

Convolution Neural Network for Audio Visual based Personality Traits Analysis

July 15, 2016

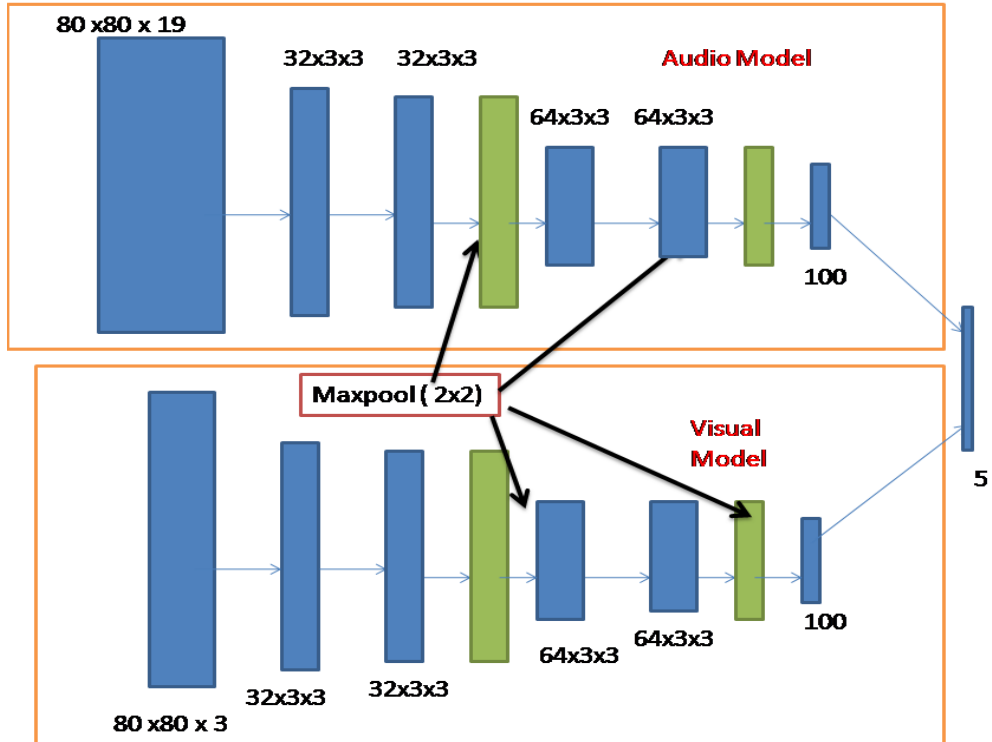
1 Team details

- Team name : Kaizoku
- Team leader name : Sonam Singh
- Team leader address, phone number and email:
D 203, VSRC, Indian Institute of Technology Kharagpur, WB, India-721302
Mob: +91-9717389804,
Email: sonamsingh19@gmail.com
- Rest of the team members : Akash Raheja
- Team website URL (if any)
- Affiliation: Indian Institute of Technology Kharagpur, WB, India

2 Contribution details

- Title of the contribution: Convolution Neural Network for Audio Visual based Personality Traits Analysis
- Final score :?
- Model Summary: We use identical Convolution Neural Network (CNN) based models for both modalities.
For Audio : We use MFCC features and represent an audio as a 3D tensor with 19 channels corresponding to MFCC levels.
For video: We take 10th frame from the video and pass it as RGB to the visual model.
Final model is concatenated features from both audio and video models (100 features each).
- References

- Representative image / diagram of the method :



- Describe data preprocessing techniques applied (if any):
All data has been normalized between 0-1 before feeding into model.

3 Multimodal Personality Trait recognition

3.1 Data Fusion Strategies

In this challenge, we have to analyze both audio and video for trait recognition.

3.1.1 Audio Feature Extraction

To extract audio features, we use MFCC coefficients with 19 levels. To keep the dimension same for all the samples, we take 6400×19 matrix and reshape it as 3D with $80 \times 80 \times 19$. This tensor is passed to a CNN model which gives 100 dimensional feature vector after series of max-pooling, LeakyReLU non-linearity layers.

3.1.2 Visual Feature Extraction

To extract video features, we use 10th frame only. In that sense, it is not representative of the whole video but it gave comparable performance with audio features. All images are resized to 80×80 with 3 RGB channels.

This tensor is passed to a CNN model which gives 100 dimensional feature vector after series of max-pooling, LeakyReLU non-linearity layers similar to audio features.

3.1.3 Fusion

Both 100 dimensional feature vectors from audio and visual CNN models are combined in late fusion strategy and finally a dense layer of 5 corresponding to traits is connected to both the CNN models.

3.1.4 Training

Final model is trained end to end with Mean Absolute error loss with Stochastic Gradient Descent with momentum.

3.2 Global Method Description

- Total method complexity: all stages
- Which pre-trained or external methods have been used (for any stage, if any) : None
- Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any) : None
- Qualitative advantages of the proposed solution: Fast execution on GPU for training and testing both.
- Results of the comparison to other approaches (if any)
- Novelty degree of the solution and if is has been previously published : Not yet assessed.

4 Other details

- Language and implementation details (including platform, memory, parallelization requirements):

Our source code is in python.

For audio, visual feature extraction, many python libraries have been used moviepy, ffmpeg, skimage, mfcc-features.

Except visual feature extraction, all other operations were done in embarrassingly parallel way with 30 processes in parallel. Final model is based

on Theano, keras (deep learning library) . We used k20 Nvidia GPU for execution for the final model.

- Human effort required for implementation, training and validation?: Majority of the effort is in extracting features. Implementation, training, and validation took minimum time due to heavy use of existing libraries.

Choosing what works and what didn't as with any machine learning competition became the challenge.

- Training/testing expended time? : Model execution for all the data (6000 training samples) is approx. 7 minutes with testing time in milliseconds per sample on K20 Nvidia GPU.

- General comments and impressions of the challenge? what do you expect from a new challenge in face and looking at people analysis?:

Challenge was nice and worth spending time. This was the first time we were doing audio and had little to no idea about extracting audio features. We started quite late in the competition and hope that such challenges are advertised more in popular mailing lists like COLT, UAI etc.