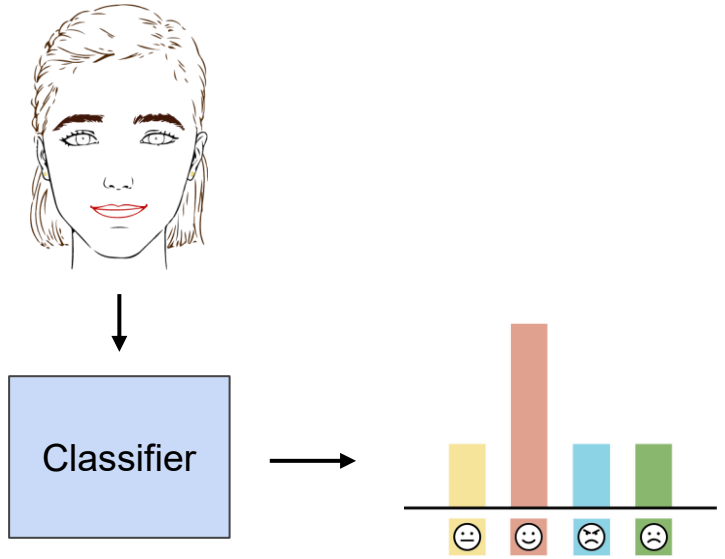
Privacy Preserving Personalization for Video Facial Expression Recognition Using Federated Learning

Ali N. Salman And Carlos Busso

Motivation

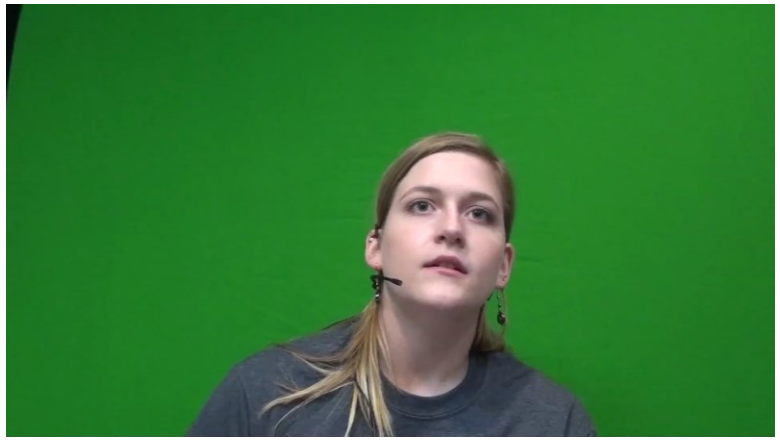
- **Emotions play an essential role in our decision making process**
- **Wide application domains**
  - Human-Computer/Robot Interaction
  - Driver distraction
  - Healthcare
  - Entertainment
- **Facial Expression Recognition (FER) is a challenging problem**
  - Different people express emotions differently
  - Different people perceive emotions differently
  - Preserving confidential information during FER

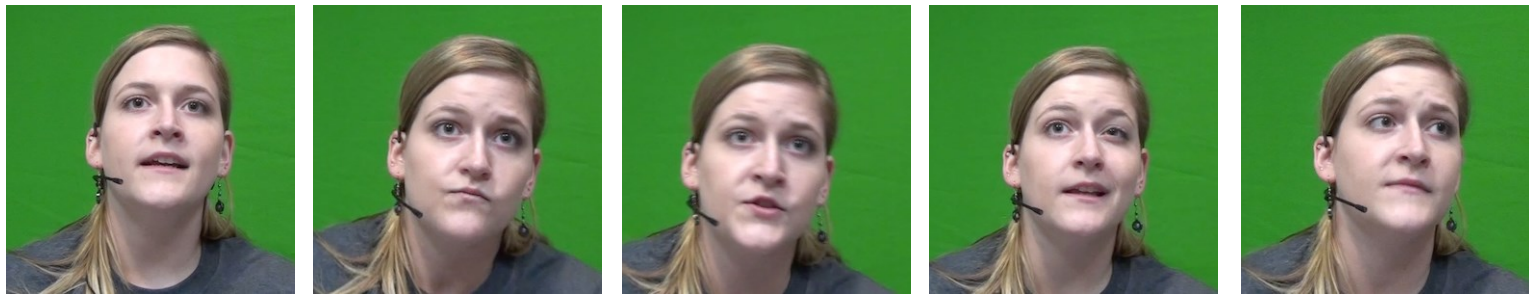


Related Work

- **Dynamic FER**
  - Emotions perceived from isolated frames is different from emotions perceived from watching corresponding video

No Audio





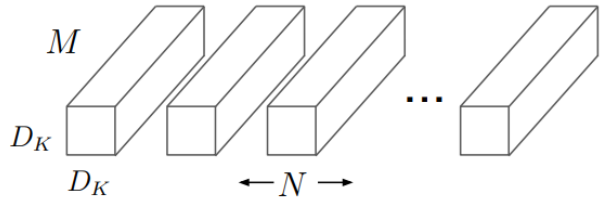
Label	Set	Precision	Recall	F1-Score
Happiness	Video/Video	0.91	0.84	0.87
	Video/Frame	0.67	0.97	0.79
Anger	Video/Video	0.73	0.67	0.70
	Video/Frame	0.55	0.14	0.22
Sadness	Video/Video	0.77	0.79	0.78
	Video/Frame	0.66	0.57	0.61
Neutral	Video/Video	0.72	0.72	0.72
	Video/Frame	0.54	0.77	0.63
Average	Video/Video	0.78	0.76	0.77
	Video/Frame	0.61	0.61	0.56

Table: Compares the perceptual evaluation between videos (different annotators) or videos compared to frames.

