Computer Vision in Presentation Analysis

Name: Mukhit Ismailov

Student ID: 20565331

Supervisor: Qu, Huamin

Abstract

Effective presentation is the essential skill in our fast-developing world. Although, a number of techniques are available to improve argumentation and logical side of a presentation, very few techniques deal with facial expressions side of the message. But the facial expressions are very important: the message might be perceived differently depending on your facial expression. In this paper, a full pipeline of detecting facial expressions during a presentation is shown. Both face detection and emotion recognition tasks are done using Deep Learning models. After emotions are detected, supporting graphs and insights are suggested to get as much as possible from the given information.

1. Introduction

Communication skills have always played a huge role in human interactions. Whether it is a work in a team, a leisure chat with your friends or a potential sales pitch. All of these require a good level of this skill. One part of this is related to logic and evidence to support your arguments. Although, it is not an easy task to master this aspect of communication, it can be developed through practice [5]. However, a more difficult aspect of an effective communication is facial expressions during the talk. They allow the talk to be more memorable and trustworthy. For example, if you are telling a sad story, your face should express sadness. Unlike, argumentation part of the speech, expressions cannot easily be practiced. But when a person sees himself presenting and gets a visualization of what kinds of expressions he had during the speech, he can make improvements by adjusting his style of speech.

This project tries to address this problem. The process is divided into four parts: 1) Frames extraction 2) face detection 3) emotion recognition 4) Infographics.

The report has the following structure: Section 2 gives background information related to deep-learning models. Section 3 describes the whole pipeline in detail. Section 4 talks about the limitations of the models. Section 5 discusses results and shows infographics. Section 6 is the conclusion.

The video used in this project is a famous TED talk given by James Veitch on spam replies [3].

2. Background

In this sections I would be talking about what architectures are used to detect faces and recognize emotions

2.1. Face detection using Haar Cascades

Face detection techniques have been developed long before Neural Networks (NN) became popular. Before NN, there were techniques of using special filters to detect faces [4].

