A Novel Approach on Multimodal Emotion Recognition in Videos

Andrei Emanuel Galbenus
Department of Computer Science
Politehnica University of Bucharest
andrei.galbenus1305@stud.acs.upb.ro

Abstract

The present paper aims to present a novel approach on the task of automated emotion recognition of humans using visual, audio or textual information as input. In the introductory part, the problem and the objectives of the implementation to be presented will be stated. In the second section, the current state-of-the-art in the domain will be illustrated as a starting point of the research. Next, the details of the proposed architectures, as well as the data it operates on will be described and finally, the obtained results and comparisons against different methods will be showcased.

1 Introduction

Automated emotion recognition is an intensely researched and utilised field of study, being in continuous development. Despite this fact, the accuracy of the current solutions still represent a problem since the domain is in a sensitive, nascent research area. The capacity of the people to recognize emotions vary widely and can be extremely imprecise, since it relies on human-specific subjectivity, most of the times. This aspect makes the formalization of this process a quite difficult challenge. Thus, using an objective emotion recognition method, such as technology is necessary in this field in order to reach a sound and impartial result. Communicating using multimodal language, verbal or non-verbal, shares a significant portion of inter-human relationship and includes face-to-face communication, video chatting and social media opinion sharing. This project aims to recognize emotions automatically, with the help of deep learning neural network architectures using auditory, visual or textual information, transmitted on a verbal or non-verbal means of communication.

1.1 Problem

The general problem the current work tries to solve is represented by the removal of the human factor in the process of emotion recognition, by automatizing the workflow. The paper proposes an implementation of a system able to eliminate the issue of human subjectivity, which in most of the times leads to inaccurate and relative results, depending on the person judging. Thus, an objective method would be better suited in order to obtain correct and impartial results. In addition, an automated system would also reduce the costly matters sentiment analysis requires, the most important being time and constant financial investments.

1.2 Objectives

As mentioned before, the general objective of the presented document is to classify the emotions of any person, using different input information. The proposed model will be compared against a set of state-of-the-art architectures, using the same data set and a number of metrics like CCC and MSE score as a distance between the predicted outputs and a golden standard manually annotated.

One of the actual usages the project could be used for is automated feedback detection. Precisely, the emotional states a person shows, voluntarily or not, through facial expressions or spoken words and interpreted by the deep learning algorithm could be used to deduce the way people feel after they have done a certain action, like watching a movie. Moreover, using this automatically analysed feedback, further information could be extracted that would predict how a specific person would react to a future product or interaction for example. Another domain the project could be used in is the medical one, where psychologists would be able to detect the mental health of a person, or if the person is likely to be in a depressed state. Lastly, in the context of the actual pandemic, where we are forced to interact in online means more due to the social distancing, emotion recognition could actually become a necessity. From inter and intra company meetings, to online interactions between teachers and students, sentiment analysis can detect the quality of an emotional connection and the engagement of the participants.

2 Related work

The current section showcases a number of existing neural network architectures aiming to solve the task of emotion recognition. All of them are using multiple types of input to extract the most important features containing the necessary information to solve the classification task.

The first approach [1] uses only video and audio cues. For the visual part a Multitask Cascaded Convolutional Network (MTCNN) [2] is firstly used to detect and align faces, then the output goes through a pre-trained SphereFace [3] for feature detection. Finally it reaches a bidirectional LSTM followed by a pooling and fully-connected (FC) layer activated by the tanh function. As for the audio information, Short-Time Fourier Transform (STFT) is applied on 3 seconds long sequences, which are then sent to a pre-trained VGG-16 on ImageNet. In the end, two FC layers split by a dropout give the final result. The activation function is tanh again. The outputs of both pipelines are then concatenated into a vector which, after a FC layer offers the final classification.

Another architecture [4] uses 4 pipelines to process all 3 types of input. The first part deals with the visual information and applies OpenFace [5] and VGG-Face to extract both frame and facial landmarks, as eye gaze and head pose. The two sets of outputs are then concatenated and sent to an LSTM and a final FC layer. The second component uses the openSMILE [6] tool to extract different audio features, like Mel Frequency Cepstral Coefficients, loudness and pitch, which then go through a FC layer. The last part of the architecture processes the textual data using two lexicons: Bing Liu [7] and MPQA Subjectivity Lexicon [8]. The extracted features reach a FC layer outputting the final result. All the information is then combined in two FC layers which finally classifies the emotion with the help of the sigmoid activation function.