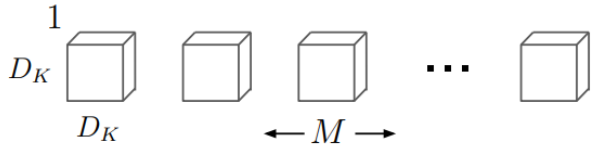
A.N. Salman and C. Busso. 2020. Dynamic versus Static Facial Expressions in the Presence of Speech. In IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). Buenos Aires, Argentina, 436-443.

Related Work

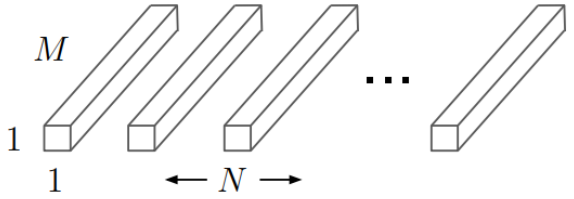
- **CNN Networks**
  - VGG 16
    - 138M parameters
    - 16 convolutional layers
  - ResNet50
    - 25M Parameters
    - 50 convolutional layers
  - MobileNetV2
    - 3.5M parameters
    - 105 layers (depth)
  - EfficientNetB0
    - 5.3M parameters
    - 132 layers (depth)



(a) Standard Convolution Filters



(b) Depthwise Convolutional Filters



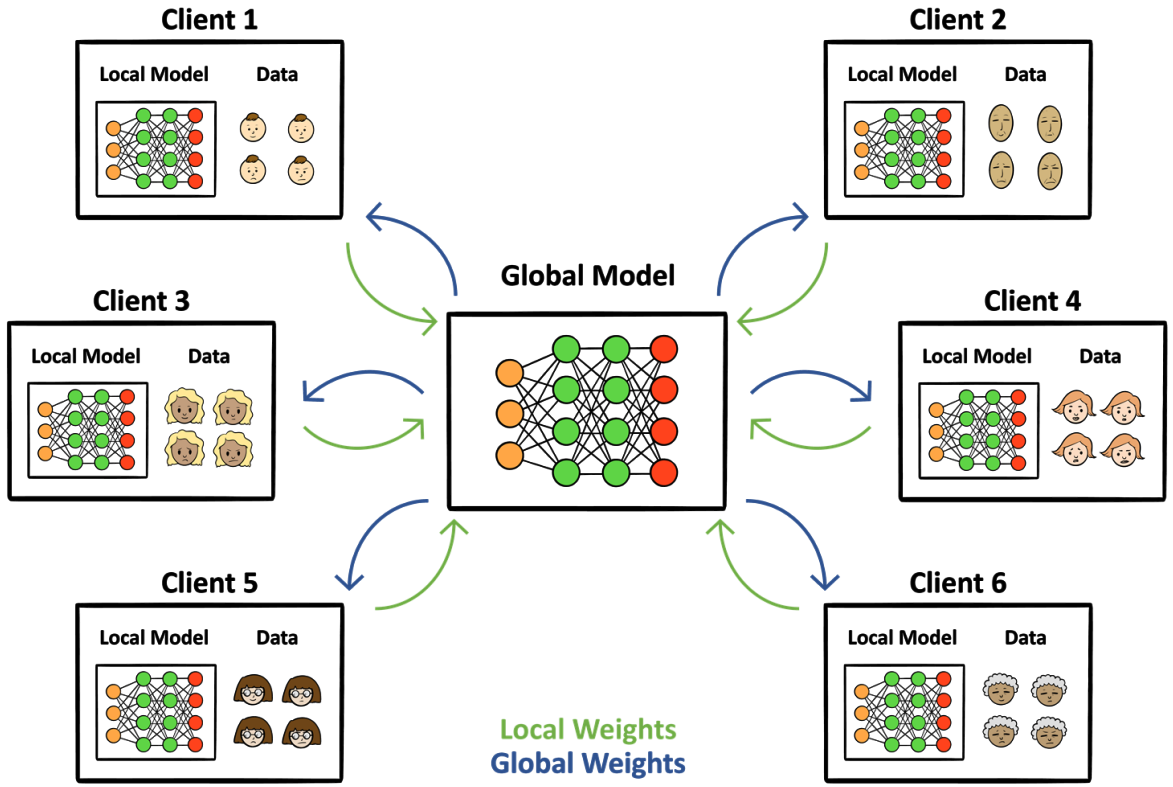
(c) 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

Figure 2. The standard convolutional filters in (a) are replaced by two layers: depthwise convolution in (b) and pointwise convolution in (c) to build a depthwise separable filter.

A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv e-prints (arXiv:1704.04861) (April 2017), 1–9. arXiv:1704.04861