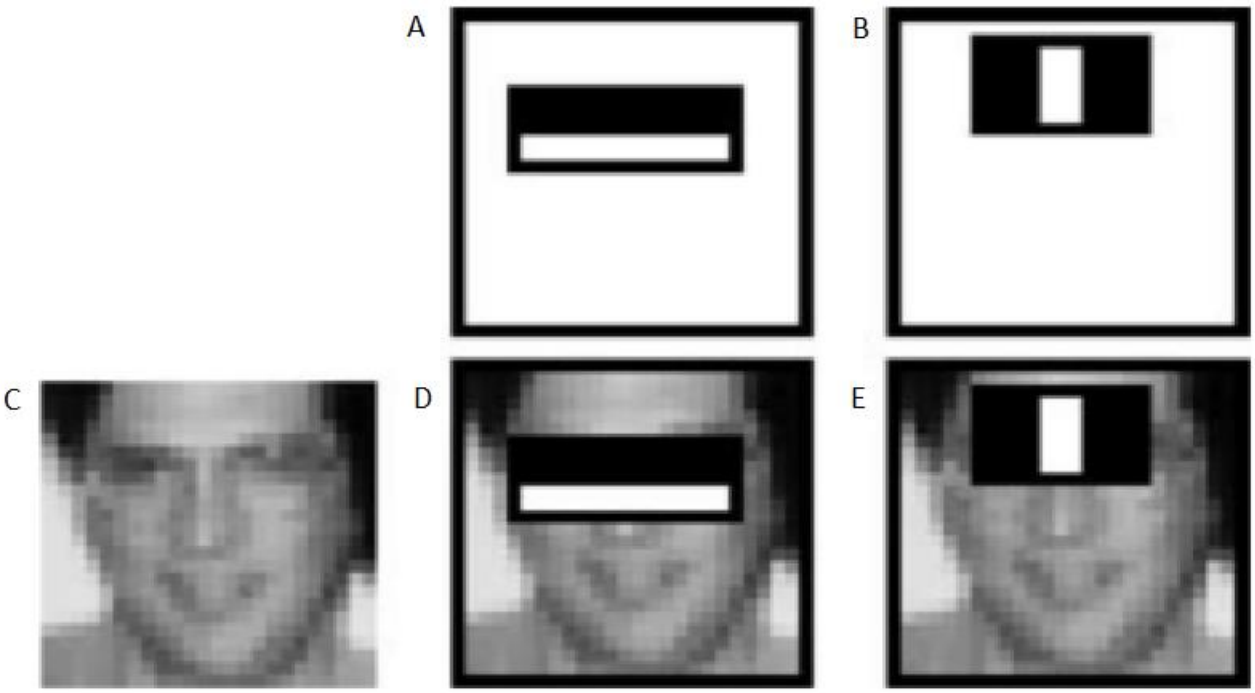

Figure 1: Face detection technique using filters

In particular, two filters are used. In the image above filter A uses the property that pixel region near eyes are darker than that around nose and cheeks. Filter B uses relies on the property of the nose being lighter than

eyes. By applying these filters across all the sub-regions and using Adaboost algorithm to select the sub-region containing a face.

The results look like this:

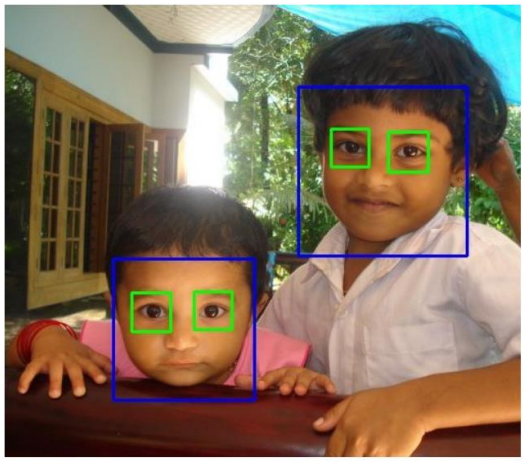

Figure 2: Sample results of using filters to detect objects: eyes and face

It is important to understand the filters and how faces were detected before Neural networks became popular. The reason because Neural networks use convolutional layers, that behave in similar fashion to filters.

2.2. Face detection using Neural Network

After neural networks gained popularity different solutions have been developed to detect faces. The one that was used in this project is based on a Single-Shot-Multibox detector (SSD) [1]. This detector can detect any object. For our purposes, we just need to tell that we are looking for a "face" object. This technique uses VGG 16 architecture in training and detecting process. This model has been trained on open source images.

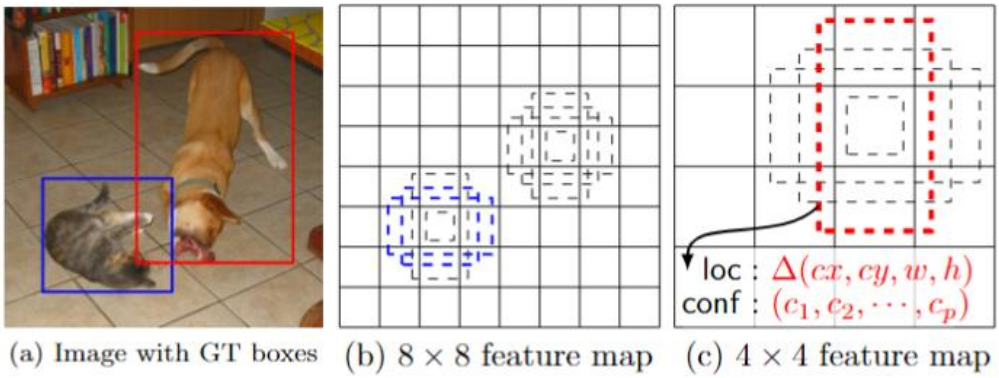

Figure 3: Intuition of how a Single-Shot-Multibox detector wokrs

SSD only needs two things to detect any type of object. 1) is the input image. 2) are the ground truth bounding boxes of those objects. During the training it uses VGG 16 as a base because it is quite fast to train since it is relatively shallow. As can be seen from the image below, after convolution layers we add fully-connected layers to extract complex information about the object.

Figure 4: a VGG 16 architecture

2.3. Emotion recognition using neural network

Emotion recognition uses similar technique to face detection. To recognize emotions attentional convolutional network is used [2]. Depending on a number (N) of emotions you want to recognize, this problem becomes a multiclass classification with N classes. Since there are not so many classes a shallower version of neural networks are used. To train the network a database of real people and FERG database (stylized characters with annotated facial expressions database) are used.

