Deep Multi-Modal Regression for Apparent Personality Analysis

July 14, 2016

1 Team details

- Team name: NJU-LAMDA
- Team leader name: Chen-Lin Zhang
- Team leader address, phone number and email

  Address:
  National Key Laboratory for Novel Software Technology
  Nanjing University
  Nanjing 210023, China

  Phone Number: +86-187-6184-0176

  E-mail: zclnjucs@gmail.com
- Rest of the team members: Hao Zhang, Xiu-Shen Wei, Jianxin Wu
- Team website URL (if any): None.
- Affiliation

  National Key Laboratory for Novel Software Technology
  Nanjing University

2 Contribution details

- Title of the contribution: Deep multi-modal regression for apparent personality analysis
- Final score: We can achieve 0.914101 mean accuracy on validation set, which ranked the first place in the learning phase.
- General method description: Our proposed deep multi-modal regression (DM2R) system employs two modalities for capturing both visual and audio information.

In the visual modality, we firstly extract about one hundred images from each original video. Then, we adopt the pre-trained VGG-face model as the initialization of the convolutional layers in our visual models. What distinguish our models from VGG-face is: the original fully connected layers are discarded, and replaced by both average- and max-pooling followed the last convolutional layer (Pool_5). Please see our previous work [1] and [2] for more details. After that, the obtained two 512-d feature vectors are concatenated as the final representation, and then a regression (fc+sigmoid) layer is added for end-to-end training. For further improving the performance, the convolutional features of ReLU_{5_2} are also incorporated as another visual model. Additionally, the popular Residual Network is fine-tuned on these extracted images, which is the third deep model of the visual modality.

For the audio modality, the log filter bank (logfbank) feature are extracted from the original audio of each video. Based on the logfbank feature, we use the linear regression model to obtain the 5 Big Five Traits.

Finally, the two modalities are fused by averaging the scores of these deep visual models and audio model, which is the final predicted Big Five Traits.

- References

    1. Wei, X.-S., Xie, C.-W., Wu, J.: Mask-CNN: Localizing parts and selecting descriptors for fine-grained image recognition. In *arXiv:1605.06878*, pages 1-9, 2016.
    2. Wei, X.-S., Luo, J.-H., Wu, J.: Selective convolutional descriptor aggregation for fine-grained image retrieval. In *arXiv:1604.04994*, pages 1-16, 2016.
    3. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In *BMVC*, pages 1-12, 2015.

- Representative image / diagram of the method: Fig. 1 presents the pipeline of our proposed DM2R system.

- Describe data preprocessing techniques applied: For each video, we extract about 100 images without any preprocessing, and meanwhile extract the log filter bank feature from videos.

3 Personality Trait recognition from Visual data

3.1 Features / Data representation

As aforementioned, we train multiple end-to-end deep regression models for visual modality. The predicted scores of these models are used as the predictions of the visual modality.

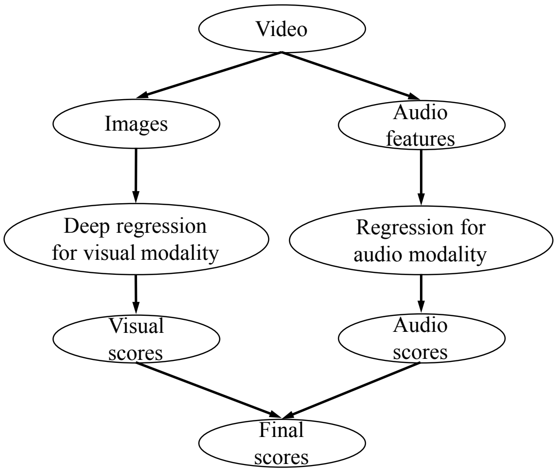
Video
Images
Audio features
Deep regression for visual modality
Regression for audio modality
Visual scores
Audio scores
Final scores

Figure 1: Pipeline of the proposed deep multi-modal regression system.

3.2 Compositional model

We use the VGG-face and ResNet models for initializing the convolutional weights. And then, we fine-tune these deep regression models on the extracted images from original videos. Note that, according to our previous studies [1] and [2], we improve the framework of the original VGG-16 model. The framework of the VGG network used in our DM2R system is shown in Fig. 2.

