大眾化 AI-從應用、演算法至硬體設計

Democratized AI – From Application, Algorithms, to Hardware Design

Y.-W. Peter Hong*[1,2], Yeong-Luh Ueng[2] , Yi-Wen Liu[2] , Chi-Chun Lee[2]
Meng-Fan Chang[2], Kea-Tiong Tang[2], Min Sun[2], Chih-Cheng Hsieh[2],
Ren-Shuo Liu[2] and Chung-Chuan Lo[3]

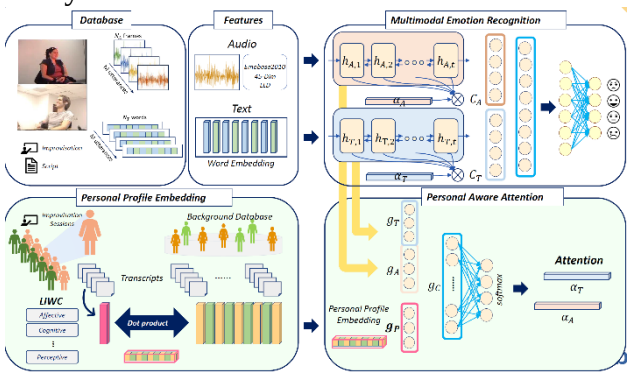
[1]Institute of Communications Engineering, National Tsing Hua University, Hsinchu, Taiwan
[2]Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
[3]Institute of Systems Neuroscience, National Tsing Hua University, Hsinchu, Taiwan
E-mail: ywhong@ee.nthu.edu.tw

The proliferation of data and the maturity of cloud computing technology has triggered the initial growth of artificial intelligence (AI) in many applications, such as autonomous driving, natural language processing, robotics, gaming, etc. To further enhance its societal impact, it is necessary to make AI technology approachable with intelligent services adopted at scale across life contexts is the essence of democratized-AI. In this project, we investigate the technology required for democratizing AI, from the application to the algorithm and hardware.

Subproject I: Realizing AI-Enabled Intelligent Services

The democratization of AI relies on the development of two personalized applications, namely, **affective computing** and **privacy-aware services**. In the former case, we developed feature learning techniques based on human emotion and user reaction, and utilized them to build human-centric intelligent services. In the latter case, we focused on facial and fingerprint recognition algorithms for financial services, and voice analysis for scam call identification.

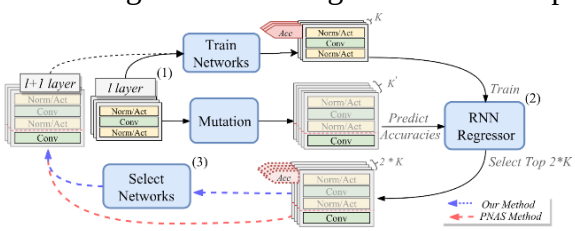
Highlight Research: In affective computing, we published a series of algorithms (at INTERSPEECH and ACII) that integrates multimodal behavior (speech, text and visual) modalities and also incorporates personalization to capture the heterogeneity across different users. The core idea is to utilize cross database transfer learning in the retrieval of personal attributes from external corpus which is integrated in learning the multimodal emotion recognition network jointly with personalized attention or graph memory networks. A demo prototype was provided at 2019 MOST Futuretech Expo and received the Futuristic Breakthrough Technology Award. The proposed system architecture is illustrated below.

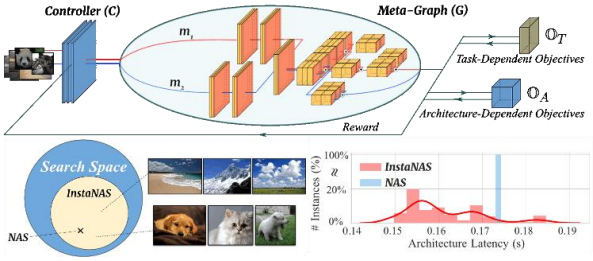


Subproject II: Efficient AI Algorithms for Edge Intelligence

Edge intelligence relies on effective learning directly at the edge computing or end-user devices, preferably without sharing of personal data with the cloud server, and is key to many democratized AI applications. This requires algorithms for distributed learning among devices and also device-aware machine learning architectures to enable efficient on-device computing.

Highlight Research: To enable device-aware machine learning, we developed several works (e.g., [1] and [2]) to support adaptive neural architecture search that adapts to the computational capabilities of the devices. At ECCV 2018 [1], we presented DPP-Net, namely, a device-aware progressive search algorithm for Pareto-optimal neural architectures. This research designed the search space to include the most common light-weighted modules, and incorporated multiple objectives including accuracy, speed, model size, and inference FLOPs in the design. The resulting architecture outperformed state-of-the-art CondenseNet in 2018.





In addition, at AAAI 2020 [2], we further proposed InstaNAS, an instance-aware neural architecture search, that further sped up the inference speed by using a controller to select the neural architecture for each input image instance.

Subproject III: Advanced Hardware Design for Democratized-AI

Deep neural networks have revitalized modern interest in AI, but require large energy consumption that may be prohibitive for end-user devices. To democratize AI, it is necessary to develop energy-efficient hardware architectures that can enable such computations locally.

Highlight Research: We developed computing-in-memory (CIM) hardware architectures to enable neural network computation directly among the memory array to reduce the energy consumption and delay required for transporting data between the memory and the processor. In addition, we will also develop an AI image processing chip that performs computing-in-sensor (CIS) to reduce the energy and bandwidth required for communicating data to backend processors. **Several key achievements include:** (a) SRAM- and ReRAM-based CIM and also CIS publications at ISSCC; (b) Student Design Contest Award at ASSCC; (c) presented a near-memory AI accelerator architecture at ESSCIRC; (d) a Student Research Competition Award at the MICRO conference; (e) finally, our highlight research outcomes were also presented at 2019 MOST Future Tech Expo, and received the Futuristic Breakthrough Technology Award.