

# Co-compressing and Unifying Deep CNN Models for Efficient Human Face and Speaker Recognition

Timmy S.T. Wan<sup>1,2</sup>, Jia-Hong Lee<sup>1,2</sup>, Yi-Ming Chan<sup>1,2</sup>, and Chu-Song Chen<sup>1,2</sup>

<sup>1</sup>Institute of Information Science, Academia Sinica, Taipei, Taiwan,

Email: Ftimmywan, honghenry.lee, yiming, songc@iis.sinica.edu.tw

<sup>2</sup>MOST Joint Research Center for AI Technology and All Vista Healthcare

## Abstract

*Deep CNN models have become state-of-the-art techniques in many application, e.g., face recognition, speaker recognition, and image classification. Although many studies address on speedup or compression of individual models, very few studies focus on co-compressing and unifying models from different modalities. In this work, to joint and compress face and speaker recognition models, a shared-codebook approach is adopted to reduce the redundancy of the combined model. Despite the modality of the inputs of these two CNN models are quite different, the shared codebook can support two CNN models of sound and image for speaker and face recognition. Experiments show the promising results of unified and co-compressing heterogeneous models for efficient inference.*

## 1. Introduction

Face recognition (FR) and speaker recognition (SR) are both important modules for applications such as access control, human/robotic interaction, and multimedia systems. Deep learning techniques have been shown promising to improve the FR and SR performance in recent years. However, most deep learning models are developed for either FR or SR tasks. In this paper, we present a deep convolutional neural network (CNN) approach that can jointly perform FR and SR in a single neural-network model. Besides, with our approach, both FR and SR tasks can be realized in one compressed neural network, and thus the storage and execution time required for the multimodal inference can be reduced.

To train a multi-task model in deep learning, a typical approach is to construct a CNN architecture with multi-task outputs in the final classification layer at first and then train this model with the union of training data from all tasks. However, such an approach requires a tedious trial-and-

error procedure because the architecture chosen could be inappropriate in the beginning and multiple re-training processes are needed in every trial. Even if neural network architecture search (NAS) [29] techniques can find appropriate architectures, it still requires taking a long time and consuming a lot of computational resources to generate satisfiable multi-task models.

One possible approach is to leverage on existing deep CNN models already trained for an individual FR or SR task. For example, with fast-growing deep CNNs, FR has gotten great performance improvement nowadays. Schroff et al. [20] propose triplet loss and design a network structure for FR. Liu et al. [12] introduce additive angular margin loss and redesign the network structure for FR. These models are publicly available with high accuracy on large face datasets (such as the LFW dataset [10]). Leveraging on well-performed single-task models has the advantage of preserving the recognition accuracy of one modality more easily. Nevertheless, it is still non-trivial to merge two well-performed models without compromising the performance of the individual task.

In this paper, we present an approach that leverages on well-trained individual models of both FR and SR. We then merge the two single-modality models into a unified one while keeping the compactness of the merged model for multi-modal inference. When deep CNN models are learned, they often have much redundancy in the network weights, and thus the models can be compressed before deployment for inference [5, 27, 7]. In our work, we not only merge the two models but also “co-compress” them into a single model, and thus the resulted model is compact (with a smaller model size) and more suitable for efficient multi-modal inference.

Our approach follows the principle of NeuralMerger [2]. In this technique, two merged layers share a common codebook consisting of a set of codewords; the individual-task













