DOI: 10.21917/ijivp.2021.0356

CO-CURING NOISY ANNOTATIONS FOR FACIAL EXPRESSION RECOGNITION

Darshan Gera and S. Balasubramanian

Department of Mathematics and Computer Science, Sri Sathya Sai Institute of Higher Learning, India

Abstract

Driven by the advancement in technology that can facilitate implementation of deep neural networks (DNNs), and due to the availability of large scale datasets, automatic recognition performance of the machines has increased by leaps and bounds. This is also true with regard to facial expression recognition (FER) wherein the machine automatically classifies a given facial image in to one of the basic expressions. However, annotations of large scale datasets in FER suffer from noise due to various factors like crowd sourcing, automatic labelling based on key word search etc. Such noisy annotations impede the performance of FER due to the memorization ability of DNNs. To address it, this paper proposes a learning algorithm called Co-curing: peer training of two joint networks using a supervision loss and a mimicry loss that are balanced dynamically, and supplemented with a relabeling module to correct the noisy annotations. Specifically, peer networks are trained independently using supervision loss during early part of the training. As training progresses, mimicry loss is given higher weightage to bring consensus between the two networks. Our Co-curing does not need to know the noise rate. Samples with wrong annotations are relabeled based on the agreement of peer networks. Experiments on synthetic as well real world noisy datasets validate the effectiveness of our method. State-of-the-art (SOTA) results on benchmark in-the-wild FER datasets like RAF-DB (89.70%), FERPlus (89.6%) and AffectNet (61.7%) are reported.

Keywords:

Noisy Annotations, Facial Expression Recognition, Co-Curing, Mimicry Loss, Peer Learning

1. INTRODUCTION

. It is said that face is the index of mind. It expresses different human emotions in our everyday life. Recognizing such expressions plays an important role in social interaction and communication [1], [2]. Ekman and Friesen [3], [4] defined six basic emotions based on a cross-culture study. These prototypical facial expressions are classified as neutral, anger, fear, disgust, happiness, surprise and sadness. Contempt was later added among basic set of emotions [5].

Traditional FER systems were built using handcrafted features [7]-[10] from facial images collected in a controlled lab environment. Examples of such datasets include CK+ [11], [12], Oulucasia [13] and Jaffe [14]. However, in reality, facial images exhibit uncontrollable factors like variations in illumination, variations in pose, presence of occlusions etc. This scenario is called as in-the-wild scenario. Traditional FER systems badly fail in in-the-wild scenario. Fortunately, due to the recent success of DNNs, and due to the availability of large scale in-the-wild datasets like RAFDB [17], [18], AffectNet [15] and FERPlus [16], FER performance has enhanced. Nevertheless, there is still a wide scope for improvement in in-the-wild scenario due to the fact that annotations of facial images in in-the-wild scenario contains noise, and most of the FER systems turn a blind eye to the influence of these noisy labels. The noisy annotations arise due to the following reasons:

i) Manual annotation is costly, laborious and requires expertise; so crowdsourcing and online key searching methods are used, resulting in noisy annotations,

ii) Facial expression images are ambiguous and may express compound emotions, for e.g. face may be happily surprised or fearfully sad,

iii) Occlusions, lighting and pose variations cause uncertainty in getting true expression,

iv) Prototypic expressions vary across cultures, situations, and across individuals under same situation [6] [19].

It has been observed that DNNs memorize noisy annotations [20]-[22] which affect the performance of modern DNN based FER methods. So, it is important to eliminate the influence of noisy samples during training.

To handle noisy annotations, this paper proposes a novel peer learning framework called Co-curing. It has two components: i) peer training of two networks using only clean samples, and ii) relabeling of noisy samples based on the agreement of both the networks. By clean samples, we identify those samples (implicitly) from the dataset that guide the learning of DNN in the early part of the training. This is because it has been observed that DNNs fit clean labels in the early part of the training, and subsequently memorize the noisy labels in the later part of the training [40]. Specifically, inspired by peer learning [31], we train two networks simultaneously for learning discriminative features based on clean samples as well as identifying the noisy samples. Training is guided by two losses: a supervision loss that uses only clean labels, and a mimicry loss [31] that aligns the probability distribution of two networks. Supervision loss is given higher weightage during early part of the learning, and subsequently, as training progresses, mimicry loss gets higher weightage [40]. Transition from supervision loss to mimicry loss happens using a dynamic balancing rule that does not require knowledge of noise rate. Further, peer networks also correct the labels of wrongly annotated samples if the predicted probability of both the networks is greater that of the ground truth probability by a certain threshold. Co-curing is a simple but effective framework for peer learning as well as curing noisy annotations.

In short, our contributions can be summarized as follows:

- A novel learning framework called Co-curing for FER in the presence of noisy annotations.
- Peer learning aided by a dynamic transition between supervision loss and mimicry loss
- Correction of noisy annotations guided by joint agreement.
- Validation of Co-curing on both synthetic as well as real noisy FER datasets. Co-curing attains state-of-the-art performance on benchmark in-the-wild FER datasets like RAF-DB (89.7%), FERPlus (89.6%) and AffectNet (61.7%).

Rest of the paper is organized as follows: section 2 presents related work. Our proposed method Co-curing is described in section 3. Experimental set up is detailed in section 4. Results and discussions are reported in section 5. Section 6 elicits the expression discrimination ability of Co-curing, and also presents few ablation studies. Comparison with recent SOTA methods is reported in section 7. Finally, conclusions are listed in section 8.

