1. Introduction

speech recognition model based on neural networks perform well;

neural networks are being widely used and do not have a high computing efficiency on CPU;

this article:

This paper modifies a CNN-based speech recognition algorithm into a binary weight neural network model where weight value is +1 or -1. Also, this paper uses Matlab quantify functions to turn float-type feature data into fix-point data level by level, whose loss of accuracy will be covered by the performance of fix-point computing on FPGA platforms. Last but not least, this paper designs a multi-PE BWN accelerator on FPGA and achieve over 300x accelerating ration compared with Matlab codes on i7-8700K.

2. Related Work / Background

BNN

Quantification

3. Speech Recognition Model

3.1 Model Architecture and Weight Binarization

fig1: nn architecture

fig2: weight binarization process

3.2 Feature Data Quantification

4. Accelerator Architecture

4.1 Parameter Storage

parameter's size

batch normalization parameter

shared storage among PEs

reduce memory cost

table1: parameter detail

fig3: parameter sharing structure

4.2 bitwidth expandsion

batch normalization needs extra bitwidth

accumulating result's value can be huge while feature's value are relatively tiny

bitwidth expandsion meets all needs

4.3 pipeline / level by level / vector mac unit

do not need to store middle data in ddr

pipeline help to rise freq and have balanced pipeline: conv1 & conv2

vector mac unit is basic compute unit

fig4: how to balance conv1 & conv2

fig5: hardware architecture

5. Experiment

5.1 Quantified model's performance

Table2: different acc under different bitwidth

8700k + matlab 1 core running time

8700k + matlab parallel lib running time

9700k muliti-node running time

Table3: perf on CPU

5.2 accelerator's performance

acc

1PE's

throughput rate, running time, utilization, power

2PE's

throughput rate, running time, utilization, power

Table4: perf on accelrator