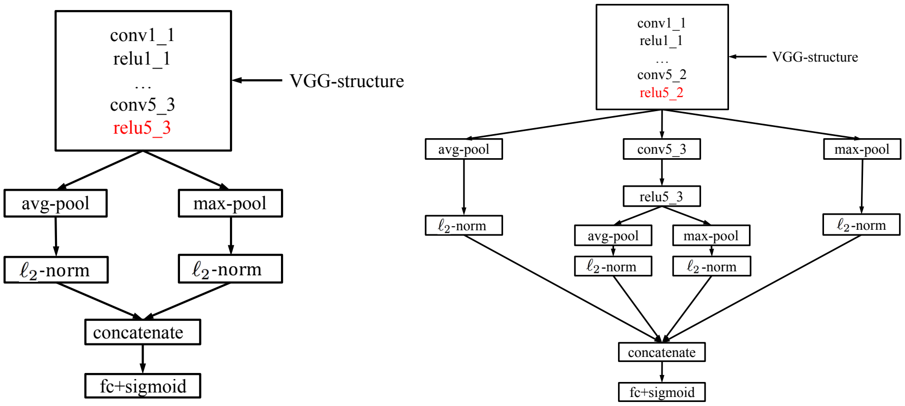

Figure 2: Framework of the VGG network used in the proposed deep multi-modal regression system.

3.3 Learning strategy

We use traditional stochastic gradient descent (SGD) algorithm for training all the end-to-end regression models. For these models, the learning rate is 10^{-3}. The number of training epochs is 2. The weight decay is 5×10^{-4}, and the momentum is 0.9.

3.4 Other techniques

Other technique/strategy used not included in previous items FOR VISUAL TRAIT RECOGNITION (if any): None.

3.5 Method complexity

It takes about 36 hours to finish the training of these visual models.

4 Personality Trait recognition from Audio data

4.1 Features / Data representation

For the audio data, we choose 44100-Hz for the sampling frequency, and $320k$bps for the audio output quality. We extract the log filter bank feature from the videos as the audio modality. After that, a model composed of a fully-connected layer followed by a sigmoid layer is employed to get the audio-based regression model. For the loss function, the ℓ_2 distance is used.

4.2 Learning strategy

We use mini-batch SGD with momentum of 0.9 to train the audio regression model. The batch size is 128. The learning rate is 8.3×10^{-4}. The weight decay is 6.5, and the learning rate decay is 1.01×10^{-6}.

4.3 Other techniques

Other technique/strategy used not included in previous items FOR AUDIO TRAIT RECOGNITION (if any): None.

4.4 Method complexity

It takes about 2.5 hours to finish the audio training.

5 Multimodal Personality Trait recognition

5.1 Data Fusion Strategies

None.

5.2 Global Method Description

- Total method complexity: Since we discard the fully connected layers of the traditional VGG models, the parameters of each deep regression models are not more than 14.72M, which is much less than the traditional VGG model (134.27M). In addition, for the audio modality, we only use the linear regression model which is also computationally efficient.

- Which pre-trained or external methods have been used (for any stage, if any): The pre-trained VGG-face model and the residual network model are used in our system.

- Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any): We did not use any additional data in the competition.

- Qualitative advantages of the proposed solution:

  1. It can reduce the parameters of deep models, and moreover improve the performance.
  2. It ensembles the multiple layers for further boosting the final regression mean accuracy.
  3. The proposed DM2R system fuses both visual and audio modalities for apparent personality analysis.

- Results of the comparison to other approaches (if any): Please refer to the [1] and [2] reference.

- Novelty degree of the solution and if is has been previously published: The novel deep model shown in Fig. 2 is the first one to aggregate the convolutional deep features by both average- and max-pooling. It not only reduce the parameters of the deep models, but also achieve the better regression performance than the traditional model with the fully connected layers. The solution is based on our previous studies [1] and [2]. They are not published, but has a preprint version in the arXiv server.

6 Other details

- Language and implementation details (including platform, memory, parallelization requirements): We used a Ubuntu 14.04 Server with 512GB memory and K80 Nvidia GPUs support. Our experiments were executed by several softwares: MATLAB R2014b, GPU with CUDA support, CUDNN, MatConvNet v1.0.20, LuaJIT, Torch7 (with cunn, cutorch), iTorch, python2.7, anaconda2, python_speech_features,[1] libxgboost, ipython and libffmpeg.

[1] https://github.com/jameslyons/python_speech_features

- Human effort required for implementation, training and validation? Beyond using the MATLAB code file to change the output prediction order and average the prediction scores, there is no manual effort required at all.

- Training/testing expended time? For training, one CNN model of the visual modality may take about 36∼40 hours in one K80 card. The audio regression model may take about 20 minutes to train. For testing, a CNN model may take about 2 hours and the audio may take 30 minutes.

- General comments and impressions of the challenge? what do you expect from a new challenge in face and looking at people analysis? Thank you for your efforts and providing such an opportunity for competition and communication.