Poster Abstract: Federated Learning for Speech Emotion Recognition Applications

Siddique Latif[*]
siddique.latif@usq.edu.au
University of Southern Queensland (USQ), Australia

Sara Khalifa
Distributed Sensing Systems Group, Data61, CSIRO
Australia

Rajib Rana
University of Southern Queensland (USQ), Australia

Raja Jurdak
Queensland University of Technology (QUT), Australia

ABSTRACT

Privacy concerns are considered one of the major challenges in the applications of speech emotion recognition (SER) as it involves the complete sharing of speech data, which can bring threatening consequences to people's lives. Federated learning is an effective technique to avoid privacy infringement by involving multiple participants to collaboratively learn a shared model without revealing their local data. In this work, we evaluated federated learning for SER using a publicly available dataset. Our preliminary results show that speech emotion recognition can benefit from federated learning by not exporting sensitive user data to central servers, while achieving promising results compared to the state-of-the-art.

CCS CONCEPTS

• **Computing methodologies → Distributed algorithms**; **Machine learning algorithms**; • **Human-centered computing → Human computer interaction (HCI)**; • **Security and privacy**;

KEYWORDS

Federated learning, deep neural networks, privacy preserving, speech emotion recognition

1 INTRODUCTION

Speech emotion recognition (SER) is an emerging area of research. It aims to automatically detect human emotion and affective states from speech. It has a widespread of applications in call centres, smart cars, forensics sciences, healthcare, etc [5]. Generally, in these applications, speech is recorded from users' devices and sent to the central server to be stored and analysed.

Speech data transmission from end devices are vulnerable to data hacking. As speech signal provides information about speaker identity, language and gender, therefore, eavesdroppers can use speech data for identification of a person and gain access to critical information. Federated learning (FL) is emerging as an efficient technique to avoid privacy infringement while allowing machine learning models to perform predictive tasks. FL environment provides greater control over users' data by incorporating privacy with distributed training and aggregation across a population of client devices.

Prior works on SER mainly focused on improving the performance of the system without considering speech privacy issues [3]. In this paper, we show the feasibility of SER using federated learning, which provides an alternative way to server-based systems, preserving the privacy of users. We train CNN and RNN based classifiers in a federated environment and achieve promising results compared to state-of-the-art. To the best of our knowledge, this is the first study to exploit the federated training paradigm for SER.

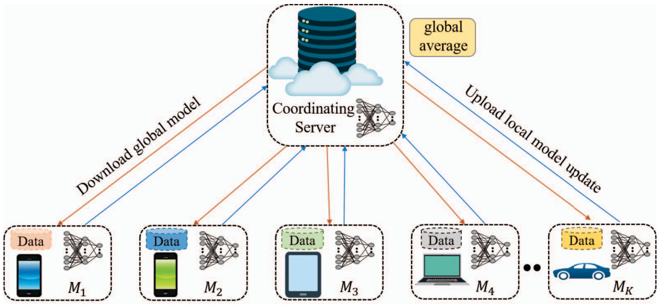


Figure 1: An illustration of the federated learning in distributed network of clients with SER application.

2 SYSTEM ARCHITECTURE

We propose to leverage federated learning (FL) for SER. FL is an efficient technique to enable distributed training of SER model by involving heterogeneous clients in a practical communication network as shown in Figure 1. In this scenario, clients can be mobile phones, laptops, autonomous vehicles, etc, which perform computing on their local emotional data without sharing to the server. This form of collaborative learning consists of a three-step protocol. In the first step, all the participating devices download the updated model from the central server. These participating clients compute an update of the downloaded model on their local speech data in the second step. In the third step, all participating clients upload their updated weights back to the server that aggregates these updates. This protocol is repeated until the convergence of certain criterion, i.e., emotion classification in this case. The benefit of this protocol is that clients only communicate updated weights instead of speech data, which remain secure locally. However, federated learning cause communication overhead and require computational resources at the clients. Recent communication technologies and the internet of things (IoT) devices fulfil these requirements.

In this work, we consider K participating devices with their local speech data. These devices have SER application for emotion detection. We used federated averaging (FedAvg) algorithm [4] to compute a global model by combining updates form clients. At training round r, coordinating server sends a global model w_t to a subset K of clients. Each participating client trains the current model

978-1-7281-5497-8/20/$31.00 ©2020 IEEE
DOI 10.1109/IPSN48710.2020.00-16

w_t on its local speech data and sends weights to the server. The sever computes global average from clients' weights to obtain a new global model w_{t+1}. In FedAvg, stochastic gradient descent (SGD) is used for optimisation and each device participates to approximate the global objective function. Other hyperparameters (i.e. learning rate and local epochs (E)) are assumed to be homogeneous among all devices in rounds r.

