs is fewer to perform training. The weakness is there still large gap for large-scale supervised pre-training. This pure transformer is self-attention-based architectures which are commonly used in natural language processing(NLP). The application of using vision transformer in computer vision is still limited. This paper evaluated the representation learning capabilities of ResNet, Vision Transformer, and the hybrid model by hand. Inspired by NLP successes, applying a standard Transformer that treat images as words would be a clever choice.

[11] presented an architecture, MLP-Mixer, which is based exclusively on multi-layer perceptrons(MLPs). MLP-Mixer attains competitive results compared to the state-of-art models when trained on large datasets. The weakness is it appears that Mixer benefits from growing dataset size even more than ViT which means small dataset may limit the performance of Mixer and large-scale dataset is always needed to achieve the state-of-art results. MLP-Mixer contains two types of layers which is simple to implement by hand. The approach here is clever to avoid complex architecture like CNN-like models and achieve comparable state-of-art performance.

3 Methods

3.1 Baseline Model

As baseline experiments, our group used ResNet-50 and ResNet-101 models [6] to build a deep CNN network, which both use bottleneck blocks with 1*1 convolutional block to reduce and increase image dimensions. The ResNet-50 model consists of 5 stages each with a convolution and identity block. Each convolution block has 3 convolution layers and each identity block also has 3 convolution layers. We also experimented with ResNet-101, which is slower but deeper.

3.2 ResNet-TCAN Hybrid Model

We propose an attention-based spatio-temporal model, ResNet-TCAN, that consists of a Residual Network and a Temporal Convolutional Attention-based Network, to explicitly capture the spatial and temporal dependencies, respectively (Figure 2). Specifically, we use a pre-trained ResNet-18 network

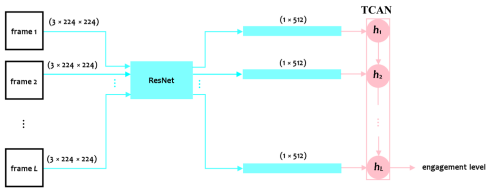

Figure 2: ResNet+TCAN Architecture. Modified from [1]

with trainable weights and the final fully-connected layer removed to extract spatial features from single frames; the spatial features are then fed into the TCAN network as the multi-dimensional input to the consecutive time steps; the output of the final time step is connected to a fully-connected layer and a softmax function, which predicts the engagement level. The TCAN network is formed by stacking several temporal blocks sequentially, where each block encapsulates repeated modules of 1d convolution, ReLU activation, and dropout layers, and is preceded with a multi-head self-attention module (Figure 3).

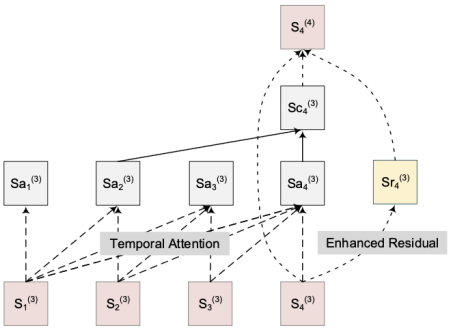

Figure 3: A TCAN Block[5]

The input to our ResNet-TCAN is a N x L × C × H × W tensor, where N, L, C, H, and W correspond to the mini-batch size, the number of frames, number of channels, frame height, and frame width, respectively. Based on the original ResNet-TCN implementation provided by [1], we have added the multi-head self-attention connection layer to the temporal blocks to further strengthen the communication of temporal information between timesteps; we have improved the connection between the residual network and the temporal network, mainly by casting the individual frame inputs across time steps as a reshaped batch input to the ResNet, which significantly reduced the number of model parameters as compared to the original ResNet-TCN model; we have also applied similar optimization techniques to the video data processing procedure.