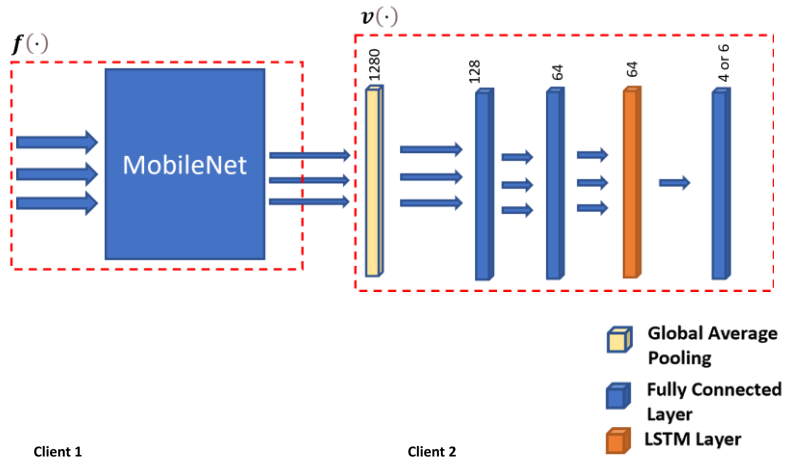
Related Work

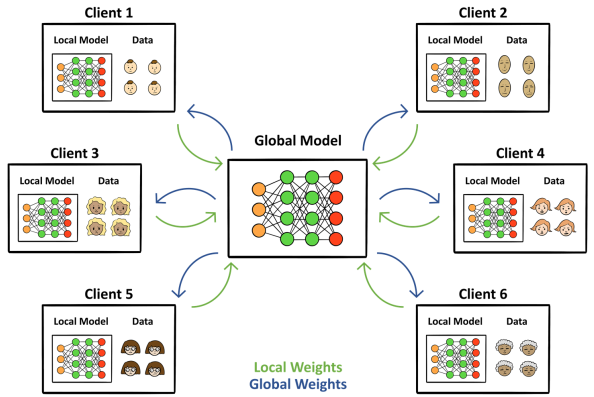
- **Federate Learning**
  - FedSGD
    - A single step of gradient per round
    - Requires more communication
  - FedAVG
    - Multiple steps of gradients per round
    - Less communication (vs FedSGD)
  - Procedure
    - 1. Start with a pre-trained or randomly initialized central model
    - 2. Distribute the central model to clients (local models)
    - 3. Train the local models on local data
    - 4. Send weights/gradients to the server
    - 5. Update the central model
    - 6. repeat steps 2-5



Goal

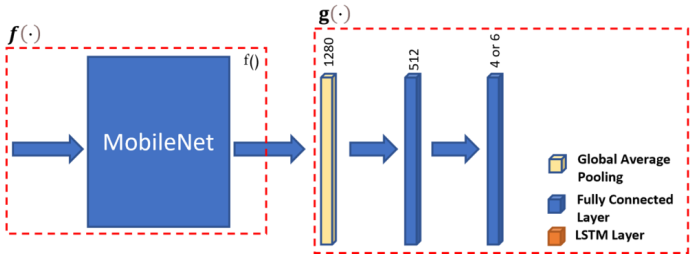
- **Dynamic FER**
  - Generate expression predictions utilizing static and temporal data
  - Personalize the model to each subject individually
  - Low computations
- **Federated Learning for FER**
  - Use FED to transmit model information instead of facial images between the client and server
  - Maintain the privacy of the users

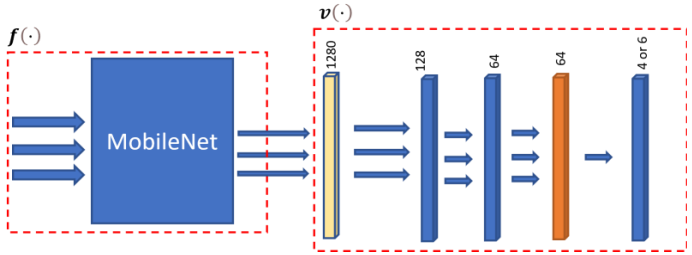




Proposed Model

- **Image FER (IFER)**
  - Generate expression predictions based on individual images
  - Leverage the large amounts of annotated images for FER (AffectNet)
- **Video FER (VFER)**
  - Generate expression predictions based on a sequence of images (video)
  - Able to capture temporal information and aggregate the data into a singular prediction
- **Federated Learning for FER**
  - Use FedAVG to maintain user privacy
  - Improve the local models by personalizing it
  - Improve the central model by only sharing a part of the local models





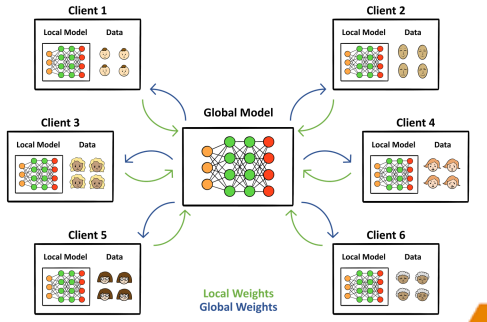
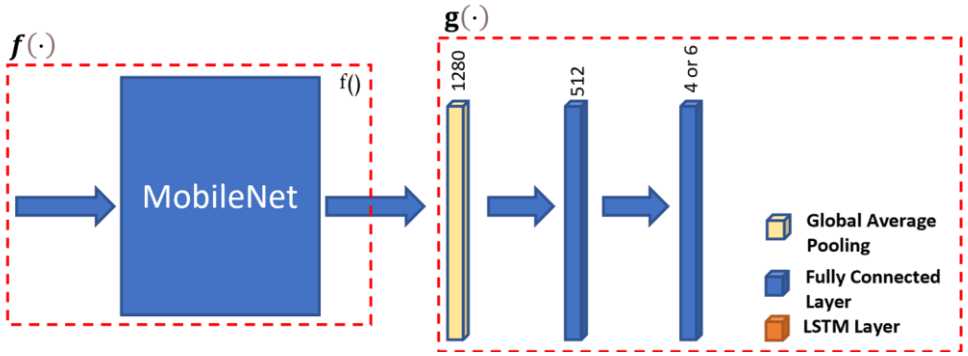


