EmoNets Analysis

The topics described below will cover the questions proposed in the exercise. I considered the paper a good starter to understand multimodal approaches and it demonstrates state of art implementations as well as points to most cited works involving this field.

Datasets

The main challenge dataset was the Acted Facial Expression in the Wild (AFEW) that contains shot video clips with the goal to classify each of them assigning to one of the seven emotion labels (angry, disgust, fear, happy, neutral, sad, surprise). The clips have approximately 1 to 2 seconds long and also contains audio track with audio voices and background music.

However, in order to explore different results and combine various models without overfitting issue, two different face datasets containing images labeled with emotions were used. The first one was the Google dataset with 35887 images whose grayscale 48 x 48 pixel version was chosen. The second one was the Toronto Face Dataset containing 4178 images labeled. Both datasets had their characteristics, so some methods were applied to reduce variation among subjects, lighting and poses.

An important observation is that the use of the extra datasets only applied to improve performance on deep neural network for face emotion classification since the training samples of original challenge dataset was not enough for deep learning approaches. The audio model was built with the AFEW dataset and combined with other models using specified experimental techniques.

Facetube extraction procedure

The video frames were extracted from competition dataset and face detection made using Google Picasa face detector. However, in some frames the Picasa could not detect faces and an alternative approach to search for the bounding box region of subsequent frames was used, named as “Facetubes”.

The method consists of searching spatial neighborhood for the bounding boxes matching histogram of color intensities. So, consecutive frames had bounding boxes extracted and associated to generate an approximate region for face recognition, also called facetubes for each subject in the video.

Bag of mouths features

In order to improve emotion recognition, a model that has mouth features was created based on cropped images around the mouth for each of them. This approach is based on the fact that a lot of emotions are based on actions using the mouth.

The method used to retrieve these features was to divide each image into 16 sections from which many 8x8 patches were extracted. After applying some transformations based on patches, for each of the 16 regions 400 centroids were found using k-means algorithm. Some calculations using Euclidean distance between patches were also applied generating final region descriptors with vector combination of 6400 dimensional features. As the number of features of each mouth image had this high amount, a "bag of mouths features" name is mentioned assimilating bag of words expression.

Experiments

The final models were used to predict two challenges (2013 and 2014 AFEW competitions) using different approaches for multimodal combination. First of all, the simple way of taking the average prediction of all 5 models achieved 40.15% in validation set. After that, some combinations were made in order to check for overfitting or similar models, per example the combination of only convolutional network model with audio model by average performed 39.90%.

A first combination technique used to mix audio and video models was the use of a learned SVM with RBF kernel with hyperparameters set accordingly to concatenated probability results of audio model and convolutional network model. This process yielded an accuracy of 42.17% in validation set and 38.46% in test set. So, the authors also tried to use all the 5 models as well as a MLP network to combine the results but this resulted in overfitting in training data.

Due to the dimensionality of hyperparameters and the analysis that different models had different performance across the emotion labels, the technique of random search through weights sampled from a uniform distribution and normalized was used performing 49.49% in validation set and achieving the best 2013 challenge result of 41.03% with the combination of all models. A similar approach was also used and performed 47.67% in 2014 challenge test set even with a different training dataset, showing the effective paper contribution.

Multimodal model

The paper describes different models combination used to form the best classification accuracy result. The models included are a convolutional neural network using deep learning, a deep neural network focused on audio stream, a K-means

consisted on bag-of-mouths features, a relational autoencoder that addresses spatio-temporal aspects of videos and mixed audio-video model.

We can consider the final work presented on this paper as multimodal since different approaches were mentioned with goal of achieve a better result by combining the models into a single prediction. In fact, this objective was done allowing the good final result in the competitions of 2013 and 2014.

Paper contributions

The work done by the authors changed the way of leading with video emotion analysis within AFEW challenge dataset because it brought combination of different models using different methods to improvement. The use of face image recognition and audio models were already being used by competitors but the autoencoder and bag of mouths models differed from the others.

The use of a specific model focused on mouth features allowed the final result to perform better mainly when predicting emotions related to the action of this region. Also, the autoencoder to address spatio-temporal aspects of videos brought a unexplored view to enhance the relation between frames.

However, all these models would not be effective trained without the use of combination by random search weights. All the methods mentioned by the authors in experimental results were essential to achieve the best position on the challenge by trying different validations.

Further development

The models developed by the authors already provide some advance techniques in emotion recognition using deep learning. However, some improvements could be tried in order to compare the results and maybe achieve a better result.

The first one is the automatic recognition of mouth keypoints for a good result before apply the “bag of mouths”. This method was mentioned in the paper but not applied. Since I believe this is a simple development technique, they could implement and test the new accuracy of this model.

Another important approach is the use of recurrent neural networks in order to capture temporal relations between frames. The authors used the autoencoder approach to do the same job but nowadays specially for deep learning, the RNN’s combined with CNN’s are the most common ones. The comparison between their model with a RNN model could assert the precision of the model or even increase the final accuracy.