Figure 5: images of a real person

Figure 6: sample images from FERG database

Important thing to note that FERG images have a clear distinction in emotions. And that helps the model to distinguish emotions better because these images are easily distinctive from each other.

3. Pipeline

The following is the detailed pipeline of this project:

1) Extract frames from a video
2) Detect a face on each frame and select only those frames where a presenter is on
3) Detect emotion of a presenter
4) Aggregate data and provide insights and infographics

3.1. Frame Extraction

Extracting frames from a video is a pretty straightforward task. There is a choice of how many frames per second you want. I chose 2 frames per second because I tried taking more but the emotion of a face does not seem to change that fast.

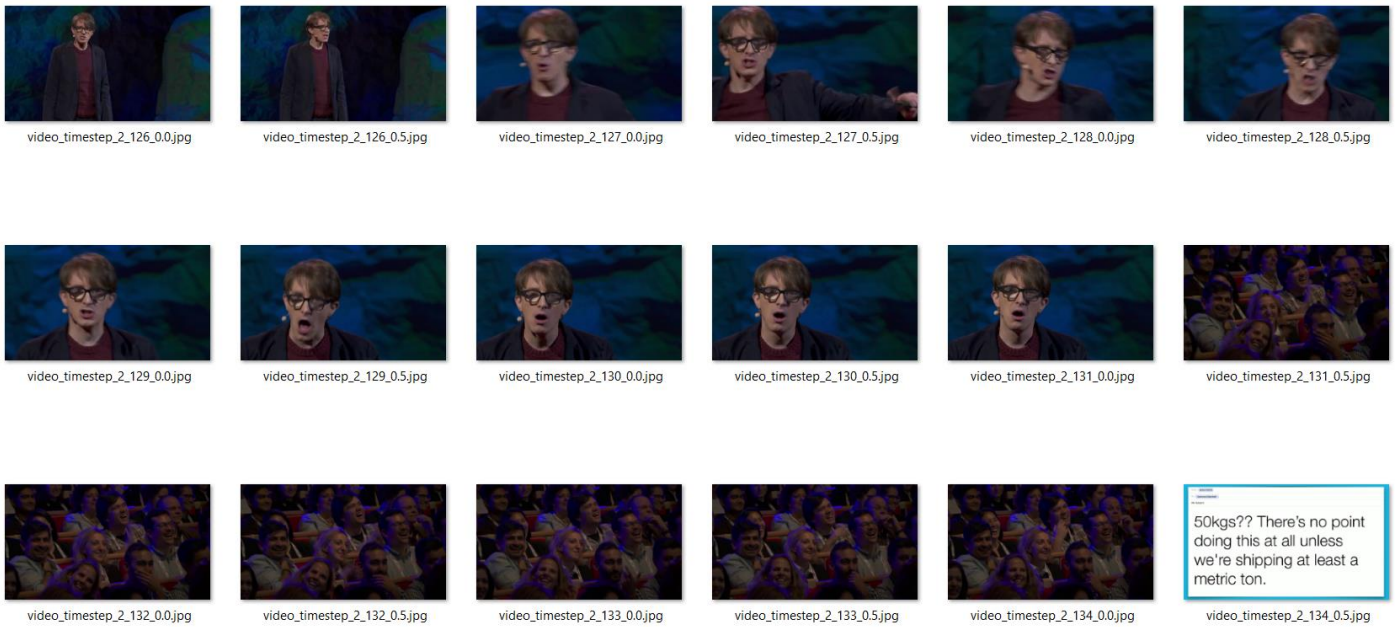

Figure 7: Examples of frames

This is approximately what we get. As can be seen there are a lot of redundant images like slides or audience. We need to remove those and make sure that only a face of a presenter is takes. This leads us to the next stage

3.2. Detecting a face

For this task a deep learning model was used to select only images with a presenter's face on it.