3.3 CNN-Transformer Hybrid Model

Because we are using video clips as input, we want to take the temporal relationship between frames into consideration. This model is a hybrid architecture which adds transformer layers to complement CNN model. First, we have self-attention layers that from basic blocks of a Transformer which are order-agnostic. Videos are ordered as sequences of frames. We did positional encoding to let our transformer model to take into account the order information. We also added an embedding layer to embed the positions of the frames inside videos to the precomputed CNN feature maps. We also implemented a subclassed layer for transformer encoder. The CNN-Transformer architecture contains 3 convolution layers and 2 types of Transformer encoder layer. We use cross entropy loss as our loss function.

$$L = \frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic})$$

The overall sample hybrid architecture is shown in Figure 4. The detailed architecture information is shown in Figure 5 and 6.

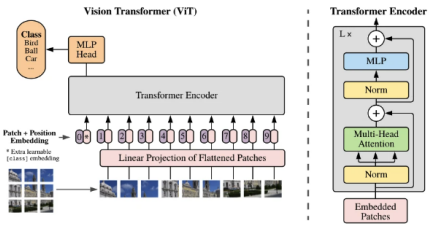

Figure 4: CNN-Transformer Hybrid Architecture

4 Dataset and Features

4.1 Dataset Description

Our model will be implemented based on a publicly available video engagement database called Dataset for Affective States in E-Environments (DAiSEE)[4], which contains 8,925 video snippets of 10 seconds (30 fps, 640 x 480) from 112 users with their engagement, frustration, boredom, and confusion level labeled from 0 (very low) to 3 (very high). We separated this dataset with 2 different ways: Untrained

Figure 5: CNN-Transformer Hybrid Architecture

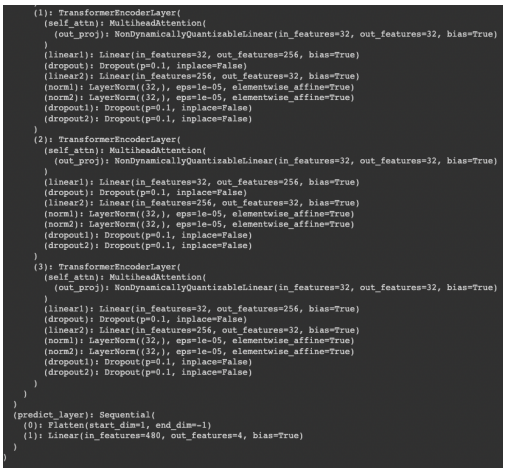

Figure 6: CNN-Transformer Hybrid Architecture

Faces split and Trained Faces split. First, we performed an Untrained Faces split, which seperates the dataset into a 5482 training set, a 1723 validation set, and a 1720 Untrained Faces test set, which contains different persons. In doing so, our model need to infer on untrained faces. Then, we shuffled the original dataset and generated a random split of 3: 1 for each person. With this Trained Faces split, our model can infer with faces of viewers who appeared in the training set.

4.2 Preprocessing

Our group extracted 15 frames from each clip. We also subtracted mean value from image, but we preserved pixel variation for data normalization.

4.2.1 Data Augmentation

Considering that class imbalance is a major problem in our dataset, we decided to augment classes with insufficient data points by randomly adjust brightness of frames, as shown in Figure 7. We augmented the data inversely to the class appearing probability, as shown in Table 2 and Figure 8. By doing so, our group aims at rebalancing the class distributions and mitigating overfitting.

Classes	Engagement Before	Engagement After	Boredom Before	Boredom After	Confusion Before	Confusion After
0	34	1232	2488	1269	3691	1274
1	214	1326	1763	1287	1301	1276
2	2649	1255	1075	437	1299	1306
3	2585	1351	156	1295	67	1276

Table 1: Imbalanced Class Resampling

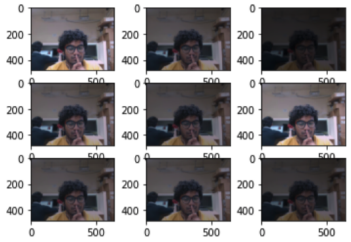

Figure 7: Data Augmentation with Random Brightness

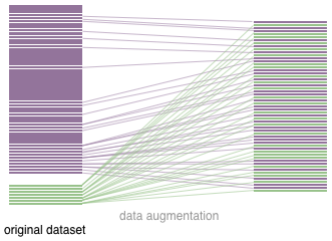

Figure 8: Imbalanced Data Sampler

4.2.2 Saliency Map and Facial Recognition

Emperically speaking, human-level classification of engagement level largely depends on facial expression of the view. Thus, our group decides to investigate whether eliminating background and posture of a viewer will improve the model accuracy. By plotting saliency maps like Figure 9, we discovered that some background settings also have high weights, for example office boxes and ceiling lamps. Although it might be a case of environmental impact on students' engagement level, our group thinks that focusing on facial expressions will make our model more reliable. We trained a facial detection model in order to crop a 144*144 detection box for each image as our new input. Results are shown in Figure 10

