Attention-based Video Classification for Engagement Detection

Zihao Song,[1] Xiaoying Yang,[2] Shenghan Chen[3]

[1] Department of Civil and Environmental Engineering, Stanford University
[2]Department of Computer Science, Stanford University
[3]Graduate School of Education, Stanford University

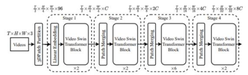
Stanford

Introduction

Attention-based Video Classification for Engagement

- Classifying engagement levels of students during remote learning

Prior works



- CNN-based models
  - DenseNet, ResNet, ShuffleNet v2, Inception v3
  - All-CNN, NiN-CNN, VD-CNN
- Spatiotemporal models
  - ResNet + Temporal Convolutional Network (Res-TCN)
  - Residual Attention Network
  - Temporal Convolutional Attention-based Network (TCAN)
- Transformer models
  - Transformer + 3D Shifted Window based MSA Module
  - Vision Transformer(ViT)
  - Multi-layer perceptrons(MLPs) mixer

Challenges

- Need the model to infer with unseen student faces
- Limited computational resources to train enough clips with temporal model
- Poor dataset human-level labeling quality

Significance of the Project

- Current tend of remote learning
- Instructors need real-time feedback from students
- Inspect how temporal features improve classification accuracy

Problem

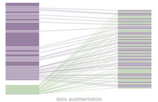
Problem Statement

- Input: Video clips of student faces captured during online classes
- Output: For each video clip: Predicted level of engagement, boredom, confusion
- Metrics: F1 score, accuracy, ROC curve, confusion matrix

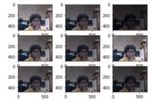
Dataset

- Dataset for Affective States in E-Environments (DAiSEE)
  - A publicly available video engagement database
  - Contains 8,925 video snippets of 10 seconds (30fps, 640 x 480) from 112 viewers
  - Categories: Engagement, Frustration, Boredom, Confusion
  - Classes: 0 (very low) to 3 (very high)
- Preprocessing
  - Data Augmentation
    - Inversely to class appearing probability
    - Augment by random brightness
  - Facial Crop
    - For each frame: detect facial area and crop into 144*144

Face Crop



Augmentation Algorithm



Random Brightness Aug

Methods & Experiments

Baseline Models

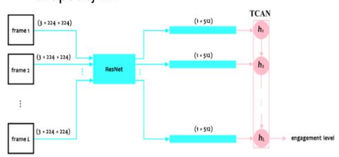
- ResNet-50 and ResNet-101
- 1*1 convolutional block
- Layer skips
- ReLU and batch normalization layer

Experiments

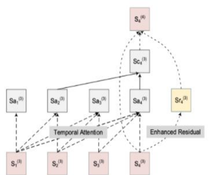
- Dataset Split
  - Trained Faces
    - Random split of video clips 3:1 for each person
    - Persons in test set appeared in train set
  - Untrained Faces
    - 5482 training, 1723 validation, 1720 test
    - Persons in test set never appeared in train set
- ResNet Models
  - ResNet-50 and ResNet-101
  - batch_size = 64, learning_rate = 0.002
- Res-TCAN Model
  - Downsampled dataset, 1284 clips
  - Kept all samples in minority classes
  - batch_size = 32, learning_rate = 0.001
- CNN-Transformer Model
  - 3 convolutional layers + 2 types of transformer encoding layers

ResNet-TCAN Hybrid Model

- Attention-based spatio-temporal model
- Pre-trained ResNet-18 + a Temporal Convolutional Attention-based Network (TCAN)
- TCAN block: 3-head self-attention + {conv1D - ReLU - dropout} x 2



ResNet+TCAN Architecture

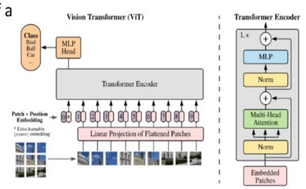


A TCAN Block

CNN-Transformer Hybrid Model

- Temporal relationship between frames
- Hybrid architecture: transformer layers + CNN model
  - Self attention layer: Order agonistic basic block of a transformer
  - Positional encoding: Take order information into account
  - Embedding layer: Embed positions of frames for CNN feature mapping
  - Subclassed layer: Transformer encoder
- Loss function: Cross Entropy