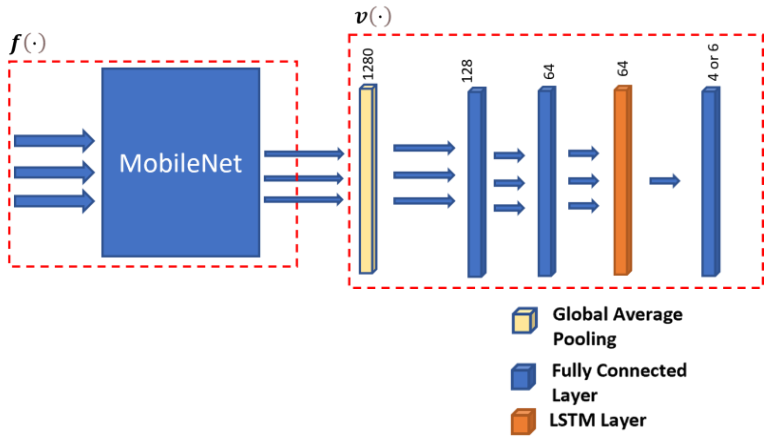
Image FER (IFER)

- **Model**
  - MobileNetV2
  - Initialize model using ImageNet weights
  - Train the model for the emotional classes using categorical cross-entropy.
- **Shared Features**
  - The output of the MobileNetV2 central average pool layer (shown in yellow)



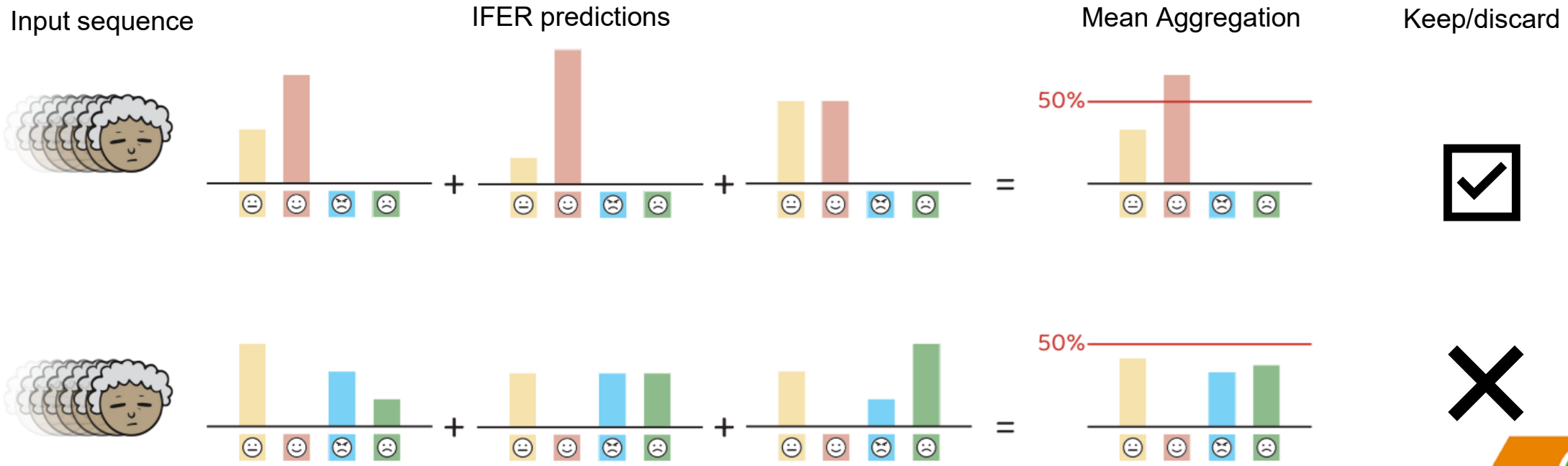
Video FER (VFER)

- **Model**
  - Use the same weights as the IFER model (MobileNetV2) up to the central average pooling layer (shown in yellow)
  - Add 2 fully connected layers (128 and 64 neurons each), followed by an LSTM layer to capture temporal information. Finally add a softmax layer.
  - Notice that the LSTM layer takes a sequence of latent features and returns a single latent vector (many-to-one)
  - Train the model for the emotional classes using categorical cross-entropy updating only the $v(.)$ weights.



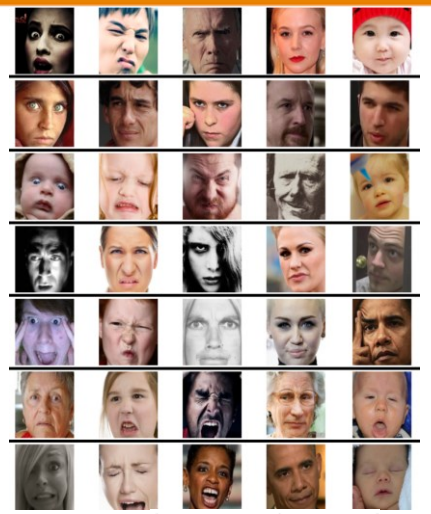
Personalization

- **Unsupervised strategy**
  - Use the IFER model to predict the FER distribution for each image in the video
  - Combine all the predictions for each video using mean aggregation
  - Keep P samples for each K emotional class ordered by highest confidence
  - Discard videos where confidence is $< P_{threshold}$
  - Train the model on the selected video samples



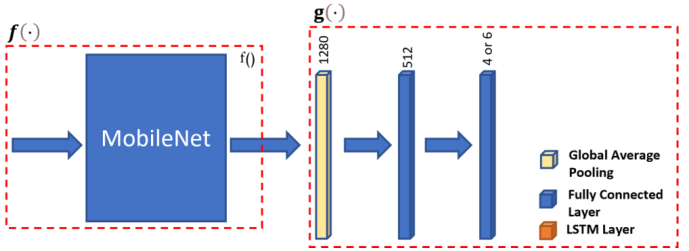
AffectNet Database

- **AffectNet [Mollahosseini et. al. 2019]**
  - Collected from the internet using major search engines
    - 1250 emotional keywords in 6 different languages
  - Over 1 million images
    - Around 440 thousand are manually annotated with seven discrete emotional labels
    - Valence and arousal annotation (not used in this study)
    - 425x425 average resolution
  - We consider 4 or 6 class formulations. Happiness, anger, sadness, and neutral state or happiness, anger, sadness, fear disgust, and neutral state
    - Downsample to 24,882 images per class (training set)
      - Random split 80/20 for training/validation
    - Validation set as testing set
  - This dataset is used to train the IFER



