ChaLearn LAP Real Versus Fake Expressed Emotion Challenge @ICCV 2017

July 2017

1 Team details

- Team name

  faceall_Xlabs

- Team leader name

  Jinchang Xu

- Team leader address, phone number and email

  Beijing University of Posts and Telecommunications, +86-18811315302, xjc1@bupt.edu.cn

- Rest of the team members

  None

- Team website URL (if any)

  None

- Affiliation

  Beijing University of Posts and Telecommunications

2 Contribution details

- Title of the contribution

  A combination of CNN and LSTM to recognize the authenticity of emotion from video

- Final score

  51.667

- General method description

  Firstly, we use a pretrained CNN network vgg16 on fer2013. Then, the vgg16 is treated as a feature extractor and use the fc7 4096 features. We extract the 128 frames per video as the feature of the video. Before using the lstm to train with all the features, we do a pca and change the final dimensions of features to 1024.

- References

  Guillaume Chevalier, LSTMs for Human Activity Recognition, 2016

- Representative image / diagram of the method

- Describe data preprocessing techniques applied (if any)

  Firstly, the videos are preprocessed to extract the frame from videos. Then the per image is resized using img_resize.m with a radio 0.5 . After that, we use face detection SDK to detect the face_rect and landmarks in per image and the results. We use the landmark points to align the faces and crop all the faces. Finally,we just select 128 frames from all the frames for per video. The 128 frames are considered as timestep of the LSTM network.

3 Recognition of fake and true emotions

3.1 Features / Data representation

Describe features used or data representation model (if any)

We extract the 128 frames per video as the feature of the video. And all of the images are via face detection, face alignment. After getting the aligned faces, we use the vgg16 model finetuned on fer2013 to extract the feature of aligned faces. The fc7 lay has 4096 dimensions. We apply the PCA to decrease the dimensions to 1024. So, the 1024-dimensions feature is the final dimension of one image.

3.2 Dimensionality reduction

Dimensionality reduction technique applied (if any)

PCA

3.3 Compositional model

Compositional model used, i.e. pictorial structure (if any)

VGG16, LSTM with 2 lstm layers and 128 time-step

3.4 Learning strategy

Learning strategy applied (if any)

AdamOptimize

3.5 Other techniques

Other technique/strategy used not included in previous items (if any)
None

4 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 training and validation data (at any stage, if any)

  None
- Qualitative advantages of the proposed solution

  By using vgg16 face model finetuned on fer2013, this can achieve high performance on emotional classifcations. Adopt 128-frame strategy for per video, we can transfer video to a 128-frame sequence. Then using LSTM network, this can learn well from a image sequence and predict the video labels better.
- Results of the comparison to other approaches (if any)

  None
- Novelty degree of the solution and if is has been previously published

5 Other details

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

  Caffe and Tensorflow 0.11.0rc0.
- Detailed list of prerequisites for compilation
- Human effort required for implementation, training and validation?

  None
- Training/testing expended time?

  3h/5 mins
- General comments and impressions of the challenge?

  This challenge can make us to apply deep learning konwledge to realistic applications.