Figure 8: Examples of a face being detected

After this phase we are left with these images:

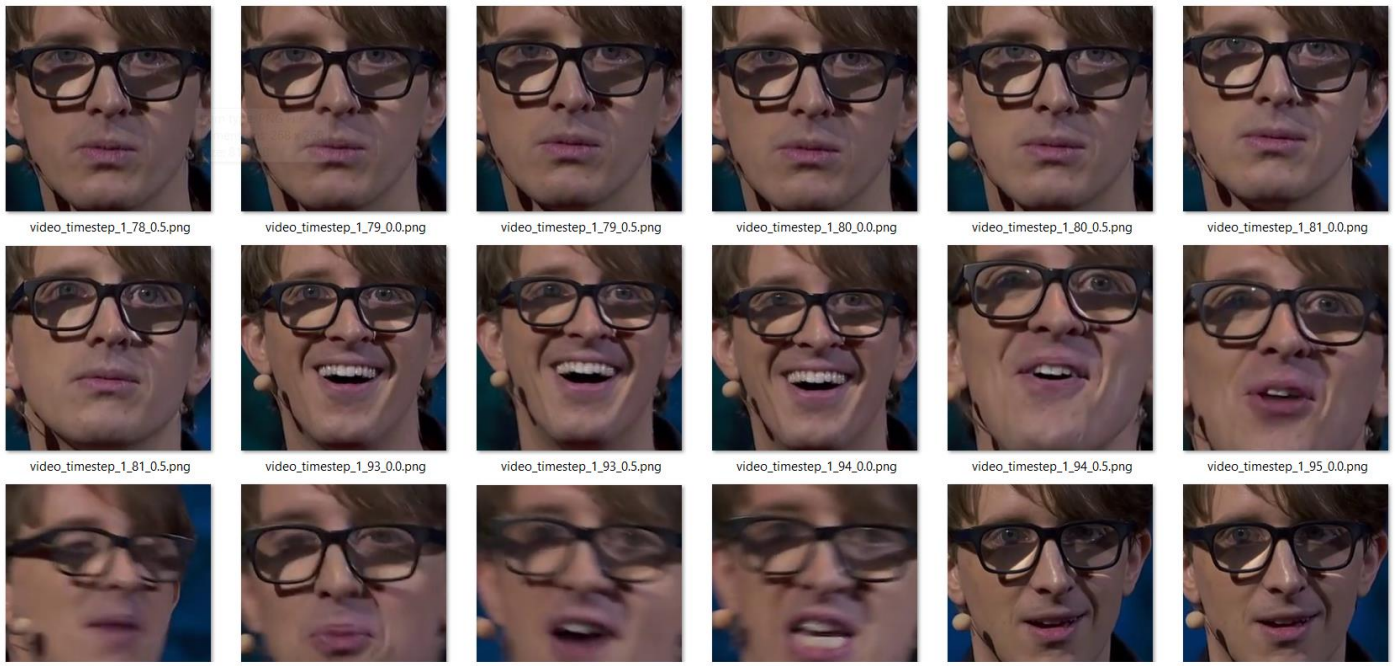

Figure 9: Examples of a set of cropped images

Now we need to detect emotion

3.3. Detecting emotion

The problem is most emotions are very close to each other. Even some human cannot distinguish if a person is happy or excited, calm or indifferent. So, for those reasons I chose to separate emotions into 3 categories, which have a positive, negative and neutral connotations. Those are “happiness”, “anger” and “calm”.

For this process of emotion detection we are using a pretrained on a public dataset of 30,000 images neural network model. At the output we get a list of emotions for each frame.

```
{'anger': 55.19169434202277,
 'calm': 16.827444999421925,
 'happiness': 27.98086065855531},
{'anger': 55.41989339368148,
 'calm': 25.784732286944205,
 'happiness': 18.795374319374325},
{'anger': 43.89052616491638,
 'calm': 1.2395263559074527,
 'happiness': 54.869947479176176},
{'anger': 56.75391943268906,
 'calm': 19.824956491981403,
 'happiness': 23.42112407532954},
{'anger': 54.75484498524503,
 'calm': 15.21984107682891,
 'happiness': 30.025313937926057},
{'anger': 47.39829569425568,
 'calm': 14.365874703476445,
 'happiness': 38.23582960226787},
{'anger': 47.62800390761291,
 'calm': 0.7431856861975069,
 'happiness': 51.62881040618959},
{'anger': 50.86578484067452,
 'calm': 20.78200594725805,
 'happiness': 28.352209212067432},
{'anger': 56.16943645724497,
 'calm': 20.023915671299967,
 'happiness': 23.806647871455066},
{'anger': 52.28614658477002,
 'calm': 11.864120100166636,
 'happiness': 35.84973331506334},
{'anger': 42.379006088930346,
 'calm': 0.898226306396274,
 'happiness': 56.722767604673386},
{'anger': 37.063665316386185,
 'calm': 10.105490565862475,
 'happiness': 52.83084411775134},
```


Figure 10: Sample of an output

As can be seen it is quite hard to extract any meaningful information from this raw list. So additional step is performed to show meaningful insights.

