Shih, JL. et al. (Eds.) (2023). Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Dynamic Facial Expression Recognition through Partial Label Learning and Federated Learning

Mohammad Alif DAFFA[a*], Manas GUPTA[b], Hao CHEN[b] & Cheryl Sze Yin WONG[b]

[a]*Singapore University of Technology and Design (SUTD), Singapore*

[b]*Institute for Infocomm Research (I2R), Agency for Science Technology and Research (A*STAR), Singapore*

{manas_gupta, chen_hao, cheryl_wong}@i2r.a-star.edu.sg

Abstract: In this paper, we present a model development pipeline for dynamic Facial Expression Recognition (FER) aimed at quantifying learning in virtual classrooms. The proposed pipeline involves the use of partial labels for training dynamic FER models, followed by the use of a self-supervised federated learning approach in further enhancing the model's performance on new subjects, addressing both continual learning needs and privacy concerns. This work ultimately contributes to advancing learning quantification in virtual classrooms by integrating partial label training and federated learning strategies for dynamic FER.

Keywords: Federated learning, Privacy preserving deep learning, Facial expression recognition, Partial Label Learning

1. Introduction

The assessment of student engagement consists of varied interactive components that can generally be divided into behavioral, affective and cognitive dimensions (Mandernach, 2015). In online learning environments, student engagement has typically been determined based on facial video recordings (Gupta et al., 2016; Dhall et al., 2020; Shen et al., 2022). Hence, dynamic facial expression recognition (FER) is an area of research that can be leveraged on for student engagement.

Despite the extensive research on both static and dynamic FER (Li & Deng, 2020), there still exists challenges that hinder the use of FER in applications. In general, FER suffers from subjective annotations and inherent similarity between emotion classes (Wang, Weijie & Sebe, Nicu & Lepri, Bruno., 2022). In addition, the variability in emotional expressions across subjects, as well as the lack of large dynamic FER datasets further enhances the difficulty of the task.

However, this raises yet another significant challenge which is the case of privacy. As FER data typically contains participant's faces, many would prefer not to have this data shared with others. A possible solution to this is the use of feature extraction methods to provide anonymity of training data. However, feature-extraction based methods generally achieve worse results when compared to end-to-end based methods (Tsalera et al., 2022). In order to overcome this issue, (Salman & Busso, 2022) developed a privacy preserving personalisation method for dynamic FER using federated learning. Their proposed method used a lightweight model that reduced the computation required for local training on an edge device. Furthermore, the use of federated learning allows for continual adaptation of the model without

comprising user privacy. The federated learning approach (Salman & Busso, 2022) allows for the local unsupervised training of a dynamic FER model, where local models are trained using pseudo labels generated by an image FER, and then used to update the central model via FedAvg. By combining this federated learning methodology with partial label learning, our work presents a generalized pipeline for developing a model which addresses the various challenges inherent to dynamic FER, which we evaluate on the CREMA-D dataset.

2. Related Work

2.1 FER

Based on the form of input data, the FER task can be further categorized as image facial expression recognition and video or dynamic facial expression recognition. In dealing with video FER, several approaches (Kahou, Michalski, Konda, Memisevic, & Pal, 2015; Lee, Choi, Kim, & Song, 2019; Lu et al., 2018) utilized convolutional neural networks architectures such as the VGG (Simonyan & Zisserman, 2014) or ResNet (He, Zhang, Ren, & Sun, 2016) to capture spatial features. These are then often paired with a Recurrent Neural Network (RNN) to handle the temporal features. The Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) model is commonly used for this purpose.

2.2 Federated Learning

Federated learning is a decentralized approach for machine learning where the training process takes place across multiple client devices rather than in a centralized server (Zhang et al. 2021). In this setup, each client holds its own local dataset and computes updates to a local model. Instead of sending raw data to a central server, only the model updates (gradients) are communicated back and aggregated by a central entity known as the aggregator. This in turn preserves privacy as raw data is not shared.

3. Methodology

3.1 Models

Similarly to (Salman & Busso, 2022), the proposed pipeline relies on two models - an image facial expression recognition model (IFER) and a dynamic video facial expression recognition model (VFER), both of which share the same feature extractor. For our feature extractor and IFER, we leverage the pretrained EmoNet model from (Toisoul et al., 2021), due to being trained for image facial expression recognition on the large diversity of faces in the AffectNet dataset, as well as having a small size of 2M parameters, making it ideal for performing inference on edge devices. We extract the final classifier module of the EmoNet as our IFER, taking only the 6 relevant emotions in the CREMA-D dataset, and the rest of the EmoNet model is used as the shared feature extractor between the IFER and VFER. Both the feature extractor and IFER are kept frozen throughout our experiments. On the other hand, the VFER consists of an LSTM model that takes in the sequences of the extracted features from the EmoNet backbone before predicting the overall emotion of a video sequence. While the VFER is used for dynamic FER, the IFER is used to generate pseudo labels necessary for self-supervised federated learning, by averaging

the output of the IFER across every frame of the input video, and assigning the most probable class as the pseudo label if its maximum confidence, c passes the threshold of c=0.5. Otherwise, the sample is discarded and not used for training.