Neutral	75374
Happy	134915
Sad	25959
Surprise	14590
Fear	6878
Disgust	4303
Anger	25382
Contempt	4250
None	33588
Uncertain	12145
Non-Face	82915
Total	420299

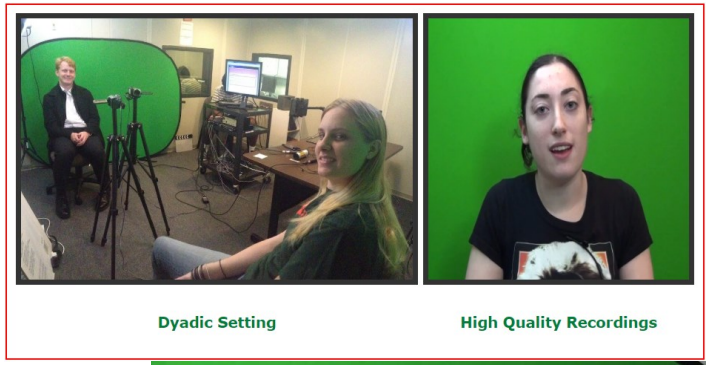
Number of images for each discrete label

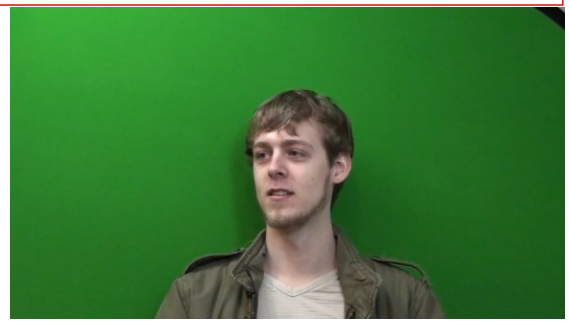


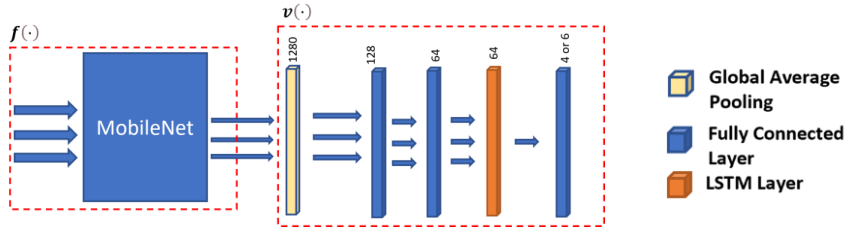
http://mohammadmahoor.com/affectnet/

MSP-IMPROV Dataset

- **MSP-IMPROV [Busso et. al. 2017]**
  - Multimodal emotional database
    - 12 subjects (six males, six females)
    - 1,440 x 1,080 resolution
    - Same sentences are spoken with different target emotions
      - Improvisations are used before/after to the target sentences to capture naturalistic data
    - We consider audio-visual annotations (happiness, anger, sadness, and neutral state)
      - 6/2/4 actors for train/validate/test (gender balanced)
  - This dataset is used to train the VFER model



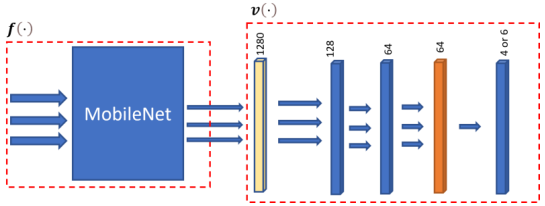




https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Improv.html

CREMA-D Dataset

- **CREMA-D [Cao et. al. 2014]**
  - Multimodal emotional dataset
    - 91 subjects (48 males, 43 females)
    - 960 x 720 resolution
    - Same sentences are spoken with different target emotions
    - Target sentences are manually annotated in different modalities
      - 7,442 annotated clips
      - We only consider audio-visual primary emotional labels (happiness, anger, sadness, fear, disgust, and neutral state)
      - 76/4/31 actors for train/validate/test (almost gender balanced)
  - This dataset is used to train the VFER model



https://github.com/CheyneyComputerScience/CREMA-D

Results - IFER

- **IFER**
  - Trained on a subset of the AffectNet database
  - Down sampled to at most 24,882 images per class
  - Results are reported on the validation set, which we use as our testing set

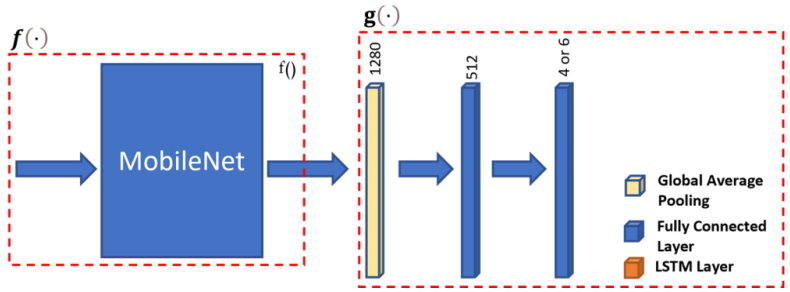


Table 1						
Emotion	Precision [%]	Recall [%]	F1-score [%]	Precision [%]	Recall [%]	F1-score [%]
Model		4class			6class	
Happiness	83.6	87.6	85.6	76	86.8	81
Anger	65.4	63.4	64.4	51.3	52.2	51.7
Sadness	67.9	61.8	64.7	53.1	53.2	53.1
Neutral	57.8	62	59.9	48.7	58.4	53.1
Fear	--	--	--	71.3	64.6	67.8
Disgust	--	--	--	60.4	44.3	51.1
Average	68.7	68.7	68.6	60.1	60	59.7

