Optimization of the Intensity Aware Loss for Dynamic Facial Expression Recognition

Davy Tec-Hinh Lau [1], Jian-Jiun Ding [2], and Guillaume Muller [3]
National Taiwan University, Taipei, Taiwan
r12942162@ntu.edu.tw +33-783258324
National Taiwan University, Taipei, Taiwan
jjding@ntu.edu.tw +886-2-33669652
Ecole des Mines de Saint-Etienne, Saint-Etienne, France
guillaume.muller@emse.fr +33-0477420271

Abstract— **Dynamic facial expression recognition (DFER) is a challenging topic in computer vision. One of the critical problems in DFER is the variation in intensity in the different scenes. Different intensities lead to different traits and in the low intensity scenarios most emotions will be misclassified as neutral expression. The intensity aware loss has been introduced as a method to put more focus on the low intensity samples and better capture the traits of low intensity cases to acquire better discriminative abilities. In this work, an alternative way to define the intensity aware loss is proposed. With the use of the logarithmic Euclidian distance, the cross-entropy loss, and the exponential decrease model, and model combination, the discriminative ability in the low-intensity case can be further improved. Moreover, several ways to further optimize the model was proposed. Experiments show that, with these proposed techniques, even better performance of DFER can be achieved.**

I. Introduction

Facial Expression Recognition is one of the most popular fields in computer vision now, because of the need of better human-computer interactions which requires the computer to understand how humans feel. After the failure of Static Facial Expression Recognition (SFER) – which provides results with high accuracy for single images within lab conditions – the focus is now on Dynamic Facial Expression Recognition (DFER), which aims to classify videos into expressions. There are a lot of difficulties in DFER because of the nature of the samples. Indeed, in the datasets (like DFEW) that has been created to develop DFER, most of the samples are in 'in-the-wild' conditions, which means that a lot of parameters (head motion, head angle, lighting, obstruction, intensity of the expression) impact the classification.

New models have been developed to tackle these problematics: Former DFER [4] introduces the use of transformer for DFER, which allows the model to have a better understanding of which channels are the more relevant for the prediction thanks to the attention mechanism. M3DFEL [10] is a multi-instance learning model, the principle is to remove the noisy frames that will have a bad influence on the classification, the frames are grouped into instances and an instance is considered noisy if all the frames within this instance are negative, in which case this instance will not be considered in the training. NR-DFERNet [12] is also a denoising method, the model captures the dynamic-static features of each frame, then each frame's features are compared to the average and the influence of one frame is determined by the gap between its features and the mean. The bigger the gap is, the lower is its influence on the training process. AEN [6] uses emotion grouping, the prediction is done in 2 steps: the first prediction is the group of emotion (positive, neutral, negative) and the second prediction is done within the group that has been predicted in the previous step. The positive group contains (happy, surprise), the neutral group contains only (neutral) and the negative one contains (sad, angry, disgust, fear). MIDAS [5] is a data augmentation method; data augmentation is useful in DFER because the number of samples are very limited and for example data in DFEW are imbalanced which cause a bias in the training.

In [1], Li *et al* introduced the intensity aware loss to tackle the problem of intensity of the expression. Indeed, they notice that all the emotions tend towards the neutral emotion when the intensity tends towards 0. Therefore, when the intensity of the emotion is low the emotions are more likely to be misclassified because the traits are more subtle and have more resemblance with the other emotions, in other words depending on the intensity the intra-class difference is bigger than the inter-class difference which will cause a bias in the classification. The intensity aware loss puts more focus on those low intensity samples to try to prevent this phenomenon. Nevertheless, the loss function that is proposed by [1] also influences the high intensity samples which are supposed to be clear predictions, therefore this can influence negatively the classification training. Also, the function put more focus on the low intensity samples but we believe that the value that it takes is not optimized and therefore the function can be improved to better assess the purpose.

To tackle these 2 issues, we design a new loss function that takes inspiration from the principle proposed by [7]. This loss function will have a influence closer to 0 when the intensity is very high and therefore the prediction is very clear, and will put more focus than the original Intensity Aware Loss on the low intensity samples.

We also design 3 ways to further optimize this loss function. Firstly, we decrease its influence on the training over the epochs because once the model has learnt to discriminate the emotions with different intensities, its influence will negatively impact the training. Then, we implement a pretrained model to tackle the data imbalance issue and lastly, we take inspiration from the MAN model (Mining Ambiguity and Noise) to combine 2 models in order to take the strength of both models to reduce the uncertainty.