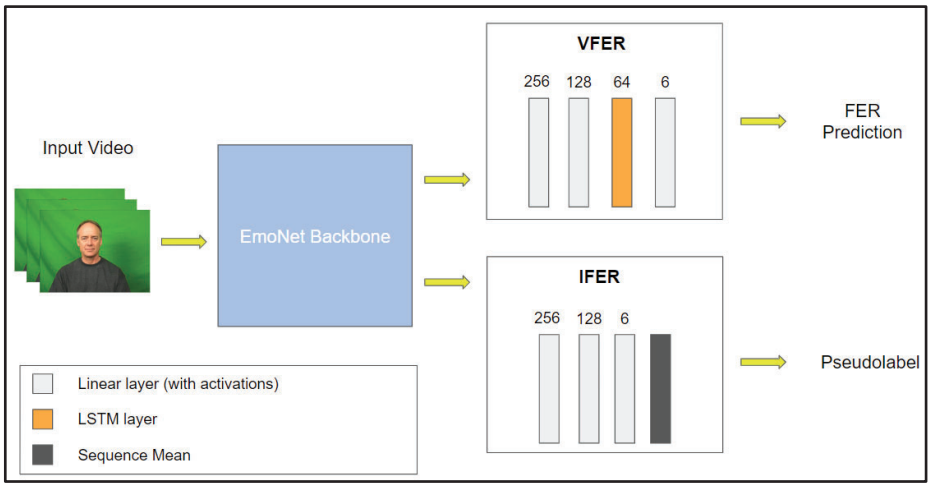

Figure 1. Model architectures for VFER and IFER

3.2 Using probability distributions as labels

Previous studies (Salman and Busso, 2022) have utilized the majority rule for allocating the final annotation to each clip. In this work, we improve upon the annotation scheme by using partial labels to reflect the relative probabilities of the annotations in the final label. As per (Wang, Weijie & Sebe, Nicu & Lepri, Bruno., 2022), modeling FER as a Partial Label Learning (PLL) task addresses subjective annotations and inherent similarities among various facial expressions. In our implementation, we convert the distribution of votes into a probability distribution using the softmax transformation. To ensure the emotions with zero votes correspond to a probability of zero, we first map every instance of zero votes to a large negative number before applying the softmax function. We then train our VFER on these probability distributions using the Kullback-Leibler Divergence loss (Kullback & Leibler, 1951).

Formula for transforming vote distributions into target probability distribution:

$$P(i) = Softmax(v_i') \; where \; v_i' = v_i \; if \; v_i \neq 0 \; else \; v_i' = -10^{-9}$$

3.3 Federated Learning

During the inference phase, we deploy separate copies of the VFER to multiple test subjects, which we refer to as the local model. In this phase, we first train the local VFER model for a given test subject by using a subset of the given subject's videos. To simulate federated learning, we discard the associated labels and assign the local videos with pseudo labels generated by the IFER instead. The local VFER copies are each then trained on the respective subject's newly labeled videos. Once all the local copies of the VFER models have been fine-tuned, FedAvg is used to aggregate the local copies into a single central model. Following this, the updated central model is evaluated on the unseen test set and the results are reported.