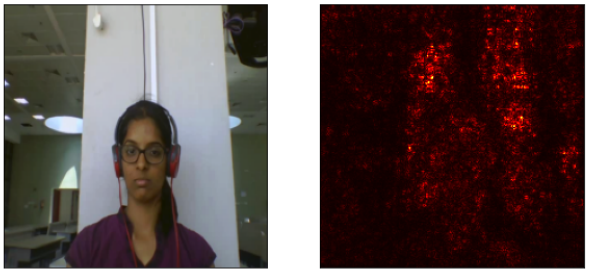

Figure 9: Saliency Map Based on ResNet-50

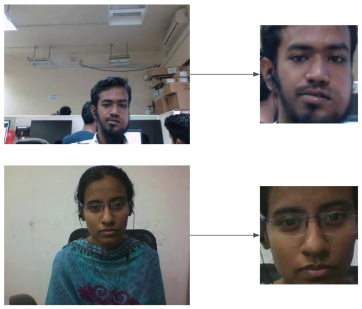

Figure 10: Facial Crop

5 Experiments

5.1 ResNet Models

With ResNet-50 and ResNet-101, we run 3 trials for the original data, balanced data, and facial crop data respectively. We also tried different hyper parameters in order to get the best model for test set. For evaluation, our group inspected the training/testing loss, confusion matrix, and accuracy on both Trained Faces testing dataset and Untrained Faces testing dataset. Hyper parameters of the model includes batch_size = 64, learning_rate = 0.002. Classification categories other than engagement is also inspected, including boredom and confusion.

5.2 Res-TCAN Model

Due to the prolonged hours required for training temporal models, we use a downsampled dataset for ResNet-TCAN. We create a medium-sized dataset out of DAiSEE by keeping all the minority class samples (with a label of engagement level 0 or 1) and randomly selecting the majority class samples (label of 2 or 3) with a probability of 0.2, resulting in a slightly more balanced dataset with 1284 training clips. We then perform a 3:1 training-test random split. The model is created with 3 attention heads and trained with a batch size of 32, an SGD optimizer with a learning rate of 0.001, and frame inputs resized to 224 x 224.

5.3 CNN-Transformer Model

With 3 convolution layers and two types of Transformer encoding layers. We run the model with CNN-Transformer only and CNN-Transformer incorporates with facial crop and data augmentation. Comparing with the results, we find CNN-Transformer model attains best result. We evaluated the results by training/test loss, confusion matrix and training/testing accuracy on both Trained Faces testing dataset and Untrained Faces testing dataset. Hyper parameters of the model includes batch_size = 64, learning_rate = 0.001.

5.4 Results

5.4.1 Model Accuracy Comparison

Model	Processing	Accuracy(Seen Faces)	Accuracy(Unseen Faces)
ResNet50	None	0.61	0.496
ResNet50	Facial Crop	0.56	0.514
ResNet50	Class Balancing	0.78	0.510
ResNet50	Facial Crop + Class Balancing	0.72	0.528
ResNet101	Class Balancing	0.81	0.458
ResNet101	Facial Crop + Class Balancing	0.74	0.480
Res-TCAN	None	0.821	0.49
CNNTransformer	None	0.71	0.67

Table 2: Comparison of model accuracy and testing loss

5.4.2 Qualitative Result

Figure 9: Sample of saliency map through ResNet-50

Table 2: Class distributions and Augmentation Result

5.4.3 ResNet-50 and ResNet-101 Performance

Figure 11: ResNet-50 Loss on Engagement.
Upper Left: No Processing; Lower Left: Class Balancing; Upper Right: Facial Crop; Lower Right: Facial Crop + Class Balancing