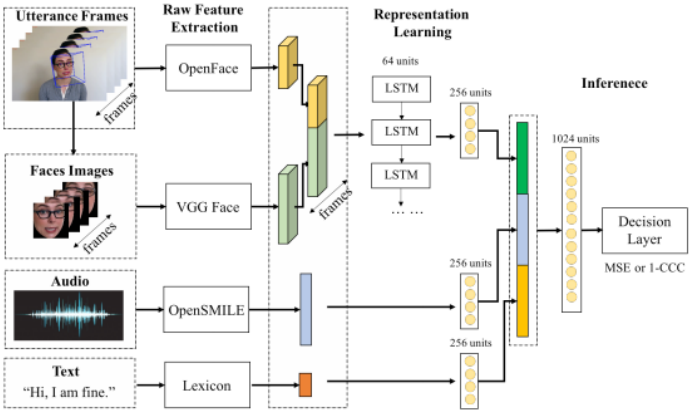

Figure 1: Model architecture

After analyzing a number of papers representing the state-of-the-art on the topic of emotion recognition it can be concluded that the general tendency is to use different pretrained architectures, which are

then fine-tuned on the specific data and then fuse of each pipelines and train them. Also, depending on the approach and available data, different types of manually cherry-picked features could be chosen.

3 Dataset

For the purpose of the current paper we will be using the same dataset as the models presented in the previous chapter. The One-Minute Gradual-Emotion Behavior Challenge Dataset [9], released in 2018, is composed of YouTube videos which are around a minute in length and are annotated taking into consideration a continuous emotional behavior. The videos were selected using a crawler technique that uses specific keywords based on long-term emotional behaviors such as monologues, auditions, dialogues and emotional scenes. The recordings contain features like facial expressions, language context and a reasonably noiseless environment. There are 420 videos, totaling around 10 hours of data, separated into clips based on utterances. Each utterance is annotated using an arousal/valence scale, as well as the categorical emotion: anger, disgust, fear, happiness, neutrality, sadness, surprise. There are 1790 utterances in the training set, 550 in the validation set and 1450 videos in the test set.

4 Proposed Methods

In the current section the proposed approach on the subject is proposed, in two sections, the first one dealing with the audio input and the respective architectures and the last one presenting the experiments for the visual information, with the purpose of finally being compared to the existing state of the art.

4.1 Audio pipeline

Each audio file was split into 3 second segments, since they variate from 0 to 47 in their raw form and sampled at 16kHz. In the next step a number of transformations were applied to them, the first one being the Short-Time Fourier Transform (STFT) to obtain the raw spectrograms, then Mel and Log Mel (LMS) spectrograms and Mel Frequency Cepstral Coefficients (MFCC). Window size was 2048, hop size 512 with 90 Mels and 40 MFCCs.

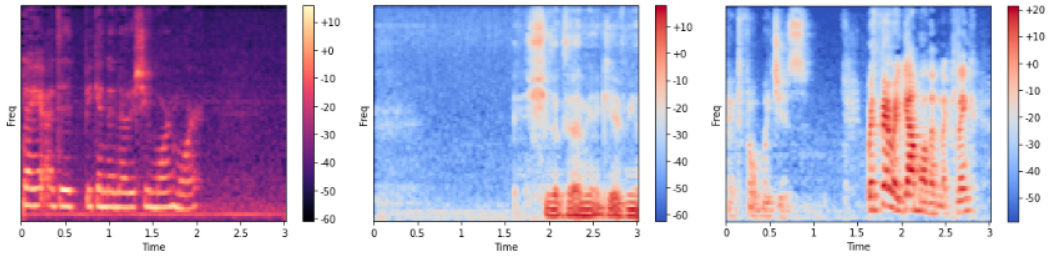

Figure 2: Log Mel Spectrograms (sadness/happyness/anger))

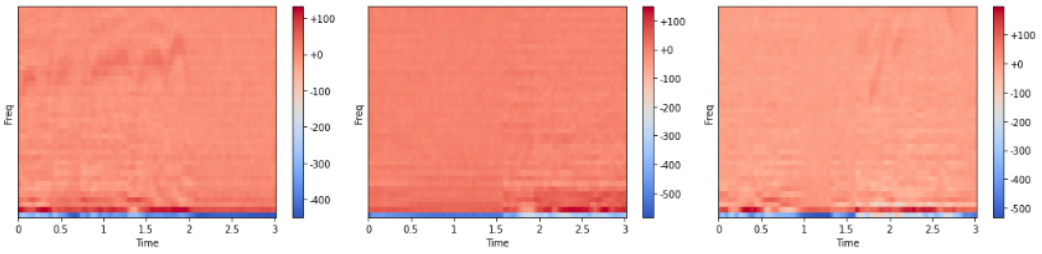

Figure 3: MFCCs (sadness/happyness/anger))

Following a number of models have been tested, from shallow to deeper ones.