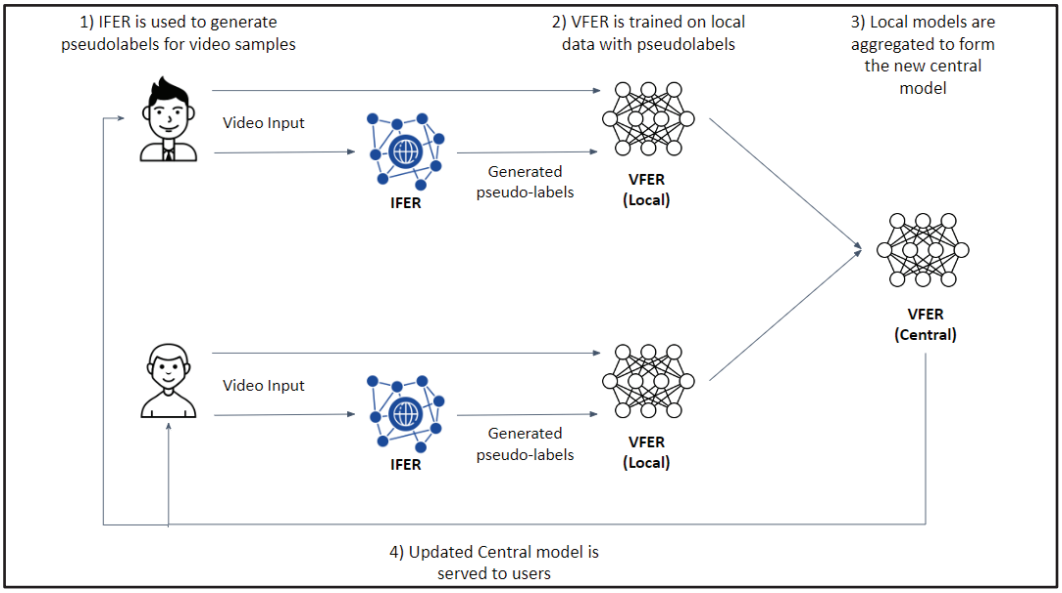

Figure 2. Federated learning framework for proposed dynamic facial expression recognition

4. Experimental evaluation

4.1 Dataset split

The dataset chosen is the CREMA-D dataset (Cao et al., 2014), which contains 7,442 audio-visual video clips from 91 actors spread across six emotion classes - happiness, sadness, neutral, anger, fear and disgust. Each clip has 4 to 12 annotations with 95% of the clips having 8 or more annotations. We select labels produced solely based on visual data, aligning with our model's exclusive use of video modality. This reduces noise from audio information unavailable to our model. To prevent data leakage and allow for an effective simulation of federated learning, we group video samples by actors before splitting the data into train, validation and test sets, ensuring that samples from the same actor only appear in one set. We also aim to keep the gender distribution in each set balanced. This results in a final split of 67 actors in our training set, 12 actors in our validation set, and 12 actors in our test set. For federated learning, the test set is split into 2 equal segments, with each segment containing half of the video samples of every actor. One segment will then be used as the adaptation set for federated learning, and the other is used as our test set for evaluating the adapted model.

4.2 Implementation

For initial training of the VFER model, we use the Adam optimizer with a learning rate of 0.0001 for 20 epochs. On each epoch, the model is evaluated on our validation set to detect for overfitting. We then save the VFER weights on the epoch which achieves the lowest validation loss. For tuning the hyperparameters for federated learning, we evaluated the performance of local models on the validation set after each epoch. This led to the observation that locally trained models experience severe catastrophic forgetting (McCloskey & Cohen, 1989), immediately demonstrating deteriorating validation loss from the first epoch. As a result, we found the ideal

hyperparameters for local adaptation to be only a single epoch of training with a low learning rate of 0.00001.

4.3 Results

We present the comparison of results on our proposed implementation of partial label learning on the CREMA-D dataset in Table 1. As shown in Table 1, using partial labels for the training of our initial VFER resulted in an improved model across measures of average precision, recall and f1 score on the testing set. Among the 6 emotions, the model trained on the partial labels achieved a higher average score in classifying Neutral, Happiness, Fear and Anger while underperforming in the classifying of Sadness and Disgust, emotions which occur the least frequently in the CREMA-D dataset.

Table 1: VFER trained using discrete vs partial labels

Emotion	Precision(%)		Recall(%)		F1_Score(%)	
	Discrete	**Partial labels**	Discrete	**Partial labels**	Discrete	**Partial labels**
Neutral	**72.3**	65.2	74.5	**75.2**	**73.4**	69.8
Happiness	81.6	**88.9**	94.1	94.1	87.4	**91.4**
Sadness	**37.5**	32.8	38.9	38.9	**38.2**	35.6
Fear	**58.2**	51.3	44.4	**55.6**	50.4	**53.3**
Disgust	61.0	**75.0**	**52.2**	43.5	**56.3**	55.0
Anger	49.4	**62.9**	**54.7**	52.0	51.9	**56.9**
Micro-mean	63.0	**64.7**	63.4	**63.6**	62.9	**63.3**
Macro-mean	60.0	**62.7**	59.8	**59.9**	59.6	**60.4**

Table 2 presents the effect of federated learning on the model initially trained on partial labels. Our results demonstrate that federated learning further enhances the performance of the initial VFER trained on partial labels with improvements in average precision, recall and f1 score.

