A Multi-resolution Training for Expression Recognition in the Wild

(Discussion Paper)

Fabio Valerio Massoli[1], Donato Cafarelli[1], Giuseppe Amato[1] and Fabrizio Falchi[1]

[1]ISTI-CNR, via G. Moruzzi 1, 56124 Pisa, Italy

Abstract

Facial expressions play a fundamental role in human communication, and their study, which represents a multidisciplinary subject, embraces a great variety of research fields, e.g., from psychology to computer science, among others. Concerning Deep Learning, the recognition of facial expressions is a task named Facial Expression Recognition (FER). With such an objective, the goal of a learning model is to classify human emotions starting from a facial image of a given subject. Typically, face images are acquired by cameras that have, by nature, different characteristics, such as the output resolution. Moreover, other circumstances might involve cameras placed far from the observed scene, thus obtaining faces with very low resolutions. Therefore, since the FER task might involve analyzing face images that can be acquired with heterogeneous sources, it is plausible to expect that resolution plays a vital role. In such a context, we propose a multi-resolution training approach to solve the FER task. We ground our intuition on the observation that, often, face images are acquired at different resolutions. Thus, directly considering such property while training a model can help achieve higher performance on recognizing facial expressions. To our aim, we use a ResNet-like architecture, equipped with Squeeze-and-Excitation blocks, trained on the Affect-in-the-Wild 2 dataset. Not being available a test set, we conduct tests and model selection by employing the validation set only on which we achieve more than 90% accuracy on classifying the seven expressions that the dataset comprises.

Keywords

Facial Expression Recognition, Deep Convolutional Neural Networks, Multi-resolution training.

1. Introduction

Facial expressions play a fundamental role in human communication. Indeed, they typically reveal the actual emotional status of people beyond the spoken language. Moreover, the comprehension of human affect based on visual patterns is a crucial ingredient for any human-machine interaction [1] system and, for such reasons, the task of Facial Expression Recognition (FER) draws both scientific and industrial interest. In recent years, Deep Learning techniques reached very high performance on FER by exploiting different architectures and learning paradigms. In such a context, we propose a multi-resolution approach to solve the FER task.

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy

fabio.massolli@isti.cnr.it (F. V. Massoli); donato.caf@gmail.com (D. Cafarelli); giuseppe.amato@isti.cnr.it (G. Amato); fabrizio.falchi@isti.cnr.it (F. Falchi)

0000-0001-6447-1301 (F. V. Massoli); 0000-0002-7575-0143 (D. Cafarelli); 0000-0003-0171-4315 (G. Amato); 0000-0001-6258-5313 (F. Falchi)

We ground our intuition on the observation that, often, face images are acquired at different resolutions. Thus, directly considering such property while training a model can help achieve higher performance on recognizing facial expressions. To our aim, we use a ResNet-like architecture, equipped with Squeeze-and-Excitation blocks, trained on the Affect-in-the-Wild 2 dataset. Not being available a test set, we conduct tests and model selection by employing the validation set only on which we achieve more than 90% accuracy on classifying the seven expressions that the dataset comprises. To let our researcher reproduce our results, we made our code publicly available on github[1]. Concerning the remaining part of the paper, we organized it as follows. In Section 2 we report several works related to the FER task, while in Section 3 and Section 4 we describe our approach and the dataset we use, respectively. Moreover, we describe the experimental campaigns we perform and the corresponding model performance in Section 5. Finally, in Section 6 we conclude our work by reporting our plans.

2. Related Works

Nowadays, the most promising approaches to the FER task are based on the use of Deep Convolutional Neural Networks (DCNN). A typical approach consists of a pre-processing phase, where the images are subject to various transformations, and a training phase where these images are iteratively given as input to a DCNN model for feature extraction and expression classification.