II. Intensity Aware Loss and MAN Models

The Intensity Aware Loss is used to put more focus on the low intensity samples during the training and is defined as:

$$L_{IA} = -\log(P_{IA}), \tag{1}$$

$$P_{IA} = \frac{e^{x_t}}{e^{x_t + e^{x_{max}}}} \tag{2}$$

where x_t is the target logit and x_{max} is the largest logit excluding the target. The intensity is calculated by the gap between x_t and x_{max}. After analyzing the function, we can notice that the values of L_{IA} are included between [0.14;0.3], which represents the max and min values for the highest intensity and the lowest. We notice that even when the intensity is at its highest, the loss function has an influence that is equal to almost half of its influence when the intensity is at the lowest. Therefore, the loss function impacts the model even when it should not because when the intensity is high the prediction is clear. Also, the values for the low intensity samples are pretty low and it could be more emphasized.

The mining ambiguity and noised (MAN) model has been developed to tackle the issue of the noise in the labeling. Indeed, it is sometimes hard even for humans to tell what expression the subject is trying to make. Therefore, it is possible that the labelling is wrong, which will make the model learn bad material and the classification will be worse. They divide the annotations in 3 categories: clean, ambiguous and noisy. In order to separate the annotations, they make 2 different models predict the emotions. If both of them predict the target emotion, the annotation is considered clean. If only one of them manage to predict the target emotion, the annotation is ambiguous and finally if none of them predict the target then it is a noisy annotation. Then to reduce the impact of the noise in the labelling, different strategies are applied according to the category of the annotation. We will take inspiration from this model to combine 2 models to take the best out of their strength and get a model that is more robust and adaptative.

III. Proposed Method

A. Overview

To further improve the performance of DFER, we first propose the improved ways to define the loss function, and then the 3 methods for optimization the improved Intensity Aware Loss are introduced.

B. New Intensity Aware Loss

To tackle the issues that have been highlighted previously, we use the Euclidian distance to get the difference between x_t and x_{max}"

$$L_{IA}^* = -log(\sqrt{(x_t{}^2 - x_{max}{}^2)}) = - log(|x_t - x_{max}|). \tag{3}$$

We then define:

$$P_{IA}^* = |x_t - x_{max}|. \tag{4}$$

If we consider a simpler model in which the prediction is binary between 2 emotions, we have:

$$x_t + x_{max} = 1, \tag{5}$$

$$L_{IA}^* = - log(|2x_t - 1|). \tag{6}$$

After analyzing this new function, we get that the values of L_{IA}^* are contained in [0.009; 1.699]. We can see that now the focus put on the low intensity samples is much bigger and the influence on the high intensity samples is almost null. Nevertheless, the influence on the low intensity samples might be too big and might have to be adjusted. The final loss is obtained by using the cross-entropy loss and L_{IA}^*.

$$L_{total} = L_{CE} + \alpha L_{IA}^*$$

where α is a hyper-parameter controlling the loss coefficients, which we will determine in the next subsection.

C. Exponential Decrease

The newly introduced L_{IA}^* has almost no influence on the training when the prediction is clear and therefore $x_t\ and\ x_{max}$ have a big gap. Nevertheless, in practice, it is unlikely to happen that the gap between $x_t\ and\ x_{max}$ will be that big because in practice, the prediction is made between 7 emotions. Therefore, we well reduce the impact of L_{IA}^* over the training because the model mostly learns the discriminative behavior at the beginning of the training, and L_{IA}^* will only cause a bias if it keeps the same influence during all the training. Therefore L_{IA}^* becomes L_{final}^*:

$$L_{final}^* = A * e^{-\lambda * n} * L_{IA}^* \tag{7}$$

where A is the amplitude of the exponential, λ is the parameter that controls the speed of the decrease, and n the current epoch of the training. These parameters will be determined in the next subsection.