Figure 12: ResNet-50 ROC Curve for Trained Faces on Engagement.
Upper Left: No Processing; Lower Left: Class Balancing; Upper Right: Facial Crop; Lower Right: Facial Crop + Class Balancing

Figure 13: ResNet-50 ROC Curve of Boredom and Confusion Classification for Trained Faces.
Left: Boredom; Right: Confusion

Figure 14: ResNet-101 ROC Curve of Engagement for Trained Faces and Untrained Faces. Left: Trained Faces; Right: Untrained Faces

Figure 15: ResNet-101 Confusion Matrices of Engagement for Trained Faces and Untrained Faces. Left: Trained Faces; Right: Untrained Faces

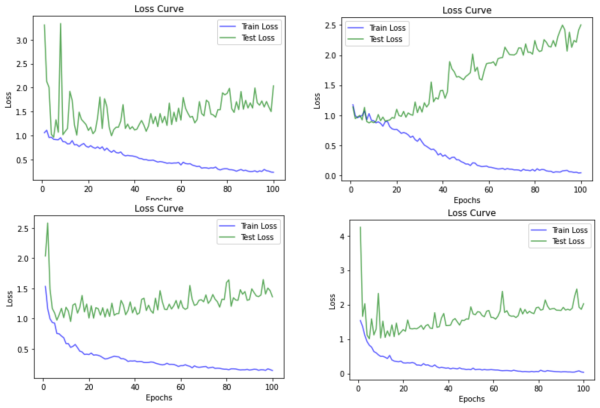

Figure 11: ResNet-50 Loss of Engagement

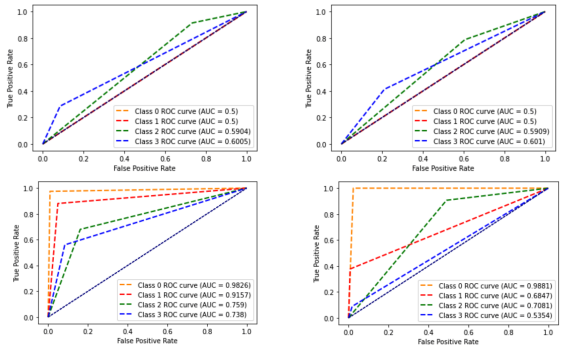

Figure 12: ResNet-50 ROC Curve of Engagement for Trained Faces

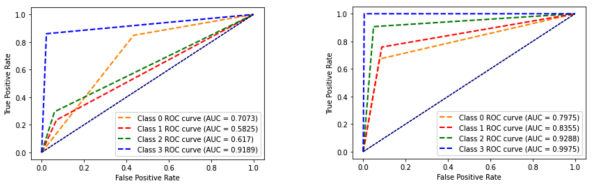

Figure 13: ResNet-50 ROC Curve of Boredom and Confusion Classification for Trained Faces

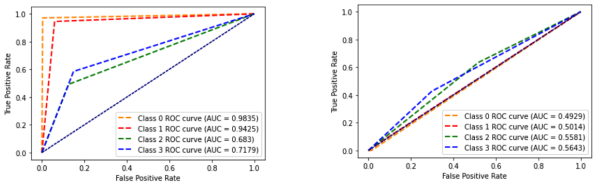

Figure 14: ResNet-101 ROC Curve of Engagement for Trained Faces and Untrained Faces

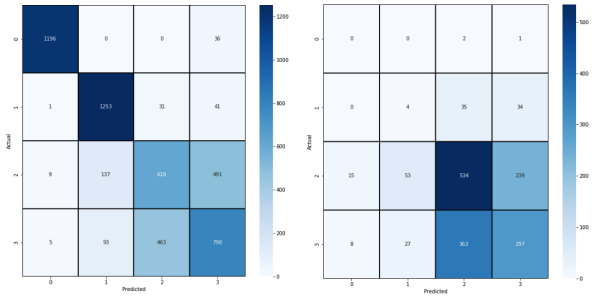

Figure 15: ResNet-101 Confusion Matrices of Engagement for Trained Faces and Untrained Faces

5.4.4 Res-TCAN Perfomance

Figure 16: Res-TCAN Loss Curve. To speed up training, model was evaluated on test set every 10 epochs.

