FedAffect: Few-shot federated learning for facial expression recognition

Debaditya Shome and T. Kar
KIIT University, Odisha, India

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.
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 \(k \in (1, \ldots, B)\) 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
  \(l(i, j) = -\log \frac{\exp(\frac{\text{sim}(x_i, x_j)}{\tau})}{\sum_{k=1}^{2M} I_{[k \neq i]} \exp(\frac{\text{sim}(x_i, x_j)}{\tau})}\)
  \(L = \frac{1}{2B} \sum_{k=1}^{B} [l(2k-1, 2k) + l(2k, 2k-1)]\)
end for
Update networks \(f\) and \(g\) to minimize \(L\)
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 $i$
      $(w_f^t)_i, (w_g^t)_i \leftarrow \text{LocalFewShot}(i, w_f^t, w_g^t)$
    end if
    if $i$ has unlabeled data then
      $(w_f^t)_i \leftarrow \text{LocalReprLearn}(i, w_f^t)$
    end if
  end for
  $w_f^{t+1} \leftarrow \sum_{k=1}^{N} \frac{D_i}{D} (w_f^t)_K$
  $w_g^{t+1} \leftarrow \sum_{k=1}^{N} \frac{D_i}{D} (w_g^t)_K$
end for
return $w_f^t, w_g^t$
### Evaluation and Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-feature ensemble [37]</td>
<td>97%</td>
</tr>
<tr>
<td>DeepExpr [4]</td>
<td>89.02%</td>
</tr>
<tr>
<td>Centralized (ours)</td>
<td>89.7%</td>
</tr>
<tr>
<td>FedAffect (proposed)</td>
<td><strong>97.3%</strong></td>
</tr>
</tbody>
</table>

Table 1: Performance comparison on FERG dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN [35]</td>
<td>65.97%</td>
</tr>
<tr>
<td>Ensemble ResMaskingNet [14]</td>
<td>76.8%</td>
</tr>
<tr>
<td>RAN-VGG16 [31]</td>
<td>89.16%</td>
</tr>
<tr>
<td>Centralized (ours)</td>
<td>87.51%</td>
</tr>
<tr>
<td>FedAffect (proposed)</td>
<td><strong>84.9%</strong></td>
</tr>
</tbody>
</table>

Table 2: Performance comparison on FER-2013 dataset
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.