TUBITAK UZAY-METU at ChaLearn LAP Real Versus Fake Expressed Emotion Challenge @ICCV 2017

Savas Ozkan, Gozde Bozdagi Akar

July 2017

1 Team details

- **Team Name:** TUBITAK UZAY-METU
- **Team Leader Name:** Savas Ozkan, Senior Researcher - PhD Student
- **Address, Phone Number, Email:** Middle East Technical University Campus, TUBITAK UZAY, 06800, Ankara, TURKEY. +903122101310-Ext:1407, savas.ozkan@tubitak.gov.tr
- **Team Members:** Savas Ozkan, Dr. Gozde Bozdagi Akar
- **Project Website:** https://github.com/savasozkan/real-fake-emotions
- **Affiliation:** TUBITAK UZAY (Image Processing Group) - Middle East Technical University (Multimedia Lab)

2 Contribution details

- **Contributions:** Our assumption to identify real-fake emotions from a video is mainly based on the observation explained in [1] and the authors show that brief emotional changes in eyes and mount movements can be distinct indicators for the problem. Based on this assumption, we propose a novel method that aggragates and models these emotional changes on the face parts as well as their sequential relations in an end-to-end learning manner.

  For this purpose, we focus to fuse two concepts, the robust video representation method [2] with the temporal model [3], to compute discriminative descriptors from small temporal windows (i.e. 150 ms) of the videos. To advance the performance, we use high-level convolution features (conv5) of Deep Convolutional Neural Network (CNN) that the model is finetuned

for the emotion detection [4]. Lastly, we pool these descriptors with Compact Bilinear Pooling (CBP) [5] to obtain one global representation for each video and a linear classifier is utilized to categorize them [6].

Our method makes two critical contributions to this domain. First, as claimed in [5], learning micro-emotional relations from the face parts and the temporal information in an end-to-end learning manner results superior performances to distinguish these real-fake emotions. Second, we propose a novel video representation method that can boost visual pooling with partially retained sequential information in the representation. Unlike the event detection [2], the temporal emotional characteristics are highly valuable in terms of the consistency [7] to categorize the human expressions properly. Therefore, our method achieves to comprise both spatial and temporal theoretical assumptions introduced for the problem in the model.

- **Final score:** Training Score 94.58%; Validation Score 68.33%; Test Score 65.00%;

- **General method description:** The overall method consists of two main stages, namely feature extraction and classification. In the first stage, we obtain robust micro-emotional visual descriptors for each emotion type that can be discriminative for the classification stage. For this purpose, we propose an end-to-end learning scheme which enables to fuse spatio-temporal pooling scheme [2] with temporal model [3] rather than the use of raw emotion features [4]. Later, these micro-emotional descriptors are pooled [5] by encoding them into a single representation for each video.

  Finally, the pooled representations are trained with a linear kernel Support Vector Machine (SVM) [6] in the classification stage.

- **References:**

  [1] M. Iwasaki and Y. Noguchi, "Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements.", in Scientific Reports, 2016.

  [2] Z Xu, Y Yang and A.G. Hauptmann, "A discriminative CNN video representation for event detection.", in IEEE CVPR, 2015.

  [3] S. Hochreiter and J. Schmidhuber, "Long short-term memory.", in Neural computation, 1997.

  [4] H.W. Ng et. al. "Deep learning for emotion recognition on small datasets using transfer learning.", in ACM ICMI, 2015.

  [5] Y. Gao et. al. "Compact bilinear pooling.", in IEEE CVPR, 2016.

  [6] https://www.csie.ntu.edu.tw/∼cjlin/libsvm/

  [7] L.C. Trutoiu et. al. "Spatial and temporal linearities in posed and spontaneous smiles.", in ACM Transactions on Applied Perception, 2014.

[8] D. Kingma and J. Ba, "Adam: A method for stochastic optimization.", in arXiv, 2014.

[9] Y Jia et. al. "Caffe: Convolutional architecture for fast feature embedding.", in ACM MM, 2014.

[10] M. Abadi et. al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems", in arXiv, 2016.