Figure 17: Res-TCAN ROC Curve

Figure 18: Res-TCAN Confusion Matrix

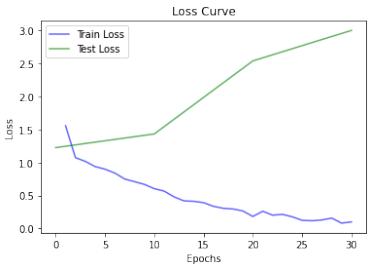

Figure 16: Res-TCAN Loss Curve. To speed up training, model was evaluated on test set every 10 epochs.

5.4.5 CNN-Transformer Performance

Figure 19: CNN-Transformer Loss

Figure 20: CNN-Transformer ROC Curve
Left: Trained Faces; Right: Untrained Faces

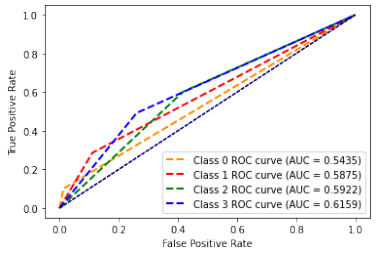

Figure 17: Res-TCAN ROC Curve

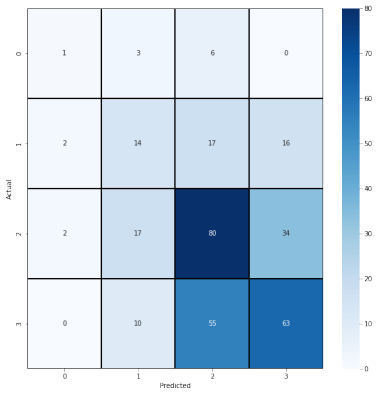

Figure 18: Res-TCAN Confusion Matrix

Figure 21:CNN-Transformer Confusion Matrix Left: Trained Faces; Right: Untrained Faces

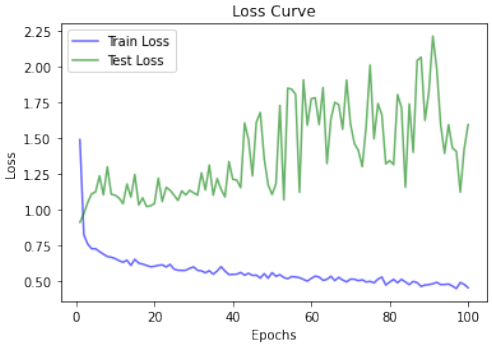

Figure 19: CNN-Transformer Loss

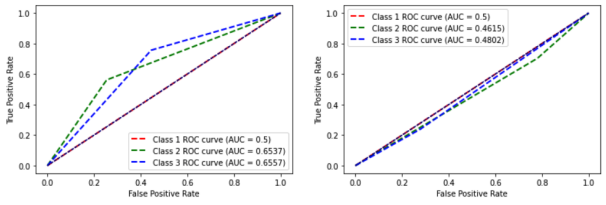

Figure 20: CNN-Transformer ROC Curve for Trained Faces and Untrained Faces

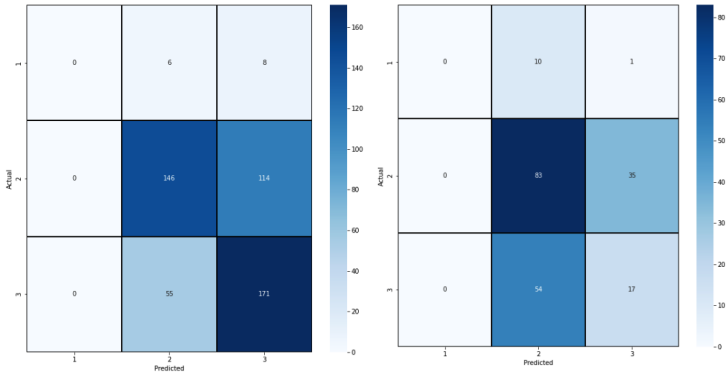

Figure 21: CNN-Transformer Confusion Matrix for Trained Faces and Untrained Faces

5.5 Discussion

5.5.1 Dataset

Because the saliency map shows that the model sometimes put high weights on the background, we decided to perform facial cropping. This improves the accuracy of the model by eliminating noisy background information and focus on the facial expression of viewers. Total training time also shows that facial cropping improve the efficient of learning because smaller size of images is used as input.

