KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY (KIIT)
Deemed to be University U/S 3 of the UGC Act, 1956

2021 ICCV OCTOBER 11-17 VIRTUAL

# FedAffect: Few-shot federated learning for facial expression recognition


Debaditya Shome and T. Kar

KIIT University, Odisha, India


TRUST VISION

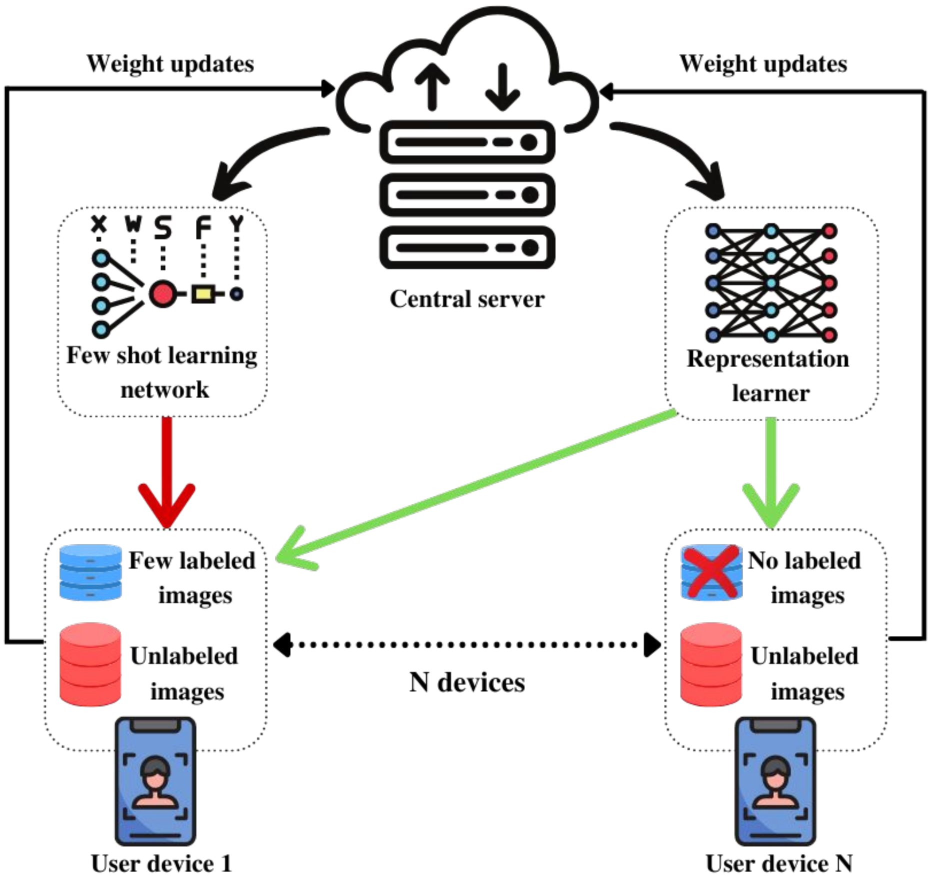
ICCV workshop: Human-centric Trustworthy Computer Vision From Research to Applications

17th October 2021

# Motivation

- Annotation of large-scale datasets in the real world is not feasible.
- Training models on large curated datasets often leads to dataset bias which reduces generalizability for real world use.
- Models fail to perform well on unseen faces.
- Fully-supervised approaches won’t scale.
- Real world user devices hold a rich collection of unlabeled facial data.
- Privacy concerns of facial data.

# FedAffect framework




**Algorithm 1** FedAffect framework

**Input:** number of devices $N$,
number of communication rounds $T$,
number of representation learner epochs $E1$,
number of classes $C$,
learning rate $\eta$

**Output:** Globally trained model weights $w_f^t$ and $w_g^t$

**Server executes:**

Initialize $w_f^0, w_g^0$
Fetch data availability information
**for** $t = 0, 1, ..., T-1$ **do**
  **for** $i = 1, 2, ..., N$ in parallel **do**
    **if** Number of labeled data samples at $i > C$ **then**
      Send $w_f^0, w_g^0$ to $i$
      $(w_f^t)_i, (w_g^t)_i \longleftarrow$ **LocalFewShot**$(i, w_f^t, w_g^t)$
    **end if**
    **if** $i$ has unlabeled data **then**
      $(w_f^t)_i \longleftarrow$ **LocalReprLearn**$(i, w_f^t)$
    **end if**
  **end for**
  $w_f^{t+1} \longleftarrow \sum_{k=1}^{N} \frac{D_i}{D} (w_f^t)_K$
  $w_g^{t+1} \longleftarrow \sum_{k=1}^{N} \frac{D_i}{D} (w_g^t)_K$
**end for**
**return** $w_f^t, w_g^t$