[11] http://dlib.net/

- **The diagram of the method:**

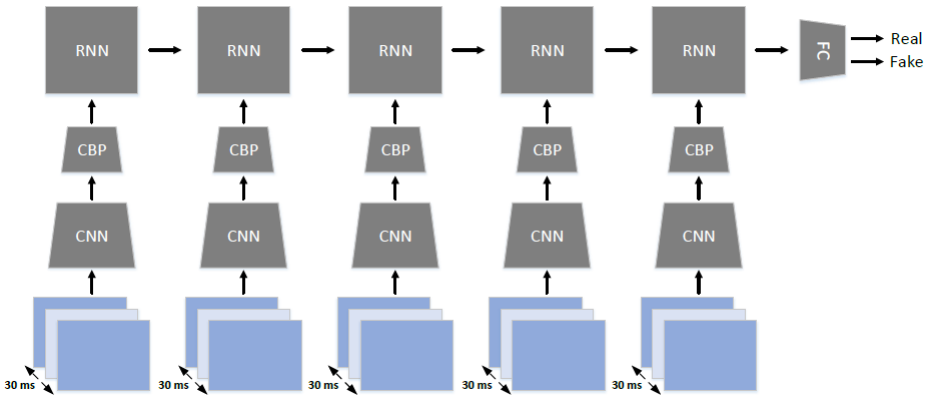

Figure 1: Learning of a micro-emotion descriptor from 150 ms temporal window of a video. Briefly, deep features (CNN) are encoded with CBP and fed into RNN to learn the robust descriptor. Later, the output of RNN is used as a descriptor in the classification step.

- **Describe data preprocessing techniques applied:** Two preprocessing steps are considered in our model. First, we applied a color mean normalization to each frame in order to center data for the CNN model as described in [4]. Also, we normalized these CNN features (conv5) in time for each person separately since we observed that the pretrained CNN model [4] still has an undesired bias for the similar faces. This step ultimately alleviates the oscillation in the learning phase.

3 Recognition of fake and true emotions

3.1 Features / Data representation

We initially compute high-level emotional features (conv5) by using a pretrained CNN model [4] for every video frames. Later, these features are encoded by the proposed spatio-temporal method and pooled for each video with [5].

3.2 Dimensionality reduction

We didn't use any dimension reduction method in our final model even though we have conducted several tests with the PCA reduced deep features. However, the obtained results are not promising compared to the original ones.

3.3 Learning strategy

For the learning step in the feature extraction, the trainable parameters in the model are optimized with gradient-based stochastic Adam [8] optimizer. Mini-batch size and learning rate are set to 64 and 0.001 respectively. Also, we set β_1 value in [8] to 0.7 to reduce oscillations. The number of training iterations varies from 50K to 200K depending on the loss variations between training and validation sets.

3.4 Fully-Connected (FC) Layer Normalization

We observed that normalization of the output of Recurrent Neural Network (RNN) with $l2$ norm yields more stable results for the micro-emotional descriptor learning stage.

3.5 Compact Bilinear Pooling (CBP) Normalization

4 Global Method Description

- **Total method complexity:** Face detection and emotion feature extraction steps consume most of the time in the method. Other steps such as feature learning and classifier training have relatively lower complexity. However, we should note that the feature learning stage is run on a NVIDIA Tesla K40 GPU card.
- **Which pre-trained or external methods have been used:** As stated, we exploited a pretrained CNN emotion model [4]. In addition, we used several open-source frameworks/libraries, namely Caffe [9], Tensorflow [10], Dlib [11], OpenCV.
- **Which additional data has been used in addition to the provided training and validation data:** No additional data is used.
- **Novelty degree of the solution and if is has been previously published:** To the best of our knowledge, our method makes critical contributions to this domain as stated in Section 2 in detail. Also, our method has not been published yet.

5 Other details

- **Language and implementation details:** Throughout this project, we used several open-source libraries/frameworks. Caffe, Tensorflow, Dlib, OpenCV, LibSVM as well as inner code infrastructure of TUBITAK UZAY can be the full list of these frameworks/libraries.
- **Detailed list of prerequisites for compilation:** The environment that we explained in the previous section should be supplied for the feature extraction stage. However, the classification stage needs only LibSVM and the precomputed descriptors from the feature extraction step.
- **Human effort required for implementation, training and validation?** Since we used different programming platforms such as C/C++ and Python, training/validation/test samples should be carefully organized in file system so as to work the method seamlessly.
- **Training/testing expended time?** The expected time in training/testing stages is quite low and they can be completed in few seconds.
- **General comments and impressions of the challenge?** I believe that the overall impressions about the challenge for our team are quite positive and everything is almost perfect.