The first one is composed of 5 convolutional layers with 5x5 kernels and a stride of 1x1, each followed by max pooling layers with the same kernel size. The feature maps are 32, 64, 128 for the first 3 layers, and 64 and 32 for the next two. ReLU activation is used for all of the convolutional layers. The network end with a classifying linear layer.

The second one uses both the raw signals and LMS as input. Initially, both of them go through a convolutional layer with kernel size of 16 and 40 feature maps, then a max pooling layer for each with a kernel size of 2 and 8 for a downsampling effect and ReLU. Next, the time signal goes through 2 convolutional layers with kernels of 512 and 1024 outputting 20 and 10 feature maps. Then the signals are reshaped to have the same sizes and added. Finally the reach a BiGRU layer with 128 units that tries to detect temporal connections and the classification layer. All the convolutional layers are followed by batch normalization and dropout layers.

The last group of models are the deep architectures, namely AlexNet, VGG16 with batch normalization and ResNet50, for which transfer learning has been applied. For the pretrained versions, the first layer was modified to accept bidimensional input and the stride was eliminated, since the input is smaller than 224x224, specifically 90x130 for LMS and 40x130 for MFCC. The second approach in this case was to retrain the models from scratch.

4.2 Visual pipeline

In the first stage 16 frames per utterance have been extracted, then then the faces have been aligned and cropped using OpenFace 2.0 [**?**], resulting in 112x112x3 images. The last frame is repeated in case the clip is shorter.

Next a series of deep pretrained convolutional networks have been applied for feature extraction. The first two models are based on VGGFace which is a framework for facial recognition, specifically VGG16 and ResNet50. For the first one, the classifier was removed, resulting in 7x7x512 feature maps for each extracted frame. Then either a flattening operation is applied, or a mapping linear layer with 4096 units for dimensionality reduction or 4 convolutional layers with kernels of 2xM, 3xM, 4xM and 5xM, where M is the number of previously resulted features, and 128 feature maps. Each of them are followed by a batch normalization layer and a max pooling one. Finally they are concatenated. For ResNet50 the output has 2048 features, from the last average pooling layer. For both of the models two BiRNN, BiGRU or BiLSTM layers with 256 units and dropout 0.3 have been applied, to extract the temporal information between the frames. Finally the classification layer has been applied. The last model experimented with is DenseNet121, resulting in 1024 features for each of the frames, followed by the same 3 types of recurrent networks, except this time the model isn't available in the VGGFace framework, so the ImageNet pretrained version was used instead.

5 Experiments

For all the networks the classifying activation function was tanh for valence and sigmoid for arousal. SGD was proved to offer the best results in comparison to Adam, with a learning rate between 1e-3 and 1e-6, momentum of 0.9 and Nesterov. Batch size ranges between 8 for models more difficult to

train to 32 samples for the faster ones. The best and most simple to use by the networks audio data format was LMS. The dataset elements are shuffled, no early stopping was used, the training was supervised manually and the maximum number or epochs 700. The double input audio model and DenseNet121 were the most difficult to train for each pipeline, with about an hour per epoch. CCC and MSE were both used as metrics as well as loss functions for training the models.

The librosa library was used for audio data processing, in Python 3.7 and PyTorch 1.9.0. For the audio pipeline, the implementation was done in Google Colab on a 12GB Tesla K80 and 12GB of RAM, as for the visual input experiments, the local environment was used, the data being to large, on a 4GB Nvidia GTX 960M and 16GB of RAM. One of the biggest problems was the size of the data for both of the pipelines which filled the capacity of memory, thus the Hickle library was used to store the data, which compresses the data and is optimized for this type of task. Also, the local GPU reached its limits while experimenting with the 4 convolutions model and DenseNet121, rebooting the system being the only option here. Also for this reason, no attempt to fuse the models and train them in the same time was done.

