Technical Summary: Face Detection and Matching System COMSYS Hackathon 5

Aritra Banerjee

July 4, 2025

Task A: Face Detection (AMR-CD)

Objective: Detect whether a given face belongs to a known identity using both grayscale and RGB inputs.

Methodology: The AMR-CD model employs a dual-branch architecture to independently process grayscale and RGB images using ResNet18 and EfficientNet-B0, respectively. Each modality contributes through separate auxiliary heads, and their features are passed through 1x1 convolutions for dimensional alignment. These representations are then fused using a learned attention mechanism that adaptively weighs grayscale and RGB features.

The fused vector, which reflects both texture-level (gray) and color-based (RGB) cues, is passed through a fully connected classifier for the final binary decision. This multi-representation approach enables robustness to lighting, modality shifts, and partial occlusions.

Architecture:

- ResNet18 for grayscale images (1-channel).
- EfficientNet-B0 for RGB images.
- Each path has its own auxiliary classifier.
- Fusion via learned attention vector.
- Final prediction through shared classifier.

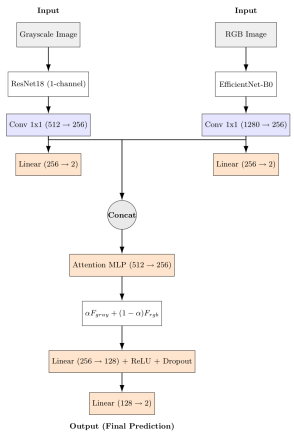

Figure 1: AMR-CD Model

Training Details:

- Loss = Main + 0.3 × (Gray Aux + RGB Aux)
- Optimizer: Adam, LR=1e-4, batch size = 32
- Early stopping: patience = 4

Best Metrics:

- **Train:** Acc 99.58%, Prec 99.67%, Rec 99.80%, F1 99.74%
- **Val:** Acc 96.21%, Prec 96.59%, Rec 98.42%, F1 97.50%
- **Train Loss:** 0.0444 **Val Loss:** 0.2415

Task B: Face Matching (Embedding)

Objective: Verify whether two face images belong to the same person (original vs distorted).

Methodology: The matching framework uses a ResNet-based backbone (ResNet18 or ResNet50) to extract deep spatial features from each image. Distorted images are transformed using a dedicated shallow CNN branch. Both original and distorted feature maps are passed through an attention mechanism that learns to assign importance to each.

The outputs are globally pooled and mapped into a 256-dimensional embedding space. During inference, the similarity between embeddings is computed using cosine similarity. This contrastive setup allows the network to learn identity-preserving representations even under noise or distortion.

Architecture:

- ResNet18 or ResNet50 backbone for feature extraction.
- Embeddings projected to 256-D space.
- Distorted view passes through a shallow transform.
- Attention-based fusion of original and distorted features.
- Cosine similarity used for inference.

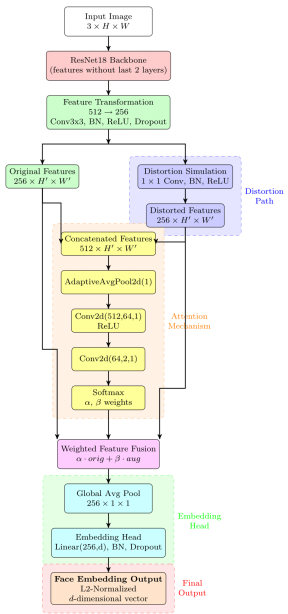

Figure 2: Face Matching Model

Training:

- Dataset of positive (same identity) and negative (different) pairs.
- Optimizer: Adam, batch size = 32
- Contrastive-style training

Best Validation Metrics:

- **Distance-Based:** Acc 97.54%, Prec 98.41%, Rec 98.78%, F1 98.60%
- **Similarity-Based:** Acc 96.45%, Prec 96.22%, Rec 99.86%, F1 98.01%
- **Train F1:** Dist 95.49%, Sim 90.80%

Summary: *Both systems leverage attention and multi-view representations to handle visual variations in real-world face recognition tasks, achieving high precision and recall across detection and matching scenarios.*