Facial Expression Recognition

Calvin GALAGAIN
ENS Paris Saclay - MVA
Calvin.galagain@ens-paris-saclay.fr

13/01/2023

(All codes are available here:
https://colab.research.google.com/drive/1YJZVdxNrWi6IgXHHHNndHOOnlh5p1-fh?usp=sharing)

Abstract

Computer vision is now able to analyze a lot of information, especially that related to facial detection. The human face has more than forty muscles that generate a wealth of expressions that can be communicated. For a human being, this understanding is innate and trying to make algorithms that reproduce it is a real challenge.

In this presentation, we presented our approach for sentiment analysis based on the use of pre-trained transformer models and networks using Pytorch. We used the AffectNet [1] database to train our models, starting with image preprocessing to improve their quality. We then used pre-trained models on ImageNet to extract features from the images. To predict the sentiments associated with the images, we used transformers, which showed good results in terms of accuracy. To understand how the models make their decision, we visualized the attention mechanisms to identify the most important areas in the images. Finally, we re-programmed a transformer from scratch and used input patches based on the important parts of the face to improve the prediction results.

For the project, we will create two libraries: Nexd and Lexd which respectively facilitate functions related to images, and functions related to landmark extraction. These are pip installable as can be seen in the colab at the top of this document.

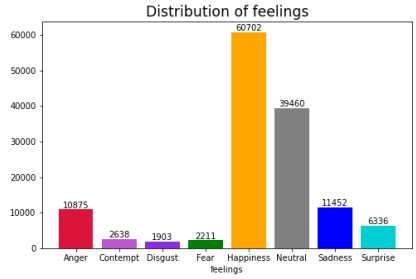
1. Introduction

In the quest to learn tasks that humans know how to do naturally, this subject has many implications. Today to answer questionnaires which aim to know our feelings during a video, we must complete everything manually and, it is very long. One use could be to incorporate a sentiment recognition system on computer cameras to track changes in facial expressions live. This could also be useful to help people with illnesses. Autistic or blind people find it difficult to read other people's facial expressions, they would need technologies like the one we are going to seek to develop. Finally, the robotics world could also be very fond of a system that can read emotions because it could interact more intelligently with its users. To go further, we could try to read the expressions of a crowd or a driver to detect dangerous situations that could be avoided with prevention. Finally, in a more recent context, many people wear masks and the face is partially hidden. It is therefore more difficult to read the expressions that the person is feeling. It is interesting to see if algorithms could overcome this barrier.

The problem is therefore very interesting because it has questioned many scientists since the 90s. But with the appearance of many databases, the problems have changed. Indeed, it is possible to have almost perfect results in controlled situations. If the subject is in optimal lighting, with an optimal background and conditions exactly the same as the other people in the database, the results are more than excellent. The real problem appears when we try to analyze faces in real situations “in the wild”. Sentiment detection promises to bring many applications that would be able to analyze our emotional state and improve our health or safety if these applications do not infringe on our freedoms.

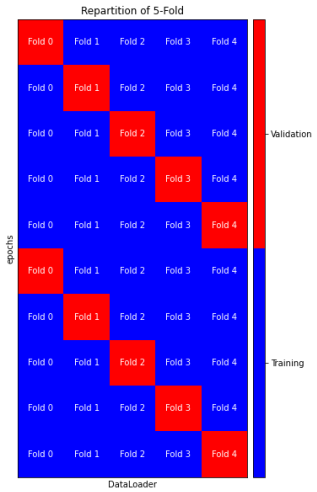
2. Data Preprocessing

The database is very poorly distributed. If we do not remedy this, we will be faced with a problem. Indeed, "Happy" faces represent 60% of the database. However, in the original paper, the overall accuracy is only 56%. Seeing this, we understand that the problem could be poorly handled and predict "Happy" faces to have a better success rate.



