HSEmotion: High-speed emotion recognition library

Andrey V. Savchenko

PII: S2665-9638(22)00117-8
DOI: https://doi.org/10.1016/j.simpa.2022.100433
Reference: SIMPA 100433

To appear in: Software Impacts

Received date: 20 September 2022
Revised date: 7 October 2022
Accepted date: 14 October 2022

Please cite this article as: A.V. Savchenko, HSEmotion: High-speed emotion recognition library, Software Impacts (2022), doi: https://doi.org/10.1016/j.simpa.2022.100433.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
HSEmotion: High-Speed Emotion Recognition Library

Andrey V. Savchenko, HSE University, Laboratory of Algorithms and Technologies for Network Analysis, Nizhny Novgorod, Russia, avsavchenko@hse.ru

Abstract
A lot of attention has been recently brought towards mitigating algorithmic bias in models of affective behavior and facial expression recognition. In this paper, a novel software is presented that provides an accurate and high-speed emotion recognition (HSEmotion) tool. Several EfficientNet-based models have been trained to classify emotions of static facial photos. Experimental study demonstrates that our pre-trained models can be used as feature extractors for fast and accurate video-based emotion recognition in various tasks without the need for fine-tuning the whole neural network.

Keywords
Facial expression recognition, affective behavior analysis in-the-wild, emotion classification, efficient convolutional neural networks

Code metadata

<table>
<thead>
<tr>
<th>Nr</th>
<th>Code metadata description</th>
<th>Please fill in this column</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Current code version</td>
<td>v0.2</td>
</tr>
<tr>
<td>C2</td>
<td>Permanent link to code/repository used for this code version</td>
<td><a href="https://github.com/HSE-asavchenko/face-emotion-recognition">https://github.com/HSE-asavchenko/face-emotion-recognition</a></td>
</tr>
<tr>
<td>C3</td>
<td>Permanent link to reproducible capsule</td>
<td><a href="https://codeocean.com/capsule/4322879/tree/v1">https://codeocean.com/capsule/4322879/tree/v1</a></td>
</tr>
<tr>
<td>C4</td>
<td>Legal code license</td>
<td>Apache 2.0 License</td>
</tr>
<tr>
<td>C5</td>
<td>Code versioning system used</td>
<td>Git</td>
</tr>
<tr>
<td>C6</td>
<td>Software code languages, tools and services used</td>
<td>Python</td>
</tr>
<tr>
<td>C7</td>
<td>Compilation requirements, operating environments and dependencies</td>
<td>PyTorch, Timm</td>
</tr>
<tr>
<td>C9</td>
<td>Support email for questions</td>
<td><a href="mailto:andrey.v.savchenko@gmail.com">andrey.v.savchenko@gmail.com</a></td>
</tr>
</tbody>
</table>

1. Introduction
Human emotions are the changes in psychological states and can be analyzed from the expressed brain signals, voice, parts of the body and face [1, 2]. Facial analytics is one of the most appropriate modalities for industrial applications because the changes of facial expressions help to predict a user's feeling without nonverbal communication. Indeed, if a person feels happy, angry, etc., one can notice the changes of expressions on his or her facial image [2]. The facial emotion analysis is still a challenging problem that is raised nowadays in various applications with man-machine interactions, such as
analysis of student engagement in e-learning and online conferences [1], advertisements and other tasks with prediction of the user satisfaction, etc.

Traditionally, facial expression recognition (FER) [3] has been performed on laboratory-controlled data that poorly represents the environment and conditions faced in real-world situations [4]. Due to the tremendous success of deep learning in computer vision, contemporary neural networks have been successfully applied to emotion recognition that works 'in the wild'. One of the most important factors of a progress in FER is an introduction of large datasets, e.g., AffectNet [5], and appearance of such challenges as EmotiW (Emotion Recognition in-the-Wild) [4] and ABAW (Affective Behavior Analysis in-the-Wild) [6, 7].

Though there exist a lot of accurate solutions for each particular dataset, there are two main obstacles for their practical usage. First of all, the best accuracy is achieved by complex models based on visual transformers, ensembles of classifiers and multi-modal techniques. Hence, they can hardly be launched by a non-specialist using only cheap laptops or mobile devices. Second, existing emotional datasets are very dirty because of the difficulties in labeling. Moreover, these datasets are an order of magnitude smaller than datasets used to train face recognition models [8]. As a result, the FER models contain the algorithmic bias: they learn features specific to a concrete dataset [6] and typically remain not robust to the diversity of recording conditions and environments [9].

