ENSEMBLE LEARNING USING TRANSFORMERS AND CONVOLUTIONAL NETWORKS FOR MASKED FACE RECOGNITION

DINH TRONG HAU

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam

What ?

A robust system for masked face recognition using a multi-model deep learning approach.

- An ensemble framework combining two CNNs and two Transformer models.
- Majority voting determines the final identity.
- Achieves **92.01%** accuracy on a custom masked LFW dataset.

Why ?

Face masks, a necessity since the COVID-19 pandemic, pose a major challenge to face recognition systems.

- Standard systems fail significantly on masked faces.
- NIST reported error rates of 5-50%.
- Masks occlude key features like the nose, mouth, and chin.

Overview

The system identifies individuals from masked face images by combining predictions from four distinct models (two FaceNet-based, two Vision Transformer-based) via majority voting for a robust final decision.

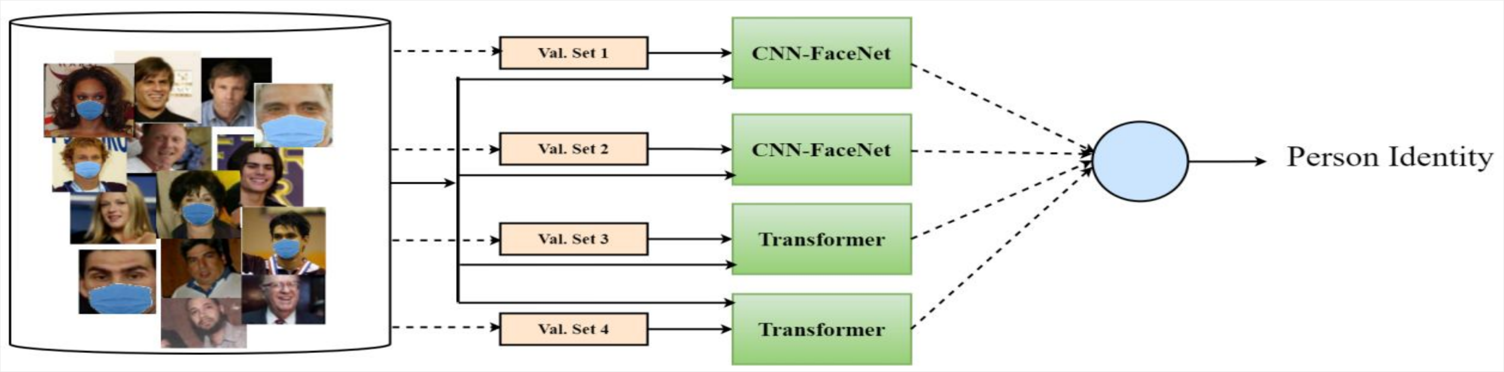


Fig. 1: The framework of the proposed ensemble learning model.

Description

Model Architectures

The approach integrates diverse model architectures to capture different feature types.

- **CNN Models**: Fine-tuned pre-trained models, with FaceNet showing superior performance due to its training on face-specific datasets.
- **Transformer Model**: An encoder-only architecture processes image patches, using self-attention to capture global contextual features.

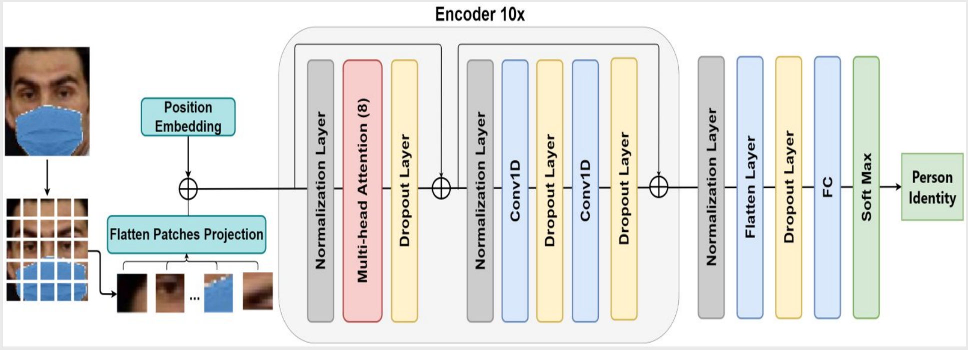


Fig. 2: The architecture of the Transformer model.

Dataset

The LFW dataset was used with synthetic masks via the MaskTheFace model, creating 26,466 images across 5,749 subjects. Data augmentation techniques were applied to address imbalance.

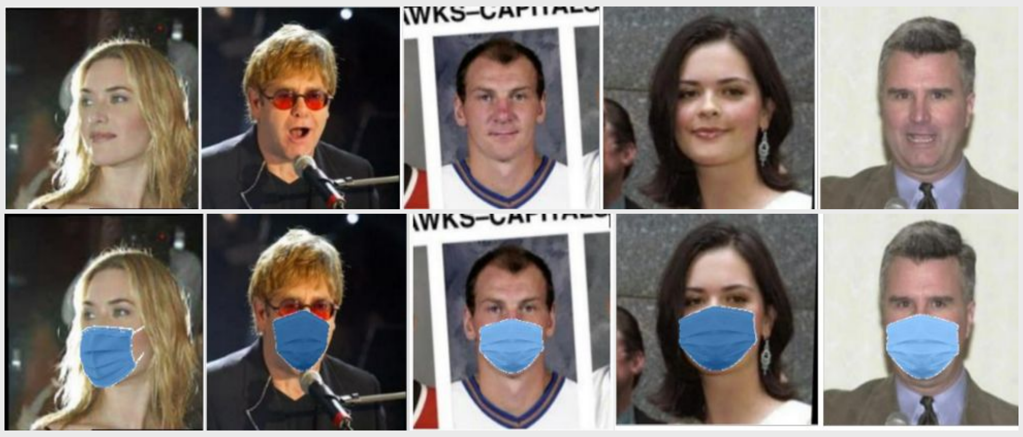
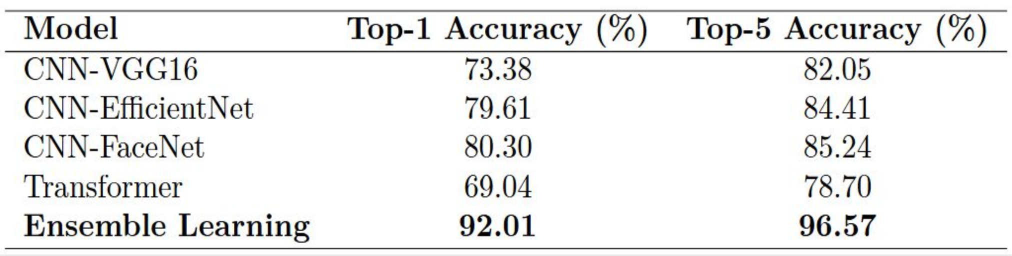


Fig. 3: Original (top) and synthetically masked (bottom) faces.

Results

Ensemble learning achieves the highest accuracy (**92.01%**), outperforming all individual models.



Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)
CNN-VGG16	73.38	82.05
CNN-EfficientNet	79.61	84.41
CNN-FaceNet	80.30	85.24
Transformer	69.04	78.70
Ensemble Learning	**92.01**	**96.57**

Conclusion

- Combining CNN and Transformer improves robustness under masked conditions
- Ensemble learning is effective for real-world face recognition

NII **Dinh Trong Hau –** University of Information Technology, Ho Chi Minh City, Vietnam
TEL : 0355752643 **Email : tronghau112@gmail.com**