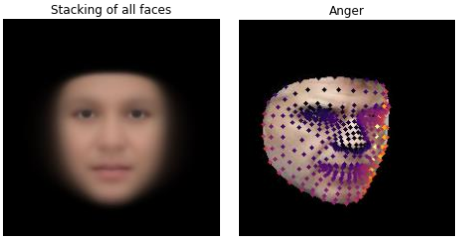
(Fig1: Distribution of feelings on AffectNet)

As we can see (Fig.1), the images are very poorly distributed and the number of occurrences for each class is completely different. For this work, the most common classes will be undersampled to form evenly sampled sets. We will then divide this work into k consecutive folds (Fig.2), keeping a percentage of the database to test our models with data that has never been seen before.



(Fig 2: for each epoch, the validation dataset is different)

The database is very large and does not only include human faces. To select the faces, we will use MediaPipe [2] here, which allows us to extract 468 3D landmarks if the face is human. The higher the detector coefficient, the greater the chance of having a human face. Since these points are always in the same position, we can extract the facial parts. This will allow us to have the position of the irises and the middle of the mouth, among other things. With these three points, we will be able to align all the faces so that the data has the least bias possible.



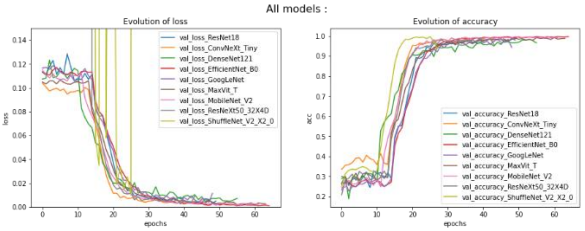
(Fig 3: (left) stacking of all the faces after our preprocessing, (right) 3D landmarks on a face)

With this treatment, the algorithms will all have the same type of images, we will reduce the biases that are related to positions or various variations.

3. Pretrained Models

The first part of the database analysis allows us to have a classic vision of understanding our database. We will use pre-trained models on Image-Net and specialize the classifier only at first. Then we will make the entire model trainable so that the other layers can also understand the emotions. The advantage of doing a cross-validation with k-folds is to reduce overfitting. As we can see on the training, the algorithms perfectly understand the sentiments but with the data set that no one has seen, the results are less good.

Indeed, the new data brings new faces, new biases that have never been seen and that is the problem. The Affect-Net paper boasts of having an accuracy of 56.4%, which we will find on the new data here. Data processing is therefore very important. Without treatment, there are 60% Happy faces so, by always putting this emotion we would have better results. This is not the goal.

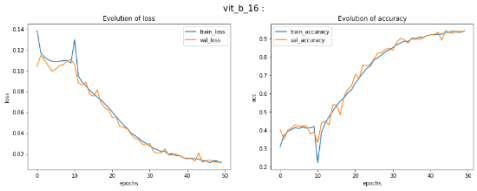


(Fig 4: Evolution of accuracy for each pretrained models)

To get better results, we will predict our data with all the models and average the results to reduce the error by weighting by the accuracy obtained during training. The goal is to reduce individual errors. For a case like ours, it is wise to weigh each sentiment by the score obtained for each model. Here, given that the final accuracy value is 100%, it is not very useful to weigh. This method gives us much more interesting results. Our success rate is the same as the paper but our variance is much lower! Each sentiment has the same percentage even when contempt is difficult to predict in the paper. We will see the results in the part 6.

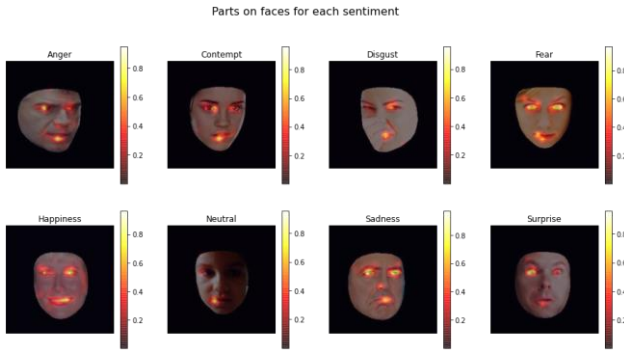
4. Attention Is All We Need

Transformer models have made significant advancements in a variety of areas, particularly in image analysis. The TransFER paper [3] demonstrates the use of a very simplistic dataset, but it shows that the attention mechanism can be visualized to better understand what our Transformers are learning. By visualizing the weights of our trained vit_b_16 model, as seen in previous models, we can gain insight into the areas of the image that the model focuses on.