3.4. Infographics

In this step I try to visualize all the important information that can be gathered from emotions. The idea is to give a presenter a summary of how he behaves during his presentation. A more detailed results would be shown in Section 5.

4. Limitations

Although the results are impressive there are still some limitations:

1) Similarity between different types of emotions (like calm or neutral). The problem is that these faces look very much similar from afar. In this case it is almost impossible to distinguish between them. To solve this, I treated those images as being of the same class – "calm"
2) Too many emotion classes. This problem is similar to the limitation described above. To address this issue, I used a clearly divisible classes – "happy", "angry" and "calm". This information is enough to get some insights of the presentation.
3) Certain frames contain image are of not a presenter. This can be addressed by taking a video of only the presenter. This is a desired action is the goal is to analyze the emotional component of the speech.

5. Results and Infographics

In this section, I discuss the results and show infographics with some useful information gathered.

5.1. Results

The model returns a normalized set of values for each of the three emotions (happy, angry and calm). Meaning, that sum of the values is equal to 100%. And from this information we can get much more than mere dominant emotion in the frame. The reasoning is that if a person is 51% angry and 49% happy is different from a person who is 100% angry. Although, the dominant emotion is anger, a person with 51% anger can be viewed as describing something passionately. Whereas, a 100% angry person is just angry.

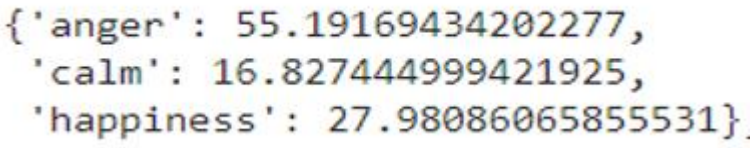

Figure 11: Output result

5.2. Infographics

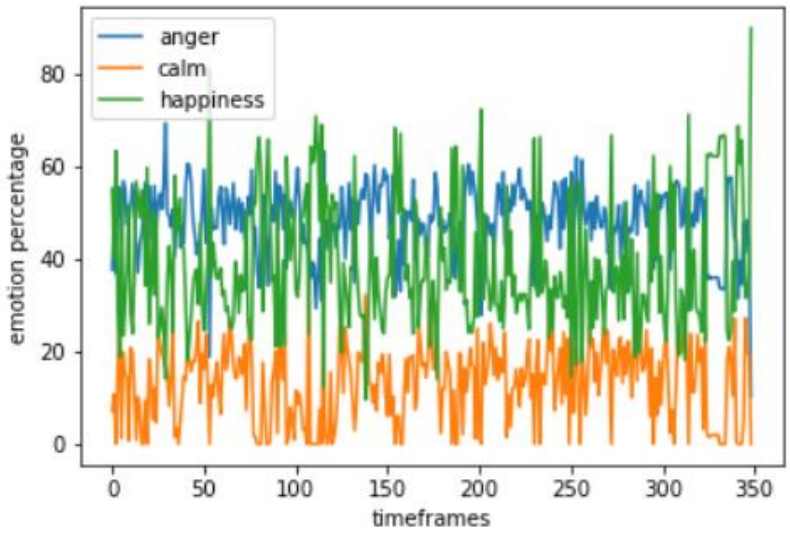

Figure 12: Model output

From this graph we can see that a presenter is rarely calm. Also, we can say that mostly he is expressing anger, then happiness. Although, it is hard to get more out of this graph, we can say that a person is describing some issue very passionately.