Results – VFER – CREMA-D

- **VFER**
  - Within corpus performance of the VFER model on CREMA-D
  - 6 emotional classes considered
  - Table reflects the performance on the test set (31 subjects)
  - Before: The performance on the central Model
  - After: The performance after the personalization step
  - Overall, all emotional classes saw an increase in F1-score except sadness, which saw a light decrease of 1.8%

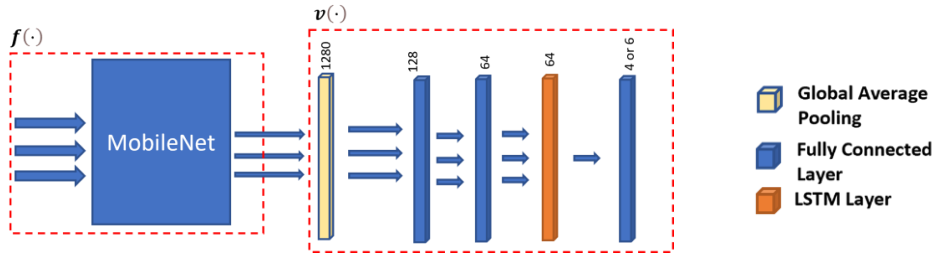


Table 2

CREMA-D Emotion	Precision [%] Before	Precision [%] After	Recall [%] Before	Recall [%] After	F1-Score [%] Before	F1-Score [%] After
Happiness	77.5	82	93.7	93.5	84.8 →	87.4
Anger	45.5	59.5	60.4	48.9	51.9 →	53.7
Sadness	34.3	38.1	29.6	24.8	31.8 ←	30
Neutral	67	62.4	51.6	73.4	58.3 →	67.4
Fear	47	49.1	46.4	48.4	46.7 →	48.7
Disgust	63.3	70.1	61.9	62.2	62.6 →	65.9
Macro mean	55.8 →	60.2	57.3 →	58.5	56 →	58.8
Micro mean	61.2 →	66.3	59 →	63.6	59.5 →	64.5

Results – VFER – MSP-IMPROV

- **VFER**
  - Within corpus performance of the VFER model on MSP-IMPROV
  - 4 emotional classes considered
  - Table reflects the performance on the test set (4 subjects)
  - Before: The performance on the central Model
  - After: The performance after the personalization step
  - Overall increase in F1-scores 2.3% (macro) and 6% (micro)

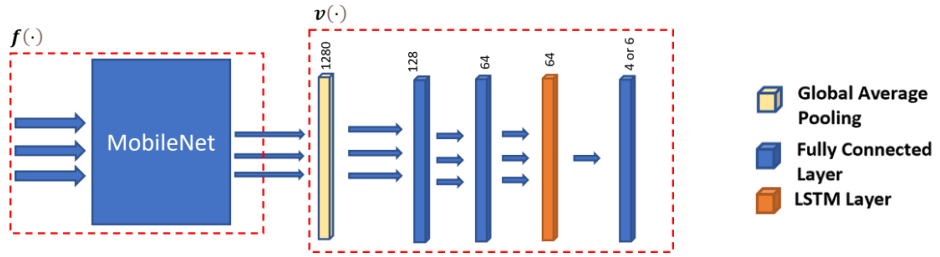
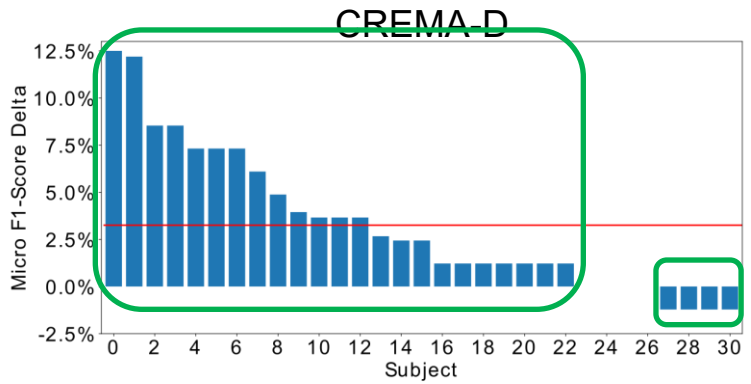


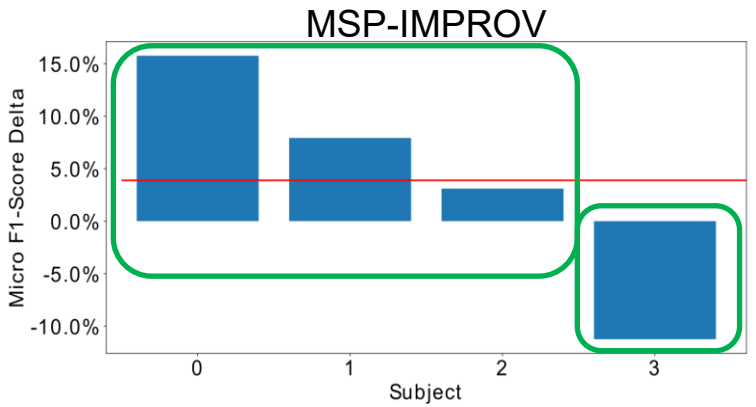
Table 3

MSP-IMPROV	Precision [%]		Recall [%]		F1-Score [%]	
Emotion	Before	After	Before	After	Before	After
Happiness	77.6	79.5	83.8	81.6	80.6	80.5
Anger	15.2	11.8	23	12.4	18.3	12.1
Sadness	20.5	25.6	60.2	51.9	30.6	34.2
Neutral	80.6	69.5	34	51.9	47.9	59.4
Macro Mean	48.5	46.6	50.2	49.4	44.3	46.6
Micro Mean	60.2	59.9	54.4	59.8	52.6	58.6