Applying class balancing is very crucial for improving class-wise accuracy for Trained Faces and avoiding overfitting, given that some classes have less than 0.1% of total instances. However, it does not improve the accuracy for Untrained Faces as much as we expected.

In addition, through qualitative analysis of our dataset, our figured that some frames have ambiguous classification. This is hard to avoid, because even for human-level labeling, it is hard to tell if a person is engaging or not. Another issue is a lot of clips share the same environment. This might be the reason why our model have high weights on environmental settings.

The difference of accuracy of models on Trained Faces testing dataset and Untrained Faces testing dataset is high. Despite that our group achieved a satisfying accuracy on seen faces, it performs bad on unseen faces. We think that the discrepancy between training data and unseen testing dataset is too big for our models to adapt to. However, given the application setting of our project, we should aim at achieving higher accuracy for unseen faces, because it is unlikely

to contain every students in the training data.

5.5.2 Models

ResNet-50 has a shorter training time and a higher accuracy for Trained Faces then ResNet-101. However, for Untrained Faces, ResNet-101 slightly outperforms ResNet-50. Trained Face inference is an easier task compared with Untrained Face inference, so a shallower model like ResNet-50 can perform well.

Res-TCAN has the highest accuracy(0.82) for Trained Faces. Temporal models are more difficult to train due to the explicitly modeled temporal dependencies, i.e. computation within the temporal blocks are performed sequentially. Our spatio-temporal hybrid model, Res-TCAN, was able to overfit a relatively small training set rather quickly (within 30 epochs). Generalization to unseen faces is again a challenge for this model, but it shows promises for a decent inference accuracy in the random training-test split setting if more computing resources would permit training on the full dataset for longer epochs.

Regarding CNN-Transformer model results, it has the best accuracy on Untrained Faces (0.67). Through the loss curve during the training process, we can know that the model is overfitting. This phenomenon may be caused by the small sample size of the dataset. If we can increase the number of samples in the dataset, the overfitting phenomenon will decrease or even disappear. Combining the ROC curve and the classification report, we can see that the CNNTransformer has poor robustness to the class imbalance problem. The classification performance of the class with a large number of class samples is very good. But for a class with a smaller number of samples, the model has poor performance without class balancing.

5.5.3 Training

By inspecting our loss plot, we can see that our model overfits at 20 epochs. Early stopping is helping our models to improve the accuracy.

6 Conclusion

This project experiments with different architectures for engagement classification of remote lecture viewers. Our proposed models show how temporal features help with inference accuracy. The CNN-Transformer model achieved the best performance for Untrained Faces, while Res-TCAN model achived the best performance for Trained Faces. For the Res-TCAN model, we introduced the attention mechanism to a state-of-the-art spatio-temporal model, Res-TCN, to better model temporal dependencies. For the CNN-Transformer model, we adopted the idea from pure transformer ViT to complement CNN architecture for better accuracy and fewer computational cost.

7 Future Work

Although facial cropping is effective, our group thinks that the posture of the viewer might contribute to the classification accuracy as well. Thus, foreground segmentation might further improve our model. Another possible direction of achieving higher accuracy for unseen testing dataset is to use more detailed segmentation like eye segmentation. Low engagement or high boredom might share some common features in eyes and mouth, which is more identifiable then the whole face analysis. In addition, from the saliency map, we can see that some environmental settings have high weights, like office background. In the ResNet, we performed facial cropping to neglect environment setting. However, the learning environment might impact the level of engagement, which will be another interest direction for future experiments.

If resource permits, we will also like to experiment more data augmenting techniques to help our model recognize unseen faces. We can also find or create a larger dataset with less imbalance issues to improve the robustness and overfitting issues. Both our proposed Res-TCAN and CNN-Transformer models will benefit from growing dataset size. A large and balanced dataset is crucial to the model performance.

