Contribution ID: **94** Type: **Paper presentation**

Activation Maps Analysis for Detected Emotions Using a Deep Convolutional Neural Network Ensemble

Friday 20 September 2024 14:40 (20 minutes)

Emotion detection is a subject of great scientific importance for multiple research fields spanning from medicine to security. This poses a great problem for Computer Vision methods like DCNNs due to the complexity of factors that need to be taken into consideration and the lack of context when determining the correct emotion category, moreover, the difficulty increases when trying to determine the emotions in large and dynamic amount of data, such as video streams. Although DCNN architectures proved that good results can be achieved by training on large datasets and being able to detect emotions from facial, posture, and text samples, there is a field that needs extensive research: analyzing the comprehension at the representation level of the machine on the entire ensemble of factors determining the correct prediction for human emotions.

The author's contribution consists of designing a novel approach at a theoretical level and from an applicative perspective by implementing a computational and visualization method for activation region analysis from extracted features of the first convolutional layer, up to the top prediction layer of a custom-trained DCNN ensemble for emotion detection. The aim is to serve as an evaluation instrument for the models that run real-time detection and classification tasks from live video streams, on human emotions, due to the specificity of the field of expertise required for validating the detections by the human factor.

The current study aims to provide a complementary method of evaluating the predictions of a custom-trained DCNN ensemble for real-time facial emotion detection by exposing the activation regions that contribute to the final prediction and extracting them. The DCNN ensemble consists of two CNNs, YOLOv.7 and RepVGGPlus. The first is trained for face detection on the WIDER FACE dataset containing 393,703 samples. The second one is trained to classify the seven primary emotions on the AffectNet dataset, which has 420,299 facial expressions. The result is fused and displayed overlayed on the source image, containing the bounding box of the detected face and the classification score corresponding to the detected emotion. We chose the YOLOv7-tiny model and trained it using transfer learning technique on the Pytorch framework due to its superior overall performance, both in detection and inference speed over YOLOv.5 existing models, as we intent to use is on real-time inference from video stream.

To conduct our research, we created a script using Python v.3.10 programming language for loading the trained models and running inference on them directly from a live video stream, using a professional USB conference camera as input, with a 1920×1080 pixels resolution and 30fps.

We further analyzed the activation zones'intensity for each layer using bicubic interpolation and extracted the corresponding feature vectors for computing and plotting the real-time histogram for each class. The activation maps are overlayed on the live video stream to help with visual observation of the model behavior. Additionally, for each detected emotion, the bounding box information is exported in .png image format as patch file to a predefined export folder. The features from these patches are extracted using the HOG algorithm and further analyzed to generate a histogram (there is also an option for real-time preview). The export is set to automatically save frames for a while that lasts until another emotion is detected. In this way, the generated set is automatically annotated. From each convolution layer of the RepVGGPlus neural network are constructed heatmaps based on activation maps for each of de identified classes during inference. The main purpose of this process is to observe and study the correlation between ground truth established by experts in the field and the relevant features extracted by the trained RepVGGPlus classification network in a dynamic way for emphasizing the variation of the predictions compared with static images.

On the test subset, comprising 125 samples, the model obtained a mAP@0.5 score of 0.922. This will be compared when considering the detection performance indicators for real-time inference and the magnitude of the dominant and relevant activation regions from overlapping predictions and convolutional layers.

Experiments are to be conducted on several subjects to show that the proposed method can be used to gain more insight on the generalization capability of a trained model for emotion prediction on live video streams and contribute to the development of attention mechanisms.

Authors: Mrs SERGHEI, Madalina Oana (National University of Science and Technology Politehnica Bucharest); Prof. ICHIM, Loretta (National University of Science and Technology Politehnica Bucharest); POPESCU, Dan (National University of Science and Technology Politehnica Bucharest)

Presenter: Mrs SERGHEI, Madalina Oana (National University of Science and Technology Politehnica Bucharest)

Session Classification: Doctoral Symposium

Track Classification: Doctoral Symposium