Results – VFER – Subject Analysis

- **CREMA-D**
  - 31 subjects total (test set)
    - 23 saw an increase in Micro F1-score from 1% up to almost 12.5%
    - 4 saw no change
    - 4 notice a decrease in performance
- **MSP-IMPROV**
  - 4 subjects in total
    - 3 show an increase in Micro F1-score
    - 1 subject show a decrease in F1-score





Results – VFER – Cross Corpus

- **CREMA-D**
  - Test on CREMA-D (4 class)
  - Before: The performance on the central model train on MSP-IMPROV
  - After: The performance after the personalization step on the local model
  - An increase for each emotional classes. Neutral being the highest (20.3%)
  - Overall increase of 13.1% macro f1-score and 9.3% micro F1-scores

Table 4

Emotion	Precision [%]		Recall [%]		F1-Score [%]	
	Before	After	Before	After	Before	After
Happiness	62.6	73.9	91.2	92.7	74.2	82.3
Anger	28	40.4	55.8	52.9	37.3	45.8
Sadness	25.7	34.9	27.6	53.3	26.6	42.2
Neutral	77.5	76.6	21.1	41.1	33.2	53.5
Macro mean	48.4	56.5	48.9	60	42.8	55.9
Micro mean	60.9	63.3	46.4	58.3	49.3	58.6

Results – VFER – Cross Corpus

- **MSP-IMPROV**
  - Test on MSP-IMPROV (4 class)
  - Before: The performance on the central model trained on CREMA-D (4 class)
  - After: The performance after the personalization step on the local model
  - Sadness and Neutral show a big improvement in F1-score, while happiness and sadness show a decrease in F1-score.
  - Increase of 1.2% macro and 3.5% micro mean F1-scores

Table 5

Emotion	Precision [%]		Recall [%]		F1-Score [%]	
	Before	After	Before	After	Before	After
Happiness	62.6	73.9	91.2	92.7	74.2 →	82.3
Anger	28	40.4	55.8	52.9	37.3 →	45.8
Sadness	25.7	34.9	27.6	53.3	26.6 →	42.2
Neutral	77.5	76.6	21.1	41.1	33.2 →	53.5
Macro mean	48.4 ←	56.5	48.9 ←	60	42.8 →	55.9
Micro mean	60.9 ←	63.3	46.4 →	58.3	49.3 →	58.6

Results – VFER FD – CREMA-D

- **CREMA-D**
  - Table 2
    - Before: The performance on the initial central model trained on CREMA-D
    - After: The performance after the personalization step on the local model
  - Table 6
    - Shows the performance after updating the central model using FedAVG
  - By updating the central model
    - All emotional classes increase, highest being neutral (3.4%) and lowest being fear (0.2%)
    - Macro and micro mean both show an improvement compared to the initial central model
    - Less performance compared to the personalized local models (After in Table 2)

Table 2 (Local Model)

CREMA-D Emotion	Precision [%]		Recall [%]		F1-Score [%]	
	Before	After	Before	After	Before	After
Happiness	77.5	82	93.7	93.5	84.8	87.4
Anger	45.5	59.5	60.4	48.9	51.9	53.7
Sadness	34.3	38.1	29.6	24.8	31.8	30
Neutral	67	62.4	51.6	73.4	58.3	67.4
Fear	47	49.1	46.4	48.4	46.7	48.7
Disgust	63.3	70.1	61.9	62.2	62.6	65.9
Macro mean	55.8	60.2	57.3	58.5	56	58.8
Micro mean	61.2	66.3	59	63.6	59.5	64.5

Table 6 (Central Model)

Emotion	Precision [%]	Recall [%]	F1-score [%]	Precision [%]	Recall [%]	F1-score [%]
Dataset	MSP-IMPROV			CREMA-D		
Happiness	80.2	79.7	79.9	79.9	92.5	85.7
Anger	12.2	12.9	12.6	59.5	50.5	54.7
Sadness	22.7	58.5	32.7	33.2	31.1	32.1
Neutral	73.2	48	58	63.1	62.2	62.7
Fear	--	--	--	45.2	47.9	46.5
Disgust	--	--	--	62.3	60.2	61.2
Macro mean	47.1	49.8	45.8	57.2	57.4	57.1
Micro mean	59.4	58	56	61.7	60.5	60.9

Results – VFER FD – MSP-IMPROV

- **MSP-IMPROV**
  - Table 2
    - Before: The performance on the initial central model trained on CREMA-D
    - After: The performance after the personalization step on the local model
  - Table 6
    - Shows the performance after updating the central model using FedAVG (2 rounds)
  - By updating the central model
    - An increase in neutral and sadness classes, with a decrease in happiness and anger
    - Macro and micro mean both shows an improvement compared to the initial central model
    - Less performance compared to the personalized local models (After in Table 2)