Table 2: VFER Performance Before vs After Federated Learning

Emotion	Precision(%)		Recall(%)		F1_Score(%)	
	Before	After	Before	After	Before	After
Neutral	65.2	**67.1**	75.2	**77.4**	69.8	**71.9**
Happiness	88.9	**89.9**	94.1	94.1	91.4	**92.0**
Sadness	32.8	**35.7**	**38.9**	37.0	35.6	**36.4**

Fear	51.3	**52.0**	**55.6**	54.2	**53.3**	53.0
Disgust	75.0	**75.5**	43.5	**53.6**	55.0	**62.7**
Anger	**62.9**	60.0	52.0	52.0	**56.9**	55.7
Micro-mean	64.7	**65.4**	63.6	**65.2**	63.3	**64.9**
Macro-mean	62.7	**63.4**	59.9	**61.4**	60.4	**61.9**

5. Conclusion

In conclusion, our study introduces a training pipeline for dynamic Facial Expression Recognition (FER). We first verify the effectiveness of using a partial label learning paradigm to address subjective labels and intra-class similarities between facial expressions (Wang, Weijie & Sebe, Nicu & Lepri, Bruno., 2022) in the context of dynamic FER, demonstrating overall improvement in our initial dynamic FER model. Subsequently, to address the need for continual learning while preserving user privacy, we validate the use of a previously established federated learning technique (Salman & Busso, 2022), demonstrating even further improvements on our previously trained model's performance.

Ultimately, by leveraging the methodologies of both partial label learning and self-supervised federated learning, we establish a formal model development pipeline and demonstrate its effectiveness in addressing the challenges inherent to FER, providing a promising technique for practical applications in the quantification of learning in virtual classrooms.

Acknowledgements

We would like to thank the Artificial Intelligence, Analytics & Informatics (AI3) and Institute for Infocomm Research (I2R), A*STAR for supporting this work.

References

Kahou, S. E., Michalski, V., Konda, K., Memisevic, R., & Pal, C. (2015). Recurrent neural networks for emotion recognition in video. In ICMI (pp. 467–474).

Lee, M. K., Choi, D. Y., Kim, D. H., & Song, B. C. (2019). Visual scene-aware hybrid neural network architecture for video-based facial expression recognition. In FG (pp. 1–8).

Lu, C., Zheng, W., Li, C., Tang, C., Liu, S., Yan, S., & Zong, Y. (2018). Multiple spatio-temporal feature learning for videobased emotion recognition in the wild. In ICMI (pp. 646–652).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In CVPR (pp. 770–778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., & Eskofier, B. (2022). Federated learning for healthcare: Systematic review and architecture proposal. ACM Transactions on Intelligent Systems and Technology (TIST), 13(4), 1-23.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Roselander, J. (2019). Towards federated learning at scale: System design. Proceedings of machine learning and systems, 1, 374-388.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. IEEE transactions on affective computing, 5(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Dhall, A., Sharma, G., Goecke, R., & Gedeon, T. (2020, October). Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In Proceedings of the 2020 International Conference on Multimodal Interaction (pp. 784-789).

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885.

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, *13*(3), 1195-1215.

Long, G., Tan, Y., Jiang, J., & Zhang, C. (2020). Federated learning for open banking. In Federated Learning: Privacy and Incentive (pp. 240-254). Cham: Springer International Publishing.

Mandernach, B. J. (2015). Assessment of student engagement in higher education: a synthesis of literature and assessment tools. Int. J. Learn. Teach. Educ. Res. 12, 1–14. doi: 10.1080/02602938.2021.1986468

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.

McMahan, H. B., Moore, E., Ramage, D., & y Arcas, B. A. (2016). Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, *2*, 2.

Pandya, S., Srivastava, G., Jhaveri, R., Babu, M. R., Bhattacharya, S., Maddikunta, P. K. R., ... & Gadekallu, T. R. (2023). Federated learning for smart cities: A comprehensive survey. Sustainable Energy Technologies and Assessments, 55, 102987.

Salman, A., & Busso, C. (2022, November). Privacy preserving personalization for video facial expression recognition using federated learning. In Proceedings of the 2022 International Conference on Multimodal Interaction (pp. 495-503).

Shen, J., Yang, H., Li, J. et al. Assessing learning engagement based on facial expression recognition in MOOC's scenario. Multimedia Systems 28, 469–478 (2022). https://doi.org/10.1007/s00530-021-00854-x

Tsalera, E., Papadakis, A., Samarakou, M., & Voyiatzis, I. (2022). Feature Extraction with Handcrafted Methods and Convolutional Neural Networks for Facial Emotion Recognition. Applied Sciences, 12(17), 8455. https://doi.org/10.3390/app12178455

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. Knowledge-Based Systems, 216, 106775.'

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The Sequential Learning Problem. *Psychology of Learning and Motivation*, 109–165. https://doi.org/10.1016/s0079-7421(08)60536-8

Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., & Pantic, M. (2021). Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, *3*(1), 42–50. https://doi.org/10.1038/s42256-020-00280-0