3 EVALUATIONS

3.1 Dataset

In this work, we used IEMOCAP [1] corpus for evaluation. We selected four basic emotions including happy, sad, angry, and neutral. IEMIOCAP data consists of five sessions where each session contains the recording of two speakers. We selected four sessions for training and utterances of one speaker from the remaining one session is used for validation and one speaker data for testing. We followed this study [2] for data augmentation to increase the size of the training set.

3.2 Learning Environment

We used convolutional neural network (CNN) for emotion classification, which consists of one convolutional layer, one max-pooling layer, and two fully connected layers followed by a softmax layer. We also implemented a recurrent neural network (RNN) based classifiers due to their popularity in speech modelling. Specifically, we used one long short term memory (LSTM) layer along with two fully connected layers and a softmax layer. We evaluated the proposed system architecture using these two classifiers.

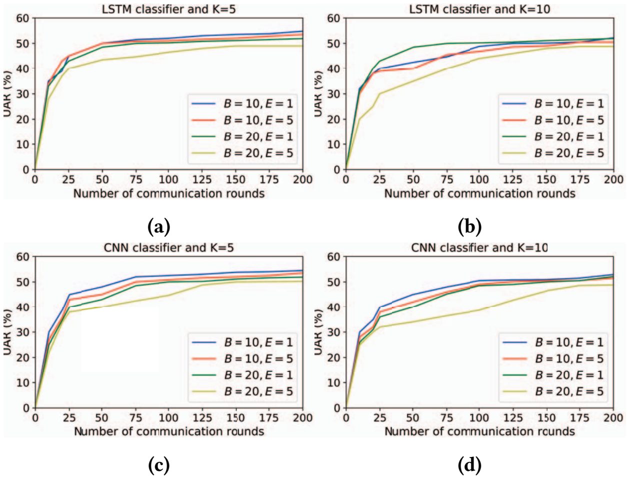
3.3 Results

We represented the speech in Mel Filterbank (MFB) features due to their popularity in speech processing applications. We used a Hamming window of 25-millisecond with a step size of 10-milliseconds to compute 40-dimensional MFB features. These features are given to both CNN and LSTM based classifiers. We computed the results using different number of clients (K=5 and 10). The training data is evenly partitioned for four classes into each set of clients. We followed the same notations for FedAvg as used in [4]. The batch size is represented with B, E shows the number of local epochs, and η represents the learning rate. We used B=10, 20, E=1 and 5, and η = 0.01. Results are reported in terms of unweighted average recall (UAR) as it is a popular evaluation metric in SER literature [2].

Figure 2 shows the results on IEMOCAP data using different parameter settings for both CNN and LSTM-based classifiers. LSTM-based classifier performs slightly better then CNNs. We achieve 54.8 UAR (%) using LSTM over 200 communication rounds. A strict comparison of results is difficult as there is no study on SER using federated learning. However, results are promising compared to benchmark results of 60.91% using variational autoencoders with LSTM [3] and 60.23 % using CNN-LSTM [2].

4 CONCLUSIONS

Federated learning (FL) is an effective technique to avoid privacy infringement in the applications of speech emotion recognition (SER) by keeping speech data local to different participating devices. In this work, we evaluated federated averaging (FedAvg) algorithms for SER on IEMOCAP dataset and achieve promising results compared to state-of-the-art studies. In our future work, we aim to utilise additional data to improve the performance of SER and also evaluate adversarial attacks in FL setting.



(a)

(b)

(c)

(d)

Figure 2: Test accuracy (UAR %) over 200 communication rounds using FedAvg algorithm with different parameters.

REFERENCES

[1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[2] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Julien Epps. 2019. Direct Modelling of Speech Emotion from Raw Speech. *Proc. Interspeech 2019* (2019), 3920–3924.

[3] Siddique Latif, Rajib Rana, Junaid Qadir, and Julien Epps. 2018. Variational autoencoders for learning latent representations of speech emotion: a preliminary study. *Interspeech 2018: Proceedings* (2018), 3107–3111.

[4] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. 2016. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629* (2016).

[5] Rajib Rana, Siddique Latif, Raj Gururajan, Anthony Gray, Geraldine Mackenzie, Gerald Humphris, and Jeff Dunn. 2019. Automated screening for distress: A perspective for the future. *European journal of cancer care* 28, 4 (2019), e13033.