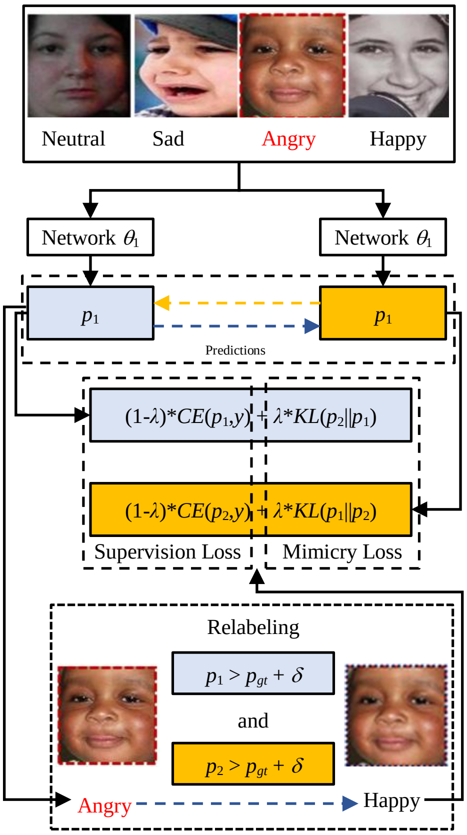


Fig.1. The pipeline of Co-curing framework. Face images are input to two identical peer networks to obtain prediction probabilities. Peer training module dynamically transitions from supervision loss to mimicry loss. Relabeling module corrects the noisy samples based on comparison of predicted probabilities with the ground truth label probability. Mislabeled sample is marked in red dotted rectangle in the input

2. RELATED WORK

2.1 DEEP LEARNING (DL) BASED NOISY METHODS

Presence of noisy labels in real-world datasets has led to development of many DL based approaches. These approaches could be summarized as below:

2.1.1 Noise Rate Estimation:

Some methods depend on the estimation of noise transition matrix [23]-[24] which encodes the probability of mislabeling between classes. For e.g. F-correction [24] estimates the noise transition matrix by using a deep network trained on the noisy dataset. The method in [23] introduces a noise adaption layer using softmax to estimate the noise transition matrix. However, success of these methods depend on quality of noise rate estimation, which is quite challenging as it requires either the prior knowledge or a subset of clean data. So, these methods may not be feasible in real-world environment.

2.1.2 Robust Loss Functions:

Many noise robust loss functions [29], [30], [32]-33] have been designed. For e.g. bootstrapping loss [33] extends the usual cross-entropy loss with a perceptual term. GCE [32] combines mean absolute loss and cross-entropy loss. But these methods do not perform well on real noisy datasets.

2.1.3 Small Loss Sample Selection

A vast collection of methods train on small loss samples as these are more likely to be associated with clean samples [20]. These methods use peer/co/joint learning by training two networks on low loss samples using cross entropy loss. The percentage of high loss samples to be dropped (depending upon noise rate) keeps increasing as a function of training epoch. In Decoupling [38], both networks are trained using small loss samples on which they disagree in their predictions. In Co-teaching [25], each model is trained using loss corresponding to other network's small loss samples whereas in Co-teaching+ [28], an extension of Co-teaching, the small loss samples are chosen among the ones where both the networks disagree in their predictions. JoCoR [39] uses the same idea of Co-teaching+ but uses agreement instead of disagreement, and computes joint loss using co-regularization. NCT [40] uses independent training of joint networks using convex combination of a supervision loss and a mimicry loss combined with a dynamic balancing scheme. It also uses target variability regularization to keep models diverged during training. Main limitation of most of above methods are: (i) They depend on accurate estimation of noise distribution in data, (ii) Hard samples are also perceived as noisy, and hence the model is biased towards easy samples which hurts the generalization capability of the model.

2.2 FER NOISY METHODS THAT TACKLE NOISY ANNOTATIONS

Very few FER methods have been developed to tackle noisy annotated labels. IPA2LT framework [41] is proposed to handle inconsistent annotations present in different FER databases. In IPA2LT, each sample is assigned multiple inconsistent human and machine predicted labels, and subsequently a network is trained to discover the true label by maximizing log-likelihood of inconsistent annotations by estimating multiple noise transition matrices. Self-cure network (SCN) [19] learns the importance of each sample using a self-attention module for loss re-weighting. Low weight samples are treated as noisy and also relabeled if predicted probability is above a certain threshold. These methods focus on correcting noisy labels. In [6], Rayleigh loss function is proposed to learn discriminative features which explicitly increases inter-class separation and intra-class compactness at feature level. Rayleigh loss is insensitive to label noises. In addition, weighted Softmax (w-Softmax) loss is also proposed to reduce the loss weight of uncertain samples based on their distances to class centers. A class specific threshold is used for reweighing loss depending upon distance between class center and nearest class center. Limitation of IPA2LT is that it is difficult to estimate noise transition matrices accurately. SCN uses single network to distinguish the uncertain samples from certain which leads to confirmation bias. Rayleigh loss and w-Softmax loss are robust only for low noise datasets.

3. PROPOSED METHOD

In this section, we describe the main components of the proposed method Co-curing, including motivation, structure, loss function and training algorithm. In addition, we discuss the relations between Co-curing and other existing approaches.