**LocalReprLearn**$(i, w_f^t)$**:**

Initialize projection network $p$
Initialize encoder network $f$ based on $w_f^t$
Set batch size B
**for** sampled minibatch $x_k$ from $k = 1$ to $B$ **do**
  **for all** $k \in (1, ..., B)$ **do**
    select two augmentation functions $t\ T, t\ T'$
    get first projection $z_{2k-1} = p(f(t(x_k)))$
    get second projection $z_{2k} = p(f(t'(x_k)))$
  **end for**
  $l(i,j) = -\log \frac{\exp(\frac{sim(z_i,z_j)}{\tau}))}{\sum_{k-1}^{2M} I_{[k \neq i]} \exp(\frac{sim(z_i,z_j)}{\tau})}$
  $L = \frac{1}{2B} \sum_1^B [l(2k-1, 2k) + l(2k, 2k-1)]$
  Update networks $f$ and $g$ to minimize $L$
**end for**
**return** updated weights of encoder, $w_f^t$

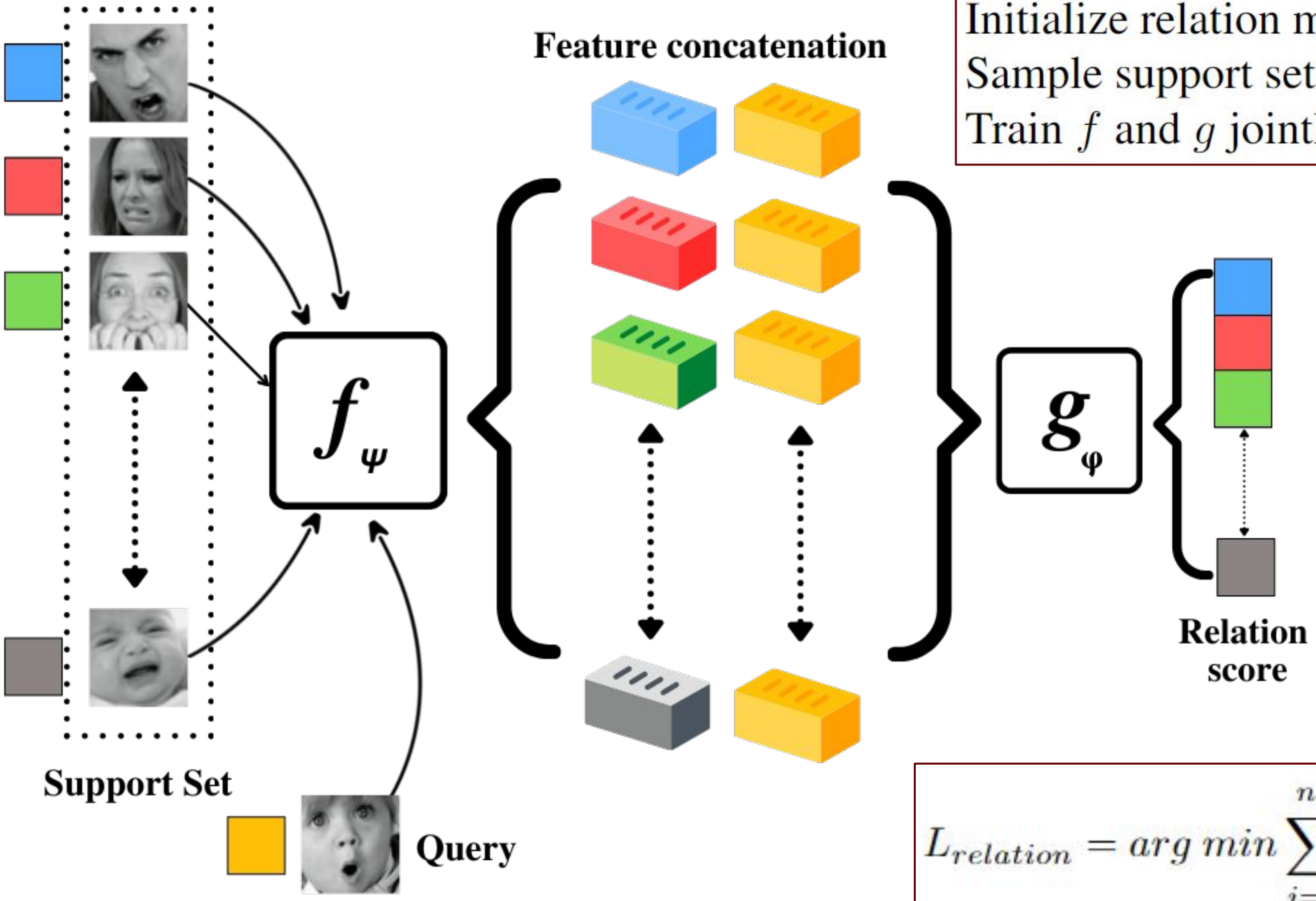
**LocalFewShot**$(i, w_f^t, w_g^t)$**:**