In addition, for now, both our Res-TCAN and CNN-Transformer models only focused on engagement detection. Both are capable of being extended to explore more classification tasks like boredom, confusion, frustration, etc.

As for more architectures, we can explore a pure transformer model to compare against the CNN-Transformer hybrid architecture. We can also explore adding the attention mechanism to the spatial subpart of the hybrid model, and compare the performance of spatio-temporal hybird models with one or both parts attention-enhanced.

8 Appendices

Some additional quantitative performance analysis shown in Figure 22, Figure 23, Figure 24

	precision	recall	f1-score	support
0.0	0.93	1.00	0.97	1294
1.0	0.94	0.38	0.54	1296
2.0	0.38	0.91	0.54	1300
3.0	0.62	0.09	0.16	1274
accuracy			0.60	5164
macro avg	0.72	0.59	0.55	5164
weighted avg	0.72	0.60	0.55	5164

Figure 22: ResNet-50 Engagement Performance

	precision	recall	f1-score	support
0.0	0.74	0.67	0.70	1274
1.0	0.74	0.76	0.75	1276
2.0	0.86	0.91	0.88	1306
3.0	0.99	1.00	0.99	1308
accuracy			0.84	5164
macro avg	0.83	0.83	0.83	5164
weighted avg	0.83	0.84	0.83	5164

Figure 23: ResNet-50 Boredom Performance

9 Contributions

Zihao Song: Implemented CNN-Transformer model. Work together with all team members with data pre-processing, report write-up.
Xiaoying Yang: ResNet models, salience map, data augmentation and facial cropping, report write-up.
Shenghan Chen: implementated Res-TCAN model based on the original ResNet-TCN [1]; worked together with all team members with data pre-processing, report write-up.

References

[1] Ali Abedi and Shehroz S Khan. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. In *2021 18th Conference on Robots and Vision (CRV)*, pages 151–157. IEEE, 2021. 2, 3

[2] Sarthak Batra, Hewei Wang, Avishek Nag, Philippe Brodeur, Marianne Checkley, Annette Klinkert, and Soumyabrata Dev. Dmcnet: Diversified model combination network for understanding engagement from video screengrabs. *Systems and Soft Computing*, 4:200039, 2022. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[4] Abhay Gupta, Richik Jaiswal, Sagar Adhikari, and Vineeth Balasubramanian. DAISEE: dataset for affective states in e-learning environments. *CoRR*, abs/1609.01885, 2016. 1, 3

[1] https://github.com/abedicodes/ResNet-TCN

	precision	recall	f1-score	support
0.0	0.39	0.85	0.53	1269
1.0	0.53	0.24	0.33	1301
2.0	0.62	0.29	0.40	1299
3.0	0.93	0.86	0.89	1295
accuracy			0.56	5164
macro avg	0.62	0.56	0.54	5164
weighted avg	0.62	0.56	0.54	5164

Figure 24: ResNet-50 Confusion Performance

[5] Hongyan Hao, Yan Wang, Yudi Xia, Jian Zhao, and Furao Shen. Temporal convolutional attention-based network for sequence modeling. *arXiv preprint arXiv:2002.12530*, 2020. 2, 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 2

[7] Zeting Jiang and Kaicheng Zhu. Engagement detection in e-learning environments using convolutional ... 1

[8] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, Jun 2021. 2

[9] Mahbub Murshed, M. Ali Akber Dewan, Fuhua Lin, and Dunwei Wen. 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, 2019. 1

[10] Devvi Sarwinda, Radifa Hilya Paradisa, Alhadi Bustamam, and Pinkie Anggia. Deep learning in image classification using residual network (resnet) variants for detection of colorectal cancer. *Procedia Computer Science*, 179:423–431, 2021. 1

[11] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 2

[12] Baozhou Zhu, Peter Hofstee, Jinho Lee, and Zaid Al-Ars. An attention module for convolutional neural networks. In *International Conference on Artificial Neural Networks*, pages 167–178. Springer, 2021. 2