3.1 OVERVIEW

The proposed Co-curing follows the principle of peer learning [31]. It involves mutual training of two networks to tackle the issue of noisy annotations. During early part of the training, both peer networks are trained independently using supervision loss because it has been observed that DNNs fit clean samples during early part of the training [20]-[21]. Further, since DNNs tend to memorize noisy samples as training progresses, mimicry loss [31] is used in the later part of the training to build consensus among predictions of both the networks. Naturally, seeking consensus between both the networks would avoid influence of noisy labels, since on samples with noisy labels, the networks will not agree. The transition between the losses is balanced dynamically [40]. Specifically, a sigmoid like ramp-up function [27] [40] is used to transition from supervision loss to mimicry loss. Label correction of noisy annotations is done if both networks prediction probabilities are greater than the ground truth probability by a certain threshold.

3.2 STRUCTURE

The proposed Co-curing framework is shown in Fig.1. It has two modules: i) peer training module and ii) relabeling module.

Given C expression categories, let $D=\{(x_i,y_i)\}_1^N$ be the dataset of N training samples where $y_i \in \{1,2,..,C\}$. Let the parameters of both the networks be denoted as θ_1 and θ_2, respectively. Let the prediction probabilities of both the networks be denoted as p_1 and p_2, respectively.

3.2.1 *Peer Training Module:*

Supervision loss is standard cross-entropy loss (L_{CE}) based on the predictions and true labels. Mimicry loss [31] is the Kullback Leibler (KL) Divergence between network predictions denoted as D_{KL}. Prediction (p_2) from second network (θ_2) is used to provide supervision to first network (θ_1) and vice versa.

The overall loss functions L_{θ_1} and L_{θ_2} for both the networks θ_1 and θ_2, respectively can be written as follows:

The overall loss functions $L_{\theta_1}$and $L_{\theta_2}$for both the networks θ_1 and θ_2, respectively can be written as follows:

$$L_{\theta_1} = (1\text{-}\lambda)*L_{CE}(p_1,y)+\lambda*D_{KL}(p_2\|p_1) \quad (1)$$

$$L_{\theta_2} = (1\text{-}\lambda)*L_{CE}(p_2,y)+\lambda*D_{KL}(p_1\|p_2) \quad (2)$$

Here, $\lambda \in [0,1]$ is dynamic balancing factor, which is a sigmoid like ramp-up function [27] defined as follows:

$$\lambda = \lambda_{\max} * e^{-\beta\left(1-\frac{t}{t_r}\right)^2} \quad (3)$$

where λ_{max} is maximum lambda value, t is the current epoch, t_r is the ramp-up length (the epoch at which λ attains its maximum value λ_{max}. Here β controls the shape of the ramp-up function. Smaller the value of β, faster it transitions to mimicry loss. We choose β = 0.65 based on the experimental study in section 6.2.2.

3.2.2 *Relabeling Module:*

Noisy annotations are relabeled based on difference between maximum predicted probabilities of both the networks and the probability of ground truth label. If the maximum prediction probabilities of both the networks for a sample are greater than the given label probability by a certain threshold, then it is relabeled using pseudo label. Formally, relabeling module is defined as:

$$y' = \begin{cases} & \quad if\ p_1^{\max} > p_1^{gt} + \delta\, and \\ l_1^{\max} & \quad p_2^{\max} > p_2^{gt} + \delta\, and \\ & \quad l_1^{\max} = l_2^{\max} \\ l^{org} & otherwise \end{cases} \quad (4)$$

Here, y' denotes new pseudo label, $p_1^{\max}$ and $p_2^{\max}$ represent the maximum probabilities of the two networks respectively, and $l_1^{\max}$ and $l_2^{\max}$ represent the index of maximum probabilities. Further, p_1^{gt} and p_2^{gt} denote the probabilities of ground truth label of the sample from networks θ_1 and θ_2 respectively. l^{org} denotes the original label of the sample.

3.2.3 *Inference:*

Average of predictions from both the networks is used to infer the label at the test time.

3.3 TRAINING ALGORITHM

Algorithm 1: Co-curing training algorithm

Input: Network f with parameters $\{\theta_1,\theta_2\}$, dataset (D), number of classes (C), batch size (b), number of epochs (t_{max}), maximum λ value (λ_{max}), learning rate (η), warm up epoch for relabeling (t_{warm}).

```
Step 1:  Initialize $\theta_1$ and $\theta_2$ randomly
Step 2:  For $t$ = 1, 2,…, $t_{max}$ epochs do
Step 3:  Sample a minimatch $D_n$ from $D$
Step 4:  Compute $p_1$=$f(x,\theta_1)$and $p_2$=$f(x,\theta_2)$ using two networks
Step 5:  Compute dynamic weighting factor $\lambda$ using Eq.(3)
Step 6:  Compute individual losses ($L_{\theta_1}$) and ($L_{\theta_2}$) using Eq.(1)
         and Eq.(2) respectively
Step 7:  Update $\theta_1 = \theta_1\text{-}\eta\nabla\theta_1$ and $\theta_2 = \theta_2\text{-}\eta\nabla\theta_2$
Step 8:  If $t > t_{warm}$, then obtain new pseudo label using Eq.(4)
Step 9:  End for
Step 10: Output: Return $\theta_1$ and $\theta_2$
```

3.4 COMPARISON WITH OTHER APPROACHES

Our algorithm Co-curing is compared with other related approaches in Table.1 across various strategies used in the literature. Decoupling [38] is based on disagreement. Co-teaching [25] uses small-loss samples and updates network parameters using cross update criteria. Co-teaching+ [28] combines cross-update along with disagreement on small loss samples to improve the performance. Limitation of disagreement strategy is that very few clean samples may be selected. So, JoCoR [39] selects small low samples using agreement between joint networks. But they may converge to consensus and fail to perform in the presence of high noise. Further, these methods need to know noise rate. Our

Co-curing does supervised training based on small loss samples implicitly and using agreement between two networks predictions in the later stage without depending upon noise rate like NCT [40]. Further, it also carefully corrects noisy annotated samples based on the agreement confidence. Both of these help to improve training using more number of clean samples.