(Fig 5: evolution of pretrained-vit_b_16)

As we can imagine, the model tends to focus on the expression that is located around the eyes and around the mouth. This is due to the fact that these areas of the face are known to convey the most emotion. Unlike other models, Transformers use an attention mechanism that is based on creating patches in the image [4]. This allows the model to focus on specific areas of the image, rather than treating the entire image as a whole.



(Fig 6: visualization of our weights)

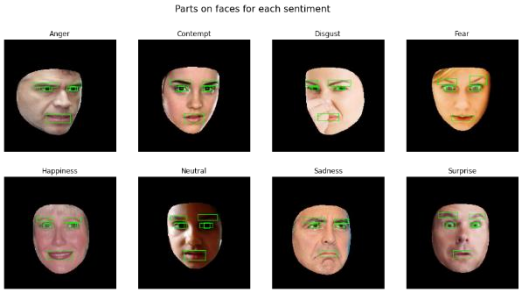
It is interesting to explore ways to select these patches and remove the least useful ones. This could potentially lead to improved performance by reducing the amount of noise in the data and allowing the model to focus on the most important areas of the image. Additionally, by understanding which areas of the image the model is focusing on, we can gain insight into the underlying mechanisms of the model and potentially improve its performance.

In conclusion, Transformer models have shown great promise in image analysis and have the potential to improve our understanding of the underlying mechanisms of image analysis. Through visualization of attention mechanisms and exploring ways to select and remove patches, we can gain insight into the model's decision making process and improve its performance.

5. Custom Transformer for Facial Expression

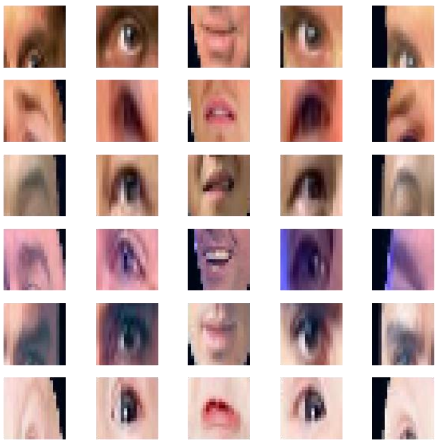
By extracting facial parts using landmarks, we can create our own patches. This allows us to target the most important areas of the image (Fig.6), thus reducing noise and improving performance. Additionally, by understanding the parts of the image the model focuses on, we can further improve performance by optimizing architecture parameters. Facial parts can be extracted using Mediapipe landmarks which are always indexed the same way. A visualization of these can be seen above (Fig.3).

To create an architecture that works perfectly, we had to recreate the Transformers and all the layers from scratch. This ensured that the model was tailored to our specific needs and was not limited by pre-existing architecture. It is now easy with our code to input anything we want to replace simple batches. This allows us to experiment with different input types and see how they affect the model's performance.



(Fig 7: parts on faces detected with Landmarks)

As shown in the figure, the parts of the face we keep are the eyes, eyebrows and mouth, slightly enlarging the area to keep more context. This allows the model to have a better understanding of the facial expression and predict emotions more accurately.



(Fig 8: patches on the input of our Custom Transformer)

The results are not as good as those seen in the other sections. Indeed, the weights are not pre-trained here and do not know in advance which parts of the image are important. However, for an architecture that is much smaller than the other models, we have the first emotions that are predicted just as well. This shows that our approach

is efficient and that with further optimization, we can achieve even better results. We will see them in the next section.