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



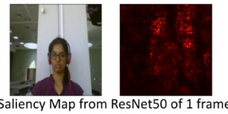
CNN-Transformer Hybrid Architecture

Results

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

Qualitative Results

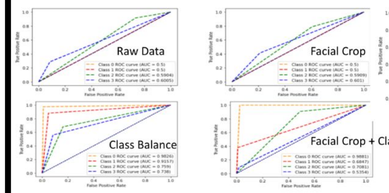


Saliency Map from ResNet50 of 1 frame

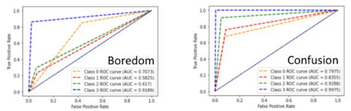
Classes	Engagement Before	Engagement After	Boredom Before	Boredom After	Confusion Before	Confusion After
0	34	1232	2488	1269	3691	1274
1	214	1326	1763	1287	1301	1276
2	2649	1255	1075	437	1299	1306
3	2585	1351	156	1295	67	1276

Number of Instances Before & After Augmentation

ResNet-50 with Trained Faces

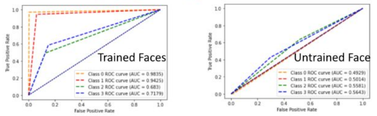


ResNet-50 ROC Curve of Engagement

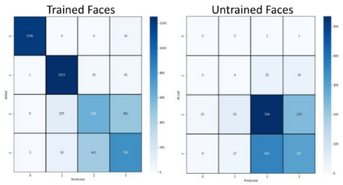


ResNet-50 ROC Curve of Boredom and Confusion

ResNet-101 (Engagement) with Trained & Untrained

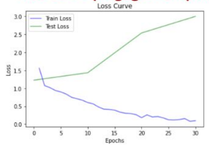


ResNet-101 ROC Curve of Engagement

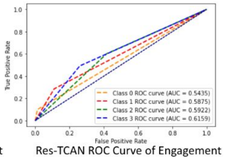


ResNet-101 Confusion Matrices of Engagement

Res-TCAN (Engagement) with Randomly Sampled Untrained

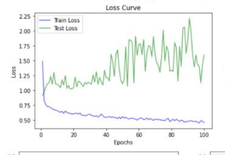


Res-TCAN Loss Curve of Engagement

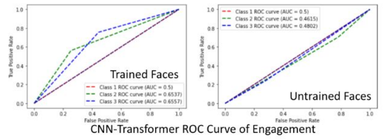


Res-TCAN ROC Curve of Engagement

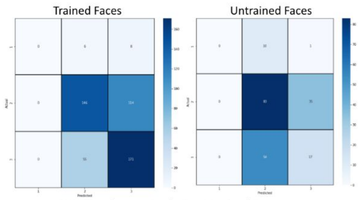
CNN-Transformer (Engagement) with Trained & Untrained



CNN-Transformer Loss of Engagement



CNN-Transformer ROC Curve of Engagement



CNN-Transformer Confusion Matrix of Engagement

Analysis

Dataset

- Saliency map: High weights occasionally on background settings
- Facial cropping: Eliminate background noise, focus on facial expression
- Class balancing: Effectively improved class-wise accuracy, avoid overfitting
- Label quality: Some human-lavel labels are ambiguous
- Video clips quality: Lots of videos with same background
- High accuracy on Trained Faces
- Low accuracy on Untrained Faces

Models

- ResNet
  - ResNet-50 is significantly faster than ResNet-101
  - ResNet-50 has higher accuracy on Untrained Faces (0.528)
  - ResNet-101 has higher accuracy on Trained Faces (0.81)
- Res-TCAN
  - Best accuracy for Trained Faces (0.82)
  - Overfits a medium-sized training set within 30 epochs
  - Struggles to generalize to unseen faces
- CNN-Transformer
  - Best accuracy for Untrained Faces (0.67)
  - Overfits due to small sample size
  - Poor robustness to the class imbalance problem
- Training
  - Easily overfits (usually at 30 epochs)
  - Early stopping is very effective

Conclusion

Main Findings

- Experimented with different architectures
- Proposed models: Temporal features
  - CNN-Transformer
    - Best accuracy for Untrained Faces (0.67)
    - Adopted idea from pure transformer ViT
    - Complement CNN architecture
    - Better accuracy and lower computational cost
  - Res-TCAN model
    - Best accuracy for Trained Faces (0.82)
    - Adds attention to Res-TCN
    - Better captures temporal dependencies

Future Work

- Better Accuracy for Untrained Faces
  - Foreground Segmentation: Taking posture into consideration
  - Background Analysis: How background environmental settings impact engagement level
  - Facial Segmentation: Analyze facial parts independently
- Model Performance
  - Larger dataset with less imbalance issues
  - Better data augmentation techniques
- Predict boredom and confusion with Res-TCAN and CNN-Transformer models
- More architectures
  - Pure transformer model
  - Adding attention to spatial subpart of Res-TCAN