Table.1. Comparison of proposed method with related approaches

Strategy	Decoupling [38]	Co-teaching [25]	Co-teaching+ [28]	JoCoR [39]	NCT [40]	Co-curing
Small loss	×	✓	✓	✓	×	×
Cross Update	×	✓	✓	✓	×	×
Disagreement	✓	×	✓	×	×	×
Agreement	×	×	×	✓	✓	✓
Joint update	×	×	×	✓	×	×
Relabeling	×	×	×	×	×	✓
Knowledge of noise rate	✓	✓	✓	✓	×	×

4. EXPERIMENTAL SETUP

4.1 DATASET

Following in-the-wild FER datasets have been used in this study:

- **RAF-DB** [17] [18] has 29762 facial images retrieved from internet. These are labelled for basic or compound expressions by 40 annotators. We use the subset with 7-basic emotions. This consists of total 15,339 images in which 12,271 are used for training and 3068 for testing.
- **FERPlus** [16] consists of 48×48 resolution images collected from Google. These have been annotated for 8-basic emotions. It is divided into training (28,709 images), validation (3589 images) and testing (3589 images).
- **AffectNet** [15] consists of 1M images obtained by querying web. 0.44M are manually annotated by 12 different annotators and remaining 0.46M images are automatically annotated for the presence of all eight facial expressions. AffectNet has 4000 images in the validation set. We use automatically annotated subset for training under real noisy conditions and manually annotated subset for training with synthetic noise.

4.2 EXPERIMENTAL DETAILS

In Co-curing, face images are detected and further aligned using MTCNN [42] and further resized to 224×224. The individual networks in Co-curing are ResNet-18 [43] which are pre-trained on large scale face recognition dataset MS-Celeb-1M [44]. We implement our method in Pytorch. Batch size is set to 128. Parameter updates are carried out by Adam optimizer. Learning rate (*lr*) is initialized to 0.001 for base networks and 0.01 for the classification layer. Learning rate is decayed exponentially on every epoch with a factor of 0.95. Data is augmented by random horizontal flip and color jitter. λ_{max} is set to 0.9 and β is se to 0.65 based on the ablation study presented in section 6. Class imbalance is overcome using oversampling on AffectNet dataset. Evaluation metric considered is overall accuracy.

5. RESULTS

5.1 EVALUATION ON SYNTHETIC NOISY DATASETS

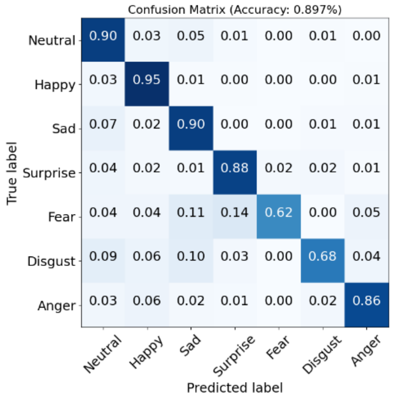
Synthetic symmetric noise of 10-40% is manually added on RAFDB, FERPlus and AffectNet datasets by randomly changing labels. Co-curing is compared with recent SOTA methods like Co-teaching [25], Co-teaching+ [28], NCT [40] and JoCoR [39]. It is to note that all of these methods employ joint networks for training to combat the influence of noisy labels but do not involve correction of noisy annotated samples. We also compare Co-curing with FER methods SCN [19] and RR [6] that relabel noisy annotations. We also compare with a baseline wherein the two networks are independently trained using only supervision loss, without having any relabelling. The inference strategy in baseline is done using average of predictions, as mentioned earlier.

Table.2. Performance comparison in the presence of synthetic noise

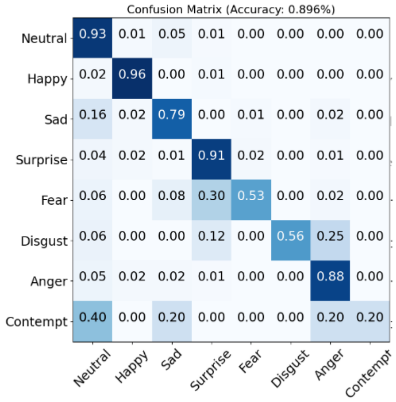
Noise level	Method	RAF-DB	FERPlus	AffectNet
10	Baseline	85.07	86.48	59.4
	Coteaching [25]	80.18	86	58.93
	Coteaching+ [28]	81.84	83.33	53.73
	JoCR [39]	84.84	85.91	58.05
	NCT [40]	87.42	87.28	59.70
	SCN [19]	82.18	84.28	58.58
	RR [6]	82.43	83.93	60.04
	Co-curing (No relabeling)	86.64	87.66	60.38
	Co-curing	88	87.85	60.58
20	Baseline	81.91	84.79	58.6
	Coteaching [25]	79.56	85.5	57
	Coteaching+ [28]	81.12	76.44	49.55
	JoCR [39]	82.79	83.71	57.28
	NCT [40]	85.29	86.42	59.28
	SCN [19]	80.01	83.17	57.25
	RR [6]	80.41	83.55	58.47
	Co-curing (No relabeling)	84.16	86.9	59.65
	Co-curing	86.7	87.38	59.15
30	Baseline	81.55	84.02	56.37
	Coteaching [25]	75	83.48	54.22
	Coteaching+ [28]	80.05	75.83	44.9
	JoCR [39]	80.96	81.51	54.38