6. Results

Our work has very good results when trying to determine the expression in real situations. Indeed, often, the expression on a face is not often unique and it is well seen. An expression is a set of micro-expressions and, during our training, we only try to predict one expression among the eight that we had. We can see that this assumption is true particularly for the Top2 accuracy.

In the following two figures, the results of the original paper are displayed in the first row. The values in green are the highest in the column and the values in red are the lowest.

Model_name_Top1	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	mean	variance
original_paper	0.6	0.32	0.64	0.57	0.68	0.49	0.6	0.63	0.566	0.011
ResNet18	0.5	0.36	0.54	0.58	0.59	0.38	0.59	0.44	0.498	0.008
ConvNeXt_Tiny	0.53	0.6	0.49	0.34	0.63	0.41	0.36	0.63	0.499	0.012
DenseNet121	0.46	0.26	0.53	0.64	0.79	0.36	0.45	0.48	0.496	0.023
vit_b_16	0.39	0.42	0.38	0.38	0.77	0.46	0.45	0.44	0.461	0.014
EfficientNet_B0	0.34	0.45	0.52	0.51	0.59	0.43	0.54	0.56	0.493	0.006
GoogLeNet	0.49	0.31	0.48	0.45	0.62	0.26	0.7	0.44	0.469	0.018
MaxVit_T	0.48	0.43	0.48	0.57	0.64	0.53	0.55	0.33	0.499	0.008
MobileNet_V2	0.45	0.48	0.51	0.37	0.67	0.42	0.41	0.43	0.468	0.007
ResNeXt50_32X4D	0.46	0.66	0.54	0.45	0.35	0.34	0.53	0.47	0.475	0.01
ShuffleNet_V2_X2_0	0.48	0.45	0.54	0.5	0.49	0.36	0.38	0.38	0.448	0.004
my_vit	0.4	0.41	0.31	0.28	0.81	0.07	0.19	0.05	0.315	0.051
all_models	0.54	0.52	0.57	0.53	0.66	0.51	0.55	0.49	0.546	0.002

(Fig 9: Top1 Norm Down-sampling result)

Model_name_Top2	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	mean	variance
original_paper	0.78	0.7	0.81	0.82	0.85	0.7	0.78	0.83	0.784	0.003
ResNet18	0.71	0.43	0.68	0.75	0.76	0.51	0.69	0.63	0.645	0.012
ConvNeXt_Tiny	0.71	0.76	0.66	0.47	0.77	0.61	0.49	0.83	0.662	0.015
DenseNet121	0.57	0.55	0.63	0.78	0.86	0.62	0.56	0.65	0.653	0.011
vit_b_16	0.54	0.57	0.66	0.53	0.85	0.79	0.57	0.71	0.628	0.017
EfficientNet_B0	0.61	0.56	0.64	0.67	0.75	0.63	0.69	0.73	0.66	0.003
GoogLeNet	0.62	0.45	0.58	0.67	0.76	0.38	0.81	0.77	0.63	0.021
MaxVit_T	0.71	0.59	0.64	0.69	0.84	0.64	0.71	0.61	0.679	0.005
MobileNet_V2	0.61	0.74	0.67	0.59	0.87	0.53	0.56	0.57	0.643	0.011
ResNeXt50_32X4D	0.7	0.79	0.72	0.64	0.46	0.52	0.67	0.65	0.644	0.01
ShuffleNet_V2_X2_0	0.67	0.64	0.65	0.73	0.65	0.57	0.56	0.64	0.639	0.003
my_vit	0.58	0.63	0.57	0.38	0.87	0.7	0.78	0.7	0.695	0.009
all_models	0.8	0.81	0.69	0.79	0.81	0.74	0.74	0.75	0.766	0.002