To address the problems mentioned above, we proposed a novel software that provides accurate and high-speed emotion recognition, called HSEmotion. It can be used either to predict emotions using facial images and videos, or extract emotional features for down-stream tasks (engagement prediction, group-wise affect prediction, etc.). The most important part of this tool is a bank of convolutional neural networks (CNNs) pre-trained by a special technique introduced in our previous paper [3] by using cropped facial images. These models are applicable to deploy even to mobile devices. Their high efficiency has been shown by the participation in recent ABAW challenges [9, 10] and the state-of-the-art results among single models in several settings.

2. Description
The main usage scenarios of the HSEmotion software are shown in Fig.1. The tool supports analysis of either static images or facial videos. In the latter case, each video frame is processed independently. The faces are detected in an input image by an appropriate external library, such as MTCNN (Multi-task CNN). Next, the facial region is fed into one of our models [1, 3]. They can either output the high-dimensional visual embeddings (emotional features) from the penultimate layer or the posterior probabilities for eight emotions, namely, Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness and Surprise [5]. In the former case, extracted visual feature vectors may be fed into the classifier for an arbitrary down-stream task. We provide examples with training of machine learning models for FER in video using the AFEW (Acted Faces in the Wild) [4], group-level affect classification using the VGAF (Video-level Group
The HSEmotion repository contains the hsemotion package for Python3, Jupyter Notebooks for training the models and solving various emotion recognition tasks, the demo mobile application for Android platform (Fig. 2) and the model zoo with pre-trained CNNs for face recognition and FER. The latter are based on PyTorch implementations of EfficientNets [12] from the timm (Pytorch Image Models) library. The following models are available in the hsemotion package: enet_b0_8_best_vgaf, enet_b0_8_best_afew, enet_b0_8_va_mtl, and enet_b2_8. All these models were trained in a similar way. At first, the EfficientNet was fine-tuned on a face identification task using the VGGFace2 dataset with faces cropped by face detector without any margins. Next, the model is fine-tuned to recognize emotions on static images from the AffectNet [5]. The details of the training procedure are provided in the paper [3].

The first three models with prefix “enet_b0” are based on EfficientNet-B0 that processes facial images with resolution 224x224 and outputs embeddings of dimensionality 1280. The last model implements EfficientNet-B2 architecture with 260x260 input image and 1408 output features. The “enet_b0_8_va_mtl” differs from all other models by predicting not only 8 basic facial expressions but also the valence and arousal, i.e., how active or passive, positive or negative is the human behavior.

We evaluate the HSEmotion models over five large-scale real-world datasets, namely, AffectNet [5] with all 8 emotional categories and 7 of them except Contempt, AFEW [4] with 7 emotions, VGAF [11] with 3 affects (positive, negative and neutral), LSD (learning from synthetic data) with 6 expressions and MTL (multi-task learning) tasks from the ABAW4 challenge [7]. The testing protocols from the authors of these datasets were used, and all the results are reported in Table 1 for their validation sets. The implementation details about first three and last two tasks are available in our papers [1] and [9], respectively. Also, we evaluate inference time for our models on Samsung Fold 3 device with Qualcomm 888 CPU and Android 12 using PyTorch Mobile runtime.

These results show that our software has excellent performance. According to the “Papers with code”¹, our “enet_b2_8” model reached state-of-the-art accuracy for 8 classes from AffectNet. Moreover, it is only 0.2% less accurate when compared to EmotionGCN [13] for AffectNet with 7 basic emotions. The accuracy of “enet_b0_8_best_afew” is 4.1% greater when compared to previously known best single model for the AFEW dataset from EmotiW 2019 competition, namely, the noisy student with iterative training from [14]. Moreover, our recent paper [15] demonstrated that fusion of this model with the wav2vec acoustic model reaches state-of-the-art accuracy (67.88%) for this dataset, outperforming the previously-known best result of LResNet50E-IR [16] by 2.3%.

Our EfficientNet-B2 is also the top single model for the VGAF dataset: it is 5.5% more accurate than the best facial model from the winner [17] of the corresponding sub-

¹ https://paperswithcode.com/sota/facial-expression-recognition-on-affectnet
challenge of the EmotiW 2020. Moreover, the F1-score for the LSD task and PMTL metric for the MTL sub-challenge, which is a sum of macro-averaged F1-scores for facial expressions, action units and average concordance correlation coefficient for valence and arousal) are also among top-performers of the ABAW4 challenge\(^2\). In fact, an ensemble of “enet_b0_8_va_mtl” with its fine-tuned version is the best model among all participants of the LSD sub-challenge. Solutions of only two participants perform better than this model for the MTL competition, among which the best value \(P_{MTL} = 1.76\) on the validation set is a special network that learns a correlation between all tasks [18].

