Facial Expression Analysis using Image/Video Data

Advanced Topics In Machine Learning Course Project

ARISTOTLE UNIVERSITY OF THESSALONIKI

Introduction

Computer aided facial expression recognition (FER) has gained popularity and transcended this research field to enable cutting-edge sentiment analysis tools in day-to-day applications.
In this project our goal is to test established models such as VGG19, ResNet, Xception and state-of-the-art transformer (DAN) based approaches using multiple cross-attention heads that achieve very high accuracy scores on testing data. We explore how these very robust models behave when presented with different datasets that consist of faces from individuals across the globe and how this diversity in facial structures affects the accuracy or even biases that are embedded during the training process. We also conceptualized and tested a pipeline for automated dataset creation from online videos.

Datasets

Two open datasets have been used and one was created using our automated method.

FER2013: Consists of 35888 different images of the same dimensions using a grayscale background. It is split to 3 sections, a training, testing and validation set with 7 folders in each where images with 7 distinct facial expressions are stored.

CK+: The CK+ database is widely regarded as the most extensively used laboratory-controlled facial expression classification database available, and is used in the majority of facial expression classification methods. It contains 981 photos split into 7 different categories (anger, disgust, fear, sad, contempt, happy, neutral).

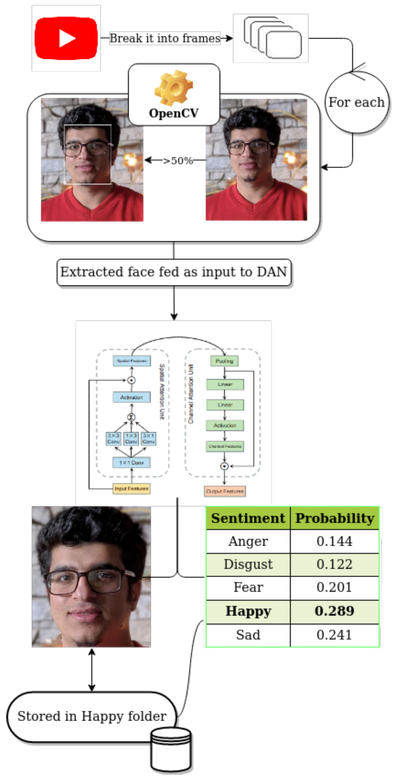
Our Dataset: Consists of 3410 different images (Fig. 1) that were extracted using our pipeline. The photos are split into the same 7 categories and derive from videos where high class actors give their best performances.

1

Automated Dataset Creation

To create our own dataset we relied on a novel approach which is reported to achieve up to 89.7% in accuracy concerning facial expression recognition. Distract-Your-Attention-Network aka DAN, is an attention based neural network that employees three core components, a Facial Clustering Network (FCN), a Multi-Head cross attention network (MAN) and an Attention Fusion Network (AFN). Combining these features results to state-of-the-art performance in most FER datasets such as AffectNet, RAF-DB and SFEW 2.0. By utilizing such a powerful model and the face extraction features of OpenCV we are able to split a video into a sequence of frames, extract faces given the probability of an area being a face is more than 50 % and finally obtain via the softmax probabilities of DANs' top layer the most probable category for every eligible frame. The framework is depicted in Figure 2.

2



Sentiment	Probability
Anger	0.144
Disgust	0.122
Fear	0.201
Happy	**0.289**
Sad	0.241

Model Construction

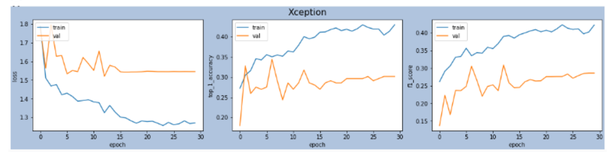
Images stored in appropriate folders named after the classes created the train, test, val sets through the utilization of the ImageDataGenerator function. Classes got different weights depending to their number of entries.

19 models were tested, each using the same ImageNet pretrained weights.

A 1080 GPU and 32 GB of ram were used. ~8 mins per model. Early stopping could further minimize the time needed.

Some final dense layers were inserted in each model to accommodate the projects needs. Images' size was altered according to each model's needs. A flexible learning rate was used. 15 warm-up and 15 training epochs. The loss, accuracy and f1 score were monitored for the train and val phases. (Fig. 3)

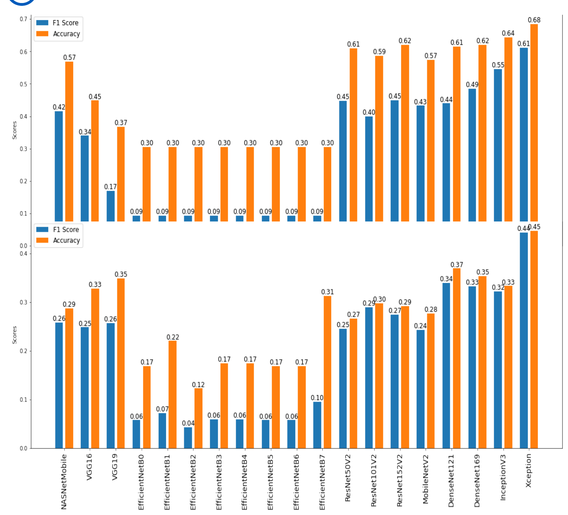
3



Model Evaluation

A comparison of the different models is being made between two different datasets. Figure 1 is the CK+ dataset while Figure 4 displays the results our dataset gave. In both cases Xception had the best results with 68% and 45% accuracy.

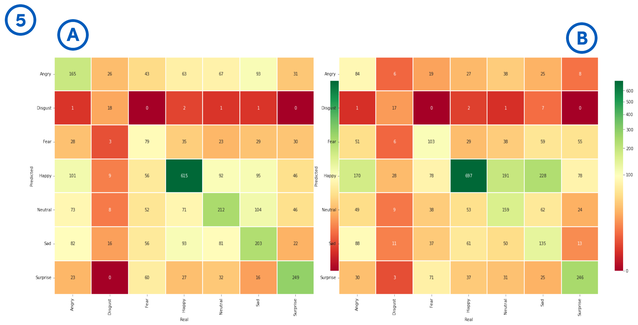
4



Testing models

Various tests were performed on all the available datasets with the ones giving the best results being ResNet and VGG (among others). The models performed well considering the number of classes and only in our dataset we had to use less classes since some of them had a small amount of photos to perform well. Figure 5 displays confusion matrix of ResNet (A) and VGG (B) for the FER2013 dataset.

5



Conclusion

Even though FER is a well studied problem there is still a lot of work to be done. State of the art transformers can just barely achieve 70% accuracy in all the classes while simpler models give lower results. The automated dataset creation is a fast and easy process to create a new dataset for testing but still there is room for improvement by deleting similar photos. A side effect of this method was the overfitting of the models since several photos where the same. Possible solution for this method is a manual split of the photos to train/test/val. Finally, an interesting result of this project is the ability to extract the general sentiment of a video using part of the automated dataset pipeline.

References

1. Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition, 2021.

2. Jie Cai, Zibo Meng, Ahmed Shehab Khan, James O'Reilly, Zhiyuan Li, Shizhong Han, and Yan Tong. Identity-free facial expression recognition using conditional generative adversarial network. In 2021 IEEE International Conference on Image Processing (ICIP), pages 1344–1348,2021

SCHOOL OF INFORMATICS
Data & Web Science MSc Program

Aristotle University of Thessaloniki
Department of Computer Science
https://dws.csd.auth.gr/