Learning Emotional-Blinded Face Representations

Alejandro Peña, Julian Fierrez, Aythami Morales
BiDA Lab, Universidad Autónoma de Madrid, Spain
Ágata Lapedriza
Universitat Oberta de Catalunya, Spain
Massachusetts Institute of Technology, USA

Abstract

We propose two face representations that are blind to facial expressions associated to emotional responses. This work is in part motivated by new international regulations for personal data protection, which enforce data controllers to protect any kind of sensitive information involved in automatic processes. The advances in Affective Computing have contributed to improve human-machine interfaces but, at the same time, the capacity to monitorize emotional responses triggers potential risks for humans, both in terms of fairness and privacy. We propose two different methods to learn these expression-blinded facial features. Our experiments demonstrate that it is possible to eliminate information related to emotion recognition tasks in the face representations, while the performance of subject verification, gender recognition, and ethnicity classification are just slightly affected.

How emotions are expressed in face images

We study facial expressions [3] related to 6 different emotional responses.

Neutral | Happy | Sad
Disgusted | Angry | Surprised

Proposed framework

We employ a privacy-preserving learning framework including domain adaption from a pre-trained face representation network to different face analysis tasks, with and without the emotional-blinded representation.

Emotional-blinded face representations

- We have adapted two existing methods to the problem of learning emotional-blinded representations, namely SensitiveNets [1] and Learning Not To Learn [2], and applied them effectively to the face recognition pre-trained model ResNet-50.
- We evaluated the performance of both the original representation $x$ and the emotional-blinded $\varphi(x)$ on subject verification in LFW, gender and ethnicity classification in DiveFace [1], and emotion classification in CFEE [3]. Our experiments demonstrate that we can remove emotion related information, while preserving the performance in the other tasks.

<table>
<thead>
<tr>
<th>Domain</th>
<th>$x$</th>
<th>$\varphi(x)$</th>
<th>Diff. SN</th>
<th>$\varphi_{SN}(x)$</th>
<th>Diff. LnL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>96.8%</td>
<td>96.3%</td>
<td>↓ 1%</td>
<td>59.4%</td>
<td>↓ 75%</td>
</tr>
<tr>
<td>Gender</td>
<td>99.2%</td>
<td>98.9%</td>
<td>↓ 1%</td>
<td>72.7%</td>
<td>↓ 53.9%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>98.8%</td>
<td>98.6%</td>
<td>↓ 1%</td>
<td>67.4%</td>
<td>↓ 47.9%</td>
</tr>
<tr>
<td>Emotion (NN)</td>
<td>88.1%</td>
<td>59.6%</td>
<td>↓ 40%</td>
<td>41.6%</td>
<td>↓ 65%</td>
</tr>
<tr>
<td>Emotion (SVM)</td>
<td>88.1%</td>
<td>16.7%</td>
<td>↓ 100%</td>
<td>25%</td>
<td>↓ 88.2%</td>
</tr>
<tr>
<td>Emotion (RF)</td>
<td>77.4%</td>
<td>58.3%</td>
<td>↓ 31%</td>
<td>44.7%</td>
<td>↓ 53.8%</td>
</tr>
</tbody>
</table>

Blind representations: Training fairer classifiers

- Case study: Attractiveness classification in presence of facial expression biases using CelebA dataset.
- We design a 40K images training set, where 70% of attractive people are smiling, while 70% of unattractive people don’t (both groups presenting 20K images).
- The emotional-blinded representations reduce the gap between both groups, and therefore improve the Equal of opportunity criterion.

Conclusions

The development of automatic systems capable of reading emotions without consent triggers a potential risk to user’s privacy. In this work, we have adapted two methods for the purpose of generating facial representations that are blind to facial expressions associated to emotional responses. Our experiments demonstrate that is possible to reduce the performance of emotion classifiers while maintaining competitive performance in other face analysis tasks.

Main references