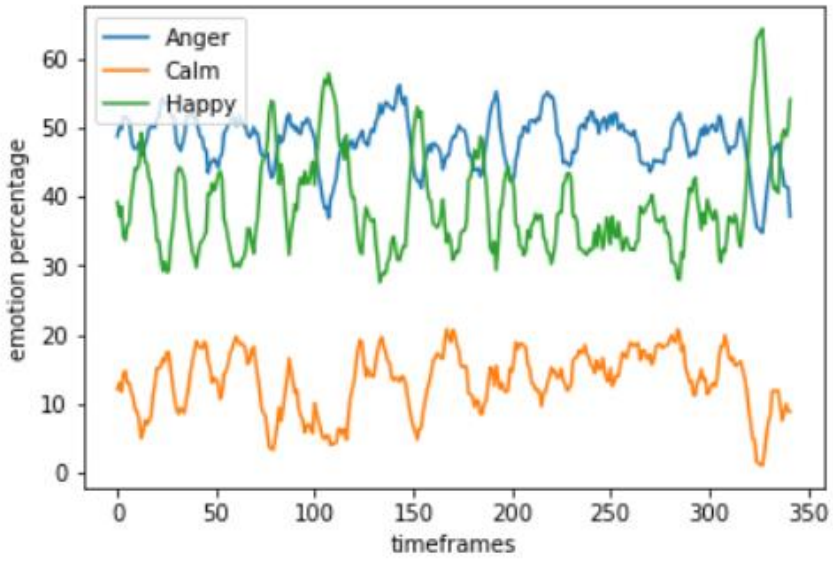

Figure 13: Smoothed model output

This is a smoother version of the previous graph. Here an 8 data-point moving average was used. Similar insights as in previous graph. But here it is much clearer.

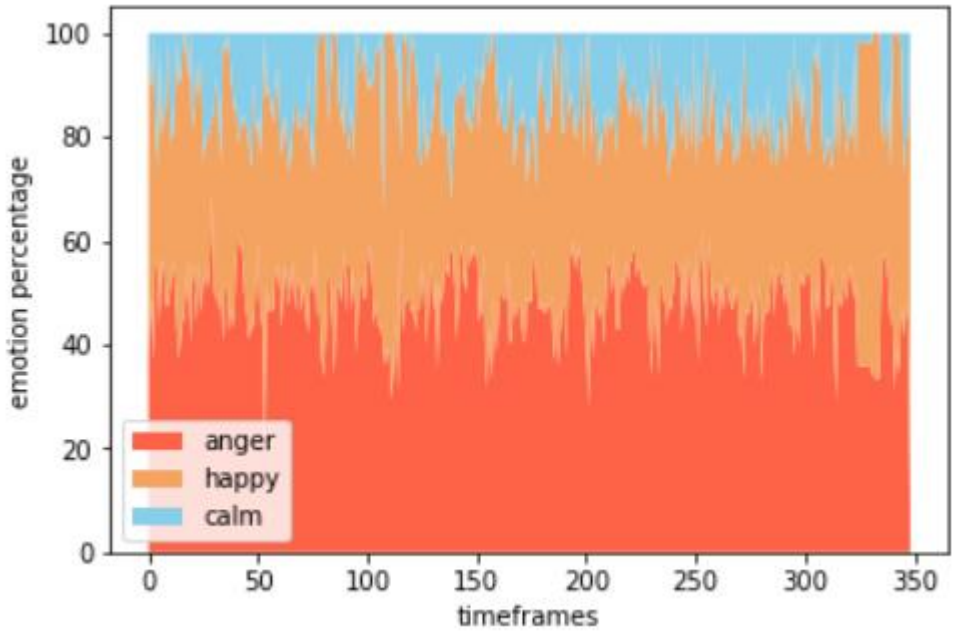

Figure 14: Stackplot of emotions

From the stackplot we can see that anger if the dominant face expression during the speech.