	NCT [40]	82.66	83.93	56.23
	SCN [19]	77.46	82.47	55.05
	RR [6]	76.77	82.75	55.82
	Co-curing (No relabeling)	81.58	84.79	56.89
	Co-curing	84.84	86.64	56.98
40	Baseline	79.95	82.81	53.57
	Coteaching [25]	61.18	80.52	50.45
	Coteaching+ [28]	80.05	75.83	44.9
	JoCR [39]	80.96	81.51	54.38
	NCT [40]	79.01	81.86	52.25
	Co-curing (No relabeling)	81.13	83.36	55.15
	Co-curing	82.06	84.41	55.35

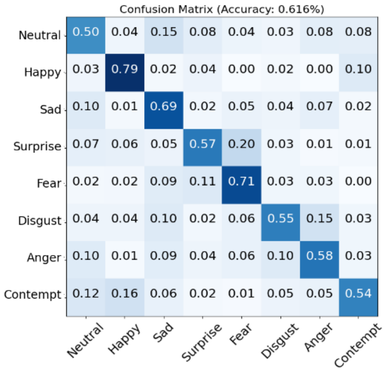
We also compare against Co-curing without including relabeling module. The Table.2 presents the results. Clearly, single network based methods fail to perform well compared to joint networks based methods. Co-curing outperforms all methods on all the datasets, and for all noise levels. The performance of Co-curing without relabelling module is comparable or superior to the SOTA method NCT.



(a)



(b)



(c)

Fig.2. Confusion plots for (a) RAF-DB, (b) FERPlus and (c) AffectNet (trained on manually annotated AffectNet)

5.2 EVALUATION ON REAL NOISE DATASET

To further validate the effectiveness of our Co-curing, we also do experiments on real-world noisy dataset. We train on automatically annotated AffectNet dataset and test the performance on clean validation set of AffectNet with 3500 images as training set has only 7 expressions (except contempt). Results are presented in Table.3. Again, Co-curing outperforms all other methods. Coteaching, Coteaching+ and JoCoR fails to perform as these are heavily dependent upon the noise rate whereas our method is superior in robustness to real noise of unknown kind and type due to dynamic switch from supervision loss to mimicry loss. CC performs superior to NCT due to relabeling module. Compared to Baseline, there is a significant improvement of 2.9% using CC.

Table.3. Performance comparison on automatically annotated noisy AffectNet dataset

Method	AffectNet
Baseline	53.85
Coteaching [25]	52.37
Coteaching+ [28]	55.08
JoCR [39]	55.00
NCT [40]	56.46
Co-curing (No relabeling)	56.02
Co-curing	56.70

6. ANALYSIS

6.1 EXPRESSION DISCRIMINATION

The performance comparison at the individual expression levels is done by plotting confusion plots for all datasets as shown in Fig.2 and Fig.3. Happiness is the easiest expression to recognize across all the datasets.

On RAF-DB, fear and disgust are most difficult whereas on FERPlus, it is contempt and fear. In manually annotated AffectNet, neutral and contempt are difficult to recognize whereas in automatically annotated AffectNet, disgust and anger cannot be recognized easily. Fear is generally confused with surprise on all datasets.

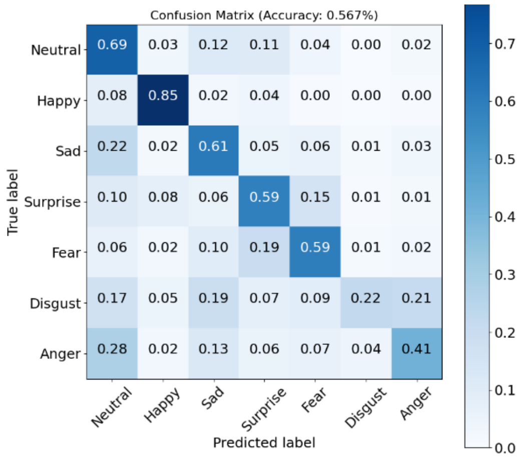


Fig.3. Confusion plots for AffectNet (trained on automatically annotated AffectNet)

6.2 ABLATION ANALYSIS

6.2.1 Comparison with Supervision Loss Training:

It is known that DNNs learn from clean samples in the early part of training, and subsequently memorize noise in the later part of the training [20]. Does our Co-curing method mitigate the memorizing effect of DNNs? We validate this by plotting the test accuracy vs. epochs on RAF-DB and FERPlus datasets in the presence of 40% synthetic noise for our Co-curing method and the Baseline that uses only CE loss. These are shown in Fig.4 and Fig.5. When trained using only CE loss, performance improves steadily during early part of the training but falls down rapidly in the later stage due to the memorization of noise. In comparison, our method Co-curing effectively prevents memorizing noisy labels as performance curve is steadily superior even in the later stage of training.