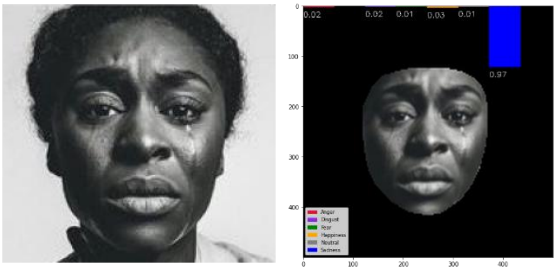
(Fig 10: Top2 Norm Down-sampling result)

There are several interesting observations to make. Independently, all models are worse than the results reported in the paper. The most interesting result concerns the voting learning. By combining the predictions of all models, we get a result that has almost the same mean prediction as the paper but, our variance is much lower (especially for Top1 accuracy). This shows that when we gather our models, they predict the entire set of emotions better, where the paper will poorly understand certain emotions.

We notice that our Custom Transformer, with our Attention mechanism, is very bad as it gets the worst scores for each possible prediction. This shows that there is still work to do on it. There is potential as for example it gets the best scores for happy faces. So it still needs more training to generalize.

Below are the results that our Voting-Learning predicts.

(Fig 11: result on a surprised face)



(Fig 12: result on a sad face)

Our model is well focused on trying to predict a single emotion as we can see. The results are therefore very good and allow us to confirm the results of the paper, understand them and even improve them.

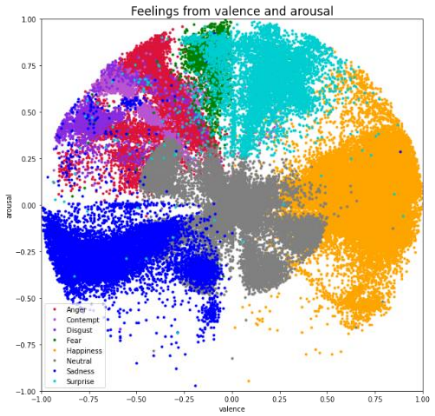
7. Ameliorations

To begin to solve this problem, it would be interesting to complicate the layers of our custom attention and adding, for example, a convolutional network that will extract on its own what our patches may not have captured. (This track has been started but not yet completed in terms of results).

The idea behind this approach is that we have a human vision that confirms what the Transformers visualize (Fig. 6) and we can easily extract these parts. However, for our model, perhaps other parts of the face that do not immediately come to mind are just as important. The idea of adding a convolutional network that learns at the same time as the Transformer is that it will create filters, that is, parts of the image for each filter. These can be added in parallel to our previous patches (Fig. 8). And, this network would learn on its own what it considers important in addition to the patches.

In a second part, the prediction of valence and arousal is to be paralleled with this work. Valence refers to how positive or negative an event is and Arousal reflects whether an event is exciting/agitating or calm/soothing. Indeed, the AffectNet database is also composed of this information which is not used at all in this work. Here, the project focused solely on image-based extraction. As we

obtain the results of the paper with just the images, we can think that the addition of this will only be beneficial and we will easily exceed our performance.



(Fig 13: emotions from Valence and Arousal)

References

[1] A.Mollahosseini, B.Hasani, M.H.Mahoor, **"AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild"**, Arxiv 2017

[2] C.Lugaresi, J.Tang, H.Nash, C.McClanahan, E.Uboweja, M.Hays, F.Zhang, C.L.Chang, M.G.Yong, J.Lee, W.T.Chang, W.Hua,M.Georg, M.Grundmann, **"MediaPipe: A Framework for Building Perception Pipelines"**, Arxiv 2019

[3] Fanglei Xue, Qiangchang Wang, Guodong Guo, **"TransFER: Learning Relation-aware Facial Expression Representations with Transformers"**, Arxiv 2021

[4] A.Dosovitskiy, L.Beyer, A.Kolesnikov, D.Weissenborn, X.Zhai, T.Unterthiner, M.Dehghani, M.Minderer, G.Heigold, S.Gelly, J.Uszkoreit, N.Houlsby, **"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"**, Arxiv 2020