Self-Supervised Facial Representation Learning with Facial Region Awareness

Focusing on Facial Expression Recognition (FER)

T.Satya Pranav

IIIT Guwahati

April 2, 2025

- **Objective**: Explore how FRA leverages self-supervised learning to enhance FER by capturing both global and local facial features.

Problem Statement and Motivation

- **FER Overview**: Recognizes emotions (e.g., happy, sad, angry) from facial images, vital for human-computer interaction, mental health analysis, etc.
- **Key Challenges**:
  - Annotation Ambiguity: Subjective interpretations lead to noisy labels.
  - Feature Representation: Requires both global (whole face) and local (eyes, mouth) features for robust FER.
  - Data Scarcity: Limited labeled datasets impede supervised learning.
- **Motivation**: Can self-supervised learning (SSL) with facial region awareness (FRA) provide a solution by learning rich representations without extensive labeling?

Background: SSL and FRA

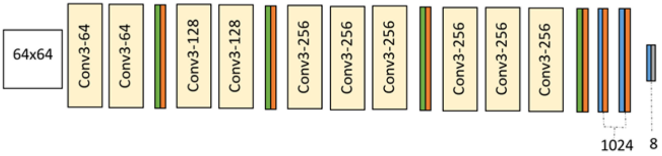
- **Self-Supervised Learning (SSL)**:
  - Learns representations via pretext tasks (e.g., contrastive learning, masked modeling).
  - Example: Momentum Contrast (MoCo), Masked Autoencoders (MAE), etc.
- **Facial Region Awareness (FRA)**:
  - Non-contrastive SSL framework enhancing FER by enforcing consistency between global and local facial representations.
  - Uses heatmaps to focus on key regions (e.g., eyes, mouth).
  - Built on BYOL

Literature Survey: Overview

- **Relevant Works**:
  1. **FRA**: Core focus—SSL with region-aware heatmaps for FER.
  2. **"Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution"**: Probabilistic modeling for noisy FER labels.
  3. **"Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for FER"**: Latent distribution and uncertainty for ambiguity in FER.
  4. **"Momentum Contrast for Unsupervised Visual Representation Learning"**: Contrastive SSL foundation for FRA.
  5. **"General Facial Representation Learning in a Visual-Linguistic Manner" (FaRL)**: Visual-linguistic SSL for facial tasks.
- **Goal**: Understand how FRA builds on these to address FER challenges.

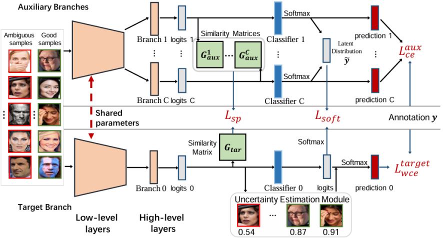
Training Deep Networks for FER with Crowd-Sourced Label Distribution

- **Objective**: Train deep CNNs for FER with noisy crowd-sourced labels.
- **Architecture**: VGG13-based DCNN:
  - 10 convolutional layers, 5 max-pooling layers.
  - Dropout (0.5), 2 dense layers (1024 nodes), softmax output.
- **Method**: Four training schemes:
  - Majority Voting, Multi-Label Learning, Probabilistic Label Drawing, Cross-entropy loss
  - Cross-Entropy: $\mathcal{L} = -\sum_{i=1}^{N} \sum_{k=1}^{8} p_k^i \log q_k^i$.
- **Network Layers**: Yellow (convolution), green (max-pooling), orange (dropout), blue (fully connected), and gray (softmax) layers.



Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for FER

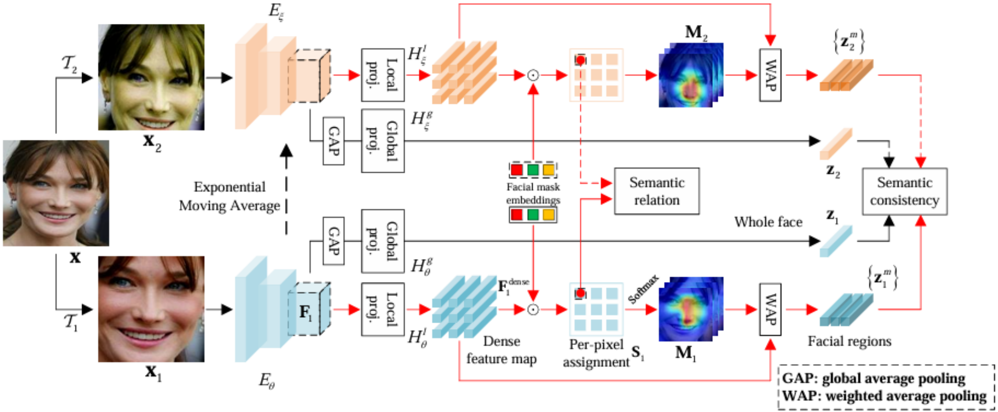
- **Objective**: Address annotation ambiguity in FER.
- **Architecture**: ResNet-18 with:
  - C auxiliary branches (each a $(C-1)$-class classifier).
  - 1 target branch for final prediction.
- **Method**:
  - Latent Distribution Mining: Predicts label probabilities for ambiguous samples using auxiliary branches. $\tilde{y}_x$.
  - Uncertainty Estimation: Assesses sample ambiguity using relationships between data (yx and yx').
  - Loss: $L_{\text{total}} = w_u(e)(L_{WCE}^{\text{target}} + \omega L_{\text{soft}} + \gamma L_{sp}) + w_d(e) L_{CE}^{\text{aux}}$.