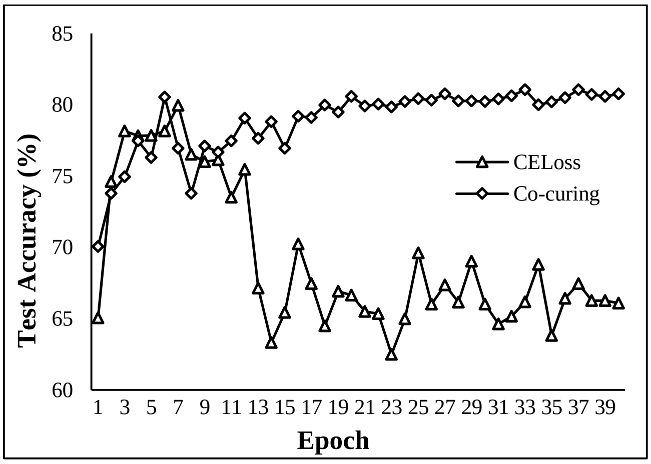


Fig.4. Accuracy vs. epoch plot on RAF-DB in the presence of 40% noise

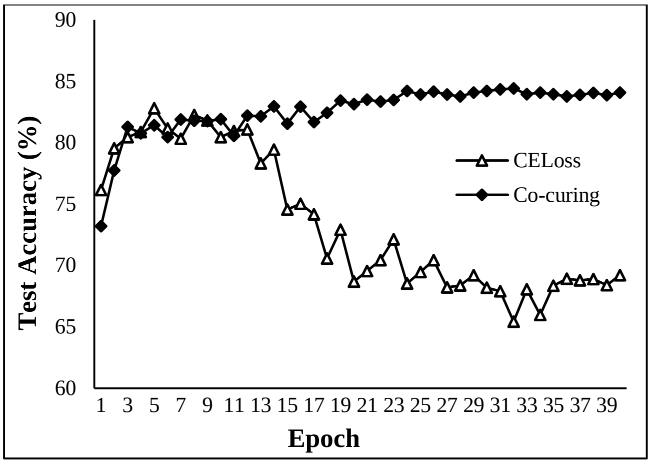


Fig.5. Accuracy vs. epoch plot on FERPlus in the presence of 40% noise

6.2.2 Influence of Dynamic Balancing Factor:

Since, as per Eq. 3, if the best β is determined, automatically the best λ gets fixed. So, we determine the best β as follows. We plot in Fig.6 the test accuracy for 0-40% noise on RAFDB dataset for different values of β. Here, $\beta = 0$ corresponds to the training dominated by mimicry loss while $\beta = 6$ corresponds to the training dominated by supervision loss. Smaller the value of β, faster it transitions from supervision to mimicry loss. Clearly, on clean (noisy) dataset, larger (smaller) β will give better performance and vice versa. But it is difficult to know beforehand whether dataset is clean or noisy. In our experiments, we found that $\beta = 0.25$ and $\beta = 1$ reported superior performance on high and low noisy datasets, respectively. We choose the intermediate $\beta = 0.65$ as it works across all noise levels and on all the datasets.

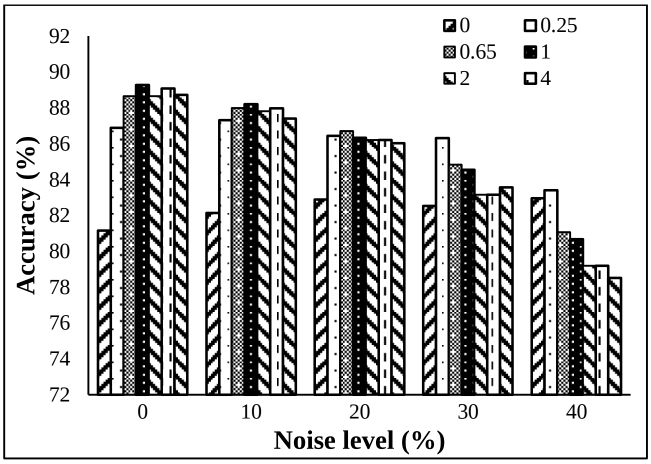


Fig.6. Influence of dynamic balancing factor (β) (refer Eq.(3))

7. COMPARISION WITH STATE-OF-THE-ART

Tables 4, 5 and 6 compare Co-curing with recent state-of-the-art methods like GACNN [37], DLP-CNN [46], IPA2LT [41], RAN [45], OADN [36], SCN [17], GCN [35] and SCAN [26] on RAF-DB, FERPlus and AffectNet datasets, respectively. Our method obtains comparable performance on RAFDB (89.70% which is lower by margin of 0.13 compared to the best performing method). It outperforms other methods on FERPlus (89.7%).

Further, it reports SOTA performance on AffectNet (61.7%). All these results demonstrate that Co-curing is a general purpose robust FER training method.