TABLE I. DISTRIBUTION OF THE EMOTIONS IN THE SUBSET

Emotion	Total of sample	%
Happy	98	12.47
Sad	119	15.14
Neutral	100	12.72
Anger	94	11.96
Surprise	107	13.61
Disgust	116	14.76
Fear	152	19.34

TABLE II. LIST OF COMBINATIONS

Combination	Without L^*_{final}	With L^*_{final}	Pretrained without L^*_{final}	Pretrained with L^*_{final}
a	x			x
b	x		x	
c		x		x
d		x	x	
e	x	x		
f			x	x

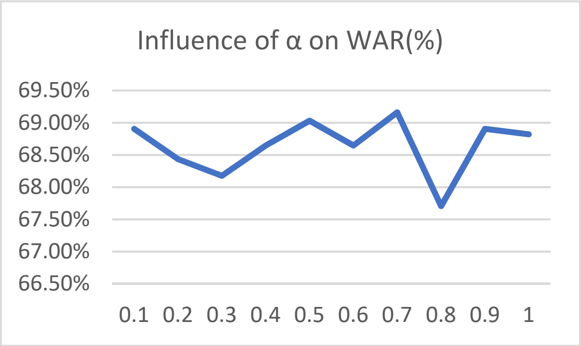

Fig. 1 Evolution of WAR with α

Moreover, we adopt 2 strategies for this method: the first one will be to have a big amplitude but also a quick decrease, whereas the other one will be to keep A = 1 but to decrease slowly the influence of the loss function.

D. Pretraining

One of the major issues of DFER is the lack of data for some emotions. This causes data imbalance in the dataset and therefore some emotions are very hard to predict. This is the case for “Disgust” that only has 146 samples in DFEW, so 1.22% of the dataset. To tackle this issue, we create a subset of the training data containing balanced data and we pretrain a model for $n_{pretrain}$ epochs.

As we can see in Table I. We use almost all the samples of “Disgust” in the subset. This could cause an overfitting. That is why we will pretrain the model with $n_{pretrain} \in \{5,10,15\}$, and for each of them we will conduct the pretraining with and without L^*_{IA}.

TABLE III. PERFORMANCES OF THE SECOND STRATEGY

Model	A	λ	WAR(%)
2.1	1	0.05	69.119
2.2	**1**	**0.03**	**69.204**
2.3	1	0.02	69.162
2.4	1	0.035	69.076
2.5	3	0.05	68.135

TABLE IV. PERFORMANCES OF THE PRETRAINED MODELS

$n_{pretrain}$	Use of L^*_{final}	WAR (%)	UAR (%)
5	Yes	68.862	54.324
5	No	68.862	54.160
10	Yes	69.247	54.399
10	**No**	**69.290**	**54.693**
15	Yes	69.204	53.927
15	No	69.162	54.184

TABLE V. BEST PERFORMANCES OF EACH COMBINATION

Combination	Best WAR	β
a	69.418	0.2
b	69.504	0.6
c	69.461	0.6
d	69.461	0.7
e	**69.717**	**0.4**
f	69.632	0.9

E. Model Combination

As mentioned previously, we take inspiration on the MAN model. We trained different models and combine them with a weighted average to have a more robust prediction.

$$\widehat{Y} = \beta\widehat{Y}_1 + (1-\beta)\widehat{Y}_2 \qquad (8)$$

where $\widehat{Y}_1$ and $\widehat{Y}_2$ are the predictions of the 2 models and β the parameter that controls the influence the importance of each model in the final prediction. We will conduct the combinations displayed on Table II.

IV. EXPERIMENTS

A. Experiment Details

We use the dataset DFEW to conduct the experiments and the same implementation details as in [1]. Therefore, the images are resized in 112x112 pixel, the optimizer is SGD, with a batch size of 40, the learning rate equal to 0.001 which decrease exponentially over 80 epochs. Also, we use the Weighted Average Recall (WAR) and Unweighted Average Recall (UAR) as metrics.

TABLE VI. COMPARISON OF THE PERFORMANCES OF THE PROPOSED MODEL.

Methods	NE	AN	SU	UAR	WAR
Former DFER [4]	67.52	70.03	56.43	53.69	65.70
M3DFEL [10]	67.88	74.24	59.69	56.10	69.25
AEN [6]	70.67	72.08	59.07	56.66	69.37
NR-DFERNET [12]	70.03	75.09	61.60	54.21	68.19
Resnet18 + MIDAS [5]	58.64	68.06	59.65	**57.45**	69.16
IAL+GCA [1]	70.10	**76.06**	62.22	55.71	69.24
Our model	**72.28**	75.12	**67.69**	55.60	**69.72**

B. *Evaluation of the Parameters*

We conduct studies on DFEW to determine the parameters of our final model.