FRA Methodology: Overview

- **Core Idea**: Learn robust FER representations by ensuring consistency between global and local features using SSL.
- **Pipeline**:
  1. Input: Two augmented views $\mathbf{x}_1, \mathbf{x}_2$ of a facial image.
  2. Encoder: ResNet extracts feature maps $\mathbf{F}_i \in \mathbb{R}^{H \times W \times D}$.
  3. Heatmap Generation: Transformer decoder outputs $\mathbf{M}_i \in \mathbb{R}^{H \times W \times N}$ for N regions.
  4. Representation: Global $\mathbf{z}_i$, local $\mathbf{z}_i^m$ embeddings.
  5. Consistency: Match embeddings across views via loss functions.
- **Online Network**: Encoder E_θ, projectors H_θ^g, H_θ^l.
- **Momentum Network**: Updated via EMA: $\xi \leftarrow 0.996\xi + 0.004\theta$.

Overview of the Proposed FRA Framework



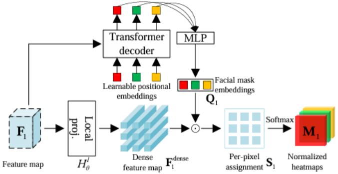
- **Process**:
  - Feature Maps: $\mathbf{F}_i$ from ResNet.
  - Per-Pixel Assignments: $\mathbf{S}_i$ computed as cosine similarity between facial mask embeddings $\mathbf{Q}_i$ and dense feature map $\mathbf{F}_i^{\text{dense}}$.
  - Heatmaps: $\mathbf{M}_i = \text{softmax}(\beta \mathbf{S}_i)$, where $\beta = 10$.
- **Semantic Consistency**:

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

- **Feature Representations**:

$$\mathbf{z}_1^m = H_\theta^l(h_1^m), \quad \mathbf{z}_2^m = H_\xi^l(h_2^m)$$

$$\mathbf{z}_1 = H_\theta^g(\text{GlobalPool}(\mathbf{F}_1)), \quad \mathbf{z}_2 = H_\xi^g(\text{GlobalPool}(\mathbf{F}_2))$$



FRA: Loss Functions

- **Semantic Consistency Loss ($\mathcal{L}_c$)**:

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

The Similarity Loss ($\mathcal{L}_{\text{sim}}$) is defined as:

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

- **Semantic Relation Loss ($\mathcal{L}_r$)**:

$$\mathcal{L}_r = \frac{1}{HW} \sum_{u,v} \left(\text{CE}(s_{u,v}^1, \hat{s}_{u,v}^1) + \text{CE}(s_{u,v}^2, \hat{s}_{u,v}^2)\right)$$

$$\text{CE}(s_{u,v}^1, \hat{s}_{u,v}^1) = -\sum_{m=1}^{N} \hat{s}_{u,v}^1[m] \log s_{u,v}^1[m]$$

- **Total Loss ($\mathcal{L}$)**:

$$\mathcal{L} = \mathcal{L}_c + 0.1\mathcal{L}_r$$

Work Done and Results

- **Method Used**: Implemented FRA:
  - **Dataset**: Pre-trained on full VGGFace2 (3.3M images), fine-tuned on AffectNet (280K training images, 7 classes).
  - **Architecture**: ResNet-50 (24M parameters), with Transformer decoder (1 layer) for heatmap generation.
  - **Parameters**: Batch size 256, epochs aligned with BYOL defaults, $\lambda_g = 0.5$, $\lambda_r = 0.1$.
- **Results on AffectNet**:

Method	Accuracy (%)	Comparison
BYOL (Baseline)	65.65	-
FRA (Fine-Tuned)	66.16	+0.51%

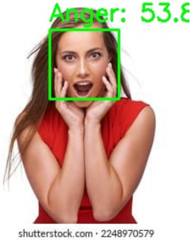
Table: Performance Comparison

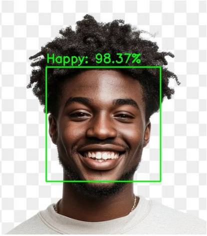
- **Analysis**: FRA's region-aware heatmaps enhance subtle expression detection, outperforming BYOL.

Conclusion

- **Summary**: Adapted ResNet50 (VGGFace2 pre-trained) for FER on reduced AffectNet7, achieving 76.3% Acc@1.
- **Work Conducted**:
  - **Dataset**:
    - VGGFace2: Reduced from 40GB to 5GB for pre-training.
    - Validation: Replaced LFW pairs with CFP for efficiency.
    - AffectNet: Reduced from 1.4GB to 240MB (approx. 9216 images, 7 classes).
  - **Architecture**: Simplified ResNet50 by removing residual blocks in layers 3 and 4.
  - **Final Parameters**: LR: 0.0000489, batch size: 256, epochs: 80.
- **Results**:
  - Achieved 76.3% Acc@1 vs. paper's 66.16%.
  - **Why the Difference?**
    - *Dataset*: Paper used full AffectNet (noisier); we used cleaner, reduced AffectNet7.
    - *Fine-Tuning*: Full layer adaptation vs. paper's possible limited tuning.
- **Future Work**: Test on full dataset, optimize reduced architecture further.

Images with Labels





Thank You!