Table.4. Performance comparison with state-of-the-art on RAF-DB dataset

Method	**Year**	**RAF-DB**
GACNN [37]	2018	85.07
DLP-CNN [46]	2019	84.22
IPA2LT [41]	2020	86.77
RAN [45]	2020	86.9
OADN [36]	2020	**89.83**
SCN [17]	2020	88.14
GCN [35]	2020	89.41
SCAN [26]	2021	89.02
Co-curing	2021	89.70

Table.5. Performance comparison with state-of-the-art on FERPlus dataset (* denotes our implementation)

Method	**Year**	**FERPlus**
GACNN [37]	2018	84.86*
RAN [45]	2020	89.26
OADN [36]	2020	88.71*
SCN [17]	2020	89.35
GCN [35]	2020	89.39
ESR [34]	2020	87.15
SCAN [26]	2021	89.42
Co-curing	2021	**89.70**

Table.6. Performance comparison with state-of-the-art on AffectNet dataset (* denotes our implementation)

Method	**Year**	**AffectNet**
GACNN [37]	2018	55.05*
RAN [45]	2020	59.5
OADN [36]	2020	58.92
SCN [17]	2020	60.23
GCN [35]	2020	60.58
ESR [34]	2020	59.3
SCAN [26]	2021	61.7
Co-curing	2021	**61.7**

8. CONCLUSION

In this paper, we propose a simple and effective method for combating noisy annotations in FER datasets called Co-curing. Our method trains two joint network using peer learning aided by a dynamic transition between supervision loss and mimicry loss. Further, it carefully corrects the noisy annotations based on joint agreement. Robustness of Co-curing is demonstrated on both synthetic as well as real noisy FER datasets. SOTA performance on benchmark in-the-wild FER datasets validates the utility of Co-curing as a general purpose FER training framework. In the future, we would like to test this method on real-world noisy datasets from other domains.

REFERENCES

[1] C. Darwin and P. Prodger, “*The Expression of the Emotions in Man and Animals*”, Oxford University Press, 1998.

[2] S. Li, W. Deng, “Deep Facial Expression Recognition: A Survey”, *IEEE Transactions on Affective Computing*, Early Access, 2020.

[3] P. Ekman and W.V. Friesen, “Constants across Cultures in the Face and Emotion”, *Journal of Personality and Social Psychology*, Vol. 17, No. 2, pp. 124-129, 1971

[4] P. Ekman, “Strong Evidence for Universals in Facial Expressions: A Reply to Russell’s Mistaken Critique”, *Psychological Bulletin*, Vol. 115, No. 2, pp. 268-287, 1994.

[5] D. Matsumoto, “More Evidence for the Universality of a Contempt Expression”, *Motivation and Emotion*, Vol. 16, No. 4, pp. 363-368, 1992.

[6] X. Fan, Z. Deng, K. Wang, X. Peng and Y. Qiao, “Learning Discriminative Representation for Facial Expression Recognition from Uncertainties”, *Proceedings of IEEE International Conference on Image Processing*, pp. 903-907, 2020.

[7] J. MA, “Facial Expression Recognition using Hybrid Texture Features based Ensemble Classifier”, *International Journal of Advanced Computer Science and Applications*, Vol. 6, pp. 1-13, 2017.

[8] C. Shan, S. Gong and P.W. Mcwoan, “Facial Expression Recognition based on Local Binary Patterns: A Comprehensive Study”, *Image and Vision Computing*, Vol. 27, No. 6, pp. 803-816, 2009.

[9] P. Hu, D. Cai, S. Wang, A. Yao and Y. Chen, “Learning Supervised Scoring Ensemble for Emotion Recognition in the Wild”, *Proceedings of ACM International Conference on Multimodal Interaction*, pp. 553-560, 2017.

[10] H. Chun Lo and R. Chung, “Facial Expression Recognition Approach for Performance Animation”, *Proceedings of IEEE International Workshop on Digital and Computational Video*, pp. 613-622, 2001.

[11] T. Kanade, J.F. Cohn and Y. Tian, “Comprehensive Database for Facial Expression Analysis”, *Proceedings of 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46-53, 2000.

[12] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, “The Extended Cohn-Kanade Dataset (ck+): A Complete Dataset for Action Unit and Emotion-Specified Expression”, *Proceedings of IEEE International Workshops on Computer Vision and Pattern Recognition*, pp. 94-101, 2010.

[13] G. Zhao, X. Huang, M. Taini, S.Z. Li and M. Pietikainen, “Facial Expression Recognition from Near-Infrared Videos”, *Proceedings of IEEE International Conference on Image and Vision Computing*, pp. 607-619, 2011.

[14] F.Y. Shih, C.F. Chuang and P.S.P. Wang, “Performance Comparisons of Facial Expression Recognition in Jaffe Database”, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 22, No. 3, pp. 445-459, 2008.

[15] A. Mollahosseini, B. Hasani and M.H. Mahoor, "A Database for Facial Expression, Valence, and Arousal Computing in the Wild", *IEEE Transactions on Affective Computing*, Vol. 10, No. 1, pp.18-31, 2017.

[16] E. Barsoum, C. Zhang, C.C. Ferrer and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution", *Proceedings of 18th ACM International Conference on Multimodal Interaction*, pp. 279-283, 2016.

[17] S. Li and W. Deng, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition", *IEEE Transactions on Image Processing*, Vol. 28, No. 1, pp. 356-370, 2018.

[18] S. Li, W. Deng and J. Du, "Reliable Crowdsourcing and Deep Locality Preserving Learning for Expression Recognition in the Wild", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 28522-2861, 2017.