We first begin by determining α: as we can see on Fig. 1. we achieve the best performances for $\alpha = 0.7$. We only consider the WAR here because what we want is the best model overall. On Table II. we can see the performances of the second strategy for the exponential decrease. The second strategy is better than the first one with an average WAR of 68.94% instead of 68.54%. Therefore, we take the model 2.2, which reduces the impact of the loss function to 9% after 80 epochs. Next, we achieved the best results with the pretrained models which has $n_{pretrain} = 10$ and without the use of L^*_{final} for the UAR and the WAR. Nevertheless, the aim of this method is not met. Indeed, the performances for “Disgust” and “Fear” are worse than before the pretraining even though it has better performances overall. The accuracy for “Fear” is still 0.00% and decreased by 4% for “Disgust”.

Finally, we achieve the best result for combination *f* which combines the model without L^*_{final} and the model with L^*_{final}. This final model has a WAR of 69.72% and a UAR of 55.60% which makes the best model among those presented before. Also, we can note that this time the performances of ‘Fear’ have been improved by 5.6%.

C. *Comparison with the Benchmark*

We compare the benchmark with our final model and the best pretrained model. We use the best pretrained model to assess the performances of the new loss function because the final model is obtained by combining a model that uses L^*_{final} and one that does not use an IAL. Therefore, the prediction might be more accurate but it will be less clear. We filmed clips of ourself doing emotions with different intensities and make the 3 models predict the emotion. We have done happiness with low, mid and high intensity and also surprise with low intensity.

We can see on Table VII. that the pretrained model has indeed clearer predictions on the “Happy” samples.

TABLE VII. PREDICTION ON THE NEW SAMPLES

	Final model	Benchmark	Pretrained
Happy low	0.466	0.437	0.563
Happy mid	0.686	0.786	0.987
Happy high	0.583	0.618	0.857
Surprise low	0.136	0.101	0.114

On all the samples, the pretrained model have clearer predictions than the benchmark model. This is especially true for “Happy mid” with 0.987% instead of 0.786% for the benchmark.

V. CONCLUSION

In this work, we develop a new loss function inspired by the Intensity Aware Loss. This new loss function combined with 3 optimization methods that we implemented provides a model that achieves great performances with an increase of 0.35% compared to the second-best model. Also, as intended this new model puts more focus on the low intensity samples than the original IAL and the influence on the high intensity samples is negligible.

To further optimize the model, it could be useful to use a data augmentation method to generate samples of ‘Fear’ and ‘Disgust’ instead of using the pretraining method.

ACKNOWLEDGE

The authors thank for the support of National Science and Technology Council under the contract of NSTC 113-2221-E-002-146.

REFERENCES

[1] H. Li, H. Niu, Z. Zhu, and F. Zhao, “Intensity-aware loss for dynamic facial expression recognition in the wild,” in *AAA*I Conf. Artificial Intelligence, vol. 37, pp.67-75, 2023.

[2] A. Psaroudakis and D. Kollias, “MixAugment & mixup: Augmentation methods for facial expression recognition,” in *IEEE*

Conf. Computer Vision and Pattern Recognition, pp. 2367-2375, 2022.

[3] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "FERV39k: A large-scale multi-scene dataset for facial expression recognition in videos," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp.20922-20931, 2022.

[4] Z. Zhao and Q. Liu, "Former-DFER: Dynamic facial expression recognition transformer," in *ACM Int. Conf. Multimedia*, pp.1553-1561, 2021.

[5] R. Kawamura, H. Hayashi, N. Takemura and H. Nagahara, "MIDAS: Mixing ambiguous data with soft labels for dynamic facial expression recognition," in *IEEE Winter Conf. Applications of Computer Vision*, pp.6552-6562, 2024.

[6] B. Lee, H. Shin, B. Ku, and H. Ko, "Frame level emotion guided dynamic facial expression recognition with emotion grouping," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp.5681-5691, 2023.

[7] Z. Zhang, X. Sun, J. Li, and M. Wang, "MAN: Mining ambiguity and noise for facial expression recognition in the wild," *Pattern Recognition Letters*, vol. 164, pp.23-29, 2022.

[8] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, issue 2, article 199, 2023.

[9] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *Int. J. Cognitive Computing in Engineering*, vol. 2, pp.57-64, 2021.

[10] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou, "Rethinking the learning paradigm for dynamic facial expression recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp.17958-17968, 2023.

[11] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu and J. Liu, "DFEW: A large-scale database for recognizing dynamic facial expressions in the wild," in *ACM Int. Conf. Multimedia*, pp.2881-2889, 2020.

[12] H. Li, M. Sui, and Z. Zhu, "NR-DFERNet: Noise-robust network for dynamic facial expression recognition," *arXiv:2206.04975*, 2022.