In [2], the authors propose a new approach for face cropping to remove useless regions in an image and a novel rotation strategy to cope with data scarcity. Furthermore, they built a simplified DCNN structure to reduce training/inference time and achieve real-time FER on devices with limited resources. Their experiments were conducted on two databases, CK+ [3] and JAFFE [4], and achieved state-of-the-art results. In [5], a novel activation function based on the ReLU function, called LS-ReLU, is presented. It exploits an adjustable log and the soft-sign functions. Neural networks based on LS–ReLU function can avoid the over-fitting problem during the training process and reduce the oscillations problem. Their experiments on JAFFE [4] and FER2013 [6] datasets showed that a DCNN based on this novel activation function has a better performance compared to most state-of-the-art activation functions. With the transition of FER datasets from laboratory-controlled to in-the-wild conditions, this task has become more challenging due to variations in pose, brightness, and background, to mention some. Therefore, in [7] the authors focus on resolving the FER task by analyzing the contribution of different face areas to different emotions, including nose, mouth, eyes, nose to mouth, nose to eyes, and mouth to eyes areas, together with the whole faces. The paper [8] addresses the problem of the class imbalance in wild FER datasets. To such an aim, the authors propose a novel Discriminant Distribution-Agnostic loss (DDA loss) to optimize the embedding space for extreme class imbalance scenarios. Specifically, DDA loss enforces inter-class separation of deep features for both majority and minority classes. In [9] the authors propose a multi-task learning framework to extract local-global and spatio-temporal information for a discriminative and robust representation of facial expressions. Their experiments achieved competitive results on the CK+ [3] and Oulu-CASIA [10] datasets. To improve the performance on the FER

[1] https://github.com/fvmassoli/affwild2-challenge.git

task, [11] proposes a novel “Masking Idea" that is implemented in a Residual Masking Network that contains several masking blocks applied across residual layers to improve the network's attention ability on relevant information. Experiments showed competitive results on the FER2013 [6] dataset.

3. Approach

Usually, face images come from heterogeneous sources [12], e.g., cameras with different resolutions or different distances from the scene. Such characteristics directly impact DL models' performance on tasks such as Face Recognition (FR) by dramatically lowering their performance [13]. Based on such an observation, we propose our approach grounded on the hypothesis that the images' resolution has a non-negligible impact on DL models' behavior when tested against the FER task. Precisely, we move our steps from [13] in which the authors explicitly take care of the multi-resolution nature of face images by designing a training technique to accommodate for such an issue adequately.

In our work, we take inspiration from the author's training procedure, and we adapted it to our case. Specifically, we experimentally notice that we do not need any Teacher-supervised signal nor curriculum learning. Thus, we simplify the training procedure by only exploiting the double random extraction to set the final image resolution. To train the models and perform model selection, we employ the Aff-Wild2 [14] dataset. We refer the reader to Section 4 for a brief description of the dataset.

Our base model is a ResNet-50 architecture [15], equipped with Squeeze-and-Excitation blocks [16], that has been pre-trained on the VGGFace2 dataset [17]. To train our models, we use the Adam [18] optimizer. We set the weight decay of $1.e^{-4}$ and the learning at $1.e^{-3}$ for the last fully connected layer and $1.e^{-4}$ for all the others. Moreover, we set the batch size to 128, and we use data augmentation techniques to avoid overfitting. Specifically, we first resize the images to have the shortest side of 256 pixels (while keeping the original aspect ratio), then we random crop a square of 224x224 pixels, and finally, we normalize the input channels. Moreover, we apply a random grayscale conversion with a probability of 0.2. We substitute the random crop with the center one, and we remove the grayscale operation to test the model on the validation set.

Concerning the random resolution extractions to train the models, we perform several experiments considering different ranges for the final image size concerning the multi-resolution training, with the minimum and maximum considered values being 8 and 256 pixels, respectively.

4. Dataset: Affect-in-the-Wild 2

The Aff-Wild2 [14] dataset is the first-ever database annotated for all three main behavior tasks: Valence Arousal (VA), Action Units (AU), and Expression (EX) classification. Concerning the last one, the dataset consists of 547 videos (collected from YouTube) that account for $\sim$2.6M of frames labeled considering seven basic expressions: neutral, anger, disgust, fear, happiness, sadness, and surprise. The annotation is made frame-by-frame by a team of seven experts. The dataset is shipped with a protocol that divides it into three non-overlapping subsets for training,

validation, and test purposes. Specifically, the three partitions consist of 253, 71, and 223 videos, respectively. The cropped-aligned version of the dataset is made of images preprocessed to have a fixed resolution of 112x112 pixels. Among the ∼2.6M available images, ∼1.2M are available for training and validation on the FER task. We report in Figure 1 an example of training images in the Aff-Wild2 [14] dataset.

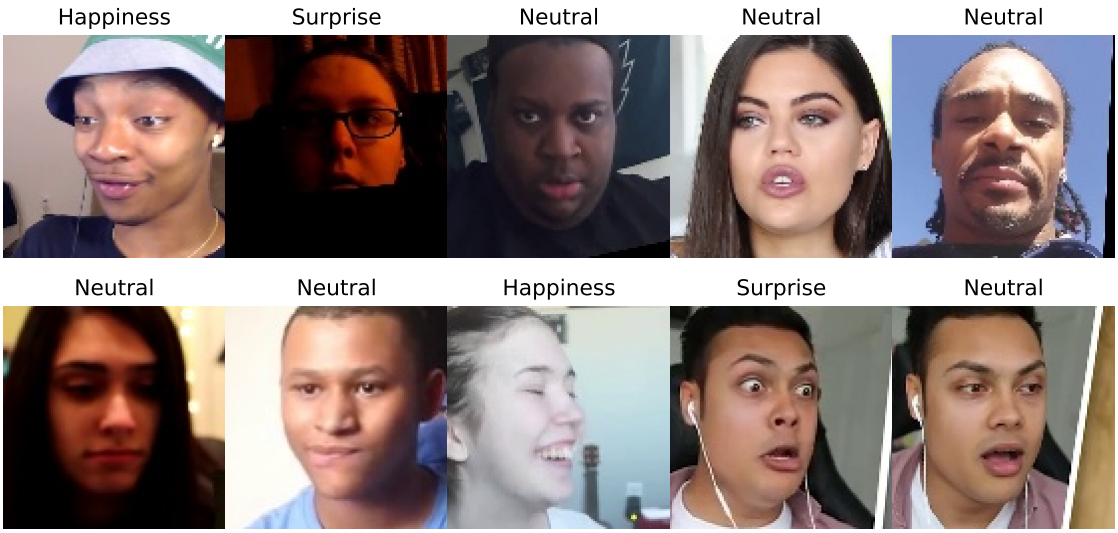

Figure 1: Example of face images from the Aff-Wild2 [14] dataset. On top of each image, we report the corresponding ground truth expression.

As we mentioned previously, the dataset comprises seven different types of expressions with a very different cardinality. In Table 1, we report the number of images for each class, both for the training and validation sets, while in Table 2, we report the classes' weight concerning the training images only.

	Expression						
	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Training	585896	23484	12497	11120	149920	100548	38564
(%)	(63.5)	(2.5)	(1.4)	(1.2)	(16.3)	(11.0)	(4.1)
Validation	181884	8003	5401	9671	52842	38534	22988
(%)	(57.0)	(2.5)	(1.7)	(3.0)	(16.5)	(12.1)	(7.2)

Table 1
Classes' cardinality for the Aff-Wild2 [14] dataset.

As one can notice from Table 1, the classes are not balanced. For that reason, we leveraged a balanced cross-entropy loss to account for the class unbalance. To such an aim we use the weights reported in Table 2.

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Aff-Wild2 [14]	0.365	0.975	0.986	0.988	0.837	0.891	0.958

Table 2
Classes' weight. The values reported are referred to the training set only. Note that weights do not need to sum up to one. The lower the weight, the higher the cardinality of the corresponding class.

5. Experimental Results

In this section, we report the experimental results we obtained on the Aff-Wild2 [14] dataset. Since the dataset is currently employed in the Affect-in-the-Wild Challenge [19], the test set's ground truth labels are not available. For such a reason, we quote the performance of our model on the validation set. Before the training, we took a small subsample of the validation set and used it for model selection purposes to avoid bias. Subsequently, we tested the best model on the entire validation set. To quote our results, we use different metrics. First, we evaluate the F1-score on each class, then we summarize the overall performance of our best model across all the seven expressions by quoting the F1-score (macro-average) and the overall accuracy. Finally, we evaluate the same score as required by the Affect-in-the-Wild Challenge [19], which is equal to:

$$s = 0.33 \cdot \text{accuracy} + 0.67 \cdot \text{f1 score}; \tag{1}$$

where the accuracy and the f1 score are relative to the whole dataset.

We report the results in Table 3 and Table 4 concerning single classes and the whole dataset, respectively.

	Expression						
	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
F1 Score	0.978	0.960	0.965	0.971	0.946	0.987	0.937

Table 3
F1 score for each class of the Aff-Wild2 [14] dataset.

Accuracy	**F1 Score (macro-average)**	**Challenge Score**
0.970	0.964	0.966

Table 4
Summary statistics on all the classes of the Aff-Wild2 [14] dataset.

From the previous tables, we can notice that our model shows promising performance on the FER task. Moreover, we acknowledge the stability of the scores among different classes even though the dataset is highly unbalanced as reported in Table 1

6. Conculsions abd Future Works

In this work, we report our first experimental campaign focused FER task. We tackle such a problem by giving more representational power to our models, assuming a multi-resolution context, and we observe promising results. As a next step, we will extend our experimental campaign to test our approach on different publicly available datasets such as FER2013 [6], RAF-DB [20], and Oulu-CASIA [10].

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research. This work was partially supported by WAC@Lucca funded by Fondazione Cassa di Risparmio di Lucca, AI4EU - an EC H2020 project (Contract n. 825619), and upon work from COST Action 16101 "Action MULTI-modal Imaging of FOREnsic SciEnce Evidence (MULTI-FORESEE)", supported by COST (European Cooperation in Science and Technology).

References

[1] V. Bettadapura, Face expression recognition and analysis: The state of the art, CoRR abs/1203.6722 (2012). URL: http://arxiv.org/abs/1203.6722. arXiv:1203.6722.

[2] K. Li, Y. Jin, M. W. Akram, R. Han, J. Chen, Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy, The Visual Computer 36 (2020) 391–404.

[3] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: 2010 ieee computer society CVPR-workshops, IEEE, 2010, pp. 94–101.

[4] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets, in: Proceedings Third IEEE international conference on automatic face and gesture recognition, IEEE, 1998, pp. 200–205.

[5] Y. Wang, Y. Li, Y. Song, X. Rong, The influence of the activation function in a convolution neural network model of facial expression recognition, Applied Sciences 10 (2020) 1897.

[6] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., Challenges in representation learning: A report on three machine learning contests, in: International conference on neural information processing, Springer, 2013, pp. 117–124.

[7] Z. Lian, Y. Li, J.-H. Tao, J. Huang, M.-Y. Niu, Expression analysis based on face regions in real-world conditions, International Journal of Automation and Computing 17 (2020) 96–107.

[8] A. H. Farzaneh, X. Qi, Facial expression recognition in the wild via deep attentive center loss, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2402–2411.

[9] M. Yu, H. Zheng, Z. Peng, J. Dong, H. Du, Facial expression recognition based on a multi-task global-local network, Pattern Recognition Letters 131 (2020) 166–171.
[10] G. Zhao, X. Huang, M. Taini, S. Z. Li, M. PietikäInen, Facial expression recognition from near-infrared videos, Image and Vision Computing 29 (2011) 607–619.
[11] P. Luan, V. Huynh, T. Tuan Anh, Facial expression recognition using residual masking network, in: IEEE 25th International Conference on Pattern Recognition, 2020, pp. 4513–4519.
[12] F. V. Massoli, F. Falchi, C. Gennaro, G. Amato, Cross-resolution deep features based image search, in: International Conference on Similarity Search and Applications, Springer, 2020, pp. 352–360.
[13] F. V. Massoli, G. Amato, F. Falchi, Cross-resolution learning for face recognition, Image and Vision Computing 99 (2020) 103927.
[14] D. Kollias, S. Zafeiriou, Aff-wild2: Extending the aff-wild database for affect recognition, arXiv:1811.07770 (2018).
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.
[16] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE CVPR, 2018, pp. 7132–7141.
[17] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age. corr abs/1710.08092 (2017), arXiv:1710.08092 (2017).
[18] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980 (2014).
[19] D. Kollias, S. Zafeiriou, First Affect-in-the-Wild Challenge, https://ibug.doc.ic.ac.uk/resources/first-affect-wild-challenge/, 2020.
[20] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: Proceedings of the IEEE CVPR, 2017, pp. 2852–2861.