[19] K. Wang, X. Peng, J. Yang, S. Lu and Y. Qiao, "Suppressing Uncertainties for Large-Scale Facial Expression Recognition", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6897-6906, 2020.

[20] D. Arpit, S. Jastrz, N. Ballas, D. Krueger, E. Bengio, M.S. Kanwal, T. Maharaj, A. Fischer, A. Courville and Y. Bengio, "A Closer Look at Memorization in Deep Networks", *Proceedings of International Conference on Machine Learning*, pp. 233-242, 2017.

[21] C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals, "Understanding Deep Learning requires Rethinking Generalization", *Proceedings of International Conference on Machine Learning*, pp. 1-13, 2017.

[22] B. Frenay and M. Verleysen, "Classification in the Presence of Label Noise: A Survey", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 25, No. 5, pp. 845-869, 2013.

[23] J. Goldberger and E. Ben-Reuven, "Training Deep Neural-Networks using A Noise Adaptation Layer", *Proceedings of International Conference on Machine Learning*, pp. 1-5, 2016.

[24] G. Patrini, A. Rozza, A. Menon, R. Nock and L. Qu, "Making Neural Networks Robust to Label Noise: A Loss Correction Approach", *Proceedings of International Conference on Machine Learning*, pp. 1-9, 2016.

[25] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang and M. Sugiyama, "Coteaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels", *Proceedings of International Conference on Machine Learning*, pp. 1-13, 2018.

[26] Darshan Gera and S. Balasubramanian, "Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information based Facial Expression Recognition", *Pattern Recognition Letters*, Vol:145, pp. 58-66, 2021.

[27] Samuli Laine and Timo Aila, "Temporal Ensembling for Semisupervised Learning", *Proceedings of International Conference on Neural and Evolutionary Computing*, pp. 1-13, 2016.

[28] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, M. Sugiyama, "How does Disagreement Help Generalization Against Label Corruption?", *Proceedings of International Conference on Machine Learning*, pp. 7164-7173, 2019.

[29] X. Wang, Y. Hua, E. Kodirov and N.M. Robertson, "Image for Noise-Robust Learning: Mean Absolute Error does not Treat Examples Equally and Gradient Magnitude's Variance Matters", *Proceedings of International Conference on Machine Learning*, pp. 1-14, 2019.

[30] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi and J. Bailey, "Symmetric Cross Entropy for Robust Learning with Noisy Labels", *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 322-330, 2019.

[31] Ying Zhang, Tao Xiang, Timothy M. Hospedales and Huchuan Lu, "Deep Mutual Learning", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4320-4328, 2018.

[32] Zhilu Zhang and Mert Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels", *Proceedings of IEEE Conference on Neural Information Processing Systems*, pp. 8778-8788, 2018.

[33] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan and Andrew Rabinovich, "Training Deep Neural Networks on Noisy Labels with Bootstrapping", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3320-3328, 2015.

[34] H. Siqueira, S. Magg and S. Wermter, "Efficient Facial Feature Learning with Wide Ensemble-Based Convolutional Neural Networks", *Proceedings of AAAI Conference on Artificial Intelligence*, pp. 5800-5809, 2020.

[35] P. Jiang, B. Wan, Q. Wang and J. Wu, "Fast and Efficient Facial Expression Recognition using a Gabor Convolutional Network", *IEEE Signal Processing Letters*, Vol. 27, pp. 1954-1958, 2020.

[36] P. Ding and R. Chellappa, "Occlusion-Adaptive Deep Network for Robust Facial Expression Recognition", *Proceedings of IEEE International Joint Conference on Biometrics*, pp. 1-9, 2020.

[37] Y. Li, J. Zeng, S. Shan and X. Chen, "Occlusion Aware Facial Expression Recognition using CNN with Attention Mechanism", *IEEE Transactions on Image Processing*, Vol. 28, No. 5, pp. 243902450, 2018.

[38] E. Malach and S. Shalev-Shwartz, "Decoupling" when to Update" from" How to Update", *Proceedings of IEEE International Conference on Advances in Neural Information Processing Systems*, pp. 1-11, 2017.

[39] H. Wei, L. Feng, X. Chen and B. An, "Combating Noisy Labels by Agreement: A Joint Training Method with Co-Regularization", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13726-13735, 2020.

[40] F. Sarfraz, E. Arani and B. Zonooz, "Noisy Concurrent Training for Efficient Learning under Label Noise", *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3159-3168, 2021.

[41] J. Zeng, S. Shan and X. Chen, "Facial Expression Recognition with Inconsistently Annotated Datasets", *Proceedings of European Conference on Computer Vision*, pp. 222-237, 2018.

[42] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment using Multitask Cascaded

Convolutional Networks", *IEEE Signal Processing Letters*, Vol. 23, No. 10, pp. 1499-1503, 2016.

[43] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[44] Y. Guo, L. Zhang, Y. Hu, X. He and J. Gao, "Ms-Celeb-1m: A Dataset and Benchmark for Large-Scale Face Recognition", *Proceedings of European Conference on Computer Vision*, pp. 87-102, 2016.

[45] K. Wang, X. Peng, J. Yang, D. Meng and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition", *IEEE Transactions on Image Processing*, Vol. 29, pp. 4057-4069, 2020.

[46] S. Li and W. Deng, 'Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition", *IEEE Transactions on Image Processing*, Vol. 28, No. 1, pp. 356-370, 2018.