### Preliminary Project Planning Form

Due day: 2:00pm 11/09/2022

One per team. Submit to the course website on Moodle.

(Grades of this form is part of final project. **Please answer with cautions!**)

TEAM Name: 男童俱樂部

(If you want to change your team name, please also specify your old team name.)

Team Leader Name: 黃昱澄

Members Name: 俞杉麒、黃冠予、王昱承、賴致文、陳奕萍

|  |  |
| --- | --- |
| **Target Application for ASPU or Duo-Core** | Index-based Kernel Transformation Architecture for Sparse CNNs |
| Please describe your target application with short motivation and key components that will be related to your application processor | [Overview]  This work proposes an index-based kernel transformation implemented on networks whose weights are quantized into an index-based representation while keeping the fixed-point precision. The proposed algorithm can eliminate redundant operations by extracting the common index patterns from different kernels to perform the identical operation only once. A specifically designed hardware is implemented to perform the convolution and rebuild the correct results from extracted patterns in parallel.  [key parts with block diagram]  The Conv-Rebuild unit is the most important component because it is essential for convolution operations—the result of which is also the basis of rebuilding. The Conv-Rebuild Unit receives and computes the data and then sends the result to the buffers after completing the operation; they are under the command of the control unit to perform convolution and rebuilding. The Conv-Rebuild unit performs rebuild and convolution work in parallel. |
| *Please describe your application with targeting specification and how the application processor will work with CPU & memory on both hardware and software sides.* | [Overview] We designed a hardware architecture to perform sparse convolution with transformed kernels that include modified and common kernels. This hardware is designed to do a convolutional computation of different kernels simultaneously. It skips computation of weights with zero value and rebuilds the correct results.  [Specification]  // This part shall have reasonable level of complexity and justification for the performance specification, i.e., why that speed values and why that power consumption.  Datasets: Cifar-10, Cifar-100  Models: NIN, VGG9, VGG16  Two hyperparameters in Clip-Q:  p (pruning rate)  : 0 to 1  b (number of bits represented by each weight index)  : FULL precision, 16, 8, 4, 2  Local memory: 3KB  Clock frequency: 500 MHz  [Work with CPU, illustrated with figures if possible]    IM  IM  CPU  DM      After training and kernel transformation, the index-based sparse convolution architecture can obtain all information needed for the convolution. Note that CPU is responsible for collecting the data from Software flow and outputs the corresponding signal and data to the index-based sparse convolution architecture. Eventually, the output will then be sent back to CPU for further use. |
| *Please provide task assignment for every member. There shall be at least one person dedicate to verification of IPs.* | 黃昱澄：  俞杉麒：  黃冠予：  王昱承：  賴致文：  陳奕萍： |
| *Please provide project time schedule by providing all members milestones for their own tasks using a chart. Note that please plan by week and DO check the dates for demo and final presentation.*  **Be aware of the schedule posted on the course** |  |
| **website.** | https://hackmd.io/@XiaXia/LBCNN |