6 Results

The following tables provide the final results of the pipelines as well as a comparison with the state-of-the-art models that participated in the competition, on the final test set.

Table 1: Audio models results comparison

Model	Valence(MSE)	Arousal(MSE)	Valence(CCC)	Arousal(CCC)
5Conv2d	0.132	0.043	0.122	0.211
ResNet50(P)	0.130	0.060	0.205	0.251
ResNet50(NP)	0.056	0.065	0.210	0.194
VGG16(P)	0.047	0.042	0.195	0.285
VGG16(NP)	0.133	0.040	0.251	0.078
AlexNet(P)	0.140	0.041	0.152	0.159
AlexNet(NP)	0.127	0.040	0.077	0.248
Conv1D + 2D	0.153	0.041	0.041	0.199

Table 2: Visual models results comparison

Model	Valence(MSE)	Arousal(MSE)	Valence(CCC)	Arousal(CCC)
ResNet50+RNN	0.145	0.042	0.322	0.112
ResNet50+GRU	0.126	0.045	0.446	0.180
ResNet50+LSTM	0.114	0.037	0.325	0.249
VGG16+RNN	0.098	0.044	0.303	0.194
VGG16+GRU	0.103	0.043	0.254	0.274
VGG16+LSTM	0.126	0.043	0.347	0.159
DenseNet121+RNN	0.142	0.145	0.133	0.083
DenseNet121+GRU	0.158	0.040	0.052	0.063
DenseNet121+LSTM	0.114	0.041	0.021	0.098
ResNet50+4Conv	0.154	0.043	0.250	0.188
VGG16+4Conv	0.125	0.044	0.345	0.113

Table 3: Models results comparison

Model	Arousal(CCC)	Valence(CCC)
GammaLab (VA)	0.361	0.498
Audeering (VA)	0.286	0.368
HKUST (VAT)	0.276	0.359
UMONS (VAT)	0.175	0.262
ADSC (VA)	0.236	0.442
ADSC (V)	0.244	0.437
Audeering (A)	0.292	0.361
iBug (V)	0.130	0.400
Baseline (A)	0.08	0.10
BaseLine (V)	0.12	0.23
VGG16(P)(A)	0.285	0.195
ResNet50(P)(A)	0.251	0.205
VGG16(NP)(A)	0.078	0.251
VGG16+GRU (V)	0.274	0.254
ResNet50+GRU (V)	0.180	0.446

As can be observed from the above tables, the best proposed audio model for arousal was VGG16 with batch normalization pretrained on the third place overall and second for audio unimodals and the same model but retrained this time for valence, on the eleventh place overall and second for audio models only. As for the visual pipeline, for arousal, the best model was VGG16+GRU on the 5th place overall and first place for video models and ResNet50+GRU for valence, placing on second overall and first for video models only.

7 Conclusion

In the current work we tested and analyzed a series of audio and video models on the task of emotion recognition. We found out the the best proposed models for both cues are the deep neural networks. We also observed that GRU cells performed better on small datasets like the one used in the experiments. Thus, the best proposed model on valence was ResNet50+GRU while the best for arousal was the pretrained VGG16. To improve the current scores of the models the best improvement would be the fusion of the best architectures.

References

[1] Songyou Peng, Le Zhang, Yutong Ban, Meng Fang, Stefan Winkler (2019) A Deep Network for Arousal-Valence Emotion Prediction with Acoustic-Visual Cues

[2] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao (2016) Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks

[3] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, Le Song (2018) SphereFace: Deep Hypersphere Embedding for Face Recognition

[4] Didan Deng, Yuqian Zhou, Jimin Pi, Bertram E.Shi (2018) Multimodal Utterance-level Affect Analysis using Visual, Audio and Text Features

[5] T. Baltrusaitis, P. Robinson, and L. P. Morency (2016) Openface: an open source facial behavior analysis toolkit. In Applications of Computer Vision (WACV), 2016

[6] Florian Eyben, Martin Wöllmer, Björn Schuller (2010) Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 9th ACM International Conference on Multimedia, MM 2010

[7] X. Ding, B. Liu, P. S. Yu (2008) A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference on web search and data mining

[8] T. Wilson, J. Wiebe, and P. Hoffmann (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics

[9] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland and S. Wermter (2018) The OMG-Emotion Behavior Dataset. In 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro