Unique Class Group Based Multi-Label Balancing Optimizer for Action Unit Detection

Ines Rieger1,*, Jaspar Pahl1,*, Dominik Seuss1
1 Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
{ines.rieger, jaspar.pahl, dominik.seuss}@iis.fraunhofer.de

Abstract—Balancing methods for single-label data cannot be applied to multi-label problems as they would also resample the samples with high occurrences. We propose to reformulate this problem as an optimization problem in order to balance multi-label data. We apply this balancing algorithm to training datasets for detecting isolated facial movements, so-called Action Units. Several Action Units can describe combined emotions or physical states such as pain. As datasets in this area are limited and mostly imbalanced, we show how optimized balancing and then augmentation can improve Action Unit detection. At the IEEE Conference on Face and Gesture Recognition 2020, we ranked third in the Affective Behavior Analysis in-the-wild (ABA W) challenge for the Action Unit Recognition 2020, we ranked third in the Affective Behavior Analysis in-the-wild (ABA W) challenge for the Action Unit detection task.

I. INTRODUCTION

Imbalanced datasets present a major challenge when training models for a multi-label multi-class problem such as facial expression recognition, text or music label categorization. Facial expression datasets often contain underrepresented classes, since different expressions do not occur naturally with the same frequency. But in contrast to single-label data, resampling methods such as simple over- or undersampling that clone over- or underrepresented samples with a majority or minority label do not work, as they would also increase or decrease potential other labels in the same sample as well.

Facial expressions such as physical pain [21] or the basic emotions anger, disgust, fear, happiness, sadness, and surprise [6] can be described by Action Units. Action Units (AUs) are distinct facial movements, which are described in the Facial Action Coding System (FACS) [7]. There are several approaches to automatically detect the occurrence and intensity of AUs [23], [24]. Deep learning models are one of the most successful approaches, but require large amounts of data to perform well. But the manual annotation time for AU occurrences and intensity is high and must be conducted by expert FACS-coders. Thus, the training data for AUs is limited. Furthermore, AU datasets are mostly imbalanced as different AUs do not occur with the same frequency for different facial expressions.

We propose a fast and simple, yet effective balancing and augmenting approach framed as optimization problem to compensate for imbalanced and scarce occurrences in a multi-label dataset. We apply this algorithm on the use case of Action Unit detection and show that our balancing approach can enhance the detection of AUs.

II. RELATED WORKS

A. Balancing Multi-Label Data

Single-label resample techniques are often extended to be also used for multi-label data [29]. Chart et al. [2] propose the LP-RUS and LP-ROS algorithms for random under- and oversampling. In our approach, we first transform the multi-label dataset in a multi-class dataset, handling each distinct combination of labels as one labelset class. They first determine minority and majority labelsets, which include a label belonging to an under- or over represented class. Then, these classes are over- or undersampled by cloning or removing samples until they reach the average mean size of the labelset classes. Chart et al. [2] further propose the ML-RUS and ML-ROS algorithms for under- and oversampling, where they evaluate the individual imbalance level of each sample and thus blocking certain samples to be removed or cloned.

MLSMOTE (Multilabel Synthetic Minority Over-sampling Technique) [3] is an extension of the SMOTE [4] algorithm in order to oversample minority classes. The SMOTE algorithm creates new samples by taking a new instance among the nearest neighbors of the desired underrepresented sample. As Chart et al. [2], they determine minority labelset classes. Chart et al. [3] create new samples for these by using the label correlation information. By experiments, they show that MLSMOTE outperforms the ML-ROS algorithm.

In contrast to these approaches, we formulate the balancing of datasets as an optimization problem and then use slight augmentation techniques to increase the number of samples.

B. Action Unit Detection

The following approaches also use the AffWild2 validation and testing dataset for evaluation (see Section V). Kollias et al. [14] use a MobileNetV2 [26] as backbone for Action Unit Detection. Deng et al. [5] propose a multitask model for Valence and Arousal, Emotion and Action Unit Detection. They use partial labels in a teacher and student training setting. For tackling the imbalanced training dataset, Deng et al. [5] employ the ML-ROS algorithm [2]. Kuhnke et al. [20] propose a two-stream multitask model for detecting Valence and Arousal, Emotions and Action Units. They use features
Fig. 1: This shows our (a) balancing pipeline and (b) training backbone.

from an aural and a visual stream employing spatial and
temporal convolutions.

III. DATASETS

We use the Aff-Wild2 [16], [18], [17], [15] as training,
validation and test dataset and as additional training datasets
the Emotionet [8], the Actor Study [27] and the Extended
Cohn-Kanade (CK+) [12], [22].

The Aff-Wild2 database is the largest in-the-wild database
consisting of videos and is partly annotated for the three tasks
valence-arousal estimation [28], basic expression recognition
[13] and Action Unit detection. All videos were scraped
from YouTube. For Action Unit detection, 56 videos with
63 subjects (32 males and 31 females) were annotated by
experts. In total, 398,835 frames are annotated. For the
challenge, this dataset is divided in training, validation and
testing. This dataset is for non-commercial research purposes
only.

The Emotionet database consists of approximately one
million images scraped from the web using emotive key-
words. Benitez-Quiroz et al. [9] analyzed the occurrence and
intensity of AUs in 90% of the images automatically. Experts
FACS-coded 10% of the dataset manually. We only use the
manually coded part of the dataset for our experiments. This
dataset is for non-commercial research purposes only.

The Actor Study contains sequences of 21 actors, in total
68 minutes of video filmed from different views and camera
speed. Each actor had the task to display specific Action
Units in different intensities and had to respond to scenarios
and enactments. The frames are annotated with Action Units
and their corresponding intensities. We use the center view
of the low speed camera for our approach. This dataset was
recently published and will be made publicly available for
commercial research.

The CK+ dataset contains 593 videos of 123 subjects. The
sequences show the facial expressions from neutral to strong.
All expressions are posed. The presence or absence of the
Action Units is coded for the peak frames, the strongest
expression, only. This dataset is for non-commercial research
purposes only.

IV. METHODS

A. Pre-processing

For the AffWild2 training, validation, and testing dataset
we use the cropped and aligned images of 112 x 112 pixel
size provided by the ABAW 2020 challenge in order to
eliminate sources of error and make our approach easier to
compare with others. Please note that there is noise in this
pre-cropped dataset, as the main person is not identified at
all frames correctly and thus, the annotated labels do not
match the facial image. The remaining training sets had
no pre-cropped images available. For the CK+ dataset and
the Actor Study dataset the Sophisticated High-speed Object
Recognition Engine SHORE™ [19] is used to crop images
of the same size. The Emotionet dataset is cropped using
OpenFace [1]. All images were converted to gray scale and
pixel values were normalized in the range of [0, 1].

B. Balancing

Combining all datasets from Section III yields a highly
imbalanced dataset (see “non-aug” in Table I). Imbalanced
datasets cause problems in the detection of under-represented
Action Units. Therefore, we implement a class balancing
algorithm (Fig. 1) which increases the allowed number of
augmentations in the under-represented Action Units. In a
multi-label multi-class problem, the occurrences of a class
cannot be increased by simply augmenting a single image
multiple times because the potential other classes in the
image would be increased by the same amount. Therefore,
we formulate the class balancing as an optimization problem:

\[
f(n_u) = \sum |z - \frac{N_u}{\sum_{i=1}^{N_u} z_i}| + \lambda Var(\frac{n_u}{N_u}) \tag{1a}
\]

\[
z = \sum_{r} n_r u_r * y_{ru_r}, \tag{1b}
\]

where \(n_u\) is the vector of number of occurrences of the
images with a unique Action Unit combination and \(y_{ru}\) is
the respective one-hot-encoded array of labels for the unique
Fig. 2: Convergence of balance optimization with $\lambda = 70$ and $n_o < n_u < 10 n_o$.

labelsets. The operator $\ast$ between $n_u_r$ and $y_u_r$ denotes a scalar multiplication for each row $r$. The variable $n_o$ is the vector with the original occurrences in the dataset non-aug, and $\lambda$ is a weighting parameter. We used $\lambda = 70$ as weighting value for our experiments (Fig. 2). Furthermore, we restricted the search domain for $n_u$ as:

$$n_o < n_u < 10 n_o.$$  \hfill (2)

As such, we can control the number of additional augmentations while stopping the optimizer from increasing the relative growth of a single unique label combination too much. The latter would cause problems due to too many augmentations on a single image, which potentially leads to overfitting.

We go through the images and augment them according to the optimized occurrences in their respective labelset class until the max occurrence is reached. The augmentations are implemented using the imgaug library [11]. We use Flipping Jr, Gaussian Blur, Linear Contrast, Additive Gaussian Noise, Multiply, and Perspective Transform. These slight augmentation methods are suited for Action Units, as the labels are still valid after the augmentation.

Table I shows the label distribution of the original training datasets and of the augmented dataset.

| Dataset | AU01 | AU02 | AU04 | AU06 | AU12 | AU15 | AU20 | AU25 |
|---------|------|------|------|------|------|------|------|------|
| non-aug | 47548 | 2271 | 32387 | 9290 | 22964 | 1537 | 3490 | 7463 |
| aug-1   | 5899 | 0 | 676 | 0 | 0 | 0 | 0 | 0 |
| Acc     | 2964 | 1875 | 3753 | 2475 | 1654 | 2115 | 1406 | 5359 |
| F1      | 11079 | 8553 | 11393 | 7798 | 215 | 2115 | 4704 | 13573 |
| AWC     | 62180 | 12699 | 48209 | 18455 | 33091 | 17748 | 28175 | 35416 |

C. Evaluation Metric

The results are measured in F1 score (3a) for each Action Unit. Furthermore, the results are measured according to the ABAW Challenge metric [14], denoted here as AWC (3b). AWC is the unweighted average of the accuracy Acc (3b) and the F1 macro score. The variable $N$ stands for all predictions.

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$ \hfill (3a)

$$\text{Acc} = \frac{N_{\text{correct predictions}}}{N}$$ \hfill (3b)

$$\text{AWC} = \frac{1}{2} \left( \frac{\sum_i N_i \text{AU}_i F1_i}{N_{\text{AU}}} + \text{Acc} \right)$$ \hfill (3c)

D. Training

Our backbone is a parameter-reduced 18-layer ResNet with input images of size 112 x 112 pixels (ResNet18-112) [25] (Fig. 1b). This modified ResNet is based on He et al. [10]. We use the ReLU activation function and changed the output layer to a sigmoid activation function with eight outputs according to the number of Action Units. The threshold for a positive detection is 0.5. We use a weighted F1 loss function with an Adam optimizer and a learning rate of 0.0001. These parameters have been determined by grid search.

V. RESULTS

Table III shows our results on the Aff-Wild2 validation dataset. The results show that our balanced augmentation algorithm can enhance the results of a multi-label problem like Action Unit Detection. Our model trained on the augmented dataset aug-1 performs better at detecting all Action Units equally, as classes are now detected that could not be detected before and the overall F1 macro score increases from 0.21 to 0.24.

Our trained models supersede by far the baseline model of Kollias et al. [14], who are also using a single-task model. Our model trained on the augmented dataset aug-1 performs slightly worse in comparison to the multitask models of Deng et al. [5] and Kuhnke et al. [20], who were ranked first and second in the ABAW 2020 challenge. In contrast to Deng et al. [5] and Kuhnke et al. [20], we use the noisy
precropped AffWild2 dataset, which has a negative effect on the performance. But what is apparent is that our model trained on aug-1 generalizes better on the testing data than the model trained on the original dataset non-aug. There is 0.5 difference between the F1 macro scores of the validation and testing dataset of the model trained on the non-aug dataset and 0.2 difference of the model trained on the aug-1 dataset.

TABLE II: Our results on the Aff-Wild2 validation dataset. The results of the AUs are measured in F1 Score. Best results are in bold.

| Model       | ResNet18-112 (ours) | ResNet18-112 (ours) |
|-------------|---------------------|---------------------|
| Train Set   | non-aug             | aug-1               |
| AU01        | 0.64                | 0.71                |
| AU02        | 0.00                | 0.00                |
| AU04        | 0.47                | 0.36                |
| AU06        | 0.21                | 0.33                |
| AU12        | 0.39                | 0.41                |
| AU15        | 0.00                | 0.10                |
| AU20        | 0.00                | 0.04                |
| AU25        | 0.00                | 0.00                |
| F1 macro    | 0.21                | 0.24                |
| Acc         | 0.94                | 0.93                |
| AWC         | 0.58                | 0.59                |

TABLE III: Comparable results on the Aff-Wild2 validation and testing dataset. Best results are in bold, second best results are in brackets.

| Dataset     | F1Macro Validation | AWC Validation | F1Macro Testing | AWC Testing |
|-------------|-------------------|---------------|----------------|-------------|
| baseline [14] | -                 | 0.31          | -              | 0.26        |
| twostream [20] | -                 | [0.59]        | [0.27]         | 0.93 [0.60] |
| multitask [5] | -                 | 0.63          | 0.31 [0.91]    | 0.61        |
| non-aug (ours) | 0.21              | 0.58          | 0.16 [0.91]    | 0.53        |
| aug-1 (ours)  | 0.24              | [0.59]        | 0.22 0.89      | 0.55        |

VI. CONCLUSIONS AND FUTURE WORKS

In this paper we present an approach based on a multi-label class balancing algorithm as a pre-processing step to overcome the imbalanced occurrences of Action Units in the training dataset. We train a ResNet with 18 layers and show an improvement in detecting single AUs better and generalizing better on unseen data when using the augmented training dataset in contrast to the original training dataset. For future work, we plan on enriching our approach with a multi-task framework in order to improve our results. Furthermore, we plan on comparing our results on the aug-1 dataset to results on even further augmented datasets created with a higher augmentation rate, i.e. different weightings and compare our results thoroughly with other resampling algorithms.
expression. In 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, pages 94–101. IEEE, 2010.

[23] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. IEEE transactions on affective computing, 2017.

[24] Ines Rieger, René Kollmann, Bettina Finzel, Dominik Seuss, and Ute Schmid. Verifying deep learning-based decisions for facial expression recognition (accepted). In Proceedings of the ESANN Conference 2020, 2020.

[25] Ines Rieger et al. Towards Real-Time Head Pose Estimation: Exploring Parameter-Reduced Residual Networks on In-the-Wild Datasets. In Conf. Proc. IEA/AIE, pages 123–134. Springer, 2019.

[26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4510–4520, 2018.

[27] Dominik Seuss, Anja Dieckmann, Teena Hassan, Jens-Uwe Garbas, Johann Heinrich Elgring, Marcello Mortillaro, and Klaus Scherer. Emotion expression from different angles: A video database for facial expressions of actors shot by a camera array. In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 35–41. IEEE, 2019.

[28] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal in-the-wild challenge. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pages 1980–1987. IEEE, 2017.

[29] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering, 26(8):1819–1837, 2013.