Attention-based Video Classification for Engagement Detection

Zihao Song, Xiaoying Yang, Shenghan Chen

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

Introduction

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)
- Contains 6,925 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 144x144

Methods & Experiments

Baseline Models
- ResNet-50 and ResNet-101
- 1D convolutional block
- LR 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
    - S4GE 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

Results

ResNet-TCAN Hybrid Model
- Attention-based spatio-temporal model
- Pre-trained ResNet-18 + Temporal Attention-based Network (TCAN)
- TCAN block: 2 hidden self-attention + (conv1D - ReLU - dropout) x 2

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

Discussion

Dataset
- Salience 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-level 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-305 has higher accuracy on Trained Faces (0.83)
- Res-TCAN
  - Best accuracy for Trained Faces (0.81)
  - Overfits a medium-sized training set within 30 epochs
  - Strategies to generalize to small samples
- CNN-Transformer
  - Best accuracy for Untrained Faces (0.47)
  - Overfits due to small sample size
- Poor robustness to the class imbalance problem

Conclusion

Better accuracy for Trained Faces
- Augment features
- Better performance on small samples
- Better results for ResNet-305 and CNN-Transformer

Future Work
- Larger dataset with less imbalance issues
- Better data augmentation techniques