3. Impact
HSEmotion is a novel software employed to improve the accuracy and performance of FER. This will make it easier and quicker for the practitioners to include the state-of-the-art emotion recognition and affect behavior analysis techniques into their software. This will also result in significant simplification of new studies as our techniques may be considered as new baselines for many tasks such as video-based FER [3], group-level affect prediction [11, 19], engagement prediction in-the-wild [1], etc.

The efficiency of our models has been demonstrated in the previous papers [1, 3]. For example, the EfficientNet-B2 reached the state-of-the-art accuracy on validation set of AffectNet [5]. It is also the best single model for various sub-challenges of EmotiW competition [3]. With help of this software, the HSE-NN team took the 3rd place in the multi-task learning challenge and 4th places in Valence-Arousal and Expression challenges [10] at ABAW3 competition [6]. Finally, it took first place in the LSD task and the 3rd place in the MTL task [9] at the recent ABAW4 challenge [7].

HSEmotion models are widely used by either independent researchers or large corporations. For example, engineers of Noah’s Ark Lab (Huawei Technologies) propagated subtle face recognition features through our FER model [20]. The fine-tuning of our models on emotional data from a concrete user makes it possible to develop very accurate personalized systems with man-machine interaction [21]. Facial emotional features extracted with help of our models enhance the e-learning software by automatic prediction of students’ distraction [22]. Moreover, our results are used as references in the papers that include experimental studies with AffectNet, AFEW, VGAF, EmotiWild and ABAW datasets [6, 7].

Declaration of competing interest
The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements
The work is supported by RSF (Russian Science Foundation) grant 20-71-10010.

References:

\(^2\) https://ibug.doc.ic.ac.uk/resources/eccv-2023-4th-abaw/
http://dx.doi.org/10.1109/TAFFC.2022.3188390


http://dx.doi.org/10.1109/SISYS2021.2021.9582908

http://dx.doi.org/10.1145/3340555.3355710


http://dx.doi.org/10.1109/CVPRW56347.2022.00259


http://dx.doi.org/10.1145/3340555.3355713


22. Andrey V. Savchenko, Ilya Makarov, Neural Network Model for Video-Based Analysis of Student’s Emotions in E-Learning, Optical Memory and Neural Networks (Information Optics), 2022, vol. 31, no. 3, pp. 237-244. http://dx.doi.org/10.3103/S1060998X22030055
Fig. 1 Typical pipeline for usage of HSEmotion tool. This cover has been designed using images from Freepik.com

Fig. 2 Sample screen of Android demo mobile application
### Tables

Table 1. Performance of the models from HSEmotion package

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy, %</th>
<th>F1-score</th>
<th>P&lt;sub&gt;MTL&lt;/sub&gt;</th>
<th>Inference time, ms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AffectNet (8 classes)</td>
<td>AffectNet (7 classes)</td>
<td>AFEW</td>
<td>VGAF</td>
</tr>
<tr>
<td>enet_b0_8_best_afew</td>
<td>60.90</td>
<td>64.71</td>
<td>59.89</td>
<td>66.80</td>
</tr>
<tr>
<td>enet_b0_8_best_vgaf</td>
<td>61.33</td>
<td>64.57</td>
<td>55.14</td>
<td>68.29</td>
</tr>
<tr>
<td>enet_b0_8_va_mtl</td>
<td>61.93</td>
<td>64.97</td>
<td>56.73</td>
<td>66.58</td>
</tr>
<tr>
<td>enet_b2_8</td>
<td>63.03</td>
<td>66.29</td>
<td>57.78</td>
<td>70.23</td>
</tr>
</tbody>
</table>
Highlights (for review)

- HSEmotion provides fast and accurate facial expression recognition models
- HSEmotion extracts emotional features for arbitrary affect behavior analysis tasks
- HSEmotion is deployed to mobile devices for real-time emotion recognition in video
Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Andrey Savchenko, Dr. of Science
Full Professor in Department of Information systems and technologies,
Senior researcher in Laboratory of Algorithms and Technologies for Network Analysis,
HSE University, Nizhny Novgorod, Russia