Self Supervised Facial Representation Learning for Facial Expression Recognition (FER)

T . Satya Pranav (2201213)

Abstract

This study uses the Facial Regional Awareness (FRA) framework with Self-Supervised Learning (SSL) to extract global and local facial features via region-aware heatmaps, offering a FER-optimized, region-based alternative to general SSL methods like MoCo or FER-specific models like Crowd FER.

Problem Statement

FER struggles with annotation ambiguity, limited labeled data, and ineffective capture of global and local facial features, reducing accuracy in real-world applications.

Philosophy

Enhance the FRA framework using SSL with region-aware heatmaps and diversity loss to improve feature extraction and emotion recognition, leading to more accurate and reliable performance.

Method/Approach

The Proposed Method: **FRA** introduces a dual-branch Siamese architecture with:

Online Network:Updated via backpropagation.

Momentum Network: Updated via EMA Exponential Moving Average: $\xi \leftarrow 0.996\xi + 0.004\theta$

Global Branch: Extracts holistic features

Local Branch: Focuses on regions using heatmaps

Heatmap Generation: Uses cosine similarity and Softmax: with β=10.

$$\mathrm{M}_i = softmax(\beta S_i) \;\; S_i = cosine\_similarity(Q_i, F_i^{dense})$$

Semantic Consistency:

$$\mathbf{h}_i^m = \mathbf{M}_i^{(m)} \otimes \mathbf{F}_i = \frac{1}{\sum_{u,v} \mathbf{M}_i[m,u,v]} \sum_{u,v} \mathbf{M}_i[m,u,v]\mathbf{F}_i[*,u,v]$$

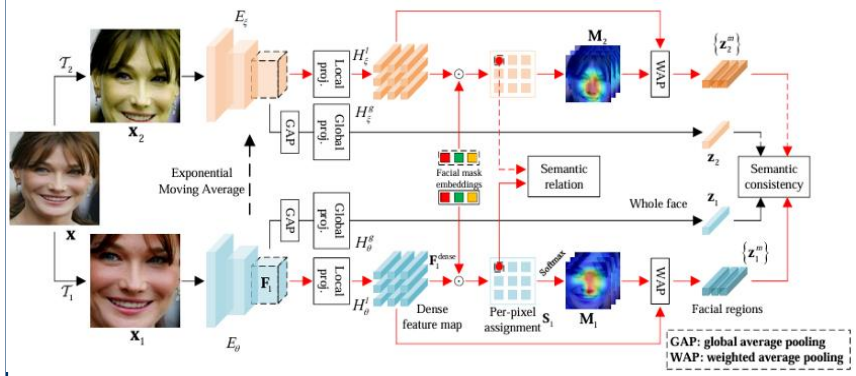
Total Loss Function: $\mathcal{L} = \mathcal{L}_c + 0.1\mathcal{L}_r$

$$\mathcal{L}_c = \mathcal{L}_{\text{sim}}(z_1, z_2) + \mathcal{L}_{\text{sim}}(z_2, z_1)$$

$$\mathcal{L}_{\text{sim}}(z_1, z_2) = -\left(\lambda_c \cdot \cos(z_1, z_2) + (1-\lambda_c) \cdot \frac{1}{N}\sum_{m=1}^{N} \cos(z_1^m, z_2^m)\right)$$

$$\mathcal{L}_{\text{r}} = \frac{1}{HW}\sum_{u,v}(CE(\mathbf{s}_1^{u,v}, \widehat{\mathbf{s}}_1^{u,v}) + CE(\mathbf{s}_2^{u,v}, \widehat{\mathbf{s}}_2^{u,v}))$$

With concise and less noisy dataset and a simplified ResNet50 implemented With full fine-tuning, achieved 76.3% Acc@1(paper's 66.16%).



Proposed improvements

Further reduction in dataset for experimentation

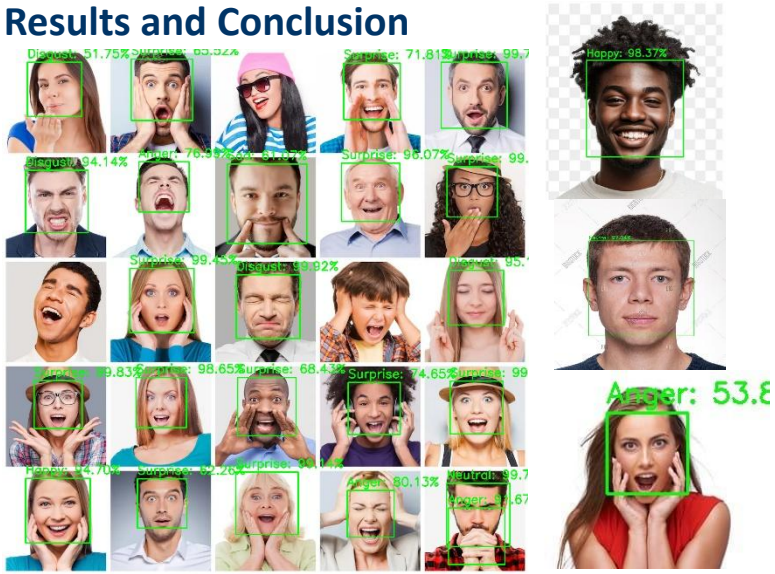
Introduced a temperature parameter τ=1.1 in heatmap generation: $M(m,u,v) = \frac{\exp(S(m,u,v)/\tau)}{\sum_{k=1}^{N}\exp(S(k,u,v)/\tau)}$,

Introduced a diversity loss to encourage attention to diverse facial regions $\delta = -\sum_{m=1}^{N} \bar{M}(m)\log\bar{M}(m)$

$\bar{M}(m) = \frac{1}{HW}\sum_{u=1}^{H}\sum_{v=1}^{W} M(m,u,v).$

Final Total Loss : $L_{\text{mod}} = L_{\text{SS}} - \lambda\,\delta$

Results and Conclusion



By integrating a diversity loss, improving robustness and feature representation through self-supervised learning. Experimental refinements over 30 epochs achieved a accuracy of 67.15% (original 66.95%).