Artificial intelligence on mobile devices

Sidan He
Strategy Dept., Huawei HiSilicon, Shenzhen, China

The first wave of smartphones with deep neural network engines has enabled basic artificial intelligence functionalities like face recognition, scene classification and voice recognition. With competing software frameworks and fast-evolving deep neural network algorithms, smartphone vendors face the tremendous challenge to select the right hardware accelerator architecture for both software compatibility and energy efficiency. Recent slowdown of Moore's Law adds even more constrains on the hardware accelerator architecture to support future generation deep neural networks. At the same time, most of the current mobile artificial intelligence (AI) applications are executing inference of pre-trained deep neural networks due to the requirement of huge amount of training data and computation cycles. There are significant barriers to adapt to the needs of individual smartphone users through continuous learning.

In this talk I will share the experience of integrating neural network processing unit with smartphone application processors and discuss the challenges and opportunities facing the next generation of mobile AI applications.