Initialize embedding module $f$ based on $w_f^t$
Initialize relation module $f$ based on $w_g^t$
Sample support set S and query Q
Train $f$ and $g$ jointly to minimize $L_{relation}$ from equation 2
**return** $(w_f^t)_i, (w_g^t)_i$

# Local self-supervised representation learning

**LocalReprLearn**($i$, $w_f^t$)**:**
Initialize projection network $p$
Initialize encoder network $f$ based on $w_f^t$
Set batch size B
**for** sampled minibatch $x_k$ from $k = 1$ to $B$ **do**
  **for all** $k \in (1, ..., B)$ **do**
    select two augmentation functions $t\ T$, $t\ T'$
    get first projection $z_{2k-1} = p(f(t(x_k)))$
    get second projection $z_{2k} = p(f(t'(x_k)))$
  **end for**
  $l(i,j) = -\log \frac{\exp(\frac{sim(x_i,x_j)}{\tau}))}{\sum_{k=1}^{2M} I_{[k\neq i]} \exp(\frac{sim(x_i,x_j)}{\tau})}$
  $L = \frac{1}{2B} \sum_1^B [l(2k-1, 2k) + l(2k, 2k-1)]$
  Update networks $f$ and $g$ to minimize $L$
**end for**
**return** updated weights of encoder, $w_f^t$

# Few-shot classifier




**LocalFewShot($i$, $w_f^t$, $w_g^t$):**
Initialize embedding module $f$ based on $w_f^t$
Initialize relation module $g$ based on $w_g^t$
Sample support set S and query Q
Train $f$ and $g$ jointly to minimize $L_{relation}$

$$L_{relation} = arg\ min \sum_{i=1}^{n_q} \sum_{j=1}^{n_s} (Y_{i,j} - 1(y_i == y_j))$$

# Global federated learning

**Server executes:**
Initialize $w_f^0, w_g^0$
Fetch data availability information
**for** $t = 0, 1, ..., T-1$ **do**
  **for** $i = 1, 2, ..., N$ in parallel **do**
    **if** Number of labeled data samples at $i > C$ **then**
      Send $w_f^0, w_g^0$ to $i$
      $(w_f^t)_i, (w_g^t)_i \longleftarrow$ **LocalFewShot**$(i, w_f^t, w_g^t)$
    **end if**
    **if** $i$ has unlabeled data **then**
      $(w_f^t)_i \longleftarrow$ **LocalReprLearn**$(i, w_f^t)$
    **end if**
  **end for**
  $w_f^{t+1} \longleftarrow \sum_{k=1}^{N} \frac{D_i}{D} (w_f^t)_K$
  $w_g^{t+1} \longleftarrow \sum_{k=1}^{N} \frac{D_i}{D} (w_g^t)_K$
**end for**
**return** $w_f^t, w_g^t$

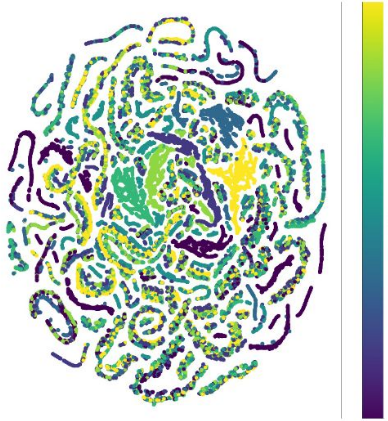
# Evaluation and Results

| Method | Overall accuracy |
|---|---|
| Multi-feature ensemble [37] | 97% |
| DeepExpr [4] | 89.02% |
| Centralized (ours) | 89.7% |
| **FedAffect (proposed)** | **97.3%** |

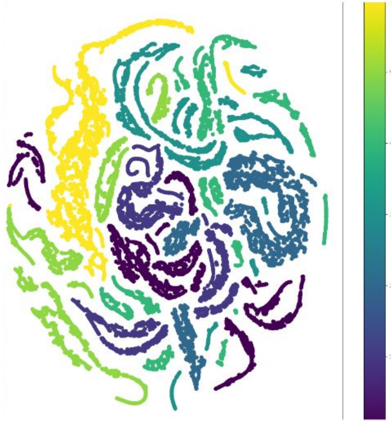
Table 1: Performance comparison on FERG dataset

| Method | Overall accuracy |
|---|---|
| CNN [35] | 65.97% |
| Ensemble ResMaskingNet [14] | 76.8% |
| RAN-VGG16 [31] | 89.16% |
| Centralized (ours) | 87.51% |
| **FedAffect (proposed)** | **84.9%** |

Table 2: Performance comparison on FER-2013 dataset



(a) Centralized learning



(b) Federated learning

# Conclusion and Future scope

- We tackle the problem of training facial expression recognition directly from decentralized privacy-sensitive data available on user devices.

- We propose FedAffect, a novel federated learning framework which collaboratively trains two disjoint neural networks for robust facial expression recognition.

- In the future, we aim to extend FedAffect to a Non-IID FL setup, with smart face cropping for dealing with in-the-wild facial expression data.