A Deep FM Acceleration Algorithm based on Baidu Cloud's Heterogeneous Multicore Platform

Southeast University

A deep FM acceleration algorithm was designed based on Baidu Cloud's heterogeneous multicore platform that contains both Intel's E5-2650 v4 CPU and Xilinx's XCKU115 FPGA.

A brief introduction of our work is as follows.

• Firstly, the model was trained using Baidu's open-source-deep-learning framework called PaddlePaddle.

• In order to better understand how the model works, the whole system were rebuilt using C++, preprocessing procedure, parameter extraction and model reconstruction included.

• What's more, computing tasks were split according to the reconstructed system. Only fully connected layers were accelerated by implementing them using Baidu's Polaris API that helps accelerate programs with FPGA.

• DMA bandwidth was utilized as full as possible, while the result, Polaris API used, was assured to be correct.

Last but not least, the performance of our algorithm is about 4 times better than the original version that runs at a PC equipped with Intel's i5-3470 CPU, with the time consumed by DMA not taken into account. The performance of our algorithm is more than 2 times of the original version, DMA time included.

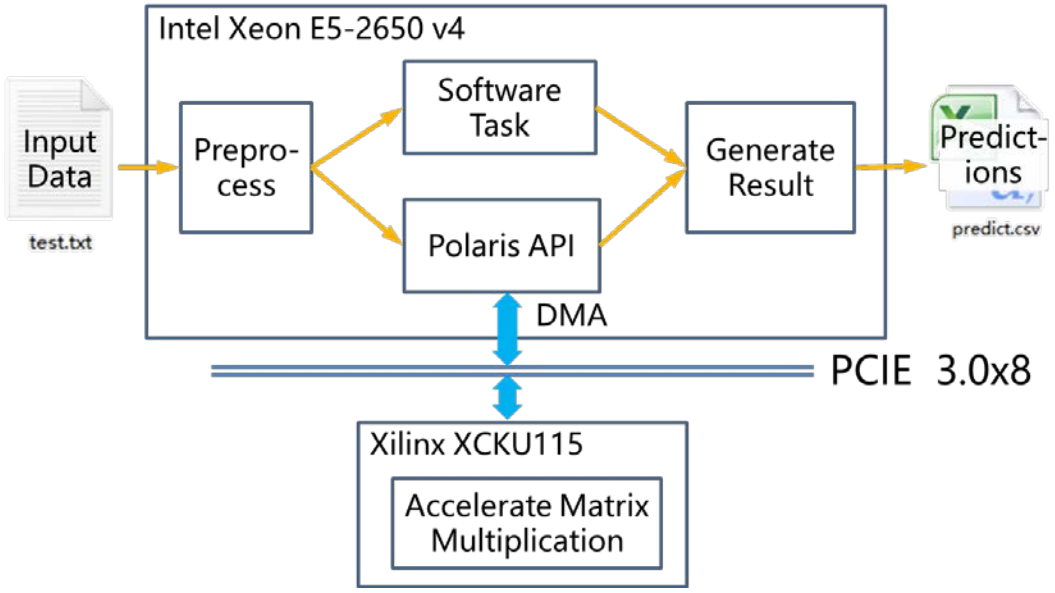

Figure 1. System Overview

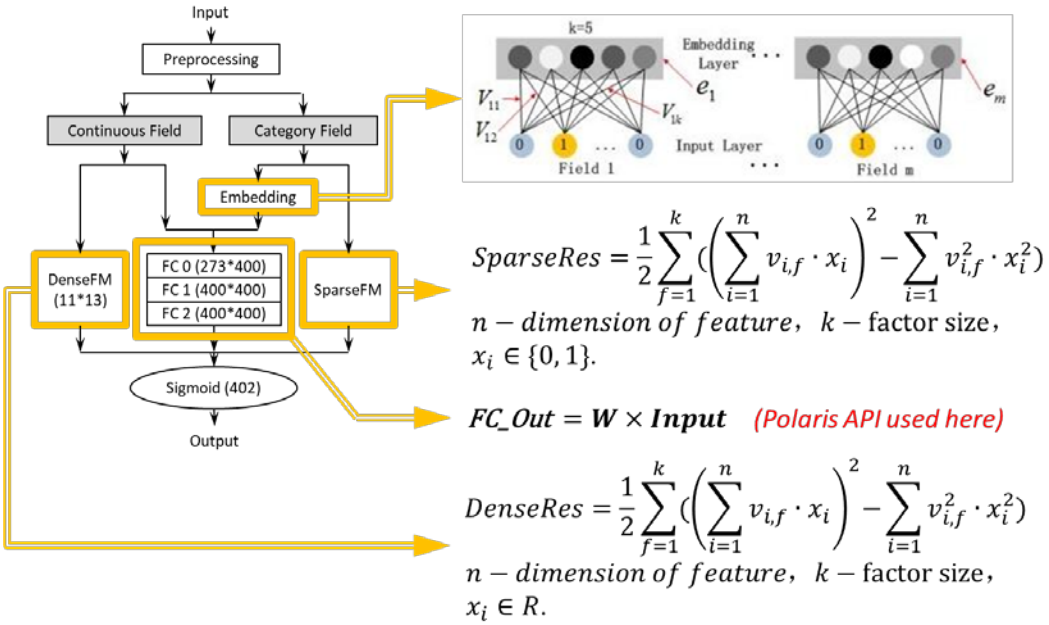

Figure 2. Algorithm Framework