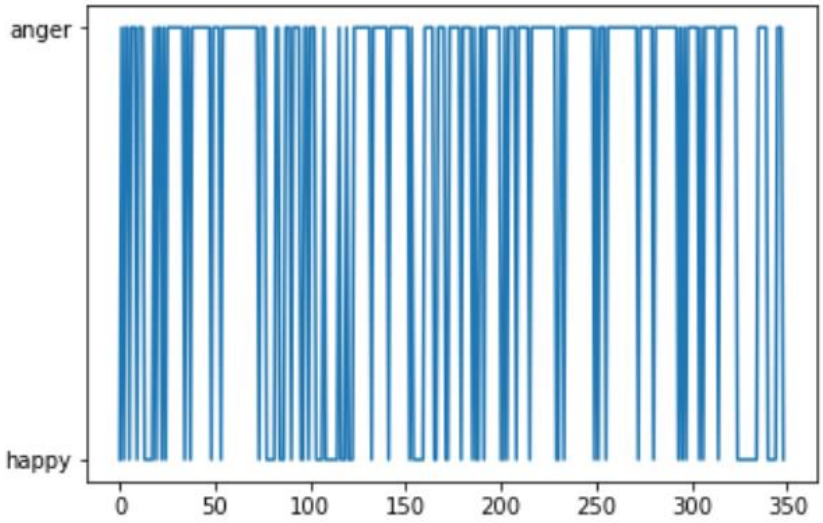

Figure 15: Frame by frame emotion

From this graph we can see that a dominant emotion is constantly changing from being angry to being happy. Additional insight is that because of constant change we can assume that he might have mixed feelings about the topic.

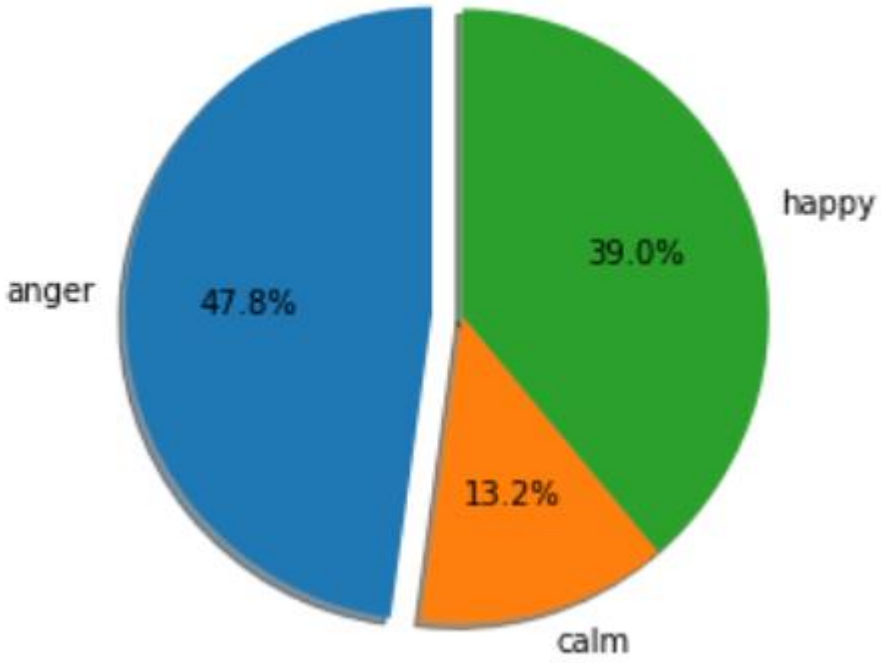

Figure 16: Pie chart of the average strength of each emotion

This is a pie-chart of the average emotion percentage points. As we can see “anger” and “happy” are quite close to each other. While he is rarely calm. That indicated that a person is very emotional in delivering his speech.

Figure 17: boxplot of emotions

From the boxplot we can see that a presenter sometimes shows extremely happy face.

6. Conclusion

This report introduces a pipeline of how presentations can be analyzed. The results of these analyses can greatly benefit those how aim to improve their public speaking. The results of the model are easy to analyze and understand, and these results can be obtained quite fast since the models use a pretrained neural networks.

7. Reference

1. Liu W, Anguelov D, Ehran D, Szegedy C, Reed S, Fu C, Berg A. "SSD: Single Shot MultiBox Detector". https://arxiv.org/abs/1512.02325 2016
2. Minaee S, Abdolrashidi A. "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network" University of California, 2019
3. Veitch J. "This is what happens when you reply to spam email" 2016. https://ted2srt.org/talks/james_veitch_this_is_what_happens_when_you_reply_to_spam_email
4. Viola P, and Jones M. "Rapid Object Detection using a Boosted Cascade of Simple Features" Conference on Computer Vision and Pattern Recognition. 2001
5. Zanaton I, Effendi Z, Tamby S, Kamisah O, Denise L, Siti M, Pramela K. "Communication skills among university students" UKM Teaching and Learning Congress, 2011