Table 3

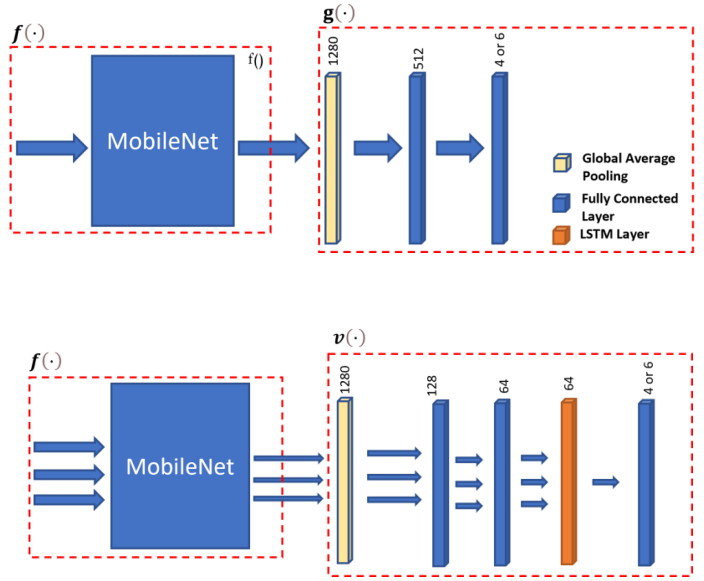
MSP-IMPROV	Precision [%]		Recall [%]		F1-Score [%]	
Emotion	Before	After	Before	After	Before	After
Happiness	77.6	79.5	83.8	81.6	80.6	80.5
Anger	15.2	11.8	23	12.4	18.3	12.1
Sadness	20.5	25.6	60.2	51.9	30.6	34.2
Neutral	80.6	69.5	34	51.9	47.9	59.4
Macro Mean	48.5	46.6	50.2	49.4	44.3	46.6
Micro Mean	60.2	59.9	54.4	59.8	52.6	58.6

Table 6 (Central Model)

Emotion	Precision [%]	Recall [%]	F1-score [%]	Precision [%]	Recall [%]	F1-score [%]
Dataset	MSP-IMPROV			CREMA-D		
Happiness	80.2	79.7	79.9	79.9	92.5	85.7
Anger	12.2	12.9	12.6	59.5	50.5	54.7
Sadness	22.7	58.5	32.7	33.2	31.1	32.1
Neutral	73.2	48	58	63.1	62.2	62.7
Fear	--	--	--	45.2	47.9	46.5
Disgust	--	--	--	62.3	60.2	61.2
Macro mean	47.1	49.8	45.8	57.2	57.4	57.1
Micro mean	59.4	58	56	61.7	60.5	60.9

Model Size

- **Model**
  - The model contains 2,223,872 trainable parameters
    - $g(.)$ contains 662,534 parameters
    - $v(.)$ contains 208,582 parameters (only 10% parameters increase)
  - Only the parameters in $v(.)$ are shared during the personalization approach
    - This reduces the bandwidth between the client and server, sharing only 208K parameters each FedAVG cycle
  - The $g(.)$ model can be discarded if training of the model is no longer necessary, further reducing the model size



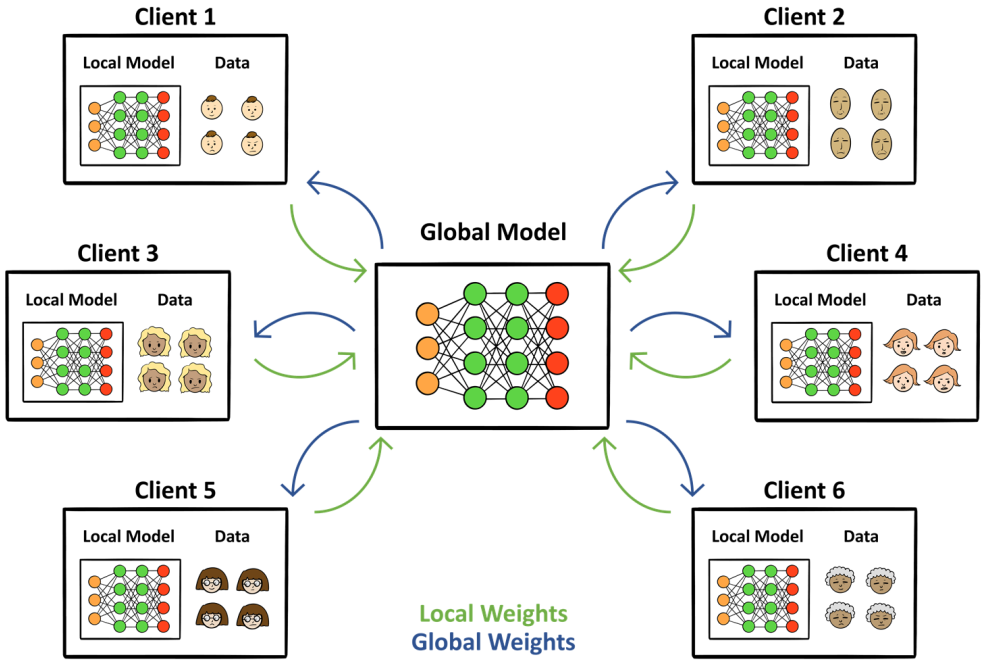
Conclusion

- **Proposed Approach**
  - A novel efficient IFER and VFER model that can predict both static and dynamic facial expressions
  - Proposed an unsupervised personalization strategy that can leverage the FedAVG approach
  - The proposed approach leads to improvements as high as 13.1% in F1-scores.

- **Future Research**
  - Explore ways to further reduce the size of the models and computations needed.
  - Explore transformer models for FER

Thank you

- **This work was funded by NEC Foundation and NSF under Grant IIS-1718944**



Our Research: msp.utdallas.edu